UNIVERSITY OF SOUTHAMPTON

FACULTY OF MEDICINE

Clinical and Experimental Sciences

BLOOD-BASED BIOMARKERS IN PSYCHIATRIC DISEASES

by

Michael S. Breen

Thesis for the degree of Doctor of Philosophy

Supervisory Team: Dr. Christopher H. Woelk, Prof. David Baldwin & Prof. Andrew Collins

May 2016

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF MEDICINE
Clinical and Experimental Sciences

Doctor of Philosophy

BLOOD-BASED BIOMARKERS IN PSYCHIATRIC DISEASES

by Michael S. Breen

Identification of blood-based biomarkers for psychiatric disease risk and development has emerged as an important area of translational research in medicine, offering a means to supplement or replace current interview-based methods for psychiatric diagnosis. The aim of this thesis is to assess the utility of genome-wide blood transcriptome profiling for the prediction, diagnosis and treatment of patients with psychiatric diseases. Some parts of this work are of a more methodological nature and geared towards the discovery of blood-based biomarkers and gene networks, while others consider mechanistic and translational implications. Overall, this work contributes to understanding the pathophysiology of major psychiatric diseases and to the development of new biomarkers and treatments.

Part I discusses the current transition from interview-based psychiatric diagnostics towards genomic-based interventions (Chapter 1) prior to introducing experimental methodologies (Chapter 2) and statistical approaches (Chapter 3) that may provide favorable translational avenues for blood biomarker discovery in psychiatry.

Part II contains four investigations (summarized in Chapter 4) that apply genome-wide transcriptome profiling of patient blood samples in pursuit of blood-based biomarkers and gene networks implicated in posttraumatic stress disorder (Chapter 5), acute psychological stress (Chapter 6), methamphetamine-associated psychosis (Chapter 7) and treatment response in bipolar disorder (Chapter 8).

Part III proposes a set of rules or postulates for accelerating the identification of reliable and accurate blood-based biomarkers in patients with psychiatric diseases (Chapter 9).

Contents

Abstract ... i

Contents ... iii

List of Figures ... xii

List of Tables ... xv

Declaration of Authorship ... xvii

Lists of Publications ... xviii

Acknowledgements ... xix

List of Abbreviations ... xxi

I Introduction 1

1 Biomarkers in Psychiatry 3

1.1. The Global Burden of Psychiatric Disorders ... 3

1.2. The ICD and DSM Classification Systems ... 4

1.2.1. The Purpose of Classification ... 4

1.2.2. The Difficulties of Classification ... 5

1.2.3. The Evolution of Classification ... 5

1.3. Research Domain Criteria (RDoC): A New Paradigm for Psychiatry ... 6

1.4. Biological Marker Strategy ... 7

1.5. Advances in Genome Technology ... 8

1.6. The Value of Transcriptome Biomarkers ... 11

1.7 .Trade-offs Between Brain and Blood Biomarkers in Psychiatry ... 11

1.7.1. Post-Mortem Brain Gene Expression ... 12

1.7.2. Peripheral Blood Gene Expression ... 13

1.7.3. Correspondence Between Brain and Blood Gene Expression ... 14

1.8. Measuring the Transcriptome ... 15

1.8.1. Microarray Technology ... 15

1.8.2. RNA-Sequencing Technology ... 16

1.9. Shapes and Forms of Transcriptome Biomarkers 16
1.10. Summary........ 17

2 Experimental Methodologies in Blood Transcriptomics 19

2.1. Experimental Design........ 21
2.2. Transcriptome RNA Diversity: Implications for Biomarker Discovery 22
2.3. RNA Isolation from Patient Blood Samples 23
2.3.1. PAXgene RNA Extraction and Globin mRNA Depletion 23
2.3.2. LeukoLOCK^{TM} RNA Extraction........ 24
2.3.3. RNA Quality Check 25
2.3.4. RNA Target Enrichment 26
2.4. Steps Required for a Microarray Experiment........ 27
2.4.1. Microarray Platforms 29
2.4.2. Strengths and Weaknesses of Microarray 29
2.5. Steps Required for an RNA-Seq Experiment........ 29
2.5.1. RNA-Seq Library Preparation........ 30
2.5.2. RNA-Seq Cluster Generation 30
2.5.3. Sequencing 31
2.6. RNA-Seq Optimization........ 32
2.6.1. Singe-End and Paired-End Sequencing........ 32
2.6.2. Sequence Depth and Coverage 33
2.6.3. Multiplexing 33
2.6.4. Strengths and Weaknesses of RNA-Seq 33
2.7. Cross-Platform Variability 34
2.7.1. Technical Assessment of Microarrays 34
2.7.2. Technical Assessment of RNA-Sequencing........ 35
2.8. Summary........ 35

3 Statistical Methodologies for Blood Biomarker Discovery 37

3.1. Standard Guidelines and Best Practices for Expression Data........ 37

3.2. Microarray Data Pre-Processing 39
3.2.1. Microarray Background Adjustment 39
3.2.2. Microarray Summarization 40
3.2.3. Microarray Normalization 40
3.2.4. Microarray Non-Specific Filtering 41
3.3. RNA-Seq Data Pre-Processing 42
3.3.1. Quality Check of Raw RNA-Seq Reads 43
3.3.2. RNA-Seq Read Trimming and Filtering Low Quality Reads 43
3.3.3. RNA-Seq Read Alignment to Reference Genome 44
3.3.4. RNA-Seq Counting Aligned (Mapped) Reads 45
3.3.5. RNA-Seq Non-Specific Filtering 46
3.3.6. RNA-Seq Normalization 46
3.4. Microarray and RNA-Seq Quality Control 48
3.5. Data Analysis: Transcriptome Exploration from Low- to High-Order 49
3.6. Supervised Machine-Learning and Classifier Construction 50
3.6.1. Statistical and Translational Implications for Gene Classifiers 53
3.7. Differential Gene Expression Analysis 54
3.7.1. Moderated *t* test 55
3.7.2. Permutation-based Testing 56
3.7.3. Functional Annotation and Enrichment Analyses 57
3.7.3.1. Semantic Similarity of Ontology Terms 58
3.7.4. Protein-Interaction Networks 59
3.8. Weighted Gene Co-expression Network Expression Analysis 59
3.8.1.Constructing a Weighted Gene Co-expression Network 61
3.8.2. Identification of Sub-Networks from the Global Network 62
3.8.3. Module Preservation and Module Differential Expression Analyses 64
3.8.4. Integration of Multi-Modal Data 66
3.8.5. Hub Genes 67
3.8.6. Statistical and Translational Implications for Gene Networks 68
3.9. Summary 69

4 Aims 71

4.1. Part II - Application of Blood Transcriptomics in Psychiatric Diseases 71

4.1.1. Chapter 5: Candidate Blood Biomarkers and Gene Networks of Posttraumatic Stress Disorder 71

4.1.2. Chapter 6: Immediate Molecular and Cellular Response to Acute Psychological Stress 72

4.1.3. Chapter 7: Candidate Blood Biomarkers and Gene Networks of Methamphetamine-Associated Psychosis 72

4.1.4. Chapter 8: Candidate Lithium Responsive Genes and Gene Networks of in Bipolar Disorder Lymphoblastoid Cell Lines 73

4.2. Part III – Moving Biomarkers Forward in Psychiatry 73

4.2.1. Chapter 9: Moving Biomarkers Forward in Psychiatry 73

II Application of Blood Transcriptomics in Psychiatric Diseases 75

5 Candidate Blood Biomarkers and Gene Networks of Posttraumatic Stress Disorder 77

5.1. Background... 77

5.2. Materials and Methods 79

5.2.1. Subject Selection and PTSD Diagnosis 79

5.2.2. RNA Isolation and Generation of Gene Expression 80

5.2.3. Data Pre-Processing 80

5.2.4. Weighted Gene Co-expression Network Analysis 83

5.2.4.1. Individual Network Construction and Module Preservation 83

5.2.4.2. Combined Network Construction and Module Differential Analysis 84

5.2.5. Functional and Cellular Enrichment Analyses 85

5.2.6. Differential Gene Expression Analyses 85

5.2.7. Supervised Machine-Learning Classification 85

5.2.8. Batch (Technology) Correction 86

5.3. Results 87

5.3.1. No Large Differences in Module Preservation 87

5.3.2. Differential Module Expression Post-Deployment in Dataset 1 89

5.3.3. Differential Module Expression Pre-Deployment in Dataset1 94

5.3.4. External Validation of Post-Deployment Modules in Dataset 2 95

5.3.5. External Validation of Pre-Deployment Modules in Dataset 2 96

5.3.6. Cross-Sectional and Longitudinal Differential Gene Expression 99

5.3.7. Putative Diagnostic PTSD Gene Expression Classifier 100

5.3.8. Putative Predictive PTSD Gene Expression Classifier 101

5.4. Discussion 102

5.4.1. Gene Networks Specific for Innate Immunity in PTSD 102

5.4.2. Predictive and Diagnostic Gene Expression Classifiers 103

5.4.3. Interpreting Blood-based Innate Immune Signatures in PTSD 103

5.4.4. The Hemostatic Response to PTSD Development 105

5.4.5. Strengths and Limitations 106

5.4.6. Concluding Remarks 106

6 Immediate Molecular and Cellular Response to Acute Psychological Stress 109

6.1. Background 109

6.2. Materials and Methods 110

6.2.1. Ethical Approval 110

6.2.2. Subject Selection and Sample Collection Schedule 111

6.2.3. RNA Isolation and Microarray Hybridization 111

6.2.4. Data Pre-Processing 112

6.2.5. Differential Gene Expression Analyses 112

6.2.6. Real-time Quantitative Reverse Transcription (RT-qPCR) 113

6.2.7. Weighted Gene Co-Expression Network Analysis 114

6.2.8. Functional and Cellular Enrichment Analyses 115

6.2.9. Protein-Interaction Networks 115

6.2.10. Flow-Cytometry Analysis 115

6.3. Results 116

6.3.1. Acute Stress Induced Physiological Responses 118

6.3.2. Candidate Acute Stress Responsive Genes 118

6.3.3. RT-qPCR Validation of Selective NK Cytotoxicity Response 120

6.3.4. Candidate Acute Stress Responsive Genes Underlying Cell Subset Fluctuations 125

6.3.5. Identification and Annotation of Gene Modules 127

6.3.6. Correlating Gene Modules with Physiological Measurements 128

6.3.7. Determining Gender-specific Immune Responses 130

6.3.8. Acute Stress Responsive Leukocytes 132

6.4. Discussion 133

6.4.1. Selective NK Cell Stress Susceptibility Genes 134

6.4.2. Putative Roles of IL-12 Signalling and TLRs 135

6.4.3. Gender-Specificity of the Acute Stress Response and Implications for Stress-Induced Pathologies 136

6.4.4. Putative Biomarkers for Discriminating Anxiety-based Stress from Neuropsychiatric and Central Nervous System Disorders 137

6.4.5. Hypoxia does not Contribute to Gene Expression Differences 138

6.4.6. Strengths and Limitations 139

6.4.7. Concluding Remarks 139

7 Candidate Blood Biomarkers and Gene Networks of Methamphetamine-Associated Psychosis 141

7.1. Background 141

7.2. Materials and Methods 143

7.2.1. Subject Selection 143

7.2.2. sMRI Acquisition and Image Processing 144

7.2.3. RNA Isolation and RNA-Seq Library Preparation 144

7.2.4. Read Trimming, Mapping and Quantification of Expression 145

7.2.5. Data Pre-Processing 145

7.2.6. Weighted Gene Co-expression Network Analyses 146

7.2.7. Differential Gene Expression Analyses 147

7.2.8. Functional and Cellular Enrichment Analyses 147

7.2.9. Supervised Machine-Learning Classification 147

7.2.10. Convergent Functional Genomic (CFG) Scoring 148

7.3. Results 150

7.3.1. Differential Analysis of Modules and Brain Volumes 153

7.3.2. Phenotypic Characterization of MAP Modules 154

7.3.3. Phenotypic Characterization of MA Modules 155

7.3.4. Putative Diagnostic Blood Biomarker for MAP 157

7.3.5. CFG Prioritization of Biomarkers 162

7.4. Discussion 165

7.4.1. Ubiquitin Proteasome System (UPS) Dysregulation 165

7.4.2. UPS and Circadian Clock Associations to sMRI Data 166

7.4.3. Putative MAP-related Gene Hunting Tools 167

7.4.4. Candidate Blood-based Diagnostic Biomarker of MAP 168

7.4.5. Candidate MA and METH Dependency Genes 169

7.4.6. Strengths and Limitations 170

7.4.7. Concluding Remarks 171

8 Candidate Lithium Responsive Genes and Gene Networks in Bipolar Disorder Lymphoblastoid Cell Lines 173

8.1. Background 173

8.2. Materials and Methods 175

8.2.1. Subject Selection 175

8.2.2. Cell Culture 176

8.2.3. RNA Isolation and RNA-Seq Library Preparation 176

8.2.4. Read Trimming, Mapping and Quantification of Expression 177

8.2.5. Data Pre-Processing 177

8.2.6. Weighted Gene Co-expression Network Analyses 178

8.2.7. Differential Gene Expression Analyses 178

8.2.8. Functional Annotation and Drug Gene-Set Testing 179

8.2.9. Protein Interaction Networks 180

8.3. Results 180

8.3.1. Differential Analysis of Module Eigengene Values and Drug-Gene Set Enrichment 182

8.3.2. Longitudinal Contrast Different Gene Expression Analysis 186

8.3.3. Cross-Sectional Differential Gene Expression Analyses 188

8.3.4. Prioritization of Lithium Responsive Genes by Cross-Tabulating Independent Transcriptome-based Evidence 189

8.4. Discussion 191

8.4.1. Gene Networks Induced by Lithium Treatment 191

8.4.2. Gene Networks Repressed by Lithium Treatment 192

8.4.3. Lithium and Clonidine: A Shared Mechanism of Action 193

8.4.4. Heterogeneity in Lithium Response 193

8.4.5. Strengths and Limitations 194

8.4.6. Concluding Remarks 196

III Moving Biomarkers Forward in Psychiatry 199

9 Moving Biomarkers Forward in Psychiatry 201

9.1. Working on the Right Problem 202

9.2. The Right Measure of Statistical Significance and Effect Size 203

9.3. Making the Right Comparisons 204

9.4. Using the Right Approach 204

9.4.1. Experimental Designs 204

9.4.2. Alternative Next Generation Sequencing Technologies 205

9.4.3. Neuroimaging 207

9.4.4. Integrating Panels of Bio-signatures 207

9.4.5. Systems-level Diversity 208

9.4.6. Statistical Design 208

9.5. Demonstrating Reproducibility in the Right Way 209

9.6. The Right Use of Biomarkers 210
9.7. Summary 211

Appendices 213

References 215

Appendix A Supplementary Data 245

A.1. Supplementary Data for Chapter 5 245
A.2. Supplementary Data for Chapter 6 254
A.3. Supplementary Data for Chapter 7 257
A.4. Supplementary Data for Chapter 8 263

Appendix B Publications 269

List of Figures

2.1. An experimental work-flow for generating gene expression data 20
2.2. Transcriptome RNA Diversity.. 22
2.3. Generation of gene expression by microarray technology.. 28
2.4. A pipeline depicting a typical RNA-Seq library preparation 30
2.5. Illumina and Ion Torrent sequencing technologies 32

3.1. A statistical work-flow for blood-based biomarker discovery 38
3.2. Affymetrix perfect match and mismatch probes.. 40
3.3. Computational pipeline for RNA-Seq data pre-processing 42
3.4. Quality control statistical methods and procedures 49
3.5. Separating and maximum margin hyperplanes based on support vector machines 52
3.6. Cross-validating prediction accuracies based on gene expression 53
3.7. Two common approaches for gene-set enrichment analysis 58
3.8. Basic principles of weighted gene co-expression network analysis 61
3.9. Balancing scale-free topology and ß-values for gene network construction 62
3.10. Utility of hierarchical cluster tree (dendrogram) 63
3.11. Statistical applications for indicating module significance... 65
3.12. Using module eigengenes for differential expression analysis 66
3.13. Integrating multi-modal data with gene co-expression modules................................. 67

5.1. Hierarchical cluster tree of the combined post-deployment network of PTSD cases and controls comprising 10,184 genes 89
5.2. Post-deployment module significance and differential expression in Dataset 1 90
5.3. Module annotation of differentially regulated modules in Dataset 1 93
5.4. Pre-deployment module significance and differential expression in Dataset 1 94
5.5. Post-deployment module significance and differential expression in Dataset 2 95
5.6. Pre-deployment module significance and differential expression in Dataset 2 96
5.7. Module annotation of differentially regulated modules in Dataset 2 97
5.8. Cross-referencing pre- and post-deployment modules in Dataset 1 and 2.................... 98

5.9. Cross-sectional differential gene expression results at pre- and the post-deployment as well as longitudinal differences between time-points 99
5.10. Construction of a Diagnostic PTSD gene expression classifier post-deployment...... 100
5.11. Construction of a Prognostic PTSD gene expression classifier pre-deployment 101

6.1. The skydive paradigm and observed physiological responses to acute stress 117
6.2. Comparative differential gene expression analyses of pre-boarding, post-landing and one-hour post-landing relative to baseline as well as heatmap and functional annotation of post-landing genes 119
6.3. Protein interaction network analysis of all differentially expressed gene 120
6.4. RT-qPCR validation of differentially expressed genes post-landing 122
6.5. Cell-type enrichment analyses at post-landing before & after cell type correction 126
6.6. Hierarchical cluster tree of gene modules across all measured time-points 127
6.7. Integration of stress-related gene co-expression modules with multi-modal data 129
6.8. Visualization of gender-specific stress responses and gene modules......................... 131
6.9. Assessment of peripheral blood leukocyte cell linage with flow cytometry 133

7.1. Multi-step analytical work-flow for identifying MAP biomarkers & gene networks 152
7.2. MAP-specific dysregulated gene networks 153
7.3. MA-specific dysregulated gene networks 154
7.4. Integration of MAP gene co-expression modules with multi-modal data 156
7.5. Differential gene expression analysis between MAP, MA and healthy controls followed by functional annotation of candidate genes 158
7.6. Two gene expression classifiers for classifying between METH dependency and controls and then for classifying between MAP from MA.. 159
7.7. Convergent functional genomic (CFG) prioritization of candidate findings and gene-disease relationships.. 164

8.1. Multi-step analytical work-flow for identifying lithium mechanism of action and surrogate markers differing in their response to lithium treatment 181
8.2. Total number of differentially expressed genes between vehicle and lithium treatment performed in a group-wise fashion and then by pooling 182

8.3. Differential module expression analysis of candidate lithium responsive gene networks 183

8.4. Drug-gene set enrichment analysis and identification of drug-targets with a similar mechanism of action 184

8.5. Detailed view of clonidine-responsive genes similar to those observed with lithium treatment 185

8.6. GO sematic similarity and STRING analyses based on differentially expressed genes differing in their response to Li between responders and non-responders 187

8.7. Cross-sectional differential gene expression analyses prior-to and then following lithium treatment 188

List of Tables

1.1. Main Differences between ICD and DSM classification systems 6

1.2. Overview of definitions used in the field of biomarker research 9

1.3. Commonly used genome-wide technologies, platforms and applications 10

5.1. Clinical and socio-demographic variables for all U.S. Marines in dataset 1 82

5.2. Clinical and socio-demographic variables for all U.S. Marines in dataset 2 83

5.3. Module preservation statistics at pre- and post-deployment within dataset 1 88

5.3. Complete network characterization of post-deployment modules in dataset 1 92

6.1. Selective regulation of NK cell cytotoxic genes at post-landing 123

7.1. Recorded clinical characteristics from all subjects 151

7.2. Brain structural volumes (mm3) from all subjects 154

7.3. Top twenty-five informative features separating controls and METH dependents 160

7.4. Top twenty informative features separating MAP from MA subjects 161

8.1. Potential surrogate marker genes differing in expression between bipolar disorder lithium responders and non-responders 190

DECLARATION OF AUTHORSHIP

I, Michael Sean Breen, declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

Blood-based Biomarkers in Psychiatric Disease

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as detailed overleaf.

Signed:

Date: 6th May 2016

List of Publications

1. Breen MS, Maihofer A, Glatt S, Tylee D, Chandler S, Tsuang M et al. Gene networks specific for innate immunity define post-traumatic stress disorder. *Nat. Molecular Psychiatry* 2015; 20: 1538-1545.

2. Breen MS, Beliakova-Bethell N, Mujica-Parodi L, Carlson J, Ensign W, Woelk C et al. Acute psychological stress induces short-term variable immune response. *Brain, Behavior, and Immunity* 2015. doi:10.1016/j.bbi.2015.10.008.

3. Breen MS, Uhlmann A, Nday CM, Mitt M, Metsalpu A et al.. Candidate gene networks and blood biomarkers of methamphetamine-associated psychosis: an integrative RNA-sequencing report. *Nat. Translational Psychiatry* (ahead of print)

Acknowledgements

I would mainly like to acknowledge, with gratitude, my debt of thanks to Dr. Christopher H. Woelk. Without his valuable time, support and open collaborations, this work would not have been completed nor would I have been integrated into this exciting area of research. I would also like to thank Prof. David Baldwin for his time, support and sharing collaborations in Cape Town, SA. Of course, many thanks to all of our collaborators at the University of California San Diego and at the University of Cape Town, and to the patients who have taken part in the studies detailed herein.

List of Abbreviations

ADHD	Attention deficit hyperactivity disorder
BD	Bipolar disorder
CFG	Convergent functional genomic
CNS	Central nervous system
CNV	Copy number variant
cTEN	Cell type enrichment
CVD	Cardiovascular disease
DLDA	Diagonal linear discriminate analysis
DSigDB	Drug Signature Database
DSM	Diagnostic and Statistical Manual
FPKM	Fragments per kilobase of transcript per million mapped reads
GCRMA	GC-robust multichip average
GO	Gene-ontology
GS	Gene significance
GWAS	Genome-wide association study
HPA	Hypothalamus-pituitary and adrenal glands
ICD	International Statistical Classification of Disease
IFN	Interferon
IL	Interleukin
KEGG	Koyoto encyclopaedia of genes and genomes
kME	Intramodular module membership
lncRNA	Long non-coding RNA
LCLs	Lymphoblastoid cell Lines
Li	Lithium
LOOCV	Leave-one-out cross-validation
MA	Methamphetamine addiction without psychosis
MAP	Methamphetamine-associated psychosis
MAQC	Microarray quality control consortium
ME	Module eigengene
METH	Methamphetamine
miRNA	Micro-RNA
ML	Machine-learning
MM	Mis-match
mRNA	Messenger RNA
MS	Module significance
NC	Nearest centroid
NGF	Nerve growth factor
NGS	Next-generation sequencing
NIMH	National Institute of Mental Health

NK	Natural killer
NN	Nearest neighbours
PBL	Peripheral blood leukocytes
PCA	Principal component analysis
PM	Perfect match
PPI	Protein-protein interaction
PGC	Psychiatric Genetics Consortium
PTM	Post-translational modifications
PTSD	Posttraumatic stress disorder
QC	Quality control
RDoC	Research domain criteria
RFE	Recursive feature elimination
RMA	Robust multichip average
RNA-Seq	RNA-Sequencing
RPKM	Reads per kilobase of transcript per million mapped reads
rRNA	Ribosomal RNA
RT-qPCR	Real-Time quantitative polymerase chain reaction
SCZ	Schizophrenia
SEQC	Sequencing quality control
snoRNA	Small nucleolar RNA
snRNA	Small nuclear RNA
sMRI	Subcortical brain structural volumes
SVM	Support vector machine
SVM RFE	Support vector machine Recursive feature elimination
TCR	T-cell receptor
TLR	Toll-like receptor
TMM	Trimmed mean of m-values
TOM	Topological overlap measure
tRNA	Transfer RNA
VST	Variance stabilizing transformation
WES	Whole-exome sequencing
WGCNA	Weighted gene co-expression network analysis

Part I

Introduction

Chapter 1

Biomarkers in Psychiatry

1.1. The Global Burden of Psychiatric Disorders

Mental and behavioural disorders are the most debilitating illnesses worldwide. The Global Burden of Disease (GBD) 2010 study identified mental and behavioural disorders as the leading cause of global disability, with an estimated 22.2% of all years lived with disability being attributable to these disorders (Murray et al., 2013). Mood and anxiety disorders, and substance abuse and drug dependence are among the top twenty conditions that result in the greatest burden of disability (Prince et al., 2007; Murray et al., 2013). Disability associated with these conditions exceeds the burden associated with other non-communicable diseases such as cancer, diabetes, and cardiovascular disease, as well as HIV/AIDS, neurological diseases (i.e. stroke, seizures), war and injuries (Murray et al., 2013). As many mental disorders emerge in adolescence and persist into adulthood, the disability associated with mental disorders has a particularly profound impact, given that these developmental years would otherwise typically be the most productive educationally, professionally and economically. Indeed, mental and behavioural disorders account for the greatest percentage of disability between ages 10 to 44 years, and this trend is comparable across low to high income countries (Grandes et al, 2011).

Mental and behavioural disorders, particularly depression, schizophrenia and bipolar disorder, are associated with increased rates of all-cause mortality risk (Craig, 2013). These disorders are also significantly associated with increased risk for suicide, which accounts for approximately 10-15% of deaths for individuals with bipolar disorder and schizophrenia: and mental disorders are a factor in approximately 90% of all completed suicides (Gvion & Apter, 2012). These data are likely to be an underestimate of the true increase in all-cause mortality associated with mental and behavioural disorders as in many cases the mental disorder serves as the longer term condition that increases risk for a more proximal cause of death. For example, impulsivity and substance abuse are associated with death by accidental injury (e.g., vehicular accidents), which would not necessarily be attributed to mental and behavioural disorders (Murray et al., 2013). Suicide reporting is sub-optimal in many countries, which results in underestimates of mortality risk associated with mental and behavioural disorders. Moreover, the relationship between mental and behavioural disorders and other health conditions is complex: mental disorders are closely associated with other health conditions that carry their own burden of disease, and comorbid mental illness incrementally increases the morbidity and mortality risk for chronic diseases such as angina, arthritis, asthma, and diabetes (Moussavi et al., 2007), cancer (Miovic & Block, 2007), and cardiovascular disease (Celano & Huffman, 2011).

1.2. The ICD and DSM Classification Systems

Improved classification of psychiatric disease is the most basic building block for advancing our understanding and treatment of mental and behavioural disorders. Without an internationally standardized and clinically useful classification system, progress in the development and distribution of evidence-based treatments would be heavily constricted. Currently, both the Diagnostic and Statistical Manual of Mental Disorders (DSM) and the International Classification of Diseases and Related Health Problems (ICD) are being re-evaluated alongside the development of an orthogonally designed system focused on dimensions and presumed underlying neurobiological mechanisms of psychiatric pathology, known as the Research Domain Criteria (RDoC).

1.2.1. The Purpose of Classification

The primary purpose of contemporary psychiatric classification systems (DSM and ICD) is three fold (*modified from* Sivkumar, 2009): (**1**) to provide nomenclature and descriptive information regarding patient entities that is essential for communication; (**2**) to guide better treatment and prevention and; (**3**) to provide comprehensive classification or understanding of the causes of psychiatric disorders and the processes involved in their development and maintenance.

1.2.2. The Difficulties of Classification

However, there are several difficulties amongst conventional interview-based diagnostic systems (*modified* from Tyrer, 2014). First, there is an uncertain threshold of psychiatric diagnosis. That is, when a clinician makes a decision about a clinical diagnosis there is little guidance in deciding on the cut-off point between disease and wellness, when this depends on a subjective report of symptoms and a clinician's observations of patient behaviours. Second, many of the existing schemes have proved to have low inter-rater reliability (assessments by different clinicians at the same point in time) and low temporal reliability (assessments carried out at different time points). Third, a reliable classification scheme is not necessarily a valid one. While there is limited empirical support for the clinical validity of most major psychiatric disorders, classification can be an 'all or none' concept. This is also problematic because symptoms are highly nonspecific and quite unstable over time. Finally, psychiatric disorders lack an objective biological basis that confirms clinical impressions of disease, unlike many other medical disorders (e.g. blood sugar in diabetes or blood pressure in hypertension). There are to date no objective clinical laboratory tests for psychiatric disease.

1.2.3. The Evolution of Classification

The ICD is the official world classification of disease and was first introduced in 1900 following the first International Conference for the Revision of the International List of Causes of Death in Paris. It has undergone ten revisions, with the most recent revision being the ICD-10 (WHO, 1992). By contrast, DSM is the official classification in the USA and was first introduced in 1952 following the Korean War, when the US military decided to create a classification of mental disorders. The DSM has undergone successive revisions in 1980, 1987, 1994, 2000, and most recently, in May of 2013 with the DSM-5 being the latest to be published (APA, 2013). In spite of differences between these two major systems (**Table 1.1**), there is much convergence between the two, and it is possible to 'convert' the diagnoses of one system into those of the other.

Table 1.1. **Main Differences between ICD and DSM classification systems.**

ICD	DSM
Official world classification	USA classification system (also used world-wide)
Intended for use by all health practitioners	Used mainly by psychiatrists
Special attention to primary care and low-income countries	Focused mainly on secondary psychiatric care in high-income countries
Major focus on clinical utility with reduction of number of diagnoses	Tends to increase the number of diagnoses with each revision
Provides diagnostic descriptions and guidance but does not use operational criteria	Diagnosis depends on operational criteria using a polythetic system for most conditions

Abbreviations; ICD, international classification of disease; DSM, diagnostic and statistical manual. Grey shading is for visualization purposes only.

The creation of each edition of ICD of DSM of psychiatry has proven enormously controversial. One consequence of the recent revisions is that there is an even higher level of symptom profile heterogeneity. A prime example of expanded symptom heterogeneity has been highlighted following the latest DSM revision (DSM-5), which permits an eight-fold expansion of the total number of possible symptom combinations for the diagnosis of posttraumatic stress disorder (PTSD) from 79,794 symptoms to 636,120 symptoms (Galatzer-Levy & Bryant, 2013): in other words, current DSM-5 criteria permit 636,120 combinations of patient symptoms for the diagnosis of PTSD.

1.3. Research Domain Criteria (RDoC): A New Paradigm for Psychiatry

Recent controversies regarding the DSM-5 reached a pinnacle with an announcement from the National Institute of Mental Health (NIMH) which would shift their efforts and funding to the development of their own psychiatric nosology, the Research Domain Criteria (RDoC) (Cuthbert & Insel, 2013). The aim of the RDoC is to identify brain mechanisms, and related biomarkers, that can explain the etiology and pathophysiology of psychiatric disorders, provide earlier and more accurate diagnosis, and predict treatment responses and outcomes (Casey et al., 2013). It incorporates genetics and behavioural science, including the influence of the environment on neurodevelopment, into a broad neuroscientific paradigm of psychiatry. By exploring the causes of mental illnesses and how these can inform interventions to modulate neural pathways, the circuit-based RDoC should offer a more satisfactory account of these illnesses than the symptom-cluster based ICD and DSM. The RDoC initiative seems to be structured around the concern that the only way to find objectivity in the classification of psychiatric disorders in psychiatry is to begin with biology and work back to symptoms.

1.4. Biological Marker Strategy

An interview-based diagnostic approach for psychiatric disorders is deficient in sensitivity and specificity, and has significant limitations for predicting diagnosis, onset, course of illness and response to treatment. The field of objective biomarkers has made tremendous recent progress, and in some instances have become well-accepted tools in guiding medical practice as in the diagnosis of myocardial infarction (Ahmad, 2012) and management of heart failure (Gaggin & Januzzi, 2013). To accelerate the identification of biomarkers for mental disorders the NIMH Strategic Plan (*modified from* NIMH, 2007) proposes that it will be important to:

1. Support the development of integrated profiles/panels of biomarkers and behavioural indicators (e.g. genes, proteins, brain images, clinical measurements, or a combination of these), creating '*biosignatures'* of disorders. A single biomarker is not likely to be sufficient to indicate the presence of a disorder, but a configuration or combination of biomarkers and behavioural indicators of small effect might do so.
2. Support studies to identify biomarkers and behavioural indicators for different stages of illness and recovery (e.g., biomarkers for onset versus relapse, biomarkers indicating risk versus resilience).
3. Support research that examines biomarkers which may be common to mental disorders and other medical disorders (e.g., inflammatory markers of heart disease) in order to identify shared molecular pathways that contribute to development of mental disorders.

Although there is no widely accepted definition of what constitutes an actual biomarker, the NIH Biomarkers Definitions Working Group defined a 'biomarker' as a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacological responses to a therapeutic intervention (Atkinson, 2001). A biomarker is an indicator of the presence or extent of a biological process that is directly linked to the clinical manifestations and outcome of a particular disease. Mueller et al. (2008) considered a biomarker to be a protein or other macromolecules that are associated with a biological process or regulatory mechanism. In this context, measurement of blood-based biomarkers might provide quantitative information that could be clinically helpful regarding such a mechanism.

Depending on the clinical application, there are different sorts of biomarkers:

- **Antecedent** markers for indicating risk of disease occurrence.
- **Screening** markers for early detection of disease.
- **Diagnostic** markers for revealing an existence of disease.
- **Staging** markers for defining the stage and severity of a disease.
- **Prognostic** makers to predict the course of disease, including treatment response.
- **Stratification** markers that predict treatment response.
- **Biomarker signatures** are indicators of a disease state that are usually linked to an ongoing pathophysiology and thus may also provide information and insights into the underlying molecular mechanisms of disease.

In order for a *diagnostic biomarker* to be useful, certain criteria need to be met (modified from Sunderland et al, 2005; Henley et al, 2005):

1. The biomarker should reflect some basic pathophysiological process, and detect a fundamental feature of the disease.
2. The biomarker should be specific for the disease compared with related disorders.
3. The biomarker can be measured repeatedly over time and should be reproducible.
4. The biomarker should be measured in noninvasive and easy-to-perform tests that can be done at the bedside or in the outpatient setting (i.e. blood tests).
5. The biomarker should not cause harm to the patient being tested.
6. The biomarker should be cost effective.

Many new concepts have arisen in the field of biomarker research (**Table 1.2**). For example, the concept of 'state' and 'trait' have enjoyed wide usage in personality psychology and in other areas of psychology (Alston, 1974; Zuckerman 1976). A *trait biomarker* is used for revealing properties of the behavioural, neuropsychological and biological processes which often play an antecedent role in the pathophysiology of the psychiatric disorder. Terms closely related to trait biomarker include elementary phenotype, intermediate phenotype, risk indicator, risk marker and vulnerability marker (Adler et al, 1999; Agarwal, 2001; Gould & Manji, 2004). A *state biomarker* is in essence, a diagnostic marker that reflects the current status of clinical manifestations in patients.

Table 1.2. **Overview of some definitions used in the field of biomarker research[1].**

Biological Marker	Measurable and quantifiable biological parameters which serve as indices for health – and physiology related assessments.
Clinical Endpoint	A characteristic or variable that reflects how a patient feels, functions or survives.
Surrogate biomarker	A laboratory measurement or physical sign that is used in therapeutic trials as a substitute for a clinically meaningful endpoint that is a direct measure of how a patient, feels, functions or survives and is expected to predict the effect of the therapy. Generally biochemical markers which are easy and quick to measure, predictive in nature (e.g. peripheral blood).
Validation	The process of assessing the assay or measurement performance characteristics and qualification is evidentiary process of linking a biomarker with biology and clinical endpoints.
Genetic marker	A single gene (DNA) for which a mutation, deletion, single nucleotide polymorphism (SNP) or some other feature provides predictive value.
Epigenetic marker	Measurable chemical modifications to DNA or histones.
Transcriptomic marker	A measurable RNA molecule that is an indicator of normal biologic processes, pathogenic processes, and/or response to therapeutic or other interventions.
Proteomic marker	A protein expression pattern which is able to discriminate or predict.
Metabolomic marker	A pattern of metabolites which is able to discriminate or predict.
Physiological marker	Endocrine or autonomic measurements indicative of physiological responses to disease.
Neuroimaging marker	A non-invasive diagnostic tool for depicting brain chemistry, function or structure.

[1] ***Based on: Biomarkers Definition Working Group 1998; Biomarkers Definition Working Group 2001; Russel, 2004; Wagner & Merck, 2004. Grey shading is for visualization purposes only.***

1.5. Advances in Genome Technology

Recent successes in psychiatric genomics are the result of a confluence of several factors, some of which were set in motion decades ago, which suggest a considerably brighter future for our ability to delineate the ‘genetic architecture’ of mental disorders. The Human Genome Project (1990-2003) and accompanying advances in technology and population genetics included a shift from targeted association studies based on candidate genes to hypothesis-free genome-wide analytical approaches. In parallel, new approaches in bioinformatics and statistics, such as high-efficiency analysis pipelines and imputation, have made possible increasingly powerful and efficient study designs and analyses of genome-wide data of neuropsychiatric disorders (McKenna et al, 2010). These advances provide an opportunity for discovery of biological markers for psychiatric diseases that is starting to bear fruit. The search for psychiatric biomarkers has encompassed several technological and methodological areas of research, including studies of genetic (DNA), epigenetic (methylation), transcriptomic (RNA) and proteomic (protein) factors (**Table 1.3**). It is anticipated that such biomarkers will, in the hands of clinical investigators, provide a dynamic and powerful approach to understanding the spectrum of psychiatric disease with obv ious applications in clinical trials and disease prevention, diagnosis, and disease management.

Table 1.3. Commonly used genome-wide technologies, platforms and applications.

Tool	Technology	Applications
Genomics	Genome-wide association study (GWAS)	Single Nucleotide Polymorphisms (SNP)
	DNA Microarray	Copy number variants (CNV)
	Whole Exome Sequencing	Rare and *de novo* variants
Epigenomics	Bisulfite Sequencing	Quantification of methylation patterns
	Methylation Array	
Transcriptomics	Microarray	Quantification of gene expression
	RNA-Sequencing	
Proteomics	Mass Spectrometry	Quantification of protein expression

Grey shading is for visualization purposes only.

1.6. The Value of Transcriptome Biomarkers

Given the high heritability of many psychiatric diseases (i.e. autistic spectrum disorder and schizophrenia), the discovery of genetic polymorphisms has rightfully encouraged much attention into their role as putative biomarkers. Results from recent genome-wide association studies (GWAS), copy number variation (CNV) and whole-exome sequencing (WES) studies are enabling, for the first time, empirical assessment of fundamental questions about the genetic architecture of psychiatric disorders (Gratten et al, 2015). However, referring back to the definition, a biomarker should be sufficiently dynamic to reflect changes in the processes it intends to index; DNA sequence variations, as ‘static elements’ do not necessarily satisfy this criterion. This is why psychiatric biomarker research tends to focus on more dynamic downstream readouts of the genome, such as the transcriptome (i.e. the RNAs transcribed from an individual’s genome) (Mirnics et al., 2006). The transcriptome has some fixed features, such as its variety of RNA sequences as dictated by the fixed DNA sequences it transcribes, but the transcriptome also has dynamic aspects, such as the amounts and combinations of RNAs expressed (i.e. gene expression) at various times within a cell in response to genetic, biological, and environmental cues. This includes different phases of the cell cycle, drug treatment, stress, aging, and diseases, all of which must be considered at the time of their determination. It is this property that makes the transcriptome a practical tool for the discovery of gene function and as a suitable molecular signature in psychiatric research.

1.7. Trade-offs Between Brain and Blood Gene Expression Biomarkers in Psychiatric Disease

As described in **Table 1.2**, a transcriptomic biomarker is a measurable RNA molecule that is an indicator of normal biologic processes, pathogenic processes, and/or response to therapeutic or other interventions. But just, ‘what’ constitutes a transcriptomic biomarker and ‘how’ are these biomarkers derived? Discovery begins with the extraction and isolation of total RNA from a target patient sample prior to experimental procedures for measuring the transcriptome (discussed in Chapter 2) and statistical applications for identification and validation of candidate biomarkers (discussed in Chapter 3). In molecular psychiatric research, RNA is commonly extracted from either patient post-mortem brain samples or peripheral blood samples, with considerable trade-offs between the two.

1.7.1. Post-Mortem Brain Gene Expression

The benefit of post-mortem brain gene expression is that it provides a mechanistic basis for directly understanding disease etiology: however several precautions need to be considered. For example, cellular diversity between and across various brain regions constitutes a serious problem for brain gene expression studies (Mirnics, Levitt et al., 2006). For most studies, the proportion of brain cell types (i.e. neurons and glia) that make up the final constituent cells for RNA extraction remains completely unknown. As gene expression between neurons and glial cells is likely to be different and related to their various functional roles, differential sampling of cells in postmortem investigations would lead to gene expression differences arising from the various proportions of cells present. Laser capture microdissection presents a potential way around this obstacle by cutting out individual cell types in order to analyze their gene expression separately (Mirnics et al., 2006). However, the use of this technology to harvest a significant number of cells for this purpose is labour intensive. Furthermore, dissection of individual cells from brain tissue will lead to sampling of some of the neuropil, which may also confound findings. Alternatively, *in silico* computational methods aim to deconvolve the frequencies of differing cell types by working backwards to estimate specific cell type contributions from the observed gene expression picture. Other challenges faced by brain gene expression studies include (**1**) variability introduced by genetic diversity, (**2**) effects of disease treatment on gene expression, (**3**) differential diagnoses, (**4**) comorbidity with other disorders, (**5**) variation in age, pH, and drug abuse between groups in a cohort, (**6**) limited sample sizes with a small number of samples yielding high-quality RNA for investigation, and (**7**) variability in platform types and methods for hybridization, (**8**) difficulty in obtaining a sufficient number of well-preserved brain samples (Mirnics, Levitt et al., 2006). (**9**) Perhaps the most obvious limitation of brain gene expression as a putative biomarker of disease etiology is the relative inaccessibility of the brain and thus the inability to sample and re-sample gene expression in an effort to monitor disease status or drug effects over extended periods of time. Thus, analysis of patient blood profiles (as presented in Chapters 5-7) offers an alternative means to investigate mechanisms relevant to psychiatric disease, providing a basis for discovery of clinically relevant biomarker signatures as a potential non-invasive surrogate of brain gene expression and function.

1.7.2. Peripheral Blood Gene Expression

Blood-based gene expression could similarly suffer from brain-based limitations (**1**), (**2**), and (**3**). In addition, lifestyle factors also significantly impact blood gene expression: diet, exercise, smoking, and time of last meal can all affect gene expression in the blood and hence matching and normalization for these factors where possible should be a standard consideration along with the other factors mentioned earlier when designing such studies (Demeaux et al., 2010). However, these variables can be controlled by the researcher (i.e. patients could be asked to give a fasting blood sample first thing in the morning). Additional questionnaires pertaining to the patient's lifestyle can be administered to glean as much information as possible regarding potential influences on gene expression changes. Importantly, blood collection by venipuncture is relatively non-invasive and can be performed as often as required in patients.

Blood contains several different cell types as its constituent components. These cell types fall primarily into three categories: erythrocytes, leukocytes and thrombocytes (Jankowsky et al., 2015). With leukocytes making up the immune component of blood, focus has been cast on assessing gene expression changes within this subcategory of blood cells. However, even within this category, several cell types exist: neutrophils, eosinophils, basophils, lymphocytes, monocytes, and macrophages. This inherent variability and difference in the functional roles of leukocytes is no doubt reflected in differences in their gene expression profiles. Furthermore, based on the immune status of an individual, different proportions of these cell types may be present. Isolation of specific lymphocyte types may be achievable through techniques such as flow cytometry analysis and immunolabeling with magnetic beads conjugated to antibodies for specific cell-surface markers followed by separation through electromagnetic columns. However, such techniques are relatively inefficient ways of yielding sufficient numbers of cells for a representative microarray investigation and place further 'stress' on the cells, so altering gene expression patterns (Beliakova-Bethell et al., 2013). In an attempt to circumnavigate problems with cell variability in blood samples, some researchers have chosen to use skin fibroblast cultures or to transduce and culture B-lymphocytes into lymphoblastoid cell lines using Epstein Barr virus prior to analysis (as presented in Chapter 8). These studies have the advantage of assessing gene expression effects in a relatively homogenous cell population free from the temporal state of individuals, but effects of exposure to virus and chromosomal alterations during culture may be considered as a confounder of these results (Iwamoto and Kato, 2006).

1.7.3. Correspondence Between Brain and Blood Gene Expression

The role of peripheral blood gene expression as a non-invasive surrogate for brain gene expression has been questioned due to the presumed indirect nature of communication between blood and brain cells. However, several lines of evidence suggest that both brain and blood cells respond to environmental stimuli at the transcriptomic level, and that this response is to some extent concordant between both tissue types. Research into the correspondence of gene expression across blood and brain compartments reveals that 35% - 80% of known transcripts are believed to be present in both tissues, with relatively low correlations ($r < 0.65$) (Tylee et al., 2013). However, the expression of a biomarker in the blood may not need to resemble the expression of the same analyte in the brain. An increasing body of literature emphasizes the fine-tuned communication between many seemingly distant bodily systems. For example, the discovery of a lymphatic vasculature reaching the meningeal linings of the brain highlights the potential role of peripheral mechanisms in brain disease etiology (Louveau et al., 2015). Results from recent blood-based gene expression studies are enabling empirical assessment of fundamental questions regarding functional genomic aspects underlying psychiatric disorders. To date, 144 publications have indicated dysregulation of immune-related genes in the periphery associated with the pathophysiology of psychiatric conditions using genome-wide transcriptome tools (Breen et al., 2016 under review). It could further be reasoned that because RNA expression can be influenced by both heritable (genetic and epigenetic) and non-heritable factors, measurements of the transcriptome in peripheral blood therefore might reflect common pathways in which untoward effects of risk genes and detrimental environmental risk factors converge. Therefore, a number of studies presented in this body of work (Chapters 5-8) evaluated the transcriptome in the blood compartment.

1.8. Measuring the Transcriptome

Once RNA has been isolated from a target patient sample, the next step is measuring the transcriptome, which consists of profiling RNA abundance on genome-wide scales. Analytical techniques for measuring the transcriptome include microarray and next generation RNA-Sequencing.

1.8.1. Microarray Technology

Microarray technology provides a powerful genome-wide approach allowing the simultaneous study of the expression of thousands of genes or their RNA products, giving an accurate picture of gene expression in the cell or the sample at the time of the study (Mirnics et al., 2006). The starting point for a microarray is the hybridization of total cellular RNA (e.g. RNA extracted from patient samples) to hundreds of thousands of short oligonucleotide probes representing genomic DNA. A typical modern microarray consists of patches of such probes complementary to the transcripts whose presence is to be investigated, bound to a solid substrate. As the design, chemistry and kinetics of microarrays advances, this technology earned its place when aiming for biomarker discovery, provided the results are independently reproducible and the findings are critically evaluated with reference to other data (often *post hoc*). Although there have been many efforts to identify gene expression biomarkers for psychiatric diseases using microarray technology, the identified genes seldom overlap across studies, and attempts to replicate previous findings in different cohorts have generally yielded disappointing results (Yao et al, 2008). Much of these discrepancies may be due to differences in RNA isolation, library preparation and technological platforms (discussed in Chapter 2). However, several computational approaches exist that can be used to find reproducible signatures across independent studies by focusing attention on groups of genes with similar functionality, cellular compartments, or correlation patterns (discussed in Chapter 3).

1.8.2. RNA-Sequencing Technology

In recent years, researchers have turned towards direct high-throughput RNA-Sequencing (RNA-Seq), which has considerable advantages over microarrays for the study of gene expression (Wang et al, 2009). The sequencing framework of RNA-Seq enables the investigation at high resolution of all the RNAs present in a sample, characterizing their sequences and quantifying their abundances at the same time (discussed in Chapter 2). In practice, millions of short strings, called ‘reads’, are sequenced from random positions of the input RNAs: these reads can then be computationally mapped on a reference genome to reveal a transcriptional repertoire for a particular sample, where the number of reads aligned to each gene gives a measure of its level of expression. With RNA-Seq, it is possible to determine the absolute quantity of every RNA molecule in a cell, and directly compare results between experiments (Wang et al., 2009). RNA-Seq has been a rich platform for deriving functional interpretation of molecular mechanisms underlying disease pathology and in providing putative predictive and diagnostic biomarkers in psychiatric disease: moreover, it has also been useful in identifying convergence across disorders (Breen et al., 2016). However, using huge and complex RNA-Seq datasets to generate biologically meaningful findings is not a trivial exercise and requires sophisticated computational strategies (discussed in Chapter 3).

1.9. Shapes and Forms of Transcriptome Biomarkers

In a statistical sense, once the transcriptome has been measured, it is possible to conceptualize three different levels of biomarker discovery (discussed in Chapter 3). First, for development of an objective test for psychiatric disease prediction and diagnosis, all clinical, biological and statistical aspects converge on the construction of a gene expression ‘classifier’: a classifier being a unique panel of cross-validated gene expression measurements (i.e. biomarkers) capable of differentiating between psychiatric conditions of interest. Second, in a strictly conventional sense a transcriptomic biomarker could also refer to a group of differentially regulated genes with sufficient statistical support. That is, when considering a two-group experimental design (e.g. PTSD versus controls) it is often of interest to consider which transcripts are significantly differentially regulated between these groups, while accounting for potential confounding factors (i.e. age, smoking, gender, therapy etc..). Third, the blood transcriptome itself is fluid and dynamic, and made of many interacting molecules and cells. Recent statistical approaches aim to capture this dynamic by identifying gene networks, or groups of coordinately expressed transcripts, which comprise functional biomarkers of disease.

At present, a consideration of the potential to derive clinical utility from gene expression should lead to reflection on the benefit, feasibility, and reproducibility, and scope for a putative prospective trial (Guest & Bahn, 2011). Benefit refers to the initial assessment of whether a predictive or diagnostic gene expression classifier is likely to improve on the accuracy or reduce the cost of any tool that may already exist for the disease. The feasibility of a classifier refers to its initial discovery and establishing its utility on a pilot study of interest. From here, a larger number of patients from the same cohort used in the pilot study should then be analyzed in the interval validation step, and once developed, needs external replication using an unrelated cohort of patients. Finally, a prospective trial should be initiated to evaluate the potential clinical utility of a gene expression classifier using prospective longitudinal studies.

1.10. Summary

In summary, peripheral blood is an ideal surrogate tissue since it is readily obtainable, provides a large RNA pool in the form of gene transcripts, and response to changes in the macro- and micro-environments is detectable as alterations in the levels of these gene transcripts. However, before blood-based gene expression biomarkers can be detected, several measures must be considered. First, *experimental methodologies* focused on RNA treatment and library preparation, as well as microarray hybridization or RNA-Sequencing need to be considered (discussed in Chapter 2). Next, *statistical methodologies* and the use of computational pipelines are required to derive potential clinical and biological significance (discussed in Chapter 3). The utility of these techniques demonstrated in a series of exploratory studies to PTSD (Chapter 5), acute psychological stress (Chapter 6), methamphetamine-associated psychosis (Chapter 7) and bipolar disorder (Chapter 8). Finally, future avenues that will accelerate the development of accurate and objective blood-based tests in psychiatry are discussed (Chapter 9).

Chapter 2

Experimental Methodologies in Blood Transcriptomics

The discovery of blood-based biomarkers could greatly advance psychiatry, however the realization of this may take some time. One reason for this is that a ‘gold-standard’ approach for generating gene expression measurements is challenged by the existence of different experimental techniques and strategies. The typical workflow for blood-based gene expression studies includes several steps (**Figure 2.1**), each with diverse parts for addressing specific experimental aims. This chapter summarises these technical aspects and dissects each into its respective sub-components. First, I discuss the importance of performing power analysis and experimental design strategies prior to reviewing transcriptome RNA diversity. Then I address RNA isolation and preparation methods from patient blood samples and consider technological platforms for generating blood-based gene expression. Finally, I describe techniques for optimizing RNA-Seq platforms, which are important for the discovery of blood-based biomarkers in psychiatric diseases.

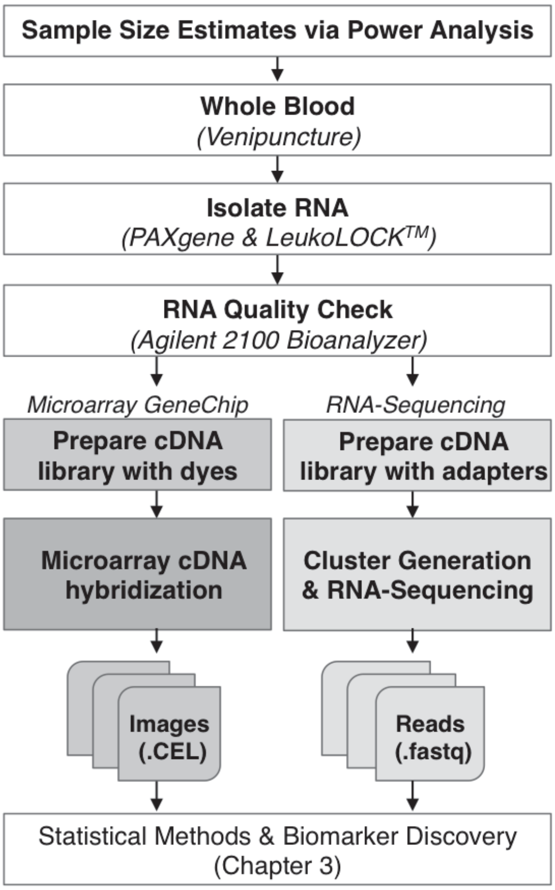

Sample Size Estimates via Power Analysis
Whole Blood
(Venipuncture)
Isolate RNA
(PAXgene & LeukoLOCK™)
RNA Quality Check
(Agilent 2100 Bioanalyzer)
Microarray GeneChip
RNA-Sequencing
Prepare cDNA library with dyes
Prepare cDNA library with adapters
Microarray cDNA hybridization
Cluster Generation & RNA-Sequencing
Images (.CEL)
Reads (.fastq)
Statistical Methods & Biomarker Discovery (Chapter 3)

Figure 2.1. A typical gene expression work-flow. Power analysis indicates the optimal number of replicates and whether there is sufficient power to detect a biological effect. Following these estimates, venipuncture is performed and RNA is isolated, quality checked and prepared for either microarray hybridization or RNA-Sequencing.

2.1. Experimental Design

Blood transcriptome studies are often complex, generate large amounts of data, and warrant careful planning. Many such studies begin with questions such as, "What is an ideal population size for discovery?" and "How can I ensure that my findings are meaningful and can be advanced upon?". The most obvious criterion that can add power to any gene expression study is the number of patients available for investigation. Determining adequate sample sizes able that will achieve the experimental goals is an important first step. To better understand the relationship between sample size and the sensitivity and specificity of a gene expression biomarker, Dobbin et al (2008) developed a power calculator specifically for gene expression studies. This tool, and ones alike, use the number of probes on the microarray platform, the likely proportion of samples in each class (i.e., PTSD vs. controls), and an estimate of the standardized fold change between classes based on the gene exhibiting the greatest difference. In addition to ensuring that enough samples are analyzed, it is best to have a balanced study design in which similar numbers of psychiatric 'cases' and controls are compared (Dupuy and Simon, 2007). Gene expression studies also generally contain two levels or replicates: first, technical replicates which provide measurement-level error estimates, and second, biological replicates that provide estimates of population-level variability. With stabilizing technological platforms, common practice is positioned towards measuring more biological replicates rather than technical replicates. It is also desired that all clinical and socio-demographic variables are balanced between the two groups being compared (e.g. equal number of males and females per group) to mitigate against confounding factors. Finally, to best evaluate the specificity of a biomarker, it is important to consider more complex experimental designs beyond the standard two group comparison. For example, if the study focus is on discovery of blood biomarkers for classifying PTSD, it may be useful to also consider the specificity of this biomarker panel relative to symptomatically similar phenotypes such as major depressive disorder, obsessive-compulsive disorder and acute stress disorder.

2.2. Transcriptome RNA Diversity: Implications for Biomarker Discovery

Blood-based biomarker discovery is mainly focused on the expression levels of messenger RNA (mRNAs). This is useful because one can derive putative protein-level implications from mRNA expression levels (i.e. DNA ➔ mRNA ➔ protein). Non-coding RNA (ncRNA) transcripts were traditionally believed to be 'by-products' derived from mRNA degradation or nonspecific polymerase activity, and therefore termed 'transcriptional noise'. However, it is now clear that ncRNAs, such as microRNA (miRNA) and long non-coding RNAs (lncRNAs), are responsible for many aspects of gene regulation including regulation of translation, of mRNA transcription, and self-regulated transcription (Jankowsky et al., 2015). Hence it is important to appreciate the range of

RNA species across the transcriptome in order to exploit this diversity in the search for putative biomarkers (**Figure 2.2**).

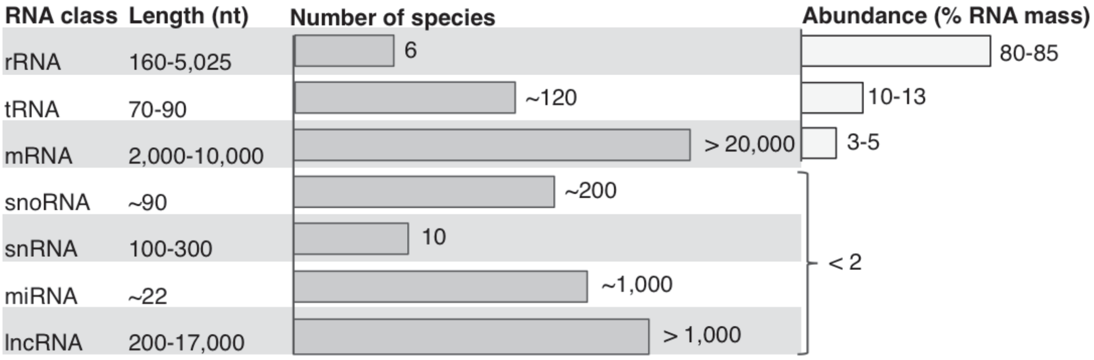


Figure 2.2. For each class of RNA, approximate length, number of different species and abundance are indicated. Nt; nucleotide. Figure adapted from (Jankowsky et al., 2015).

The primary division of total RNA in the transcriptome is between protein-coding (mRNA) and ncRNAs. Human cells generally contain ~20,000-25,000 different mRNAs, comprising ~4% of total RNA (Jankowsky et al., 2015). mRNA diversity is further increased by alternative splicing (Nilsen et al., 2010) and by chemical modifications (Liu et al., 2013). In addition to mRNAs, cells can also express thousands of species of lncRNAs and hundreds miRNAs, transfer RNAs (tRNAs) and small nucleolar RNAs (snoRNAs). Conversely, there are only a few ribosomal RNA (rRNA) and small nuclear RNA (snRNA) species. Despite this, rRNAs account for approximately 80–85% of the cellular RNA mass, followed by tRNAs (~10%): mRNAs comprise only ~4% of total RNA, and all other RNAs together account for less than 2% of the mass. Another factor contributing to the disparity in cellular RNA mass is that RNAs vary greatly in length, from more than 10,000 nucleotides (mRNAs and lncRNAs) to only 22 nucleotides (miRNAs).

Each RNA class has a functional role:

- rRNAs carry out protein synthesis.
- tRNAs carry amino acids to the ribosome and ensure that amino acids are linked together in the order specified by the nucleotide sequence of the mRNA that is being translated.

- mRNAs transport genetic information from DNA to the ribosome, where they specify amino acid sequence of the protein products of gene expression.
- snoRNAs guide chemical modification of rRNAs and tRNAs.
- snRNAs are mainly involved in splicing.
- miRNAs regulate the expression of individual genes.
- lncRNAs are important regulators of transcription and translation but have additional only partly defined roles.

2.3. RNA Isolation from Patient Blood Samples

Following blood collection via venipuncture, the first experimental step for blood transcriptome biomarker discovery involves isolating and purifying cellular RNA. RNA is more labile than DNA, and the moment blood is drawn cells begin to die and RNA begins to be degraded. If RNA isolation is not done immediately it is advisable that blood samples are frozen (e.g. liquid nitrogen or -80°C) with a RNA stabilizer, such as RNAstable® (Biomatrica) or RNAlater® (Qiagen). Blood-based RNA isolation systems are typically aimed at isolating total RNA. PAXgene Blood RNA tubes and LeukoLOCK^{TM} Filters are commonly used methods for mRNA extraction from whole blood.

2.3.1. PAXgene RNA Extraction and Globin mRNA Depletion

- **Cell Lysis**. As specified by the PAXgene RNA kit (Qiagen, CA, USA), whole blood samples (2.5ml) are concentrated by centrifugation. The pellet is then washed with 4ml RNase-free water, re-suspended with 350µl re-suspension buffer and incubated in buffers containing 300µl binding buffer and 40µl Proteinase K for protein digestion.
- **RNA isolation**. Another centrifugation step is done to remove residual cell debris. After the addition of 350µl ethanol, the lysate is applied to a silica-gel membrane/column. Upon centrifugation, RNA that remains bound to the membrane and contaminants is removed by three washes using washing buffer PBS (700µl, 500µl and 500µl respectively). The total extracted RNA is then eluted using 40µl elution buffer, incubated at 65°C for 5 minutes, then immediately chilled on ice.

- **Removal of Globin mRNA**. Here, RNA is extracted from whole blood and thus will be affected by globin interference, which includes >95% of globin mRNA. Globin mRNA is commonly expressed at high levels in red blood cells (RBCs) and reticulocytes. Therefore, globin mRNA is depleted from blood samples using the GLOBINclear - Human Kit (Life Technologies, USA), by mixing total RNA with the biotinylated Capture Oligo Mix that is specific for human globin mRNA. The mixture is incubated (15 min) to allow biotinylated oligonucleotides to hybridize with globin mRNA. Streptavidin magnetic beads are then added, and the mixture is incubated for 30 min, before capturing the beads on the side of the tube using a magnet stand. Total RNA (depleted of the globin mRNA) is transferred to another tube followed by RNA purification using a rapid magnetic bead based purification method. This consists of adding a bead re-suspension mix buffer to the RNA sample. The magnetic beads are captured, washed and GLOBINclear RNA is eluted.

2.3.2. LeukoLOCK^{TM} RNA Extraction

- **Sample Collection and Capture of Leukocytes**. As specified by the LeukoLOCK^{TM} RNA kit (Qiagen, CA, USA), whole blood samples (10ml) are collected into an EDTA-coated collection. Blood is passed over a LeukoLOCK^{TM} filter using an evacuated tube as a vacuum source, which is flushed with PBS and then fully saturated with 3mL of RNAlater®. Where as PAXgene tubes require a globin depletion step, the LeukoLOCK^{TM} filter is a filter-based leukocyte-depletion step aimed to isolate leukocytes from whole blood and RNAlater® solution and to stabilize the cells on the filter, thereby avoiding a globin depletion step. At this point, if cell lysis and RNA is not ready to be performed the filters may be stored at -80°C for up to 50 months.
- **Filter Processing and Cell Lysis**. Subsequently, filters are flushed with 2.5mL cell lysis solution and incubated for 5 minutes in buffers containing 2.5ml nuclease-free water and 25µl Proteinase K for protein digestion.
- **RNA Isolation**. Following, 50µl RNA binding beads are added along with 2.5ml isopropanol and incubated at room temperature for 5 minutes prior to recovering to centrifugation and discarding supernatant. Upon centrifugation, RNA binding beads are recovered and washed three times with PBS wash buffer (1.2ml, 750µl, 750µl,

respectively) and left to air dry on the third wash. Finally, the total extracted RNA is eluted using 50µl elution buffer.

2.3.3. RNA Quality Check

RNA concentration and purity can be measured using a spectrophotometer. RNA concentration is measured at absorbance of 260nm (A_{260}) adjusting the (A_{260}) measurement for turbidity (measured by absorbance at 320nm), multiplying by the dilution factor, and by 40, because 1.0 unit of A_{260} correlate with 40µg/ml of pure RNA (Formula 2.1).

$$\text{RNA concentration } (\mu g/ml) = (A_{260} - A_{320}) \times \text{dilution factor} \times 40\mu g/ml$$

(Formula 2.1)

The RNA purity is estimated from the A_{260}/A_{280} ratio, between 1.7 and 2.0 generally represents a high quality RNA sample. The ratio could be calculated after correcting for turbidity (absorbance at 320nm) (Formula 2)

$$\text{RNA purity } (A_{260}/A_{320}) = (A_{260} - A_{320}) / (A_{280} - A_{320})$$

(Formula 2.2)

Following, the integrity of the total RNA can be assessed for signs of degradation by running the Agilent 2100 Bioanalyzer (Agilent; CA, USA) per manufacturer's recommendation. This produces an RNA integrity number (RIN) ranging from 1 (totally degraded) to 10 (good-quality RNA). A general rule of thumb is to only proceed with gene expression generation with samples that have a RIN of 7 or higher (Guest & Bahn, 2011). It is sometimes necessary to digest DNA using DNase I in the event of DNA contamination.

2.3.4. RNA Target Enrichment

Total RNA recovered from PAXgene and LeukoLOCK methods consists of >80% ribosomal RNA (rRNA) (Raz et al. 2011). If this portion of RNA is not removed from the sample the majority of final sequenced reads would be enriched for rRNA and this would provide a major limitation to transcriptomic exploration. Therefore, it is often necessary to further enrich the total RNA sample for RNA targets of interest. Two target enrichment methods applied to experiments in the context of this thesis include:

Selection of target RNAs via hybridization. During RNA processing, a poly-adenylated (poly-A) tail (i.e. a long chain of adenine nucleotides) is added to mature mRNAs to increase the stability of the molecule. The method of target RNA selection via hybridization uses oligo-dT primers to selectively fish out mature mRNAs by duplexing their poly-A tails. At the same time, the proportions of RNA classes that do not have long poly-A stretches will be reduced. As this method only recovers poly-A$^+$ RNAs it is useful for characterizing the levels of mature mRNAs, but other RNAs such as immature mRNAs and poly-A$^-$ ncRNAs will be lost. Moreover, some mitochondrial mRNAs are also poly-adenylated and will be enriched by oligo-dT.

Removal of ribosomal sequences via hybridization. This step is used to selectively remove rRNA from total RNA samples. In contrast to poly-A enrichment, this approach preserves poly-A$^-$ RNAs allowing investigation of broader classes of RNAs including immature mRNAs and poly-A$^-$ ncRNAs. This technique uses oligos that are complementary to highly conserved rRNA sequences to bind and remove rRNA from solution via binding beads. Different commercial kits use different technologies to capture the bound complex, however all kits are capable of removing the majority of rRNA from a total RNA sample. The oligos in the Ribominus (Invitrogen/Life Technologies) and Ribo-Zero (Epicentre/Illumina) kits have a biotin tag that can be captured using streptavidin coated magnetic beads. Moreover, since these kits rely upon a limited number of oligos they only work well if the input RNA is not degraded.

Additional target enrichment methods, not used in the context of this thesis, have also gained popularity and deserve attention:

Copy-number normalization via duplex-specific nuclease digestion (**DSN**). DSN-normalization is a technique that partially normalizes the concentrations of each mRNA by selectively removing many of the most abundant transcripts, which effectively increases the relative concentration of low abundance transcripts (Zhulidov 2004 and Zhulidov 2005). The concentration of specific mRNAs varies dramatically within a cell. Some transcripts may be present at relatively high concentrations (>10,000 copies per cell) while for others there may be only a few copies. Therefore much deeper sequencing is required to interrogate the low abundance transcripts. DSN can therefore be very useful for RNA-Seq experiments that have annotative goals such as gene-discovery and characterization of transcript architectures (Ekblom et al. 2012).

Target enrichment via size-selection. This method is generally reserved for enrichment of small ncRNAs such as miRNA and siRNA. Since these RNAs are much smaller than mRNA and rRNA they can be separated by electrophoresis of the total RNA through an agarose or acrylamide gel and then by cutting out the region that corresponds to the size of interest. Although effective, this method is laborious and recoveries can be low. Several companies now offer small RNA purification kits that are based on this solid phase extraction approach.

2.4. Steps Required for a Microarray Experiment

Microarrays are analog, hybridization-based, high-throughput assays that can measure a complete set of transcripts in a given biological sample (i.e. the blood transcriptome). Microarrays consist of an orderly arrangement of oligonucleotide probes bound to a glass or silicon slide (i.e. gene chip) that are complementary to fluorescently labeled cDNA. In this way, microarrays provide supervised detection of hundreds of thousands of probes for genes per array.

Figure 2.3 represents the experimental protocol to generate microarray gene expression data. Two different conditional groups are being tested (psychiatric cases and healthy controls). Sample preparation starts by isolating and purifying RNA containing mRNA that ideally represents a quantitative copy of genes expressed at the time of sample collection (as discussed above). Subsequently, mRNA is converted into cDNA using a reverse-transcriptase enzyme. This step also requires a short primer to initiate cDNA synthesis. Next, each cDNA is labeled with a different tracking molecule, often red and green fluorescent cyanine dyes (i.e. Cy3 and Cy5). The labeled cDNA (psychiatric and healthy conditions) are then mixed together, and purified to remove contaminates using the Quaquick from Qiagen. After purification, the mixed/labeled cDNA is hybridized to probes fixed on the surface of the microarray chip. Each gene target will bind to a probe on the array that contains its complementary DNA sequence. A laser scanner (special for red and green dyes) is used to read the fluorescent intensity off each probe and the intensities represent the relative abundance of targets for each probe. The raw data from a microarray experiment is therefore an image (.CEL file). For example, in **Figure 2.3**, if the psychiatric condition for a particular gene was in lesser abundance than the healthy condition, one would find the spot to be green. If it was the other way around, the spot would be red or if the green was in similar abundance it would be yellow.

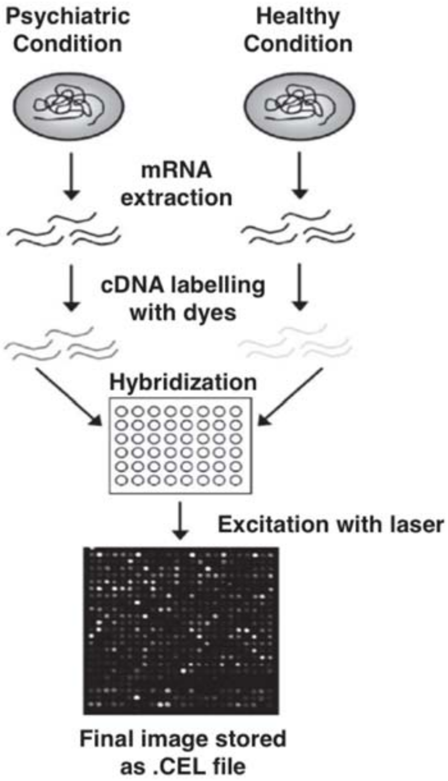


Figure 2.3. Scheme depicting generation of gene expression by microarray. In short, mRNA is isolated, reverse transcribed, hybridized, excited with laser technology and the final image stored as a .CEL file. Figure adapted from Babu (2006).

2.4.1. Microarray Platforms

Some of the results described within this thesis (Chapter 5-6) have been generated by the *Affymetrix Human Gene 1.0ST Array* and the *Illumina HumanHT12 v4 BeadChip Array* and therefore a short comparison between these two platforms is presented here. The latest Affymetrix Human Gene 1.0ST Array provides coverage for 32,020 coding transcripts and 2,967 non-coding transcripts, and further covers 466 lncRNA transcripts (RefSeq annotated). The Illumina HumanHT12 v4 BeadChip Array covers 31,000 annotated protein-coding genes using > 47,000 probes (RefSeq annotated), permitting up to 12 samples to be run on one chip. For the Affymetrix gene chip, probes are defined as gene-specific ~25mer oligonucleotides, whereas the Illumina BeadChips utilize 50mer probe sets which are can be more sensitive for detecting mRNA signal. Affymetrix arrays, however, also contain perfect match and mismatch probe pairs which can be used to help eliminate background noise and interpretation of down-stream results (discussed in Chapter 3).

2.4.2. Strengths and Weaknesses of Microarray

Microarray hybridization-based approaches are high-throughput and relatively inexpensive. Methodological limitations include: (**1**) reliance upon existing knowledge about genome sequences; (**2**) high background levels owing to cross-hybridization (i.e. hybridization between sequences that are not strictly complementary); (**3**) and a limited dynamic range of detection owing to both background and saturation of signal. Moreover, (**4**) comparing expression levels across independent experiments is often difficult and can require complicated normalization methods.

2.5. Steps Required for an RNA-Seq Experiment

RNA-Seq workflows include 3 basic steps: (**1**) RNA-Seq Library Preparation, (**2**) Cluster Generation and (**3**) Sequencing. Unlike different microarray technologies, different RNA-Seq platforms necessitate special RNA library preparations, sequencing techniques and initial data pre-processing steps. Some of the results presented within this thesis have been generated by the *Illumina Hi-Seq 2000* (Chapters 5 and 7) and the *Ion Torrent*

(Chapter 8) sequencing platforms and therefore details for both of these technologies are presented.

2.5.1 RNA-Seq Library Preparation

A typical RNA-Seq library preparation is described in the following **Figure 2.4**. Following RNA isolation and target enrichment the RNA is first fragmented. Unlike short RNAs, mRNAs are typically fragmented to smaller pieces of RNA to enable sequencing. Protocols differ as to when the fragmentation is performed (i.e. fragmenting before converting RNA into cDNA) but regardless of which method is used, it is important that the shearing is random and produces a fairly tight symmetrical size distribution (~200-300bp) depending on experimental goals.

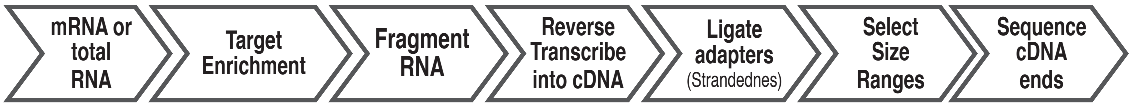


Figure 2.4. A typical RNA-Seq library preparation.

Library Preparation by Illumina. Following fragmentation, first and second strand cDNA is reverse transcribed from fragmented RNA using random hexamers or oligo-dT primers followed by a adapter ligation step. The 5’ and 3’ ends of cDNA are repaired and adapters (containing unique sequences to allow hybridization to a sequencing flow cell) are ligated. Then libraries are enriched for correctly ligated cDNA fragments and amplified by PCR to add any remaining sequencing primer sequences.

Library Preparation by Ion Torrent. Following fragmentation, partly degenerate guide adapters hybridize the fragmented target RNA to allow splint ligation of 5’ and 3’ adapter with defined sequences. Next, cDNA is synthesized and amplified by PCR to add additional required sequences followed by emulsion PCR on microbeads.

2.5.2. RNA-Seq Cluster Generation

Cluster Generation by Illumina. For cluster generation, once libraries have been prepared they are loaded into a flow cell where fragments are captured on a ‘lawn’ of surface-bound oligos complementary to the library adapters. Once hybridized, the captured oligonucleotide primes DNA polymerase extension activity resulting in a covalently bound full-length complementary copy of the cDNA fragment that is subjected to several rounds of 'solid-phase' PCR amplification, composed of two basic steps: (**1**) initial priming and extending of the single-stranded, single-molecules template and (**2**) bridge amplification of the immobilized template with immediate adjacent primers to form clusters. When cluster generation is complete, the templates are ready for sequencing.

Cluster Generation by Ion Torrent. Unlike Illumina, the standard Ion Torrent library protocol is strand-specific by default. Moreover, instead of relying on solid-phase PCR amplification, Ion Torrent uses 'emulsion PCR' to prepare fragmented RNA sequencing templates in a cell-free system. First, beads with complementary oligonucleotides are mixed with PCR reagents and a dilute solution of cDNA library and oil added to make an emulsion. Ideally, each micro-droplet of emulsion will contain one bead and one cDNA fragment along with PCR reagents to allow for clonal amplification. Following 16-18 cycles of PCR the emulsion is then broken by organic extraction, beads purified and loaded on to a disposable semiconductor sequencing chip.

2.5.3. Sequencing

Sequencing by Illumina. Illumina employs a sequencing by synthesis technology and utilizes a reversible terminator-based method which detects single bases as they are incorporated into DNA template strands (**Figure 2.5A**). Sequence reagents, including fluorescently labeled nucleotides, are added and the first base is incorporated. The flow cell is imaged and the emission from each cluster is recorded. The emission wavelength and intensity are used to identify the base. This cycle is repeated 'y' times to create a read length of 'z' bases. The result is highly accurate base by-base sequencing.

Sequencing by Ion Torrent. Unlike Illumina's fluorescence-based sequencing by synthesis, Ion Torrent determines sequence identity by detecting pH alterations due to hydrogen ion release following nucleotide incorporation (**Figure 2.5B)**. Since the dNTPs are not differentially labeled by a fluorophore, they must be added successively so that ion release can be associated with a particular nucleotide. However, whereas Illumina utilizes reversible terminator chemistry to restrict dNTP incorporation to once per cycle and sequence through homopolymers, Ion Torrent relies on the number of hydrogen ions released as being proportional to the number of dNTPs incorporated.

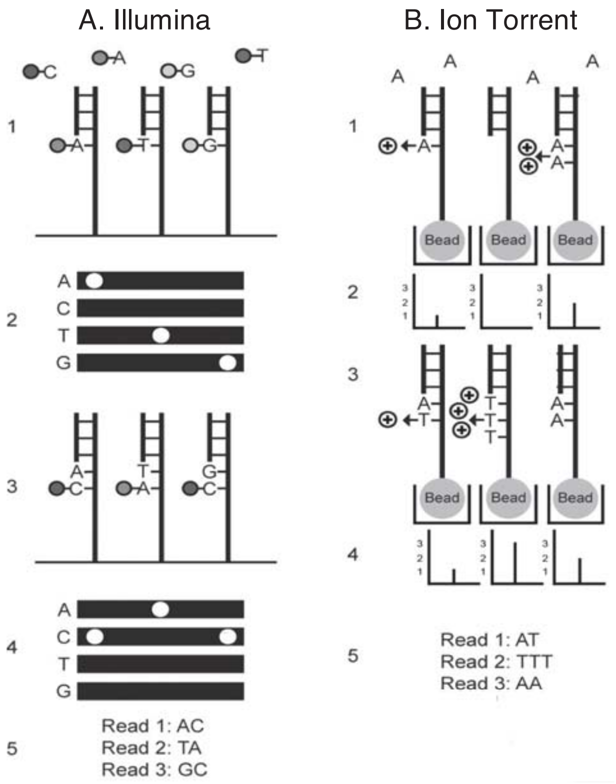


Figure 2.5. Sequence detection methods of Illumina and Ion Torrent. (**A**) Illumina detection is fluorescence-based using reversible terminator dNTPs, resulting in one nucleotide incorporation per cycle. cDNA fragments are covalently linked to a flow cell and fluorescence detected with addition of each nucleotide. (**B**) Ion Torrent sequence by synthesis relies on detection of hydrogen ions for base calling. Each pH detector well contains one clonally amplified cDNA fragments on a micro-bead. Nucleotides are added sequentially.

2.6. RNA-Seq Optimization

2.6.1. Single-End and Paired-End Sequencing

In single-end sequencing, the sequencer reads an RNA fragment from one end of the fragment to the other, generating a sequence of base pairs. In paired-end sequencing, the sequencer starts at one read end, finishes this direction at the specified read length (e.g. 50bp or 100bp), and then starts another round of reading at the opposite end of the fragment (Corney, 2013). Sequencing both ends of the fragmented read is a more efficient use of the cDNA library and having pairs of reads improves read alignment by improving the ability to resolve chromosomal rearrangements such as insertions, deletions and inversions. Paired-end reads can also be particularly useful for the identification of alternatively spliced isoforms and viral integration sites. The information about the expected distance of the reads sequenced from these two ends, is estimated from the distribution of fragment lengths and can be exploited to increase mapping or assembly accuracy.

2.6.2. Sequence Depth and Coverage

In RNA-Seq, sequencing 'depth', or the number of times a transcriptome has been sequenced during the sequencing process, has direct implications for coverage and costs (Wang et al., 2009). Sequence coverage is expected to be a function of the prevalence of the transcript in the sample and of the depth of sequencing. Higher sequence coverage necessitates more sequencing depth, and to identify rare transcripts (or variants) considerable depth is needed. By increasing or decreasing the number of sequencing reads, researchers can tune the sensitivity of an experiment to accommodate various study objectives, also called dynamic range. The dynamic range is adjustable and nearly unlimited, permitting detection of subtle gene expression changes with high sensitivity.

2.6.3. Multiplexing

Multiplexing allows large numbers of libraries to be pooled and sequenced simultaneously during a single sequencing run. Unique index sequences are attached to each cDNA library (i.e. each sample), during the library preparation phase. Subsequently, all libraries are pooled and lowed into the same flow cell lane. Then, libraries are sequences together during a single instrument run. All sequences are exported together. Finally a 'de'-multiplexing algorithm sorts the unique index sequences into different files according to indexes (Corney, 2013).

2.6.4. Strengths and Weaknesses of RNA-Seq

RNA-Seq has shown strong potential to replace microarrays for whole-genome transcriptome profiling. Contrary to microarrays, RNA-Seq: (**1**) is not limited to detecting transcripts that correspond to existing genomic sequences (i.e. unsupervised); (**2**) provides information regarding novel transcripts, splicing events and sequence variations (although bioinformatic pipelines for these analyses need large improvement); (**3**) a low background signal indicates that there is no upper limit for quantification and thus a large dynamic range of expression (i.e. highly quantitative). Importantly (**4**), RNA-Seq has an improved ability to compare transcription levels across different genes, samples, experiments, time points and platforms.

2.7. Cross-Platform Variability

2.7.1. Technical Assessment of Microarrays

A potential problem with microarrays lies in cross-platform variability, and even more so in the variability in analytic tools used by investigators (discussed in Chapter 3). Recent technological advances have attempted to converge techniques to find better correlation between platforms, but even a subtle ~5-10% divergence can lead to differences in ~1,250-2,500 genes between platforms. Linked with this drift between platforms is the major concern of cross-hybridization on microarrays and the relative inability to detect gene expression changes present at low abundance. For example, if Gene A is expressed at 10 copies/cell, and Gene B is expressed at 1000 copies/cell, at 2% cross-hybridization of Gene B to microarray target A, 65% of the observed target A signal will originate from Gene B. In these cases, even if Gene A is differentially expressed, due to the obliteration of the specific signal, the assay may not be sufficiently sensitive to uncover it (Mirnics et al, 2006). This example illustrates the need for validation of microarray results by techniques such as Real-time quantitative PCR (RT-qPCR). However, the ability to detect subtle changes in transcripts of low abundance will also depend on the length of the oligonucleotide probes present on microarray chips. This is a potential consideration in RNA extraction methodology as well as in choosing a platform for analysis, as some, such as Codelink, use longer 60–70mer probes as opposed to Affymetrix who uses shorter 20–30mer probes on their chips.

2.7.2. Technical Assessment of RNA-Sequencing

Studies benchmarking RNA-Seq results using various NGS platforms and utilizing similar RNA isolation and technological optimization steps, indicate a high correlation of gene expression variation generated by the same technologies amongst different laboratory sites (correlation coefficient > 0.86) as well as high correlations across different technologies in laboratory different sites (correlation coefficient > 0.83) (Sheng et al., 2010). Read depth also represents a limiting factor when experimental goals include detection of lowly expressed genes and coverage of introns. This is because standard read depth (i.e. 10X-30X) is insufficiently sensitive to detect differences amongst low abundance transcripts and splice junctions (Sheng et al., 2010). This is somewhat problematic because deep sampling is not currently cost-effective with long-read platforms. Despite this, short-read platforms could circumvent this obstacle as they are able to cover a wider dynamic range and thereby generate more reads per sample.

2.8. Summary

Following venipuncture, a number of RNA isolation and optimization techniques as well as gene expression quantification protocols may be used. In the context of this thesis, RNA isolation from whole blood was performed using LuekoLOCK^{TM} RNA isolation (Chapters 5,6 and 8) and PAXgene RNA tubes (Chapter 7). Gene expression was quantified using RNA-Seq methods Illumina (Chapters 5 and 7) and Ion Torrent (Chapter 8) as well as Microarray methods Affymetrix GeneChips (Chapter 5) and Illumina BeadChips (Chapter 6).

Chapter 3

Statistical Methodologies for Blood Biomarker Discovery

Statistical and methodological clarity and rigor in terms of biomarker discovery, validation and testing are critical steps for realizing blood-based markers in psychiatric disease. Computational work-flows for deriving clinical and biological significance from a blood-based gene expression study are multifaceted (**Figure 3.1**) and commonly seek to address one of three main objectives: (**1**) construction of a gene expression classifier (i.e. a unique panel of cross-validated biomarkers) for disease prediction or diagnosis; (**2**) characterization of molecular factors involved in disease pathology; (**3**) identification of gene networks as functional biomarkers of disease. The following chapter reviews statistical aspects that are relevant in realizing these aims.

3.1. Standard Guidelines and Best Practices for Expression Data

Microarray and RNA-Seq represent core technologies in biomarker discovery but before these technologies can be used reliably in clinical practice and regulatory decision-making, standards and quality measures need to be developed. The MicroArray Quality Control (MAQC) and Sequencing Quality Control (SEQC) projects are helping improve microarray and RNA-Seq technologies and foster their proper applications in discovery and development for FDA rated tests (i.e. accurate blood tests). The MAQC project, focused on microarray technologies, has undergone three main phases (www.fda.gov/MicroArrayQC): MAQC-I (2006) to establish QC metrics and guidelines for data analysis; MAQC-II (2010) to assess the capabilities and limitations of various data analysis methods in developing and validating predictive models and genotyping for personalized medicine; and MAQC-III and SEQC-I (2014) to examine latest tools for measuring gene activity and establish best practice for reproducibility across different technologies and laboratory sites. It is important to emphasize that RNA-Seq is not a ‘mature’ technology but is undergoing rapid evolution in biochemistry of sample preparation, of sequencing platforms of multiple RNA-Seq computational pipelines and of subsequent analysis methods that include statistical treatments and transcript model building, evaluated here. Collectively, these points make standard guidelines and best practices for RNA-Seq data complex.

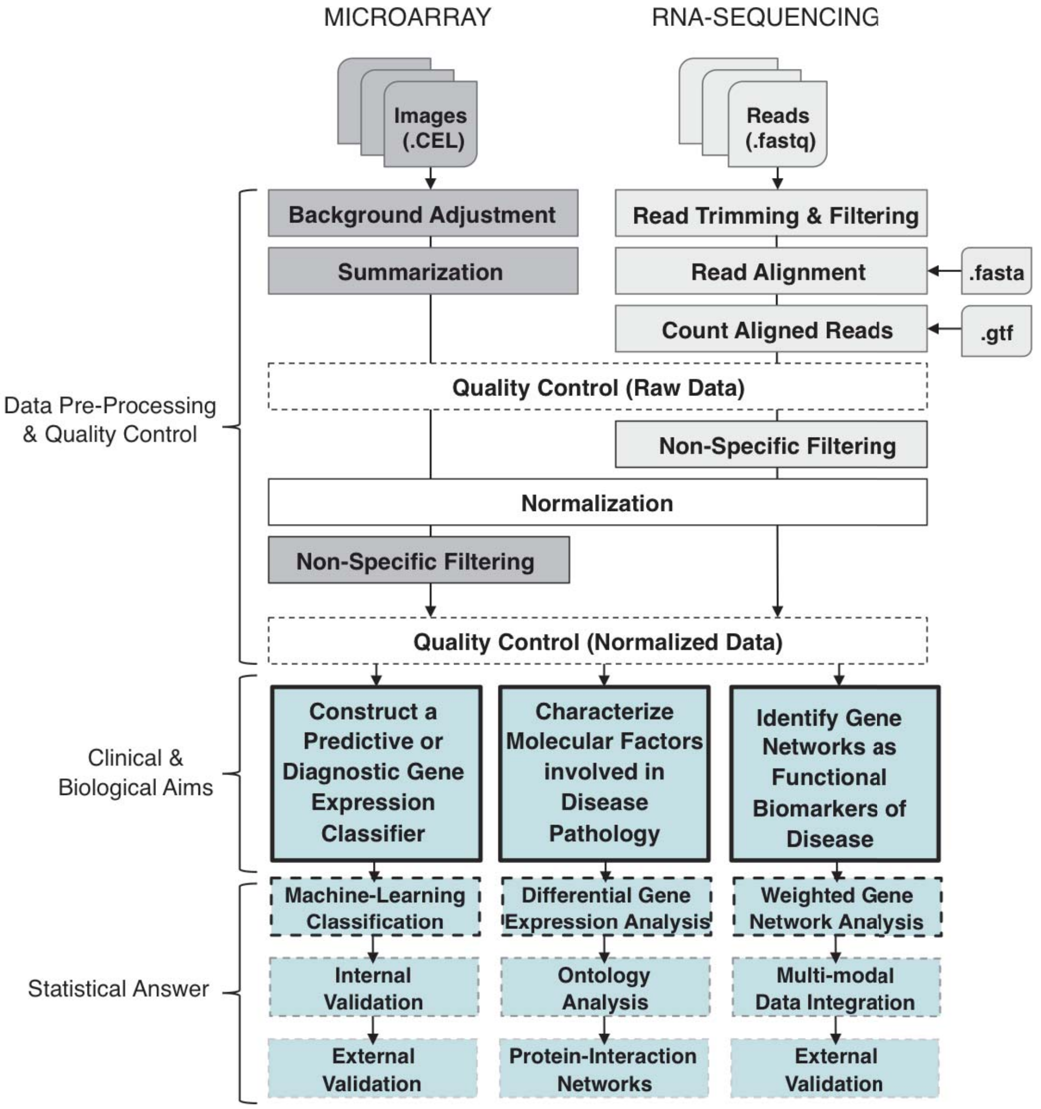

MICROARRAY
RNA-SEQUENCING
Images (.CEL)
Reads (.fastq)
Background Adjustment
Summarization
Read Trimming & Filtering
Read Alignment
.fasta
Count Aligned Reads
.gtf
Quality Control (Raw Data)
Non-Specific Filtering
Normalization
Non-Specific Filtering
Quality Control (Normalized Data)
Data Pre-Processing & Quality Control
Clinical & Biological Aims
Construct a Predictive or Diagnostic Gene Expression Classifier
Characterize Molecular Factors involved in Disease Pathology
Identify Gene Networks as Functional Biomarkers of Disease
Statistical Answer
Machine-Learning Classification
Differential Gene Expression Analysis
Weighted Gene Network Analysis
Internal Validation
Ontology Analysis
Multi-modal Data Integration
External Validation
Protein-Interaction Networks
External Validation

Figure 3.1. A workflow of data pre-processing, normalization, non-specific filtering, quality control and downstream data analysis options for microarray and RNA-Seq gene expression. Abbreviations; .CEL, raw microarray data; .fastq, raw RNA-Seq data; .fasta, human genome assembly; .gtf, human genome annotations.

3.2. Microarray Data Pre-Processing

After the image is taken, probe intensities are generally used as raw data for microarrays. However, these data contain ‘noise’ from many sources and initial data pre-processing and quality control is compulsory.

3.2.1. Microarray Background Adjustment

The desired reading on each probe is the amount of fluorescence from molecules that are complementary to the probe (i.e. the intended RNA target). However, the hybridization sample consists of a mixture of nucleotide molecules, and non-complementary sequences also bind to the probes. This phenomenon is referred to as *non-specific binding* and is a major reason for background noise in microarray data (Olson, 2006). Optical noise is another source of background noise, but is usually smaller than non-specific binding and appears not to be probe specific. As a result, the observed probe intensity is a sum of the above components. The relative quantity of the target RNA across samples can be seriously biased if background is not accounted for.

One way to tackle background noise when using Affymetrix microarray is by using perfect match (PM) and mismatch (MM) probes (**Figure 3.2**) (Olson, 2006). The PM probe has a sequence exactly complimentary to the particular gene and thus measures the expression of the gene. The MM probe differs from the perfect match probe by a single base substitution at the center base position, disturbing the binding of the target gene transcript. This helps to determine the background and non-specific hybridization that contributes to the signal measured for the PM oligo. These direct measurements are subtracted from the PM intensities to adjust for the additive background in several generations of preprocessing methods provided by Affymetrix. Other approaches, which do not include such direct measurement, involve more complicated statistical analysis using an empirical Bayes approach to borrow information across probes on the same array. This step shrinks the background estimate for either the entire sample or probes sharing a similar sequence structures (Silver, 2009).

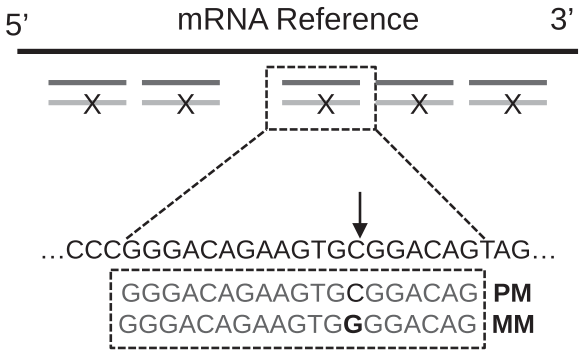


Figure 3.2. For Affymetrix GeneChips, each gene is represented on the array by two different probes – a perfect match (PM) and mismatch (MM) probe pair. These are used to help determine the extent of background noise on the array prior to down-stream analysis.

3.2.2. Microarray Summarization

For some platforms multiple probes are used to quantify the same genomic target. For example, Affymetrix GeneChips use a set of 11-20 probes to measure expression levels of a gene and on average 4 probes for an exon. Illumina arrays use one probe for each gene but include technical replicates (approximately 30) of the same probe. After preprocessing, a summary of these multiple probe-level measurements are combined into individual gene expression values (i.e. mean or median expression) (Herber, 2013). This reduces the number of expression measurements across many probes, and multiple comparisons which differential expression will deal with in further analysis.

3.2.3. Microarray Normalization

In addition to background noise, other sources of *technical variation* can affect the observed intensities that are not themselves of biological interest. For example, arrays on one scanner could in general give higher readings than those from another scanner. Degree of similarity, array scanning dates, changes in RNA isolation reagents and calibration of equipment can all induce technical variation and should be removed through normalization. Most normalization methods equalize some summary statistics of the distribution of measurements across arrays. The simplest ones, such as MAS 5.0, scale the arrays so that each array has the same mean or median intensity (Affymetrix, 2002): this scaling normalization implicitly assumes that biological variations of interest may affect a number of measurements but should not change the mean or mode of the distribution of intensities on each array. Since non-linear relationships between arrays are common, normalization methods that use a non-linear smooth curve have also been introduced. Using a baseline array, a smooth normalization curve can be estimated from the scatter plot of two arrays (Li & Wong, 2001). Without it, one could use all arrays available in a dataset, and iterate over pairwise combinations of arrays so that all arrays are normalized to an 'average' array. The *cyclic loess normalization* (Dudoit et al., 2002) is a good example of this approach.

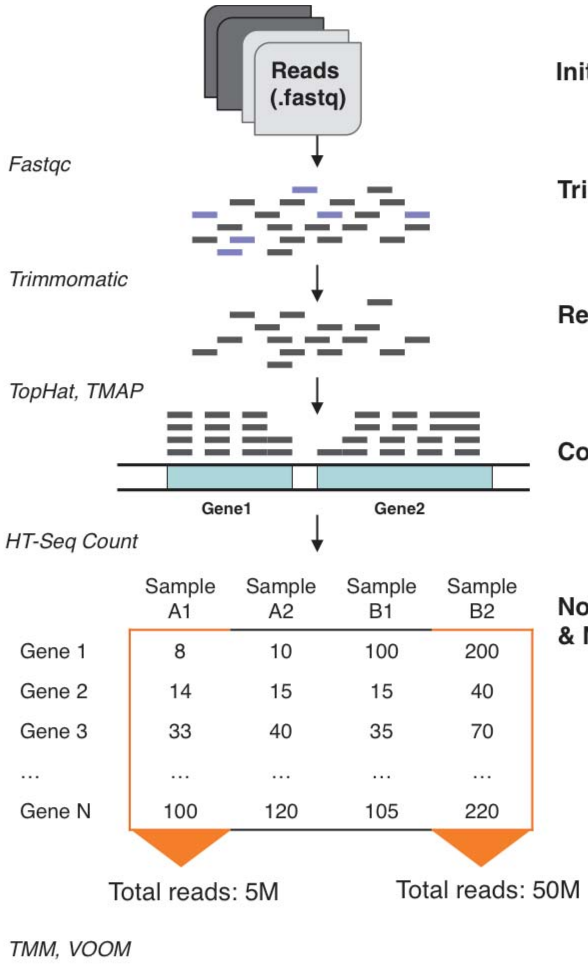
The most popular and well-established normalization methods seek *quantile normalization* ***(***Amaratunga et al., 2001***)***, which makes all arrays have the same empirical distribution of intensities after normalization. A baseline array can be used for the reference distribution, or all arrays used to generate an average distribution for reference. Quantile normalization is used in robust multichip average (RMA) (Bolstad, 2002) and GC robust multichip average (GCRMA) normalization (Wu, 2014). RMA normalization performs background adjustment, quantile normalization and summarization of these data all in one. The GCRMA function uses the same normalization and summarization methods as RMA however utilizing information obtained from the MM probes to estimate probe affinity to non-specific binding. Because GC-rich probes seem to have higher non-specific signal, GCRMA models GC content over the probes which dictate the binding affinity to target.

3.2.4. Microarray Non-Specific Filtering

Following normalization, the next step of the microarray data pre-process, before statistical analysis, is the non-specific filtering of probe intensities. With tens of thousands of genes represented on an array, and with one or more hypotheses being tested for each gene, a multiple testing adjustment is certainly warranted. The aim of non-specific filtering is to reduce the number of multiple comparisons by filtering out lowly expressed probes and probes with low variation which are assumed to be unable to achieve statistical significance in down-stream analyses. Two common hard-threshold filtering methods are routinely applied to the global gene expression picture while ignoring group labels (Hackstadt & Hess, 2009). (**1**) Filtering by coefficient of variation (i.e. relative standard deviation) which is the removal of genes with low variance across all available samples. The rationale is that expression for equally expressed genes should not differ greatly between treatment groups, hence leading to small overall variance. (**2**) Filtering by average expression which is the removal of genes with low average expression across all available samples.

3.3. RNA-Seq Data Pre-Processing

RNA-Seq pre-processing involves several computational steps not found in microarray data handling (**Figure 3.3**), including; (1) determining the quality of raw RNA-Seq reads, (2) trimming and filtering reads, (3) read alignment to a reference genome, and (4) counting mapped reads to determine overall measure of gene expression.



	Sample A1	Sample A2	Sample B1	Sample B2
Gene 1	8	10	100	200
Gene 2	14	15	15	40
Gene 3	33	40	35	70
…	…	…	…	…
Gene N	100	120	105	220

Figure 3.3. Computational pipeline for RNA-Seq data pre-processing. All steps (right) and tools (left) for generating a matrix of RNA-Seq gene expression measurements. Fastqc is a quality assessment tool for RNA-Seq raw .fastq files; trimmomatic is a tool for filtering poor quality reads and read trimming; TopHat and TMAP are tools for mapping fragmented reads to the transcriptome; HT-Seq is a tool for counting (quanitifying) expression of genes mapped to the genome; TMM and VOOM are normalization tools for removing differences in library effect sizes. All citations found below. Abbreviations: QC, quality control; TMM, trimmed median values; M, million.

