University of Southampton

FACULTY OF MEDICINE

HUMAN DEVELOPMENT & HEALTH

Identification of genetic factors associated with myeloid neoplasms

DOI: 10.5258/SOTON/T0048

Gabriella Galatà

Supervisory Team: Prof Nicholas Cross & Dr William Tapper

ORCID ID 0000-0001-7990-7964

A thesis submitted for the degree of Doctor of Philosophy

June 2021

To Salvo and my parents

University of Southampton

Abstract

FACULTY OF MEDICINE

HUMAN DEVELOPMENT & HEALTH

Doctor of Philosophy

Identification of genetic factors associated with myeloid neoplasms

by

Gabriella Galatà

Myeloid neoplasms are clonal haematopoietic disorders characterised by the abnormal proliferation of specific myeloid cell types. The first part of this thesis focuses on mastocytosis, a rare haematological neoplasm characterised by the uncontrolled proliferation of mast cells. To test the hypothesis that germline variants can alter the risk of developing mastocytosis, a two-stage case-control genome-wide association study was conducted in five European populations with 1,035 *KIT*^{D816V}-positive cases and 17,960 controls. This analysis identified three genome-wide significant SNPs: rs4616402 (P_{meta}=1.37×10^{-15}, I^2=4.2), rs4662380 (P_{meta}=2.11×10^{-12}, I^2=0) and rs13077541 (P_{meta}=2.10×10^{-9}, I^2=0). Expression and methylation quantitative trait loci analysis were used to identify candidate genes located near the SNPs, specifically *CEBPA*, *TEX41* and *TBL1XR1*. Statistical analysis with available clinical data, showed that rs4616402 was associated with age at presentation (P = 0.009; beta = 4.41; n = 422) in patients with non-advanced disease. Additional focused analysis identified suggestive associations between mastocytosis and genetic variation at *TERT*, *TPSAB1/TPSB2*, and *IL13*. Finally, a gene-based analysis was performed using the summary statistics of the stage 1 meta-analysis and multiple regression which suggested that the *VEGFC* gene is also associated with mastocytosis. The findings described in this thesis demonstrate that multiple inherited common risk variants predispose to *KIT*^{D816V} positive mastocytosis and provide novel avenues for functional investigation.

In the second part of this thesis, the genetics of somatically acquired uniparental disomy (aUPD) in myeloid malignancies was investigated. Several regions of recurrent aUPD have been identified in patients affected with haematological neoplasms, many of which harbour somatic mutations that drive clonal proliferation. Similar regions of aUPD have also been identified in apparently healthy individuals, especially the elderly, which confer a tenfold increased risk of developing haematological malignancies. Large-scale sequencing initiatives of individuals unselected for cancer therefore represent a valuable resource to identify novel regions of aUPD and the

underlying somatic mutations which drive clonal haematopoiesis (CH). Whole-exome sequence (WES) data for 49,996 individuals from the UK biobank (mean age = 56.5 years) was used to develop an automated pipeline for identifying aUPD regions and a new scoring system (gg score) to select aUPD regions with high confidence for manual review. Precision and recall were used to evaluate the gg score. The recall (or sensitivity) showed that it correctly identifies 55% of the predicted aUPD regions, although the model can also produce false negatives. On the other hand, the score performed well in term of precision and indicated that 90% of the aUPD regions were correctly classified. The methodology was then applied to WES data from a Swedish Case-Control study of Schizophrenia consisting of 12,380 samples and with a mean age of 65. Genes targeting the aUPD regions identified in the Swedish cohort are known (*MPL*, 1p; *TET2*, 4q; *EZH2*, 7q; *JAK2*, 9p; *FLT3*, 13q; *MEG3-DLK1*, 14q). Regions of aUPD were screened for somatic mutation if they were overlapping in two or more samples. However, only *JAK2*^{V617F} was confirmed in all five samples with UPD9p and new aUPD regions with unknown gene target were not identified. This work showed that the frequency of sample with aUPD regions identified by WES data is lower (0.2-0.3%) than expected (1-2%) and provides an estimate what is needed in term of sample size to detect aUPD regions from WES data.

Table of Contents

List of Figures

List of Tables

List of Tables

List of Supplementary Tables

Research Thesis: Declaration of Authorship

I, Gabriella Galatà 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.

Title of thesis: Identification of genetic factors associated with myeloid neoplasms

I confirm that:

This work was done wholly or mainly while in candidature for a research degree at this University;

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;

Where I have consulted the published work of others, this is always clearly attributed;

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;

I have acknowledged all main sources of help;

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;

Parts of this work have been published as:

Galatà G, García-Montero AC, Kristensen T, Dawoud AAZ, Muñoz-González JI, Meggendorfer M, Guglielmelli P, Hoade Y, Alvarez-Twose I, Gieger C, Strauch K, Ferrucci L, Tanaka T, Bandinelli S, Schnurr TM, Haferlach T, Broesby-Olsen S, Vestergaard H, Møller MB, Bindslev-Jensen C, Vannucchi AM, Orfao A, Radia D, Reiter A, Chase AJ, Cross NCP, Tapper WJ. Genome-wide association study identifies novel susceptibility loci for KIT D816V positive mastocytosis. Am J Hum Genet. 2021 Feb 4;108(2):284-294. doi: 10.1016/j.ajhg.2020.12.007. Epub 2021 Jan 8. PMID: 33421400; PMCID: PMC7895845.

Signature: Date:30/06/2021

Acknowledgements

I would like to acknowledge the fantastic people that have helped me to go through the fulfilling adventure of my PhD.

Firstly, I would like to thank my supervisors Professor Nicholas Cross and Dr William Tapper for accepting me on their team and for supervising this research project. Without their essential help and contribution this work would not have been possible.

A great thank you goes to Nick for the academic and moral support, for encouragement and for inspiring me with his knowledge and passion. Nick, I feel so lucky to have been involved in this exciting research and for the opportunity you have given me. Your support has been essential for me to complete this work.

A special thanks to Will for also supporting me. You guided me through the basics of bioinformatics and data analysis, from command line tools for using the Iridis computer cluster to performing a GWAS. You have encouraged me to improve my programming skills and you pushed me out of my comfort zone. I will always be grateful for everything I have learnt from you.

The GWAS study has been a great example of a collaborative success and I thank everyone who contributed to this work, who generated and provided the data, and all the participants that took part in this study.

Many thanks to Professor Andrew Collins and Dr Faisal Rezwan for the fruitful discussions and advice during my progression review.

I would like to express my gratitude to the Iridis Team at the University of Southampton for the support to run the jobs smoothly on the Iridis cluster; in particular Dr Elena Vagata who helped me to set up the job array in the bioinformatic pipeline.

Thank you also to the colleagues I have met in the Genomic Informatics lab, especially Alejandra, Dareen, Enrico and Reza for creating an inspiring environment. I have been very lucky to be surrounded by a bunch of fantastic people who added laughter and good times to this journey.

A big thank you to the Student & Postdoc Interactive Network (SPiN) for creating a space within HDH where students and postdocs could access a sense of community and support. Being part of the SPiN team has had a positive and invaluable impact on me, especially during the difficult time of isolation. It has been a great honour working together in the last year and half. Together we made Anna's idea happen and I am very proud of each one of you.

Thanks to Chiara and my brother Salvo for the encouraging words and advice, I always feel in a safe place with them. A special thanks to Salvo for always supporting me, for the long chats about science and life; you are such an inspiration to me. Thank you also to little Beatrice who has filled much time of this journey with smiles and happiness. Last but not least, I want to thank my parents, for always being there for me, for believing in me and for their unconditional love and support.

Ethics approval

This study was approved by the University Ethics Committee through the ERGO system (Submission ID: 45559, Submission Title: The molecular pathogenesis of atypical chronic myeloproliferative neoplasms and related disorders, Submitter Name: Nicholas Cross).

Funders

This research is funded by University of Southampton VC Award and Blood Cancer UK Programme Grant Research Studentship.

Definitions and Abbreviations

aCML ..... Atypical chronic myeloid leukaemia

AD ..... Allelic Depth

AI ..... Allelic Imbalance

ALL ..... Acute Lymphoblastic Leukaemia

AML ..... Acute Myeloid Leukaemia

ARCH ..... Age-Related Clonal Haematopoiesis

AUC ..... Area Under the Curve

aUPD ..... Acquired Uniparental Disomy

BAF ..... B Allele Frequency

BAM ..... Binary Alignment/Map

BCR-ABL1 ..... Breakpoint Cluster Region-ABL proto-oncogene 1

BM ..... Bone Marrow

Bp ..... basepairs

BST ..... Baseline Serum Tryptase

BWA ..... Burrows-Wheeler Aligner

CALR ..... Calreticulin

CBS ..... Circular binary segmentation

CEBPA ..... CCAAT Enhancer Binding Protein Alpha

CH ..... Clonal haematopoiesis

CSC ..... Cancer Stem Cell

CEL ..... Chronic Eosinophilic Leukaemia

CEPH ..... Centre d'Etude du Polymorphisme Humain

ceRNA ..... Competing endogenous RNA

CEU ..... CEPH Utah residents with ancestry from Northern and Western Europe

CHB ..... Han Chinese in Beijing

CHIP ..... Clonal Haematopoiesis of Indeterminate Potential

CI ..... Confidence interval

CM ..... Cutaneous Mastocytosis

CML ..... Chronic Myeloid Leukaemia

CMML ..... Chronic Myelomonocytic Leukaemia

CNL ..... Chronic Neutrophilic Leukaemia

CNN ..... Copy Number Neutral

CNV ..... Copy Number Variant

CV ..... Cross-Validation

C6orf10 ..... Chromosome 6 Open Reading Frame

dbGaP ..... Database of Genotype and Phenotypes

dbSNP ..... SNP database DNA

DNA ..... Deoxyribonucleic Acid

DNBC ..... Danish National Birth Cohort

DP ..... Depth

ECD ..... Extracellular Domain

eQTL ..... Expression quantitative trait loci

FDP ..... Familial platelet disorder

GATK ..... Genome Analysis Toolkit

GH ..... Genotype Harmoniser

GPC ..... Genetic Power Calculator

GT ..... Genotype

GWAS ..... Genome-wide Association Study

Hbf ..... Fetal haemoglobin

HSC ..... Haematopoietic Stem Cell

HWE ..... Hardy-Weinberg equilibrium

IBD ..... Identity by descent

IBS ..... Identity By State

ICD-O-3 ..... International Classification of Diseases for Oncology, Third Edition

InCHIANTI ..... Invecchiare in Chianti

ISM ..... Indolent systemic mastocytosis

JAK2 ..... Janus kinase 2

JMD ..... Juxtamembrane domain

JMML ..... Juvenile myelomonocytic leukaemia

KID ..... Kinase Insert Domain

KIT ..... KIT Proto-Oncogene Receptor Tyrosine Kinase

LD ..... Linkage Disequilibrium

LINC01412 ..... Long Intergenic Non-Protein Coding RNA 1412

lncRNA ..... Long non-coding RNA

LOH ..... Loss of Heterozygosity

LR ..... Logistic Regression

LRR ..... Log R Ratio

MAF ..... Minor Allele Frequency

mBAF ..... mirrored B Allele Frequency

MC ..... Mast Cell

MCAD ..... Mast Cell Activation Disease

MCL ..... Mast Cell Leukaemia

MCP ..... Mast cell progenitor

MDS ..... Myelodysplastic syndrome

MPL ..... Myeloproliferative Leukaemia Proto-Oncogene

MPN ..... Myeloproliferative Neoplasm

MPN-U ..... MPN Unclassifiable

NCBI ..... National Centre for Biotechnology Information

SNDNAB ..... Spanish National DNA Bank (Spanish controls)

NGS ..... Next-generation sequencing

NOTCH4 ..... Notch receptor 4

PB ..... Peripheral Blood

PCA ..... Principal Component Analysis

PMF ..... Primary Myelofibrosis

PRS ..... Polygenic risk score

PV ..... Polycythaemia vera

QC ..... Quality Control

QQ ..... Quantile-Quantile

Rb ..... Retinoblastoma

RET ..... Receptor Tyrosine Kinase

RNA ..... Ribonucleic Acid

ROC ..... Receiver Operating Characteristics

rs ..... Reference SNP

SCF ..... Stem Cell Factor

SD ..... Standard deviation

SFS ..... Sequential Feature Selection

SM ..... Systemic Mastocytosis

SLC7A10 ..... Solute Carrier Family 7 Member 10

SNP ..... Single Nucleotide Polymorphism

T2D ..... Type 2 diabetes

TBL1XR1 ..... Transducin Beta Like 1 X-Linked Receptor 1

TF ..... Transcription factor

TKD ..... Tyrosine Kinase Domain

TMD ..... Transmembrane Domain

TPSAB1 ..... Tryptase Alpha/Beta 1

TPSD1 ..... Tryptase Delta 1

TSBP ..... Testis expressed basic protein 1

UPD ..... Uniparental Disomy

VAF ..... Variant Allele Frequency

VCF ..... Variant Call Format

VEGFC ..... Vascular endothelial-derived growth factor C

VQSR..... Variant Quality Score Recalibration

WES ..... Whole Exome Sequencing

WGS..... Whole Genome Sequencing

WHO ..... World Health Organization

WTCCC2..... Wellcome Trust Case Control Consortium 2

YRI ..... Yoruban in Ibadan, Nigeria

ZEB2-AS1 ..... *ZEB2* Antisense RNA 1

Chapter 1 Introduction

1.1 Cancer overview

Cancers are a group of diseases that are characterised by uncontrolled cell division and decreased cellular death, both of which are influenced by genetic and epigenetic control (Strachan and Read, 2011). Mechanisms have evolved, such as apoptosis and deoxyribonucleic acid (DNA) repair, in part to protect the human body from malignancy, and these processes can be impaired by both germline and somatic mutations (Strachan and Read, 2011; Stratton et al., 2009). Germline mutations occur in sex cells and can therefore be passed onto offspring where they will be present in every cell. Somatic mutations occur in non-germ tissue and are not inherited. They are clonal in nature, so a clone of cells can be defined by a founding mutation and separated into subclones by subsequent mutations. Somatic mutations can be further categorised into drivers and passengers (Stratton et al., 2009). Driver mutations confer a growth advantage so they are positively selected and give rise to the hallmarks of cancer such as cell proliferation, immortalisation, metastasis, angiogenesis and evasion of growth suppressors (Hanahan and Weinberg, 2011). Passengers on the other hand are selectively neutral and not required for the initiation or maintenance of carcinogenesis. Most likely they simply happened to be present in a cell that acquired a driver mutation. Distinguishing between driver and passenger mutations has become one of the central goals of cancer genomics, although this is complicated by the observation that some tumours can contain up to 100,000 passenger mutations and fewer than 20 driver mutations. However, haematological malignancies are much simpler, and fewer driving mutations are required to generate a tumour (Stratton et al., 2009). In general, the mutational rate across cancers is highly heterogeneous: a study of 7,664 tumours across 29 cancer types showed that 1 to 10 driver mutations are needed to convert a normal cell into a cancer cell (Martincorena et al., 2017).

1.1.1 Oncogenes, tumour suppressor genes and the two-hit hypothesis

Mutations target specific genes, traditionally known as oncogenes and tumour suppressor genes, resulting in the conversion of a normal cell into a malignant tumour (Strachan and Read, 2011). Proto-oncogenes are present in normal cells and generally encode for proteins promoting cell proliferation, arresting cell death or inhibiting cell differentiation. Proto-oncogenes are usually activated in somatic cells by dominant genetic changes such as point mutations, gene amplifications and translocations (Figure 1.1) (Chial et al., 2008). Point mutations can be found within a promoter or a gene. The human telomerase reverse transcriptase (*TERT*) gene, for

example, has been implicated in a wide range of cancers, and single nucleotide substitution in the promoter of this gene can enhance mRNA expression (Horn et al., 2013; Huang et al., 2013). Point mutations within genes can instead produce normal protein with constitutive activity (e.g., *KIT* and *JAK2*) or degrade protein function (Gnanasambandan et al., 2010; Laine et al., 2011). Gene amplifications and overexpression of the amplified gene can lead to malignant transformation both in solid cancers (e.g., *HER2* mainly in breast cancer) and haematologic malignancies (e.g., *MYC* in lymphoid leukaemia) (L'Abbate et al., 2018; Neve et al., 2001; Zakrzewski et al., 2019). Proto-oncogenes activated by chromosomal translocations have been associated with gene hyperactivation as a consequence of new super-enhancers (e.g., *MYC* in multiple myeloma) or fusion genes (e.g., *BCR-ABL1* in chronic myeloid leukaemia) (Hnisz et al., 2014; Lancho and Herranz, 2018; Peiris et al., 2019). In particular, chromosomal translocations involved in haematological cancer will be discussed in Section 1.2 of this thesis. Other mechanisms such as hypomethylation of long interspersed nuclear element-1 (LINE-1) have been associated with the activation of proto-oncogenes in various human cancers (Bae et al., 2012; Hur et al., 2014; Roman-Gomez et al., 2005). Activated proto-oncogenes, called oncogenes, promote cell proliferation and differentiation (Strachan and Read, 2011).

Tumour suppressor genes encode for proteins involved in several mechanisms such as the inhibition of cell proliferation, apoptosis, replication and DNA repair. According to the two- hit hypothesis proposed by Knudson in retinoblastoma (Rb), carcinogenesis in some cases can initiate when the cell has mutations in both alleles of a tumour suppressor gene; i.e., they are recessive (Knudson, 2001). If a tumour suppressor is inactivated, mechanisms that control the normal cell cycle will be lost (Strachan and Read, 2011). Familial Rb (accounting for 25–35% of Rb cases) is an autosomal dominant disease where one mutated allele is inherited (Jagadeesan et al., 2016). For most tumour suppressor genes, however, inactivation of both alleles corresponds to somatic events. Inactivation of tumour suppressor genes is often caused by whole-gene deletion of one allele, mitotic recombination or duplication of the mutant allele, which may be detected by loss of heterozygosity (LOH) of informative markers upon a comparison of tumour and normal tissue.

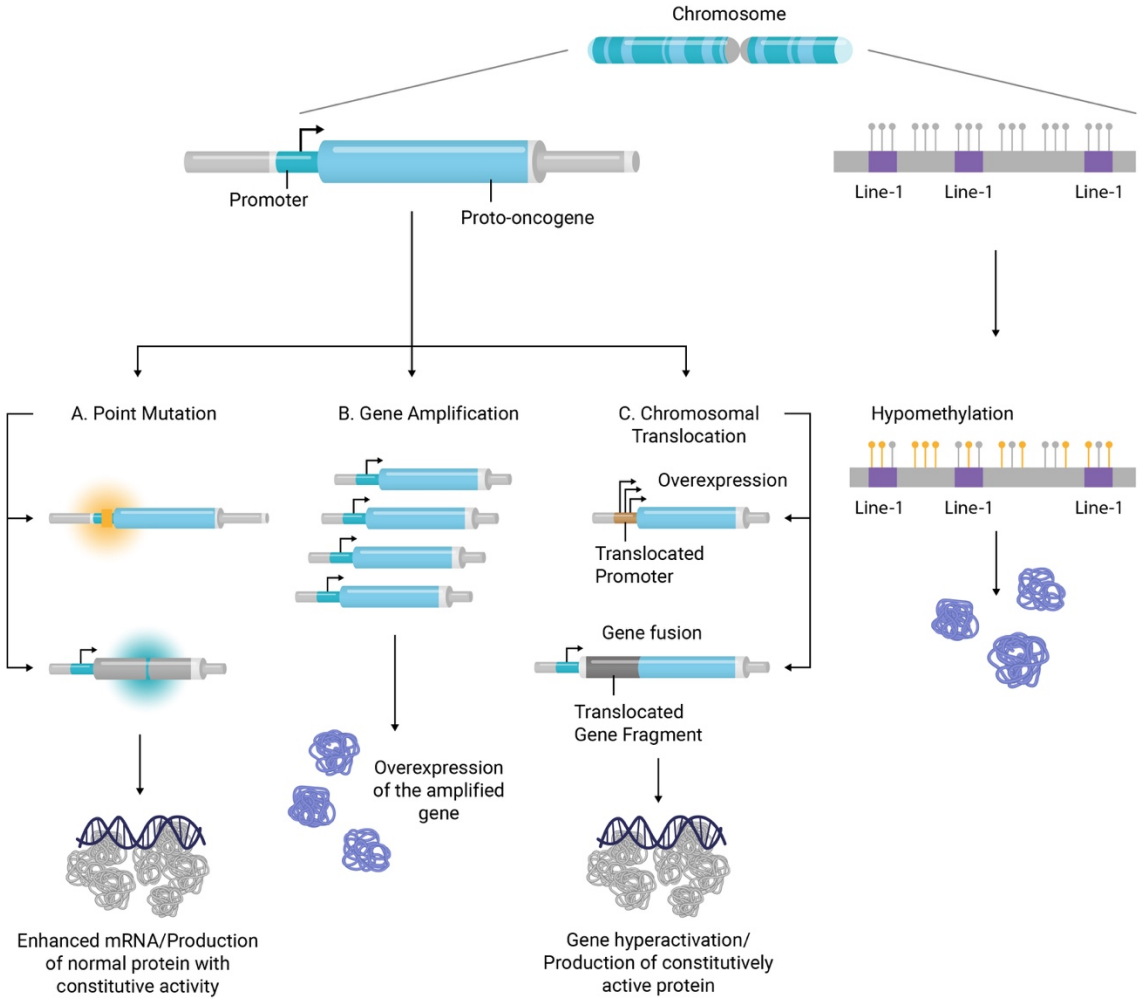


Figure 1.1 **Proto-oncogene activation mechanisms.**

Proto-oncogenes are genes involved in the regulation of the cell cycle. Genetic changes such as point mutations, gene amplification, chromosomal translocation and hypomethylation can activate proto-oncogenes to become oncogenes.

As more mutated genes have been discovered in cancer it has become apparent that the model of dominant oncogenes and recessive tumour repressor genes is rather simplistic with many genes in fact having both dominant and recessive characteristics at the cellular level (Soussi and Wiman, 2015).

1.1.2 Clonal evolution in cancer

Cancers evolve by clonal evolution, a concept formulated in the 1970s by Nowell (Nowell, 1976). He proposed that most neoplasms are the result of an evolutionary process initiated by a single, previously normal cell. An initial event gives rise to a proliferative advantage and clonal outgrowth. Acquisition of additional mutations, possibly in the context of genomic instability,

gives rise to further subclones. This evolutionary process will lead to a selection of more aggressive subclones which, due to their growth advantage over the normal cell, will begin to predominate and, for solid tumours, metastasise and invade local tissue. The rate of acquired epigenetic changes has been estimated to have a crucial role in genetic changes in clonal evolution (Siegmund et al., 2009). Overall, both genetic and epigenetic changes and subclonal selection processes result in advanced human malignancies characterised by uncontrolled proliferation (Flavahan et al., 2017; Nowell, 1976). More recently, it has become apparent that cancer evolution is not a simple linear process, but involves branched evolution and complex interactions between subclones (Greaves and Maley, 2012). Data on acute lymphoblastic leukaemia (ALL) in childhood revealed more dynamic clonal expansions, which occur without a preferential order. In fact, dominance and the architecture of subclones change constantly before subclones begin to dominate in early cancer development (Anderson et al., 2011).

1.1.3 Heterogeneity and hierarchical organisation in cancer

Cancer is characterised by genetic heterogeneity whereby different tumour cells have unique mutation profiles that form a hierarchical organisation (Caldas, 2012). Mouse model experiments gave, for the first time, evidence of heterogeneous subpopulations in a single tumour line. The analysis of isolated sub-clones showed their different metastatic potential consistent with the heterogeneity of cancer (Harris et al., 1982). This heterogeneity and hierarchy of tumour subpopulations was also demonstrated in human acute myeloid leukaemia (AML) using non-obese diabetic mice with severe combined immunodeficiency disease (NOD/SCID) transplanted with human leukaemic cells (Bonnet and Dick, 1997). AML is composed of multiple distinct cell types and maintained by slow-cycling leukaemic stem cells (Clevers, 2011). Dick and Bonnet demonstrated *in vivo* the hierarchical organisation of a leukaemic clone by comparing the organisation of the normal and the AML haematopoietic system in humans. In fact, theywere able to detect a primitive leukaemic stem cell that produces clonogenic leukaemic progenitors (AML-CFU) and leukaemic blasts (Bonnet and Dick, 1997).

1.1.4 Insight into clonal evolution

Genetic diversity and epigenetic plasticity in cancer can lead to clonal evolution, drug resistant subclones, therapeutic failure and tumour relapse (Greaves, 2015). Advanced technologies such as single nucleotide polymorphism (SNP) microarrays and next-generation sequencing (NGS) have been widely used to investigate clonal evolution and genetic heterogeneity, and have improved our knowledge of the genotypic and phenotypic evolution of tumour cells (Ding et al., 2012; Landau et al., 2014; McGranahan and Swanton, 2017). For example, SNP arrays were used for genome-wide analysis of copy number variant (CNV) and LOH analysis on diagnostic and relapse

bone marrow (BM) samples of 61 patients affected with ALL (Mullighan et al., 2008). The results of the study showed no difference in CNVs betweenrelapse and diagnostic samples on 8% of the patients, while 34% of the relapse samples showed clonal evolution of the diagnostic clone. This study demonstrated how a common ancestral clone can give rise to major and minor clones that are both present at diagnosis, and how the minor clone can acquire new genetic alterations and generate a new clone that is positively selected and responsible for relapse. In an illustrative case, two deletions at relapse were reported, one of which was present in a minor clone of the diagnostic sample, whereas the second one was acquired during a different stage of the evolution of the relapse clone (Mullighan et al., 2008). In patients with AML, clonal evolution can also be a cause of death after tumour relapse. In a study performed with 8 patients from different French-American-British subtypes of AML, the primary tumour, relapse, and matched normal skin samples were sequenced using NGS, and new clonal mutational patterns in tumour relapse were identified (Ding et al., 2012). In order to investigate the cytotoxic effect of the chemotherapy on the evolution of the tumour, transversion (substitution of a purine for a pyrimidine or vice versa) and transition (changes from purine/pyrimidine to another purine/pyrimidine) in the relapse-specific tumour were compared with the changes identified in the primary tumour. The comparison revealed an increase in transversions for relapse-specific mutations. Although the primary tumour sub-clones were eradicated by therapy and therefore absent at relapse in 50% of the cases, this study showed that the cytotoxicity of the therapeutic treatment alters the clonal structure of the tumour and allows a more aggressive clone to dominate and contribute to drug resistance at relapse (Ding et al., 2012). Overall, these results show that different classes of mutations can be responsible for clonal evolution and need to be investigated in further studies. Furthermore, targeting with future therapies needs to consider not only the primary clone but also its subclones, and this should be one of the main foci of cancer research in order to minimise the impact of relapse after treatment (Mullighan et al., 2008).

1.1.5 Cancer stem cell model

According to the cancer stem cell (CSC) concept, the growth of tumours is driven by a group of slow-cycling CSC with pluripotency, self-renewal and chemo-resistance capabilities. The CSC model presumes that the tumour is composed of two groups of cell; differentiated cells that have lost their proliferative capability, and CSCs, which represent the tumourigenic part of the tumour. Therefore, this feature contributes to relapse and supports the hypothesis of using CSCs as the target for new strategies in cancer therapy (Clevers, 2011). Moreover, recent identification of several markers and an understanding of signalling pathways associated with CSC proliferation, apoptosis and differentiation have given insight into the development of drugs that used in

combination with traditional treatment are under evaluation in preclinical and clinical studies (Dragu et al., 2015).

1.1.6 Cancer classification

Based on the International Classification of Diseases for Oncology, Third Edition (ICD-O-3) (Fritz et al., 2013), cancers can be named according to the type of tissue where they originate. The NIH National Cancer Institute (https://training.seer.cancer.gov/disease/categories/classification.html) lists approximately 200 types of cancers, which can be grouped into six main categories based on histological type (NIH National Cancer Institute):

- carcinoma, cancer that originates in epithelial tissue;
- sarcoma, malignancies of connective tissue (bone, cartilage, smooth muscle, skeletal muscle, blood vessels, adipose tissue, etc.);
- myeloma, a type of cancer that affects plasma cells. Plasma cells are leucocytes involved in immunoglobulin secretion and originated from B-cell differentiation(Oracki et al., 2010);
- leukaemia, liquid cancers usually affecting leucocytes. Red blood cells can also be affected;
- lymphoma, solid cancers that originate in the lymphatic system. The main lymphomas are Hodgkin's disease and non-Hodgkin's lymphoma;
- mixed types, containing different cell types.

Haematological malignancies (leukaemia, lymphoma, myeloma) can be defined as myeloid or lymphoid depending on which cell lineage in haematopoiesis is affected, and acute or chronic depending on the tempo of onset and degree of differentiation. The classification of lymphoid and myeloid neoplasms was summarised in the fourth edition of the World Health Organization (WHO) classification of tumours of haematopoietic and lymphoid tissues (Swerdlow et al., 2008). In 2016, new clinical, prognostic, diagnostic and genetic findings derived from gene expression and sequencing studies led to a further revision of the WHO classification. For instance, systemic mastocytosis (SM) was reclassified and is no longer considered a subgroup of myeloproliferative neoplasms (Arber et al., 2016a; Swerdlow et al., 2016). Details of the new classification for SM are described on paragraph 1.2.5.1.

1.2 Myeloid Neoplasms

Myeloid neoplasms are clonal haematopoietic disorders that are characterised by constitutive activation of signal-transduction pathways and other changes which lead to transformation and abnormal proliferation of haematopoietic stem cells (HSC), overproduction of one or more cell types in the myeloid lineage in the BM, and an increase in specific myeloid cells in the peripheral blood (Korn and Méndez-Ferrer, 2017). In the most recent WHO classification, these malignancies are categorised into major subtypes which include myeloproliferative neoplasm (MPN), myelodysplastic syndromes/myeloproliferative neoplasms (MDS/MPN), myelodysplastic syndromes (MDS) and AML (Arber et al., 2016b). Recently, progress has been made through the identification of new driver mutations that can be used for diagnosis and to estimate the prognosis of these disorders (Arber et al., 2016a; Patel et al., 2017). However, despite an updated classification and increased understanding of their molecular pathogenesis, these are heterogeneous disorders and some overlapping features remain. Myeloid malignancies are mainly sporadic; however, a small group of cases associated with germline mutations have been reported both in children and adults. Germline mutations associated with familial myeloid neoplasms will be discussed in the following paragraphs together with the description of the disease subtypes (Arber et al., 2016a; Baptista et al., 2017). A distinct group of myeloid neoplasms known as therapy-related myeloid neoplasms (t-MNs) can arise in patients that follow chemotherapy or radiotherapy for a primary tumour or an autoimmune disease (Arber et al., 2016b). Cytotoxic treatments are known to play an important role in the pathogenesis of these diseases (Hasan et al., 2008). However, data have shown that familial predisposition has also been found to be involved in the development of t-MNs (Churpek et al., 2016).

1.2.1 Myeloproliferative neoplasms

MPNs are clonal haematological diseases that are characterised by an excess production of several haematopoietic lineages (e.g., erythroid, megakaryocytic and granulocytic cells), BM fibrosis and symptoms related to peripheral blood (PB) cell abnormalities (Kim et al., 2015). According to the latest WHO classification, MPNs are grouped into seven main malignancies: chronic myeloid leukaemia (CML), chronic neutrophilic leukaemia (CNL), polycythaemia vera (PV), essential thrombocythaemia (ET), primary myelofibrosis (PMF), chronic eosinophilic leukaemia (CEL) and MPN unclassifiable (MPN-U) (Arber et al., 2016a; Skoda et al., 2015). Evidence in the literature demonstrates that genes encoding a protein with tyrosine kinase activity are mutated in many haematologic malignancies and most MPN (Klampfl et al., 2013; Tefferi and Vardiman, 2008). The defining molecular marker used for the diagnosis of CML is the fusion gene between the

breakpoint cluster region gene (*BCR*) and *ABL1* proto-oncogene 1 (*BCR-ABL1)* resulting from a translocation between chromosomes 9 and 22. The derivative chromosome 22, called the Philadelphia chromosome, is usually identified using molecular genetics techniques or by karyotype investigation. MPN cases without *BCR-ABL1* are known as *BCR-ABL1* negative MPN and their identification together with other factors have diagnostic and prognostic importance; *CSF3R*^{T618I} or other *CSF3R* activating mutations together with other diagnostic criteria are strongly associated with CML, and the presence of *JAK2* (Janus kinase 2)^{V617F} is usually associated with PV, ET or PMF. Occasional PV cases have *JAK2* exon 12 mutations, but the majority of *JAK2*^{V617F} negative ET and PMF cases are characterised by the presence of myeloproliferative leukaemia proto-oncogene (MPL) or calreticulin (*CALR*) mutations. The small proportion of ET and PMF cases that test negative for *JAK2*^{V617F}, *MPL* and *CALR* mutations are referred to as triple-negative MPN (Arber et al., 2016a; Kim et al., 2015). In recent studies, other disease-causing genes have been revealed to be mutated in MPN and, as shown in Table 1.1, different mutations can affect signalling, epigenetic abnormalities, splicing factors, DNA repair/tumour suppressor gene (Patel et al., 2017). Many of these genes are also mutated in MDS/MPN, MDS and AML.

1.2.2 Myelodysplastic syndromes

MDS is a myeloid malignancy and one of the most frequent haematopoietic disorders, especially in the elderly (Arber et al., 2016a). It is characterised by peripheral cytopenia, impaired haematopoiesis, dysplasia of haematopoietic cells and elevated risk of developing AML. Cytopenia is an essential diagnostic feature and according the WHO it is defined by specific thresholds of haemoglobin, platelet and neutrophil counts. According to the WHO classification, the degree of dysplasia and blast percentage also need to be considered in order to define specific MDS subtypes. The threshold of dysplastic cells is 10% in MDS; however, some individuals may have levels of dysplasia greater than 10%, so alternative causes of dysplasia need to be taken into account before a definite diagnosis can be made. Recurrent acquired mutations in *SF3B1, TET2, SRSF2, ASXL1, DNMT3A, RUNX1, U2AF1, TP53, EZH2* and many other genes have been identified in patients affected with MDS (Haferlach et al., 2014; Papaemmanuil et al., 2013). Some mutations can be useful for prognosis. For example, *TP53* mutation if present in patients with del(5q) is a predictive factor of poor response if the patients undergo specific treatment such as lenalidomide (Mallo et al., 2013).

Table 1.1 **Disease-causing genes/mutations in MPN.**

GENE	MUTATION	GENOMIC LOCATION	ESTIMATED FREQUENCY (%)		
			PV	ET	PMF
Signal Transduction					
JAK2	V617F	9p24	95-97	50-60	55-60
JAK2	Exon 12; missense, indels	9p24	2	Rare	Rare
MPL	Exon 10; missense	1p24.2	<1	3-5	5-10
CALR	Exon 9; indels	19p13.13	<1	20-25	25-30
SH2B3	Exon 2; missense, deletion	12q24	Rare	Rare	Rare
CBL	Exon 8-9; missense in codons 366-420	11q23	Rare	0-2	5-10
Epigenetic Modification					
TET2	All exons; indels, nonsense and missense	4q24	10-20	5	10-20
IDH1 *IDH2*	*IDH1*:missense *R132*; *IDH2*:missense R140Q;R172	2q33/ 15q26	~2	<1	3-5
DNMT3A	Exon 7-23; missense R882; nonsense, frameshift or splice site	2p23	5-10	1-5	5-12
ASXL1	Exon 13; frameshift or nonsense	20q11	2-7	5-10	15-35
EZH2	All exons; nonsense or frameshift	7q35-q36	~2	~2	5-10
Splicing Factors					
SF3B1	Exon 12-16; missense in codons 622-781	2q33	~1	~1	5-10
SRSF2	Exon 1; missense	17q25	Rare	Rare	Rare
U2AF1	Exon 2-7; missense	21q22	<1	<1	5-16
DNA Repair/Tumour Suppressor					
TP53	Exons 4-9; nonsense, frameshift, splice site, missense	17p13.1	<1	<1	2-4

Common somatic mutations in MPN, the genomic locations of the genes and estimated frequency in disease subtypes (PV, ET, PMF). The mutated genes are grouped by function. Genes regulating signalling, epigenetic modification, splicing and DNA repair mechanisms are most frequently affected (Patel et al., 2017).

1.2.3 Myelodysplastic/myeloproliferative neoplasms

MDS/MPN is a group of diseases with clinical, laboratory and morphological features of both MPN and MDS. The karyotype may present the same abnormalities as seen in MDS (Arber et al., 2016a). Targeted sequencing of genes often mutated in myeloid disorders identifies variants in 80% of patients affected with chronic myelomonocytic leukaemia (CMML), the most common MDS/MPN subtype. The most commonly affected genes are *SRSF2, TET2* and *ASXL1*, while mutations in *SETBP1, NRAS/KRAS, RUNX1, CBL* and *EZH2* are identified at a lower rate. All these genes are mutated in other MDS/MPN subtypes, with broadly different mutational patterns associated with four specific entities being highly relevant for their diagnosis (Meggendorfer et al., 2018). For example, atypical CML (aCML) is a rare MDS/MPN subtype that, similarly to CNL, is characterised by neutrophilia, but it is associated with *SETBP1* and *ETNK1* mutations. In most cases, *JAK2*, *CALR*, *MPL* are generally not present in this MDS/MPN subtype but *CSF3R* mutations are seen in 10% of aCML cases (Arber et al., 2016a; Wang et al., 2014a). Juvenile myelomonocytic leukaemia (JMML) is another MDS/MPN subtype initiated by RAS-activating mutations and characterised by overproduction of monocytes and granulocytes (Chang et al., 2014). JMML occurs in children, and almost 90% of the patients have somatic and sometimes germline changes in *PTPN11*, *KRAS*, *NRAS*, *CBL* and *NF1* (Arber et al., 2016a). MDS/MPN with ring sideroblasts and thrombocytosis (MDS/MPN-RS-T) and MDS/MPN-Unclassifiable (MDS/MPN-U) are other MDS/MPN subtypes under the 2016 WHO classification. MDS/MPN-RS-T in most cases (70%–90%) is strongly associated with mutations in the spliceosome gene *SF3B1* co-existing with an MPN driver mutation, such as *JAK2*^{V617F} (50%-65%), *CALR* or *MPL* mutations (<10%) (Arber et al., 2016a; Reinig and He, 2017). MDS/MPN-U is a very rare, heterogeneous neoplasm that comprises less than 5% of MDS/MPN. It can sometimes not be distinguished from aCML and not much is known about the disease (Chaudhury et al., 2015).

1.2.4 Myeloid neoplasms with germline predisposition

A small group of familial myeloid neoplasms associated with germline mutations have been reported and a common finding is that genes mutated in sporadic cases are also found to be mutated in familial cases. For instance, 10–15% of sporadic AML have normal karyotype and somatic mutations in the *CEBPA* gene, which is also mutated in familial AML, an autosomal dominant condition with nearly complete penetrance (Baptista et al., 2017). For example, sequence analysis of the germline DNA in three family members (two siblings and their father) affected with AML revealed c.212delC mutation (Smith et al., 2004). Mutational analysis in 3 families with a familial platelet disorder (FPD/AML) revealed heterozygous *RUNX1* missense

mutations which segregate with the disorder in all of the family members tested (Michaud et al., 2002). Subsequently many other germline mutations predisposing to MDS/AML have also been reported throughout the *RUNX1* gene, including missense, nonsense, frameshift and indel mutations (Baptista et al., 2017). Inherited *GATA2* mutations associated with familial MDS/AML have also been reported in several studies (Gao et al., 2014; Hahn et al., 2011) and the growing list of predisposition genes associated with myeloid neoplasms also includes *DDX41*, *ANKRD26* and *ETV6* (Obrochta and Godley, 2018).

1.2.5 Mastocytosis

Mast cells (MCs) originate from the multipotent HSC that, after leaving the haematopoietic tissue as mast cell progenitors (MCPs), migrate through the peripheral blood to the connective or mucosal tissue, and then proliferate and differentiate into MCs (Kitamura et al., 1979). Once differentiated, MCs maintain high expression of the KIT receptor, also known as CD117 (Chen and George, 2018). MC granules mainly store mature tryptase, a tetrameric serine protease, and activation of MCs can lead to an elevated basal serum tryptase level, which has been established to be clinically significant in mastocytosis as well as other myeloid neoplasms (Arber et al., 2016a; Khoury and Lyons, 2019; Payne and Kam, 2004). Mastocytosis is a heterogeneous neoplasm that is characterised by abnormal growth and accumulation of clonal MCs in the BM and/or other tissues/organs. Mastocytosis can occur during childhood or adulthood. In most childhood cases, mastocytosis is limited to the skin, whereas in adults a systemic condition is more common with less than 5% of cutaneous forms in adults. The disease will present itself in males and females in equal ratios, although affected males are more predominant during childhood and female predominance is more likely to happen in adulthood (https://rarediseases.info.nih.gov/diseases/6987/mastocytosis#ref_8371).

Classification and diagnostic criteria of mastocytosis were revised in 2016 by the WHO (Arber et al., 2016a). Mastocytosis represents a specific disease category, and due to its peculiar features is no longer considered a subgroup of MPNs, although it is clearly related to these disorders (Arber et al., 2016a). Mastocytosis is currently subclassified into three groups; cutaneous mastocytosis (CM), SM and mast cell leukaemia (MCL). CM occurs more frequently during childhood and is considered a skin disease. In contrast, SM occurs with a higher incidence in adults and the neoplastic MCs form focal and/or diffuse infiltrates in several tissues/organs such as bone marrow (BM), liver and spleen, leading to their functional impairment (Kristensen et al., 2011). The research described in the following two chapters focuses on SM, and in the following paragraph I will give more insights into this disease.

1.2.5.1 Systemic mastocytosis

SM is a rare disease with a worldwide prevalence estimated to be between 1/20,000 and 1/40,000 (https://www.orpha.net/). According to the WHO diagnostic criteria, a biopsy of the BM or sections of other extracutaneous organs is needed to detect aggregates of MCs, where an aggregate contains at least 15 MCs. This represents the major diagnostic criteria for SM. Depending on which organ is primarily affected by MC accumulation, five different SM subtypes have been identified. Indolent systemic mastocytosis (ISM) is the most common phenotype and is associated with a normal life expectancy. ISM only rarely develops into a more advanced phenotype. Smouldering systemic mastocytosis (SSM) also has a relatively benign phenotype but it can transform into an advanced subtype. The remaining SM subtypes are associated with a shorter life expectancy; systemic mastocytosis with an associated haematological neoplasm (SM-AHN), aggressive systemic mastocytosis (ASM), and mast cell leukaemia (MCL) (Arber et al., 2016a). Sometimes, the major criteria may not be sufficient for the final diagnosis and therefore the following minor diagnostic criteria have been established for SM: biopsy sections of BM or extracutaneous organs showing more than 25% MCs with atypical or spindle-shaped morphology; detection of a *KIT* point mutation at codon 816 in the BM or another extracutaneous organ; MCs in BM, blood or other extracutaneous organ expressing CD2 and/or CD25 which are not expressed under healthy physiological conditions; baseline serum tryptase (BST) level greater than 20 ng/mL (assuming the absence of an unrelated myeloid neoplasm). Based on WHO 2016 guidelines, the diagnosis of SM is established following the detection of the major criteria and one minor criterion, or at least three minor criteria (Arber et al., 2016b; Chen and George, 2018; Valent et al., 2017a).

1.2.5.2 The *KIT* gene and driver mutations that activate the KIT receptor

The *KIT* gene encodes the receptor tyrosine kinase KIT. The extracellular domain of KIT contains five Ig-like modules that bind stem cell factor (SCF), a cytoplasmic region containing a regulatory juxtamembrane domain (JMD) and a tyrosine kinase domain (TKD). The extra and intracellular domains are connected by a hydrophobic transmembrane domain (TMD). *KIT* is expressed throughout the entire development of MCs and is essential for their survival (Kitamura et al., 2007). KIT is normally activated by stem cell factor (SCF) binding, which induces dimerisation of the receptor and upregulation of the tyrosine kinase activity and subsequent downstream signalling pathways (Figure 1.2).

Gain-of-function mutations in the *KIT* gene constitutively activate the KIT receptor, causing continuous growth and survival of MCs in the absence of SCF (Kitamura et al., 2007). Approximately

90% (Table 1.2) of adult SM patients have a specific somatic driver mutation (c.71763A>T p.D816V, substitution of an aspartate with a valine) in *KIT* which is significant for both diagnosis of mastocytosis and therapeutic decision-making (Baird and Gotlib, 2018). Other rare somatic *KIT* mutations (e.g. D815K, D816Y, D816F, D816H, D820G, V560G, A502_Y503dup) have been detected in less than 5% of patients (Table 1.2) (Conde-Fernandes et al., 2017; Manthri et al., 2020; Mital et al., 2011; Ustun et al., 2016). Somatic mutations have been identified in other genes in advanced SM patients (e.g. *TET2, SRSF2, ASXL1, CBL, RUNX1, RAS, EZH2* and *JAK2*^{V617F}) and some of these additional mutations confer a poor prognosis, notably *SRSF2*, *ASXL1* and *RUNX1* (Jawhar et al., 2015; Manthri et al., 2020; Valent et al., 2017b). Although the KIT receptor is considered a target of the tyrosine kinase inhibitor (TKI) imatinib, this compound is ineffective against D816V, as this mutation locks the receptor into an active conformation that imatinib is unable to access (Frost et al., 2002). Other *KIT* mutations, however, may be responsive to imatinib (Manthri et al., 2020; Mital et al., 2011) and encouraging clinical results have been obtained in mastocytosis using the alternative KIT inhibitor midostaurin (Gotlib et al., 2016), and more recently avapritinib (Gilreath et al., 2019). Cladribine is a non TKI-based chemotherapy, and although it has been effective for mastocytosis patients, its use has declined with the advent of TKIs targeting KIT; however, it still remains a safe drug to consider during pregnancy (Gilreath et al., 2019). *KIT* mutations are not only seen in mastocytosis but also characterise gastrointestinal stromal tumours (GST) and are often seen in AML with core binding factor fusion genes *RUNX1-RUNX1T1* and *CBFB-MYH11* (Faiyaz-Ul-Haque et al., 2018; Hirota et al., 1998; Ishikawa et al., 2020; Liu et al., 2020). In GSTs, the most common mutations are localised on exon 11 (70% of cases); these mutations have also been reported to be involved in the development of liver metastasis (Liu et al., 2020; Tanaka et al., 2010). However, other *KIT* mutations in GST patients are found on exon 9 (5–10%),13 (1–3%) and 17 (<1–3%) as well as a novel cyclin Y like 1 (*CCNYL1*)-*BRAF* gene fusion (Liu et al., 2020).

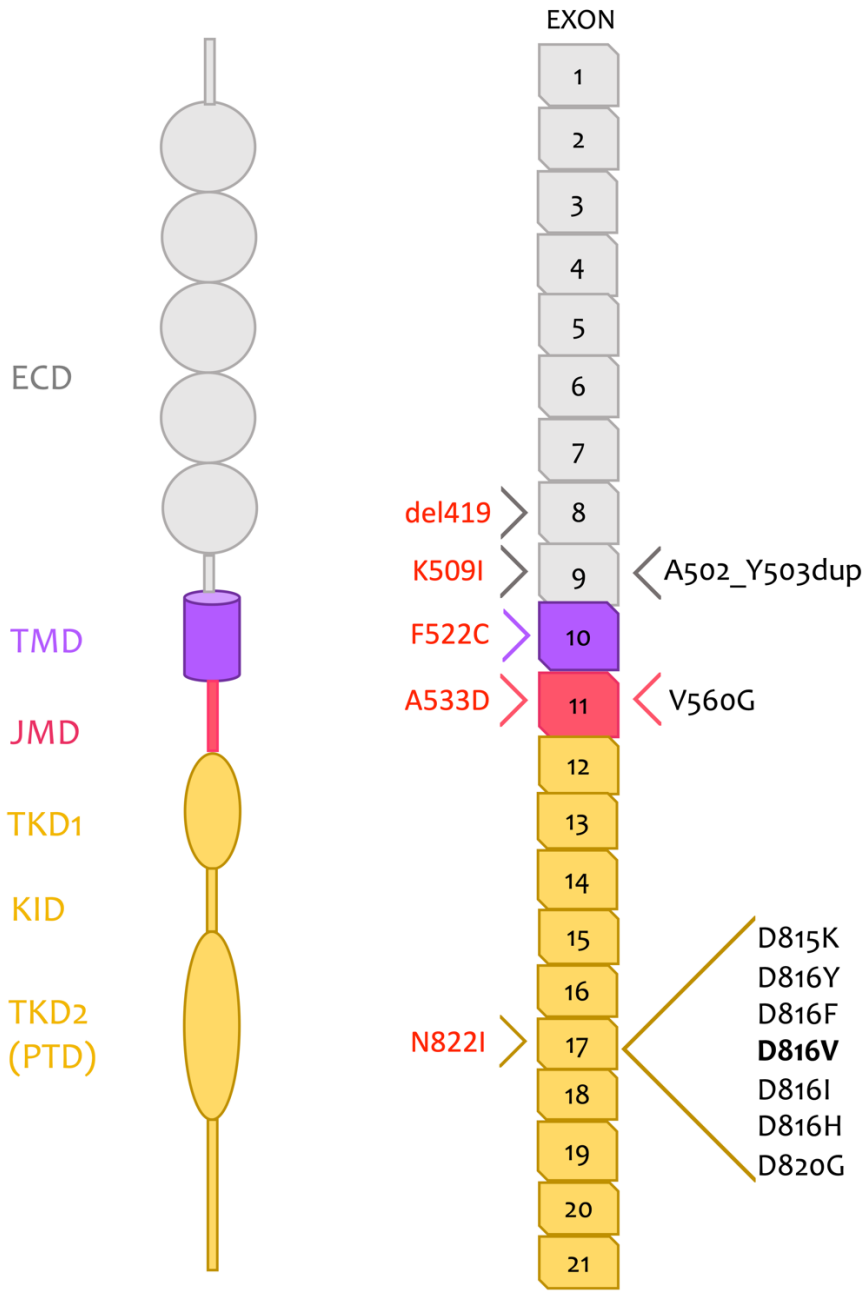


Figure 1.2 **Schematic of KIT receptor and localisation of main somatic and germline mutations observed in the sequence of the KIT gene in association with SM.**

The figure shows the KIT receptor tyrosine kinase in its monomeric form. *KIT* is a proto-oncogene of 21 exons, located on chromosome 4, that encodes the KIT transmembrane receptor comprised of 976 amino acids. The receptor is composed of an extracellular domain (ECD) (in light grey), a TMD (in purple) and an intracellular domain. The ECD contains five Ig-like modules which are crucial for positioning KIT dimers in the correct orientation during the dimerisation of the receptor. The cytoplasmic region contains a JMD (in magenta) and a TKD (in yellow) composed of TKD1 and TKD2, and linked by a kinase insert domain (KID). The most common activating mutation (D816V) highlighted in bold, occurs in TKD2 and affects the cytoplasmic phosphotransferase domain's (PTD) activation loop (A-loop). The ligand binding site and dimerisation site are in the ECD. In the figure, the exon numbers are shown in boxes and the main somatic (in black) and germline (in red) mutations identified in SM are indicated in the corresponding exonic regions (Baird and Gotlib, 2018; Ustun et al., 2016). Mutations marked as germline may also be acquired somatically.

Familial cases of CM and SM have also been reported (Hartmann et al., 2005; Wasag et al., 2011), with some families testing positive for inherited *KIT* mutations. A case study reported a father and two children affected with CM and harbouring a *KIT* p.N822I missense mutation. It was shown that this mutation constitutively activated the KIT receptor and also that N822I is resistant to imatinib but sensitive to dasatinib (Wasag et al., 2011). A study described a novel *KIT* germline mutation in exon 8 (del419) in a German family affected with gastrointestinal stromal tumour and mastocytosis. This mutation is a deletion affecting the extracellular domain of the receptor and was previously reported in one case of AML as well as childhood CM. *In vitro* experiments demonstrated that the constitutive phosphorylation of KIT was inhibited by imatinib (Hartmann et al., 2005). Another interesting study reported a K509I mutation associated with familial SM. A woman and her daughter harboured the same mutation and after sequencing both parents of the woman, the mutation was identified as an acquired *de novo* mutation, which was transmitted to the daughter. *In vitro* experiments showed that imatinib was able to induce apoptosis of MCs harbouring the *KIT* K509I mutation. The clinical condition of both patients improved remarkably after three months of treatment with imatinib (de Melo Campos et al., 2014).

