READ ME File For ‘Appendices for Investigating Cellular Origins to Identify Peptide Vaccine Targets in two Independent Transmissible Tumours Circulating in the Tasmanian Devil (Sarcophilus harrisii)’

Dataset DOI: 10.5258/SOTON/D0982

ReadMe Author: RACHEL OWEN, University of Southampton

This dataset supports the thesis entitled:

Investigating Cellular Origins to Identify Peptide Vaccine Targets in two Independent Transmissible Tumours Circulating in the Tasmanian Devil

AWARDED BY: University of Southampton

DATE OF AWARD: 2019

This dataset contains:

**Appendix C.1:**

**Raw mass spectrometry data for triplicate non-quantified whole cell proteomes for DFT2\_RV, DFT1\_4906 (+IFNγ) and fibroblast (Salem) cell lines.**

Columns “ensembl\_peptide\_id”, “ensembl\_gene\_id” and “name” contain the information on peptide (and hence, protein) identity. DFT1 replicate data is presented in the sheets titled “DFT1 Replicate 1”, “DFT1 Replicate 2” and “DFT1 Replicate 3”, DFT2 replicate data is presented in the sheets titled “DFT2 Replicate 1”, “DFT2 Replicate 2” and “DFT2 Replicate 3”, fibroblast replicate data is presented in the sheets titled “Salem Replicate 1”, “Salem Replicate 2” and “Salem Replicate 3”. Data is presented as an Excel file with multiple sheets. Data generated using Peaks mass spectrometry software.

**Appendix C.2:**

**Raw mass spectrometry data for label-free quantification of proteins identified in the whole cell proteomes of DFT2\_RV, DFT1\_4906 and fibroblast (Salem) cell lines.**

All quantified proteins were detected in all three replicates of the given cell line of the proteome presented in Appendix 3A. Columns “x\_1\_intensity”, “x\_2\_intensity” and “x\_3\_intensity” represent the raw expression values for a given protein, and “Average” represents the average expression values used for functional analysis. Columns “ensembl\_peptide\_id”, “ensembl\_gene\_id” and “name” contain the information on peptide (and hence, protein) identity. The sheets titled “Protein\_replicates.csv” and “Unmodified\_proteins.csv” represent the data used to generate the heatmaps in Figure 3.5 and Appendix B.1. Triplicate data is presented for DFT1, DFT2 and devil fibroblasts in the sheets titled “Quantified and annotated DFT1”, “Quantified and annotated DFT2” and “Quantified and annotated Salem” respectively. The sheet titled “Fibroblasts\_Karbiener2017” contains proteomic data on sheep oral mucosa fibroblasts previously published in Karbiener et al. 2017 which was used as a control dataset during functional analysis. Proteins were ranked based on an average of the values presented in the columns titled “LFQintensity.OMF.BR1”, “LFQintensity.OMF.BR2”, “LFQintensity.OMF.BR3” and “LFQintensity.OMF.BR4”. Sheets titled “Unmodified\_proteins.csv” and “Protein\_replicates.csv” represent the csv files used to generate the heatmaps in Figure 3.5 as described in Chapter 2.3.4 and Appendix B.1. Data is presented as an Excel file with multiple sheets. Data generated using Peaks mass spectrometry software.

**Appendix C.3:**

**Full results for functional analysis of the full, quantified proteomes of each cell line against the Gene Ontology: Biological process, Gene Ontology: Cellular compartment, Reactome and Human Protein Atlas databases.**

Proteins were ranked by expression from highest to lowest using the average values presented in Appendix C.2. The “source” column indicates the database each enriched functional term is derived from, where GO:BP = Gene Ontology: Biological process, GO:CC = Gene Ontology: Cellular compartment, REAC = Reactome and HPA = Human protein atlas. “term\_name” and “term\_id” are the identifiers for each enriched process. “adjusted\_p\_value” is the significance of enrichment of each term, adjusted for multiple testing using the g:SCS algorithm (Chapter 2.3.4). “term\_size” is the number of proteins in the database associated with a given process. “query\_size” is the number of recognised proteins submitted for analysis. “intersection\_size” is the number of proteins in the query which are associated with a given process, and these proteins are listed in “intersections”. Functional analysis for DFT1, DFT2, devil fibroblasts and sheep fibroblasts (Karbiener et al. 2017) is presented in the sheets titled “DFT1 full functional analysis”, “DFT2 full functional analysis”, “Salem full functional analysis” and “SheepFibs\_Karbiener2017” respectively. The sheets titled “Biological\_process\_full.csv”, “Cell\_compartment\_full.csv”, “Reactome\_full.csv” and “Human\_protein\_ontology\_full.csv” correspond to the data used to generate the heatmaps in Figure 3.6 as described in Chapter 2.3.4 and Appendix B.1. Data is presented as an Excel file with multiple sheets. Data generated using the online ontology software gProfiler.

**Appendix C.4:**

**Cross cell line comparisons of the functional analysis presented in Appendix C.3.**

Adjusted p-values from Appendix C.3 are used to compare functional analysis between cell lines. NA indicates a functional term was not identified as enriched in that cell line. GO:BP and GO:CC analysis is further annotated to identify terms which give cell-type specific information in order to identify DFT2s cell type of origin, as discussed in Chapter 3.3.2. Yellow cell highlight indicates functional terms identified as providing cell-type specific information in the context of this project. These terms have been further categorised according to the key. Sheet titles indicate the functional database used in analysis. Data is presented as an Excel file with multiple sheets. Data generated in Excel.