3.3.1. Quality Check of Raw RNA-Seq Reads

RNA-Seq reads and the corresponding base call qualities (i.e. the probability of a correct call) are typically delivered to the user as a FASTQ file (extension .fastq or .fq). FASTQ files contain a four-line record for each read, including its nucleotide sequence, a "+" sign separator (optionally with the read identifier repeated), and a corresponding ASCII string of quality characters. Each ASCII character corresponds to an integer *i ranging* from -5 to 41 (depending on the version of software used for base-calling), and may be translated to *p*, the probability that a given base is incorrectly called, using the Phred scale (Formula 3.1).

$$i = -10 \times \log_{10}(p)$$

(Formula 3.1)

Base calling error rate is highest during the final cycles of sequencing and it is not uncommon for per base quality score to be low. The FastQC command-line tool (Andrews, 2010), and similar tools, use the raw sequences provided in FASTQ format and display basic statistics to allow the quick evaluation of whether sequences are as expected. Outputted parameters include number of reads and GC percentage, per base sequence quality score (a measure of confidence of correct base calling), per base sequence content (a representation of each nucleotide at each base position to visualize position/sequence bias), per base N content (a plot of uncalled nucleotides at each base position), duplicate reads (typically a result of PCR over-amplification during library preparation) and overrepresented sequences and K-mers. It is important to evaluate the report in the context of the anticipated results, since QC programmes assume sequencing of a random and diverse library, which may not be the case depending on experimental design and library preparation.

3.3.2. RNA-Seq Read Trimming and Filtering Low Quality Reads

Following an initial quality check, the first step of RNA-Seq data pre-processing is read trimming and filtering. During the sequencing process, considerable amounts of RNA-Seq reads are generated, a process which typically encompasses some errors. Artefact sequences and low quality reads make up a minority proportion of sequences in the

FASTQ file and including these in the mapping stage will introduce mapping errors and in some instances creates artificial indels. An easy way to improve mapping of sequenced reads is by confirming the sequence quality of raw reads by evaluating base quality, the GC content distribution and the duplication rate (Guo 2013, Patel and Jain 2012). Many end users also elect to discard reads suspected to contain sequencing errors or low quality Phred scores. Reads likely to contain multiple sequencing errors provide less biological information and are expected to hinder alignment. Reads generated using Illumina Stranded Kits are known to have a higher error rate towards the 3′-end of the read, so if a reduction in quality is detected within the read, it is normal to trim off the rest of the read. Absolute minimum, average, and sliding-window-average quality scores are commonly used as criteria for discarding and/or trimming reads. These steps are achievable through use of command line tools Trimmomatic (Bolger et al., 2014)and Fastx-Toolkit (Patel & Jain, 2012).

3.3.3. RNA-Seq Read Alignment to Reference Genome

Following read trimming and filtering, the second step of RNA-Seq data pre-processing is the alignment (mapping) of high quality trimmed short reads to a reference genome. Read alignment to a reference genome generates a dictionary of the genomic features represented in each RNA-Seq sample. That is, aligned reads become annotated, and the highly fragmented data is thus connected to the gene families, individual transcripts, small RNAs, or individual exons encompassed by the original tissue sample (i.e. blood). Read mapping generally requires two basic steps: the reference genome (.fasta) is first converted to an *indexed reference* to allow fast read mapping, which is the second step. There are two commonly used annotated backbone reference genomes for *homo sapiens*; hg19 (also known as GRCh37) and hg38 (also known as CRCh38) (Team TBD) annotation provided by UCSC genome. The latest release (hg38) provides a comprehensive gene annotation for all protein-coding transcript sequences as well as lncRNAs and pseudo-genes.

Features of the reference genome such as repetitive regions, assembly errors, and missing information can render alignment impossible (un-mappable) for a subset of the newly generated reads (< 2%). Likewise, sequencing errors, polymorphisms, and limited complexity within the short reads can also act as obstacles. Alignment algorithms must therefore be flexible and allow for *approximate* matches when applying mapping criteria. Just how approximate the matches are depend on the features of the aligner, which is generally defined through user-specified parameters. For example, the number of allowed mis-matches and minimal score of match obtained from the RNA target to reference genome can all have an effect on mapping. Similarly, treatment of multiple mapping reads and establishing a cap on the number of genomic coordinates to which multiple reads were distributed (e.g. only loci < 10 reads are reported) also has an effect. For 'split-reads', establishing constraints regarding the location of the splits (i.e. within the same chromosome; within a certain genomic interval) and regarding the sequences at the split (allowed only at the conical junctions etc.) is also an important parameter for mapping. Indicating paired-end sequences and stranded information, when available, aids the alignment process. The powerful TopHat and TMAP (designed for Ion Torrent reads) short read mapping algorithms require these pieces of information for accurate and precise mapping (Trapnell et al., 2009; Caboche et al., 2014). Both tools output aligned reads in the format of a .bam file (binary form of .sam), which generally need converting to a .sam file prior to counting mapped reads. Alignment parameter settings that are extremely stringent will result in only a small subset of reads being mapped; whereas liberal settings will result in lost specificity, and many reads will map to multiple features of the reference. It therefore takes some experimentation to achieve the optimal balance of sensitivity and specificity for a given data set.

3.3.4. RNA-Seq Counting Aligned Reads to Measure Gene Expression

After mapping, the third step of RNA-Seq data pre-processing is counting the number of reads that have been aligned (i.e. mapped) uniquely to each genomic coordinate – exon, transcript or gene level. Given a sorted SAM file with aligned sequencing reads and a list of genomic features (.gtf file), a common task is to count how many reads map to each feature. Here, a feature can be an interval (i.e. range of positions) on a chromosome or the union of such intervals. The most used approach for computing counts considers the total number of reads overlapping the exons of a gene, in which features are typically genes, where each gene is considered the union of all its exons. However, even in well-annotated organisms, a fraction of reads map outside the boundaries of known exons (Pickrell et al., 2010). Thus, an alternative strategy considers the whole length of a gene, also counting reads from introns. Moreover, if correctly handled in the mapping phase, spliced reads can be used to model the abundance of different splicing isoforms of a gene (Trapnell et al., 2010). Particular attention should be paid to genes with overlapping sequences. HT-Seq count, which is a user-friendly Python package (Anders et al., 2014), implements a flexible approach permitting the user to select the direct model for read counting in the presence of overlapping features for both Illumina and Ion Torrent aligned reads.

3.3.5. RNA-Seq Non-Specific Filtering

Unlike microarray data, RNA-Seq gene expression data undergo non-specific filtering prior-to normalization. This is because RNA-Seq read counts represent a quantitative measure of gene expression and some genes may not contain any counts due to the level of sequencing depth and breath of coverage. As a result, the inclusion of hundreds of gene measurements with very low to absent gene expression values would skew normalization and down-stream analysis. Similar filtering strategies by the coefficient of variation and average expression can be applied to RNA-Seq data. Additionally, to take advantage of missing gene expression calls, an alternative filtering strategy for RNA-Seq data include removal of lowly expressed genes using a combined count and sample threshold (i.e. any genes < 20 counts in at least a third of the samples).

3.3.6. RNA-Seq Normalization

Similar to microarray, RNA-Seq data require normalization to remove unwanted non-biological variation between samples. The first bias to be taken into account is variation in sequencing depth across samples, here defined as the total number of mapped reads. Consider a hypothetical sample A and B, two RNA-Seq experiments with no differentially regulated genes. If experiment A generates ten times as many reads as experiment B (as in **Figure 3.3**), it is likely that the counts from experiment A will also be doubled. Hence, a common practice is that of scaling counts in each experiment *j* by the sequencing depth d_j estimated for that sample, where d_j was computed by counting the total number of reads mapped in sample j (i.e. global scaling) (Marioni et al., 2008). Other applications consider counts depending on the whole RNA population of the sequenced sample. If there is a set of highly expressed genes in a sample, they will inevitably 'consume' the total sequenced reads, and the expression level of the remaining genes will be largely underestimated. A similar issue may result from the presence of contaminates. Bullard et al (2010) suggest a quantile normalize similar to that used for microarray data, and an alternative global scaling that adjusts counts distributions with respect to their third quartile, so to reduce the effects of genes with high-counts. Another popular approach is the proposed Trimmed Mean of M-values (TMM) normalization to account for differences in library composition between samples (Robinson & Oshlack, 2010). To reduce bias due to high-count genes, TMM is computed removing the 30% of genes that are characters by the most extreme M-value (i.e. log-fold changes) for the compared samples. This normalization factor is then used to correct for differences in library sizes. VOOM normalization, a variance stabilization transformation method has also gained popularity in its ability to transform negative binomial distributed data into a normal distribution (Anders et al., 2014). Interestingly, studies benchmarking the effects of normalization approaches have demonstrated the potential for different approaches to produce varying end results. In the context of this thesis, VOOM normalization was used as the appropriate normalization method.

RNA-Seq counts also show a gene length bias: the expected number of reads mapped to a gene is proportional to both the abundance and length of the isoforms transcribed from that gene. Longer genes produce more reads than shorter ones, resulting in higher power for detection. To tackle this problem RPKM (reads per kilobase of transcript per million mapped reads) and FPKM (fragments per kilobase of transcript per million mapped reads) are widely used normalization metrics, the latter often used for paired-end sequencing (Mortazavi, 2008). However, most recent comparative normalization approaches have demonstrated that RPKM normalization may not be the most appropriate normalization technique due its simplistic nature to normalize read counts by gene length and the total number of mapped reads in the sample (Dillies, 2013). It is advisable to test a number of different normalization techniques and continue them through down stream analyses in order to best assess the normalization approach that is best suited for an independent study.

3.4. Microarray and RNA-Seq Quality Control

The identification of outlying samples can bias down-stream analysis and therefore should be discarded using biological and statistical reasoning. The search for outlying samples should occur before and after data normalization, as the inclusion of outlying samples may heavily bias normalization. For these purposes, visualization of the data in every way, shape and form is critical (**Figure 3.4**). Qualitative visualization plots such as boxplots and histograms of raw and normalized data are routine. MA-plots are used to plot of the distribution of the red/green intensity ratio (‘M’) plotted by the average intensity (‘A’), which attempt to visualize the intra and inter array spread of the data on a sample-to-sample basis. The underlying assumption is that most genes are not differentially expressed and therefore the majority of M values (i.e. difference of log-ratios) should be located at 0. Relative log expression (RLE) and normalized un-scaled standard error (NUSE) plots are also useful for identifying outlying samples in a similar fashion. Principal Component Analysis (PCA) works to reduce the dimensionality of the thousands of gene expression measurements into two or three main components using a linear scale. Hierarchical clustering dendrograms (Euclidean and Pearson’s correlation coefficients) are also ways to visualize pairwise distance between samples. A heuristic for these approaches is that if the median PCA or hierarchal clustering coefficient of a sample is beyond a given number of standard deviations from the average, it may be removed from the analysis. Another point of interest focuses on non-biological experimental variation, or batch effects, which are commonly observed across multiple batches of microarray or RNA-Seq data, which makes combining batches a difficult task. Combat is a method whereby a flexible empirical Bayes framework is used to adjust for additive, multiplicative and exponential batch effects (or standardize across expression measurements) and is also robust to outliers in small and large sample size (Johnson, 2006).

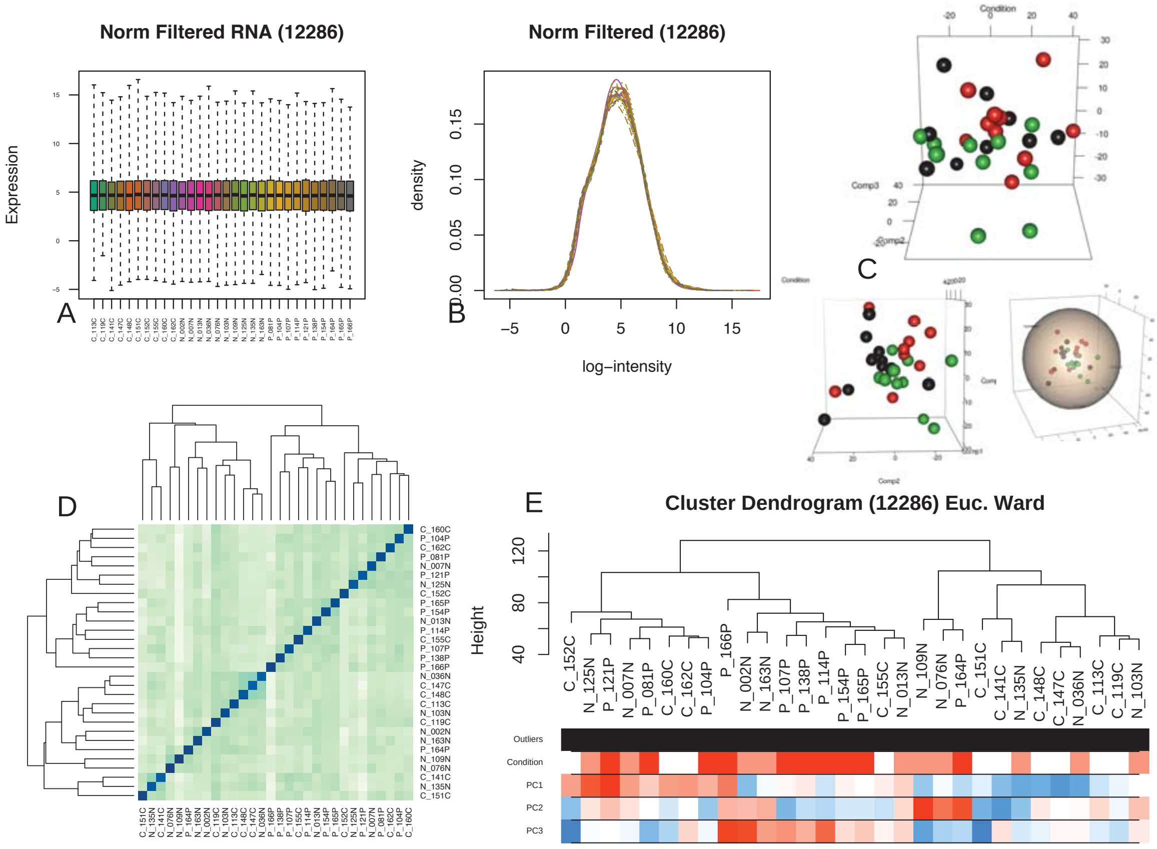


Figure 3.4. Quality control methods and procedures. (A) Normalized boxplots for 30 samples where the y-axis represents normalized expression values and the x-axis represents each sample. (B) Histograms of the same 30 samples where the x-axis represents normalized expression and the y-axis represents density (or frequency). (C) PCA analysis colored to delineate three experimental groups being tested in this example rotated for complete visualization and investigated by placing 2 standard deviations around the centre of all samples. (D) Correlation matrix plot and (E) hierarchical clustering to demonstrate pairwise sample-to-sample relationships using Euclidean (Euc.) coefficients and wards distance metric. Here, 12286 reflect the number of genes post-filtering.

3.5. Data Analysis: Transcriptome Exploration from Low- to High-Order

Proper data pre-processing and quality control should be done prior to statistical analyses. The aims of gene expression studies vary from experiment to experiment but can broadly be categorized in terms of transcriptomic exploration, i.e. the amount of transcriptomic activity that is characterised and defined by the data analysis. Broadly speaking, there are three generally aims when it comes to analysing blood transcriptome data, each with varying degrees of transcriptomic exploration. The first, and smallest order of transcriptomic exploration, is *supervised machine-learning* (ML). ML is a powerful tool for constructing gene expression classifiers for disease prediction and diagnosis. While this approach is most feasible for the eventual development of a blood-based test, ML 'cares' little about the higher-order functionality of the features which it uses for prediction as long as they perform well across experimental groups. The next highest level of transcriptome exploration is *differential expression analysis*. This approach has little regard for gene-gene relationships which is a considerable pitfall when attempting to model complex clinical phenotypes using a heterogonous tissue source. Nonetheless, this approach is often able to provide more mechanistic insights than ML approaches by mapping genes onto gene ontology categories and protein-interaction networks. The highest order of transcriptomic exploration is that provided by network-based approaches. *Weighted gene co-expression network analysis* (WGCNA) is a correlation network based approach suitable for modeling complex systems (i.e. blood transcriptome) and phenotypes (i.e. psychiatric diseases). Moreover, it allows for multi-modal data integration interfacing with differentially expressed genes, protein-interaction information, clinical traits and other sources of high-throughput biology (e.g. proteomics, methylation etc…). The following section presents a detailed account of these three broad statistical applications, based on transcriptomic exploration (low-to-high order), to give a better understanding of how they may be used in the context of psychiatric biomarker discovery.

3.6. Supervised Machine-Learning and Classifier Construction

In a disease context, supervised machine-learning (ML) is concerned with identifying a small panel of biomarkers with maximum accuracy able to either '*predict'* disease status or treatment outcomes (i.e. prognostic classifier), or to '*diagnose'* a disease status (i.e. diagnostic classifier). BRB-Array Tools (Simon et al., 2007) is a useful package for the construction of gene expression classifiers and uses methods of *class prediction*: a supervised approach that incorporates the sample labels to identify the genes whose expression can be used to predict which group of a blinded sample belongs to. ML employs numerous statistical and optimization techniques permitting algorithms to 'learn' based on presented data from past examples, and to detect hard-to-discern patterns from large complex data-sets, such as transcriptomic data. ML must first 'learn' the data using information (i.e. expression measurements) presented from a training set to develop a classifier, and subsequently evaluate classifier accuracy to predict blinded samples on a withheld test set. ML applications consist of four essential steps:

*(**1**) Data filtering*: ML is used to identify the most important features in a gene expression data-set that are able to accurately discriminate between two or more experimental groups. It is often a good first step to subset the data by removing less informative features and retaining the more informative ones using *p*-value, log-fold change or standard deviation thresholds. However, the eventual predictor may be less biologically interpretable and clinically applicable, if fewer genes are included. The main aim of this step is to cast a wide net to gather all informative genes while false-positives will be pared off in subsequent feature selection and ML optimization steps.

*(**2**) Feature Selection:* The criteria for feature selection are based on the concept that predictive accuracy is important, and not necessarily the statistical significance of features. Feature selection is performed to remove 'noise' features via filtering based on gene rank accordingly to differential expression, thereby reducing thousands of genes down to hundreds. Recursive feature elimination (RFE) (Simon et al., 2007) provides a more statistically involved filter where genes are assigned weights and the number of genes desired within a classifier may be selected *a priori* to evaluate predictive accuracies across classifiers containing different numbers of features (i.e. accuracy of a classifier including 20 genes versus one including 50 genes). This is a powerful approach for identifying the smallest number of features needed to make accurate predictions.

*(**3**) Select Machine-Learning Algorithm*: The utility of RFE can be assessed by numerous multivariate classification methods including support vector machines (SVM), diagonal linear discriminate analysis (DLDA), nearest centroid (NC) and three-nearest neighbors (3-NN). Additional algorithms that excel in multi-class comparisons (2 or > groups) include decision trees, random forest and neural networks. SVM represents the 'start of the art' approach for gene expression prediction problems. The simplest type of

SVM is linear classification, which tries to draw a straight lines (also called hyperplanes) that separates data with two dimensions (**Figure 3.5**). Many linear classifiers are able to separate data but only one achieves maximum separation. Vapnik and Lerner (1963) proposed a linear classifier as an original optimal hyperplane algorithm. The replacement of dot product by a non-linear kernel function allows the algorithm to fit the maximum-margin hyperplane in the transformed feature space. SVM finds a linear separating hyperplane with the maximal margin in this higher dimensional space.

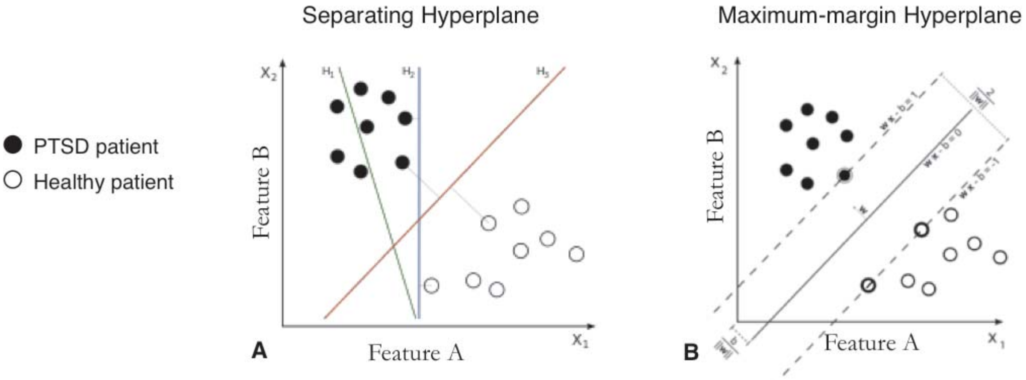


Figure 3.5. Example of linearly separable data:. (A) Three possible hyperplanes to linearly separate data, (B) Centered line demonstrates the optimal hyperplane with maximum margin from two classes.

(4) Estimating Prediction Accuracy: To evaluate classifier performance, a *2x2* contingency table needs to be populated so that the numbers of true positives (TP), true negatives (TN), and false positives (FP), and false negatives (FN) may be estimated in order to calculate classifier accuracy, sensitivity and specificity (**Figure 3.6**). There are various internal cross-validation methods for estimating prediction accuracy, and these are arguably the most important aspects of ML. Internal cross-validation can be divided into two categories; split-sample (i.e. hold-out) validation and leave-one-out cross-validation (LOOCV) (Simon et al., 2007). Split-sample validation (**Figure 3.6A**) splits the data into a training set where ranked features are selected, an ML algorithm is chosen and parameters and cut-off thresholds are determined. Here, the test set is withheld until a single clinical classifier is fully specified using the training set and then applied to samples in the test set to predict group status. LOOCV (**Figure 3.6B**) divides data into training and test sets where the test set contains only one withheld sample and a classifier is constructed on the training and predicts the class of the left out sample. This

process is repeated until every sample has been left out at least once and the predictions of each sample are used to populate a *2x2* contingency table. Alternatively, this process can leave a tenth or a fifth out (10-fold, 5-fold) for cross-validating classifier accuracies.

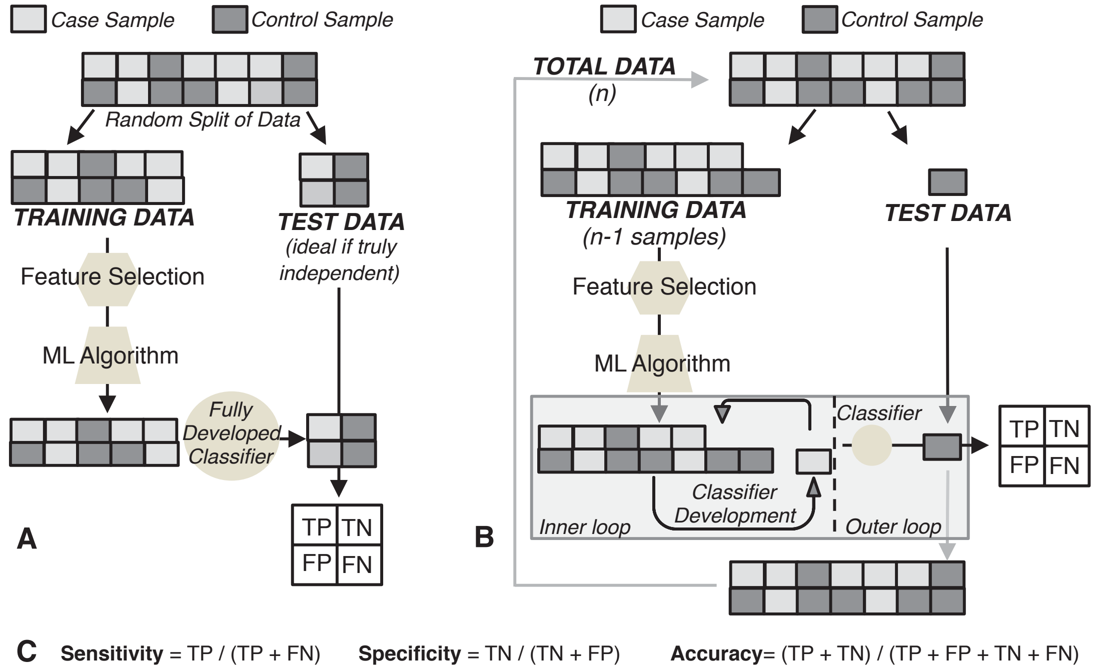


Figure 3.6. Cross-validating prediction accuracies of gene expression classifiers. (A) Split-sample classification randomly splits data into training and test sets. (B) LOOCV implemented with nested cross-validation where the inner-loop evaluates the efficacy of the classifier on each sample and the outer loop tests the classifier on the withheld sample. Feature selection and the ML algorithm are cross-validated in each loop. (C) 2x2 contingency tables containing the true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) are used to assess the true positive rates, (sensitivity) true negative rates (specificity) and final classifier accuracy.

3.6.1. Statistical and Translational Implications for Gene Classifiers

The following statistical and translational observations for implementing ML and applying the results into clinical settings may be useful. First, the fit of a model to the same data used to develop a classifier is not evidence of prediction accuracy for independent data. In other words, cross-validation is only valid if the test set is not used in the development of the model. Using the complete set of samples to select features violates this assumption and so invalidates cross-validation. In fact, an ideal test set is a truly separate, independent cohort of patients. Second, 'over-fitting' occurs when a ML algorithm is trained on too few features and becomes specific to the training set resulting in poor classification on independent test data. Alternatively, as the number of features increases, more training data is required to ensure there are enough training instances with each combination of the feature values. Third, for small sample sizes (n < 50) LOOCV is less biased than split-sample whereas for moderate sample sizes 10-fold or split-sample validation is preferred. In either case, nested cross-validation should be considered as this avoids optimistically biased prediction accuracies by further dividing training data into an outer-loop (optimization of feature selection) and an inner-loop (specificity of the model is fit prior to applying to the withheld test set) (Cawley and Talbot, 2010). Fourth, longitudinal experimental designs are critical in fully elucidating mechanistic trends as well as in the prediction of disease onset, relapse or response to treatment strategies. Finally, ML is confounded by the heterogeneity of a biological tissue whereby assuming that all genes within a biological system (blood transcriptome) are independent from each other and identically distributed. To overcome tissue complexity, various statistical approaches have been recommended, focusing on integrating physical protein-protein interaction information or summarizing gene-sets to predict pathway level to obtain a higher-order understanding of the relationships between and amongst the final biomarker panel.

3.7. Differential Gene Expression Analysis

Standard differential gene expression analysis, or *class comparison*, is one of the most widely used approaches to gain a snapshot of transcriptional activity across two or more biological conditions. The aim of differential gene expression analysis is to quantify the relative change of each gene between two or more groups with a *p*-value, to adjust these *p*-values for multiple comparisons using a false-discovery rate (FDR) and subsequently to choose an appropriate *cut-off* to create a candidate list of differentially expressed genes. To elaborate, given that microarrays and RNA-Seq often are comparing the expression of thousands of genes at once, some genes may exhibit significant differences in gene expression by chance. This is called the FDR (Benjamin and Hochberg 1995) and may result through experimental bias, testing of multiple hypotheses, inadequate sample-size, and improper use of statistical analyses (Pawitan et al. 2005; Broadhurst and Kell 2006). Therefore, a cut-off is generally established based on overall change threshold (i.e. fold-change) and FDR adjusted *p*-value.

It is important to use both biological and statistical reasoning when defining cut-offs based on candidate lists of differentially expressed genes. For example, suppose that one selects a significance level of *p*-value < 0.001 for including genes in the gene list. If there are 8000 normalized genes in the experiment, then by chance 8 genes on the gene list could be false positives. If one obtains 24 genes on the gene list, then about one-third of them may be false positives. If one obtain 80 genes on the gene list, then about one-tenth of them may be false positives, and so on. Traditional multiple comparison corrections used in statistical analyses are more stringent, usually requiring that the chance of any false positives be very small, but for most gene expression studies, such conservatism is not appropriate. However, gene lists will not be a useful basis for further experimentation if they are heavily populated by false positives. The presence of false positives also makes interpretation or planning confirmatory experiments very problematic. Using a biological and statistical justified *p*-value criteria is important. There are many differential expression methodologies for modeling RNA-Seq and microarray data, based on differing conceptual positions. In their most basic form, moderated *t*-tests (i.e. *limma*, *edgeR* and *DESeq*) and permutation-tests (i.e. *SAM*, SAMSeq, PoissonSeq) are able to compare gene expression levels with sufficient data between two or more groups.

3.7.1. Moderated *t* test

The core component of a moderated *t* test is the ability to fit gene-wise linear models to gene expression measurement to assess differential expression. As in *limma* (suitable for RNA-Seq and microarray), *p*-values are computed by adjusting the standard error through employing an empirical Bayes framework to borrow information across all genes to improve inference on any single gene and by modifying degrees of freedom, adjusting a term that represents the *a priori* number of degrees of freedom for the model (Ritchie et al., 2015). The method is sufficiently flexible to fit almost any experimental design: including experiments with two or more groups, factorial and time-course designs. Where appropriate, ‘nuisance variables’ such as batch and dye effects can also be modeled as continuous covariates within the linear regression. Once a linear model is fitted, forming a contrast matrix allows for making a series of comparisons between groups. The fitted model object and contrast matrix are used to compute $\log_2$-fold-changes and *t* statistics for the comparisons of interest, allowing all possible pairwise comparisons between treatments to be made (Ritchie et al., 2015).

EdgeR (Robinson, 2009) is a closely analogous procedure to *limma*, but is mathematically more complex, takes RNA-Seq count data as input and then employs an overdispersed Negative Binomial model (since count data follows a Poisson / Negative Binomal distribution), before using an empirical Bayes procedure similar to *limma* to moderate the degree of 'overdispersions' across genes. To clarify what overdispersions mean, a negative binomial model is for RNA-Seq data because there is typically more variation in these data than can be accounted for by using a Poisson model, and this is termed overdispersion. EdgeR estimates the mean (mu) of the counts for each gene, which corresponds to the abundance of that gene in the RNA sample. EdgeR models the mean for a gene as the (library size × concentration). This model has a second parameter, called the dispersion parameter. This parameter is very important, as it determines how to model the variance for each gene using a variance function (Formula 3.2), where each gene has a distinct value for mu. Under the *common dispersion model* the *same* value for the dispersion is used when modeling the variance for each gene. Under a *tagwise dispersion model* a different value for the dispersion is used for each gene (Robinson, 2009).

$$V = mu\ (1 + dispersion \times mu)$$

(Formula 3.2)

3.7.2. Permutation-based Testing

Multivariate permutation-tests are another powerful approach for differential gene expression and are based on random permutation, or re-sampling, of the group labels (Ritchie et al., 2015). For each permutation, either parametric or nonparametric tests are averaged over several re-samplings of the data using sample permutation strategy to estimate a FDR, desirable for moderately sized two group comparisons. Yet, permutation testing has drawbacks: for example, when applied to small sample sizes ($N \leq 7$), permutation testing often results in low power to detect differences and is also inappropriate for multi-level experimental designs (i.e. > 2 group comparisons). Permutation also assumes that all samples are independent and identically distributed, which is perhaps unrealistic. Moreover, when samples are correlated, permutation testing may be misleading. Nonetheless, both statistical approaches, (moderated *t* tests and permutation testing) have been criticized for ignoring biological knowledge regarding the higher-order relationships of how genes work together.

3.7.3. Functional Annotation and Enrichment Analyses

Gene-set enrichment proposes to incorporate biological insight to a list of candidate genes by utilizing *a priori* defined functional gene-sets. Gene-sets are formed by grouping genes that are parts of the same cellular components, biological processes or molecular factors (i.e. GO terms), or more important biological pathways (i.e KEGG, Reactome). Recent advances also permit the grouping of genes that are coordinately expressed in specific cell types or those which have been linked to therapeutic drug treatments. There are two major techniques for performing gene-set enrichment (**Fig 3.8**).

Over-representation gene-set analyses are traditionally based on calculating the over-representation of candidate genes (i.e. differentially expressed genes) within a list of genes assigned to a particular functional vocabulary, such as a GO-term or KEGG pathway (Chen et al., 2009). This is commonly done by performing either a one-tailed hyper-geometric test, Fishers exact test, or the chi-squared test as a statistical measure of significance for over-representation. In some instances, it may be suggested to perform enrichment analysis of gene-sets for up- and down-regulated genes independently. Whilst this is a simple approach requiring very little computational time, it ignores genes which lie outside the candidate gene list, so is highly dependent on user-defined cut-offs when identifying differentially expressed genes.

Functional class scoring gene-set analysis considers an entire data set belonging to two groups as input. All genes are ranked based on correlation between their expression and group status using a suitable rank metric. Subsequently, the method computes an enrichment score based on whether gene members of an *a priori* defined gene-set are randomly distributed throughout the ranked dataset or primarily located towards the extremes, irrespective of whether they are differentially expressed (i.e. GSEA) (Subramanian et al., 2005). Significance of this score is based on permutation of sample labels and adjusted for multiple comparisons. A sub-grouping of these methods aim to correct this ranked test for inter-gene correlations, given that expression measurements of individual genes in a gene-set are almost always correlated, which take into consideration both extent and directional fold-change information (i.e. CAMERA, Qu-SAGE) (Ritchie et al., 2015; Yaarie et al., 2013).

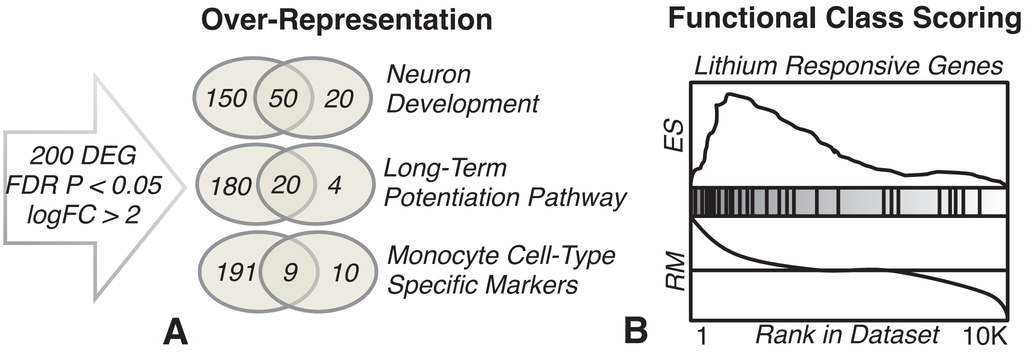


Figure 3.7 Two common approaches to gene-set enrichment analysis based on a hypothetical candidate list of 200 differentially expressed genes (DEG). (A) Over-representation tests for a significant over-representation of differentially expressed genes in a prior defined genes lists representing biological processes, KEGG pathways or cell-type specificity. Numbers in venn diagram indicate the overlap of DEG onto genes annotated as neuron development, long-term potentiation and monocyte specific markers. (B) Functional class scoring first uses a rank metric (RM) to correlate global expression to group status and then computes an enrichment score (ES) by ranking the entire dataset from high-to-low expression and testing for enrichment of a priori defined gene-sets towards the extremes of the list.

3.7.3.1. Semantic Similarity of Ontology Terms

Semantic similarity modularization integration (SSIM) is used to make sense of large lists of significant over-represented gene-sets generated from a list of differentially expressed genes, by organizing lists of biological vocabularies based on GO semantic similarity revealing structured biological processes involved in disease processes. A composite set of over-represented gene-sets is often used to create gene-set pairwise similarity matrices based on GO semantic similarities between gene pairs using GoSemSim (Yu et al., 2010). The similarity matrix can be subjected to hierarchical clustering analysis (i.e. Euclidean or Pearson correlation coefficients) to reveal ‘mini’ modules (i.e. groups of genes) with essential biological functions able to discriminate between groups.

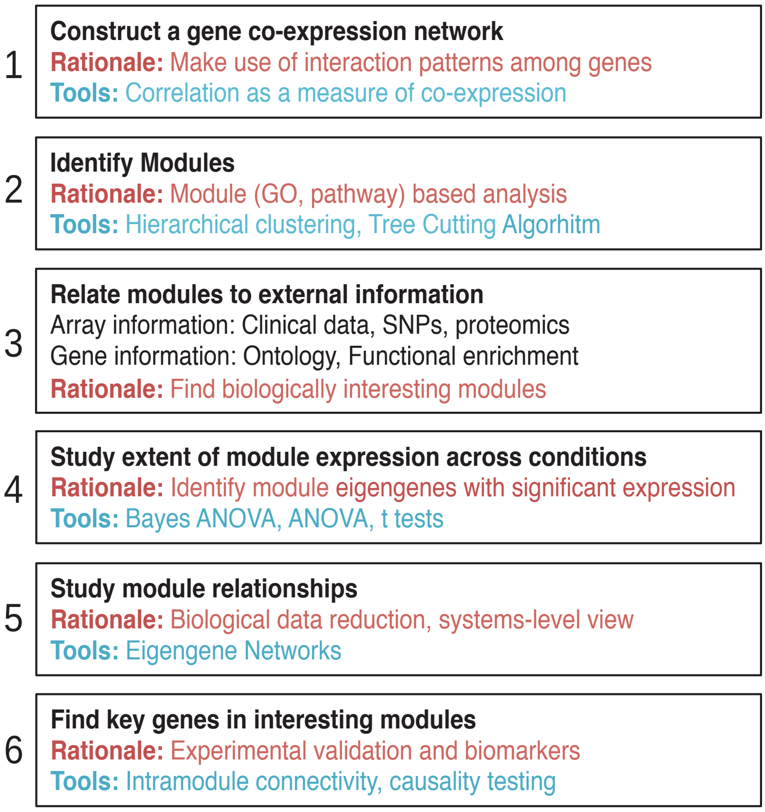
3.7.4. Protein-Interaction Networks

While genes within gene-sets interact in some shape or form, especially at the pathway level, researchers are often left to interpret the potential implications of these molecular profiles using their knowledge regarding the biology of the functionally enriched vocabularies. The physical mappings of complex biological networks provide a conceptual framework to interpret candidate gene lists as interactome network maps. This effort has resulted in the creation of open source protein-protein interaction (PPI) repositories which provide user-friendly ways to extract physical protein-protein interaction information for a list of candidate genes (i.e. STRING) (Franceschini et al., 2012). Mapping PPIs onto gene expression data provides a useful framework to understand functional implications of novel genes and gene-sets and discovering physical links between biologically meaningful genes and gene-sets. This is especially the case when a candidate gene list is composed of differentially expressed genes because such genes generally co-function in specific biological processes and pathways. Integration and visualization of PPI and gene expression data is routinely accomplished with the open source software CytoScape (Shannon et al., 2003).

PPI repositories are validated in a variety of ways ranging from meta-mining literature to computational predictions based on physical/biochemical interactions and extending to a range of *in vitro* (i.e. protein microarrays, affinity chromatography) and *in silico* (i.e. gene fusion, phylogenetics, gene expression) methods. Unfortunately, there is poor overlap in PPIs across databases despite overlapping proteins, as well as large differences in PPI annotations through the use of alternative vocabulary terms across repositories (Mathivanan et al., 2006). Given the poor consistency across databases it is recommended to make sure that core PPI network findings can be validated across independent PPI databases.

3.8. Weighted Gene Co-expression Network Analysis

Contrary to supervised ML and differential expression analysis, weighted gene co-expression network analysis (WGCNA) provides a means to move beyond single gene approaches and provide a systems-biology perspective for understanding biological disturbances underlying disease etiology. WGCNA is able to aggregate gene expression measurements from across the entire blood transcriptome in an unbiased fashion, to focus analysis on discrete groups of genes with highly correlated expression patterns (i.e. co-expression modules) (Langfelder, 2008a). The probability for multiple transcripts to follow a complex pattern of expression across dozens or even hundreds of samples only by chance is low and such sets of genes should therefore constitute coherent and biologically meaningful transcriptional modules. Because of the large number of comparisons (usually >10,000) within conventional approaches (e.g. differential expression), these results are far less permissive to 'noise', so enhancing biomarker discovery and interpretation. Transcriptional modules can be annotated for specific molecular functions, peripheral blood cell type specificity and can be further be associated to disease status, clinical measurements and external biological data. Modules with likely biological origins and direct clinical associations reflect gene regulatory networks of the blood transcriptome and act as functional biomarkers of disease rather than a panel of unique blood-based biomarkers (i.e. ML). Basic principles of WGCNA are outlined in **Figure 3.9,** from constructing a global co-expression network, to the identification of sub-networks (i.e. co-expression modules), external data integration, the study of co-expression module relationships and identification of network hub genes.



1 **Construct a gene co-expression network**
Rationale: Make use of interaction patterns among genes
Tools: Correlation as a measure of co-expression

2 **Identify Modules**
Rationale: Module (GO, pathway) based analysis
Tools: Hierarchical clustering, Tree Cutting Algorhitm

3 **Relate modules to external information**
Array information: Clinical data, SNPs, proteomics
Gene information: Ontology, Functional enrichment
Rationale: Find biologically interesting modules

4 **Study extent of module expression across conditions**
Rationale: Identify module eigengenes with significant expression
Tools: Bayes ANOVA, ANOVA, t tests

5 **Study module relationships**
Rationale: Biological data reduction, systems-level view
Tools: Eigengene Networks

6 **Find key genes in interesting modules**
Rationale: Experimental validation and biomarkers
Tools: Intramodule connectivity, causality testing

Figure 3.8 Basic principles behind weighted gene co-expression network analysis (WGCNA). First a pairwise correlation matrix is drawn and raised to some power Beta. Following, modules are detected with hierarchical clustering and implementing a dynamic branch cut algorithm to identify discrete groups of co-regulated genes (modules). Modules are subsequently interrogated for associations to clinical traits, enrichment of differentially expressed genes as well as functional annotation. Modules can also be used associated to each other and key hub genes driving the formation (clustering) of such co-expressed modules can be identified and labeled as therapeutic or putative biomarkers depending on the experimental context.

3.8.1. Constructing a Global Weighted Gene Co-expression Network

Deriving a gene network from a matrix of gene expression measurements constitutes a multi-step analytical process. First, expression data is filtered by coefficient of variation and normalized across all experimental groups and samples. Second, gene co-expression is measured with a correlation coefficient across all possible gene pairs. This is done with a Pearsons correlation for sample sizes greater than 20. For smaller sample sizes it is advisable to use a bi-midweight coefficient as a more robust means to measure correlations across small sample sizes. The end goal is to create a correlation matrix, a_{ij}, a symmetric $n \times n$ matrix with entries in [0, 1] whose component a_{ij} encodes the network connection strength between nodes *i* and *j*. Subsequently, an adjaceny matrix is computed by defining a *co-expression similarity* s_{ij} as the absolute value of the correlation coefficient between the profiles of genes *i* and *j* (Formula 3.3)

$$s_{ij} = |cor(x_i, x_j)|$$

(Formula 3.3)

Third, WGCNA takes the absolute value of the correlations and raises them to the power ß, in order to emphasize strong correlations and punish weak correlations on an exponential scale. This is because un-weighted networks do not reflect the continuous nature of the underlying co-expression information and consequently produce a loss of information. Moreover, while expression data can be noisy and the number of samples is often small, this step is useful for both consolidating and removing transcriptional noise (i.e. technical and non-biological variation) (Formula 3.4).

$$a_{ij} = s_{ij}^{\beta}$$

(Formula 3.4)

There is a trade-off between maximizing the scale-free topology model fit (scale free fitting parameter R^2) and maintaining a high mean number of connections (**Figure 3.10**). That is, high values of ß often lead to higher values of $R^{2,}$ but the higher power of ß, the lower is the mean connectivity of the network. Consequently, a good rule of thumb is to consider those powers that lead to a network satisfying scale-free topology at least approximately (e.g. $R^2 > 0.80$) so the mean connectivity is high and the network contains enough information (i.e. module detection).

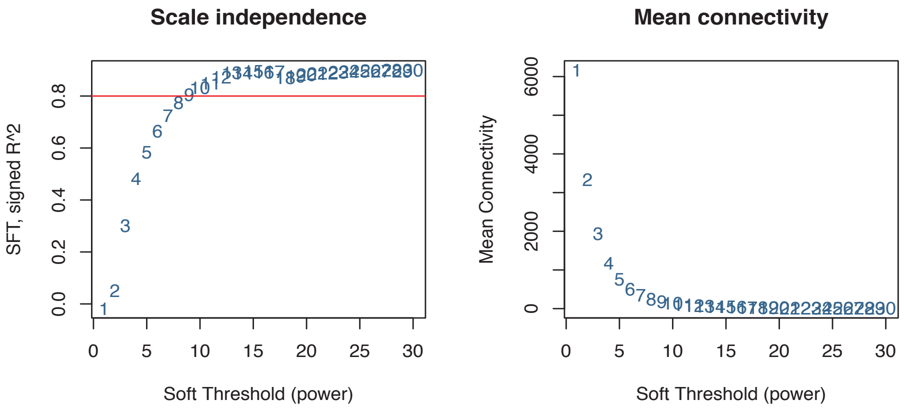


Figure 3.9 How to create a weighted co-expression network and choose a proper ß for your dataset. The higher the ß, the better the scale free-topology (SFT) (left). However, the higher the ß also causes depletion of network connectivity (right). As a rule of thumb, ß values with a SFT higher than 0.8 are optimal.

3.8.2. Identification of Sub-networks from the Global Network

Once a global weighted gene co-expression network is created, the next step is the identification of sub-networks from the global weighted network. These sub-networks, or gene co-expression modules, are discrete groups of genes with highly correlated expression patterns. In WGNCA, there are many options in module identification. One of the more robust and powerful approaches is hierarchical clustering using the standard R function *hclust* (Langfelder, 2008a) (**Figure 3.11**); branches of the hierarchical clustering dendrogram correspond to modules and can be identified using one of a number of available branch cutting methods, for example the constant-height cut or two Dynamic Branch Cut methods (Langfelder, 2008b). Although the height and shape parameters of the Dynamic Tree Cut method provide improved flexibility for branch cutting and module detection, it remains uncertain how to choose optimal cutting parameters or how to estimate the number of clusters in the data set.

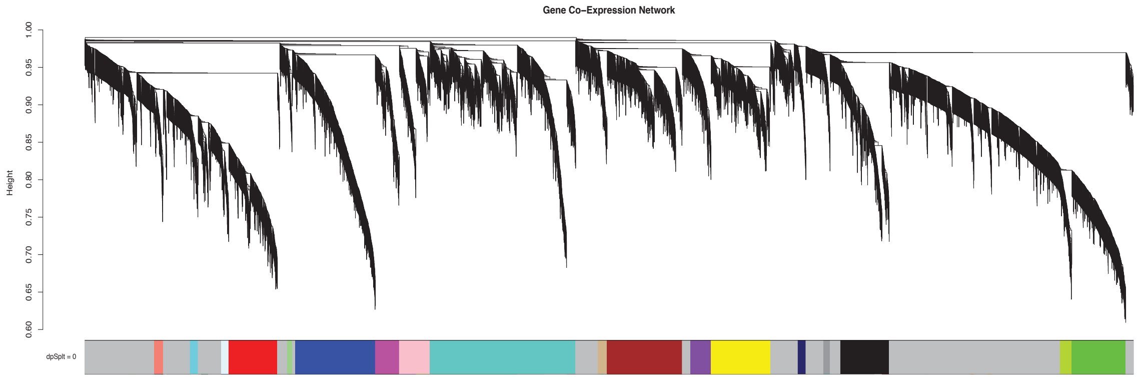


Figure 3.10. A cluster dendrogram of 28 identified modules in a network. Each hanging line represents a gene (leaf) on the tree and each group of genes (branch) represents a group of co-expressed genes. Numerous discrete modules have been identified in the colour band below the tree where the grey colour reflects genes which do not correlate well with densely interconnected genes.

As aforementioned, clusters of coordinately expressed genes may reflect biological signal such as GO terms, KEGG pathways, or cell type specific signatures (or even batch effects or contamination). When interpreting co-expression networks, it is therefore helpful to focus on modules with likely biological origins instead of those which may be associated to technical effects. To test whether the identified modules are biologically meaningful, functional enrichment analysis can be used on each module independently. If a significant proportion of genes within a co-expression module relate to functional or cellular properties (i.e. over-representation gene-set enrichment) via ‘guilt-by-association’, the remaining genes in a module are expected to be of that function.

3.8.3. Module Preservation and Module Differential Expression Analyses

The identification of disease-related gene co-expression networks are commonly identified in one of two ways. First, in network applications, one is often interested in studying whether modules are preserved across multiple networks. For example, to determine whether a pathway of genes is perturbed in cases relative to healthy controls, one can study whether its connectivity pattern is no longer preserved in one group compared to the other. Non-preserved modules can either be biologically uninteresting (e.g., reflecting data outliers) or interesting (e.g., reflecting fundamental co-regulatory differences). Here, the creation or disruptions of co-expression patterns within modules are examined transitioning from a healthy to disease state through ‘module preservation statistics’. There are both internal and external indexes of module preservation including density, connectivity and cross-tabulation based module preservation statistics. However, based on a global view of modular structure, it may be advantageous to aggregate multiple module statists into ‘summary preservation’ statistics based on a permutation testing implemented in the modulePreservation R function (Langfelder et al., 2011). The modulePreservation function in R implements a permutation test involving several powerful network based statistics for evaluating module preservation. These statistics are summarized into the composite preservation called Zsummary. For each module in one dataset (a disease dataset), the function calculates the Zsummary statistic in the second dataset (a control dataset). For a given module, Zsummary > 10 indicates strong evidence for preservation in the test data set. Zsummary < 2 indicates no evidence of module preservation. An advantage of the preservation Z statistic is that it makes few assumptions regarding module definition and module properties.

Since biologists are often more familiar with *p*-values as opposed to Z statistics, this R function also calculates empirical *p*–values ($P_{summary}$). The smaller the *p*-value, the stronger the evidence that the module is preserved. It is important to note that module preservation and module disruption are related and complementary concepts and they can both hold for a given module. Even though modules might be highly preserved across biological conditions, this does not preclude the emergence of subtle changes in network structure that are not enough to render the module non-preserved, but nevertheless are statistically significant and, potentially, biologically meaningful.

Thus, alternatively, networks can be constructed of case and control data collectively and the resulting network modules subjected to statistical testing. This represents a test of 'module significance', and is an approach which is able to complement standard gene differential expression analysis, at the gene network level (Langfelder, 2008a). For example, if a gene co-expression network is identified across two experimental groups (i.e. disease vs. control) let *GS* (gene significance) represent the $-\log_{10}$ p-value for every gene following a conventional differential expression analysis between disease and control group labels, as a measure of strength of differential expression (Formula 3.5).

$$GS = -log_{10}\ p\text{-}value$$

(Formula 3.5)

Let *MS* be calculated as the average *GS* within each module (Formula 3.6). This test will allow for the identification of co-expressed modules that are enriched for a large number of differentially expressed genes **(Figure 3.12).**

$$MS = {}_{u}GS$$

(Formula 3.6)

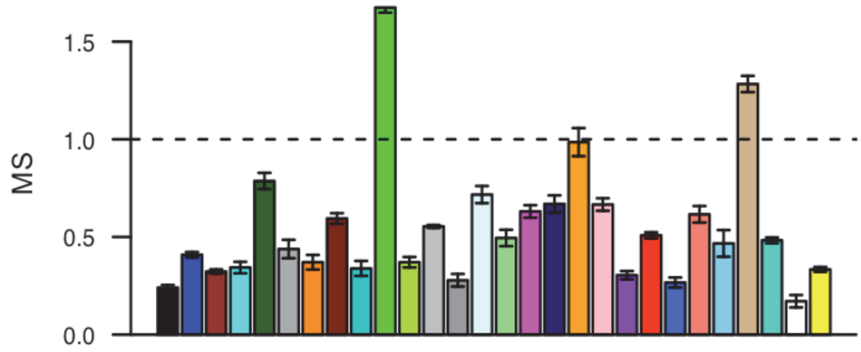


Figure 3.11. Integrating differential expression analysis into co-expression analysis through module significance measures. A total of 28 modules were identified. MS values averaged across three group-wise testing (Disease1, Disease2 and Controls) on the y-axis. This plot shows an enrichment of differentially expressed genes within the green and tan colored modules.

While the direction of change (i.e. up- or down-regulation) in one group relative to the other with **Figure 3.12** is left to interpretation, it is often wise to investigate these matters further.

Once a module of interest has been identified it can be summarized down to its first principal component termed the module eigengene (ME). The ME summarizes the main trend of gene expression across samples for a particular module of interest. ME values for all identified modules can be subjected to differential module expression analysis, correcting *p*-values for multiple comparisons and visualized in a boxplot **(Figure 3.13)**. Now the direction of fold-change can be interpreted across the green and tan modules and subjected to statistical testing. This approach drastically reduces the multiple comparison problems from thousands of genes to tens of modules.

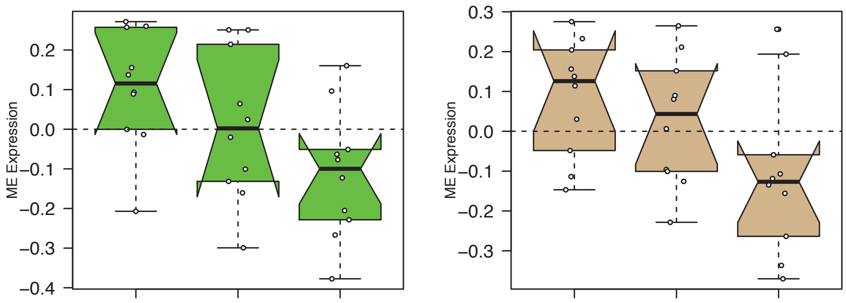


Figure 3.12. Summarizing module ME values for differential module expression. Here green and tan colored modules from Figures 3.11 and 3.12 are displayed across Disease 1 (far left), Disease 2 (middle), Control (left) groups.

3.8.4 Integration of Multi-Modal Data

Co-expression networks also provide a statistically sound framework for data integration of external clinical traits, PPI information and multiple omic data-types. The identification of clinical trait-related co-expression modules adds another layer of information to each module, bringing module discovery closer to phenotypic alterations and recorded clinical manifestations. When diagnosing psychiatric illness numerous clinical findings and laboratory measurements may be collected, but only infrequently incorporated into the analysis. ME values can be associated to external data-types **(Figure 3.14).** In this case, these data can be correlated, through Pearson's or Euclidean's correlation coefficients, to ME values and significance is drawn with a students asymptotic P-value for significance. This provides a sophisticated approach for identifying gene co-expression networks which may be associated to potential confounding factors (e.g. age, smoking, gender etc..).

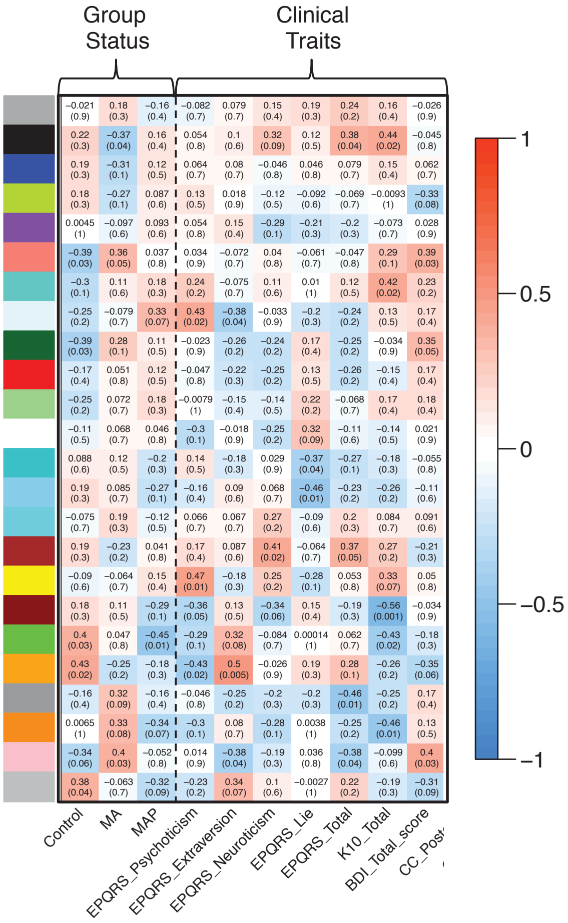


Figure 3.13. Summarizing modules by ME values (y-axis) and correlating them with external clinical trait information and neuroimaging data (x-axis). Red signifies a positive correlation and blue a negative correlation. Inside each box is the *r* value as the strength of correlation above, and its associated p-value below.

3.8.5 Hub Genes

A key aspect of WGCNA is the ability to find centrally located intramodular hub genes. Genes with the highest correlation to other genes within a module, i.e. those that are highly connected, are labeled hub genes and are predicted to be of essential function to the co-expression module. Hub genes are explained simply with the following equation where x_i is the profile of gene *i* and $E^{(q)}$ is the module eigengene of module *q* (Langfelder, 2008b). If for example, a hub gene and a clinical trait are both highly correlated to a ME (of a particular function), this hub may represent a putative marker with putative implications and association to the trait being measured (Formula 3.7)

$$K_{cor,\, i} = cor(x_i, E^{(q)})$$

(Formula 3.7)

3.8.6 Statistical and Translational Implications for Gene Networks

Moving from an unsupervised gene-network approach to clinical utility requires a multi-step process. First, the biology underlying disease etiology should be fully explored: comprehensive molecular characterizations at the systems-level only enhance and guide future prognostic and diagnostic hypotheses. For example, ML often fails to place blood-based biomarkers into a coherent biological framework making it difficult to derive practical and mechanistic insights of ML derived single gene biomarkers. Systems approaches permit the placing of single biomarkers into an empirically derived gene-network with likely biological origins. Second, reproducibility is a necessity in genetic and biomarker testing. While there is a need to increase the likelihood that findings will prove reproducible and have predictive power in independent cohorts, a key advantage of systems-level analyses is that they are often more robust and reproducible compared to ML (Chaussabel, 2015). Repeated studies, which follow up functional characterization of prioritized candidates through network models are needed, ideally capitalizing on emerging systems-immunology technologies. Third, network analyses are particularly useful in pharmacogenomics: for example, the identification of a co-expression module differentiating cases from controls before symptom development may represent a feasible drug target for limiting disease development. Alternatively, the identification of drug-induced co-expression modules may be able to predict novel gene functions and provide new insights regarding drug-induced mechanisms and provide leads for drug repositioning. Moreover, when drug-induced responses are placed into the context of specialized immune subsets, opportunities to understand pharmacological and toxicological chemical properties may unfold. Fourth, systems-level analyses permit the integration of genetic variants, neuroimaging findings, and clinical measurements with blood transcriptome data. Integrating such data across multiple scales could lead to more informed decisions for personalized, predictive and preventive medicine. Finally, the inclusion of multiple disease-types is a key step towards placing results into a broader context. Determining how gene networks interact and converge across psychiatric diseases supports the discovery of gene networks which might drive critical neurobiological processes involved in the pathophysiology of many psychiatric disorders.

3.9. Summary

Following initial pre-processing, quality control and normalization of gene expression data, three broad analytical themes have emerged to address specific clinical and biological aims. First, the identification of a unique panel of biomarkers with putative prognostic or diagnostic clinical value – this is a supervised machine-learning classification problem and is applied in Chapters 5 and 7. Second, the identification of differentially expressed genes and the mapping of these genes on to dysregulated pathways and PPI information – this accords with a conventional bioinformatics pipeline and varying aspects of these approaches are applied throughout Chapters 5-8. Third, the identification of functional biomarkers (i.e. gene networks) of disease and treatment response – this analytical challenge aligns with more holistic WGCNA applications and is used throughout Chapter 5-8.

Chapter 4

Aims

The central purpose of this work is to evaluate the utility of genome-wide blood-based gene expression measurements for the prediction, diagnostics and treatment of patients diagnosed with psychiatric diseases. This is particularly relevant given that blood-based transcriptome gene expression biomarkers sufficiently reflect changes in the amounts and combinations of RNAs expressed at various times in response to gene-environment interplay, health-to-disease transitions and mechanisms underlying therapeutic treatment. Here, I present a summary of the specific intentions of each primary research chapter and specific author contributions below.

4.1. PART II – Application of Blood Transcriptomics in Psychiatric Diseases

4.1.1. Chapter 5: Candidate Blood Biomarkers and Gene Networks of Posttraumatic Stress Disorder

Chapter 5 contains the generation and subsequent analysis of blood-based RNA-Seq gene expression measurements collected from U.S. Marines (N=188) prior-to and following deployment to conflict zones (i.e. Iraq and Afghanistan). The collected sample size was enriched for U.S. Marines whom developed posttraumatic stress disorder (PTSD) following deployment. The aim of this study was to identify blood-based gene networks and biomarkers capable of characterizing PTSD risk (at pre-deployment) and PTSD development (at post-deployment). I further sought to reproduce relevant gene signatures in an independent cohort of U.S. Marines for which blood-based microarray gene expression measurements were generated (N=96).

This chapter is predominately my work with significant input from Dr. Christopher H. Woelk and Dr. Caroline Nievergelt at the University of California San Diego (UCSD). As first and corresponding author I was responsible for RNA isolation from whole blood, RNA quality check, design and application of appropriate statistical analyses, data interpretation, manuscript write-up, submission and handling of reviewer comments.

4.1.2. Chapter 6: Immediate Molecular and Cellular Response to Acute Psychological Stress

Chapter 6 consists of analyzing blood-based microarray gene expression data and integration with physiological measurements (endocrine and autonomic) throughout the sequence of events leading up to, during and following a first-time tandem skydive in otherwise healthy participants (N=13). Gene expression results were compared to a second cohort of healthy participants (N=26) for which peripheral blood was subjected to flow-cytometry. The aim of this work was to describe the molecular and cellular response of the human innate and acquired immune system in reaction to physical danger.

As first and corresponding author of this work my contribution was design and application of appropriate statistical analyses, data interpretation, manuscript write-up, submission and handling of reviewer comments. Dr. Nadia Beliakova-Bethell (UCSD) was responsible for RNA isolation from whole blood and RNA quality check and Drs. Christopher H. Woelk and Brinda Rana (UCSD) provided significant input.

4.1.3. Chapter 7: Candidate Blood Biomarkers and Gene Networks of Methamphetamine-Associated Psychosis

Chapter 7 contains the analysis of blood-based of RNA-Seq gene expression data and integration with subcortical brain structural volumes and numerous clinical parameters of subjects diagnosed with methamphetamine-associated psychosis (MAP) (N=30). The clinical presentation, course and treatment of MAP are similar to that observed in schizophrenia (SCZ) and subsequently MAP has been hypothesized as a pharmacological and environmental model of SCZ. The central aim of this work was to accurately identify and characterize MAP with the given data and to validate the MAP model as an exemplar for SCZ biomarker discovery.

As first and corresponding author of this work my contribution was design and application of appropriate statistical analyses to all data, data interpretation, manuscript write-up, submission and handling of reviewer comments with significant input from Dr. Dan Stein at the University of Cape Town (UCT). Dr. Christiane Nday (UCT) was responsible for RNA isolation and quality check and Dr. Anne Uhlmann was responsible for brain scanning.

4.1.4. Chapter 8: Candidate Lithium Responsive Genes and Gene Networks in Bipolar Disorder Lymphoblastoid Cell Lines

Chapter 8 contains an analysis of RNA-Seq expression profiles of bipolar disorder (BD) patient primary cell transformed lymphoblastoid cell lines prior-to and following lithium treatment using three experimental groups; (**1**) BD patients that respond to lithium treatment (responders) (N=8); (**2**) BD patients that do not respond to lithium treatment (non-responders) (N=8); (**3**) healthy controls (N=8).The intent of this study was to explore the mechanism of action and heterogeneity in clinical response to lithium treatment in BD.

For this work, as first and corresponding author my contribution was design and application of appropriate statistical analyses to all data, data interpretation, manuscript write-up, submission and handling of reviewer comments with significant input from Drs. Christopher H. Woelk and John Kelsoe (UCSD). Ms. Tantyana Shekhtman was responsible for cell culture work.

4.2. Part III – Moving Biomarkers Forward in Psychiatry

4.2.1. Chapter 9: Moving Biomarkers Forward in Psychiatry

Finally, Chapter 9 extends new ideas and postulates for what constitutes a good biomarker and guidelines towards achieving accurate and objective blood-based biomarkers for psychiatric disease.

Part II

Application of Blood Transcriptomics in Psychiatric Diseases

Chapter 5

Candidate Blood Biomarkers and Gene Networks of Posttraumatic Stress Disorder

5.1. Background

There is much scope for studying molecular factors that determine risk and subsequent development of post-traumatic stress disorder (PTSD). Significant numbers of men and women exposed to severe emotional trauma and loss emerge from these events with persistent PTSD symptoms, such as intrusive imagery, avoidance and hyperarousal, as well as other long-term physical health problems. PTSD affects 7-8% of the general United States (US) population, and is more common among troops recently returned from military service in Iraq and Afghanistan, with estimates of prevalence as high as 20% (Ramchand et al., 2010). Annual health care costs associated with PTSD in the US have been estimated to be 180 million dollars (Heinzelmann & Gill, 2013). Heterogeneity in susceptibility to PTSD suggests that differences at the molecular level (i.e. gene-expression level) may influence an individual's physiological and psychological response to trauma and thus the development of PTSD. A clear understanding of the molecular mechanisms underlying this response to trauma is required to reduce the substantial morbidity and mortality associated with this disorder.

A number of studies have analyzed blood gene expression and glucocorticoid activity to build more effective models for identifying molecular factors associated with PTSD (Ziker et al., 2007; Yehuda et al., 2009; Neylan et al., 2011; Sarpas et al., 2011; Mehta et al., 2011; Pace et al., 2012; van Zuiden et al., 2012a; van Zuiden et al., 2012b; Matić et al., 2013; Glatt et al., 2013). These studies were recently reviewed by Heinzlemann and Gill (2013), who summarized that the increased expression of inflammatory genes and decreased expression of genes that regulate inflammation contribute to the onset of PTSD. Specifically, when considering the overlap in results from transcriptomic studies, decreased expression of *FKBP5* and *STAT5B*, which both regulate inflammation, is evident (Yehuda et al., 2009; Sarpas et al., 2011; Mehta et al., 2011; van Zuiden et al., 2012a). While suggestive, the majority of these reviewed studies centered transcriptomic analysis on pre-determined targets (contrary to genome-wide applications) in subjects already diagnosed with PTSD, and thus lacked a prospective study design. Consequently, the identification of gene networks and blood biomarkers that confer risk to and resilience against PTSD remains an inadequately researched area.

In the current investigation, RNA-Seq and microarray gene expression profiling was applied to peripheral blood taken from two independent cohorts of U.S. Marines (N=188, N=96), both before and after deployment to conflict zones (Iraq and Afghanistan). These rare samples provide an opportunity to better understand the molecular factors involved in the pathophysiology of PTSD and to identify blood-based biomarkers and gene networks implicated in PTSD risk and development. To do so, four main aims were tested. **Aim1:** First, to determine whether *large* changes in the underlying gene-gene connectivity (i.e. co-expression) in peripheral blood provide a basis for the pathology of PTSD. This aim included searching for gene co-expression networks (i.e. modules) that were either created or disrupted in PTSD cases relative to controls, and vice versa, by testing for module preservation. **Aim 2:** Second, to determine whether *subtle* changes in the underlying gene-gene co-expression patterns in peripheral blood provide a basis for PTSD pathology. This aim included searching for modules using a combination of PTSD cases and controls and testing them for association with PTSD. **Aim 3:** Third, focusing analysis on the individual gene level to identify differentially expressed genes between PTSD cases and controls cross-sectionally at pre- and then post-deployment, and subsequently testing longitudinally between time-points. **Aim 4:** Finally, to construct gene expression classifiers (unique panels of biomarkers) for predicting the PTSD development at pre-deployment and for classifying PTSD at post-deployment, while cross-validating prediction accuracies using an independent withheld test dataset.

5.2. Materials and Methods

5.2.1. Subject Selection and PTSD Diagnosis

All subjects were male and participants in the Marine Resilience Study (MRS), a prospective study of well-characterized U.S. Marines scheduled for combat deployment to Iraq or Afghanistan, with longitudinal follow-up to track the effect of combat stress. At the time of each blood draw, PTSD symptoms were assessed using a structured diagnostic interview, the Clinician Administered PTSD Scale (CAPS) (Blake et al., 1995; King et al., 1998; Weathers et al., 2001). Using criteria from the Diagnostic and Statistical Manual of Mental Disorders, fourth edition (2000) (APA, 2000), diagnosis for partial or full PTSD was defined as a threat to life, injury, or physical integrity (Criterion A1) and the presence of at least one re-experiencing symptom and either three avoidance symptoms or two hyperarousal symptoms, or two avoidance symptoms plus two hyperarousal symptoms (Blanchard et al., 1995a; Blanchard et al., 1995b; Blanchard et al., 1996). Symptoms must have occurred at least once within the past month (frequency ≥ 1) and caused a moderate amount of distress (intensity ≥ 2). PTSD co-morbidities (e.g. depression, acute stress disorder, agoraphobia) according to MINI International Neuropsychiatric Interview (MINI) criteria were not recorded for these participants.

A subset of MRS study participants, enriched for PTSD post-deployment, were pre-selected for gene expression analysis. Participants had to meet the following criteria. First, at pre-deployment, all participants had to be symptom free, with no PTSD diagnosis and a CAPS score ≤ 25. Second, at post-deployment, participants who fulfilled criteria for partial or full PTSD diagnosis were designated the PTSD group. Third, participants with post-deployment CAPS score ≤ 25 that matched the post-deployment PTSD group on variables of combat exposure, age and ethnicity were designated the ’control‘ group. Under these criteria, all paired subjects were stratified into two groups based upon CAPS scores at 3-months post-deployment. If a participant developed PTSD following trauma-exposure at 3-months post-deployment, their pre-deployment sample would be included in the ‘PTSD-risk’ group. Likewise, if a subject avoided PTSD symptoms at 3 months post-deployment their sample at pre-deployment was included in the ‘control’ group.

5.2.2. RNA Isolation and Generation of Gene Expression

Dataset 1: Whole blood was obtained from 124 U.S. Marines who served a 7-month deployment. Blood was drawn one month prior to deployment and again at 3-months post-deployment for each participant. Each blood sample (10ml) was collected into an EDTA-coated collection tube, RNA was isolated from peripheral blood using LeukoLOCK Total RNA Isolation Kit and all samples passed a RNA integrity number (RIN) >7. mRNA was subject to Poly-A enrichment and libraries were prepared for sequencing using standard Illumina Tru-Seq protocols and subjected to 50bp paired-end sequencing on the Illumina Hi-Seq 2000. RNA-Sequenced reads were trimmed for adaptor sequence, and masked for low-complexity or low quality sequence, then mapped to hg19 RefSeq Human Genome using RSEM and counted using HT-Seq count. **Dataset 2:** For external validation, data were compared to an independently generated gene expression data-set from a separate, non-overlapping, group of 50 MRS Marine participants (Glatt et al., 2013, previously published pre-deployment data). Similarly, whole blood was obtained from U.S. Marines who served a 7-month deployment at both one month prior to deployment and again at 3-months post-deployment. RNA samples were treated in an identical fashion as described above, however final RNA was hybridized to the Affymetrix Hu-Gene 1.0 ST Array.

5.2.3. Data Pre-Processing

Dataset 1: Quality control metrics were used to identify poor quality mRNA reads and potential outliers which may bias downstream analysis. First, GT content and library sizes were compared across samples to identify samples with large sequencing error. Second, genes with low read counts were filtered in a non-specific manner using edgeR. Genes having > 20 counts per million in at least 50% of the samples were retained. Third, resulting count data was normalized using the edgeR VOOM function (Robinson et al., 2010) and subjected to clustering analysis and principal component analysis (PCA) to identify outliers beyond 2 standard deviations from the average. From a total of

124 participants (52 with PTSD and 72 without PTSD) outlier analysis identified 10 outliers (5 with PTSD and 5 without PTSD), so yielding a total of 47 with PTSD and 67 without PTSD. Subsequently, we sought to obtain a balanced experimental design by matching subjects at baseline by CAPS scores. Our final cohort consisted of 47 participants with PTSD and 47 without PTSD, matched for baseline anxiety-like symptoms, sampled both prior to and following deployment to conflict zones (**Table 5.1**).

Table 5.1. Recorded clinical parameters from U.S. Marines assessed at pre- and post-deployment in Dataset

Time point	**Pre-Deployment**			**Post-Deployment**		
	PTSD Cases (N=47)	Controls (N=47)	P-Value	PTSD Cases (N=47)	Controls (N=47)	P-Value
Age	22.15 ± 2.53	22.42 ± 3.92	0.682	23.14 ± 2.52	23.42 ± 3.92	0.685
Alcohol	2.08 ± 1.55	1.62 ± 1.33	0.124	1.79 ± 1.32	1.54 ± 1.11	0.318
Tobacco	1.75 ± 1.62	0.97 ± 1.51	0.02	1.69 ± 1.69	1.02 ± 1.47	0.042
WC adj.	1.65 ± 0.13	1.72 ± 0.13	0.015	1.68 ± 0.14	1.75 ± 0.12	0.012
PCL	21.29 ± 4.72	18.33 ± 2.27	0.0001	42.38 ± 11.09	20.94 ± 3.87	5.37E-22
CAPS total	11.39 ± 7.23	6.75 ± 6.90	0.002	53.17 ± 15.08	10.04 ± 7.26	5.99E-32
CAPSBs	1.00 ± 1.91	0.54 ± 1.92	0.245	14.9 ± 7.25	1.54 ± 2.37	6.29E-21
CAPSCAs	0.54 ± 1.11	0.10 ± 0.51	0.015	5.31 ± 4.57	0.85 ± 2.08	1.88E-08
CAPSCN1s	1.10 ± 2.23	0.97 ± 2.88	0.813	9.17 ± 5.32	1.19 ± 2.87	1.21E-14
CAPSDs	8.39 ± 5.66	4.58 ± 4.98	0.001	22.6 ± 6.7	6.42 ± 4.79	5.97E-24
CAPSCs	2.00 ± 2.73	1.62 ± 3.66	0.571	15.67 ± 7.23	2.08 ± 3.66	7.15E-20
Prior Deployment	19	16	0.6699	-	-	-
TBI	-	-	-	30	21	0.097
CES PBE mean	-	-	-	0.63 ± 0.25	0.53 ± 0.12	0.02
Caucasian	26	26	1	-	-	-
African American	4	4	1	-	-	-
Native American Mexican	13	15	0.822	-	-	-
Asian & Other	5	3	0.714	-	-	-

Abbreviations: Alcohol = alcohol consumption; Tobacco = tobacco use; WC adj. = waist circumference was adjusted for height; PCL = PTSD symptom check list, CAPS total = CAPS total score, CAPSBs = re-experiencing subscale, CAPSCAs = symptoms of avoidance, CAPSCN1s = symptoms of numbing, CAPSCs = subtotal C subscale, CAPSDs = hyper-arousal subscale, TBI = traumatic brain injury, CES = combat exposure scale, PBE = post battle experience; -, not applicable. Significance was assessed with a Student's two-tailed *t* test for continuous variables and fishers exact test of proportions for binary variables. (Average ± standard deviation).Grey shading is for visualization only.

Dataset 2: Microarray files (.CEL) were read using library affy (Gautier et al., 2004) and background adjusted, normalized and summarized to the probe level using RMA normalization (Irizarry et al., 2003). If two or more probes mapped to the same gene identifier, one was selected on the basis of having higher average expression across all samples. Non-specific filtering by average expression removed the lower 10% of all probes on the array. Outliers were identified in an identical fashion (as above) producing a final cohort of 24 participants with PTSD and 24 without PTSD, at pre- and post-deployment (**Table 5.2)**.

Table 5.2. Recorded clinical parameters from U.S. Marines assessed at pre- and post-deployment for *Dataset 2.*

Time Point	**Pre-Deployment**			**Post-Deployment**		
	PTSD Cases (N=24)	Controls (N=24)	P-Value	PTSD Cases (N=24)	Controls (N=24)	P-Value
Age	22.52 ± 3.16	22.01 ± 3.19	0.58	-	-	-
CAPs	22.63 ± 12.02	13.33 ± 8.92	0	64 ± 18.42	10.75 ± 9.57	1.80E-16
PCL	24.58 ± 6.43	22.75 ± 3.34	0.22	49.25 ± 12.55	21.38 ± 5.33	3.80E-13
Prior Deployment	12	12	1			
CES	-	-	-	18.04 ± 13.24	19.25 ± 15.09	0.77
BPE	-	-	-	7.35 ± 4.59	7.96 ± 4.03	0.63
CES injured	-	-	-	9	2	0.04
TBI	-	-	-	11	4	0.06
Caucasian	17	18	1	-	-	-
African American	4	2	0.67	-	-	-
Native American	2	3	1	-	-	-
Asian Other	1	1	1	-	-	-

Grey shading is for visualization only. For abbreviations and p-value calculations see **Table 5.1**.

To compare findings from RNA-Seq data in *Dataset 1* to microarray data in *Dataset 2*, normalized gene expression measurements found across both platforms (*N*=10,184) passed into the subsequent analysis.

5.2.4. Weighted Gene Co-expression Network Analysis

5.2.4.1. Individual Network Construction and Module Preservation

WGCNA (Langfelder & Horvath, 2008) was used to create global gene co-expression networks for cases and controls independently at pre-deployment and post-deployment for *Dataset 1,* comprising a total of four co-expression networks. The ß power of 12 was reached for all four networks. The dynamic tree-cut algorithm was used to identify sub-networks (i.e. co-expression modules) from each global network, setting minimum module size and the minimum height for merging modules to 60 and 0.2. Subsequently, for each co-expression module in each network, the extent of co-expression preservation between cases and controls was assessed using a permutation-based preservation statistic, $Z_{summary}$, implemented within WGCNA with 1000 random permutations of the data. $Z_{summary}$ is used as a connectivity-based preservation-statistic able to determine whether the connectivity pattern between genes in a reference network is similar to that in a test network (Langfelder et al., 2011). A $Z_{summary}$ score < 2 indicates no evidence of preservation, 2< $Z_{summary}$ <10 implies weak preservation and $Z_{summary}$ > 10 suggests strong preservation. It is important to note that module preservation and module disruption are related and complementary concepts and they can both hold for a given module.

5.2.4.2. Combined Network Construction and Module Differential Analysis

Global gene co-expression networks were created with a combination of cases and controls at pre-deployment and post-deployment for *Dataset 1* and *Dataset 2,* comprising 4 global co-expression networks in total*.* Similarly, the power of 12 was reached for both networks in *Dataset 1* and the power of 30 was reached for both networks in *Dataset 2*. Network connectivity in the microarray data was less than that of the RNA-Sequencing data and a higher ß value was used to reach a more satisfactory scale-free topology for the networks. Similarity, the dynamic tree-cut algorithm was used to identify sub-networks (i.e. co-expression modules) from each global network, setting minimum module size and the minimum height for merging modules to 60 and 0.2. Here, module eigengenes (ME) for all modules were correlated to clinical parameters such as PTSD-risk status, control status, age, alcohol consumption, tobacco usage, CAPS scores and criteria, traumatic brain injury (TBI) and ethnicity which provides a complementary assessment of these potential confounders to that performed in standard differential expression analysis. For each gene in a module, module membership (*kME*) was defined as the correlation between gene expression values and *ME* expression. Genes with high *kME* inside co-expression modules are labeled as hub genes (Langfelder & Horvath, 2008). *GS* was calculated as the $–\log_{10}$ of the *p*-value generated for each gene within a particular module using a moderated *t* test and is a measure of the strength of differential gene expression between PTSD cases and controls. *MS* was calculated as the average *GS* within each module.

5.2.5. Functional and Cellular Enrichment Analyses

Module enrichment was assessed by over-representation analyses. First, broad over-arching module functions were determined using GO biological processes from the DAVID database (Huang et al., 2009) (http://david.abcc.ncifcrf.gov/). Second, more precise and exact module functions were determined using Reactome NCBI Biosystems pathways and terms (Geer et al., 2010). Finally, since co-expression patterns may also represent specific cell-types from a larger heterogeneous population (i.e. peripheral blood leukocytes), we undertook cell-type module enrichment using the highly expressed, cell specific (HECS) gene expression database compiled by Shoemaker et al. (2012). All module genes were used for enrichment analyses using a FDR corrected p-value < 0.05 as significant.

5.2.6. Differential Gene Expression Analyses

Differentially expressed genes were assessed using the moderated t-test in edgeR (Robinson et al., 2010) and LIMMA (Smyth et al., 2005) packages for RNA-Seq and microarray data, respectively. Our multi-level experimental design permitted us to test gene differential expression in three main ways. First, a cross-sectional analysis compared PTSD cases to controls at post-deployment. Second, a cross-sectional analysis compared PTSD risk cases to controls pre-deployment. Third, a longitudinal contrast analysis was performed utilizing the paired nature of these data searching for genes responding differently within one group across time-points, from pre- to post-deployment. The significance threshold for theses analyses was set to a nominal p-value < 0.05. A nominally significant p-value was used to yield a reasonable number of genes to include within network analyses.

5.2.7. Supervised Machine-Learning Classification

BRB-Array Tools (Simon et al., 2007) supervised multivariate classification methods were used to construct gene expression classifiers at pre-deployment (to predict PTSD development) and post-deployment (to diagnose PTSD). Each model consisted of three steps. First, all genes with $P < 0.05$ (comparing PTSD cases to controls) from Dataset 1 were subjected to classifier construction. These criteria were used to cast a wide net to catch all potentially informative genes, while false-positives could be discarded by subsequent optimization and cross-validation steps. Second, classifiers composed of different numbers of genes were constructed by recursive feature elimination (RFE). RFE provided feature selection, model fitting and performance evaluation via identifying the optimal number of features with maximum predictive accuracy. Third, the ability for RFE to predict group outcome was assessed by different multivariate classification methods including diagonal linear discriminant analysis (DLDA), support vector machine (SVM), nearest centroid (NC) and three-nearest neighbors (3NN). Prediction accuracies were cross-validated using a leave-one-out cross-validation (LOOCV) approach using Dataset 1 and subsequently performing external cross-validation of prediction accuracies on *Dataset 2*.

5.2.8. Batch (Technology) Correction

Despite applying similar supervised multivariate classification algorithms to construct gene expression classifiers from Dataset 1 (RNA-Seq) and Dataset 2 (Microarray), the ability to cross-validate classification accuracies using external data (i.e. Dataset 2) is hampered by the use of different technological platforms. Prior to classifier construction it is necessary to find means to merge these two different data distributions into one common distribution. To do so, normalized gene expression measurements from Dataset 1 and Dataset 2 were subjected to 'Combat correction' (Johnson & Rabinovic, 2007), a gene standardization approach, using the two datasets as independent batches for correction (**Supplementary Figure 5.1**). We could then proceed to classifier construction.

5.3. Results

5.3.1. No Large Differences in Module Preservation

WGCNA was used in *Dataset 1* to assess module preservation between PTSD cases (N=47) and controls (N=47) for the pre- and then the post-deployment time point. These analyses aim to identify large differences in gene co-regulatory patterns, as being disrupted or created in PTSD cases relative to controls, or vice versa. However, following 1000 random permutations of co-expression modules, we observed strong preservation statistics ($Z_{summary} > 10$) for all modules at pre and post-deployment for PTSD cases and controls (**Table 5.3**). This indicates similar fundamental gene co-regulation within PTSD cases and controls, suggesting that major changes in the underlying gene-gene connectivity are unlikely to be the basis for the pathology of this disorder. However, even though modules might be highly preserved across PTSD cases and controls, this does not preclude the emergence of subtle changes in gene network structure that are not strong enough to render the module fully non-preserved, but still may differ in gene expressions that are statistically significant and, potentially, biologically meaningful.

Table 5.3. Preservation statistics at pre- and post-deployment within *Dataset 1.*

a) Pre-Deployment : Case (reference) to Control			**b)** Pre-Deployment : Control (reference) to Case		
Module	**Module Size**	**$Z_{summary}$**	**Module**	**Module Size**	**$Z_{summary}$**
1	407	95.84	1	743	94.35
2	338	95.83	2	654	79.21
3	818	57.39	3	1000	49.3
4	225	37.11	4	233	40.55
5	95	34.87	5	119	34.89
6	171	33.05	6	524	31.06
7	168	31.77	7	1000	30.54
8	259	30.11	8	237	28.75
9	122	30.1	9	116	24.33
10	1000	29.29	10	150	23.81
11	79	22.27	11	254	21.59
12	481	22.22	12	281	17.59
13	114	19.13	13	399	17.09
14	73	18.66	14	164	16.47
15	213	17.55	15	100	11.13
16	79	16.53	16	118	10.2
17	70	13.21	-	-	-
18	157	12.63	-	-	-
19	105	11.48	-	-	-
20	100	10.38	-	-	-
21	196	10.69	-	-	-
c) Post-Deployment : Case (reference) to Control			**d)** Post-Deployment : Control (reference) to Case		
Module	**Module Size**	**$Z_{summary}$**	**Module**	**Module Size**	**$Z_{summary}$**
1	487	52.67	1	1000	44.3
2	1000	47.61	2	191	43.56
3	106	41.41	3	122	33.47
4	1000	27.42	4	198	32.84
5	119	26.18	5	125	29.84
6	68	20.9	6	1000	29.46
7	87	17.48	7	91	17.27
8	201	13.47	8	158	16.55
9	122	12.95	9	100	11.73
10	100	11.26	10	151	11.14

Gene network modules were constructed from Dataset 1 pre-deployment PTSD risk cases and assessed for preservation within a control network (**a**) and *vise versa* (**b**). The same test was performed within Dataset 1 post-deployment PTSD cases and assessed for preservation within controls (**c**) and *vise versa* (**d**). $Z_{summary}$ is the summary preservation statistic, using either the PTSD modules or control modules as reference. The preservation statistic describes the preservation of the corresponding module as compared to the reference. $Z_{summary} < 2$ implies no evidence for module preservation, $2 < Z_{summary} < 10$ implies weak evidence of preservation and $Z_{summary} > 10$ implies strong evidence for module preservation. $Z_{summary}$ was assessed using 200 permutations of the data. $Z_{summary}$ scores are ranked high to low. Grey shading is for visualization purposes only.

5.3.2. Differential Module Expression Post-Deployment in Dataset 1

WGCNA was used to construct a global gene co-expression network from a combination of PTSD cases (N = 47) and controls (N = 47) at post-deployment using RNA-Seq expression data from *Dataset 1* (**Figure 5.1**). This analysis identified nine modules which were further examined for enrichment of differentially expressed genes and subjected to clinical and functional annotation.

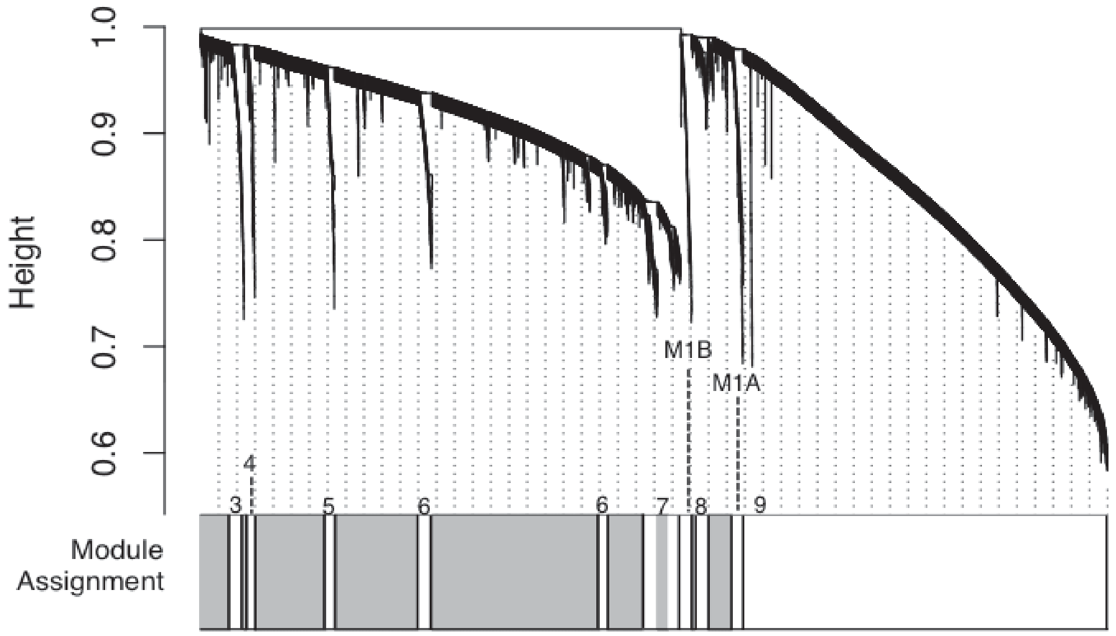


Figure 5.1. Hierarchical cluster tree (dendrogram) of the combine post-deployment network of PTSD cases (N=47) and controls (N=47) comprising 10,184 genes. Each line represents a gene (leaf) and each low-hanging cluster represents a group of co-expressed genes with similar network connections (branch) on the tree. The first band underneath the tree indicates the nine detected network modules.

Two modules (M1A and M1B) were enriched for genes identified as differentially expressed between PTSD cases and controls, reflected by an elevated module significance (*MS*) value (**Figure 5.2A**). To determine if the overall expression of modules M1A and M1B were significantly associated with PTSD group status, we calculated differences in module expression using module eigengene (*ME*) values. Consistent with results using MS, expression of module M1B was significantly higher in the PTSD resilient control group (*p*=0.004 and **Figure 5.2B**), meanwhile expression of module M1A was significantly higher in the PTSD group (*p*=0.02, **Figure 5.2B**).

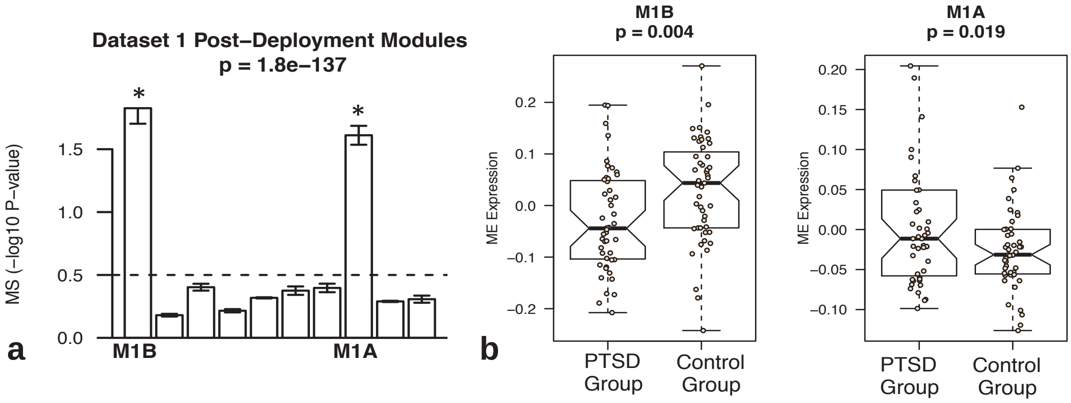


Figure 5.2. Module significance (MS) and module eigengene (ME) expression boxplots. (**a**) MS was measured across all post-deployment modules in Dataset 1. Here, a Kruskal-Wallis p-value was used only for descriptive purposes and not inferential. (**b**) Significant differences in ME expression were observed in post-deployment modules M1B and M1A. Differences in ME expression were measured using a two-tailed student's t test on and a p-value < 0.05 is considered significant.

Subsequently, *ME* values for each module were subjected to clinical annotation to determine module-trait relationships (**Table 5.4**). The *ME* for module M1B was significantly correlated to post-deployment resilient controls (*r*=0.29, *p*=0.005), negatively correlated to post-deployment CAPS and PCL (CAPs, *r* =-0.27, *p* =0.009; PCL r=-0.28, p=0.007) and negatively correlated with other measures of CAPS (**Table 5.4**) but not correlated to any other measured clinical variable, suggesting that differential gene expression in M1B was not confounded by recorded measurements such as body-mass-index, smoking, or alcohol consumption. Conversely, the *ME* for module M1A was significantly correlated to PTSD cases (*r* = 0.23, *p* = 0.03), post-deployment CAPs criteria of avoidance (CAPSCA, *r* = 0.32, *p* = 0.002) and post-deployment CAPs criteria of re-experiencing (CAPSBs, *r*=0.2, *p*=0.05) but to no other variables (**Table 5.4**).

Genes in M1B were expressed to a greater extent in resilient controls (**Figure 5.2B**) while enrichment analysis revealed a significant association to terms including hemostasis, platelet activation and wound healing (**Figure 5.3A**). Further, enrichment for cell-type specificity revealed an over-representation of erythroid expression markers (blood platelets). Hub genes are those most strongly correlated to the *ME* value for a particular module and represent possible disease associated markers[13], in this case putative PTSD-resiliency markers. The top 5 hub genes in M1B (*C6orf25, CTDSPL*, *ITGB3, PRKAR2B* and *TUBB1*) were are all associated with hemostasis and in particular, with platelet regulation and function (Zarbock et al., 2007; Beck et al., 2014; Daly, 2010; Raslova et al., 2007) (**Figure 5.3B**). Additionally, enrichment analysis for M1B revealed a significant association with immune response as exemplified by innate responses mediated by interferon (IFN) signaling (**Figure 5.3C**), as well as with monocyte specific markers. The top 5 hub genes in M1A included *IFI35, IFIH1, PARP14, RSAD2 and UBE2L6*; all well described interferon stimulated genes (Rusinova et al., 2013) and here considered putative PTSD-associated markers (**Figure 5.3D**).

Table 5.4. Complete network characterisation of post-deployment modules within *Dataset 1*.

Module	Genes(*n*):	Top Significant Biological Process	Top Significant Pathway	Top Significant Cell-Type	Significant ME Correlations Condition or Trait	(*R*, *P*-value)
M1A	115 (*69)	Immune Response	Interferon Signalling	CD14$^+$ Monocytes	PTSD Group CAPsBs CAPsSCAs	(0.23, 0.03) (0.2, 0.05) (0.32, 0.002)
M1B	118 (*74)	Coagulation	Hemostasis	(Blood Platelets) CD71+ Early Erythroid	Control Group PCL CAPs CAPsBs CAPsSCN1 CAPsDs CAPsCs	(0.29, 0.005) (-0.28, 0.007) (-0.27, 0.009) (-0.24, 0.02) (-0.23, 0.02) (-0.23, 0.03) (-0.25, 0.01)
3	146 (*3)	-	-	-	CES PBE	(0.29, 0.005)
4	80 (*1)	M Phase	Cell Cycle	(Blood Platelets) CD71+ Early Erythroid	-	-
5	85 (*1)	B cell activation	-	CD19$^+$ B Cells	Tobacco CES	(-0.22, 0.03) (-0.21, 0.04)
6	217	Cellular Defence Response	-	CD56$^+$ NK Cells	Tobacco Prior Deployment Ethnicity (C)	(-0.2, 0.05) (0.25, 0.01) (-0.25, 0.02)
7	283	Translation Elongation	Eurkaryotic Translation Elongation	CD4$^+$ T cells	Prior Deployment	(0.2, 0.05)
8	146 (*2)	-	-	CD14$^+$ Monocytes	Alcohol Ethnicity (AA)	(0.26, 0.01) (-0.26, 0.01)
9	4090 (*60)	Intracellular Signalling Cascade	Signalling by Interleukins	CD33$^+$ Myeloid	Prior Deployment	(-0.26, 0.01)

The first column represents the identified modules. The second column represents the number of genes within each module and numbers denoted as (*) reflect the number of significantly differentially expressed genes within each particular module. The third, fourth and fifth column represent significantly overrepresented biological processes (annotated with DAVID), pathways (annotated with REACTOME) and cell-types (annotated with CTen) for each module, respectively. All significant terms were not included as to reduce redundancy. In the sixth column, ME values were correlated to clinical parameters, and only the significant correlations ($p < 0.05$) are reported. Abbreviations: Ethnicity (C) = Caucasian, Ethnicity (AA) = African American; Ethnicity (AM) = American Mexican; Ethnicity (A) = Asian. All other abbreviations found in Table 1. Grey shading is for visualisation purposes only.

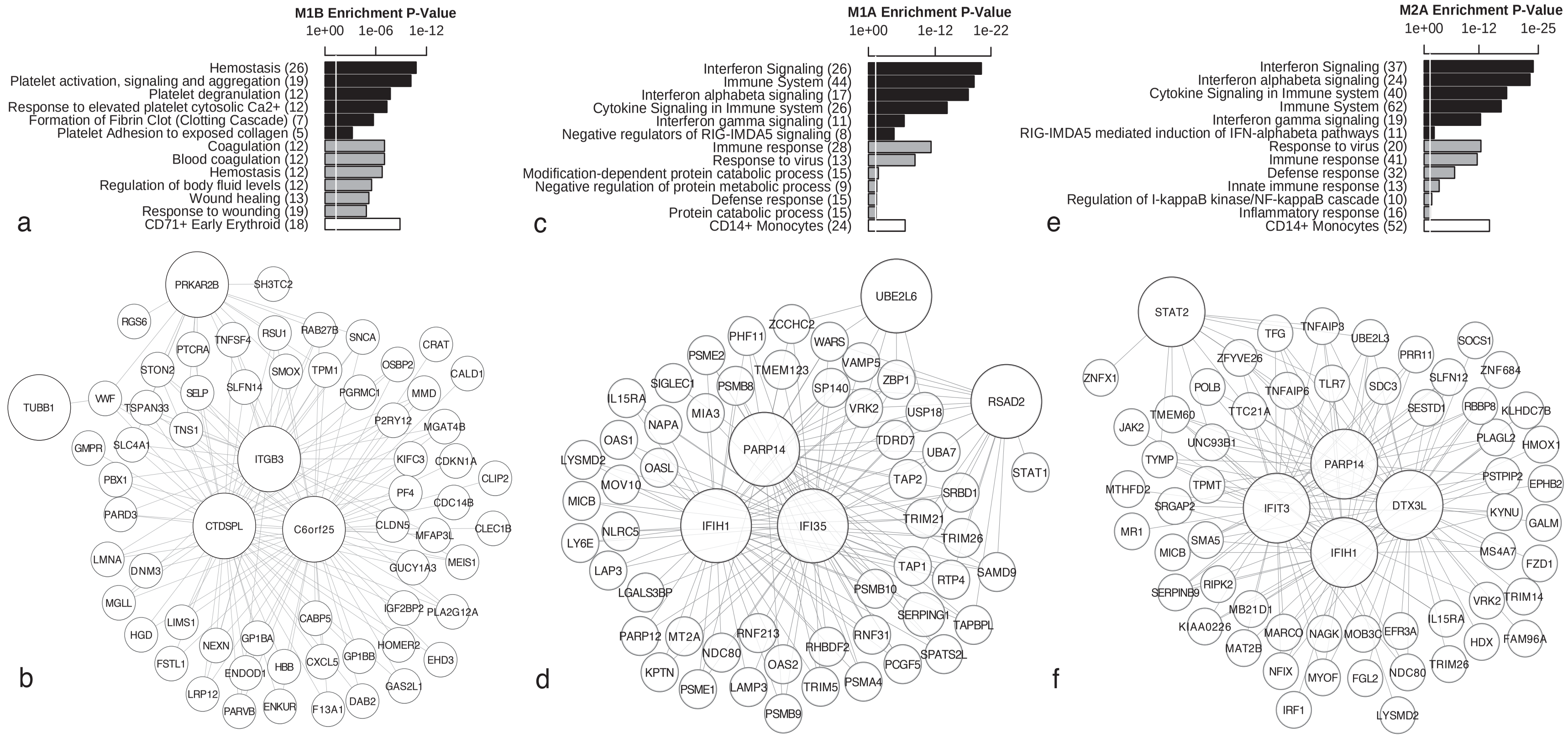


Figure 5.3. Module characterization for *Dataset 1.* Enrichment analysis and correlation networks for modules M1B (**a & b**) and M1A (**c & d**) identified post-deployment, and module M2A (**e & f**) identified pre-deployment in *Dataset 1*. Enrichment analysis was used to identify the top 6 ‘specific’ REACTOME ontology terms (black bars), the top 6 ‘broad’ DAVID ontology terms (grey bars) and the most significant cell-type signature (white bar) over-represented in the list of genes within each module. All terms were deemed significant as assessed by a hypergeometric test FDR corrected *p*-value <0.05 displayed as a white line. The total number of genes within each significant term is denoted within the brackets associated with that term. Gene-networks were constructed selecting the top 150 most significant connections ranked by *kME*. Nodes represent genes and edges represent correlations. The top 5 hub genes, those most correlated to *ME* values, are shown in larger sizes.