Mastocytosis is considered clinically to be part of a wider range of mast cell activation disorders (MCAD), including mast cell activation syndrome (MCAS). MCAS is a poorly understood immunological condition in which mast cells inappropriately and excessively release chemical mediators, resulting in a range of chronic symptoms, including anaphylaxis. Thus far no clearly recurrent genetic abnormalities have been described in MCAS; however, a review of familial cases showed that approximately 75% of mast cell activation disease (MCAD) patients had at least one first-degree relative with MCAD, which indicates a significant germline contribution (Molderings et al., 2013).

Table 1.2 **Mutations in the *KIT* gene sequence in patients with mastocytosis.**

DOMAIN	MUTATION	LOCATION	FREQUENCY (ADULTS)	FREQUENCY (CHILDREN)
Extracellular domain Ig-like module 5	Missense, indels	Exon 8	5%	20%
	Missense, duplication	Exon 9	<5%	25%
Transmembrane domain	Missense	Exon 10	rare	rare
Intracellular domain	Missense, deletions	Exon 11	<5%	<5%
Tyrosine kinase domain	D816V	Exon 17	80–90%	40%
	D816 other	Exon 17	<5%	5%
	missense	Exon 18	<5%	rare

1.3 Clonal haematopoiesis in healthy people

Clonal haematopoiesis (CH) refers to the clonal expansion of any haematopoietic cells which have acquired somatic mutations or chromosomal abnormalities over time (Jaiswal and Ebert, 2019). Several studies have shown that the expansion of haematopoietic cell clones in the general population, termed age-related clonal haematopoiesis (ARCH)/clonal haematopoiesis of indeterminate potential (CHIP), is common in healthy elderly individuals and is associated with an increased risk of developing haematologic cancer as well as other cancers, cardiovascular disease and other age-related diseases (Bick et al., 2020; Busque et al., 2012; Genovese et al., 2014; Jacobs et al., 2012; Jaiswal et al., 2014, 2017; Laurie et al., 2012; Xie et al., 2014). For the purpose of this thesis, I will focus on the association between CHIP and haematological malignancies. A study performed on blood-derived DNA from a Swedish cohort identified genes that are most frequently mutated in association with clonality and observed CH with somatic mutations in 10% of individuals aged 65 or older. The 12,380 samples were unselected for blood cancer and their health condition was followed for up to 7 years after sample collection. Interestingly, 42% of participants who developed haematological malignancy during the study period had CH at study entry (Genovese et al., 2014). These results were confirmed by a second study of 17,182 samples coming from five different populations (African-American, East Asian, European, Hispanic, South Asian) (Jaiswal et al., 2014). Although individuals with CH are clearly at risk of developing a haematological malignancy, the rate of progression was only about 1% per annum.

1.3.1 CHIP/ARCH and associated mutations

DNMT3A, *ASXL1* and *TET2* are frequently mutated in patients with AML and MDS, and are also the most frequently mutated genes in apparently healthy individuals with CH (Busque et al., 2012; Genovese et al., 2014; Jaiswal et al., 2014). Findings from another study suggested that *DNMT3A*

R882H is particularly common in driving clonal events. They also demonstrated that CHIP is far more frequent, as they observed CH with *DNMT3A* and *TET2* mutation in 95% of the healthy samples aged 50–60 (Young et al., 2016). Other frequently mutated genes are *PPM1D*, *JAK2*, *TP53*, *GNAS*, *BCORL1* and *SF3B1* (Genovese et al. 2014; Jaiswal et al. 2014; Xie et al. 2014). These genes can be used as a marker for early detection of CH in individuals that have not developed clinical symptoms for haematologic cancers (Genovese et al., 2014).

1.4 Chromosomal abnormalities in myeloid neoplasms

Genome instability and mutations are one of the hallmarks of cancer (Hanahan and Weinberg, 2011). Chromosomal abnormalities in cancer were first described between 1890 and 1914 by Hansemann and Boveri who were performing microscopic analysis of cancer cells (Calkins et al., 1914; Hansemann, 1890). Both chromosomal and molecular abnormalities can be responsible for the initiation of a malignant event and they can also be identified as clonal markers (Nowell, 1976). Not all the cells in the malignant tissue acquire genomic anomalies and some cells do not acquire proliferative advantage (Heim and Mitelman, 2015). A study conducted on 50K samples from the general population using SNP microarray data showed that detected mosaic chromosomal anomalies associated with CH tend to overlap with the same regions of copy-number variants or copy-number neutral events as those that are seen in haematological malignancies (Laurie et al., 2012). The investigation of chromosomal abnormalities in the genome is particularly important for the genetic diagnostic and clinical management of haematological malignancies (Arber et al., 2016a; Swerdlow et al., 2016).

1.4.1 Chromosomal abnormalities

Balanced chromosomal translocations are key abnormalities in the diagnosis of leukaemia and lymphoma and for understanding the pathogenesis of these diseases. Translocations may generate dominantly acting fusion genes that act as primary drivers of the disease process, or may result in aberrant expression of neighbouring genes. The prime example of a reciprocal translocation giving rise to a fusion gene is the Philadelphia chromosome, the smaller derivative of a translocation event between chromosome 9 and chromosome 22. The *BCR-ABL1* fusion gene resulting from this rearrangement encodes a deregulated tyrosine kinase protein and is associated with the development of CML, as well as up to 50% of ALL and 1% of AML (Johansson and Harrison 2015; Kang et al. 2016).

1.4.2 Loss of heterozygosity

LOH is a common genetic event occurring in cancer (Ryland et al., 2015). It involves the conversion of heterozygous loci to a homozygous state by a variety of mechanisms. LOH can cause the loss of normal function of one allele in a tumour suppressor gene or oncogene in which the other allele was already inactivated. Alternatively, LOH can convert a heterozygous driver mutation to homozygosity, which may provide an additional clonal advantage. Regions of LOH may span entire chromosomes or short sections of DNA, and they can occur due to copy number losses (CNV-LOH) or they can be copy number neutral (CNN-LOH), associated with acquired uniparental disomy (UPD) (O'Keefe et al., 2010; Ryland et al., 2015).

1.4.3 Acquired uniparental disomy

UPD is a type of LOH event whereby both copies of a chromosome pair or parts of chromosomes have originated from one parent (Engel, 1980). Inherited UPD, where both chromosome copies are inherited from one parent, occurs due to errors in meiosis and is associated with developmental disorders resulting from abnormal expression of imprinted genes (Robinson, 2000). In contrast, somatically acquired UPD (aUPD) occurs in cancer as a result of mitotic errors, either non-disjunction resulting in aUPD of a whole chromosome, or more commonly recombination involving a whole chromosome arm or terminal segments followed by disjunction and DNA replication resulting in aUPD/LOH in the recombined region (Tuna et al., 2009) (Figure 1.3).

1.4.4 Regions of aUPD in healthy people

Regions of aUPD in healthy individuals represent another form of CHIP/ARCH. Only 0.5% of people under the age of 50 are affected but this rises to 2–3% of individuals over 50 and 10% of elderly individuals aged 65 and older. Importantly, the chromosomal regions affected are almost identical to those seen in patients with haematological malignancies and involve the same mutant genes (Genovese et al., 2014; Jacobs et al., 2012; Laurie et al., 2012). The finding of aUPD in an otherwise haematologically normal individual is associated with a tenfold increased risk of subsequently developing haematological neoplasia (Jaiswal et al., 2014; Laurie et al., 2012). This observation suggests that large genomic datasets accumulated in the study of benign conditions could be used to facilitate the detection of rare abnormalities associated with haematological neoplasms.

1.4.5 Detection of aUPD

Regions of aUPD cannot be detected by conventional chromosome analysis but can be detected using SNP arrays or NGS to examine the status of polymorphisms (Afyounian et al., 2017; Score and Cross, 2012). Several bioinformatics tools, such as B allele frequency (BAF) segmentation (Figure 1.4), ExomeAI and Segmentum, have been developed for identifying regions of allelic imbalance (AI) in cancer cells from both SNP array and NGS data (Afyounian et al., 2017; Nadaf et al., 2015; Staaf et al., 2008).

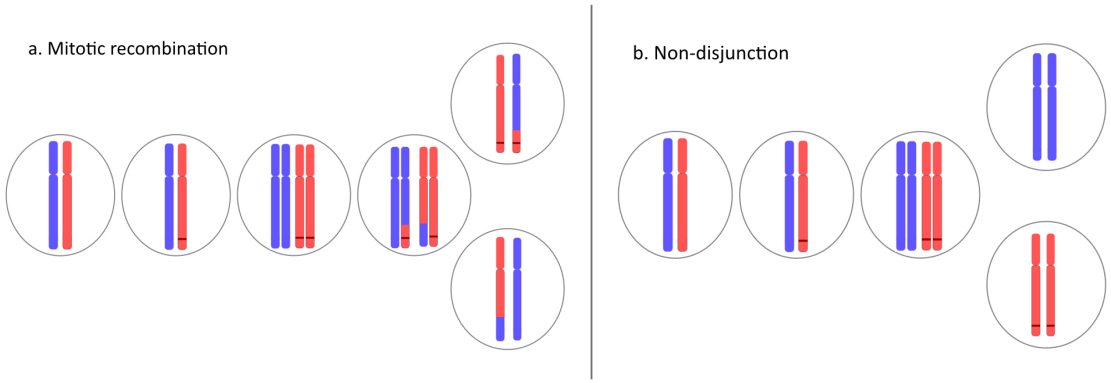


Figure 1.3 **Mechanism of acquired UPD.**

Mechanisms leading to segmental and numerical aUPD. Segmental aUPD can be either telomeric or interstitial. Telomeric aUPD can occur following a single mitotic recombination event leading to exchange of chromatids (a). This mechanism can also generate interstitial aUPD although two consecutive or simultaneous homologous recombination steps are required (Makishima and Maciejewski, 2011). Numerical aUPD can also be a result of mitotic errors, such as chromosomal non-disjunction, in which cohesin complexes holding the chromatids fail to be removed and sister chromatids are incorporated into the same daughter cell (b). Another mechanism causing numerical aUPD is anaphase lag if during the anaphase a chromosome is delayed in its movement and fails to be incorporated into one of the two daughter nuclei. Anaphase lag can be followed by degradation of the chromosome not entering the nucleus and replication of the remaining chromosome (Strachan and Read, 2011).

To generate SNP array data, fragmented single-stranded sample DNA is hybridised to the array, which consists of up to one million or more nucleotide probe sequences using modern platforms. SNP array genotyping generates two intensity values, one for each allele, for each SNP on the array. After hybridisation, the signal intensity, which is associated with the quantity of target DNA in the sample, is measured. The intensity values are transformed to give normalised intensity values (R) and allelic intensity ratios (θ) which are used to calculate BAF and log R ratio (LRR) for identifying structural chromosomal variation. The BAF reflects the probe intensity for a SNP

relative to the expected probe intensity for AA, AB and BB genotypes. The BAF plot is the amount of B allele observed in a probe that should concentrate at zero for zero copy (genotype AA), at 0.5 for one copy (genotype AB) and at 1 for two copies (genotype BB). BAF values of 1 or 0 are therefore expected in LOH regions. However, in a tumour sample with LOH the BAF values may not reach 0 or 1 because of mosaicism; i.e., the tumour consists of a mixed population of cells with and without LOH. In these cases, the BAF values need to be significantly different from 0.5 in order to identify LOH regions.

LRR is the ratio between observed normalised intensity of the experimental sample versus the expected intensity. In a LRR plot, copy number gains and losses are indicated by values that are significantly greatly or lower than zero respectively (Illumina, 2010; Staaf et al., 2008). In NGS data from paired tumour/normal samples, the BAF and LRRs are calculated using read depth for reference and alternate alleles, which is extracted from the binary alignment/map (BAM) file. Different algorithms are then used to calculate the BAF and LRR based on the read depth data (Afyounian et al., 2017; Nadaf et al., 2015). Regions of aUPD are identified as regions with BAF that are significantly different from 0.5 and that have two copies (copy number neutral) and therefore look normal in the LRR plots (Illumina, 2010) (Figure 1.4).

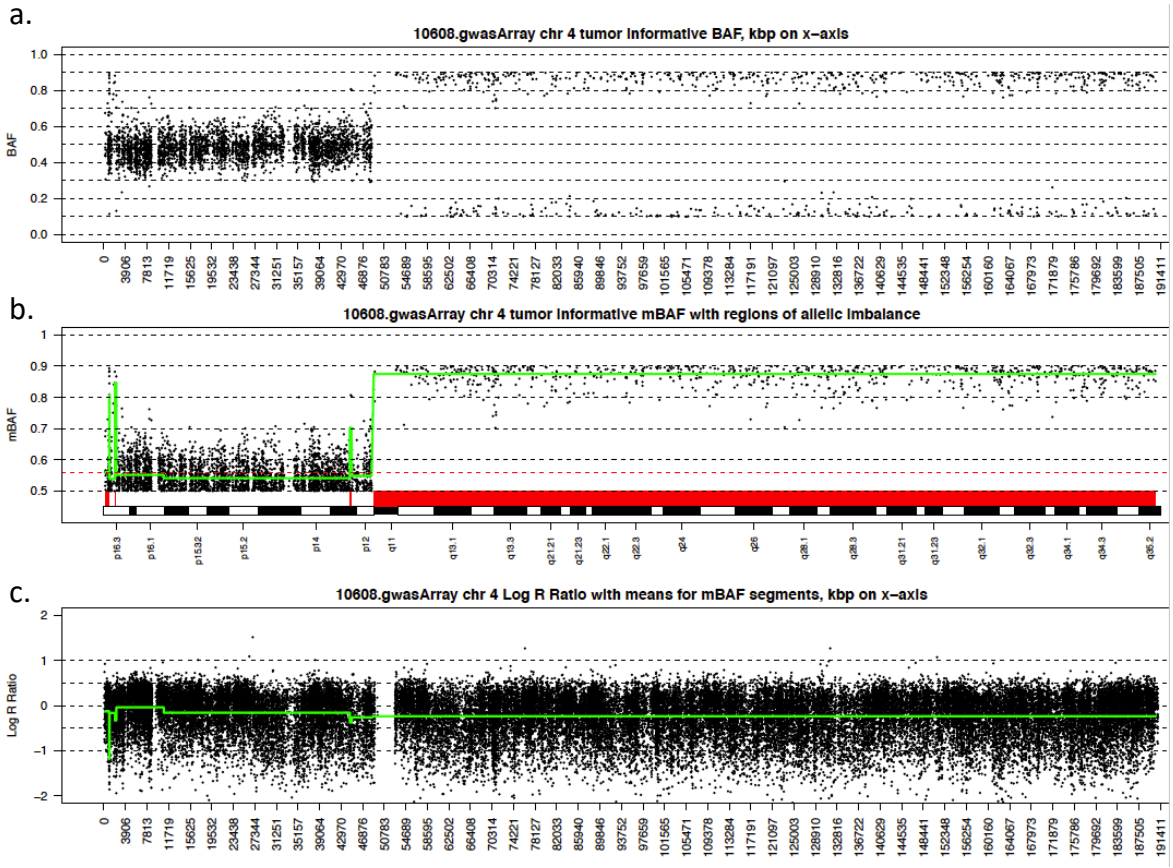


Figure 1.4 **BAF, mBAF and LRR plots obtained with BAF segmentation.**

Panel A represents the BAF plot for chromosome 4. Panel B shows the transformation of BAF values, which are reflected along the 0.5 axis to give mirrored BAF (mBAF). Regions of AI are identified where the segmented mBAF is > 0.56 (red dashed line) and highlighted by a red rectangle. The green line shows the circular binary segmentation (CBS) profile applied to the mBAF values to identify regions of similar allelic proportions. The plot in panel C is the copy number profile with CBS in green used to merge regions with similar level of LRR.

1.4.6 Identification of genes underlying aUPD in haematological neoplasms

The identification of regions of recurrent aUPD has led to the discovery of both novel driver genes and imprinted loci associated with haematological neoplasms (Chase et al., 2015; O'Keefe et al., 2010). For example, SNP array profiling revealed a minimal recurrent region of aUPD on chromosome 11q in 58 patients with aCML, *JAK2* mutation-negative myelofibrosis or *JAK2* mutation-negative PV. Subsequently, the *CBL* gene on 11q23.3 was identified as a candidate gene and sequenced in patients with 11q aUPD and a bigger cohort of MPN patients. These sequencing studies identified a causal somatic *CBL* mutation in 3 of the 11q aUPD patients and in 26 patients from the wider MPN cohort (Grand et al., 2009). Similarly recurrent regions of aUPD and mutation screening have been used to identify *TET2* on chromosome 4q24 in MDS patients (Langemeijer et al., 2009; Massé et al., 2009; Mohamedali et al., 2009), *EZH2* on 7q36.1 (Ernst et al., 2010; Nikoloski et al., 2010), *JAK2* on 9p (Kralovics et al., 2002; Tiedt et al., 2005), *MPL* on 1p and *FLT3*

on 13q (Kralovics et al., 2002; Raghavan et al., 2008; Score and Cross, 2012). In contrast, the imprinted *MEG3-DLK1* locus was identified as a target of 14q aUPD after demonstrating the consistent loss of maternal chromosome 14 and gain of paternal chromosome 14 (Chase et al., 2015).

1.5 Genome-wide association studies

Genome-wide association studies (GWAS) have given much insight into the genetic basis of complex and multifactorial diseases and have generated many scientific discoveries over the last 15 or more years (Ferrari et al., 2014; Ku et al., 2010; Tapper et al., 2015; Visscher et al., 2012). The aim of a GWAS is to identify genes which predispose to a trait of interest. The method involves genotyping approximately 1 million SNPs spread across the genome in as many unrelated cases and controls as possible. The SNPs are then tested for association with the trait of interest by comparing their allele frequencies in cases and controls. SNPs with significantly different allele frequencies can then be used to pinpoint the causal gene(s) (Visscher et al., 2017). In contrast to rare variants related to Mendelian disease, which can be identified using linkage and sequencing technologies (Boycott et al., 2013), GWAS are more suited to detecting common variants underlying polygenic disorders (Smith and Newton-Cheh, 2009). One of the strengths of this technique is that no prior hypothesis of likely candidate genes or disease pathogenesis is needed. Therefore, GWAS may discover novel pathways and genes that would not have been considered based on their function.

To date a one stage GWAS of mastocytosis (Nedoszytko et al., 2020) and two GWAS of MPN have been reported (Hinds et al., 2016; Tapper et al., 2015). Tapper et al. demonstrated that genetic variation at *MECOM*, *TERT, JAK2* and *HBS1L-MYB* predisposes to *JAK2*-unmutated MPN and that *HBS1L-MYB* and the *JAK2* 46/1 haplotype influences whether *JAK2*^{V617F} mutated cases presented with PV or ET. They also showed that SNPs in *TERT* are associated with MPN and that additional SNPs in *SH2B2, ATM*, *CHEK2*, *GFI1B*, and *PINT* predispose to *JAK2*^{V617F}-positive MPNs (Tapper et al., 2015). A second GWAS was performed to identify germline alleles predisposing to Philadelphia chromosome-negative MPNs and *JAK2*^{V617F} CH in the general population. As a result, inherited genome-wide significant loci were found in or near *TERT*, *SH2B3* and *TET2*. The joint analysis of the stage 1 and replication results identified additional germline risk factors associated with age-related *JAK2*^{V617F} CH as well as *JAK2*^{V617F}-negative MPN (Hinds et al., 2016). These studies have recently been extended with the identification of new risk loci for MPN, and functional data

indicating that selected risk loci modulate the function of HSCs (Bao et al., 2020). GWAS have also been very successful in lymphoid disorders with the identification of multiple loci predisposing to ALL, chronic lymphocytic leukaemia and myeloma (Di Bernardo et al., 2008; Chubb et al., 2013; Crowther-Swanepoel et al., 2010; Papaemmanuil et al., 2009).

GWAS of rare diseases with a non-Mendelian pattern of inheritance have also been very successful, especially among neurodegenerative disease (e.g., amyotrophic lateral sclerosis, frontotemporal dementia and corticobasal degeneration) and cancer (Campa et al., 2020; Chio et al., 2009; Ferrari et al., 2014; Kouri et al., 2015). For example, a recent study conducted on European individuals affected with a rare malignant tumour of the eye identified a risk allele in a region associated with overexpression of the *CLPTM1L* gene (Mobuchon et al., 2017).

1.5.1 Study design and population structure

Study design and population structure need to be considered before sampling and genotyping based on the disease prevalence and how the disease segregates in the family. The main study designs are population-based or family-based. Population-based studies include case-control studies of unrelated people, cross-sectional studies, prospective and retrospective cohort studies and studies in population isolates.

Case-control studies are sensitive to population stratification; for this reason both cases and controls should be selected from a homogeneous population (Lieb, 2013; Smith and Newton-Cheh, 2009). In a population study design, even though population stratification can be adjusted during the analysis, it is more opportune to minimise these types of errors during the study design by sampling cases and controls from the same population (Zondervan and Cardon, 2007). In contrast, family-based studies are performed within the family and will not present problems due to population stratification (Hong and Park, 2012). However, it can be difficult to accumulate a large number of affected pedigrees. Family studies may therefore lack power to detect genetic effects due to their small sample size. However, studies have shown that case-control and family-based designs give relatively similar estimates of association (Evangelou et al., 2006).

Successful GWAS requires sufficient statistical power and appropriate sample size in order to reduce spurious results (Jones, 2003). The power of a genetic study measures the probability of

detecting the hypothesised association between a SNP and the disease. A large number of samples are more informative and give more strength to the study. In some case-control studies, the number of cases is limited because the disease is rare. In this situation, study power can be increased by increasing the number of controls (Mobuchon et al., 2017; Smith and Newton-Cheh, 2009).

1.5.2 Single nucleotide polymorphisms

GWAS uses high-density arrays to genotype approximately 1,000,000 SNPs in a single reaction, screening many patients on a genome-wide scale (LaFramboise, 2009; Manolio et al., 2009). SNPs are the most common genetic variation in the genomic DNA, and are selected to have a frequency greater than 1% in the entire population (LaFramboise, 2009). SNPs consist of single base-pair (bp) change in certain genome positions and, on average, they occur once in every 300 base pairs of the human genome (Strachan and Read, 2011). SNPs are bi-allelic, and the less common allele is known as the minor allele (Bush and Moore, 2012). The SNP database (dbSNP), has catalogued a total of 364,060,923 human SNPs for build 151, which are identified with an unique "reference SNP" (rs) number (https://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi?view+summary=view+summary&build_id=151).

1.5.3 Data quality control

Assessment of data quality is an important step during GWAS. Quality control (QC) can be considered under two different aspects. The first is related to technical quality; e.g., missing SNPs data due to intensity measurement issues when performing genotype calls. The other aspect is downstream QC, aimed at evaluating the different sources of error. Genotype QC is performed by analysing raw intensity data using specific genotype-calling algorithms that estimate the probability for AA, AB or BB genotypes. Only the genotypes whose probability is over a set threshold are selected as 'called', whereas the remaining genotypes are indicated as 'missing'. It is important to apply the correct threshold depending on the study. For instance, in GWAS a high calling threshold could generate a high rate of missing genotypes and reduce genomic coverage and genetic power, which are important factors for detecting association. Genotype quality can be manually inspected using cluster plots, and it is recommended that after association testing these plots are checked for all SNPs taken forward for replication (Anderson et al., 2010).

Downstream QC is applied after the genotypes have been called. The main purpose of QC is to remove samples and SNPs with poor genotyping that can lead to false positive associations.

Usually, up to one million markers are tested for GWAS; therefore, even a low rate of poor genotypes can lead to numerous false positives that should not be selected for replication. These downstream steps analyse data from samples (subject-based quality measures aimed at assessing genotyping errors such as contamination, duplication or poor DNA quality) and SNPs (variant-based quality measures) and those poorly genotyped will be removed (Anderson et al., 2010).

1.5.4 Linkage disequilibrium

Linkage disequilibrium (LD) describes the non-random association between alleles at two or more linked loci on a contiguous stretch of genomic sequence. The term was coined to describe the correlation between genetic variation in a population over time. Considering a haplotype block as a genomic region with linked sets of alleles, LD describes also the low probability of altering the haplotype structure through recombination events (Andrew, 2007). LD patterns are a result of population size, natural selection, genetic distance, rate of recombination and mutation events over many generations.

As a result of LD, the frequencies of two alleles observed in the same haplotype compared to the frequency expected if the alleles are independent may show positive or negative LD (Goode, 2011). Based on their allele frequencies, positive LD occur when two alleles exist on the same haplotype more often than expected, whereas a negative LD means that alleles can occur together less frequently than expected (Earp and Goode, 2017). The difference between the observed and the expected frequencies can be measured by different LD metrics. D' and r^2 represent the most commonly used measures of LD (Devlin and Risch, 1995). The covariance (D) represents the difference between expected and observed haplotype frequencies and, since it is sensitive to allelic frequencies, it is not calculated at the extreme values of 0 or 1 (Goode, 2011). In order to reduce frequency dependence, Lewontin used the measure D', which is a normalised D ranging from 0 to 1 that can represent complete linkage equilibrium or no recombination between the two markers respectively. LD between genetic variants is more often measured using the Pearson correlation coefficient, also termed squared correlation coefficient (r^2), which is also scaled from 0 to 1 for completely independent and dependent (co-inherited) polymorphisms, respectively (Bush and Moore, 2012; Lewontin, 1964). LD r^2 is dependent on allele frequencies, so in order to increase the likelihood of detecting disease association, it is important to take into account the maximum difference in allele frequencies between two loci when selecting candidate SNPs (Wray, 2005). LD measures represent an essential tool in several steps of a GWAS, such as

imputation, detection of strand issues, selection of independent SNPs and clumps of correlated SNPs, and defining regions of interest.

1.5.5 Association tests

In GWAS, the association between genotype and phenotype is performed for each SNP using contingency tables or regression to assess differences in the distribution of alleles or genotypes (Bush and Moore, 2012). The null hypothesis of no association with the disease is true when no significant difference is detected in allelic or genotypic frequencies between cases and controls.

In a case-control study, genotypic tests use a 2×3 contingency table of genotypic counts which has 2 degrees of freedom (df). These tables can be collapsed to test for both dominant and recessive models. Allelic tests may also be applied, which use a 2×2 contingency table with one df. Allelic tests are considered to be most powerful statistic for testing a multiplicative model of penetrance.

In some studies, the association test needs to account for the effects of population, epidemiological risk factor (e.g., gender, diet or geographic location) and clinical variables (e.g., treatment, body mass index). In these situations the factors can be treated as covariates, using linear regression for quantitative traits and logistic regression for binary traits such as case or control status (Clarke et al., 2011). For instance, spurious association signals can occur if there are differences in ethnicity between cases and controls since allele frequencies may vary as a result of ethnicity rather than association with disease risk. To minimise spurious association due to ethnic differences, principal component analysis (PCA) or multidimensional scaling can be used to either identify and remove outliers or to generate principal components that can be used as covariates in the statistical association tests that account for difference due to population stratification (Anderson et al., 2010).

The statistical analysis involves multiple independent tests, a fraction of which may produce false positive association signals (type 1 errors). For instance, if a total of 10^6 alleles need to be tested in a GWAS, a strict control for type 1 error is required (Dudbridge and Gusnanto, 2008; Pe'er et al., 2008). This is generally accounted for by adjusting the threshold needed in a single test for the null hypothesis to be rejected. The Bonferroni correction represents a widely accepted approach to adjusting the P-value threshold for genome-wide significance and to minimise the number of

spurious positive results due to multiple comparisons. This corresponds to dividing the P-value threshold (0.05) by the total number of markers (N) used before testing the association, and the resulting conventional threshold for genome-wide significance is 5×10^{-8} (=0.05/1 million) (Khoury and Yang, 1998).

1.5.6 Follow-up of results: Replication studies

Because of possible errors (e.g., systematic genotyping errors, statistical errors) that may arise during GWAS, replication studies in independent samples are required to validate the association observed at stage 1 (Chanock et al., 2007). It is important that the replication study has sufficient power to confirm or refute findings. Sample size and genetic power therefore need to be considered during the replication and discovery stages (Jones, 2003; Smith and Newton-Cheh, 2009). In order to confirm that observed association is not due to genotyping artefacts, in stage 2 SNPs should ideally be genotyped on a different platform and reanalysed. The selected SNPs could be highly correlated with the phenotype in one cohort used in stage 1 of the analysis, but the same SNPs could be poorly correlated in a different ancestry group (Smith and Newton-Cheh, 2009). This can be determined using the tool Tagger implemented within the program Haploview; this is a SNP haplotype-tagging method based on HapMap samples (de Bakker et al., 2005).

1.5.7 Meta-analysis

To increase power and give new insight into the aetiology of diseases, meta-analysis can be used to combine evidence from separate GWAS. Because of the larger sample size and independent cohorts, this approach can reduce the number of false positives and increase the significance of true positives (Smith and Newton-Cheh, 2009). Recent studies have demonstrated that meta-analysis of GWAS data can identify new susceptibility loci involved in complex diseases (Nalls et al., 2014; Pharoah et al., 2013). Before performing the meta-analysis, any kind of heterogeneity (e.g., sample structure, individual ancestry, population structure, results) between studies must be considered. Heterogeneity of results can be examined using forest plots and statistics such as the χ^2-based Cochran's Q test and I^2 (Smith and Newton-Cheh, 2009). The former is used to detect whether there is a statistically significant heterogeneity between the combined studies (Zeggini and Ioannidis, 2009). The I^2 test is able to analyse whether the percentage of variation is attributed to heterogeneity or to chance (Higgins et al., 2003; Zeggini and Ioannidis, 2009). Once the statistical variation has been detected and the results from each study have been weighted, the data can be jointly analysed. In meta-analysis, a model termed random-effect allows the effect size to be different between cohorts and can be used if the variation is due to heterogeneity.

On the other hand, if the between-studies variation occurs by chance, a fixed-effect model is most appropriate, as it assumes that the variant has one true effect size (Smith and Newton-Cheh, 2009).

1.5.8 Data imputation and the HapMap project

The choice of markers that are representative of the LD pattern of the genome is an important part in the design of GWAS, with a key parameter being the proportion of common variation that is tagged by the subset of genotyped SNPs. Markers that are not directly genotyped but are in high LD with the genotyped markers can be recovered through imputation. In this way, the causal allele could still be detected through its correlation with marker loci genotyped in the assay and associated with the disease. Imputation can be used to improve the resolution of GWAS by estimating what the genotype should be for SNPs with missing genotypes and for SNPs that were not genotyped on the array (Chanock et al., 2007; Sherry, 2001). Genotypes are estimated using the LD pattern in sequenced reference datasets (e.g. HapMap) and known genotypes from the study. The International HapMap Project is the main source for LD information and has produced a map of common human DNA variants that cluster together to form haplotype blocks. The HapMap Consortium used data from healthy individual including African ancestry from Nigeria (Yoruban in Ibadan, YRI), Chinese (Han Chinese from Beijing, CHB), Japanese (Japanese in Tokyo, JPT) and European (Utah residents with ancestry from northern and western Europe, CEU) ancestry to catalogue population-specific differences in genetic variation. The project was completed in 2009, having genotyped 3,000,000 SNPs from 1,301 individuals from 11 human populations (Altshuler et al., 2010b). Another resource in use for genotype imputation is the Human Reference Consortium (HRC), a large reference panel mainly of European ancestry of 64,976 human haplotypes with 39,235,157 SNPs derived from whole exome sequencing (WES) data. A total of 20 studies have been added in the panel and these also include the 1000 Genomes Project Phase 3 cohort. The increased number of SNPs, haplotypes and populations coming from the HRC has enabled an increase of marker density in GWAS samples and therefore the accuracy to infer initially unobserved genotypes (Iglesias et al., 2017; McCarthy et al., 2016).

1.5.9 Strength and weaknesses of GWAS

GWAS have been very successful, having identified nearly 157,000 robust associations involved in a wide range of complex disorders, which are highly replicable within and between populations (MacArthur et al., 2017). However, despite these successes, the GWAS approach has some limitations that need to be considered along with their design and analysis. The detection of false positives is one of the main weaknesses of GWAS. For this reason, study design, QC, correction for multiple testing and replication are all critical steps to optimise the chance of detecting true

positive association whilst maintaining the power of the study (Pearson and Manolio, 2008). Applying stringent significance thresholds is one way of minimising false positives, but multi-stage studies performed on breast cancer and multiple sclerosis have showed that the most robust findings are not necessarily the most significant signals in the discovery stage (Hunter et al., 2007; Strachan and Read, 2011; Verma, 2012).

Since the development of high-throughput SNP arrays approximately twenty years ago, the costs have fallen and the number of SNPs in the arrays have increased. It is now possible to genotype between 200,000 to 2,000,000 SNPs in a single array (Chee et al., 1996; Visscher et al., 2017). These improvements have helped to reduce false negatives through increased SNP coverage and by making genotyping of more samples affordable.

Typically, SNPs identified by GWAS are not causal but in LD with the causal variant(s). Furthermore, risk SNPs are typically located in intronic or intergenic regions. As a result, the biological and functional role of associated SNPs is often unclear and further studies involving fine mapping and functional analyses are required to identify the causal mutation and gene involved, a task that is often very difficult.

When GWAS started there were high expectations of discovering the genetic factors accounting for the heritability of complex traits (Visscher et al., 2008). However, despite huge GWAS for adult height involving 253,288 individuals, which identified 697 variants with genome-wide significance, their combined effect could only explain 20% of the heritability (Genovese et al., 2014). The so-called “hidden heritability problem” can be explained by at least three factors. The first is that the susceptibility in the great majority of complex traits is attributed to a large number of variants with subtle effects that will require enormous sample sizes to detect (Strachan and Read, 2011). Indeed, by considering all common variants the majority (60%) of heritability in adult height could be explained (Visscher et al., 2017).

Second, disease susceptibility may be due to a highly heterogeneous collection of rare variants that display Mendelian inheritance and play a major role in the development of the disease (Strachan and Read, 2011). This is the case of atopic dermatitis or eczema, a common and

complex trait caused by common variants that, in contrast to polymorphisms, cause complete loss of function of the filaggrin gene (FLG). Almost 10% of the European population carries one of the five specific FLG variants. Different sets of variants are common within different populations (Irvine and McLean, 2006). Other approaches should be considered, such as WES or whole genome sequencing (WGS), to discover gene variants that cause susceptibility to complex disease with monogenic Mendelian inheritance patterns (Strachan and Read, 2011). In other cases, rare forms of common diseases with Mendelian patterns can be caused by highly penetrant variants with low (0.5%<MAF<5%) or rare minor allele frequency (MAF<0.5%) which could explain part of the missing heritability (Gibson, 2012). Since these variants are not covered by conventional genome-wide genotyping arrays, new methodologies such as a rare variant association study (RVAS) can be adopted to identify rare variants associated with phenotypic variation (Auer and Lettre, 2015).

The third factor that may account for part of missing heritability in GWAS is represented by additive epigenetic changes (e.g., histone modifications, DNA methylations) transmitted for more generations and that are not taken into account by GWAS (Strachan and Read, 2011; Trerotola et al., 2015).

Finally, some have considered that another weakness of GWAS is that the identified variants tend to have small effect sizes which limit or prevent clinical utility. However clinical utility is only one consideration, and even small effect sizes may provide important new biological insights into disease pathology.

1.6 Aims of study

As reviewed above, genome-wide genetic studies have revealed the importance of germline variation and somatic mutations in the pathogenesis of haematological malignancies. Furthermore, ongoing WES and WGS sequencing projects targeted at specific disorders or conducted at a population level are generating increasingly large datasets of sequence variation. I hypothesise that further insights into the pathogenesis of myeloid neoplasms may be obtained by focusing on genetic predisposition to specific, genetically-defined subtypes of disease. In addition, I hypothesise that large sequence datasets from individuals unselected for a malignant phenotype can be mined to gain new insights into blood cell clonality as a precursor to haematological malignancies. In this context I aim to:

(i) Identify genetic predisposition to mastocytosis using the GWAS approach, and using the somatically acquired *KIT*^{D816V} marker to help to ensure homogeneity of cases, and

(ii) Utilise WES datasets to identify regions of AI and aUPD, and explore the potential of this approach to identify novel driver mutations associated with CH and myeloid neoplasms.

Chapter 2 A Genome-Wide Association Study of Systemic Mastocytosis

2.1 Introduction

Most occurrences of SM are sporadic and over 80% of SM patients have a somatic *KIT*^{D816V} mutation. Familial cases of SM are rare and little is known about the contribution of germline predisposition. However, several familial cases have been reported involving rare highly penetrant germline mutations in the *KIT* gene (Hartmann et al., 2005) or acquisition of somatic *KIT* mutations including D816V (Broesby-Olsen et al., 2012; Zanotti et al., 2013), S849I and M835K (Molderings et al., 2013) by multiple family members. The simultaneous occurrence of these somatic mutations, which includes one pair of monozygotic twins, is unlikely to occur by chance and suggests the involvement of inherited predisposition to acquired somatic *KIT* mutations similar to those seen in MPN involving the somatic mutation *JAK2*^{V617F} (Broesby-Olsen et al., 2012; Jones et al., 2009). Further evidence from family-based studies has suggested that SM has a heritable component following the observation that 74% of patients with systemic MCAD (n=62/84) had at least one first degree relative with suspected MCAD based on a self-reported questionnaire (Molderings et al., 2013). Furthermore, several constitutional genetic variants have been associated with the development of different mastocytosis phenotypes in relatively small candidate gene studies (Daley et al., 2001; Lange et al., 2017; Nedoszytko et al., 2009, 2018; Rausz et al., 2013).

When this study was started, no GWAS had been undertaken to test for germline predisposition to SM. However, other GWAS had demonstrated that germline variation at several loci is associated with the risk of developing MPN and can influence whether MPN patients develop ET or PV (Tapper et al., 2015). Our hypothesis is that inherited genetic factors also predispose to SM. To test this hypothesis, I conducted a two-stage GWAS of SM. To limit genetic heterogeneity and increase power, the GWAS focused on SM patients with somatic *KIT*^{D816V} mutations only. The identification of genetic markers associated with SM may have a clinical impact and will provide insights into understanding whether inherited markers are key factors for predisposing to or protecting from the development of the disease. At stage 1, 479 *KIT* ^{D816V}-positive SM patients were recruited from the United Kingdom and Germany. For comparison, publicly available control cohorts were obtained, consisting of 9,597 healthy controls from the Wellcome Trust Case Control Consortium 2 (WTCCC2) and the Cooperative Health Research in the Region Augsburg (KORA)

study (Burton et al., 2007; Holle et al., 2005). A replication cohort of 666 Spanish, Danish and Italian SM cases with *KIT*^{D816V} mutations were recruited and compared against matching controls to replicate selected SNPs from stage 1.

2.2 Materials and Methods

2.2.1 Discovery and replication cohorts

Careful ethnicity matching of cases and controls at the design stage of the GWAS was aimed at reducing the chance of heterogeneity both in the primary and in the replication study. Prior to quality control (QC), the stage 1 discovery cases consisted of 479 SM cases (hereafter referred to as SM-1). All of these patients had a somatic *KIT*^{D816V} mutation and were recruited from the UK (n=329) and Germany (n=150). At stage 2, 666 independent *KIT*^{D816V} replication patients were recruited from Spain (n=399), Denmark (n=185) and Italy (n=82). Participants provided informed consent for sampling according to the Declaration of Helsinki. All mastocytosis cases were adults diagnosed using standard procedures. The stage 1 discovery cohorts were recruited from two diagnostic laboratories (Wessex Regional Genetics Laboratory, UK and Munich Leukaemia Laboratory, Germany) based on (i) referral for investigation of mastocytosis and (ii) testing positive for *KIT*^{D816V}. A detailed breakdown of WHO-defined clinical subtypes and other clinical information was not available for these cases, but <10% were known to have advanced SM. Clinical subtypes were available for stage 2 cases whose diagnosis was simplified into two main disease groups, non-advanced and advanced. Non-advanced cases (MCAS=mast cell activation syndrome, CM=cutaneous mastocytosis, ISM=indolent systemic mastocytosis, SSM=smouldering systemic mastocytosis) have a good life expectancy and very few of them are likely to develop advanced disease. The advanced disease group is characterised by shorter life expectancy and a more severe phenotype. As described by the WHO classification, only three subtypes (ASM=aggressive systemic mastocytosis, SM-AHN= SM with an associated haematologic neoplasm and MCL=mast cell leukaemia) are included in the advanced disease group (Arber et al., 2016b). A breakdown by subtype for stage 2 cases is given in Table 2.1. Additional diagnostic and clinical variables were only available for the Spanish and Italian cohorts due to ethical limitations regarding consent. The study was approved by UK NRES Committee South West reference 10/H0102/61; Germany: MLL cohort, BLAEK ethics commission, reference 05117; Spain: ethics committee of the University Hospital of Salamanca reference 2016/PI16/00642; Italy: local Ethics

Committee, March 12, 2019, protocol number 14560_OSS. The Danish SM study was performed in accordance with the Danish National Committee on Health Ethics.

Table 2.1 **Breakdown of stage 2 patients cohorts by disease subtype.**

		Non-advanced				Advanced			
Cohort	n	MCAS	CM	ISM	SSM	ASM	SM-AHN	MCL	N/A
Spain	399	6	4	368	3	9	8	1	0
Denmark	185	0	13	152	5	0	12	0	3
Italy	82	2	0	64	4	8	1	1	2

MCAS: mast cell activation syndrome; CM: cutaneous mastocytosis; ISM: indolent SM; SSM: smouldering SM; ASM: aggressive SM; SM-AHD: SM with associated haematologic neoplasm; MCL: mast cell leukaemia; N/A: Data not available

2.2.2 Description of control cohorts

For comparison, 5,200 UK controls from WTCCC2 and 4,397 German controls from KORA were used (Table 2.1). Both WTCC2 and KORA control cohorts comprised two separate studies. The WTCCC2 cohort consisted of participants from the 1958 British birth cohort (BBC, n=2,699) and participants from the National Blood Service (NBS, n=2,501) (Burton et al., 2007), while the KORA controls were KORA_A (n=1,938), representing a subset of follow-up F3 of the population-based survey KORA S3, and KORA_B (n=2,459) (Holle et al., 2005), representing an independent subset of KORA S3/F3.

The stage 2 replication controls were obtained in collaboration with the Spanish National DNA Bank Carlos III (SNDNAB, n=1,062) (Bosch, 2004; Julià et al., 2013), a Danish study of ischaemic heart disease (Inter99, n=6,184) (Jørgensen et al., 2003; Pisinger et al., 2005) and the Italian Invecchiare in Chianti study (InCHIANTI, n=1,210) (Ferrucci et al., 2000; Tanaka et al., 2009).

The Spanish individuals were all adults, gave informed consent and were determined to be healthy based on self-reported health status obtained from personal interviews. See http://www.bancoadn.org for further details.

The Inter99 study is a randomised, non-pharmacological intervention study for the prevention of ischaemic heart disease (Husemoen et al., 2003; Jørgensen et al., 2003). In brief, more than 13,000 individuals between 30 and 60 years of age and from 11 municipalities in the south-western part of Copenhagen were randomly selected from the Danish Civil Registration System. Overall, baseline examinations were attended by 6,784 (52%) individuals and genotype information was available for 6,184 individuals. The Inter99 study was approved by the Scientific Ethics Committee of the Capital Region of Denmark (KA98155) and registered as a clinical trial

(ClinicalTrials.gov; ID-no: NCT00289237). The study protocols were in accordance with the Helsinki declaration and approved by the local ethical committees.

The InCHIANTI study is a population-based epidemiological study aimed at evaluating the factors that influence mobility in the older population living in the Chianti region in Tuscany, Italy. The details of the study have been previously reported (Ferrucci et al., 2000). Briefly, 1616 residents were selected from the population registry of Greve in Chianti (a rural area: 11,709 residents with 19.3% of the population greater than 65 years of age), and Bagno a Ripoli (Antella village near Florence; 4,704 inhabitants, with 20.3% greater than 65 years of age). The participation rate was 90% (n=1453), and the subjects ranged between 21–102 years of age. The study protocol was approved by the Italian National Institute of Research and Care of Aging Institutional Review, the internal Review Board of the National Institute for Environmental Health Sciences (NIEHS) and by the Medstar Research Institute (Baltimore, MD).

The number of samples that were recruited and used for analysis after QC (see 2.2.5 and 2.2.7) in the discovery and replication stages is shown in Table 2.3 and Table 2.6 respectively. An overview of the two-stage study design and sample numbers is shown in Figure 2.1.

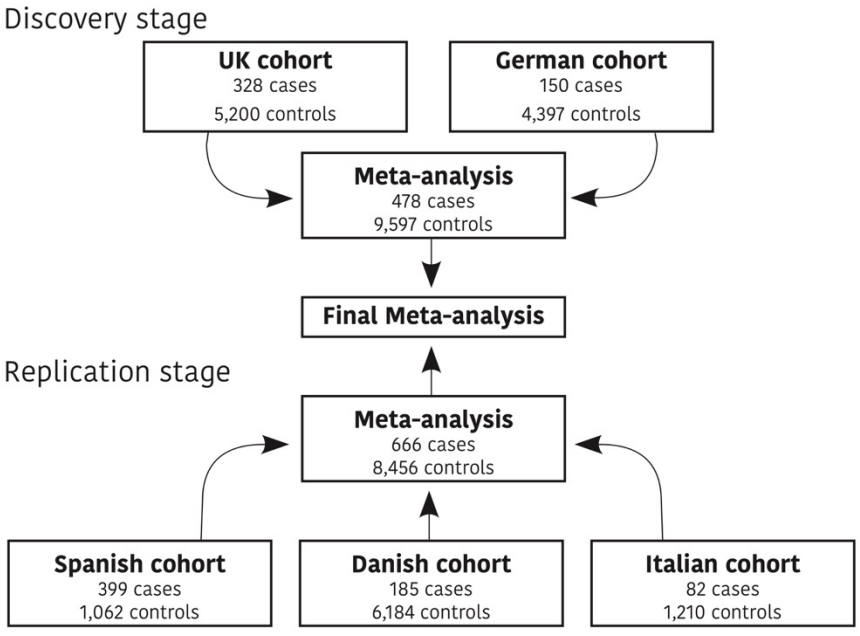


Figure 2.1 **Two-stage study design.**

An overview of the two-stage case control study design and sample numbers, before QC, that were used to investigate inherited predisposition to SM. In the discovery stage, SM patients and healthy controls from the UK and Germany were tested for association using binary logistic regression. Evidence from these separate cohorts was combined using a fixed-effect meta-analysis. SNPs selected for replication were tested in three European cohorts (Spanish,

Danish and Italian) using binary logistic regression. Another fixed-effects meta-analysis was used to determine the final effect size and significance levels by combining evidence from the discovery (stage 1) and replication stage (stage 2).

2.2.3 Genotyping

DNA was extracted from peripheral blood or bone marrow. The stage 1 cases were genotyped for 960,919 SNPs using Infinium OmniExpress exome chips (version 8_1.4_A1) and the Genome Studio software (GSGT Version 1.9.4) at the Clinical Research Facility in Edinburgh. These data are available on request from ArrayExpress (accession number E-MTAB-9358). The stage 2 cases (n=666) were genotyped for 92 SNPs, selected from the stage 1 analysis, using custom designed Kompetitive Allele Specific PCR (KASP) at LGC Genomics Limited (Hertfordshire, UK) (He et al., 2014). Briefly, KASP is a fluorescence resonant energy transfer (FRET) PCR based assay. Genotypic data for the control cohorts were obtained from published studies (Bosch, 2004; Ferrucci et al., 2000; Jørgensen et al., 2003; Julià et al., 2013; Pisinger et al., 2005; Tanaka et al., 2009).

For the WTCCC2 stage 1 controls, the NBS and BBC subsets were separately genotyped using the Illumina 1.2M Duo chips platform and Illumina's programme was used to call SNPs with a posterior probability >0.95 (Teo et al., 2007). The German controls from KORA_A (a subset of follow-up F3 of the population based survey KORA S3) were genotyped using Illumina human Omni chip (version 2.5-4v1_B) for 2,443,177. KORA_B controls (an independent subset of KORA S3/F3) were genotyped for 730,372 SNPs using Illumina human Omni express chips (version 12v1_H) (Holle et al., 2005).

Controls from SNDNAB, Inter99 and InCHIANTI were genotyped using Illumina Global Screening arrays, Illumina HumanOmniExpress-24 (versions 1.0A and 1.1A) and Illumina Infinium HumanHap 550K SNP arrays which include 18, 90 and 45 of the SNPs selected for replication respectively. Genotypes for the remaining SNPs were determined by imputation (Appendix Table A.6).

2.2.4 Imputation

Imputation of the discovery cohorts was used to increase SNP density and enable fine mapping around significant loci. SNPs were imputed using the Sanger imputation server (McCarthy et al., 2016) which used EAGLE2 for pre-phasing into the Haplotype Reference Consortium (HRC release 1.1), and positional Burrows-Wheeler transform (PBWT) for imputation. Imputed genotypes were quality controlled by excluding SNPs with info score <0.80, posterior genotype probabilities less

than 0.99, minor allele frequency less than 1%, greater than 10% missing genotypes or extreme deviation from HWE (P-value ≤ $1x10^{-10}$).

In the stage 2 control cohorts, genotypes for the remaining SNPs were determined by imputation. In brief, SNPs and/or samples were removed from SNDNAB due to a low call rate (<98%), significant deviation from HWE (P-value <0.0001), extreme heterozygosity (|F|>0.10) or evidence of second-degree relatedness (IBD>0.25). Genotypes for additional SNPs were obtained by imputation, which involved a two-step process. In the first step, the observed data were phased using SHAPEIT (version 2.r837). In the second step, the phased data were imputed using IMPUTE2 (version 2.3.0) with default settings, an effective population size (-Ne) of 20,000 which is recommended for achieving high accuracy across all population groups and reference haplotypes from phase 3 of the 1,000 Genomes Project (Auton et al., 2015). Imputation was performed in 5Mb chunks, as recommended, and then joined (Howie et al., 2009). Genotypes with an uncertainty greater than 0.1 were set to missing and the remainder were used as hard calls. SNPs with low imputation quality were excluded (INF0 score < 0.6).

Genotyping and QC of the InCHIANTI study has previously been described (Tanaka et al., 2009). In brief, SNPs and/or samples were removed due to low call rate (<97%), HWE (P-value <10^{-4}), heterozygosity (> 0.3), MAF (<1%) and sex mismatches, leaving 1,210 samples and 495,343 autosomal SNPs that passed quality control. SNPs were imputed using the Michigan Imputation Server, HRC haplotype reference panel (HRC r1.1 2016) and SNPs with low quality score were removed (INFO ≤0.7).

Genotyping and QC of the Inter99 study have previously been described (Graae et al., 2018). Individuals were genotyped using the Illumina HumanOmniExpress-24 SNP arrays (versions v1.0_A and v1.1_A) and the GenomeStudio software. QC filtering was applied before imputation, which involved selection of non-monomorphic SNPs, samples with a call rate ≥98%, and SNPs in HWE (P-value > 10^{-5}). Additional SNP genotypes were imputed using Eagle for pre-phasing autosomal SNPs and imputed to the Haplotype Reference Consortia panel (HRC version r1.1) by following the standard protocol on the Michigan imputation server (https://imputationserver.sph.umich.edu/index.html) (Das et al., 2016). All variants included in this study were in HWE (p > 0.05) and had high imputation quality scores (INF0≥0.9).