**Appendix C.5:**

**Full results for functional analysis of proteins unique to and quantified in a single cell line against the Gene Ontology: Biological process, Gene Ontology: Cellular compartment, Reactome and Human Protein Atlas databases.**

Proteins were ranked by expression from highest to lowest using the average values presented in Appendix C.2. The “source” column indicates the database each enriched functional term is derived from, where GO:BP = Gene Ontology: Biological process, GO:CC = Gene Ontology: Cellular compartment, REAC = Reactome and HPA = Human protein atlas. “term\_name” and “term\_id” are the identifiers for each enriched process. “adjusted\_p\_value” is the significance of enrichment of each term, adjusted for multiple testing using the g:SCS algorithm (Chapter 2.3.4). “term\_size” is the number of proteins in the database associated with a given process. “query\_size” is the number of recognised proteins submitted for analysis. “intersection\_size” is the number of proteins in the query which are associated with a given process, and these proteins are listed in “intersections”. Functional analysis of proteins unique to DFT1, DFT2 and fibroblasts are presented in sheets titled “DFT1 Unique Funct. Analysis”, “DFT2 Unique Funct. Analysis” and “Salem Unique Funct. Analysis” respectively. Data is presented as an Excel file with multiple sheets. Data generated using the online ontology software gProfiler.

**Appendix C.6:**

**Cross cell line comparisons of the functional analysis presented in Appendix C.5.**

Adjusted p-values from Appendix C.5 are used to compare functional analysis between cell lines. NA indicates a functional term was not identified as enriched in that cell line. GO:BP, GO:CC and Reactome analysis is further annotated to identify terms which give cell-type specific information in order to identify DFT2s cell type of origin, as discussed in Chapter 3.1.2. Yellow cell highlight indicates functional terms identified as providing cell-type specific information in the context of this project. These terms have been further categorised according to the key. Sheet titles indicate the functional database used in analysis. Data presented as an Excel file with multiple sheets. Data generated in Excel.

**Appendix C.7:**

**Functional analysis of tumour specific and ubiquitous proteins.**

1. **Proteins unique to DFT2, plus proteins quantified in DFT1 and DFT2 but not in fibroblasts, presented in sheet labelled “DFT2 not Salem Funct. Analysis”**
2. **Proteins unique to DFT1, plus proteins quantified in DFT1 and DFT2 but not in fibroblasts, presented in sheet labelled “DFT1 not Salem Funct. Analysis”**
3. **Proteins unique to fibroblasts, plus proteins quantified in all three cell lines, presented in sheet titled “Salem Unique All Cells Funct.”**

**against the Gene Ontology: Biological process, Gene Ontology: Cellular compartment, Reactome and Human Protein Atlas databases.**

Proteins were ranked by expression from highest to lowest using the average values presented in Appendix C.2. The “source” column indicates the database each enriched functional term is derived from, where GO:BP = Gene Ontology: Biological process, GO:CC = Gene Ontology: Cellular compartment, REAC = Reactome and HPA = Human protein atlas. “term\_name” and “term\_id” are the identifiers for each enriched process. “adjusted\_p\_value” is the significance of enrichment of each term, adjusted for multiple testing using the g:SCS algorithm (Chapter 2.3.4). “term\_size” is the number of proteins in the database associated with a given process. “query\_size” is the number of recognised proteins submitted for analysis. “intersection\_size” is the number of proteins in the query which are associated with a given process, and these proteins are listed in “intersections”.

**Appendix G.3:**

**Raw mass spectrometry data and sequenced peptide information following immunoaffinity purification of the MHC class I complex.**

Peptide sequence is given in the column labelled “peptide\_cln”, peptide length is given in the column “length”. “exp” refers to the replicate each peptide was detected in, MYC\_1, MYC\_2 and MYC\_3 are the three replicates of myc immunoaffinity purification, B2M\_1, B2M\_2 and B2M\_3 are the corresponding β2-m immunoaffinity replicates. Peptide and gene Ensembl identifiers for the source protein of detected peptides are presented in columns “accession”, “ensembl\_peptide\_id” and “ensembl\_gene\_id” where available. Official gene symbol and protein description are presented in “name” and “description” where available. RefSeq protein ID is given in “IDs” where available. Genome scaffold, chromosome and start and end of protein sequences in the genome of the source protein are presented in the columns “seqnames”, “chr”, “start” and “end” where available. Data presented as an Excel file. Data generated using Peaks mass spectrometry software.

**Appendix G.4:**

**Raw mass spectrometry data and sequenced peptide information following immunoaffinity purification of the MHC class I complex performed by Annalisa Gastaldello et al. (in preparation).**

Peptide sequence is given in the column labelled “peptide\_cln”, peptide length is given in the column “length”. “exp” refers to the replicate each peptide was detected in, B2M\_4, B2M\_5 and B2M\_7 are the corresponding β2-m immunoaffinity replicates. B2M\_6 was a failed experiment and is not included in this analysis. Peptide and gene Ensembl identifiers for the source protein of detected peptides are presented in columns “accession”, “ensembl\_peptide\_id” and “ensembl\_gene\_id” where available. Official gene symbol and protein description are presented in “name” and “description” where available. RefSeq protein ID is given in “IDs” where available. Genome scaffold, chromosome and start and end of protein sequences in the genome of the source protein are presented in the columns “seqnames”, “chr”, “start” and “end” where available. Data presented as an Excel file. Data generated using Peaks mass spectrometry software.

Licence: CC BY-NC-ND

Date of data collection: October 2015 – March 2019

Information about geographic location of data collection: University of Southampton, U.K.

Related projects/Funders: Leverhulme Trust (RPG-2015-103)

Date that the file was created: June, 2019