5.3.3. Differential Module Expression Pre-deployment in Dataset 1

It is unclear whether the modules identified post-deployment are involved in causing PTSD development or are simply a consequence of the disorder. To determine if any post-deployment modules could be re-identified and thus denoted as causal modules, we constructed a gene co-expression network combining RNA-Seq gene expression data from PTSD-risk cases (N = 47) and controls (N = 47) pre-deployment in *Dataset 1*. Twenty-two pre-deployment modules were identified, examined for enrichment of differentially expressed genes and then subjected to functional and clinical annotation (**Supplementary Table 5.1**). A single module (M2A) was enriched for differentially expressed genes between PTSD-risk participants and controls as reflected by an elevated MS value (**Figure 5.4A**). Along the same lines, M2A module expression was significantly higher in the PTSD risk group (*p*=0.001 and **Figure 5.4B**). Module M2A *ME* was significantly correlated to one variable, PTSD-risk (*r*=0.32, *p*=0.002). Similar to module M1A that was identified post-deployment, enrichment analysis of genes in M2A revealed a significant association with innate immune responses, IFN signalling and monocyte specificity (**Figure 5.3E**). The top 5 hub genes were again associated with IFN signalling (*DTX3L, IFIH1, IFIT3, PARP14* and *STAT2)* (**Figure 5.3F**). Gene-set overlap analysis compared all of the genes in M2A pre-deployment (n=245) to those in M1A post-deployment (n=115) to reveal a significant overlap (∩ = 108, *p = 6.7e-181*).

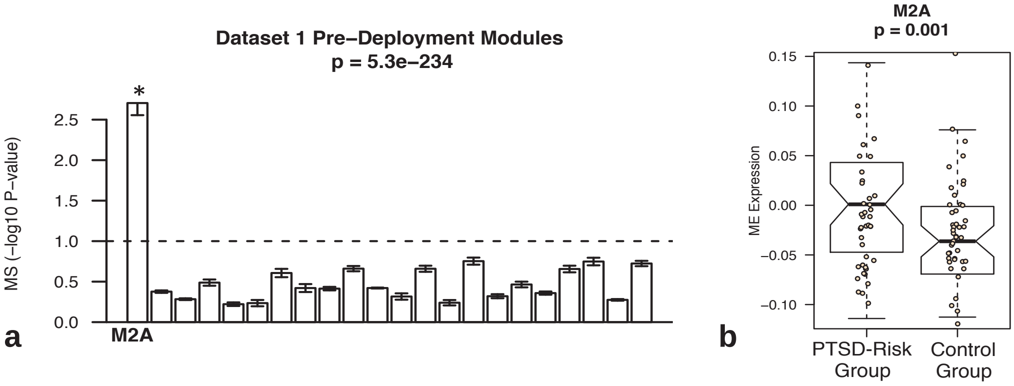


Figure 5.4. Module significance (MS) and module eigengene (ME) expression boxplots. (**a**) MS was measured across all pre-deployment modules in Dataset 1. Here, a Kruskal-Wallis p-value was used only for descriptive purposes and not inferential. (**b**) Significant differences in ME expression were observed in pre-deployment module M2A. Differences in ME expression were measured using a two-tailed student's t test on and a p-value < 0.05 is considered significant.

5.3.4. External Validation of Post-Deployment Modules in Dataset 2

To validate post-deployment findings in *Dataset 1* we assessed *Dataset 2* for similar network properties in a combined network analysis of PTSD cases (N = 24) and controls (N = 24) post-deployment. Out of 8 modules (full characterisation **Supplementary Table 5.2**), a single module (M3A) contained an enrichment of differentially expressed genes (**Figure 5.5A**) demonstrating a modest, yet insignificant, increase in module expression within the PTSD group (*p* = 0.1, **Figure 5.5B**). The *ME* was significantly correlated to post battle experience (*r* = 0.4, *p* = 0.004) and post-deployment CAPS (*r*=0.32, *p*=0.03) but weakly correlated to PTSD caseness (*r* = 0.21, *p* = 0.1). The genes in this module were over-expressed in PTSD cases relative to controls (**Figure 5.5B**) and enrichment analysis revealed a significant association with innate immune responses, IFN signalling and monocytes (**Figure 5.7A**). The top 5 hub genes (*DDX58, IFI35, IFIT5*, *PARP9 and ZBP1*) were again all associated with IFN signalling (**Figure 5.7B**). A highly significant overlap in post-deployment module genes across M1A (n=115) in *Dataset 1* and M3A (n=83) in *Dataset 2* (∩ = 63, *p* = 2.0E-105) confirmed the identification of a dysregulated innate immune module related to PTSD cases across two independent datasets.

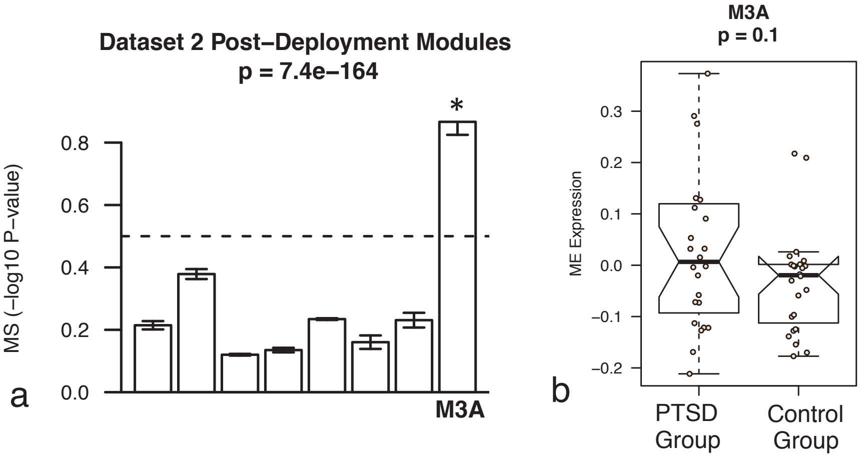


Figure 5.5. Module significance (MS) and module eigengene (ME) expression boxplots. (**a**) MS was measured across all post-deployment modules in Dataset 2. Here, a Kruskal-Wallis p-value was used only for descriptive purposes and not inferential. (**b**) Significant differences in ME expression were observed in post-deployment module M3A. Differences in ME expression were measured using a two-tailed student's t test on and a p-value < 0.05 is considered significant.

5.3.5. External Validation of Pre-Deployment Modules in Dataset 2

To re-confirm pre-deployment findings from *Dataset 1*, PTSD-risk cases (N=24) and controls (N=24) pre-deployment were combined from *Dataset 2* and subjected to network analysis which identified 11 modules (full characterisation in **Supplementary Table 5.3**). A single module (M4A) was enriched for differentially expressed genes between PTSD-risk cases and controls (**Figure 5.6A**). The PTSD-risk group displayed a significant over-expression of module expression (*p* = 0.01, **Figure 5.6B**). The *ME* for M4A was significantly correlated to PTSD-risk (*r* = 0.36, *p* = 0.01) and CAPs (*r*=0.44, *p*=0.002). Moreover, enrichment analysis of M4A revealed a significant association with innate immune responses, IFN signaling and monocytes (**Figure 5.7C**), and the top 5 hub genes (*PARP9*, *UBE2L6, STAT2, TRIM22* and *GBP1*) were again all associated with IFN signaling (**Figure 5.7D**). All pairwise gene-set overlap analyses across modules M1A, M2A, M3A and M4A revealed a highly significant overlap (**Figure 5.8**) and hub gene expression for these modules showed elevated expression in PTSD groups when compared to controls both pre- and post-deployment across both datasets. These results demonstrate the association of a dysregulated innate immune module, related to IFN signaling, which appears to define at least part of the pathophysiology of PTSD through causal association to PTSD development.

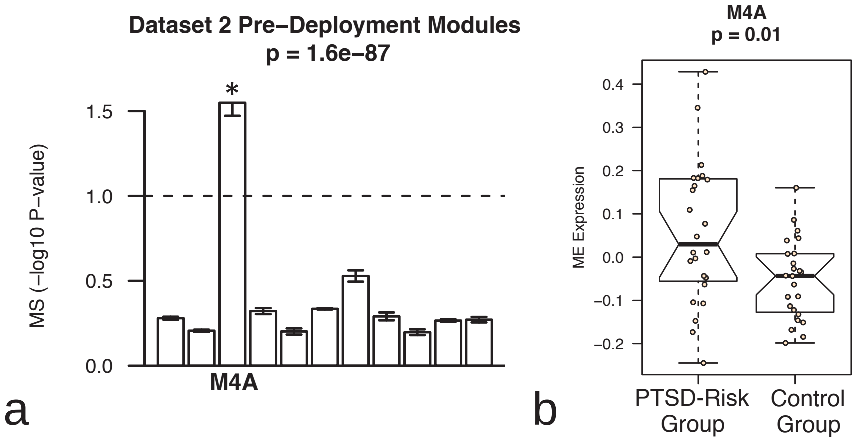


Figure 5.6. Module significance (MS) and module eigengene (ME) expression boxplots. (**a**) MS was measured across all pre-deployment modules in Dataset 2. Here, a Kruskal-Wallis p-value was used only for descriptive purposes and not inferential. (**b**) Significant differences in ME expression were observed in pre-deployment module M4A. Differences in ME expression were measured using a two-tailed student's t test on and a p-value < 0.05 is considered significant.

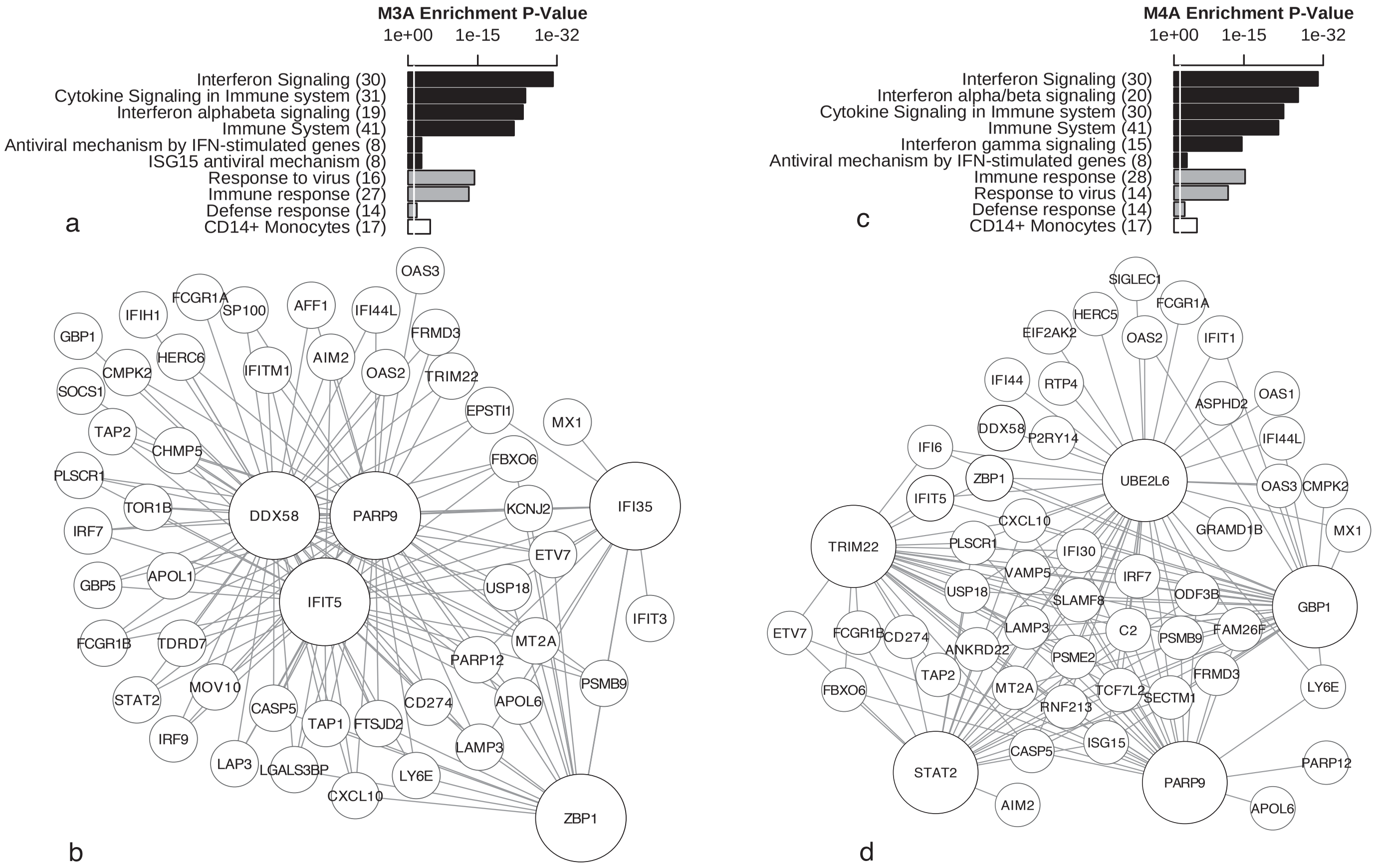
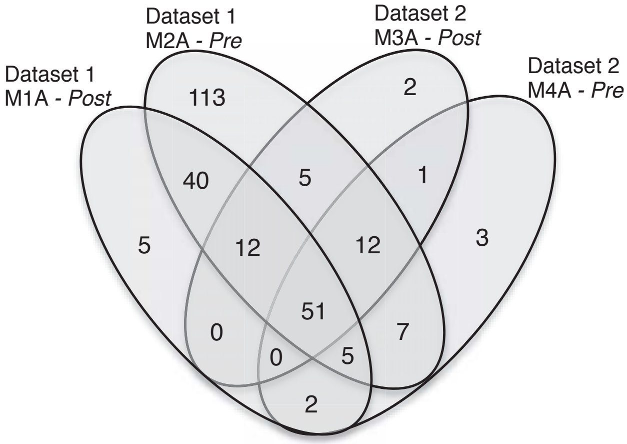


Figure 5.7. Enrichment analysis and correlation networks for module M3A (**a & b**) identified post-deployment and module M4A (**c & d**) identified pre-deployment in Dataset 2. Enrichment analysis was used to identify the top 6 'specific' Reactome ontology terms (black bars), top 6 'general' DAVID ontology terms (grey bars) and the top cell-type signature (white bar) over-represented in the list of genes in each module. All terms were deemed significant as assessed by a hypergeometric test (FDR corrected p-value <0.05). Gene-networks were constructed selecting the top 150 most significant connections. Nodes represent genes and edges represent correlations. The top 5 hub genes with the highest correlation with *ME* are shown in larger sizes.



a

	M1A	M2A	M3A	M4A
M1A	-	*6.7E-181*	*2.0E-105*	*1.0E-134*
M2A	**108** ∩	-	*6.3E-134*	*2.4E-121*
M3A	**63** ∩	**80** ∩	-	*8.8E-152*
M4A	**58** ∩	**75** ∩	**69** ∩	-

b

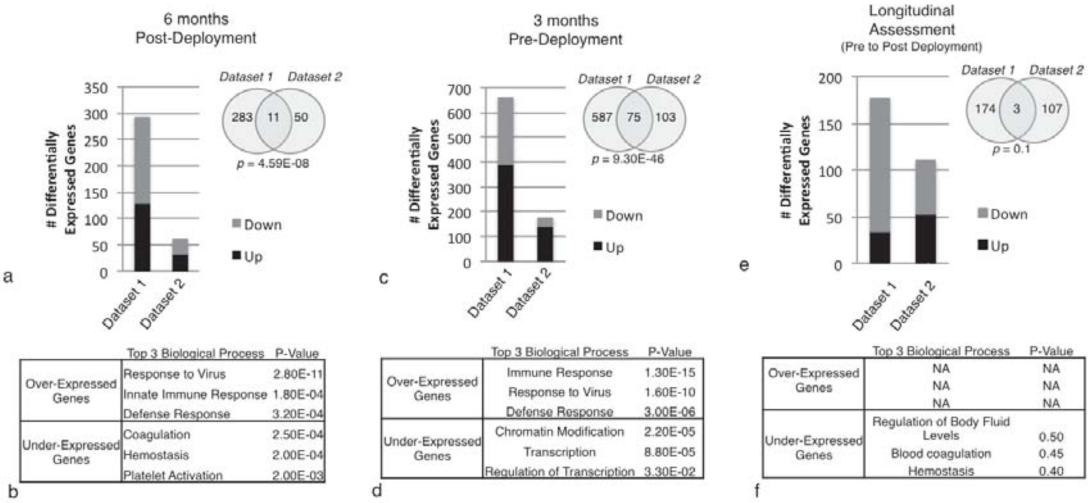
	kME Rank			
Gene Symbol	M1A	M2A	M3A	M4A
IFIH1	**3**	**1**	12	**7**
STAT2	**6**	**2**	19	**3**
PARP14	**4**	**3**	22	40
DTX3L	39	**4**	46	39
IFIT3	**10**	**5**	**8**	**10**
IFI35	**2**	**6**	**4**	52
UBE2L6	**1**	**7**	18	**2**
PARP9	47	**8**	**2**	**1**
TRIM22	31	**9**	14	**4**
DDX58	36	**10**	**3**	26
TRIM5	34	11	41	16
CMPK2	34	12	51	34
IFIT5	17	13	**1**	21
RSAD2	**5**	14	35	48
HERC5	11	15	17	14
IFI6	13	17	32	43
OAS3	15	18	15	44
IRF9	25	19	44	62
IFIT2	37	20	24	45
IFIT1	16	25	26	49
SERPING1	30	26	34	**6**
STAT1	21	27	30	23
GBP1	43	28	**10**	**5**
IFI44	24	32	**9**	20
SAMD9L	14	33	13	12
PML	**8**	34	28	32
ZBP1	51	35	**5**	38
APOL6	31	36	36	11
APOL1	30	37	67	75
MX1	38	38	**7**	37
IFI44L	28	39	58	65
DDX60	37	40	16	28
BATF2	32	43	83	50
OASL	40	44	62	69
EPSTI1	40	45	42	25
FBXO6	42	47	56	15
LAP3	70	50	11	41
OAS2	46	52	21	18
TAP1	63	58	31	**8**
PARP12	33	63	23	27
RTP4	55	65	27	13
TAP2	45	67	53	70
SPATS2L	67	69	64	81
CXCL10	35	70	57	68
LY6E	65	75	37	64
OAS1	74	84	43	66
DHX58	38	93	47	29
USP18	66	97	82	82
CD274	33	121	55	30
MOV10	94	123	60	31
ETV7	41	128	79	74
GBP5	44	151	38	**9**

c

Figure 5.8. Venn Diagram of Innate Immune Co-expression Modules across *Dataset 1* and *Dataset 2*. Venn Diagram (**a**) depicting significant overlap in genes belonging to modules M1A post-deployment and M2A pre-deployment in *Dataset 1* as well as modules M3A post-deployment and M4A pre-deployment in *Dataset 2*. Gene overlap (∩) with associated hypergeometric p-value, in italics, are depicted for all pairwise comparisons of module genes (**b**). The overlap identified 51 genes found across all four analyses (**c**) which are displayed in the table along with the corresponding kME rank (i.e. rank of connectivity) for each gene within a particular module. A high rank indicates hub gene status (i.e. PTSD risk and PTSD associated markers). Numbers in bold outline the top 10 hub genes across each module, respectively. Genes are ordered accordingly to M2A kME.

5.3.6. Cross-Sectional and Longitudinal Differential Gene Expression

Differential gene expression analyses revealed that most changes were observed cross-sectionally at pre-deployment and post-deployment, rather than occurring across time-points in a longitudinal fashion (**Figure 5.9**). These genes were also subjected to functional annotation and the top three most significant terms are reported based on over- and under-expressed genes for each time-point. As anticipated, all functional ontology terms overlap with findings from differential module expression analyses (Aim 2).



	Top 3 Biological Process	P-Value
Over-Expressed Genes	Response to Virus	2.80E-11
	Innate Immune Response	1.80E-04
	Defense Response	3.20E-04
Under-Expressed Genes	Coagulation	2.50E-04
	Hemostasis	2.00E-04
	Platelet Activation	2.00E-03

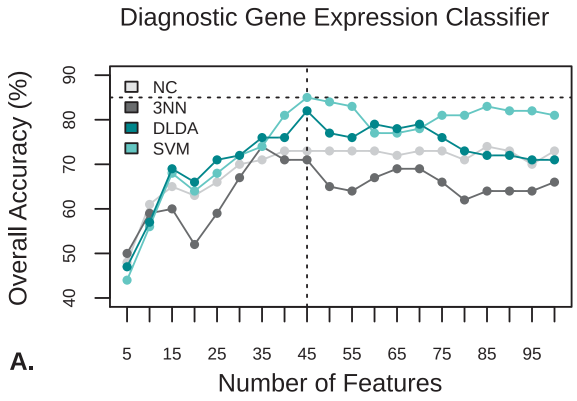
	Top 3 Biological Process	P-Value
Over-Expressed Genes	Immune Response	1.30E-15
	Response to Virus	1.60E-10
	Defense Response	3.00E-06
Under-Expressed Genes	Chromatin Modification	2.20E-05
	Transcription	8.80E-05
	Regulation of Transcription	3.30E-02

	Top 3 Biological Process	P-Value
Over-Expressed Genes	NA	NA
	NA	NA
	NA	NA
Under-Expressed Genes	Regulation of Body Fluid Levels	0.50
	Blood coagulation	0.45
	Hemostasis	0.40

Figure 5.9. Differential gene expression analyses were performed using a moderated *t* statistic within *Dataset 1* and *Dataset 2*. A cross-sectional analysis compared PTSD cases to controls post-deployment in *Dataset 1* (47 PTSD cases vs. 47 controls) and subsequently in *Dataset 2* (24 PTSD cases vs. 24 controls) to reveal 294 and 61 differentially expressed genes, respectively, with a significant overlap (**a**). The top 3 most significant biological processes (annotated with DAVID) based on over-expressed and under-expressed genes identified from *Dataset 1* are reported (**b**). Subsequently, the same paired data were analyzed pre-deployment for *Dataset 1* and *Dataset 2* revealing 662 and 178 differentially expressed genes, respectively, with a significant overlap (**c**). The top 3 most significant biological processes from *Dataset 1* are reported (**d**). Utilizing, the paired structure of the data, a longitudinal contrast analysis was applied to identify genes behaving differently across time within the PTSD and control groups. This analysis revealed a total of 177 genes in *Dataset 1* and 110 genes in *Dataset 2*, with minimal overlap **(e)** and no significant functional annotation **(f)**. Up and down symbols are relative to the PTSD group. Significance threshold for genes was set to a nominal p < 0.05 where as we used a more strict threshold for functional annotations with a Bonferroni corrected p-value < 0.05.

5.3.7. Putative Diagnostic PTSD Gene Expression Classifier

To identify a panel of biomarkers capable of confirming PTSD at post-deployment, we used four different supervised multivariate classification algorithms using a LOOCV on Dataset 1 (N=94) and subsequently externally cross-validated prediction accuracies on a left out test-set, Dataset 2 (N=48). Classification accuracies reached 85% when the expression of 45 genes was used with SVM multivariate classification method using LOOCV on Dataset 1 (**Figure 5.10A&B, Supplementary Table 5.4**). Classification accuracies for the 45 gene classifier were subjected to external validation on Dataset 2 where classification accuracies only reached 45% (**Figure 5.10C**). A total of 5 genes from the 45 gene classifier overlapped with post-deployment module M1A.



B.	DLDA	3NN	NC	SVM
Sensitivity	0.809	0.787	0.766	0.809
Specificity	0.83	0.638	0.702	0.894
PPV	0.826	0.685	0.75	0.884
NPV	0.812	0.75	0.72	0.824
Overall Accuracy	0.71	0.66	0.73	0.81

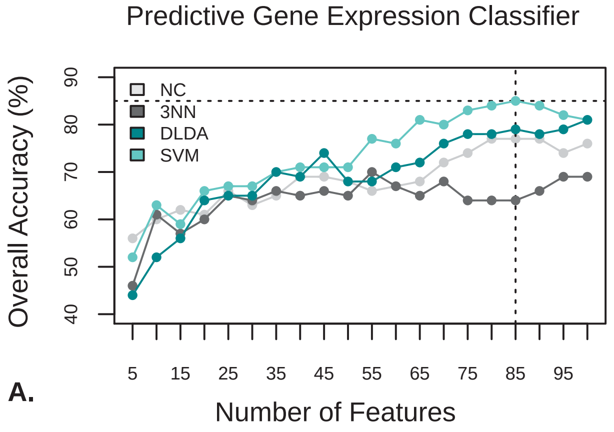
C. External Validation on Dataset 2

Correct Calls: PTSD	12	14	9	11
Correct Calls: Control	13	10	11	11
Overall Accuracy	0.52	0.5	0.41	0.458

Figure 5.10. Diagnostic gene expression classifier construction on dataset 1 post-deployment participants. (**a**) Four different multivariate classification algorithms were used with RFE feature selection and (**b**) accuracies were evaluated with a LOOCV and subsequently (**c**) using a left out test set, Dataset 2. Abbreviations; PPV, positive predictive value; NPV, negative predictive value; DLDA, diagonal linear discriminate analysis; 3NN, three nearest neighbors; NC, nearest centroid; SVM, support vector machines.

5.3.8. Putative Predictive PTSD Gene Expression Classifier

Similarity, we sought to identify a unique panel of biomarkers capable of predicting the eventual development of PTSD at pre-deployment using the same methodologies as at post-deployment. Classification accuracies reached 85% when the expression of 85 genes was used with SVM multivariate classification method and LOOCV on Dataset 1 (**Figure 5.11A&B**). Classification accuracies for the 85 gene classifier were subjected to external validation on Dataset 2 where classification accuracies reached 70% (**Figure 5.11C, Supplementary Table 5.5**). A total of 34 genes from the 85 gene classifier overlapped with pre-deployment module M2A, more than expected by chance (*p*=0.005).



B.	DLDA	3NN	NC	SVM
Sensitivity	0.723	0.574	0.766	0.851
Specificity	0.851	0.702	0.766	0.851
PPV	0.829	0.659	0.766	0.851
NPV	0.755	0.623	0.766	0.851
Overall Accuracy	0.79	0.64	0.77	0.85

C. External Validation on Dataset 2

Correct Calls: PTSD	15	14	18	17
Correct Calls: Control	16	14	11	17
Overall Accuracy	0.64	0.58	0.6	0.7

Figure 5.11. Predictive gene expression classifier construction on dataset 1 pre-deployment participants. (**a**) Four different multivariate classification algorithms were used with RFE feature selection and (**b**) accuracies were evaluated with a LOOCV and subsequently (**c**) using a left out test set, Dataset 2. Abbreviations; PPV, positive predictive value; NPV, negative predictive value; DLDA, diagonal linear discriminate analysis; 3NN, three nearest neighbors; NC, nearest centroid; SVM, support vector machines.

5.4. Discussion

Gene expression data were generated by RNA-Seq (*Dataset 1 N=188*) and microarray (*Dataset 2 N=96*) using peripheral blood samples isolated from U.S. Marines pre- and post-deployment to conflict zones (Iraq and Afghanistan). Our prospective experimental design allowed for the identification of candidate PTSD biomarkers, and permitted the re-confirmation of findings in an independent dataset. Our methodological aims focused our genome-wide analysis at the higher-order gene network level, with further investigation of differences at the individual gene level. We were able to rule out *large* changes in the underlying gene-gene connectivity within peripheral blood as a basis for the pathology of PTSD however *subtle* changes in the expressions of gene networks may provide a useful indicator for PTSD risk and development. More specifically, these tests revealed, for the first time, the identification of dysregulated modules specific for innate immunity capable of characterizing causal and consequential molecular signatures of PTSD, and then further replicated these findings across independent datasets.

5.4.1. Gene Networks Specific for Innate Immunity in PTSD

Our central finding was the identification of a dysregulated innate immune module associated with the development of PTSD (**Figures 5.2-5.6, Supplementary Figure 5.2**), illuminated by the replication of modules post-deployment (M1A and M3A) and those pre-deployment (M2A and M4A) that could be associated with PTSD. These findings suggest that differences in innate immunity modules were not simply a consequence of the PTSD state after deployment but also have causal relevance for PTSD development and may therefore at least partly explain the pathophysiology of the disorder, exemplified by their identification pre-deployment. These results highlight our differential expression analyses (**Figure 5.9**) and our previous reports of C-reactive protein (CRP), a general marker of immune activation and inflammation, and 5'-oligoadenylate synthetase genes (*i.e. OAS1, OAS2, OAS3)* as markers of the antiviral interferon response, that were associated with an increased risk of developing PTSD (Eraly et al., 2014; Glatt et al., 2013). However, our current findings dramatically extend these results by showing that the IFN response is being modulated to a much greater extent than previously thought pre- and post-deployment. A number of single case studies have reported that treatment of PTSD subjects infected with hepatitis C virus (HCV) with recombinant interferon (IFN- α2b) worsened PTSD symptoms (Maunder et al., 1998; Dieperink et al., 2008). In our study, where subjects were not receiving IFN therapy, it is unclear what is stimulating the IFN response.

5.4.2. Predictive and Diagnostic Gene Expression Classifiers

Both a predictive and a diagnostic biomarker panel were identified through supervised multivariate classification methods. Overall classifier accuracies for a diagnostic classifier for Dataset 1 at post-deployment were high using SVM (85%), but external cross-validation on Dataset 2 were sub-optimal with 48% accuracy. Predictive classifier accuracies for Dataset 1 at pre-deployment were also high using SVM (85%), while external cross-validation on Dataset 2 were better than expected by chance at 70% - where the translational value of preventing PTSD development before onset is relevant. Pre-deployment results also re-affirm genes specific to innate immunity that were identified from our network analyses. Indeed, a total of 34 genes from a unique panel of 85 putative predictive genes overlapped with pre-deployment module M2A, more than expected by chance (*p*=0.005). However immune-gene dysregulation may be only one piece of the biological puzzle of PTSD susceptibility, as many genes comprising the best-performing PTSD-predictive and –diagnostic classifiers were not immune-system genes (**Supplementary Tables 5.4-5.5**). Additionally, because of the heavy amount of statistical analysis prior to biomarker discovery, these results should be interpreted cautiously as a fair amount of technical variation exists between Dataset 1 (RNA-Seq) and Dataset 2 (Microarray) and needed to be accounted for prior to classifier construction (**Supplementary Figure 5.1**). Furthermore, classifiers constructed of both clinical and molecular for predictive and diagnostic purposes may provide interesting future avenues. For example, PCL scores differed significantly between eventual PTSD and controls at pre-deployment (*p*=0.001) in Dataset 1, and this may yield clinically relevant sensitivity for a putative classifier.

5.4.3. Interpreting Blood-based Innate Immune Signatures in PTSD

Our observations lead to several questions and some potential answers. First, how does one interpret the over-expression of innate immunity genes found prior-to trauma? One possible explanation is that both acute and severe stress, predictors in their own right for PTSD, are also associated with hyper-activation of the immune system and subsequent inflammation (Butcher et al., 2004; Clark et al., 2014). An alternative hypothesis is that stress, pathogens and/or high viral loads may ‘prime’ the immune system, driving the IFN response, altering a subsequent response to trauma. Studies focusing on the gut-brain barrier have shown that intestinal mucosal dysfunction, defined as increased translocation of gram-negative bacteria (‘leaky gut’), plays a role in the inflammatory pathophysiology of depression suggesting that differences in gut flora may stimulate an IFN response (Maes et al., 2008). Second, does a dysregulated innate immune module pre-deployment hold predictive value? Previous work constructing a prognostic classifier from *Dataset 2* pre-deployment participants (Glatt etl al., 2013) suggests that immune-related genes hold predictive value although these results have not yet been replicated across larger datasets using machine-learning methods. Inferring the prognostic relevance of network-based applications remains challenging. However, cross-referencing our findings with this previous work suggests that network statistics, and our innate immune modules, have predictive potential. Third, out of the entire network of pairwise correlations between genes across the transcriptome, are the most informative genes interconnected within similar modules or spread out across numerous modules? A possible limitation of this study was that by analyzing co-regulated modules of genes we may have missed individual genes which do not correlate within our modules of interest although are of functional relevance to PTSD. For example, previous reports specifically target *FKBP5* and *STAT5B* as differentially expressed biomarkers (Ziker et al., 2007; Yehuda et al., 2009; Neylan et al., 2011; Sarpas et al., 2011; Mehta et al., 2011; Pace et al., 2012) although they were not assigned to co-expressed modules nor found to be significantly differentially expressed between PTSD cases and controls. Finally, of what relevance is PBL gene expression for a disorder primarily associated with the brain? In this study we identify innate immunity and IFN signaling genes whose expression was elevated in PBLs both before and after the development of PTSD. Although the recruitment of such signaling could be triggered by various factors, they ultimately release toxic compounds including degradative enzymes and reactive oxygen species that can impair cellular processes (Aiboshi et al., 2001; Veldhuis et al., 2003; Bhatia et al., 2004). It could be hypothesized that the accumulation of these compounds in the blood prior-to-deployment may be detrimental to the brain if the integrity of the blood-brain-barrier (BBB) was compromised by injury. An increasing body of evidence indicates that changes in the blood may seed pathology in the brain across various disorders. Investigation in multiple sclerosis (Minagar and Alexander, 2003) of the association of INF with the BBB suggests that IFN-γ and other proinflammatory cytokines (TNF-α and IL-1β) disrupt the BBB through a variety of mechanisms. Further, Alzheimer's disease models suggest that breaches in the BBB can lead to 'leakage' into the brain of blood-borne molecules that are toxic to neurons and cause neurodegenerative changes (Carmeliet & Strooper, 2012). Future studies investigating the role of the BBB in PTSD may provide a detailed explanation for a specific course of PTSD development.

5.4.4. The Hemostatic Response to PTSD Development

A novel finding was the identification that modules related to hemostasis and wound responsiveness were expressed to a greater extent post-deployment in US Marines who did not develop PTSD (**Figure 5.2**), as in module M1B (**Figure 5.3A&B**). Interestingly, the three other network analyses also detected modules related to hemostasis and wound response with significant overlap (M16 pre-deployment *Dataset 1*; M7 and M6 indented post- and pre-deployment in *Dataset 2*; **Supplementary Figure 5.3**). These other modules revealed patterns of heterogeneous gene expression irrespective of group status and time-point suggesting that these modules and corresponding processes may infer wound resilience in only a sub-set of individuals. It has been well documented that different degrees of stress will elicit different stress responses (Pacak, 2001), and in particular, a response involving blood platelets has been shown to be a critical biomarker of hemostatic, thrombotic, and inflammatory challenges to an organism and a key player in cardiovascular disease and chronic stress, as in PTSD (Bray et al., 2013; Austin et al., 2013). Moreover, in a review of a large number of studies examining various tissue types, it was found that different types of psychological stress were associated with impaired wound healing (Walburn et al., 2009). A meta-analysis found an inverse correlation (r = -0.42) between psychological stress and wound healing (Goulin et al., 2011) supporting the positive association between wound healing and resilience against PTSD (r =0.29, p =0.005) found in this study. This suggests that high levels of stress may hinder proper wound healing during/after battlefield trauma, although the degree of such stress appears to be a key factor for establishing associations with the hemostatic system.

5.4.5. Strengths and Limitations

A main strength of this study is the longitudinal and multi-level (two independent cohorts) experimental design which permitted the testing of numerous hypothesizes. Additionally, our cohorts were comprised of well-defined groups and equally balanced for pre-deployment CAPs scores. Although there were limitations. First, gene significance was low for a genome-wide study. Most genes we call ‘statistically significant’ pass only a nominally significant $p < 0.05$. Despite, analysis of higher-order gene co-expression modules permitted for the identification of reproducible small collective changes in modules specific to innate immunity. Second, our putative gene expression classifiers at pre- and post-deployment should be interpreted cautiously due to necessary statistical adjustments made in order to merge the two datasets, which took place prior to classifier construction. Third, we were unable to rule out the possibility of PTSD co-morbidities (e.g. depression, agoraphobia, etc…) contributing to the observed gene expression results. Future studies should focus on obtaining a more clinically heterogeneous cohort, or recording clinically relevant co-morbidities, which may permit for a biomarker to discriminate between lesser/greater degrees of illness and relative co-morbidities.

5.4.6. Concluding Remarks

Our data provide a broad framework for previously unknown molecular aspects of PTSD and provide a new context concerning the complex nature of PTSD development. Specifically, modules of co-expressed genes associated with the innate immune response and IFN signaling appear to be implicated in the development of PTSD and persist once the disorder is established. Modules associated with hemostasis and wound healing may contribute to resilience against developing PTSD. This study may encourage further work examining differences in innate immune factors for the development of PTSD and the potential role of platelets in the stress response. This could in turn allow for advanced PTSD prevention and detection, by identifying susceptible service members prior to deployment to conflict zones, by removing the causal path (i.e. trauma exposure), or through implementation of novel targeted therapies to modulate the interferon signature.

Contributions. *These results are predominately my own work. I was fully responsible for all* *blood handling, RNA treatment, statistical design, data analysis, data interpretation and writing.*

Chapter 6

Immediate Molecular and Cellular Response to Acute Psychological Stress

6.1. Background

While chronic stress-related effects upon the immune system are deleterious, acute stress appears to have both protective and adverse effects. For example, acute stress can enhance the acquisition and expression of immunoprotection by activation of bodily defences prior to wounding or infection (Ackerman et al., 2002; Amkraut et al., 1971; Carney, 2004; Dhabhar, 2009), or alternatively induce immunopathology via exacerbating autoimmune inflammation, with respiratory and cardiovascular consequences (Al'Abadie et al, 1994; Black, 2006; Bosch et al., 2003; Dhabhar et al., 1995; Garg et al., 2001). A more detailed understanding of immunomodulation throughout acute stress in humans is necessary not only to clinically reduce immunopathology, but also to harness stress-related immunoprotective effects.

One primary mechanism by which acute psychological stress induces an immune response is through rapid changes in leukocyte distributions in the peripheral circulation (Bosch et al., 2008). Studies investigating acute short-term stressors in humans, such as public speaking, have reported brief increases of natural killer (NK) cell numbers and other leukocyte subtype cell numbers, a reduction in lymphocyte proliferation, an increase in pro-inflammatory cytokine production, and reduced healing capacity of the skin (Altemus et al., 2001; Segerstrom and Miller, 2004). Studies of acute (psychological) stress due to physical danger have used first-time tandem skydive (Schedlowski et al., 1993), as this challenge has the advantage of representing real risk and eliciting reliable effects, yet permitting a high degree of experimental control. Studies using this paradigm have found transient increases of T cells and NK cells in the blood, as well as a parallel increase in NK cell cytotoxic activity. This suggests that changes in leukocyte numbers may be an important mediator of apparent changes in leukocyte activity. Comparably, an equivalent study of bungee jumping reported increases in neutrophils, pro-inflammatory monocytes, and CD8$^+$ T cell numbers following the jump (van Westerloo et al., 2001). While these studies are suggestive, one important limitation has been the lack of molecular and computational approaches for large-scale immune system monitoring. As a result, the molecular and cellular response underlying the rapid adaption to the immune system to an acute stressor is still incompletely defined.

To this end, this exploratory study examined the detailed molecular and cellular response of the immune system throughout the sequence of events leading up to, during, and after short-term exposure to physical danger in humans. Peripheral blood-based microarray transcriptome profiles were analysed and integrated with physiological measurements (endocrine and autonomic) collected longitudinally from 13 healthy participants (7 male, 6 female) at four different time-points throughout a first-time tandem skydive; (**1**) baseline, (**2**) pre-boarding, (**3**) post-landing, and (**4**) one-hour post-landing. These changes were compared to a second cohort of 26 healthy participants (17 male, 9 female) for which blood was collected and subjected to a detailed flow-cytometry analysis. This comprehensive and prospective experimental design allowed four main aims to be tested. **Aim 1:** First, to identify acute stress responsive genes (i.e. differentially expressed genes) in peripheral blood at pre-boarding, post-landing, and one-hour after the acute stress response relative to baseline measurements. **Aim 2:** Second, to identify patterns of co-expression in peripheral blood throughout these sequence of events and to determine their relationships with physiological measurements. **Aim 3:** Third, to assess gender-specific stress response differences at the molecular, cellular and physiological levels. **Aim 4**: Finally, to provide a comprehensive characterization of acute stress responsive peripheral blood leukocyte (PBL) cell types.

6.2. Materials and Methods

6.2.1. Ethical Approval

State University of New York at Stony Brook and the University of California San Diego Institutional Review Boards approved this study. Thirty-nine skydivers participated in this study consisting of 13 subjects for RNA expression profiles (7 male, 6 female) and 26 subjects for flow cytometry (17 male, 9 female). All skydivers provided written consent prior to participation. Participants were recruited from individuals who independently contacted an area skydiving school (Skydive Long Island, Calverton, NY) to schedule their first-time tandem skydive. Skydivers were healthy adult subjects with no history of cardiac or mental illness, as determined by physical examination, medical history, and screening using the Structured Clinical Interview for DSM-IV.

6.2.2. Subject Selection and Sample Collection Schedule

The study protocol adhered to a strict timeline for sample and data collection. Baseline blood samples were collected at 9:15 am within one week prior to or after the day of the skydive during a hospitalized testing that was time-locked to data collection during the skydive day and therefore served as a baseline and control. On the skydive day, all skydivers awoke at 6:30 am and arrived at Stony Brook University Hospital at 7:30 am. “Pre-boarding” samples were collected at 9:15 am, one hour before take-off. Take-off occurred at 10:15 am, and the jump occurred at 10:30 am when the airplane reached an altitude of 11,500 feet (3,505.2 meters). Skydivers landed at about 10:35 am and “post-landing” samples were collected at 10:45 am. Skydivers were immediately transported to Stony Brook University Hospital for a final blood draw at 11:30 am (“one hour post-landing” sample). Saliva was collected every 15 minutes from 9:15 am to 11:30 am on both the skydive and baseline hospital day.

6.2.3. RNA Isolation and Microarray Hybridization

Whole blood was obtained for 13 participants at baseline, pre-boarding, post-landing and one-hour post-landing. In total, 52 blood samples were collected across all time-points (i.e. 13 participants across 4 time-points). Each blood sample (10ml) was collected into an EDTA-coated collection tube, RNA was isolated from peripheral blood using LeukoLOCK Total RNA Isolation Kit and all samples passed a RNA integrity number (RIN) > 6. Synthesis of cDNA and biotinylated cRNA and hybridization of cRNA to Illumina HumanHT12 v4 BeadChips (47,231 probes). Because the integrity of RNA was of low quality for 2 samples (1 sample at pre-boarding and 1 sample at one-hour post-landing), these data were discarded and partially paired data were analyzed (50 samples total).

6.2.4. Data Pre-Processing

Quality control of microarray data, variance-stabilizing transformation (vst) normalization and removal of genes not expressed in any of the samples was performed in the R statistical computing environment version 2.8.0, using the Bioconductor package *lumi* (Du et al., 2008). Probes lacking gene symbol annotations were removed while probes with duplicate gene symbols were selected on the basis of having a higher average expression across all samples. This final filtering step left a total of 18,129 probes that passed into our subsequent analyses. We used two methods to identify outlier samples (2.5 standard deviations + mean) for quality control: clustering analysis based on Pearson correlation and average distance metric and principal component analysis (PCA) using the first three components. In total, 5 outliers were identified; 3 baseline, 1 pre-boarding, and 1 one-hour post-landing. This reduced our sample size from 50 samples to a total of 45 subjects (**Supplementary Table 1**). The resulting quality-control treated data were used as input for differential expression and WGCNA analyses.

6.2.5. Differential Gene Expression Analyses

We measured differential expression with respect to gene expression at baseline for each time point using 18,129 probes, correcting for gender differences. Differentially expressed genes were assessed using the moderated t-test in LIMMA (Smyth, 2005), and unless otherwise specified, a highly statistically significant threshold of p-value < 0.01 was used. To ensure that genes which were found to be significantly differentially expressed post-landing were not solely a consequence of increased proportion of NK cells, we used a multivariate linear model to regress individual gene expression levels against NK-cell specific marker genes. The criteria for classification as a NK-cell marker were that the genes must be these particular genes needed to be: 1) identified in multiple publications linking them to the NK-cell type; and 2) found intersecting across three independent cell type specific expression databases [CTen (Shoemaker et al., 2010), IRIS (Abbas et al., 2005), and HaemAtlas (Watkins et al., 2009)]. Like others whom have made similar corrections (Miller et al., 2013), we note that the model is fairly robust to choice of marker genes for cell type.

6.2.6. Real-time Quantitative Reverse Transcription (RT-q PCR)

Twenty-two targets were chosen for RT-qPCR confirmation of gene expression. To rule out *false positives*, 15 components of NK cell-mediated cytotoxicity pathway were selected: (**1**) killer cell immunoglobulin-like receptor, three domains, long cytoplasmic tail, 1 (*KIR3DL1*); (**2,3**) killer cell immunoglobulin-like receptor, two domains, long cytoplasmic tail, 1 and 4 (*KIR2DL1* and *4*); (**4**) killer cell lectin-like receptor subfamily D, member 1 (*KLRD1*); (**5**) killer cell lectin-like receptor subfamily C, member 2 (*KLRC2*); (**6**) natural cytotoxicity triggering receptor 3 (*NCR3*); (**7**) Fas ligand (*FASLG*); (**8**) perforin 1 (*PRF1*); (**9**) granzyme B (*GZMB*); (**10**) lymphocyte-specific protein tyrosine kinase (*LCK*); (**11**) zeta-chain (*TCR*) associated protein kinase 70kDa (*ZAP70*); (**12**) linker for activation of T cells (*LAT*); (**13**) SH2 domain containing 1B (*SH2D1B*); (**14**) interferon gamma (*IFNG*); (**15**) CD247 molecule (*CD247*). An additional 3 transcription factors were also selected to rule out *false positives*: (**1**) runt-related transcription factor 3 (*RUNX3*); (**2**) FBJ murine osteosarcoma viral oncogene homolog (*FOS*); (**3**) interferon regulatory factor 1 (*IRF1*). To rule out *false negatives*, 3 targets were selected: (**1**) killer cell lectin-like receptor subfamily K, member 1 (*KLRK1*); (**2**) cathepsin C (*CTSC*); (**3**) transcription factor T-box 21(*TBX21*, also known as T-bet).

One gene not detected by microarray was selected to test possibility of the presence of faulty probes – natural cytotoxicity triggering receptor 1 (*NCR1*). When available, TaqMan® Gene Expression Assays (Applied Biosystems by Life Technologies, Carlsbad, CA) were selected that matched the region of the RNA targeted by the corresponding Illumina probe as closely as possible; otherwise, custom assays were designed and ordered from Integrated DNA Technologies, Inc. (Corallville, IA). Reverse transcription reactions were performed using qScript™ cDNA SuperMix (Quanta Biosciences, Inc., Gaithersburg, MD). *GAPDH* control assay was used as a normalizer. Fold changes were obtained using DataAssist software version 3.01 (Applied Biosystems by Life Technologies, Carlsbad, CA) using the $2^{-\Delta\Delta CT}$ method. To determine significance, a paired *t*-test or Wilcoxon test (depending on the normality of the distribution as assessed by Shapiro test) was performed using normalized *Ct* values (target Ct - *GAPDH* Ct) between the time point of interest and baseline samples.

6.2.7. Weighted Gene Co-expression Network Analysis

The process of identifying discrete groups of co-regulated genes can be divided into two steps. First, a signed global co-expression network was built with weighted gene co-expression network analysis (WGCNA) in R using normalized expression data of 18,129 probes. In this study, we found that our microarray data needed a ß of 9 to reach a scale-free fit. Second, the adjacency matrix was used to calculate the topological overlap measure (TOM), representing the overlap in shared neighbors and hybrid tree-cut algorithm was used to identify sub-networks (i.e. co-expression modules) from the global network (**Figure 6.6**) (Langfelder and Horvath, 2008). With minimal module size set to 15 probes and merging threshold set to 0.1, 20 modules were detected.

To integrate physiological measurements with these co-expression modules, we ran singular value decomposition of each module's expression matrix and used the resulting module eigengene (ME), equivalent to the first principal component, to represent the overall expression profiles for each module. Subsequently, MEs for all modules were correlated to recorded clinical and physiological parameters such as nerve growth factor, epinephrine, norepinephrine, beta endorphin, heart rate, state anxiety trait and cortisol levels which provide a complementary assessment of these potential confounders to that performed in standard differential expression analysis. Further, a Bayes ANOVA (parameters: conf=12, bayes=1, winSize=5) (Kayala and Baldi, 2012) was used to compare ME expression values for modules of interest across time-points while taking into account gender differences. Similarly as previously described, each gene in a module, intramodular membership (kME) was defined as the correlation between gene expression values and ME expression. Genes with high kME inside co-expression modules are labeled as hub genes and are predicted to be essential to the function of the module.

6.2.8. Functional and Cellular Enrichment Analyses

All differentially expressed genes passing a p-value < 0.01 and all network modules with genes passing a kME > 0.50 were subjected to functional annotation. First, the ToppFunn module of ToppGene Suite software (Division of Biomedical informatics) (Chen et al., 2009) was used to assess enrichment of GO ontology terms associated with relevant biological processes and pathways based on a one-tailed hyper geometric distribution with a Bonferroni correction. Second, to predict the involvement of key cell types we utilized the cell specific (HECS) gene expression database from the cell type enrichment (CTen) analysis web-based tool compiled by Shoemaker et al., 2011 for a broad characterization of cell type specific expression. For each gene list supplied, the significance of cell type specific expression is determined using the one-tailed hyper-geometric distribution with a Bonferroni correction across all cell/tissue types.

6.2.9. Protein-Interaction Networks

Protein-protein and protein-DNA interactions for products of differentially expressed genes at pre-boarding, post-landing and one hour post-landing were determined using the direct interactions algorithm in MetaCore^{TM} (GeneGo, St. Joseph, MI). The interactions documented in MetaCore^{TM} have been manually curated and supported by the literature. When protein networks are constructed, they often reveal hub genes which represent transcription factors that control the regulation of multiple target genes. Visualization of a direct protein interaction network was facilitated by use of Cytoscape (Shannon et al., 2003).

6.2.10. Flow-Cytometry

Two blood samples were collected from an additional cohort of 26 first-time tandem skydivers for flow cytometry analysis (one for complete blood counts and a second tube for flow cytometry data analysis). Aliquots from each blood sample were placed into 8 tubes (panels) and incubated with the mAb combinations using the manufacture’s recommended procedures. After incubation, sample processing for the flow cytometry analysis followed the manufacture’s instruction using red blood cell (RBC’s) lysing solution (Becton Dickinson, San Jose, CA). After lysing the RBC's, the white blood cells were washed in phosphate buffered saline (PBS) and re-suspended in PBS buffer and analyzed using a FACS Caliber 4-colour flow cytometer (Becton Dickinson, San Jose, CA). Expression of cell-surface proteins labeled with R-Phycoerythin (PE) was quantified using the geometric means of the mean florescence intensity (MFI) (Shapiro, 2003). All mAb's were purchased from BD Biosciences Pharmingen (San Diego, CA).

6.3. Results

In this exploratory study, we induced 'real-world' acute psychological stress in response to a first-time tandem skydive. Subjects reached altitude in fifteen minutes, jumped at 13,000 feet (4km), fell at terminal velocity for one minute, and parachuted for another four minutes prior to landing. PBL samples and circulating hormone measurements from thirteen participants (7 male and 6 female) were collected at baseline (9:15 am one week before/after the skydive day), pre-boarding (9:15 a.m. skydive day), post-landing (10:45 a.m. skydive day, immediately after landing) and one hour post-landing (11:45 a.m. skydive day) (**Figure 6.1A**).

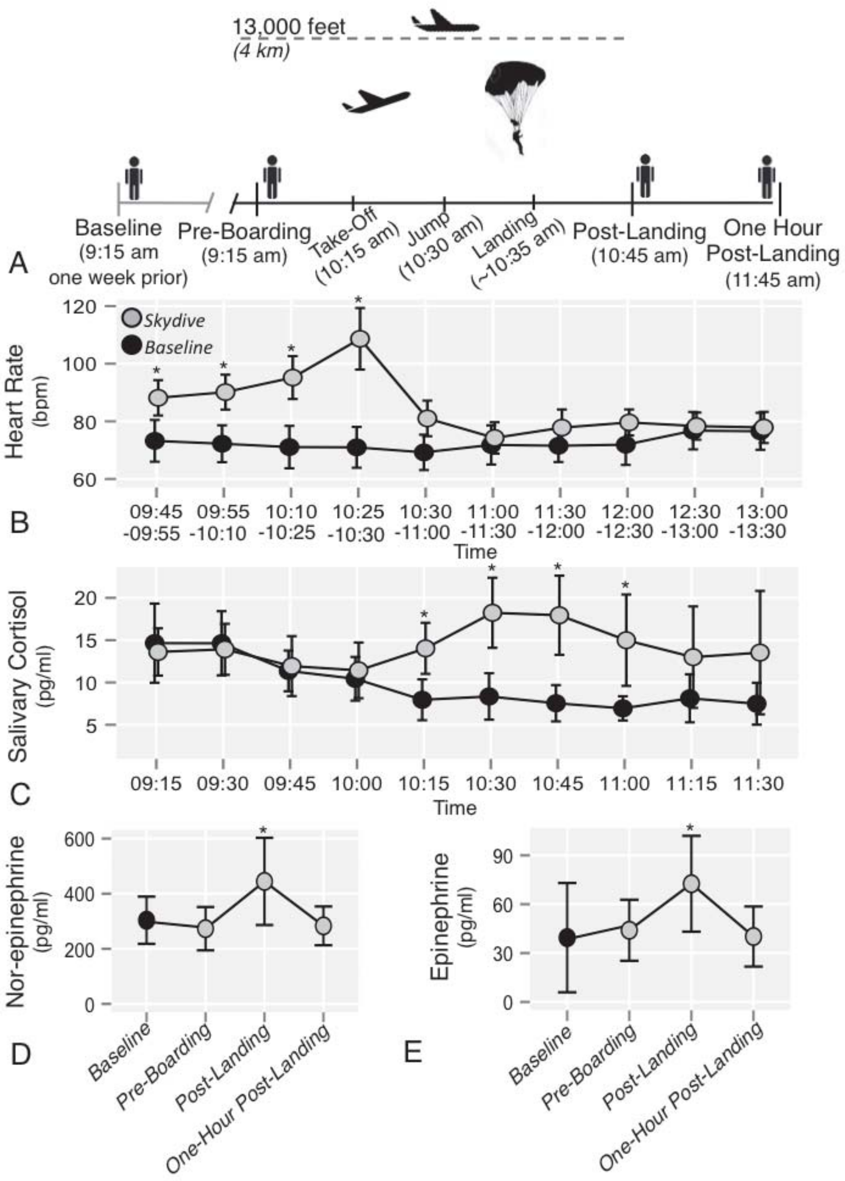


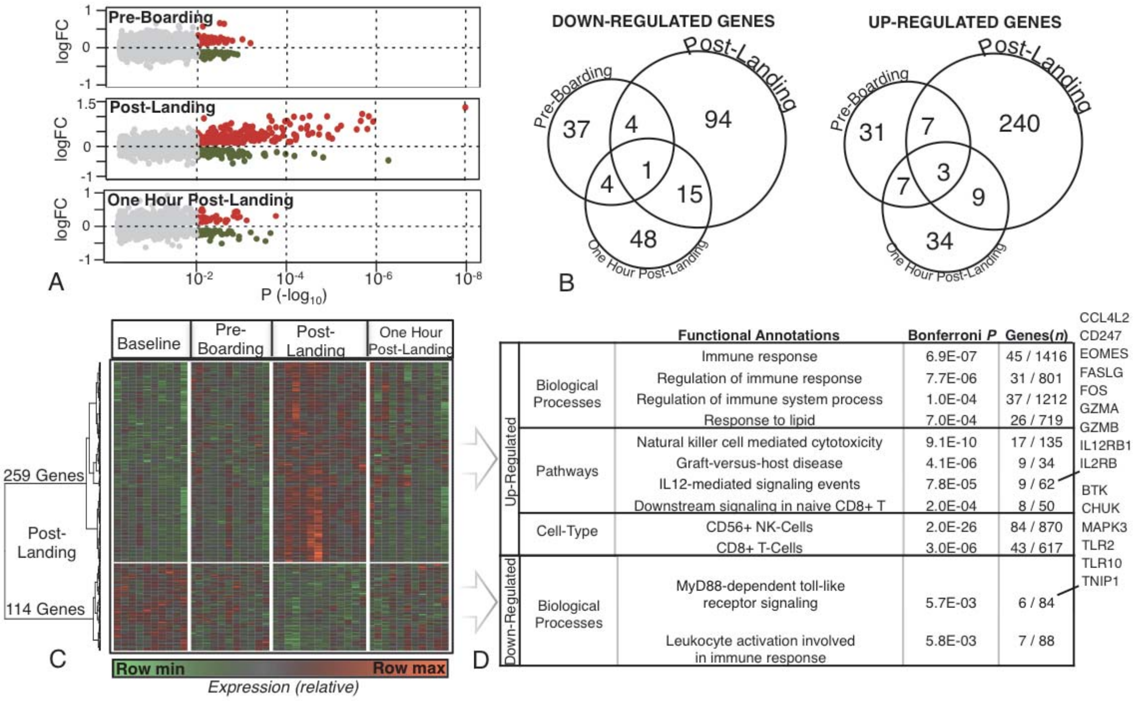
Figure 6.1. Physiological changes observed throughout the sequence of events leading up to, during, and after a first time tandem skydive jump. (**A**) The skydiving paradigm and relevant time-points. (**B**). Heart rate measurements (bpm) were obtained throughout the course of both baseline and skydive days. (**C**) Salivary cortisol (pg/ml) was collected every 15 minutes from 9:15 am until 11:45 am on both baseline and skydive days. (**D**) Norepinephrine (pg/ml) and (**E**) epinephrine (pg/ml) were measured in duplicate and averaged at the corresponding four time points. Dark blue represents baseline day and light-blue represents skydive day. Error bars represent 95% confidence interval and (*) indicates p-value < 0.05 based on non-parametric Mann-Whitney U test.

6.3.1. Acute Stress Induced Physiological Responses

Testosterone, norepinephrine, epinephrine, beta-endorphin, nerve growth factor (NGF), salivary cortisol and heart rate were monitored throughout both the baseline and skydive days as well-established biomarkers for HPA-axis activation consequent to acute psychological stress. Heart rates were elevated on the skydive day relative to baseline as early as pre-boarding the airplane (09:45-09:55) and remained elevated until 30 minutes post-landing (10:30-11:00), peaking immediately before exiting the airplane (10:25-10:30, *p*=6.04E-05) (**Figure 6.1B**). Salivary cortisol measurements were taken every 15 minutes, starting pre-boarding (09:15) to one hour post-landing (11:35) at both the baseline and skydive days. On the skydive day, a significant increase in salivary cortisol was observed immediately before exiting the plane (10:15, *p*=8.0E-03) and peaked between jumping and one hour post-landing (10:30 *p*=5.0E-04; 10:45, *p*=5.0E-03; 11:00, *p*=2.0E-02) (**Figure 6.1C**) compared to the same time-points at baseline. Moderate, yet insignificant, increases in circulating testosterone, beta-endorphin and NFG were observed from baseline to post-landing **(Supplementary Table 6.1)**. Circulating levels of norepinephrine and epinephrine increased post-landing relative to baseline (*p*=4.0E-02, *p*=3.0E-02) (**Figure 6.1D&E**). Heart rate, salivary cortisol and catecholamine levels returned to baseline levels one-hour post-landing. These patterns support stress-induced HPA activation that occurred in response to the stress of skydive. Therefore, gene expression signatures that closely followed changes in these physiological responses were expected.

6.3.2. Candidate Acute Stress Responsive Genes

To identify stress response genes that were non-gender specific, PBL gene expression profiles were corrected for gender differences at pre-boarding, post-landing and one-hour post-landing relative to baseline. Differentially expressed genes (all $p < 0.01$) were identified pre-boarding (*N*=94), post-landing (*N*=373) and one-hour post-landing (*N*=121) relative to baseline (**Figure 6.2 A-B**). The majority of gene expression differences were detected at post-landing and visualized on a heatmap to compare expression levels of these genes at other time-points (**Figure 6.2C**). Genes modulated pre-boarding and one-hour post-landing displayed no functional characteristics or leukocyte cell type specificity. However, of the 373 differentially expressed genes identified from baseline to post-landing, genes relating to NK cell cytotoxicity and IL-12 signalling, including IFN-γ, were up-regulated (**Figure 6.2D**). Genes related to MyD88-dependent toll-like receptor (TLR) signalling tended to show decreased expression. Additionally, cell type enrichment analysis revealed a significant enrichment of up-regulated genes post-landing specific to CD56$^+$ NK cells, and to a lesser extent CD8$^+$ T cells.



		Functional Annotations	Bonferroni *P*	Genes(*n*)
Up-Regulated	Biological Processes	Immune response	6.9E-07	45 / 1416
		Regulation of immune response	7.7E-06	31 / 801
		Regulation of immune system process	1.0E-04	37 / 1212
		Response to lipid	7.0E-04	26 / 719
	Pathways	Natural killer cell mediated cytotoxicity	9.1E-10	17 / 135
		Graft-versus-host disease	4.1E-06	9 / 34
		IL12-mediated signaling events	7.8E-05	9 / 62
		Downstream signaling in naive CD8+ T	2.0E-04	8 / 50
	Cell-Type	CD56+ NK-Cells	2.0E-26	84 / 870
		CD8+ T-Cells	3.0E-06	43 / 617
Down-Regulated	Biological Processes	MyD88-dependent toll-like receptor signaling	5.7E-03	6 / 84
		Leukocyte activation involved in immune response	5.8E-03	7 / 88

Figure 6.2. Comprehensive depiction of gender corrected differentially expressed genes (all p < 0.01) leading up to and following acute psychological stress. (**A**) Volcano plots for differentially expressed genes display extent of log fold-change compared to the –log10 p-value significance at pre-boarding, post-landing and one-hour post-landing respective to baseline. (**B**) Overlap of down-regulated and up-regulated genes across time-points. (**C**) All differentially expressed genes identified from baseline to post-landing. (**D**) Functional annotation of differentially expressed genes identified baseline to post-landing performed separately for up- and down-regulated genes. The top 4 most significant annotations (all p < 0.05 Bonferroni corrected) are shown for categories of biological processes and pathways (annotated with ToppGene) and cell types (annotated with CTen). Genes involved in IL-12 signalling and MyD88-dependent pathway are displayed for quick referencing.

Key genes, including those encoding transcription factors, involved in mediating stress-immune interactions were discovered through interactome analysis of all differentially expressed genes, utilizing validated direct protein-protein interaction (PPI) information from MetaCore^{TM} (**Figure 6.3**). This analysis revealed the up-regulation of transcription factors

RUNX3, *FOS*, *JUN* of the innate immune system, and cyclin-dependent kinase inhibitor 1A (*CDKN1A)* and zeta-chain (TCR) associated protein kinase 70kDa (*ZAP70)* of the acquired immune system. Mitogen-activated protein kinase 3 (*MAPK3*), malic enzyme 2 (*ME2*) and guanine nucleotide binding protein (*GNAI*) mediating innate immune events were down-regulated.

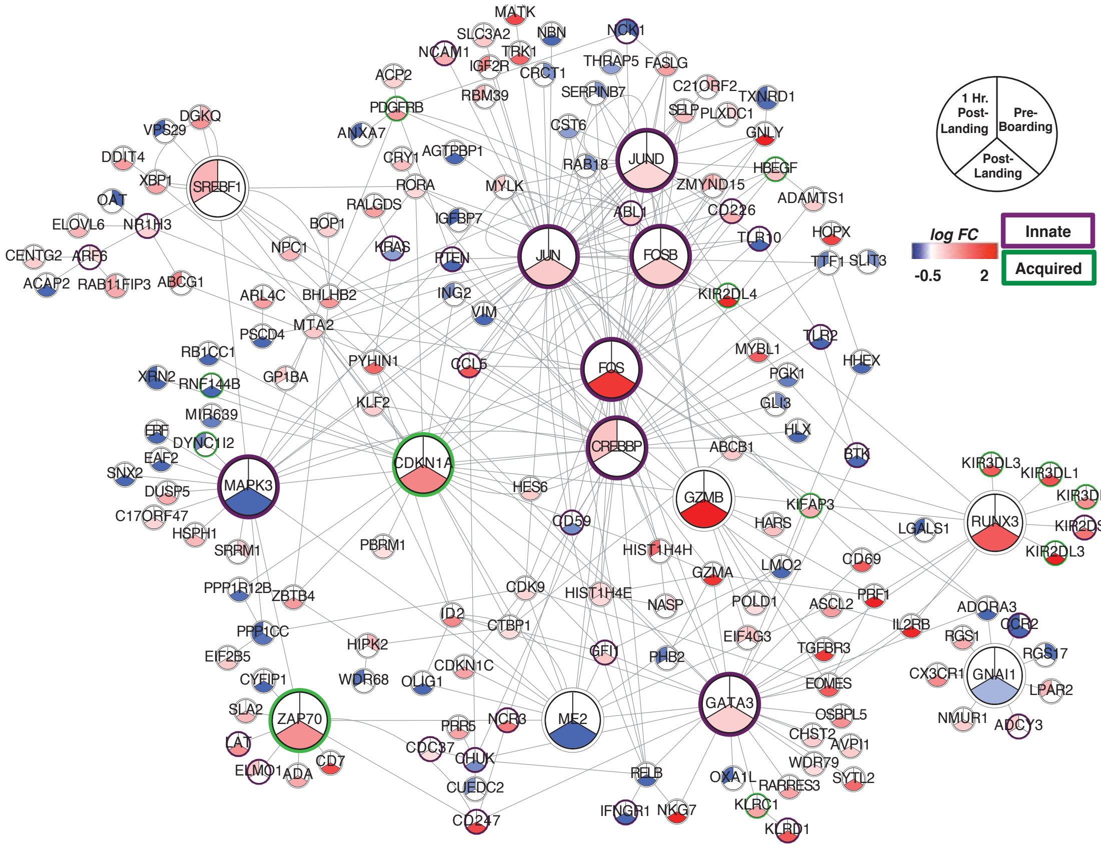


Figure 6.3. Protein interaction network (PIN) of differentially expressed genes. This PIN reflects differentially expressed genes pre-boarding, post-landing, and one-hour post-landing as delineated by the pie chart. Large node sizes reflect key transcription factors with more than 10 validated interactions. Red, up-regulation; blue, down-regulation, on the scale presented by the colour bar; white, no change. Purple circle, genes related to innate immunity and green circle, genes related to acquired immunity.

6.3.3. RT-q PCR Validation of Selective NK Cytotoxicity Response

A set of independent RT-qPCR assays was used to verify differentially expressed genes (from microarray data) post-landing. The RT-qPCR analysis was conducted on 22 of the differentially expressed genes that play a key role in the NK cell cytotoxicity response

(**Figure 6.4**). These genes include those that encode inhibitory receptors (*KIR2DL1*, *KIR3DL1)* and activating receptors (*KIR2DL4*, *KLRC2*, *KLRD1*, *NCR3*), classical MHC class 1 molecules (*HLA-C, B, E, G*) which bind to the receptors, adapter molecules for activating receptors (*SH2D1B, CD247*), signal transduction molecules (*LAT, LCK, ZAP70*) important for NK and T cell activation, cytolytic granules (*PRF1, GZMB*), and transcription factors (*RUNX3, FOS*). Based on previous reports of NK cell mobilization into blood in response to acute stress, it was probable that a significant number of genes would map to NK cell mediated cytotoxicity pathway (Altemus et al., 2001; Schedlowski et al., 1993). However, not all well characterized NK cell related molecules, pro- and anti-inflammatory cytokines, receptors and transcription factors were differentially expressed (**Table 6.1**). For example, activating receptors *NCR1* and *KLRK1*, cytolytic granule *CTSC* and transcription factor *TBX21* were not dysregulated; gene expression was confirmed by RT-qPCR (**Figure 6.4**). These results suggest a precise and selective regulation of NK cell molecules and inflammatory properties of the innate and acquired immune system during acute stress, which are not accounted solely by an influx of NK cells into the periphery.

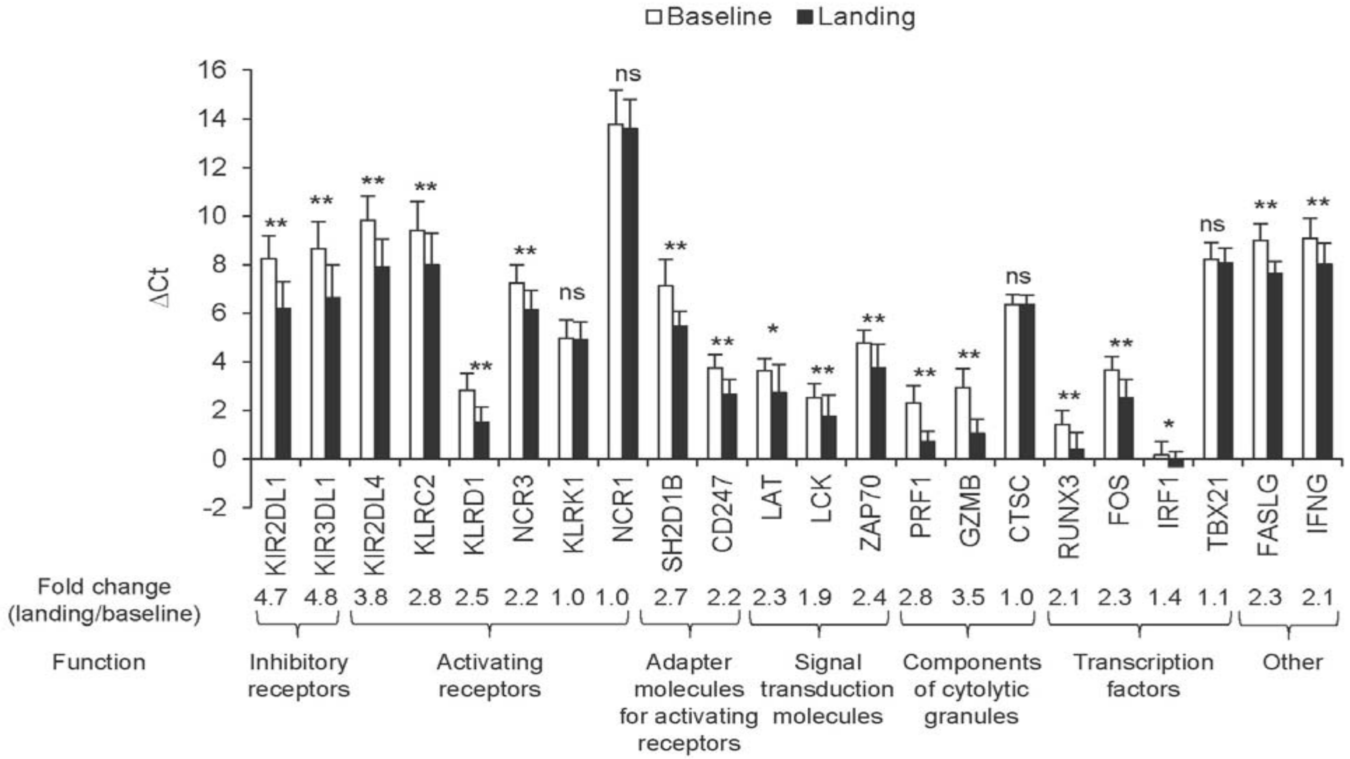


Figure 6.4. RT-qPCR confirmation of differentially expressed genes. Gene expression analysis by RT-qPCR was performed for 18 genes found as differentially expressed between baseline and post-landing, three genes that were not differentially expressed and one gene that was below the limit of detection in the microarray analysis. ΔCt values (Ct – Ct_{GAPDH}) were calculated for each gene at baseline (white bars) and post-landing (black bars). Error bars represent standard deviation. The smaller ΔCt values indicate presence of greater number of mRNA molecules in a sample. All the differentially expressed genes identified by microarray were also differentially expressed by RT-qPCR (** $p < 0.01$, *$0.01 < p < 0.05$) between baseline and post-landing time points. All genes that were not differentially expressed by microarray were also not differentially expressed by RT-qPCR (ns, not significant).

Table 6.1. Selective regulation of NK cell cytotoxic genes at post-landing

Function	Modulated	Non-modulated
Inhibitory NK receptors	KIR2DL3 ↑	LILRB1
	KIR3DL1 ↑	TIGIT
	KIR3DL3 ↑	LAIR-1
	KLRG1 ↑	CEACAM-1
	KIR2DL1 ↑	KLRC1
		KIR3DL2
		KIR2DL5A
		KIR2DL5B
Activating NK receptors	KIR2DL4 ↑	Ly9
	KLRF1 ↑	KIR2DS1
	KLRC2 ↑	KIR2DS2
	KLRD1 ↑	KIR2DS5
	NCR3 ↑	KLRK1
		FCGR3A (CD16)
Adaptor molecules for activating NK receptors	CD247 ↑	HCST (DAP10)
	SH2D1A ↑	TYROBP (DAP12)
	SH2D1B ↑	
Components of cytolytic granules	PRF1 ↑	CALR
	GZMB ↑	CTSC
	GZMA ↑	SRGN
	GZMH ↑	
	GLNY ↑	
	CTSW ↑	
Chemotactic receptors	S1PR5 ↑	CCR2
		CCR7
		CXCR1
		CXCR4
		CXCR6
		CX3CR1

Table 6.1 Continued...

Table 6.1 Continued...

Function	Modulated	Non-modulated
Pro-inflammatory Cytokines and Receptors	*IFNGR1* ↓	*CCL3*
	IFNAR1 ↓	*IL-1*
	TNF ↓	*IL1B*
	TNFAIP8L2 ↓	*IL-2*
		IL-6
	IFN-g ↑	*IL-8*
	IL12RB1 ↑	*IL-12*
	IL2RB ↑	*IL12RB2*
	IL21R ↑	*IL-14*
	IL18BP ↑	*IL-21*
		TNFa
		ILR2
		IL2RA
		Il2RG
Anti-Inflammatory Cytokines and Receptors	*CCL4L2* ↑	*IL-4*
	IL10RA ↑	*IL-10*
	IL10RB ↑	*TGF-beta*
	IL5RA ↓	
Transcription factors	*RUNX3* ↑	*TBX21* (T-bet)
	EOMES ↑	*STAT4*
	JUND ↑	*STAT1*
	FOS ↑	
	GATA3 ↑	
	ME2 ↓	
	MAPK3 ↓	

Modulation is according to post-landing significance with a $p < 0.01$. Arrows indicate direction of change. Grey shading indicates genes modulated at post-landing.

6.3.4. Candidate Acute Stress Responsive Genes Underlying Cell Subset Fluctuations

To account for NK cell type differences underlying differential gene expression changes from baseline to post-landing, a linear regression model was created taking into account expression of major NK cell markers. Four NK cell markers were selected that were consistently found across three different cell type-specific expression databases (Abbas et al., 2005; Shoemaker et al., 2010; Watkins et al., 2009): *CLIC3*, *KLRF1*, *KIR2DL3* and *KIR3DL1.* Accounting for NK cell type composition at post-landing indicate that ~15% of the previously identified differentially expressed genes remained significant. Genes encoding for *FOS* and *GZMB* were among the most up-regulated genes surviving this correction*,* whereas *CLC* and *PAPSS1* were among the most down-regulated. Functional enrichment analysis revealed that genes corresponding to NK cell mediated cytotoxicity and graft-vs.-host pathways were no longer significant. However, a significant up-regulation of genes enriched for *IL-12* mediated signalling (*FOS, RELB, CD247, GZMB, IL2RB*), cytotoxic T-lymphocyte (CTL) mediated immune response (*CD247, PRF1, GZMB*) and downstream signalling in naive CD8$^+$ T cells remained significant albeit to a lesser extent. A most interesting finding resulting from this correction was a significant enrichment of genes specific to the adrenal cortex, a key mediator of the stress response **(Figure 6.5)**.

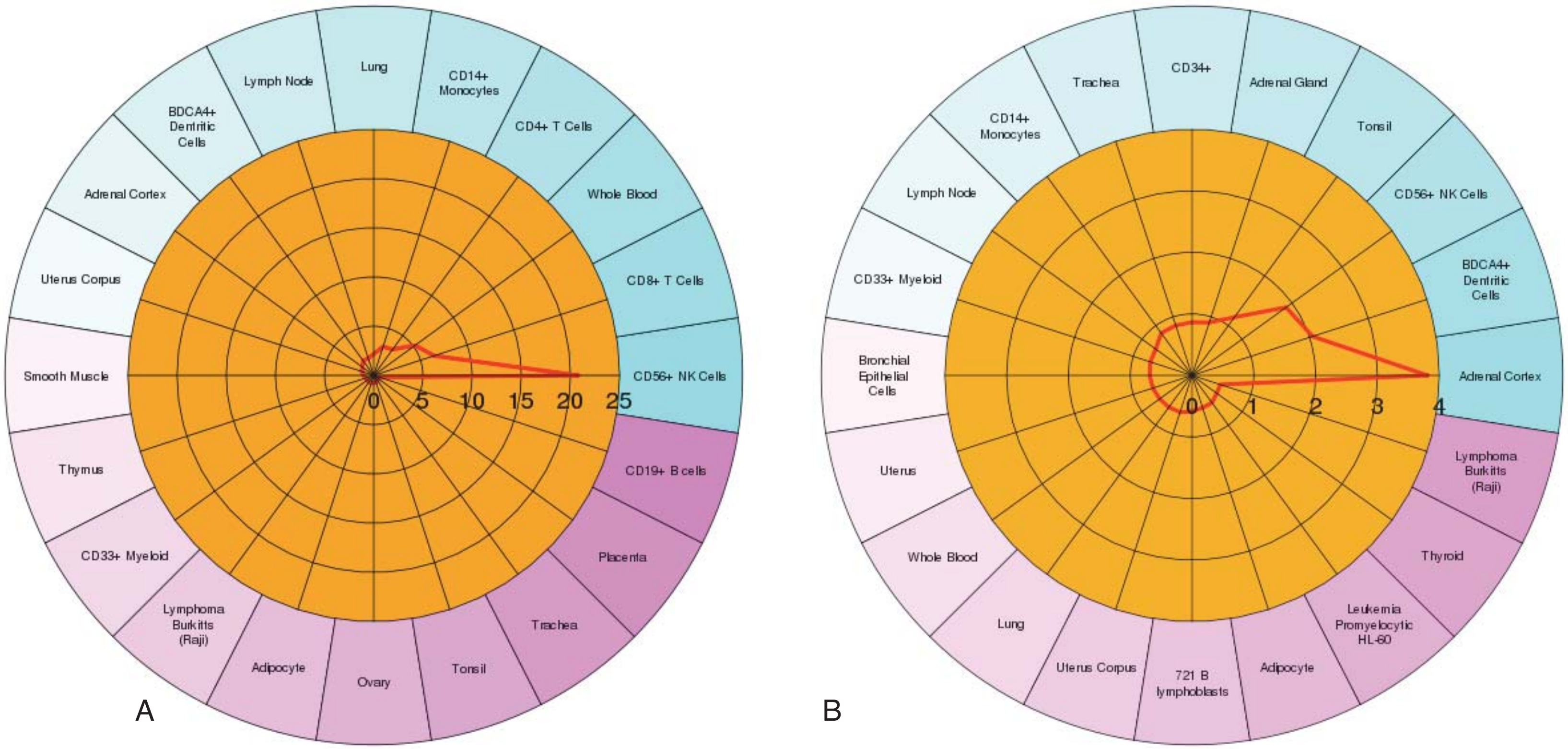


Figure 6.5. CTen cell type enrichment analyses based on differentially expressed genes identified from baseline to post-landing. (**A**) Using all identified genes from Figure 2C, without correcting for NK cell types. (**B**) Enrichment using genes that survived cell type correction with multivariate linear model.

6.3.5. Identification and Annotation of Gene Co-expression Modules

To identify coordinately expressed genes (modules) involved in the short-term variable immune response to acute stress, unsupervised WGCNA was performed. The analysis identified 19 distinct co-expression modules and 1 module representing all background genes that could not be clustered into any module **(Figure 6.6),** each with a distinct expression pattern across all four time-points. Subsequently, all modules were functionally annotated using the top significant biological process, pathway and cell type for each individual module (all Bonferroni $p < 0.05$) **(Supplementary Table 6.2).**

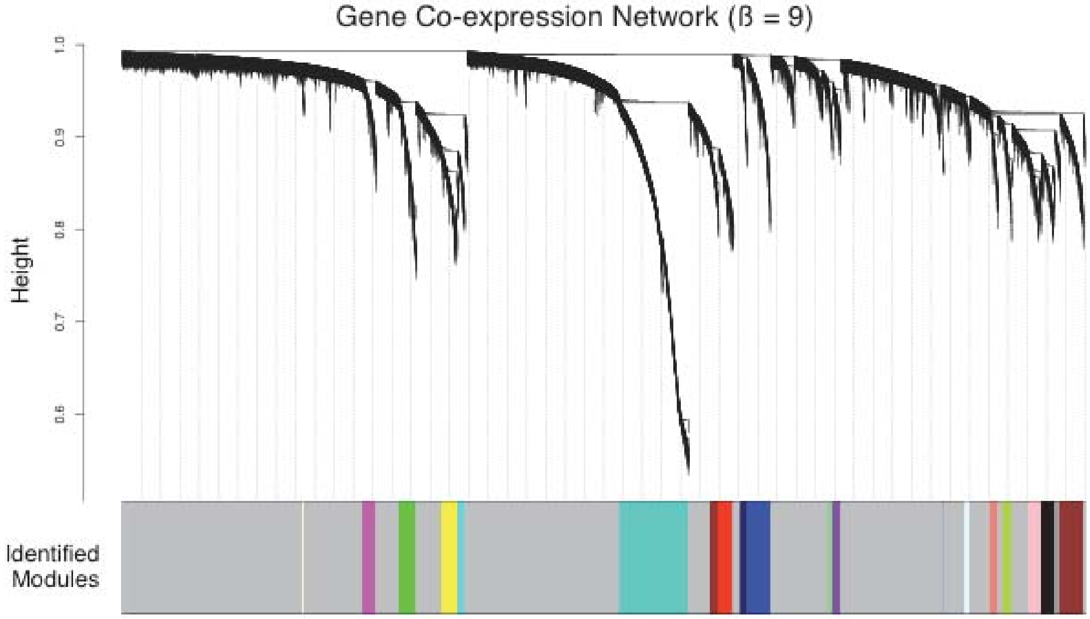


Figure 6.6. Identification and organization of gene co-expression modules. WGCNA cluster dendrogram and network modules with corresponding information bars. The network was raised to the beta power of 9 to satisfy scale-free topology. The bar represents the identified modules (as denoted by colours), the grey module corresponds to genes which do not cluster into any other module. Each line represents a gene (leaf), and multiple genes clustered together represent a group of co-regulated genes (low hanging branches) on the cluster dendrogram (tree). The *y*-axis corresponds to distance determined by the extent of topological overlap measure (1-TOM).

6.3.6.Correlating Gene Modules with Physiological Measurements

Next, we sought to determine the relationships between the 20 modules identified above and the observed physiological and hormonal fluctuations throughout the stress response. To integrate these multi-scale data types, module eigengene (ME) values were correlated to each time-point and all recorded subjective and physiological traits (**Figure 6.7**). Briefly, ME value is the first PC of module expression and summarizes the main trend of expression within a module. Among the modules with high association with time-points and physiological traits, the *ME* of a module specific for 'Cytokine Production' was negatively correlated to post-landing (r = -0.29, p = 0.05) as well as fluctuations in circulating norepinephrine (r = -0.32, p = 0.03). The ME of modules associated to 'T Cell Receptor (TCR) Signalling Pathway' and 'NK Cell Mediated Cytotoxicity' were positively correlated to post-landing (r = 0.28, p = 0.06; r = 0.57, p = 4E-05 respectively). Moreover, the 'NK Cell Mediated Cytotoxicity' module was positively correlated to norepinephrine (r = 0.39, p = 0.007) which was expected given elevated norepinephrine and NK cell specific gene expression peak post-landing and return to baseline levels one hour later (**Figure 6.1D** & **Figure 6.2C**). The expression pattern of each marker gene used in our linear model to correct differential gene expression analysis (*CLIC3*, *KLRF1*, *KIR2DL3* and *KIR3DL1*) showed a strong correlation to the ME of this particular module, confirming that the genes for our linear model were appropriately chosen. The ME of a 'Hemostasis' module showed a gradual change from negative to positive correlation from baseline to one-hour post-landing and was significantly correlated to beta-endorphin fluctuations (r = 0.32 p = 0.03). Additionally, the ME for a module involved in 'Oxygen Uptake and Carbon Dioxide Release' was positively correlated to both heart rate (r=0.38, p=0.01) and salivary cortisol levels (r=0.43, p=0.003), highlighting the interaction between the cardiovascular and respiratory systems. Most interestingly, including gender as a discrete measure revealed that many modules were either positively or negatively correlated to gender differences (**Figure 6.7**) suggesting gender-specific expression patterns within each of these modules.

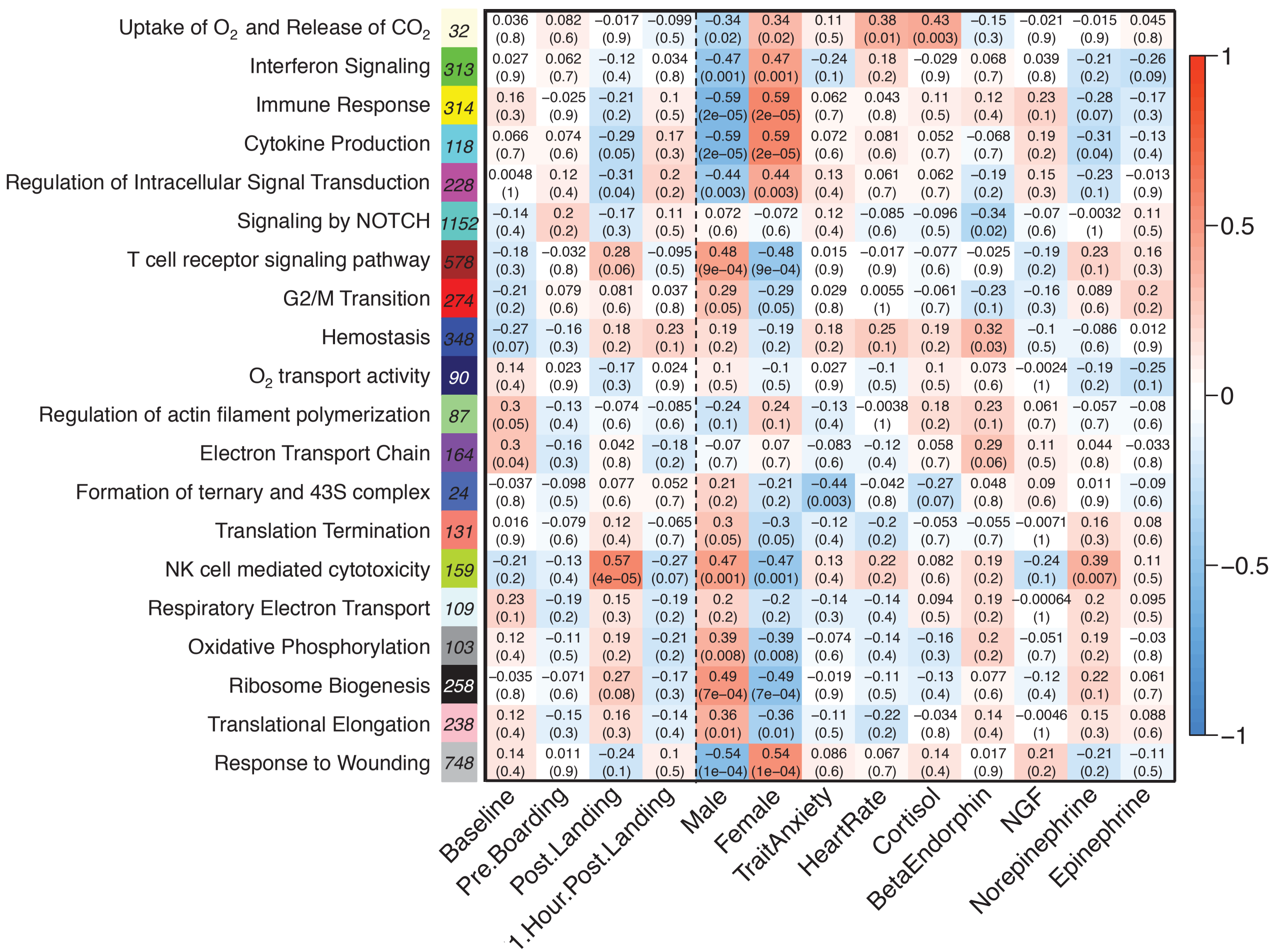


Figure 6.7. Functional PBL gene module – stress hormone relationships throughout the stress response. ME values of functional PBL gene modules were correlated to underlying stress hormone measurements. The measure of correlation, *r*, is the top value in each box and the related *p*-value is designated below within brackets. Red signifies a positive correlation and blue signifies a negative correlation as indicated by scale. The number next to each functional PBL gene module signifies the number of genes within that module with kME > 0.05 used for functional annotation. Abbreviations: NGF, nerve growth factor; NK, natural killer.

6.3.7. Determining Gender-specific Immune Responses

The extent of co-expression differences was visualized throughout the stress response considering gender, averaging ME values for seven males and six females at each time-point. A Bayes ANOVA was used to compare *ME* expression values for modules of interest across time-points while taking into account gender differences (**Figure 6.8**). The ‘NK cell mediated cytotoxicity’ and ‘Ribosome Biogenesis’ modules showed intensified expression post-landing in males relative to females (**Figure 6.8A-B**), whereas the expression of the ‘TCR Signalling Pathway’ module was highest one-hour post-landing in males relative to females (**Figure 6.8C**). Co-regulated genes specific to ‘Hemostasis’, which includes genes for blood coagulation, showed a gradual increase in expression (**Figure 6.8D**) for both males and females peaking one-hour post-landing relative to baseline. Strikingly, four modules specific to ‘Immune/Defense Response’, ‘Response to Wounding’, ‘Cytokine Production’ and ‘Interferon Signalling’ (**Figure 6.8E-H**) were down-regulated in males post-landing and one-hour post-landing relative to females, while ME expression either increased or remained unchanged.

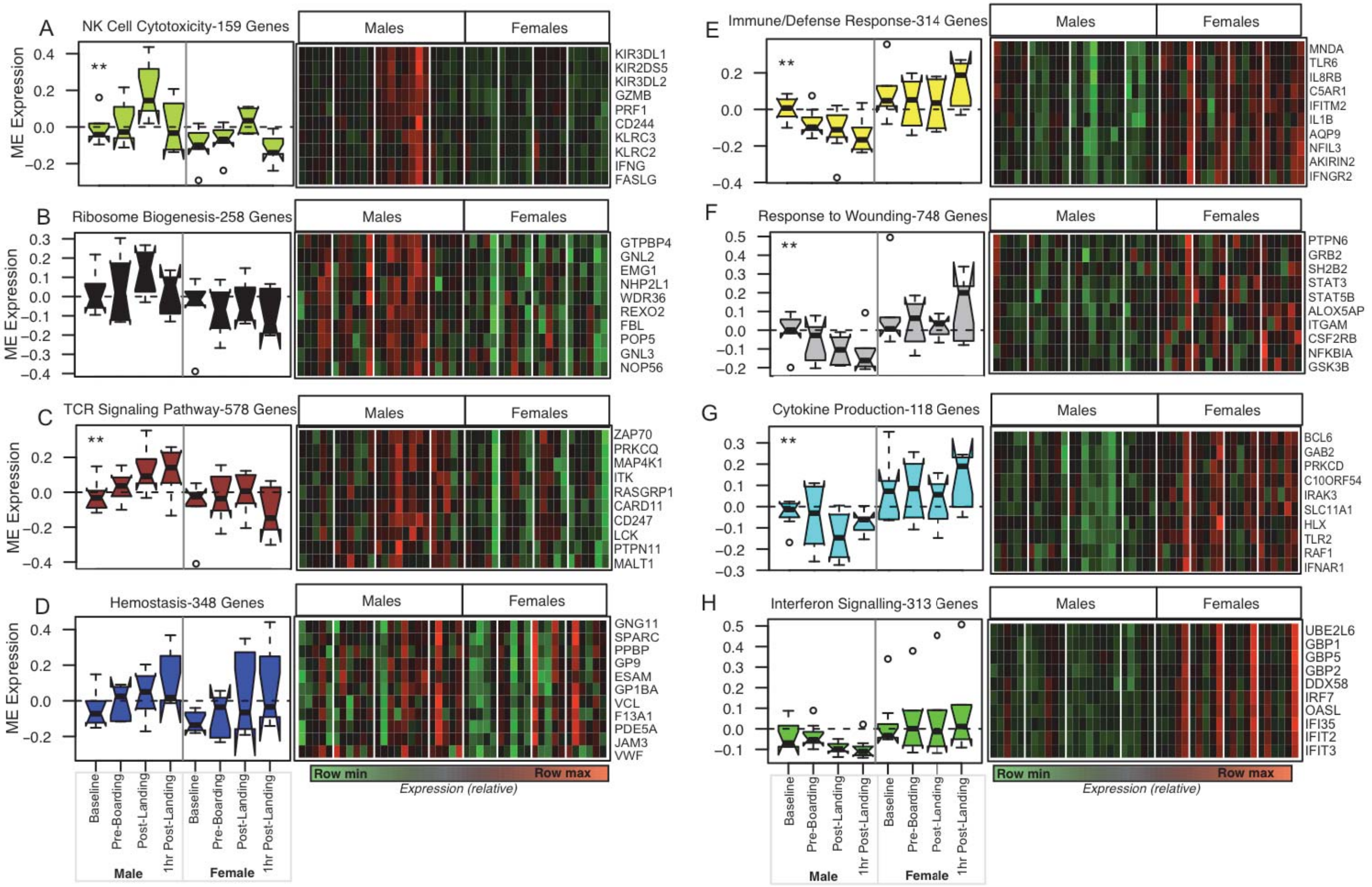


Figure 6.8. Gender specific differences in functional gene co-expression modules. ME values for modules of interest are evaluated across the four time-points comparing males and females. Modules specific to (**A**) NK cell cytoxicity, (**B**) ribosome biogenesis, (**C**) TCR signalling pathway, (**D**) hemostasis, (**E**) immune/defense response, (**F**) response to wounding, (**G**) cytokine production (**H**), interferon/cytokine signalling are displayed. Heatmaps display the extent to which expression profiles of the top 10 functional hub genes, for each corresponding module, change in males and females across different time-points. White line spacers in heatmaps indicate the four time-points. The functional annotation and number of genes within each module are displayed above the boxplots. A Bayes ANOVA was used on ME values to test for significance between males and females, (**) indicates $p < 0.001$ implying strong gender-specific differences throughout course of the stress response.

6.3.8. Acute Stress Responsive Leukocytes

Acute stress has been shown to cause a redistribution of leukocytes throughout the periphery (Dhabhar, 2009). To fully characterize changes in peripheral leukocyte and lymphocyte subsets throughout acute psychological stress in this study, a second cohort consisting of 26 participants (17 male and 9 female) was recruited under the same matching experimental design as the gene expression cohort. Subsequent blood samples were subjected to flow cytometry analysis. These quantitative cell-type data were also used to better understand the extent of which gene expression results may be affected by migrating cell types. Changes within leukocyte and lymphocyte subsets were measured and displayed as both percentages and absolute cell counts combined across both males and females **(Figure 6.9)**, as there were no strong differences in cell type fluctuations between genders **(Supplementary Table 6.3)**.

Total leukocytes significantly increased from baseline to pre-boarding and post-landing, returning to baseline levels one-hour post-landing. There was a marked increase in the proportion and absolute count of neutrophils pre-boarding, while the post-landing proportion (albeit significantly greater than baseline) was significantly smaller than pre-boarding. Eosinophil proportion and absolute count reduced pre-boarding and remained low post-landing and one-hour post-landing relative to baseline. Monocytes and total lymphocytes showed similar patterns with the lowest proportion and absolute cell counts pre-boarding.

Changes in lymphocyte subsets were also investigated (**Figure 6.9**). The percentage of CD19$^+$ B lymphocytes and absolute B cell numbers were significantly reduced post-landing. Conversely, NK cells (defined as CD3$^-$CD16$^+$CD56$^+$) were significantly increased pre-boarding and post-landing. The percentage of CD3$^+$ T lymphocytes were significantly reduced post-landing while absolute number of T lymphocytes was significantly decreased pre-boarding compared to baseline. Of the CD3$^+$ lymphocytes, CD8$^+$ and CD4$^+$ T cell absolute counts significantly increased post-landing relative pre-boarding, while CD4$^+$ T cell proportions decreased post-landing.

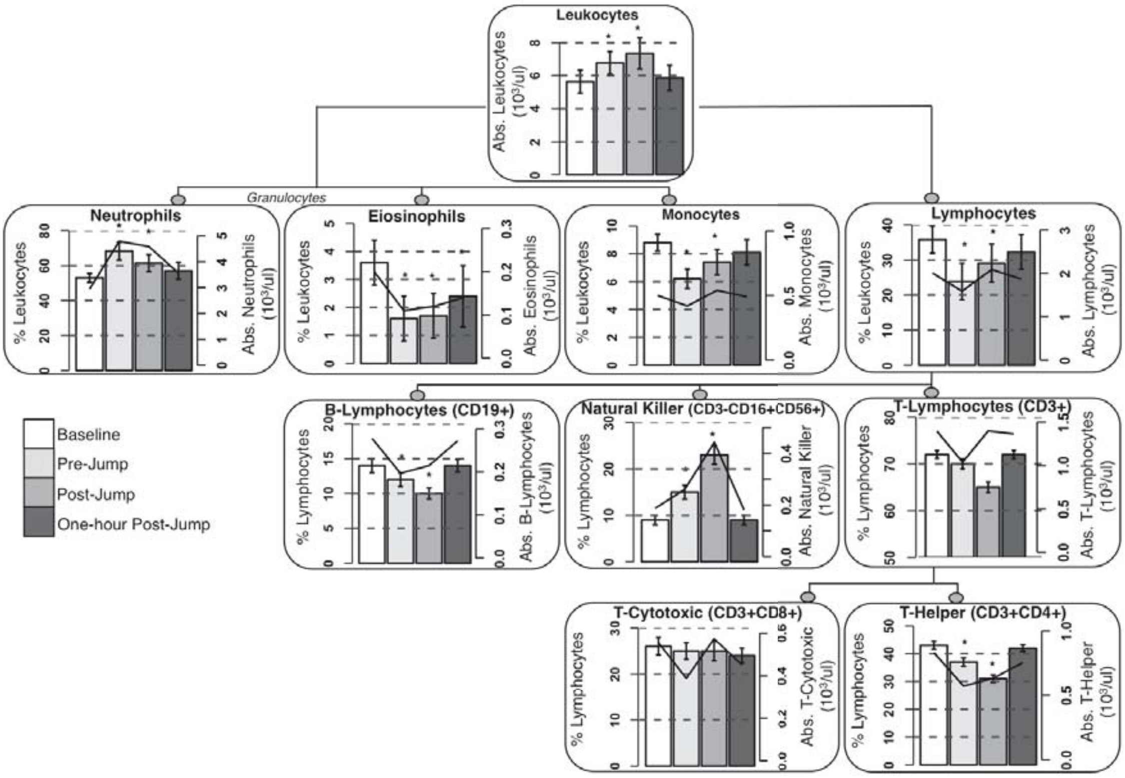


Figure 6.9. A quantitative measurement of the PBL cell lineage via flow cytometry. The analysis used a gating strategy based on the forward scatter/side characteristics of immune cells from total leukocytes; granulocytes (CD45$^+$), monocytes (CD14$^+$), T cells (CD3$^+$, CD4$^+$, CD8$^+$), B lymphocytes (CD19$^+$) and NK cells (CD3$^-$CD56$^+$CD16$^+$). The raw flow data is presented as a percentage of gated cells (as indicated by the bar plots). To determine the absolute immune cell counts (as indicated by the line), leukocyte differential counts from the complete blood counts results were used to produce estimates of the actual number of immune cells in the peripheral blood samples. Statistical analysis was based on a Dunnet's Test multiple comparison of means was used, comparing measurements back to baseline.