2.2.5 Quality control of the stage 1 data

Prior to analysis, the quality of the genotypic data was assessed and cleaned using standard QC procedures for GWAS (Anderson et al., 2010). Plink v1.90p was used to check genotype missingness (per sample and per SNP), MAF, HWE, sex mismatches, heterozygosity (Figure 2.3), cryptic relatedness, strand orientation and ancestry as detailed below (Chang et al., 2015). Duplicate markers are deliberately included in raw genotyping data to access the concordance rate of genotype calls for a specific array. Therefore, as an additional QC step, these duplicate SNPs were identified and removed prior to testing for association (Gogarten et al., 2012). Plink was used to merge datasets together and to flip those SNPs detected as not bi-allelic; this step ensures that strand orientation is concordant in each dataset. Strand assignment for palindromic SNPs (A/T-G/C) were checked and when necessary assigned to the correct strand using Genotype Harmonizer (GH) (Deelen et al., 2014). A manifest file for the Omni express exome chip (version 8_1.4_A1), developed by Will Rayner (Wellcome Centre for Human Genetics, University of Oxford), was used to update strand orientation, genomic location, SNP name and chromosome in the SM-1 dataset (Rayner and Mccarthy, 2011). In the KORA datasets, the SNP name was updated using the Illumina rsID-conversion file which is specific for each genotyping platform (KORA_A rsID Conversion File; KORA_B rsID Conversion File). The number of SNPs and samples removed by these QC measures in the stage 1 data is shown in Table 2.3 and Table 2.4.

2.2.5.1 Per-individual missingness

QC of the stage 1 genotypes involved the removal of samples with a large proportion of missing genotypes, which indicates poorly genotyped samples possible due to low quality DNA. Since GWAS aims to associate SNPs with disease, removing one marker might have a greater effect on the study than removing one individual (Smith and Newton-Cheh, 2009). This approach maximises the number of SNPs in the study and avoids removal of markers due to a subset of poorly genotyped individuals. For this reason, the QC on individual missingness was performed before the per-marker QC. Individuals with missing genotypes for 10% or more SNPs were excluded from the analysis. The proportion of missing genotypes per individual was determined using Plink and plotted in R Studio to visualise the distribution. For the unimputed KORA dataset per individual call rate ≥97% was applied by the KORA-study Group (Holle et al., 2005).

2.2.5.2 Per-SNP missingness

SNP-specific missingness rate is used to detect and exclude poorly genotyped SNPs, which could reduce the possibility of identifying a real association with the disease phenotype (Anderson et al., 2010). After poorly genotyped individuals were removed, the per-marker missingness QC were carried out, and SNPs with missing genotypes greater than 10% were detected and removed using

Plink (Chang et al., 2015). For the unimputed KORA dataset per SNP call rate ≥98% was applied by the KORA-study Group (Holle et al., 2005).

2.2.5.3 SNP minor allele frequency

Rare SNPs (MAF < 5%) can frequently produce false positive results due to small sample size and sampling errors. In general, for case-control GWAS with modest sample size a MAF threshold of 1–2% or higher in studies with smaller sample size is recommended (Anderson et al., 2010). In this step, SNPs with a MAF less than 5% were excluded both from cases and controls.

2.2.5.4 Hardy–Weinberg equilibrium

HWE states that there is a predictable relationship between allele and genotype frequencies under the assumptions of no mutation, random mating, no gene flow, infinite population size, and no selection. When these assumptions are met and case/control cohorts have been genotyped at the same time using the same genotyping array, SNPs with significant deviation from HWE (exact test P-value ≤0.001) in controls are indicative of genotyping error and should be removed from both cases and controls (Wigginton et al., 2005). However, since our cases and controls were genotyped separately, HWE was assessed separately in cases and controls. SNPs were excluded if they had modest deviation from HWE in controls (P-value <0.001) or extreme deviation in cases (P-value $\leq 1x10^{-10}$) which most likely reflects poor genotyping rather than disease association (Marees et al., 2018; Turner et al., 2011). A higher P-value threshold was used in cases because modest deviations from HWE might occur due to association with the disease while extreme deviations are most likely due to genotyping error (Affymetrix, 2011; Hammerschlag et al., 2017; Tapper et al., 2015). For the unimputed KORA dataset, HWE P-value < $1x10^{-10}$ filter was initially applied by the KORA-study Group (Holle et al., 2005).

2.2.5.5 Sex check

As a crude check of sample provenance and quality, and to avoid sex inconsistencies that could arise from data handling issues, the genotypic data was used to infer sex. Samples were removed if the inferred and reported sex were discordant. To infer sex the X chromosome homozygosity rate was calculated for each individual using Plink and plotted in R Studio to visualise the distribution (Chang et al., 2015). Male calls were made if the X-chromosome homozygosity was greater than 0.8 and female calls were made if it was below 0.2 (Figure 2.4). Individuals with discordant reported and inferred sex were removed (Anderson et al., 2010). Although males have one copy of the X chromosome they are not expected to have 100% homozygosity due to the pseudoautosomal regions (PARs). PARs are terminal regions of homology between chromosomes X

and Y which act like autosomes in the sense that they can recombine and contain both heterozygous and homozygous variants (Strachan and Read, 2011).

2.2.5.6 Sample heterozygosity

Another important QC step is to assess the evidence for DNA sample contamination or potential consanguinity using per sample heterozygosity. Excess heterozygosity is suggestive of DNA contamination or recent admixture, whereas deficiencies may indicate failed hybridisation, large chromosomal deletions or inbreeding. To identify samples with outlying levels of heterozygosity the autosomal heterozygosity rate (het_rate) per sample was calculated using the following formulae in Plink: het_rate=[N_HOM-N_NM]/N_NM), where N_HOM is the number of homozygous genotypes and N_NM is the total number of non-missing genotypes per sample. The heterozygosity rate for all samples versus the proportion of missing genotypes was plotted in R studio to visualise the distribution, and samples with mean heterozygosity values ± 3 standard deviations (SD) from the mean were excluded (Figure 2.3).

2.2.5.7 Approaches for data merging and strand orientation check

To carry out further QC, the case control datasets were merged despite significant challenges due to them being genotyped by different facilities using different SNP arrays. The issues involved in merging such datasets were highlighted by the electronic Medical Records and Genomics (eMERGE-I) Research Network which include: mismatched genotyping (strand forward or reverse orientation), the use of different SNP names and locations and errors introduced by the merging procedure, which have the potential of creating significant array or batch effects (Zuvich et al., 2011). Additional QC checkpoints were therefore used to address these issues. Firstly, SNP names and locations were updated in the SM-1 cohort using curated strand files for the respective SNP arrays that were downloaded from the McCarthy Group (Rayner and Mccarthy, 2011). To update the SNP name in KORA controls, an rsID-conversion file was downloaded from the Illumina website and used to convert the Illumina identifiers (kgp) to the corresponding rsID (KORA_A rsID Conversion File; KORA_B rsID Conversion File). After updating the SNP names and location, the case and controls datasets were merged using Plink and any mismatched strands for non AT/GC were detected as triallelic SNPs. The genotyping strand for these SNPs were corrected using the flip option in Plink or removed if unresolved (Chang et al., 2015). Despite these updates, SNPs with the same location but different names may still be identified when merging. To correct these 'same position' warnings, SNPs with the same location and alleles were combined using the '--merge-equal-pos' option in Plink.

2.2.5.8 Relatedness

Bias could be introduced if duplicates or related individuals are tested for association, as their genotype may be over-represented and the allele frequencies would not reflect the real population frequencies. If during recruitment of cases and controls some related individuals were inadvertently collected, checking for evidence of relatedness between samples is a standard QC procedure to ensure that duplicates, sample mix-ups, and related samples (first and second-degree relatives) are removed from the analysis. Pairwise values of genome-wide average identity by state (IBS), which describe the number of shared alleles between a pair of individuals, were therefore used to check for evidence of relatedness. To calculate IBS a set of autosomal SNPs in LD were selected using LD-based SNP pruning in Plink (Figure 2.2). SNPs in LD were selected using a maximum pairwise genotypic correlation (r^2 <0.5) within a window size of 50kb that was shifted in steps of 5 SNPs across the genome (Chang et al., 2015). SNPs in linkage equilibrium with a maximum pairwise genotypic correlation (r^2 threshold <0.5) were selected. SNPs selected after pruning were used to calculate genome-wide average IBS between each pairof individuals that passed QC. For sample pairs with evidence of relatedness (IBS >=0.86) the sample with the lowest genotyping rate for all SNPs passing QC was excluded (Burton et al., 2007).

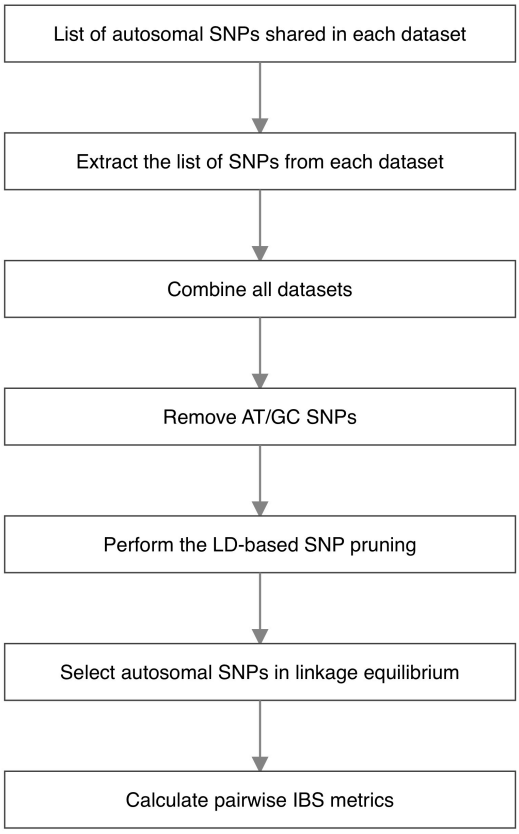


Figure 2.2 **Method to select independent SNPs for IBS metrics and multidimensional scaling.**

The flow diagram outlines all the steps performed to select autosomal SNPs in linkage equilibrium. The list of SNPs that are shared between all the datasets are listed and extracted in all datasets. The autosomal markers remaining after one round of LD-based SNP pruning (--indep-pairwise 50 5 0.5) were extracted from the merged dataset and used to calculate the pairwise IBS. Palindromic A/T and G/C SNPs were removed from the Hapmap data to facilitate combining these samples with the cases and controls.

2.2.5.9 Population stratification

In order to examine population stratification, infer ancestry and to check if the cases and controls form a homogeneous population, a multi-dimensional scaling analysis was performed using Plink. All merged datasets were combined with genotype data from the HapMap study that had already been quality controlled. The HapMap samples are from three reference populations consisting of 55 samples with ancestry from northern and western Europe (CEU) from the Centre d'Etude du Polymorphisme Humain (CEPH), 43 Han Chinese samples from Beijing, China (CHB) and 55 Yoruban samples from Ibadan, Nigeria (YRI). For analysis, a subset of uncorrelated markers (SNPs not in LD) were selected by LD pruning (Figure 2.2) and used to calculate a matrix of IBS values between all pairs of individuals. These pairwise IBS values were used as the input for multi-dimensional scaling analysis which generated five principal components. To examine the results and infer ethnicity, R Studio was used to make a scatter plot from the first two principal

components (C1 and C2; Figure 2.5). Samples with outlying values for C1 (±3 SD from the mean for stage 1 cases and controls and HapMap CEU) were considered ancestry outliers and excluded from further analysis.

2.2.6 Preliminary analysis of the stage 1 data

At this point of the QC, a preliminary case versus control analysis of the stage 1 data was performed and summarised using a quantile-quantile plot to determine if the test statistic was inflated and whether this could be related to problems with the merging process such as unresolved strand issues at AT/GC SNPs. After this test, the GH software was used to detect strand issues at AT/GC SNPs in the pre-merged controls based on differential LD patterns in comparison with the cases (Deelen et al., 2014). Strand mismatches were called by GH when the number of negative SNP correlations exceeded positive ones and these SNPs were then flipped in the control dataset using Plink. SNPs failing alignment were removed. Unknown strand assignment can also be addressed by comparing the MAF between datasets (Deelen et al., 2014). For further evaluation, the MAF difference between cases and controls was checked for AT/GC SNPs and SNPs with MAF difference greater than 0.34 were removed (Table 2.5). After correcting these mismatched AT/GC SNPs, the merging and preliminary case controls analysis were repeated.

2.2.7 Quality control of the stage 2 data

The same QC measures described in stage 1 were applied to the stage 2 cases, with the exceptions that per sample QC measures for heterozygosity, sex-mismatch, cryptic relatedness and non-Caucasian ancestry were not performed due to the small number of SNPs genotyped. In cases, QC was performed at the marker level only using per locus missingness whereby SNPs with greater than 10% missing genotypes were excluded (see 2.2.5.2) and SNPs with extreme deviation from HWE ($p<1\times10^{-10}$) were excluded (see 2.2.5.4). QC and imputation of the stage 2 controls has previously been described (Ferrucci et al., 2000; Jørgensen et al., 2003; Julià et al., 2013; Pisinger et al., 2005; Tanaka et al., 2009). The control datasets were obtained from a previous GWAS in collaboration with the SNDNAB, INCHIANTI and the University of Copenhagen who had performed their own QC and imputation (see 2.2.4). These data were further scrutinised using per locus missingness and HWE, and the number of samples and SNPs removed from the stage 2 data by these QC measures are shown in Table 2.6 and Table 2.7.

2.2.8 Statistical analysis

2.2.8.1 Genetic power calculation

The power to detect SNPs associated with SM was estimated using the Genetic Power Calculator (GPC) (Purcell et al., 2003) with the following parameters and assumptions. The sample size was determined by the number of cases and controls that passed QC both in stage 1 and stage 2. Although the incidence of SM is estimated to be approximately 1–9 in 100,000 (Coltoff and Mascarenhas, 2019) a minimum value of 1 in 10,000 had to be used. The controls were labelled as unselected as they had not been screened to confirm the absence of disease. To account for possibility of misclassified controls the power calculation assumed that a proportion of controls equal to the incidence may be misclassified. The genotyped SNPs were assumed to act via a multiplicative disease model and to be in linkage disequilibrium (D'= 1) with the causal variant. A range of minor allele frequencies (0.05, 0.1, 0.2, 0.3, 0.4) and effect sizes ($1.1 \leq OR \leq 2$ in 0.1 increments) were then used to estimate the power to detect genetic effects at a genome-wide level of significance (P-value $\leq 5\times10^{-8}$) (Figure 2.14).

2.2.8.2 Logistic regression model of association

After QC, the stage 1 data were tested for disease trait SNP association using binary logistic regression in Plink. Samples from the UK and Germany were tested as two separate populations, and samples with evidence of non-Caucasian ancestry were excluded rather than adjusting the association analysis for population stratification. A fixed-effects inverse variance-weighted meta-analysis was then used in Plink to combine evidence from the stage 1 cohorts (UK and Germany) and to determine the final effect sizes and significance levels by combining evidence across stages 1 and 2. To examine the effect of this decision, the ancestry outliers were retained, and the stage 1 analyses were repeated. In this second analysis, the first two principal components from the multi-dimensional scaling analysis were used as covariates in the logistic regression to account for the effect of population stratification (see 2.2.5.9, Figure 2.13 and Table 2.9).

To ensure that the separate and pooled analyses generated results that relate to the same risk allele, a file containing the minor allele in the pooled data was used to specify the risk allele in both the pooled and separate analyses. Results for SNPs that were only genotyped in one control population were obtained from the initial analysis, as a minimum of two cohorts are needed for meta-analysis. To examine the effectiveness of the QC measures and assess evidence for any systematic biases, the GWAS results from the stage 1 analysis of the UK and German cohorts and the stage 1 meta-analysis were visualised and interpreted using quantile-quantile plots (QQ plots) (Figure 2.7) and a Manhattan plot (Figure 2.8). The QQ plots were generated using a custom R script, and the qqman R Studio package was used to construct the Manhattan plot (Turner, 2018).

Results from the final meta-analysis of stages 1 and 2 were displayed in a forest plot using Stata (Figure 2.12). The FUMA software was used to generate regional plots of the stage 1 association results obtained with the imputed data (Watanabe et al., 2017). Heterogeneity of results in the meta-analysis was examined through the χ^2-based Cochrane's Q and I^2 statistics, which describe the percentage of variation across studies that is due to heterogeneity rather than chance.

2.2.8.3 Conditional analysis

Several *TERT* SNPs have been identified as risk factors for the development of haematological malignancies, including MPN (Tapper et al., 2015), as well as some solid tumours (Hung et al., 2019; Rafnar et al., 2009). Putative secondary signals were evaluated in Plink by performing conditional analysis on the index variant in the *TERT* locus. (Chang et al., 2015).

2.2.9 Clumping

To minimise false positives and the potential for overlooking signals with compelling functional evidence but modest significance, the following method was used to select SNPs for follow-up at stage 2. A clumping procedure was used to shortlist SNPs for follow-up at stage 2 using Plink software (Chang et al., 2015). For this analysis, results from the meta-analysis were used unless the SNP had been tested in one population only. Meta-analysis was prioritised since it favours SNPs that are significant in both populations, which reduces potential false positives and increases the likelihood of replication. The clumping procedure was used to identify clusters of correlated SNPs that contained at least one SNP with a P-value < 0.001 (P1). The most significant SNP within a clump is hereafter referred to as an index SNP. Clumps were formed by identifying all other SNPs in LD ($r^2 \geq 0.5$) and within 500kb from an index SNP. A greedy algorithm was used to construct these clumps so that each SNP could only appear in a single clump. Finally, index SNPs were only shortlisted for follow-up if the clump included at least one other correlated SNP with a P-value less than 0.01 (P2). This procedure ensured that only the most significant independent loci (index SNP with P<0.001) with supporting evidence from at least one correlated SNP ($r^2 \geq 0.5$, kb <500kb and P<0.01) were considered for follow-up at stage 2. This strategy and the parameters used are similar to those applied by previous studies (Chang et al., 2015; Tapper et al., 2015) and the default settings were used in Plink (Table 2.2). In relation to the default values, the P-value for selecting index SNPs was raised to 0.001 to account for the fairly modest sample sizes at stage 1, which limit study power. The distance between correlated SNPs was increased to 500 kb to accommodate long range LD and limit the number of shortlisted SNPs in close proximity to each other.

Table 2.2 **Clumping parameters in Plink.**

	P1	P2	r^2	Kb
Default parameters	0.0001	0.01	0.5	250
Applied parameters	0.001	Default	Default	500

The table shows the parameters used to determine the level of clumping: P1 = P-value threshold for the *index* SNPs; P2 = P-value threshold for the SNPs in the clumps; r^2 = LD threshold for clumping; Kb is the physical distance in kilobases from the index SNP for clumping. The first row shows the parameters applied in Plink by default, the second row shows the parameters applied to determine clumps for the GWAS analysis.

2.2.10 Functional annotation and criteria for SNP selection

Following the clumping procedure, gene-based annotation of all the index SNPs eligible for replication was performed using ANNOVAR (Wang et al., 2010). The list of the nearest genes was submitted to GeneAlaCart, a tool that extracts information from the GeneCards database to generate a spreadsheet containing all the functional annotations associated with the list of genes (Stelzer et al., 2011). Genes were retained if their biological function from GeneAlacart was related to kinase activity (Receptor Tyrosine Kinase (RTK) or KIT), haematopoiesis, myeloid leukaemia, or myeloproliferative or MC conditions such as mastocytosis (Appendix Table A.2). To minimise false positives and the potential for overlooking signals with compelling functional evidence but modest significance, the following method was used to select 92 index SNPs for follow-up at stage 2. First index SNPs that according to annotation from GeneAlacart (Stelzer et al., 2011) were located within or adjacent to a gene with functional relevance were given priority. The number of selected SNPs was then infilled to 82 by selecting the remaining most significant index SNPs. To add support and to guard against failed or problematic genotyping, additional SNPs were selected as backups for each of the most promising index SNPs in terms of either their biological relevance, individual significance or level of support from correlated SNPs.

2.2.11 Identification of clonal mosaicism using BAF segmentation

DNA from SM patients was extracted from peripheral blood leukocytes, which are expected to consist of a mixture of clonal and non-clonal cells. To assess the frequency of somatic changes, which could affect the association analysis, BAF segmentation was therefore used to analyse all of the stage 1 cases and to identify genomic regions of AI that were subsequently categorised as either aUPD, copy number gains or losses using a separate script. Since this analysis requires genome-wide data, BAF segmentation was only applied to the SM patients from stage 1 (n=479) (Staaf et al., 2008).

Briefly, to identify AI regions using BAF segmentation, non-informative markers with BAF less than 0.1 or greater than 0.9 were excluded and the remaining BAF values were mirrored at 0.5 to give mirrored BAF values (mBAF). The data were further cleaned using triplet filtering to remove SNPs where the absolute difference between preceding or succeeding SNPs was greater than 0.6. Finally, circular binary segmentation (CBS) was used to identify regions with similar mBAF values that were classified as a region of AI if the mean mBAF value was greater than 0.56.

A custom script was then used to categorise the AI regions as likely aUPD if the region was greater than 2Mb in length, extended to the telomere and had a neutral copy number (LRR between −0.15 to 0.065) (68). AI regions greater than 2Mb were classified as a copy number gains if LRR was greater than 0.073 or loss if LRR was less than −0.14 (Staaf et al., 2008). An automated method was used to extract regions of AI involving *KIT* (hg19 chr4:55,524,095 – 55,606,881). Acquired UPDs tend to be greater than 1Mb in size and extend to the telomere, and we used a custom program to identify telomeric AI regions. There are numerous interstitial regions of AI which may be interesting if they overlap in multiple samples. Furthermore, these regions may help to narrow down large candidate regions of aUPD that extend to the telomere. After identifying telomeric AI regions, an automated method was used to detect internal AI regions greater than 3Mb from regions that passed QC. To identify minimal recurrent regions, internal AI regions were converted to bed files and intersected using bedtools. AI regions overlapping in 3 or more samples were selected and added to the ideogram used to examine and visualise the regions of AI. In the scatterplot (Figure 2.9), per sample metrics for the total number of AI regions and percentage of the autosome consisting of AI regions were used to make a scatter plot and to identify any sample outliers. To calculate the autosomal AI percentage, the length of the autosome was defined by the Illumina Infinium OmniExpress exome chip that was used to genotype the SM cases which came to 2.792GB (Appendix Table A.4). In the ideogram (Figure 2.11), regions of aUPD were plotted on a chromosome ideogram to identify recurrent regions.

2.2.12 Replication and final meta-analysis

The replication data included 666 SM patients from Spain, Denmark and Italy and 8,456 controls (Figure 2.1). Logistic regression, as described in Section 2.2.8.2, was used to test the SNPs that were selected for replication and passed QC. To determine the final significance and effect size, a fixed effects meta-analysis was used to combine the evidence from stages 1 and 2. Regional plots

of the stage 1 imputed data were generated using FUMA (Watanabe et al., 2017) to investigate the candidate region surrounding SNPs that reached genome-wide significance in the final meta-analysis.

2.3 Results

2.3.1 Quality control of cases at stage 1

Following QC, a total of 39 patients (Table 2.3) were removed from the SM-1 cohort due to these samples having either more than 10% missing genotypes (n=19, Figure 2.3), mismatches between inferred and reported gender (n=2, Figure 2.4), autosomal heterozygosity exceeding +/- 3SD from the mean (n=9, Figure 2.3) or evidence of cryptic relatedness (IBS>=0.86, n=9 UK). Before performing the heterozygosity check, 1,699 non-autosomes (X, Y, mitochondrial chromosomes and pseudo-autosomal region of X) were removed. Both mean (0.29) and SD (0.0056) calculations were based on 743,882 autosomal variants scanned in Plink. Sex was inferred using X-chromosome homozygosity and two samples with a mismatch between the inferred and reported sex were removed. For subsequent analyses, inferred sex was used for samples where the reported sex was unknown (n=9).

At the marker level, a total of 368,912 SNPs were removed from the SM-1 dataset during QC (Table 2.4). These SNPs include those with more than 10% missing genotypes (n=1,725); MAF less than 0.05 (n=340,313), extreme deviation from HWE ($p<1\times10^{-10}$, n=240), duplicate markers (n=14,939), not bi-allelic SNPs (n=4) and SNPs failing genotyping (n=3). After these QC measures were applied, 592,007 SNPs (449,874 in the UK and 583,528 in the German cohort) and 414 cases remained for further analysis (Table 2.3).

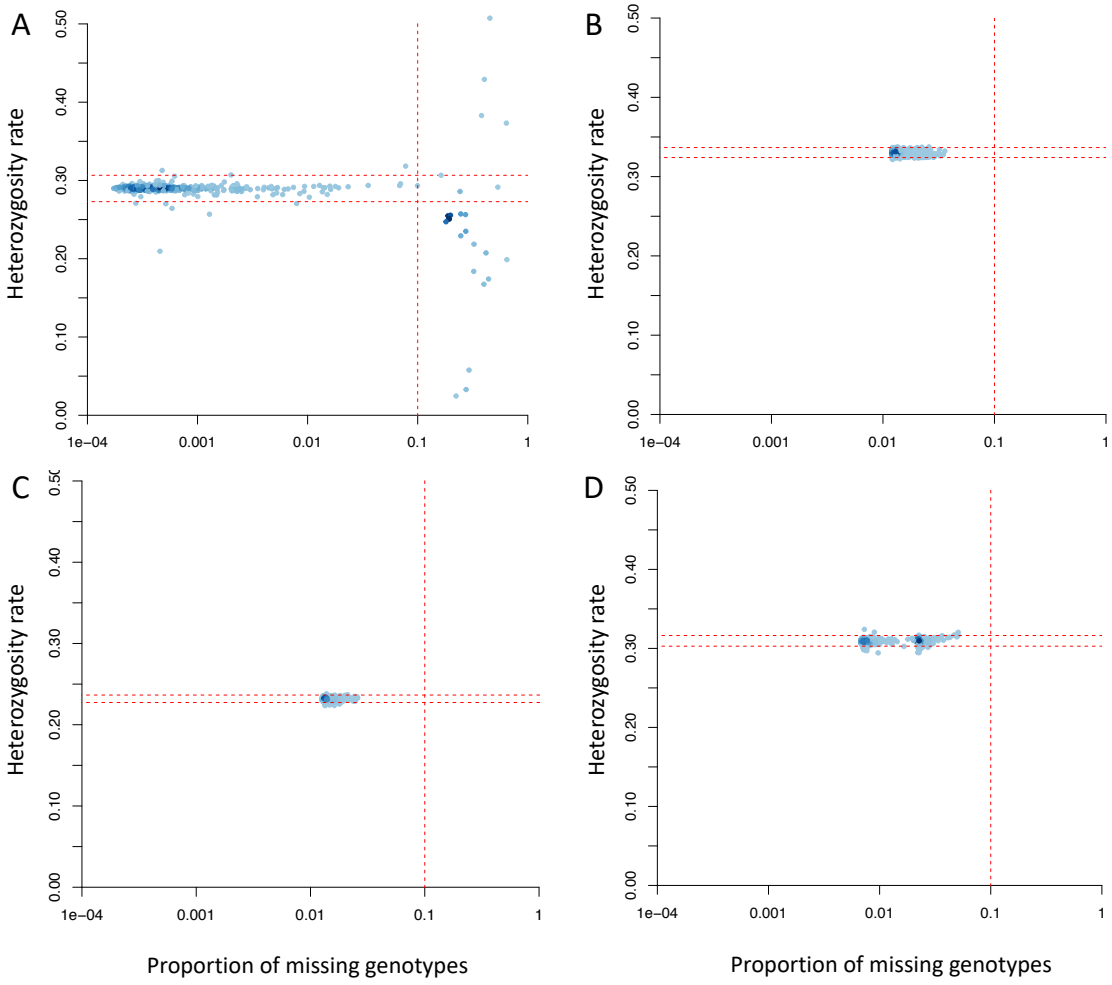


Figure 2.3 **Quality control for autosomal heterozygosity and per sample missingness.**

Horizontal dashed lines indicate the thresholds used to identify samples with outlying levels of heterozygosity in the stage 1 SM patients (±3 SD from the mean). Vertical dashed lines show the threshold used to remove samples with more than 10% missing genotypes. **A.** SM-1 patients from the UK and German cohorts. The upper dashed line corresponds to 0.30 (het mean +3SD), the lower one corresponds to 0.27 (het mean −3SD). **B.** Healthy controls from the WTCCC2 cohort. The upper dashed line corresponds to 0.34 (het mean +3SD), the lower red line corresponds to 0.33 (het mean −3SD). **C.** Healthy controls from the KORA_A cohort. The upper dashed line corresponds to 0.24 (het mean +3SD), the lower red line corresponds to 0.23 (het mean −3SD). **D**. Healthy controls from the KORA_B cohort. The upper dashed line corresponds to 0.24 (het mean +3SD), the lower red line corresponds to 0.23 (het mean −3SD).

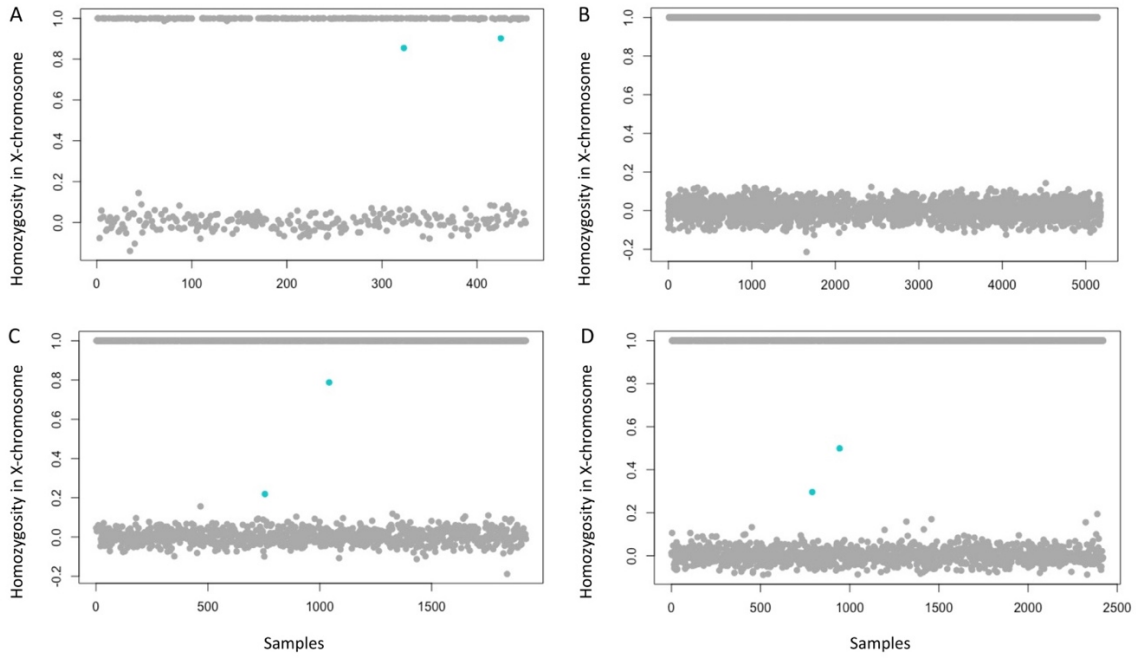


Figure 2.4 **Sex inference based on X chromosome homozygosity.**

The expected homozygous rates are greater than 0.8 for males and less than 0.2 for females (Anderson et al., 2010). Samples highlighted in light blue were removed because the inferred and reported sex were not concordant. **A.** In the SM-1 cohort 2 samples were removed. **B.** No discordance between reported and inferred sex was identified in the healthy controls from the WTCCC2 cohort. **C.** KORA_A: Two samples with X chromosome homozygosity rate between 0.8 and 0.2 were removed; **D.** KORA_B: Sex inconsistency was identified in two samples and these were removed.

2.3.2 Quality control in control datasets at stage 1

Although QC had already been applied to the genotypic data for controls from the WTCCC2 and KORA cohorts, they were tested again using our own QC thresholds. The second round of QC removed 84 individuals (Table 2.3) due to autosomal heterozygosity exceeding +/- 3SD from the mean (n=23 WTCCC2, n=17 KORA_A, n=38 KORA_B; Figure 2.3), evidence of relatedness (IBS>=0.86, n=1 WTCCC2, n=1 KORA_A) or mismatches between inferred and reported gender (n=2 KORA_A, n=2 KORA_B; Figure 2.4). To perform the heterozygosity check, non-autosome SNPs (X chromosome, Y chromosome, pseudo-autosomal region of X, mitochondrial chromosome) were removed (WTCCC2=40,355, KORA_A=49,888, KORA_B=16,833). Both mean (WTCC2=0.33, KORA_A=0.23, KORA_B=0.31) and SD (WTCCC2=0.0021, KORA_A=0.0015, KORA_B=0.0022) calculations were based on autosomal variants (WTCCC2=887,903, KORA_A=1,846,164, KORA_B=656,562) scanned in Plink.

At the marker level, a total of 504,270 SNPs were removed from WTCCC, 1,813,477 in KORA_A and 155,872 in KORA_B following specific exclusion filters (Table 2.4).

Table 2.3 **Sample sizes before and after quality control in stage 1.**

	Stage 1 cases		Stage 1 controls		
Quality control measure	UK	Germany	WTCCC	KORA A	KORA B
Total samples pre-QC	329	150	5200	1938	2459
≥10% missing genotypes	18	1	0	0	0
Patients with outlying heterozygosity 3SD	5	4	23	17	38
Patients with gender mismatch	2	0	0	2	2
Patients with relatedness	9	0	1	1	0
Ancestry outliers	21	5	0	5	4
Samples remaining	274	140	5176	4328	

After sample QC, 414 cases remained at stage 1. The 26 ancestry outliers were retained when the stage 1 analyses were repeated with adjustment for population stratification. QC: quality control, SD: standard deviation.

Table 2.4 **SNP number before and after quality control in stage 1.**

	Stage 1 cases		Stage 1 controls		
Quality control measure	UK	Germany	WTCCC	KORA A	KORA B
Total observed SNPs pre-QC	960919		954144	2380310	721694
SNPs failed genotyping	3		0	0	0
SNPs with ≥10% missing genotypes	1725		24263	29469	19016
SNPs with MAF ≤ 5%	340313		71631	1085092	120853
SNPs failing HWE*	240		3598	2376	1250
Not bi-allelic SNPs	4		3	4	0
Unknown strand	0		372	439	291
Duplicates/triplicates	14939		2	4378	1
MAF difference >0.34	0		7	1	1
Not in cases and controls	153821	20167	404394	691718	14460
Total observed SNPs passing QC	449874	583528	449874	583528	
Imputed SNPs with info score >0.8 and MAF>0.01	7397922	7253056	7397922	7253056	
HWE*	200212	195134	200212	195134	
Duplicates	5816	5396	5816	5396	
Total imputed and observed SNP remaining	7191894	7052526	7191894	7052526	

In total 592,007 SNPs were tested at stage 1. Of these, 441,395 were tested in both the UK and Germany cohorts, 8,479 were tested in the UK only, and 142,133 were tested in the German cohort only. QC: quality control, MAF: minor allele frequency, HWE: Hardy-Weinberg equilibrium. *HWE P-value <1×10-10 in cases, P-value <0.001 in controls.

2.3.3 Merging of cases and controls

To aid merging a strand file for the Illumina Infinium OmniExpress exome chip was downloaded from Will Rayner's website (https://www.well.ox.ac.uk/~wrayner/strand/) and used to update the chromosome and genomic locations for 603,592 SNPs and to flip the genotyping strand for 301,568 SNPs. An rsID conversion file was also downloaded and used to update 603,319 SNP names. Following these measures, the cases and controls were merged and any non AT-GC SNPs that generated two or more alleles were detected and flipped (n=707 in WTCCC2, n=163 in KORA_A, n=171 in KORA_B). During merging, SNPs with 'same position' warnings were detected in KORA_A (n=1,273) and KORA_B (n=10) and resolved using the '--merge-equal-pos' option in Plink.

2.3.4 Relatedness and population stratification

Multidimensional scaling analysis was performed to assess the evidence for population substructure, which can generate false positive and false negative results. For this analysis the 440 stage 1 patients, 9,513 controls (n=5,176 WTCCC2, n=4337 KORA) that passed QC were used, and 153 individuals from HapMap (n=55 CEU, n=43 CHB, n=55 YRI). A total of 331,793 SNPs present in each dataset were extracted from the merged dataset, and 150,381 variants were removed using LD-based SNP pruning (Figure 2.2). Pairwise measures of IBS were then determined using 181,411 autosomal SNPs in linkage equilibrium. Based on these IBS measures, during QC we removed 11 samples (cases=9, controls=2) with evidence for cryptic relatedness (IBS>0.86) (Table 2.3). A multidimensional scaling analysis was also performed using the pairwise measures of IBS. When plotting the first and second components from multidimensional scaling, most cases and controls formed a single cluster overlapping with the Caucasian reference population from HapMap (Figure 2.5 A). The close ancestral relationship between most cases and controls suggests they have European ancestry and are suitable for comparison. However, there was evidence of non-Caucasian ancestry in 26 cases (21 UK, 5 German) and 9 KORA controls. The mean of the C1 values of the European groups (WTCCC2, KORA, SM-1, CEU) was calculated and samples with ±3 SD (0.0023) or more from the mean (0.0013) were considered ancestry outliers in further analysis.

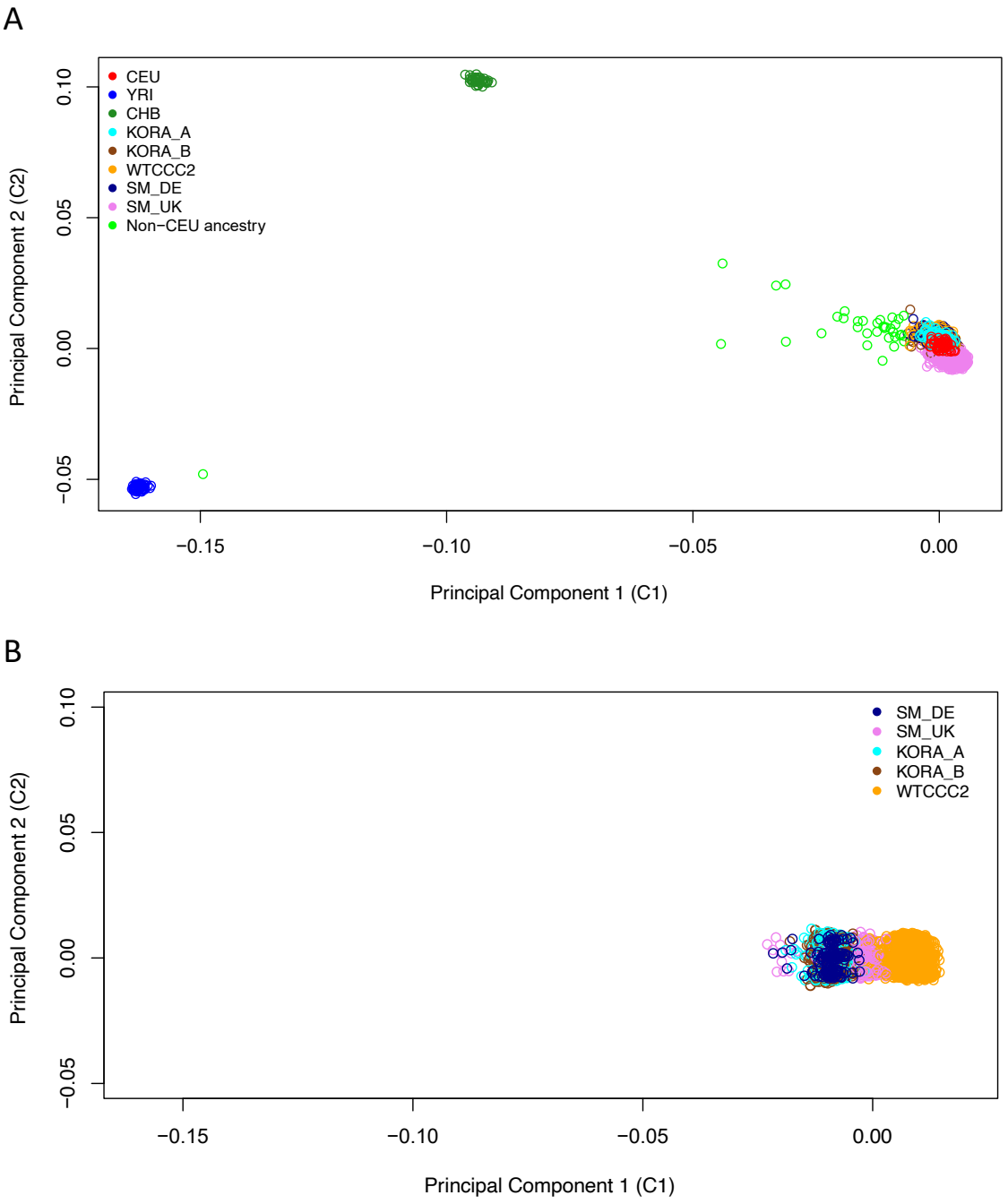


Figure 2.5 **Multidimensional scaling plots.**

Multidimensional scaling plot generated by plotting the first two components (C1 and C2). A. SM patients from the UK (pink circles) and Germany (dark blue circles), KORA controls (KORA_A turquoise, KORA_B brown) and WTCCC2 controls (orange), reference populations from HapMap for Utah residents with Northern and Western European ancestry (CEU, red circles), Yoruban individuals from Ibadan, Nigeria (YRI, blue circles), Han Chinese in Beijing, China (dark green circles). Samples with outlying values for C1 (±3 SD from the mean for stage 1 cases and controls and HapMap CEU) were considered ancestry outliers and excluded from further analysis (light green circles). B. The MDS plot is showing C1 and C2 components for patients (pink and dark blue circles) and controls (turquoise, brown and orange circles).

Multidimensional scaling analysis was again performed on the study cohorts only (stage-1 cases=440, controls=9,513) to assess substructure of Caucasian population, the reference populations from HapMap were excluded from the analysis. A total of 151,907 SNPs were removed after LD-based SNP pruning and the remaining genotype data (n=180,332 SNPs) were used to calculate a genome-wide pairwise IBS distance matrix and to perform the multidimensional scaling analysis. As shown in Figure 2.5 B, one major cluster was identified. This shows that the small population substructure in this study should not have appreciable effect on the final results. It is worth noting that WTCCC2 controls are shifted slightly to the right with limited overlap with the UK cases; also, the overlap between German cases and controls suggests some point of difference between UK cases and controls, which can be due to several factors, such as residual QC issue or different genotyping chips.

2.3.5 Preliminary analysis of the stage 1 data

To assess the merging process, a preliminary analysis of the stage 1 data was performed as a pooled analysis that tested the UK and German cohorts as a single European cohort. Logistic regression was used to compare the pooled set of cases (n=440) and controls (n=9,513) for all the SNPs that passed the initial QC. For this analysis, the ancestry outliers were retained and the first five principal components from the multi-dimensional scaling were used to correct for population stratification, and a QQ plot was used to inspect the results. Although the QQ plot showed no evidence for systematic biases between the cases and controls (genomic inflation factor λ=0.96) there were 174 SNPs (Appendix Table A.1) that reached genome-wide significance (P-value < 5×10^{-8}), which is more than expected given the modest sample size and estimated study power (Figure 2.14). To investigate further, these significant SNPs were stratified by their alleles, which showed that 77% had palindromic alleles, either AT/TA (20.1%) or GC/CG (56.9%). This is more than expected given the proportion of AT/GC SNPs that were tested (Figure 2.6), and suggests that unresolved strand issues at palindromic SNPs may account for the excess of significant SNPs.

Following this observation, the GH program was used to assess the evidence for strand mismatches at AT/GC SNPs by comparing the LD pattern between cases and controls (Deelen et al., 2014). This analysis identified 853 palindromic AT/GC SNPs with potential strand issues (n=687 in WTCCC2, n=159 in KORA_A, n=7 in KORA_B) that were flipped using Plink (Table 2.4). Strand assignments could not be resolved for 1,102 AT/GC SNPs (n=372 in WTCCC2, n=439 in KORA_A, n=291 in KORA_B) because of a lack of SNPs that are in LD in the surrounding area, and so these ambiguous SNPs were removed (Table 2.4). AT/GC SNPs were evaluated further, SNPs with MAF

difference > 0.34 between cases versus controls were identified as outliers and removed from further analysis (Table 2.4 and Table 2.5).

Table 2.5 **SNPs with highest MAF differences between case and control datasets.**

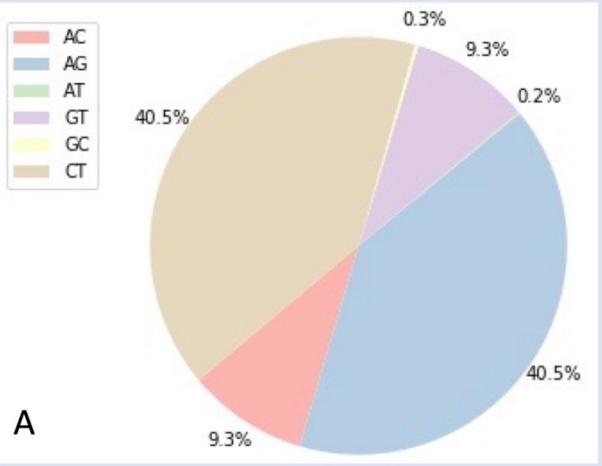
CHR	SNP	BP	MAF_A	MAF_U	MAF difference	A1	A2	Ctrl Dataset
6	rs6553229	153316274	0.1036	0.8934	0.7898	G	C	WTCCC2
6	rs6553229	153316274	0.1036	0.9049	0.8013	G	C	KORA_A
6	rs6553229	153316274	0.1036	0.90257	0.79897	G	C	KORA_B
13	rs10507391	31312097	0.3244	0.674	0.3496	A	T	WTCCC2
23	rs28861531	1374728	0.1127	0.8682	0.7555	G	C	WTCCC2
23	rs17881232	1464821	0.255	0.7248	0.4698	C	G	WTCCC2
23	rs17808080	2591888	0.1824	0.7715	0.5891	T	A	WTCCC2
23	rs731477	155228954	0.1167	0.8792	0.7625	G	C	WTCCC2
23	rs731478	155229100	0.1183	0.8789	0.7606	G	C	WTCCC2

CHR: chromosome; SNP: SNP identifier; BP: base pair; MAF_A: minor allele frequency in affected individuals; MAF_U: minor allele frequency in unaffected individuals; A1: alternative or minor allele; A2: reference allele.

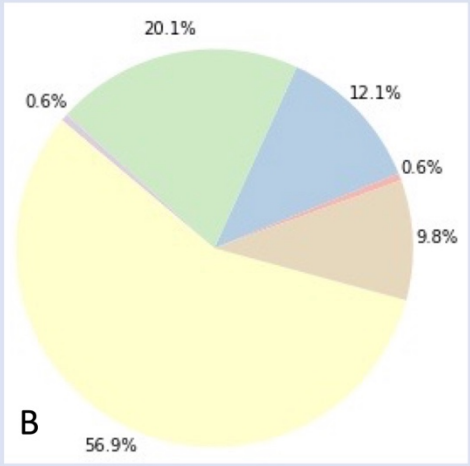
After processing the AT/GC SNPs, the preliminary analysis was repeated and resulted in only one SNP with genome-wide significance. The significance threshold was therefore reduced to P-value $<10^{-4}$ and the SNPs reaching this level of significance were stratified by their alleles. This showed that the proportion of significant SNPs by allele were similar to those in the total tested, suggesting that the strand issues have been resolved.

Preliminary analysis

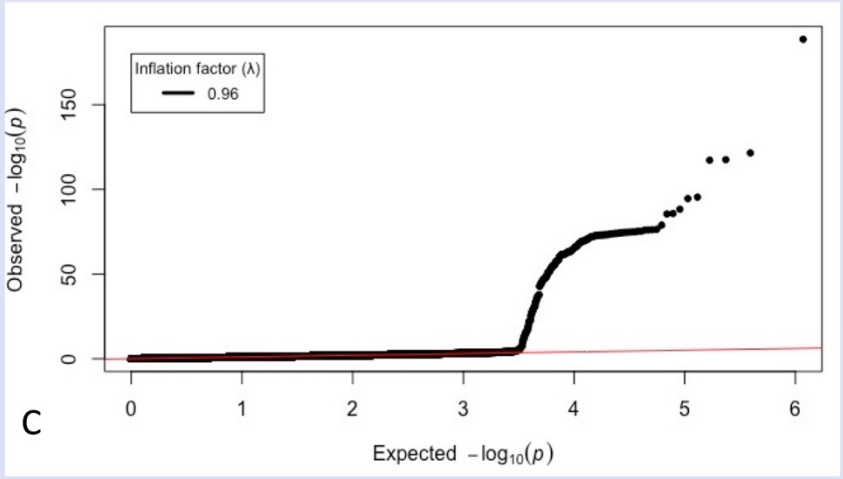
Proportion of total tested SNPs by allele



Proportion of significant SNPs by allele (p-value < 5 x 10^{-8})

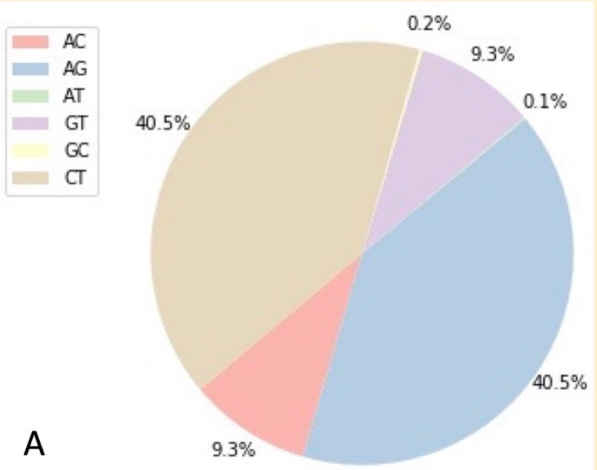


QQ-plot pre-AT/GC QC

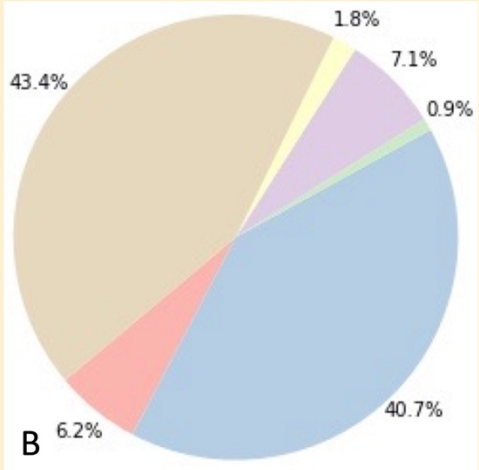


Final analysis

Proportion of total tested SNPs by allele



Proportion of significant SNPs by allele (p-value < 1 x 10^{-4})



QQ-plot post-AT/GC QC (GH)

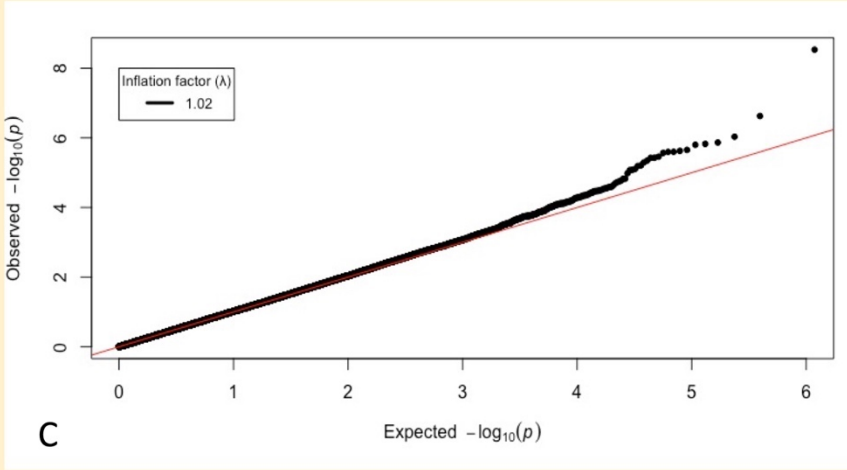


Figure 2.6 **Effect of the strand orientation QC on the association analysis results.**

This figure shows the effect of the strand orientation QC for reducing type 1 errors when testing for association. In the preliminary analysis (top of the figure), the allelic proportion of significant SNPs (B) was highly different from the allelic proportion of total tested SNPs (A) and when strand orientation QC for palindromic SNPs was not applied, the QQ-plot (C) show an early and abrupt deviation of the test statistics from the null hypothesis. In comparison, in the final analysis (bottom of the figure) performed after addressing the additional strand orientation QC for palindromic SNPs, the allelic proportion was very similar between significant SNPs (B) and total number of tested SNPs (A); as a result, the QQ-plot (C) shows instead good agreement between observed and expected P-values until SNPs with P-values $<10^{-4}$ beginning to show modest deviation from the null distribution.

2.3.6 Logistic Regression

In the preliminary analysis, logistic regression was used to compare all of the cases (n=440) and controls (n=9,513) and test all the SNPs that passed QC. In this analysis the first five principal components from the multidimensional scaling analysis were used as covariates to correct for population stratification.

After quality control of the stage 1 data and after resolving the residual strand issues for AT/GC SNPs, binary logistic regression was used to test the stage 1 data as two separate populations from the UK and Germany. In this analysis, 35 ancestry outliers were removed (UK = 21, German = 5, KORA = 9) before testing the UK (274 cases versus 5176 controls) and German (140 cases versus 4328) populations (Table 2.3). A fixed effects meta-analysis was then used to combine summary statistics from the separate analyses of the UK and German cohorts. Results from the meta-analysis are available at LocusZoom (http://locuszoom.org/) under "Mastocytosis GWAS" (Pruim et al., 2011).

At stage 1, a total of 592,007 SNPs were tested for association with *KIT*^{D816V} positive mastocytosis. Of these, 441,395 were tested in both the UK and Germany. An additional 150,703 SNPs were not genotyped in both control populations and could not be combined by the meta-analysis, which needs a minimum of two cohorts. Of these SNPs 8,479 were tested in the UK only, and 142,133 were tested in the German cohort only (Table 2.4). The quantile-quantile (QQ) plots for each analysis and their low genomic inflation factors ($\lambda \leq 1.038$) demonstrate a close agreement with the null hypothesis up to the tail of the distribution, where SNPs with *P*-values less than 10^{-4} become more significant than expected by chance alone (Figure 2.7).

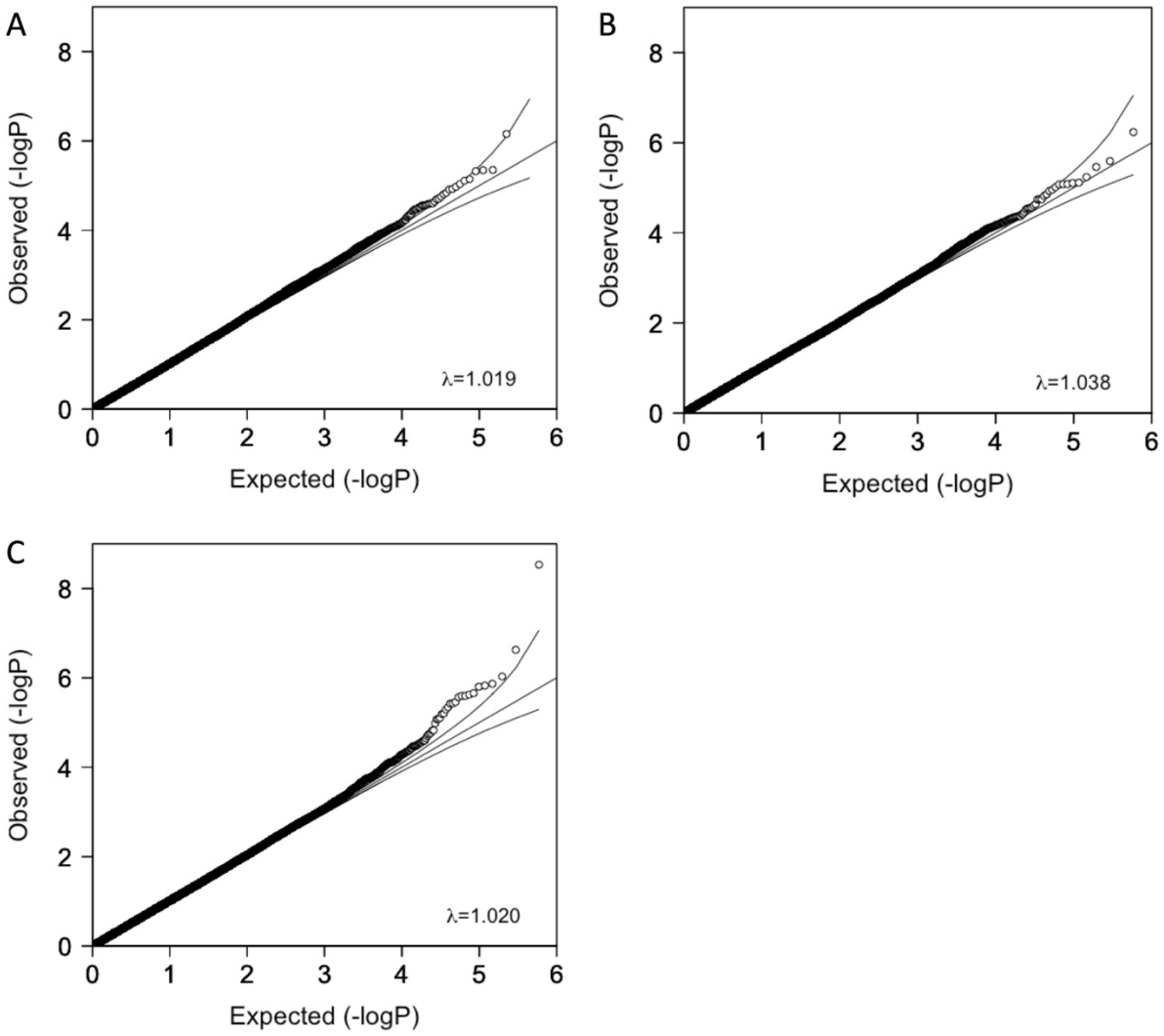


Figure 2.7 **QQ plots of P-values from the stage 1 analyses.**

The QQ plots of the observed versus expected P-values when testing the association with SM at stage 1 for separate analysis of the UK (A) and German (B) cohorts and meta-analysis (C). The black diagonal indicates expected QQ plot under null hypothesis when no SNPs are associated with SM. The area between the curved lines represents the 95% confidence interval (CI) of the expected P-values on the plot. The −log P-values are mostly within the 95% CI until SNPs with P-values $<10^{-4}$ start deviating from the levels of significance that are expected by chance alone (C). The −log P-values of UK (A) and German (B) analysis are mostly within the 95% CI.

A Manhattan plot summarising the results of the stage 1 meta-analysis is shown in Figure 2.8. A total of 18 SNPs were identified with the less stringent threshold of suggestive significance (P-value < 1×10^{-5}). The Manhattan plots showed that there were several peaks of significant SNPs with support from nearby SNPs, most notably on chromosomes 2, 3, 4 and 11. However, only one SNP, on chromosome X, surpassed the genome-wide level of significance.

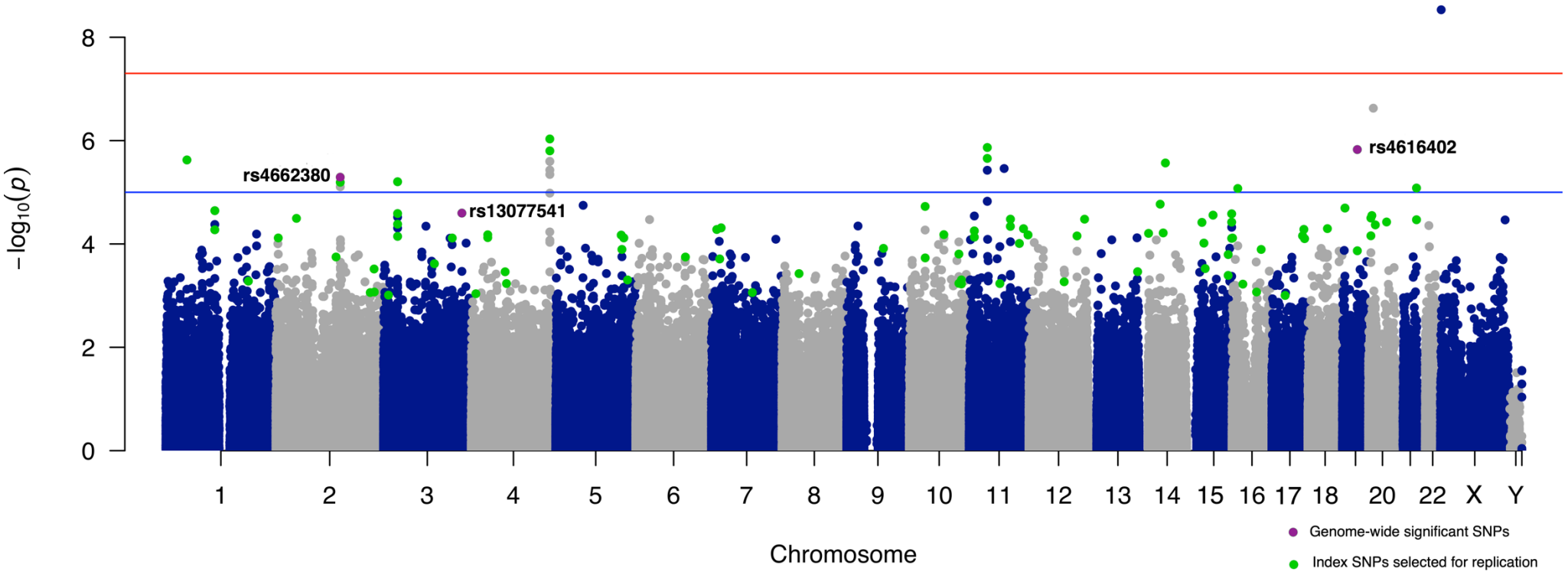


Figure 2.8 **Manhattan plot.**

Manhattan plots representing results of stage 1 meta-analysis GWAS for all 24 chromosomes. In total 592,007 SNPs are plotted, and each typed SNP is shown in alternate blue and grey. Results are plotted as −log10 of the meta-analysis P-values on the y-axis against their physical chromosomal position on the x-axis. Horizontal lines were added to indicate the threshold for genome-wide significance (P-value $<5\times10^{-8}$, red line) and a suggestive level of significance (P-value $<1\times10^{-5}$, blue line). All the 92 SNPs selected for follow-up are highlighted in green and the three SNPs that reached genome-wide significance after meta-analysis of stages one and two are highlighted in purple. One SNP meets the genome-wide significance level and a further 18 SNPs were identified with suggestive P-values. Consistent signals are shown on chromosomes 2, 3, 4 and 11.

2.3.7 Clumping

A clumping procedure was used to select the most promising SNPs for replication analysis at stage 2. Results for 441,395 SNPs obtained from the meta-analysis were used, as they will help to select SNPs with a similar trend in both the UK and German cohorts that are more likely to be replicated. A limited number of index SNPs (n=79) were identified using the default parameters in Plink (Chang et al., 2015). For this reason, less stringent parameters were used, and identified a total of 441 index SNPs with a P-value less than 0.001 and support from at least one correlated SNP (r^2<0.5) with a P-value less than 0.01 and within 500 Kb of the *index* SNP.

2.3.8 Functional annotation and selection of SNPs for replication

A gene-based annotation of the 441 index SNPs was submitted to ANNOVAR and a list of 560 genes was generated (Wang et al., 2010). The list of genes was submitted to GeneAlaCart and reduced to 50 genes with biological relevance, which include kinase activity (receptor tyrosine kinase (RET) or KIT), haematopoiesis, myeloid leukaemia, myeloproliferative or MC conditions such as mastocytosis (Appendix Table A.2) (Stelzer et al., 2011). It is plausible to speculate that the strategy applied to a shortlist of only 50 genes, may have overlooked some interesting signals. As discussed in Chapter 5, this can be due to missing knowledge at the time the analysis was performed. The following criteria were then used to select SNPs for replication at stage 2. First, 44 index SNPs were selected located in or flanked by a functionally relevant gene, with a moderate significance threshold (P≤0.001) and supported from correlated SNPs (Appendix Table A.6). The list of selected SNPs was then infilled to 82 by selecting the 38 most significant index SNPs and with support from correlated SNPs (Appendix Table A.6). To add support and to guard against failed or problematic genotyping, 10 additional SNPs were selected as backups for each of the most promising index SNPs in terms of either their biological relevance, individual significance, or level of support from correlated SNPs (Appendix Table A.6). After these selection criteria were applied, a total of 92 SNPs were selected for replication. To ensure that no interesting association signals were overlooked, the shortlisted SNPs were highlighted in a final Manhattan plot (Figure 2.8). One SNP achieved genome-wide significance in the stage 1 analysis, rs7884433, but it was not selected for replication because it lacked support from any of the SNPs in strong LD (r^2>0.8) and is thus likely to be a technical artefact.

2.3.9 Identification of clonal mosaicism using BAF segmentation

To assess the frequency of somatic changes, which could affect the association analysis, BAF segmentation (Staaf et al., 2008) was used to analyse all of the stage 1 cases (n=478) and to identify genomic regions of AI. The raw output from this analysis includes a text file that lists all the AI regions that were detected in each sample (Appendix Table A.3). To examine the raw output from BAF segmentation and to identify any sample outliers, the total number of AI regions and the percentage of autosomal AI regions in each sample was determined and plotted (Figure 2.9). Visual inspection of the plot identified 24 outlying samples that had either 95% or more of their autosome being called as regions of AI (n=21) or more than 3,000 separate regions of AI (n=3). During the QC steps, 19 of these outliers were found to have more than 10% missing genotype. Therefore, of the 414 individuals tested for association with mastocytosis, five other samples were excluded for the BAF segmentation analysis (Appendix Table A.5).

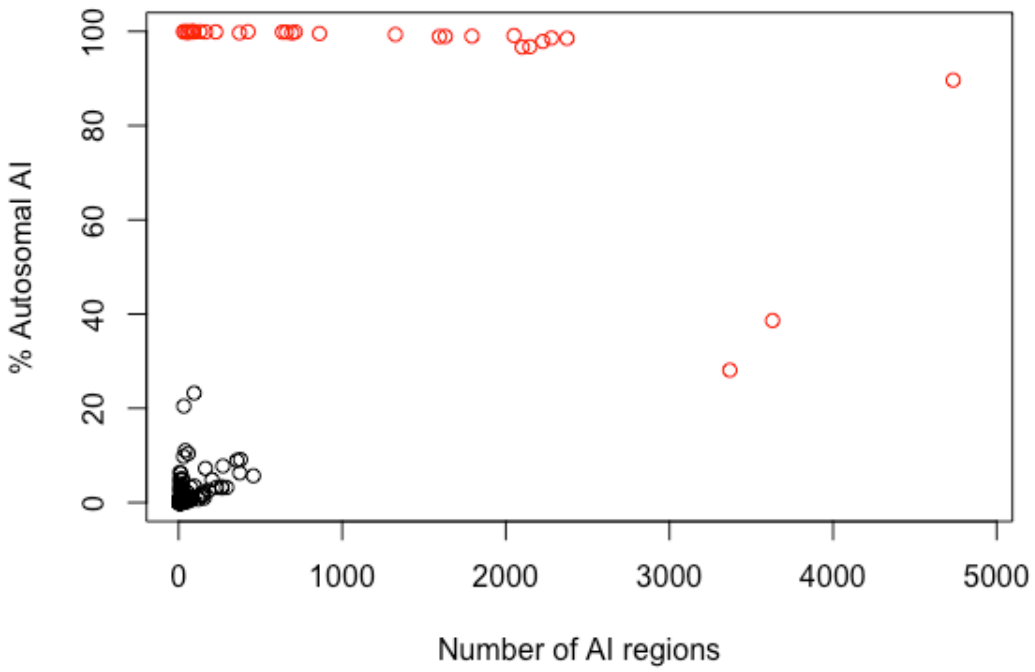


Figure 2.9 **Scatter plot showing the percentage of AI coverage versus the number of AI regions**

The scatter plot of the number of AI regions versus the percentage of autosomal AI shows that samples with either > 95% of autosomal AI or > 3,000 AI regions are outliers due to noisy array results. The 24 outliers are displayed in red.

The BAF plots for each of these samples were examined, which showed that the SNPs in these samples had a wide range of BAF values and did not form the expected genotypic clusters (Figure 2.10 B).

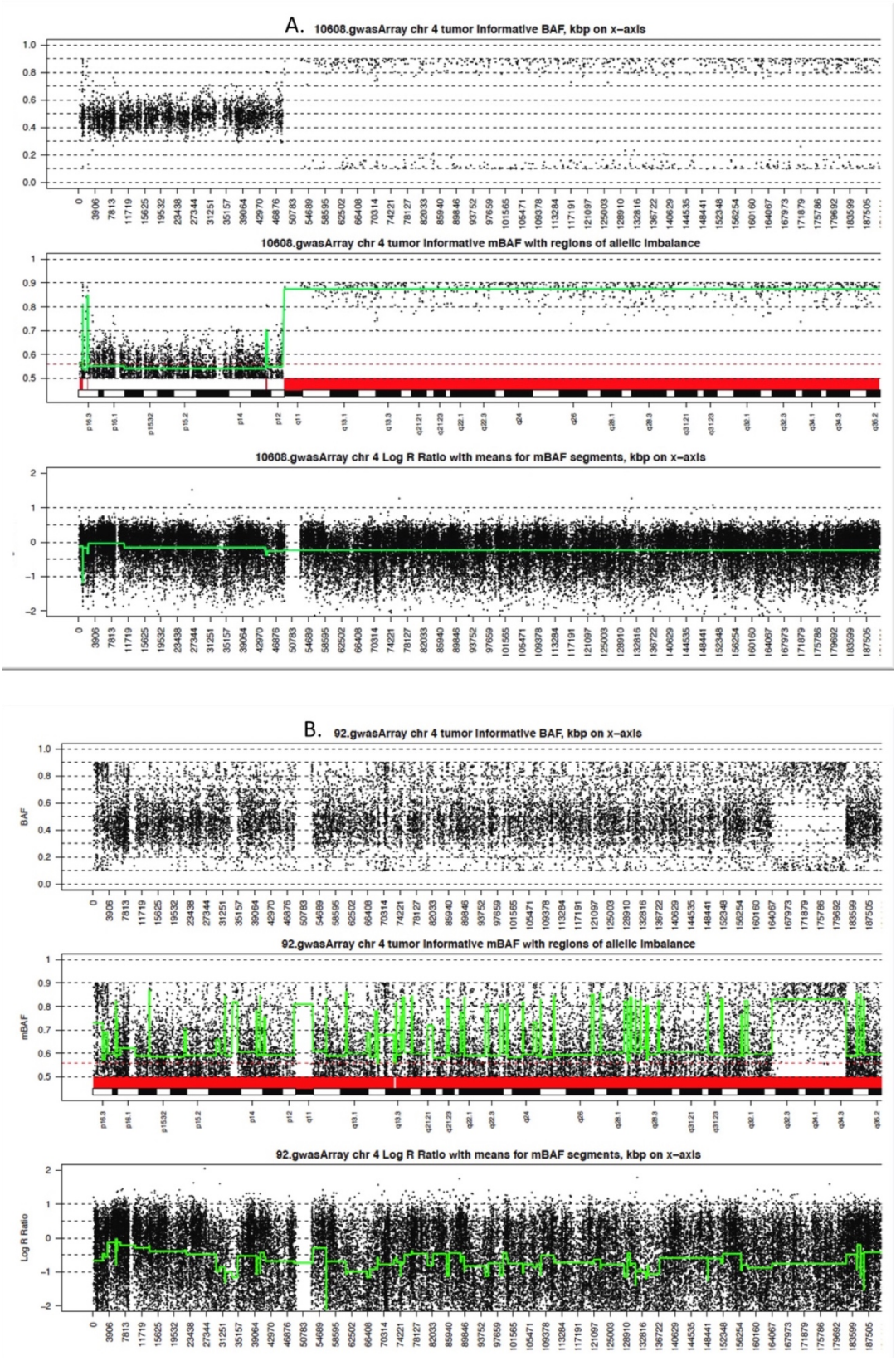


Figure 2.10 **BAF, mBAF and LRR plots of two samples for chromosome 4.**

A. The 4q region represent a region of aUPD. This is detected by the clear shift away from the heterozygous BAF value of 0.5 compared to the p arm that shows a normal 4p region. **B.** In noisy samples, the plotted data show a messy array for each chromosome of the same sample.

After removing the samples with noisy arrays, a custom script was used to categorise the remaining AI regions as either copy number neutral regions of aUPD (LRR between −0.15 and 0.065) if they were greater than 2Mb in length and extended to the telomere, copy number gains if the LRR was greater than 0.073 or copy number losses if the LRR was less than −0.14 (Staaf et al., 2008). This analysis showed that SM genomes are relatively simple with only 51 cases showing likely somatic copy number changes or aUPD (Figure 2.11). Large regions of aUPD and copy number alterations were rare, occurring with a similar frequency to that observed in MPN (Geyer, 2019; Tapper et al., 2015). Since these abnormalities are rare and do not overlap in a large proportion of patients it is unlikely that they will affect the association tests in the GWAS. Furthermore, apart from isolated cases the genomic regions with somatic changes did not include the risk factors that were identified and none of these regions were excluded from further analysis.

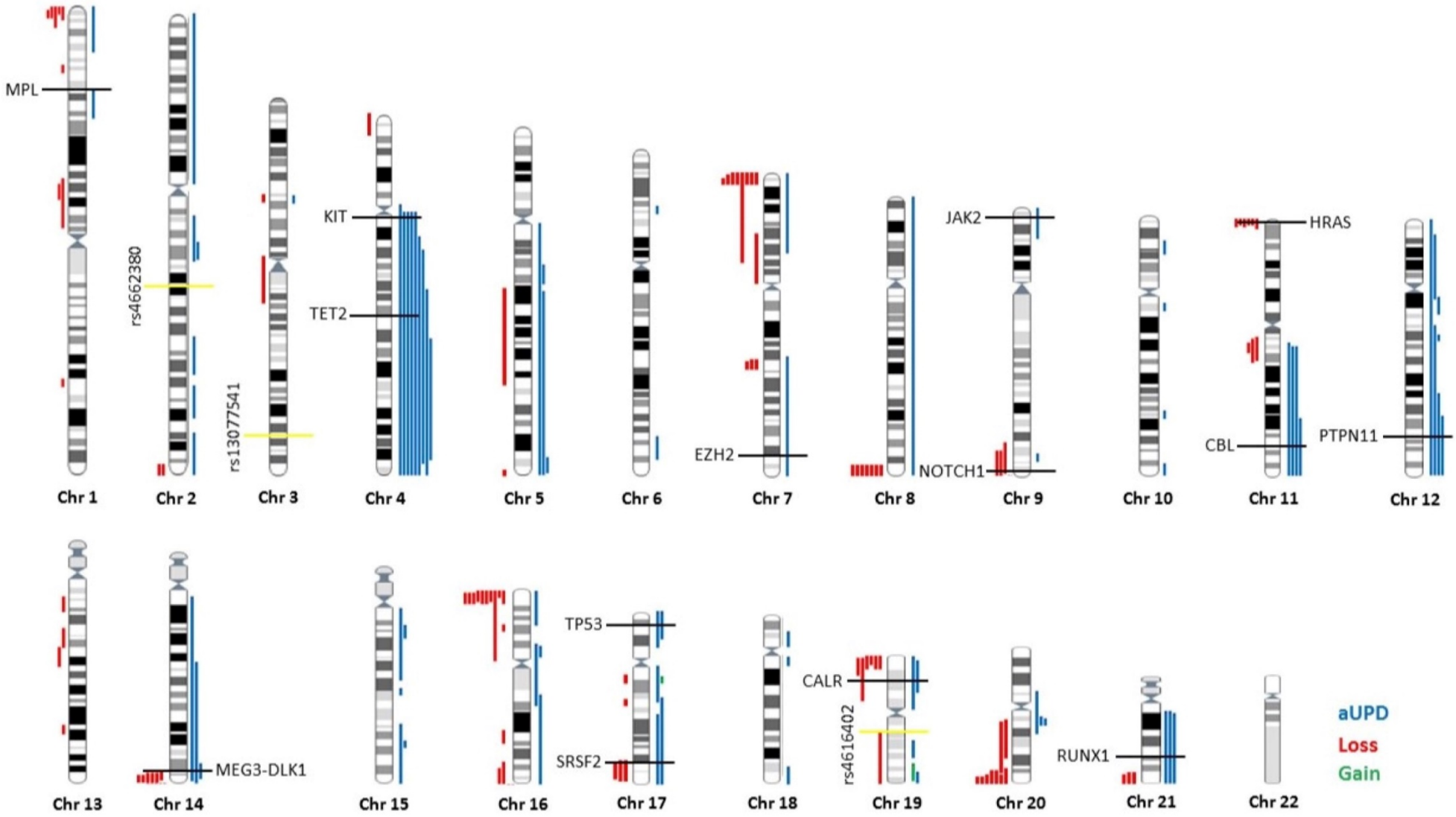

MPL
rs4662380
rs13077541
KIT
TET2
EZH2
JAK2
NOTCH1
HRAS
CBL
PTPN11
MEG3-DLK1
TP53
SRSF2
CALR
rs4616402
RUNX1
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
Chr 11
Chr 12
Chr 13
Chr 14
Chr 15
Chr 16
Chr 17
Chr 18
Chr 19
Chr 20
Chr 21
Chr 22
aUPD
Loss
Gain

Figure 2.11 **Copy number changes and regions of acquired uniparental disomy in the 409 stage 1 cases.**

The ideogram shows CNV and aUPD regions detected by Illumina Infinium OmniExpress array in the mastocytosis cases and mapped by chromosome. On the right: the blue bars depict regions of aUPD, gain regions are in green; on the left: the red bars indicate deletions. The horizontal yellow bars depict the location of the three genome-wide significant SNPs in the chromosome. The horizontal black bars indicate the location of recurrent aUPD regions targeting driver genes and an imprinted locus (*MEG3-DLK1* on chromosome 14) associated with myeloid neoplasms.

2.3.10 Replication in mastocytosis GWAS

Of the 92 SNPs selected, 75 were successfully genotyped in 666 *KIT*^{D816V} mastocytosis cases from Spain, Denmark and Italy. Additional controls (n=8,456) from the same populations that had previously been genotyped were used for comparison. After QC, 621 cases and all the controls remained for analysis (Table 2.6). All SNPs passed QC in cases although 19 were excluded from the Spanish controls due to per SNP missingness (≥10%) following imputation (Table 2.7).

Table 2.6 **Sample sizes before and after quality control in stage 2.**

	Stage 2 cases			**Stage 2 controls**		
Quality control measure	**Spanish**	**Danish**	**Italian**	**SNDNAB**	**Inter99**	**InCHIANTI**
Total samples pre-QC	399	185	82	1062	6184	1210
≥10% missing genotypes	30	14	1	0	0	0
Patients with outlying heterozygosity 3SD	0	0	0	0	0	0
Patients with gender mismatch	0	0	0	0	0	0
Patients with relatedness	0	0	0	0	0	0
Ancestry outliers	0	0	0	0	0	0
Samples remaining	369	171	81	1062	6184	1210

Table 2.7 **SNP number before and after quality control in stage 2.**

	Stage 2 cases			**Stage 2 controls**		
Quality control measure	**Spanish**	**Danish**	**Italian**	**SNDNAB**	**Inter99**	**InCHIANTI**
Total observed SNPs pre-QC	92			92	92	92
SNPs failed genotyping	17			0	0	0
SNPs with ≥10% missing genotypes	0			19	0	0
SNPs with MAF ≤ 5%	0			0	0	0
SNPs failing HWE*	0			0	0	0
Not bi-allelic SNPs	0			0	0	0
Unknown strand	0			0	0	0
Duplicates/Triplicates	0			0	0	0
MAF difference >0.34	0			0	0	0
Not in cases and controls	0			0	0	0
Total observed SNPs passing QC	75			73	92	92

QC: quality control, MAF: minor allele frequency, HWE: Hardy-Weinberg equilibrium.
*HWE P-value < 1×10^{-10} in cases, P-value <0.001 in controls

Samples were tested for association with SM as three separate cohorts using binary logistic regression. The final significance levels and effect sizes were determined using a fixed effects inverse variance-weighted meta-analysis to combine evidence from stages 1 and 2. This meta-analysis identified three intergenic SNPs with genome-wide significance, rs4616402 (P_{meta}=1.37×10^{-15}), rs4662380 (P_{meta}=2.11×10^{-12}) and rs13077541 (P_{meta}=2.10×10^{-9}) (Table 2.8).

Table 2.8 **Summary of the most significant SNPs from meta-analysis of stages 1 and 2.**

SNP	Chr	Location (hg19)	Alleles	RAF	Gene	P_{META}	OR (CI)	I^2
rs4616402	19q13	33,753,555	A/G	0.240	*SLC7A10-CEBPA*	1.37×10^{-15}	1.52 (1.37–1.68)	4.2
rs4662380	2q22	145,316,407	C/T	0.189	*LINC01412*	2.11×10^{-12}	1.46 (1.32–1.63)	0
rs13077541	3q26	176,925,740	G/A	0.464	*TBL1XR1-LINC00501*	2.10×10^{-9}	1.33 (1.21–1.45)	0

SNP, rs identifier from dbSNP; Alleles, risk associated/non-risk associated allele; RAF, risk allele frequency in Europeans from 1000 genomes; P_{META}, fixed effects meta-analysis of stages 1 and 2; OR, odds ratio; CI, 95% confidence interval; I^2, heterogeneity index (0–100).

Results for the three SNPs reaching genome-wide significance are summarised in a forest plot which shows that each SNP is significant in four of the five cohorts tested and that there is evidence for the same trend in the remaining population (Figure 2.12). I^2 statistics showed that for each SNP there was no evidence of heterogeneity between cohorts (Table 2.8). Results from the meta-analysis of stages 1 and 2 for all SNPs tested are shown in Appendix Table A.6.

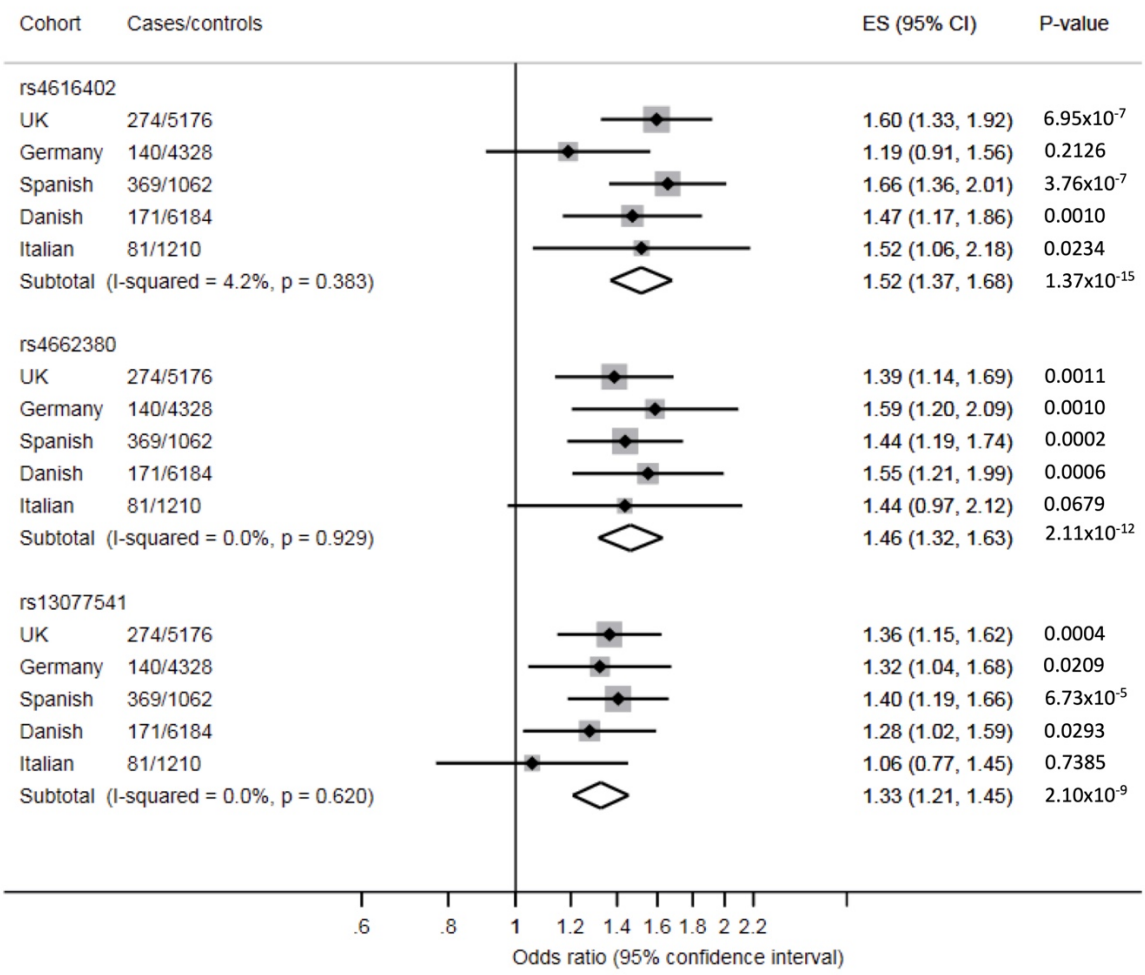


Figure 2.12 **Forest plots and meta-analysis for three SNPs reaching genome-wide significance.**

Forest plots for each SNP associated with SM at a genome-wide level of significance. Odds ratios (OR = ES) and 95% confidence intervals (CI) are displayed on the x-axis. Results are shown for each cohort (UK, German, Spanish, Danish and Italian) and the combined analysis. The SNP subtotals and diamond show the final OR and CI for a fixed effects meta-analysis of all five cohorts and uses I-squared to assess heterogeneity in effect sizes between cohorts.

2.3.11 Comparison of the stage 1 analyses

To investigate the possibility of residual population stratification, the stage 1 analyses were repeated without removing 26 samples with evidence of outlying ancestry (Table 2.3) and adjusting the association analysis using the first two principal components from MDS. The top three SNPs retained genome-wide significance, with rs4662380 and rs13077541 becoming slightly more significant (Table 2.9), which suggests an absence of residual population stratification in the original analysis. The results for all SNP tested in both stage 1 analyses were viewed side-by-side in QQ plots (Figure 2.13). Their genomic inflation factors ($\lambda \leq 1.02$) showed that both analyses generated similar significance profiles and demonstrated a close agreement between the

observed and expected P-values up to the tail of the distribution, where SNPs with P-values less than 10^{-4} began to deviate from the null distribution. Consequently, systematic biases such as separate genotyping of cases and controls, population stratification, or clonal somatic changes in the SM cases are therefore considered to be unlikely to contribute to the significance of these SNPs.

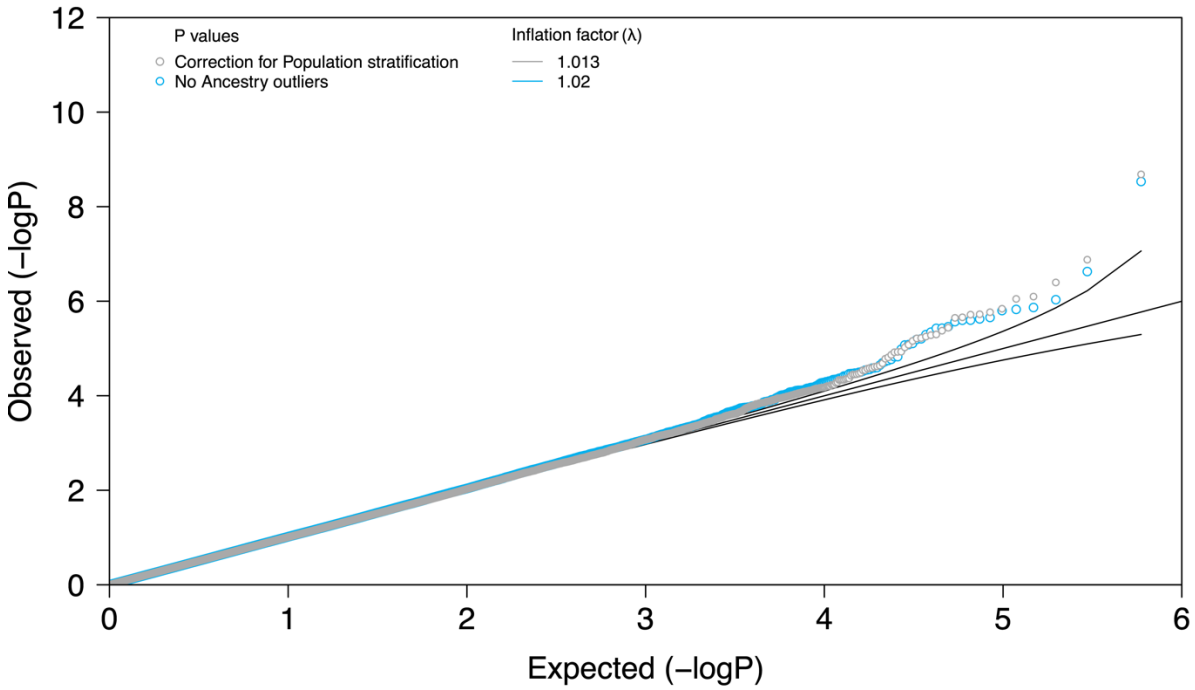


Figure 2.13 **QQ plot of the stage 1 meta-analysis with and without correction for population stratification.**

The analysis without correction excluded 26 ancestry outliers. These samples were included in the analysis, which corrected for population stratification using the first two principal components from the MDS analysis.

Table 2.9 **Summary of the most significant SNPs from meta-analysis with adjustment for population stratification.**

SNP	Chr	Location (hg19)	Alleles	RAF	Gene	P_{META}	OR (CI)	I^2
rs4616402	19q13	33,753,555	A/G	0.240	*SLC7A10-CEBPA*	5.26×10^{-15}	1.5 (1.36–1.66)	6.68
rs4662380	2q22	145,316,407	C/T	0.189	*LINC01412*	7.17×10^{-13}	1.47 (1.32–1.64)	0
rs13077541	3q26	176,925,740	G/A	0.464	*TBL1XR1-LINC00501*	5.32×10^{-10}	1.34 (1.22–1.47)	0

SNP, rs identifier from dbSNP; Alleles, risk associated/non-risk associated allele; RAF, risk allele frequency in Europeans from 1000 genomes; P_{META}, fixed effects meta analysis of stages 1 and 2; OR, odds ratio; CI, 95% confidence interval; I^2, heterogeneity index (0–100).

2.3.12 Genetic power calculation

Following QC, the stage 1 and stage 2 analyses involved 1,035 mastocytosis cases and 17,960 controls. According to these sample sizes and using a multiplicative disease model, this study is estimated to have 80% power to detect rare SNPs (MAF=0.1) with a relative risk of 1.82, and common SNPs (MAF=0.4) with a relative risk of 1.56 (Figure 2.14). Although these power estimates are encouraging, only one SNP with genome-wide significance was identified by the stage 1 analysis. The lack of genome-wide significant SNPs is most likely due to the relatively small number of cases and the power estimates being somewhat inflated by the comparatively large number of controls. Despite the small number of SNPs reaching genome-wide significance in stage 1 there were 18 SNPs with suggestive levels of significance and several of these formed well supported peaks on the Manhattan plots. Furthermore, three genome-wide significant SNP were replicated at stage 2. Due to the potential to overlook SNPs with smaller effect sizes, we used a set of selection criteria rather than significance alone (see 2.2.10) to identify 92 SNPs for replication.

2.3.13 Association with *TERT*

The stage 1 analysis included rs2853677, which has been linked to both MPN and *JAK2*^{V617F} associated CH (Hinds et al., 2016). This SNP is within *TERT* at 5p15 and marginally failed to meet the criteria for analysis at stage 2; however, the stage 1 meta-analysis for directly genotyped UK and German cases showed P_{meta}=0.0011, suggesting the possibility of an association. To examine this in more detail genotypes for 64 additional SNPs spanning *TERT* were imputed and tested for association with mastocytosis. As shown in Appendix Table A.7, 7 SNPs achieved P values of <0.001. The strongest of these was for rs7726159 (P_{meta}=8×10^{-5}), an established risk SNP for multiple cancer types (Wang et al., 2014b). One secondary association at *TERT* was identified for rs2853677, which remained significant after conditioning on rs7726159 ($P_{conditional}$= 0.035). No associations were seen with other SNPs that predispose to other MPN (Bao et al., 2020) or CHIP (Bick et al., 2020) in the stage 1 data (Table 2.10).

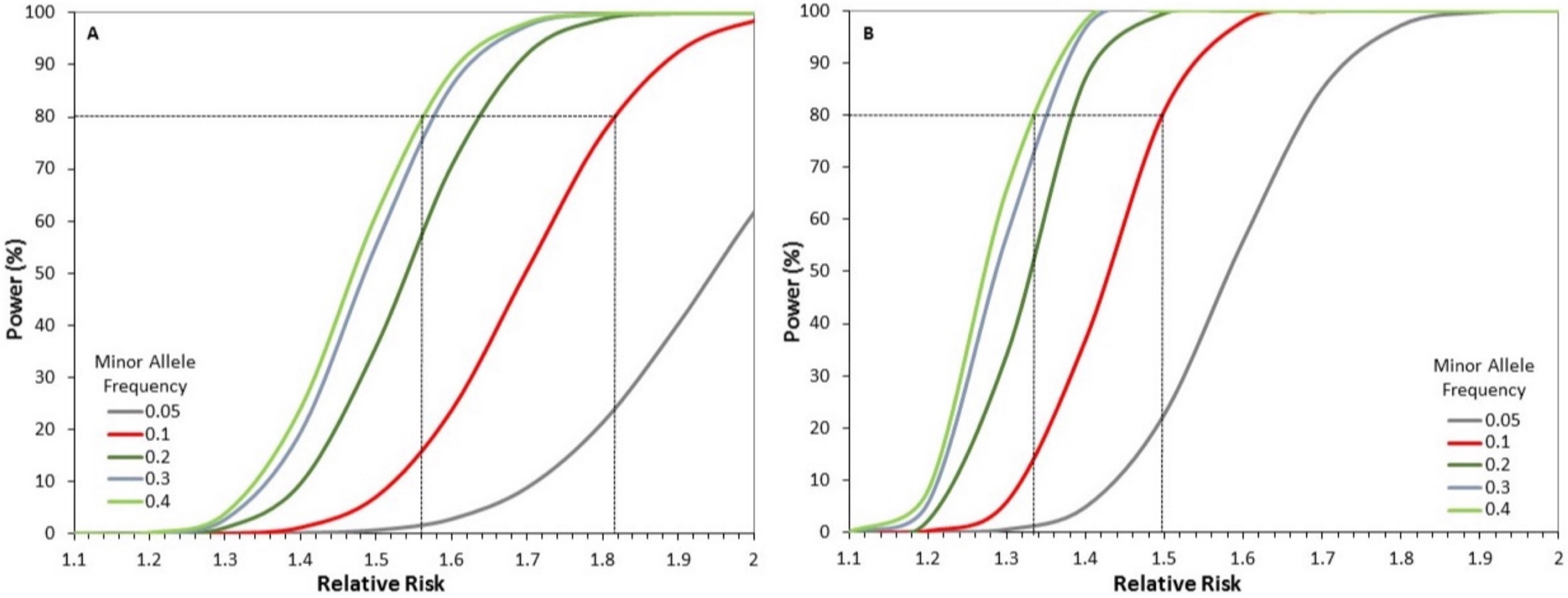


Figure 2.14 **Estimation of power to detect genetic effects in association with mastocytosis.**

The graphs show the relationship between relative risk and statistical power for SNPs with various minor allele frequency. The minor allele frequencies are represented by different colours (grey = 0.05, red = 0.1, dark green = 0.2, purple = 0.3, light green = 0.4). **A.** Stage 1 meta-analysis involving 414 cases versus 9,504 healthy controls. **B.** Meta-analysis of stages 1 and 2 involving 1,035 cases versus 17,960 healthy controls

Table 2.10 **Recent published genetic associations with MPN and CHIP**.

						Our stage 1 meta-analysis				Published associations			
CHR	SNP	BP (hg19)	A1	A2	Candidate gene	Observed P	Imputed P	OR	I^2	Published P value	OR	Disorder	Publication
3	rs74676712	160284736	C	T	*KPNA4*	.	0.6253	1.0591	0	2.64E-11	1.3	MPN	Bao et al. (2020)
3	rs9847631	168832107	T	G	*MECOM*	0.07679	.	1.1356	36.49	4.89E-10	1.17	MPN	Bao et al. (2020)
3	rs9864772	128316939	A	G	*GATA2*	0.6957	.	0.9719	0	2.64E-08	1.15	MPN	Bao et al. (2020)
3	rs77249081	159633461	C	G	*SCHIP1*	.	.	.	.	9.53E-07	3.7	MPN	Bao et al. (2020)
3	rs1210060191	160180516	GT	G	*KPNA4-TRIM59*	.	.	.	.	5.30E-10	1.16	CHIP	Bick et al. (2020)
4	rs62329718	105758059	A	T	*TET2*	.	0.4067	1.1656	0	2.72E-34	2.11	MPN	Bao et al. (2020)
4	rs144418061*	105728982	G	A	*TET2*	NA	NA	NA	NA	4.00E-09	2.4	CHIP	Bick et al. (2020)
6	rs116466979	34235378	T	C	*NUDT3*	.	0.07999	1.3258	0	2.31E-12	1.5	MPN	Bao et al. (2020)
7	rs62471615	130746955	A	C	*MKLN1*	.	.	.	.	2.31E-17	1.3	MPN	Bao et al. (2020)
9	rs1327494	4999303	G	A	*JAK2*	.	0.5219	1.0523	27.89	1.11E-170	2	MPN	Bao et al. (2020)
9	rs1633768	135879138	T	C	*GFI1B*	.	0.06677	1.1522	0	2.15E-12	1.2	MPN	Bao et al. (2020)
11	rs1800057	108143456	C	G	*ATM*	.	0.3938	1.1979	69.28	7.27E-10	1.65	MPN	Bao et al. (2020)
12	rs7310615	111865049	G	C	*SH2B3*	.	0.6168	1.0363	0	2.91E-21	1.27	MPN	Bao et al. (2020)
13	rs8002412	41331497	C	T	*MRPS31*	.	0.5268	1.0592	0	5.23E-10	1.2	MPN	Bao et al. (2020)
14	rs2887399	96180695	G	T	*TCL1A*	0.06372	.	0.8413	22.48	3.90E-09	1.23	DNMT3A /TET2 CHIP	Bick et al. (2020)
18	rs9946154	22810619	T	C	*ZNF521*	.	0.9973	0.9998	0	1.50E-08	1.15	MPN	Bao et al. (2020)
21	rs55857134	36347627	T	C	*RUNX1*	.	0.7773	0.9789	0	1.05E-09	1.17	MPN	Bao et al. (2020)
22	rs17879961	29121087	A	G	*CHEK2*	.	.	.	.	5.22E-06	2.23	MPN	Bao et al. (2020)

*This variant is not present in non-African ancestry

See Appendix Table A.7 for details on SNPs spanning *TERT*

2.3.14 Association with *TPSAB1* and *TPSB2*

Copy number variation at *TPSAB1*, the gene at 16p13 encoding α-tryptase, is associated with elevated serum tryptase levels in hereditary α-tryptasaemia (Lyons et al., 2016). My analysis did not include direct copy number analysis of this gene; however, a recent study linked *TPSAB1* duplications with three SNPs including rs58124832 (Lyons et al., 2018). This SNP was genotyped at stage 1 and met our criteria for analysis at stage 2, yielding a suggestive overall association with mastocytosis (P_{meta}=9.03×10^{-6}). The Cochran's Q test and I^2 statistics showed no evidence of heterogeneity between cohorts; however, the association was significant in only two cohorts (P_{German}=0.0058, P_{UK}=0.0042) and borderline in a third cohort ($P_{Spanish}$=0.05; Appendix Table A.6).

2.3.15 Associations with other genetic factors

A thorough search of the relevant literature yielded 14 SNPs that have been associated with the development of or phenotype of human mastocytosis (Daley et al., 2001; Lange et al., 2017; Nedoszytko et al., 2009, 2018, 2020; Rausz et al., 2013). Of these, 11 were directly genotyped or could be imputed from the stage 1 data (Table 2.11) but only one of these was significant; rs1800925 in the promoter region of *IL13* at 5q31 ($P_{imputed}$ = 0.008). This SNP has been linked to the development of adult SM and serum interleukin-13 levels (Nedoszytko et al., 2009) and inflammatory disorders such as chronic obstructive pulmonary disease (Ahmadi et al., 2019).

Table 2.11 **Published genetic associations with mastocytosis.**

						Our stage 1 meta-analysis				Published associations		
CHR	**SNP**	**BP (hg19)**	**A1**	**A2**	**Candidate gene**	**Observed P**	**Imputed P**	**OR**	**I^2**	**Published P value**	**OR**	**Publication**
11	rs10838094	5443893	A	G	*OR51Q1*	0.6678	0.8636	1.0124	11.23	2.21×10^{-29}	0.2071	Nedoszytko et al. (2020)
9	rs80138802	139915940	C	A	*ABCA2*	.	.	.	.	1.98×10^{-27}	5.739	Nedoszytko et al. (2020)
14	rs11845537	57446273	A	G	*OTX2-AS1*	.	0.7306	0.9442	54.16	1.60×10^{-18}	5.625	Nedoszytko et al. (2020)
3	rs9828758	73718136	T	C	*Near RP11*	0.126	0.08278	0.8713	70.62	2.94×10^{-7}	0.1467	Nedoszytko et al. (2020)
19	rs2279343	41515263	G	A	*CYP2B6*	.	0.2459	0.9069	0	2.32×10^{-10}	0.2795	Nedoszytko et al. (2020)
6	rs1611207	29759876	A	G	*HLA-V*	.	0.3148	0.9277	70.08	7.25×10^{-8}	2.105	Nedoszytko et al. (2020)
1	rs76015112	152129094	G	A	*RPTN*	.	0.7584	1.0288	0	2.94×10^{-7}	0.2965	Nedoszytko et al. (2020)
1	rs1778155	144874815	T	C	*PDE4DIP*	.	.	.	.	3.26×10^{-6}	2.032	Nedoszytko et al. (2020)
21	rs61735841	47558552	A	G	*FTCD*	.	0.05358	1.2361	71.2	4.34×10^{-5}	0.02573	Nedoszytko et al. (2020)
16	rs1801275	27374400	G	A	*IL-4RA*	0.7497	0.9594	1.0045	88.6	.	.	Daley et al. (2001)
5	rs1800925	131992809	T	C	*IL-13*	.	0.008089	0.77	0	0.0001	.	Nedoszytko et al. (2009)
1	rs2228145	154426970	C	A	*IL-6R*	0.5606	0.6291	1.0355	69.35	0.0088	2.488	Rausz et al. (2012)
4	rs5743708	154626317	A	G	*TLR-2*	.	.	.	.	0.01	4.22	Nedoszytko et al. (2018)
12	rs6489188	122660776	A	G	*IL-31*	.	0.7115	0.9721	89.92	0.045	4.04	Lange et al. (2016)

2.4 Discussion

This study represents the first two-stage case-control GWAS of mastocytosis. Although the disease is defined in most cases by the presence of a somatic *KIT*^{D816V} driver mutation, mastocytosis is in fact a complex disorder with diverse clinical phenotypes and outcomes. In this study, constitutional genotype has been identified as an additional factor that predisposes to mastocytosis. The use of molecular criteria to define cases in this study, rather than clinically-defined subtypes, plus careful matching of cases and controls with regard to ethnicity aimed to reduce the chance of heterogeneity in both the discovery and replication cohorts. Matching of cases and controls, despite the use of small sample sizes, has been successful in other GWAS investigating genetic predisposition of rare diseases in European ancestry (Mobuchon et al., 2017). Thus, with a relatively modest cohort size by current GWAS standards, it was possible to identify and validate 3 SNPs that achieved genome-wide significance, and identify further SNPs with suggestive associations at *TERT*, *IL3* and *TPSAB1/TPSB2*. Importantly, except for rs1800925 (*IL13*), none of the previously published associations were confirmed (Table 2.11). These publications included several candidate gene studies plus a recent GWAS that did not include a replication cohort (Nedoszytko et al., 2020). Both these approaches are highly prone to false positive results, although it is also possible that differences in genetic predisposition between populations may account for the lack of replication. An example of failure to replicate, due to difference in allele frequencies and reduced genetic power, was seen in GWAS of major depressive disorder when comparing populations from Europe and Asia (Cai et al., 2015).