6.4. Discussion

This study describes the molecular and cellular response of the human innate and acquired immune system in reaction to physical danger. A first-time tandem skydive was used as a short-term longitudinal design to induce acute psychological stress in a controlled environment; the stressor induces a severe form of emotional response aligned with distress related to fear (Carter and Goldstein, 2011). This exploratory study took a dual approach. First, comparative analysis of PBL gene expression profiles between time-points identified that most gene expression changes occurred during/immediately after the stress response. Immediate immunomodulation is observed as a selective up-regulation of NK cytotoxicity genes, further validated with RT-qPCR assays. Correcting for changes in NK cells post-stressor revealed a molecular signature specific to the adrenal cortex. Second*,* focusing analysis on co-expressed modules revealed gender-specific peripheral immune activation evident by hundreds of co-regulated genes within several biologically annotated modules whose expression differed between males and females. These findings provide a useful characterization of acute stress-induced immune system alterations with implications for the understanding and treatment of stress-related disorders and gender vulnerability to stress-induced pathologies. Major changes in blood-based gene expression were confirmed in a second cohort where blood was subjected to a detailed flow-cytometry analysis.

6.4.1. Selective NK Cell Stress Susceptibility Genes

Although our flow cytometry data showed significant changes in leukocyte subtypes in the course of the stressor, we also showed that changes in observed gene expression profiles could not be explained solely by the fluctuation of different leukocyte subsets. For example, peripheral neutrophils were elevated and peripheral eosinophils were reduced in the periphery pre-boarding in anticipation of the stressor. The changes in cell composition were paralleled by the up-regulation of 48 genes and the down-regulation of 46 genes, which were not associated to any functional annotations or leukocyte cell type specificity.

One unexpected finding of our study is the selective up-regulation of only a subset of NK cell genes post-landing (confirmed by RT-qPCR **Figure 6.4**), despite a pronounced 2.5 fold increase of NK cells in the periphery (**Figure 6.9**). This result may be explained through four phenomena. First, it is possible that a subset of NK genes that displayed no change in expression, were down-regulated in individual NK cells: NK cell activity may be regulated post-transcriptionally, including increases in translation and redistribution of receptors to the cell surface, which is a likely mechanism due to a fast nature of the response. Second, it is also possible that a specialized, characterized (e.g. CD56^{Lo} (Bosch et al., 2005)) or not-yet characterized subset of NK cells, expressing only a subset of specific markers is mobilized into the periphery in response to stress. Third, since gene expression was profiled from the mixture of cells, contribution of other leukocyte subsets that express overlapping sets of genes cannot be ruled out. In particular, gene expression markers for CD8$^+$ T cells were slightly elevated post-landing compared to baseline despite no change in CD8$^+$ T cell frequency in blood (**Figure 6.9**). Even though NK cell-related genes are also expressed at lower levels in these cells, a large change in their expression in T cells can contribute to their expression change in total leukocytes. Finally, although differential gene expression analyses were gender corrected, it could be that that the NK cell response is modulated to differing degrees between males and females as suggested by WGCNA observed gender-specific differences (**Figure 6.8**).

While only ~15% of the originally identified differentially expressed genes were found to be dysregulated after correcting for NK cells, the consistent up-regulation of cell toxicity transcript *GZMB* and transcription factor *FOS* was evident. Proteolytic granzymes, such as *GZMB*, and granulysin delivered from cytotoxic cells via granule exocytosis cause activation of caspase-dependent apoptosis in stressed or pathogenic target cells (Bernard et al., 1999), which helps to explain functional annotations such as CTL mediated immune response and apoptosis following the correction. The up-regulation of *FOS*, an early immediate gene which is expressed in the brain (Bernard et al., 1999), blood (Torres and Lotfi, 2007) and adrenal cortex and mediates physiological adrenocorticotropic hormone-induced responses in adrenal cortical cells (Rui et al., 2014; Verstrepen et al., 2008), is consistent with the enrichment of differentially expressed genes following NK cell correction associated with the adrenal cortex and the production of cortisol (**Figure 6.5**). This important observation may have been difficult to detect if gene expression was measured for each cell type isolated independently.

6.4.2. Putative Roles of IL-12 Signalling and TLRs

The most pronounced effect following multivariate linear regression to adjust for an influx of NK cells into the periphery post-landing, was the consistent up-regulation of genes involved in IL-12 mediated signalling (*CD247*, *FOS*, *GZMB*, *IL2RB*), and the minor production of IFN-γ. The IL-12 signalling pathway determines the type and duration of innate and adaptive immune response in part by promoting NK cell cytoxicity as well as the differentiation of naive CD4+ T cells into T helper 1 (Th1) cells via the production of IFN-γ. Up-regulation of IL-12 signalling may indicate priming of the pro-inflammatory arm of the immune system. Such immunomodulation creates an advantage during events such as vaccination since a primed pro-inflammatory state is important for vaccine-mediated T cell immune responses, which are induced by most anti-bacterial and anti-viral vaccination strategies (Dhabhar, 2009). These data suggest a more focused adaptive immune response which under further emotional distress or antigen presentation may provide a cytokine environment favorable for Th1 polarization of the immune system.

These data also show the down-regulation of MyD88-dependent pathway including signalling molecules *MAPK3*, *CHUK* (i.e. *IKK-α*) and toll-like receptors (TLRs) 2, 6 and 10. In homeostatic conditions, TLRs lead to NFkB activation and production of pro-inflammatory cytokines IL1β, IL6 and TNFα, all involved in different pathways for innate immune activation and defense (Rui et al., 2014; Verstrepen et al., 2008). Down-regulation of TLRs is consistent with previous reports suggesting that increased cortisol levels during acute stress may inhibit the NFkB, JAK-STAT and MAPK signalling pathways (Kadmiel and Cidlowski, 2013; Reichardt et al., 2002; Rui et al., 2014; Webster et al., 2002). Under repeated bouts of acute stress or chronic exposure to psychosocial stress (and continued emotional activation), the response of HPA axis to sustained stress is diminished and subsequently the effectiveness of glucocorticoids (e.g. cortisol) to regulate the inflammatory response is altered as immune cells become insensitive to its regulatory effects (Cohen et al., 2012). Consequently, inflammatory pathways may become activated and initiate a negative feedback loop driving inflammation and promoting the development of many diseases.

6.4.3. Gender-Specificity of the Acute Stress Response and Implications for Stress-induced Pathologies

Another unexpected result stemming from our exploratory gene co-expression approach was the gender-specific immune response to acute stress (**Figure 6.8**) despite similar cellular and hormonal alterations (**Supplementary Table 6.1, 6.3**), which may have relevant translational avenues. While gender-specific differences in the psychobiological stress response have not been clearly identified, they may provide an insight towards understanding the differential cardiovascular risk in men and women. Processes associated with cardiovascular disease, such as TCR signalling, defense response, response to wounding, cytokine production and interferon signalling (Mehra et al., 2005) were differently regulated by acute stress in males and females in our study (**Figure 6.8**). These findings may help to explain gender-specific predisposition to CVD and emphasize these genes and pathways as potential tools which may be able to measure an entire facet of CVD risk, the impact of maladaptive molecular response to psychological stress in both sexes and among women in particular. Moreover, since many inflammatory disorders that are more common in women, such as many autoimmune conditions, are also exacerbated by psychological stress (Whitacre, 1999), gender differences in cytokine response to stress (**Figure 6.8G**) could mark an important underlying mechanism.

Women, more frequently than men, suffer from chronic forms of stress such as post-traumatic stress disorder (PTSD) (Becker et al., 2007), but the reasons for this disparity are not entirely clear. It has been proposed that these differences are not explained solely on the basis of exposure type and/or severity (Sherin and Nemeroff, 2011) and that modulation of sex steroids such as estrogen and progesterone may exert effects on neurotransmitter systems involved in the stress response. Factors other than exposure may play a role in the development of an intermediate state in which gender may determine vulnerability to PTSD, and these may include transcriptomic level differences.

6.4.4. Putative Biomarkers for Discriminating Anxiety-based Stress from Neuropsychiatric and Central Nervous System (CNS) Disorders

An important task for studies investigating peripheral mechanisms of CNS disorders (multiple sclerosis, stroke and seizure) as well as panic attacks, myocardial ischemia, and related rodent models of such disease (Achiron et al., 2004; Kim et al., 2014; Samad et al., 2014; Yang et al., 2005; Yang et al, 2001), is the ability to disentangle molecular mechanisms which are most closely associated with the clinical presentation of disease.

For example, there is a need for biomarkers for discriminating between ‘psychogenic’, non-epileptic seizures and true epilepsy is needed (Testa et al., 2012). In our study the most down-regulated gene post-stressor and one-hour post-stressor was *IMAP2*, and the most down-regulated transcription factor post-stressor was *ME2*, as indicated by interactome analysis (**Figure 6.3**). Interestingly, in genome-wide association studies, both genes *IMPA2 and ME2* have been reported as susceptibility genes in febrile seizures and idiopathic generalized epilepsy (Arai et al., 2007; Greenberg et al., 2005; Mas et al., 2004; Prasad et al., 2013). However, gene expression studies report the lack of *IMPA2* and *ME2* dysregulation in the blood of humans (Piro et al., 2011; Yang et al., 2005; Yang et al, 2001) as well as in the brains of rodents post-seizure (Harald et al., 2001). While these results should be interpreted cautiously, the general inconsistencies between these studies of epilepsy and the results presented here describe *IMPA2* and *ME2* as genes warranting further investigation as putative biomarkers discriminating psychogenic from non-psychogenic seizures.

6.4.5. Hypoxia does not Contribute to Gene Expression Differences

Studies using an exaggerated 12 hours sustained poikilocapnic hypoxic model system have noted the dysregulation of mRNA expression specific to hypoxia-inducible factor 1 (*HIF1A), GAPDH*, *EPO*, and *VEG* within the first two-hours (Pialoux et al., 2009). Thus, there was a slight possibility that factors attributable to a short-term exposure (i.e. 20 minutes) to high altitude (i.e. 13,000 ft.), such as hypobaric hypoxia, could influence gene expression in subjects during the skydive. Therefore, the expression of these mRNA species was investigated*. HIF1A* was measured on the microarrays by three probes: none of these probes were detected as significant in our differential gene expression analysis (all p > 0.1). None of the probes for other genes associated with hypoxic conditions such as *GAPDH*, *EPO* or *VEG* (Pialoux et al., 2009; Zhong et al., 1999) were dysregulated*.* We observed the differential expression of *HIPK2* among the identified anticipatory genes at pre-boarding, known to suppress *HIF1A* in hypoxia-mimicking conditions (Nardinocchi et al., 2009). The early activation of *HIPK2* may reflect increased anticipatory heart rate and early rapid breathing in anticipation to the skydive which may be working to suppress ‘hypoxia-mimicking’ conditions in the PBL microenvironment.

6.4.6. Strengths and Limitations

A strength of this study was the prospective longitudinal design. However several limitations remained. For example, while we adjusted for cell type changes affecting global gene expression, clear limitations are the lack of transcriptomic investigation on individual cell types and the ability to perform transcriptomic analysis and flow cytometric data analysis on the same cohort of individuals. Additionally, while gender specific differences were observed across a small number of samples, the evidence of hundreds of co-expressed functional modules throughout the skydive is significantly robust. An important future direction would be to extend and replicate this exploratory study using a larger cohort of participants. Another putative psychological variable to consider while interpreting gender-specific stress responses is that the tandem skydive master was always male, which may provide a different environment for male and female participants.

6.4.7. Concluding Remarks

This exploratory study profiled the PBL transcriptome throughout a first-time tandem skydive, as a measure of intense acute psycho- logical stress, to reveal a detailed response to acute stress at the molecular level. A novel finding of the study is the degree of specificity of the immune response with respect to up-regulation of a subset of NK cell genes that cannot be solely attributed to the influx of NK cells into the periphery in response to stress parallel by increases in cortisol and catecholamines. Correcting differential gene expression analysis post-stressor revealed a molecular signature specific to the adrenal cortex. Network analysis stratified by gender identified hundreds of genes within several functional co- expression modules responding to stress in a gender-specific manner. These findings have potential implications for future research aimed at identifying therapeutic targets of stress-related disorders, and underscore the importance of gender-specific molecular profiles which could be used to better understand patterns of gender vulnerability to stress-induced disease.

Contributions. *These results are an effect of a large team effort. I played a role in the experimental design by incorporating a second cohort for which PBL samples were subjected to flow-cytometry. I was fully responsible for statistical design, data analysis, data interpretation and writing.*

Chapter 7

Candidate Blood Biomarkers and Gene Networks of Methamphetamine-Associated Psychosis

7.1. Background

Methamphetamine (METH) is a N-methyl derivative of amphetamine and a highly addictive psychostimulant (Yang et al., 2008). METH use is at epidemic levels in several areas of the world and its global prevalence is estimated at 15-16 million people with several pockets of increased use in the USA, Europe and Africa (UNO, 2004; Kapp, 2008). Recent evidence ranked METH fourth out of 20 of the most harmful drugs due to self-harm to the user (Nutt et al., 2010). One reason for this is that METH provokes psychotic reactions in 72-100% of all abusers (Srisurapanont et al., 2003; Smith et al., 2009).

Methamphetamine-associated psychosis (MAP) has been considered a pharmacological and environmental 'model' of schizophrenia (SCZ) due to similarities in clinical presentation (i.e. paranoia, hallucinations, disorganized speech and negative symptoms), response to treatment (neuroleptics), and presumed neuromechanisms (central dopaminergic neurotransmission) (Bousman et al., 2009; Hsieh et al., 2014; Srisurapanont et al., 2011). Better understanding of the molecular mechanisms underlying SCZ may be achieved through examination of human models with similarities to the disease. In this context, the MAP model could quicken the discovery of risk biomarkers, screening for sub-clinical disease, prognostics, diagnostics, or disease staging. However, several challenges currently exist in terms of accurately diagnosing MAP on a molecular and cognitive level before the MAP model could contribute to the discovery of SCZ biomarkers.

Genome-wide blood transcriptome profiling coupled with network analyses provide a platform for identifying functionally relevant biological markers of disease, permitting multi-scale data integration. This is a critical point as acute and chronic effects of METH use are widespread across the body and an integrative technique determining relationships of biological markers with magnetic resonance imaging (MRI), life events (i.e. stress, culture) and psychometric measurements could provide key insights into cognitive and molecular mechanisms of MAP, and into the versatility of the MAP model in molecular psychiatry research. Complimentary machine-learning provides a useful tool for *in silico* prediction of candidate biomarkers, and confirmation and validation of these biomarkers may be accomplished by utilizing convergent functional genomics (CFG) evidence. The CFG approach has proven highly successful for reducing false-positives and false-negatives moderately sized psychiatric cohorts by drawing on multiple disparate yet 'convergent' sources of external functional genomic information across independent human studies (Niculescu et al., 2000; Ogden et al., 2004; Patel et al., 2010; Le-Niculescu et al., 2009; Rodd et al., 2007; Le-Niculescu et al., 2011; Ayalew et al., 2012; Le-Niculescu et al., 2013; Kurian et al., 2009). Collectively, these techniques hold great promise for the prioritization and validation of candidate biomarkers for MAP and its relatedness to SCZ.

In the current investigation, we present a preliminary integrative RNA-Sequencing report exploring peripheral blood gene expression amongst subjects diagnosed with METH-associated psychosis (MAP) (N=10), METH-dependency without psychotic symptoms (MA) (N=10), and healthy control subjects (N=10). Additionally, subcortical brain structural volumes (sMRI) and a battery of self-reported psychometric measurements were collected for each subject. The primary objective of this study was to assess sMRI and clinical parameters of MAP within the framework of an integrative genome-wide RNA-Seq blood transcriptome analysis. Our cross-sectional experimental design allowed us to test three principal aims. **Aim 1:** First, to identify gene co-expression networks associated with MAP, later subjected to functional annotation and multi-modal data integration collected from the same subjects. **Aim 2**: Second, to perform supervised machine-learning classification based on differentially expressed genes to identify candidate blood-based biomarkers able to differentiate between MA, MAP and healthy control subjects. **Aim 3:** Finally, to validate the role of candidate blood biomarkers and gene networks in the pathophysiology of MAP using CFG information, and to confirm their shared association to psychotic disorders and SCZ in independent studies with the absence of METH.

7.2. Materials and Methods

7.2.1. Subject Selection

A total of 10 MAP subjects, 10 subjects with METH dependence (MA) without psychotic symptoms, and 10 healthy control subjects were enrolled in this study. Gender (male) and age matched (25.8 ± 6 years) right-handed subjects were recruited from drug rehabilitation facilities, hospitals and communities in Cape Town, South Africa where all subjects received detailed study information and provided written consent. Each subject underwent two assessment sessions. The first session consisted of a detailed psychiatric interview and demographic and substance variables were recorded. During the second session, approximately one week later, patients were asked to fast and refrain from smoking overnight, before blood was collected between 9:00-11:00. This was followed by a brain scan. Clinical assessment was performed using the Structured Diagnostic Interview for DSM-IV Axis I Disorders (SCID-I) (First et al., 2001), which classifies MAP based on the following criteria; (i) onset of psychosis within 1 month of last use and (ii) 1 month maximum duration of psychosis. All patients also completed a battery of self-report questionnaires including the Life Events Questionnaire (Brugha et al., 1990), Kessler Psychological Distress Scale (K10) (Kessler at al., 2002), Beck Depression Inventory (BDI-II) (Beck et al., 1996), Behavioural Inhibition System/Behavioural Activation System (BIS/BAS) scale (Carver et al., 1994), Eysenck Personality Questionnaire – Revised short scale (EPQR-S) (Eysenck et al., 1985). Positive and negative symptoms within the MAP group were rated using the Positive and Negative Syndrome Scale (PANSS) (Kay et al., 1987); PANSS positive subscale (14.5 ± 6.1), negative subscale (22.0 ± 11.5) and total score (66.8 ± 26.1).

Exclusion criteria comprised: 1) additional substance dependencies other than nicotine and METH for the MA and MAP groups, and any substance dependence other than nicotine in the control group; 2) lifetime and current diagnosis of any psychiatric disorder (other than MA dependence and MAP in the MA and MAP groups); 3) a history of psychosis prior to MA abuse; 4) a medical or neurological illness or head trauma; 5) a seropositive test for HIV; 6) MRI incompatibilities or known claustrophobia. All participants in the MAP group were receiving treatment with antipsychotic medication (haloperidol) at the time of testing. Polysubstance use was allowed to facilitate participant recruitment including nicotine, cannabis, and alcohol for all study groups. This study was approved (HREC REF 340/2009) by the University of Cape Town Faculty of Health Sciences Human Research Ethics Committee.

7.2.2. sMRI Acquisition and Image Processing

Subjects in this study form part of a larger project investigating fronto-temporal cortical and subcortical gray matter structures in MA and MAP. Images were acquired on a 3T Magnetom Allegra scanner (Siemens, Erlangen, Germany) at the Cape Universities Brain Imaging Centre (CUBIC). A high-resolution, T1-weighted, 3D-multi-echo MPRAGE sequence (scan parameters: TR=2530ms; graded TE=1.53, 3.21, 4.89, 6.57ms; flip angle=7°; FOV=256mm) produced 160 sagittal images of 1mm thickness. By acquiring four separate structural scans with graded TEs and averaging those into a final high contrast image (van der Kouwe et al., 2008), the MEMPRAGE method creates structural images with low distortion and high signal-to-noise ratio.

MRI scans were analysed using the FreeSurfer software package v5.1 (http://surfer.nmr.mgh.harvard.edu/). Regional estimates of subcortical volumes were assessed with a specialized surface-based reconstruction and automatic labelling tool, described in detail elsewhere (Fischl et al., 2004). FreeSurfer processing includes motion correction, skull-stripping, Talairach transformation, segmentation of subcortical white matter and deep gray matter volumetric structures, intensity normalization, tessellation of the gray matter/ white matter boundary, automated topology correction, and surface deformation.

7.2.3. RNA Isolation and RNA-Seq Library Preparation

Blood was collected using PAXgene RNA tubes (Qiagen, CA, USA) and total RNA was extracted and purified in accordance with the PAX gene RNA kit per manufacturer's instructions. Globin mRNA was depleted from samples using the GLOBINclear – Human Kit (Life Technologies, USA). Subsequently, the quantity of all purified RNA samples was measured on a nanodrop (56.6 ± 16.7ng/µl) and the quality and integrity measured with the Agilent 2100 Bioanalyzer (Agilent; CA, USA). All RNA passed integrity numbers > 7 (8.4 ± 0.7).

The Illumina TruSeq Stranded Total RNA kit (Ilumina, Inc.) was used for library preparation accordingly to manufacturer instructions without any modifications. The 30 indexed RNA libraries were pooled and sequenced using long paired-end chemistry (2x93 bp) on 7 lanes using the Illumina HiSeq2500. All replicates were run for 2x40 million reads per sample and all reads were primary processed using Casava v1.8.2 to transform primary base call files into fastq files.

7.2.4. Read Trimming, Mapping and Quantification of Expression

All fragmented RNA-Seq reads were trimmed to 90 bp and low quality reads were discarded using Trimmomatic (Bolger et al., 2014) options SLIDINGWINDOW:90:10 MINLEN:90 CROP:90. Subsequently, all high quality trimmed reads were mapped to UCSC *Homo sapiens* reference genome (build hg19) using TopHat v2.0.0 (Trapnell et al., 2009). I used the estimated mean inner distance and standard deviation between mate paired-ends as the -r and --mate-std-dev parameters, respectively. TopHat calls Bowtie v1.1.1 (Langmead et al., 2009) to perform alignment with no more than two mismatches. I used the pre-built index files of UCSC *H. sapiens* hg19, downloaded from the TopHat homepage (https://ccb.jhu.edu/software/tophat/igenomes.shtml). Samtools (Li et al., 2009) was used to convert bamfiles to samfiles and HTseq v0.6.0 (Anders et al., 2015) was used to count all of the mapped reads by htseq-count using parameters –stranded=reverse –q.

7.2.5. Data Pre-Processing

Raw count data measured 23,345 transcripts across 30 subjects. Unspecific filtering removed lowly expressed genes which did not meet the requirement of a minimum of 20 reads in at least 10 subjects. A total of 12,281 transcripts were retained, then subjected to edgeR VOOM normalization (Law et al., 2013), a variance-stabilization transformation method. Normalized data were inspected for outlying samples using unsupervised hierarchical clustering of subjects (based on Pearson coefficient and average distance metric) and principal component analysis to identify potential outliers outside two standard deviations from these averages. No outliers were present in these data and resulting normalised values were used as input for down-stream analyses.

7.2.6. Weighted Gene Co-expression Network Analyses

Signed co-expression networks were built using weighted gene co-expression network analysis (WGCNA) (Langfelder & Horvath, 2008) in R, as previously described. A total of 12,281 transcripts were used to construct a global network of all 30 subjects. To construct a network, the absolute values of Pearson correlation coefficients were calculated for all possible gene pairs and resulting values were transformed using a ß power of 9 so that the final correlation matrix followed an approximate scale-free topology. The WGCNA cut-tree hybrid algorithm was used to detect sub-networks, or co-expression modules, within the global network optimizing minimum module size to 15, deep split of 2, and a tree-cut height of 0.2 in order to merge neighbouring network modules with similar expression profiles. For each identified module, we ran singular value decomposition of each module's expression matrix and used the resulting module eigengene (ME), equivalent to the first principal component, to represent the overall expression profiles for each module. Differential expression of MEs was performed using a Bayes ANOVA (Kayala & Baldi et al., 2012) (parameters: conf=12, bayes=1, winSize=5) testing between groups, and *P* values were corrected for multiple comparisons (post-hoc Tukey correction). Subsequently, to determine which modules were most associated with clinical parameters and potential confounding variables, MEs for all modules were correlated to external subjective and objective data using a Pearson correlation and a Student's asymptotic *P* value for significance. MEs were also used to determine module membership (kME) values for each gene in a specified module, defined as the correlation between gene expression values and ME expression. Genes with the highest intramodular kME were labelled hub genes and predicted to be essential to the function of the module.

7.2.7. Differential Gene Expression Analyses

A moderated *t* test, implemented through the *limma* (Smyth, 2005) package, assessed differential gene expression between the three groups in a group-wise fashion across 12,281 transcripts. Significance threshold was set to a nominal *P* value < 0.01 to permit sufficient genes to move forward with functional characterization and supervised classification methods. Differentially expressed genes corresponding to WGCNA modules which were significantly associated with polysubstance abuse were excluded and removed from functional annotation and supervised classification methods, as a robust and complimentary strategy of adjusting for confounding factors.

7.2.8. Functional and Cellular Enrichment Analyses

All differentially expressed genes passing a *P* value < 0.01 and all network modules with genes passing a kME > 0.50 were subjected to functional annotation. First, the ToppFunn module of ToppGene Suite software (Chen et al., 2009) (https://toppgene.cchmc.org/) was used to assess enrichment of GO ontology terms relevant to cellular components, molecular factors, biological processes, metabolic pathways and well annotated drug-compounds from the comparative toxicogenomics database (Davis et al., 2015) (CTD), using a one-tailed hyper geometric distribution with a Bonferroni correction. A minimum of a two gene overlap per gene-set was necessary to be allowed for testing. The human cell specific (HECS) gene expression database from the cell type enrichment (CTen) (Shoemaker et al., 2011) analysis web-based tool was used to predict the involvement of key cell types within candidate gene lists. For each supplied gene list, the significance of cell type specific expression are determined using the one-tailed Fishers-exact test with a Bonferroni correction across all available cell/tissue types.

7.2.9. Supervised Machine-Learning Classification

BRB-Array Tools (Simon et al., 2007) supervised classification methods were used to construct gene expression classifiers. Two models were specified: (1) controls vs. METH dependents and (2) MA vs. MAP subjects. Each model consisted of three steps. First, all genes with $P < 0.01$ were subjected to classifier construction. These criteria were used to cast a wide net to catch all potentially informative genes, while false-positives could be discarded by subsequent optimization and cross-validation steps. Second, classifiers composed of different numbers of genes were constructed by recursive feature elimination (RFE). RFE provided feature selection, model fitting and performance evaluation via identifying the optimal number of features with maximum predictive accuracy. Third, the ability for RFE to predict group outcome was assessed by diagonal linear discriminant analysis (DLDA) and compared to three different multivariate classification methods (i.e. support vector machine (SVM), nearest centroid (NC), three-nearest neighbors (3NN)) in a leave-one-out cross-validation (LOOCV) approach. Additionally, a permutation *P* value, based on 1000 random permutations, for the cross-validated misclassification error rate for each classification method was implemented. This *P* value indicates the proportion of the random permutations that gave as small a cross-validated misclassification rate as was obtained with the real class labels.

7.2.10. Convergent Functional Genomic (CFG) Scoring

Convergent functional genomics (CFG) represents a translational methodology that integrates multiple lines of external evidence from human and animal model studies in a Bayesian-like fashion. This approach increases the ability to distinguish signal from noise in limited size cohorts and is routinely applied to support the identification of blood biomarkers across neuropsychiatric disorders (Niculescu et al., 2000; Ogden et al., 2004; Patel et al., 2010; Le-Niculescu et al., 2009; Rodd et al., 2007; Le-Niculescu et al., 2011; Ayalew et al., 2012; Le-Niculescu et al., 2013; Kurian et al., 2009). The principal aim of the CFG approach is to increase the likelihood that findings will prove reproducible and have predictive power in independent cohorts. Our CFG scoring paradigm for prioritization of MAP biomarkers is an adaptation of previous techniques, representing a two-step process (**Supplementary Figure 7.1**):

Internal lines of evidence: All genes assigned a *P* value < 0.05 were included in the CFG scoring. These liberal criteria were used to cast a wide net to find all potentially informative genes which may be involved in the pathophysiology of MAP, while false-positives would be pared off by subsequent CFG scoring and optimization steps. Each gene was given 3 *P* values (based on 3 group-wise differential expression analyses). Subsequently, a score of 1 was given to genes passing *P* < 0.001, a score of 0.5 was given to genes passing 0.001 > *P* < 0.01, and a score of 0.2 was given for genes passing 0.01 > *P* < 0.05, permitting a maximum score of 3 and a minimum score of 0.2. A bonus point of 0.5 was awarded for genes passing *P* < 0.01 occurring in both MAP vs. controls and MAP vs. MA comparisons as well as genes found to be members of MAP associated modules. Thus, a max score of 4 is attainable (3 + 0.5 + 0.5).

External lines of evidence: CFG evidence was scored for a gene if there were published reports of human data including post-mortem brain expression, peripheral blood expression and/or genetic evidence (association and linkage) utilizing two large databases. The first database represents a recently built in-house database specific to human blood transcriptome studies using PubMed (http://www.ncbi.nlm.nih.gov/pubmed) search queries and combinations of key words (e.g. blood transcriptomics AND psychosis) (Breen et al., 2016 under revision). To do so, we performed the following PubMed queries:

*(blood OR PBMC OR PBMCs OR PBL OR PBLs OR peripheral blood leukocytes OR peripheral blood OR leukocytes OR blood-based OR blood-based biomarker) AND (transcriptome OR transcriptomics OR RNA-Sequencing OR RNA-Seq OR RNAseq OR RNASeq OR RNA Sequencing OR microarray OR microarrays OR blood gene expression OR peripheral blood gene expression OR leukocyte gene expression) AND (“**disease term x**”)*

A total of 4 independent queries were performed in which terms ‘psychosis’, ‘schizophrenia’, ‘depression’ and ‘neurocognitive impairment’ were substituted with the above-stated “***disease term x***” and results were pooled. The search was limited to human studies published in the last 10 years (up to August 2015). Studies using transcriptomic platforms to profile miRNA, mRNA or lnRNA were included while studies using qualitative real-time PCR (RT-qPCR) as a means to investigate a targeted panel of candidate genes were removed. We excluded studies investigating molecular mechanisms in lymphoblastoid cell lines derived primary cells and skin fibroblast cultures in order to retain true peripheral blood signatures. Review papers and secondary data integration analyses were not included. Publications were grouped based on disease type.

Second, we sought to consider functional support across divergent technological platforms and human post-mortem brain samples. To do so, we accessed DisGenNet (Piñero et al., 2015), a comprehensive database of human gene-disease associations from various expert curated databases and text-mining derived associations. These database searches included gene-disease relationships focusing specifically on psychosis, SCZ, depression/stress and neurocognitive impairment, in order to consider comorbidity in MAP in our study. Studies containing a METH component were excluded in order to validate MAP biomarkers in drug-free (METH) models. For the CFG analysis and scoring, external cross-validating lines of evidence were weighted such that findings in human peripheral blood specific to psychosis were given an additional 1 point. A maximum of 5 external lines of evidence were allowed. Thus, the total maximum CFG score that a candidate biomarker gene could have was 10 (4 for threshold + 5 for external evidence + 1 blood presence in psychosis). Like other studies using this approach, there are other ways of scoring blood biomarkers based on CFG which may give slightly different results in terms of prioritization. Given the past utility of this approach, this empirical scoring system allows for advantageous separation of genes based on our focus for identifying human MAP blood biomarker and by default, biomarkers of psychosis and SCZ.

7.3. Results

We conducted a preliminary integrative RNA-Sequencing study profiling peripheral blood gene expression from a primary cohort of 10 MA, 10 MAP and 10 healthy controls (**Table 7.1**). To identify and prioritize diagnostic blood biomarkers of MAP, a multimodal translational approach was used (**Figure 7.1**). A global gene co-expression network was first constructed using all available subjects and identified 24 co-expression modules, which were functionally annotated to molecular factors, biological processes, cellular compartments, metabolic pathways, well characterized drug compounds and cell type specificity (**Supplementary Table 7.1**).

Table 7.1. Recorded clinical characteristics from all subjects (N=30).

	Healthy Controls (N=10)	**MA (N=10)**	**MAP (N=10)**	**ANOVA**		**Post-Hoc Significance**
	Mean ± SD	Mean ± SD	Mean ± SD	X^2 *(df=2)*	*P* value	Bonferroni *P* value
Age	25.5 ± 5.8	24.8 ± 3.9	27.2 ± 8.3	0.040	0.980	
Education Level	12.2 ±1.2	10.7 ± 2.1	9.3 ± 1.7	10.788	0.005	Contol > MAP
METH Age Started Using	-	18.6 ± 3.9	18.8 ± 6.8	0.191	0.662	
METH Abstinence (days)	-	53.1 ± 82.9	45.5 ± 36.2	0.593	0.441	
METH Duration of use (years)	-	5.8 ± 2.3	7.1 ± 3.0	0.688	0.407	
Nicotene Use Last 30 Days	5	6	9	2.400	0.121	
Cannabis Use Last 30 Days	2	2	1	0.529	0.467	
Alcohol Use Last 30 Days	3	4	2	1.347	0.246	
EPQRS-Psychoticism	2.3 ± 1.7	1.6 ± 1.2	3 ± 2.1	1.880	0.391	
EPQRS-Extraversion	10.3 ± 2.5	8.2 ± 3.5	6.6 ± 2.5	7.039	0.030	Contol > MAP
EPQRS-Neuroticism	2.6 ± 1.8	4.6 ± 2.9	5.6 ± 3.2	4.624	0.099	
EPQRS-Lie	5.6 ± 2.3	4 ± 1.9	5.1 ± 3.3	1.902	0.386	
EPQRS Total Score	20.8 ± 5.3	18.5 ± 2.3	20.4 ± 4.7	1.876	0.391	
BIS	15.1 ± 1.5	15.8 ± 3.1	13.1 ± 3.6	3.018	0.221	
BAS Drive	7.4 ± 2.5	8.3 ± 2.6	6.5 ± 1.3	2.267	0.322	
BAS Fun Seeking	7.1 ± 1.5	8.1 ± 1.6	6 ± 1.2	7.014	0.030	MA > MAP
BAS Reward Responsiveness	7.7 ± 1.9	7.2 ± 1.8	6.2 ± 1.7	3.859	0.145	
BIS-BAS Total Score	44.8 ± 5.8	47 ± 7.9	38.4 ± 5.6	6.269	0.044	
BDI Total Score	4.3 ± 3.0	17.3 ± 10.3	16.6 ± 12.5	10.363	0.006	Control > MAP; Control > MA
K10 Total Score	14 ± 3.8	18.2 ± 7.7	23.5 ± 8.2	7.944	0.019	Control > MAP
LEQ - Sum of life events ($\leq$ 6 months)	2.6 ± 1.7	4.4 ± 2.0	4.7 ± 1.6	5.663	0.059	
LEQ - Sum of life events (> 6 months)	2.2 ± 2.2	4.2 ± 3.5	4.1 ± 2.0	3.643	0.162	

Abbreviations; MA, Methamphetamine dependent subjects with no psychiotic events; MAP, Methamphetamine-associated psychosis; EPQR-S, Eysenck Personality Questionnaire; BIS/BAS,Behavioural Inhibition System/Behavioural Activation System,BDI, Beck Depression Inventory; K10, Kessler Psychological Distress Scale; LEQ,Life Events Questionnaire; PANSS, Positive and Negative Syndrome Scale. Shapiro wilk test was used to assess normality of variables and either a one-way ANOVA or KRUSKAL-Wallis ANOVA with post-hoc Bonferroni correction was implemented accordingly. Grey shading is for visualization purposes only.

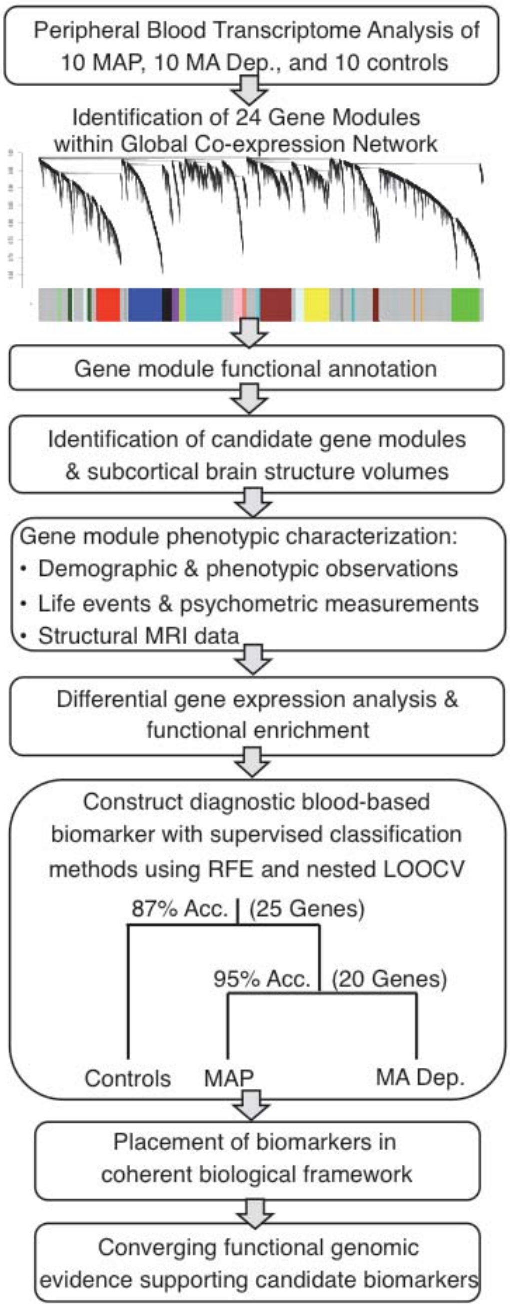

Peripheral Blood Transcriptome Analysis of 10 MAP, 10 MA Dep., and 10 controls
Identification of 24 Gene Modules within Global Co-expression Network
Gene module functional annotation
Identification of candidate gene modules & subcortical brain structure volumes
Gene module phenotypic characterization:
• Demographic & phenotypic observations
• Life events & psychometric measurements
• Structural MRI data
Differential gene expression analysis & functional enrichment
Construct diagnostic blood-based biomarker with supervised classification methods using RFE and nested LOOCV
87% Acc.
(25 Genes)
95% Acc.
(20 Genes)
Controls
MAP
MA Dep.
Placement of biomarkers in coherent biological framework
Converging functional genomic evidence supporting candidate biomarkers

Figure 7.1. A multi-step translational work-flow for identifying MAP biomarkers. First, WGCNA analysis built a global co-expression network and identified 24 co-expression modules. On the hierarchical cluster tree each line represents a gene (leaf) and each group of lines represents a discrete group of co-regulated genes, or gene modules (branch) on the clustering gene tree. Each gene module is indicated by the colour bar below the dendrogram, and subsequently functionally annotated then integrated with recorded clinical and biological data to identify candidate gene modules representing functional biomarkers of MAP. Second, differential gene expression and class prediction methods identified 20 candidate MAP biomarkers (14 were recycled from the second split on the tree). A Bayesian-like convergent functional genomic (CFG) approach prioritized our panel of biomarkers specific to MAP and biomarkers were placed within an empirically derived biological framework.

7.3.1. Differential Analysis of Modules and Brain Volumes

All ME values were subjected to a Bayes ANOVA[32] testing to compare the extent of module expression between groups and *P* values were correcting for multiple comparisons. MAP associated findings included significant decreases of ME expression in modules specific to ‘ubiquitin-mediated proteolysis’ (767 genes) and ‘RNA degradation’ (1156 genes) in MAP subjects compared to controls (*P*=0.01, *P*=0.03, respectively), and MA subjects compared to controls (*P*=0.07, *P*=0.055, respectively) (**Figure 7.2A-B**). Further, an increase of ME expression in a module annotated as ‘circadian clock’ (332 genes) was observed in MAP compared to controls (*P*=0.04) (**Figure 7.2C**). MA associated findings included the increase of ME expression in modules specific to ‘chloride transporter activity’ (106 genes), ‘interferon signalling’ (263 genes), and ‘cytokine signalling’ (186 genes), and a decrease of ME expression in modules associated with ‘generic transcription’ (48 genes), and ‘ribosome pathway’ (281 genes) in MA subjects relative to healthy controls **(Figure 7.3).** The same methodology was extended to compare brain structural volumes *(mm3)* across the three groups, which revealed bilaterally reduced hippocampus volumes in MAP subjects (left, *P*=0.04; right, *P*=0.02) (**Table 2**).

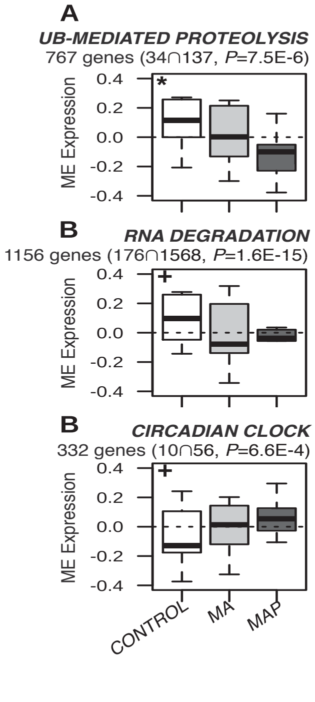


Figure 7.2. Significant MAP findings from differential analysis of module eigengene (ME) values and brains structural volumes (mm3) across controls (white), MA subjects (light grey) and MAP subjects (dark grey). Modules specific to MAP include **(A)** ubiquitin(UB)-mediated proteolysis, **(B)** RNA degradation and **(C)** circadian clock. Indicated for each module are, number of overlapping genes from the module ∩ out of total genes in the term. Enrichment *P* values are Bonferroni corrected for multiple comparisons. A Bayes ANOVA (parameters: conf=12, bayes=1, winSize=5) was used on ME values to test for significance between groups and *P* values were corrected multiple comparisons where (*) implies post-hoc corrected *p*-value significance < 0.05 and ($^+$) indicates *p*-value significance < 0.05 without post-hoc correction.

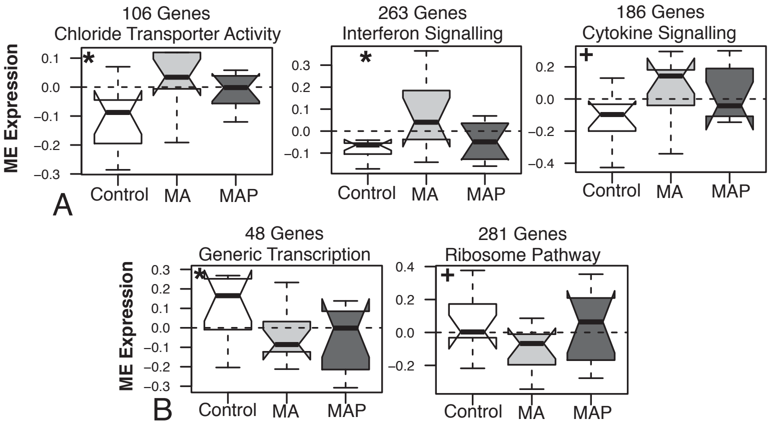


Figure 7.3. Module eigengene (ME) differential expression analysis across controls (white), MA dependent subjects (light grey) and MAP subjects (dark grey). (**A**) ME expression values over-expressed in MA subjects relative to controls include modules enriched for chloride transporter activity, interferon signalling and cytokine signalling. (**B**) ME expression values under-expressed in MA subjects relative to controls include modules enriched for generic transcription and ribosome pathway. A Bayes ANOVA (parameters: conf=12, bayes=1, winSize=5) was used on ME values to test for significance between groups and corrected for multiple comparisons where (*) implies post-hoc corrected *p*-value significance < 0.05 and ($^{+}$) indicates *p*-value significance < 0.05 without post-hoc correction.

Table 7.2. Brain structural volumes (mm3) from all subjects (N=30).

Brain Region	**Healthy Controls (N=10)** Mean ± SD	**MA (N=10)** Mean ± SD	**MAP (N=10)** Mean ± SD	**Bayes ANOVA** *F (df=2)*	*P* value	**Post-Hoc Significance**
L. Hippocampus	3950.11 ± 463.71	3790 ± 297.51	3521.71 ± 173.43	3.538	0.041	Control > MAP
R. Hippocampus	4067.56 ± 414.08	4005.43 ± 196.29	3645.29 ± 189.97	4.261	0.029	Control > MAP
L. Accumbens	690.56 ± 80.38	689.14 ± 128.15	651.57 ± 99.24	0.343	0.714	
R. Accumbens	669.33 ± 100.54	673.00 ± 199.23	694.71 ± 91.48	0.076	0.927	
L. Caudate	4116.89 ± 340.84	4078.57 ± 293.78	3906.71 ± 177.23	1.149	0.337	
R. Caudate	4211.22 ± 251.11	4283.86 ± 314.36	4119 ± 163.64	0.760	0.481	
L. Putamen	6606.78 ± 408.97	6633.14 ± 667.17	6718.57 ± 661.5	0.078	0.925	
R. Putamen	6313.33 ± 371.03	6274.43 ± 596.45	6506.71 ± 672.14	0.373	0.694	
L. Ventral DC	4551.33 ± 247.16	4295.71 ± 273.56	4323.71 ± 204.25	2.715	0.091	
R. Ventral DC	4473.44 ± 377.34	4340.43 ± 78.7	4369.86 ± 278.58	0.485	0.623	
CC Anterior	938.78 ± 125.96	1056.14 ± 194.83	1016.57 ± 100.31	1.389	0.272	
CC Posterior	966.00 ± 191.65	912.29 ± 139.86	956.29 ± 135.16	0.236	0.792	

Bayes ANOVA parameters: conf=12, bayes=1, winSize=5. *P* values corrected for multiple comparisons. Abbreviations:L., left; R., right; DC, diencephalon; CC, corpus callosum. Grey shading is for visualization purposes only.

7.3.2. Phenotypic Characterisation of MAP Modules

ME values for MAP specific modules were correlated with all phenotypic traits in this study (brain structural volumes, life history and psychometric measures) to gain insight into the role that each module may play in the pathophysiology of the disorder **(Figure 7.4)**. *P* values < 0.002 pass the most conservative multiple comparison correction

(Bonferroni). The ME of a ‘ubiquitin-mediated proteolysis’ module was negatively associated with MAP status (*r*=-0.45, *P*=0.01) as well as K10 total score (*r*=-0.43, *P*=0.02). Interestingly, this module was also negatively associated with brain structure volumes in areas of the anterior corpus callosum (CC) (*r*=-0.55, *P*=0.002), right accumbens area (*r*=-0.40, *P*=0.03) and positively associated with areas in the left caudate (*r*=0.37, *P*=0.04) and left ventral DC (*r*=0.48, *P*=0.007). The ‘RNA degradation’ module was negatively associated with the CC anterior (*r*=-0.48, *P*=0.008) and left accumbens (*r*=0.50, *P*=0.005), while positively associated with the left ventral DC (*r*=0.37, *P*=0.04). The ‘circadian clock’ module, was positively correlated with EPQRS measure of psychoticism (*r*=0.43, *P*=0.02) and negatively associated to extraversion (*r*=-0.36, *P*=0.04).

7.3.3. Phenotypic Characterisation of MA Modules

A similar strategy was chosen to characterise MA specific modules **(Figure 7.4)**. The ME of the ‘interferon signalling’ module was positively associated with MA status (*r*=0.40, *P*=0.03), BDI total score (*r*=0.40, *P*=0.03) and with structural information from both left (*r*=0.54, *P*=0.002) and right putamen areas (*r*=0.41, *P*=0.03). This module was negatively associated to EPQRS measure of extraversion (*r*=-0.38, *P*=0.04) and EPQRS total score (*r*=-0.38, *P*=0.04). Further, the ME of the ‘chloride transporter activity’ module was positively associated with both MA status (*r*=0.36, *P*=0.05) and METH dependency (*r*=0.39, *P*=0.03), in addition to BDI total score (*r*=0.39, *P*=0.03) and brain volume in the left putamen (*r*=0.53, *P*=0.003). This module was also negatively associated with control status (*r*=-0.39, *P*=0.03) and the left ventral diencephalon (DC) (*r*=-0.40, *P*=0.03). The ‘ribosome pathway’ module was negatively associated with MA status (*r*=-0.37, *P*=0.04) and positively associated with EPQRS total score (*r*=0.38, *P*=0.04), and K10 total score (*r*=0.44, *P*=0.02). The ‘cytokine signalling’ module was positively associated with both left accumbens (*r*=0.37, *P*=0.04) and right accumbens (*r*=0.55, *P*=0.002), while the ‘generic transcription’ module was negatively associated with these areas (*r*=-0.49, *P*=0.006; *r*=-0.60, *P*=5e-04, respectively).

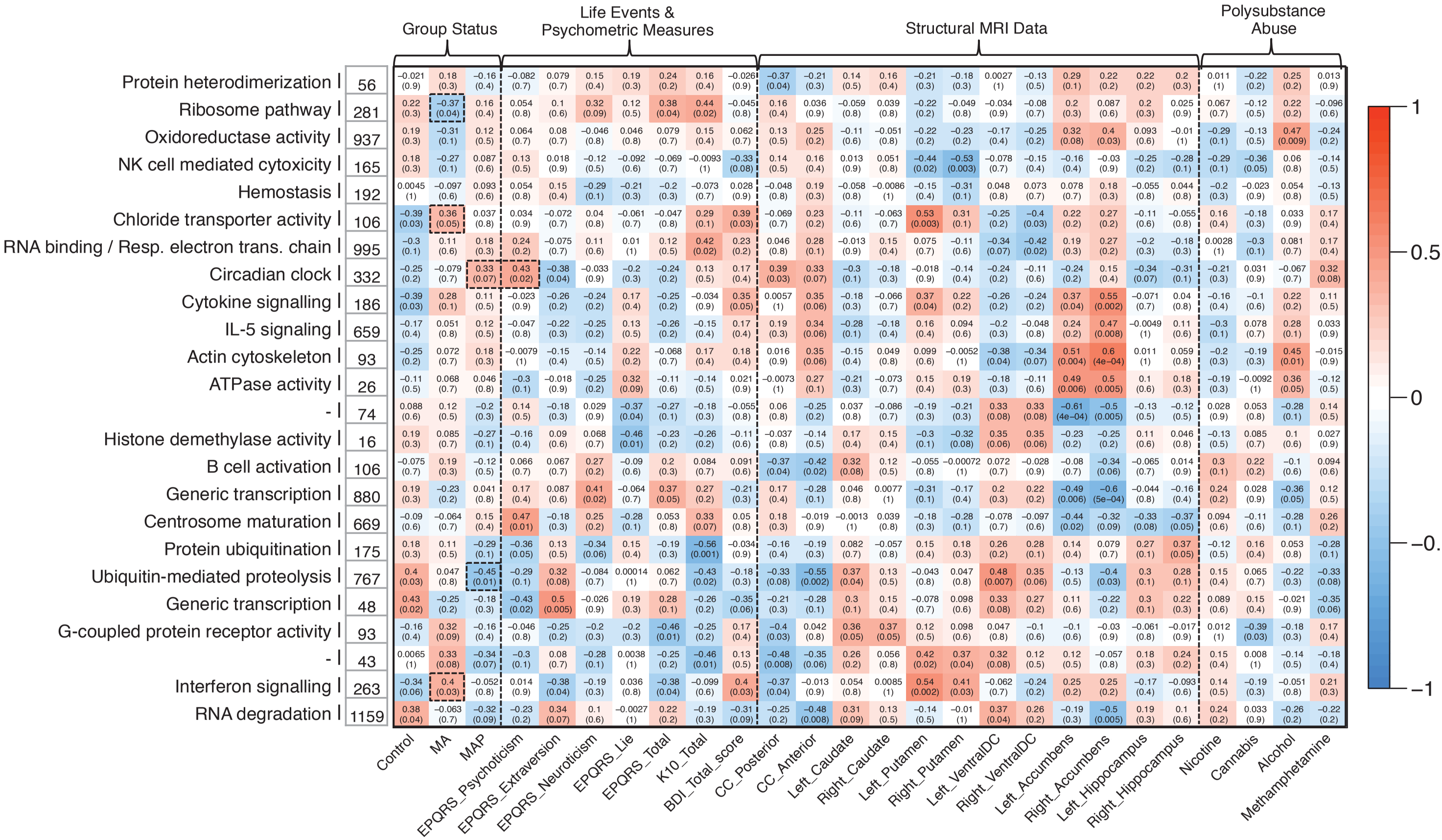


Figure 7.4. Module eigengene (ME) associations. Functional gene modules were associated with external data including self-reported measurements, life history, behavioural and depression scores, structural MRI data and polysubstance abuse (confounding factors). The primary function of each gene co-expression module is labelled on the y-axis with the corresponding number of genes with kME > 0.5 within each module. The measure of correlation, *r*, is the top value in each box and the related *p*-value is designated below within brackets. Red signifies a positive correlation and blue signifies a negative correlation as indicated by colour scale. Significant MA and MAP associations are outlined in boxes. Due to the high number of self-reported measures and structural MRI data we only report on those variables that drew at least one significant ME association. '-' indicates no principal function was identified.

7.3.4. Putative Diagnostic Blood Biomarker for MAP

Supervised class prediction methods were used to identify any single important gene(s) which may have been over-looked in our network analysis. First, differentially expressed genes (all $P < 0.01$) were identified between control and MA subjects (N=197), control and MAP subjects (N=409) and between MA and MAP subjects (N=79) (**Figures 7.5A-D**). To control for confounding factors, genes corresponding to WGCNA modules significantly associated with polysubstance abuse were excluded. Gene lists were annotated for functionality at the pathway level and cross-referenced with drug-induced gene signatures from the CTD database (**Figure 7.5E-F**).

Subsequently, differentially expressed genes ($P < 0.01$) were pooled from across the three candidate gene lists and subjected to RFE feature selection and different multivariate classification methods in a LOOCV approach. Two models were built for separating classes. First, when separating healthy controls from METH dependents (MA and MAP subjects) classification accuracy reached 87% when the expression of 25 genes was used with DLDA multivariate classification method (**Figure 7.6A-B & Table 7.3**). Second, when separating MA from MAP, classification accuracy reached 95% when the expression of 20 genes (recycling 14 genes from the first model) was used with DLDA (**Figure 7.6C-D & Table 7.4**).

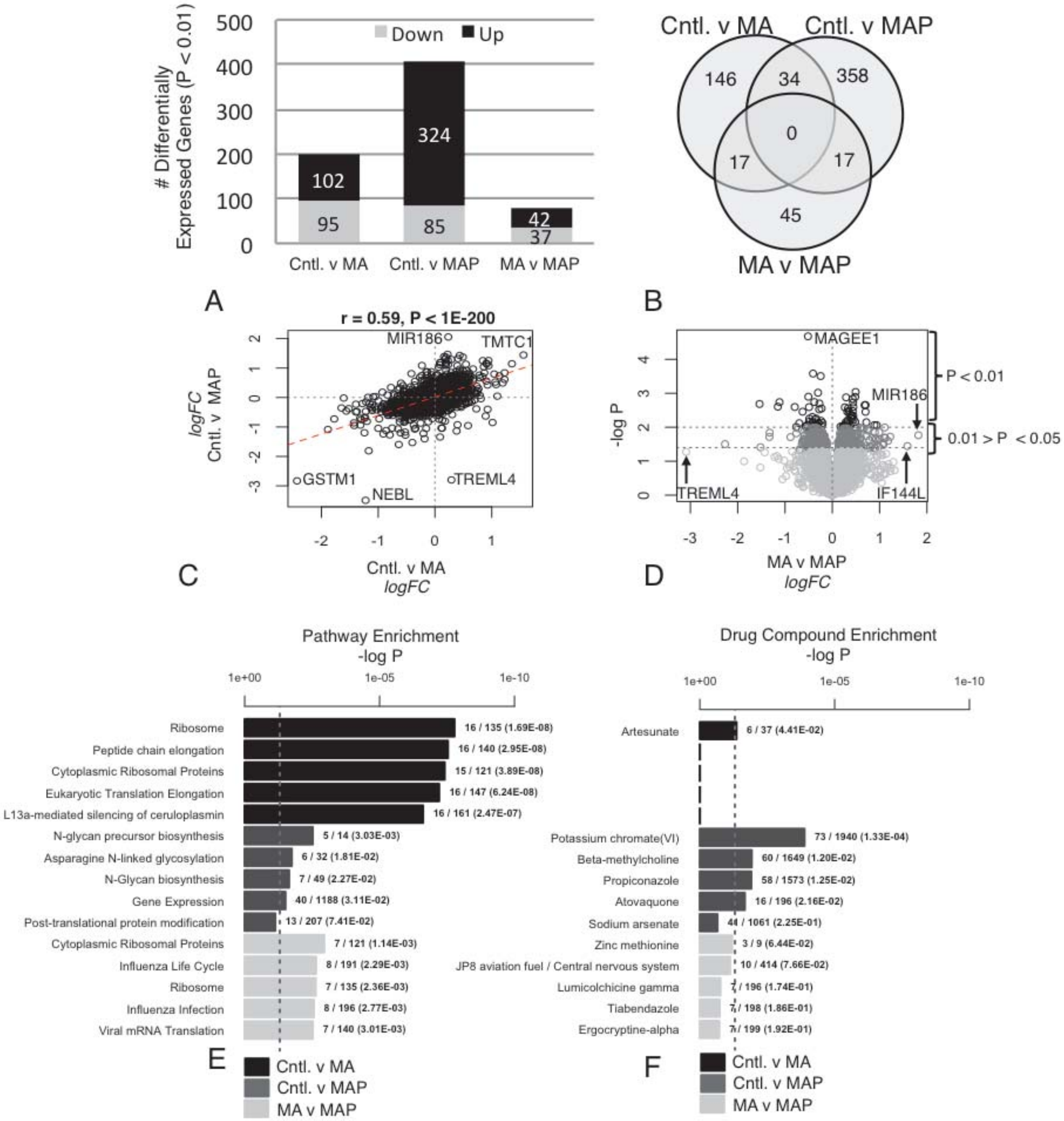
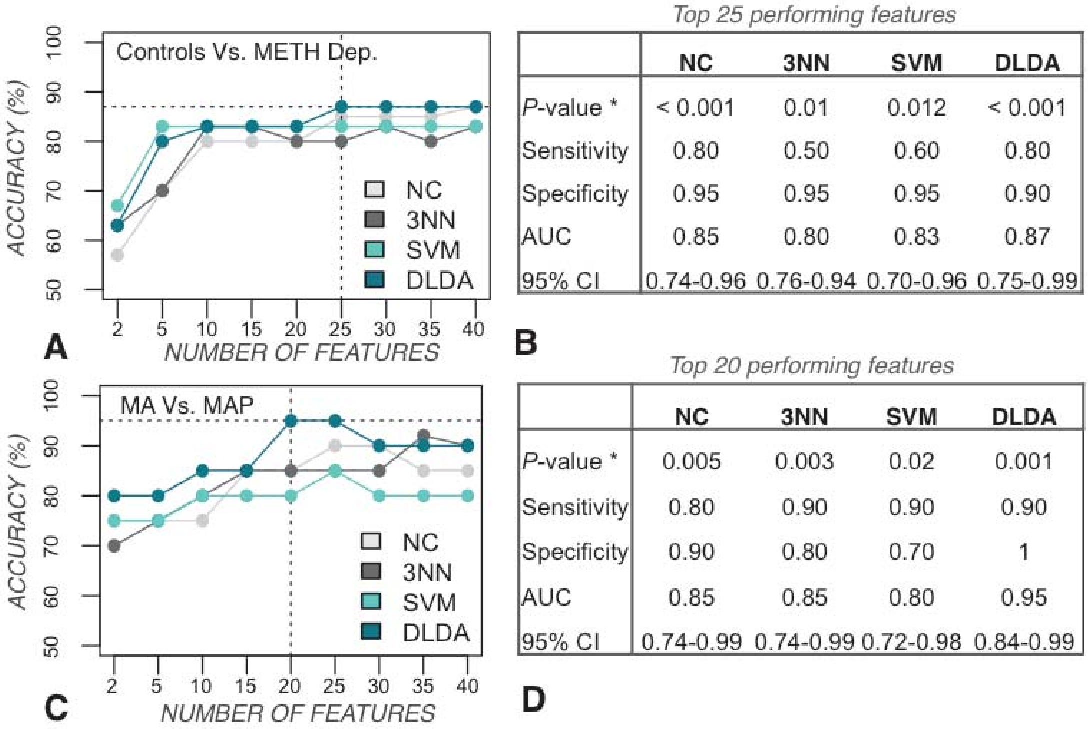


Figure 7.5. Differential gene expression analyses. (**A**) The total number of over- and under-expressed genes ($P < 0.01$) are shown for each pair-wise group comparison and subsequently (**B**) the overlap of all identified genes are displayed. (**C**) Log fold-change (logFC) of all genes between controls and MAP subjects were associated with logFC values for genes between controls and MA subjects, MAP or MA specific genes are labelled. (**D**) Volcano plot (logFC vs. log *p*-value) of dysregulated genes between MA and MAP subjects, *P* values coloured by significance. (**E**) The top 5 most significantly enriched pathways and (**F**) drug-compounds for each pair-wise comparison (Bonferroni $p < 0.05$).



Top 25 performing features

	NC	3NN	SVM	DLDA
P-value *	< 0.001	0.01	0.012	< 0.001
Sensitivity	0.80	0.50	0.60	0.80
Specificity	0.95	0.95	0.95	0.90
AUC	0.85	0.80	0.83	0.87
95% CI	0.74-0.96	0.76-0.94	0.70-0.96	0.75-0.99

Top 20 performing features

	NC	3NN	SVM	DLDA
P-value *	0.005	0.003	0.02	0.001
Sensitivity	0.80	0.90	0.90	0.90
Specificity	0.90	0.80	0.70	1
AUC	0.85	0.85	0.80	0.95
95% CI	0.74-0.99	0.74-0.99	0.72-0.98	0.84-0.99

Figure 7.6. Two separate models were used to predict group outcomes. (**A**) Gene expression classifier accuracies achieved when discriminating between healthy control and METH dependent subjects (MA + MAP groups) and (**B**) results of the top performing model containing 25 genes are displayed. (**C**) Gene expression classifier accuracies achieved when discriminating between MA subjects from MAP subjects and (**D**) results of the top performing model containing 20 genes are displayed. In each case supervised class prediction was performed using different combinations of genes with Recursive Feature Elimination and evaluated with four different multivariate classification methods. Abbreviations; P-value *, result of 1000 random permutations to class labels; AUC, overall balanced accuracy; CI, confidence interval; NC, nearest centroid; 3NN, three-nearest neighbors; SVM, support vector machine; DLDA, diagonal linear discriminate analysis.

To understand the biology represented by these MAP biomarkers and to derive mechanistic insights, our multi-step approach permitted taking each biomarker and returning to our network analysis to retrieve guilt-by-association biological information from empirically derived functional gene modules. The majority of these genes were found in a module annotated to ‘RNA degradation’ (*CLN3*, *FBP1*, *TBC1D2*, *ZNF821*, *ADAM15*, *ARL6*, *FBN1* and *MTHFSD*) (**Table 7.3 & 7.4**). However, two top scoring biomarkers were found to be implicated in ‘circadian clock’ dysfunction (*ELK3* and *SINA3*) and three other top scoring biomarkers were found in the module annotated to ‘ubiquitin-mediated proteolysis’ (*PIGF*,*UHMK1* and *C7orf11*).

Table 7.3. Top twenty-five informative features separating controls and METH dependents.

Gene Symbol	Parametric p-value	% CV support	Module Correspondence	Significant Positive Correlations	Significant Negative Correlations
ELK3†	0.0175377	97	Circadian Clock	•EPQRS Psychoticism (*r* = 0.43, *P*=0.02) •CC Posterior (r = 0.39, p=0.03)	•EPQRS Extraversion (*r* = -0.38, *P*=0.04)
CRTAM	0.03485	97	Generic Transcription	•EPQRS Neuroticisim (*r* = 0.41, *P*=0.02) •EPQRS Total (*r* = 0.37, *P*=0.05)	•Left Accumbens (*r* = -0.49, *P*=0.0006) •Right Accumbens (*r* =- 0.6, *P*=0.00005)
MAGEE1	0.0158379	100			
RNF138P1	0.0078459	87	RNA Degradation	•Control Status (*r* = 0.38, *P*=0.04) •Left Ventral DC (*r* = 0.37, *P*=0.04)	•CC Anterior (*r* = -0.48, *P*=0.008) •Right Accumbens (*r* = -0.5, *P*=0.005)
MFN1	0.0070206	87			
TBC1D2*	0.000805	90			
ZNF286B	0.000065	90			
MRPL50	0.0001267	93			
ADAM15†	0.000731	97			
DDRGK1	0.0055266	97			
MTHFSD	0.0612426	97			
ARL6	0.0037554	97			
GKAP1	0.0009407	97			
FAM169A	0.0008839	97			
KBTBD6	0.0003548	97			
ZSCAN5A	0.0012892	100			
FBN1†*	0.0168567	100			
ZNF821	0.0193724	100			
FBP1†	0.6900462	100			
CDK7	0.0054503	100			
CDYL2	0.0000834	93	RNA-Binding	•K10 Total (*r* = 0.42, *P*=0.02)	•Right Ventral DC (*r* = -0.42, *P*=0.02)
TOMM34	0.0019291	100			
C7orf11	0.0587445	80	Ubiquitin-Mediated Proteolysis	•Control status (*r* = 0.4, *P*=0.03) •Left Caudate (*r* = 0.37, *P*=0.04) •Left Ventral DC (*r* = 0.48, *P*=0.0007)	•Control status (*r* = 0.4, *P*=0.03) •Left Caudate (*r* = 0.37, *P*=0.04) •Left Ventral DC (*r* = 0.48, *P*=0.0007)
UHMK1†	0.6057577	97			
PHLDB2	0.0007613	100			

Parametric p-value indicates significance in a strict sense following 1000 random permutations to group labels using small N. %CV support denotes the number of correctly passed cross-validations for each gene. Module correspondence is the module membership of each gene and the subsequent significant correlations for each module are depicted. Genes in bold are those which were used in classification for nodes 1 and 2 (14 genes total). Genes found dysregulated in the blood of psychosis studies (*), or as genetic variants in SCZ studies (†).

Table 7.4. Top twenty informative features separating MAP from MA subjects.

Gene Symbol	Parametric p-value	% CV support	Module Correspondence	Significant Positive Correlations	Significant Negative Correlations
SIN3A*	0.0926295	70	Circadian Clock	•EPQRS Psychoticism (r = 0.43, P=0.02) •CC Posterior (r = 0.39, P=0.03)	•EPQRS Extraversion (r = -0.38, P=0.04)
ELK3†	0.0002902	90			
MAGEE1	0.0001558	100	Generic Transcription	•EPQRS Neuroticisim (r = 0.41, P=0.02) •EPQRS Total (r = 0.37, P=0.05)	•Right Choroid Plexus (r = -0.39, P=0.03) •Left Accumbens (r = -0.49, P=0.0006) •Right Accumbens (r =- 0.6, P=0.00005)
MFSD7	0.0440767	85	Interferon Signalling	•MA Dep. Status (r = 0.4, P=0.03) •BDI Total (r = 0.4, P=0.03) •Left Putamen (r = 0.54, P=0.002) •Right Putamen (r = 0.41, P=0.03)	•EPQRS Extraversion (r = -0.43, P=0.02) •EPQRS Total (r = -0.43, P=0.02) •CC Posterior (r = -0.43, P=0.02)
SLC41A3	0.0018933	100	Ribosome Pathway	•EPQRS Total (r = 0.38, P=0.04) •BDI Total (r = 0.44, P=0.02)	•MA Dep. status (r =- 0.37, P=0.04)
MTHFSD	0.0002405	90	RNA Degradation	•Control Status (r = 0.38, P=0.04) •Left Ventral DC (r = 0.37, P=0.04)	•CC Anterior (r = -0.48, P=0.008) •Right Accumbens (r = -0.5, P=0.005)
ZNF821*	0.11798	90			
FBP1*	0.3549855	90			
RNF138P1	0.0195014	90			
ARL6	0.0317958	95			
ETFA	0.0235683	95			
TBC1D2*	0.157939	100			
FAM169A	0.0132909	100			
ZSCAN5A	0.0112376	100			
CLN3	0.0087815	100			
DDRGK1	0.0082958	100			
FBN1†*	0.0070075	100			
PIGF	0.1818898	90	Ubiquitin-Mediated Proteolysis	•Control status (r = 0.4, P=0.03) •Left Caudate (r = 0.37, P=0.04) •Left Ventral DC (r = 0.48, P=0.0007)	•Control status (r = 0.4, P=0.03) •Left Caudate (r = 0.37, P=0.04) •Left Ventral DC (r = 0.48, P=0.0007)
C7orf11	0.0028377	90			
PHLDB2	0.1302545	95			

Abbreviations and symbols same as in **Table 7.3.**

7.3.5. CFG Prioritization of Biomarkers

Biomarkers were prioritized using a Bayesian-like CFG approach (**Supplementary Figure 5.1**) integrating previously published human evidence based on genetics (e.g. GWAS, copy number variants), post-mortem brain gene expression and peripheral blood gene expression specific to psychosis, SCZ, depression/stress as well as neurocognitive impairment (August 2015). This is a way of validating relevant blood transcriptome biomarkers from moderately sized datasets, extracting generalizable signal out of potential cohort-specific noise. Using the CFG approach, we first focused attention on the 'ubiquitin-mediated proteolysis' annotated module, which in this study represents a functional biomarker of MAP. This module was enriched with 61 genes having CFG evidence (*P*=4.8E-10), including those found to be dysregulated in the blood of patients with a psychotic disorder (∩=29) as well as in the blood and/or post-mortem brain of SCZ patients (∩=32) across independent human studies (**Supplementary Table 7.2**). Notably, of the 29 CFG genes found in the blood of a psychotic disorder, 21 pertained to one single study (Lee et al., 2012). We found a significant enrichment of 39 genes holding CFG evidence (*P*=7.0E-12) within the module annotated as 'circadian clock' (**Supplementary Table 7.3)**. Similarly, these genes were also previously associated with psychosis and/or SCZ in independent studies. Two genes within the 'ubiquitin-mediated proteolysis' annotated module (*TMEM106B* and *SCAMP1*) and one within the 'circadian clock' annotated module (*DCTN1*) overlap with a previous study which had used CFG based approach to validate blood biomarkers for delusions, a core symptom of psychotic disorders (Kurian et al., 2009). An additional gene (*RAB18*) within the 'ubiquitin-mediated proteolysis' module was also validated as a SCZ biomarker using the CFG approach (Ayalew et al., 2012).

Applying the CFG approach to our panel of 31 discriminative biomarkers confirmed 8 candidate biomarkers for MAP (**Table 7.3 & 7.4**) which had a CFG score of 3 or above, meaning either maximal score from the *p*-value threshold cut-offs or at least two other lines of prior independent evidence (**Figure 7.7A**). Indeed, CFG evidence for 8 out of 31 discriminatory biomarkers is a significant overlap (*P*=0.01), beyond what would be expected by chance. Of these validated MAP biomarkers, four were previously reported to predict psychosis in an independent human blood transcriptome investigation (*FBP1*, *ZNF821*, *TBC1D2* and *SIN3A*), one of which was previously labelled a genetic variant for SCZ risk (*FBP1*). In addition, one other biomarker had been implicated in SCZ risk across two independent studies (*UHMK1*). Subsequently, a gene-disease network was built using all CFG validated biomarkers, either in the form of a functional biomarker (gene modules) or single biomarkers, to visualize unique gene signatures of MAP and consensus signatures of MAP, psychosis and SCZ **(Figure 7.7B)**. In this study, we found that MAP shares 69 genes with SCZ, 39 genes with other psychotic disorders and six genes are shared across all three conditions. Importantly, cross-referencing all candidate MAP genes onto possible haloperidol gene expression signatures from the CMap and CDT provided preliminary evidence for the lack of neuroleptic-associations across our candidate findings (**Figure 7.7B**).

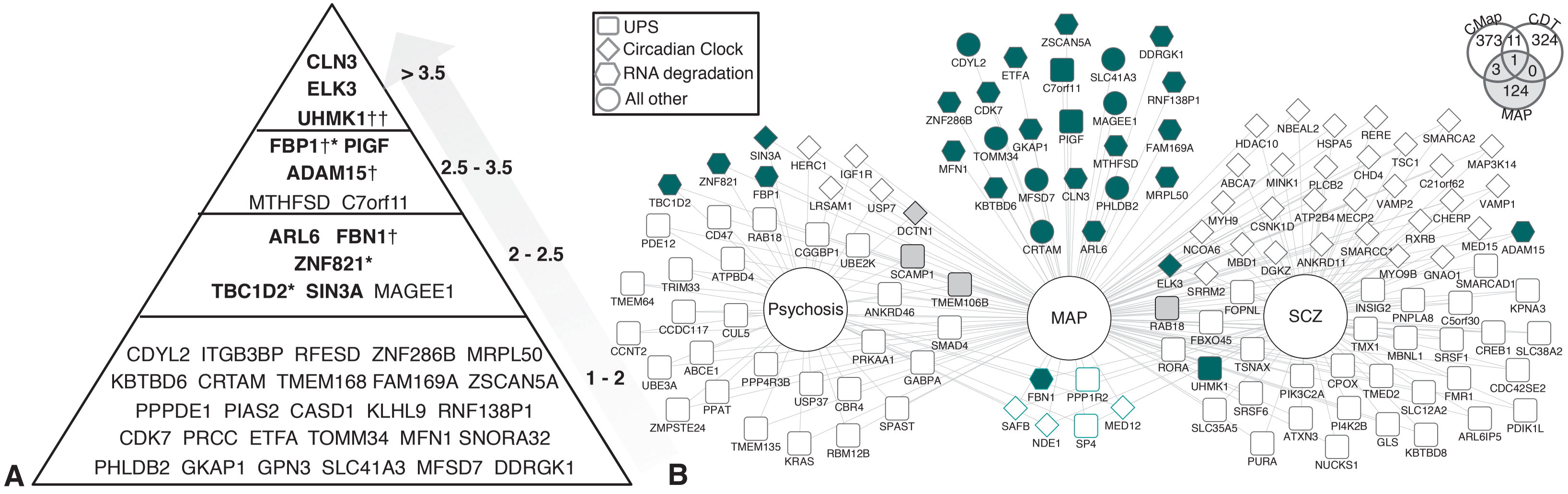


Figure 7.7. Top candidate blood biomarkers for MAP. (**A**) CFG evidence and scoring are depicted on the right side of the pyramid. Genes in bold have been found in external publications. Genes found in METH-free studies investigating SCZ (†) and psychosis (*) are as indicated. (**B**) Overlapping gene-disease relationships including CFG validated genes within gene modules (ubiquitin-mediated proteolysis & circadian rhythm) and single gene biomarkers. Nodes represent genes and edges indicate gene-disease relationships. Node shape denotes empirically derived functions from our network analysis. Green shading indicates biomarkers from our machine-learning analysis including 14 unique genes separating controls from METH dependents. Grey nodes represent CFG validated biomarkers of delusion (psychosis) or SCZ. Node border colour in turquoise indicates gene signatures across MAP, general psychosis and SCZ studies. Venn diagram depicts lack of overlap from curated haloperidol-gene signatures onto the 128 candidate MAP genes (61 UPS + 39 clock + 25 + 20 =128 genes (while accounting for overlap across lists)).

7.4. Discussion

This preliminary report describes gene networks and blood biomarkers of MAP, further validating the MAP model as an exemplar for discovery of biomarkers related to SCZ susceptibility and clinical course. In essence, this pharmacogenomics approach is a tool for identifying genes that contain pathophysiological relevance to psychotic disorders and SCZ. Considering the variable environmental component of MAP, it is possible that not all subjects would show changes in all biomarker genes. Hence, our approach incorporated blood gene expression, clinical assessment, psychometric measures and structural MRI data revealing several mechanistic insights regarding the pathophysiology of MAP and its overlapping mechanistic nature with psychotic disorders and SCZ. First, we identified a functional biomarker of MAP in the form of a co-expression module annotated to ubiquitin-mediated proteolysis, further enriched with 61 genes containing CFG evidence. We also revealed a psychoticism associated module implicated in the circadian clock, enriched with 39 genes containing CFG evidence. Second, we identified 25 genes that were able to distinguish healthy controls from METH dependents with high accuracy, while only 20 genes (recycling 14 genes from the previous split) were able to differentiate between MA and MAP subjects. A significant proportion of these single blood biomarkers also contained CFG evidence. Further, cross-referencing these results onto haloperidol-specific gene expression signatures reduced the likelihood of these genes being neuroleptic-related. These high overlaps suggest similar biological mechanisms detectable in peripheral blood underlying the pathophysiology of psychosis, regardless of substance abuse. These findings also suggest new avenues for exploring the utility of the MAP model in SCZ research.

7.4.1. Ubiquitin Proteasome System (UPS) Dysregulation

A central finding from the network analysis was the identification of a functional biomarker (gene module) annotated to ubiquitin-mediated proteolysis expressed to a lesser extent in MAP subjects (**Figure 7.2**). The ubiquitin proteasome system (UPS) is a highly complex and tightly regulated process that plays major roles in a variety of basic cellular processes, specifically degradation of intracellular proteins and modulation of cellular responses to inflammation and oxidative stress (Ciechanover et al., 2000). The UPS has been identified in genetic reports as a major pathway associated with psychosis (Bousman et al., 2010a; Lee et al., 2012), SCZ and bipolar disorder (Bousman et al., 2010b; Vawter et al., 2001; Vawter et al., 2002; Middleton et al., 2002; Altar et al., 2005; Konradi et al., 2004), as well as with neurodegenerative conditions including Alzheimer's disease (Lam et al., 2000) and Parkinson's disease (Shimura et al., 2000). Studies using post-mortem brain gene expression to investigate mechanisms of psychosis and SCZ provide consistent evidence for the down-regulation of UPS-related genes in these conditions (Vawter et al., 2002; Middleton et al., 2002; Altar et al., 2005). It was also recently shown that UPS abnormalities disrupt expression at the protein-level in SCZ (Ikeda et al., 2013). Studies using peripheral blood gene expression have also found that the UPS pathway was consistently dysregulated across bipolar, SCZ and psychosis patient groups (Bousman et al., 2010a). A later study used a targeted approach associating blood expression measurements of UPS pathway gene members with Scales for Assessment of Positive and Negative Symptoms (SAPS-SANS) and determined *UBE2K* (also a gene member of our 'ubiquitin-mediated proteolysis' module), was 1 of 3 genes most significantly associated with positive symptoms of psychosis (Bousman et al., 2010b). Another independent report built a diagnostic blood-based classifier able to distinguish first-episode psychosis from controls with 400 genes (Lee et al., 2012), 21 of which were found within our UPS annotated module **(Supplementary Table 7.2).** It is interesting that genes with a well-established role in brain functioning also show changes in peripheral blood in relationship to psychiatric symptom states, and moreover that the direction of change should be concordant with that reported in human post-mortem brain studies. As a consequence of the overlapping nature of UPS dysfunction found across mental diseases, the proteasome system has emerged as a putative candidate highlighting both mRNA and protein-level changes in psychosis and SCZ.

7.4.2. UPS and Circadian Clock Associations to sMRI Data

In determining relationships between blood gene expression and structural MRI data, we revealed a significant association of the ubiquitin-mediated proteolysis module with the anterior corpus callosum (CC) (*r*=-0.55, *P*=0.002) (**Figure 7.4**). Conversely, the circadian clock module, expressed to a greater extent in MAP subjects (**Figure 7.2**), was significantly associated with the EPQRS measure of 'psychoticism' (*i.e.* aggression, egocentrism and impulsiveness) (*r*=0.43, *P*=0.02) and the posterior CC (*r*=0.39, *P*=0.03) (**Figure 7.4**). There is considerable evidence suggesting that global white matter abnormalities (*i.e.* disruptions in connectivity in intra- and interhemispheric pathways) play a role in the pathophysiology of psychiatric disorders (White et al., 2008). The CC is the largest white matter tract containing highly packed neuronal fibres, and abnormalities in this structure have frequently been reported in patients with SCZ (Whitford et al., 2010), including first-episode SCZ and psychosis patients (Price et al., 2007), often relating to the severity of psychotic symptoms. It has been hypothesised that less efficient connectivity and resulting aberrant signal transmission between brain regions may be a pivotal factor in the manifestation of psychotic symptoms, including delusions and hallucinations, and of cognitive dysfunctions (Friston & Frith, 1995; Kubicki et al., 2007). However, these disturbances have not been fully elucidated in the context of MAP nor in its relationship to blood gene expression differences. But we also observed significantly lower bilateral hippocampal volumes in MAP subjects (**Table 7.2**). While correlates of blood gene expression to hippocampal volumes relate mainly to processes of protein ubiquitination (*r*=0.37, *P*=0.05), reductions in hippocampal volumes are consistent with previous reports of pathological hippocampus changes in MAP (Orikabe et al., 2011), in first episode and chronic schizophrenia (Velakoulis et al., 2006), and in individuals at high risk for psychosis (Fusar-Poli et al., 2009). Taken together, these findings suggest that changes in the blood occur in parallel to structural changes in the brain of MAP subjects and that they are also most likely involved in the pathophysiology of psychotic disorders and SCZ in the absence of METH.

7.4.3. Putative MAP-related Gene Hunting Tools

Interrogation of the comparative toxicogenomics database (CTD) (Davis et al., 2015) with a signature query composed of the genes in our 'ubiquitin-mediated proteolysis' annotated module revealed an enrichment of sodium arsenate gene signatures (**Supplementary Table 7.1**). Sodium arsenate is one of the most toxic metals derived from the natural environment, but has been used as a therapeutic medication in acute promylocytic leukaemia based on its mechanism to induce apoptotic effects via release of apoptosis-inducing factor (*AIF*) (Schenk & Stolk, 1967). However, arsenic is mainly a contaminant and interestingly is known to cause clinical features such as psychosis, toxic cardiomyopathy, and seizures (Lebrun et al., 2010). This exploratory result suggests arsenic, and chemically similar compounds, may be a useful gene-hunting tool for investigating future mechanisms of psychosis in either primary or patient-derived lymphoblast cell lines to elucidate further these effects in search for more verifiable biomarkers.

7.4.4. Candidate Blood-based Diagnostic Biomarkers of MAP

Topping our list of candidate MAP biomarkers we found 8 genes involved in RNA degradation (*CLN3*, *FBP1*, *TBC1D2*, *ZNF821*, *ADAM15*, *ARL6*, *FBN1* and *MTHFSD*), two which are specific to circadian rhythm (*ELK3* and *SINA3*) and three involved in ubiquitin-mediated proteolysis (*PIGF*,*UHMK1* and *C7orf11*) (**Table 7.3 & 7.4**). Some of the gene expression changes detected in this moderately sized cohort (N=30) may represent biological or technical artefacts, butto minimize such effects, our candidate MAP biomarkers were selected based on having a line of evidence (CFG) score of two or higher (**Figure 7.7A**). Proper cross-validation both *in silico* and across-literature (CFG), minimised the likelihood of having identified false positives while increasing sensitivity and specificity in the ability to distinguish true signal (biomarkers) from noise through a fit-to-disease Bayesian-like methodology (Niculescu et al., 2000; Ogden et al., 2004; Patel et al., 2010; Le-Niculescu et al., 2009; Rodd et al., 2007; Le-Niculescu et al., 2011; Ayalew et al., 2012; Le-Niculescu et al., 2013; Kurian et al., 2009).

CLN3 (Ceroid-Lipofuscinosis, Neuronal 3) was the top scoring gene in our study and is involved in lysosome function. Mutations in this gene are well-known to cause neurodegenerative diseases such as Batten disease (Lebrun et al., 2010; Mitchison et al., 1994), which impairs mental and motor development during childhood, causing difficulty with walking, speaking and intellectual functioning. Patients with a CLN3 mutation are also prone to recurrent seizures, epilepsies, visual impairment, and occasionally psychosis. It is hypothesised that mutations in CLN3 disrupt lysosome function resulting in build-up of lipopigments, which may induce apoptotic effects in brain neurons. This gene has not yet been discussed in the context of psychosis, but may represent a putative biomarker of MAP. Additionally, variants in the gene *FBP1* (fructose-1,6-bisphosphatase 1) have previously provided genetic support for the view that alterations in glucose metabolism are intrinsic to SCZ pathology (Olsen et al., 2008). However, in our study this gene was found co-expressed in the 'RNA degradation' module. Other top scoring genes included genes annotated to a circadian clock module (**Supplementary Table 7.3**), which are involved in sleep-wake cycles and previously identified as risk factors for psychosis (Niculescu et al., 2000), anxiety disorders (Le-Niculescu et al., 2011), suicidality (Le-Niculescu et al., 2013) and mood disorders (Le-Niculescu et al., 2007). *ELK3* (ETS-Domain Protein (SRF Accessory Protein 2)) encodes a transcriptional factor that may switch from activator to repressor in the presence of Ras whereas *SIN3A* (SIN3 Transcription Regulator Family Member A) encodes a transcriptional repressor with known roles in circadian clock negative feedback (Duong et al., 2011). While *SIN3A* has well known association to circadian clock function, an advantage of our approach was to be able to derive guilt-by-association co-expression interpretation of biomarkers, such as *ELK3*, by indicating module membership status. Dysregulation of circadian clock genes in post-mortem brain of SCZ patients have previously been observed (Monti et al., 2013).

7.4.5. Candidate MA and METH Dependency Genes

The MA associated findings also allow us to speculate on molecular mechanisms of psychosis. MA discoveries mainly included elevated expression in modules specific to interferon and cytokine signalling. While cytokine signalling was positively associated with METH-dependency (i.e. MA and MAP subjects) (*r*=0.39, *P*=0.03), a module specific to 'interferon signalling' was significantly over-expressed in the blood of MA subjects relative to controls, rather than MAP subjects relative to controls (**Figure 7.3**). Previous work has highlighted a weak or absent immune stress response, specific to HPA axis activation (van Venrooji et al., 2012) and cortisol measurements (Mondelli et al., 2015), in medication-naive first-onset psychosis patients. Our results are consistent with these findings suggesting a potentially diminished defence and innate immune response involved in the pathophysiology of psychosis, which may not necessarily be due to environmental variables or METH abuse. In addition to the cytokine signalling module, modules specific to IL-5 signalling, actin cytoskeleton and ATPase activity all showed a strong association to both the left and right accumbens area (**Figure 7.4**). Due to high levels of dopaminergic innervations, the nucleus accumbens, together with other subcortical structures, plays a pivotal role in several neurocircuits involved in reward, motivation, drug-reinforcement and drug seeking behaviour, mood regulation, and sleep wake cycles (Le Moal & Koob, 2007; Qiu et al., 2012). Such neurocircuit functions are similarly affected by drug exposure as well as by stressors, life events, or social pressure, with increased dopamine release in the nucleus accumbens triggered by the stimulant in addiction and by glucocorticoid hormones in stress (Le Moal & Koob, 2007). Furthermore, there is emerging evidence that cytokines circulating in blood may target subcortical dopamine function, with potential implications on behaviour, sleep patterns, and the progression of psychiatric disorders, such as depression (Felger et al., 2012).

7.4.6. Strengths and Limitations

While it appears that the identification of blood-based biomarkers may be accomplished by systems-level and machine-learning approaches, it remains uncertain which approach provides the most favourable translational avenue. A strength of this study is the identification of gene networks which were subjected to multi-modal data integration, and later enriched for CFG evidence. Additionally, a significant proportion of MAP single gene biomarkers identified by machine-learning were also validated by CFG evidence. Another strength of the computational approach is the ability to place single-gene biomarkers into a coherent biological framework in order to derive mechanistic insights. Limitations of our study include its small sample size, cross-sectional (lack of longitudinal) experimental design and the lack of a disease control group. For example, to emphasize the MAP model as an exemplar for SCZ biomarker discovery it would be useful to incorporate a SCZ diagnosed group into the experimental design.

7.4.7. Concluding Remarks

Overall, our results support the MAP model for identification of biomarkers involved in psychosis and SCZ. Our findings suggest that genes involved in UPS and circadian clock dysregulation are potential players in psychosis and are reflected in both peripheral blood and post-mortem brain profiles. Specifically, UPS abnormalities have emerged as a common denominator across a variety of independent studies investigating psychosis, SCZ and bipolar disorder. Our results shed light on biological mechanisms of psychosis, regardless of polysubstance abuse, medication or other confounding factors and further emphasize the value of moving towards comprehensive empirical profiling. These results also suggest empirical avenues for future field trials, clinical testing and validation in various at-risk populations.

Contributions. *I was solely responsible for all statistical design, data analysis, data interpretation and writing.*

Chapter 8

Candidate Lithium Responsive Genes and Gene Networks in Bipolar Disorder Lymphoblastoid Cell Lines

8.1. Background

Bipolar disorder (BD) is a common psychiatric illness affecting 1-4% of the population worldwide (Merikangas et al., 2011). Its pathophysiology is still largely unknown. BD is characterized by recurrence of depressive, hypomanic, or manic episodes with intervening intervals of partial or full remission (Garnham et al., 2007). Lithium (Li) treatment is the mainstay medication in the long-term treatment for BD and is one of only two medications known to reduce risk of suicide (Garnham et al., 2007; Baldessarini & Tondo, 2000; Lewitzka et al., 2015). However, clinical response to Li is variable and the mechanisms by which this drug stabilizes mood are multifaceted.

Previous research has identified predictive factors such as clinical presentation and family history (Kleindienst et al; 2005), DNA polymorphisms (Serretti et al., 1999; Turecki et al., 1998; Rybakowski et al., 2005; Serretti et al., 2001; Masui et al., 2006), or common genetic variants through genome-wide association studies (Perlis et al., 2009) to identify subsets of BD patients who might respond more or less favorably to Li treatment. However, to date no single factor has been fully reproduced nor able to accurately predict treatment response. Much effort has been focused upon obtaining more information on underlying neurobiological processes associated with Li response, as well as the mechanisms that influence gene expression and molecular pathways. Evidence from both *in vitro* and *in vivo* studies demonstrates that Li exerts multiple effects on ion transport, signal transduction cascades, neurotransmitter/receptor mediated signaling, hormonal and circadian regulation and greatly alters gene expression patterns (Lenox et al., 2003). Actions on the phosphoinositide pathway and on glycogen synthase kinase-3 represent two of the many Li-responsive biological processes (Klein & Melton, 1996; Berridge et al., 1989). In addition to these mechanisms, transcriptomic reports have shown Li up-regulates anti-apoptotic genes and down-regulates pro-apoptotic genes in Li responders relative to Li non-responders (Lowthert et al., 2012; Beech et al.., 2013). These results suggest differential changes in the balance of pro- and anti-apoptotic gene expression may partly explain the heterogeneity in clinical response to treatment. Overall, it is clear that the processes that underpin the therapeutic actions of Li are complex and most likely inter-related. Accordingly, it is hypothesized that a gene network approach should be well-suited to identify subsets of genes underlying therapeutic and nontherapeutic actions of Li (Lenox et al., 2003) and model changes in the expressions of genes which differ based on treatment outcome occurring after treatment initiation.

As BD is thought to be a brain disorder, access to live human brain tissue is unlikely, so a model system is necessary. In this study we chose to study human cells, rather than using animal models for BD and treatment response, which have some limitations (Overstreet, 2012; Harro, 2013). In our case, we are investigating the response to Li in BD in a peripheral cell line, namely lymphoblast cell lines (LCLs), because unlike PBMCs, they more closely mimic neuronal cells in their gene expressions and regulation (Abe et al., 1991; Gutekunst et al., 1995; Kobayashi et al., 2003; Koide et al., 1999). Moreover, neuronal cells from live patients are also difficult to acquire, while LCLs are readily available and also produce a resource that can be used in follow-on studies. While transformation of peripheral B lymphocytes by EBV is the method of choice for generating LCLs, it is expected that there will be changes in gene expression caused by the virus and culturing. However, comparing the biological effects of Li on paired samples (i.e. untreated vs. Li-treated LCLs) should eliminate most biases. As such, we, and others (McEachin et al., 2010; Hunsberger et al., 2015), believe that LCLs represent the most appropriate model for this work.

The current investigation aimed to study Li responsive genes and gene networks using a genome-wide transcriptomic approach, rather than examining changes in expression levels of presumed relevant targets. We compared human cell lines from BD patients classified as responders or non-responders to Li treatment, together with healthy control donors. Using RNA-Sequencing, we profiled the expression of mRNA of human lymphoblastoid cell lines (LCLs) exposed to a therapeutic course of Li. In this study, three main aims were established. **Aim 1:** First, to identify gene networks which were either induced or repressed with Li treatment, which may reveal new targets for the molecular mechanisms underlying Li action. **Aim 2**: Second, to better understand the relationship of gene expression signatures of Li identified in our study with other gene expression perturbations caused by similar small bioactive drugs/compounds. **Aim 3:** Finally, to identify individual genes that differ in their response to Li between BD patients that demonstrate relapse and non-relapse symptoms, providing a mechanistic basis for therapeutic heterogeneity.

8.2. Materials and Methods

8.2.1. Subject Selection

A total sample of 125 BD patients participated in a prospective relapse prevention trial of Li. From here, a subset of age-matched 8 BD Li responders (22.62 ± 8.01 yrs.), 8 BD Li non-responders (22.87 ± 6.57 yrs.) and 7 healthy controls (22.53 ± 7.23 yrs.) were enrolled in this study. All participants were Caucasian males and recruited from the San Diego Veterans Affairs Medical Center. Participants diagnosed by a psychiatrist or clinical psychologist with a DSM-III or DSM-IV diagnosis of BD were included in this study. After screening for eligibility and initial assessment, patients entered a multi-phase clinical design to determine BD Li responders from BD Li non-responders (*for detailed experimental design see* **Supplementary Figure 8.1**). First, patients were started on Li and entered the first phase (that is, stabilization phase). The goal in this phase was to stabilize patients within three months on Li monotherapy. Following, patients entered the second phase (that is, observation phase) and were observed for one month to assure stabilization after discontinuation of other medications. Finally, patients then entered the final phase (that is, maintenance phase) and were followed at two to four month intervals for two years. Positive Li response (i.e. non-relapse) was defined as reaching the end of the maintenance phase of the study without relapse of BD symptoms and Li re-administration. In this context, patients classified as BD Li responders denote non-relapse patients and patients classified as BD Li non-responders denote relapse patients. Average number of weeks in the study for BD patients characterized as BD responders and non-responders was 69 and 11 weeks, respectively.

8.2.2. Cell Culture

Blood samples from all patients were obtained at the beginning of the trial, before the initiation of Li treatment, and subsequently transformed by infecting buffy coat with Epstein-Barr virus (EBV). A total of 23 lymphoblastoid cell lines (LCLs) of low passage were revived and cultured in RPMI 1640 supplemented with 10% fetal bovine serum and 1X antibiotics to reach density of 1M cells/mL and split in half. Subsequently, each half was then placed in the continuous presence of 1mM Li chloride in the vehicle (i.e. Li treated) or vehicle alone (i.e. untreated) for seven days, which is considered to closely mimic chronic exposure and treatment concentrations of Li in patients' brains (Asai et al., 2013; Cruceanu et al., 2012; Sugawara et al., 2010).

8.2.3. RNA Isolation and RNA-Seq Library Preparation

Total RNA was extracted from 2-5x10^6 cells of each treated and untreated (vehicle-treated) cell suspensions using QIAamp RNA blood mini kit (Qiagen, CA, USA) accordingly to the manufacturer's protocol. Quality and quantity of obtained RNA was assessed by NanoDrop and Agilent 2100 Bioanalyzer using RNA 6000 Labchip (RNA Integrity Numbers: 9.7 ± 0.3). Subsequently, to isolate the mRNA transcriptome, the Dynabeads® mRNA DIRECT™ Micro Purification Kit (Life Technologies, USA) was modified for lower total RNA input. This step utilizes RNase inhibitors in the lysis/binding buffer, combined with stringent hybridization and washing steps to isolate intact mRNA and to deplete ribosomal and small RNA molecules. Only polyadenylated RNA species are captured, resulting in cleaner preparations and more sensitive results.

Libraries were prepared using the Ion Torrent RNA Seq Kit V2 with ERCC control RNA according to the manufacturer's user guide (publication 447286 Rev D). Each subject's Li and vehicle-treated paired samples were barcoded (Ion Xpress RNA-Seq barcode 01-16 Kit) and run together on 318 chips. Data was available as aligned BAM and raw fastq files (average 5.6 ± 0.3 million reads/chip). For data availability (in raw fastq format) contact corresponding authors.

8.2.4. Read Trimming, Mapping and Quantification of Expression

Raw reads from the Thermo Fisher Scientific's Ion PGM sequencer were filtered using the Fastx-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html) developed by the Hannon lab (Pearson et al., 1997). All reads less than 12 base pairs in length were removed due to their propensity to map to multiple locations by chance. Due to read quality degradation near the end of the reads, they were trimmed to 100 base pairs and filtered for artifact reads composing only 3 total base pairs. The reads were mapped against the hg19 genome using TMAP version 3.0.1, a Smith-Waterman alignment optimization (Li et al., 2010). TMAP was specifically designed for the Ion Torrent data and has shown more robust results. TMAP was run in mapall mode with the map1, map2, and map 3 functions and their default settings. The map1 function is well suited for short reads whereas the map2 and map3 settings are more well designed for longer reads. These settings allowed for the mapping of reads ranging from 12 to 100 base pairs with varying levels of mismatches based upon the size of the read. Reads were then counted against the hg19 genome using HTSeq version 0.6.1p2 with the default settings (Anders et al., 2014).

8.2.5. Data Pre-Processing

Raw count data measured 23,349 transcripts across all patients before and after Li treatment (i.e. 8 BD Li responders, 8 BD Li non-responders and 7 healthy controls). Non-specific filtering requiring more than 10 counts per million in at least 5 subjects retained 9,128 transcripts subjected to VOOM normalization in EdgeR (Law et al., 2013), a variance-stabilization transformation method. Normalized data were inspected for outlying samples using unsupervised hierarchical clustering of subjects (based on Pearson coefficient and average distance metric) and principal component analysis to identify potential outliers outside two standard deviations from these averages. Three outliers were present in these data (1 control subject prior to Li treatment, 1 control subject following Li treatment, 1 BD Li non-relapse subject following Li treatment) and resulting normalized values were used as input for WGCNA, differential gene expression and drug-gene activity analyses.

8.2.6. Weighted Gene Co-expression Network Analyses

Signed co-expression networks were built using weighted gene co-expression network analysis (WGCNA) (Langfelder & Horvath, 2008) in R. A total of 9,128 transcripts were used to construct a global network of all 43 subjects. To construct a global weighted gene co-expression network, the absolute values of Pearson correlation coefficients were calculated for all possible gene pairs and resulting values were transformed using a ß power of 13 so that the final correlation matrix followed an approximate scale-free topology. Subsequently, to identify sub-networks from the global network (i.e. co-expression modules), the WGCNA cut-tree hybrid algorithm was used optimizing minimum module size to 15, deep split of 3, and a tree-cut height of 0.35 in order to merge neighboring network modules with similar expression profiles. Once individual modules were identified, we sought to perform module differential expression to determine which modules were most affected by Li treatment. We ran singular value decomposition of each module's expression matrix and used the resulting module eigengene (ME), equivalent to the first principal component, to represent the overall expression profiles for each module. As previously described, differential expression of MEs was performed using a Bayes ANOVA (Kayla & Baldi et al., 2012) (parameters: conf=12, bayes=1, winSize=5), comparing between healthy controls, non-responders and responders, correcting *P* values for multiple comparisons with post-hoc Tukey tests. MEs are also used to determine module membership (kME) values for each gene in a specified module, defined as the correlation between gene expression values and ME expression. Genes with the highest intramodular kME are labeled hub genes and are predicted to be essential to the function of the module. Gene significance (*GS*) was calculated as the $-\log_{10}$ of the *P* value generated for each gene within a particular module using a moderated t test, and is a measure of the strength of differential expression between vehicle and Li treatment. Module significance (*MS)* was calculated as the average *GS* within each module, to identify modules enriched with differentially expressed genes.