The genomic DNA used in my study was extracted from peripheral blood leukocytes, which can potentially have both clonal and non-clonal origin. The possibility that clonal somatic changes might affect the GWAS analysis was considered. For example, recurrent somatic chromosomal changes or small copy number variants at high levels of clonality would lead to systematic errors in the assignment of constitutional genotypes in the affected regions. To exclude any spurious association due to somatic changes in the clonal lineage, an analysis of aUPD and copy number was performed on the discovery cohort, which showed that genomes of mastocytosis cases are relatively simple with only 51 cases having somatic copy number changes or aUPD (Figure 2.11). This is not unexpected since the size of the neoplastic clone in mastocytosis is often very small, and expected to be well below the resolution of SNP arrays (Arock et al., 2015). Chromosome 4q was the most frequent region of aUPD, although this was only present in 2.2% (9/414) of patients (Figure 2.11). In addition, no recurrent copy number changes or regions of aUPD were seen at the same location as the genome-wide significant SNPs identified in this study. The low incidence

overall of recurrent aUPD and copy number changes suggests that they are unlikely to have influenced the GWAS. Of interest, there was no evidence that genetic variation at *KIT* was associated with *KIT*^{D816V}-positive mastocytosis, in contrast to MPN in which the somatic *JAK2*^{V617F} mutation is more likely to arise on certain *JAK2* haplotypes (46 and 1) (Jones et al., 2009).

Lastly, the genotyping data were thoroughly quality-controlled and the QQ plots and their low genomic inflation factors from the stage 1 analyses showed no evidence for systematic biases between cases and controls such as recurrent somatic changes or population substructure ($\lambda \leq 1.02$; Figure 2.7). Consequently, clonal somatic changes are unlikely to account for the significant GWAS findings.

At stage 1 after the logistic regression analysis, 92 SNPs were selected for follow-up using the *clumping* procedure (Table 2.2). A stage 2 analysis was performed on the discovery cohort comprising of 666 cases and 8456 controls. Following the meta-analysis of five mastocytosis *KIT*^{D816V}-positive cohorts from stage 1 and stage 2, three SNPs with genome-wide significance were identified (rs4616402, rs4662380 and rs13077541).

rs4616402 (P-value = 1.37×10^{-15}) was the most significant marker, associated with a 1.52-fold increased risk of development of *KIT*^{D816V}-positive mastocytosis. This singleton SNP at 19q13 was tested in all 5 populations (Figure 2.12). It is located 36.8 kb downstream of solute carrier family 7 member 10 (*SLC7A10*) and 37.3 kb upstream of CCAAT enhancer binding protein alpha (*CEBPA*) (Figure 3.1 A). *CEBPA* is a single exon gene that encodes a leucine zipper transcription factor (C/EBPα) that binds CCAAT motifs in the promoter region of target genes (www.ncbi.nlm.nih.gove/gene). It is expressed in myeloid progenitor cells and involved in the proliferation arrest and differentiation of several types of cell lines including the myeloid lineage (Boyd and Arber, 2011). Several studies have defined a critical role for C/EBPα in myeloid development as well as malignant transformation of myeloid cells (Avellino and Delwel, 2017). *CEBPA* mutations have been shown to play an important role in inhibition of the wild-type C/EBPα tumour suppressor protein. About 13% of adults and 20% of children affected with AML harbour mutations in the *CEBPA* gene, usually in cases with a normal karyotype (Naeim et al., 2018). Cases with biallelic *CEBPA* mutations have a favourable outcome and, therefore, *CEBPA* testing is recommended in AML patients with normal karyotype (Griffith et al., 2017). Familial AML with

germline *CEBPA* mutations have been identified and are characterised by an autosomal dominant inheritance and 10-year-survival rate of 67% (Geyer, 2019). Interestingly, *CEBPA* was also a target of somatic mutations in an adult patient diagnosed with SM with associated CH non-mast cell lineage disease (SM-AHNMD) (Jayakumar and Xie, 2018) suggesting that *CEBPA* mutations might co-operate with KIT^{D816V} in disease progression. Of interest, two other deregulated tyrosine kinases in haematological malignancies are known to interact with *CEBPA* or C/EBPα: the BCR-ABL1 fusion protein downregulates *CEBPA* by a post-transcriptional mechanism (Perrotti et al., 2002) and oncogenic FLT3 mutants disrupt C/EBPα function by ERK1/2-mediated phosphorylation (Radomska et al., 2006). *CEBPA* is thus a strong candidate gene associated with the signal at rs4616402. The right gene *SLC7A10* flanking rs4616402 is a protein-coding gene and to date has not been related to myeloid malignancies or relevant biological process.

The second most significant SNP, rs4662380, is located at chromosome 2q22 within *LINC01412* and 109 kb upstream of testis expressed 41 (*TEX41*). Both of these genes are long non-coding RNAs (lncRNA) of unknown function, but due to the possibility of long-range interactions between GWAS signals and target genes it is unclear if either is directly relevant to SM. The competing endogenous RNA (ceRNA) hypothesis was outlined in 2011 to explain how a large proportion of RNAs from the transcriptome (protein coding genes, pseudogenes and lncRNAs) can communicate with each other via microRNAs, which may be considered as letters of a new RNA language (Qi et al., 2015; Salmena et al., 2011). lncRNAs are known to play an important role in cancer progression by modulating the expression of miRNAs or target proteins (Rathinasamy and Velmurugan, 2018) and additional studies have revealed their important role in proliferation, apoptosis and differentiation of leukaemia cells (Liu et al., 2019). Notably, a study conducted to investigate the role of *COMMD6* in tumourigenesis and malignant progression led to the proposal of ceRNAs networks on the basis of differentially expressed transcriptome from the cancer genome atlas database. In addition, a *TEX41*-miR-340-*COMMD6* ceRNA network in head and neck squamous cell carcinoma (HNSC) identified a potential tumour-promoting role for *TEX41* and *COMMD6* (Yang et al., 2019). The same role for *TEX41* in promoting tumour progression was also identified in cervical cancer (Li et al., 2018), confirming it as a potential gene involved in molecular mechanisms in several human tumours.

Zinc finger enhancer-box (E-box) homeobox 2 (*ZEB2*) and *ZEB2* antisense RNA 1 (*ZEB2-AS1*) are other genes near rs4662380. *ZEB2* is a gene encoding for a transcription factor with a zinc finger motif of about 23 amino acids that binds E-box-like sequences (CANNTG, where N is not a specific nucleotide) in the promoters of target genes (Strachan and Read, 2011). This protein is a complex

transcription factor with several functional domains that can also interact with other proteins to form a transcriptional complex that can activate or repress transcription of target genes (Remacle et al., 1999). *ZEB2* has been linked to both myeloid and lymphoid leukaemias (Bolouri et al., 2018; Goossens et al., 2019) and plays a critical oncogenic role in the malignant transformation of several tumours such as breast cancer (Duan et al., 2019) and glioblastoma (Safaee et al., 2021). In a recent study of a subgroup of immature acute leukaemias, four types of translocation involving *BCL11B* were identified with *ZEB2-BCL11B* being the only rearrangement producing a fusion gene (Di Giacomo et al., 2021). Also, *ZEB2-AS1* is another lncRNA that promotes the cell proliferation and invasion of several types of cancers (Gao et al., 2018; Guo et al., 2018; Wu et al., 2017; Xu et al., 2019; Zhang et al., 2019). The overexpression of *ZEB2-AS1* was demonstrated to be highly associated with poor clinical outcomes in patients affected with AML, particularly a shorter overall survival rate (Shi et al., 2019). In addition, a recent study showed both *in vitro* and *in vivo* using a mouse model, that cell proliferation was suppressed and apoptosis of AML cells increased when silencing *ZEB-AS1*. They were able to identify a regulatory role for *ZEB-AS1* in the proliferation of AML cells through the *ZEB2-AS1*/miR-122-5p/*PLK1* ceRNA network (Guan et al., 2020; Salmena et al., 2011). Thus, there are a number of possible candidate functional mechanisms to explain my GWAS findings at rs4662380 that merit further investigation.

The final significant SNP rs13077541 (P-value = 1.224×10^{-9}) is located on chromosome 3 (Figure 3.1 C). The association signal is located 10,692 bp downstream of transducin beta like 1 X-linked receptor 1 (*TBL1XR1*) and 234 kb upstream of lncRNA 501 (*LINC00501*). *TBL1XR1* is a member of the WD repeat-containing gene family and encodes a protein required for transcriptional activation. It shares sequence similarity with TBL1X, a component of both histone deacetylase 3 and nuclear receptor corepressor complexes that are required for transcriptional activation by a variety of transcription factors. *TBL1XR1* is also involved in rare translocation events such as the *TBL1XR1-PIK3CA* fusion in breast cancer and prostate cancer where the *TBL1XR1* sequence contributes only the 5' untranslated region, which drives the overexpression of its partner (Stransky et al., 2014). A *TBL1XR1-PDGFRB* fusion was also identified in association with myeloid malignancies and marked eosinophilia. The pathogenic fusion results in an in-frame rearrangement containing the tyrosine kinase domain of *PDGFRB* and the N-terminal of *TBL1XR1*, which promotes protein dimerisation (Campregher et al., 2017). Interestingly, several other fusions involving the receptor tyrosine kinase *PDGFRB* have been found in association with MPN, including another WD repeat family member, suggesting a strong association between these gene fusions and myeloid cancers (Hidalgo-Curtis et al., 2010).

GWAS have been very successful in identifying genome-wide significant associations (Tam et al., 2019). Up to January 2019 3,730 GWAS have been published and have successfully identified risk loci for several traits including rare diseases and cancer (Chio et al., 2009; Ferrari et al., 2014; Kouri et al., 2015; Mobuchon et al., 2017). However, GWAS often require large sample sizes and it can be difficult to acquire sufficient numbers of cases when dealing with rare diseases such as SM (estimated prevalence only 1–9/100,000). In fact, only relatively small case cohorts were available for the study and thus the power to detect SNPs with small effect sizes (OR<1.82) was limited (Figure 2.14). In this study controls are unselected, meaning that they are randomly selected from the population, and they have not been screened for disease. In this case, the power calculation will assume that a proportion of controls in relation to the disease prevalence will develop the disease, so the statistical power will be reduced. The higher the disease prevalence, the lower will be the power of detecting genetic effect. However, we are investigating a rare disease and this reduction in power is expected to be very limited since there is an increase in power in relation to the prevalence of the disease when using unselected controls. In my study, cases from multiple populations were used to accrue a sufficient sample size. However, the presence of multiple populations could also introduce some limitations such as reducing the genetic similarities between individuals and therefore introducing genetic heterogeneity. To minimise heterogeneity, we only selected *KIT*^{D816V}-positive cases both in stage 1 and stage 2 of this study and performed case–control comparisons in separate populations followed by meta-analysis. Another limitation of the study is the use of many control cohorts that had been genotyped by different facilities using different genotyping arrays. Strand inconsistency, and different locations and SNP names are some of the issues that can make the analysis very challenging if additional QC checkpoints are not addressed before the data are merged. The use of independent cohorts coming from 5 different populations could potentially make the replication of the selected SNPs more difficult. However, all the cohorts chosen for this study belong to the Caucasian population, therefore they are not dramatically different. According to previously published GWAS findings, it was expected that a reasonable percentage (1–3%) of the selected SNPs should replicate with adjusted P-value <0.05 (accounting for the number of SNPs tested at stage 2). In my study, three genome-wide significant SNPs from the 75 that were successfully genotyped at stage 2 were tested in all five populations and significant in 2/3 of the replication cohorts (Figure 2.12), providing compelling evidence for real effects.

Variants passing the threshold of genome-wide significance (P-value < 5×10^{-8}) were investigated further to assess their association with mastocytosis. In the following chapter, *in silico* functional

follow-up such as expression quantitative trait loci (eQTL) will be presented and discussed. This analysis is very important for identifying potential target genes and investigating the effect of genotype on gene expression levels that are likely to affect the disease phenotype (Nica and Dermitzakis, 2013; Spain and Barrett, 2015).

To conclude, this chapter describes the background of the two-stage GWAS of mastocytosis, as well as the methods, results obtained from the stage 1 and stage 2 analysis, and a discussion of the limitations of the study. Consideration of the three signals has identified three strong candidate genes, *CEBPA, TEX41* and *ZEB2*, plus other genes of potential interest. Translating the new findings into causal variants and providing proof for target genes is the most challenging step in a GWAS, especially for those SNPs in intronic or intergenic regions of the genome with unknown function. For the SM-GWAS, additional analysis in other independent cohorts will help to confirm these findings, and detailed functional and genetic studies will be needed to provide insights into their biological significance, to localise candidate causal variants and to analyse gene–gene or protein–protein interactions.

Chapter 3 Post-GWAS analysis

3.1 Introduction

Genome-wide association studies have been very successful in identifying thousands of unique common variants that influence individuals' predisposition to complex traits (Buniello et al., 2019). In most instances, understanding the underlying mechanism by which these variants impact the associated phenotype is still limited, because most of these variants are located in non-coding regions of the genome and are more likely to have regulatory functions rather than disrupting the reading frame for a protein. Although the variant-to-function translation remains challenging, many research groups have made further steps in identifying key genes in biological processes, diseases underlying causal variants, and biological pathways associated with altering the risk of developing the respective disease (Gallagher and Chen-Plotkin, 2018). The identification of target genes represents the first step in tackling the link between a genetic association and the biological function.

For example, the first GWAS looking at the association between blood disorders (β-thalassemia and sickle cell disease) and fetal haemoglobin (Hbf), (tetramer of two adult α-globin and two fetal γ-globin subunits; after birth two β-globins will replace the fetal ones), identified a strong association (rs11886868) with the *BCL11A* gene on chromosome 2 in disparate population studies (Lettre et al., 2008; Uda et al., 2008). β-thalassemia samples with mild phenotype and carrying the risk allele were found to have elevated Hbf levels compared with those with a severe form of the disease (Uda et al., 2008). Follow-up studies examining the role of *BCL11A* in modulating Hbf levels found that this gene serves as a key regulator of haemoglobin production. By examining the expression of *BCL11A* in adult erythroid cells, they saw that cells carrying two risk alleles (associated with high Hbf) showed a reduced expression of this gene compared to those homozygous for the non-risk allele (low Hbf). Knockdown of *BCL11A* in differentiated erythroid precursors showed an increase in γ-globin levels showing a clear molecular function of *BCL11A* in silencing the γ-globin genes (Sankaran et al., 2008). This information has provided biological insights for better understanding of haematopoiesis and has led to the identification of two major transcriptional repressors, *BCL11A* and *ZBTB7A*, regulating γ-globin genes and haemoglobin switching (Martyn et al., 2018; Sankaran et al., 2009). Decades of research have ultimately led to ongoing clinical trials to suppress *BCL11A* and reactivate the developmentally silenced γ-globin genes to increase the amount of Hbf in patients affected with sickle cell disease (e.g. ClinicalTrials.gov Identifier: NCT03282656).

Following the identification of a large spectrum of variants associated with MPN (Bao et al., 2020; Bick et al., 2020; Hinds et al., 2016; Jones et al., 2009; Kilpivaara et al., 2009; Olcaydu et al., 2009; Tapper et al., 2015), potential target genes were also identified using a variety of approaches such as genetic fine-mapping and targeted variant-to-function assay (Bao et al., 2020; Ulirsch et al., 2019). Functional investigation of *CHEK2*, for example, showed that loss-of-function variants within this gene are associated with increased risk of CH. In addition, researchers proved the involvement of *CHEK2* in stem cell expansion, as they showed that suppression of this gene allowed increased expansion of human haematopoietic progenitor cells (Bao et al., 2020).

An important step for understanding the role played by the identified loci is also the identification of the key cellular type and tissue for mediating disease risk (Cano-Gamez and Trynka, 2020; Nandakumar et al., 2020). For instance, following the identification of over 400 independent risk factors associated with type 2 diabetes (T2D), the Human Islet Biobank was established as part of a collaborative effort (Fuchsberger et al., 2016; Thurner et al., 2018). This allowed a detailed characterisation of the tissue to be performed and consequent understanding of the human pancreatic islets and regulatory mechanisms using different omics data (van de Bunt et al., 2015; Gaulton et al., 2010; Viñuela et al., 2020). To facilitate identification of the most specific tissue for each identified association, a tool named TACTICAL (Tissue of ACTion scores for Investigating Complex trait-Associated Loci) has recently been developed to obtain what are called tissue of action (TOA) scores and select key tissues in the pathogenesis of complex phenotypes (Torres et al., 2020).

The GWAS described in Chapter 2 has led to the identification of three genome-wide significant SNPs associated with increased risk of developing *KIT*^{D816V}-positive mastocytosis. The aim of the post-analytical interrogation was to take advantage of this variation to better understand mastocytosis. The majority of variants reported by GWAS are in noncoding regions of the genome; these lead variants may not be causal but might be in high LD (r^2>0.8) with the casual variant. This chapter will describe the *in silico* approaches that were used to explore the relationship between the regions containing genome-wide significant SNPs and mastocytosis, and to nominate a number of target genes that are potentially impacting key mechanisms in patients (Boyle et al., 2012; Ward and Kellis, 2016; Watanabe et al., 2017). Additionally, this chapter describes a gene-based analysis of the stage 1 data which used multiple SNPs to generate P-values for individual genes.

3.2 Materials and Methods

3.2.1 Post-analytical interrogation of SNPs

3.2.1.1 Functional annotation using HaploReg

SNPs in LD are inherited together in the population and can be used to define causal regions. The genome-wide significant associations were therefore investigated to see if SNPs in LD are also associated with mastocytosis. Lead SNPs and variants in high LD (r^2 ≥0.8) with the lead SNPs were identified and annotated to determine their biological relevance using HaploReg (version 4.1) (Ward and Kellis, 2016). This tool integrates a range of databases (e.g. ENCODE Project, Roadmap Epigenomics Project, dbSNP, EBI-NHGRI GWAS Catalog) and uses data from the 1000 Genomes Project Phase 1 release to calculate LD for four ancestral populations, including Europeans (Altshuler et al., 2012; Buniello et al., 2019; Kheradpour and Kellis, 2014; Kundaje et al., 2015; Sherry, 2001). The SNPs were listed and annotated with respect to genomic features, such as mammalian evolutionary sequence conservation elements, using SiPhy and GERP statistics and epigenomic features. These are described in the following paragraph (Davydov et al., 2010; Lindblad-Toh et al., 2011). HaploReg v4.1 is updated to November 2015, and annotations with respect to previously identified GWAS associations are only available up to this date (Ward and Kellis, 2016). Many other SNPs associated with various blood traits and MPN have been identified in more recent studies (Bao et al., 2020; Bick et al., 2020), Therefore, for a more accurate annotation, the EBI-NHGRI GWAS Catalog was also explored using FUMA GWAS (Watanabe et al., 2017).

3.2.1.2 HaploReg approach for epigenomic annotation

The DNA inside the nucleus is associated with proteins to form chromatin, a complex, dynamic molecular structure (Van Steensel, 2011). When scientists first started to look at these structures, microscopic analysis revealed only two chromatin states (Baker, 2011). Heterochromatin or ‘closed’ chromatin is a highly condensed state, within which genes are not accessible to the transcriptional machinery. On the other end, chromatin can have an extended state known as euchromatin or ‘open’ chromatin, which enables active transcription (Strachan and Read, 2011). Following the application of computational analyses, researchers used a set of five histone modification marks (Table 3.1) and recurrent ones have been grouped into several different conformations or chromatin states that have mapped several regulatory elements as the critical elements in gene expression (Baker, 2011). To interpret GWAS results and to investigate whether the SNPs identified in this study are in regulatory elements of the genome, the 15-state chromatin model was used in this study, which is based on a multivariate hidden Markov model and captures

all the key interaction that occur between the chromatin marks (Ernst and Kellis, 2017; Ward and Kellis, 2016).

Table 3.1 **Histone modification marks.**

Marks	Abbreviation	Function
Histone H3 lysine 4 trimethylation	H3K4me3	Associated with promoter regions
Histone H3 lysine 4 monomethylation	H3K4me1	Associated with enhancer regions
Histone H3 lysine 36 trimethylation	H3K36me3	Associated with transcribed regions
Histone H3 lysine 27 trimethylation	H3K27me3	Associated with Polycomb repression (gene silencing)
Histone H3 lysine 9 trimethylation	H3K9me3	Associated with heterochromatin regions (gene silencing)

H3K4me3 and H3K4me1 (Heintzman et al., 2007; Igolkina et al., 2019); H3K36me3 and H3K27me3 (Bonasio et al., 2010; Li et al., 2007); H3K9me3 (Li et al., 2007; Peters et al., 2003).

The 15-state model consists of 8 active states associated with gene transcription (states 1–8 in Table 3.2) and 7 repressed states (states 9–15 in Table 3.2) that take into account DNA methylation, transcription factors binding, evolutionary conservation and DNA accessibility (Kundaje et al., 2015).

Table 3.2 **15 Chromatin states.**

State No	Chromatin state	Abbreviation
1	Active transcription start site (TSS)	TssA
2	Flanking active TSS	TssAFlnk
3	Transcription at gene 5’ and 3’	TxFlnk
4	Strong transcription	Tx
5	Weak transcription	TxWk
6	Genic enhancers	EnhG
7	Enhancers	Enh
8	Zinc finger protein genes and repeats	ZNF/Rpts
9	Heterochromatin	Het
10	Bivalent/poised TSS	TssBiv
11	Flanking bivalent TSS/Enhancers	BivFlnk
12	Bivalent enhancer	EnhBiv
13	Repressed Polycomb	ReprPC
14	Weak repressed Polycomb	ReprPCWk
15	Quiescent/low	Quies

The 15 different states are available for 147 cell or tissue types on the basis of the epigenomic information generated from the 111 reference human epigenomes of the NIH Roadmap

Epigenomics Consortium and the 16 additional epigenomes from the Encyclopedia of DNA Elements (ENCODE) (Dunham et al., 2012; Javierre et al., 2016; Kundaje et al., 2015; Ward and Kellis, 2016). For the post-GWAS analysis, chromatin accessibility was interrogated in two cell lines relevant to mastocytosis, the blood cell lines E035 (primary haematopoietic stem cell) and E123 (K562 chronic myeloid leukaemia cell line), using the HaploReg tool (version 4.1) (Figure 3.1) (Ward and Kellis, 2016).

3.2.1.3 RegulomeDB for interpretation of regulatory variants

Most GWAS associations are located in non-coding regions of the genome and are more likely to have regulatory functions (Gallagher and Chen-Plotkin, 2018). RegulomeDB was used to interpret the functional effect that variants mapping to regulatory regions may have on protein binding (Boyle et al., 2012). RegulomeDB is a database that combines 962 experimental datasets from several sources including ENCODE and across more than 100 tissues and cell lines. Briefly, ENCODE transcription factors (TF) for chromatin Immunoprecipitation sequencing (ChIP-seq) and the modified version ChIP-exo, histone ChIP-seq, formaldehyde-assisted isolation of regulatory elements (FAIRE), DNase I hypersensitive site data and a collection of eQTL and dsQTL data were all included. These data are integrated together into a tool that assigns a RegulomeDB score to each variant in order to estimate their potential regulatory effect and identify functional variants. According to this heuristic scoring system (Table 3.3) the lower scores represent higher confidence for a variant to be located in a region of the genome with functional relevance. Variants with lower scores show increased confidence for their functional relevance. Variants scoring 1 have been associated with expression of target genes and are likely to affect binding; variants scoring 2 are likely to affect binding; variants scoring 3 are less likely to affect binding; variants scoring 4, 5 and 6 have minimal binding evidence.

Table 3.3 **RegulomeDB scoring system.**

Category	Subcategory	Description of the score
1	A	eQTL + TF binding + matched TF motif + matched DNase footprint + DNase peak
	B	eQTL + TF binding + any motif + DNase footprint + DNase peak
	C	eQTL + TF binding + matched TF motif + DNase peak
	D	eQTL + TF binding + any motif + DNase peak
	E	eQTL + TF binding + matched TF motif
	F	eQTL + TF binding/DNase peak
2	A	TF binding + matched TF motif + matched DNase footprint + DNase peak
	B	TF binding + any motif + DNase footprint + DNase peak
	C	TF binding + matched TF motif + DNase peak
3	A	TF binding + any motif + DNase peak
	B	TF binding + matched TF motif
4	NA	TF binding + DNase peak
5	NA	TF binding or DNase peak

Category	Subcategory	Description of the score
6	NA	Motif hit

eQTL: expression quantitative trait loci; TF: transcription factor; DNase: enzyme deoxyribonuclease. Adapted from Boyle et al., 2012.

3.2.1.4 Long non-coding RNA investigation

Long noncoding RNA (lncRNA) are known to promote the proliferation of several types of cancer (Huarte, 2015). This large class contains non-coding RNA genes longer than 200 nucleotides that have known regulatory functions, and studies have suggested that they can regulate expression of nearby genes (Marchese et al., 2017). For instance, the overexpression of *ZEB2-AS1* is associated with poor clinical outcomes in patients affected with lung cancer, AML and breast cancer (Gourvest et al., 2019; Guo et al., 2018; Zhang et al., 2019). To annotate and investigate causally relevant lncRNAs, the fifth release of LNCipedia (Volders et al., 2019) and version 1 of the Cancer LncRNA Census (CLC) (Carlevaro-Fita et al., 2020) were used. LncRNAs located in proximity to our genome-wide significant SNPs were queried by their names using the publicly available LNCipedia resource built with a web-interface and containing 21,488 unique transcripts. In contrast to the other databases, the CLC has the advantage of including only high confidence genes that have strong genetic and functional causal roles in cancer.

3.2.1.5 Pleiotropy/GWAS catalog

Pleiotropy is association between the same variants with multiple traits (Gratten and Visscher, 2016; Solovieff et al., 2013). Genomic research has shown that this phenomenon is common for many complex traits, including cancer (Wu et al., 2018). The NHGRI-EBI GWAS Catalog was used to examine whether the annotated lead variants or their proxies also influence blood counts or apparently unrelated phenotypic traits. The GWAS Catalog, a publicly available resource of SNP-trait association and summary statistics from early July 2019 contains more than 150,000 unique SNP associations for 17 trait categories and over 4,000 publications (Buniello et al., 2019).

3.2.1.6 Quantitative trait locus analysis (QTL)

To test for association between genetic variation and transcript level of a gene, expression quantitative trait loci (eQTL) analysis of the lead SNPs and their proxies ($r^2 \geq 0.8$) was performed in blood using GTEx v8 and QTLbase (Carithers and Moore, 2015; Zheng et al., 2020).

To gain further functional insights, methylation quantitative trait loci (mQTL) analysis was performed with QTLbase to study the association between the annotated SNPs and epigenetic regulation in non-diseased human whole blood (Zheng et al., 2020). Most of the QTL results that are displayed in QTLbase come from studies conducted within European populations, however in some instances results derive from groups of combined ethnicity, which are indicated as mixed populations (Zheng et al., 2020). The results from the eQTL contain important statistical values, including P-value, effect size and tested allele. If the tested allele is not specified, as in GTEx, the result reported refers to the expression of the alternative or minor allele compared with the reference. The effect size is normalised according to the statistical method applied, beta in GTEX and normalised effect size (NES) in QTLbase, where magnitude has no direct biological interpretation. Negative NES/beta indicates that the tested allele is associated with a reduction in gene expression, whereas a positive NES/beta indicates increased gene expression.

3.2.1.7 CADD score

Combined annotation-dependent depletion (CADD) is a score used to estimate deleteriousness of SNVs and indels in any location of the human genome. This is a machine learning method freely available to give an estimate of pathogenicity. CADD scores were used because they were shown to have greater predictive accuracy (AUC) when compared with other metrics used for prediction of pathogenic mutations that are reported in the ClinVar database (Landrum et al., 2020). The CADD score takes into account many different features such as sequence conservation across species, structural and biochemical features of the protein (Kircher et al., 2014). The higher the score, the more deleterious the predicted consequences of the SNP is, and 12.37 represents the suggestive threshold for estimating whether a SNP should be considered deleterious or not (Kircher et al., 2014). The score for the lead SNPs and their proxies was generated using the most recent version v1.4 for the human genome build GRCh37 (https://cadd.gs.washington.edu).

3.2.2 Data analysis

3.2.2.1 Description of clinical features in the Spanish and Italian cohort

Diagnostic and phenotype variables at initial diagnosis (advanced disease = ASM, SM-AHN, MCL; non-advanced disease = all other subtypes, Table 2.1), the presence or absence of skin lesions (yes/no), sex, BST (ng/mL) and age were available for the majority of the Spanish (n=369) and Italian (n=81) cohorts, but not other cohorts. Bone marrow involvement and D816V mutation burden were only available for some of the Spanish cases.

3.2.2.2 Association with clinical features

Statistical analysis was performed on a cohort of 450 individuals (n=81 from Italy, n=369 from Spain) after removing those individuals with more than 10% missing genotypes (n=31). Three categorical variables (initial diagnosis, skin lesions and sex) were tested for association with allelic counts of the three significant SNPs using Fisher's exact test. A fixed-effect inverse variance-weighted meta-analysis was used to combine evidence from the two cohorts. Normal distribution of continuous variables (tryptase, age and D816V mutation burden) was checked using Kolmogorov-Smirnov and tryptase levels were normalised using quantile transformation. Following normalisation, continuous variables were tested using linear regression following Kolmogorov-Smirnov checks for normal distribution and normalisation of tryptase levels using quantile transformation.

3.2.2.3 Gene-based test

Gene-based analysis as well as the single-marker association test represent valuable approaches when investigating complex traits. The gene-based approach allows the joint effect of weakly-associated markers seen by single-SNP analysis to be considered collectively (de Leeuw et al., 2015). The summary statistics from the stage 1 meta-analysis (N_{cases}=414; $N_{controls}$=9,504;) were used as input in FUMA to perform a gene-based analysis of association with SM which uses MAGMA (version 1.08) to apply multiple linear regression and obtain gene-based P-values. For the gene-based test, the P-value is computed for each gene using all the SNPs located within genes and including SNPs in a 10kb window in both directions around genes. MAGMA takes into account gene size, number of SNPs in a gene, and from the reference data with similar ancestry, corrects for LD between markers. The 1000 genomes phase 3 data of European ancestry was used as the reference to account for LD between SNPs (Altshuler et al., 2010a). The default settings in FUMA were used to determine the number of independent loci from the meta-analysis (Watanabe et al., 2017). The sample size was specified for each SNP as n=9,918 if the SNP was tested in both populations, n=4,468 if the SNP was only tested in the German population or n=5,450 if the SNP was tested in the UK population only. This analysis maps all the input SNPs against all the protein coding genes across the genome, and to identify significant genes, a Bonferroni-adjusted P-value was used to correct for multiple testing.

3.3 Results

3.3.1 Functional annotation and candidate gene mapping

The functional relevance of the three regions associated with mastocytosis were explored using RegulomeDB and HaploReg to see if the risk SNP or their proxies (r^2 ≥0.8) were located in regions that might have regulatory functions based on alteration of transcription factor (TF) binding motifs, chromatin modification or DNA methylation profiles (Appendix Table A.8). Additional functional insights were gained by performing eQTL and mQTL analysis on the lead SNP and proxies using GTEx v8 and QTLbase (Carithers and Moore, 2015; Zheng et al., 2020). Lastly, SNPs were imputed and the stage 1 meta-analysis was repeated to fine map around the lead SNPs and generate association results for SNPs in high LD which had not been directly genotyped.

The most significant SNP, rs4616402, is located in an intergenic region at chromosome 19q13.11 between *SLC7A10* (36.8Kb downstream), a solute carrier gene, and *CEBPA* (37.2kb downstream), a gene encoding a transcription factor that co-ordinates differentiation and proliferation of myeloid progenitor cells (Figure 3.1A). QTLbase analysis showed that that rs4616402 is strongly associated with *CEBPA* expression in blood cells in three independent eQTL studies (P_{eQTL}=2.30×10^{-14}; P_{eQTL}= 2.96×10^{-11}; P_{eQTL}= 9.20×10^{-9}) (Lloyd-Jones et al., 2017; Võsa et al., 2018a; Westra et al., 2013). No additional SNPs were identified in high LD with rs4616402, but the RegulomeDB score for this SNP was 4, suggesting the possibility that it might have functional consequences. Specifically, the risk allele is predicted to alter three TF binding motifs (Arnt_1, Gm397 and Hmx_1, Appendix Table A.8). The chromatin structure surrounding rs4616402 shows an enrichment of H3K4me1 in primary haematopoietic stem cells, a histone mark (7_Enh) that is often associated with primed enhancers (Yao et al., 2020). No association between rs4616402 and expression of *SLC7A10* was found and there is no published evidence to suggest that *SLC7A10* has a role in the development or pathogenesis of cancer, including leukaemia.

The second most significant SNP, rs4662380, is located in the first intron of *LINC01412*, a lincRNA gene (Figure 3.1B) at chromosome 2q22.3. This SNP increases the risk of developing mastocytosis by 1.46. Twelve other SNPs in *LINC01412* were found to be in strong LD (r^2>0.8) with rs4662380 and were thus considered as proxies. Three of these are located in candidate enhancers (7_Enh: rs13413446, rs6722387, rs16823865) in primary haematopoietic stem cells and one (rs16823855) is located in the flanking region of an active transcription start site (2_TssAFlnk) in K562 cells (Table 3.3). The RegulomeDB scores suggest that two proxies affect binding of TFs; rs4662227

(score=2c) and rs13413446 (score=3a). The remaining proxy SNPs only had weak or no evidence for functional consequences. However, the GWAS catalog (Buniello et al., 2019) indicates that one of the proxies, rs16823866, has been strongly associated with white blood cell counts in two previous studies ($P=4\times10^{-18}$ and $P=6\times10^{-11}$) (Astle et al., 2016; Chen et al., 2020; Kanai et al., 2018). Lastly, QTLbase analysis indicated that the lead SNP rs4662380 ($P_{eQTL}=2.55\times10^{-11}$) and four proxies including rs16823866 ($P_{eQTL}=2.55\times10^{-11}$) were strongly associated with expression of a closely located gene, *TEX41*, in neutrophils (Chen et al., 2016).

The third SNP, rs13077541, is located at chromosome 3q26.32 in an intergenic region between transducin beta like 1 X-linked receptor 1 (*TBL1XR1*, 10.6kb upstream) and another lncRNA gene (*LINC00501*, 86.5kb upstream) (Figure 3.1C). This SNP is associated with a 1.33-fold increase in the risk of developing mastocytosis. Fifty-three additional proxy SNPs were identified to be in strong LD ($r^2>0.8$) with rs13077541, a number that includes 27 *TBL1XR1* intronic SNPs (Appendix Table A.8). Eleven of these proxies are located in active region of chromatin, including three in transcription start sites (1_TssA: rs34302523, rs12493005, rs12486557) and two in the 5' transcribed region (3_TxFlnk: rs34311793, rs35072945) in K562 cells. The RegulomeDB scores identified five proxies that are likely to affect TF binding (score2a-c: rs7616138, rs1920131, rs6790639, rs34302523 and rs6772872). Of these 5 SNPs, rs6790639 is particularly interesting, as the PU.1 TF, encoded by the Spi-1 proto-oncogene (*SPI1*), has been shown to bind to this region in K562 cells using chromatin immunoprecipitation analysis (Dunham et al., 2012). PU.1, together with other TFs, in known to regulate the expression of genes that are critical to myelopoiesis (Van Riel and Rosenbauer, 2014). Using QTLbase, the lead SNP ($P_{eQTL}=5.70\times10^{-8}$) and one of the proxies, rs16823866 ($P_{eQTL}=9.52\times10^{-9}$), were found to be strongly associated with *TBL1XR1* expression in CD4+ naïve T cells (Chen et al., 2016). In addition, there is evidence for the lead SNP being an mQTL, supported by results from a study conducted on five independent European populations (Gaunt et al., 2016a), showing that rs13077541 is associated ($p<1.03\times10^{-13}$) with a specific CpG site (cg001132484, chr3:176916496, rs1025797382) in blood (Appendix Table A.10). cg001132484 is located at the *TBL1XR1* promoter, 2KB upstream *TBL1XR1* and according to K562 methylation 450K Bead Array data from the ENCODE project, this is a partially methylated (200 < methylation score < 600) CpG site (UCSC). *LINC00501* has not been functionally described on LNCipedia database or CLC.

A SNP, rs58124832, yielding a suggestive association with mastocytosis was also investigated and the eQTL analysis indicated it to be strongly associated with the expression of both *TPSAB1* (P_{eQTL}<1.9×10^{-58}) and *TPSB2* (tryptase-β2; P_{eQTL}=1.96×10^{-75}) in blood (Lloyd-Jones et al., 2017; Võsa et al., 2018b).

3.3.2 Association with clinical features

To determine whether the three significant SNPs associated with mastocytosis are also associated with particular clinical features, Fisher's exact tests and linear regression were used to correlate allelic counts with clinical phenotypes in the Spanish and Italian cohorts (Table 3.4). These were the only cases with available clinical information. A significant association was found between rs4616402 and age at presentation (n=422; P=0.009; beta=4.41) in patients with non-advanced disease (Table 2.1) that remained significant after correction for multiple testing. No association with age was seen in the much smaller group of cases (n=26) with advanced disease, and it is important to note that this a subgroup for which additional mutations may be a confounding factor. In cases with non-advanced disease, the age of onset was estimated to increase by 4.41 years per risk allele. No associations were found by comparing allelic counts with gender, skin lesions, baseline tryptase levels, or disease phenotype.

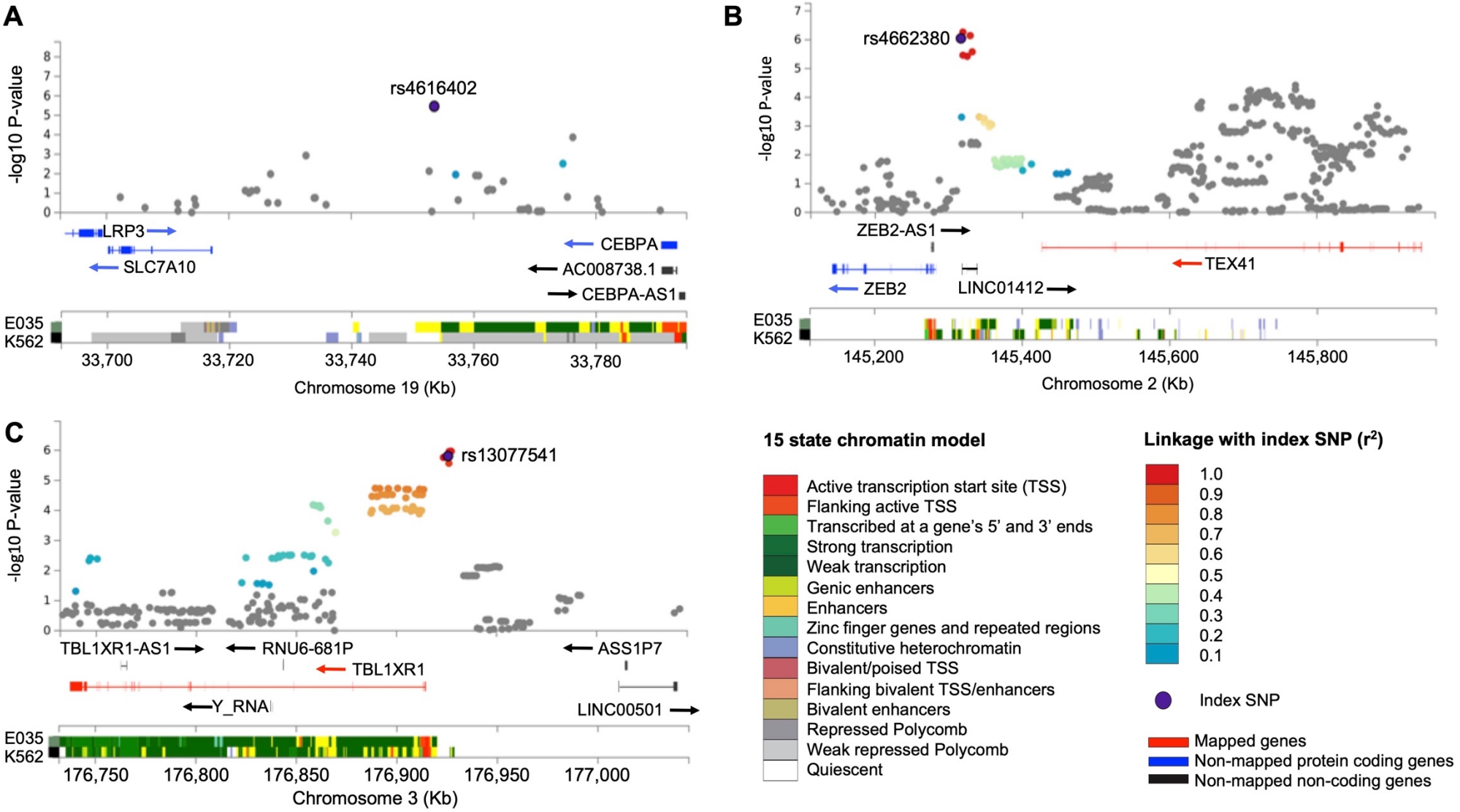

A
-log10 P-value
rs4616402
LRP3
SLC7A10
CEBPA
AC008738.1
CEBPA-AS1
E035
K562
33,700
33,720
33,740
33,760
33,780
Chromosome 19 (Kb)
B
-log10 P-value
rs4662380
ZEB2-AS1
TEX41
ZEB2
LINC01412
E035
K562
145,200
145,400
145,600
145,800
Chromosome 2 (Kb)
C
-log10 P-value
rs13077541
TBL1XR1-AS1
RNU6-681P
ASS1P7
TBL1XR1
Y_RNA
LINC00501
E035
K562
176,750
176,800
176,850
176,900
176,950
177,000
Chromosome 3 (Kb)
15 state chromatin model
Active transcription start site (TSS)
Flanking active TSS
Transcribed at a gene's 5' and 3' ends
Strong transcription
Weak transcription
Genic enhancers
Enhancers
Zinc finger genes and repeated regions
Constitutive heterochromatin
Bivalent/poised TSS
Flanking bivalent TSS/enhancers
Bivalent enhancers
Repressed Polycomb
Weak repressed Polycomb
Quiescent
Linkage with index SNP (r²)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Index SNP
Mapped genes
Non-mapped protein coding genes
Non-mapped non-coding genes

Figure 3.1 **Regional plots of the imputed stage 1 meta-analysis for SNPs reaching genome-wide significance in the final meta-analysis.**

Results from stage 1 meta-analysis using imputed SNPs in regions surrounding three lead SNPs (A, rs4616402; B, rs4662380 and C, rs13077541). In each plot, the lead SNP is indicated by a purple circle and the colours of other SNPs represent the strength of LD strength (r^2) with the lead SNP as indicated by the key. Protein coding genes and RNA genes are shown in the lower track with arrows to indicate the direction of transcription, and thick lines represent the location of exons. The lower part of the panel shows the 15 state chromatin track (chromHMM) in primary haematopoietic stem cells (E035) and K562 cells using data from the NIH Roadmap Epigenomics Consortium (Kundaje et al., 2015). Physical positions relate to build 37 (hg19) of the human genome. At the bottom right of the figure the colour of each candidate-state is indicated, followed by a chromatin state description.

Table 3.4 **Association between the most significant SNPs and clinical phenotypes in the Spanish and Italian cohorts.**

		rs4662380		rs13077541		rs4616402	
Phenotype	**No Cases**	**P value**	**Effect size (CI)**	**P value**	**Effect size (CI)**	**P value**	**Effect size (CI)**
Initial diagnosis (indolent/advanced)	422/26	0.175	0.58 (0.26–1.27)	0.646	0.88 (0.50–1.54)	0.238	0.60 (0.25–1.40)
Sex (F/M)	235/214	0.266	1.18 (0.88–1.60)	0.384	1.12 (0.86–1.46)	0.904	1.03 (0.65–1.61)
Skin lesions (+/−)	275/122	0.638	1.08 (0.77–1.51)	0.151	0.81 (0.60–1.08)	0.406	1.23 (0.75–2.00)
Age at diagnosis	422	0.668	0.55 (−1.97–3.07)	0.625	0.67 (−2.02–3.35)	0.009	4.41 (1.09–7.73)
Tryptase	417	0.452	−0.08 (−0.29–0.13)	0.136	−0.17 (−0.39–0.05)	0.249	0.17 (−0.12–0.45)
KIT^{D816V} Mutation burden	109	0.946	−0.16 (−4.96–4.63)	0.163	3.43 (−1.41–8.28)	0.648	−1.47 (−7.88–4.93)

Categorical phenotypes: Initial diagnosis (422 indolent vs 26 advanced mastocytosis cases), sex (235 female vs 214 male cases) and skin lesions (275 cases with skin phenotype vs 122 cases without skin phenotype); P value, fixed effects meta-analysis of Italian and Spanish Fisher's exact test; effect size, odds ratio; CI, 95% confidence interval**.** Continuous phenotypes: Age at diagnosis, tryptase levels and mutation burden tested in cases with non-advanced phenotype; P value, linear regression; effect size, regression coefficient beta; CI, 95% confidence interval**.**

3.3.3 Gene-based test

The gene-based analysis identified the vascular endothelial-derived growth factor C (*VEGFC*) gene as significantly (P-value=2.34 × 10^{-6}) associated with mastocytosis (Figure 3.2). A further 8 genes were found with P-value < 0.001, with the most significant one being *TPSAB1*. In Table 3.5 the results are shown for the 20 most significant genes. Input SNPs were mapped to 19,540 protein coding genes. The Bonferroni adjusted P-value of 2.559 × 10^{-6} was used after correcting for multiple testing; i.e. only *VEGFC* was significant after the multiple testing correction. At stage 1, the most significant SNP mapping to *VEGFC* is rs6820170 (P-value = 9.3×10^{-7}, P-value$_{meta}$= 1.69×10^{-4}; Appendix Table A.6), which is located in an intronic region (Figure 3.3). This signal is supported by 10 other SNPs in the clump (P <0.001). rs6820170 was tested at stage 2 but failed to replicate in the Spanish (P-value=0.65), Danish (P-value=0.44) and Italian cohorts (P-value=0.15). The second most significant SNP in the clump (P-value = 1.58×10^{-6}), rs11131764, is intergenic and was selected as a backup signal, however it failed genotyping at stage 2. Sixty three additional proxy SNPs were in strong LD (r^2>0.8) with rs6820170 (Appendix Table A.9). Of these, ten proxies had a RegulomeDB score of 2b (rs4146612, rs13132761) and 3a (rs3822038, rs1692787, rs1471813, rs1995083, rs2877967, rs7694268, rs2333530, rs3755972), suggesting that they could affect protein binding (Table 3.3). The lead SNP is predicted to alter two TF binding motifs (Hbp1; PRDM1_known1). The chromatin structure surrounding rs3755972, one of the SNPs in strong LD (r^2=0.89), shows an enrichment of H3K4me3 in primary haematopoietic stem cells, a mark which is often associated with promoter regions (Heintzman et al., 2007; Igolkina et al., 2019), and a bivalent enhancer (Table 3.2) characterised by two histone marks that can be associated with both activation or repression of transcriptional events, both of which are crucial during cell differentiation (Blanco et al., 2020). One SNP rs13122901 is in strong LD (r^2=0.91) with the lead SNP, and in the stage 1 meta-analysis of the imputed data it surpassed the genome-wide significance (P-value = 1.37 × 10^{-12}). QTLbase analysis showed that the lead SNP is strongly associated with chr4:177628507-177628507 methylation in blood (P-value$_{mQTL}$=3.1×10^{-6}) (McClay et al., 2015). An association between rs6820170 and VEGFC expression is only detected in the thyroid (GTEx2015_v6).

Table 3.5 **Results for gene-based association with mastocytosis.**

CHR	START	STOP	N SNPS	N PARAM	N	P	GENE
4	177594689	177723881	15	3	8959	2.34×10^{-6}	*VEGFC*
16	1280697	1302555	3	2	8101	0.00023405	*TPSAB1*
12	21907889	21938515	1	1	9918	0.0004032	*KCNJ8*
6	111398781	111562397	15	3	9555	0.00047077	*SLC16A10*
4	79798281	79870592	2	1	9918	0.00073203	*PAQR3*
5	179068298	179089445	1	1	4468	0.0007667	*AC136604.1*
19	4219495	4247528	4	1	9918	0.00084005	*EBI3*
19	35605417	35643355	12	3	9918	0.00094844	*LGI4*
5	133474633	133522729	5	3	8828	0.00097557	*SKP1*
20	5272317	5307378	20	4	9373	0.0010449	*PROKR2*
5	1307859	1355214	11	2	9016	0.0010621	*CLPTM1L*
23	135034229	135066222	6	1	9918	0.0011011	*MMGT1*
3	31689382	32129072	128	42	8613	0.0011201	*OSBPL10*
7	140362953	140406061	9	2	9918	0.001205	*ADCK2*
6	111570551	111602370	5	2	9918	0.0012344	*KIAA1919*
1	182859000	182932660	5	2	7934	0.0013396	*SHCBP1L*
11	93201638	93286674	28	8	7972	0.0016447	*SMCO4*
14	23379720	23408794	5	1	9918	0.0018362	*PRMT5*
19	45302328	45334673	8	4	9918	0.0019452	*BCAM*
14	23430383	23461851	4	2	9918	0.0020915	*AJUBA*

CHR: Chromosome; START/STOP: Gene boundaries annotated on build hg19; N SNPS: Number of SNPs mapping the gene including SNPs that are in 10Kb window both directions; N PARAM: Number of relevant parameters used in the model; N: Sample size used for analysing the gene; P: The gene P-value computed using MAGMA; GENE: Gene name.

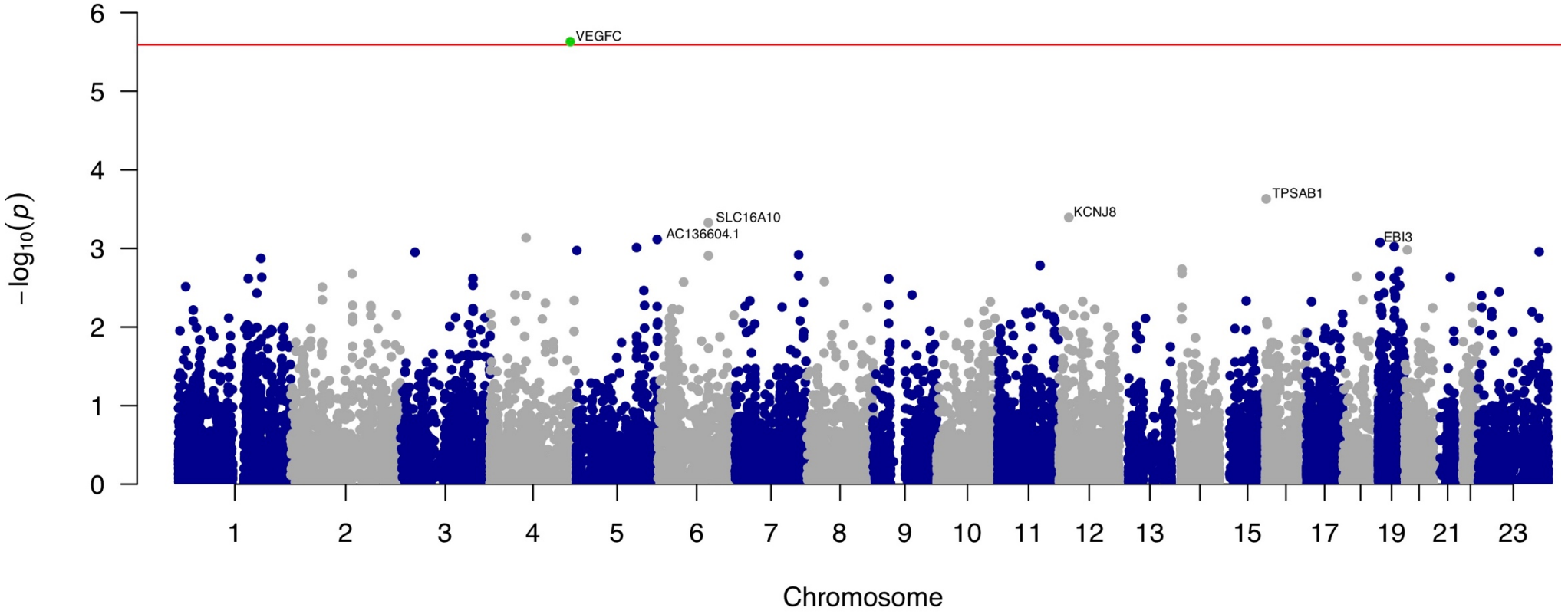


Figure 3.2 **Results of the gene-based associations of *KIT*^{D816V}-positive mastocytosis.**

Manhattan plot showing results of the gene-based test as computed by MAGMA and based on the summary statistics of the stage 1 meta-analysis for all 24 chromosomes. Genome-wide significance (red line in the plot) was defined at an adjusted P-value of 2.6×10^{-6}. Results for 19,540 genes are plotted as −log10 of the meta-analysis P-values on the y-axis against genomic location on the x-axis. *VEGFC* gene, highlighted by the green circle, was identified with genome-wide significance (P-value $<2.6\times10^{-6}$). As shown in Table 3.5, a further 8 genes were identified with P-value $<1\times10^{-3}$ and the top genes per chromosome that exceed this threshold are labelled in black.

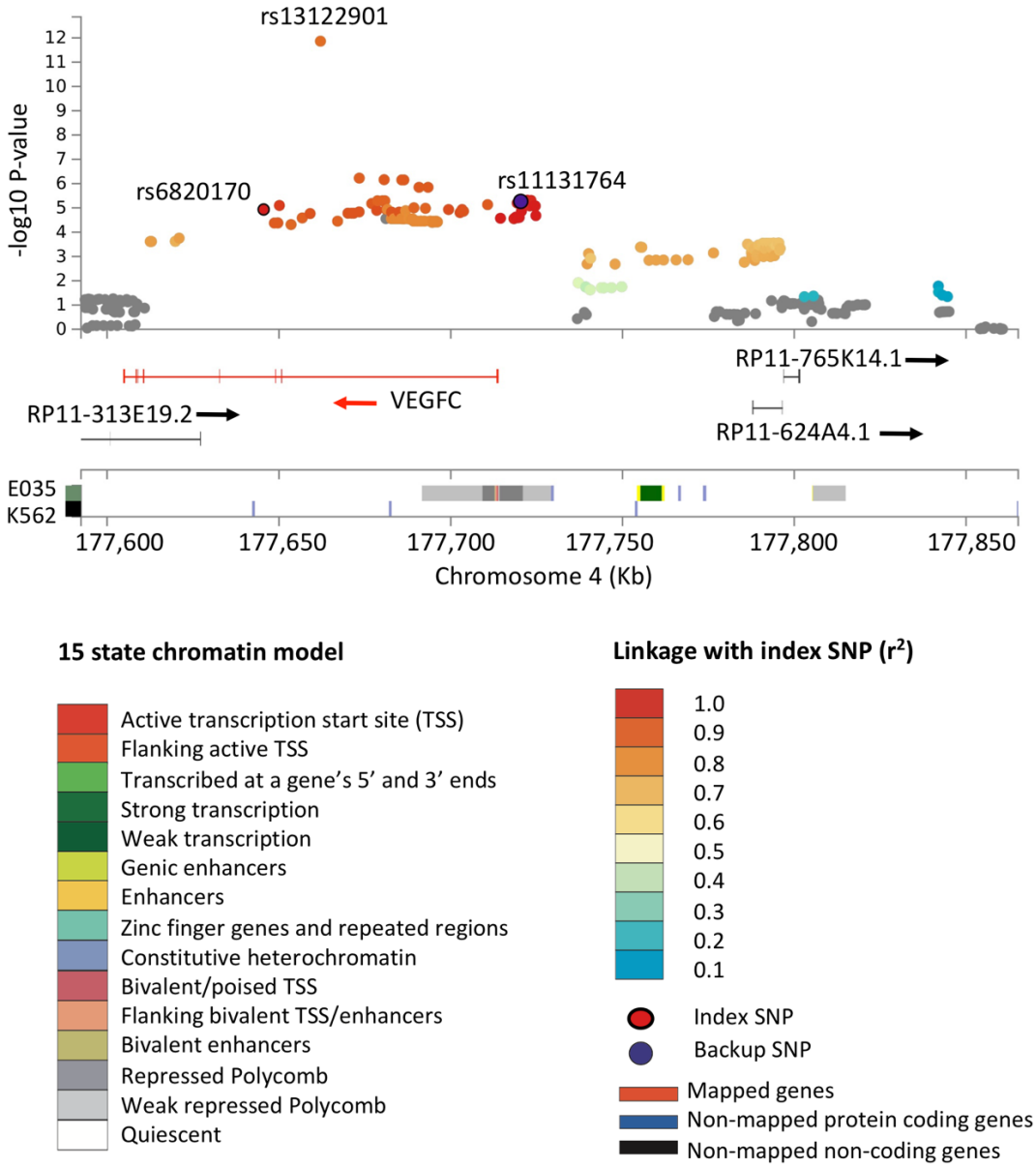


Figure 3.3 **Regional plot of the imputed stage 1 meta-analysis for VEGFC SNPs selected for stage 2.**

Results from stage 1 meta-analysis using imputed SNPs in regions surrounding the *VEGFC* signals. The lead SNP (rs6820170) is identified by a red circle and the purple circle indicates the backup SNP (rs13122901). The colours of other SNPs represent the strength of LD (r^2) with the lead SNP as indicated by the key. The most significant SNP is also labelled rs11131764 and the colour shows that it is in strong LD (r^2>0.8) with the index SNP. Protein coding genes and non-coding genes are shown in the lower track with arrows to indicate the direction of transcription, and thick lines represent the location of exons. The bottom of the panel shows the 15 state chromatin track (chromHMM) in primary haematopoietic stem cells (E035) and K562 cells using data from the NIH Roadmap Epigenomics Consortium (Kundaje et al., 2015). Physical positions relate to build 37 (hg19) of the human genome. At the bottom left of the figure the colour of each candidate state is indicated, followed by a chromatin state description.

3.4 Discussion

In theory, common inherited genetic variation might influence the development and diagnosis of mastocytosis by a number of different mechanisms. First, genetic variation could promote or favour the outgrowth of a *KIT*^{D816V}-positive clone that arose by random mutation in a haematopoietic stem cell (fertile ground hypothesis). Second, it might increase the probability of a *KIT*^{D816V} mutation arising in a stem cell, possibly as a consequence of a generally increased mutation rate (hypermutability hypothesis). Third, it might promote or exacerbate the development of clinical symptoms such as rash, itching or abnormal blood counts in a patient with a *KIT*^{D816V}-positive clone, thereby increasing the chance that the patient might seek medical help (phenotypic hypothesis). These potential mechanisms are considered below for each association.

Focusing on the three significant SNPs identified associated with mastocytosis, the strongest association was found for rs4616402 at chromosome 19q13. Interestingly, this SNP was associated with age at diagnosis for patients with non-advanced disease. rs4616402 is located in a predicted enhancer, and the risk allele is associated with lower expression of *CEBPA* (Lloyd-Jones et al., 2017), which is located 37.3kb upstream. Another SNP at 19q13, rs78744187, has been linked to basophil counts in a previous study and also been shown to affect the activity of another *CEBPA* enhancer (Guo et al., 2017), but this variant is in weak LD ($r^2 = 0.22$) with rs4616402 (Arnold et al., 2015) and therefore cannot account for the association observed in this study. Of potential relevance, high C/EBPα expression is known to inhibit the generation of mast cells from mast/basophil common progenitors, while low C/EBPα expression inhibits the generation of basophils (Bick et al., 2020). Although the consequence of different expression levels of C/EBPα in the presence of *KIT*^{D816V} remains to be defined experimentally, it is plausible that reduced *CEBPA* expression linked to the rs4616402 risk allele may be relevant to both the fertile ground and phenotypic hypotheses defined above by promoting a cellular environment that favours mast cell production. Low *CEBPA* expression is a common feature of AML, although the underlying mechanism is unclear (Avellino and Delwel, 2017). Potentially, rs4616402 might be relevant to this observation and it would be interesting to genotype this SNP in AML and relate the findings to *CEBPA* expression. Overall, it is clear that detailed functional studies are required to understand the relationship between *KIT*^{D816V}-driven mastocytosis and *CEBPA* expression.

The second most significant SNP, rs4662380, at chromosome 2q22.3, is associated with elevated expression of the nearby lncRNA *TEX41*. The role of *TEX41* in promoting tumour progression has

been described in other human cancers (Li et al., 2018; Yang et al., 2019) although the mechanism is not understood. For the first time, we have associated *TEX41* with myeloid cancer, however the functional involvement of *TEX41* in mastocytosis is not clear and needs to be investigated.

ZEB2 is another nearby gene but no association was found between rs4662380 and *ZEB2* expression. It is important to note that mRNA expression does not always reflect protein abundance. Identifying pQTLs also becomes necessary for understanding the direct effect of genetic variants on protein abundance (Robins et al., 2021). In fact, Robins et al., after comparing eQTLs and pQTLs in the brain, identified an overlap between pQTLs and eQTLs but not *vice versa*. Interestingly, similar results were also reported in human blood, making this evidence more generalisable to other tissues and cell types (Emilsson et al., 2018; Sun et al., 2018).

In another study, rs16823866, a SNP in LD with rs4662380 ($r^2 = 0.99$), was associated with elevated white blood cells and specifically with elevated basophil counts in three independent population-based studies (Astle et al., 2016; Kanai et al., 2018; Vuckovic et al., 2020). Mast cells and basophils are highly related, with basophils being found mainly in the peripheral blood whereas mast cells are resident within tissues. Although the mechanism underlying this association is unclear, this finding suggests that an association between rs4662380 and mastocytosis may be relevant to the phenotypic hypothesis, since individuals with abnormal blood counts may be more likely to be investigated clinically.

The risk allele for the third SNP, rs13077541, at chromosome 3q26.32, is linked to reduced expression of *TBL1XR1* (Chen et al., 2016). The same SNP is also significantly related to the methylation level of a specific CpG site (cg001132484, chr3:176916496, rs1025797382) located at the *TBL1XR1* promoter (Gaunt et al., 2016b), suggesting that change in gene expression could be affected by methylation of this site. This gene fuses to *PDGFRB, ROS1*, *RARA* and *RARB* as a consequence of rare chromosomal translocations in myeloid malignancies (Campregher et al., 2017; Murakami et al., 2018; Osumi et al., 2018) but the significance of altered expression in relation to mastocytosis remains to be established.

MAGMA, a gene-based association method widely used with GWAS summary statistics (Marioni et al., 2018), was applied to the stage 1 mastocytosis GWAS to combine stage 1 association statistics from all SNPs within a gene (de Leeuw et al., 2015). Following this statistical analysis, *VEGFC* was the only significant gene after Bonferroni correction (P-value of 2.6×10^{-6}). Two SNPs mapping *VEGFC* were selected for replication. The most significant SNP in *VEGFC* (P-value = 9.3×10^{-7}, P-value$_{meta}$= 1.69×10^{-4}; Appendix Table A.6), rs6820170, was tested at stage 2 but failed to replicate in the Spanish (P-value = 0.65), Danish (P-value = 0.44) and Italian (P-value = 0.15) cohorts. This could be due to a number of reasons such as a lack of power at stage 2, or heterogeneity between cohorts. Therefore, it does not rule out the stage 1 result and this signal should be tested in other independent cohorts.

VEGFs are members of a family of proteins (e.g. VEGFA, VEGFB, VEGFC, VEGFD, VEGFE) that are very important in vasculogenesis and angiogenesis; VEGFC and VEGFD are known to be mainly involved in lymphangiogenesis in hyperplasia of the skin (Apte et al., 2019; Jeltsch et al., 1997). In the last 20 years, the role of VEGFs in the pathogenesis of cancer and non-malignant disorders such as ophthalmic diseases has become clear, since the continuous growth of blood vessels carrying nutrients is crucial to maintain homeostasis within the tissue environment (Apte et al., 2019). Interestingly, a study conducted with 64 mastocytosis cases and 64 healthy controls evaluated the serum concentration of three VEGFs and identified that both VEGFA and VEGFC levels were significantly higher in mastocytosis patients. VEGFD did not show the same pattern (Marcella et al., 2021). Whether elevated VEGF levels are a cause or consequence of mastocytosis, and whether SNP genotype within *VEGFC* is linked to expression levels in serum should be investigated further.

Mast cell activation followed by degranulation leads to the release of several bioactive molecules, including histamine, tryptase and proinflammatory cytokines (Frenzel and Hermine, 2013). A GWAS was performed in relation to levels of circulating cytokines and growth factors to gain insight into inflammatory diseases that might share common causal pathways and underlying pathology. Ahola-Olli et al. identified a SNP, rs6921438, in the *VEGFA* locus associated with concentration of five cytokines (VEGF, IL-7, IL-12p70, IL-10 and IL-13) (Ahola-Olli et al., 2017). Another GWAS in a European population also identified the same lead SNP, rs6921438 (P-value=6.11×10^{-506}), associated with circulating VEGF levels (Debette et al., 2011). Conditional analysis on rs6921438 identified another SNP in the same locus, rs12214617, located in the promoter flanking region of *VEGFA*, which suggests a potential role in regulation of transcription. The role of VEGF as an upstream regulator has been supported by Mendelian randomisation

performed using both SNPs (Ahola-Olli et al., 2017). Particularly interesting in relation to mastocytosis is the association identified between IL-13 and *VEGFA* and the role that IL-13 could play in the pathogenesis of mastocytosis. As described in Chapter 2 (see 2.3.15), rs1800925 at IL-13 was linked to mastocytosis in a previous study and this link was supported by my stage 1 analysis (Table 2.11).

The statistical analyses conducted with mastocytosis patients identified significant associations with tryptase alpha/beta-1 (*TPSAB1*) and tryptase beta-2 (*TPSB2*), two genes located at 16p13.3, a region known as the human tryptase locus. The GWAS identified a suggestive association between rs58124832 and mastocytosis. The eQTL analysis revealed this locus to be associated with the gene expression level for *TPSAB1* and *TPSB2*. The association between *TPSAB1* and mastocytosis was also revealed from the gene-based analysis. *TPSAB1* and *TPSB2* encode serine protease produced largely by mast cells. While *TPSAB1* only encodes the α-tryptase, both *TPSAB1* and *TPSB2* encode the β–tryptase (Schwartz et al., 1981). Levels of BST >20 ng/mL represent one of the minor diagnostic criteria for SM that were confirmed by the WHO in 2008 and updated in 2016 (Valent et al., 2017a). However, this is not always the case. Elevated BST level in association with clinical features (e.g., gastrointestinal and cutaneous symptoms) are seen in 4–6% of the general population with no mastocytosis or mast cell activation. A study performed in 35 families linked *TPSAB1* duplication and triplication to a significant increase in BST level. The correlation between phenotype and gene dose, an inherited phenotype known as hereditary α–tryptasaemia, was demonstrated by designing a digital droplet polymerase chain reaction (ddPCR) genotyping assay to identify duplication or triplication of α–tryptase (Lyons et al., 2018). In mastocytosis, an increased BST level reflects the increased mast cell burden in mastocytosis patients (Schwartz et al., 1995). Another study, consistent with Lyons et al., used the ddPCR assay to assess the *TPSAB1* CNV and compare with tryptase levels in mastocytosis patients. They showed that the prevalence of hereditary α–tryptasaemia and associated BST levels were significantly higher in mastocytosis cases compared to control cohorts. After demonstrating the correlation between *TPSAB1* CNV and mastocytosis, *TPSAB1* CNV was proposed as a novel genetic biomarker to predict the risk of severe anaphylaxis in patients with mastocytosis (Greiner et al., 2021).

In my analysis, a SNP located in the exonic region of *CACNA1H*, rs58124832, reached a suggestive level of significance after meta-analysis. This SNP is part of a haplotype that co-segregates with

the *TPSAB1* CNV in Caucasian families (Lyons et al., 2018) and is strongly associated with higher expression of both *TPSAB1* and *TPSB2* (Lloyd-Jones et al., 2017; Võsa et al., 2018b). Interestingly, *TPSAB1* is also the second most significant gene (P-value=2.3×10^{-4}) identified from my gene-based test analysis. However, it did not retain significance after correcting for multiple testing, and thus this result must be confirmed in an independent cohort with genome-wide genotyping. These results need to be investigated further to understand the functional involvement of *TPSAB1* in modulating disease severity in mastocytosis, and potentially to identify therapeutic approaches to modulate the α-tryptase-dependent response. Although the *TPSAB1* CNV is associated with disease severity in mastocytosis (Greiner et al., 2021), it seems likely that elevated BST levels associated with the CNV are also related to the phenotypic hypothesis since patients with high BST may be more likely to be investigated for KIT^{D816V} and/or be diagnosed with mastocytosis.

Chapter 4 Identification of genetic targets of acquired uniparental disomy

4.1 Introduction

Uniparental disomy (UPD), described in 1980 by Engel, is usually associated with congenital abnormalities and arises when two copies of a chromosome or part of a chromosome are inherited from one parent (Engel, 1980). However, UPD can be somatically acquired (aUPD) through mitotic recombination or non-disjunction errors (Figure 1.3) and is strongly associated with the presence of cancer driver mutations in the affected region (Tuna et al., 2009). It is believed that an initial somatically acquired driver mutation promotes clonal expansion, but subsequent aUPD converts this mutation to a homozygous state which then confers an additional clonal advantage. Array based studies have indicated that aUPD is widespread in cancer, including up to 30% of cases of myeloid malignancy. Specific gene targets have been identified for the most common recurrent regions, for example *JAK2* mutations are associated with aUPD of the short arm of chromosome 9 (9p), *TET2* mutations with aUPD 4q, *EZH2* mutations with aUPD7q, *CBL* mutations with aUPD 11q and several others (Chase et al., 2015; Ernst et al., 2010; Grand et al., 2009; Kralovics et al., 2002; Langemeijer et al., 2009; Massé et al., 2009; Mohamedali et al., 2009; Nikoloski et al., 2010; O'Keefe et al., 2010; Raghavan et al., 2008; Score and Cross, 2012; Tiedt et al., 2005; Wang et al., 2016). However, there are other regions for which the genetic targets have not been identified. Thus, I hypothesise that better definition of recurrent aUPD in myeloid disorders will help to identify regions of the genome that harbour novel cancer driver genes.

Myeloid neoplasms are relatively uncommon and SNP array analysis of large numbers of individuals is expensive. However SNP microarray data from diverse GWAS using DNA extracted from blood cells have identified mosaic abnormalities such as aUPD in individuals unselected for haematological malignancies (Jacobs et al., 2012; Laurie et al., 2012). These studies were extended by WES analysis of 29,562 individuals with the finding that somatically acquired myeloid driver mutations (particularly *DNMT3A, TET2* and *ASXL1*) are unexpectedly common in the population at large (Genovese et al., 2014; Jaiswal et al., 2014). Although aUPD occurs at a lower frequency in these individuals it has been shown to increase with age (only 1% in individuals less than 50 years, 2-3% in individuals over 50 years old and 10% in the elderly aged 65 and older) (Genovese et al., 2014; Jacobs et al., 2012; Laurie et al., 2012). Furthermore, clonal mosaicism in

the elderly is associated with a tenfold increased risk of developing haematological cancer and these regions of aUPD and the underlying somatic mutation are the same as those identified in both mature B-cell neoplasms and myeloid malignancies (Laurie et al., 2012). Owing to these features, aUPD in apparently healthy individuals is now recognized as a specific condition which is termed CHIP or ARCH (Genovese et al., 2014; Jacobs et al., 2012; Laurie et al., 2012). Increasingly large publicly available datasets of WES and WGS data derived from blood cells are also being accumulated which can be interrogated directly for aUPD and mutations. I propose to exploit such datasets to identify regions of aUPD and associated genetic targets.

As a proof of principle, the Cross/Tapper research group identified five cases with aUPD22q, of which three cases had a known myeloid malignancy and two were identified from a Swedish population-based study of elderly men. WES analysis identified a novel gene, *PRR14L,* as the target of aUPD22q. Although the function of *PRR14L* is still unknown, functional studies suggested its involvement in cell division (Chase et al., 2019).

Although aUPD is predominantly associated with somatic mutations in specific genes, other mechanisms have been described. The Cross/Tapper research group identified a minimal recurrent region of aUPD involving 11.2 Mb on chromosome 14q which contained an imprinted region (*DLK1-MEG3)*. WES failed to identify any recurrently mutated genes in affected individuals but testing the *DLK1-MEG3* methylation status in cases with aUPD14q in blood cells showed an increase in methylation, which is associated with the gain of the paternal chromosome, and demonstrated for the first time that aUPD14q can target an imprinted locus and can promote clonal haemopoiesis either as an initiating event or as a secondary change (Chase et al., 2015). This represents the first imprinted locus targeted by both somatically acquired UPD as well as constitutional UPD in association with the developmental disorders Temple syndrome and Kagami-Ogata syndrome.

Another study has demonstrated that aUPD may be associated with the loss of a deleterious germline variant. Specifically, two families with cytopenia and predisposition to MDS showed the coexistence of rare (0.00003% of frequency reported in the Genome Aggregation database (gnomAD)) inherited mutations and aUPD. Germ line gain-of-function heterozygous mutations were identified on *SAMD9L* a tumour suppressor gene located at chromosome 7q. Mutated *SAMDL9* is associated with impaired haemopoiesis; UPD7q in this context leads to clonal

restoration of homozygous wild type *SAMDL9* with a selective advantage over the mutant background (Tesi et al., 2017).

To test the hypothesis that large population cohorts which are unselected for haematological malignancies can be used to identify recurrent regions of aUPD and associated mutations, I have focused on WES data obtained from the UK Biobank (Van Hout et al., 2020) and a Swedish case control study of Schizophrenia (Purcell et al., 2014) consisting of 49,996 and 12,380 individuals respectively. The UK Biobank dataset was used to develop a step-wise method for identifying aUPD from WES data based on extended regions of AI that were detected using B allele frequency (BAF) segmentation (Staaf et al., 2008). The Biobank data are ideal for method development due to the availability of both WES and array based genotype data, which are optimal for BAF segmentation and have previously been used to identify aUPD (Dawoud et al., 2020). These array-based aUPD calls were used for comparison. The Schizophrenia data were used for further validation of the method and were selected as an exemplar WES cohort as (i) the study group was unselected for haematological disorders, (ii) DNA from peripheral blood cells was used for analysis, and (iii) the median age of the study was relatively old at 65 years and this would be expected to be enriched in clonal abnormalities compared to younger populations. The specific aims of this work are to develop a method for identifying regions of likely aUPD using WES data and identify candidate mutated genes in the affected regions that could be responsible for the development of myeloid malignancies and associated with clonal proliferation. The final analysis was focused on known genes with relatively frequent mutations, specifically *MPL*, *TET2*, *EZH2*, *JAK2* and *FLT3* mutations as well as exploring the possibility of discovering new aUPD regions overlapping in multiple samples.

4.2 Materials and Methods

4.2.1 The data sample

The large-scale cohort used for the analysis of aUPD regions comes from the Sweden-Schizophrenia Population-Based Case-Control Exome Sequencing study (dbGaP Study Accession: phs000473.v2.p2). The dataset is publicly available through the Database of Genotypes and Phenotypes (dbGaP, https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000473.v2.p2) distributing genotype datasets from studies which

investigate the interaction between genotypes and phenotypes (Mailman et al., 2007) and developed by the National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov/). The second version of Sweden- Schizophrenia Population-Based Case-Control cohort used for this study was released in October 2016 and in this report, I will refer to this cohort as Schizo-WES02. Details of the schizophrenia data had been described previously and a brief description is here provided (Ganna et al., 2016; Purcell et al., 2014; Ripke et al., 2013). The peripheral blood (PB) sampling of 12,380 subjects (6,135 cases and 6,245 controls) aged between 19 and93 years old (mean age 65) took place between 2005 and 2013 (Genovese et al., 2014) and were selected either from the Swedish National Hospital Discharge Register or from Swedish population registers.

The UK Biobank (UKB) is an open access resource available to the scientific community that wish to conduct health-related research studies for a wide range of diseases and without establishing collaborations. This large population-based prospective study combines baseline, genotypic and phenotypic data from 500,000 participants aged between 40-69 (mean age 56.5), recruited between 2006 and 2010 and assessed in 22 centres in the UK (Sudlow et al., 2015). The UKB contains different sources of genetic data: 1) genome-wide array based genotyping performed on all UKB participant and in coordinates relative to GRCh37, this data has allowed novel discoveries though population genetic analyses (Bycroft et al., 2018); 2) whole exome sequencing (WES) is performed on 49,997 participants and in coordinates relative to GRCh38. These participants are prioritized because of more complete phenotype data and are available since March 2019. One sample was removed because it did not have enough DNA for sequencing. Data for an additional 150,000 participants was made available in October 2020, after starting this analysis. For this reason, WES data of the first released 50,000 samples were used for this research project and in this thesis this cohort will be referred to as UKB-WES50.

Initially, an exemplar dataset consisting of 120 samples from the UKB-WES50 were selected for method development. According to the matched array data these samples included 17 with aUPD (chr2=1; chr6=4; chr9=9; chr13=2; chr17=1) and 40 with the somatic *JAK2*^{V617F} mutation, which are more likely to have 9p aUPD. Most of the exemplar samples were free from all types of cancer (n=64) while 35 were diagnosed with a haematological malignancy and 21 were diagnosed with other types of cancer.

4.2.2 Whole-Exome Sequencing

Whole-Exome Sequencing is a NGS technique to capture the whole exonic sequences in the genome that are involved in coding for proteins. Following the drop of sequencing costs, WES data have been produced in numerous genetic studies, for which array-based genotyping were previously available. In the context of aUPD analysis, WES data is expected to be more useful than SNP array data because it should enable researchers to investigate somatic mutations in specific genes associated with regions of aUPD.

Sequencing, alignment and variant calling of the Swedish-WES02 cohort were all performed at the Broad Institute. The samples were sequenced using either the Agilent SureSelect Human All Exon Kit targeting 29 Mb of the human genome or the Agilent SureSelect Human All Exon v.2 Kit targeting 33 Mb of the human genome. Sequencing was performed on IlluminaGAII, Illumina HiSeq2000 or Illumina HiSeq X Ten instruments, with pair ended sequencing reads of 75 base pairs and mean target coverage of 90x (Ganna et al., 2016). After completing the sequencing step, the Picard/Burrows-Wheeler Aligner (BWA)/ Genome Analysis Toolkit (GATK) pipeline was used to analyse the raw read data (BAM file). During alignment, a bioinformatic tool called Picard (http://broadinstitute.github.io/picard/) was used to perform data pre-processing and intermediate analyses (manipulation of FASTQ and SAM files, marking duplicate reads, filtering, sorting). BWA is another bioinformatic tool used during the alignment to map reads against the reference human genome (version GRCh37) and generate outputs in the Sequence Alignment Map (SAM) format (Li and Durbin, 2009; Teo et al., 2007). For the downstream analysis, GATK tool was used to process the SAM files and calling variants (Depristo et al., 2011) in the Variant Call File (VCF) format (Danecek et al., 2011). Variant calls were made on the entire sample creating a single multi-sample VCF that, following relevant approvals, was downloaded from the online dbGaP through Aspera Connect v3.6.2 and NCBI SRA Toolkit.

The UKB-WES50 VCFs were released in March 2019, and were pre-processed by Regeneron Genetics Center and GlaxoSmithKline using two protocols, Functional Equivalence (FE) (Regier et al., 2018) and Regeneron Seal Point Balinese (SPB) (Van Hout et al., 2020). In August 2019, the UKB reported an issue in marking duplicate reads, this was only limited to the exome data processed with the SPB pipeline. Therefore, for this analysis that started in November 2019, only the data produced using the FE pipeline was used (http://www.ukbiobank.ac.uk/wp-

content/uploads/2019/08/UKB-50k-Exome-Sequencing-Data-Release-July-2019-FAQs.pdf). Genomic DNA samples were transferred from the UKB to the Regeneron Genetics Center and stored at -80 C prior to sample preparation. Exome capture was performed using a fully-automated approach developed at the Regeneron Genetics Center. A slightly modified version of IDT's xGen probe library was used and supplemental probes were added to capture regions of the genome poorly covered by the standard xGen probes. In total, 39 Mbp of the human genome (19,396 genes) were included in the targeted regions. The multiplexed samples were sequenced using 75 bp paired-end reads with two 10 bp index reads on the Illumina NovaSeq 6000 platform using S2 flow cells. Complete sequencing protocols are described in detail by the summary manuscript (Van Hout et al., 2020). Following the completion of sequencing, raw data were converted into FASTQ files using the DNAnexus platform. During the alignment, BWA-mem was used to align the FASTQ-formatted reads to the GRCh38 reference human genome in the BAM file (Li and Durbin, 2009), Picard MarkDuplicates tool was used to flag duplicate reads. GATK 3.0 was used for the variant calling and a gVCF was generated for each sample. Then files were subject to hard filtering of variants with inbreeding coefficient<-0.03 or without at least one variant genotype of DP≥10, GQ≥20 and, if heterozygous, AB≥0.20 **(**http://biobank.ctsu.ox.ac.uk/showcase/label.cgi?id=170)**.**

4.2.3 Variant Quality Score Recalibration

Variant Quality Score Recalibration (VQSR) was used to improve the accuracy of the confidence score for each variant in the WES VCFs. An automated pipeline was developed to prepare the UKB-WES50 data and apply VQSR. Initially the gVCFs were indexed using IndexFeatureFile (McKenna et al., 2010). The single sample gVCFs were loaded into a datastore using GenomicDBImport, then GenotypeGVCFs was used to generate multi-sample VCFs in which all samples have been jointly genotyped (Auwera et al., 2014). The process of merging generated batches of 100 samples, except for the last batches which contained 174 and 122 samples. A total of 499 multi-sample VCF files were generated. The VQSR was performed separately for SNPs and indels. The recalibrated scores were then used to exclude low quality variants by applying a minimum threshold of phred>=20 for both SNPs and indels which is equivalent to 1% chance of error. These settings are in-line with the GATK best practice guidelines which recommended applying VQSR on at least 30 WES samples so that there are enough variant sites to apply the Gaussian mixture model.

To efficiently manage file preparation and VQSR on such a large number of VCFs, job arrays and dependencies were used to submit a maximum of 64 simultaneous jobs to the university's high

performance computer (HPC, IRIDIS4). To avoid interference between the parallel jobs, an empty directory with a unique name was generated for each job that was used as the workspace for GenomicsDBImport.

4.2.4 WES data processing

Having created multi-sample VCFs and performed VQSR, the next steps were to extract the data into single-sample VCFs, perform QC filtering and to generate the input files for BAF segmentation software (Staaf et al., 2008) that was used to identify regions of AI. BCFtools (Danecek et al., 2021) was used to extract single sample VCFs, to annotate variants' name based on the information from dbSNP build 151 (available from https://ftp.ncbi.nlm.nih.gov/snp/organisms/) and to apply the QC measures aimed at excluding variants that were: mitochondrial, had missing genotypes, had a mean depth (DP) less than 10, quality score (Qual) less than 20, a recalibrated quality score less than 20, minor allele frequency less than 1%. Filtered VCFs were compressed using BGZip and indexed with Tabix, which is a standard way to store VCFs to aid efficient manipulation (Li, 2011). Finally, input files for BAF segmentation were generated using VCFtools (Danecek et al., 2011) to extract the allelic depth (AD), genotype (GT) and rsID for each variant in the filtered VCFs. Variants without an rsID were named according to their genomic location (chr:position) using a custom python script (mkVAF.py). It is widely reported in the literature that genotypes generated in WES data can be prone to higher level of genotyping errors compared to array-based technologies (Carson et al., 2014; Koboldt et al., 2010; Ledergerber and Dessimoz, 2011; Nielsen et al., 2011). Studies have also demonstrated that variants with these type of errors may remain after applying GATK's VQSR filter (Van der Auwera and O'Connor, 2020; O'Rawe et al., 2013). To address this issue the mkVAF.py script was used to check the AD and GT variables and remove false positives that were incorrectly called as heterozygous (0/1 and 1/2) or homozygous (0/0 and 1/1) in the absence of reads supporting either the reference or alternate alleles. Variants were removed if their genotype information was discordant with the AD number supporting either the reference or alternate allele. Furthermore, the python script checks for multiple entries at the same location and excludes duplicate variants. As shown at the bottom of Figure 4.1, the generated input file contains four columns: marker name, chromosome, base pair location and variant allele frequency (VAF=A/D, where A is the number of reads with the alternate allele and D is the total depth). The input for BAF segmentation usually contains a measure of copy number for each variant which can be used to categorise AI regions as either copy number gain, loss or copy number neutral (aUPD). In SNP-array data, copy number is measured by the log R ratio (LRR) which compares observed and reference probe intensities [$\log_2$(observed intensity/reference intensity)]. The same calculation of LRR is not possible for unpaired sequence

data although copy number could be determined in these samples by comparing sequencing depth from the Binary Alignment/Map (BAM) file with reference regions from elsewhere in the exome (Straver et al., 2017). However, at this stage of the study, alignment file in BAM format were not available for CNV classification.

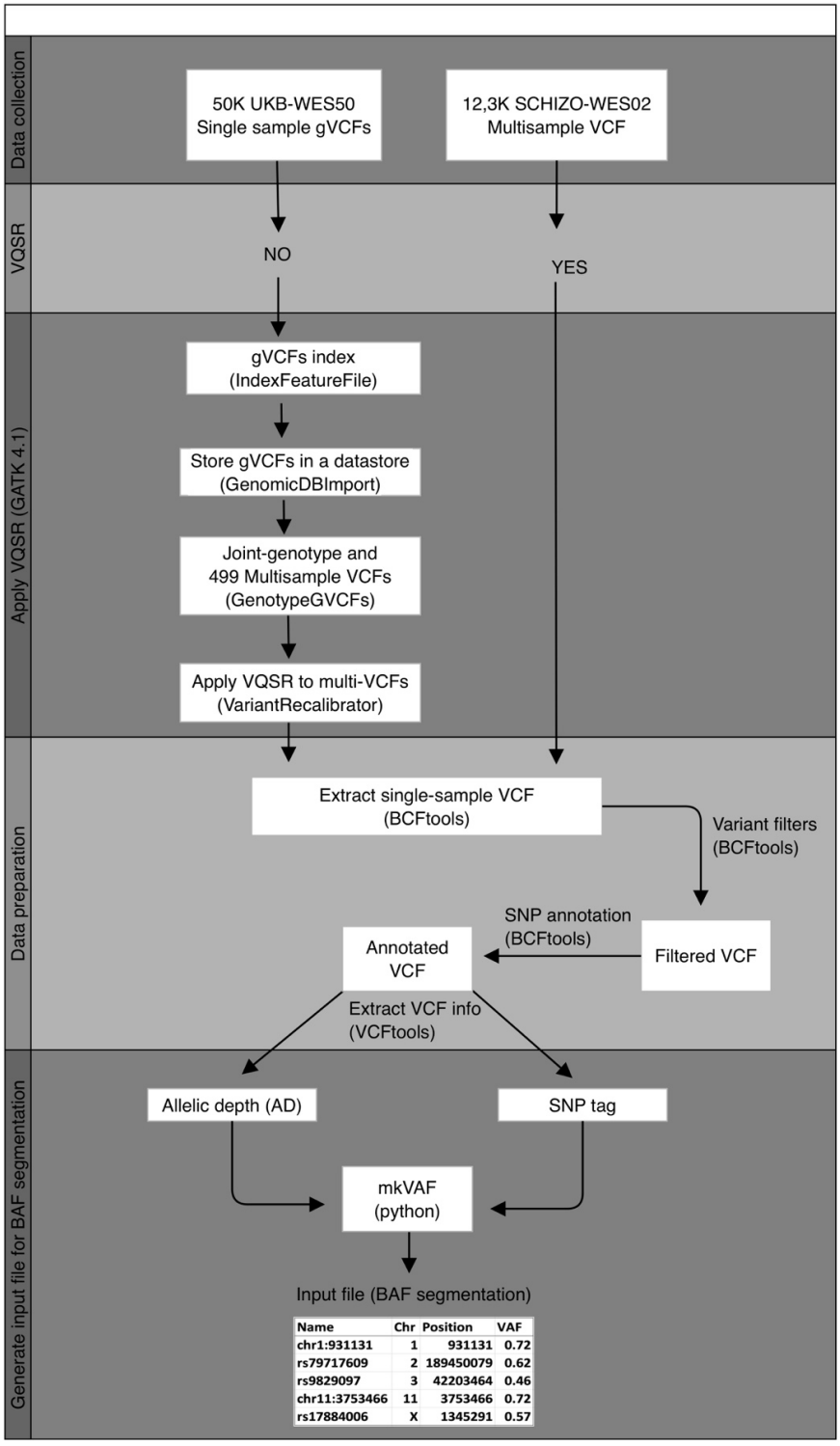


Figure 4.1 **Automated pipeline to process WES data.**The flowchart outlines all the steps performed to process the WES data and generate input files for BAF segmentation: data collection; if VQSR method has not been applied to the data, the variant recalibration procedure was applied to produce recalibrated files in multi-VCF format; the extraction of the single sample VCF for each individual of the study; the SNP tag annotation was added to the VCF file; the filtering of the VCF; the extraction of SNP tag and AD from the

VCF file; VCF file info are then used from the python program (mkVAF.py) to generate the input file for BAF segmentation where each row contains information for different variants.

4.2.5 Run BAF segmentation using WES data

BAF segmentation (Staaf et al., 2008), a method for identifying regions of AI from B allele frequencies (BAF) obtained from SNP array genotyping, was used and described in chapter 2 (Section 2.2.11). However, for this work WES datasets were used to extract and store the B allele frequencies (referred to as VAF) in files generated from the automated pipeline (Figure 4.1). These input files were run through the BAF segmentation program to identify AI regions in the WES data. The raw outputs are saved in a text file which uses five features to define the segmented AI regions detected, these features are described in Table 4.1.

Table 4.1 **Features to define a segmented AI region.**

Feature by BAF seg	**Description**
No. of informative SNPs	The number of SNPs with an mBAF value < 0.9 and > mBAF threshold
Total No. of SNPs	Total number of SNPs after triplet filter is applied and non-informative SNPs are removed
mBAF	Mirrored BAF data along the 0.5 axis
Heterozygosity rate	The heterozygosity rate for each AI region
Size	Size of the AI region in bp

The table lists five AI-characterizing features and their description. The number of informative SNPs can be the same as or a fraction of the total number of SNPs. Therefore, during the data processing was checked that the number of informative SNPs was less than or equal to the total number of SNPs.

4.2.5.1 High and low stringency settings

BAF segmentation provides several parameters that can be used to alter the stringency of quality control and sensitivity of AI detection. To optimise this tool for application to sequencing data, the exemplar UK Biobank data (n=120) were analysed using high and low stringency settings (Table 4.2).

Table 4.2 **BAF segmentation settings.**

BAF segmentation parameters used/default value	**Description of the parameters**	**Low stringency settings**	**High stringency settings**
--ai_threshold/0.56	mBAF threshold for calling regions of AI based on segmented mBAF values.	0.6	0.65

BAF segmentation parameters used/default value	Description of the parameters	Low stringency settings	High stringency settings
--non_informative/0.97	mBAF threshold for removing putatively non-informative homozygous SNPs.	0.9	0.9
--triplet/0.8	Threshold for triplet filtering used to improve removal of putatively non-informative homozygous SNP.	0.6	0.6
--ai_size/4	Minimal number of SNPs a segmented region should contain in order to be called as AI. Segments with less numbers of SNPs are removed from further analysis.	4 (Default)	10

BAF segmentation parameters applied to the WES data and their description. The two columns on the right show the low and high stringency settings used in this work.

To evaluate sensitivity and specificity of the proposed settings, the results obtained with BAF segmentation were cross-validated against the array results for the same 120 samples. The BAF segmentation results obtained from the genome-wide array data on the UK Biobank participants are publicly available (Dawoud et al., 2020), and were used for this project to understand how well results from the exome analysis agree with the SNP-array results. Overlaps between the WES and SNP array results were identified using bedtools intersect (Quinlan and Hall, 2010) and samples with AI regions that overlapped by at least 2 Mb where classified as true positive (TP). Samples with AI regions identified only by SNP array were classified as false negative (FN), whereas samples with no AI regions were identified as true negative (TN) if there was concordance by both genotyping methods. False positive (FP) were present if samples with AI regions were not identified by Dawoud and colleagues. A confusion matrix, with rows representing true class and columns representing the predicted class, was used to display this information (James et al., 2013). Sensitivity (TP/[TP+FN]) and specificity (TN/[TN+FP]) were calculated considering the array results as a truth dataset which is the preferred method to identify AI regions and were used to determine which BAF segmentation settings performed best (Table 4.6).

Following manual review of the BAF plots, a number of compelling AI regions were identified that had not been shortlisted in the comparative array-based analysis (Dawoud et al., 2020). These regions were large (> 5.4 Mb), extended to the telomere, had high mBAF scores (> 0.81) and were therefore considered to be likely regions of high level aUPD. As a result, these regions were added to the list of true positives and the sensitivity/specificity values were recalculated (Table 4.7). The

BAF segmentation parameters, either high or low stringency setting, with the highest sensitivity and specificity were identified and applied in a further analysis of the complete UKB-WES50 and Schizo-WES02 datasets.

4.2.5.2 Assessment of VQSR filter

VQSR assigns an accurate confidence score to each putative variant call and was applied to the UKB-WES50 data as part of the QC process (Section 4.2.4). The Swedish data were downloaded after VQSR had been applied so this step was not repeated (Figure 4.1). To investigate the effect of VQSR, the exemplar UKB-WES50 subset of 120 samples were analysed with and without recalibration using BAF segmentation and the low stringency settings. The raw results were compared using a scatter plot to observe the relationship between the total number of AI regions per sample and the percentage of the autosome covered by regions of AI per sample.

4.2.5.3 Processing UKB-WES50 and Schizo-WES02

Following the VQSR assessment and the evaluation of best parameters for BAF segmentation, both UKB-WES50 and Schizo-WES02 datasets were run through BAF segmentation using low stringency settings.

4.2.6 Identify and remove low quality samples

A scatterplot was made to examine the raw output from BAF segmentation and to identify and exclude low quality samples. Per sample metrics for the total number of AI regions and percentage of the autosome covered by AI regions were used to make the scatterplot and to identify any sample outliers in both the Swedish-WES02 and UKB-WES50 cohorts. To calculate the autosomal AI percentage, the total length of the 23 chromosomes was defined using python dictionaries which store the chromosome length for both hg19 (2.881 GB) and hg38 (2.875 GB). Outlying samples with an excessive number of AI regions and/or large proportion of the genome composed of AI were considered to be indicative of low sample quality and were removed.

4.2.7 Filtering strategy and data preparation

The frequency of AI events detected by BAF segmentation were higher than expected based on published studies (Jacobs et al., 2012; Laurie et al., 2012). To bring the frequency of AI events in-line with these expectations, a stepwise method was developed to select AI regions with strong supporting evidence and properties associated with aUPD events. Dawoud et al. developed a custom script to exclude FP regions identified by BAF segmentation in SNP array data. The filtering

method, applied to array data only involved the following steps: merge regions <2Mb apart; drop regions with a size <2Mb; keep regions with a density of at least one marker per 20 Kb. This strategy was adapted by applying thresholds that were more appropriate for WES data, adding new features for filtering (listed in Table 4.3 and in *italics* in the text) and supplementing the output file to aid the removal of FP regions and enable statistical analysis. First of all, consecutive AI regions were merged using bedtools (Quinlan and Hall, 2010) when the distance between them was less than 4 Mb. The distance between two regions was increased because of the low density marker that are used with WES compared to array data. As a result of this step, the values for mBAF, heterozygosity rate, physical size of AI region, number of informative SNPs and total number of SNPs were recalculated and saved in the output file. Second, TP and FN calls that were identified during the cross-validation process (4.2.5.1) were used to investigate potential causes of the FN results and to establish the minimum size of AI regions (>=5 Mb). Merged regions that were smaller than this were removed.

Samples from the exemplar dataset that carried the *JAK2*^{V617F} mutation, a somatic mutation known to be associated with 9p chromosomal abnormalities, were used to determine the best SNP density threshold (*bases per marker*) (Table 4.3). Several cut-offs were used (100, 250, 400, 550, 700, 850 Kb) and the minimum value that resulted in detection of all 9p aUPD positive samples was defined as the best threshold for this filter.

Three more features were added to the output file: *Bases per informative marker*, which defines the rate of Kb per informative marker; *coverage* of the merged region; *centromere overlap*, which represents the percentage of each AI region overlapping the centromere (Table 4.3). Annotation files, downloaded from the Genome Reference Consortium (https://www.ncbi.nlm.nih.gov/grc/human), were used to extract centromere locations and chromosome lengths for both hg19 (Church et al., 2011) and hg38 (Schneider et al., 2017). The new features generated at this stage were used for the subsequent logistic regression (LR) model that was developed for prediction of AI regions (4.2.9).

Table 4.3 **New features generated to aid the filtering of FP calls.**

New Features	**Description**	**Threshold value**
new_size	The new size of an AI event following the merging of consecutive regions	5 Mb

New Features	Description	Threshold value
Bases per marker	Define the rate of kb per marker (size/Total No of SNPs)	850 Kb
Bases per informative marker*	Define the rate of kb per informative marker (size/No of informative SNPs)	NA
Coverage*	(size of individual events/new size)	NA
Centromere overlap*	AI region percentage overlapping the centromeric region	NA

NA: Not applicable; *These features were used to build the LR model (Section 4.2.9) and not for filtering.

This work has given the opportunity to analyse over 60,000 WES genomes. The size of the AI regions and FP calls presented the need to handle computationally the stepwise method described above. Therefore, a user-friendly tool, BRawO (BAF Raw Output), was developed to manipulate the output file, to generate new features (Table 4.3), to calculate the empirical score (*heterozygosity rate* x *number of informative SNPs* x *coverage*) (Dawoud et al., 2020) and to apply a number of filters to the BAF raw output file from BAF segmentation. BRawO facilitated the stepwise approach in a number of ways: by selecting the genome build (hg19 or hg38) of the dataset; by removing noisier samples with too many AI regions and/or a large proportion of AIs in their genome; by defining a maximum distance (Mb) to allow the merging of consecutive regions; by filtering regions according to their size (Mb); by removing AI regions with low marker density; by selecting only telomeric regions that falls within a defined distance (Mb) from the end of the chromosome. Positional and optional arguments of this tool are descried in Table 4.4 and Table 4.5.

Table 4.4 **Positional argument of BRawO.**

Positional arguments	Description
`ai_regions_file`	The file containing the AI data to be analyzed. The file must be tab delimited.
`hg_ref`	The Human Genome (hg) reference to be used. You can choose one among "hg19" and "hg38"

Table 4.5 **Optional argument of BRawO.**

Optional arguments	Description
`--max-perc-AI-regions-per-sample`	Max allowed percentage of AI regions per sample. Samples with more percentage of AI regions than this threshold are filtered out

Optional arguments	Description
`--max-AI-regions-per-sample`	Max allowed number of AI regions per sample. Samples with more AI regions than this threshold are filtered out
`--region-merge-distance-Mb`	The maximum distance in Mb between two regions for them to be merged by bedtools
`--min-ai-region-size-Mb`	Minimum size, in mega bases units, of the merged AI regions that will be selected. Smaller regions are filtered out.
`--max-bases-per-marker-kb`	Max bases, in kilo-bases units, per marker. Regions with more bases per marker are filtered out.
`--telomeric-keep-width-Mb`	Width in mega bases of the telomeric regions to be kept. All events falling outside this width after the chromosome starts or before the chromosome ends are dropped.
`-h, --help`	Show the help message and exit

4.2.8 Visual inspection of selected AI regions

BAF segmentation generates, for each sample, a file with two plots per chromosome, a BAF plot representing the BAF values and a mBAF plot with the mBAF values reflected along the 0.5 axis. (Figure 4.2). After running BAF segmentation and filtering the raw output, a visual inspection of the segmented regions was carried out to annotate them as likely aUPD, false positives or negatives independently of the published array based results (Dawoud et al., 2020). AI regions were annotated as false positives when the markers on the plot were not clearly and consistently separated over the entire region or when they were identified by a small number of SNPs with low density and large gaps in coverage that were frequently associated with centromeric regions (Figure 4.2 A). Regions were annotated as likely aUPD when the markers in the BAF plot showed a clear and consistent shift away from the expected heterozygous BAF value of 0.5, indicating a clonal LOH which can be more or less pronounced (Figure 4.2 B). The level of clonality can be inferred by the average mBAF value of the AI region where subclonal events have a more subtle shift away from 0.5 and mBAF values that are closer to 0.5.

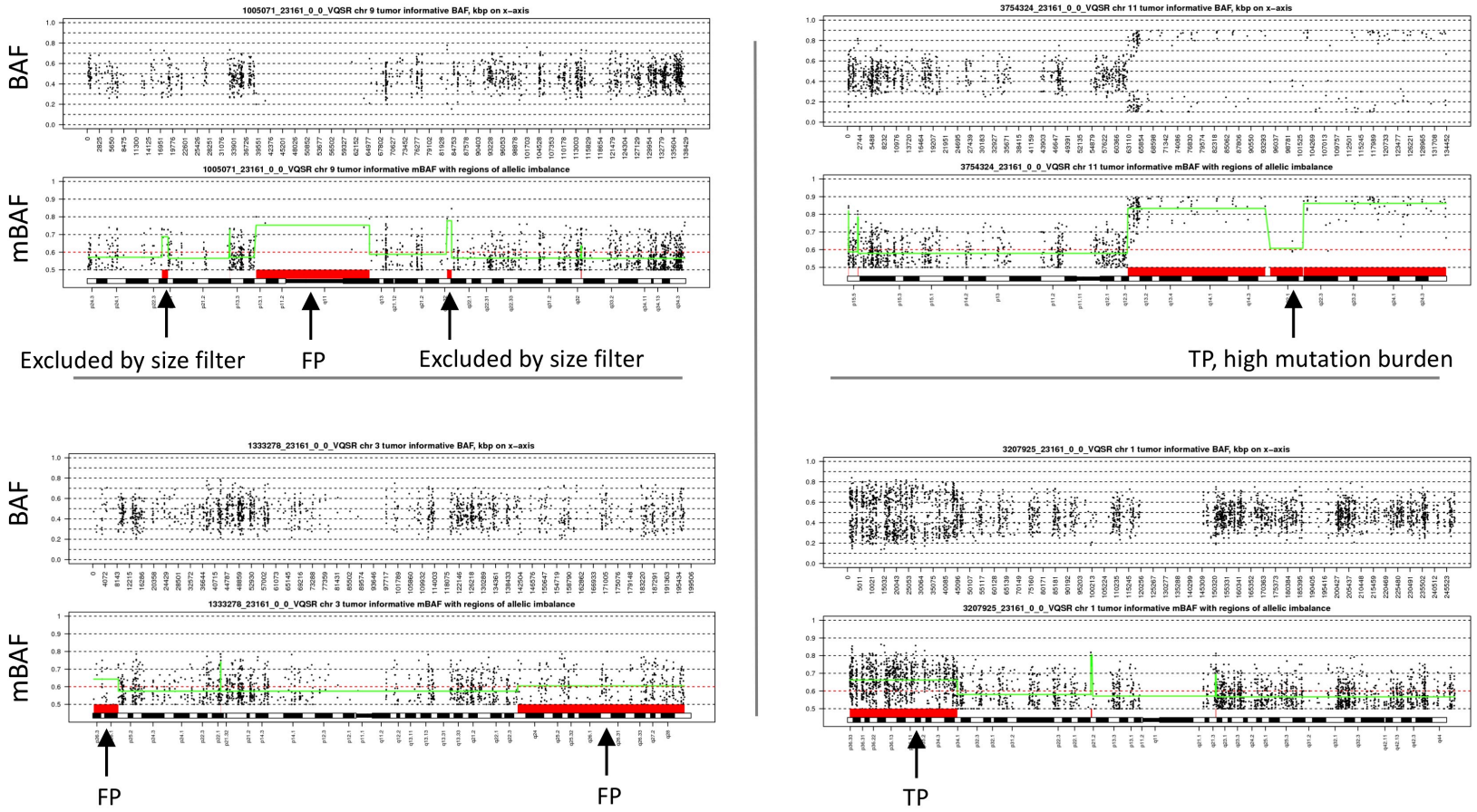


Figure 4.2 **BAF segmentation plots.**

BAF and mBAF plots for four samples with AI regions that are classified as either FP (n=3) or TP (n=2) after manual review. BAF plots show the BAF and mBAF plots are the transformation of BAF values along the 0.5 axis. Circular binary segmentation is applied to mBAF values to identify regions of similar allelic proportions (green profile). Regions of AI are identified where the segmented mBAF is > 0.6 (red dashed line) and are highlighted by a red rectangle. **A.** The mBAF plot displays three FP regions on chromosome 9: the centromeric region is characterised by low marker density; the two small regions did not pass the size filter. The lower panel shows two telomeric AI regions on chromosome 3 that were classified as FP because the supporting markers do not split clearly from the 0.5 axis and all markers on the chromosome have a broad spread of BAF values. **B.** Both panels show AI regions that were categorised as TP (aUPD) due to the clear shift in mBAF values compared to background markers, their size and proximity to the telomere. The high mBAF value of the AI region in the top panel (mBAF=0.82) is suggestive of a high mutation burden.