8.2.7. Differential Gene Expression Analyses

Differentially expressed genes were assessed between groups using a moderated *t*-test in the *limma* package (Smyth et al., 2005). The multi-level experimental design permitted the testing of multiple hypothesizes. First, we sought to explore the effect of Li on gene expression, relative to vehicle treatment: longitudinal group-wise comparisons were made considering differences between vehicle and Li treatment by testing non-responders, responders and controls each independently. Subsequently, the effects of Li on gene expression was analyzed by pooling all subjects relative to treatments (i.e. comparing vehicle to Li). *P*-value significance was set to a FDR $P < 0.05$ due. Second, we sought to explore standing variation in gene expression profiles prior-to and following Li treatment, independently: cross-sectional group-wise comparisons were made at vehicle treatment and again separately at Li treatment. *P*-value significance was set to a nominal $P < 0.01$ to permit sufficient enough information to carry on with down-stream functional enrichment. Third, we sought to identify genes whose expression differed between vehicle and Li treatment between BD responders and non-responders: we performed a mixed linear contrast analysis with significance set to a nominal $P < 0.05$. Similarity, this assumption was relaxed to cast a wide net to permit for down-stream functional interpretation of the candidate gene-list.

8.2.8 Functional Annotation and Drug Gene-Set Testing

All identified network modules and differentially expressed genes were subjected to functional annotation. The ToppFunn module of ToppGene Suite software (Chen et al., 2009) (Division of Biomedical Informatics) was used to assess enrichment of GO ontology terms specific to biological processes and molecular factors using a one-tailed hyper geometric distribution with a Bonferroni correction. Because differentially expressed genes identified from our mixed contrast analysis produced a large amount of enriched GO terms, GO semantic similarity analysis was used using G-Sesame (Du et al., 2009) semantic similarity metrics and default semantic contribution factors (“is_a” relationship: 0.8 and “part_of” relationship: 0.6). This analysis results in a symmetric matrix in which each value represents a score for similarity between GO term pairs. Then, we undertook hierarchical clustering based on semantic similarity matrix to group together all GO terms with common GO ‘parent’. Finally, to identify drug activated gene expression signatures from cultured human cells which are most similar to those of Li identified in our study, we utilized the Drug Signatures Database (DSigDB) (Yoo et al., 2015), a resource linking gene expressions with > 20K drugs/compounds for translational research. The QuSage software (Yaari et al., 2013) (version 1.9.0) was used to perform drug gene-set testing to identify drugs/compounds eliciting similar up- and down-regulated gene expression profiles.

8.2.9. Protein Interaction Networks

When protein interaction networks are constructed from candidate gene-lists, they can reveal key genes and transcription factors that control the regulation of multiple target genes. Protein-protein interactions (PPI) were obtained from the STRING (Mering et al., 2003) database with a signature query of differentially expressed genes identified from our mixed contrast analysis with a combined STRING score higher than 0.4. For visualization, the STRING network was imported into CytoScape (Shannon et al., 2003).

8.3. Results

We conducted an exploratory RNA-Sequencing study profiling LCLs before and after Li exposure from a primary cohort of BD Li responders, BD Li non-responders, and healthy controls. To compliment this multi-level experimental design and to identify candidate Li responsive genes and gene networks, a multi-step analytical approach was used (**Figure 8.1**). An unsupervised global gene co-expression network was first constructed using all available subjects to identify groups of coordinately expressed genes (i.e. co-expression modules) involved in the overall molecular response to Li. This analysis identified 22 co-expression modules, which were functionally annotated to GO molecular factors and GO biological processes (**Supplementary Table 8.1**).

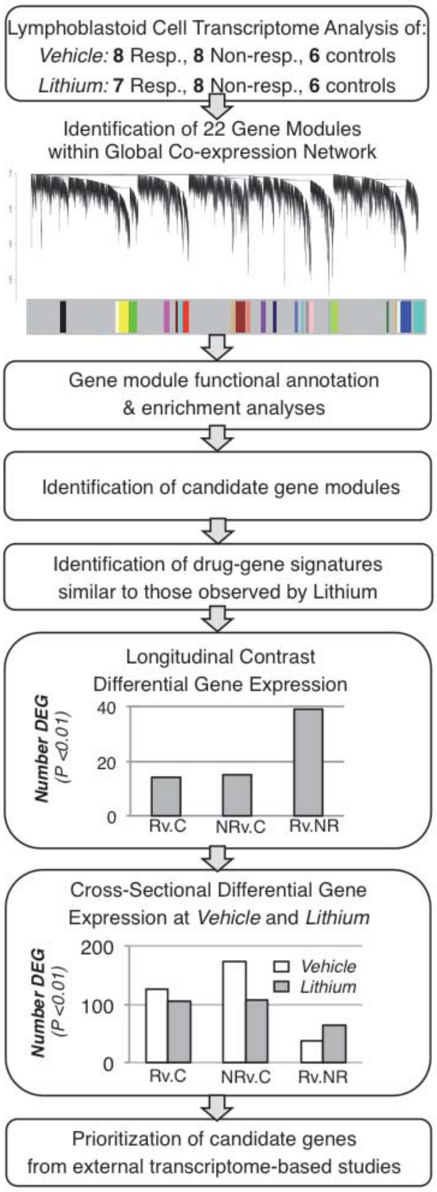

Lymphoblastoid Cell Transcriptome Analysis of:
Vehicle: **8** Resp., **8** Non-resp., **6** controls
Lithium: **7** Resp., **8** Non-resp., **6** controls
Identification of 22 Gene Modules within Global Co-expression Network
Gene module functional annotation & enrichment analyses
Identification of candidate gene modules
Identification of drug-gene signatures similar to those observed by Lithium
Longitudinal Contrast Differential Gene Expression
Number DEG (P <0.01)
40
20
0
Rv.C
NRv.C
Rv.NR
Cross-Sectional Differential Gene Expression at *Vehicle* and *Lithium*
Number DEG (P <0.01)
200
100
0
□ Vehicle
▨ Lithium
Rv.C
NRv.C
Rv.NR
Prioritization of candidate genes from external transcriptome-based studies

Figure 8.1. A multi-step analytical work-flow was used for identifying candidate Li responsive genes and gene networks. Quality control identified three outliers (1 control subject prior to Li treatment, 1 control subject following Li treatment, 1 BD Li non-relapse subject following Li treatment). First, WGCNA analysis built a global co-expression network and identified 22 co-expression modules. On the hierarchical cluster tree each line represents a gene (leaf) and each group of lines represents a discrete group of co-regulated genes, or gene modules (branch) on the clustering gene tree. Each gene module is indicated by the colour bar below the dendrogram, and subsequently functionally annotated using GO biological processes and molecular factors. Second, drug-gene signatures of Li were compared to those of other small molecule compounds to identify similar mechanisms of action. Third, a series differential gene expression analyses were used to identify single gene(s) involved in Li's therapeutic effects. Finally, candidate genes were prioritized using external lines of transcriptome-based evidence. For each step the corresponding figure and/or table is listed providing a quick reference. Abbreviations: NR, non-responders; R, responders; C, healthy controls; DEG, differentially expressed genes.

8.3.1. Differential Analysis of Module Eigengene Values and Drug-Gene Set Enrichment

Next, modules were examined for over-representation of genes identified as significantly differentially expressed between vehicle and Li treatment. These gene-based results were analogous across groups and exhibited a high degree of overlap, and as a result, modules were specifically enriched for differentially expressed genes identified from pooling all groups together at vehicle and then at Li treatment (**Figure 8.2**).

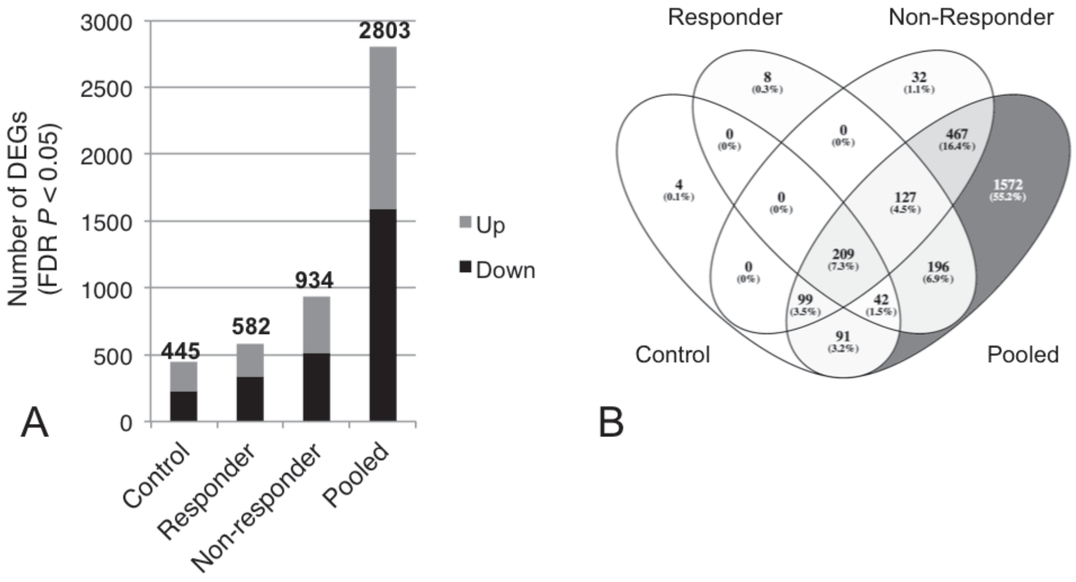


Figure 8.2. Longitudinal differential gene expression analysis. (**A**) Within group differential expression analysis between vehicle and Li treatment identified gene expression signatures from healthy controls, responders, non-responders, and when all groups were pooled at both treatments. (**B**) Overlap analysis determined that most of these genes were found in relation to Li treatment and did not exhibit group specificity.

We identified seven modules containing an over-representation of differentially expressed genes between vehicle and Li treatment, as reflected by an elevated module significance (MS) value (**Figure 8.3A**). To determine the extent and significance of fold-change, this enrichment was further quantified by assessing differential expression of

ME values using Bayes ANOVA testing and correcting *P* values for multiple comparisons (**Figure 8.3B&C**). ME expression was significantly up-regulated with Li treatment in modules annotated to immune response (M2), apoptosis signaling (M3), defense response to virus (M7) and response to ER stress (M12). This analysis identified ME expression significantly down-regulated with Li treatment in modules annotated to ER (M5), translation initiation (M22), and one module of unknown function (M11), as well as a module showing a distinct expression pattern implicated to phophatidylserine metabolism (M17).

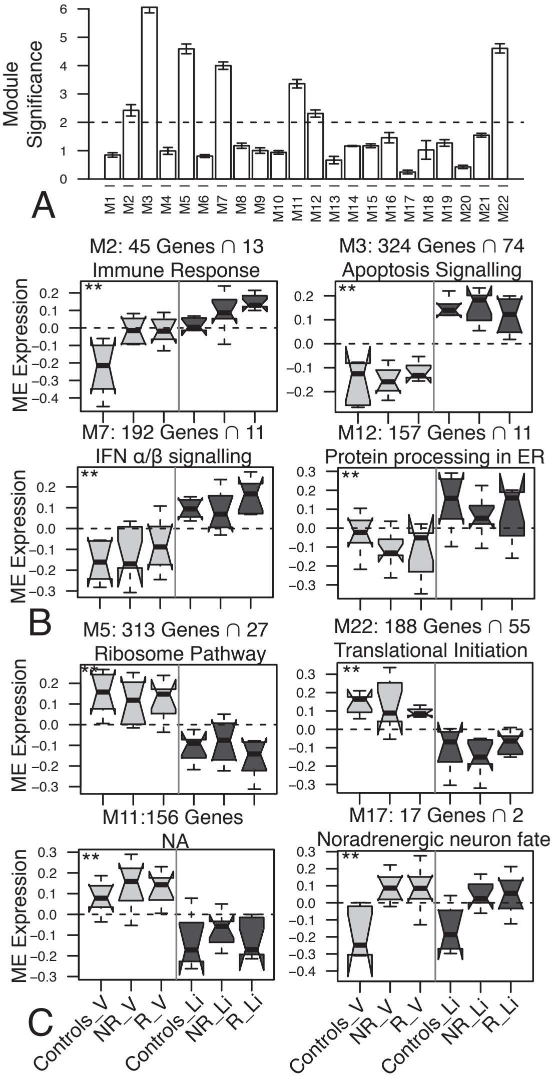


Figure 8.3. Module significance and module eigengene (ME) expression boxplots. (**A**) MS was measured across all 22 modules. The y-axis indicates MS by calculating the average $-\log_{10}$ P values generated by a moderate *t* statistic for each gene within a particular module, when assessing differential expression between vehicle and Li treatment. (**B**) ME values for modules induced with Li treatment (**C**) and those repressed with Li treatment. For each module the total number of genes including the overlap (∩) of module genes onto respective ontology terms and the putative ontology term (i.e function) are displayed above each boxplot. A Bayes ANOVA (parameters: conf=12, bayes=1, winSize=5) was used on ME values to test for significance between treatments and conditions, (**) indicates *P* < 0.05 Bonferroni multiple test corrected implying strong Li effects.

We next sought to map Li gene expression effects that we observed in our study onto gene expression perturbations elicited by other small molecule drug/compounds with physiochemical properties that can be effectively administered to patients. To do so, we interrogated DSigDB with the QuSage software using a signature query composed of Li gene expression changes (by pooling all groups at vehicle and then at Li). This test revealed significant associations of up- and down-regulated genes similar to those that were observed in our study with several different treatments including clonidine, isoprenaline and colchicine treatment (**Figure 8.4**). However, under closer inspection, genes both induced and repressed by clonidine treatment were most similar to those observed of Li treatment in this study (**Figure 8.5**).

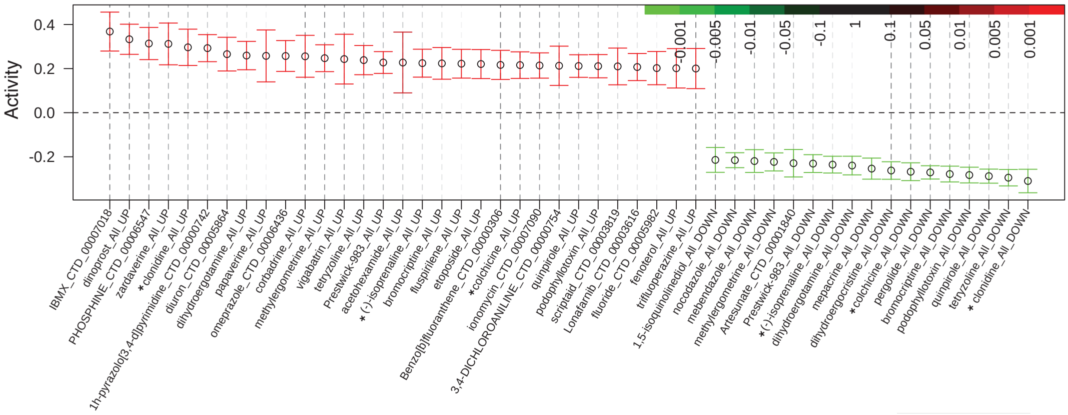


Figure 8.4. QuSage analysis of Li gene expression signatures comparable to those in DSigDB. Summary of drug gene set activity and corresponding mean and 95% confidence intervals plotted and colored-coded according to their False discovery rate (FDR)-corrected *P* values when compared to zero. Drug gene set activity passing a mean fold change of 0.2 are displayed. Asterisks indicate drug signatures overlapping with both up- and down-regulated Li signatures.

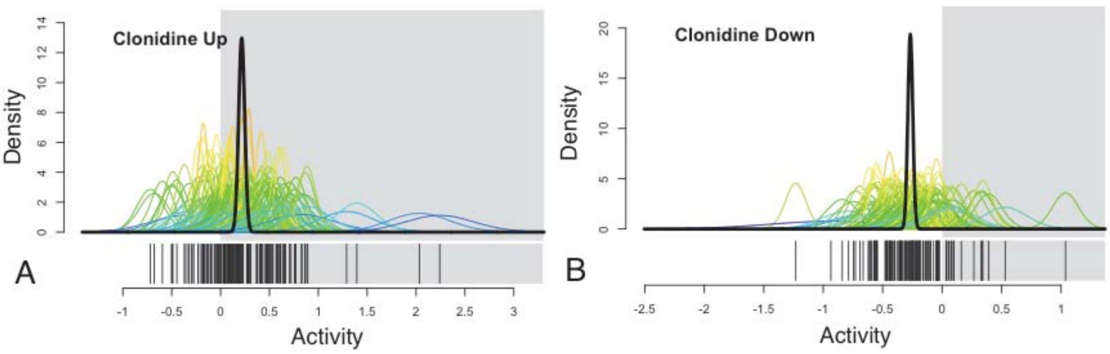


Figure 8.5. A clonidine treatment gene expression signature is compared to genes both (**A**) induced and (**B**) repressed Li treatment. Differential expression probability density functions (comparing vehicle to Li treatment) are shown for genes (thin curves color-coded by standard deviation), along with aggregated estimate the clonidine signature after taking into account gene-gene correlation (thick black curve). The mean differential expression for individual genes in the set are indicated as line barcodes below each panel.

8.3.2. Longitudinal Contrast Differential Gene Expression Analysis

Following, we re-focused our analyses on the individual gene level. A fundamental question to ask is whether any single gene(s) differ in their expression patterns between responders and non-responders over the course of Li (i.e. before and after Li treatment) and thus could be used as potential surrogate markers to explain heterogeneity in clinical response to treatment. A linear mix contrast approach was used to address this aim (*See Materials and Methods for more details)* (*P* < 0.01) and identified 28 genes down-regulated in BD non-responders compared to responders following Li treatment including genes *HSPE1*, *LYPLAL1*, *ORC3*, *GAR1*, *LSM5* and *PEX13*. This analysis also revealed 10 genes up-regulated in BD non-responders compared to responders including genes *ZNF48*, *ILVBL*, *GBA*, *TBC1D10A* and *SLC50A1*. Yet, to permit for sufficient enough information to move onto functional enrichment and protein interaction analyses, we relaxed our assumption of significance to *P* < 0.05. Subsequent functional annotation of these genes revealed processes associated with cell-cycle, nucleotide-excision (DNA) repair, protein deacylation, cellular response to stress, CoA thioesterase activity and cellular localization specific to the nucleoplasm (**Figure 8.6A**). We also analyzed whether these candidate genes that are dysregulated together also interact with each other at the protein level using the STRING database (**Figure 8.6B**). This analysis revealed differential regulation of hub genes *RANBP2*, *RBBP7, UTY*, *HDAC2*, *POLR3B*, *UMPS* and *ERCC2;* all of which have a putative role in mediating Li responsive effects differing between non-responders and responders. In total, 206 interactions were observed between 244 genes, more than expected by chance (*P*=4.6E-6).

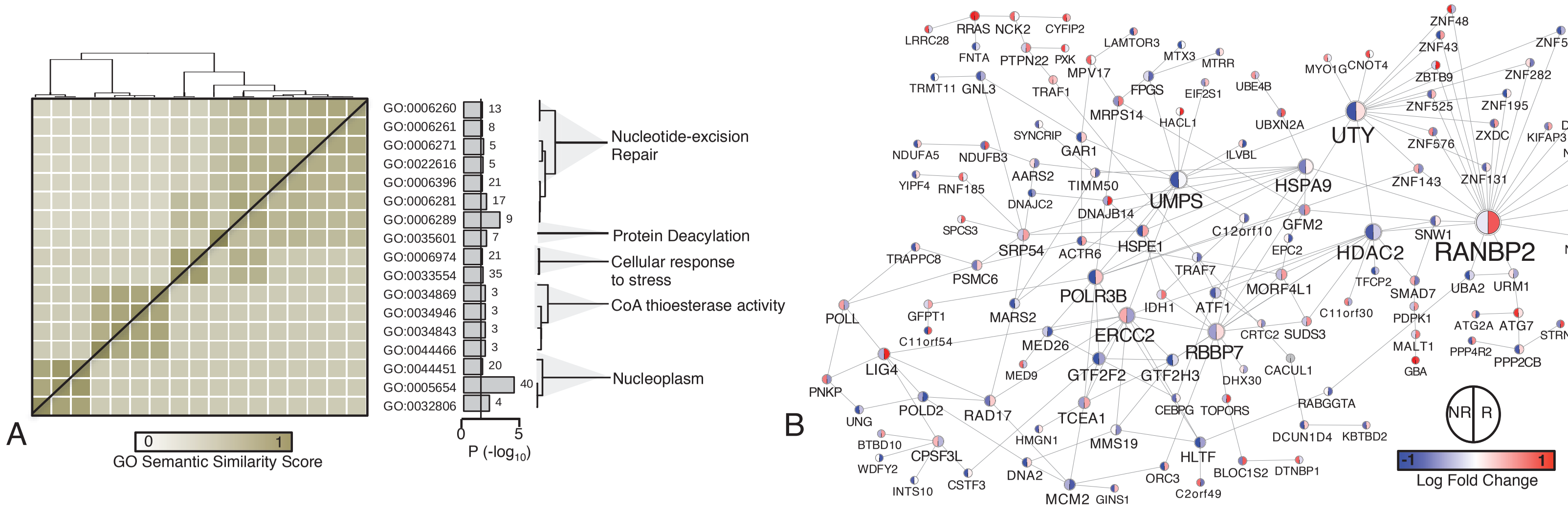


Figure 8.6. G-Seasame and STRING analyses based on differentially expressed genes differing in their response to Li between responders and non-responders. (**A**) Semantic similarity scores for all gene-ontology (GO) term pairs clustered by hierarchical clustering method (left) with adjusted *P* values for each GO term (middle), as well as common ancestors (right). (**B**) STRING protein interactions among candidate genes visualized within CytoScape software. Edges represent direct interactions and nodes represent genes. Each node is multi-colored; left portion indicates non-responder expression and the right portion indicates responder expression as compared between vehicle and Li for each particular gene. Relative expression is displayed as red (high) and blue (low).

8.3.3. Cross-Sectional Differential Gene Expression Analyses

Cross-sectional analyses at vehicle treatment revealed an overlap of 51 differentially expressed genes ($P < 0.01$) found in common between responders and controls and non-responders and controls (**Figure 8.7A**). This *baseline* BD signature was more specifically characterized by the down-regulation of MHC II protein complex genes including *HLA-DMA*, *HLA-DPB1* and *HLA-DRA* (**Figure 8.7B**). Interestingly, the same cross-sectional analyses following Li treatment revealed a consistent down-regulation of MHC II protein complex genes in BD relative to healthy controls (**Figure 8.7C&D**), further including genes *HLA-DMB*, *HLA-DOA* and *HLA-DRB1*.

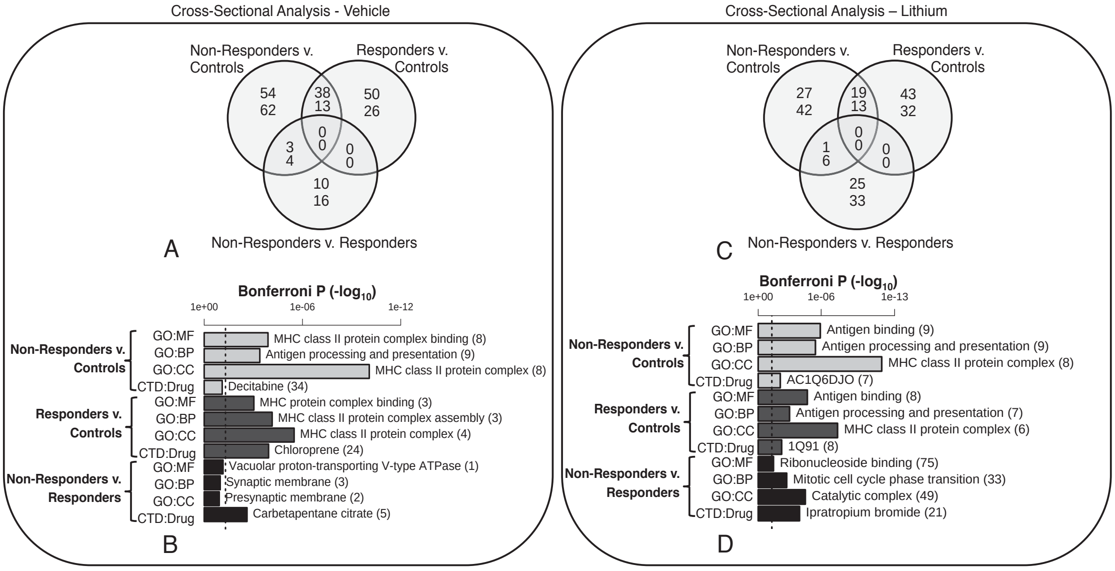


Figure 8.7. Cross-sectional differential gene expression analysis. Group-wise differential expression analysis identified unique gene expression signatures to non-responders, responders and controls among both (**A**) vehicle treated samples and (**C**) Li treated samples. Most significant GO biological process, molecular factor and cellular compartment are reported for each group-wise comparison among (**B**) vehicle treated samples and (**D**) Li treated samples. Bonferroni corrected *P* values for gene sets identified between non-responders compared to controls (light grey), responders compared to controls (dark grey), and non-responders compared to responders (black) are displayed with bar-plots. Gene-overlaps are broken down into over-expressed (top number) and under-expressed (bottom number) in each Venn-diagram.

8.3.4. Prioritization of Lithium Responsive Genes by Cross-Tabulating Independent Transcriptome-based Evidence

Longitudinal differential gene expression analyses (before and after Li treatment) revealed 2803 differentially regulated genes (FDR P < 0.05) (**Figure 8.2A**). Therefore, to contextualize these individual targets, genes modulated upon treatment with Li in our study were overlapped with Li responsive gene expression signatures in peripheral blood and LCLs from previous transcriptomic reports (Lowthert et al., 2012; Beech et al., 2013; McEachin et al., 2010; Hunsberger et al., 2015; Wantanabe et al., 2014). This curation of literature resulted in six studies and an overlap of 48 Li responsive genes (15 down-regulated, 33 up-regulated) were identified within two or more studies (**Supplementary Table 8.2**). Among the most down-regulated genes on this candidate gene list included genes *STC2*, *HADH*, *GAMT*, *MAT2A*, and *HSP90AA1*, while those most up-regulated included genes *CRIP1*, *CKB*, *FOS*, *LAX1*, and *RSAD2*. We also sought to characterize Li responsive gene expression signatures accordingly to Li responder and non-responder specificity by cross-referencing our results with an independent LCL study (Hunsberger et al., 2015).. This approach identified an overlap of 9 Li-responsive genes (7 down-regulated, 2 up-regulated) specific to BD Li responders including genes *FANCE*, *STOML1* and *SLC37A4* and 14 genes (11 down-regulated, 3 up-regulated) specific to BD Li non-responders including genes *DNAJC2*, *KLHL5*, and *NREP* (**Table 8.1**).

Table 8.1A. Nine lithium responder specific differentially expressed genes found across two studies.

	Non-Responders		***Responders***	
Gene Symbol (Gene name)	**Log Fold-Change**	**P-Value**	**Log Fold-Change**	**P-Value**
FANCE (Fanconi anemia, complementation group E)	-0.4235	0.2759	-1.4894	0.0181
STOML1 (stomatin (EPB72)-like 1)	-0.1259	0.6156	-0.8350	0.0434
SLC37A4 (solute carrier family 37 (glucose-6-phosphate transporter), member 4)	-0.2608	0.1168	-0.6267	0.0189
MED26 (mediator complex subunit 26)	-0.1478	0.2228	-0.5710	0.0035
EXOSC2 (exosome component 2)	-0.2124	0.1634	-0.5076	0.0415
TRAF7 (TNF receptor-associated factor 7, E3 ubiquitin protein ligase)	-0.0431	0.6949	-0.3848	0.0341
ANAPC5 (anaphase promoting complex subunit 5)	-0.1121	0.2679	-0.3343	0.0437
YWHAZ (tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, zeta)	0.1947	0.1097	0.3962	0.0448
BPNT1 (3'(2'), 5'-bisphosphate nucleotidase 1)	0.1918	0.3244	0.6470	0.0430

Table 8.1B. Fourteen lithium non-responder specific differentially expressed genes found across two studies.

	Non-Responders		***Responders***	
Gene Symbol (Gene name)	**Log Fold-Change**	**P-Value**	**Log Fold-Change**	**P-Value**
DNAJC2 (DnaJ (Hsp40) homolog, subfamily C, member 2)	-1.0530	0.0025	-0.2867	0.2426
KLHL5 (kelch-like family member 5)	-0.9451	0.0029	-0.2849	0.2032
NREP (neuronal regeneration related protein)	-0.7943	0.0271	-0.1647	0.5113
TSNAX (translin-associated factor X)	-0.7347	0.0086	-0.2228	0.2574
ARGLU1 (arginine and glutamate rich 1)	-0.6899	0.0412	-0.2367	0.3153
TAF1 (TAF1 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 250kDa)	-0.6455	0.0164	-0.2150	0.2571
UFL1 (UFM1-specific ligase 1)	-0.6226	0.0321	-0.1508	0.4558
ZNF544 (zinc finger protein 544)	-0.6124	0.0014	-0.2093	0.1221
PAK4 (p21 protein (Cdc42/Rac)-activated kinase 4)	-0.6076	0.0433	-0.1547	0.4597
IBTK (inhibitor of Bruton agammaglobulinemia tyrosine kinase)	-0.5984	0.0363	-0.2038	0.3077
LARS (leucyl-tRNA synthetase)	-0.4589	0.0208	-0.2489	0.0789
RAB7A (RAB7A, member RAS oncogene family)	0.3259	0.0075	0.1530	0.0781
RAB11FIP4 (RAB11 family interacting protein 4 (class II))	0.5201	0.0209	0.2319	0.1465
PIP4K2C (phosphatidylinositol-5-phosphate 4-kinase, type II, gamma)	0.5321	0.0336	0.2590	0.1414

Grey shading is for visualization purposes only.

8.4. Discussion

In this exploratory study we compared the effects of Li using LCLs derived from BD and healthy control donors to further elucidate the mechanism of action of lithium. We were able to identify several recurring themes regarding its mode of action and provide further insight into gene expression patterns differing between BD Li responders and non-responders. First, we identified several gene networks whose expression changed differentially upon treatment with Li, indicating widespread effects of Li on diverse cellular signaling systems including apoptotic signaling. Second, we identified genes differing in expression between responders and non-responders upon treatment of Li involved in the cell-cycle, nucleotide-excision repair, cellular response to stress CoA thioesterase activity and cellular localization specific to the nucleoplasm. These processes may explain at least part of the heterogeneity in clinical response to treatment. Third, comparing Li gene expression signatures identified in our study to other small molecule perturbations observed in cultured human cell lines revealed a strong enrichment for changes produced by clonidine (an anti-hypertensive drug) treatment. Finally, we were able to identify high overlaps of Li-regulated gene expression found in our study with previously published transcriptomic reports and further refine genes specific to Li responders and non-responders. These results represent a step towards better understanding the mechanisms underlying Li treatment in BD and identification of key genes involved Li's therapeutic action.

8.4.1. Gene Networks Induced by Lithium Treatment

The most striking observations from this study were the discovery of several Li induced and repressed gene modules stemming from our network analysis (**Figure 8.1**). The identification of a module specific to apoptosis signaling (M3: 324 genes) up-regulated with Li treatment appears to be a common denominator across many studies investigating therapeutic effects of Li (Lowthert et al., 2012; Beech et al., 2013; Zhang et al., 2005). Li has been reported to influence apoptosis in several cell types and previous reports demonstrate that Li up-regulates anti-apoptotic genes and down-regulates pro-apoptotic genes in Li responders relative to Li non-responders (Lowthert et al., 2012; Beech et al., 2013). Interestingly, module M3 was significantly enriched for BCL-2 family proteins (*P*=2.8E-03), which included both pro-apoptotic members (e.g. *BAD*, *BAX, BAK1, BMF*) and anti-apoptotic (e.g. *BCL2L1* and *MCL1*) proteins members. Further, regulatory effects of Li on apoptosis-controlling proteins occur in both the mitochondria and the ER, while ER stress is known to result in apoptosis (Ghribi et al., 2002; Yeste et al., 2006; Hiroi et al., 2005), supporting the identification of a module up-regulated by Li implicating response to ER stress (M12: 157 genes). Indeed, impairment of ER function has also been linked to the neuropathology of a variety of neurodegenerative diseases that involve neuronal apoptosis, such as cerebral ischemia and Alzheimer's disease (Mattson et al., 2001; Sherman & Goldberg, 2001; Paschen, 2003). Additionally, the up-regulated modules involved in immune response (M2: 45 genes) and defense response to virus (M7: 192 genes) following Li treatment provide further support for genes underlying cell proliferation and immune response. These findings may point to specific pathways via which Li acts to produce granulocytosis and lymphopenia while activating both phagocytic cells and lymphocytes (Lenox et al., 2003).

8.4.2. Gene Networks Repressed by Lithium Treatment

The identification of Li repressed gene modules include the down-regulation of two modules specific to protein targeting to the ER (M5: 313 genes) and translational initiation (M22: 188 genes), which may reflect the mechanism by which Li affects protein synthesis, reducing translation. This is consistent with previous reports indicating Li's effect in lowering the protein translation. Indeed, Li can promote proteasome-mediated degradation (Jing et al., 2013) and influence components of the translational machinery (Bosetti et al., 2002; Karyo et al., 2010) consequently interfering with protein turnover and thus affecting neurological function. However, in this context it has been proposed that Li may have therapeutic benefits for neurodegenerative disorders that are caused by over-expression of proteins (e.g. alpha-synuclein in Parkinson's disease). Further, we observed the down-regulation of a module, which was unable to be functionally annotated (M11: 156 genes). However, top hub genes for this module included K(lysine) acetyltransferase 2A (*KAT2A*), known to regulate hippocampal gene expression linked to memory formation (Stilling et a., 2014) as well as FK506 binding protein (*FKBP5*), involved in regulating glucocorticoid receptor sensitivity (Semba et al., 2000). Hyperactivity of the stress hormone system that is consistently found in chronically depressed and manic patients may be linked to an impaired negative feedback regulation of the HPA axis through the glucocorticoid receptor (Semba et al., 2000). Protein kinase 3 (*PKC3*), also representing a down-regulated hub gene, is implicated in the regulation of neurotransmitter release, neuron excitability and long-term changes in PCK-regulated protein function (Manji et al., 1994). An important cofactor for PKC is phosphatidylserine (PS) (Vance & Steenbergen, 2005), which was found in a module (M17: 17 genes) consistently up-regulated in BD patients relative to healthy controls in both vehicle and Li treatments. The exposure of PS on the outside surface of cells is widely believed to play a key role in the removal of apoptotic cells (Vance & Steenbergen, 2005).

8.4.3. Lithium and Clonidine Treatment: A Shared Mechanism of Action

The interrogation of DSigDB with a signature query composed of our Li gene expression signature revealed that clonidine treatment elicits the most similar effects of up- and down-regulated gene expression as the effects of Li identified in our study (**Figure 8.4**). Clonidine, as monotherapy or adjunctive therapy, is reported to be efficacious in treating attention deficit hyperactivity disorder (ADHD) symptoms and anxiety disorders in children and adolescents with or without comorbid disorders[45]. Clonidine treatment has been attributed with improvements in inattention, impulsivity and hyperactivity (Ming et al., 2011; Jaselskis et al., 1992). It has been reported that that clonidine treatment, rather than Li, is associated with symptomatic development of hypotension and depression (Zubenko et al., 1984). These results suggest clonidine, and chemically similar compounds, as a putatively useful gene-hunting tool for elucidating mechanistic mood stabilizing affects in either primary or lymphoblast derived cell lines to further elucidate these effects in search for more verifiable biomarkers.

8.4.4. Heterogeneity in Lithium Response

Genes identified as being differentially regulated with Li treatment between responders and non-responders included genes encoding cell-cycle and nucleotide-excision repair (**Figure 8.3**). A recent report investigating interaction networks of Li and valproate (an alternative medication for treating BD) revealed that valproate (but not Li) induced a highly enhanced recruitment of nuclear lumen processes enriched for the cell cycle, nucleotide excision repair and DNA replication pathways (Gupta et al., 2011). Conversely, our results suggest that subtle differences in these processes between Li responders and non-responders may explain part of the heterogeneity in clinical response to treatment. RAN binding protein 2 (*RANBP2*) in particular displayed a large difference in response between treatment groups and constitutes a key hub gene controlling a network enriched for cell cycle and related processes.

8.4.5. Strengths and Limitations

An obvious strength of our study is the hypothesis-free nature of genome-wide expression studies. Moreover, we examined mRNA expression in LCLs across haplotype matched BD Li responders, non-responders and healthy control donors, so that confounders caused by the presence of unique polymorphic DNA sequence alleles are unlikely to contribute to our observations. The inclusion of these three groups measured before and after Li treatment represents an important experimental design strategy in an attempt to elucidate Li's mechanism of action and to better understand its therapeutic effects. While the exact and precise mechanisms of Li's effects still remain clouded, we revealed several novel candidate Li responsive gene networks and displayed their differences across all treatment groups. Moreover, we were also able to confirm 48 candidate Li responsive genes (**Supplementary Table 8.2**), 9 genes specific to responders and 14 specific to non-responders (**Table 8.1**) found in our study which were also present in independent transcriptomic reports. Administration of Li monotherapy allowed us to rule out alternative concurrent medications affecting gene expression.

Our exploratory study has some limitations. First, it was apparent that the general effects of Li on gene expression were analogous across responders, non-responders and healthy controls (**Figure 8.2**). Indeed, while 1mM exposure of Li represents a true clinical dosage, future studies may benefit from measuring the effects of Li on gene expression patterns in a dose-response manner. Second, we utilized LCLs from BD and healthy unrelated donors to study the transcriptional effects of therapeutic Li exposure. While LCLs have been instrumental in pharmacogenomics discovery due to their ability to capture the natural variation of the human genome and by reducing environmental influences and cell type heterogeneity which may affect gene expression results (Wheeler & Dolan, 2012; Shim et al., 2012; Sie et al., 2009), there are limitations. One has to keep in mind that transcriptomic drug effects may differ in the primary cells compared to EBV transformed cell lines. In particular, transforming and culturing LCLs under laboratory conditions may not represent natural gene expression *in vivo* due to a large percentage of pauciclonality and widespread monoallelic expression (Min et al., 2010). Moreover, comparative studies between primary B lymphocytes and LCLs on the same subjects have found disagreeing changes at both the gene expression and DNA methylation levels (Caliskan et al., 2011). Additionally, despite the ability of Li to induce apoptotic signalling in our study (**Figure 8.3B**), EBV transformation has also been demonstrated to alter processes of apoptosis in response to certain drugs, which should be considered when LCLs are used in pharmacogenomics studies (Liu, 2004). This should also be considered when interpreting Li-induced apoptotic signalling found in our LCL samples, as well as in other studies (Lowthert et al., 2012; Beech et al., 2013). While a model system is clearly needed, future studies using patient-derived neuronal cultures differentiated from induced pluripotent stem cells may represent a more disease-relevant cell type. Third, the results from this early exploratory study can not yet be used to full understand the mode of action of Li nor its therapeutic variability and require validation in a larger cohort of BD patients. Finally, as a result of our modest sample size, the statistical significance of gene expression changes is low for a genome-wide transcriptome study. Apart from the comparison of vehicle to Li treated samples, cross-sectional and longitudinal contrast differential gene expression analyses revealed few genes passing an FDR corrected *P*-value of significance. As a result, we used varying *P*-value cut-offs to gather sufficient enough information for down-stream functional enrichment and interaction analyses. In spite of this, our exploratory approach was able to characterize putative direct PPI occurring within these candidate genes and found a network strongly enriched with physical interactions (**Figure 8.6B**), supporting these relaxed assumptions of significance.

8.4.6. Concluding Remarks

This explorative study used RNA-Seq gene expression of LCLs derived from BD responders, non-responders and healthy donors to investigate the expression levels of genes under the influence of Li treatment compared to vehicle. Li treatment was associated with the induction and repression of several cellular signaling pathways. These pronounced gene module differences are most likely to be a consequence of Li treatment and represent non-therapeutic cellular reprogramming of gene expression, rather than representing putative pathways differing between treatment groups. However, focusing analysis on individual genes differing in expression between BD responders and non-responders following Li treatment identified differential dysregulation of genes encoding for cell-cycle, nucleotide-excision repair and cellular response to stress. The implications of the genes reported here for the etiology and treatment of BD should ideally be examined with the blood samples of large cohorts of BD patients, before and after several weeks of treatment with Li. Comparing such transcriptomic changes between good and poor Li responders may contribute to the personalized treatment of BD.

Contributions. *I was solely responsible for all statistical design, data analysis, data interpretation and writing.*

Part III

Moving Biomarkers Forward in Psychiatry

Chapter 9

Moving Biomarkers Forward in Psychiatry

The preceding four chapters depict the potential for genome-wide transcriptome profiling of patient blood samples for biomarker discovery in psychiatry. First, the identification of blood-based gene networks capable of characterizing PTSD risk (at pre-deployment) and PTSD development (at post-deployment) implicating innate immunity were replicated across two independent cohorts of U.S. Marines (Chapter 5). Second, identification of a selective regulation of NK cell cytotoxicity events and gender-specific transcriptional responses at the gene network level portrayed the molecular response to short-term acute psychological stress (Chapter 6). Third, in application to MAP, blood-based gene networks and single gene biomarkers implicated in ubiquitin-mediated proteasome and circadian clock dysfunction were identified, and results were supported by CFG evidence (Chapter 7). Finally, in application in BD patient derived LCLs, a putative mechanism of action for lithium treatment was delineated along with several surrogate markers differing in response to lithium between BD responders and non-responders (Chapter 8). Overall, the results from these applications indicate a promising role for genome-wide blood transcriptome tools for biomarker discovery in psychiatry spanning prognostics, diagnostics and treatment responses to therapeutic interventions.

Such early developmental and exploratory blood-based biomarker research is useful as long as it is understood that intriguing preliminary insights may not accurately predict the eventual utility of the marker in clinical practice. If the putative biomarker demonstrates potential, the hypotheses generated in this early developmental phase should then be evaluated in a series of subsequent validation studies, each with increasing methodological rigor. Thus, building on results and insights gained from previous Chapters 5-8, it seems reasonable to outline a set of conditions for the further evaluation of blood-based biomarkers in patients with psychiatric disease.

9.1. Working on the Right Problem

It is important to be clear about the population which is being accessed for discovery research; for which well-defined phenotype, in what clinical diagnostic group, of what gender, the biomarker(s) have been identified. Any error in diagnosis or other means of categorizing participants is a form of measurement error and can invalidate or lead to inconsistent results across studies. As within Chapters 5-8, when defining the target population, it is useful to characterise the patient's diagnosis, stage of illness, age, gender and other features which are through to be relevant to the research question. In many instances, we were able to gather detailed clinical data (Chapter 5), physiological measurements (Chapter 6) and neurocognitive data (Chapter 7) which allowed us to gleam as much information as possible about the target populations being tested. It further permitted for correlating clinical data with molecular data (e.g. **Table 5.4**, **Figure 6.7**, **Figure 7.4**), accelerating new basic discoveries and the translation of research results in clinical practice. Long-term benefits of such multi-modal approaches may include improved diagnosis, reduced costs and the avoidance of jumping to premature negative and/or positive conclusions.

On the other hand, while reducing sample heterogeneity may increase the likelihood of finding an effect, it may also reduce the potential generalizability of the biomarker. That is, fractionizing disease states into more numerous and homogenous categories, without *a priori* biological validation, could make it harder to find specific biomedical tests that might diagnose or predict the disorder (Kapur et al., 2012). For example, considering and accounting for disease-relevant co-morbidities of a target population could reduce sample heterogeneity. For instance, if a significant fraction of PTSD participants from Chapter 5 had also been diagnosed with depression, this co-morbidity information may reduce the sensitive of a putative classifier to generalize across the entire PTSD cohort. On the other hand, this information may present a useful strategy to inform between finer gradients of the illness. This concept also carries over to patients in Chapter 7 diagnosed with MAP, a condition were patients may also show symptoms of hypomania, depression or schizophrenia. Despite, the concept of clear categories of psychiatric disorders as long been questioned and a dimensional spectrum may provide a better representation of clinical reality (Helzer et al., 2006).

9.2. The Right Measure of Statistical Significance and Effect Size

A recurring theme in Chapters 5-8 is the execution and interpretation of a measure of statistical significance between cases and controls based on clinical, molecular and/or neurocognitive data. Thus, it could be thought that strong statistical significance indicates high clinical utility, since any putative biomarker of clinical interest must demonstrate a strong degree of significance. However, many biological findings in psychiatry are of only small or moderate effect size, even though many of them meet the "$P < 0.05$" test of statistical significance. It has been argued that most initial reports are statistically significant but of small-effect size and have never been substantiated (Ioannidis, 2005); in findings which have been replicated, effect sizes are often lower than originally thought (Ioannidis, 2008). Given that efforts to replicate an initial finding usually involve a different clinical setting, a different participant selection and slightly different methods, the chance of replication after an original finding with a $P < 0.05$ is often low (Cumming, 2008; Miller, 2009). Most studies in the field of molecular psychiatry tend to be underpowered in statistical terms (Rothpearl et al., 1981; Allen et al., 2009). Similarly, despite measures of statistical significance, sample numbers from the exploratory investigations in Chapters 5-8 are also underpowered, and results should be interpreted cautiously.

This problem is analogous with that of 'approximate' replications of candidate findings (Maxwell, 2004). An initial underpowered study is often followed by another study of a similar size but with some additional measures and variables to give it some novelty and distinction. These subsequent studies usually have only modest statistical power to definitively confirm or refute the original findings, but have sufficient new measures to generate another significant finding – even though not precisely the one observed in the first study. For example, decades of research have attempted to better understand the mechanism of action of Li (as in Chapter 8). Analogous to our study (Chapter 8), a recent study profiled the transcriptome of LCLs derived from BD patients classified as responders and non-responders (Hunsberger et al., 2015). However, our results were distinct due to the inclusion of healthy control patients and comparing drug treatments with similar mechanism of action of Li using public databases. However, there can also be advantages to 'approximate' replications and repeated studies. In this case, we were able to cross-reference our results with this previous independent study to further refine large candidate gene-lists (e.g. **Table 8.1**).

9.3. Making the Right Comparisons

Another methodological issue is the inability to transfer a putative biomarker into clinical practice, when it was derived from making an extreme-type of comparison. Studies in biological psychiatry often assess the utility of a biomarker in a cohort with a uniform diagnosis by comparing it to healthy controls with no psychiatric or neurological history. While this approach is useful for detecting an effect or a relationship, a diagnostic test validated in this manner may be impractical when the time comes to apply it in wider clinical samples. Experimental designs implemented in Chapter 7 and Chapter 8 accounted for extreme-type of comparisons to some extent. For example, Chapter 7 included two-levels of biological controls for the identification of MAP biomarkers: healthy control and MA patients. Similarly, Chapter 8 also contained two-levels of controls for determining genes specific to BD Li responders; healthy controls and BD Li non-responder patients. This level of information provides further disease-related sensitivity and specificity, which could renderer the biomarker able to discriminate between lesser degrees of illness in the general patient population. Studies that fail to provide a description of their population or the test examined most likely lead to inflated estimates of accuracy.

9.4. Using the Right Approach

9.4.1. Experimental Designs

Prospective cohort studies in which individuals begin free of the outcome at baseline and followed longitudinally, are most useful in studies of prognostic markers. Again, a primary example of such an experimental design is found in Chapter 5. By profiling paired blood samples of U.S. Marines both before and after exposure to conflict zones we were able to identify prognostic signatures (**Figure 5.3**, **Figure 5.11**) capable of predicting the eventual development of PTSD at pre-deployment. Prospective study designs can also be used to better understand the physiological responses to disease, and when followed over extended periods of time may permit for discovery of disease biomarkers that are associated with disease screening and staging. However, repeated measures on the same group over extended periods can be costly and difficult. In such cases, special attention needs to be focused on reducing technical variation from initial venesection to RNA preparation and sequencing, as these steps could involve differing reagents across different dates and sites. When appropriate, short-term longitudinal studies, as in Chapter 7, may alleviate technical variation observed in longer-term studies. By contrast, retrospective cohort studies select patients based on previously recorded exposures (or measurements) and assess outcomes in the present. The case-control study can be a retrospective study design where the recruitment starting point for cases is the current presence of a desired outcome. These studies are frequently used to assess biomarkers as diagnostic tests, and are often are unable to provide information regarding causality. For example, the cross-sectional study design in Chapter 7 permitted for the identification of candidate blood biomarkers and gene networks of MAP that represent diagnostic markers, and causality was unable to be determined.

9.4.2. Alternative Next Generation Sequencing Technologies

Another aspect of the right approach for biomarker discovery is determining which biological/neurobiological entity deserves measuring. All of the primary research performed in the context of this thesis (Chapters 5-8) was done at the transcriptome level, however optimization of alternative technologies may also provide favorable biomarkers for psychiatric diseases. For example, a bottom-up approach to understanding and treating mental disorders could begin at the DNA level and the associated genetics. Genome-wide association studies (GWAS) and copy number variation (CNV) analysis usually contrast the frequencies of genetic variants between cases and controls for a large set of genetic markers (usually 500K-1M) distributed across the genome. A genetic contribution to psychiatric disorders has been established from both clinical (Kendler & Gardner, 1997) and epidemiological (Lichtenstein et al., 2010) studies, showing increased risk of disorders in relatives of affected individuals. Moreover, patterns of epigenetic modifications (e.g. DNA methylation) serve as epigenetic biomarkers to represent gene activity and expression as well as chromatin state (Heinkoff et al., 1997). Similar to the transcriptome, the epigenome is more dynamic than its DNA sequence, and may be altered environment, stochastic events and genetic background (Weaver et al., 2004). Whole-genome bisulfite sequencing and methylation-based arrays have revealed DNA methylation patterns implicated in various neural processes, from learning and memory to seizures and neurogenesis, and to suicide (Labonte et al., 2012), depression (Perroud et al, 2011) and chronic stress (Tyrka et al, 2012). Furthermore, several additional advances in mass spectrometry are giving promise to genome-wide proteome and metabolome profiling. Pineaar and collogues (2008), portray ‘neuroproteomics’ as an emerging tool to establish disease-associated protein profiles, while also generating a greater understanding as to how these proteins interact at the post-translational level. Metabolome profiling aims to quantitatively measure all small molecule metabolites found within a cell and use this information to understand the response to pathophysiological stimuli or genetic modification. Advantages of metabolomics include the relatively small number of biomarkers (~2,500-3,000) to be profiled, which is cost-effective. Overall, it remains unclear as to which NGS approach may provide the most favorable outcomes for biomarker discovery. Metabolomics might be more direct (albeit more limited) than proteomics, which in turn, if used in an unbiased discovery fashion, may be more powerful than gene expression, which in turn is more powerful than genetics, as thousands of single-nucleotide polymorphisms can converge in the regulation of expression of a gene.

Future studies may also consider optimizing transcriptomic procedures in order to measure the entire landscape of RNA species. For example, based on the significant dysregulation of genes encoded for RNA degradation and ubiquitin-mediated proteolysis in the context of MAP (Chapter 7), future studies of MAP may benefit by investigating the miRNA fraction of the transcriptome. Such an approach could identify candidate miRNAs responsible for directing mRNAs towards a pathway of degradation.

9.4.3. Neuroimaging

Many forms of neuroimaging have the potential to be used as biomarkers in psychiatric illness. Positron emission tomography (PET) can be used to (i) characterize resting-state metabolic signatures; or (ii) to measure the density of neurotransmitter receptors or transporters for which a radioligand exists. Magnetic resonance imaging (MRI) can be used to measure: (i) subcortical brain structural volumes (sMRI) as in Chapter 7; (ii) white matter integrity and density [diffusion tensor imaging (DTI)]; or (iii) functional metabolic activity patterns (fMRI), either in the resting state or in response to a certain challenge or task. fMRI patterns reflect states of brain metabolic activity (Mayberg, 2014). Greater metabolic activity in a brain region is accompanied by increases in blood flow, which is detected as alterations within the magnetic field of the MRI scanner. Furthermore, fMRI may be used to examine activity in single brain regions or in coordinated temporal patterns of activity across multiple regions (functional connectivity MRI [fcMRI] (Fox et al., 2007)). Although neuroimaging methods have yielded important research findings about psychiatric disorders, the routine use of these methods is not yet justifiable in the diagnostic evaluation of individual patients (First et al., 2012). Future avenues of research may benefit by combining blood-based measurements with brain-based measurements, as in Chapter 7. In this context we were able to identify candidate blood-based predictors of brain function and permit for mechanistic implications to be made based purely on blood-based gene dysregulation. Building off of this work, measuring the entire transcriptome (e.g. mRNAs, miRNAs, lncRNAs etc…) and several brain-based measurements (e.g. sMRI, fRMI, DTI etc…) may provide a useful strategy for identifying blood-based predictors of brain structure, function and/or chemistry.

9.4.4. Integrating Panels of Bio-signatures

As psychiatric disorders are complex illnesses, it could be argued that it is unlikely that any single biomarker could accurately predict outcomes in individual patients. Rather, a combination or an integrated panel of the biomarkers could provide sufficient information (Schunemann et al., 2008), as mentioned above when considering blood predictors of brain status. Similar to what was achieved in Chapter 7, a range of potential ‘readouts’ could be combined, including biochemical, cognitive, electrophysiological, genetic and neuroimaging markers. However, some important factors should be taken into account when considering the generation of such ‘multi-marker’ tests based on the use of a composite of several biomarkers (measured in parallel) for predicting disease risk and patient outcomes. One factors is ‘multi-collinearity’, or the inherent correlation between biomarkers which track the same process in one individual. If the correlation between two markers is very high, then measuring one of them is sufficient to capture an event. Another factor is ascertaining the incremental utility of adding a new biomarker to a panel; at some point, including and measuring additional factors will not improve diagnosis accuracy or change patient management or clinical outcome.

9.4.5. Systems-level Diversity

It does not inevitably follow that a biological tests for psychiatric illnesses would provide the most informative or effective methods for identifying them. If we consider a biological root of disease, we have to consider the complexity of a system and that genes, pathways, cells, and neuronal circuits have to work via dozens of mechanistic levels. For gene expression studies, co-expression network analysis (as performed in Chapters 5-8) leverages the fact that gene expression reflects the stat of the cellular or tissue system that is being analysed. It is also unlikely that a single biological alteration will have a dominant one-to-one mapping with a DSM- or ICD-defined mental disorder. Most psychiatric diseases are likely to be associated with perturbations in complex neurobiological networks spanning a hierarchy of different molecular levels (genome, epigenome, transcriptome, proteome). An alternative explanation for why biological psychiatry findings have proven difficult to reproduce and hence to translate into novel therapies is because the complexity of the system being measured has been under-appreciated. Complex systems, such as the human immune system, are generally democratized into their individual parts in biomedical research to allow for a more detailed understanding of limited aspects of the system and to describe mechanistic detail (Brodin, 2013).

9.4.6. Statistical Design

Choice of the appropriate statistical tests for analysis of biomarkers depends on the purpose of the study, as well as the variables under examination. For transcriptome-based studies, while it appears that the identification of biomarkers may be accomplished by gene network approaches (Chapters 5-8) and machine-learning approaches (Chapters 5&7), it remains uncertain which approach provides the most favorable translational avenues. Statistical approaches that consider system complexity (e.g. WGCNA) are particularly useful in providing comprehensive characterizations of the molecular factors in a given disease state and for multi-scale data integration, and are statistically robust in terms of reproducibility (Langfelder & Horvath, 2008). Machine-learning applications, while often fit-to-cohort, rank genes by importance producing a unique predictive or diagnostic panel of biomarkers (Simon et al., 2007). Machine-learning applications are unable to provide a systems-level picture of molecular mechanisms underpinning disease pathology, but they represent a useful step forward in the construction and validation of a multi-marker signature (i.e. biosignatures) for psychiatric illness. It would be useful for future studies to provide comprehensive empirical evaluations of novel data. As data is passed through these various tools (i.e. gene networks, differential expression, machine-learning) it may also be useful to establish a user-friendly interface capable of storing each set of results in a simple and retrievable format. This would ensure that each independent study uses the same, yet diverse, sets of methodologies and that the subsequent results are formatted in a manner that allow for quick cross-referencing. The convergent functional genomic (CFG) approach, as in Chapter 7, represents a positive interim step in this direction. For example, the CFG approach permitted for identifying relevant blood-based biomarkers of MAP which had been validated in independent genetic and genomic reports of psychosis and schizophrenia.

9.5. Demonstrating Reproducibility in the Right Way

If a biomarker(s) appear(s) to be linked to an outcome of interest, technical details and feasibility are commonly assessed during the developmental and exploratory phases. Research at this stage can be useful as long as it is understood that it does not accurately reflect the usefulness of the marker in clinical practice. If the putative biomarker demonstrates potential, the hypotheses generated in this early developmental phase should be evaluated in a series of succeeding rigorous validation studies. Subsequently, external validation of the *a priori* selected biomarker(s) in a completely independent cohort is a requirement before making any sustainable claims. Various strategies for achieving reproducibility have been performed in the context of this thesis; the CFG approach (Chapter 7) and cross-referencing results with previous studies (Chapter 8). Perhaps the most straightforward interpretation of biomarker reproducibility was achieved using supervised multivariate machine-learning methods to develop a prognostic classifier on a training-set (Dataset 1) and cross-validating prediction accuracies using a completely independent test-set (Dataset 2) as in Chapter 5. In this context, if a PTSD a biomarker demonstrates adequate sensitivity, specificity and predictive value in the validation stage (as through machine-learning approaches), its usefulness in practice can then be assessed; ideally within a randomized controlled trial where a group of patients is randomized to either undergo the biomarker test or not, and its effect on diagnostic or prognostic outcomes and patient and societal implications are assessed.

To accelerate reproducible biomarker signatures across independent cohorts, international research efforts collecting biological samples across independent cohorts sent for joint technical processing may bypass technical biases. Since the sample sizes needed for discovery and replication are beyond the reach of single groups, multiple consortia have emerged to foster scientific discovery. A good example of this are the ongoing research efforts of the Psychiatric Genomic Consortium (PGC) comprised of approximately 300 investigators and more than 75,000 patients with GWAS data under analysis (PGC, 2009).

9.6. The Right Use of Biomarkers

When deciding on whether or not to use a biomarker test, it is important to consider whether the patient will be better off having had the test than they would be if they had not undergone it. This requires consideration of whether the sample and setting in which the biomarker has been tested is sufficiently similar to the situation of interest in the clinic to justify applying it. If the test is unlikely to apply to your patient, it is not worth performing. Alternatively, if the prognostic or diagnostic test is associated with a mental illness currently lacking solid evidence-based treatments, the test may be unfit. Outcomes such as quality of life and peace of mind of the patient should also be considered. Unfortunately, the quality and type of data required to make decisions based on biomarker tests is often not available for most, if not all, psychiatric biomarkers. Demonstrating that the biomarkers have predictive ability for future clinical course is necessary for the field before adopting and incorporating them into clinical practice.

9.7. Summary

Naturally, as evidence evolves, any biomarker strategy will evolve accordingly. The delay of accurate and objective biomarker tests in psychiatry is expected given a later start than in other areas of medicine, the inherent neurobiological complexity, and the changing nature of psychiatric nosology. Optimistically, the opportunity allowed by the substantial progress in NGS and neuroimaging combined with advances in computer science, mathematics and systems-biology are unprecedented and may deliver useful clinical tests in the not too distant future. These tests could identify homogenous populations for whom targeted new therapeutics could be developed, thereby realizing a vision of a new stratified or personalized practice in psychiatry to either replace or supplement current diagnostic criteria.

Appendices

References

1. Abbas AR, Baldwin D, Ma Y, Ouyang W, Gurney A, Martin F et al. Immune Response in Silico (IRIS): Immune-specific Genes Identified from a Compendium of Microarray Expression Data. *Genes and Immunity* 2005; 6.4:319-31.

2. Abe K, St George-Hyslop PH, Tanzi RE, Kogure K. 1991. Induction of amyloid precursor protein mRNA after heat shock in cultured human lymphoblastoid cells. *Neurosci Lett* 125:169–171.

3. Achiron A, Gurevich M, Friedman N, Kaminski N, Mandel M. Blood Transcriptional Signatures of Multiple Sclerosis: Unique Gene Expression of Disease Activity. *Annals of Neurology* 2004; 55.3: 410-17.

4. Ackerman KD, et al. Stressful life events precede exacerbations of multiple sclerosis. *Psychosom Med* 2002; 64:916-20.

5. Adler LE, Freedman R, Ross RG, et al. Elementary phenotypes in the neurobiological and genetic study of schizophrenia. *Biol Psychiatry* 1999; 46:8–18.

6. Affymetrix I. Statistical Algorithms Description Document. 2002. http://www.affymetrix.com/support/technical/whitepapers.affx.

7. Agarwal DP. The genetics of alcohol metabolism and alcoholism. *Indian J Hum Genet* 2001; 1:25–32.

8. Ahmad M. Biomarkers in Acute Myocardial Infarction. *J Clin Exp Cardiolog* 2012; 03. doi:10.4172/2155-9880.1000222.

9. Aiboshi J, Moore EE, Ciesla CJ, Silliman CC. Blood transfusion and the two-insult model of post-injury multiple organ failure. *Shock* 2001; 15; 302–306.

10. Al'Abadie MS, Kent GG, Gawkrodger DJ. The relationship between stress and the onset and exacerbation of psoriasis and other skin conditions. Br *J Dermatol* 1994;130:199-203.

11. Allen AJ, Griss ME, Folley BS, Hawkins KA, Pearlson GD. Endophenotypes in schizophrenia: a selective review. *Schizophr Res* 2009; 109 : 24--37.

12. Alston WP. Traits, consistency, and conceptual alternatives for personality theory. *J Theor Soc Behav* 1975; 5:17–48.

13. Altar CA, Jurata LW, Charles V, Lemire A, Liu P, Bukhman Y. et al. Deficient

hippocampal neuron expression of proteasome, ubiquitin, and mitochondrial genes in multiple schizophrenia cohorts. *Biol Psychiatry* 2005; 58: 85–96.

14. Altemus M, Rao B, Dhabhar FS, Ding W, Granstein RD. Stress-induced changes in skin barrier function in healthy women. *J. Invest. Dermatol*. 2001;117:309-317.

15. Amaratunga D, Cabrera J. Analysis of Data From Viral DNA Microchips. *Journal of the American Statistical Association* 2001; 96: 1161-1170.

16. American Psychiatric Association (2013) *Diagnostic and Statistical Manual of Mental Disorders (5th edn) (DSM-5).* APA.

17. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders: 4th edition.* American Psychiatric Press: Washington DC, 2000.

18. Amkraut AA, Solomon CF, Kraemer HC. Stress, early experience and adjuvant-induced arthritis in the rat. *Psychosom* Med 1971; 33: 203-14.

19. Anders S, Pyl P, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 2014; 31: 166-169.

20. Arai R, Ito K, Ohnishi T, Ohba H, Akasaka R, Bessho Y et al. Crystal Structure of Human Myo-inositol Monophosphatase 2, the Product of the Putative Susceptibility Gene for Bipolar Disorder, Schizophrenia, and Febrile Seizures. *Proteins: Structure, Function, and Bioinformatics* 2007; 67(3):732-42.

21. Asai T, Bundo M, Sugawara H, Sunaga F, Ueda J, Tanaka G et al. Effect of mood stabilizers on DNA methylation in human neuroblastoma cells. *Int J Neuropsychopharmacology* 2013; 16: 2285–2294.

22. Atkinson AEA. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther* 2001; 69:89–95.

23. Austin AW, Wissmann T, Von Kanel R. Stress and Hemostasis: An Update. *Seminars in Thrombosis and Hemostasis* 2013; 39.08: 902-12.

24. Ayalew M, Le-Niculescu H, Levey DF, Jain N, Changala B, Patel SD et al. Convergent functional genomics of schizophrenia: from comprehensive understanding to genetic risk prediction. *Mol Psychiatry* 2012; 17: 887–905.

25. Baldessarini R, Tondo L. Does Lithium Treatment Still Work?. *Arch Gen Psychiatry* 2000; 57: 187.

26. Beck F, Geiger J, Gambaryan S, Veit J, Vaudel M, Nollau P *et al.* Time-resolved characterization of cAMP/PKA-dependent signaling reveals that platelet inhibition is a concerted process involving multiple signaling pathways. *Blood* 2014;123(5)e1-e10.

27. Beck, AT, Steer RA, Brown GK. Manual for the Beck Depression Inventory-II. San Antonio, TX: Psychological Corporation. 1996.

28. Becker JB, Monteggia LM, Perrot-Sinal TS, Romeo RD, Talylor JR et al. Stress and disease: is being female a predisposing factor? *J Neurosci*. 2007; 27:11851-11855.

29. Beech R, Leffert J, Lin A, Sylvia L, Umlauf S, Mane S et al. Gene-expression differences in peripheral blood between lithium responders and non-responders in the Lithium Treatment-Moderate dose Use Study (LiTMUS). *Pharmacogenomics J* 2013; 14: 182-191.

30. Beliakova-Bethell N, Massanella M, White C, Lada S, Du P, Vaida F et al. The effect of cell subset isolation method on gene expression in leukocytes. *Cytometry* 2013; 85: 94-104.

31. Benjamini Y, Hochberg Y.Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing *Journal of the Royal Statistical Society. Series B (Methodological)* 1995; 57: 289-300

32. Bernard K, Cambiaggi A, Guia S, Bertucci F, Granjeaud S, et al. Engagement of Natural Cytotoxicity Programs Regulates AP-1 Expression in the NKL Human NK Cell Line. *The Journal of Immunology* 1999; 162(7):4062-4068.

33. Berridge M, Downes C, Hanley M. Neural and developmental actions of lithium: A unifying hypothesis. *Cell* 1989; 59: 411-419.

34. Bhatia M, Moochhala S. Role of inflammatory mediators in the pathophysiology of acute respiratory distress syndrome. *J. Pathol.* 2004; 202:145–156.

35. Biomarkers Definition Working Group, 1998, Gregory Downing, NIH Initiatives in Surrogate Endpoints and Endpoint Analysis, Non-clinical Studies Subcommittee, Advisory Committee for Pharmaceutical Science presentation, 2000.

36. Biomarkers Definitions Working Group: Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther* 2001; 69:89–95.

37. Black PH. The inflammatory consequences of psychologic stress: Relationship to insulin resistance, obesity, atherosclerosis and diabetes mellitus, type II. *Medical Hypotheses* 2006; 67(4):879-891.

38. Blake, DD, Weathers FW, Nagy LM, Kaloupek DG, Gusman FD, Charney DS *et al.* The development of a Clinician-Administered PTSD Scale. *J. Trauma. Stress* 1995; 8: 75-90.

39. Blanchard EB, Hickling EJ, Taylor AE, Loos WR. Psychiatric morbidity associated with motor vehicle accidents. *J. Nerv. Ment. Dis.* 1995;183:495-504.

40. Blanchard EB, Hickling EJ, Vollmer AJ, Loos WR, Buckley TC, Jaccard, J. Short-term follow-up of post-traumatic stress symptoms in motor vehicle accident victims. *Behaviour Research and Therapy* 1995;33:369-77.

41. Blanchard EB, Hickling EJ., Barton KA, Taylor AE, Loos WR, Jones-Alexander J. One-year prospective follow-up of motor vehicle accident victims. *Behaviour Research and Therapy* 1996; 34: 775-86.

42. Bolger A, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014; 30: 2114-2120.

43. Bolstad B, Irizarry R, Astrand M, Speed T. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 2003; 19: 185-193.

44. Bosch JA, Berntson GG, Cacioppo JT, Dhabhar FS, Marucha PT. Acute stress evokes selective mobilization of T cells that differ in chemokine receptor expression: a potential pathway linking immunologic reactivity to cardiovascular disease. *Brain Behav Immun* 2003; 17(4):251-259.

45. Bosch JA, Berntson GG, Cacioppo JT, Marucha PT. Differential Mobilization of Functionally Distinct Natural Killer Subsets During Acute Psychological Stress. *Psychosomatic Medicine* 2005; 67(3):366-375.

46. Bosetti F, Seemann R, Rapoport S. Chronic lithium chloride administration to rats decreases brain protein level of epsilon (ϵ) subunit of eukaryotic initiation factor-2B. *Neuroscience Letters* 2002; 327: 71-73.

47. Bousman CA, Chana G, Glatt SJ, Chandler SD, Lucero GR, Tatro E et al. Preliminary Evidence of Ubiquitin Proteasome System Dysregulation in Schizophrenia and Bipolar Disorder: Convergent Pathway Analysis Findings From Two Independent Samples. *Am J Med Genet Part B* 2010; 153B: 494–502.

48. Bousman CA, Chana G, Glatt SJ, Chandler SD, May T, Lohr J et al. Positive symptoms of psychosis correlate with expression of ubiquitin proteasome genes in peripheral blood. *Am J Med Genet Part B* 2010; 153B: 1336–41.

49. Bousman CA, Glatt SJ, Everall IP, Tsuang MT. Genetic association studies of methamphetamine use disorders: a systematic review and synthesis. *Am J Med Genet B Neuropsychiatr Genet* 2009; 150B: 1025–49.

50. Bray PF, Mckenzie SE, Edelstein LC, Nagalla S, Delgrosso K, Ertel A, The Complex Transcriptional Landscape of the Anucleate Human Platelet. *BMC Genomics* 2013; 4.1:1.

51. Breen M, Beliakova-Bethell N, Mujica-Parodi L, Carlson J, Ensign W, Woelk C et al. Acute psychological stress induces short-term variable immune response. *Brain, Behavior, and Immunity* 2015. doi:10.1016/j.bbi.2015.10.008.

52. Breen M, Maihofer A, Glatt S, Tylee D, Chandler S, Tsuang M et al. Gene networks specific for innate immunity define post-traumatic stress disorder. *Molecular Psychiatry* 2015; 20: 1538-1545.

53. Breen MS, Stein DJ, Baldwin DS. A Systematic Review of Blood Transcriptomics and Complex Brain Disorders: Moving Beyond 'Surrogate Marker' Status. *Human Psychopharmacology: Clinical and Experimental* 2015. (*in review*)

54. Broadhurst D, Kell D. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics* 2006; 2: 171-196.

55. Brodin P, Valentini D, Uhlin M, Mattsson J, Zumla A, Maeurer M. Systems level immune response analysis and personalized medicine. *Expert Review of Clinical Immunology* 2013; 9: 307-317.

56. Brugha TS, Cragg D. The List of Threatening Experiences: the reliability and validity of a brief life events questionnaire. *Acta Psychiatr Scand* 1990; 82: 77-81.

57. Bullard J, Purdom E, Hansen K, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 2010; 11: 94.

58. Butcher SK, Lord JM. Stress Responses and Innate Immunity: Aging as a Contributory Factor. *Aging Cell* 2004; 3.4: 151-60.

59. Caboche S, Audebert C, Lemoine Y, Hot D. Comparison of mapping algorithms used in high-throughput sequencing: application to Ion Torrent data. *BMC Genomics* 2014; 15: 264.

60. Caliskan M, Cusanovich D, Ober C, Gilad Y. The effects of EBV transformation on gene expression levels and methylation profiles. *Human Molecular Genetics* 2011; 20: 1643-1652.

61. Carmeliet P, Strooper BD. Alzheimer's Disease: A Breach in the Blood–brain Barrier. *Nature* 2012; 485.7399: 451-52.

62. Carter JR, Goldstein DS. Sympathoneural and Adrenomedullary Responses to Mental Stress. *Comprehensive Physiology* 2011; 119-46.

63. Carver C, White T. Behavioral inhibition, behavioral activation, and affective

responses to impending reward and punishment: the BIS/BAS scales. *Journal of personality and social psychology* 1994*;* 67(2):319-33.

64. Casey B, Craddock N, Cuthbert B, Hyman S, Lee F, Ressler K. DSM-5 and RDoC: progress in psychiatry research?. *Nature Reviews Neuroscience* 2013; 14: 810-814.

65. Cawley D, Talbot NLC. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *J of Machine Learning Research* 2010; 11: 2079-2107.

66. Celano C, Huffman J. Depression and Cardiac Disease. *Cardiology in Review* 2011; 19: 130-142.

67. Charney DS. Psychobiological mechanisms of resilience and vulnerability: implications for successful adaptation to extreme stress. *Am J Psychiatry* 2004; 161:195-216.

68. Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res*. 2009; 37:W305-11.

69. Ciechanover A, Orian A, Schwartz AL. Ubiquitin-mediated proteolysis: Biological regulation via destruction. *Bioessays* 2000; 22: 442–51.

70. Clark SM, San J, Francis TC, Nagaraju A, Michael KC, Keegan AD *et al.* Immune status influences fear and anxiety responses in mice after acute stress exposure. *Brain, behavior, and immunity* 2014; 38: 192-201.

71. Cohen S, Janicki-Deverts D, Doyle W, Miller GE, Frank E et al. Chronic Stress, Glucocorticoid Receptor Resistance, Inflammation, and Disease Risk. *Proceedings of the National Academy of Sciences* 2012; 109(16):5995-999.

72. Corney D. RNA-seq Using Next Generation Sequencing. *Materials and Methods* 2013; 3. doi:10.13070/mm.en.3.203.

73. Cruceanu C, Alda M, Grof P, Rouleau GA, Turecki G. Synapsin II is involved in the molecular pathway of lithium treatment in bipolar disorder. *PloS One* 2012; 7: e32680.

74. Cumming G. Replication and p intervals: P values predict the future only vaguely, but confidence intervals do much better. *Perspect Psychol Sci* 2008; 3: 286--300.

75. Cuthbert B, Insel T. Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC Medicine* 2013; 11: 126.

76. Daly, ME. Determinants of Platelet Count in Humans. *Haematologica* 2010;

96.1: 10-13.

77. Davis AP, Grondin CJ, Lennon-Hopkins K, Saraceni-Richards C, Sciaky D, King BL et al. The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. *Nucleic Acids Res* 2015; 43(Database issue):D914-20.

78. Dhabhar FS, Miller AH, McEwen BS, Spencer RL. Effects of stress on immune cell distribution-dynamics and hormonal mechanisms. *J Immunol* 1995; 154:5511-27.

79. Dhabhar FS. Enhancing versus Suppressive Effects of Stress on Immune Function: Implications for Immunoprotection and Immunopathology. *Neuroimmunomodulation* 2009; 16.5:300-17.

80. Dieperink E, Leskela J, Dieperink ME, Evans B, Thuras P, Ho SB. The Effect of Pegylated Interferon-α2b and Ribavirin on Posttraumatic Stress Disorder Symptoms. *Psychosomatics* 2008; 49.3: 225-29.

81. Dillies M, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics* 2012; 14: 671-683.

82. Dobbin K, Zhao Y, Simon R. How Large a Training Set is Needed to Develop a Classifier for Microarray Data?. *Clinical Cancer Research* 2008; 14: 108-114.

83. Du P, Kibbe WA, Lin SM. Lumi: a pipeline for processing Illumina microarray. *Bioinformatics* 2008; 24(13):1547-1548.

84. Du Z, Li L, Chen C, Yu P, Wang J. G-SESAME: web tools for GO-term-based gene similarity analysis and knowledge discovery. *Nucleic Acids Research* 2009; 37: W345-W349.

85. Dumeaux V, Olsen K, Nuel G, Paulssen R, B√∏rresen-Dale A, Lund E. Deciphering Normal Blood Gene Expression Variation,ÄîThe NOWAC Postgenome Study. *PLoS Genetics* 2010; 6: e1000873.

86. Duong HA, Robles MS, Knutti D, Weitz CJ. A molecular mechanism for circadian clock negative feedback. *Science* 2011; 332(6036): 1446-9.

87. Ekblom R, Slate J, Horsburgh G, Birkhead T, Burke T. Comparison between Normalised and Unnormalised 454-Sequencing Libraries for Small-Scale RNA-Seq Studies. *Comparative and Functional Genomics* 2012; 2012: 1-8.

88. Eraly SA, Nievergelt CM, Maihofer AX, Barkauskas DA, Nilima Biswas N, Agorastos A *et al.* Assessment of Plasma C-Reactive Protein as a Biomarker of Posttraumatic Stress Disorder Risk. *JAMA Psychiatry* 2014; 71.4: 423.

89. Eysenck S, Eysenck H, Barrett P. A revised version of the psychoticism scale. *Personality and Individual Differences* 1985; 6(1): 21-9.

90. Felger JC, Miller AH. Cytokine Effects on the Basal Ganglia and Dopamine Function: the Subcortical Source of Inflammatory Malais. *Front Neuroendocrinol* 2012; 33(3): 315–27.

91. First MB, Gibbon M, Spitzer RL, Williams JBW. User's Guide for the Structured Clinical Interview for DSM-IV-TR Axis I Disorders – Research Version - (SCID-I for DSM-IV-TR, November 2002 Revision).

92. Fischl B, Salat DH, van der Kouwe AJ, Makris N, Segonne F, Quinn BT et al. Sequence-independent segmentation of magnetic resonance images. *Neuroimage* 2004; 23(Suppl.1): S69-S84.

93. Fox MD., Raichle ME. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nat Rev Neurosci.* 2007;8:700–711

94. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research* 2012; 41: D808-D815.

95. Fridhandler BM. Conceptual note on state, trait, and the state-trait distinction. *J Pers Soc Psychol* 1986; 50:169–174.

96. Friston KJ, Frith CD. Schizophrenia: a disconnection syndrome? *Clin Neurosci* 1995; 3: 89–97.

97. Fusar-Poli, P, Howes OD, Allen P, Broome M, Valli I, Asselin M-C et al. Abnormal Prefrontal Activation Directly Related to Pre-synaptic Striatal Dopamine Dysfunction in People at Clinical High Risk for Psychosis. *Molecular Psychiatry Mol Psychiatry* 2009; 16(1): 67-75.

98. Gaggin H, Januzzi J. Biomarkers and diagnostics in heart failure. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* 2013; 1832: 2442-2450.

99. Galatzer-Levy I, Bryant R. 636,120 Ways to Have Posttraumatic Stress Disorder. *Perspectives on Psychological Science* 2013; 8: 651-662.

100. Garg A, Chren MM, Sands LP, Matsui MS, Marenus KD, Feingold KR et al. Psychological stress perturbs epidermal permeability barrier homeostasis: implications for the pathogenesis of stress- associated skin disorders. *Arch Dermatol* 2001; 137:53-9.

101. Garnham J, Munro A, Slaney C, MacDougall M, Passmore M, Duffy A et al. Prophylactic treatment response in bipolar disorder: Results of a naturalistic observation study. *Journal of Affective Disorders* 2007; 104: 185-190.

102. Geer LY, Marchler-Bauer A, Geer RC, Han L, He J, He S *et al.* The NCBI BioSystems database. *Nucleic Acids Res.* 2010; 38:D492-6.

103. Gharaibeh R, Fodor A, Gibas C. Background correction using dinucleotide affinities improves the performance of GCRMA. *BMC Bioinformatics* 2008; 9: 452.

104. Ghribi O, Herman M, Spaulding N, Savory J. Lithium inhibits aluminum-induced apoptosis in rabbit hippocampus, by preventing cytochrome c translocation, Bcl-2 decrease, Bax elevation and caspase-3 activation. *Journal of Neurochemistry* 2002; 82: 137-145.

105. Giannoudis PV. Current concepts of the inflammatory response after major trauma: an update. *Injury* 2003; 34:397–404.

106. Glatt SJ, Tylee DS, Chandler SD, Pazol J, Nievergelt CM, Woelk CH *et al.* Blood-based gene-expression predictors of PTSD risk and resilience among deployed marines: A pilot study. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 2013; 162.4: 313-326.

107. Gouin J-P, Kiecolt-Glaser JK. The Impact of Psychological Stress on Wound Healing: Methods and Mechanisms. *Immunology and Allergy Clinics of North America* 2011; 31.1: 81-93.

108. Gould TD, Manji HK. The molecular medicine revolution and psychiatry: bridging the gap between basic neuroscience research and clinical psychiatry. *J Clin Psychiatry* 2004; 65:598–604.

109. Grandes G, Montoya I, Arietaleanizbeaskoa M, Arce V, Sanchez A. The burden of mental disorders in primary care. *European Psychiatry* 2011; 26: 428-435.

110. Greenberg DA, Cayanis E, Strug L, Marathe S, Durner M, Pal DK et al. Malic enzyme 2 may underlie susceptibility to adolescent-onset idiopathic generalized epilepsy. *Am. J. Hum. Genet.* 2005; 76:139-46.

111. Guest P, Bahn S. *Biomarkers of neurological and psychiatric disease*. Elsevier: Amsterdam, 2011.

112. Guo Y, Ye F, Sheng Q, Clark T, Samuels D. Three-stage quality control strategies for DNA re-sequencing data. *Briefings in Bioinformatics* 2013; 15: 879-889.

113. Gupta A, Schulze T, Nagarajan V, Akula N, Corona W, Jiang X et al. Interaction networks of lithium and valproate molecular targets reveal a striking enrichment of apoptosis functional clusters and neurotrophin signaling. *Pharmacogenomics J* 2011; 12: 328-341.

114. Gutekunst CA, Levey AI, Heilman CJ, Whaley WL, Yi H, Nash NR, Rees HD,

Madden JJ, Hersch SM. 1995. Identification and localization of huntingtin in brain and human lymphoblastoid cell lines with anti-fusion protein antibodies. *Proc Natl Acad Sci U S A* 92:8710–8714

115. Gvion Y, Apter A. Suicide and suicidal behaviour. Public Health Rev 2012; 34:1-20.

116. Hackstadt A, Hess A. Filtering for increased power for microarray data analysis. *BMC Bioinformatics* 2009; 10: 11.

117. Harald S, Badan I, Fischer B, Wagner AP. Dynamics of Gene Expression for Immediate Early- and Late Genes after Seizure Activity in Aged Rats. *Archives of Gerontology and Geriatrics* 2001; 32(3):199-218.

118. Harro J. Animal models of depression vulnerability. *Curr Top Behav Neuroscience* 2013; 14: 29-54.

119. Heber S, Sick B. Quality Assessment of Affymetrix GeneChip Data. *OMICS: A Journal of Integrative Biology* 2006; 10: 358-368.

120. Heinzelmann M, Gill J. Epigenetic Mechanisms Shape the Biological Response to Trauma and Risk for PTSD: A Critical Review. *Nursing Research and Practice* 2013; 2013: 1-10.

121. Helzer JE, Kraemer HC, Krueger RF. The feasibility and need for dimensional psychiatric diagnosis. *Psychol Med* 2006;36:1671–1680

122. Henikoff S, Matzke MA. Exploring and explaining epigenetic effects. *Trends in Genet*ics 1997;13:293-295

123. Henley SM, Bates GP, Tabrizi SJ. Biomarkers for neurode- generative diseases. *Curr Opin Neurol* 2005; 18:698–705.

124. Hiroi T, Wei H, Hough C, Leeds P, Chuang D. Protracted lithium treatment protects against the ER stress elicited by thapsigargin in rat PC12 cells: roles of intracellular calcium, GRP78 and Bcl-2. *Pharmacogenomics J* 2005; 5: 102-111.

125. Hsieh JH, Stein DJ, Howells FM. The neurobiology of methamphetamine induced psychosis. *Front Hum Neurosci* 2014; 8: 537.

126. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protoc. 2009;4(1):44-57.*

127. Hunsberger J, Chibane F, Elkahloun A, Henderson R, Singh R, Lawson J et al. Novel integrative genomic tool for interrogating lithium response in bipolar disorder. *Translational Psychiatry* 2015; 5: e504.

128. Ikeda M, Okahisa Y, Aleksic B, Won M, Kondo N, Naruse N, et al. Evidence for shared genetic risk between methamphetamine-induced psychosis and schizophrenia. *Neuropsychopharmacology* 2013; 38(10): 1864-70.

129. Ioannidis JP. Why most discovered true associations are inflated. *Epidemiology* 2008;19: 640--648.

130. Ioannidis JP. Why most published research findings are false. *PLoS Med* 2005;2:e124.

131. Jankowsky E, Harris M. Specificity and nonspecificity in RNA,Äìprotein interactions. *Nature Reviews Molecular Cell Biology* 2015; 16: 533-544.

132. Jaselskis C, Cook E, Fletcher K, Leventhal B. Clonidine Treatment of Hyperactive and Impulsive Children with Autistic Disorder. *Journal of Clinical Psychopharmacology* 1992; 12: 322-327.

133. Jing P, Zhang J, Ouyang Q, Wu J, Zhang X. Lithium treatment induces proteasomal degradation of over-expressed acetylcholinesterase (AChE-S) and inhibit GSK3β. *Chemico-Biological Interactions* 2013; 203: 309-313.

134. Johnson W, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2006; 8: 118-127.

135. Kadmiel M, Cidlowski JA. Glucocorticoid receptor signaling in health and disease. *Trends in Pharmacological Sciences* 2013; 34(9):518-530.

136. Kapp C. Crystal meth boom adds to South Africa's health challenges. *Lancet* 2008; 371(9608): 193–194.

137. Kapur S, Phillips A, Insel T. Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it?. *Molecular Psychiatry* 2012; 17: 1174-1179.

138. Karyo R, Eskira Y, Pinhasov A, Belmaker R, Agam G, Eldar-Finkelman H. Identification of eukaryotic elongation factor-2 as a novel cellular target of lithium and glycogen synthase kinase-3. *Molecular and Cellular Neuroscience* 2010; 45: 449-455.

139. Kay SR, Fiszbein A, Opler LA. The Positive and Negative Syndrome Scale (PANSS) for schizophrenia. Schizophrenia Bulletin 1987*;* 13: 261-76.

140. Kayala M, Baldi P. Cyber-T web server: differential analysis of high-throughput data. *Nucleic Acids Research* 2012; 40: W553-W559.

141. Kendler, K.S. & Gardner, C.O. The risk for psychiatric disorders in relatives of schizo- phrenic and control probands: a comparison of three independent studies. *Psychol. Med.* **27,** 411–419 (1997).

142. Kessler RC, Andrews G, Colpe LJ, Hiripi E, Mroczek DK, Normand SL et al. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychol Med* 2002; 32: 959-76.

143. Kim J, Ghasemzadeh N, Eapen DJ, Chung NC, Storey JD et al. Gene expression profiles associated with acute myocardial infarction and risk of cardiovascular death. *Genome Med*. 2014; 6:40.

144. King DW, Leskin GA, King LA ,Weathers FW. Confirmatory factor analysis of the Clinician-Administered PTSD Scale: Evidence for the dimensionality of posttraumatic stress disorder. *Psychol. Assess.* 1998;10: 90-96.

145. Klein P, Melton D. A molecular mechanism for the effect of lithium on development. *Proceedings of the National Academy of Sciences* 1996; 93: 8455-8459.

146. Kleindienst N, Engel R, Greil W. Which clinical factors predict response to prophylactic lithium? A systematic review for bipolar disorders. *Bipolar Disorders* 2005; 7: 404-417.

147. Kobayashi H, Kruger R, Markopoulou K, Wszolek Z, Chase B, Taka H,Mineki R, Murayama K, Riess O, Mizuno Y, Hattori N. 2003. Haploinsufficiency at the alpha-synuclein gene underlies phenotypic severity in familial Parkinson's disease. *Brain* 126:32–42.

148. Koide R, Kobayashi S, Shimohata T, Ikeuchi T, Maruyama M, Saito M, Yamada M, Takahashi H, Tsuji S. 1999. A neurological disease caused by an expanded CAG trinucleotide repeat in the TAT-binding protein gene: a new polyglutamine disease? *Hum Mol Genet* 8:2047– 2053.

149. Konradi C, Eaton M, MacDonald ML, Walsh J, Benes FM, Heckers S. Molecular evidence for mitochondrial dysfunction in bipolar disorder. *Arch Gen Psychiatry* 2004; 61: 300–08.

150. Kubicki M, McCarley R, Westin CF, Park HJ, Maier S, Kikinis R et al. A review of diffusion tensor imaging studies in schizophrenia. *J Psychiatr Res* 2007; 41(1-2): 15–30.

151. Kurian SM, Le-Niculescu H, Patel SD, Bertram D, Davis J, Dike C et al. Identification of Blood Biomarkers for Psychosis Using Convergent Functional Genomics. *Molecular Psychiatry Mol Psychiatry* 2009; 16(1): 37-58.

152. Labonte B, Yerko V, Gross J, Mechawar N, Meaney MJ, Szyf M, et al. Differential glucocorticoid receptor exon 1(B), 1(C), and 1(H) expression and methylation in suicide completers with a history of childhood abuse. *Biol Psychiatry* 2012;72:41- 8.

153. Lam YA, Pickart CM, Alban A, Landon M, Jamieson C, Ramage R et al. Inhibition of the ubiquitin-proteasome system in Alzheimer's disease. *Proc Natl Acad Sci* 2000; 97: 9902–06.

154. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008; 9: 559.

155. Langfelder P, Luo R, Oldham M, Horvath S. Is My Network Module Preserved and Reproducible?. *PLoS Comput Biol* 2011; 7: e1001057.

156. Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* 2007; 24: 719-720.

157. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome*. Genome Biol* 2009; 10(3): R25.

158. Law C, Chen Y, Shi W, Smyth G. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* 2014; 15: R29.

159. Law CW, Chen Y, Shi W, Smyth GK. Voom Precision weights unlock linear model analysis tools for RNA-seq read counts. Technical report Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia; 2013 [http://www. statsci. org/smyth/pubs/13 5 1-voom-techreport]

160. Le Moal M, Koob GF. Drug addiction: Pathways to the disease and pathophysiological perspectives. *European Neuropsychopharmacology* 2007; 17: 377–93.

161. Le-Niculescu H, Balaraman Y, Patel S, Tan J, Sidhu K, Jerome RE et al. Towards understanding the schizophrenia code: an expanded convergent functional genomics approach. *Am J Med Genet B Neuropsychiatr Genet* 2007; 144B: 129–58.

162. Le-Niculescu H, Balaraman Y, Patel SD, Ayalew M, Gupta J, Kuczenski R et al. Convergent functional genomics of anxiety disorders: translational identification of genes, biomarkers, pathways and mechanisms. *Transl Psychiatr* 2011; 1:e9.

163. Le-Niculescu H, Levey DF, Ayalew M, Palmer L, Gavrin LM , Jain N et al. Discovery and Validation of Blood Biomarkers for Suicidality. *Molecular Psychiatry Mol Psychiatry* 2013; 18(12): 1249-64.

164. Le-Niculescu H, Patel SD, Bhat M, Kuczenski R, Faraone SV, Tsuang MT et al. Convergent functional genomics of genome-wide association data for bipolar disorder: comprehensive identification of candidate genes, pathways and mechanisms. *Am J Med Genet B Neuropsychiatr Genet* 2009; 150B: 155–81.

165. Lebrun AH, Storch S, Pohl S, Streichert T, Mole SE, Ullrich K, et al.

Identification of Potential Biomarkers and Modifiers of CLN3-disease Progression. *Neuropediatrics* 2010; 41(02): n. pag. Web.

166. Lee J, Goh L-K, Chen G, Verma S, Tan C-H, Lee T-S. Analysis of blood-based gene expression signature in first-episode psychosis. *Psychiatry Res* 2012; 200(1): 52–4.

167. Lenox R, Wang L. Molecular basis of lithium action: integration of lithium-responsive signaling and gene expression networks. *Molecular Psychiatry* 2003; 8: 135-144.

168. Lewitzka U, Severus E, Bauer R, Ritter P, Müller-Oerlinghausen B, Bauer M. The suicide prevention effect of lithium: more than 20 years of evidence—a narrative review. *Int J Bipolar Disord* 2015; 3. doi:10.1186/s40345-015-0032-2.

169. Li C, Wong W. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Sciences* 2001; 98: 31-36.

170. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N et al. The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* 2009; 25, 2078-79

171. Li H, Homer N. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics* 2010; 11: 473-483.

172. Li S, Tighe S, Nicolet C, Grove D, Levy S, Farmerie W et al. Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. *Nat Biotechnol* 2014; 32: 915-925.

173. Lichtenstein, P., Carlstrom, E., Rastam, M., Gillberg, C. & Anckarsater, H. The genetics of autism spectrum disorders and related neuropsychiatric disorders in childhood. *Am. J. Psychiatry* **167,** 1357–1363 (2010).

174. Liu M, Chen Y, Chen S, Hu C, Lin C, Chang Y et al. Epstein–Barr virus latent membrane protein 1 induces micronucleus formation, represses DNA repair and enhances sensitivity to DNA-damaging agents in human epithelial cells. *Oncogene* 2004; 23: 2531-2539.

175. Louveau A, Smirnov I, Keyes T, Eccles J, Rouhani S, Peske J et al. Structural and functional features of central nervous system lymphatic vessels. *Nature* 2015; 523: 337-341.

176. Lowthert L, Leffert J, Lin A, Umlauf S, Maloney K, Muralidharan A et al. Increased ratio of anti-apoptotic to pro-apoptotic Bcl2 gene-family members in lithium-responders one month after treatment initiation. *Biology of Mood & Anxiety Disorders* 2012; 2: 15.

177. M. First et al., Neuroimaging Markers of Psychiatric Disorders: Consensus Report of the APA Work Group," In Resource Documents,American Psychiatric Association, 2012, accessed January 21, 2014,http://www.psychiatry.org/learn/library--archives/resource-documents.

178. G. M. Macqueen, Will There Be a Role for Neuroimaging in Clinical Psychiatry? *Journal of Psychiatry & Neuroscience* 35, no. 5 (2010): 291-93.

179. Maes M, Kubera M, Leunis J. The gut-brain barrier in major depression: Intestinal mucosal dysfunction with an increased translocation of LPS from gram negative enterobacteria (leaky gut) plays a role in the inflammatory pathophysiology of depression. *Nueroendocrinology Letters* 2008; 29(1):117-124.

180. Manji H, Lenox R. Long-term action of lithium: A role for transcriptional and posttranscriptional factors regulated by protein kinase C. *Synapse* 1994; 16: 11-28.

181. Marioni J, Mason C, Mane S, Stephens M, Gilad Y. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research* 2008; 18: 1509-1517.

182. Mas C, Taske N, Deutsch S, Guipponi M, Thomas P, Covanis A, et al. Association of the connexin36 gene with juvenile myoclonic epilepsy. *J. Med. Genet*. 2004; 41:e93.

183. Masui T, Hashimoto R, Kusumi I, Suzuki K, Tanaka T, Nakagawa S et al. Lithium response and Val66Met polymorphism of the brain-derived neurotrophic factor gene in Japanese patients with bipolar disorder. *Psychiatric Genetics* 2006; 16: 49-50.

184. Mathers C, Fat DM, Boerma J. The Global Burden of Disease: 2004 Update. Geneva, Switzerland: World Health Organization 2008.

185. Mathivanan S, Periaswamy B, Gandhi T, Kandasamy K, Suresh S, Mohmood R et al. An evaluation of human protein-protein interaction data in the public domain. *BMC Bioinformatics* 2006; 7: S19.

186. Matić G, Milutinović DV, Nestorov J, Elaković I, Jovanović SM, Perišić T *et al.* Lymphocyte glucocorticoid receptor expression level and hormone-binding properties differ between war trauma-exposed men with and without PTSD. *Progress in Neuro-Psychopharmacology and Biological Psychiatry* 2013; 43: 238-245.

187. Mattson M, LaFerla F, Chan S, Leissring M, Shepel P, Geiger J. Calcium signaling in the ER: its role in neuronal plasticity and neurodegenerative disorders. *Trends in Neurosciences* 2000; 23: 222-229.

188. Maunder RG, Hunter JJ, Feinman SV. Interferon Treatment of Hepatitis C Associated With Symptoms of PTSD. *Psychosomatics* 1998; 39.5: 461-64.

189. Maxwell SE. The persistence of underpowered studies in psychological research: causes, consequences, and remedies. *Psychol Methods* 2004; 9: 147--163.

190. Mayberg, Helen S. Neuroimaging And Psychiatry: The Long Road From Bench To Bedside. *Hastings Center Report* 44.s2 (2014): S31-S36. Web.

191. McEachin R, Chen H, Sartor M, Saccone S, Keller B, Prossin A et al. A genetic network model of cellular responses to lithium treatment and cocaine abuse in bipolar disorder. *BMC Systems Biology* 2010; 4: 158.

192. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 2010; 20: 1297-1303.

193. Mehra VC, Ramgolam VS, Bender JR. Cytokines and Cardiovascular Disease. *Journal of Leukocyte Biology* 2005; 78(4):805-18.

194. Mehta D, Gonik M, Klengel T, Rex-Haffner M, Menke A, Rubel J *et al.* Using Polymorphisms in FKBP5 to Define Biologically Distinct Subtypes of Posttraumatic Stress Disorder: Evidence From Endocrine and Gene Expression Studies. *Archives of General Psychiatry* 2011; 68.9: 901-910.

195. Merikangas K, Jin R, He J, Kessler R, Lee S, Sampson N et al. Prevalence and Correlates of Bipolar Spectrum Disorder in the World Mental Health Survey Initiative. *Arch Gen Psychiatry* 2011; 68: 241.

196. Mering C. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Research* 2003; 31: 258-261.

197. Middleton FA, Mirnics K, Pierri JN, Lewis DA, Levitt P. Gene expression profiling reveals alterations of specific metabolic pathways in schizophrenia. *J Neurosci* 2002; 22: 2718–29.

198. Miller J, Woltjer RL, Goodenbour JF, Horvath S, Geschwind DH. Genes and pathways underlying regional and cell type changes in Alzheimer's disease. *Genome Medicine* 2013; 5:48.

199. Miller J. What is the probability of replicating a statistically significant effect? *Psychon Bull Rev* 2009;16: 617--640.

200. Miller JA, Cai C, Langfelder P, Geschwind DH, Kurian SM, Salomon DR, Horvath S. Strategies for aggregating gene expression data: the collapse Rows R function. *BMC Bioinformatics* 2011;12:322.

201. Min J, Barrett A, Watts T, Pettersson F, Lockstone H, Lindgren C et al. Variability of gene expression profiles in human blood and lymphoblastoid cell lines. *BMC Genomics* 2010; 11: 96.

202. Minagar A, Alexander JS. Blood-brain Barrier Disruption in Multiple Sclerosis. *Multiple Sclerosis* 2003; 9.6: 540-49.

203. Ming X, Mulvey M, Mohanty S, Patel V. Safety and efficacy of clonidine and clonidine extended-release in the treatment of children and adolescents with attention deficit and hyperactivity disorders. *AHMT* 2011; : 105.

204. Miovic M, Block S. Psychiatric disorders in advanced cancer. *Cancer* 2007; 110: 1665-1676.

205. Mirnics K, Levitt P, Lewis D. Critical Appraisal of DNA Microarrays in Psychiatric Genomics. *Biological Psychiatry* 2006; 60: 163-176.

206. Mitchison HM, Taschner PEM, O'Rawe AM, De Vos N, Phillips HA, Thompson AD et al. Genetic Mapping of the Batten Disease Locus (CLN3) to the Interval D16S288-D16S383 by Analysis of Haplotypes and Allelic Association. *Genomics* 1994; 22(2): 465-68.

207. Mondelli V, Ciufonlini S, Murri MB, Bonaccorso S, Di Fortio M, Giordano A et al. Cortisol and Inflammatory Biomarkers Predict Poor Treatment Response in First Episode Psychosis. *Schizophr Bull* 2015; 41(5): 1162-70.

208. Monti JM, BaHammam AS, Pandi-Perumal SR, Bromundt V, Spence DW, Cardinali DP et al. Sleep and Circadian Rhythm Dysregulation in Schizophrenia. *Progress in Neuro-Psychopharmacology and Biological Psychiatry* 2013; (43): 209-16.

209. Mortazavi A, Williams B, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* 2008; 5: 621-628.

210. Moussavi S, Chatterji S, Verdes E, Tandon A, Patel V, Ustun B. Depression, chronic diseases, and decrements in health: results from the World Health Surveys. *The Lancet* 2007; 370: 851-858.

211. Mueller C, Müller B, Perruchoud AP. Biomarkers: past, present, and future. *Swiss Med Wkly* 2008;138:225–229.

212. Murray C, Vos T, Lozano R, Naghavi M, Flaxman A, Michaud C et al. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990,Äì2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet* 2012; 380: 2197-2223.

213. Nardinocchi L, Puca R, Sacchi A, D'orazi G. Inhibition of HIF-1alpha Activity by Homeodomain-interacting Protein Kinase-2 Correlates with Sensitization of Chemoresistant Cells to Undergo Apoptosis. *Molecular Cancer* 2009; 8(1).

214. Neylan TC, Sun B, Rempel H, Ross J, Lenoci M, O'Donovan A *et al*. Suppressed monocyte gene expression profile in men versus women with PTSD. *Brain, Behavior, and Immunity* 2011; 25.3: 524-531.

215. Niculescu A, Segal D, Kuczenski R, Barrett T, Hauger R, Kelsoe J. Identifying a series of candidate genes for mania and psychosis: a convergent functional genomics approach. *Physiol Genomics* 2000; 4: 83–91.

216. Nilsen T, Graveley B. Expansion of the eukaryotic proteome by alternative splicing. *Nature* 2010; 463: 457-463.

217. Nutt DJ, King LA, Phillips LD. Drug harms in the UK: a multicriteria decision analysis. *Lancet* 2010; 376: 1558–65.

218. Ogden CA, Rich ME, Schork NJ, Paulus MP, Geyer MA, Lohr JB et al. Candidate genes, pathways and mechanisms for bipolar (manic-depressive) and related disorders: an expanded convergent functional genomics approach. *Mol Psychiatr* 2004; 9: 1007–29.

219. Olsen L, Hansen T, Jakobsen KD, Djurovic S, Melle I, Agartz I et al. The Estrogen Hypothesis of Schizophrenia Implicates Glucose Metabolism: Association Study in Three Independent Samples. *BMC Medical Genetics* 2008; 9: 39.

220. Olson N. The microarray data analysis process: From raw data to biological significance. *NeuroRX* 2006; 3: 373-383.

221. Orikabe L, Yamasue H, Inoue H, Takayanagi Y, Mozue Y, Sudo Y et al. Reduced Amygdala and Hippocampal Volumes in Patients with Methamphetamine Psychosis. *Schizophrenia Research* 2011; 132(2-3): 183-89.

222. Overstreet DH. Modeling depression in animal models. *Methods Mol Biol.* 2012; 829: 125-44.

223. Pacak, K. Stressor Specificity of Central Neuroendocrine Responses: Implications for Stress-Related Disorders. *Endocrine Reviews* 2001; 22.4: 502-48.

224. Pace TW, Wingenfeld K, Schmidt I, Meinlschmidt G, Hellhammer DH, Heim CM. Increased peripheral NF-KB pathway activity in women with childhood abuse-related posttraumatic stress disorder. *Brain, Behavior, and Immunity* 2012; 26.1: 13-17.

225. Paschen W. Endoplasmic reticulum: a primary target in various acute disorders

and degenerative diseases of the brain. *Cell Calcium* 2003; 34: 365-383.

226. Patel R, Jain M. NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data. *PLoS ONE* 2012; 7: e30619.

227. Patel SD, Le-Niculescu H, Koller DL, Green SD, Lahiri DK, McMahon FJ et al. Coming to grips with complex disorders: genetic risk prediction in bipolar disorder using panels of genes identified through convergent functional genomics. *Am J Med Genet B Neuropsychiatr Genet* 2010; 153B: 850–77.

228. Pawitan Y, Michiels S, Koscielny S, Gusnanto A, Ploner A. False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics* 2005; 21: 3017-3024.

229. Pearson W, Wood T, Zhang Z, Miller W. Comparison of DNA Sequences with Protein Sequences. *Genomics* 1997; 46: 24-36.

230. Perlis R, Smoller J, Ferreira M, McQuillin A, Bass N, Lawrence J et al. A Genomewide Association Study of Response to Lithium for Prevention of Recurrence in Bipolar Disorder. *American Journal of Psychiatry* 2009; 166: 718-725.

231. Perroud N, Paoloni-Giacobino A, Prada P, Olie´ E, Salzmann A, Nicastro R, et al. Increased methylation of glucocorticoid receptor gen (NR3C1) in adults with a history of childhood maltreatment: a link with the severity and type of trauma. *Transl Psychiatry* 2011;1:e59.

232. Pialoux V, Mounier R, Brown AD, Steinback CD, Rawling JM et al. Relationship between oxidative stress and HIF-1_ mRNA during sustained hypoxia in humans. *Free Radical Biology and Medicine* 2009;46(2):321-326.

233. Pickrell J, Marioni J, Pai A, Degner J, Engelhardt B, Nkadori E et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 2010; 464: 768-772.

234. Pienaar I, Daniels W, Götz J. Neuroproteomics as a promising tool in Parkinson’s disease research. *Journal of Neural Transmission* 2008; 115: 1413-1430.

235. Piñero J, Queralt-Rosinach N, Bravo A, Deu-Pons J, Bauer-Mehren A, Baron M et al. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database* 2015; Vol. 2015: article ID bav028; doi:10.1093/database/bav028

236. Piro RM, Molineris I, Ala U, Di Cunto F. Evaluation of Candidate Genes from Orphan FEB and GEFS Loci by Analysis of Human Brain Gene Expression Atlases. Ed. Takeo Yoshikawa. PLoS ONE 2011; 6(8):e23149.

237. Prasad DK, Satyanarayana U, Munshi A. Genetics of Idiopathic Generalized Epilepsy: An Overview. *Neurology India* 2013; 61(6): 572.

238. Price G, Cercignani M, Parker GJ, Altmann DR, Barnes TR, Barker GJ et al. Abnormal Brain Connectivity in First-episode Psychosis: A Diffusion MRI Tractography Study of the Corpus Callosum. *NeuroImage* 2007; 35(2): 458-66.

239. Prince M, Patel V, Saxena S, Maj M, Maselko J, Phillips M et al. No health without mental health. *The Lancet* 2007; 370: 859-877.

240. Psychiatric GWAS Consortium. A framework for interpreting genomewide association studies of psychiatric disorders. *Mol Psychiatry* 2009; 14: 10–7.

241. Qiu MH, Liu W, Qu WM, Urade Y, Lu J, Huang ZL. The role of nucleus accumbens core/shell in sleep-wake regulation and their involvement in modafinil-induced arousal. *PLoS One* 2012; 7(9):e45471.

242. Ramchand R, Schell TL, Karney BR, Osilla KM, Burns RM, Caldarone LB. Disparate prevalence estimates of PTSD among service members who served in Iraq and Afghanistan: possible explanations. *J. Trauma. Stress* 2010; 23: 59–68.

243. Raslova H, Kauffmann A, Sekkai D, Ripoche H, Larbret F, Robert T *et al.* Interrelation between Polyploidization and Megakaryocyte Differentiation: A Gene Profiling Approach. *Blood* 2007;109.8: 3225-234.

244. Ratnaike, RN. Acute and chronic arsenic toxicity *Postgrad Med J 2003; 79: 391-96.*

245. Raz T, Kapranov P, Lipson D, Letovsky S, Milos P, Thompson J. Protocol Dependence of Sequencing-Based Gene Expression Measurements. *PLoS ONE* 2011; 6: e19287.

246. Reichardt HM, Tuckermann JP, Göttlicher M et al. Repression of inflammatory responses in the absence of DNA binding by the glucocorticoid receptor. *EMBO Journal* 2002; 20(24):7168-7173.

247. Ritchie M, Phipson B, Wu D, Hu Y, Law C, Shi W et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* 2015; 43: e47-e47.

248. Robinson M, McCarthy D, Smyth G. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2009; 26: 139-140.

249. Robinson M, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* 2010; 11: R25.

250. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010; 26:139-140.

251. Rodd ZA, Bertsch BA, Strother WN, Le-Niculescu H, Balaraman Y, Hayden E et al. Candidate genes, pathways and mechanisms for alcoholism: an expanded convergent functional genomics approach. *Pharmacogenomics J* 2007; 7: 222–56.

252. Rothpearl AB, Mohs RC, Davis KL. Statistical power in biological psychiatry. *Psychiatry Res* 1981; 5: 257--266.

253. Rui Tian T,Hou G,Li D, Yuan TF. A Possible Change Process of Inflammatory Cytokines in the Prolonged Chronic Stress and Its Ultimate Implications for Health. *The Scientific World Journal Article* 2014; 1155:780616.

254. Rusinova I, Forster S, Yu S, Kannan A, Masse M, Cumming H *et al.* INTERFEROME V2.0: An Updated Database of Annotated Interferon-regulated Genes. *Nucleic Acids Research* 2012; 41.D1: D1040-1046.

255. Russell K. Biomarkers and surrogate markers: an FDA perspective. *NeuroRx* 2004; 1:189–195.

256. Rybakowski J, Suwalska A, Skibinska M, Szczepankiewicz A, Leszczynska-Rodziewicz A, Permoda A et al. Prophylactic Lithium Response and Polymorphism of the Brain-Derived Neurotrophic Factor Gene. *Pharmacopsychiatry* 2005; 38: 166-170.

257. Samad Z, Boyle S, Ersboll M, Vora AN, Zhang Y et al. Sex Differences in Platelet Reactivity and Cardiovascular and Psychological Response to Mental Stress in Patients With Stable Ischemic Heart Disease: Insights From the REMIT Study. *J Am Coll Cardiol.* 2014; 64(16):1669-1678.

258. Sarapas C, Cai G, Bierer LM, Golier JA, Galea S, Ising M *et al.* Genetic Markers for PTSD Risk and Resilience Among Survivors of the World Trade Center Attacks. *Disease Markers* 2011; 30.2-3: 101-110.

259. Schedlowski M, Jacobs R, Stratmann G, Richter S, Hadicke A et al. Changes of natural killer cells during acute psychological stress. *J Clin Immunol* 1993; 13(2):119-126.

260. Schenk VW, Stolk PJ. Psychosis following arsenic (possibly thalium) poisoning. *Psychiatr Neurol Neurochir 1967; 70: 31–7.*

261. Schunemann HJ, Oxman AD, Brozek J, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ* 2008;336:1106–1110.

262. Segerstrom SC, Miller GE. Psychological stress and the human immune system: a meta-analytic study of 30 years of inquiry. *Psychol. Bull* 2004; 130: 1-37.

263. Semba J, Watanabe H, Suhara T, Akanuma N. Chronic lithium chloride injection increases glucocorticoid receptor but not mineralocorticoid receptor mRNA expression in rat brain. *Neuroscience Research* 2000; 38: 313-319.

264. Serretti A, Lilli R, Mandelli L, Lorenzi C, Smeraldi E. Serotonin transporter gene associated with lithium prophylaxis in mood disorders. *Pharmacogenomics J* 2001; 1: 71-77.

265. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT et al. Cytoscape: a software environment for integrated modules of biomolecular interaction networks. *Genome Research* 2003; 13(11):2498-504.

266. Sherin JE, Nemeroff CB. Post-traumatic stress disorder: the neurobiological impact of psychological trauma. *Dialogues in Clinical Neuroscience* 2011; 13(3):263-278.

267. Sherman M, Goldberg A. Cellular Defenses against Unfolded Proteins. *Neuron* 2001; 29: 15-32.

268. Shim S, Nam H, Lee J, Kim J, Han G, Jeon J. MicroRNAs in human lymphobastoid cell lines. Crit Rev Eukayoti Gene Expr. 2012; 22(3); 189-96.

269. Shimura H, Schlossmacher MG, Hattori N, Frosch MP, Trockenbacher A, Schneider R et al. Ubiquitination of a new form of alpha-synuclein by Parkin from human brain: Implications for Parkinson's disease. *Science* 2001; 293: 263–69.

270. Shoemaker JE, Tiago L, Ghosh S, Matsuoka Y, Kawaoka Y, Kitano H. CTen: A Web-based Platform for Identifying Enriched Cell Types from Heterogeneous Microarray Data. *BMC Genomics* 2012; 13.1: 460.

271. Sie L, Loong S, Tan E. Utility of lymphoblastoid cell lines. *J Neurosci Res* 2009; 87: 1953-1959.

272. Silver J, Ritchie M, Smyth G. Microarray background correction: maximum likelihood estimation for the normal-exponential convolution. *Biostatistics* 2008; 10: 352-363.

273. Simon R, Lam A, Li M-C, Ngan M, Menenzes S, Zhao Y. Analysis of gene expression data using BRB-Array tools. *Cancer Inform* 2007; 2:11–17

274. Sivakumar T, Dalal P. Moving towards ICD-11 and DSM-V: Concept and evolution of psychiatric classification. *Indian Journal of Psychiatry* 2009; 51: 310.

275. Smith MJ, Thirthalli J, Abdallah AB, Murray RM, Cottler LB. Prevalence of psychotic symptoms in substance users: a comparison across substances. *Compr Psychiatry* 2009; 50(3): 245–50.

276. Smyth GK. Limma: linear models for microarray data. In Gentleman R, Carey V, Dudoit S, Irizarry R and Huber W Eds., Bioinformatics and Computational Biology Solutions Using R and Bioconductor (Springer, New York) 2005, pp. 397-420.

277. Smyth GK. Limma: linear models for microarray data. In Gentleman R, Carey V, Dudoit S, Irizarry R and Huber W Eds., Bioinformatics and Computational Biology Solutions Using R and Bioconductor (Springer, New York), 2005, pp. 397-420.

278. Sorel E. *21st century global mental health*. Jones & Bartlett Learning: Boston, 2013.

279. Srisurapanont M, Ali R, Marsden J, Sunga A, Wada K, Monteiro M. Psychotic symptoms in methamphetamine psychotic in-patients. *Int J Neuropsychopharmacol* 2003; 6(4):347–52.

280. Srisurapanont M, Arunpongpaisal S, Wada K, Marsden J, Ali R, Kongsakon R. Comparisons of methamphetamine psychotic and schizophrenic symptoms: a differential item functioning analysis. *Prog Neuro-Psychopharmacol Biol Psychiatry* 2011; 35: 959–964.

281. Stilling R, Ronicke R, Benito E, Urbanke H, Capece V, Burkhardt S et al. K-Lysine acetyltransferase 2a regulates a hippocampal gene expression network linked to memory formation. *The EMBO Journal* 2014; 33: 1912-1927.

282. Subramanian A, Tamayo P, Mootha V, Mukherjee S, Ebert B, Gillette M et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* 2005; 102: 15545-15550.

283. Sugawara H, Iwamoto K, Bundo M, Ishiwata M, Ueda J, Kakiuchi C et al. Effect of mood stabilizers on gene expression in lymphoblastoid cells. *J Neural Transm* 2010; 117: 155–164.

284. Sunderland T, Gur RE, Arnold SE. The use of biomarkers in the elderly: current and future challenges. *Biol Psychiatry* 2005; 58:272–276.

285. Takahashi M, Hayashi H, Watanabe Y, Sawamura K, Fukui N, Watanabe J et al. Diagnostic classification of schizophrenia by neural network analysis of blood-based gene expression signatures. *Schizophrenia Research* 2010; 119: 210-218.

286. Testa SM, Krauss GL, Lesser RP, Brandt J. Stressful Life Event Appraisal and Coping in Patients with Psychogenic Seizures and Those with Epilepsy. *Seizure* 2012; 21(4):282-87.

287. The National Institute of Mental Health Strategic Plan. NIMH, 2007, p. 37. http://www.nimh.nih.gov/about/strategic- planning-reports/index.shtml

288. Torres TEP and Lotfi CFP. Distribution of cells expressing Jun and Fos proteins and synthesizing DNA in the adrenal cortex of hypophysectomized rats: regulation by ACTH and FGF2. *Cell and Tissue Research* 2007; 329(3): 443-455.

289. Trapnell C, Pachter L, Salzberg S. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 2009; 25: 1105-1111.

290. Trapnell C, Williams B, Pertea G, Mortazavi A, Kwan G, van Baren M et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 2010; 28: 511-515.

291. Tsuang M, Nossova N, Yager T, Tsuang M, Guo S, Shyu K et al. Assessing the validity of blood-based gene expression profiles for the classification of schizophrenia and bipolar disorder: A preliminary report. *Am J Med Genet* 2005; 133B: 1-5.

292. Turecki G, Grof P, Cavazzoni P, Duffy A, Grof E, Ahrens B et al. Evidence for a role of phospholipase C-γ1 in the pathogenesis of bipolar disorder. *Molecular Psychiatry* 1998; 3: 534-538.

293. Tylee D, Kawaguchi D, Glatt S. On the outside, looking in: A review and evaluation of the comparability of blood and brain ‚Äú-omes‚Äù. *Am J Med Genet* 2013; 162: 595-603.

294. Tyrer P. A comparison of DSM and ICD classifications of mental disorder. *Advances in Psychiatric Treatment* 2014; 20: 280-285.

295. Tyrka AR, Price LH, Marsit C, Walters OC, Carpenter LL. Childhood adversity and epigenetic modulation of the leukocyte glucocorticoid receptor: preliminary findings in healthy adults. *PLoS One* 2012;7:e30148

296. United Nations Office on Drugs and Crime (2004) World Drug Report 2004. Vienna: UN Office on Drugs and Crime.

297. van der Kouwe AJ, Benner T, Salat DH, Fischl B. Brain Morphometry with Multiecho MPRAGE. *NeuroImage* 2008; 40(2): 559-69.

298. van Venrooij JA, Fluitman SB, Lijmer JG, Kavelaars A, Heijnen CJ, Westenberg HG et al. Impaired neuroendocrine and immune response to acute stress in medication-naive patients with a first episode of psychosis. *Schizophr Bull* 2012; 38(2): 272–79.

299. van Westerloo DJ, Choi G, Lowenberg EC, Truijen J, de Vos AF, et al. Acute stress elicited by bungee jumping suppresses human innate immunity. *Mol. Med.* 2001; 17, 180-188

300. van Zuiden M, Geuze E, Willemen HL, Vermetten E, Maas M, Amarouchi K *et al.* Glucocorticoid Receptor Pathway Components Predict Posttraumatic Stress Disorder Symptom Development: A Prospective Study. *Biological Psychiatry* 2012: 71.4; 309-316.

301. van Zuiden M, Heijnen CJ, Maas M, Amarouchi K, Vermetten E, Geuze E *et al.* Glucocorticoid sensitivity of leukocytes predicts PTSD, depressive and fatigue symptoms after military deployment: A prospective study. *Psychoneuroendocrinology* 2012; 37.11: 1822-1836.

302. Vance JE, Steenbergen R. Metabolism and functions of phophatidylserine. *Prog Lipid Res.* 1994; 44: 207-34.

303. Vapnik V, Lerner A. Pattern recognition using generalized portrait method. *Automation and Remote Control* 1963; 24: 774–780.

304. Vawter MP, Barrett T, Cheadle C, Sokolov BP, Wood WH III, Donovan DM et al. Application of cDNA microarrays to examine gene expression differences in schizophrenia. *Brain Res Bull* 2001; 55: 641–50.

305. Vawter MP, Crook JM, Hyde TM, Kleinman JE, Weinberger DR, Becker KG et al. Microarray analysis of gene expression in the prefrontal cortex in schizophrenia: A preliminary study. *Schizophr Res* 2002; 58: 11–20.

306. Velakoulis D, Wood SJ, Wong MT, McGorry PD, Yung A, Phillips L et al. Hippocampal and amygdala volumes according to psychosis stage and diagnosis: a magnetic resonance imaging study of chronic schizophrenia, first-episode psychosis, and ultra-high-risk individuals. *Arch Gen Psychiatry* 2006; 63(2): 139-49.

307. Veldhuis TB, Floris T, van der Meide PH, Vos IM, de Vries HE, Dijkstra CD et al. Interferon-beta prevents cytokine-induced neutrophil infiltration and attenuates blood–brain barrier disruption. *J. Cerebral Blood Flow Metab.* 2003; 23;1060–1069.

308. Verstrepen L, Bekaert T, Chau LT, Tavernier J, Chariot A, Beyaert R. TLR-4, IL-1R and TNF-R signaling to NF-_B: variations on a common theme. *Cellular and Molecular Life Sciences* 2008; 65(19): 2264-2978.

309. Visscher P, Brown M, McCarthy M, Yang J. Five Years of GWAS Discovery. *The American Journal of Human Genetics* 2012; 90: 7-24.

310. Wagner J, Merck A. Conference on Biomarkers Discovery and Validation, Oct. 14–18, 2004 http://bigdaddy.scripps.edu/darlene/Asilomar/pages/abstracts/jwagner.htm

311. Walburn J, Vedhara K, Hankins M, Rixon L, Weinman J. Psychological stress and wound healing in humans: a systematic review and meta-analysis. *J Psychosom Res* 2009; 67(3): 253–71.

312. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009; 10: 57-63.

313. Watanabe S, Iga J, Nishi A, Numata S, Kinoshita M, Kikuchi K et al. Microarray analysis of global gene expression in leukocytes following lithium treatment. *Human Psychopharmacology: Clinical and Experimental* 2014; 29: 190-198.

314. Watkins NA, Gusnanto A, De Bono B, De S, Miranda-Saavedra et al. A HaemAtlas: Characterizing Gene Expression in Differentiated Human Blood Cells. *Blood* 2009;113.19: E1-E9.

315. Weathers FW, Keane TM, Davidson JR Clinician-administered PTSD scale: a review of the first ten years of research. *Depress. Anxiety* 2001;13:132-56.

316. Weathers FW., Ruscio AM, Keane TM. Psychometric properties of nine scoring rules for the clinician-administered posttraumatic stress disorder scale. *Psychol. Assess.* 1999;11:124-133.

317. Weaver IC, Cervoni N, Champagne FA, et al. Epigenetic programmingby maternal behavior. *Nat Neurosci*. 2004;7:847-854.

318. Webster JI, Tonelli L, and Sternberg EM. Neuroendocrine regulation of immunity. *Annual Review of Immunology* 2002; 20: 125-163.

319. Wheeler H, Dolan M. Lymphoblastoid cell lines in pharmacogenomic discovery and clinical translation. *Pharmacogenomics* 2012; 13: 55-70.

320. Whitacre CC. A gender gap in autoimmunity. *Science* 1999; 283:1277-1278.

321. White T, Nelson M, Lim KO. Diffusion tensor imaging in psychiatric disorders. *Top Magn Reson Imaging* 2008; 19: 97–109.

322. Whitford TJ, Kubicki M, Schneiderman JS, O'Donnell LJ, King R, Alvarado JL et al. Corpus Callosum Abnormalities and Their Association with Psychotic Symptoms in Patients with Schizophrenia. *Biological Psychiatry* 2010; 68(1): 70-7.

323. World Health Organization (1992) *ICD-10: Classification of Mental and Behavioural Disorders.* WHO.

324. Yaari G, Bolen C, Thakar J, Kleinstein S. Quantitative set analysis for gene expression: a method to quantify gene set differential expression including gene-gene correlations. *Nucleic Acids Research* 2013; 41: e170-e170.

325. Yaari G, Bolen C, Thakar J, Kleinstein S. Quantitative set analysis for gene expression: a method to quantify gene set differential expression including gene-gene correlations. *Nucleic Acids Research* 2013; 41: e170-e170.

326. Yang MH, Jung MS, Lee MJ, Yoo KH, Yook YJ, Park EY et al. Gene expression profiling of the rewarding effect caused by methamphetamine in the mesolimbic dopamine system. *Mol Cells* 2008; 26(2): 121-30.

327. Yang T, Gilbert DL, Glauser TA, Hershey AD, Sharp FR. Blood Gene Expression Profiling of Neurologic Diseases. *Archives of Neurology Arch Neurol* 2005; 62(2): 210.

328. Yang T, Lu A, Aronow BJ, Sharp FR. Blood Genomic Responses Differ after Stroke, Seizures, Hypoglycemia, and Hypoxia: Blood Genomic Fingerprints of Disease. *Annals of Neurology Ann Neurol*. 2001; 50(6):699-707.

329. Yao Y, Schr√∂der J, Karlsson H. Verification of proposed peripheral biomarkers in mononuclear cells of individuals with schizophrenia. *Journal of Psychiatric Research* 2008; 42: 639-643.

330. Yehuda R, Holsboer F, Buxbaum JD, Miller-Myhsok B, Schmeidler J, Rein T *et al.* Gene Expression Patterns Associated with Posttraumatic Stress Disorder Following Exposure to the World Trade Center Attacks. *Biological Psychiatry* 2009; 66.7: 708-711.

331. Yeste M, Alvira D, Verdaguer E, Tajes M, Folch J, Rimbau V et al. Evaluation of acute antiapoptotic effects of Li+ in neuronal cell cultures. *Journal of Neural Transmission* 2006; 114: 405-416.

332. Yoo M, Shin J, Kim J, Ryall K, Lee K, Lee S et al. DSigDB: drug signatures database for gene set analysis. *Bioinformatics* 2015; 31: 3069-3071.

333. Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* 2010; 26: 976-978.

334. Zarbock A, Polanowska-Grabowska RK, Ley K. Platelet-neutrophil-interactions: Linking Hemostasis and Inflammation. *Blood Reviews* 2007; 21.2: 99-111.

335. Zhang W, Jüllig M, Connolly A, Stott N. Early gene response in lithium chloride induced apoptosis. *Apoptosis* 2005; 10: 75-90.

336. Zhong H, Simons JW. Direct Comparison of GAPDH, Beta-Actin, Cyclophilin, and 28S rRNA as Internal Standards for Quantifying RNA Levels under Hypoxia. *Biochemical and Biophysical Research Communications* 1999; 259(3):523-26.

337. Zhulidov P, Bogdanova E, Shcheglov A, Shagina I, Wagner L, Khazpekov G et al. A method for the preparation of normalized cDNA libraries enriched with full-length sequences. *Russian Journal of Bioorganic Chemistry* 2005; 31: 170-177.

338. Zhulidov P. Simple cDNA normalization using kamchatka crab duplex-specific nuclease. *Nucleic Acids Research* 2004; 32: 37e-37.

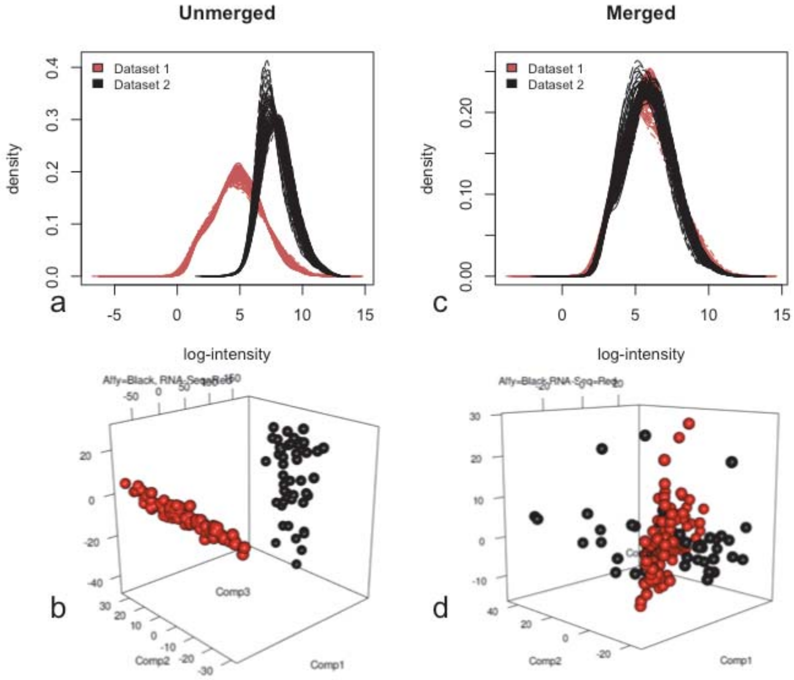
339. Zieker J, Zieker D, Jatzko A, Dietzsch J, Niesel K, Schmitt A *et al.* Differential gene expression in peripheral blood of patients suffering from post-traumatic stress disorder. *Molecular Psychiatry* 2007*;* 12.2: 116-118.

340. Zubenko G, Cohen B, Lipinski J, Jonas J. Use of clonidine in treating neuroleptic-induced akathisia. *Psychiatry Research* 1984; 13: 253-259.

341. Zuckerman M. General and situation-specific traits and states: new approaches to assessment of anxiety and other constructs. In: Zuckerman M, Spielberger CD (eds), Emotion and anxiety: new concepts, methods, and applications. Erlbaum, Hillsdale, NJ; 1976, pp. 133–174.

Appendix A Supplementary Data

A.1. Supplementary Data for Chapter 5



Supplementary Figure 5.1. Combat adjustment controlling for technical variation in Dataset 1 (red) and Dataset 2 (black). Filtered and normalized data are visualized by boxplot and histogram both (**a** & **b**) before and (**c** & **d**) after the adjustment and merging of data distributions.

Supplementary Table 5.1. Characterization of Pre-Deployment Modules in Dataset 1

Characterisation of Pre-Deployment Modules in Dataset 1

		Top Significant	Top Significant	Top Significant	Significant ME Correlations	
Module	Genes (*n*):	Biological Process	Pathway	Cell-Type	Condition or Trait	(r-value, p-value)
1	488	Intracellular Signalling Cascade	Signalling by Rho GTPase	CD14+ Monocytes	Caucasian Eth. CAPSBs	(0.22, 0.03) (-0.22, 0.04)
M2A	245 (*177)	Response to Virus	Interferon Signalling	CD14+ Monocytes	PTSD risk	(0.21, 0.002)
3	555	Transcription	mRNA 3'-end processing	CD8+ T cells	Caucasian	(-0.24, 0.02)
4	132	B cell activation	Collagen biosynthesis	CD19+ B cells	-	-
5	65	-	-	CD8+ T cells	-	-
6	51	Defense Response to Bacterium	Activation of Matrix Metalloproteinase	-	Audit Combined BMI WC adj Asian Eth.	(0.29, 0.005) (0.25, 0.01) (0.26,.01) (-0.22, 0.03)
7	70	-	-	-	-	-
8	58	-	RAF Activation	CD33+ Myeloid	-	-
9	277	Regulation of Natural Killer Cell Mediated Cytotoxicity	Natural Killer cell mediated cytotoxicity	CD56+ NK cells	Tobacco Caucasian Eth. American Mexican Eth. CAPsBs	(-0.25, 0.01) (-0.34, 8E-04) (0.28,0.007) (0.22, 0.04)
10	238	Negative regulation of cellular protein metabolic process	Regulation of Apoptosis	721 B lymphoblasts	CAPsBs	(0.29, 0.004)
11	92	-	-	-	Audit Combined	(0.23, 0.03)
12	187	RNA processing	-	-	-	-
13	73	M Phase	Cell Cycle	(Blood Platelets) CD71+ Early Eythroid	Caucasian Eth. African American Eth. CAPsBs	(-0.22, 0.03) (0.2,0.05) (0.26,0.01)
14	163	-	-	-	CAPs_tots CAPsBs CAPsDs	(-0.2, 0.0) (-0.22, 0.04) (-0.22, 0.03)
15	163	Transcription	Activation of AP-1 family of transcription factors	-	American Mexican Eth.	(0.2, 0.05)
16	156	Blood Coagulation	Platelet Degranulation	(Blood Platelets) CD71+ Early Eythroid	-	-
17	270	Translation Elongation	Eukaryotic Translation Termination	CD4+ T cells	African American Eth.	(0.21, 0.05)
18	139	NA	-	CD4+ T cells	-	-
19	145	Protein Transport	Signalling by TGF-beta Receptor Complex in Cancer	-	CAPsBs	(-0.24, 0.02)
20	741	Intracellular Signalling Cascade	Signalling by Interleukins	CD33+ Myeloid	-	-
21	334	Endocytosis	MHC Class II antigen presentation	CD14+ Monocytes	Caucasian Eth. African American Eth.	(0.22, 0.03) (-0.22, 0.04)

Supplementary Table 5.2. Characterization of Post-Deployment Modules in Dataset 2

Characterisation of Post-Deployment Modules Dataset 2

		Top Significant	Top Significant	Top Significant	Significant ME Correlations	
Module	Genes (*n*):	Biological Process	Pathway	Cell-Type	Condition or Trait	(r-value, p-value)
1	187	Cell adhesion	Collagen formation	-	Age CES BPE	(-0.28, 0.05) (-0.29, 0.05) (-0.56, 3E-05)
2	861	Protein transport	Membrane trafficking	CD33+ Myeloid	Age CES BPE	(0.42, 0.003) (0.34, 0.02) (0.54, 8E-05)
M3A	83 (*13)	Response to virus	Interferon signalling	CD14+ Monocytes	CAPs PCL BPE	(0.32, 0.03) (0.33, 0.02) (0.4, 0.004)
4	766	Intracellular signalling cascade	Attenuation phase	CD33+ Myeloid	BPE	(0.49, 5E-04)
5	294	Chromatin modification	-	CD8+ T cells	Age CES BPE	(0.28, 0.05) (0.29, 0.05) (0.57, 2E-05)
6	230	regulation of gene expression	--	CD4+ T cells	Age CES BPE	(0.42, 0.003) (0.34, 0.02) (0.54, 2E-05)
7	2243	rNA processing	Mitochondrial translation initiation	721 B lymphoblast	Age CES BPE	(0.36, 0.01) (0.32, 0.02) (0.57, 3E-05)
8	362	Cell adhesion	Degradation of the extracellular matrix	-	Age BPE	(-0.4, 0.005) (-0.45, 0.001)

Supplementary Table 5.3. Characterization of Pre-Deployment Modules in Dataset 2

Characterisation Pre-Deployment Modules Dataset 2

		Top Significant	Top Significant	Top Significant	Significant ME Correlations	
Module	Genes (*n*):	Biological Process	Pathway	Cell-Type	Condition or Trait	(r-value, p-value)
1	658	RNA processing	Gene expression (generic)	CD8+ T Cells	CAPS	(0.3, 0.04)
2	800	Protein transport	-	CD56+ NK Cells	CAPS PCL	(0.32, 0.03) (0.36, 0.01)
3	209	Cell adhesion	Collagen degradation	NA	CAPS PCL	(-0.43, 0.003) (-0.35, 0.01)
M4A	82 (*49)	Immune response	Interferon signalling	CD14+ Monocytes	PTSD Risk Group CAPS	(0.36, 0.01) (0.43, 0.002)
5	121	T cell activation	Repression of WNT target genes	CD8+ T cells	-	-
6	70	Blood coagulation	Hemostasis	-	-	-
7	135	Transcription	-	CD4+ T cells	-	-
8	101	Cell adhesion	-	-	-	-
9	1020	Intracellular signalling cascade	Signalling by interluekins	CD14+ Monocytes	PCL Caucasian	(0.32, 0.02) (0.31, 0.03)
10	218	Aerobic respiration	Citric acid cycle	-	-	-

Supplementary Table 5.4. Diagnostic PTSD 45 Gene Expression Classifier at Post-Deployment

Gene Symbol	Parametric p-value	Log Fold-change	% CV support
CCDC134	0.0431	0.93	55
NDFIP1	0.0236	1.06	57
EXOC5	0.0248	0.94	62
NOTCH4	0.0054	0.89	63
EHD3	0.0274	1.12	63
MBD5	0.0047	0.93	69
OTUD6B	0.0311	0.93	71
CLEC12B	0.0148	0.76	74
CD300A	0.0137	0.9	77
POLR2J2	0.0357	0.81	79
NUDT1	0.0266	1.07	80
NGFRAP1	0.0033	1.16	80
VAMP5	0.0169	0.9	82
COQ2	0.0207	0.93	82
METTL21B	0.0232	1.13	83
GIPC3	0.0005	1.23	83
POLR2I	0.0207	1.09	84
C2orf49	0.0044	0.92	89
SLC35B3	0.0024	0.93	90
TSPAN4	0.0343	1.11	90
ZNF347	0.0332	1.1	90
TM2D2	0.0485	0.93	91
PSMD10	0.0276	0.94	95
PAXIP1	0.0316	0.94	95
MAP7	0.0529	1.12	96
PDK1	0.044	1.07	96
DHX32	0.011	1.1	96
NUDT7	0.0218	1.15	97
TRAPPC4	0.0481	0.92	98
AHI1	0.0542	0.87	98
PPP2R2D	0.044	1.06	98
MYBL2	0.0222	1.26	98
S1PR2	0.0179	1.07	98
LEPR	0.0021	1.15	98
TMEM45B	0.0061	0.77	100
RSRC1	0.014	0.92	100
BFAR	0.0164	0.93	100
EFCAB4A	0.0293	1.2	100
ATP9A	0.0285	1.18	100
ZNF738	0.0212	1.08	100
SNX15	0.0185	1.06	100
TRPM3	0.0096	1.11	100
KCNN4	0.0064	1.15	100
COPS3	0.0057	1.11	100
HPCAL4	0.0007	1.31	100

Abbreviations: CV, cross-validation.

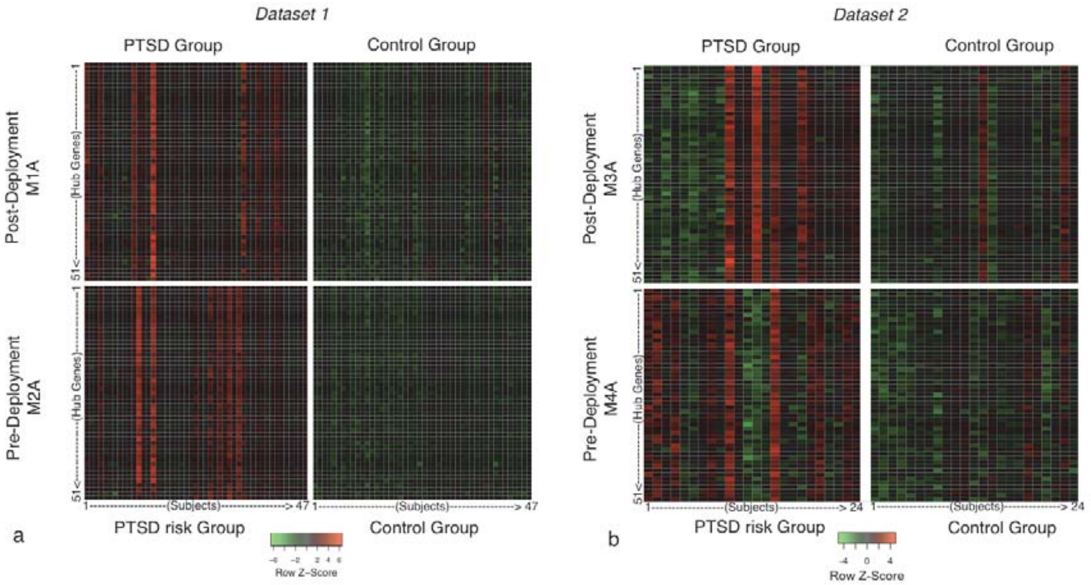
Supplementary Table 5.5. Predictive PTSD 85 Gene Expression Classifier at Pre-Deployment

Gene Symbol	Parametric p-value	Log Fold-change	% CV support
SFXN4	0.0115	0.9	43
H2AFX	0.0217	1.08	43
RETSAT	0.0309	1.06	44
ISG15	0.0009	0.69	57
DYNC1H1	0.0389	1.07	59
IRF7	0.0232	0.88	62
NOD2	0.0036	0.84	64
UACA	0.011	0.86	64
TMEM60	0.0414	0.93	64
SOX12	0.0202	1.1	64
METTL21B	0.0129	1.15	67
RBM43	0.0079	0.89	68
ALKBH5	0.0378	1.05	69
ZKSCAN5	0.0101	0.94	70
SIGLEC16	0.029	0.87	72
HARBI1	0.0207	0.92	74
TFB2M	0.0481	0.93	74
ZNF100	0.0503	0.9	74
LOC338799	0.0357	1.08	74
RAB39A	0.0277	0.88	76
PAOX	0.0118	1.1	77
COX18	0.0314	0.95	81
VPS13D	0.0265	1.06	81
LCN10	0.0184	1.19	82
PDZD11	0.0067	0.92	83
ZNF595	0.0041	1.22	85
TTF2	0.0104	0.92	86
TADA2A	0.0186	0.92	86
RPP14	0.0161	0.92	87
EFCAB4A	0.0057	1.21	87
TSFM	0.0105	0.94	88
C11orf95	0.0475	1.08	88
TUFM	0.0083	1.08	88
DPAGT1	0.0222	0.93	89
AHI1	0.0544	0.87	89
CLEC12A	0.011	0.73	93
ITM2A	0.014	1.16	93
FMNL2	0.0016	0.83	95
SGSM1	0.018	1.09	95
INTS4	0.0178	0.93	96
GTF2H5	0.0199	0.93	96
UBR4	0.0042	1.08	96
FPR3	0.0128	0.79	96
KIAA1279	0.0063	0.92	97
HMGN4	0.0204	0.94	97
ACYP1	0.0298	1.09	97
HIPK2	0.0111	1.08	97

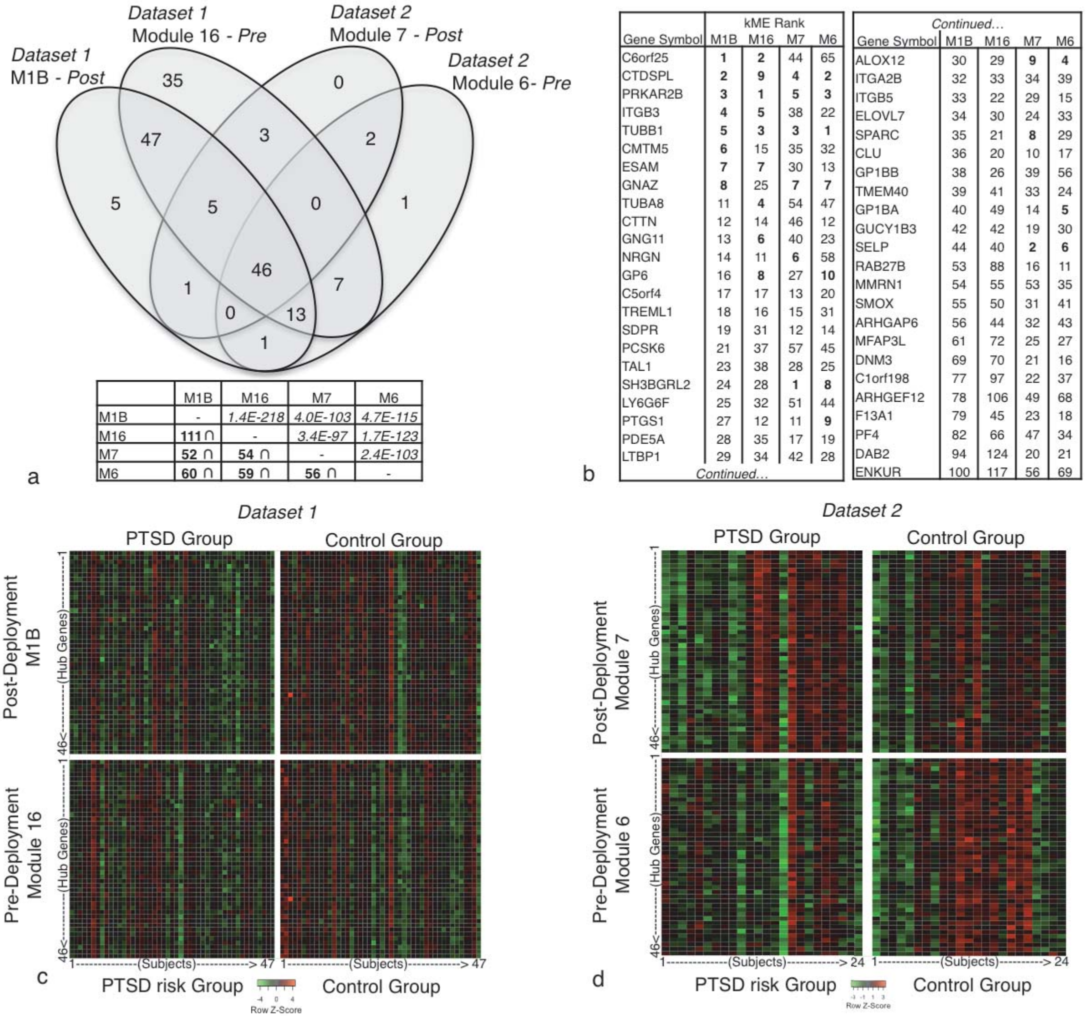
Supplementary Table 5.5. Continued…

AMFR	0.002	1.1	97
ENPP2	0.0036	0.85	98
FRMD4B	0.0064	0.88	98
IFT52	0.0149	0.92	98
TMEM45B	0.0463	0.83	98
CECR1	0.0497	0.91	98
LTBP2	0.018	1.1	98
ZNF273	0.0112	1.06	98
XRCC2	0.0241	0.92	99
MYH7B	0.0331	0.92	99
INSIG1	0.0364	0.93	99
PMS2P1	0.0476	0.93	99
AGAP4	0.0529	1.36	99
FAM19A2	0.0499	1.21	99
COPS3	0.0439	1.07	99
TWSG1	0.0363	1.07	99
FAM86C1	0.0071	1.14	99
EHHADH	0.0007	0.84	100
PCBD1	0.0012	0.87	100
NOP16	0.0043	0.92	100
SLC35E2	0.008	0.85	100
ZNF584	0.0109	0.92	100
TNFSF12	0.0165	0.61	100
ZNF485	0.0186	0.92	100
TAGLN	0.0188	0.82	100
FAM86DP	0.0261	0.89	100
CRYBB2P1	0.0292	0.94	100
MTHFD2	0.0345	0.92	100
CD86	0.0377	0.89	100
COL8A2	0.0432	0.88	100
PEX10	0.0434	0.93	100
MLF1IP	0.0456	0.89	100
LEPR	0.0415	1.1	100
PDCL	0.0374	1.08	100
CEP192	0.0304	1.1	100
WASH3P	0.0107	1.13	100
FGFR1	0.0089	1.15	100
HPCAL4	0.0007	1.28	100

Abbreviations: CV, cross-validation.



Supplementary Figure 5.2. Heatmaps representation of the 51 overlapping genes from module M1A, M2A, M3A, and M4A. Heatmaps are divided into four main parts. Dataset 1 modules (**a**), M1A on the top row and M2A on the bottom row, are split into PTSD cases (left column) and controls (right column). Dataset 2 modules (**b**), M3A on the top row and M4A on the bottom row, are split into PTSD cases (left column) and controls (right column). All subjects have been sorted in an identical fashion in order to properly visualize differences in expression as occurring within each individual sample across the two time-points (i.e. Subject 1 post-deployment is directly above Subject 1 pre-deployment). Red represents over-expression, green represents under-expression and black represents the mean.



	M1B	M16	M7	M6
M1B	-	*1.4E-218*	*4.0E-103*	*4.7E-115*
M16	**111 ∩**	-	*3.4E-97*	*1.7E-123*
M7	**52 ∩**	**54 ∩**	-	*2.4E-103*
M6	**60 ∩**	**59 ∩**	**56 ∩**	-

	kME Rank			
Gene Symbol	M1B	M16	M7	M6
C6orf25	**1**	**2**	44	65
CTDSPL	**2**	**9**	**4**	**2**
PRKAR2B	**3**	**1**	**5**	**3**
ITGB3	**4**	**5**	38	22
TUBB1	**5**	**3**	**3**	**1**
CMTM5	**6**	15	35	32
ESAM	**7**	**7**	30	13
GNAZ	**8**	25	**7**	**7**
TUBA8	11	**4**	54	47
CTTN	12	14	46	12
GNG11	13	**6**	40	23
NRGN	14	11	**6**	58
GP6	16	**8**	27	**10**
C5orf4	17	17	13	20
TREML1	18	16	15	31
SDPR	19	31	12	14
PCSK6	21	37	57	45
TAL1	23	38	28	25
SH3BGRL2	24	28	**1**	**8**
LY6G6F	25	32	51	44
PTGS1	27	12	11	**9**
PDE5A	28	35	17	19
LTBP1	29	34	42	28

Continued...

Continued...

Gene Symbol	M1B	M16	M7	M6
ALOX12	30	29	**9**	**4**
ITGA2B	32	33	34	39
ITGB5	33	22	29	15
ELOVL7	34	30	24	33
SPARC	35	21	**8**	29
CLU	36	20	10	17
GP1BB	38	26	39	56
TMEM40	39	41	33	24
GP1BA	40	49	14	**5**
GUCY1B3	42	42	19	30
SELP	44	40	**2**	**6**
RAB27B	53	88	16	11
MMRN1	54	55	53	35
SMOX	55	50	31	41
ARHGAP6	56	44	32	43
MFAP3L	61	72	25	27
DNM3	69	70	21	16
C1orf198	77	97	22	37
ARHGEF12	78	106	49	68
F13A1	79	45	23	18
PF4	82	66	47	34
DAB2	94	124	20	21
ENKUR	100	117	56	69

Supplementary Figure 5.3. Venn Diagram (**a**) depicting significant overlap in hemostasis and blood coagulation genes belonging to modules M1B post-deployment and Module 16 pre-deployment in *Dataset 1* as well as modules Module X post-deployment and Module 6 pre-deployment in *Dataset 2*. The overlap identified 46 genes found across all four analyses which are displayed in table format (**b**) along with the corresponding kME rank (i.e. rank of connectivity) for each gene within a particular module. Higher rank indicates hub gene status, the top 10 genes for each module are in bold. Heatmaps representing these 46 genes are shown across both datasets and are divided into four main parts. *Dataset 1* modules (**c**), M1B on the top row and Module 16 on the bottom row, are split into PTSD cases (left column) and controls (right column). *Dataset 2* modules (**d**), Module X on the top row and Module 6 on the bottom row, are split into PTSD cases (left column) and controls (right column). All subjects have been sorted in an identical fashion in order to properly visualize differences in expression as occurring within each individual sample across the two time-points (i.e. Subject 1 post-deployment is directly above Subject 1 pre-deployment). Red represents over-expression, green represents under-expression and black represents the mean according to the scale bar.

A.2 Supplementary Data for Chapter 6

Supplementary Table 6.1A. Physiological and hormonal measurements of all participants. (average + standard deviation)

	Baseline			Pre-Boarding			Post-Landing			One-Hour Post-Landing		
	Pooled	Male	Female	Pooled	Male	Female	Pooled	Male	Female	Pooled	Male	Female
Subjects *n*:	10	5	5	11	6	5	13	7	6	11	5	6
Age	23.1 + 4.61	24.6 + 5.98	21.6 + 2.51	22.27 + 4.45	23.17 + 2.77	21.2 + 2.77	22.31 + 4.42	23 + 5.26	21.2 + 2.77	21.09 + 2.63	21.2 + 3.03	21 + 2.53
BMI	22.86 + 1.18	23 + 1.13	22.72 + 1.35	24.17 + 3.99	25.62 + 2.51	22.44 + 2.51	23.18 + 1.97	23.65 + 1.55	22.44 + 2.51	23.23 + 2.09	23.94 + 1.78	22.63 + 2.3
Weight	149.68 + 14.73	158.68 + 14.92	140.68 + 7.94	159.05 + 25.45	177 + 18.15	137.52 + 18.15	152.81 + 18.32	162.36 + 10.77	137.52 + 18.15	153.11 + 23.61	170.92 + 18.26	138.27 + 16.34
Height	5.66 + 0.27	5.72 + 0.37	5.6 + 0.12	5.66 + 0.39	5.75 + 0.18	5.56 + 0.18	5.63 + 0.31	5.68 + 0.38	5.56 + 0.18	5.66 + 0.32	5.8 + 0.43	5.55 + 0.16
TraitAnxiety	8.5 + 1.96	9 + 2.55	8 + 1.22	8.82 + 2.86	8.17 + 3.21	9.6 + 3.21	10.46 + 3.57	10.88 + 4.29	9.8 + 2.28	7.64 + 2.42	6.2 + 0.45	8.83 + 2.79
HeartRate	73.23 + 10.45	64.94 + 7.41	81.51 + 4.1	88.17 + 9.38	85.17 + 8.48	90.57 + 8.48	108.66 + 18.08	106.06 + 10.05	111.78 + 25.85	77.81 + 9.68	76.82 + 6.68	78.61 + 12.33
Cortisol	14.65 + 7.55	13.91 + 7.47	15.39 + 8.42	13.61 + 4.93	11.28 + 6.09	16.4 + 6.09	18.24 + 7.62	16.93 + 5.25	20.86 + 11.61	13.24 + 9.2	12.46 + 10.14	14.01 + 9.29
Testosterone	254.5 + 240.93	477.2 + 80.17	31.8 + 13.68	281.45 + 247.68	490.5 + 10.76	30.6 + 10.76	316.77 + 239.39	494 + 69.74	33.2 + 8.53	219.91 + 224.58	446 + 93.99	31.5 + 9.38
BetaEndorphin	12.84 + 5.38	15.52 + 5.19	10.16 + 4.52	12.41 + 4.41	14.39 + 0.93	10.03 + 0.93	17.34 + 6.19	17.26 + 4.68	17.47 + 8.77	16.18 + 6.83	12.12 + 2.85	21.26 + 7.19
NGF	38.14 + 27.91	54.85 + 31.5	24.78 + 17.67	25.63 + 17.18	27.8 + 21.87	23.03 + 21.87	23.23 + 15.28	22.5 + 17.5	24.34 + 13.69	127.71 + 134.26	31.21 + 29.71	192.04 + 139.65
Norepinephrine1	303.72 + 138.2	293.5 + 165.27	313.94 + 124.08	273.08 + 133.03	320.93 + 59.58	215.66 + 59.58	490.67 + 254.57	570.55 + 296.43	362.86 + 86.38	281.18 + 120.54	298.2 + 148.54	267 + 104.31
Norepinephrine2	298.13 + 154.9	299.44 + 194.01	296.82 + 127.82	276.4 + 138.74	325.77 + 69.34	217.16 + 69.34	496.07 + 283.44	578.21 + 336.19	364.64 + 90.17	204.42 + 186.15	320.36 + 162.14	107.8 + 153.72
Epinephrine1	41.23 + 53.63	27.98 + 16.28	54.48 + 75.95	42.63 + 31	40.62 + 38.57	45.04 + 38.57	75.66 + 52	96.38 + 53.14	42.52 + 30.62	45.65 + 41.09	42.42 + 23.61	48.35 + 53.96
Epinephrine2	37.75 + 54.65	23.78 + 10.11	51.72 + 78.29	45.42 + 32.42	43.1 + 40.34	48.2 + 40.34	70.23 + 55.33	93.84 + 59.23	32.46 + 12.05	48.14 + 41.95	47.22 + 29.06	48.9 + 53.31

Supplementary Table 6.1B. Non-Parametric Mann-Whitney U test used for comparisons from Baseline to Pre-Boarding, Post-Landing and One-Hour Post-Landing. P < 0.05 was considered significant.

	P-Values Relative To Baseline								
	Pre-Boarding			Post-Landing			One-Hour Post-Landing		
	Pooled	Male	Female	Pooled	Male	Female	Pooled	Male	Female
TraitAnxiety	1	0.6435	0.4606	0.179	0.5549	0.1127	0.216	0.05735	0.7808
HeartRate	0.01522	0.02857	0.01587	0.0001058	0.009524	0.01587	0.5414	0.1143	0.9048
Cortisol	0.8094	0.7922	0.8413	0.1802	0.2844	0.4127	0.5288	0.5476	1
Testosterone	0.8047	0.7922	1	0.7327	0.9416	0.6723	0.7512	0.6905	1
BetaEndorphin	0.9159	0.6473	0.6905	0.1151	0.6216	0.2222	0.2428	0.2222	0.01587
NGF	0.3702	0.1714	0.8413	0.211	0.06667	0.9048	0.2428	0.3429	0.01732
Norepinephrine1	0.7564	0.9307	0.3095	0.03584	0.04507	0.5476	0.7564	1	0.5368
Norepinephrine2	0.9177	0.9307	0.4206	0.05746	0.1709	0.5476	0.1734	1	0.08225
Epinephrine1	0.5116	0.5368	0.8413	0.02136	0.01088	0.4206	0.3867	0.3095	0.5368
Epinephrine2	0.1734	0.4286	0.3095	0.02552	0.02953	0.3095	0.1734	0.3095	0.4286

Abbreviations; NGF, nerve growth factor.

Supplementary Table 6.2. A comprehensive functional characterization for all identified WGCNA modules

	Genes (*n*) in:		***Characterization***		
Module Color	**Module**	**kME > 0.5**	**Biological Process**	**Pathway**	**Cell Type**
Black	258	258	Ribosome Biogenesis	Ribosome Biogenesis in Eukaryotes	-
Pink	239	239	Translational elongation	Ribosome	-
Salmon	131	131	Translational termination	Cytoplasmic Ribosomal Proteins	-
Royal-blue	24	24	-	Formation of the ternary comple and 43S complex	-
Green-yellow	159	159	Immune Response	NK cell mediated cytotoxicity	NK Cells
Brown	578	578	T-cell receptor signaling	T cell receptor signaling pathway	CD8 T Cells
Red	274	274	RNA processing	G2/M Transition	CD4 T Cells
Yellow	314	314	Immune response	Thrombin signaling	CD14 Monocytes
Cyan	118	118	Cytokine production in immune response	-	CD14 Monocytes
Green	313	313	Innate Immune Response	Interferon Signaling	CD14 Monocytes
Blue	443	348	Blood Coagulation	Hemostasis	Blood Platelets
Light yellow	32	32	-	Uptake of Oxygen and Release	Blood Platelets
Midnight blue	113	90	Oxygen transporter activity	-	CD71 Early Erythroid
Grey60	103	103	Respiratory electron transport chain	Parkinson's disease	-
Light cyan	109	109	Respiratory electron transport chain	Respiratory electron transport	-
Purple	164	164	Respiratory electron transport chain	Electron Transport Chain	-
Light green	87	87	Regulation of actin filament polymerization	Role of PI3K subunit p85 in regulation of Actin Organization	CD14 Monocytes
Magenta	228	228	Regulation of intracellular signal transduction	RORA Activates Circadian Expression	CD33 Myeloid
Turquoise	1289	1152	Protein Ubiquitination	Signaling by NOTCH	-
Grey	13,152	748	Response to Wounding	IL-5 Signaling Pathway	-

The top most significant functional annotations of all co-expression modules. Listed in the table are the total number of genes corresponding to each module, the total number of genes with a kME > 0.5 used for enrichment analyses, the top most significant biological process and pathway (as indicated by ToppGene(65)) and cell type (as indicated by CTen (18)) for each corresponding module. kME specifies the strength of association of a gene to its corresponding ME value. Only genes with kME > 0.05 were used for enrichment analyses. All annotations must have passed a Bonferroni correction $p < 0.05$. Grey shading is for visualization purposes only.

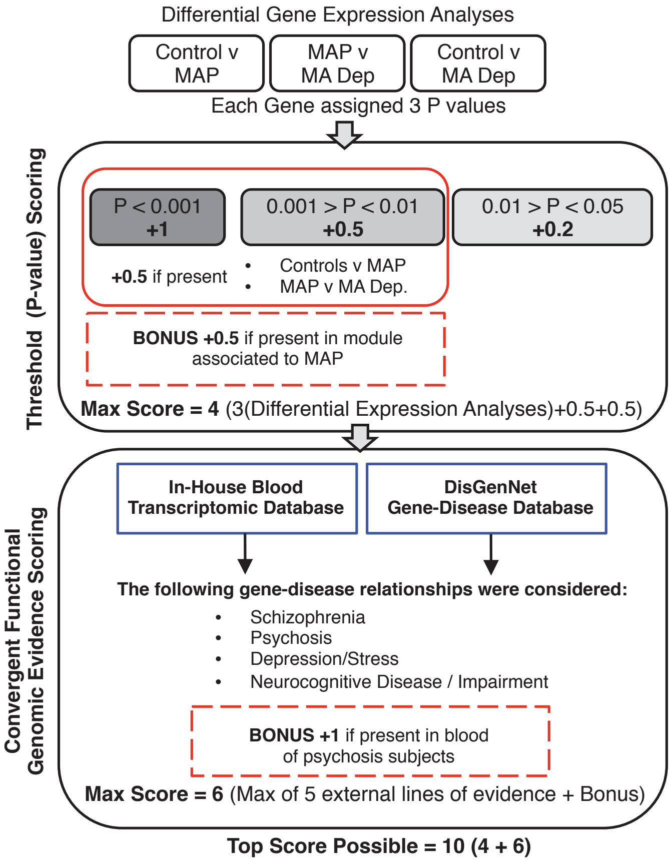
Supplementary Table 6.3A. Flow cytometry data from 26 participants (17 Male and 9 Female) on peripheral blood luekocyte subsets. (average + standard deviation)

	Baseline			Pre-Boarding			Post-Landing			One-hour Post-Landing		
Cell Types	POOLED	MALE	FEMALE	POOLED	MALE	FEMALE	POOLED	MALE	FEMALE	POOLED	MALE	FEMALE
LEUK	5.625 + 1.271	6.008 + 1.247	4.914 + 1.048	6.814 + 1.341	7.171 + 1.341	6.1 + 1.098	7.438 + 1.603	7.643 + 1.765	7.029 + 1.235	6.052 + 1.411	6.371 + 1.466	5.414 + 1.123
NEUT	52.629 + 5.132	53.877 + 5.331	50.6 + 4.352	67.4 + 9.757	70.507 + 10.574	61.963 + 4.984	60.727 + 9.018	63.129 + 9.287	56.525 + 7.215	56.459 + 9.03	59.586 + 9.591	50.988 + 4.474
EOS	3.606 + 1.85	3.854 + 2.115	2.96 + 0.631	1.55 + 1.703	1.714 + 1.985	1.167 + 0.723	1.66 + 1.622	1.807 + 1.873	1.317 + 0.816	2.4 + 1.974	2.429 + 2.296	2.333 + 1.048
MONO	8.816 + 1.384	9.208 + 1.322	8.30 + 1.198	6.206 + 1.357	6.321 + 1.317	5.8 + 1.623	7.444 + 1.901	7.679 + 1.881	6.625 + 1.996	8.106 + 1.84	8.307 + 1.963	7.4 + 1.278
TOTAL LYMPH	34.042 + 6.566	32.638 + 6.123	37.083 + 7	22.33 + 8.57	20.971 + 8.737	26.833 + 5.354	28.525 + 8.96	27.471 + 8.757	30.983 + 9.766	31.255 + 8.234	28.579 + 7.281	37.5 + 7.275
B LYMPH	13.515 + 3.912	12.472 + 3.438	15.9 + 4.126	11.525 + 3.769	10.418 + 2.829	13.185 + 4.553	9.875 + 3.102	9.31 + 2.154	10.721 + 4.179	13.46 + 3.374	13.238 + 3.068	13.793 + 3.987
NK	8.434 + 4.631	8.606 + 3.412	8.069 + 6.833	14.248 + 6.718	15.658 + 5.858	11.026 + 7.892	22.34 + 9.296	22.901 + 8.627	20.977 + 11.389	8.585 + 4.299	8.918 + 3.991	7.779 + 5.226
CD3 LYMPH	72.233 + 5.846	73.363 + 5.164	69.81 + 6.883	69.646 + 8.316	70.638 + 6.532	67.663 + 11.583	64.295 + 8.797	65.623 + 7.476	61.418 + 11.398	72.206 + 6.72	72.728 + 5.808	71.075 + 8.909
CD4 LYMPH	43.456 + 6.159	42.461 + 5.214	45.942 + 8.077	39.339 + 7.643	37.083 + 6.36	44.494 + 8.275	33.267 + 7.025	32.082 + 6.713	36.144 + 7.44	43.433 + 6.164	42.878 + 5.781	44.78 + 7.316
CD8 LYMPH	61.41 + 13.341	63.121 + 12.444	57.132 + 15.734	61.7 + 13.12	61.071 + 13.017	63.14 + 14.287	59.546 + 12.886	59.926 + 12.027	58.621 + 15.791	60.157 + 11.889	60.03 + 11.583	60.464 + 13.561

Supplementary Table 6.3 B. A Dunnett Test for simulaneous comparison of means was used to compare pre-boarding, post-landing and one-hour post-landing flow cytometry data from Supp. Table 5A back to baseline.

	Pre-Boarding			Post-Landing			One-hour Post-Landing		
Cell Types	POOLED	MALE	FEMALE	POOLED	MALE	FEMALE	POOLED	MALE	FEMALE
LEUK	0.023	0.112	0.148	0.001	0.016	0.005	0.646	0.855	0.746
NEUT	0.001	0.001	0.001	0.007	0.027	0.061	0.319	0.242	0.998
EOS	0.002	0.027	0.005	0.004	0.035	0.010	0.104	0.190	0.461
MONO	0.001	0.001	0.060	0.020	0.052	0.421	0.302	0.359	0.898
TOTAL LYMPH	0.001	0.001	0.063	0.084	0.217	0.380	0.542	0.397	0.999
B LYMPH	0.188	0.177	0.508	0.003	0.016	0.056	1.000	0.855	0.659
NK	0.008	0.003	0.828	0.001	0.000	0.014	1.000	0.997	1.000
CD3 LYMPH	0.585	0.608	0.959	0.000	0.006	0.024	1.000	0.996	0.991
CD4 LYMPH	0.124	0.041	0.987	0.001	0.001	0.084	1.000	0.998	0.991
CD8 LYMPH	0.999	0.934	0.778	0.925	0.782	0.993	0.973	0.799	0.951

A.3 Supplementary Data for Chapter 7


Differential Gene Expression Analyses
Control v MAP
MAP v MA Dep
Control v MA Dep
Each Gene assigned 3 P values
Threshold (P-value) Scoring
P < 0.001 +1
0.001 > P < 0.01 +0.5
0.01 > P < 0.05 +0.2
+0.5 if present
Controls v MAP
MAP v MA Dep.
BONUS +0.5 if present in module associated to MAP
Max Score = 4 (3(Differential Expression Analyses)+0.5+0.5)
Convergent Functional Genomic Evidence Scoring
In-House Blood Transcriptomic Database
DisGenNet Gene-Disease Database
The following gene-disease relationships were considered:
Schizophrenia
Psychosis
Depression/Stress
Neurocognitive Disease / Impairment
BONUS +1 if present in blood of psychosis subjects
Max Score = 6 (Max of 5 external lines of evidence + Bonus)
Top Score Possible = 10 (4 + 6)

Supplementary Figure 7.1. Convergent Functional Genomic (CFG) scoring scheme. First, each gene received a score based on *p*-value threshold. A score of 1 was given for $P < 0.001$, a score of 0.5 for $0.001 > P < 0.01$ and a score of 0.2 for $0.01 > P < 0.05$. A gene was given an additional score of 0.5 if $P < 0.01$ between MAP and controls subjects as well as MAP and MA subjects. A bonus 0.5 point was given if this gene was found in a functional module associated to MAP or psychosis (i.e. ubiquitin-mediated proteolysis or circadian rhythm modules). Thus, the maximum score based on this first series of thresholds is 4 (3(differential expression analyses) + 0.5 + 0.5). Second, we used CFG evidence as identified from two databases; (i) an in-house blood transcriptomic database and (ii) DisGenNet database. We only used gene-disease relationships for the following diseases: schizophrenia, psychosis, depression/stress and neurocognitive impairment. A maximum of 5 external lines of evidence were allowed. A bonus point of 1 was granted if present in the blood of previous psychosis studies. Thus, the maximum score attainable is 6 (5 lines of evidence + 1) and the top score possible for each gene considering all possible combinations of points is 10 (4 + 6).

Supplementary Table 7.1. Annotation of co-expression modules including GO functional components, KEGG and Reactome pathways, and CTD drug compounds as well as cell type specificity.

Consensus Function (Module Name)	Total # Genes	kME > 0.5	GO: Molecular Factor	GO: Biological Process	GO: Cellular Compartment	KEGG and Reactome Pathways	CTD Drug Response	C-Ten Cell type specificity
Protein Heterodimerization	56	56	protein heterodimerization activity	nucleosome assembly	chromosome	Alcoholism	Lucanthone	CD71+ Early Erythroid
Ribosome Pathway	281	281	structural constituent of ribosome	translational termination	ribosomal subunit	Ribosome	guanosine 5'-diphosphate disodium salt	-
Oxidoreductase activity	937	937	oxidoreductase activity, acting on peroxide as acceptor	proteasomal protein catabolic process	mitochondrial part	Senescence-Associated Secretory Phenotype (SASP)	Selenium	CD71+ Early Erythroid
Natural Killer cell mediated cytoxicity	165	165	tubulin binding	immune response (T Cell activation)	membrane raft	Natural killer cell mediated cytotoxicity	abrine	CD56+ NK cells
Hemostasis	192	192	fibrinogen binding	blood coagulation	platelet alpha granule	Hemostasis	U46619	Cardiac Myocytes
Chloride transporter Activity	106	106	chloride transmembrane transporter activity	regulation of phosphate metabolic process	lytic vacuole	-	Dietary Carbohydrates	CD14+ Monocytes
RNA binding / Resp. electron trans. Chain	995	995	RNA binding	ncRNA metabolic process	nucleolus	The citric acid (TCA) cycle and respiratory electron transport	Selenium	-
Circadian Clock	332	332	chromatin binding	chromatin modification	nucleoplasm part	Circadian Clock	-	-
Cytokine Signalling	186	186	enzyme binding	defense response	cortical cytoskeleton	Cytokine Signaling in Immune system	Selenium	CD14+ Monocytes
IL-5 Signalling	659	659	kinase activity	response to wounding	focal adhesion	IL-5 Signaling Pathway	Aspirin	Whole Blood
Actine cytoskeleton	93	93	-	single-organism organelle organization	actin cytoskeleton	-	Selenium	-
ATPase activity	26	26	proton-transporting ATPase activity, rotational mechanism	interspecies interaction between organisms	-	Viral carcinogenesis	-	-
-	74	74	-	-	-	-	-	-
Histrone demethylase activity	16	16	histone demethylase activity	histone lysine demethylation	-	-	Chromium	-
B cell activation	106	106	MHC class II protein complex binding	B cell activation	external side of plasma membrane	B Cell Receptor Signaling Pathway	Chloroprene / Dextran Sulfate	CD19+ B cells
Generic Transcription	880	880	-	-	-	Generic Transcription Pathway	-	-
Centrosome maturation	699	669	adenyl nucleotide binding	-	nucleolus	Centrosome maturation	beta-methylcholine	-
Protein ubiquination	175	175	ubiquitin-protein transferase activity	protein ubiquitination	endosome membrane	Translocation of GLUT4 to the Plasma Membrane	Chlorodiphenyl (54% Chlorine)	Whole Blood
Ubiquitin mediated proteolysis	767	767	RNA binding	ER to Golgi vesicle-mediated transport	catalytic complex	Ubiquitin mediated proteolysis	sodium arsenate	-
Generic Transcription	48	48	sequence-specific DNA binding transcription factor activity	-	-	Generic Transcription Pathway	Clorgyline	-
G-coupled protein receptor activity	90	93	G-protein coupled pyrimidinergic nucleotide receptor activity	-	-	Nucleotide-like (purinergic) receptors	-	-
-	43	43	-	-	-	-	-	-
Interferon Signalling	263	263	double-stranded RNA binding	defense response to virus (innate immune response)	host	Interferon Signaling	Zidovudine	CD14+ Monocytes
RNA degradation	5094	1159	RNA binding	RNA degradation	nucleolus	Gene Expression	potassium chromate(VI)	-

The top most significant functional annotations of all co-expression modules. Listed in the table are the total number of genes corresponding to each module, the total number of genes with kME > 0.5 used for enrichment analyses, the consensus function (i.e. Module Name), the top most significant GO terms and Pathways (as indicated by ToppGene), drug compound (as indicated by CTD) and cell type (as indicated by CTen) for each corresponding module. All annotations must have passed a Bonferroni correction p < 0.05. Abbreviations: (-) signifies no enchriment. Modules are ordered as they are presented in Supplementary Figure 1.

Supplementary Table 7.2. Converging Evidence of Ubiquitin-Proteasome System Dysfunction in Psychosis and SCZ (61 genes) across 2 or more studies

Gene Symbol (Gene Name)	kME Rank	Controls compared to MAP subjects (P-Value)	MA dep. compared to MAP subjects (P-Value)
SLC35A5 (solute carrier family 35, member A5)*	402	0.005036068	0.009066804
TMEM135 (transmembrane protein 135)*	541	0.004280073	0.02070815
SRSF1 (serine/arginine-rich splicing factor 1)*	130	0.020365344	0.027750422
KRAS (Kirsten rat sarcoma viral oncogene homolog)*	748	0.005468547	0.035304303
CBR4 (carbonyl reductase 4)*	297	0.011896041	0.03795962
ATXN3 (ataxin 3)*	248	0.005862784	0.038045436
SCAMP1 (secretory carrier membrane protein 1)*	50	0.009233944	0.046212855
PDE12 (phosphodiesterase 12)*	189	0.003623366	0.049340743
UHMK1 (U2AF homology motif (UHM) kinase 1)	122	0.000770142	0.050781828
PI4K2B (phosphatidylinositol 4-kinase type 2 beta)	471	0.070296417	0.052068539
SLC12A2 (solute carrier family 12 (sodium/potassium/chloride transporter), member 2)	397	0.002913315	0.053049548
PIK3C2A (phosphatidylinositol-4-phosphate 3-kinase, catalytic subunit type 2 alpha)	515	0.016284186	0.05371985
CREB1 (cAMP responsive element binding protein 1)	705	0.05103424	0.055487382
TMX1 (thioredoxin-related transmembrane protein 1)	126	0.041730028	0.072693538
TMED2 (transmembrane emp24 domain trafficking protein 2)	304	0.013018731	0.076380753
SRSF6 (serine/arginine-rich splicing factor 6)	661	0.014962981	0.081968547
INSIG2 (insulin induced gene 2)	521	0.023312974	0.097937444
GLS (glutaminase)	52	0.00345909	0.111942995
PPP4R3B (protein phosphatase 4, regulatory subunit 3B)	153	0.006801293	0.115199044
TMEM106B (transmembrane protein 106B)	23	0.00791962	0.121807757
ZMPSTE24 (zinc metallopeptidase STE24)	664	0.045687915	0.126663613
CD47 (CD47 molecule)	127	0.021245395	0.127631979
SMAD4 (SMAD family member 4)	428	0.01431481	0.143685381
TRIM33 (tripartite motif containing 33)	704	0.064622229	0.14606847
SPAST (spastin)	637	0.03563803	0.149721331
FMR1 (fragile X mental retardation 1)	710	0.211420859	0.150840516
CUL5 (cullin 5)	42	0.009944088	0.167495521
UBE3A (ubiquitin protein ligase E3A)	234	0.018753135	0.174383469
PPP1R2 (protein phosphatase 1, regulatory (inhibitor) subunit 2)	138	0.012123043	0.18965167
SMARCAD1 (SWI/SNF-related, matrix-associated actin-dependent regulator of chromatin, subfamily a, containing DEAD/H box 1)	395	0.002426846	0.191423652

Supplementary Table 7.2 Continued…			
KBTBD8 (kelch repeat and BTB (POZ) domain containing 8)	450	0.045333109	0.195495324
CGGBP1 (CGG triplet repeat binding protein 1)	27	0.015986201	0.210443621
GABPA (GA binding protein transcription factor, alpha subunit 60kDa)	167	0.056065324	0.211705588
UBE2K (ubiquitin-conjugating enzyme E2K)	148	0.009114176	0.213537693
TSNAX (translin-associated factor X)	168	0.054280285	0.214208409
TMEM64 (transmembrane protein 64)	298	0.012574054	0.219020364
PURA (purine-rich element binding protein A)	351	0.022901138	0.219326715
ABCE1 (ATP-binding cassette, sub-family E (OABP), member 1)	131	0.006996746	0.222064969
PPAT (phosphoribosyl pyrophosphate amidotransferase)	671	0.049250888	0.227668205
CCNT2 (cyclin T2)	337	0.07940372	0.236109857
USP37 (ubiquitin specific peptidase 37)	463	0.008354336	0.236890265
ARL6IP5 (ADP-ribosylation factor-like 6 interacting protein 5)	751	0.085090314	0.259326929
RORA (RAR-related orphan receptor A)	550	0.004138795	0.26325037
ATPBD4 (ATP-Binding Domain-Containing Protein 4)	640	0.134404836	0.268554194
RBM12B (RNA binding motif protein 12B)	292	0.024960346	0.281060893
MBNL1 (muscleblind-like splicing regulator 1)	139	0.022403533	0.282922302
FBXO45 (F-box protein 45)	507	0.039659814	0.319519337
CDC42SE2 (CDC42 small effector 2)	611	0.01752612	0.353651265
FOPNL (FGFR1OP N-terminal like)	655	0.010828004	0.365110172
PNPLA8 (patatin-like phospholipase domain containing 8)	701	0.038406574	0.386508032
CCDC117 (coiled-coil domain containing 117)	288	0.016813123	0.390250961
CPOX (coproporphyrinogen oxidase)	626	0.059787423	0.395058907
SLC38A2 (solute carrier family 38, member 2)	374	0.011860001	0.480423705
PRKAA1 (protein kinase, AMP-activated, alpha 1 catalytic subunit)	523	0.111568576	0.508335014
ANKRD46 (ankyrin repeat domain 46)	687	0.005836655	0.526368714
C5orf30 (chromosome 5 open reading frame 30)	745	0.229324132	0.534481255
NUCKS1 (nuclear casein kinase and cyclin-dependent kinase substrate 1)	454	0.064902255	0.619381343
SP4 (Sp4 transcription factor)	546	0.071196324	0.69576249
PDIK1L (PDLIM1 interacting kinase 1 like)	442	0.042933639	0.827291808
RAB18 (RAB18, member RAS oncogene family)*	758	0.022318388	0.046208386
KPNA3 (karyopherin alpha 3 (importin alpha 4))	474	0.122081291	0.876145255

Abbreviations; kME, intramodule connectivity – lower values indicated more connections.

Supplementary Table 7.3. Converging Evidence of Circadian Clock Dysfunction in Psychosis and SCZ (39 genes) across 2 or more studies.

Gene Symbol (Gene Name)	kME Rank	Controls compared to MAP subjects (P-Value)	MA dep. compared to MAP subjects (P-Value)
ELK3 (ELK3, ETS-domain protein (SRF accessory protein 2))*	279	0.000373891	0.002799746
SIN3A (SIN3 transcription regulator family member A)	269	0.056621475	0.000960396
NCOA6 (nuclear receptor coactivator 6)	15	0.00819083	0.065782976
HSPA5 (heat shock 70kDa protein 5 (glucose-regulated protein, 78kDa))	161	0.019425242	0.038156813
LRSAM1 (leucine rich repeat and sterile alpha motif containing 1)	282	0.028265776	0.624673386
NDE1 (nudE neurodevelopment protein 1)	83	0.038033941	0.124204979
DCTN1 (dynactin 1)	224	0.049859396	0.712786026
CHERP (calcium homeostasis endoplasmic reticulum protein)	184	0.074477147	0.387195703
CHD4 (chromodomain helicase DNA binding protein 4)	163	0.077288498	0.193406776
MYO9B (myosin IXB)	70	0.082222837	0.560870768
SRRM2 (serine/arginine repetitive matrix 2)	31	0.082936525	0.041678231
MINK1 (misshapen-like kinase 1)	182	0.086827984	0.288562178
SMARCC1 (SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily c, member 1)	271	0.094353099	0.219343362
SMARCA2 (SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 2)	202	0.107412613	0.037344599
MED15 (mediator complex subunit 15)	54	0.111453087	0.181283044
SAFB (scaffold attachment factor B)	230	0.133032076	0.66423428
MAP3K14 (mitogen-activated protein kinase kinase kinase 14)	199	0.133291767	0.482811868
VAMP2 (vesicle-associated membrane protein 2 (synaptobrevin 2))	187	0.16882053	0.154038143
RERE (arginine-glutamic acid dipeptide (RE) repeats)	111	0.186683998	0.389227155
MED12 (mediator complex subunit 12)	27	0.194281831	0.340704704
ATP2B4 (ATPase, Ca++ transporting, plasma membrane 4)	126	0.206310855	0.180437354
GNAO1 (guanine nucleotide binding protein (G protein), alpha activating activity polypeptide O)	232	0.246287946	0.055338767
PLCB2 (phospholipase C, beta 2)	101	0.258614082	0.920323146
NBEAL2 (neurobeachin-like 2)	32	0.296695445	0.756020353
CSNK1D (casein kinase 1, delta)	130	0.313725274	0.59197451
MECP2 (methyl CpG binding protein 2)	110	0.329054985	0.030184796

Supplementary Table 7.3. Continued….

TSC1 (tuberous sclerosis 1)	128	0.370942204	0.3112294
IGF1R (insulin-like growth factor 1 receptor)	173	0.383151949	0.209443873
VAMP1 (vesicle-associated membrane protein 1 (synaptobrevin 1))	269	0.398909388	0.607844156
MYH9 (myosin, heavy chain 9, non-muscle)	209	0.406930252	0.720077891
ABCA7 (ATP-binding cassette, sub-family A (ABC1), member 7)	154	0.423083402	0.907774563
MBD1 (methyl-CpG binding domain protein 1)	250	0.438154677	0.817735487
ANKRD11 (ankyrin repeat domain 11)	36	0.528331631	0.73245932
USP7 (ubiquitin specific peptidase 7 (herpes virus-associated))	201	0.565622579	0.468172617
HDAC10 (histone deacetylase 10)	332	0.606352092	0.458737712
C21orf62 (chromosome 21 open reading frame 62)	200	0.751897518	0.834000846
DGKZ (diacylglycerol kinase, zeta)	252	0.765018474	0.278595433
RXRB (retinoid X receptor, beta)	139	0.782877301	0.665337942
HERC1 (HECT and RLD domain containing E3 ubiquitin protein ligase family member 1)	236	0.790071043	0.157318358

Abbreviations; kME, intramodule connectivity – lower values indicated more connections.

A.4 Supplementary Data for Chapter 8

Supplementary Table 8.1. Functional Enrichment of WGCNA Modules using ToppGene

Module	Number of Genes	GO: Biological Process	GO: Molecular Function
M1	49	RNA-Splicing	NA
M2	45	Immune Response	Hydrolase Acticity
M3	324	Apoptotic Process	Rho GTPase Activity
M4	60	Viral Transcription	NADH Dehydrogenase
M5	313	Protein Targeting to ER	RNA Binding
M6	266	RNA Binding	Membrane Organization
M7	192	Defense Response to Virus	Double-Stranded RNA Binding
M8	182	NA	NA
M9	93	Pyruvate Metabolism	Monosaccharide Binding
M10	212	Organelle Organization	Zinc Ion Binding
M11	156	NA	NA
M12	157	Response to ER Stress	Protein Disulfide Isomerase
M13	16	NA	Dynein Light Chain Binding
M14	6030	NA	NA
M15	352	Cell Cycle	NA
M16	18	Response to Unfolded Protein	Unfolded Protein Binding
M17	17	Phosphatidylserine Metabolism	Protein Kinase C Binding
M18	17	Cellular Amino Acid Metabolism	Aminoacyl-tRNA ligase Activity
M19	85	Axonal Growth Stimulation	Poly(A) RNA Binding
M20	15	Signal Attenuation	NA
M21	341	RNA processing	RNA Binding
M22	188	Translational Initiation	RNA Binding

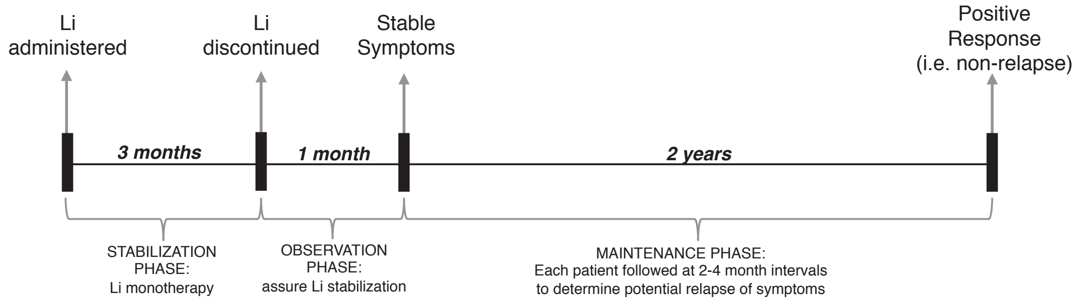
All reported GO terms pass Bonferroni Corrected P-value < 0.05. Grey shading is for visualization purposes only.

Supplementary Table 8.2. Lithium responsive genes (48 genes) found across two or more transcriptome-based studies

Gene Symbol (Gene Name)	Log Fold-Change	FDR P-Value
STC2 (stanniocalcin 2)	-0.787642412	0.004431163
HADH (hydroxyacyl-CoA dehydrogenase)	-0.742905941	5.43E-08
GAMT (guanidinoacetate N-methyltransferase)	-0.741196314	0.000765779
MAT2A (methionine adenosyltransferase II, alpha)	-0.480038186	0.000706098
HSP90AA1 (heat shock protein 90kDa alpha (cytosolic), class A member 1)	-0.473267319	5.85E-05
SERPINH1 (serpin peptidase inhibitor, clade H (heat shock protein 47), member 1, (collagen binding protein 1))	-0.469304794	0.015807785
SAMD1 (sterile alpha motif domain containing 1)	-0.401903954	0.004792479
UBE2O (ubiquitin-conjugating enzyme E2O)	-0.370350649	4.93E-05
RCL1 (RNA terminal phosphate cyclase-like 1)	-0.350496614	0.015438112
CD274 (CD274 molecule)	-0.340682942	0.021924838
RBM3 (RNA binding motif (RNP1, RRM) protein 3)	-0.33533595	0.001021757
WDR74 (WD repeat domain 74)	-0.328864046	0.002098612
PTGES3 (prostaglandin E synthase 3 (cytosolic))	-0.30906055	0.004631893
PRPF19 (pre-mRNA processing factor 19)	-0.244685246	0.010914539
SF3A1 (splicing factor 3a, subunit 1, 120kDa)	-0.150228772	0.042112352
GPBP1 (GC-rich promoter binding protein 1)	0.215204406	0.045129837
GSK3B (glycogen synthase kinase 3 beta)	0.223931229	0.020924478
ZMAT2 (zinc finger, matrin-type 2)	0.246070174	0.002962116
RNF10 (ring finger protein 10)	0.247009827	0.001067108
OAS1 (2'-5'-oligoadenylate synthetase 1, 40/46kDa)	0.318289165	0.010329817
LY6E (lymphocyte antigen 6 complex, locus E)	0.343537333	0.000739946
OAS3 (2'-5'-oligoadenylate synthetase 3, 100kDa)	0.346065696	0.011863165
ICMT (isoprenylcysteine carboxyl methyltransferase)	0.346234993	0.002864987
ZCCHC2 (zinc finger, CCHC domain containing 2)	0.365761581	0.013891424
RSU1 (Ras suppressor protein 1)	0.398498594	0.001268264
SPAG9 (sperm associated antigen 9)	0.420805708	0.000393742
OAS2 (2'-5'-oligoadenylate synthetase 2, 69/71kDa)	0.424841647	8.22E-05
IFIT1 (interferon-induced protein with tetratricopeptide repeats 1)	0.432200676	0.047500448
TARSL2 (threonyl-tRNA synthetase-like 2)	0.460413488	0.019991193
ATF4 (activating transcription factor 4)	0.477965693	1.73E-06
KLF6 (Kruppel-like factor 6)	0.485935959	6.22E-05
PCYOX1 (prenylcysteine oxidase 1)	0.488646586	0.002532514

Supplementary Table 8.2 Continued…

MAP2K3 (mitogen-activated protein kinase kinase 3)	0.576294946	0.000141255
BAK1 (BCL2-antagonist/killer 1)	0.641622163	1.13E-05
BMF (Bcl2 modifying factor)	0.650090637	4.35E-08
IFI6 (interferon, alpha-inducible protein 6)	0.654402152	0.00030134
EPSTI1 (epithelial stromal interaction 1 (breast))	0.682009503	3.34E-05
ETHE1 (ethylmalonic encephalopathy 1)	0.687180204	5.53E-06
BCL2L1 (BCL2-like 1)	0.716512969	6.24E-07
APOL6 (apolipoprotein L, 6)	0.817789836	2.06E-05
RGS2 (regulator of G-protein signaling 2)	0.826971027	0.049235509
ACP5 (acid phosphatase 5, tartrate resistant)	0.925079439	8.79E-05
ETV7 (ets variant 7)	1.06684254	0.004754349
RSAD2 (radical S-adenosyl methionine domain containing 2)	1.190676058	3.18E-05
LAX1 (lymphocyte transmembrane adaptor 1)	1.223269007	0.002584398
FOS (FBJ murine osteosarcoma viral oncogene homolog)	1.297603269	1.70E-06
CKB (creatine kinase, brain)	2.275784906	2.67E-10
CRIP1 (cysteine-rich protein 1 (intestinal))	2.46875341	4.60E-13


Li administered
Li discontinued
Stable Symptoms
Positive Response (i.e. non-relapse)
3 months
1 month
2 years
STABILIZATION PHASE: Li monotherapy
OBSERVATION PHASE: assure Li stabilization
MAINTENANCE PHASE: Each patient followed at 2-4 month intervals to determine potential relapse of symptoms

Supplementary Figure 8.1. Clinical design used to determine BD Li responders (i.e. non-relapse) and BD Li non-responders (i.e. relapse). First, blood samples were taken and following, Li monotherapy was administered to each patient for 3 months (stabilization phase). Second, Li was discontinued and BD patients were followed for 1 month to determine whether symptoms were stable (observation phase). Third, BD patients were followed for 2 years at 2-4 month intervals to determine potential relapse of BD symptoms (i.e. non-responders). Li response (i.e. relapse status) was determined by the patients ability to reach the end of the 2 year follow-up without relapse of BD symptoms and Li administration (maintenance phase).

Molecular Psychiatry (2015), 1–8
© 2015 Macmillan Publishers Limited All rights reserved 1359-4184/15
www.nature.com/mp

ORIGINAL ARTICLE

Gene networks specific for innate immunity define post-traumatic stress disorder

MS Breen[1], AX Maihofer[2], SJ Glatt[3], DS Tylee[3], SD Chandler[2], MT Tsuang[2,4,5,6,7], VB Risbrough[2,4], DG Baker[2,4], DT O'Connor[6,8], CM Nievergelt[2,4,9] and CH Woelk[1,9]

The molecular factors involved in the development of Post-Traumatic Stress Disorder (PTSD) remain poorly understood. Previous transcriptomic studies investigating the mechanisms of PTSD apply targeted approaches to identify individual genes under a cross-sectional framework lack a holistic view of the behaviours and properties of these genes at the system-level. Here we sought to apply an unsupervised gene-network based approach to a prospective experimental design using whole-transcriptome RNA-Seq gene expression from peripheral blood leukocytes of U.S. Marines (*N* = *188*), obtained both pre- and post-deployment to conflict zones. We identified discrete groups of co-regulated genes (i.e., co-expression modules) and tested them for association to PTSD. We identified one module at both pre- and post-deployment containing putative causal signatures for PTSD development displaying an over-expression of genes enriched for functions of innate-immune response and interferon signalling (Type-I and Type-II). Importantly, these results were replicated in a second non-overlapping independent dataset of U.S. Marines (*N* = *96*), further outlining the role of innate immune and interferon signalling genes within co-expression modules to explain at least part of the causal pathophysiology for PTSD development. A second module, consequential of trauma exposure, contained PTSD resiliency signatures and an over-expression of genes involved in hemostasis and wound responsiveness suggesting that chronic levels of stress impair proper wound healing during/after exposure to the battlefield while highlighting the role of the hemostatic system as a clinical indicator of chronic-based stress. These findings provide novel insights for early preventative measures and advanced PTSD detection, which may lead to interventions that delay or perhaps abrogate the development of PTSD.

Molecular Psychiatry (2015) **00,** 000–000. doi:10.1038/mp.2015.9

INTRODUCTION

The study of the molecular factors that determine risk and subsequent development of Post-traumatic stress disorder (PTSD) are at the forefront of molecular psychiatric research. A significant number of men and women exposed to severe emotional trauma and loss emerge from these events with persistent PTSD symptoms, such as intrusive imagery, avoidance and hyperarousal, as well as other long-term physical health problems. PTSD affects 7–8% of the general United States (US) population, and is higher among troops recently returned from the wars in Iraq and Afghanistan, with estimates of prevalence as high as 20%.[1] Annual health care costs associated with PTSD in the US have been estimated to be 180 million dollars.[2] Heterogeneity in susceptibility to PTSD suggests that differences at the molecular level (i.e. gene-expression level) may influence an individual's physiological and psychological response to trauma and thus the development of PTSD. A clear understanding of the molecular mechanisms underlying this aberrant response to trauma is required to reduce the substantial morbidity and mortality associated with this disorder.

A number of studies have analyzed blood gene expression and glucocorticoid activity to build more effective models for identifying molecular factors associated to PTSD.[3–12] These studies were recently reviewed by Heinzlemann and Gill,[2] who summarized that the increased expression of inflammatory genes and decreased expression of the genes that regulate inflammation contribute to the onset of PTSD. Specifically, when considering the overlap in results from transcriptomic studies, the decreased expression of *FKBP5* and *STAT5B*, which regulate inflammation, is evident.[4,6,7,9] The majority of these reviewed studies[3–8,11,12] centered transcriptomic analyses on subjects already diagnosed with PTSD, and thus lacked a prospective study design, as well as independent datasets for validation purposes. These studies employ gene expression analysis on pre-determined targets, focusing analyses on the individual gene-level and the putative clinical utilities of the resulting gene-list, without studying the connectivity of these genes at the system-level.

Recent gene-expression network analyses, such as weighted gene co-expression network analysis (WGCNA), aim to integrate expression data across thousands of genes into a higher-order

Q1 [1]Clinical and Experimental Sciences, Faculty of Medicine, University of Southampton, UK; [2]Department of Psychiatry, University of California San Diego, California, USA; [3]Psychiatric Genetic Epidemiology and Neurobiology Laboratory (PsychGENe Lab), Departments of Psychiatry and Behavioral Sciences and Neuroscience and Physiology, Medical Genetics Research Center, SUNY Upstate Medical University, Syracuse, New York, USA; [4]Veterans Affairs Center of Excellence for Stress and Mental Health, San Diego, California, USA; [5]Veterans Affairs San Diego Healthcare System, San Diego, California, USA; [6]Institute of Genomic Medicine, University of California, San Diego, La Jolla, California, USA; [7]Center for Behavioral Genomics, Department of Psychiatry, University of California San Diego, California, USA and [8]Departments of Medicine and Pharmacology, University of Q2 California San Diego, California, USA. Correspondence: MS Breen, University of Southampton, Faculty of Medicine, Room LE57, MP813, Southampton General Hospital, Tremona Road, Southampton, SO16 6YD, Hampshire, USA or C Nievergelt, School of Medicine, Department of Psychiatry, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093 0737, USA.
E-mail: msb1g13@soton.ac.uk or cnievergelt@ucsd.edu

[9]These authors contributed equally to this work.

Received 22 August 2014; revised 25 November 2014; accepted 19 December 2014

system-level context to identify groups of genes within a network whose expressions are highly correlated (i.e. co-expression modules).[13] In doing so, WGCNA provides a powerful unsupervised approach to tackle the molecular complexity that occurs in neurodevelopmental and psychophysiological disorders,[14–19] although has never before been applied to PTSD.

We applied WGBNA to RNA-Seq and microarray peripheral blood leukocyte (PBL) gene expression taken from two independent groups of U.S. Marines, both pre- and post-deployment to conflict zones. The primary goal of this analysis was to best characterise the prognostic and diagnostic molecular signatures defining both 'PTSD risk' and 'PTSD' states, while demonstrating the robustness and reproducibility of WGCNA findings across datasets. Instead of identifying differentially expressed genes on a gene-by-gene basis, we constructed unsupervised gene co-expression networks from a combination of case and control data and identified gene co-expression modules within these networks. Modules were first assessed for containing differentially expressed genes, tested for their association with PTSD, and finally subjected to functional enrichment analysis. In this manner, we then assessed whether the PTSD-associated modules were detected in our second non-overlapping dataset of U.S. Marines to demonstrate a significant and consistent association of our findings. We conclude that prospectively profiling the transcriptome of U.S. Marines pre- and post-deployment to conflict zones, using a co-expression analysis approach is a promising strategy for identifying and studying the functions of causal and consequential molecular factors in PTSD development, with particular value in reproducing results across independent datasets of U.S. Marines.

SUBJECTS AND METHODS

Sample collection and datasets

All subjects were male and participants in the Marine Resilience Study (MRS), a prospective study of well-characterized U.S. Marines scheduled for combat deployment to Iraq or Afghanistan, with longitudinal follow-up to track the effect of combat stress.

Dataset 1—Whole blood was obtained from 124 U.S. Marines who served a seven month deployment. Blood was drawn 1-month prior to deployment and again at 3-months post-deployment for each participant. Each blood sample (10 ml) was collected into an EDTA-coated collection tube, RNA was isolated from peripheral blood leukocytes using LeukoLOCK Total RNA isolation and sequenced using the Illumina Hi-Seq 2000.

Dataset 2—For validation, data were compared to an independently generated gene expression data-set from a separate, non-overlapping, group of 50 MRS Marine participants (*Glatt et al. 2013*, previously published pre-deployment data[12]). Blood samples were treated in an identical fashion as described above, however final RNA was hybridized to the Affymetrix Hu-Gene 1.0 ST Array.

PTSD diagnosis

At the time of each blood draw, PTSD symptoms were assessed using a structured diagnostic interview, the Clinician Administered PTSD Scale (CAPS).[20–23] Using the criteria from the Diagnostic and Statistical Manual of Mental Disorders, fourth edition (2000),[24] diagnosis for partial or full PTSD was defined as a threat to life, injury, or physical integrity (Criterion A1) and the presence of at least one re-experiencing symptom and either three avoidance symptoms or two hyperarousal symptoms, or two avoidance symptoms plus two hyperarousal symptom.[25–27] Symptoms must have occurred at least once within the past month (frequency $\geqslant 1$) and caused a moderate amount of distress (intensity $\geqslant 2$).

Subject selection

A subset of MRS study participants were pre-selected for RNA-Seq analysis. First, at pre-deployment, all participants had to be symptom free, with no PTSD diagnosis and a CAPS $\leqslant 25$. Second, at post-deployment, participants who fulfilled criteria for partial or full PTSD diagnosis were designated the PTSD group. Third, participants with post-deployment CAPS $\leqslant 25$ that matched the post-deployment PTSD group on variables of combat exposure, age and ethnicity were designated the 'control' group. Under these criteria, all paired subjects were stratified into two groups based upon CAPS scores at 3-months post-deployment (Table 1, Supplementary Table 1). If a Marine participant developed PTSD following trauma-exposure at 3-months post-deployment, their pre-deployment sample would be included in the 'PTSD-risk' group. Likewise, if a subject avoided PTSD symptoms at 3 months post-deployment their sample at pre-deployment was included in the 'control' group.