4.2.9 Logistic regression model for predicting likely aUPD

Both Schizo-WES02 and UK-WES50 were run through BAF segmentation and the stepwise method was applied to remove low quality samples and FP calls. Following visual inspection of the mBAF plots, filtered AI regions from the UK-WES50 data were labelled as likely aUPDs or false positives. The labelled data were split into a training (70%) and test dataset (30%) with the same ratio of UPD classes (real to FP) to reduce sampling error and to maintain heterogeneity of both sets. The heterogeneity is a fundamental feature to train the model on a balanced dataset. The training set was used to fit the logistic regression (LR) model to implement a scoring system, subsequently referred to as the *gg score*, that estimates the probability of each AI region being a real UPDs and is used to rank all of the AI regions. The LR model was optimized using an L2 regularization also known as ridge regression, to improve numerical stability and to prevent overfitting. The optimal regularization parameter C used in the L2 regularization was found using cross-validation (CV). The k-fold CV was used to split the train set into k smaller sets and at each computed loop, one part of them was used as "validation set" so that the final optimal value is the mean of the values computed at each loop. The 5-fold CV thus optimizes the use of the available data, with respect to a classical train-test split. The final and unbiased evaluation of the predictive performance of the LR model was obtained by applying the fitted LR model to the test set. A flow chart describing the design of this study is shown in Figure 4.3. The LR model was built from the logistic function included in python (Baranwal et al., 2011).

4.2.9.1 Sequential feature selection

Feature selection consists in finding the set of features that produces the best-performing predictive model. Sequential Feature Selection (SFS) can be used to perform a forward selection or a backward selection which, respectively, iteratively adds the best features or removes the worst ones, on the basis of the CV score of the estimator (Ferri et al., 1994). In both cases, n-1 features were ranked based on the average score obtained on a 5-fold CV splitting of the data. In a first step, backward SFS was used to remove the pair of features that affected the performance the least. Then, forward selection was applied to the remaining features, by adding the predictors one by one, until the CV score stopped improving significantly. After performing forward SFS, a further manual check was conducted, which resulted in adding one more feature to the final set, that was deemed to bring relevant improvement to the performance. Further details are explained in Section 4.3.6 (Results).

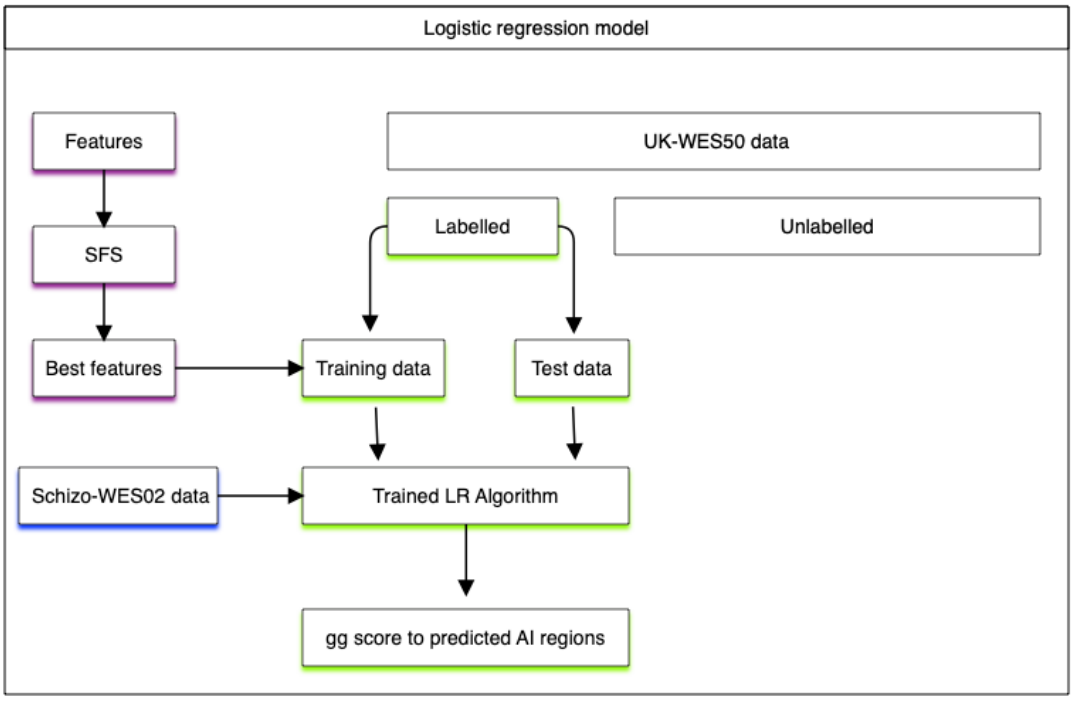


Figure 4.3 **Flow chart of the logistic regression model.**

Schematic representation of the procedure and data used for training the logistic regression model that generates the *gg score*. Feature selection (purple) and model training and application (green). The blue box indicates the WES data processed through the LR algorithm. The black box indicates the entire UK-WES50 data that was annotated only in part.

4.2.9.2 Evaluation metrics

Commonly used metrics such as area under the receiver operating characteristic curve (AUC), precision, recall and F1-score were calculated on the test set and used to determine the performance of the LR model.

The Receiving Operating Characteristics (ROC) curve is a graphic that shows the benefit of applying a certain statistical test. It simultaneously traces out the two types of errors for every possible threshold as these vary from 0 to 1. The true positive rate or sensitivity on the y-axis represents the fraction of AI regions that are correctly identified. The x-axis shows the false positive rate, which represents here the fraction of false positive calls that are incorrectly classified as AI regions. The ideal ROC curve hugs the top left corner of the graph, indicating a high true positive rate and a low false positive rate (James et al., 2013). The performance of the model is given by the AUC, which ideally indicates excellent and good predictions for values >0.9 and >0.8 respectively. An AUC=0.5 indicates that the model is not discriminative and its performance can be attributed to chance alone (Swets, 1988).

Precision measures how many of the predicted positive cases are correctly classified. On the other hand, recall (or sensitivity) measures the proportion of true positive instances that are correctly predicted, which in this case is the primary aim of the model. Models can be optimized using a measure called F1-score which is a weighted average of both precision and recall. In the optimal scenario, with perfect precision and recall, the highest F1-score is 1, whereas if either precision or recall is zero, it can reach its lowest value, zero (Powers, 2007). Due to the pronounced imbalance between the positive and negative classes, I chose to optimize the model using the ROC-AUC metric.

4.2.9.3 Validation of the gg score

To validate the *gg score*, I compared its performance to an existing scoring system that was developed using SNP-array and BAF segmentation (Dawoud et al., 2020). This score is defined as the product of *bases per marker*, *heterozygosity rate* and *coverage* and empirical threshold of ≥9 was used to select likely somatic events. Having applied this scoring system to the labelled UKB-WES50 data set I examined the distribution of scores across false positives and likely aUPD events and compared these results with the distribution of the gg scoring system. The python plt.hist() function was used to plot the histograms of both scores and compare their distribution for false positive and likely aUPD.

4.2.10 Identification of candidate somatic driver variants from WES data

Finally, the gg scoring system was applied to the Schizo-WES02 data and a score greater than 0.5 was used to identify AI regions that were likely to be real UPDs. If we assume that the logistic regression output is the probability that the example belongs to class 1, the 0.5 threshold corresponds to choosing the class (0 or 1) with the highest probability in the binary classification. Ideograms were generated with karyoploteR to visualize the likely aUPDs (Gel and Serra, 2017) and if these were overlapping in two or more samples, the single sample VCF was extracted and searched for novel and/or known somatic mutations in the target genes. First of all, SAMtools/Bcftools was used to filter the VCF file to remove variants that did not pass the VQSR filter and with low read depth (DP<10) (Danecek et al., 2021). Then, SnpSift was used to extract variants that intersect a specific gene known to be the target of aUPD and the file was annotated using wANNOVAR (Cingolani et al., 2012; Yang and Wang, 2015). Putative somatic mutations were screened for in five genes known to be aUPD targets: *MPL* (aUPD1p), *TET2* (aUPD4q), *EZH2* (aUPD7q), *JAK2* (aUPD9p) and *FLT3* (aUPD13q).

4.3 Results

4.3.1 BAF segmentation parameters for WES data

To determine appropriate software settings for WES data the exemplar dataset, consisting of 120 selected samples from the UKB cohort, were analysed using BAF segmentation with either low or high stringency settings. A total of 961 (Appendix Table B.2) and 3,075 (Appendix Table B.3) AI regions were identified using the high and low stringency settings respectively (Table 4.2) which, for both settings used, correspond to AI regions present in all samples. To select AI regions that are likely to represent real occurrences of aUPD and to bring these frequencies in-line with those expected from published studies (Jacobs et al., 2012; Laurie et al., 2012) the raw outputs were filtered to select regions that were greater than 5 Mb; to keep regions with 850 Kb per marker; to merge consecutive regions that were <4 Mb apart. Following the filtering of the raw outputs, a total of 212 and 61 AI regions were selected from the low and high stringency analysis respectively. BAF plots for these selected AI regions were visually inspected and manually annotated as samples with either likely aUPD (n=38) or negative (n=82) according to the criteria described in Section 4.2.8. The annotated results were cross-referenced against those from Dawoud et al. 2020 using bedtools to identify overlaps. Results from the previous study, which used corresponding SNP-arrays and an empirical score to select regions of mosaic chromosome abnormalities (mCA), were treated as true positives and negatives and used to determine the performance of the low and high stringency settings in terms of sensitivity and specificity which are displayed in a confusion matrix (Table 4.6). The low stringency settings offered much higher sensitivity (36.8% versus 28.9%) and only a slight reduction in specificity (91.5% versus 92.7%) compared with the high stringency settings.

Following manual annotation of the BAF plots, seven samples with mBAF values ranging from 0.81-0.87 were annotated as likely aUPD due to the presence of long runs of AI (5.4-35.4 Mb) which extended to the telomere and involved large numbers of informative SNPs (8-75 kb per SNP). For example, Figure 4.4 shows a region of likely aUPD following manual annotation involving chromosome 9 that, in the comparative SNP-array results, was below the empirically defined threshold of 9 and therefore was labelled as negative.

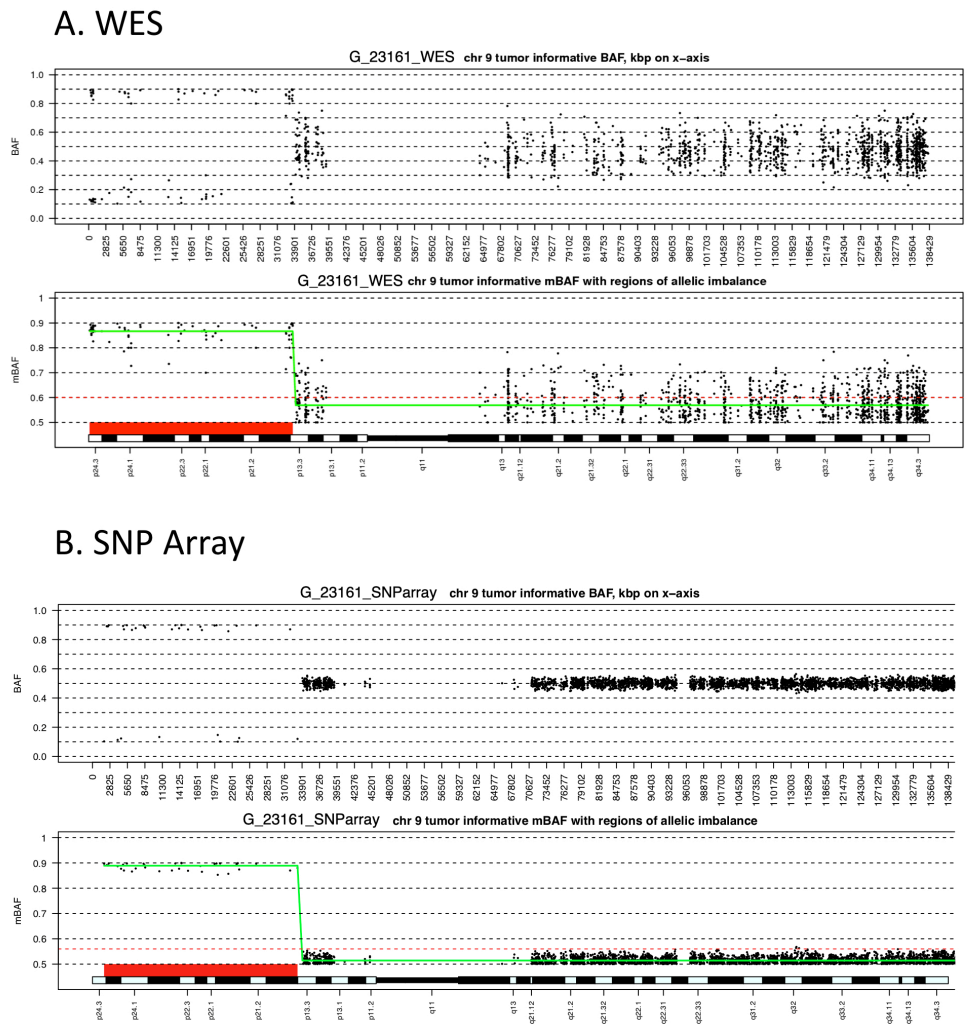


Figure 4.4 **AI region detected after visual reassessment.**

The figure shows examples of BAF and mBAF plots of the same sample from the WES (A) and the SNP-array result (B). The AI region of 33.4 Mb is located on chromosome 9p and following the visual reassessment was annotated as likely aUPD due to its size and proximity to the telomere. Both panels show a telomeric AI region with high level of AI.

Previous studies have shown that mBAF values are directly related to the proportion of cells with aUPD (Chase et al., 2015). The large mBAF values therefore suggest that these aUPD events have high frequency in the major clone. Furthermore, 5 out of 7 of these samples were shown to have a known somatic mutation (*JAK2*^{V617F}) in the aUPD regions. This in combination with the properties of these regions lead to their reclassification as TP and recalculation of the specificity and sensitivity which showed an overall improvement for both the low (46.6% and 100% respectively) and high stringency settings (38.6% sensitivity and 100% specificity) (Table 4.7).

Table 4.6 **Performance metrics for WES based detection of aUPD.**

	Predicted aUPD status based on WES data					
	Low stringency			High stringency		
	Likely aUPD	**No aUPD**	**Performance**	**Likely aUPD**	**No aUPD**	**Performance**
Positive	14 (TP)	7 (FP)	Sensitivity = 36.8%	11 (TP)	6 (FP)	Sensitivity = 28.9%
Negative	24 (FN)	75 (TN)	Specificity = 91.5%	27 (FN)	76 (TN)	Specificity = 92.7%
	Precision = 66.6%	NPV = 75.7%	Accuracy = 74.2%	Precision = 64.7%	NPV = 73.8%	Accuracy = 72.5%

TP: True positive; TN: True negative; FP: False positive; FN: False negative; NPV: Negative predictive value. Information displayed in the confusion matrix shows that BAF segmentation with low stringency settings identified 21 samples with AI and only 7 samples were incorrectly labelled. On the other hand, high stringency identified 17 samples with AI regions and 6 of them were not identified correctly.

Table 4.7 **Confusion matrix for the computational comparison and visual reassessment.**

	Predicted aUPD status based on WES data and visual inspection					
	Low stringency			High stringency		
	Likely aUPD	**No aUPD**	**Performance**	**Likely aUPD**	**No aUPD**	**Performance**
Positive	21 (TP)	0 (FP)	Sensitivity = 46.6%	17 (TP)	0 (FP)	Sensitivity = 38.6%
Negative	24 (FN)	75 (TN)	Specificity = 100%	27 (FN)	76 (TN)	Specificity = 100%
	Precision = 100%	NPV = 75.7%	Accuracy = 80%	Precision = 100%	NPV = 73.8%	Accuracy = 80%

TP: True positive; TN: True negative; FP: False positive; FN: False negative; NPV: Negative predictive value. A confusion matrix compares the results obtained from BAF segmentation with both low and high stringency settings after manual annotation. Information here displayed shows that low stringency settings identified 21 samples with AI and none of them were labelled incorrectly. On the other hand, high stringency identified 17 samples with AI regions and no FPs.

After the evaluation of sensitivity and specificity of BAF segmentation carried out in the UK Biobank subset of 120 samples, low stringency settings showed higher sensitivity and 100% specificity. Therefore, they were confirmed as the best parameters to use in any further analysis.

4.3.2 Investigation of the FN results

As part of these analyses, I also investigated the source behind the FN results, which have been identified in the array samples but not in the WES data. A box plot was used to investigate the size of the AI regions that were classified as TP and FN compared with the revised SNP-array based results (Figure 4.5). The TP calls (n=22) were larger in size (median=29.8 Mb) compared with the FN calls (median=2.9Mb) whose size is extracted from the SNP-array results (Dawoud et al., 2020). The boxplot shows that all 30 of the FN regions that were not detected across 24 samples (Table 4.7) of the NGS exemplar cohort are, in term of size, all smaller than 9 Mb in size with a median span of 2.9 Mb. This suggests that small regions of aUPD are difficult to detect using WES data due to the lower density of variants compared with SNP arrays. It is important to note that following manual annotation of the BAF plots, evidence of aUPDs in the WES data was not identified for all 30 FNs. The minimum size of a TP region was 5.4 Mb and this knowledge was used to further refine the automated filtering of AI regions. The identification of TP and FN classes guided to establish 5 Mb as the threshold for the minimum AI size (Mb).

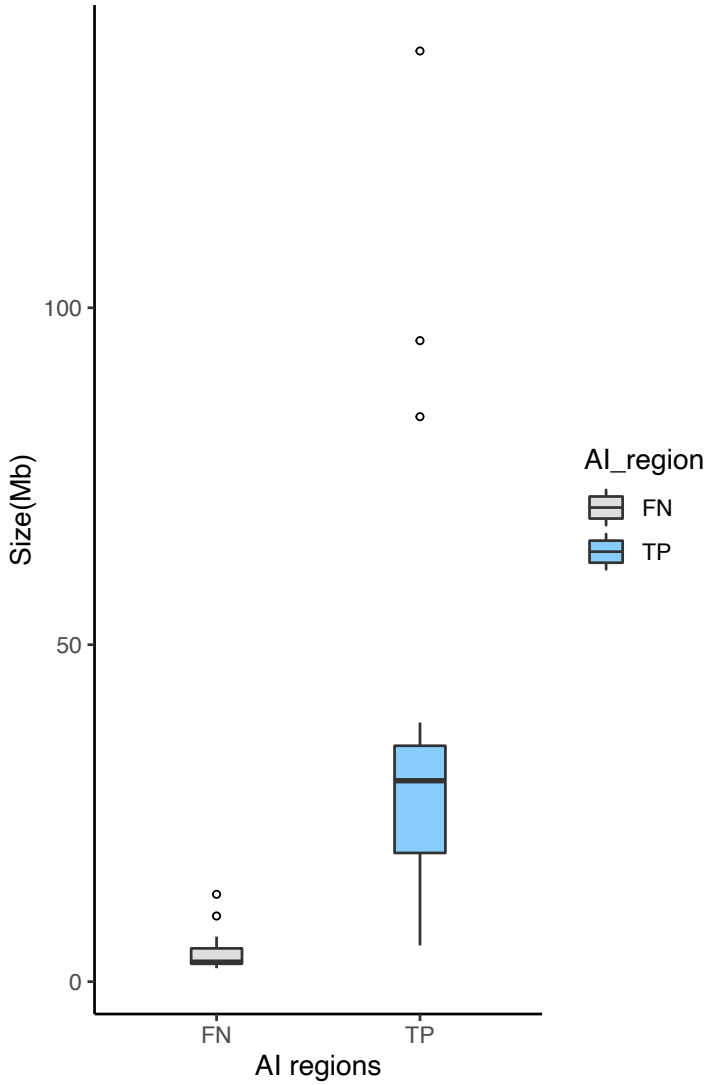


Figure 4.5 **Boxplot of size (Mb) for AI regions labelled as FN and TP.**

The region's sizes (Mb) across FN (grey) and TP (light blue) are summarized as boxplots where the middle black like is the median. The boxplots show that regions identified by BAF segmentation tent to be bigger in size when using WES data.

4.3.3 Evaluate the effect of VQSR on the BAF results

The UKB-WES50 data was generated by the UK Biobank using the GATK-based pipeline without applying VQSR. In general, the frequency of detectable AI events in healthy individuals is between 0.23% and 1.91%, with a slightly higher frequency in cancer patients (Jacobs et al., 2012). Therefore, the subset of 120 UKB-WES50 samples was used to assess whether the exclusion of low quality variants detected through the VQSR filter could help to generate less noise in the BAF segmentation raw output file. The subset cohort without and with recalibration was run through BAF segmentation using low stringency settings, and results from the two groups were compared. The scatter plot in Figure 4.6 shows the comparison between the raw output from BAF segmentation of the UKB-WES50 data before and after VQSR is applied. The final results show that the main effect of VQSR is to reduce the number of AI regions called per sample which is

demonstrated by the right to left shift of samples in the scatter plot (Figure 4.6 A). VQSR has a more subtle effect on AI coverage resulting in a slight reduction in coverage and a downward shift on the scatter plot.

Samples form 2 clusters with either high or low autosomal coverage, which did not appear to be related to the effect of VQSR. This clustering could be due to the low number of samples plotted, which reduces the chance of having a more uniform distribution of points and, therefore, separate clusters rather than a spread of samples. The clustering was investigated further (Figure 4.6 B) by colouring samples according to their AI status, either positive or negative, and whether or not they had a *JAK2*^{V617F} mutation. AI negative samples and samples without a *JAK2*^{V617F} mutation were expected to have a lower percentage of autosomal AI. However, these four groups of samples were randomly distributed between the clusters and were therefore ruled out as the cause of this clustering.

I did not identify the cause of this clustering, but it was not due to either AI status (either positive or negative based on manual review) or *JAK2* carrier status. Additional factors that were not investigated but could be relevant include cancer status, age and sex.

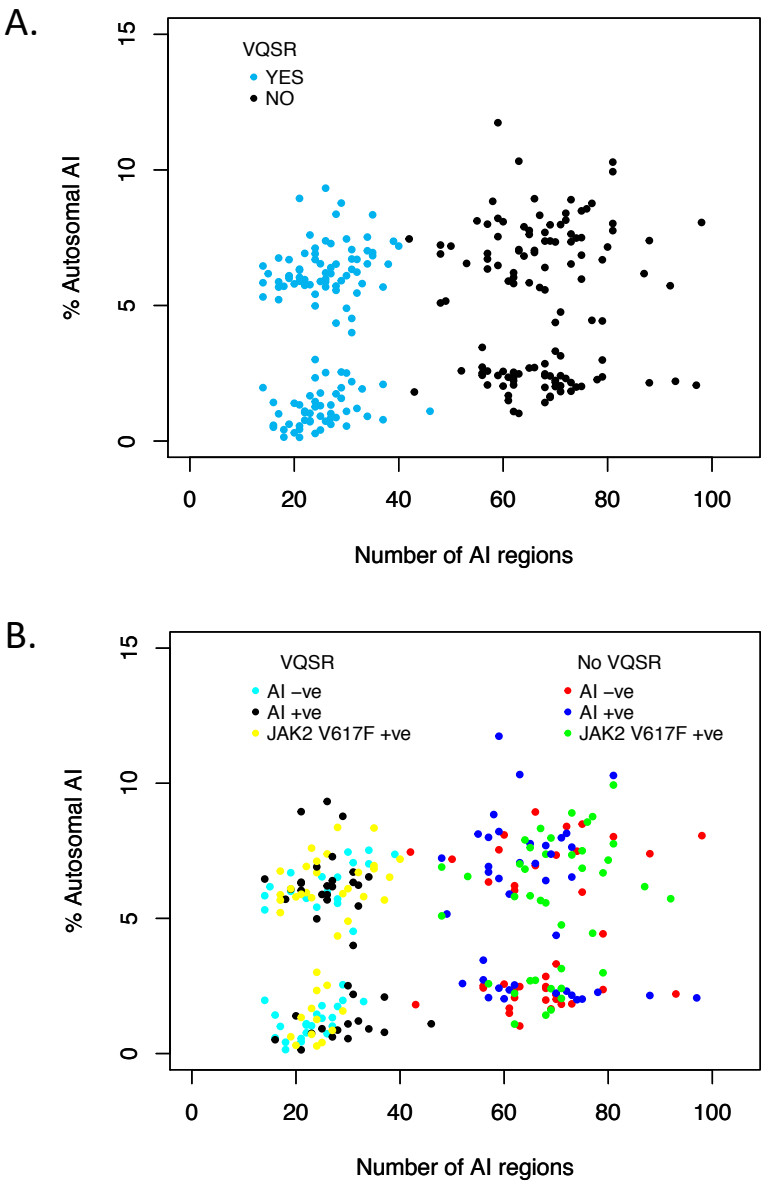


Figure 4.6 **Scatterplot comparing BAF results from WES data with and without VQSR.**

The scatterplots are produced using RStudio. **A.** The plotted points show the comparison of the BAF results obtained from WES data with (blue points) and without VQSR (black points). The two left-right clusters show that VQSR reduces number of AI regions and percentage of autosome covered by AI regions. **B.** The two up-down clusters were here investigated by grouping the samples in AI positive, AI negative and *JAK2*^{V617F} positive. The three groups appear randomly distributed and thus ruled out as a cause for this clustering.

4.3.4 Quality control of WES data

The UKB-WES50 dataset contains a total of 49,996 samples distributed across 499 multi-sample VCF files (an average of 100 samples per multi-VCF). After extracting single samples, an average of n=377,054 variants were called per sample. After the QC steps were applied to each single-sample VCF, most variants were removed due to MAF<0.01. About 34,000 markers per sample were used to detect AI regions through BAF segmentation (Table 4.8).

A total of n=1,811,204 variants were called in the whole Schizo-WES02 multi-sample VCF file containing 12,380 samples. After processing the multi-sample VCF, the QC measures (4.2.7) are applied to each single-sample VCF. Following the data preparation, an average of 23,000 markers per sample were kept as input to BAF segmentation (Table 4.8).

Table 4.8 **Filters applied for variant exclusion in WES datasets.**

		Total	MT	Missingness	FORMAT/DP>10	QUAL>20	FILTER=PASS	MAF>0.01	Multiple entries	HET FP
UKB-WES50	**mean**	377053.8	377053.8	372702.92	284214.30	284208.91	260206.03	35980.02	35976.23	35968.49
	std	18522.02	18522.02	18486.85	28321.02	28304.21	26851.52	3291.94	3291.20	3289.37
	min	336925	336925	302417	184734	184734	169266	20660	20658	20653
	max	482074	482074	480058	473383	469264	433488	87726	87697	87688
Schizo-WES02	**mean**	1811204	1809746	1771659.92	1572462.02	1572462.02	1483445.48	22945.30	NaN	22926.57
	std	0	0	22056.35	66677.52	66677.52	62410.58	939.30	NaN	937.61
	min	1811204	1809746	1515157	720238	720238	676115	13664	NaN	13660
	max	1811204	1809746	1791810	1760612	1760612	1659244	32897	NaN	32882

MT: Mitochondrial variants; DP: Read depth in the format field of the VCF file; QUAL: Quality score; FILTER=PASS: A flag indicating that the variant has passed all set of filters; MAF: Minor allele frequency; HET FP: False positive heterozygosity; NaN: No multiple entries at the same location were identified in the Schizo-WES02 data. Sample count: 49,996 UKB-WES50 and 12,380 Schizo-WES02.

4.3.5 BAF segmentation and filtering strategy

The input files generated in the previous step were run through BAF segmentation using low stringency settings (Table 4.2) which identified 615,401 (12,380 samples) and 1,281,943 (49,996 samples) AI regions (Staaf et al., 2008) in the Schizo-WES02 and the UKB-WES50 respectively. In both the Swedish-WES02 and the UKB-WES50 cohorts, AI events were found in 100% of the samples, these unexpected pre-filtering results confirmed the need for data filtering.

The first part of this work focussed on the UKB-WES50 samples carrying a *JAK2*^{V617F} mutation which is the most frequent cause of 9p aUPD and is associated with haematological malignancies (Wang et al., 2016). In the UKB-WES50 cohort, 0.08% of the samples (n=40) are *JAK2*^{V617F} positive and this group was used to determine the best threshold for *bases per marker*. Following manual review of the BAF segmentation plots for each of the 40 samples, 37.5%(n=15) of them have aUPD of chromosome 9p positive. These regions are all telomeric and have a size range between 13Mb and 138Mb and thus more likely to be aUPD. No other aUPD events were detected in the remaining 25 *JAK2*^{V617F} positive UKB-WES50 samples.

Several features were used to select AI regions that are likely to be real aUPDs including the *bases per marker* parameter which was optimised before being applied. The best threshold was determined to be the minimum density that resulted in detection of all 15 *JAK2*^{V617F} positive samples with 9p aUPD. Several cut-offs in Kb density (from 100 to 850) were investigated and 1 SNP every 850 Kb was identified as the best threshold to apply to the WES data as it was the minimum density that identified 9p aUPD in the 15 *JAK2*^{V617F} positive samples.

The aim for the next step was the identification and removal of low quality samples in both UKB-WES50 and Schizo-WES02. The raw output file from BAF segmentation was examined to identify any sample outliers in terms of the total number of AI regions per sample and the percentage of the autosome composed of AI regions in each sample. These metrics were determined and presented in a scatterplot (Figure 4.7). Following the visual inspection of the scatterplot, outlier samples were identified and removed. The majority of samples have less than 100 AI regions and less than 30% of the autosome covered by AI in both cohorts. Samples with more than 100 AI regions and/or greater than 30% of the autosome covered by aUPD were identified as outliers and removed. These per sample thresholds for hard filtering identified six samples with more than 100 AI regions and 5,047 samples with autosomal AI coverage above 30% in the UKB-WES50

cohort. These samples, representing 10.1% of the total, were identified as outliers and removed from further analysis (Figure 4.7 A). Visual inspection of the plots did not identify any outlier sample in the Schizo-WES02 cohort. The Schizo-WES02 genotype data is of high-quality that before being submitted to dbGaP, underwent through QC check and curation by dbGaP (Figure 4.7 B).

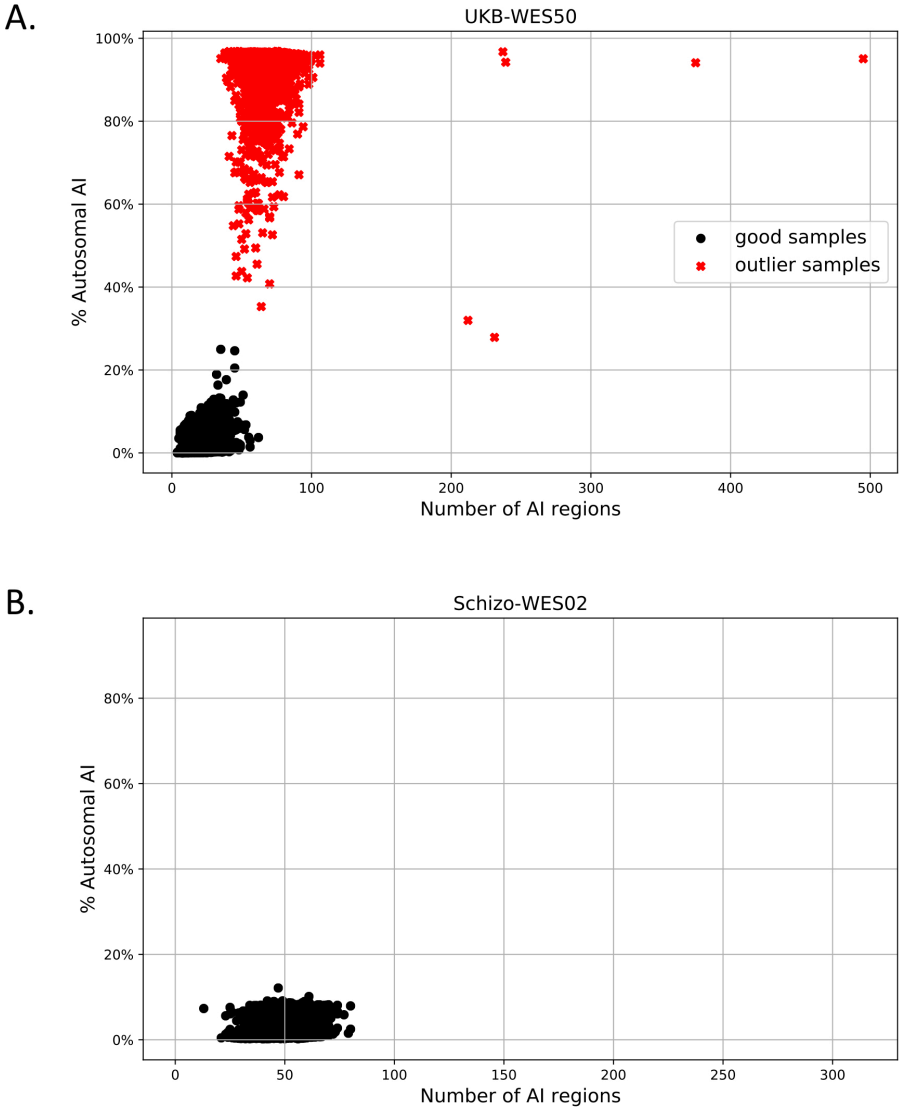


Figure 4.7 **Per sample metrics identify low quality samples.**

The scatterplots show the number of AI regions versus the percentage of autosomal AI both in the UKB-WES50 and in the Schizo-WES02 datasets. **A.** The UKB-WES50 cohort presents outliers samples (red 'x' markers) with either >30% of autosomal AI or >100 AI regions. **B.** The scatterplot shows the 12,380 samples of the Shizo-WES02 dataset forming a distinct cluster with <100 AI regions and <20% autosomal AI.

A custom program, BRawO (see methods Section 4.2.7), was used to create the input files for BAF segmentation and to apply filters that are designed to remove FN regions and select AI regions

whose properties are more likely to be associated with real aUPD events. The filters applied are as follows: *base per marker* threshold of 850 Kb; 5 Mb as the minimum allowed AI size; maximum number of AI regions and autosomal AI percentage; 4 Mb as the maximum distance to allow merging between consecutive regions.

4.3.6 Logistic regression model and feature selection

The filtering strategy identified 63,088 potential AI regions in 33,535 samples in the UKB-WES50 for further analysis. After visual inspection of the mBAF plots, a total of 3,800 regions (n=3,193 telomeric, n=607 interstitial) were labelled as either FP (n=3,643) or likely aUPD (n=157). The labelled data were split into a training (70%, n=2,660) and test (30%, n=1,140) dataset (Table 4.9) and used to develop a logistic regression (LR) model for estimating the probability of aUPD for each AI region (Figure 4.3) which is hereafter referred to as the *gg score* (Appendix Table B.4). The LR model was optimized using the regularization parameter (C=23) which was found using 5-fold CV.

The labelled data were first used to find the best variables to include in the model and to minimize the noise caused by non-informative features (Figure 4.3). SFS was used to fit a separate LR using L2 regularization and C=23 as optimal regularization parameter (Ferri et al., 1994). Initially, a backward SFS was applied by adding all ten features (Table 4.1 and Table 4.3) to the model (*number of informative SNPs*, *total number of SNPs*, *mBAF*, *heterozygosity rate*, *original size*, *merged size*, *bases per marker*, *bases per informative marker, coverage* and *centromere overlap*). Six of these features (*number of informative SNPs,* range: 4-2,320; *total number of SNPs*, range: 6-2,487; *original size*, range: 0.3 Kb-186.2 Mb; *merged size*, range: 5 Mb-186.2Mb; *bases per marker,* range:15.6-849,558; *bases per informative marker*, range: 23.3-2,217,937) were log transformed to stabilize the spread of large values (Keene, 1995). *Original size* and *merged size* were determined to be the two weakest features based on the feature ranking and were excluded from the LR model.

In the next step the forward SFS was applied, the model was fit using one feature, then two features, and so forth. The selection was made by selecting the most predictive features until the performance of the model did not show further improvement. Figure 4.8 shows the four steps of the forward SFS up to the selection of six features (*mBAF*, *bases per informative marker*, *coverage*, *centromere overlap*, *number of informative marker* and *bases per marker*) that were ranked by

the model as the most important ones. The labelled data was then fit and tested using these selected features (Figure 4.9 A).

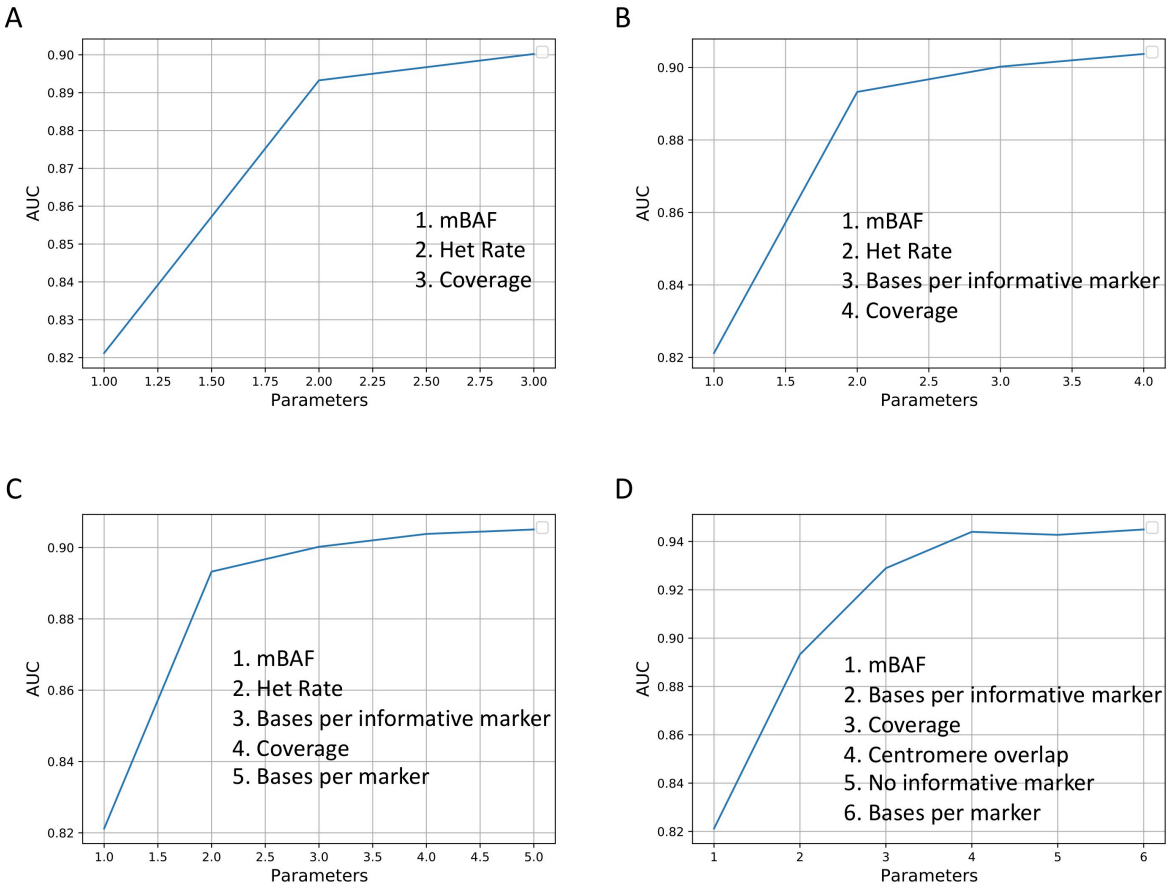


Figure 4.8 **Steps of the forward SFS.**

The figure describes the last four steps up to the selection of the most important features. X-axis indicates the rank order for each parameter. The y-axis displays the AUC obtained when each parameter is added to the model. The curve reaches the plateau (D) when feature 5 and 6 are added to the model and the AUC is stabilized.

At each of the first three steps of the forward SFS (Figure 4.8 A, B and C), *heterozygosity rate* was ranked as the second best feature of this model as this together with the mBAF allowed the model to reach AUC just under 0.90. Thus, alongside the information aided by SFS, the model was trained by adding the *heterozygosity rate* to the six features previously selected. Finally, the fitted model was applied to the test data (Figure 4.9 B, Appendix Table B.4).

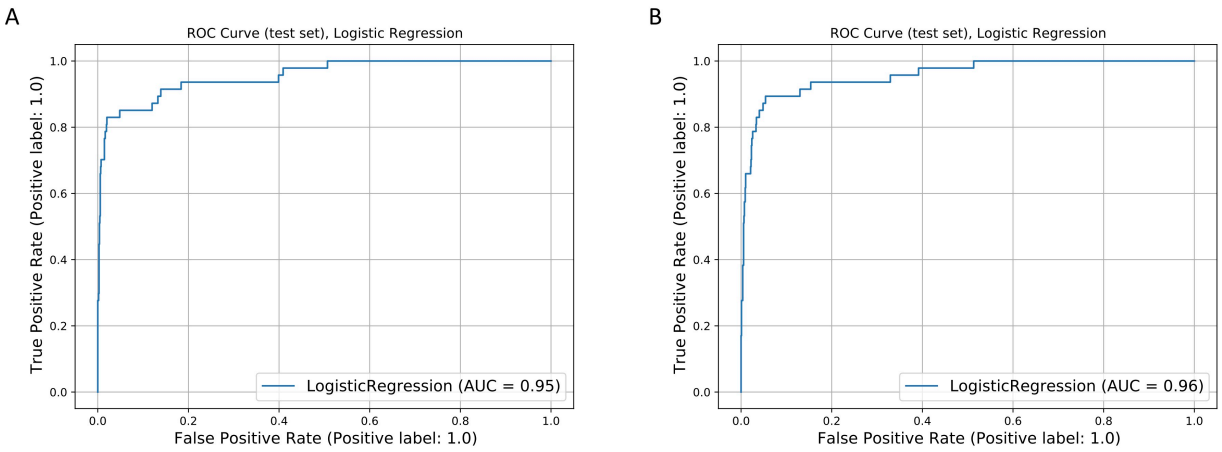


Figure 4.9 **Comparison of the ROC curves for the AI regions classifier on the testing data.**

The ROC curves for the test set of the UKB-WES50 show some difference between the LR model trained with six features (A) and seven features (B). The LR model show the best results on the training data when seven features are added to the model.

The two models described above were optimized using the ROC-AUC metric. The two curves were compared and the greatest AUC was observed when *heterozygosity rate* is added to the model (AUC=0.96). Thus, this was chosen as the best model to predict AI regions. As already mentioned in the methods (Section 4.2.9.2) it is important to reiterate that the annotated data present more negative regions (3,643) than likely aUPD (157) (Table 4.9). Therefore, it is possible that the classification problems highlighted by the recall metric, can be due to imbalanced data. The recall improves to 0.55 when seven features are used, meaning that 55% of the likely aUPD regions are correctly predicted (Table 4.10). This explains the high ROC-AUC (Figure 4.9) and that the model is better trained at identifying FP regions. Results from precision also indicated an overall improvement in the prediction accuracy. Specifically, when *heterozygosity rate* is included in the set of features the model is able to correctly classify 90% of the regions that are present in the test set (Table 4.10). These results represent just an example of the improvement that can be provided in a classification problem when different avenues are tested.

Table 4.9 **Confusion matrix for the logistic regression model using total and test data.**

Manual annotation	**Dataset**			**Classification by *gg score***			
				Total dataset		**Test dataset**	
	Total	**Training**	**Test**	**Correct**	**Incorrect**	**Correct**	**Incorrect**
Likely UPD	157	110	47	84 (TP)	73 (FN)	26 (TP)	21 (FN)
Negative	3643	2550	1093	3619 (TN)	24 (FP)	1090 (TN)	3 (FP)

Total	3800	2660	1140	

TP: True positive; TN: True negative; FP: False positive; FN: False negative

Table 4.10 **Performance of the two models.**

Evaluation metrics	6 features	6 features + het rate
Number of correctly classified on test data	1106	1116
Fraction of correctly classified on test data	0.97	0.98
Precision on AI zones	0.76	0.9
Recall on AI zones	0.4	0.55
F1-score	0.52	0.68
ROC AUC	0.95	0.96

4.3.7 Score Validation

The empirical score was calculated (Section 4.2.7) and applied to the results from the WES labelled data (n=3,800, Appendix Table B.4) (Dawoud et al., 2020). Then its distribution over likely aUPD regions (AI) and FP regions was compared with the *gg score*. The score distribution was plotted using histograms. Among the 3,800 labelled regions, 3,643 were classified as FP (n=24 have *gg score* >0.5, n=3,266 have empirical score >9) and only 157 (n=84 have *gg score* >0.5, n=139 have empirical score >9) were likely aUPD. To make the difference between the two classes visible on the frequency plot, the AI regions were weighted up by a factor 23 (Figure 4.10). Dawoud's AUC score with annotated data was 0.59 and the plot (Figure 4.10 A) shows that the bulk of AI events is superimposed to the bulk of the FP regions. About 40 (1000/23) AI events lie near score zero and FP events have a score ranging between 0 and 500, thus the two classes are less distinguishable. On the other hand, the *gg score* identified only four events with a score of zero. Distribution of FP events is right skewed, as expected, and most of them score <0.5 (Figure 4.10 B).

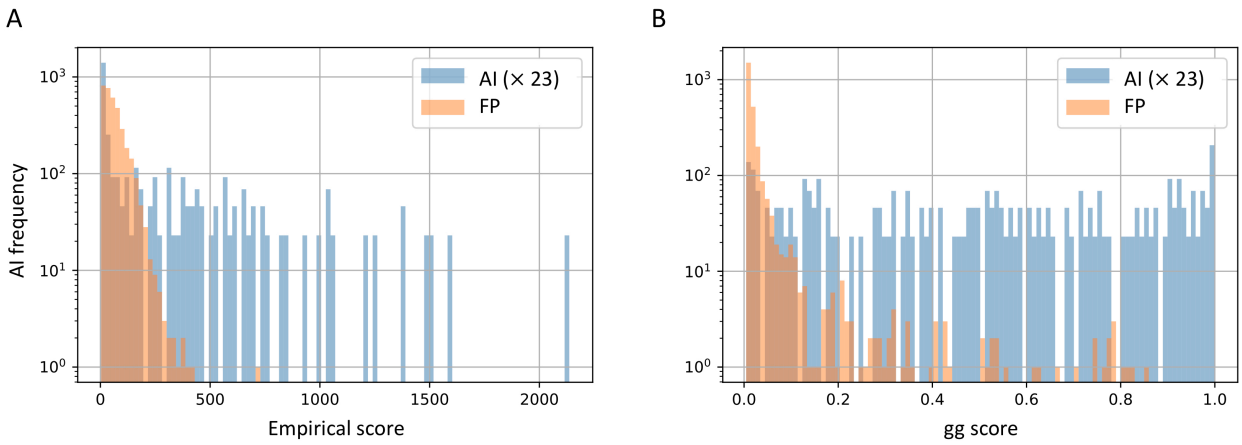


Figure 4.10 **Distribution of scores with labelled data.**

The plots show the distribution of the two scores with AI (orange) and FP (light blue) regions. **A.** Distribution of the empirical score (Dawoud et al., 2020) with labelled data. **B.** Distribution of *gg score* with labelled data.

4.3.8 Apply the gg score system to Schizo-WES02

After applying the filtering strategy described above (4.2.7), the Schizo-WES02 was left with 14,619 potential AI regions in 8,871 samples. Thus, the filtered data was passed to the LR model and manual revision of the BAF plots was performed on the regions with a *gg score* >0.5 (n=172) to confirm those regions that have been correctly identified as real AI events. As a result of the manual annotation, 29 of the 172 regions were identified as likely aUPD (Table B.1). These were distributed across 26 samples which represents a tiny (0.21%) proportion of the whole set. These results show a fall in pickup rate from 53.3% (84/157) in the UKB-WES50 to 16.9% (29/172) in the Schizo-WES02.

4.3.9 Identification of putative somatic mutations

The 29 AI regions that were identified as likely aUPDs were plotted on an ideogram which shows their distribution across chromosomes and overlap between separate samples (Figure 4.11). Putative somatic mutations in known target genes were checked if the AI region were overlapping in two or more patients. Thus, chromosomes 1 (*MPL*), 4 (*TET*2), 7 (*EZH2*), 9 (*JAK2*) and 13 (*FLT3*) were examined. Four samples overlapped the region 14q, but these were not checked as this region target the imprinted *MEG3-DLK1* locus at 14q32 (Chase et al., 2015). Results are shown in Table 4.11.

Table 4.11 **Variants identified in target genes of known aUPD.**

Chr	Start	End	Ref	Alt	Func.refGene	Gene.refGene	ExonicFunc.refGene	GT:0/1	GT:1/1	dbSNP	ClinVar SIG	SIFT pred	CADD phred	No. Samples with AI
1	43812075	43812075	G	A	intronic	*MPL*		1	1	rs1760670	Benign	.	.	4
7	148504717	148504717	G	-	UTR3	*EZH2*		0	1	rs397889421	Benign|Benign	.	.	2
7	148504717	148504717	-	G	UTR3	*EZH2*		0	1					2
7	148507534	148507534	-	T	intronic	*EZH2*		1	0	rs146223228	.	.	.	2
7	148507534	148507534	C	A	intronic	*EZH2*		1	0	rs73469687	.	.	.	2
7	148543525	148543525	A	G	intronic	*EZH2*		1	1	rs10274535	.	.	.	2
7	148543694	148543695	AA	-	intronic	*EZH2*		1	1	rs745733123	Likely benign	.	.	2
7	148543695	148543695	A	-	intronic	*EZH2*		1	1					2
7	148543695	148543695	-	A	intronic	*EZH2*		1	1					2
9	5073770	5073770	G	T	exonic	*JAK2*	nonsynonymous SNV	5	0	rs77375493	Pathogenic	D	33	5
9	5050706	5050706	C	T	exonic	*JAK2*	synonymous SNV	4	1	rs2230722	Benign|Benign	.	.	5
9	5081780	5081780	G	A	exonic	*JAK2*	synonymous SNV	3	2	rs2230724	Benign|Benign	.	.	5
9	5090934	5090934	A	T	intronic	*JAK2*		4	1	rs2274649	.	.	.	5
13	28624294	28624294	G	A	exonic	*FLT3*	nonsynonymous SNV	1	1	rs1933437	not provided	Deleterious	23.8	2
13	28636084	28636084	G	A	exonic	*FLT3*	synonymous SNV	1	1	rs7338903	.	.	.	2
13	28589267	28589267	C	T	intronic	*FLT3*		0	2	rs4073630	.	.	.	2
13	28592546	28592546	T	C	intronic	*FLT3*		1	0	rs17086226	.	.	.	2
13	28607989	28607989	T	G	intronic	*FLT3*		1	1	rs2491223	.	.	.	2
13	28609825	28609825	A	G	intronic	*FLT3*		1	1	rs2491227	.	.	.	2
13	28609846	28609846	A	T	intronic	*FLT3*		1	1	rs2491228	.	.	.	2
13	28610183	28610183	A	G	intronic	*FLT3*		1	1	rs2491231	.	.	.	2
13	28623699	28623699	G	T	intronic	*FLT3*		1	0	rs9507985	.	.	.	2
13	28623938	28623938	-	A	intronic	*FLT3*		1	0	rs869135449	.	.	.	2

The four samples with 1p aUPD (Figure 4.11) were screened and the analysis was focused on *MPL* W515 (rs121913615) a common change in myeloid phenotypes. However, no mutations were identified on codon 515 (Rumi et al., 2013), only one intronic variant (rs1760670) was present in two samples reported on ClinVar as benign.