Table 1. Recorded clinical parameters from U.S. Marines assessed at pre- and post-deployment for Dataset 1

Time point	*Pre-Deployment*			*Post-Deployment*		
	PTSD Cases (N = *47)*	*Controls* (N = *47)*	*P-value*	*PTSD Cases* (N = *47)*	*Controls* (N = *47)*	*P-value*
Age	22.15 ± 2.53	22.42 ± 3.92	0.682	23.14 ± 2.52	23.42 ± 3.92	0.685
Alcohol	2.08 ± 1.55	1.62 ± 1.33	0.124	1.79 ± 1.32	1.54 ± 1.11	0.318
Tobacco	1.75 ± 1.62	0.97 ± 1.51	0.02	1.69 ± 1.69	1.02 ± 1.47	0.042
WC adj.	1.65 ± 0.13	1.72 ± 0.13	0.015	1.68 ± 0.14	1.75 ± 0.12	0.012
PCL	21.29 ± 4.72	18.33 ± 2.27	0.0001	42.38 ± 11.09	20.94 ± 3.87	5.37E-22
CAPS total	11.39 ± 7.23	6.75 ± 6.90	0.002	53.17 ± 15.08	10.04 ± 7.26	5.99E-32
CAPSBs	1.00 ± 1.91	0.54 ± 1.92	0.245	14.9 ± 7.25	1.54 ± 2.37	6.29E-21
CAPSCAs	0.54 ± 1.11	0.10 ± 0.51	0.015	5.31 ± 4.57	0.85 ± 2.08	1.88E-08
CAPSCN1s	1.10 ± 2.23	0.97 ± 2.88	0.813	9.17 ± 5.32	1.19 ± 2.87	1.21E-14
CAPSDs	8.39 ± 5.66	4.58 ± 4.98	0.001	22.6 ± 6.7	6.42 ± 4.79	5.97E-24
CAPSCs	2.00 ± 2.73	1.62 ± 3.66	0.571	15.67 ± 7.23	2.08 ± 3.66	7.15E-20
Prior Deployment	19	16	0.6699	—	—	—
TBI	—	—	—	30	21	0.097
CES PBE mean	—	—	—	0.63 ± 0.25	0.53 ± 0.12	0.02
Caucasian	26	26	1	—	—	—
African American	4	4	1	—	—	—
Native American Mexican	13	15	0.822	—	—	—
Asian & Other	5	3	0.714	—	—	—

Abbreviations: Alcohol, alcohol consumption; CAPS total, CAPS total score; CAPSBs, re-experiencing subscale; CAPSCAs, symptoms of avoidance; CAPSCN1s, symptoms of numbing; CAPSCs, subtotal C subscale; CAPSDs, hyper-arousal subscale; CES, combat exposure scale; PBE, post battle experience; PCL, PTSD symptom check list; TBI, traumatic brain injury; Tobacco, tobacco use; WC adj., waist circumference was adjusted for height; -, not applicable. Significance was assessed with a Student's two-tailed *t* test for continuous variables and fishers exact test of proportions for binary variables. (Average ± standard deviation).

Data pre-processing

All data were pre-processed by normalization, filtering genes with low expression values, and removing any outliers which may bias down-stream analysis. Final subject numbers resulted in 94-paired subjects (47 paired cases and 47 paired controls) in *Dataset 1* and 48 paired subjects (24 paired cases and 24 paired controls) in *Dataset 2*. To compare findings from RNA-Seq data in *Dataset 1* to microarray data in *Dataset 2*, genes found only on both platforms ($N = 10\,184$) passed into our subsequent analysis (see Supplementary File for more detailed information).

Differential gene expression analyses

Differentially expressed genes were assessed using the moderated *t*-test in edgeR[28] and LIMMA[29] packages for RNA-Seq and microarray data, respectively, and unless otherwise specified, the significance threshold was a nominal *P*-value < 0.05. A nominally significant *P*-value was used to yield a reasonable number of genes to include within network analyses. Differential expression analyses were performed on 10 184 genes between pre-deployment PTSD case and control groups, and again between post-deployment PTSD case and control groups (see Supplementary File for more detailed information).

Gene network construction and module detection

Signed co-expression networks were built using weighted gene co-expression network analysis (WGCNA)[13] in R. A total of 10 184 genes were used to construct each network. To construct the networks, the absolute values of Pearson correlation coefficients were calculated for all possible gene pairs and resulting values were transformed so that the final matrix followed an approximate scale-free topology (see Supplementary File for detailed information). The WGCNA dynamic tree-cut algorithm was used to detect network modules. In order to determine which modules, and corresponding processes were most associated to PTSD related states, we ran singular value decomposition on each module's expression matrix and used the resulting module eigengene (ME), which is equivalent to the first principal component,[13] to represent the overall expression profiles for each module. For each gene in a module, module membership (*kME*) was defined as the correlation between gene expression values and *ME* expression. Genes with high *kME* inside co-expression modules are labeled as hub genes.[13] *GS* was calculated as the $-\log_{10}$ of the *P*-value generated for each gene within a particular module using a moderated *t* test and is a measure of the strength of differential gene expression between PTSD cases and controls. *MS* was calculated as the average *GS* within each module (see Supplementary File for more information).

Statistical analyses

All gene-set overlap analyses were performed by assessing the cumulative hypergeometric probability using the *phyper* function in R.

Enrichment analyses

Module enrichment was assessed three ways. First, general module enrichment categories were obtained using gene ontology biological processes from the DAVID database[30] (http://david.abcc.ncifcrf.gov/). Second, specific module enrichment categories were obtained using the WGCNA function userlistEnrichment[31] using modules as input-lists and curated Reactome NCBI Biosystems pathways and terms[32] as user-defined lists. Finally, we downloaded the highly expressed, cell specific (HECS) gene expression database compiled by Shoemaker *et al.*[33] to assess cell-type specific enrichment results, here cell-type marker lists were used as a user-defined lists. All module genes were used for enrichment analyses using a FDR corrected *P*-value < 0.05 as significant.

Data availability

RNA-Sequencing and microarray gene expression data for *Dataset 1* and *Dataset 2* are freely available at the Gene Expression Omnibus. (http://www.ncbi.nlm.nih.gov/geo/). RNA-Seq gene expression data from *Dataset 1* can be found under accession number GSEXXXX and microarray gene expression data from *Dataset 2* can be found under accession number GSEXXXX.

Full Methods and any associated references are available in Supplementary Methods.

RESULTS

We analyzed two different gene expression datasets generated from RNA-Seq (*Dataset 1*, Table 1) and microarray (*Dataset 2*, Supplementary Table 1) using peripheral blood leukocyte (PBL) samples taken from U.S. Marines pre- and post-deployment. We aimed to characterise the prognostic and diagnostic molecular signatures of PTSD by studying transcriptional differences at the systems-level at pre-deployment and post-deployment separately. Initially, WGCNA was used in *Dataset 1* to assess module preservation between PTSD cases ($N = 47$) and controls ($N = 47$) for the pre- and then the post-deployment time point (see Supplementary File for complete description). This analysis identifies large differences in gene co-regulatory patterns, as being disrupted or created in PTSD cases relative to controls, or vis-versa. However, we observed strong preservation statistics between the two groups indicating similar fundamental gene co-regulation within PTSD cases and controls, suggesting that major changes in the underlying gene-gene connectivity are not a basis for the pathology of this disorder (Supplementary Table 2). As a result we used the higher confidence and completeness of a combined network of case and control data.

Differential module expression post-deployment in Dataset 1

We constructed a gene co-expression network from a combination of PTSD cases ($N = 47$) and controls ($N = 47$) post-deployment using RNA-Seq expression data from *Dataset 1* (Figure 1). This analysis identified nine modules (fully characterised in Supplementary Table 3) that were first examined for enrichment of differentially expressed genes. Two modules (M1A and M1B) were enriched for genes identified as differentially expressed between PTSD cases and controls, reflected by an elevated module significance (*MS*) value (Figure 2a). To determine if the overall expression of modules M1A and M1B were significantly associated with PTSD group status, we calculated differences in module expression using module eigengene (*ME*) values *(See Materials and Methods for complete description of ME)*. Consistent with results using MS, expression of module M1B was significantly higher in the PTSD resilient control group ($P = 0.004$ and Figure 2b) suggesting a positive correlation to PTSD resiliency, meanwhile expression of module M1A was significantly higher in the PTSD

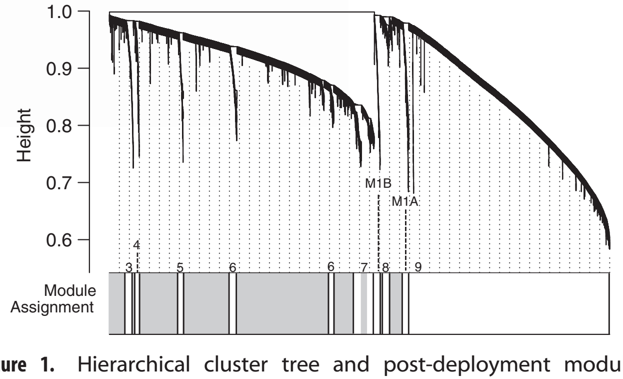


Figure 1. Hierarchical cluster tree and post-deployment module structure in *Dataset 1*. Hierarchical cluster tree (dendrogram) of the combine post-deployment network of PTSD cases ($N = 47$) and controls ($N = 47$) comprising 10 184 genes. Each line represents a gene (leaf) and each low-hanging cluster represents a group of co-expressed genes with similar network connections (branch) on the tree. The first band underneath the tree indicates the nine detected, and subsequently analyzed, network modules. Genes shaded in grey were not assigned to a particular module and represent background noise. For a comprehensive functional annotation of each module and calculation of all significant module-trait relationships see Supplementary Table 3.

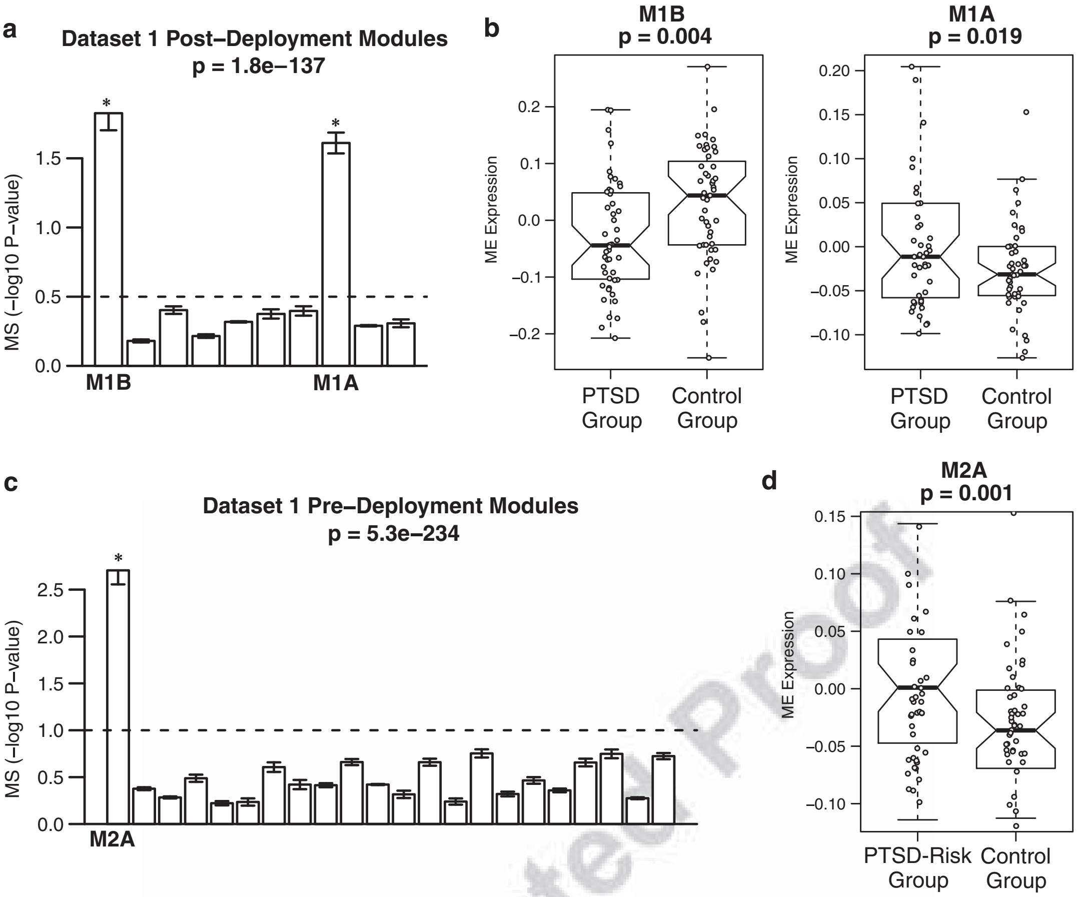


Figure 2. Module significance (MS) and module eigengene (ME) expression boxplots. MS was measured across all pre- and post-deployment modules in Dataset 1. WGCNA detected ten modules post-deployment from a combination of PTSD cases and control (**a**) and twenty-two modules at pre-deployment from a combination of PTSD risk cases and controls (**c**). The y-axis indicates MS by calculating the average $-\log_{10}$ *P*-values, generated by a moderated *t* test, for each gene within a particular module, when assessing differential expression between PTSD cases and controls. Here, a kruskal-wallis *P*-value was used only for descriptive purposes and not inferential. Modules denoted with an asterisk (*) have ME values significantly correlated to conditional states (i.e. PTSD cases or controls). Representative modules with high MS at post-deployment and pre-deployment were investigated for module expression differences. Differences in ME expression were measured using a two-tailed student's *t* test on and a *P*-value < 0.05 is considered significant. Boxplots are displayed for each main group. Significant differences in ME expression were observed in post-deployment modules M1B and M1A **(b)** and in pre-deployment module M2A **(d).**

group ($P = 0.02$, Figure 2b). Subsequently, *ME* values for each module were correlated to all clinical parameters, found in Table 1, to determine module-trait relationships. The *ME* for module M1B was significantly correlated to post-deployment PTSD resilient controls ($r = 0.29$, $P = 0.005$), negatively correlated to post-deployment CAPs and PCL (CAPs, $r = -0.27$, $P = 0.009$; PCL $r = -0.28$, $P = 0.007$) and negatively correlated other measures of CAPS (Supplementary Table 3) but not correlated to any other measured clinical variable, suggesting that differential gene expression in M1B was not confounded by recorded measurements such as body-mass-index, smoking, or alcohol consumption. Genes in M1B were expressed to a greater extent in PTSD resilient controls (Figure 2b) while enrichment analysis revealed a significant association with hemostasis, platelet activation and wound healing (Figure 3a). Further, enrichment for cell-type specificity revealed on over-representation of erythroid expression markers (blood platelets). Hub genes are those most strongly correlated to the *ME* value for a particular module and represent possible disease associated markers,[13] in this case putative PTSD-resiliency markers. The top 5 hub genes in M1B (*C6orf25, CTDSPL, ITGB3, PRKAR2B* and *TUBB1*) were are all associated with hemostasis and in particular, with platelet regulation and function[34–37] (Figure 3b).

The *ME* for module M1A was significantly correlated to PTSD cases ($r = 0.23$, $P = 0.03$), post-deployment CAPs criteria of avoidance (CAPSCA, $r = 0.32$, $P = 0.002$) and post-deployment CAPs criteria of re-experiencing (CAPSBs, $r = 0.2$, $P = 0.05$) but to no other variables (Supplementary Table 3). Genes in M1A were over-expressed in PTSD cases (Figure 2b) while enrichment analysis revealed a significant association with immune response as exemplified by innate responses mediated by interferon (IFN) signalling (Figure 3c), as well as with monocyte specific markers. The top 5 hub genes in M1A included *IFI35, IFIH1, PARP14, RSAD2 and UBE2L6*; all well described interferon stimulated genes[38] and here considered putative PTSD-associated markers (Figure 3d).

Differential module expression pre-deployment in Dataset 1

It is unclear whether the modules identified post-deployment are causal of PTSD development or are simply a consequence of the disorder. To determine if any post-deployment modules could be re-identified and thus associated as causal modules, we constructed a gene co-expression network combining RNA-Seq gene expression data from PTSD-risk cases ($N = 47$) and controls ($N = 47$) pre-deployment in *Dataset 1*. Twenty-two pre-deployment

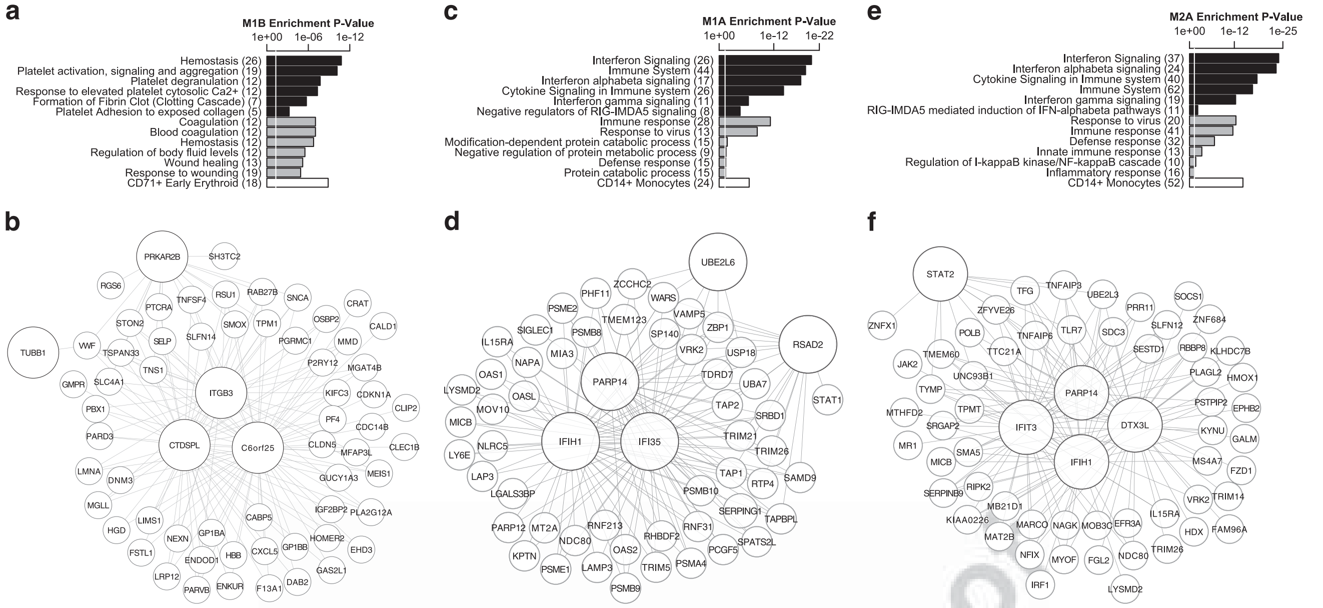


Figure 3. Module characterization for *Dataset 1*. Enrichment analysis and correlation networks for modules M1B (**a** & **b**) and M1A (**c** & **d**) identified post-deployment, and module M2A (**e** & **f**) identified pre-deployment in *Dataset 1*. Enrichment analysis was used to identify the top 6 REACTOME ontology terms (black bars), the top 6 DAVID ontology terms (grey bars) and the most significant cell-type signature (white bar) over-represented in the list of genes within each module. All terms were deemed significant as assessed by a hypergeometric test FDR corrected *P*-value < 0.05 displayed as a white line. The total number of genes within each significant term is denoted within the brackets associated with that term. Gene-networks were constructed selecting the top 150 most significant connections ranked by *kME*. Nodes represent genes and edges represent correlations. The top 5 hub genes, those most correlated to *ME* values, are shown in larger sizes.

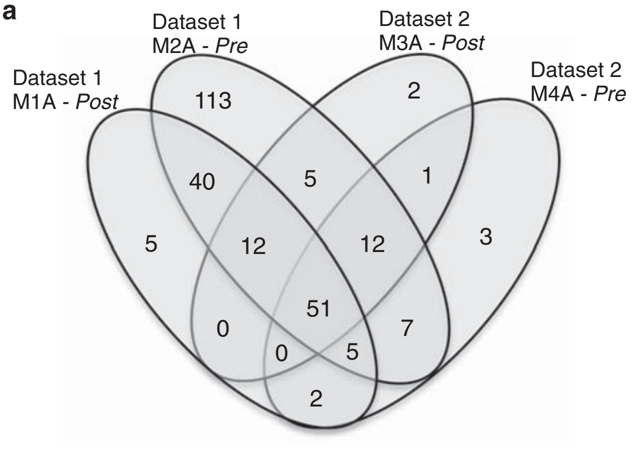
modules were identified (fully characterised in Supplementary Table 4) whereby a single module (M2A) was enriched for differentially expressed genes between PTSD-risk participants and controls as reflected by an elevated MS value (Figure 2c). Along the same lines, M2A module expression was significantly higher in the PTSD risk group ($P = 0.001$ and Figure 2d). Module M2A *ME* was significantly correlated to one variable, PTSD-risk ($r = 0.32$, $P = 0.002$, Supplementary Table 4). Similar to module M1A that was identified post-deployment, enrichment analysis of genes in M2A revealed a significant association with innate immune responses, IFN signalling and monocyte specificity (Figure 3e). The top 5 hub genes were again associated with IFN signalling (*DTX3L, IFIH1, IFIT3, PARP14* and *STAT2)* (Figure 3f). Gene-set overlap analysis compared all of the genes in M2A pre-deployment ($n = 245$) to those in M1A post-deployment ($n\ = 115$) to reveal a significant overlap (∩ = 108, *P = 6.7e-181*, Figure 4).

Validation of differential module expression post-deployment in Dataset 2

To validate post-deployment findings in *Dataset 1* we assessed *Dataset 2* for similar network properties in a combined network analysis of PTSD cases ($N = 24$) and controls ($N = 24$) post-deployment. Out of 8 modules (full characterisation Supplementary Table 5), a single module (M3A) contained an enrichment of differentially expressed genes (Supplementary Figure 2A) demonstrating a modest, yet insignificant, increase in module expression within the PTSD group ($P = 0.1$, Supplementary Figure 2B). The *ME* was significantly correlated to post battle experience ($r = 0.4$, $P = 0.004$), post-deployment CAPS ($r = 0.32$, $P = 0.03$) and weakly correlated to a PTSD cases ($r = 0.21$, $P = 0.1$, Supplementary Table 5). The genes in this module were over-expressed in PTSD cases relative to controls (Supplementary Figure 2B) and enrichment analysis revealed a significant association with innate immune responses, IFN signalling and monocytes (Supplementary Figure 3A). The top 5 hub genes (*DDX58, IFI35, IFIT5, PARP9 and ZBP1*) were again all associated with IFN signalling (Supplementary Figure 3B). A highly significant overlap in post-deployment module genes across M1A ($n = 115$) in *Dataset 1* and M3A ($n = 83$) in *Dataset 2* (∩ = 63, $P = 2.0\text{E-}105$, Figure 4b) confirmed the identification of a dysregulated innate immune module related to PTSD cases across two independent datasets.

Validation of differential module expression pre-deployment in Dataset 2

To re-confirm pre-deployment findings from *Dataset 1*, PTSD-risk cases ($N = 24$) and controls ($N = 24$) pre-deployment were combined from *Dataset 2* and subjected to network analysis which identified 11 modules (full characterisation in Supplementary Table 6). A single module (M4A) was enriched for differentially expressed genes between PTSD-risk cases and controls (Supplementary Figure 2C). The PTSD-risk group displayed a significant over-expression of module expression ($P = 0.01$, Supplementary Figure 2D). The *ME* for M4A was significantly correlated to PTSD-risk ($r = 0.36$, $P = 0.01$) and CAPs ($r = 0.44$, $P = 0.002$, Supplementary Table 6). Moreover, enrichment analysis of M4A revealed a significant association with innate immune responses, IFN signalling and monocytes (Supplementary Figure 3C), and the top 5 hub genes (*PARP9*, *UBE2L6*, *STAT2*, *TRIM22* and *GBP1*) were again all associated with IFN signalling (Supplementary Figure 3D). All pairwise gene-set overlap analyses across modules M1A, M2A, M3A and M4A revealed a highly significant overlap (Figure 4b) and hub gene expression for these modules showed elevated expression in PTSD groups when compared to controls both pre- and post-deployment across both datasets (Supplementary Figure 4). These results demonstrate the association of a dysregulated innate immune module, related to IFN signalling, which appears to define at least part of the pathophysiology of PTSD through causal association to PTSD development.



b

	M1A	M2A	M3A	M4A
M1A	-	*6.7E-181*	*2.0E-105*	*1.0E-134*
M2A	**108** ∩	-	*6.3E-134*	*2.4E-121*
M3A	**63** ∩	**80** ∩	-	*8.8E-152*
M4A	**58** ∩	**75** ∩	**69** ∩	-

c

	kME Rank			
Gene Symbol	M1A	M2A	M3A	M4A
IFIH1	**3**	**1**	12	**7**
STAT2	**6**	**2**	19	**3**
PARP14	**4**	**3**	22	40
DTX3L	39	**4**	46	39
IFIT3	**10**	**5**	**8**	**10**
IFI35	**2**	**6**	**4**	52
UBE2L6	**1**	**7**	18	**2**
PARP9	47	**8**	**2**	**1**
TRIM22	31	**9**	14	**4**
DDX58	36	**10**	**3**	26
TRIM5	34	11	41	16
CMPK2	34	12	51	34
IFIT5	17	13	**1**	21
RSAD2	**5**	14	35	48
HERC5	11	15	17	14
IFI6	13	17	32	43
OAS3	15	18	15	44
IRF9	25	19	44	62
IFIT2	37	20	24	45
IFIT1	16	25	26	49
SERPING1	30	26	34	**6**
STAT1	21	27	30	23
GBP1	43	28	**10**	**5**
IFI44	24	32	**9**	20
SAMD9L	14	33	13	12
PML	**8**	34	28	32
ZBP1	51	35	**5**	38
APOL6	31	36	36	11
APOL1	30	37	67	75
MX1	38	38	**7**	37
IFI44L	28	39	58	65
DDX60	37	40	16	28
BATF2	32	43	83	50
OASL	40	44	62	69
EPSTI1	40	45	42	25
FBXO6	42	47	56	15
LAP3	70	50	11	41
OAS2	46	52	21	18
TAP1	63	58	31	**8**
PARP12	33	63	23	27
RTP4	55	65	27	13
TAP2	45	67	53	70
SPATS2L	67	69	64	81
CXCL10	35	70	57	68
LY6E	65	75	37	64
OAS1	74	84	43	66
DHX58	38	93	47	29
USP18	66	97	82	82
CD274	33	121	55	30
MOV10	94	123	60	31
ETV7	41	128	79	74
GBP5	44	151	38	**9**

Figure 4. Venn Diagram of Innate Immune Modules across *Dataset 1* and *Dataset 2*. Venn Diagram (**a**) depicting significant overlap in genes belonging to modules M1A post-deployment and M2A pre-deployment in *Dataset 1* as well as modules M3A post-deployment and M4A pre-deployment in *Dataset 2*. Gene overlap (∩) with associated hypergeometric *P*-value, in italics, are depicted for all pairwise comparisons of module genes (**b**). The overlap identified 51 genes found across all four analyses (**c**) which are displayed in the table along with the corresponding kME rank (i.e. rank of connectivity) for each gene within a particular module. A high rank indicates hub gene status (i.e. PTSD risk and PTSD associated markers). Numbers in bold outline the top 10 hub genes across each module, respectively. Genes are ordered accordingly to M2A kME. All 51 genes are displayed via heatmap in Supplementary Figure 4.

DISCUSSION

We investigated the high-order system-level properties of PTSD using an unsupervised network-based approach (WGCNA) to identify differences at the gene co-expression level, rather than investigating at the individual gene level. Gene expression data were generated by RNA-Seq (*Dataset 1*) and microarray (*Dataset 2*) using PBL samples isolated from U.S. Marines pre- and post-deployment to conflict zones (i.e. Iraq and Afghanistan). Our comprehensive and prospective experimental design allowed the investigation of both biological processes that define PTSD and those driving the development of this disorder, and further, allowed the re-confirmation of findings in an independent dataset. This is the first time dysregulated gene networks specific for innate immunity have been used to characterise causal and consequential molecular signatures of PTSD and then to further replicated these findings across independent datasets.

A novel finding from our network analyses was the identification of modules related to hemostasis and wound responsiveness expressed to a greater extent post-deployment in US Marines who did not develop PTSD (Figure 2b), as in module M1B (Figure 3a). Interestingly, the three other network analyses also detected modules related to hemostasis and wound response with significant overlap (M16 pre-deployment *Dataset 1*; M7 and M6 indented post- and pre-deployment in *Dataset 2*; Supplementary Figure 5, Supplementary Tables 4). These other modules revealed patterns of heterogeneous gene expression irrespective of group status and time-point suggesting that these modules and corresponding processes may infer wound resilience in only a small subset of individuals. Along these lines, it has been well documented that different degrees of stress will elicit different stress responses (review[39]), and in particular, a response involving blood platelets, has been shown to be a critical biomarker of hemostatic, thrombotic, and inflammatory challenges to an organism and a key player in cardiovascular disease and chronic based stress, as in PTSD.[40,41] Moreover, in a review of a large number of studies examining various tissue types, it was found that different types of psychological stress were associated with impaired wound healing.[42] A meta-analysis found an inverse correlation ($r = -0.42$) between psychological stress and wound healing[43] supporting the positive association between wound healing and PTSD resilience ($r \ = 0.29$, $P = 0.005$) found in this study. This suggests that high levels of stress may hinder proper wound healing during/after battlefield trauma, although the degree of such stress appears to be a key factor for establishing associations with the hemostatic system. Our central finding was the identification of a dysregulated innate immune module associated with the development of PTSD (Figures 2 and 3, Supplementary Figure 3), illuminated by the replication of modules post-deployment (M1A and M3A) and those pre-deployment (M2A and M4A) that could be associated with PTSD. These findings suggest that differences in innate immunity modules were not simply a consequence of the PTSD state post-deployment but also have causal relevance for PTSD development and explain at least part of the pathophysiology of the disorder, exemplified by their identification pre-deployment. These results highlight our differential expression analyses (Supplementary Figure 1) and our previous reports of C-reactive protein (CRP), a general marker of immune activation and inflammation, and 5′-oligoadenylate synthetase genes (*i.e. OAS1, OAS2, OAS3)* as markers of the antiviral interferon response, that were associated with an increased risk of developing PTSD.[44,12] However, our current findings dramatically extend these results by showing that the IFN response is being modulated to a much greater extent than previously thought pre- and post-deployment. Notably, a number of single case studies have reported that treatment of hepatitis C virus (HCV) infected PTSD subjects with recombinant interferon (IFN- α2b) precipitated PTSD symptoms.[45,46] In our study, where subjects were not receiving IFN therapy, it is unclear what is stimulating the IFN response.

Our observations lead to several fundamental questions and some putative solutions. First, how does one interpret the over-expression of innate immunity genes found prior-to trauma? One possible explanation is that both acute and severe stress, predictors in their own right for PTSD, are also associated with the hyper-activation of the immune system and subsequent inflammation.[47,48] An alternative hypothesis is that stress, pathogens and/or high viral loads may 'prime' the immune system, driving the IFN response, altering a subsequent response to trauma. Along these lines, studies focusing on the gut-brain barrier have shown that intestinal mucosal dysfunction, defined as increased translocation of gram-negative bacteria (leaky gut), plays a role in the inflammatory pathophysiology of depression suggesting that differences in gut flora may stimulate an IFN response.[49] Second, does a dysregulated innate immune module pre-deployment hold predictive value? Our previous work constructing a prognostic classifier from *Dataset 2* pre-deployment participants[12] suggests that immune-related genes do hold predictive value although these results have not yet been replicated across larger datasets using machine-learning methods. Inferring the prognostic relevance of network-based applications remains challenging. However, cross-referencing our findings with this previous work suggests that network statistics, and our innate immune modules, do have potential to contain predictive value. Third, out of the entire network of pairwise correlations between genes across the transcriptome, are the most informative genes interconnected within similar modules or spread out across numerous modules? A possible limitation of this study was that by analyzing co-regulated modules of genes we may have missed individual genes, which do not correlate within our modules of interest although are of functional relevance to PTSD. For example, previous reports specifically target *FKBP5* and *STAT5B* as differentially expressed biomarkers[3–8,11,12] although they were not assigned to co-expressed modules nor found to be significantly differentially expressed between PTSD cases and controls. Finally, of what relevance is PBL gene expression for a disorder primarily associated with the brain? In this study we identify innate immunity and IFN signalling genes whose expression was elevated in PBLs both before and after the development of PTSD (Figure 2 and Supplementary Figure 4). Although the recruitment of such signalling could be triggered by various factors, they ultimately release toxic compounds including degradative enzymes and reactive oxygen species that can impair cellular processes.[50–53] It could be hypothesized that the accumulation of these compounds in the blood prior-to-deployment may be detrimental to the brain if the integrity of the blood-brain-barrier (BBB) was then compromised by injury (e.g. TBI). An increasing body of evidence indicates that changes in the blood may seed pathology in the brain across various disorders. In a recent Multiple Sclerosis study, Minagar and Alexander[54] investigate the association of INF with the BBB suggesting that IFN-γ and other proinflammatory cytokines (TNF-α and IL-1β) disrupt the BBB through a variety of mechanisms. Further, Alzheimer's disease models suggest that breaches in the BBB lead to leakage into the brain of blood-borne molecules that are toxic to neurons and cause neurodegenerative changes.[55] Future studies investigating the role of the BBB in PTSD may provide a detailed explanation for a specific course of PTSD development. In summary, our data provide a global framework for previously unknown molecular aspects of PTSD and describe a new context concerning the complex pathophysiological nature of PTSD development. Specifically, modules of co-expressed genes associated with the innate immune response and IFN signalling appear to be implicated in the development of PTSD and continue to persist once the disorder is established. Modules associated with hemostasis and wound healing may contribute to resilience against developing PTSD. It is hoped that this study will lead to future work confirming the importance of differences in innate immune factors to the development of PTSD and the role of platelets in the stress response. Ideally, these findings will allow for advanced PTSD detection, which could delay or abrogate PTSD development by identifying susceptible service members prior to deployment to conflict zones by either removing the causal path (i.e. trauma exposure) or through early intervention of new therapies to modulate the interferon signature.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

We thank and extend our deepest gratitude to Dr Daniel T. O'Connor whose passion and intellect of clinical, translational, and basic research brought considerable knowledge and insight to various realms of this research project and numerous others. This work was supported by Headquarters Marine Corps funds, administered through Navy BUMED Contract # N62645-11-C-4027, for MRS-II (DGB), and this Demonstration Project (CN and DOC). Replication of findings on a non-overlapping cohort was supported in part by both the National Institute of Mental Health R21 (MH085240) and R01 (MHO85560). Further support was provided by R01 (MH085521), the Gerber Foundation, the Sidney R Baer, Jr Foundation and NARSAD: The Brain and Behavior Research Foundation. We acknowledge assistance of the MRS-II administrative core, A Patel, A De La Rosa and other members of the MRS-II Team. Likewise, we acknowledge administrative support from the Veterans Medical Research Foundation (VMRF) and valuable input from M.E. Polak, D Baldwin and A Collins who assisted in critical reading of the manuscript. We also thank the Marine and Navy volunteers for their military service and for their participation in this study.

AUTHOR CONTRIBUTIONS

DGB, CN, CHW and DOC obtained the funding for this study. AXM curated clinical information regarding all participants. SJG, DST and SDC generated microarray data. MSB participated in generating RNA-Seq data and quality testing both RNA-Seq and microarray data, designed/conducted the study and wrote the manuscript with the participation of remaining authors.

AUTHOR INFORMATION

RNA-Sequencing and microarray gene expression data for Dataset 1 and Dataset 2 will be deposited and freely available in Gene Expression Omnibus (GEO) online database upon acceptance. The authors declare no conflict of interest. Correspondence and requests for materials should be addressed to msb1g13@soton.ac.uk.

REFERENCES

1 Ramchand R, Schell TL, Karney BR, Osilla KM, Burns RM, Caldarone LB. Disparate prevalence estimates of PTSD among service members who served in Iraq and Afghanistan: possible explanations. *J. Trauma. Stress* 2010; **23**: 59–68.

2 Heinzelmann M, Gill J. Epigenetic Mechanisms Shape the Biological Response to Trauma and Risk for PTSD: A Critical Review. *Nursing Research and Practice* 2013; **2013**: 1–10.

3 Zieker J, Zieker D, Jatzko A, Dietzsch J, Niesel K, Schmitt A *et al.* Differential gene expression in peripheral blood of patients suffering from post-traumatic stress disorder. *Molecular Psychiatry* 2007; **12.2**: 116–118.

4 Yehuda R, Holsboer F, Buxbaum JD, Miller-Myhsok B, Schmeidler J, Rein T *et al.* Gene Expression Patterns Associated with Posttraumatic Stress Disorder Following Exposure to the World Trade Center Attacks. *Biological Psychiatry* 2009; **66.7**: 708–711.

5 Neylan TC, Sun B, Rempel H, Ross J, Lenoci M, O'Donovan A *et al.* Suppressed monocyte gene expression profile in men versus women with PTSD. *Brain, Behavior, and Immunity* 2011; **25.3**: 524–531.

6 Sarapas C, Cai G, Bierer LM, Golier JA, Galea S, Ising M *et al.* Genetic Markers for PTSD Risk and Resilience Among Survivors of the World Trade Center Attacks. *Disease Markers* 2011; **30.2-3**: 101–110.

7 Mehta D, Gonik M, Klengel T, Rex-Haffner M, Menke A, Rubel J *et al.* Using Polymorphisms in FKBP5 to Define Biologically Distinct Subtypes of Posttraumatic Stress Disorder: Evidence From Endocrine and Gene Expression Studies. *Archives of General Psychiatry* 2011; **68.9**: 901–910.

8 Pace TW, Wingenfeld K, Schmidt I, Meinlschmidt G, Hellhammer DH, Heim CM. Increased peripheral NF-KB pathway activity in women with childhood abuse-related posttraumatic stress disorder. *Brain, Behavior, and Immunity* 2012; **26.1**: 13–17.

9 van Zuiden M, Heijnen CJ, Maas M, Amarouchi K, Vermetten E, Geuze E *et al.* Glucocorticoid sensitivity of leukocytes predicts PTSD, depressive and fatigue symptoms after military deployment: A prospective study. *Psychoneuroendocrinology* 2012; **37.11**: 1822–1836.

10 van Zuiden M, Geuze E, Willemen HL, Vermetten E, Maas M, Amarouchi K *et al.* Glucocorticoid Receptor Pathway Components Predict Posttraumatic Stress Disorder Symptom Development: A Prospective Study. *Biological Psychiatry* 2012; **71.4**: 309–316.

11 Matić G, Milutinović DV, Nestorov J, Elaković I, Jovanović SM, Perišić T *et al.* Lymphocyte glucocorticoid receptor expression level and hormone-binding properties differ between war trauma-exposed men with and without PTSD. *Progress in Neuro-Psychopharmacology and Biological Psychiatry* 2013; **43**: 238–245.

12 Glatt SJ, Tylee DS, Chandler SD, Pazol J, Nievergelt CM, Woelk CH *et al.* Blood-based gene-expression predictors of PTSD risk and resilience among deployed marines: A pilot study. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 2013; **162.4**: 313–326.

13 Langfelder P, Horvath S. WGCNA: An R Package for Weighted Correlation Network Analysis. *BMC Bioinformatics* 2008; **9.1**: 559.13.

14 Miller JA, Oldham MC, Geschwind DH. A Systems Level Analysis of Transcriptional Changes in Alzheimer's Disease and Normal Aging. *Journal of Neuroscience* 2008; **28.6**: 1410–1420.

15 Saris C, Horvath S, van Vught PWJ, vanEs MA, Blaue HM, Fuller TF *et al.* Weighted gene co-expression network analysis of the peripheral blood from Amyotrophic Lateral Sclerosis patients. *BMC Genomics* 2009; **10.1**: 405.

16 Voineagu I, Wang X, Johnston P, Lowe JK, Tian Y, Horvath S *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* 2011; **474.7351**: 380–384.

17 Hwang Y, Kim J, Shin JY, Kim JI, Seo JS, Webster MJ *et al.* Gene expression profiling by mRNA sequencing reveals increased expression of immune/inflammation-related genes in the hippocampus of individuals with schizophrenia. *Translational Psychiatry* 2013; **3.10**: e321.

18 Chen C, Cheng L, Grennan K, Pibiri F, Zhang C, Badner JA *et al.* Two gene co-expression modules differentiate psychotics and controls. *Molecular Psychiatry* 2012; **18.12**: 1308–1314.

19 Torkamani A, Dean B, Schork NJ, Thomas EA. Coexpression Network Analysis of Neural Tissue Reveals Perturbations in Developmental Processes in Schizophrenia. *Genome Research* 2010; **20.4**: 403–412.

20 Blake DD, Weathers FW, Nagy LM, Kaloupek DG, Gusman FD, Charney DS *et al.* The development of a Clinician-Administered PTSD Scale. *J. Trauma. Stress* 1995; **8**: 75–90.

21 King DW, Leskin GA, King LA, Weathers FW. Confirmatory factor analysis of the Clinician-Administered PTSD Scale: Evidence for the dimensionality of posttraumatic stress disorder. *Psychol. Assess* 1998; **10**: 90–96.

22 Weathers FW, Keane TM, Davidson JR. Clinician-administered PTSD scale: a review of the first ten years of research. *Depress. Anxiety* 2001; **13**: 132–156.

23 Weathers FW, Ruscio AM, Keane TM. Psychometric properties of nine scoring rules for the clinician-administered posttraumatic stress disorder scale. *Psychol. Assess* 1999; **11**: 124–133.

24 American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders* 4th edition. American Psychiatric Press: Washington DC, 2000.

25 Blanchard EB, Hickling EJ, Taylor AE, Loos WR. Psychiatric morbidity associated with motor vehicle accidents. *J. Nerv. Ment. Dis.* 1995; **183**: 495–504.

26 Blanchard EB, Hickling EJ, Vollmer AJ, Loos WR, Buckley TC, Jaccard J. Short-term follow-up of post-traumatic stress symptoms in motor vehicle accident victims. *Behaviour Research and Therapy* 1995; **33**: 369–377.

27 Blanchard EB, Hickling EJ, Barton KA, Taylor AE, Loos WR, Jones-Alexander J. One-year prospective follow-up of motor vehicle accident victims. *Behaviour Research and Therapy* 1996; **34**: 775–786.

28 Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010; **26**: 139–140.

29 Smyth GK. Limma: linear models for microarray data. In Gentlemen R, Carey V, Dudoit S, Irizarry R, Huber W (ed) *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. Springer: New York, 2005, pp 397–420.

30 Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protoc* 2009; **4**: 44–57.

31 Miller JA, Cai C, Langfelder P, Geschwind DH, Kurian SM, Salomon DR, Horvath S. Strategies for aggregating gene expression data: the collapse Rows R function. *BMC Bioinformatics* 2011; **12**: 322.

32 Geer LY, Marchler-Bauer A, Geer RC, Han L, He J, He S *et al.* The NCBI BioSystems database. *Nucleic Acids Res.* 2010; **38**: D492–D496.

33 Shoemaker JE, Tiago L, Ghosh S, Matsuoka Y, Kawaoka Y, Kitano H. CTen: A Web-based Platform for Identifying Enriched Cell Types from Heterogeneous Microarray Data. *BMC Genomics* 2012; **13.1**: 460.

34 Zarbock A, Polanowska-Grabowska RK, Ley K. Platelet-neutrophil-interactions: Linking Hemostasis and Inflammation. *Blood Reviews* 2007; **21.2**: 99–111.

35 Beck F, Geiger J, Gambaryan S, Veit J, Vaudel M, Nollau P *et al.* Time-resolved characterization of cAMP/PKA-dependent signaling reveals that platelet inhibition is a concerted process involving multiple signaling pathways. *Blood* 2014; **123**: e1–e10.

36 Daly ME. Determinants of Platelet Count in Humans. *Haematologica* 2010; **96.1**: 10–13.

37 Raslova H, Kauffmann A, Sekkai D, Ripoche H, Larbret F, Robert T *et al.* Interrelation between Polyploidization and Megakaryocyte Differentiation: A Gene Profiling Approach. *Blood* 2007; **109.8**: 3225–3234.

38 Rusinova I, Forster S, Yu S, Kannan A, Masse M, Cumming H *et al.* INTERFEROME V2.0: An Updated Database of Annotated Interferon-regulated Genes. *Nucleic Acids Research* 2012; **41.D1**: D1040–D1046.

39 Pacak K. Stressor Specificity of Central Neuroendocrine Responses: Implications for Stress-Related Disorders. *Endocrine Reviews* 2001; **22.4**: 502–548.

40 Bray PF, Mckenzie SE, Edelstein LC, Nagalla S, Delgrosso K, Ertel A. The Complex Transcriptional Landscape of the Anucleate Human Platelet. *BMC Genomics* 2013; **4.1**: 1.

41 Austin AW, Wissmann T, Von Kanel R. Stress and Hemostasis: An Update. *Seminars in Thrombosis and Hemostasis* 2013; **39.08**: 902–912.

42 Walburn J, Vedhara K, Hankins M, Rixon L, Weinman J. Psychological stress and wound healing in humans: a systematic review and meta-analysis. *J Psychosom Res* 2009; **67**: 253–271.

43 Gouin J-P, Kiecolt-Glaser JK. The Impact of Psychological Stress on Wound Healing: Methods and Mechanisms. *Immunology and Allergy Clinics of North America* 2011; **31.1**: 81–93.

44 Eraly SA, Nievergelt CM, Maihofer AX, Barkauskas DA, Nilima Biswas N, Agorastos A *et al.* Assessment of Plasma C-Reactive Protein as a Biomarker of Posttraumatic Stress Disorder Risk. *JAMA Psychiatry* 2014; **71.4**: 423.

45 Maunder RG, Hunter JJ, Feinman SV. Interferon Treatment of Hepatitis C Associated With Symptoms of PTSD. *Psychosomatics* 1998; **39.5**: 461–464.

46 Dieperink E, Leskela J, Dieperink ME, Evans B, Thuras P, Ho SB. The Effect of Pegylated Interferon-α2b and Ribavirin on Posttraumatic Stress Disorder Symptoms. *Psychosomatics* 2008; **49.3**: 225–229.

47 Butcher SK, Lord JM. Stress Responses and Innate Immunity: Aging as a Contributory Factor. *Aging Cell* 2004; **3.4**: 151–160.

48 Clark SM, San J, Francis TC, Nagaraju A, Michael KC, Keegan AD *et al.* Immune status influences fear and anxiety responses in mice after acute stress exposure. *Brain, behavior, and immunity* 2014; **38**: 192–201.

49 Maes M, Kubera M, Leunis J. The gut-brain barrier in major depression: Intestinal mucosal dysfunction with an increased translocation of LPS from gram negative enterobacteria (leaky gut) plays a role in the inflammatory pathophysiology of depression. *Nueroendocrinology Letters* 2008; **29**: 117–124.

50 Aiboshi J, Moore EE, Ciesla CJ, Silliman CC. Blood transfusion and the two-insult model of post-injury multiple organ failure. *Shock* 2001; **15**: 302–306.

51 Veldhuis TB, Floris T, van der Meide PH, Vos IM, de Vries HE, Dijkstra CD *et al.* Interferon-beta prevents cytokine-induced neutrophil infiltration and attenuates blood–brain barrier disruption. *J. Cerebral Blood Flow Metab* 2003; **23**: 1060–1069.

52 Bhatia M, Moochhala S. Role of inflammatory mediators in the pathophysiology of acute respiratory distress syndrome. *J. Pathol.* 2004; **202**: 145–156.

53 Giannoudis PV. Current concepts of the inflammatory response after major trauma: an update. *Injury* 2003; **34**: 397–404.

54 Minagar A, Alexander JS. Blood-brain Barrier Disruption in Multiple Sclerosis. *Multiple Sclerosis* 2003; **9.6**: 540–549.

55 Carmeliet P, Strooper BD. Alzheimer's Disease: A Breach in the Blood–brain Barrier. *Nature* 2012; **485.7399**: 451–452.

Supplementary Information accompanies the paper on the Molecular Psychiatry website (http://www.nature.com/mp)

ARTICLE IN PRESS

Brain, Behavior, and Immunity xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Brain, Behavior, and Immunity

journal homepage: www.elsevier.com/locate/ybrbi

Acute psychological stress induces short-term variable immune response

Michael S. Breen [a,*,1], Nadejda Beliakova-Bethell [b,1], Lilianne R. Mujica-Parodi [c], Joshua M. Carlson [d], Wayne Y. Ensign [e], Christopher H. Woelk [a,2], Brinda K. Rana [f,g,*,2]

[a] *Clinical and Experimental Sciences, Faculty of Medicine, University of Southampton, Southampton SO16 6YD, UK*
[b] *Department of Medicine, University of California San Diego, La Jolla, CA 92093, USA*
[c] *Department of Biomedical Engineering, State University of New York at Stony Brook, Stony Brook, NY 11794-5281, USA*
[d] *Department of Psychology, Northern Michigan University, Marquette, MI 49855, USA*
[e] *Space and Naval Warfare Systems Center – Pacific, Applied Sciences Division, San Diego, CA 92152, USA*
[f] *Department of Psychiatry, University of California San Diego, La Jolla, CA 92093, USA*
[g] *VA San Diego Center for Stress and Mental Health, La Jolla, CA 92093, USA*

ARTICLE INFO

Article history:
Received 20 August 2015
Received in revised form 1 October 2015
Accepted 13 October 2015
Available online xxxx

Keywords:
Immune system
Periphery
Leukocytes
Skydive
Brain
Gene co-expression network

ABSTRACT

In spite of advances in understanding the cross-talk between the peripheral immune system and the brain, the molecular mechanisms underlying the rapid adaptation of the immune system to an acute psychological stressor remain largely unknown. Conventional approaches to classify molecular factors mediating these responses have targeted relatively few biological measurements or explored cross-sectional study designs, and therefore have restricted characterization of stress–immune interactions. This exploratory study analyzed transcriptional profiles and flow cytometric data of peripheral blood leukocytes with physiological (endocrine, autonomic) measurements collected throughout the sequence of events leading up to, during, and after short-term exposure to physical danger in humans. Immediate immunomodulation to acute psychological stress was defined as a short-term selective up-regulation of natural killer (NK) cell-associated cytotoxic and IL-12 mediated signaling genes that correlated with increased cortisol, catecholamines and NK cells into the periphery. In parallel, we observed down-regulation of innate immune toll-like receptor genes and genes of the MyD88-dependent signaling pathway. Correcting gene expression for an influx of NK cells revealed a molecular signature specific to the adrenal cortex. Subsequently, focusing analyses on discrete groups of coordinately expressed genes (modules) throughout the time-series revealed immune stress responses in modules associated to immune/defense response, response to wounding, cytokine production, TCR signaling and NK cell cytotoxicity which differed between males and females. These results offer a spring-board for future research towards improved treatment of stress-related disease including the impact of stress on cardiovascular and autoimmune disorders, and identifies an immune mechanism by which vulnerabilities to these diseases may be gender-specific.

1. Introduction

Chronic psychosocial and emotional distress impact immune function which leads to increased risk for disease. Current estimates forecast that by year 2030, stress-related pathologies will lead as the most debilitating and widespread health disorders (Mathers et al., 2008). At the same time, while chronic stress-related effects upon the immune system are uniformly deleterious, acute stress appears to have both protective and adverse effects. For example, acute stress can enhance the acquisition and expression of immunoprotection by activation of bodily defences prior to wounding or infection (Ackerman et al., 2002; Amkraut et al., 1971; Charney, 2004; Dhabhar, 2009), or alternatively induce immunopathology via exacerbating autoimmune inflammation, with respiratory and cardiovascular consequences (Al'Abadie et al., 1994; Black, 2006; Bosch et al., 2003; Dhabhar et al., 1995; Garg et al., 2001). The dissociation between excitatory and inhibitory molecular mechanisms remains incomplete. A more detailed

* Corresponding authors at: University of Southampton, Faculty of Medicine, Room LE57, MP813, Southampton SO16 6YD, USA (M.S. Breen). University of California San Diego, 9500 Gilman Drive, MC-0738, La Jolla, CA 92093-0738, USA (B.K. Rana).

E-mail addresses: msb1g13@soton.ac.uk (M.S. Breen), bkrana@ucsd.edu (B.K. Rana).

[1] Both authors contributed equally as first authors to this manuscript.

[2] Both authors contributed equally as senior authors to this manuscript.

http://dx.doi.org/10.1016/j.bbi.2015.10.008
0889-1591/© 2015 Published by Elsevier Inc.

understanding of immunomodulation throughout acute stress in humans is necessary not only to clinically reduce immunopathology, but also to harness stress-related immunoprotective effects.

One primary mechanism by which acute psychological stress induces immune response is through rapid changes in leukocyte distributions in the peripheral circulation (Bosch et al., 2005). Studies investigating acute short-term stressors in humans, such as public speaking, have reported brief increases of natural killer (NK) cell numbers and other leukocyte subtype cell numbers, a reduction in lymphocyte proliferation, an increase in pro-inflammatory cytokine production, and reduced healing capacity of the skin (Altemus et al., 2001; Segerstrom and Miller, 2004). Studies of acute (psychological) stress due to physical danger have used first-time tandem skydive (Mujica-Parodi et al., 2014; Schedlowski et al., 1993), as this challenge has the advantage of representing real risk, eliciting reliable effects, and yet permitting a high degree of experimental control. Studies using this paradigm report transient increases of T cells and NK cells in the blood, as well as a parallel increase in NK cell cytotoxic activity. This suggests that changes in leukocyte numbers may be an important mediator of apparent changes in leukocyte activity. Comparably, an equivalent study of bungee jumping reported increases in neutrophils, pro-inflammatory monocytes, and CD8$^+$ T cell numbers following the jump (van Westerloo et al., 2001).

While these studies are suggestive, one important limitation until recently has been the lack of computational and molecular approaches for large-scale immune system monitoring. Microarray analysis of blood transcriptional profiles offers a means to investigate immunological mechanisms relevant to acute psychological stress on a genome-wide scale. To complement these data, network analyses have been used in the field of immunology to identify the groups of coordinately expressed transcripts (modules) that are involved in the response of immune cells to immunomodulatory factors (i.e. acute stress). Indeed, the probability for multiple transcripts to follow a complex pattern of expression across dozens of participants throughout a time-series only by chance is low, and such sets of genes should therefore constitute coherent biologically meaningful transcriptional modules.

To exploit these capabilities, we performed a detailed molecular and cellular analysis upon two cohorts of participants undergoing their first-time tandem skydives. We first applied a comparative analysis of peripheral blood leukocyte (PBL) gene expression profiles between the four time-points (i) baseline, (ii) leading up to, (iii) during, and (iv) after each skydive to identify a unique panel of candidate stress responsive genes, which were validated by RT-qPCR assays. An unsupervised network analysis was then used to identify coordinately expressed genes (modules) involved in the short-term variable immune response to acute stress while considering gender-specific effects. Finally, the implications of gene expression analysis with respect to cell subset changes were validated by flow cytometry on a second cohort of participants.

2. Materials and methods

2.1. Ethical approval

State University of New York at Stony Brook and the University of California San Diego Institutional Review Boards approved this study. Thirty-nine skydivers participated in this study consisting of 13 subjects for RNA expression profiles (7 male, 6 female) and 26 subjects for flow cytometry (17 male, 9 female). All skydivers provided written consent prior to participation. Participants were recruited from individuals who independently contacted an area skydiving school (Skydive Long Island, Calverton, NY) to schedule their first-time tandem skydive. Skydivers were healthy adult subjects with no history of cardiac or mental illness, as determined by physical examination, medical history, and screening using the Structured Clinical Interview for DSM-IV.

2.2. Subjects and sample collection schedule

The study protocol adhered to a strict timeline for sample and data collection. Baseline blood samples were collected at 9:15 am within one week prior to or after the day of the skydive during a hospitalized testing that was time-locked to data collection during the skydive day and therefore served as a baseline and control. On the skydive day, all skydivers awoke at 6:30 am and arrived at Stony Brook University Hospital at 7:30 am. "Pre-boarding" samples were collected at 9:15 am, 1 h before take-off. Take-off occurred at 10:15 am, and the jump occurred at 10:30 am when the airplane reached an altitude of 11,500 feet (3,505.2 m). Skydivers landed at about 10:35 am and "post-landing" samples were collected at 10:45 am. Skydivers were immediately transported to Stony Brook University Hospital for a final blood draw at 11:30 am ("1 h post-landing" sample). Saliva was collected every 15 min from 9:15 am to 11:30 am on both the skydive and baseline hospital day.

2.3. RNA isolation and microarray gene expression analysis

Ten milliliters of blood were collected for each blood draw in an EDTA coated vacutainer blood collection tube and leukocytes were fractionated by passing the blood through LeukoLOCK filters. RNA isolation was performed using the LeukoLOCK Total RNA Isolation Kit and 100 ng of total RNA were used as starting material. RNA with a 260/280 ratio >1.7 and a RIN >6 was considered suitable for microarray analysis. Synthesis of cDNA and biotinylated cRNA and hybridization of cRNA to Illumina HumanHT12 v4 BeadChips (47,231 probes). Because the integrity of RNA was of low quality for three subjects, partially paired data was analyzed (Table S1).

2.4. Data pre-processing

Quality control of microarray data, variance-stabilizing transformation (vst), robust-spline normalization and removal of genes not expressed in any of the samples was performed in the R statistical computing environment version 2.8.0, using the Bioconductor package *lumi* (Du et al., 2008). Probes lacking gene symbol annotations were removed while probes with duplicate gene symbols were selected on the basis of having a higher average expression across all samples. This final filtering step left a total of 18,129 probes that passed into our subsequent analyses. We used two methods to identify outlier samples (2.5 standard deviations ± mean) for quality control: clustering analysis based on Pearson correlation and average distance metric and principal component analysis (PCA) using the first three components. This reduced our sample size from 50 subjects to a total 45 subjects (Table S1). The resulting quality-control treated data were used as input for differential expression and WGCNA analyses.

2.5. Differential gene expression analysis

We measured differential expression with respect to gene expression at baseline for each time point using 18,129 probes, correcting for gender differences. Differentially expressed genes were assessed using the moderated *t*-test in LIMMA (Smyth, 2005), and unless otherwise specified, a highly statistically significant threshold of *p*-value <0.01 was used. To ensure that genes found significantly differentially expressed post-landing were not solely a consequence of increased proportion of NK cells, we used a multivariate linear model to regress individual gene expression levels against NK-cell specific marker genes. The criteria for classification as a NK-cell marker were that genes needed to be: (1) identified in multiple publications linking them to the NK-cell type; and (2) found intersecting across three independent cell type specific expression databases [CTen (Shoemaker et al., 2011), IRIS (Abbas et al., 2005), and HaemAtlas (Watkins et al., 2009)]. Like others who have made similar corrections (Miller et al., 2013), we note that the model is fairly robust to choice of marker genes for cell type.

2.6. Weighted gene co-expression network analysis and module characterization

The process of identifying discrete groups of co-regulated genes can be divided into two steps. First, a signed global co-expression network was built with weighted gene co-expression network analysis (WGCNA) in R using normalized expression data of 18,129 probes. For each set of probes, a pair-wise correlation matrix was computed using the Pearson correlation. WGCNA weights the Pearson 'correlation matrix' by taking their absolute value and raising them to the power β, producing an 'adjacency matrix' (Langfelder and Horvath, 2008). This step emphasizes strong correlations and punishes weak correlations on an exponential scale. We only consider those powers that lead to a network satisfying scale-free topology at least approximately ($R^2 > 0.80$) so the mean connectivity is high and the network contains enough information (e.g. for module detection). We found that our microarray data needed a β of 9 to reach a scale-free fit. Second, the adjacency matrix was used to calculate the topological overlap measure (TOM), representing the overlap in shared neighbors. The dissimilarity TOM was used as input for the gene dendrogram (i.e. gene tree of closest pairwise neighbors), and co-expression modules were detected as branches of the gene dendrogram using the hybrid tree-cut algorithm (Fig. S4) (Langfelder and Horvath, 2008). With minimal module size set to 15 probes and merging threshold set to 0.1, 20 modules were detected.

To integrate physiological measurements with these co-expression modules, we ran singular value decomposition of each module's expression matrix and used the resulting module eigengene (ME), equivalent to the first principal component, to represent the overall expression profiles for each module. Subsequently, MEs for all modules were correlated to recorded clinical and physiological parameters such as nerve growth factor, epinephrine, norepinephrine, beta endorphin, heart rate, state anxiety trait and cortisol levels which provide a complementary assessment of these potential confounders to that performed in standard differential expression analysis. MEs are also useful for decreasing the amount of sample space tested in terms of reducing the number of multiple comparisons. A Bayes ANOVA (parameters: conf = 12, bayes = 1, winSize = 5) (Kayala and Baldi, 2012) was used to compare ME expression values for modules of interest across time-points while taking into account gender differences. For each gene in a module, intramodular membership (kME) was defined as the correlation between gene expression values and ME expression. Genes with high kME inside co-expression modules are labeled as hub genes and are predicted to be of essential to the function of the module.

2.7. Gene enrichment analyses

All differentially expressed genes passing a p-value <0.01 and all 20 network modules with genes passing a kME >0.50 were subjected to functional annotation. First, the ToppFunn module of ToppGene Suite software (Division of Biomedical Informatics) (Chen et al., 2009) was used to assess enrichment of GO ontology terms associated to relevant biological processes and pathways based on a one-tailed hyper geometric distribution with a Bonferroni correction. All annotations must have contained at least two genes to be allowed for testing. Second, to predict the involvement of key cell types we utilized the cell specific (HECS) gene expression database from the cell type enrichment (CTen) analysis web-based tool compiled by Shoemaker et al. (2011) for a broad characterization of cell type specific expression. For each gene list supplied, the significance of cell type specific expression is determined using the one-tailed hyper-geometric distribution with a Bonferroni correction across all cell/tissue types.

2.8. Protein interaction networks

Protein–protein and protein–DNA interactions for products of differentially expressed genes at pre-boarding, post-landing and 1 h post-landing were determined using the direct interactions algorithm in MetaCore™ (GeneGo, St. Joseph, MI). The interactions documented in MetaCore™ have been manually curated and are supported by citations in the literature record. When protein networks are constructed, they often reveal hub genes which represent transcription factors that control the regulation of multiple target genes. Visualization of a direct protein interaction network was facilitated by use of Cytoscape (Shannon et al., 2003).

2.9. Real time RT-qPCR

Twenty-two targets were chosen for RT-qPCR confirmation of gene expression. To rule out false positives, 15 components of NK cell-mediated cytotoxicity pathway and 3 transcription factors were selected: killer cell immunoglobulin-like receptor, three domains, long cytoplasmic tail, 1 (*KIR3DL1*), killer cell immunoglobulin-like receptor, two domains, long cytoplasmic tail, 1 and 4 (*KIR2DL1* and *4*), killer cell lectin-like receptor subfamily D, member 1 (*KLRD1*), killer cell lectin-like receptor subfamily C, member 2 (*KLRC2*), natural cytotoxicity triggering receptor 3 (*NCR3*), Fas ligand (*FASLG*), perforin 1 (*PRF1*), granzyme B (*GZMB*), lymphocyte-specific protein tyrosine kinase (*LCK*), zeta-chain (*TCR*) associated protein kinase 70 kDa (*ZAP70*), linker for activation of T cells (*LAT*), SH2 domain containing 1B (*SH2D1B*), interferon gamma (*IFNG*), CD247 molecule (*CD247*), runt-related transcription factor 3 (*RUNX3*), FBJ murine osteosarcoma viral oncogene homolog (*FOS*), interferon regulatory factor 1 (*IRF1*). To rule out false negatives, 3 targets were selected: killer cell lectin-like receptor subfamily K, member 1 (*KLRK1*), cathepsin C (*CTSC*) and transcription factor T-box 21(*TBX21*, also known as T-bet). One gene not detected by microarray was selected to test possibility of the presence of faulty probes – natural cytotoxicity triggering receptor 1 (*NCR1*). When available, TaqMan® Gene Expression Assays (Applied Biosystems by Life Technologies, Carlsbad, CA) were selected that matched the region of the RNA targeted by the corresponding Illumina probe as closely as possible; otherwise, custom assays were designed and ordered from Integrated DNA Technologies, Inc. (Corallville, IA). Reverse transcription reactions were performed using qScript™ cDNA SuperMix (Quanta Biosciences, Inc., Gaithersburg, MD). *GAPDH* control assay was used as a normalizer. Fold changes were obtained using DataAssist software version 3.01 (Applied Biosystems by Life Technologies, Carlsbad, CA) using the $2^{-\Delta\Delta CT}$ method. To determine significance, a paired t-test or Wilcoxon test (depending on the normality of the distribution as assessed by Shapiro test) was performed using normalized *Ct* values (target Ct – *GAPDH* Ct) between the time point of interest and baseline samples. Genes with p-values <0.05 were considered significant.

2.10. Flow cytometry

Two blood samples were collected from an additional cohort of 26 first-time tandem skydivers for flow cytometry analysis (one for complete blood counts and a second tube for flow cytometry data analysis). Aliquots from each blood sample were placed into 8 tubes (panels) and incubated with the mAb combinations using the manufacture's recommended procedures. After incubation, sample processing for the flow cytometry analysis followed the manufacture's instruction using red blood cell (RBC's) lysing solution (Becton Dickinson, San Jose, CA). After lysing the RBC's, the white blood cells were washed in phosphate buffered saline (PBS), re-suspended in PBS buffer and analyzed using a FACS Caliber 4-color flow cytometer (Becton Dickinson, San Jose, CA). Expression of cell-surface proteins labeled with R-Phycoerythin (PE) was quantified using the geometric means of the mean florescence intensity (MFI). All mAb's were purchased from BD Biosciences Pharmingen (San Diego, CA).

3. Results

In this exploratory study, we induced 'real-world' acute psychological stress in response to a first-time tandem skydive. Subjects reached altitude in fifteen minutes, jumped at 13,000 feet (4 km), fell at terminal velocity for one minute, and parachuted for another four minutes prior to landing. PBL samples and circulating hormone measurements from thirteen participants (7 male and 6 female) were collected at baseline (9:15 am one week before/after the skydive day), pre-boarding (9:15 am skydive day), post-landing (10:45 am skydive day, immediately after landing) and 1 h post-landing (11:45 am skydive day) (Fig. 1A).

3.1. Fluctuations in endocrine and autonomic measurements in response to acute stress

Testosterone, norepinephrine, epinephrine, beta-endorphin, nerve growth factor (NGF), salivary cortisol and heart rate were monitored throughout both the baseline and skydive days as well-established biomarkers for HPA-axis activation consequent to acute psychological stress. Heart rates were elevated on the skydive day relative to baseline as early as pre-boarding the airplane (09:45–09:55) and remained elevated until 30 min post-landing (10:30–11:00), peaking immediately before existing the airplane (10:25–10:30, $p = 6.04\text{E}{-}05$) (Fig. 1B). Salivary cortisol measurements were taken every 15 min, starting pre-boarding (09:15) to 1 h post-landing (11:35) at both the baseline and skydive days. On the skydive day, a significant increase in salivary cortisol was observed immediately before exiting the plane (10:15, $p = 8.0\text{E}{-}03$) and peaked between jumping and 1 h post-landing (10:30 $p = 5.0\text{E}{-}04$; 10:45, $p = 5.0\text{E}{-}03$; 11:00, $p = 2.0\text{E}{-}02$) (Fig. 1C) compared to the same time-points at baseline. A moderate, yet insignificant, increase of circulating testosterone, beta-endorphin and NFG was observed from baseline to post-landing (Table S1). Circulating levels of norepinephrine and epinephrine increased post-landing relative to baseline ($p = 4.0\text{E}{-}02$, $p = 3.0\text{E}{-}02$) (Fig. 1D–E). Heart rate, salivary cortisol and catecholamine levels returned to baseline levels one-hour post-landing. These patterns support stress-induced HPA activation that occurred in response to the stress of skydive. Therefore, gene expression signatures that closely followed changes in these physiological responses were expected.

3.2. Identification of candidate acute stress responsive genes

To identify stress response genes that were non-gender specific, PBL gene expression profiles were corrected for gender differences at pre-boarding, post-landing and one-hour post-landing relative to baseline. Differentially expressed genes (all $p < 0.01$) were identified pre-boarding ($N = 94$), post-landing ($N = 373$) and one-hour post-landing ($N = 121$) relative to baseline (Fig. 2A and B; for lists of differentially expressed genes see Table S2). The majority of gene expression differences were detected at post-landing and visualized on a heatmap to compare expression levels of these genes at other time-points (Fig. 2C). Genes modulated pre-boarding and one-hour post-landing displayed no functional characteristics or leukocyte cell type specificity. However, of the 373 differentially expressed genes identified from baseline to post-landing, NK cell cytotoxicity and IL-12 signaling genes, including IFN-γ, were up-regulated (Fig. 2D). Genes related to MyD88-dependent toll-like receptor (TLR) signaling tended to show decreased expression. Additionally, cell type enrichment analysis revealed a significant enrichment of up-regulated genes post-landing specific to CD56^{+} NK cells, and to a lesser extent CD8^{+} T cells (Fig. S3A).

Key genes, including those encoding transcription factors, involved in mediating stress–immune interactions were discovered through interactome analysis of all differentially expressed genes, utilizing validated direct protein–protein interaction (PPI) information from MetaCore™ (Fig. S1). This analysis revealed the up-regulation of transcription factors *RUNX3*, *FOS*, *JUN* of the innate immune system and cyclin-dependent kinase inhibitor 1A (*CDKN1A)* and zeta-chain (TCR) associated protein kinase 70 kDa (*ZAP70)* of the acquired immune system. Mitogen-activated protein kinase 3 (*MAPK3*), malic enzyme 2 (*ME2*) and guanine nucleotide binding protein (*GNAI*) mediating innate immune events were down-regulated.

3.3. RT-qPCR validation of selective NK cell cytoxocity response

A set of independent RT-qPCR assays were used to verify differentially expressed genes (from microarray data) post-landing. The RT-qPCR analysis was conducted on 22 of the differentially expressed genes that play a key role in the NK cell cytotoxicity response (Fig. S2). These genes include those that encode inhibitory receptors (*KIR2DL1*, *KIR3DL1)* and activating receptors (*KIR2DL4*, *KLRC2*, *KLRD1*, *NCR3*), classical MHC class 1 molecules (*HLA-C*, *B*, *E*, *G*) which bind to the receptors, adapter molecules for activating receptors (*SH2D1B*, *CD247*), signal transduction molecules (*LAT*, *LCK*, *ZAP70*) important for NK and T cell activation, cytolytic granules (*PRF1*, *GZMB*), and transcription factors (*RUNX3*, *FOS*). Based on previous reports of NK cell mobilization into blood in response to acute stress, it was probable that a significant number of genes would map to NK cell mediated cytotoxicity pathway (Altemus et al., 2001; Schedlowski et al., 1993). However, not all well characterized NK cell related molecules, pro- and anti-inflammatory cytokines, receptors and transcription factors were differentially expressed (Table S3). For example, activating receptors *NCR1* and *KLRK1*, cytolytic granule *CTSC* and transcription factor *TBX21* were not dysregulated; gene expression was confirmed by RT-qPCR (Fig. S2). These results suggest a precise and selective regulation of NK cell molecules and inflammatory properties of the innate and acquired immune system during acute stress, which are not accounted solely by an influx of NK cells into the periphery.

3.4. Identification of molecular alterations beyond NK cell subset differences post-landing

To account for NK cell type differences underlying differential gene expression changes from baseline to post-landing, a linear regression model was created taking into account expression of major NK cell markers. In total, four NK cell markers were selected that were consistently found across three different cell type-specific expression databases (Abbas et al., 2005; Shoemaker

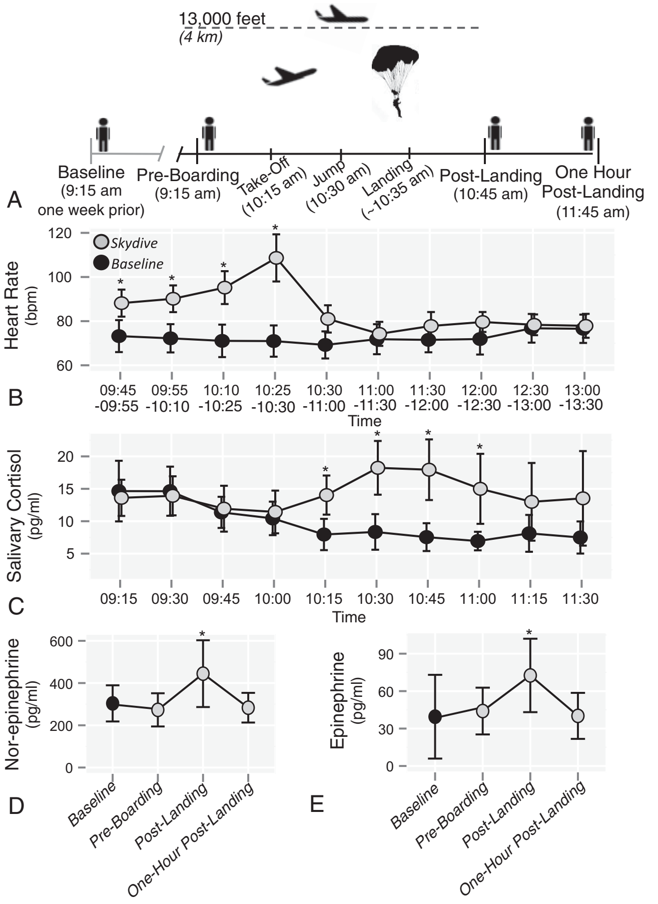


Fig. 1. Physiological changes observed throughout the sequence of events leading up to, during, and after a first time tandem skydive jump. (A) The skydiving paradigm and relevant time-points. (B). Heart rate measurements (bpm) were obtained throughout the course of both baseline and skydive days. (C) Salivary cortisol (pg/ml) was collected every 15 min from 9:15 am until 11:45 am on both baseline and skydive days. (D) Norepinephrine (pg/ml) and (E) epinephrine (pg/ml) were measured in duplicate and averaged at the corresponding four time points. Black represents baseline day and grey represents skydive day. Error bars represent 95% confidence interval and (⁎) indicates *p*-value <0.05 based on non-parametric Mann–Whitney *U* test. All *p*-values are reported in Table S1.

et al., 2011; Watkins et al., 2009): *CLIC3*, *KLRF1*, *KIR2DL3* and *KIR3DL1*. Accounting for NK cell type composition at post-landing revealed ~15% of the previously identified differentially expressed genes remained significant. Genes encoding for *FOS* and *GZMB* were among the most up-regulated genes surviving this correction, whereas *CLC* and *PAPSS1* were among the most down-regulated (Table S2). Functional enrichment analysis revealed that genes corresponding to NK cell mediated cytotoxicity and graft-vs.-host pathways were no longer significant. However, a significant up-regulation of genes enriched for *IL-12* mediated signaling (*FOS*, *RELB*, *CD247*, *GZMB*, *IL2RB*), cytotoxic T-lymphocyte (CTL) mediated immune response (*CD247*, *PRF1*, *GZMB*) and downstream signaling in naive CD8$^+$ T cells remained significant albeit to a lesser extent (Table S2E). A most interesting finding resulting from this correction was a significant enrichment of genes specific to the adrenal cortex, a key mediator of the stress response (Fig. S3).

3.5. Identification and functional annotation of gene co-expression modules

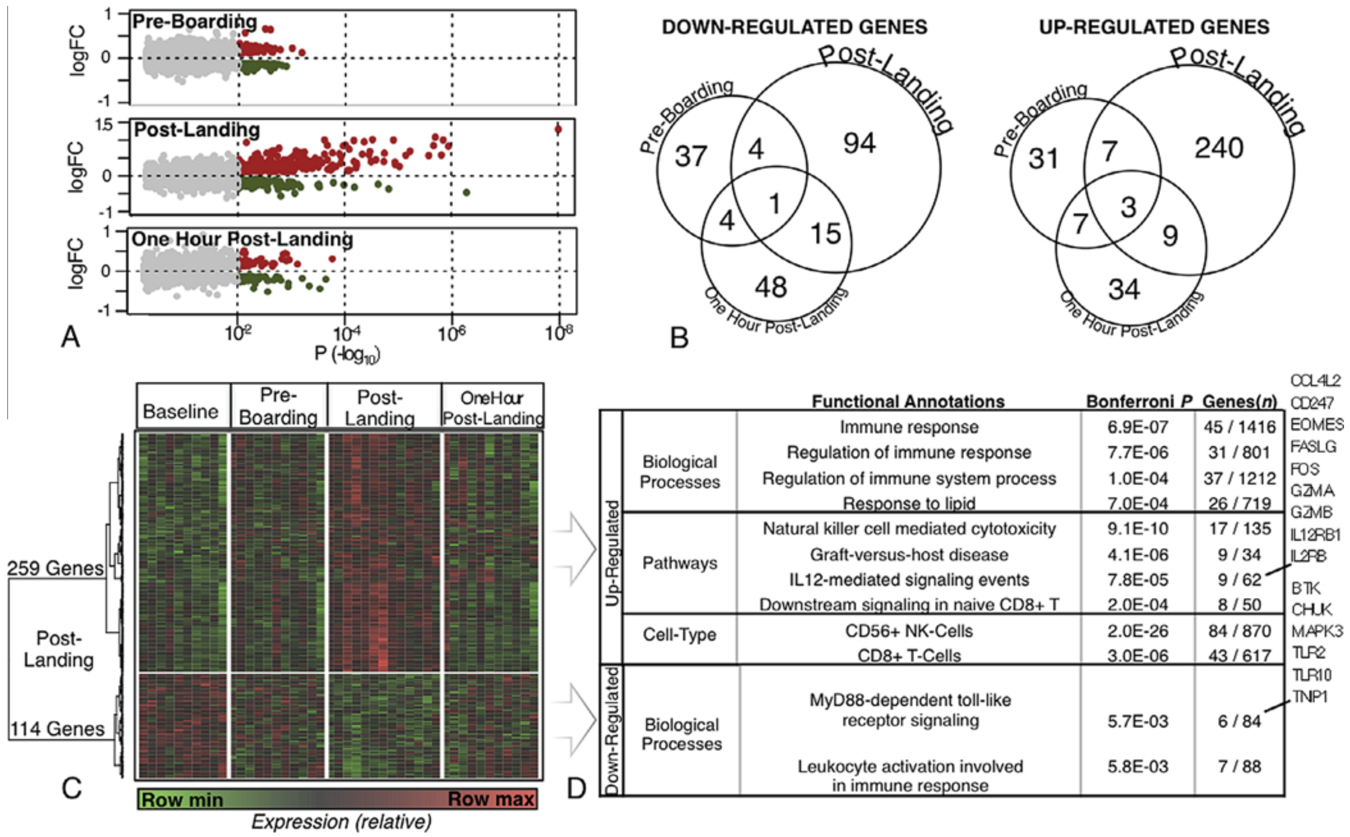
To identify coordinately expressed genes (modules) involved in the short-term variable immune response to acute stress, unsupervised WGCNA was performed. The analysis identified 19 distinct co-expression modules and 1 module representing all background genes that could not be clustered into any module (Fig. S4), each with a distinct expression pattern across all four time-points. Subsequently, all modules were functionally annotated using the top significant biological process, pathway and cell type for each individual module (all Bonferroni $p < 0.05$) (Table S4).

3.6. Functional gene co-expression modules correlate with stress induced changes in stress hormones

Next, we sought to determine the relationships between the 20 modules identified above and the observed physiological and hormonal fluctuations throughout the stress response. To integrate these multi-scale data types, module eigengene (ME) values were correlated to each time-point and all recorded subjective and physiological traits (Fig. S5). Briefly, ME value is the first PC of module expression and summarizes the main trend of expression within a module. Among the modules with high association with time-points and physiological traits, the *ME* of a module specific for ‘Cytokine Production’ was negatively correlated to post-landing ($r = -0.29$, $p = 0.05$) as well as fluctuations in circulating norepinephrine ($r = -0.32$, $p = 0.03$). The ME of modules associated to ‘T Cell Receptor (TCR) Signaling Pathway’ and ‘NK Cell Mediated Cytotoxicity’ were positively correlated to post-landing ($r = 0.28$, $p = 0.06$; $r = 0.57$, $p = 4E{-}05$ respectively). Moreover, the ‘NK Cell Mediated Cytotoxicity’ module was positively correlated to norepinephrine ($r = 0.39$, $p = 0.007$) which was expected given elevated norepinephrine and NK cell specific gene expression peak post-landing and return to baseline levels one-hour later (Figs. 1D and 2C). Of interest, the expression pattern of each marker gene used in our linear model to correct differential gene expression analysis (*CLIC3*, *KLRF1*, *KIR2DL3* and *KIR3DL1*) showed strong correlation to the ME of this particular module, confirming that the genes for our linear model were appropriately chosen. The ME of a ‘Hemostasis’ module showed gradual change from negative to positive correlation from baseline to one-hour post-landing and was significantly correlated to beta-endorphin fluctuations ($r = 0.32$ $p = 0.03$). Additionally, the ME for a module involved in ‘Oxygen Uptake and Carbon Dioxide Release’ was positively correlated to heart rate ($r = 0.38$, $p = 0.01$) and salivary cortisol levels ($r = 0.43$, $p = 0.003$), highlighting the interaction between the cardiovascular and respiratory systems. Most interestingly, including gender as a discrete measure revealed that many modules were either positively or negatively correlated to gender differences (Fig. S5) suggesting gender-specific expression patterns within each of these modules.

3.7. Gender-specific peripheral immune activation evident by divergent expression profiles within functional co-expression modules

The extent of co-expression differences was visualized throughout the stress response considering gender, averaging ME values for seven males and six females at each time-point. A Bayes ANOVA was used to compare *ME* expression values for modules of interest across time-points while taking into account gender differences (Fig. 3). The ‘NK cell mediated cytotoxicity’ and ‘Ribosome Biogenesis’ modules showed intensified expression post-landing in males relative to females (Fig. 3A and B), whereas the expression of the ‘TCR Signaling Pathway’ module was highest one-hour post-landing in males relative to females (Fig. 3C). Co-regulated genes specific to ‘Hemostasis’, which includes genes for blood coagulation, showed a gradual increase in expression (Fig. 3D) for both males and females peaking one-hour post-landing relative to baseline. Strikingly, four modules specific to ‘Immune/Defense Response’, ‘Response to Wounding’, ‘Cytokine Production’ and ‘Interferon Signaling’ (Fig. 3E–H) were down-regulated in males



		Functional Annotations	Bonferroni *P*	Genes(*n*)
Up-Regulated	Biological Processes	Immune response	6.9E-07	45 / 1416
		Regulation of immune response	7.7E-06	31 / 801
		Regulation of immune system process	1.0E-04	37 / 1212
		Response to lipid	7.0E-04	26 / 719
	Pathways	Natural killer cell mediated cytotoxicity	9.1E-10	17 / 135
		Graft-versus-host disease	4.1E-06	9 / 34
		IL12-mediated signaling events	7.8E-05	9 / 62
		Downstream signaling in naive CD8+ T	2.0E-04	8 / 50
	Cell-Type	CD56+ NK-Cells	2.0E-26	84 / 870
		CD8+ T-Cells	3.0E-06	43 / 617
Down-Regulated	Biological Processes	MyD88-dependent toll-like receptor signaling	5.7E-03	6 / 84
		Leukocyte activation involved in immune response	5.8E-03	7 / 88

Fig. 2. Comprehensive depiction of gender corrected differentially expressed genes (all $p < 0.01$) leading up to and following acute psychological stress. (A) Volcano plots for differentially expressed genes display extent of log fold-change compared to the –log10 *p*-value significance at pre-boarding, post-landing and one-hour post-landing respective to baseline. (B) Overlap of down-regulated and up-regulated genes across time-points. (C) All differentially expressed genes identified from baseline to post-landing. (D) Functional annotation of differentially expressed genes identified baseline to post-landing performed separately for up- and down-regulated genes. The top 4 most significant annotations (all $p < 0.05$ Bonferroni corrected) are shown for categories of biological processes and pathways (annotated with ToppGene) and cell types (annotated with CTen). Genes involved in IL-12 signaling and MyD88-dependent pathway are displayed for quick referencing.

post-landing and one-hour post-landing relative to females, while ME expression either increased or remained unchanged.

3.8. Stress induces changes in leukocyte and lymphocyte subset differential counts

Acute stress has been shown to cause a redistribution of leukocytes throughout the periphery (Dhabhar, 2009). To fully characterize changes in peripheral leukocyte and lymphocyte subsets throughout acute psychological stress in the present study, a second cohort consisting of 26 participants (17 male and 9 female) was recruited under the same matching experimental design as the gene expression cohort. Subsequent blood samples were subjected to flow cytometry analysis. These quantitative cell-type data were also used to better understand the extent of which gene expression results may be affected by migrating cell types. Changes within leukocyte and lymphocyte subsets were measured and displayed as both percentages and absolute cell counts combined across both males and females (Fig. 4), as there were no strong differences in cell type fluctuations between genders (Table S5).

Total leukocytes significantly increased from baseline to pre-boarding and post-landing, returning to baseline levels one-hour post-landing. There was a marked increase in the proportion and absolute count of neutrophils pre-boarding, while the post-landing proportion, albeit significantly greater than baseline, was significantly smaller than pre-boarding. Eosinophil proportion and absolute count reduced pre-boarding and remained low post-landing and one-hour post-landing relative to baseline. Monocytes and total lymphocytes showed similar patterns with the lowest proportion and absolute cell counts pre-boarding.

Changes in lymphocyte subsets were also investigated (Fig. 4 and Table S5). The percentage of $CD19^+$ B lymphocytes and absolute B cell numbers were significantly reduced post-landing. Conversely, NK cells (defined as $CD3^-CD16^+CD56^+$) were significantly increased pre-boarding and post-landing. The percentage of $CD3^+$ T lymphocytes were significantly reduced post-landing while absolute number of T lymphocytes was significantly decreased pre-boarding compared to baseline. Of the $CD3^+$ lymphocytes, $CD8^+$ and $CD4^+$ T cell absolute counts significantly increased post-landing relative pre-boarding, while $CD4^+$ T cell proportions decreased post-landing.

4. Discussion

This study describes the molecular and cellular response of the human innate and acquired immune system in reaction to physical danger. A first-time tandem skydive was used as a short-term longitudinal design to simulate acute psychological stress in a controlled environment; the stressor induces a severe form of emotional response aligned with distress related to fear (Carter and Goldstein, 2011). Our exploratory study took a dual approach. First, comparative analysis of PBL gene expression profiles between time-points identified that most gene expression changes occurred during/immediately after the stress response. Here, immediate immunomodulation is observed as a selective up-regulation of NK cytotoxicity genes, further validated with RT-qPCR assays. Correcting for changes in NK cells post-stressor revealed a molecular

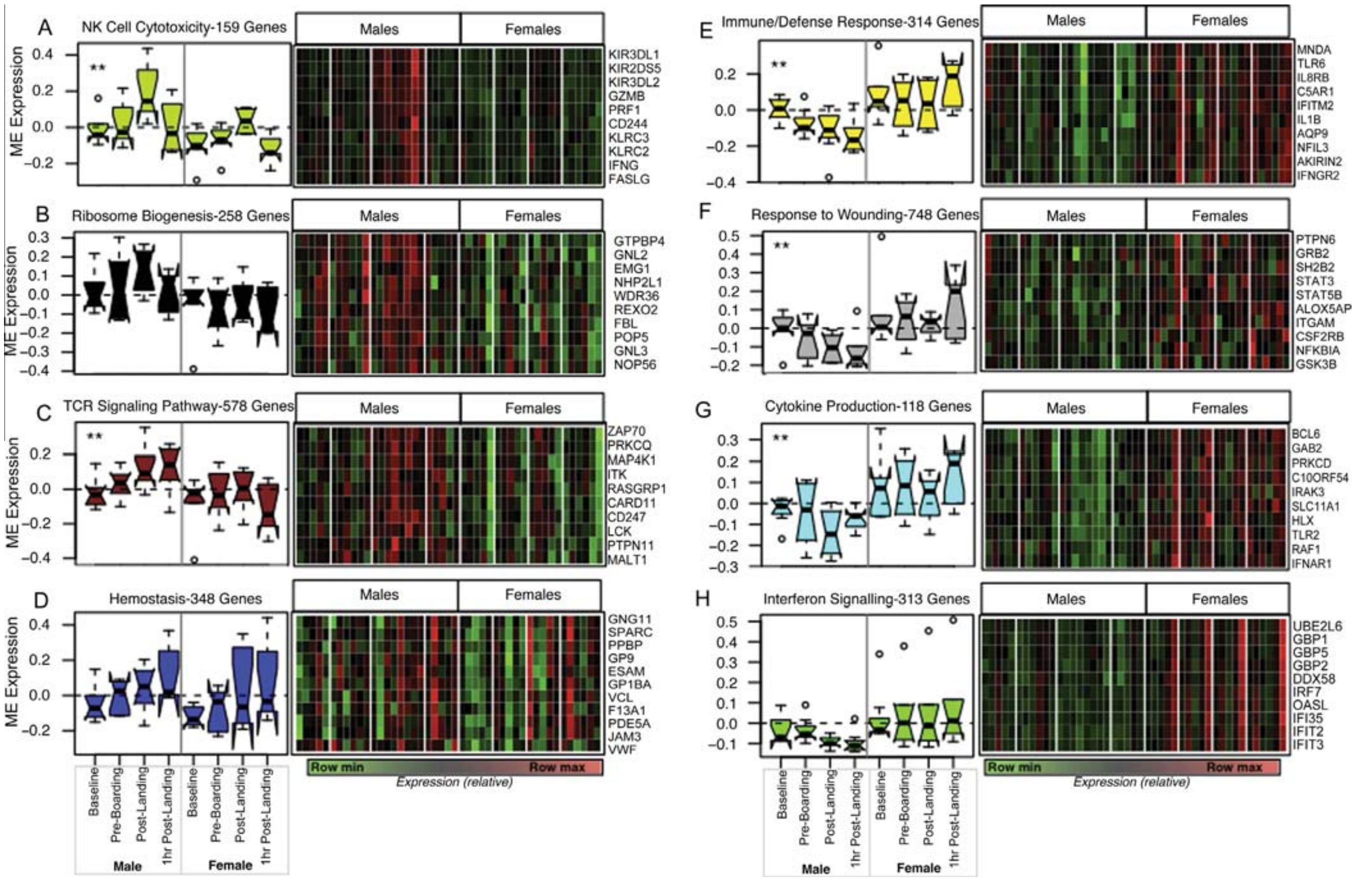


Fig. 3. Gender specific differences in functional gene co-expression modules. ME values for modules of interest are evaluated across the four time-points comparing males and females. Modules specific to (A) NK cell cytoxicity, (B) ribosome biogenesis, (C) TCR signaling pathway, (D) hemostasis, (E) immune/defense response, (F) response to wounding, (G) cytokine production (H), interferon/cytokine signaling are displayed. Heatmaps display the extent to which expression profiles of the top 10 functional hub genes, for each corresponding module, change in males and females across different time-points. White line spacers in heatmaps indicate the four time-points. The functional annotation and number of genes within each module are displayed above the boxplots. A Bayes ANOVA was used on ME values to test for significance between males and females, (**) indicates $p < 0.001$ implying strong gender-specific differences throughout course of the stress response.

signature specific to the adrenal cortex. Second, focusing our analysis on co-expressed modules revealed gender-specific peripheral immune activation evident by hundreds of co-regulated genes within several biologically annotated modules whose expression differed between males and females. These discoveries provide a useful characterization of acute stress–induced immune system alterations with implications for the understanding and treatment of stress-related disorders and gender vulnerability to stress-induced pathologies.

4.1. NK cell stress susceptibility and selective regulation of NK cell cytotoxic signaling

Although our flow cytometry data showed significant changes in leukocyte subtypes in the course of the stressor, we also showed that changes in observed gene expression profiles cannot be explained solely by the fluctuation of different leukocyte subsets. For example, peripheral neutrophils were elevated and peripheral eosinophils were reduced in the periphery pre-boarding in anticipation of the stressor. The changes in cell composition were paralleled by the up-regulation of 48 genes and the down-regulation of 46 genes, which were not associated to any functional annotations or leukocyte cell type specificity.

One unexpected finding of our study is the selective up-regulation of only a subset of NK cell genes post-landing (confirmed by RT-qPCR Fig. S2), despite a pronounced 2.5-fold increase of NK cells in the periphery (Fig. 4). The possible implications of this result may be explained through four phenomena. First, it is possible that a subset of NK genes that displayed no change in expression, were down-regulated in individual NK cells. NK cell activity may be regulated post-transcriptionally, including increases in translation and redistribution of receptors to the cell surface, which is a likely mechanism due to a fast nature of the response. Second, it is also possible that a specialized, characterized (e.g. CD56^{Lo} (Bosch et al., 2005)) or not-yet characterized subset of NK cells, expressing only a subset of specific markers is mobilized into the periphery in response to stress. Third, since gene expression was profiled from the mixture of cells, contribution of other leukocyte subsets that express overlapping sets of genes cannot be ruled out. In particular, gene expression markers for CD8$^+$ T cells were slightly elevated post-landing compared to baseline despite no change in CD8$^+$ T cell frequency in blood (Fig. 4). Even though NK cell-related genes are also expressed at lower levels in these cells, a large change in their expression in T cells can contribute to their expression change in total leukocytes. Finally, although differential gene expression analyses were gender corrected, it could be that the NK cell response is modulated to differing degrees between males and females as suggested by WGCNA observed gender-specific differences (Fig. 3A).

While only ~15% of the originally identified differentially expressed genes were found to be dysregulated after correcting for NK cells, the consistent up-regulation of cell toxicity transcript

Please cite this article in press as: Breen, M.S., et al. Acute psychological stress induces short-term variable immune response. Brain Behav. Immun. (2015), http://dx.doi.org/10.1016/j.bbi.2015.10.008

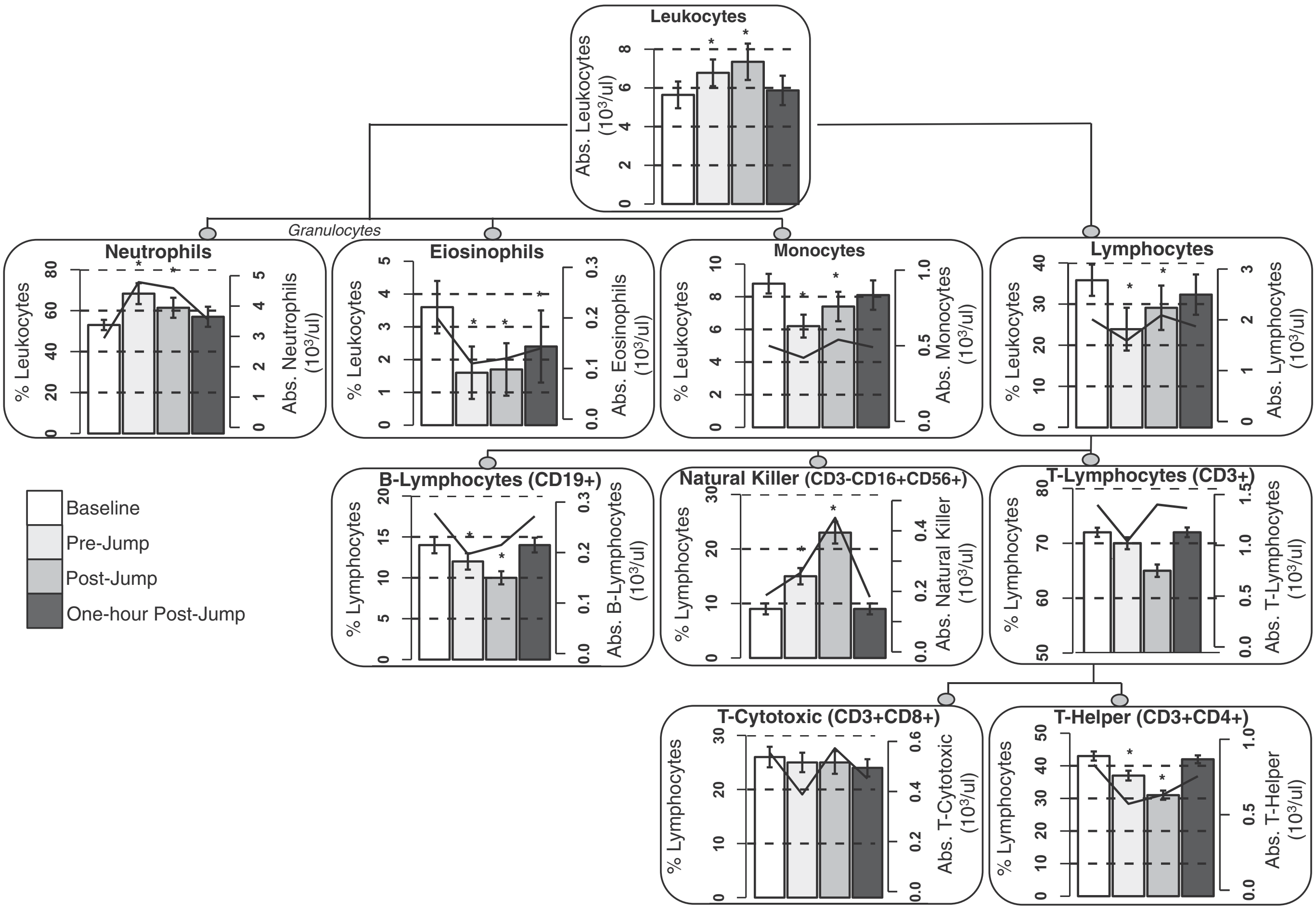


Fig. 4. A quantitative measurement of the PBL cell lineage via flow cytometry. The analysis used a gating strategy based on the forward scatter/side characteristics of immune cells from total leukocytes; granulocytes (CD45$^+$), monocytes (CD14$^+$), T cells (CD3$^+$, CD4$^+$, CD8$^+$), B lymphocytes (CD19$^+$) and NK cells (CD3$^-$CD56$^+$CD16$^+$). The raw flow data is presented as a percentage of gated cells (as indicated by the bar plots). To determine the absolute immune cell counts (as indicated by the line), leukocyte differential counts from the complete blood counts results were used to produce estimates of the actual number of immune cells in the peripheral blood samples. Statistical analysis was based on a Dunnet's test multiple comparison of means, comparing measurements back to baseline. All corresponding *p*-values are presented in Table S5.

GZMB and transcription factor *FOS* was evident. Proteolytic granzymes, such as *GZMB*, and granulysin delivered from cytotoxic cells via granule exocytosis cause activation of caspase-dependent apoptosis in stressed or pathogenic target cells (Bernard et al., 1999), which helps to explain functional annotations such as CTL mediated immune response and apoptosis following the correction. The up-regulation of *FOS*, an early immediate gene which is turned on in brain (Bernard et al., 1999), blood (Torres and Lotfi, 2007) and adrenal cortex and mediates physiological adrenocorticotropic hormone-induced responses in adrenal cortical cells (Rui Tian et al., 2014; Verstrepen et al., 2008), is consistent with the enrichment of differentially expressed genes following NK cell correction associated with the adrenal cortex and the production of cortisol (Fig. S3). This is an important observation and one that may have been difficult to detect if gene expression was measured for each cell type isolated independently.

4.2. Potential roles of IL-12 signaling and TLRs in response to acute stress

The most pronounced effect following multivariate linear regression to adjust for an influx of NK cells into the periphery post-landing, was the consistent up-regulation of genes involved in IL-12 mediated signaling (*CD247*, *FOS*, *GZMB*, *IL2RB*), and the minor production of IFN-γ. The IL-12 signaling pathway determines the type and duration of innate and adaptive immune response in part by promoting NK cell cytoxicity as well as the differentiation of naive CD4$^+$ T cells into T helper 1 (Th1) cells via the production of IFN-γ. Here, up-regulation of IL-12 signaling may indicate priming of the pro-inflammatory arm of the immune system. Such immunomodulation creates an advantage during events such as vaccination since a primed pro-inflammatory state is important for vaccine-mediated T cell immune responses, which are induced by most anti-bacterial and anti-viral vaccination strategies (Dhabhar, 2009). Thus, these data suggest a more focused adaptive immune response which under further emotional distress or antigen presentation may provide a cytokine environment favorable for Th1 polarization of the immune system.

These data also show the down-regulation of MyD88-dependent pathway including signaling molecules *MAPK3*, *CHUK* (i.e. *IKK-α*) and toll-like receptors (TLRs) 2, 6 and 10. In homeostatic conditions, TLRs lead to NFκB activation and production of pro-inflammatory cytokines IL1β, IL6 and TNFα, all involved in different pathways for innate immune activation and defense (Rui Tian et al., 2014; Verstrepen et al., 2008). Down-regulation of TLRs is consistent with previous reports suggesting that increased cortisol levels during acute stress may inhibit the NFκB, JAK-STAT and MAPK signaling pathways (Kadmiel and Cidlowski, 2013; Reichardt et al., 2002; Rui Tian et al., 2014; Webster et al., 2002). Under repeated bouts of acute stress or chronic exposure to psychosocial stress (and continued emotional activation), the response of HPA axis to sustained stress is diminished and subsequently the effectiveness of glucocorticoids (e.g. cortisol) to regulate the inflammatory response is altered as immune cells become insensitive to its regulatory effects (Cohen et al., 2012). Consequently, inflammatory pathways may become activated and initiate a negative feedback loop driving inflammation and promoting the development of many diseases.

4.3. Gender-specificity of the acute stress response at the transcriptional level and implications for stress-induced pathologies more frequent in women

Another unexpected result stemming from our exploratory gene co-expression approach was the gender-specific immune response to acute stress (Fig. 3) despite similar cellular and hormonal alterations (Tables S1 and S5), which may have relevant translational avenues. For example, it is widely accepted that among individuals experiencing chronic mental stress, cardiovascular disease (CVD) affects women more than men and gender-specific effects of mental stress on the heart is a main component of this disparity (Samad et al., 2014). While gender-specific differences in the psychobiological stress response have not been clearly identified, they may provide valuable insight towards understanding the differential cardiovascular risk in men and women. Processes associated to CVD, such as TCR signaling, defense response, response to wounding, cytokine production and interferon signaling (Mehra et al., 2005) were differently regulated by acute stress in males and females in our study (Fig. 3). These findings may help to explain gender-specific predisposition to CVD and emphasize these genes and pathways as potential tools which may be able to measure an entire facet of CVD risk, the impact of maladaptive molecular response to psychological stress in both sexes and among women in particular. Moreover, since many inflammatory disorders that are most common in women, such as autoimmunity, are also exacerbated by psychological stress (Whitacre, 1999), gender differences in cytokine response to stress (Fig. 3G) could mark an important underlying mechanism.

It is widely accepted that women suffer from chronic forms of stress such as post-traumatic stress disorder (PTSD) (Becker et al., 2007) more frequently than men, yet the reasons for this disparity are not entirely clear. It has been proposed that these differences are not explained solely on the basis of exposure type and/or severity (Sherin and Nemeroff, 2011) and that modulation of sex steroids such as estrogen and progestoreone have implicated changes in neurotransmitter systems involved in the stress response. However, factors other than exposure must play a role in the development of the disorder that might determine gender vulnerability to PTSD, and these may include transcriptomic level differences. While personalized medicine for such ubiquitous pathologies confronts numerous biomedical and financial challenges, gender-based medicine may provide a more appropriate medical platform, at the least for evaluating gender vulnerability to stress-induced pathologies.

4.4. Putative blood-based biomarkers for discriminating anxiety-based stress from related neuropsychiatric and central nervous system (CNS) disorders

An important task for studies investigating peripheral mechanisms of CNS disorders (multiple sclerosis, stroke and seizure) as well as panic attacks, myocardial ischemia, and related rodent models of such disease (Achiron et al., 2004; Kim et al., 2014; Samad et al., 2014; Yang et al., 2001, 2005), is the ability to disentangle molecular mechanisms more closely associated with the clinical presentation of disease rather than differences which are psychogenic in nature. For example, in our study the most down-regulated gene post-stressor and one-hour post-stressor was *IMAP2*, and the most down-regulated transcription factor post-stressor was *ME2*, as indicated by interactome analysis (Fig. S1). In genome-wide studies, both genes *IMPA2 and ME2* have been reported as susceptibility genes in febrile seizures and idiopathic generalized epilepsy (Arai et al., 2007; Greenberg et al., 2005; Mas et al., 2004; Prasad et al., 2013). Recently, seizures have been reported to occur following acute emotional stress (i.e. psychogenic non-epileptic seizures) rather than the result of abnormal electrical activity in the brain, as with epilepsy (Testa et al., 2012). However, baseline human blood gene expression signatures of epilepsy prior to drug treatment do not include dysregulation of *IMPA2* or *ME2* (Piro et al., 2011; Yang et al., 2001, 2005). Moreover, dysregulation of these genes was not observed in the brains of rodents post-seizure (Harald et al., 2001). While these results should be interpreted cautiously, the general inconsistencies between these studies and the results presented here may provide evidence for a role of *IMPA2* and *ME2* in differentiating between psychogenic non-epileptic seizures from true epileptic seizures.

4.5. Hypoxia does not contribute to observed gene expression profiles

Studies using an exaggerated 12 h sustained poikilocapnic hypoxic model system have noted the dysregulation of mRNA expression specific to hypoxia-inducible factor 1 (*HIF1A)*, *GAPDH*, *EPO*, and *VEG* within the first two-hours (Pialoux et al., 2009). Thus, there was a slight possibility that factors attributable to a short-term exposure (i.e. 20 min) to high altitude (i.e. 13,000 ft.), such as hypobaric hypoxia, could influence gene expression in subjects during the skydive. Therefore, the expression of these mRNA species was investigated. *HIF1A* was measured on the microarrays by three probes: none of these probes were detected as significant in our differential gene expression analysis (all $p > 0.1$). None of the probes for other genes associated with hypoxic conditions such as *GAPDH*, *EPO* or *VEG* (Pialoux et al., 2009; Zhong and Simons, 1999) were dysregulated. We did observe the differential expression of *HIPK2* among the identified anticipatory genes at pre-boarding, known to suppress *HIF1A* in hypoxia-mimicking conditions (Nardinocchi et al., 2009). The early activation of *HIPK2* may reflect increased anticipatory heart rate and early rapid breathing in anticipation to the skydive which may be working to suppress ‘hypoxia-mimicking’ conditions in the PBL microenvironment.

4.6. Limitations and future direction

While we adjusted for cell type changes affecting global gene expression, clear limitations of this study are the lack of transcriptomic investigation on individual cell types and the ability to perform transcriptomic analysis and flow cytometric data analysis on the same cohort of individuals. While gender specific differences were observed across a small number of samples, the evidence of hundreds of co-expressed functional modules throughout the skydive is significantly robust. However, one important future direction would be to extend and replicate this exploratory study using a larger cohort of participants.

4.7. Conclusion

Molecular mechanisms underlying the rapid adaptation of the immune system to an acute stressor are still incompletely defined. Our exploratory study profiled the PBL transcriptome throughout a first-time tandem skydive, as a measure of intense acute psychological stress, to reveal a detailed response to acute stress at the molecular level. A novel finding of the study is the degree of specificity of the immune response with respect to upregulation of a subset of NK cell genes that cannot be solely attributed to the influx of NK cells into the periphery in response to stress parallel by increases in cortisol and catecholamines. Correcting differential gene expression analysis post-stressor revealed a molecular signature specific to the adrenal cortex. Network analysis stratified by gender identified hundreds of genes within several functional co-expression modules responding to stress in a gender-specific manner. These results offer a spring-board for future research aimed towards identifying therapeutic targets of stress-related disorders, while underscoring the importance of gender-specific molecular profiles which could be used to better understand gender vulnerability to stress-induced disease.

Data availability

The microarray data have been submitted to the NCBI Gene Expression Omnibus database (www.ncbi.nlm.nih.gov/geo/) under accession number GSE69172.

Authors' contribution

BKR, LMP and WYE conceived and designed the study. LMP and JMC recruited the participants and conducted phenotyping and samples collection on the skydivers. WYE conducted flow cytometry assays. CHW provided gene expression guidance. NBB prepared RNA samples and conducted gene expression assays. MSB conducted gene expression, hormone, and flow-cytometry data analysis and interpretation of data. All authors contributed to writing the manuscript.

Acknowledgments

We thank Dr. Sharon Chandler for her advice in RNA stabilization and extraction protocols used in this study, Rachael Leong for assistance with RNA extractions and Anna Tocheva for thoughtful immunology discussion.

This work was supported by grants from the National Institute of Health (5MO1-RR-10710, Stony Brook University Hospital General Clinical Research Center; AG035031, BKR; 2T32AI007384-21A1, NBB) and from the Office of Naval Research Stress Physiology and Biophysics Program (N00014-09-1-0351, BKR, LMP and WYE; N0014-04-1-005, LMP; and N00014-09A-F00002, WYE). BKR was also supported by the VA Center of Excellence for Stress and Mental Health. Furthermore, this work was performed with the support of the Genomics Core at the Center for AIDS Research (AI36214) at the University of California San Diego, the San Diego Veterans Medical Research Foundation, and the Pendleton Charitable Trust. This material is based upon work supported in part by the Department of Veterans Affairs, Veterans Health Administration, Office of Research and Development. The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the United States government.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.bbi.2015.10.008.

References

Abbas, A.R., Baldwin, D., Ma, Y., Ouyang, W., Gurney, A., Martin, F., et al., 2005. Immune response in silico (IRIS): immune-specific genes identified from a compendium of microarray expression data. Genes Immun. 6 (4), 319–331.

Achiron, A., Gurevich, M., Friedman, N., Kaminski, N., Mandel, M., 2004. Blood transcriptional signatures of multiple sclerosis: unique gene expression of disease activity. Ann. Neurol. 55 (3), 410–417.

Ackerman, K.D. et al., 2002. Stressful life events precede exacerbations of multiple sclerosis. Psychosom. Med. 64, 916–920.

Al'Abadie, M.S., Kent, G.G., Gawkrodger, D.J., 1994. The relationship between stress and the onset and exacerbation of psoriasis and other skin conditions. Br. J. Dermatol. 130, 199–203.

Altemus, M., Rao, B., Dhabhar, F.S., Ding, W., Granstein, R.D., 2001. Stress-induced changes in skin barrier function in healthy women. J. Invest. Dermatol. 117, 309–317.

Amkraut, A.A., Solomon, C.F., Kraemer, H.C., 1971. Stress, early experience and adjuvant-induced arthritis in the rat. Psychosom. Med. 33, 203–214.

Arai, R., Ito, K., Ohnishi, T., Ohba, H., Akasaka, R., Bessho, Y., et al., 2007. Crystal structure of human myo-inositol monophosphatase 2, the product of the putative susceptibility gene for bipolar disorder, schizophrenia, and febrile seizures. Proteins 67 (3), 732–742.

Becker, J.B., Monteggia, L.M., Perrot-Sinal, T.S., Romeo, R.D., Talylor, J.R., et al., 2007. Stress and disease: is being female a predisposing factor? J. Neurosci. 27, 11851–11855.

Bernard, K., Cambiaggi, A., Guia, S., Bertucci, F., Granjeaud, S., et al., 1999. Engagement of natural cytotoxicity programs regulates AP-1 expression in the NKL human NK cell line. J. Immunol. 162 (7), 4062–4068.

Black, P.H., 2006. The inflammatory consequences of psychologic stress: relationship to insulin resistance, obesity, atherosclerosis and diabetes mellitus, type II. Med. Hypotheses 67 (4), 879–891.

Bosch, J.A., Berntson, G.G., Cacioppo, J.T., Dhabhar, F.S., Marucha, P.T., 2003. Acute stress evokes selective mobilization of T cells that differ in chemokine receptor expression: a potential pathway linking immunologic reactivity to cardiovascular disease. Brain Behav. Immun. 17 (4), 251–259.

Bosch, J.A., Berntson, G.G., Cacioppo, J.T., Marucha, P.T., 2005. Differential mobilization of functionally distinct natural killer subsets during acute psychological stress. Psychosom. Med. 67 (3), 366–375.

Carter, J.R., Goldstein, D.S., 2011. Sympathoneural and adrenomedullary responses to mental stress. Compr. Physiol., 119–146

Charney, D.S., 2004. Psychobiological mechanisms of resilience and vulnerability: implications for successful adaptation to extreme stress. Am. J. Psychiatry 161, 195–216.

Chen, J., Bardes, E.E., Aronow, B.J., Jegga, A.G., 2009. ToppGene suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. 37, W305-11.

Cohen, S., Janicki-Deverts, D., Doyle, W., Miller, G.E., Frank, E., et al., 2012. Chronic stress, glucocorticoid receptor resistance, inflammation, and disease risk. Proc. Natl. Acad. Sci. 109 (16), 5995–5999.

Dhabhar, F.S., Miller, A.H., McEwen, B.S., Spencer, R.L., 1995. Effects of stress on immune cell distribution-dynamics and hormonal mechanisms. J. Immunol. 154, 5511–5527.

Dhabhar, F.S., 2009. Enhancing versus suppressive effects of stress on immune function: implications for immunoprotection and immunopathology. NeuroImmunoModulation 16 (5), 300–317.

Du, P., Kibbe, W.A., Lin, S.M., 2008. Lumi: a pipeline for processing Illumina microarray. Bioinformatics 24 (13), 1547–1548.

Garg, A., Chren, M.M., Sands, L.P., Matsui, M.S., Marenus, K.D., Feingold, K.R., et al., 2001. Psychological stress perturbs epidermal permeability barrier homeostasis: implications for the pathogenesis of stress-associated skin disorders. Arch. Dermatol. 137, 53–59.

Greenberg, D.A., Cayanis, E., Strug, L., Marathe, S., Durner, M., Pal, D.K., et al., 2005. Malic enzyme 2 may underlie susceptibility to adolescent-onset idiopathic generalized epilepsy. Am. J. Hum. Genet. 76, 139–146.

Harald, S., Badan, I., Fischer, B., Wagner, A.P., 2001. Dynamics of gene expression for immediate early- and late genes after seizure activity in aged rats. Arch. Gerontol. Geriatr. 32 (3), 199–218.

Kadmiel, M., Cidlowski, J.A., 2013. Glucocorticoid receptor signaling in health and disease. Trends Pharmacol. Sci. 34 (9), 518–530.

Kayala, M.A., Baldi, P., 2012. Cyber-T web server: differential analysis of high-throughput data. Nucleic Acids Res. 40 (W1), W553–W559.

Kim, J., Ghasemzadeh, N., Eapen, D.J., Chung, N.C., Storey, J.D., et al., 2014. Gene expression profiles associated with acute myocardial infarction and risk of cardiovascular death. Genome Med. 6, 40.

Langfelder, P., Horvath, S., 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinf. 9, 559.

Mas, C., Taske, N., Deutsch, S., Guipponi, M., Thomas, P., Covanis, A., et al., 2004. Association of the connexin36 gene with juvenile myoclonic epilepsy. J. Med. Genet. 41, e93.

Mathers, C., Fat, D.M., Boerma, J., 2008. The Global Burden of Disease: 2004 Update. World Health Organization, Geneva, Switzerland.

Mehra, V.C., Ramgolam, V.S., Bender, J.R., 2005. Cytokines and cardiovascular disease. J. Leukoc. Biol. 78 (4), 805–818.

Miller, J., Woltjer, R.L., Goodenbour, J.F., Horvath, S., Geschwind, D.H., 2013. Genes and pathways underlying regional and cell type changes in Alzheimer's disease. Genome Med. 5, 48.

Mujica-Parodi, L.R., Carlson, J.M., Cha, J., Rubin, D., 2014. The fine line between 'brave' and 'reckless': Amygdala reactivity and regulation predict recognition of risk. NeuroImage 103, 1–9.

Nardinocchi, L., Puca, R., Sacchi, A., D'orazi, G., 2009. Inhibition of HIF-1alpha activity by homeodomain-interacting protein kinase-2 correlates with sensitization of chemoresistant cells to undergo apoptosis. Mol. Cancer 8 (1).

Pialoux, V., Mounier, R., Brown, A.D., Steinback, C.D., Rawling, J.M., et al., 2009. Relationship between oxidative stress and HIF-1_mRNA during sustained hypoxia in humans. Free Radical Biol. Med. 46 (2), 321–326.

Piro, R.M., Molineris, I., Ala, U., Di Cunto, F., 2011. Evaluation of candidate genes from orphan FEB and GEFS loci by analysis of human brain gene expression atlases. Ed. Takeo Yoshikawa. PLoS One 6 (8), e23149.

Prasad, D.K., Satyanarayana, U., Munshi, A., 2013. Genetics of idiopathic generalized epilepsy: an overview. Neurol. India 61 (6), 572.

Reichardt, H.M., Tuckermann, J.P., Göttlicher, M., et al., 2002. Repression of inflammatory responses in the absence of DNA binding by the glucocorticoid receptor. EMBO J. 20 (24), 7168–7173.

Rui Tian, T., Hou, G., Li, D., Yuan, T.F., 2014. A possible change process of inflammatory cytokines in the prolonged chronic stress and its ultimate implications for health. Sci. World J. Article 1155, 780616.

Samad, Z., Boyle, S., Ersboll, M., Vora, A.N., Zhang, Y., et al., 2014. Sex differences in platelet reactivity and cardiovascular and psychological response to mental stress in patients with stable ischemic heart disease: insights from the REMIT study. J. Am. Coll. Cardiol. 64 (16), 1669–1678.

Schedlowski, M., Jacobs, R., Stratmann, G., Richter, S., Hadicke, A., et al., 1993. Changes of natural killer cells during acute psychological stress. J. Clin. Immunol. 13 (2), 119–126.

Segerstrom, S.C., Miller, G.E., 2004. Psychological stress and the human immune system: a meta-analytic study of 30 years of inquiry. Psychol. Bull. 130, 1–37.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., et al., 2003. Cytoscape: a software environment for integrated modules of biomolecular interaction networks. Genome Res. 13 (11), 2498–2504.

Sherin, J.E., Nemeroff, C.B., 2011. Post-traumatic stress disorder: the neurobiological impact of psychological trauma. Dialogues Clin. Neurosci. 13 (3), 263–278.

Shoemaker, J.E., Fukuyama, S., Sakabe, S., Kitano, H., Kawaoka, Y., 2011. CTen: a web-based platform for identifying enriched cell types from heterogenous microarray data. BMC Genomics 13, 460.

Smyth, G.K., 2005. Limma: linear models for microarray data. In: Gentleman, R., Carey, V., Dudoit, S., Irizarry, R., Huber, W. (Eds.), Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer, New York, pp. 397–420.

Testa, S.M., Krauss, G.L., Lesser, R.P., Brandt, J., 2012. Stressful life event appraisal and coping in patients with psychogenic seizures and those with epilepsy. Seizure 21 (4), 282–287.

Torres, T.E.P., Lotfi, C.F.P., 2007. Distribution of cells expressing Jun and Fos proteins and synthesizing DNA in the adrenal cortex of hypophysectomized rats: regulation by ACTH and FGF2. Cell Tissue Res. 329 (3), 443–455.

van Westerloo, D.J., Choi, G., Lowenberg, E.C., Truijen, J., de Vos, A.F., et al., 2001. Acute stress elicited by bungee jumping suppresses human innate immunity. Mol. Med. 17, 180–188.

Verstrepen, L., Bekaert, T., Chau, L.T., Tavernier, J., Chariot, A., Beyaert, R., 2008. TLR-4, IL-1R and TNF-R signaling to NF-_B: variations on a common theme. Cell. Mol. Life Sci. 65 (19), 2264–2978.

Watkins, N.A., Gusnanto, A., De Bono, B., De, S., Miranda-Saavedra, et al., 2009. A HaemAtlas: characterizing gene expression in differentiated human blood cells. Blood 113 (19), E1–E9.

Webster, J.I., Tonelli, L., Sternberg, E.M., 2002. Neuroendocrine regulation of immunity. Annu. Rev. Immunol. 20, 125–163.

Whitacre, C.C., 1999. A gender gap in autoimmunity. Science 283, 1277–1278.

Yang, T., Gilbert, D.L., Glauser, T.A., Hershey, A.D., Sharp, F.R., 2005. Blood gene expression profiling of neurologic diseases. Arch. Neurol. 62 (2), 210.

Yang, T., Lu, A., Aronow, B.J., Sharp, F.R., 2001. Blood genomic responses differ after stroke, seizures, hypoglycemia, and hypoxia: blood genomic fingerprints of disease. Ann. Neurol. 50 (6), 699–707.

Zhong, H., Simons, J.W., 1999. Direct comparison of GAPDH, beta-actin, cyclophilin, and 28S rRNA as internal standards for quantifying RNA levels under hypoxia. Biochem. Biophys. Res. Commun. 259 (3), 523–526.

Please cite this article in press as: Breen, M.S., et al. Acute psychological stress induces short-term variable immune response. Brain Behav. Immun. (2015), http://dx.doi.org/10.1016/j.bbi.2015.10.008

OPEN

Citation: Transl Psychiatry (2016) **6**, e●●; doi:10.1038/tp.2016.67

www.nature.com/tp

ORIGINAL ARTICLE

Candidate gene networks and blood biomarkers of methamphetamine-associated psychosis: an integrative RNA-sequencing report

MS Breen[1], A Uhlmann[2], CM Nday[3], SJ Glatt[4], M Mitt[5], A Metsalpu[5], DJ Stein[2] and N Illing[3]

The clinical presentation, course and treatment of methamphetamine (METH)-associated psychosis (MAP) are similar to that observed in schizophrenia (SCZ) and subsequently MAP has been hypothesized as a pharmacological and environmental model of SCZ. However, several challenges currently exist in diagnosing MAP accurately at the molecular and neurocognitive level before the MAP model can contribute to the discovery of SCZ biomarkers. We directly assessed subcortical brain structural volumes and clinical parameters of MAP within the framework of an integrative genome-wide RNA-Seq blood transcriptome analysis of subjects diagnosed with MAP ($N = 10$), METH dependency without psychosis (MA; $N = 10$) and healthy controls ($N = 10$). First, we identified discrete groups of co-expressed genes (that is, modules) and tested them for functional annotation and phenotypic relationships to brain structure volumes, life events and psychometric measurements. We discovered one MAP-associated module involved in ubiquitin-mediated proteolysis downregulation, enriched with 61 genes previously found implicated in psychosis and SCZ across independent blood and post-mortem brain studies using convergent functional genomic (CFG) evidence. This module demonstrated significant relationships with subcortical brain structure volumes including the anterior corpus callosum (CC) and the nucleus accumbens. Furthermore, a second MAP and psychoticism-associated module involved in circadian clock upregulation was also enriched with 39 CFG genes, further associated with the CC. Subsequently, a machine-learning analysis of differentially expressed genes identified single blood-based biomarkers able to differentiate controls from methamphetamine dependents with 87% accuracy and MAP from MA subjects with 95% accuracy. CFG evidence validated a significant proportion of these putative MAP biomarkers in independent studies including *CLN3*, *FBP1*, *TBC1D2* and *ZNF821* (RNA degradation), *ELK3* and *SINA3* (circadian clock) and *PIGF* and *UHMK1* (ubiquitin-mediated proteolysis). Finally, focusing analysis on brain structure volumes revealed significantly lower bilateral hippocampal volumes in MAP subjects. Overall, these results suggest similar molecular and neurocognitive mechanisms underlying the pathophysiology of psychosis and SCZ regardless of substance abuse and provide preliminary evidence supporting the MAP paradigm as an exemplar for SCZ biomarker discovery.

Translational Psychiatry (2016) **6,** e●●; doi:10.1038/tp.2016.67; published online xx xxx 2016

INTRODUCTION

Methamphetamine (METH) is an *N*-methyl derivative of amphetamine and a highly addictive psychostimulant severely affecting the central nervous system.[1] METH use is at epidemic levels in several areas of the world and its global prevalence is estimated at 15–16 million people with several pockets of increased use in the USA, Europe and Africa.[2,3] Recent evidence ranked METH fourth out of 20 of the most harmful drugs due to self-harm to the user.[4] One reason for this is that METH provokes psychotic reactions in an estimated 72–100% of all abusers.[5,6]

Methamphetamine-associated psychosis (MAP) has been considered a pharmacological and environmental model of schizophrenia (SCZ) due to similarities in clinical presentation (that is, paranoia, hallucinations, disorganized speech and negative symptoms), response to treatment (neuroleptics) and presumed neuromechanisms (central dopaminergic neurotransmission).[7–9] It is hypothesized that a better understanding of the molecular mechanisms underlying SCZ may be accelerated via examination of human models related to the disease. In this context, the MAP model could quicken the discovery of risk biomarkers, screening for subclinical disease, prognostics, diagnostics or disease staging. However, several challenges currently exist in terms of accurately diagnosing MAP on a molecular and cognitive level before the MAP model can contribute to the discovery of SCZ biomarkers.

Genome-wide blood transcriptome profiling coupled with network analyses provide a platform for identifying functionally relevant biological markers of disease, permitting multi-scale data integration.[10] This is a critical point as acute and chronic effects of MAP are widespread across the body and an integrative technique determining relationships of biological markers with magnetic

[1]Department of Clinical and Experimental Sciences, Faculty of Medicine, University of Southampton, Southampton, UK; [2]Department of Psychiatry and MRC Unit on Anxiety and Stress Disorders, Groote Schuur Hospital (J-2), University of Cape Town, Cape Town, South Africa; [3]Department of Molecular and Cellular Biology, University of Cape Town, Cape Town, South Africa; [4]Psychiatric Genetic Epidemiology and Neurobiology Laboratory, Departments of Psychiatry and Behavioral Sciences and Neuroscience and Physiology, Medical Genetics Research Center, SUNY Upstate Medical University, Syracuse, NY, USA and [5]The Estonian Genome Center, University of Tartu, Tartu, Estonia. Correspondence: Dr MS Breen, Department of Clinical and Experimental Sciences, Faculty of Medicine, University of Southampton, Tremona Road, Room LE57, MP813 Southampton General Hospital, Southampton SO16 6YD, UK or Professor DJ Stein, Department of Psychiatry and MRC Unit on Anxiety and Stress Disorders, Groote Schuur Hospital (J-2), University of Cape Town, Anzio Road, Observatory, Cape Town 7925, South Africa.

E-mail: MSB1G13@soton.ac.uk or Dan.Stein@uct.ac.za

Received 9 October 2015; revised 21 January 2016; accepted 24 January 2016

resonance imaging (MRI), life events (that is, stress, culture) and psychometric measurements could provide key insights towards cognitive and molecular mechanisms of MAP, and the versatility of the MAP model in molecular psychiatry research. Complimentary, machine learning provides a useful tool for *in silico* prediction of candidate biomarkers.[11] Further confirmation and validation of these biomarkers may be accomplished by utilizing convergent functional genomics (CFG) evidence. The CFG approach has proven highly successful for moderately sized psychiatric cohorts in reducing false positives and false negatives by drawing on multiple disparate yet 'convergent' sources of external functional genomic information across independent human studies.[12–20] Collectively, these techniques hold great promise for the prioritization and validation of candidate genes for MAP and their relatedness to SCZ.

We present a preliminary integrative RNA-sequencing report exploring peripheral blood gene expression among subjects diagnosed with METH-associated psychosis (MAP), METH dependency without psychotic symptoms (MA) and healthy control subjects. The primary goal of this analysis was to best characterize the molecular signatures defining MAP at the systems level and again at the individual gene level to reveal a novel panel of MAP blood biomarkers. An unbiased weighted gene co-expression network analysis (WGCNA) was first used to identify co-expression modules that were subjected to functional annotation and multi-scale data integration collected from the same subjects. Subsequently, a multi-class machine-learning approach was used to identify candidate blood biomarkers able to differentiate between MA, MAP and healthy control subjects. CFG information was used to validate the role of candidate gene networks and blood biomarkers in the pathophysiology of MAP and confirm their shared association to psychotic disorders and SCZ in independent studies with the absence of METH.

MATERIALS AND METHODS

Participants

A total of 10 MAP subjects, 10 subjects with METH dependence without developing psychotic symptoms (MA), and 10 healthy control subjects were enrolled in this study. Gender (male) and age-matched (25.8 ± 6 years) right-handed subjects were recruited from drug rehabilitation facilities, hospitals and communities in Cape Town, South Africa where all the subjects were provided detailed study information and gave written consent. Each subject underwent two assessment sessions. The first session consisted of a detailed psychiatric interview and demographic and substance variables were recorded. During the second session, approximately 1 week later, the patients were asked to fast and refrain from smoking overnight, before blood was collected between 0900 and 1100 h. This was followed by a brain scan. Clinical assessment was performed using the Structured Diagnostic Interview for DSM-IV Axis I Disorders[21] and the patients completed a battery of self-report questionnaires including the Life Events Questionnaire,[22] Kessler Psychological Distress Scale (K10),[23] the Beck Depression Inventory,[24] behavioural inhibition system/behavioural activation system scale,[25] Eysenck Personality Questionnaire—Revised short scale[26] (For detailed information regarding each of these measures, see Supplementary File). Positive and negative symptoms within the MAP group were rated using the PANSS (Positive and Negative Syndrome Scale):[27] PANSS positive subscale (14.5 ± 6.1), negative subscale (22.0 ± 11.5) and total score (66.8 ± 26.1). Exclusion criteria comprised the following: (1) additional substance dependencies other than nicotine and METH for the MA and MAP groups, and any substance dependence other than nicotine in the control group; (2) lifetime and current diagnosis of any psychiatric disorders (other than MA dependence and MAP in the MA and MAP groups); (3) a history of psychosis before MA abuse; (4) a medical or neurological illness or head trauma; (5) a seropositive test for HIV; (6) MRI incompatibilities or known claustrophobia. All the participants in the MAP group were on treatment with neuroleptic medication (haloperidol) at the time of testing. Polysubstance use was allowed to facilitate participant recruitment including nicotine, cannabis and alcohol for all the study groups. This study was approved (HREC REF 340/2009) by the University of Cape Town Faculty of Health Sciences Human Research Ethics Committee.

MRI acquisition and image processing

The subjects in this study form part of a larger project investigating fronto-temporal cortical and subcortical grey matter structures in MA and MAP. The images were acquired on a 3 T Magnetom Allegra scanner (Siemens, Erlangen, Germany) at the Cape Universities Brain Imaging Centre. A high-resolution, T1-weighted, three-dimensional multi-echo MPRAGE sequence (scan parameters: repetition time = 2530 ms; graded echo time = 1.53, 3.21, 4.89, 6.57 ms; flip angle = 7°; field of view = 256 mm) produced 160 sagittal images of 1 mm thickness. By acquiring four separate structural scans with graded echo times and averaging those into a final high contrast image,[28] the MEMPRAGE method creates structural images with low distortion and high signal-to-noise ratio.

The MRI scans were analysed using the FreeSurfer software package v5.1 (http://surfer.nmr.mgh.harvard.edu/). Regional estimates of subcortical volumes were assessed with a specialized surface-based reconstruction and automatic labelling tool, which is described in detail elsewhere.[29] In summary, FreeSurfer processing includes motion correction, skull-stripping, Talairach transformation, segmentation of subcortical white matter and deep grey matter volumetric structures, intensity normalization, tessellation of the grey matter/white matter boundary, automated topology correction and surface deformation.

RNA isolation, library preparation and data availability

Blood was collected using PAXgene RNA tubes (Qiagen, Valencia, CA, USA) and total RNA was extracted and purified in accordance with the PAX gene RNA kit per manufacturer's instructions. Globin mRNA was depleted from samples using the GLOBINclear—Human Kit (Life Technologies, Carlsbad, CA, USA). Subsequently, the quantity of all purified RNA samples was measured on a nanodrop (56.6 ± 16.7 ng μl^{-1}) and the quality and integrity measured with the Agilent 2100 Bioanalyzer (Agilent, Santa Clara, CA, USA). All RNA integrity numbers were greater than 7 (8.4 ± 0.7).

The Illumina TruSeq Stranded Total RNA kit (Ilumina, San Diego, CA, USA) was used for library preparation accordingly to manufacturer instructions without any modifications. The 30 indexed RNA libraries were pooled and sequenced using long paired-end chemistry (2x93 bp) on seven lanes using the Illumina HiSeq2500. All the replicates were run for 2 × 40 million reads per sample and all the reads were primary processed using Casava v1.8.2 to transform primary base call files into fastq files. These raw RNA-sequencing fastq data have been submitted to Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE74737.

Read trimming, mapping and quantification of gene expression

All the fragmented RNA-Seq reads were trimmed to 90 bp and low quality reads were discarded using Trimmomatic[30] options SLIDINGWINDOW:90:10 MINLEN:90 CROP:90. Subsequently, all high-quality trimmed reads were mapped to UCSC *Homo sapiens* reference genome (build hg19) using TopHat v2.0.0.[31] We used the estimated mean inner distance and standard deviation between mate paired-ends as the -r and --mate-std-dev parameters, respectively. TopHat calls Bowtie v1.1.1[32] to perform alignment with no more than two mismatches. We used the pre-built index files of UCSC *H. sapiens* hg19, downloaded from the TopHat homepage (https://ccb.jhu.edu/software/tophat/igenomes.shtml). Samtools[33] was used to convert bamfiles to samfiles and HTseq v0.6.0[34] was used to count all of the mapped reads by htseq-count using parameters –stranded = reverse –q.

Data pre-processing

Raw count data measured 23 345 transcripts across 30 subjects. Unspecific filtering removed lowly expressed genes that did not meet the requirement of a minimum of 20 reads in at least 10 subjects. A total of 12 281 transcripts were retained, then subjected to edgeR VOOM normalization,[35] a variance-stabilization transformation method. Normalized data were inspected for outlying samples using unsupervised hierarchical clustering of subjects (based on Pearson coefficient and average distance metric) and principal component analysis to identify potential outliers outside two standard deviations from these averages. No outliers were present in these data and resulting normalized values were used as input for downstream analyses.

Gene co-expression network construction and module detection

Signed co-expression networks were built using WGCNA[10] in R, as previously described.[36,37] A total of 12 281 transcripts were used to construct a global network of all 30 subjects. To construct a network, the absolute values of Pearson correlation coefficients were calculated for all the possible gene pairs and resulting values were transformed using a β-power of 9 so that the final correlation matrix followed an approximate scale-free topology.[10] The WGCNA cut-tree hybrid algorithm was used to detect sub-networks, or co-expression modules, within the global network optimizing minimum module size to 15, deep split of 2 and a tree-cut height of 0.2 to merge neighbouring network modules with similar expression profiles. For each identified module, we ran singular value decomposition of each module's expression matrix and used the resulting module eigengene (ME), equivalent to the first principal component, to represent the overall expression profiles for each module. Differential expression of MEs was performed using a Bayes analysis of variance[38] (parameters: conf = 12, bayes = 1, winSize = 5) testing between groups and *P*-values were corrected for multiple comparisons (*post hoc* Tukey correction). Subsequently, to determine which modules were most associated to recorded clinical parameters and potential confounding variables in this study, MEs for all modules were correlated to external subjective and objective data using a Pearson correlation and a Student's asymptotic *P*-value for significance. MEs were also used to determine module membership (kME) values for each gene in a specified module, defined as the correlation between gene expression values and ME expression. Genes with the highest intramodular kME were labelled as hub genes and predicted to be essential to the function of the module.

Differential gene expression analyses

A moderated *t*-test, implemented through the *limma*[39] package, assessed differential gene expression between the three groups in a group-wise manner across 12 281 transcripts. Significance threshold was set to a nominal *P*-value <0.01 to permit sufficient enough genes to move forward with functional characterization and supervised classification methods. Differentially expressed genes corresponding to WGCNA modules which were significantly associated with polysubstance abuse were excluded and removed from functional annotation and supervised classification methods, as a robust and complimentary strategy of adjusting for confounding factors.

Functional enrichment analyses

All differentially expressed genes passing a *P*-value <0.01 and all network modules with genes passing a kME >0.50 were subjected to functional annotation. First, the ToppFunn module of ToppGene Suite software[40] (https://toppgene.cchmc.org/) was used to assess enrichment of GO ontology terms relevant to cellular components, molecular factors, biological processes, metabolic pathways and well-annotated drug compounds from the comparative toxicogenomics database[41] using a one-tailed hyper-geometric distribution with a Bonferroni correction. A minimum of a two-gene overlap per gene-set was necessary to be allowed for testing. The human cell-specific gene expression database from the cell type enrichment[42] analysis web-based tool was used to predict the involvement of key cell types within candidate gene lists. For each supplied gene list, the significance of cell type-specific expression are determined using the one-tailed Fisher's exact test with a Bonferroni correction across all the available cell/tissue types. For information pertaining to curating haloperidol gene signatures, see Supplementary File.

Construction of diagnostic blood classifier for MAP

BRB-Array Tools[11]-supervised classification methods were used to construct gene expression classifiers. Two models were specified: (1) controls vs METH dependents and (2) MA vs MAP subjects. Each model consisted of three steps. First, to ensure a fair comparison and to decrease computational time, all genes with $P<0.01$ were subjected to classifier construction. This heuristic rule of thumb approach was used to cast a wide net to catch all potentially informative genes, while false positives would be pared off by subsequent optimization and cross-validation steps. Second, classifiers composed of different numbers of genes were constructed by recursive feature elimination. Recursive feature elimination provided feature selection, model fitting and performance evaluation via identifying the optimal number of features with maximum predictive accuracy. Third, the ability for recursive feature elimination to predict group outcome was assessed by diagonal linear discriminant analysis and compared with three different multivariate classification methods (that is, support vector machine, nearest centroid, three-nearest neighbours) in a leave-one-out cross-validation approach. In addition, a permutation *P*-value, based on 1000 random permutations, for the cross-validated misclassification error rate for each classification method was implemented. This *P*-value indicates the proportion of the random permutations that gave as small a cross-validated misclassification rate as was obtained with the real class labels.

Converging functional genomic scoring

CFG represents a translational methodology that integrates multiple lines of external evidence from human and animal model studies in a Bayesian-like manner. This approach increases the ability to distinguish signal from noise in limited size cohorts and is routinely applied to support the identification of blood biomarkers across neuropsychiatric disorders.[12–20] The principal aim of the CFG approach is to increase the likelihood that findings will prove reproducible and have predictive power in independent cohorts. Our CFG scoring paradigm for prioritization of MAP biomarkers is an adaptation of previous techniques, representing a two-step process (Supplementary Figure 6) as given below.

Internal lines of evidence: All genes assigned a *P*-value <0.05 were included in the CFG scoring. These liberal criteria were used to cast a wide net of all potentially informative genes, which may be involved in the pathophysiology of MAP, while false positives would be pared off by subsequent CFG scoring and optimization steps. Each gene was given three *P*-values (based on three group-wise differential expression analyses). Subsequently, a score of 1 was given to genes passing $P<0.001$, a score of 0.5 was given to genes passing $0.001>P<0.01$, and a score of 0.2 was given for genes passing $0.01>P<0.05$, permitting a maximum score of 3 and a minimum score of 0.2. A bonus point of 0.5 was awarded for genes passing $P<0.01$ occurring in both MAP vs controls and MAP vs MA comparisons, as well as genes found to be members of MAP-associated modules. Thus, a max score of 4 is attainable (3+0.5+0.5).

External lines of evidence: CFG evidence was scored for a gene if there were published reports of human data including post-mortem brain expression, peripheral blood expression and/or genetic evidence (association and linkage) utilizing two large databases. One database represents a recently built in-house database specific to human blood transcriptome studies using PubMed (http://www.ncbi.nlm.nih.gov/pubmed) search queries and combinations of key words (e.g. blood transcriptome and psychosis).[43] To consider functional support across divergent technological platforms and human post-mortem brain samples, we accessed DisGenNet,[44] a comprehensive database of human gene–disease associations from various expert curated databases and text-mining-derived associations. These database searches included gene–disease relationships focusing specifically on psychosis, SCZ, depression/stress and neurocognitive impairment to consider comorbid effects of MAP in our study. Importantly, studies containing a METH component were excluded in order to validate MAP biomarkers in drug-free (METH) models. For the CFG analysis and scoring, external cross-validating lines of evidence were weighted such that findings in human peripheral blood specific to psychosis were given an additional 1 point. A maximum of five external lines of evidence were allowed. Thus, the total maximum CFG score that a candidate biomarker gene could have was 10 (4 for threshold+5 for external evidence+1 blood presence in psychosis). Like other studies using this approach,[12–20] we appreciate there are other ways of scoring blood biomarkers based on CFG which may give slightly different results in terms of prioritization.[12–20] Given the past utility of this approach, we and others believe that this empirical scoring system allows for advantageous separation of genes based on our focus for identifying human MAP blood biomarker and by default, biomarkers of psychosis and SCZ.

RESULTS

We conducted a preliminary integrative RNA-sequencing study profiling peripheral blood gene expression from a primary cohort of 10 MA, 10 MAP and 10 healthy controls (Table 1 and Supplementary Figure 1). To identify and prioritize diagnostic blood biomarkers of MAP, a multimodal translational approach was used (Figure 1). A global gene co-expression network was first constructed using all the available subjects and identified 24 co-expression modules, which were functionally annotated to molecular factors, biological processes, cellular compartments, metabolic pathways, well-characterized drug compounds and cell type specificity (Supplementary Table 1).

Table 1. Recorded clinical characteristics from all subjects ($N = 30$)

	Healthy controls (N = 10)	*MA (N = 10)*	*MAP (N = 10)*	*ANOVA*		Post hoc *significance*
	Mean ± s.d.	*Mean ± s.d.*	*Mean ± s.d.*	X^2*(df = 2)*	P-*value*	*Bonferroni* P-*value*
Age	25.5 ± 5.8	24.8 ± 3.9	27.2 ± 8.3	0.040	0.980	
Education level	12.2 ± 1.2	10.7 ± 2.1	9.3 ± 1.7	10.788	0.005	Contol > MAP
METH age started using	—	18.6 ± 3.9	18.8 ± 6.8	0.191	0.662	
METH abstinence (days)	—	53.1 ± 82.9	45.5 ± 36.2	0.593	0.441	
METH duration of use (years)	—	5.8 ± 2.3	7.1 ± 3.0	0.688	0.407	
Nicotene use last 30 days	5	6	9	2.400	0.121	
Cannabis use last 30 days	2	2	1	0.529	0.467	
Alcohol use last 30 days	3	4	2	1.347	0.246	
EPQRS psychoticism	2.3 ± 1.7	1.6 ± 1.2	3 ± 2.1	1.880	0.391	
EPQRS extraversion	10.3 ± 2.5	8.2 ± 3.5	6.6 ± 2.5	7.039	0.030	Contol > MAP
EPQRS neuroticism	2.6 ± 1.8	4.6 ± 2.9	5.6 ± 3.2	4.624	0.099	
EPQRS lie	5.6 ± 2.3	4 ± 1.9	5.1 ± 3.3	1.902	0.386	
EPQRS total score	20.8 ± 5.3	18.5 ± 2.3	20.4 ± 4.7	1.876	0.391	
BIS	15.1 ± 1.5	15.8 ± 3.1	13.1 ± 3.6	3.018	0.221	
BAS drive	7.4 ± 2.5	8.3 ± 2.6	6.5 ± 1.3	2.267	0.322	
BAS fun seeking	7.1 ± 1.5	8.1 ± 1.6	6 ± 1.2	7.014	0.030	MA > MAP
BAS reward responsiveness	7.7 ± 1.9	7.2 ± 1.8	6.2 ± 1.7	3.859	0.145	
BIS/BAS total score	44.8 ± 5.8	47 ± 7.9	38.4 ± 5.6	6.269	0.044	
BDI total score	4.3 ± 3.0	17.3 ± 10.3	16.6 ± 12.5	10.363	0.006	Control > MAP; Control > MA
K10 total score	14 ± 3.8	18.2 ± 7.7	23.5 ± 8.2	7.944	0.019	Control > MAP
LEQ—sum of life events (⩽6 months)	2.6 ± 1.7	4.4 ± 2.0	4.7 ± 1.6	5.663	0.059	
LEQ—sum of life events (>6 months ago)	2.2 ± 2.2	4.2 ± 3.5	4.1 ± 2.0	3.643	0.162	

Abbreviations: BDI, Beck Depression Inventory; BIS/BAS, behavioural inhibition system/behavioural activation system; EPQRS, Eysenck Personality Questionnaire; K10, Kessler Psychological Distress Scale; LEQ, Life Events Questionnaire; MA, methamphetamine-dependent subjects with no psychotic events; MAP, methamphetamine-associated psychosis; PANSS, Positive and Negative Syndrome Scale. Shapiro wilk test was used to assess normality of variables and either a one-way analysis of variance (ANOVA) or KRUSKAL–Wallis ANOVA with *post hoc* Bonferroni correction was implemented accordingly.

Differential analysis of ME values and brain structure volumes

To reduce the number of multiple testing corrections and false positives arising from standard differential gene expression analyses, we calculated differences in module expression using ME values (See Materials and methods for complete description of ME). All the ME values were subjected to a Bayes analysis of variance[32] testing to compare the extent of module expression between the groups and the *P*-values were corrected for multiple comparisons. MAP-associated findings included significant decreases of ME expression in modules specific to 'ubiquitin-mediated proteolysis' (767 genes) and 'RNA degradation' (1156 genes) in MAP subjects compared with controls ($P = 0.01$, $P = 0.03$, respectively; Figures 2a and b). Further, an increase of ME expression in a module annotated as 'circadian clock' (332 genes) was observed in MAP compared with controls ($P = 0.04$; Figure 2c). MA-associated findings included the increase of ME expression in modules specific to 'chloride transporter activity' (106 genes), 'interferon signalling' (263 genes) and 'cytokine signalling' (186 genes), and a decrease of ME expression in modules associated to 'generic transcription' (48 genes) and 'ribosome pathway' (281 genes) in MA subjects relative to healthy controls (Supplementary Figure 2). The same methodology was extended to compare the brain structural volumes (mm^3) across the three groups, which revealed bilaterally reduced hippocampus volumes in MAP subjects (left, $P = 0.04$; right, $P = 0.02$; Table 2).

Phenotypic characterization of MAP modules

The ME values for MAP-specific modules were correlated with all phenotypic traits in this study (brain structural volumes, life history and psychometric measures) to gain insight into the role that each module may have in the pathophysiology of the disorder (Supplementary Figure 3). The *P*-values < 0.002 pass the most conservative multiple comparison correction (Bonferroni). The ME of a 'ubiquitin-mediated proteolysis' module was negatively associated to MAP status ($r = -0.45$, $P = 0.01$) as well as K10 total score ($r = -0.43$, $P = 0.02$). Interestingly, this module was also negatively associated with brain structure volumes in areas of the anterior CC ($r = -0.55$, $P = 0.002$), right accumbens area ($r = -0.40$, $P = 0.03$) and positively associated to areas in the left caudate ($r = 0.37$, $P = 0.04$) and left ventral diencephalon (DC, $r = 0.48$, $P = 0.007$). The 'RNA degradation' module was negatively associated with the CC anterior ($r = -0.48$, $P = 0.008$) and left accumbens ($r = 0.50$, $P = 0.005$), while positively associated with the left ventral DC ($r = 0.37$, $P = 0.04$). The 'circadian clock' module, was positively correlated with EPQRS measure of psychoticism ($r = 0.43$, $P = 0.02$) and negatively associated to extraversion ($r = -0.36$, $P = 0.04$).

Phenotypic characterization of MA modules

A similar strategy was chosen to characterize MA-specific modules (Supplementary Figure 3). The ME of the 'interferon signalling' module was positively associated to MA status ($r = 0.40$, $P = 0.03$), BDI total score ($r = 0.40$, $P = 0.03$), as well as structural information from both left ($r = 0.54$, $P = 0.002$) and right putamen areas ($r = 0.41$, $P = 0.03$). This module was negatively associated to EPQRS measure of extraversion ($r = -0.38$, $P = 0.04$) and EPQRS total score ($r = -0.38$, $P = 0.04$). Further, the ME of the 'chloride transporter activity' module was positively associated with both MA status ($r = 0.36$, $P = 0.05$) and METH dependency ($r = 0.39$, $P = 0.03$), in addition to BDI total score ($r = 0.39$, $P = 0.03$) and brain volume in the left putamen ($r = 0.53$, $P = 0.003$). This module was also negatively associated to control status ($r = -0.39$, $P = 0.03$) and the left ventral DC ($r = -0.40$, $P = 0.03$). The 'ribosome pathway' module was negatively associated to MA status ($r = -0.37$, $P = 0.04$) and positively associated to EPQRS total score ($r = 0.38$, $P = 0.04$) and K10 total score ($r = 0.44$, $P = 0.02$). The

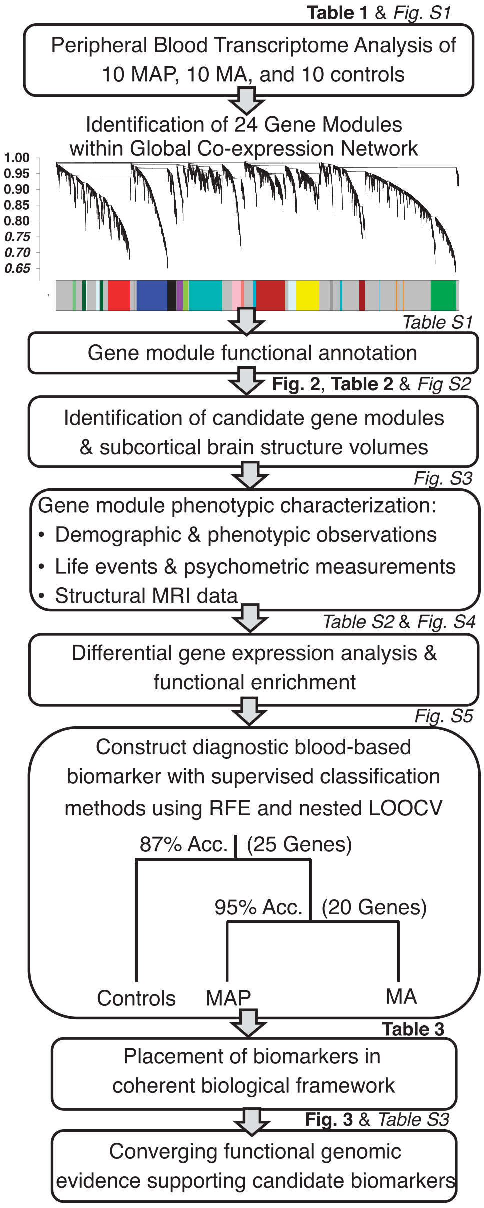


Table 1 & Fig. S1
Peripheral Blood Transcriptome Analysis of 10 MAP, 10 MA, and 10 controls
Identification of 24 Gene Modules within Global Co-expression Network
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
Table S1
Gene module functional annotation
Fig. 2, Table 2 & Fig S2
Identification of candidate gene modules & subcortical brain structure volumes
Fig. S3
Gene module phenotypic characterization:
• Demographic & phenotypic observations
• Life events & psychometric measurements
• Structural MRI data
Table S2 & Fig. S4
Differential gene expression analysis & functional enrichment
Fig. S5
Construct diagnostic blood-based biomarker with supervised classification methods using RFE and nested LOOCV
87% Acc.
(25 Genes)
95% Acc.
(20 Genes)
Controls
MAP
MA
Table 3
Placement of biomarkers in coherent biological framework
Fig. 3 & Table S3
Converging functional genomic evidence supporting candidate biomarkers

Figure 1. A multi-step translational work-flow for identifying methamphetamine-associated psychosis (MAP) biomarkers. First, weighted gene co-expression network analysis (WGCNA) analysis built a global co-expression network and identified 24 co-expression modules. On the hierarchical cluster tree, each line represents a gene (leaf) and each group of lines represents a discrete group of co-regulated genes or gene modules (branch) on the clustering gene tree. Each gene module is indicated by the colour bar below the dendrogram, and subsequently functionally annotated then integrated with recorded clinical and biological data to identify candidate gene modules representing functional biomarkers of MAP. Second, differential gene expression and class prediction methods identified 20 candidate MAP biomarkers (14 were recycled from the second split on the tree). A Bayesian-like convergent functional genomic (CFG) approach prioritized our panel of biomarkers specific to MAP and biomarkers were placed within an empirically derived biological framework. For each step, the corresponding figure and/or table is listed providing a quick reference. LOOCV, leave-one-out cross-validation; MA Dep., MA, methamphetamine-dependency without psychotic symptom; RFE, recursive feature elimination.

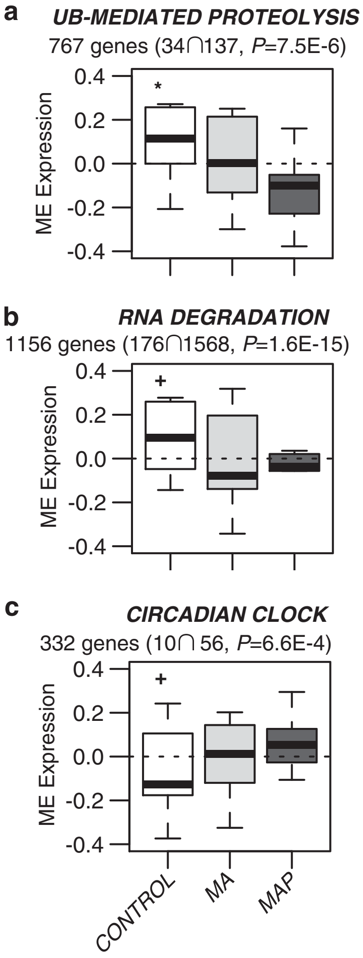


Figure 2. Significant methamphetamine-associated psychosis (MAP) findings from differential analysis of module eigengene (ME) values across controls (white), MA subjects (light grey) and MAP subjects (dark grey). Modules specific to MAP include (**a**) ubiquitin (UB)-mediated proteolysis, (**b**) RNA degradation and (**c**) circadian clock. Indicated for each module are number of overlapping genes from the module ∩ out of total genes in the term. Enrichment *P*-values are Bonferroni corrected for multiple comparisons. A Bayes analysis of variance (parameters: conf = 12, bayes = 1, winSize = 5) was used on the ME values to test for significance between the groups and *P*-values were corrected for multiple comparisons where (*) implies *post hoc*-corrected *P*-value significance < 0.05 and ($^{+}$) indicates *P*-value significance < 0.05 without *post hoc* correction. MA, methamphetamine-dependency without psychotic symptom.

'cytokine signalling' module was positively associated with both left accumbens ($r = 0.37$, $P = 0.04$) and right accumbens ($r = 0.55$, $P = 0.002$), whereas the 'generic transcription' module was negatively associated to these areas ($r = -0.49$, $P = 0.006$; $r = -0.60$, $P = 5\text{e-}04$, respectively).

Putative diagnostic blood biomarker for MAP

Supervised class prediction methods were used to identify any single important gene(s) that may have been over-looked in our network analysis. First, differentially expressed genes (all $P < 0.01$) were identified between the control and MA subjects ($N = 197$), control and MAP subjects ($N = 409$) and between the MA and MAP subjects ($N = 79$; Supplementary Table 2, Supplementary Figures 4A–D). To control for confounding factors, genes corresponding to WGCNA modules significantly associated to polysubstance abuse

Table 2. Brain structural volumes (mm^3) from all the subjects ($N=30$)

Brain region	*Healthy controls* (N = *10)*	*MA* (N = *10)*	*MAP* (N = *10)*	*Bayes ANOVA*		Post hoc *significance*
	Mean ± s.d.	*Mean ± s.d.*	*Mean ± s.d.*	X^2 *(df = 2)*	P-*value*	*Bonferroni* P-value
L hippocampus	3950.11 ± 463.71	3790 ± 297.51	3521.71 ± 173.43	3.538	0.041	Control > MAP
R hippocampus	4067.56 ± 414.08	4005.43 ± 196.29	3645.29 ± 189.97	4.261	0.029	Control > MAP
L accumbens	690.56 ± 80.38	689.14 ± 128.15	651.57 ± 99.24	0.343	0.714	
R accumbens	669.33 ± 100.54	673.00 ± 199.23	694.71 ± 91.48	0.076	0.927	
L caudate	4116.89 ± 340.84	4078.57 ± 293.78	3906.71 ± 177.23	1.149	0.337	
R caudate	4211.22 ± 251.11	4283.86 ± 314.36	4119 ± 163.64	0.760	0.481	
L putamen	6606.78 ± 408.97	6633.14 ± 667.17	6718.57 ± 661.5	0.078	0.925	
R putamen	6313.33 ± 371.03	6274.43 ± 596.45	6506.71 ± 672.14	0.373	0.694	
L ventral DC	4551.33 ± 247.16	4295.71 ± 273.56	4323.71 ± 204.25	2.715	0.091	
R ventral DC	4473.44 ± 377.34	4340.43 ± 78.7	4369.86 ± 278.58	0.485	0.623	
CC anterior	938.78 ± 125.96	1056.14 ± 194.83	1016.57 ± 100.31	1.389	0.272	
CC posterior	966.00 ± 191.65	912.29 ± 139.86	956.29 ± 135.16	0.236	0.792	

Abbreviations: CC, corpus callosum; DC, diencephalon; L, left; R, right. Bayes analysis of variance (ANOVA) parameters: conf = 12, Bayes = 1, winSize = 5. *P*-values corrected for multiple comparisons.

were excluded. Gene lists were annotated for functionality at the pathway level and cross-referenced with drug-induced gene signatures from the comparative toxicogenomics database (Supplementary Figure 4E and F; See Supplementary File for detailed information).

Subsequently, differentially expressed genes ($P<0.01$) were pooled from across the three candidate gene lists and subjected to recursive feature elimination feature selection and different multivariate classification methods in a leave-one-out cross-validation approach (See Materials and Methods for complete description). Two models were built for separating classes. First, when separating healthy controls form METH dependents (MA and MAP subjects) classification accuracy reached 87% when the expression of 25 genes was used with diagonal linear discriminant analysis multivariate classification method (Supplementary Figures 5a and b). Second, when separating MA from MAP, classification accuracy reached 95% when the expression of 20 genes (recycling 14 genes from the first model) was used with diagonal linear discriminant analysis (Supplementary Figures 5c and d).

We next sought to understand the biology represented by these MAP biomarkers and derive mechanistic insights. Our multi-step approach permitted taking each single biomarker and returning to our network analysis to retrieve guilt-by-association biological information from our empirically derived functional gene modules. Majority of these genes were found in a module annotated to 'RNA degradation' (*CLN3*, *FBP1*, *TBC1D2*, *ZNF821*, *ADAM15*, *ARL6*, *FBN1* and *MTHFSD*; Table 3). However, two top-scoring biomarkers were found to be implicated in 'circadian clock' dysfunction (*ELK3* and *SINA3*) and three other top-scoring biomarkers were found in the module annotated to 'ubiquitin-mediated proteolysis' (*PIGF*,*UHMK1* and *C7orf11*).

Prioritization and biological interpretation of blood biomarkers

Biomarkers were prioritized using a Bayesian-like CFG approach (Supplementary Figure 6) integrating previously published human evidence based on genetics (for example, GWAS, copy number variants), post-mortem brain gene expression and peripheral blood gene expression specific to psychosis, SCZ, depression/stress as well as neurocognitive impairment at the time of our analysis (August 2015). This is a way of validating relevant blood transcriptome biomarkers from moderately sized data sets, extracting generalizable signal out of potential cohort-specific noise.[12–20] Using the CFG approach, we first focused our attention on the 'ubiquitin-mediated proteolysis' annotated module, which in this study represents a functional biomarker of MAP. This module was enriched with 61 genes having CFG evidence ($P=4.8\text{E}-10$), including those found to be dysregulated in the blood of a psychotic disorder ($n=29$) as well as in the blood and/or post-mortem brain of SCZ patients ($n=32$) across independent human studies (Supplementary Table 3A). Notably, of the 29 CFG genes found in the blood of a psychotic disorder, 21 pertained to one single study.[45] We further found a significant enrichment of 39 genes holding CFG evidence ($P=7.0\text{E}-12$) within the module annotated as 'circadian clock' (Supplementary Table 3B). Similarly, these genes were also previously associated to psychosis and/or SCZ in independent studies. Of interest, two genes within the 'ubiquitin-mediated proteolysis' annotated module (*TMEM106B* and *SCAMP1*) and one within the 'circadian clock' annotated module (*DCTN1*) overlap with a previous study that had used CFG-based approach to validate blood biomarkers for delusions, a core symptom of psychotic disorders.[20] An additional gene (*RAB18*) within the 'ubiquitin-mediated proteolysis' module was also validated as a SCZ biomarker using the CFG approach.[18]

Applying the CFG approach to our panel of 31 discriminative biomarkers confirmed 8 candidate biomarkers for MAP (Table 3) which had a CFG score of 3 or above, meaning either maximal score from the *P*-value threshold cut-offs or at least two other lines of prior independent evidence (Figure 3a). Indeed, CFG evidence for 8 out of 31 discriminatory biomarkers is a significant overlap ($P=0.01$), beyond what would be expected by chance. Of these validated MAP biomarkers, four were previously reported to predict psychosis in an independent human blood transcriptome investigation (*FBP1*, *ZNF821*, *TBC1D2* and *SIN3A*), one of which was previously labelled a genetic variant for SCZ risk (*FBP1*). In addition, one other biomarker had been implicated in SCZ risk across two independent studies (*UHMK1*). Subsequently, a gene–disease network was built using all the CFG-validated biomarkers, either in the form of a functional biomarker (gene modules) or single biomarkers, to visualize unique gene signatures of MAP and consensus signatures of MAP, psychosis and SCZ (Figure 3b). In this study, we found that MAP shares 69 genes with SCZ, 39 genes with other psychotic disorders and six genes are shared across all the three conditions. Importantly, cross-referencing all the candidate MAP genes onto query haloperidol gene expression signatures from the CMap and CDT provided preliminary evidence

Table 3. Top informative features for separating controls from METH subjects (25 genes) and MA from MAP subjects (20 genes)

Gene symbol	*Parametric P-value*	*%CV support*	*Module correspondence*	*Significant positive correlations*	*Significant negative correlations*
Top 25 informative features separating controls from METH subjects					
ELK3†	0.0175377	97	Circadian clock	EPQRS psychoticism (*r* = 0.43, *P* = 0.02) CC posterior (*r* = 0.39, *P* = 0.03)	EPQRS extraversion (*r* = − 0.38, *P* = 0.04)
CRTAM	0.03485	97	Generic transcription	EPQRS neuroticisim (*r* = 0.41, *P* = 0.02) EPQRS total (*r* = 0.37, *P* = 0.05)	Left accumbens (*r* = − 0.49, *P* = 0.0006) Right accumbens (*r* = − 0.6, *P* = 0.00005)
MAGEE1	0.0158379	100	Generic transcription		
RNF138P1	0.0078459	87	RNA degradation	Control status (*r* = 0.38, *P* = 0.04) Left ventral DC (*r* = 0.37, *P* = 0.04)	CC anterior (*r* = − 0.48, *P* = 0.008) Right accumbens (*r* = − 0.5, *P* = 0.005)
MFN1	0.0070206	87	RNA degradation		
TBC1D2***	0.000805	90	RNA degradation		
ZNF286B	0.000065	90	RNA degradation		
MRPL50	0.0001267	93	RNA degradation		
ADAM15†	0.000731	97	RNA degradation		
DDRGK1	0.0055266	97	RNA degradation		
MTHFSD	0.0612426	97	RNA degradation		
ARL6	0.0037554	97	RNA degradation		
GKAP1	0.0009407	97	RNA degradation		
FAM169A	0.0008839	97	RNA degradation		
KBTBD6	0.0003548	97	RNA degradation		
ZSCAN5A	0.0012892	100	RNA degradation		
FBN1†,*	0.0168567	100	RNA degradation		
ZNF821	0.0193724	100	RNA degradation		
FBP1†	0.6900462	100	RNA degradation		
CDK7	0.0054503	100	RNA degradation		
CDYL2	0.0000834	93	RNA-binding	K10 total (*r* = 0.42, *P* = 0.02)	Right ventral DC (*r* = − 0.42, *P* = 0.02)
TOMM34	0.0019291	100	RNA-binding		
C7orf11	0.0587445	80	Ubiquitin-mediated proteolysis	Control status (*r* = 0.4, *P* = 0.03) Left caudate (*r* = 0.37, *P* = 0.04) Left ventral DC (*r* = 0.48, *P* = 0.0007)	Control status (*r* = 0.4, *P* = 0.03) Left caudate (*r* = 0.37, *P* = 0.04) Left ventral DC (*r* = 0.48, *P* = 0.0007)
UHMK1†	0.6057577	97	Ubiquitin-mediated proteolysis		
PHLDB2	0.0007613	100	Ubiquitin-mediated proteolysis		
Top 20 informative features separating MA from MAP subjects					
*SIN3A**	0.0926295	70	Circadian clock	EPQRS psychoticism (*r* = 0.43, *P* = 0.02) CC posterior (*r* = 0.39, *P* = 0.03)	EPQRS extraversion (*r* = − 0.38, *P* = 0.04)
ELK3†	0.0002902	90	Circadian clock		
MAGEE1	0.0001558	100	Generic transcription	EPQRS neuroticisim (*r* = 0.41, *P* = 0.02) EPQRS total (*r* = 0.37, *P* = 0.05)	Left accumbens (*r* = − 0.49, *P* = 0.0006) Right accumbens (*r* = − 0.6, *P* = 0.00005)
MFSD7	0.0440767	85	Interferon signalling	MA dep. status (*r* = 0.4, *P* = 0.03) BDI total (*r* = 0.4, *P* = 0.03) Left putamen (*r* = 0.54, *P* = 0.002) Right putamen (*r* = 0.41, *P* = 0.03)	EPQRS extraversion (*r* = − 0.43, *P* = 0.02) EPQRS total (*r* = − 0.43, *P* = 0.02) CC posterior (*r* = − 0.43, *P* = 0.02)
SLC41A3	0.0018933	100	Ribosome pathway	EPQRS total (*r* = 0.38, *P* = 0.04) BDI total (*r* = 0.44, *P* = 0.02)	MA status (*r* = − 0.37, *P* = 0.04)
MTHFSD	0.0002405	90	RNA degradation	Control status (*r* = 0.38, *P* = 0.04) Left ventral DC (*r* = 0.37, *P* = 0.04)	CC anterior (*r* = − 0.48, *P* = 0.008) Right accumbens (*r* = − 0.5, *P* = 0.005)
ZNF821*	0.11798	90	RNA degradation		
FBP1*	0.3549855	90	RNA degradation		
RNF138P1	0.0195014	90	RNA degradation		
ARL6	0.0317958	95	RNA degradation		
ETFA	0.0235683	95	RNA degradation		
TBC1D2*	0.157939	100	RNA degradation		
FAM169A	0.0132909	100	RNA degradation		
ZSCAN5A	0.0112376	100	RNA degradation		
CLN3	0.0087815	100	RNA degradation		
DDRGK1	0.0082958	100	RNA degradation		
FBN1†,*	0.0070075	100	RNA degradation		
PIGF	0.1818898	90	Ubiquitin-mediated proteolysis	Control status (*r* = 0.4, *P* = 0.03) Left caudate (*r* = 0.37, *P* = 0.04) Left ventral DC (*r* = 0.48, *P* = 0.0007)	Control status (*r* = 0.4, *P* = 0.03) Left caudate (*r* = 0.37, *P* = 0.04) Left ventral DC (*r* = 0.48, *P* = 0.0007)
C7orf11	0.0028377	90	Ubiquitin-mediated proteolysis		
PHLDB2	0.1302545	95	Ubiquitin-mediated proteolysis		

Abbreviations: BDI, Beck Depression Inventory; CC, corpus callosum; DC, diencephalon; EPQRS, Eysenck Personality Questionnaire. Parametric *P*-value indicates significance in a strict sense following 1000 random permutations to group labels using small *N*. %CV support denotes the number of correctly passed cross-validations for each gene. Module correspondence is the module membership of each gene and the subsequent significant correlations for each module are depicted. Genes in bold are those that were used in classification for nodes 1 and 2 (14 genes total). (*) indicates genes found dysregulated in the blood of psychosis studies; (†) indicates genes found as genetic variants in SCZ studies.

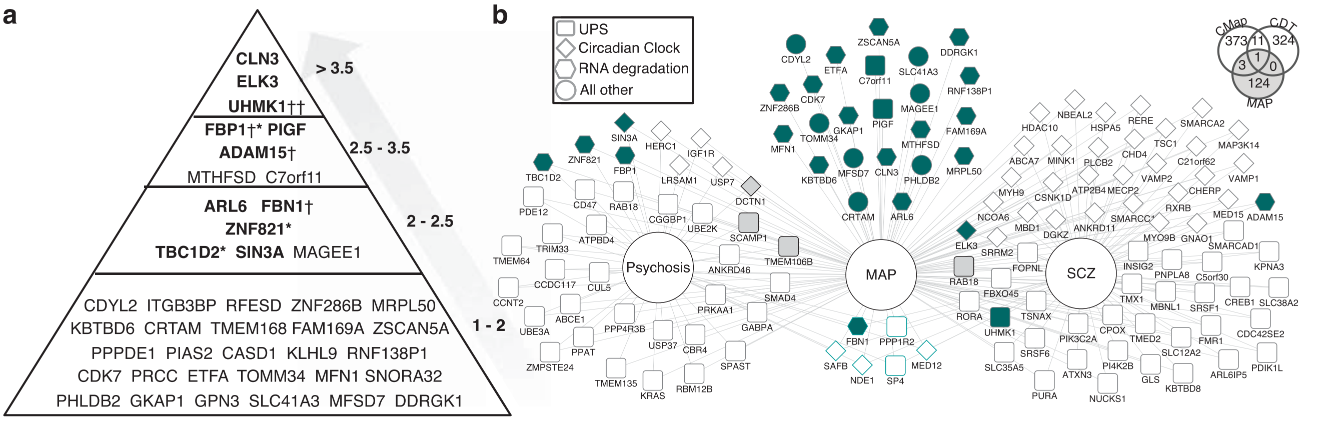


Figure 3. Top candidate blood biomarkers for methamphetamine-associated psychosis (MAP). (**a**) Convergent functional genomic (CFG) evidence and scoring are depicted on the right side of the pyramid. Genes in bold have been found in external publications. Genes found in methamphetamine (METH)-free studies investigating schizophrenia (SCZ, †) and psychosis (*) are as indicated. (**b**) Overlapping gene–disease relationships including CFG-validated genes within gene modules (ubiquitin-mediated proteolysis and circadian rhythm) and single-gene biomarkers. Nodes represent genes and edges indicate gene–disease relationships. Node shape denotes empirically derived functions from our network analysis. Green shading indicates biomarkers from our machine-learning analysis including 14 unique genes separating controls from METH dependants. Grey nodes represent CFG-validated biomarkers of delusion (psychosis) or SCZ.[11,17] Node border colour in turquoise indicates gene signatures across MAP, general psychosis and SCZ studies. Venn diagram depicts lack of overlap from curated haloperidol gene signatures onto the 128 candidate MAP genes (61 UPS+39 clock+25+20 = 128 genes (while accounting for overlap across lists)).

for the lack of neuroleptic-associations across our candidate findings (Figure 3b).

DISCUSSION

This preliminary report describes gene networks and blood biomarkers of MAP, further validating the MAP model as an exemplar for discovery of biomarkers related to SCZ susceptibility and clinical course. In essence, this pharmacogenomics approach is a tool for identifying genes that contain pathophysiological relevance to psychotic disorders and SCZ. Considering the variable environmental component of MAP, it is possible that not all subjects would show changes in all the biomarker genes. Hence, our multimodal approach incorporated blood gene expression, clinical assessment of life history, psychometric measures and structural MRI data revealing several mechanistic insights regarding the pathophysiology of MAP and its overlapping mechanistic nature with psychotic disorders and SCZ. First, we identified a functional biomarker of MAP in the form of a co-expression module annotated to ubiquitin-mediated proteolysis, further enriched with 61 genes containing CFG evidence. We also revealed a psychoticism-associated module implicated in circadian clock, enriched with 39 genes containing CFG evidence. Second, we identified 25 genes that were able to distinguish healthy controls from METH dependents with high accuracy, while only 20 genes (recycling 14 genes from the previous split) were able to differentiate between MA and MAP subjects. A significant proportion of these single blood biomarkers also contained CFG evidence. Further, cross-referencing these results onto haloperidol specific gene expression signatures reduced the likelihood of these genes being neuroleptic-related. These high overlaps suggest similar biological mechanisms detectable in peripheral blood underlying the pathophysiology of psychosis, regardless of substance abuse. These findings also outline new avenues regarding how the MAP model may function in SCZ research.

A central finding from our network analysis was the identification of a functional biomarker (gene module) annotated to ubiquitin-mediated proteolysis expressed to a lesser extent in MAP subjects (Figure 2a). The ubiquitin proteasome system (UPS) is a highly complex and tightly regulated process that has major roles in a variety of basic cellular processes, specifically degradation of intracellular proteins and modulation of cellular responses to inflammation and oxidative stress.[46] The UPS has been identified in genetic reports as a canonical pathway associated to psychosis,[45,47] SCZ,[48–52] bipolar disorder,[48,53] as well as neurodegenerative diseases such as Alzheimer's[54] and Parkinson's.[55] Studies using post-mortem brain gene expression to investigate mechanisms of psychosis and SCZ provide consistent evidence for the downregulation of UPS-related genes in these conditions.[50–52] It was also recently shown that UPS abnormalities disrupt expression at the protein level in SCZ.[56] Interestingly, studies using peripheral blood gene expression also found that the UPS pathway was consistently dysregulated across bipolar, SCZ and psychosis patient groups.[48] A later study used a targeted approach associating blood expression measurements of UPS pathway gene members with Scales for Assessment of Positive and Negative Symptoms and determined *UBE2K* (also a gene member of our 'ubiquitin-mediated proteolysis' module), was one of three genes most significantly associated to positive symptoms of psychosis.[47] Another independent report built a diagnostic blood-based classifier able to distinguish first-episode psychosis from controls with 400 genes,[45] 21 of which were found within our UPS annotated module (Supplementary Table 3A). Indeed, it is interesting that genes that have a well-established role in brain functioning should also show changes in peripheral blood in relationship to psychiatric symptom states, and moreover that the direction of change should be concordant with that reported in human post-mortem brain studies. As a consequence of the overlapping nature of UPS dysfunction found across mental diseases, the proteasome system has emerged as a putative candidate highlighting both mRNA and protein-level changes in psychosis and SCZ. This clearly is an area that deserves attention and mechanistic elucidation by future hypothesis-driven research.

In determining relationships between blood gene expression and structural MRI data, we revealed a significant association of the ubiquitin-mediated proteolysis module to the anterior CC ($r = -0.55$, $P = 0.002$; Supplementary Figure 3). Conversely, the circadian clock module, expressed to a greater extent in MAP subjects (Figure 2), was significantly associated to EPQRS measure of psychoticism (that is, aggression, egocentrism and impulsiveness; $r = 0.43$, $P = 0.02$) and the posterior CC ($r = 0.39$, $P = 0.03$; Supplementary Figure 3). There is considerable evidence suggesting that global white matter abnormalities (that is, disruptions in connectivity in intra- and interhemispheric pathways) have a role in the pathophysiology of psychiatric disorders.[57] With the CC being the largest white matter tract containing highly packed neuronal fibres, abnormalities in this structure have frequently been reported in patients with SCZ,[58] including first-episode SCZ and psychosis patients,[59] often relating to the severity of psychotic symptoms. It has been hypothesized that less efficient connectivity and resulting aberrant signal transmission between the brain regions may be a pivotal factor in the manifestation of psychotic symptoms, including delusions and hallucinations, and of cognitive dysfunctions.[60,61] However, these disturbances have not been fully elucidated in the context of MAP nor in its relationship to blood gene expression differences. Yet most interestingly, we also observed significantly lower bilateral hippocampal volumes in MAP subjects (Table 2). Although correlates of blood gene expression to hippocampal volumes relate mainly to processes of protein ubiquitination ($r = 0.37$, $P = 0.05$), reductions in the hippocampal volumes are consistent with previous reports of pathological hippocampus changes in MAP,[62] in first-episode and chronic schizophrenia,[63] and in individuals at high risk for psychosis.[64] Taken together, these results suggest that changes in the blood occur in parallel to structural changes in the brain of MAP subjects and that they are also most likely involved in the pathophysiology of psychotic disorders and SCZ in the absence of METH.

The interrogation of the comparative toxicogenomics database[41] with a signature query composed of the genes in our 'ubiquitin-mediated proteolysis' annotated module revealed an enrichment of sodium arsenate gene signatures (Supplementary Table 1). Although sodium arsenate is one of the most toxic metals derived from the natural environment,[65] it has been used as a therapeutic medication in acute promylocytic leukaemia based on its mechanism to induce apoptotic effects via release of apoptosis-inducing factor.[65] However, arsenic is mainly a contaminator and interestingly is known to cause clinical features such as psychosis, toxic cardiomyopathy and seizures.[66,67] This exploratory result suggests arsenic, and chemically similar compounds, as a putatively useful gene-hunting tool for investigating future mechanisms of psychosis in either primary or patient-derived lymphoblast cell lines to elucidate further these effects in search for more verifiable biomarkers.

Topping our list of candidate MAP biomarkers, we found eight genes involved in RNA degradation (*CLN3*, *FBP1*, *TBC1D2*, *ZNF821*, *ADAM15*, *ARL6*, *FBN1* and *MTHFSD*), two specific to circadian rhythm (*ELK3* and *SINA3*) and three involved in ubiquitin-mediated proteolysis (*PIGF*,*UHMK1* and *C7orf11*; Table 3). Indeed it is possible that some of the gene expression changes detected in this moderately sized cohort ($N = 30$) may represent biological or technical artefacts. To minimize such effects, our candidate MAP biomarkers were selected based on having a line of evidence (CFG) score of two or higher (Figure 3a). Proper cross-validation both *in silico* and across-literature (CFG), minimized the likelihood of having identified false positives while increasing sensitivity and specificity in the ability to distinguish true signal (biomarkers) from noise through a fit-to-disease Bayesian-like methodology.[12–20]

CLN3 (Ceroid-Lipofuscinosis, Neuronal 3) was the top-scoring gene in our study and is conventionally involved in lysosome function. Mutations in this gene are well known to cause neurodegenerative diseases such as Batten disease,[66,68] which impairs mental and motor development during childhood, causing difficulty with walking, speaking and intellectual functioning. Patients with a CLN3 mutation are also prone to recurrent seizures, epilepsies, vision impairment and occasionally psychosis. It is hypothesized that mutations in CLN3 disrupt lysosome function resulting in build-up of lipopigments, which may induce apoptotic effects in brain neurons. Although this gene has not yet been discussed in the context of psychosis, it may represent a putative biomarker of MAP. In addition, variants in the gene *FBP1* (fructose-1,6-bisphosphatase 1) have previously provided genetic support for the view that alterations in glucose metabolism are intrinsic to SCZ pathology.[69] However, in our study, this gene was found co-expressed in the 'RNA degradation' module. Other top-scoring genes included genes annotated to a circadian clock module (Supplementary Table 3B), which are involved in sleep–wake cycles and previously identified as risk factors for psychosis,[12] anxiety disorders,[17] suicidality[19] and mood disorders.[70] *ELK3* (ETS-Domain Protein (SRF Accessory Protein 2)) encodes a transcriptional factor that may switch from activator to repressor in the presence of Ras, whereas *SIN3A* (SIN3 Transcription Regulator Family Member A) encodes a transcriptional repressor with known roles in circadian clock negative feedback.[71] Although *SIN3A* has well-known association to circadian clock function, an advantage of our approach was to be able to derive guilt-by-association co-expression interpretation of biomarkers, such as *ELK3*, by indicating module membership status. Dysregulation of circadian clock genes in post-mortem brain of SCZ patients have previously been observed,[72] however, reports in the blood are less frequent.

Of note, MA-associated findings also allow us to speculate on molecular mechanisms of psychosis. MA discoveries mainly included elevated expression in modules specific to interferon and cytokine signalling. Although cytokine signalling was positively associated to METH dependency (that is, MA and MAP subjects; $r = 0.39$, $P = 0.03$), a module specific to 'interferon signalling' was significantly overexpressed in the blood of MA subjects relative to controls, rather than MAP subjects relative to controls (Supplementary Figure 2). Previous work has highlighted a weak or absent immune stress response, specific to HPA axis activation[73] and cortisol measurements,[74] in medication-naive first-onset psychosis patients. Moreover, modules specific to IL-5 signalling, actin cytoskeleton and ATPase activity all showed a strong association to both the left and right accumbens area (Supplementary Figure 3). Owing to high levels of dopaminergic innervations, the nucleus accumbens, together with other subcortical structures, has a pivotal role in several neurocircuits involved in reward, motivation, drug-reinforcement and drug-seeking behaviour, mood regulation and sleep–wake cycles.[75,76] Such neurocircuit functions are similarly affected by drug exposure as well as stressors, life events or social pressure, with increased dopamine release in the nucleus accumbens triggered by the stimulant in addiction and by glucocorticoid hormones in stress.[75] Furthermore, there is emerging evidence that cytokines circulating in blood may target subcortical dopamine function, with potential implications on behaviour, sleep patterns and the progression of psychiatric disorders, such as depression.[77]

Although it appears that the identification of blood-based biomarkers may be accomplished by systems level and machine-learning approaches, it remains an open empirical question for future work, which approach provides the most favourable translational avenues. Systems approaches are particularly useful in providing comprehensive characterizations of the molecular factors for a given disease state, multi-scale data integration and are statistically robust in terms of reproducibility. Machine-learning applications, while often fit-to-cohort, rank genes by importance producing a unique predictive or diagnostic panel of biomarkers. This dual approach permitted the placement of MAP single-gene biomarkers into an empirically derived biological framework (that is, gene network) to derive mechanistic insights. Pragmatically, these results provide a proof of principle for joint statistical analysis providing complimentary and comprehensive molecular characterizations in pursuit of blood biomarkers for MAP. A limitation of this study is that our findings cannot yet be used to change the clinical practice. Notwithstanding that many of our MAP single-gene biomarkers identified by machine learning were supported by CFG evidence, these findings need to be replicated in an independent MAP sample.

Overall, our results support the MAP model for the identification of biomarkers involved in psychosis and SCZ. Our most significant findings suggest that genes involved in UPS and circadian clock dysregulation are prominent players in psychosis and are reflected in both peripheral blood and post-mortem brain profiles. Specifically, UPS abnormalities have emerged as a common denominator across a variety of independent studies investigating psychosis, SCZ and bipolar disorder. Indeed in clinical practice there is a high degree of overlap and comorbidity between psychotic disorders, MAP and SCZ. Our results were able to shed light on the biological mechanisms of psychosis, regardless of polysubstance abuse, medication or other confounding factors and further emphasize the value of moving towards comprehensive empirical profiling. These results also open empirical avenues for future field trials, clinical testing and validation in various at-risk populations.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

DJS is funded by the Medical Research Council of South Africa. This work is partially supported through participation in EU Marie Curie International Staff Exchange Scheme grant for the European South African Research Network in Anxiety Disorders (EUSARNAD; PIRSES-GA-2010-269213). We thank Dr Baldwin with EUSARNAD collaboration, Dr P-A Kivistik for help in study design and Ms K Grand and Mr V Soo for technical help.

REFERENCES

1 Yang MH, Jung MS, Lee MJ, Yoo KH, Yook YJ, Park EY *et al.* Gene expression profiling of the rewarding effect caused by methamphetamine in the mesolimbic dopamine system. *Mol Cells* 2008; **26**: 121–130.

2 United Nations Office on Drugs and Crime. World Drug Report 2004. UN Office on Drugs and Crime: Vienna, Austria.

3 Kapp C. Crystal meth boom adds to South Africa's health challenges. *Lancet* 2008; **371**: 193–194.

4 Nutt DJ, King LA, Phillips LD. Drug harms in the UK: a multicriteria decision analysis. *Lancet* 2010; **376**: 1558–1565.

5 Srisurapanont M, Ali R, Marsden J, Sunga A, Wada K, Monteiro M. Psychotic symptoms in methamphetamine psychotic in-patients. *Int J Neuropsychopharmacol* 2003; **6**: 347–352.

6 Smith MJ, Thirthalli J, Abdallah AB, Murray RM, Cottler LB. Prevalence of psychotic symptoms in substance users: a comparison across substances. *Compr Psychiatry* 2009; **50**: 245–250.

7 Bousman CA, Glatt SJ, Everall IP, Tsuang MT. Genetic association studies of methamphetamine use disorders: a systematic review and synthesis. *Am J Med Genet B Neuropsychiatr Genet* 2009; **150B**: 1025–1049.

8 Hsieh JH, Stein DJ, Howells FM. The neurobiology of methamphetamine induced psychosis. *Front Hum Neurosci* 2014; **8**: 537.

9 Srisurapanont M, Arunpongpaisal S, Wada K, Marsden J, Ali R, Kongsakon R. Comparisons of methamphetamine psychotic and schizophrenic symptoms: a differential item functioning analysis. *Prog Neuropsychopharmacol Biol Psychiatry* 2011; **35**: 959–964.

10 Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008; **9**: 559.

11 Simon R, Lam A, Li M-C, Ngan M, Menenzes S, Zhao Y. Analysis of gene expression data using BRB-Array tools. *Cancer Inform* 2007; **2**: 11–17.

12 Niculescu A, Segal D, Kuczenski R, Barrett T, Hauger R, Kelsoe J. Identifying a series of candidate genes for mania and psychosis: a convergent functional genomics approach. *Physiol Genomics* 2000; **4**: 83–91.

13 Ogden CA, Rich ME, Schork NJ, Paulus MP, Geyer MA, Lohr JB *et al.* Candidate genes, pathways and mechanisms for bipolar (manic-depressive) and related disorders: an expanded convergent functional genomics approach. *Mol Psychiatr* 2004; **9**: 1007–1029.

14 Patel SD, Le-Niculescu H, Koller DL, Green SD, Lahiri DK, McMahon FJ *et al.* Coming to grips with complex disorders: genetic risk prediction in bipolar disorder using panels of genes identified through convergent functional genomics. *Am J Med Genet B Neuropsychiatr Genet* 2010; **153B**: 850–877.

15 Le-Niculescu H, Patel SD, Bhat M, Kuczenski R, Faraone SV, Tsuang MT *et al.* Convergent functional genomics of genome-wide association data for bipolar disorder: comprehensive identification of candidate genes, pathways and mechanisms. *Am J Med Genet B Neuropsychiatr Genet* 2009; **150B**: 155–181.

16 Rodd ZA, Bertsch BA, Strother WN, Le-Niculescu H, Balaraman Y, Hayden E *et al.* Candidate genes, pathways and mechanisms for alcoholism: an expanded convergent functional genomics approach. *Pharmacogenomics J* 2007; **7**: 222–256.

17 Le-Niculescu H, Balaraman Y, Patel SD, Ayalew M, Gupta J, Kuczenski R *et al.* Convergent functional genomics of anxiety disorders: translational identification of genes, biomarkers, pathways and mechanisms. *Transl Psychiatr* 2011; **1**: e9.

18 Ayalew M, Le-Niculescu H, Levey DF, Jain N, Changala B, Patel SD *et al.* Convergent functional genomics of schizophrenia: from comprehensive understanding to genetic risk prediction. *Mol Psychiatry* 2012; **17**: 887–905.

19 Le-Niculescu H, Levey DF, Ayalew M, Palmer L, Gavrin LM, Jain N *et al.* Discovery and validation of blood biomarkers for suicidality. *Mol Psychiatry* 2013; **18**: 1249–1264.

20 Kurian SM, Le-Niculescu H, Patel SD, Bertram D, Davis J, Dike C *et al.* Identification of blood biomarkers for psychosis using convergent functional genomics. *Mol Psychiatry* 2009; **16**: 37–58.

21 First MB, Gibbon M, Spitzer RL, Williams JBW. User's Guide for the Structured Clinical Interview for DSM-IV-TR Axis I Disorders—Research Version—(SCID-I for DSM-IV-TR, November 2002 Revision).

22 Brugha TS, Cragg D. The List of Threatening Experiences: the reliability and validity of a brief life events questionnaire. *Acta Psychiatr Scand* 1990; **82**: 77–81.

23 Kessler RC, Andrews G, Colpe LJ, Hiripi E, Mroczek DK, Normand SL *et al.* Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychol Med* 2002; **32**: 959–976.

24 Beck AT, Steer RA, Brown GK. *Manual for the Beck Depression Inventory-II*. Psychological Corporation: San Antonio, TX, USA, 1996.

25 Carver C, White T. Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: the BIS/BAS scales. *J Pers Soc Psychol* 1994; **67**: 319–333.

26 Eysenck S, Eysenck H, Barrett P. A revised version of the psychoticism scale. *Pers Individ Diff* 1985; **6**: 21–29.

27 Kay SR, Fiszbein A, Opler LA. The Positive and Negative Syndrome Scale (PANSS) for schizophrenia. *Schizophr Bull* 1987; **13**: 261–276.

28 van der Kouwe AJ, Benner T, Salat DH, Fischl B. Brain morphometry with multiecho MPRAGE. *Neuroimage* 2008; **40**: 559–569.

29 Fischl B, Salat DH, van der Kouwe AJ, Makris N, Segonne F, Quinn BT *et al.* Sequence-independent segmentation of magnetic resonance images. *Neuroimage* 2004; **23**: S69–S84.

30 Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina Sequence Data. *Bioinformatics* 2014; **30**: 2114–2120.

31 Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 2009; **25**: 1105–1111.

32 Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009; **10**: R25.

33 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N *et al.* The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* 2009; **25**: 2078–2079.

34 Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* 2015; **31**: 166–169.

35 Law CW, Chen Y, Shi W, Smyth GK. *Voom Precision Weights Unlock Linear Model Analysis Tools for RNA-Seq Read Counts. Technical Report Bioinformatics Division*. Walter and Eliza Hall Institute of Medical Research: Melbourne, VIC, Australia, 2013; [http://www.statsci.org/smyth/pubs/13 5 1-voom-techreport].

36 Breen MS, Beliakova-Bethell N, Mujica-Parodi LR, Carlson JM, Ensign WY, Woelk CH *et al.* Acute psychological stress induces short-term variable immune response. *Brain Behav Immun* 2015; **53**: 172–182.

37 Breen MS, Maihofer AX, Glatt SJ, Tylee DS, Chandler SD, Tsuang MT *et al.* Gene networks specific for innate immunity define post-traumatic stress disorder. *Mol Psychiatry* 2015; **20**: 1538–1545.

38 Kayala MA, Baldi P. Cyber-T web server: differential analysis of high-throughput data. *Nucleic Acids Res* 2012; **40**: W553–W559.

39 Smyth GK. Limma: linear models for microarray data. In: Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W (eds). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer: New York, NY, USA, 2005, pp 397–420.

40 Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res* 2009; **37**: W305–W311.

41 Davis AP, Grondin CJ, Lennon-Hopkins K, Saraceni-Richards C, Sciaky D, King BL *et al*. The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. *Nucleic Acids Res* 2015; **43**, (Database issue) D914–D920.
42 Shoemaker JE, Fukuyama S, Sakabe S, Kitano H, Kawaoka Y. CTen: a web-based platform for identifying enriched cell types from heterogenous microarray data. *BMC Genomics* 2011; **13**: 460.
43 Breen MS, Stein DJ, Baldwin DS. A systematic review of blood transcriptomics and complex brain disorders: moving beyond 'surrogate marker' status. *Hum Psychopharmacol* 2015 (in review).
44 Piñero J, Queralt-Rosinach N, Bravo A, Deu-Pons J, Bauer-Mehren A, Baron M *et al*. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database* 2015; **2015**: abav028.
45 Lee J, Goh L-K, Chen G, Verma S, Tan C-H, Lee T-S. Analysis of blood-based gene expression signature in first-episode psychosis. *Psychiatry Res* 2012; **200**: 52–54.
46 Ciechanover A, Orian A, Schwartz AL. Ubiquitin-mediated proteolysis: biological regulation via destruction. *Bioessays* 2000; **22**: 442–451.
47 Bousman CA, Chana G, Glatt SJ, Chandler SD, May T, Lohr J *et al*. Positive symptoms of psychosis correlate with expression of ubiquitin proteasome genes in peripheral blood. *Am J Med Genet Part B* 2010; **153B**: 1336–1341.
48 Bousman CA, Chana G, Glatt SJ, Chandler SD, Lucero GR, Tatro E *et al*. Preliminary evidence of ubiquitin proteasome system dysregulation in schizophrenia and bipolar disorder: convergent pathway analysis findings from two independent samples. *Am J Med Genet Part B* 2010; **153B**: 494–502.
49 Vawter MP, Barrett T, Cheadle C, Sokolov BP, Wood WH III, Donovan DM *et al*. Application of cDNA microarrays to examine gene expression differences in schizophrenia. *Brain Res Bull* 2001; **55**: 641–650.
50 Vawter MP, Crook JM, Hyde TM, Kleinman JE, Weinberger DR, Becker KG *et al*. Microarray analysis of gene expression in the prefrontal cortex in schizophrenia: a preliminary study. *Schizophr Res* 2002; **58**: 11–20.
51 Middleton FA, Mirnics K, Pierri JN, Lewis DA, Levitt P. Gene expression profiling reveals alterations of specific metabolic pathways in schizophrenia. *J Neurosci* 2002; **22**: 2718–2729.
52 Altar CA, Jurata LW, Charles V, Lemire A, Liu P, Bukhman Y *et al*. Deficient hippocampal neuron expression of proteasome, ubiquitin, and mitochondrial genes in multiple schizophrenia cohorts. *Biol Psychiatry* 2005; **58**: 85–96.
53 Konradi C, Eaton M, MacDonald ML, Walsh J, Benes FM, Heckers S. Molecular evidence for mitochondrial dysfunction in bipolar disorder. *Arch Gen Psychiatry* 2004; **61**: 300–308.
54 Lam YA, Pickart CM, Alban A, Landon M, Jamieson C, Ramage R *et al*. Inhibition of the ubiquitin-proteasome system in Alzheimer's disease. *Proc Natl Acad Sci USA* 2000; **97**: 9902–9906.
55 Shimura H, Schlossmacher MG, Hattori N, Frosch MP, Trockenbacher A, Schneider R *et al*. Ubiquitination of a new form of alpha-synuclein by Parkin from human brain: implications for Parkinson's disease. *Science* 2001; **293**: 263–269.
56 Rubio M, Wood K, Haroutunian V, Meador-Woodruff J. Dysfunction of the Ubiquitin Proteasome and Ubiquitin-Like Systems in Schizophrenia. *Neuropsychopharmacology* 2013; **38**: 1910–1920.
57 White T, Nelson M, Lim KO. Diffusion tensor imaging in psychiatric disorders. *Top Magn Reson Imaging* 2008; **19**: 97–109.
58 Whitford TJ, Kubicki M, Schneiderman JS, O'Donnell LJ, King R, Alvarado JL *et al*. Corpus callosum abnormalities and their association with psychotic symptoms in patients with schizophrenia. *Biol Psychiatry* 2010; **68**: 70–77.
59 Price G, Cercignani M, Parker GJ, Altmann DR, Barnes TR, Barker GJ *et al*. Abnormal brain connectivity in first-episode psychosis: a diffusion MRI tractography study of the corpus callosum. *Neuroimage* 2007; **35**: 458–466.
60 Friston KJ, Frith CD. Schizophrenia: a disconnection syndrome? *Clin Neurosci* 1995; **3**: 89–97.
61 Kubicki M, McCarley R, Westin CF, Park HJ, Maier S, Kikinis R *et al*. A review of diffusion tensor imaging studies in schizophrenia. *J Psychiatr Res* 2007; **41**: 15–30.
62 Orikabe L, Yamasue H, Inoue H, Takayanagi Y, Mozue Y, Sudo Y *et al*. Reduced amygdala and hippocampal volumes in patients with methamphetamine psychosis. *Schizophr Res* 2011; **132**: 183–189.
63 Velakoulis D, Wood SJ, Wong MT, McGorry PD, Yung A, Phillips L *et al*. Hippocampal and amygdala volumes according to psychosis stage and diagnosis: a magnetic resonance imaging study of chronic schizophrenia, first-episode psychosis, and ultra-high-risk individuals. *Arch Gen Psychiatry* 2006; **63**: 139–149.
64 Fusar-Poli P, Howes OD, Allen P, Broome M, Valli I, Asselin M-C *et al*. Abnormal prefrontal activation directly related to pre-synaptic striatal dopamine dysfunction in people at clinical high risk for psychosis. *Mol Psychiatry* 2009; **16**: 67–75.
65 Schenk VW, Stolk PJ. Psychosis following arsenic (possibly thalium) poisoning. *Psychiatr Neurol Neurochir* 1967; **70**: 31–37.
66 Lebrun AH, Storch S, Pohl S, Streichert T, Mole SE, Ullrich K *et al*. Identification of potential biomarkers and modifiers of CLN3-disease progression. *Neuropediatrics* 2010; **41**: V1240.
67 Ratnaike RN. Acute and chronic arsenic toxicity. *Postgrad Med J* 2003; **79**: 391–396.
68 Mitchison HM, Taschner PEM, O'Rawe AM, De Vos N, Phillips HA, Thompson AD *et al*. Genetic mapping of the batten disease locus (CLN3) to the interval D16S288-D16S383 by analysis of haplotypes and allelic association. *Genomics* 1994; **22**: 465–468.
69 Olsen L, Hansen T, Jakobsen KD, Djurovic S, Melle I, Agartz I *et al*. The estrogen hypothesis of schizophrenia implicates glucose metabolism: association study in three independent samples. *BMC Med Genet* 2008; **9**: 39.
70 Le-Niculescu H, Balaraman Y, Patel S, Tan J, Sidhu K, Jerome RE *et al*. Towards understanding the schizophrenia code: an expanded convergent functional genomics approach. *Am J Med Genet B Neuropsychiatr Genet* 2007; **144B**: 129–158.
71 Duong HA, Robles MS, Knutti D, Weitz CJ. A molecular mechanism for circadian clock negative feedback. *Science* 2011; **332**: 1446–1449.
72 Monti JM, BaHammam AS, Pandi-Perumal SR, Bromundt V, Spence DW, Cardinali DP *et al*. Sleep and circadian rhythm dysregulation in schizophrenia. *Prog Neuropsychopharmacol Biol Psychiatry* 2013; **43**: 209–216.
73 van Venrooij JA, Fluitman SB, Lijmer JG, Kavelaars A, Heijnen CJ, Westenberg HG *et al*. Impaired neuroendocrine and immune response to acute stress in medication-naive patients with a first episode of psychosis. *Schizophr Bull* 2012; **38**: 272–279.
74 Mondelli V, Ciufonlini S, Murri MB, Bonaccorso S, Di Fortio M, Giordano A *et al*. Cortisol and inflammatory biomarkers predict poor treatment response in first episode psychosis. *Schizophr Bull* 2015; **41**: 1162–1170.
75 Le Moal M, Koob GF. Drug addiction: pathways to the disease and pathophysiological perspectives. *Eur Neuropsychopharmacol* 2007; **17**: 377–393.
76 Qiu MH, Liu W, Qu WM, Urade Y, Lu J, Huang ZL. The role of nucleus accumbens core/shell in sleep-wake regulation and their involvement in modafinil-induced arousal. *PLoS One* 2012; **7**: e45471.
77 Felger JC, Miller AH. Cytokine effects on the basal ganglia and dopamine function: the subcortical source of inflammatory malais. *Front Neuroendocrinol* 2012; **33**: 315–327.

Supplementary Information accompanies the paper on the Translational Psychiatry website (http://www.nature.com/tp)