Samples with 4q (n=2) and 7q (n=2) aUPD did not harbour any mutations in *TET2* and *EZH2*, respectively.

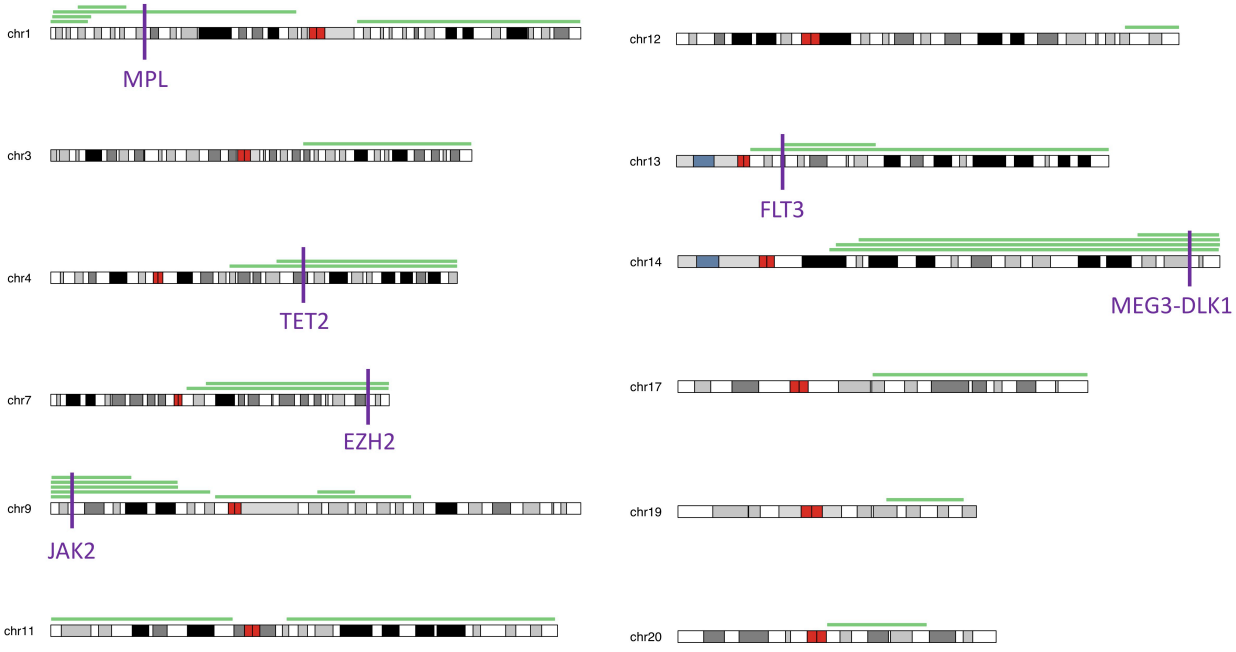


Figure 4.11 **Ideogram of the likely aUPD regions.**
The ideogram shows the physical position of likely aUPD across the autosomes.

Five samples detected by *gg score* as UPD 9p-positive, harbour the *JAK2*^{V617F} mutation. The events detected have a size range between 5.6 Mb and 42.2 Mb and are more likely to be aUPD. Also, in all five samples there is evidence of 3 known silent polymorphisms including the previously reported rs2230724, this is expected as the 46/1 haplotype is strongly associated with *JAK2*^{V617F} and particularly in association with aUPD (Jones et al., 2010). rs2230724 is also in strong LD (r^2=0.83) with rs10974944 a germline SNP know to predispose to the development of *JAK2*^{V617F}-positive MPN (Kilpivaara et al., 2009). To check whether the *gg score* identified all the 9p aUPD events we looked at the whole cohort and identified 0.14% of the samples (n=18) harbouring a

JAK2^{V617F} mutation. Following a visual inspection of the BAF plot for all 18 samples, in the remaining 13 *JAK2*^{V617F} positive Schizophrenia samples no AI events were detected by BAF segmentation, meaning that almost 1/3 of *JAK2*^{V617F} positive have aUPD. The higher prevalence of the mutation compared to the UKB-WES50 (0.08%) is expected in an older cohort such as the Schizo-WES02 with a mean age of 65 (Genovese et al., 2014).

Two UPD13q-positive samples were also analysed for the presence of *FLT3* putative somatic mutations and a missense variant c.C680T:pT227M (rs1933437) predicted to be deleterious by *in silico* tools was present in both samples (Table 4.11). The same somatic variant has been previously reported in samples with myeloid phenotypes however UPD13q is generally associated with *FLT3* internal tandem duplications which would be missed by WES.

4.4 Discussion

Somatically acquired UPDs (aUPD) are chromosomal abnormalities that have been associated with driver mutations in various cancers (Tuna et al., 2009). Identifying these regions has been established as a useful technique which has helped to identify novel cancer driver genes via targeted sequencing analysis of the aUPD regions (Tuna and Amos, 2010). For example, somatic mutations driving clonal proliferation have been identified in association with the most recurrent regions of aUPD in patients affected with myeloid malignancies (O'Keefe et al., 2010). Evidence from SNP array analysis of large cohorts has shown that aUPD occurs in apparently healthy individuals aged 65 or older and confers a tenfold increased risk of developing haematological malignancies. Furthermore, the regions of aUPD detected in the elderly are very similar to those identified in patients affected with myeloid malignancies (Genovese et al., 2014; Jacobs et al., 2012; Laurie et al., 2012). In the last twenty years, the significant advance of molecular genetics and bioinformatics has allowed scientists to use powerful techniques, such as SNP arrays and NGS, to identify these regions (Makishima and Maciejewski, 2011; Tuna et al., 2009). Since the underlying genetic abnormalities in several regions of aUPD remain unidentified, large-scale sequencing data of individuals unselected for cancer represents a valuable resource to assess the possibility of identifying mutated genes in novel affected regions of aUPD driving CH.

Therefore, this chapter focused on the development of an automated method aimed at the identification of likely aUPD regions from publicly available WES data. BAF segmentation, a tool

established in our lab to detect AI regions from whole-genome SNP arrays, was used and optimised (Section 4.2.4) to analyse sequencing data (Staaf et al., 2008).

This study used WES data from the UK biobank (49,996 individuals) and dbGaP (12,380 individuals). These cohorts are referred to as UKB-WES50 and Schizo-WES02, respectively, and were analysed to discover new aUPD regions and to search for associated target genes/mutations that may drive clonal proliferation and myeloid malignancies. To ensure the datasets were analysed in an unbiased manner, WES data were harmonised, and VQSR was applied to the UKB-WES50 cohort. The GATK VQSR is a critical QC step that enables exclusion of potential false-positive variants and selection of high-quality variants based on a single VQSLOD score. Subsequently, I developed a pipeline to filter and process the multi-sample VCFs and to generate one BAF segmentation input file per sample (Figure 4.1). To select BAF segmentation settings that were appropriate for WES data, an exemplar cohort consisting of 120 individuals from the UKB-WES50 with matched SNP arrays were analysed using high and low stringency settings (Section 4.2.5.1). Following computational cross-validation of the WES and matched array results, that were filtered using an empirical score (Dawoud et al., 2020), the low stringency settings were determined to improve sensitivity while not affecting specificity when applied to the WES data.

Upon manual review, eight regions of likely aUPD (belonging to 7 samples) that were not shortlisted by the SNP array-based analysis and empirical score were reclassified as TP results (Figure 4.4) which further improved the sensitivity of the WES based analysis. The visual inspection and re-classification of these eight AI regions indicated that the empirical score appeared to overlook AI regions with high mBAF values (>0.8) that are indicative of either high mutation burden or possible germline inheritance. It is important to note that the rationale behind the empirical score developed by Dawoud et al. was to select for somatic events and exclude potential constitutional runs of homozygosity which is probably why the empirical scoring system was less discriminative when applied to WES data than the *gg score* (Section 4.3.7). However, visual inspection of these AI regions and their correlation with *JAK2*^{V617F}, a known somatic driver mutation, suggests that a proportion at least are indeed high level aUPDs. This observation led to the decision of using manual annotation of shortlisted BAF plots as the gold standard for aUPD detection. It is also important to consider what proportion of the AI regions that pass automated filtering and manual review might have germline origins which could be

determined by analysing another DNA source in parallel as a germline source (e.g. DNA extracted from fibroblasts or cultured T-cells).

After analysing the entire UKB-WES50 and removing low quality samples, the frequency of AI positive samples was 67.1% which is much higher than expected (1-2%) (Genovese et al., 2014; Jacobs et al., 2012; Laurie et al., 2012). This highlighted the need to develop an automated filtering method for selecting putative AI regions that would align more closely with expectations. For this purpose, the AI regions were manually reviewed and categorised as either FP (n=3,643) or likely aUPD (n=157) (Table 4.9). The data were split into training (70%) and test (30%) sets that contained equal ratios of FP to likely aUPD regions (110/47) and logistic regression was used to develop a classifier, the *gg score*, which models the probability of aUPD. The highest ROC-AUC (96%) was obtained using a model consisting of seven features (*mBAF*, *bases per informative marker*, *coverage*, *centromere overlap*, *number of informative markers*, *bases per marker* and *heterozygosity rate*). Although the model performed well, this was largely due to the correct prediction of TN (3619/3643) which accounted for 90% of the observations in the test set that were correctly classified. On the other hand, only 55% of the likely aUPD regions were correctly predicted (84/157). These differences are thought to result from the unbalanced dataset which contained far more FP than likely aUPD observations.

To validate the method, BAF segmentation, quality control and the *gg score* filter were applied to an independent case control cohort consisting of 12,380 samples (Schizo-WES02, Section 4.2.7). A total of 172 likely aUPD regions were identified with a *gg score* above 0.5. Of these, 29 regions in 26 samples were confirmed as aUPD by visual inspection of the BAF plots. The frequency of CH in the Schizophrenia cohort was therefore determined to be 0.21% (26/12380) which, given the cohort's mean age of 65 (Genovese et al., 2014), is significantly lower than the expected frequency of 2-3% in individuals over 50 years old and 10% in the elderly aged 65 and older (Genovese et al., 2014; Jacobs et al., 2012; Laurie et al., 2012). In the Schizophrenia cohort, WES identified an average of 22,926 variants per sample (Table 4.8) that were used to detect AI regions which is significantly lower than SNP arrays, which typically provide between 500K to 1 million SNPs per sample. Furthermore, WES based estimates of per SNP BAF, which form the raw input for AI detection, are less accurately calculated than SNP arrays because they are determined by the sequencing depth which is limited and varies across the genome. This combination of low variant density and imprecise estimates of per SNP BAF is likely to make WES much less sensitive than arrays and might explain why this technique detected fewer regions of aUPD than expected. During the filtering steps of the VCF file, most variants were excluded due to MAF<0.01 and low

read depth (Table 4.8). Considering that the power to identify aUPD could be related to the number of variants these filters should be reassessed although there is likely to be a trade-off between SNPs density and high-quality variants with accurate VAF calculation.

After identifying likely aUPD regions that overlapped in at least two samples, five genes (*MPL*, *TET2*, *EZH2*, *JAK2*, and *FLT3*) were screened for potential somatic driver mutations in the WES VCFs. The *JAK2*^{V617F} mutation was identified in five out of five samples with UPD9p (Appendix Table B.1). A missense variant in *FLT3*, c.C680T:pT227M (rs1933437), which is predicted to be deleterious based on SIFT (Table 4.11), was identified in two samples with UPD13q-positive samples (Appendix Table B.1). However, somatic driver mutations in *FLT3* seen in myeloid neoplasms such as internal tandem duplications and *FLT3*^{D835Y} (Nguyen et al., 2017) are activating. It is therefore unlikely that the deleterious mutation (rs1933437) is the underlying cause of UPD13q seen in these samples. Despite the fact that target genes of the aUPD regions considered here are well known, driver mutations in most of the samples were not identified. The absence of detectable mutations in these regions might be due to several reasons such as genetic heterogeneity. For example, in an analysis of patients with myeloid neoplasia, only 7/12 cases with aUPD7q had discernible mutations of *EZH2* (Ernst et al., 2010). Other studies have also described the absence of somatic driver mutations in apparently healthy individuals with CH, as determined by WGS. Despite the absence of driver mutations, CH was still a risk factor for the development of haematological malignancies and overall survival (Holstege et al., 2014; Zink et al., 2017). One possible explanation for these findings might be a proliferative advantage provided by clonally inherited epigenetic states, and it is possible that such states might also be related to regions aUPD.

When this study was initially conceived it was hoped that large WES datasets would provide a useful resource to identify regions of aUPD and associated mutations. Following the development of this work and the opportunity of analysing over 60,000 exomes, it became apparent that much bigger samples sets would be required and I was able to estimate that we need at least 4-5 times the sample size to detect likely aUPD for a population of comparable age to UK Biobank (median age = 58 years at recruitment) or the Schizophrenia cohort (median age = 65 years).

Chapter 5 Conclusions and future work

This study began with the design of a GWAS to identify germline predisposition to mastocytosis. The study was built on two observations. First, results from previous GWAS of MPN, which are a rare group of blood cancers that are loosely related to mastocytosis, indicated that inherited common variants can influence the risk of developing MPN (Hinds et al., 2016; Kilpivaara et al., 2009; Tapper et al., 2015). Second, evidence in the literature showed some familial clustering of mastocytosis cases (Broesby-Olsen et al., 2012; Hartmann et al., 2005; Molderings et al., 2013; Zanotti et al., 2013) which suggested a heritable component in this disorder. Given these lines of evidence, it was hypothesised that germline factors influence the risk of developing mastocytosis and that these factors would be identified by a GWAS. To this end, a two-stage case-control GWAS of mastocytosis was conducted in five European populations which consisted of 1,035 patients with *KIT*^{D816V}-positive disease and 17,960 healthy controls. This represents the first two-stage mastocytosis GWAS and the largest cohort assessed to date. According to these sample sizes the study was estimated to have 80% power to detect common SNPs (MAF=0.4) with a relative risk of 1.56 and rare SNPs (MAF=0.1) with a relative risk of 1.82.

The mastocytosis GWAS identified three genome-wide significant SNPs that replicated in independent cohorts without evidence of heterogeneity, thus providing strong evidence that inherited common genetic variants increase the risk (OR<1.52) of developing mastocytosis in European populations. To begin to understand how these SNP predispose to mastocytosis, a range of *in silico* analyses (functional and epigenomic annotation, eQTL and mQTL in blood) which identified *TEX41*, *CEBPA* and *TBL1XR1* as the potential target genes involved. The involvement of these genes in mastocytosis was discussed in detail in **Chapter 3**. The association between reduced expression of *CEBPA* and rs4616402 is likely to promote a cellular environment that is more favourable to mast cell growth. Based on its known roles in normal and abnormal haematopoiesis, *CEBPA* is a very strong candidate that is ripe for evaluation in model systems in conjunction with mutant *KIT*. On the other hand, the potential role of *TEX41* and *TBL1XR1* in mastocytosis was less clear. A separate gene-based analysis of the stage 1 data identified *VEGFC* as an additional significant gene after correcting for multiple testing. A recent study has shown that *VEGFC* is significantly expressed in mastocytosis patients (Marcella et al., 2021), and the link between *VEGFC* and mastocytosis deserves further investigation. A small single-stage GWAS with only 234 cases has recently reported several genetic variants predisposing to systemic mastocytosis (Nedoszytko et al., 2020). Of these (Table 2.11), only one association was confirmed

in our GWAS (rs1800925, P-value$_{imputed}$=0.008). This observation supports the critical importance of robust replication to confirm the signals identified in the discovery stage.

Despite producing novel findings, my GWAS had some limitations. The sample size for both the discovery and replication cohorts was small compared to most published studies in other conditions, but it is important to reiterate that mastocytosis is a rare disease (prevalence 1–9/100,000). Given the constraint on the sample size, more effort with a larger sample size would be statistically more powerful and probably generated more significant results. In addition, it would have been highly desirable to have had cohorts with much more complete annotation to enable a more systematic comparison between associated SNPs and clinical features, outcomes and laboratory data. For some of the populations I studied this information was not available, either because it had not been collected or due to constraints with regard to patient consent.

It is known that large-scale genomic studies, such as GWAS, have predominantly been performed in European (52%) and Asian populations (21%) (Sirugo et al., 2019). Some populations (e.g., African, Hispanic and other minority groups) are under-studied and under-represented in genomic databases (Popejoy and Fullerton, 2016). Several examples in the literature have demonstrated that novel risk variants can be identified through GWAS in ethnically diverse populations (Adeyemo et al., 2019; Bick et al., 2020; Kilpeläinen et al., 2019). For example, a GWAS performed to investigate inherited predisposition to T2D in Africans confirmed several known markers and identified a novel *ZRANB3* locus predisposing to T2D (Adeyemo et al., 2019; Bick et al., 2020; Kilpeläinen et al., 2019). This variant is specific to Africans and would had not been discovered in studies performed only with persons of European ancestry. Such variants can be identified only in certain populations either because some variants have a higher frequency or are only present in those populations, or markers can have significant differences in LD across different ethnicities (Sirugo et al., 2019). One relevant study performed by Bick et al. identified three *TET2* variants associated with CHIP status, and one specific locus (rs144418061) was specific to individuals of African ancestry. This SNP was presented in **Chapter 2** (Table 2.10) when the stage 1 results were scrutinised to see if associations from the stage 1 data were seen with other SNPs that predispose to MPN or CHIP (Bao et al., 2020; Bick et al., 2020).

The under-representation of certain populations from genetic research will ultimately lead to persistent bias when discoveries are translated into clinical applications. For instance, the derivation of polygenic risk scores (PRS) from European-based studies may be inaccurate in under-studied populations. PRS is a common tool for predicting the genetic predisposition to a disease. The score represents a metric based on cumulative effect sizes of large numbers of common SNPs discovered by GWAS and it can be used to stratify population into individuals that have a higher or lower risk of developing the trait of interest (Lambert et al., 2019). It has been claimed that these tools can result in cumulative risks that are comparable to monogenic disease for some conditions and this emphasises their potential utility in clinical practice and the need for more focus on other populations (Khera et al., 2018; Peprah et al., 2015). My study was not sufficiently powered to identify a large number of associated SNPs that would be required for a mastocytosis PRS, but larger studies that include clinically annotated, ethnically diverse groups should be considered in future mastocytosis GWAS to facilitate the identification of new genetic variants associated with this rare blood cancer.

In **Chapter 2** the criteria for selection of SNPs to take forward in stage 2 were outlined (Section 2.2.10). The most significant SNPs and less significant index SNPs mapping close to a list of functionally relevant genes (Appendix Table A.2) were selected for further analysis. This strategy aimed to maximise the selection of likely relevant SNPs whilst minimising the number that were selected for analysis at stage 2, both for reasons of cost but also statistical power taking into account the need to correct for multiple testing. However, this strategy may have overlooked important SNPs due to lack of relevant knowledge at the time. For example, a recent study on the immunoregulatory roles of members of the human leukocyte immunoglobin-like receptor (LILR) family identified LILRB3 as a novel myeloid checkpoint receptor with immunosuppressive functions (Yeboah et al., 2020). Members of the LILR family are categorised in activating subfamily A (LILRA1-6) or inhibitory subfamily B (LILRB1-5) (van der Touw et al., 2017). The stage 1 results of the mastocytosis GWAS identified an intergenic SNP rs422948 (P-value = 2.2×10^{-4}) located on chromosome 19 between *LILRA6* and *LILRB5*. *LILRB3* could potentially be relevant to mastocytosis, and with this knowledge of functional relevance, rs422948 would have certainly been selected in our GWAS for replication and should be considered for future replication studies.

Reproducibility has always been key in the scientific method, and replication in GWAS has been highlighted to improve the credibility of the study while controlling for biases and spurious associations (Kraft et al., 2009). The Manhattan plot of the stage 1 meta-analysis (Figure 2.8) showed consistent signals on chromosome 4 and 11, which serve as a good example to illustrate

why replication is so important. The most significant SNP on chromosome 11 and its backup SNP failed genotyping in stage 2 cases, but should be considered for inclusion in future studies. The most significant SNP on chromosome 4 (rs6820170, P-value=9.3×10^{-7}) and its backup SNP (rs11131764, P-value=1.58×10^{-6}) were both selected for replication. rs6820170 was successfully genotyped in all five stage 2 cohorts; however, it was not significant in any of them. It is possible that the variant is population-specific, which could explain the high heterogeneity identified between cohorts (I^2=68.78) and the failure of replication. Alternatively, it is possible that the association seen at stage 1 was simply a random sampling effect which would also be consistent with the high heterogeneity between cohorts. This result from **Chapter 2** makes it clear that the lack of replication can lead to false positive results and highlights the importance of confirming signals that have been identified at the discovery stage. It is strongly recommended to always include independent replication cohorts in GWAS.

The statistical analysis outlined in **Chapter 2** also identified an interesting suggestive association between mastocytosis and rs58124832 (P-value$_{meta}$=9.03×10^{-6}, Appendix Table A.6), a SNP that our eQTL analysis showed to be associated with *TPSAB1* and *TPSB2* expression (Lloyd-Jones et al., 2017; Võsa et al., 2018b) in blood. This association is also supported by the gene-based test analysis presented in **Chapter 3**, which identified *TPSAB1* as the second most significant gene (P-value=2.3×10^{-4}, Table 3.5); however, it did not maintain significance following Bonferroni correction. The same SNP has also been associated with *TPSAB1* duplication (Lyons et al., 2018), and importantly a study has recently linked *TPSAB1* to mastocytosis (Greiner et al., 2021). Our current analysis does not include copy number analysis of *TPSAB1* in mastocytosis patients, but these important findings have generated further questions on whether rs58124832 and *TPSAB1* copy number are correlated in *KIT*^{D816V}-positive cases and what the mechanism behind it might be. As a result of this observation, further investigation is needed to explore the relationship between *KIT*^{D816V} and *TPSAB1* duplications in mastocytosis patients. In the context of the gene-based test analysis, *VEGFC* was the only gene significantly associated with mastocytosis after adjusting for multiple corrections. It is important to reiterate that the test was performed using the stage 1 summary statistics and to note that this association should be tested in an independent mastocytosis cohort.

A set of new human cell lines called ROSA$^{KIT\ WT}$ and ROSA$^{KIT\ D816V}$ are cell lines of MC established from normal haematopoietic progenitors (Saleh et al., 2014). ROSA has been shown to be a

valuable tool in mastocytosis studies (Marcella et al., 2021) and its use would facilitate the investigation of genes selected in this study as well as responses to targeted drugs. The majority of mastocytosis patients (over 90%) carry the *KIT*^{D816V} mutation and only limited treatment options targeting this mutation are effective and available; midostaurin (Arock et al., 2015) and avapritinib (ClinicalTrials.gov Identifier: NCT03580655, DeAngelo et al., 2021). Further validation of the genes identified by the GWAS as well as more detailed studies to link genetic variation with specific clinical features will help to better understand the pathogenesis of this disease and might potentially aid the development of targeted therapies that could also be effective for *KIT*^{D816V}-negative patients. For example, my study linked *TPSAB1* to mastocytosis, and tryptase encoded by *TPSAB1* is a potential therapeutic target (Caughey, 2016).

The second part of my study focused on the potential of large population-based genomic datasets to yield new information that is relevant to cancer. Regions of aUPD are known contributors to cancer since they are associated with driver gene mutations in both haematological malignancies and solid tumours (Score and Cross, 2012; Torabi et al., 2019; Tuna et al., 2012; Walsh et al., 2008). The discovery of recurrent regions of UPD9p, for example, facilitated the identification of *JAK2*^{V617F} mutation in MPN patients (Tiedt et al., 2005). The search for common regions of aUPD has been used as a research tool to identify many driver genes. However, the presumptive target gene or genes remains unidentified for many regions of aUPD. WES datasets were utilised to develop an effective method for detection of regions of AI (CNV and CNN-LOH), working under the hypothesis that uncovering novel and recurrent region of aUPD would facilitate the identification of novel drivers of myeloid neoplasms.

Based on the results presented in **Chapter 4,** it can be concluded that using WES data, more than 250,000 samples from older individuals are needed to identify novel recurrent regions of aUPD in association with CHIP and haematological malignancies. In my study, novel recurrent AI regions were not identified, and *JAK2*^{V617F} was the only causing-disease mutation associated with aUPD identified in the Schizo-WES02 cohort. Regions of aUPD with unknown driver genes were only detected in single samples, which due to the size of AI regions makes it challenging to shortlist potential somatic mutations.

In the effort to identify novel AI regions, some challenges were encountered. Specifically, following QC, most markers were removed due to MAF<0.01, leaving with an average marker density per sample of only 23,000 in the Schizo-WES02 cohort and 36,000 in UKB-WES50 (Table

4.8). This ultimately could explain what makes WES data much less sensitive than arrays and why AI regions could be identified in only 0.3% of samples. This frequency is much lower than expected. The rationale behind retaining only common SNP was to make the data available more similar to the SNP array data. However, it is known that the power to identify likely aUPD regions is dependent on the number of variants, and as a result of this observation, the MAF filter should be reassessed. A further investigation would be beneficial to avoid a drastically reduced number of markers and instead guaranteeing that only high-quality markers are kept. This would enable accurate calculation of VAF and calling of AI regions.

The limitations observed in the analysis presented in **Chapter 4** allow the opportunity to consider alternative approaches for identifying AI regions. The adoption of software developed specifically for WES data (e.g. ExomeAI and hapLOHseq) seems to be a plausible solution to explore whether a genomic resolution more similar to the SNP array could be reached (Nadaf et al., 2015; San Lucas et al., 2016). For instance, a tool initially developed for SNP array data (hapLOH) has also been implemented for the detection of AI from WES data (hapLOHseq) and it can discriminate between CNV-LOH and CNN-LOH (San Lucas et al., 2016; Vattathil and Scheet, 2013). The hapLOHseq algorithm identifies AI events of 10 Mb or more in 16% of samples using WES data with depth coverage of 80× (San Lucas et al., 2016) and has been used in some very recent studies to determine regions of AI (Lee et al., 2020; Semaan et al., 2021; Sivakumar et al., 2021). As described in the methods of **Chapter 4** (Section 4.2.2), the Schizo-WES02 has a mean coverage of 90× (Ganna et al., 2016). Thus, the sequencing depth would be sufficient to accurately detect AI regions in a greater number of samples. This approach might help in the future to overcome the limitations that were encountered with BAF segmentation. The use hapLOHseq and ExomeAI could also be an interesting opportunity to compare the performance of BAF segmentation versus other tools that have been developed for WES data.

The *gg score* described in **Chapter 4** has helped to identify with high confidence AI regions from the Schizo-WES02 and part of the UKB-WES50 datasets. However, regions identified in both datasets were not combined. To increase the power of the study to detect new somatic signatures in the genome, the *gg score* could be applied to the entire UKB-WES50 cohort and additional datasets from public databases. Furthermore, integrating the results from different sample sets will represent a valuable resource to detect other recurrent regions of AI to screen for

novel driver genes. Application of these tools to new, rapidly growing WGS datasets will provide much greater resolution and power to detect new abnormalities.

In conclusion, this thesis presents different genetic approaches to better understand genetic factors associated with myeloid neoplasms. First, I presented the results of a GWAS investigating the inherited predisposition to mastocytosis. Following this analysis, novel genetic variants associated with mastocytosis were identified, and several genes that emerged from this work were nominated for further investigation. In the second part of my work, I took advantage of WES datasets to develop a method that could aid the identification of AI regions. I highlighted the limitations that were encountered in both studies and provided potential avenues for future research. This is an exciting time to be studying cancer genomics since the advancement of powerful genomic technologies and large population biobanks have enabled the scientific community to pave the way toward personalised medicine.

Appendix A Supplementary Data for Chapter 2 and 3

Table A.1 **Genome-wide significant results caused by AT/GC unresolved strand issues**

CHR	BP	SNP	A1	A2	P
6	153316274	rs6553229	C	G	2.52E-189
13	95903610	rs7335275	T	C	3.13E-122
5	3266396	rs1661068	T	C	3.00E-118
1	159969008	rs6676862	C	T	6.00E-118
2	104115536	rs6736012	T	C	3.91E-96
6	107929025	rs9373956	G	A	2.74E-95
11	16300057	rs16932876	T	C	5.31E-89
1	29475394	rs2230677	G	C	1.80E-86
5	66139238	rs2441109	A	G	3.34E-86
5	71331087	rs7711863	A	C	1.26E-79
14	39556185	rs8018720	C	G	3.68E-77
4	7802227	rs28406288	G	C	6.09E-77
3	102157365	rs6784362	A	T	8.86E-77
2	128379563	rs61744148	C	G	1.08E-76
7	21628242	rs62441683	C	G	4.93E-76
17	48595988	rs8064455	G	C	5.02E-76
5	129040056	rs11749126	A	T	1.27E-75
5	122685727	rs1047437	C	G	1.33E-75
19	36303664	rs3848666	G	C	1.79E-75
6	36446975	rs2239808	G	C	2.24E-75
14	39722023	rs10134365	G	C	2.38E-75
3	38633923	rs11708996	G	C	3.52E-75
20	43836173	rs2301366	T	A	4.79E-75
8	130572110	rs10956483	G	C	8.37E-75
19	58213773	rs2188736	G	C	1.01E-74
9	133936571	rs3739510	G	C	1.17E-74
6	4087934	rs619483	G	C	1.59E-74
19	5778517	rs2305925	A	T	1.69E-74
12	53189696	rs28721426	C	G	4.23E-74
5	151775064	rs4958535	C	G	5.83E-74
17	35743010	rs1714987	C	G	6.35E-74
11	20648364	rs3740870	G	C	6.42E-74
6	31842598	rs12661281	T	A	8.94E-74
7	139138950	rs17160911	C	G	8.96E-74
11	5510598	rs7950082	A	T	9.05E-74
5	145894896	rs7709485	G	C	1.19E-73
7	105615426	rs34426483	G	C	1.26E-73
7	89938588	rs7803620	G	C	1.76E-73
11	56000403	rs10791893	G	C	4.42E-73
8	101648164	rs2187016	C	G	5.40E-73
21	48063476	rs10854485	G	C	9.04E-73
7	139415775	rs7456421	G	C	1.25E-72
11	62911079	rs1939748	C	G	2.72E-72
17	42164885	rs228757	C	G	1.50E-71
17	7606722	rs7640	C	G	2.44E-71
9	1056959	rs17641078	G	C	2.95E-71
11	108044091	rs4144901	T	A	1.97E-70
19	8160897	rs12150963	G	C	2.01E-70
4	110914427	rs4698803	T	A	4.06E-70
19	7606908	rs17854645	G	C	4.34E-70
16	57738810	rs58373934	T	A	5.37E-70
20	40805278	rs733976	G	A	2.40E-69
6	6728616	rs9504905	A	G	5.39E-69
17	4574751	rs9436	T	A	3.01E-68
7	7545691	rs10486176	G	C	7.03E-68
2	238243464	rs2270669	G	C	4.98E-67
3	169539812	rs61738871	C	G	5.31E-67
11	94326765	rs57607909	G	C	7.76E-67
14	75248652	rs45617140	C	G	5.50E-66
16	48204078	rs61739606	T	A	1.14E-65
1	54681920	rs13571	C	G	2.60E-65
17	29159404	rs999796	G	C	8.41E-65
4	100274157	rs4147541	G	C	3.82E-64
9	90501514	rs4076794	C	G	4.19E-64
1	26608896	rs757886085	C	G	4.59E-64
6	152647681	rs2306916	T	A	5.56E-64
12	96292170	rs75959092	C	G	7.43E-64
19	19823270	rs12973901	T	A	1.26E-63
11	55563336	rs76383258	A	T	2.94E-63
8	6118677	rs7834259	A	G	5.77E-63
11	56237722	rs605734	T	A	9.63E-63
5	53815240	rs61739378	G	C	1.66E-62
7	102574920	rs1057066	G	C	2.33E-62
14	55448409	rs61741224	G	C	2.65E-62
11	55999950	rs12221615	G	C	2.68E-62
12	32137512	rs3759299	C	G	2.89E-62
9	100823135	rs1058446	G	C	2.93E-62
6	29589666	rs29220	C	G	1.83E-61
11	5021055	rs61734126	C	G	2.76E-61
15	75498744	rs1873379	G	C	3.93E-61
7	150935430	rs3748098	G	C	3.16E-59
3	124731689	rs78680419	T	A	5.67E-59
5	75591710	rs2270927	C	G	5.93E-59
19	38378539	rs10422056	C	G	2.58E-58
17	649505	rs4968104	T	A	2.94E-58
19	57839567	rs1968090	A	T	3.60E-58
16	64453857	rs9673844	C	T	1.45E-57
6	96053922	rs35772543	T	A	1.86E-56
6	32680640	rs7764856	T	A	7.71E-56
14	21993638	rs2242527	G	C	7.74E-56
7	135082953	rs77841106	G	C	7.86E-56
5	134364518	rs479632	C	G	1.21E-55
11	124789828	rs78859654	A	T	1.57E-55
17	4689313	rs2279961	G	C	8.39E-55
20	47246077	rs3936192	C	G	3.73E-54
3	14174427	rs4685076	A	T	4.77E-54
20	746197	rs35655964	G	C	4.22E-53
6	37990758	rs259678	G	A	3.24E-52
12	9346792	rs12230214	G	C	7.42E-52
11	993907	rs11538725	C	G	8.26E-52
22	31535872	rs2074735	G	C	1.22E-51
9	116132092	rs818711	G	C	1.23E-51
9	131588888	rs6478854	G	C	3.12E-50
17	71197323	rs62621249	G	C	6.38E-50
19	12154799	rs67102109	G	C	3.33E-49
3	107417178	rs11918431	A	T	1.50E-48
12	122689181	rs7136356	C	G	1.55E-48
20	62492922	rs2281534	T	A	2.83E-48
16	71509796	rs8050871	G	C	3.33E-48
2	238247734	rs36104025	C	G	8.09E-48
8	68421768	rs17853192	G	C	1.80E-47
6	38953292	rs4380739	G	A	2.39E-47
10	81697868	rs3088308	A	T	5.79E-47
1	161161284	rs41270041	G	C	1.69E-46
12	121615131	rs2230911	C	G	2.37E-46
21	45945648	rs35028190	G	C	1.53E-45
11	6644427	rs35599968	C	G	3.41E-45
1	214818223	rs3748693	T	A	5.20E-45
11	59132798	rs7941190	C	G	3.19E-44
3	39373902	rs2853699	C	G	1.53E-43
15	89169703	rs8026929	C	G	1.84E-43
1	119575818	rs3790549	C	G	1.55E-38
8	6464034	rs2454518	C	T	3.21E-38
5	113279595	rs6894635	A	G	3.30E-38
19	44833851	rs4280359	C	G	1.03E-37
8	133733145	rs10216529	A	G	2.97E-37
16	88781614	rs1058158	G	C	6.68E-37
19	34584137	rs574004	A	G	5.30E-36
1	236700807	rs1126407	A	T	2.05E-35
19	37677748	rs45626541	T	A	1.30E-34
15	40655845	rs1898883	G	C	5.54E-34
1	153390542	rs3006412	G	C	1.50E-33
5	9021310	rs1598826	C	T	1.24E-31
7	130413311	rs3909552	C	T	1.68E-31
11	34378381	rs1925368	G	C	2.01E-31
18	28611061	rs276937	A	T	8.24E-31
2	225588620	rs388591	A	G	2.35E-30
1	179562740	rs61310274	G	C	1.15E-29
6	32636866	rs3134996	T	A	2.73E-29
2	165476253	rs61748245	T	A	5.33E-29
3	195515617	rs2641776	C	G	3.47E-28
11	124440362	rs55861866	C	G	5.56E-28
9	112398754	rs1358917	G	A	1.70E-27
22	25601196	rs9608378	G	C	2.82E-26
4	100516022	rs2306985	C	G	4.95E-26
11	34483894	rs554576	A	T	7.16E-24
15	42371752	rs4924618	A	T	3.64E-23
5	66538400	rs11951571	A	G	5.74E-23
1	206231264	rs33985287	A	G	7.13E-23
16	4790204	rs61731839	C	G	1.77E-22
1	19565344	rs709683	C	G	2.31E-22
7	158117269	rs10949716	A	G	4.42E-20
12	11339020	rs35969491	A	T	6.56E-20
14	93005721	rs10498634	A	G	9.08E-19
16	5294643	rs2333764	T	C	6.53E-18
20	57538175	rs16982339	G	A	2.32E-17
9	136494425	rs2519765	T	C	2.55E-17
13	106477029	rs7981276	T	C	6.42E-17
17	61901197	rs2727288	A	T	9.97E-17
2	98844674	rs7601049	G	C	4.65E-16
1	223949314	rs28370127	C	G	8.12E-16
7	141673345	rs713598	C	G	8.61E-16
4	40722850	rs2200061	A	G	2.12E-15
7	129979092	rs7797072	T	C	1.31E-14
15	37163043	rs1450421	A	G	2.94E-14
6	52112717	rs608137	T	C	1.23E-13
12	127402542	rs1683739	A	G	3.41E-13
17	60810875	rs35432569	C	T	1.26E-12
11	21240934	rs1791869	C	T	4.67E-12
13	75339229	rs2094437	T	C	2.01E-11
11	19189086	rs10833066	G	T	3.58E-11
11	8947283	rs3751066	C	G	4.33E-10
14	96777468	rs3759601	G	C	3.25E-08
22	41726053	rs4822021	A	G	4.19E-08

The table lists the extremely significant results obtained during the preliminary analysis, when the AT/GC strand check is not addressed.

Appendix A

Table A.2 **List of genes with functional relevance**

Gene selected	Evidence for biological relevance
BMP2	J. Cancer genetics and cytogenetics: Bone morphogenetic protein antagonist gene NOG is involved in myeloproliferative disease associated with myelofibrosis.
CAMK2D	Reactome: RET signaling
CCND1	GO: regulation of protein kinase activity; KEGG: Acute/Chronic myeloid leukaemia
CCNG1	GO: regulation of cyclin-dependent protein serine/threonine kinase activity
CDH5	J. Cancer: Derivation of a new haematopoietic cell line with endothelial features from a patient with transformed myeloproliferative syndrome: a case report.
CDK6	KEGG: Chronic myeloid leukaemia
CEBPA	Civic: AML with mutated CEBPA' is a provisional entity in the WHO classification of AML and is recommended to be tested for in patients with AML. CEBPA mutations are particularly associated with cytogenetically normal AML (CN-AML). CEBPA mutations are associated with a favourable prognosis, however, NPM1 and FLT3 mutations should also be assessed in CN-AML patients as concurrent mutations may have prognostic implications. HPO: Acute myeloid leukaemia
CLNK	Entrez: MIST is a member of the SLP76 family of adaptors (see LCP2, MIM 601603; BLNK, MIM 604515). Swiss-Prot: MIST plays a role in the regulation of immunoreceptor signaling, including FC-epsilon R1 (see FCER1A, MIM 147140)-mediated MC degranulation (Cao et al., 1999 [PubMed 10562326]; Goitsuka et al., 2000, 2001 [PubMed 10744659] [PubMed 11463797]).[supplied by OMIM, Mar 2008]
DRD2	Go: activation of protein kinase activity
DUSP5	Reactome: RET signaling
EBPA	Reactome: RET signaling
EPHA4	GO: positive regulation of protein tyrosine kinase activity
ERBB4	Reactome:R-HSA-1433557, Signaling by SCF-KIT. Swiss Prot: Binding of a cognate ligand leads to dimerisation and activation by autophosphorylation on tyrosine residues. In vitro kinase activity is increased by Mg(2+). Inhibited by PD153035, lapatinib, gefitinib (iressa, ZD1839), AG1478 and BIBX1382BS.

Gene selected	Evidence for biological relevance
FYN	Reactome: Regulation of KIT signaling
FZD8	GO: positive regulation of JUN kinase activity
GFRA1	GO: RET signaling
HDAC9	Entrez Gene: Histones play a critical role in transcriptional regulation, cell cycle progression, and developmental events. This encoded protein may play a role in haematopoiesis. Journal Leukaemia & lymphoma: Increased gene expression of histone deacetylases in patients with Philadelphia-negative chronic myeloproliferative neoplasms.
HRH1	Entrez Gene: Histamine is a ubiquitous messenger molecule released from MCs, enterochromaffin-like cells, and neurons. Its various actions are mediated by histamine receptors H1, H2, H3 and H4. The protein encoded by this gene is an integral membrane protein and belongs to the G protein-coupled receptor superfamily. Multiple alternatively spliced variants, encoding the same protein, have been identified. [provided by RefSeq, Jan 2015]
IBTK	Entrez Gene: Bruton tyrosine kinase (BTK) is a protein tyrosine kinase that is expressed in B cells, macrophages, and neutrophils. The protein encoded by this gene binds to BTK and downregulates BTK's kinase activity. This gene has a pseudogene on chromosome 18. Alternative splicing results in multiple transcript variants encoding distinct isoforms. [provided by RefSeq, Jul 2014]. Swiss Prot: Acts as an inhibitor of BTK tyrosine kinase activity, thereby playing a role in B-cell development. Down-regulates BTK kinase activity, leading to interference with BTK-mediated calcium mobilisation and NF-kappa-B-driven transcription.
JAG1	Entrez Gene: The jagged 1 protein encoded by JAG1 is the human homolog of the Drosophilia jagged protein. Jagged 1 signalling through notch 1 has also been shown to play a role in haematopoiesis. [provided by RefSeq, Jul 2008]. Swiss Prot: Ligand for multiple Notch receptors and involved in the mediation of Notch signaling. May be involved in cell-fate decisions during haematopoiesis.
KCNJ2	DISEASES||HGMD||GeneCards: chronic myeloproliferative disorder
KIAA1804	GO: activation of JUN kinase activity
LRRC4C	GO: negative regulation of protein kinase activity
LRRK1	Swiss Prot: Binding of GTP stimulates kinase activity.
LRRTM4	GO: negative regulation of protein kinase activity
LTK	UniProt: Receptor with a tyrosine-protein kinase activity. The exact function of this protein is not known. Studies with chimeric proteins (replacing its extracellular region with that of several known growth factor receptors) demonstrate its ability to promote growth and cell survival. Signaling appears to involve the PI3 kinase pathway.

Appendix A

Gene selected	Evidence for biological relevance
MMP2	J. Cancer research: The effect of CXCL12 processing on CD34+ cell migration in myeloproliferative neoplasms.
NOG	J. Cancer genetics and cytogenetics: Bone morphogenetic protein antagonist gene NOG is involved in myeloproliferative disease associated with myelofibrosis.
NRTN	Reactome: RET signaling
PAQR3	Reactome: RET signaling; GO: negative regulation of MAP kinase activity
PDGFRA	Tocris: Platelet-derived growth factor receptors (PDGFRs) are catalytic receptors that have intracellular tyrosine kinase activity. They have roles in the regulation of many biological processes including embryonic development, angiogenesis, cell proliferation and differentiation; Reactome: RET signalling. HPO: Myeloproliferative disorder
PDGFRB	Entrez Gene: The protein encoded by this gene is a cell surface tyrosine kinase receptor for members of the platelet-derived growth factor family. The identity of the growth factor bound to a receptor monomer determines whether the functional receptor is a homodimer (PDGFB or PDGFD) or a heterodimer (PDGFA and PDGFB). A translocation between chromosomes 5 and 12, that fuses this gene to that of the ETV6 gene, results in chronic myeloproliferative disorder with eosinophilia. [provided by RefSeq, Aug 2017]; Reactome: RET signalling. HPO: Myeloproliferative disorder
PLA2G4A	BioSystems: Fc-epsilon receptor I signaling in mast cells
PPEF2	GO: regulation of MAP kinase activity
PRKCE	UniProt: Calcium-independent, phospholipid- and diacylglycerol (DAG)-dependent serine/threonine-protein kinase that plays essential roles in the regulation of multiple cellular processes linked to cytoskeletal proteins, such as cell adhesion, motility, migration and cell cycle, functions in neuron growth and ion channel regulation, and is involved in immune response, cancer cell invasion and regulation of apoptosis. During cytokinesis, forms a complex with YWHAB, which is crucial for daughter cell separation, and facilitates abscission by a mechanism which may implicate the regulation of RHOA. In differentiating erythroid progenitors, is regulated by EPO and controls the protection against the TNFSF10/TRAIL-mediated apoptosis, via BCL2.
RASGRP1	UniProt: Functions as a calcium- and diacylglycerol (DAG)-regulated nucleotide exchange factor specifically activating Ras through the exchange of bound GDP for GTP (PubMed:15899849, PubMed:23908768). Regulates T-cell/B-cell development, homeostasis and differentiation by coupling T-lymphocyte/B- lymphocyte antigen receptors to Ras (PubMed:10807788, PubMed:12839994). Functions in MC degranulation and cytokine secretion, regulating FcERI-evoked allergic responses (By similarity). May also function in differentiation of other cell types (PubMed:12845332); Reactome: RET signaling
RASSF2	GO: positive regulation of protein kinase activity
RBBP6	J. Blood: Germline RBBP6 mutations in familial myeloproliferative neoplasms.

Gene selected	Evidence for biological relevance
RGMA	DISEASES||HGMD||GeneCards: Leukaemia, Acute Myeloid
SOX9	GeneCards||OMIM||ClinVar||Orphanet||Swiss-Prot||GeneTests||HGMD||Novoseek||DISEASES: chronic myeloproliferative disorder
SYT1	HGMD||GeneCards||DISEASES: Mast-Cell Leukaemia
TACC1	J. Cancer genetics and cytogenetics: Combined translocation with ZNF198-FGFR1 gene fusion and deletion of potential tumor suppressors in a myeloproliferative disorder.
TBL1XR1	Orphanet||DISEASES: Leukaemia, Acute Promyelocytic, Somatic
TLE1	GeneCards: Core Binding Factor Acute Myeloid Leukaemia
TPSAB1	Entrex gene: Tryptases comprise a family of trypsin-like serine proteases, the peptidase family S1. Beta tryptases appear to be the main isoenzymes expressed in MCs; whereas in basophils, alpha tryptases predominate. [provided by RefSeq, Jul 2008]; HGMD||GeneCards||Novoseek: Systemic/Cutaneous Mastocytosis and Mast cell Disease. Swiss Prot: Tryptase is the major neutral protease present in MCs and is secreted upon the coupled activation-degranulation response of this cell type.
TPSD1	UniProt: Tryptase is the major neutral protease present in mast cells and is secreted upon the coupled activation-degranulation response of this cell type.
TRIM27	GO: negative regulation of protein kinase activity
VEGFC	GO: positive regulation of mast cell chemotaxis
YSK4	GO: activation of protein kinase activity
ZBTB20	UniProt: May be a transcription factor that may be involved in haematopoiesis, oncogenesis, and immune responses (PubMed:11352661).

Table A.3 **All regions of AI for one sample ID:10138**

Chr	Start	End	StartSNP	EndSNP	mBAF	HetRate	Median LRR	BpSize	NbrSNPsMBAF	NbrSNPsFull
1	1222596	1305561	exm2135	rs17160669	0.81	0.08	-0.36	82966	12	144
1	11850927	11863057	rs2274976	rs2066470	0.63	0.21	0.08	12131	8	38
1	97743805	98050656	rs641805	rs2811205	0.63	0.06	-0.39	306852	16	287
1	103165230	103578334	rs6684108	rs7543626	0.83	0.03	-0.56	413105	4	119
2	116894086	116905270	rs7579948	rs11903740	0.7	1	-0.15	11185	4	4
5	74798156	75514986	rs10055011	rs11960832	0.58	0.12	-0.06	716831	28	237
6	29782470	29789190	exm-rs1736959	exm-rs1610678	0.62	0.71	0.3	6721	12	17
6	32428285	32652359	exm-rs6903608	rs3021058	0.6	0.46	0.01	224075	66	143
6	42932200	43013046	exm547609	exm548040	0.67	0.06	0.2	80847	5	90
6	57761561	62673145	rs4236163	rs1192457	0.88	0.05	-0.24	4911585	8	151
8	144946092	145003862	exm728897	exm729894	0.77	0.09	-0.37	57771	12	141
9	140093908	140141794	exm802290	rs11497277	0.85	0.07	-0.07	47887	7	106
11	1017085	1018657	exm873676	exm873963	0.87	0.18	0.22	1573	7	39
12	80699475	81074138	exm1023804	rs11114567	0.85	0.09	-0.53	374664	13	140
14	48847571	49140883	rs1905824	rs946626	0.69	0.06	-0.34	293313	4	68
16	825003	855732	exm1199126	exm1199724	0.84	0.04	-0.2	30730	4	109
16	1538464	1559399	rs2745103	rs3829558	0.76	0.13	-0.33	20936	5	38
16	3598190	3763179	exm1211487	rs129988	0.81	0.02	-0.06	164990	4	221
17	19648316	21189598	exm1303404	rs1466314	0.85	0.01	-0.01	1541283	4	285
20	8186186	8206986	rs6055645	rs6133556	0.74	0.5	-0.42	20801	4	8
22	38065655	38822300	rs12628135	rs196057	0.58	0.19	0.05	756646	72	387
X	2655180	28817458	rs11575897	rs9786224	0.64	0.74	-3.96	26162279	1028	1389

The whole file contains all the regions of AI for each sample that was run through BAF segmentation. Each row of the table contains information for each segmented region identified as AI. The information are reported in each column as follows: Chr: chromosome; Start: start of the AI breakpoint; End: end of the AI breakpoint; StartSNP: SNP name where the breakpoint starts; EndSNP: SNP name where the breakpoint ends; mBAF: mirrored BAF value; HetRate: heterozygosity rate per segmented region; Median LRR: median Log R Ratio of the segment; BpSize: length of the segmented region measured in base pair; NbrSNPsMBAF: number of SNPs used to estimate the mirrored BAF; NbrSNPsFull: count of the total number of SNPs that are present in the segmented region.

Table A.4 **Per chromosome regions spanned by SNPs**

Chr	From SNP	To SNP	From Location	To Location	Length
1	rs4477212	rs12746903	82154	249218992	249136838
2	rs10195681	rs12478296	18674	243048760	243030086
3	rs13060385	rs10433653	61495	197838262	197776767
4	rs13125929	rs3903261	71566	190963766	190892200
5	rs9313223	rs876154	25328	180693127	180667799
6	rs412135	rs12530134	108666	170919470	170810804
7	rs7456436	rs1124425	44935	159119486	159074551
8	rs11780869	rs6599566	164984	146293414	146128430
9	rs10814410	rs9314655	46587	141066491	141019904
10	rs11252127	rs11528930	98087	135477883	135379796
11	exm869284	rs12294124	193146	134934063	134740917
12	rs11063263	exm1054977	191619	133810935	133619316
13	rs2762261	rs17067959	19058717	115103529	96044812
14	rs28842485	rs10149476	19255726	107287663	88031937
15	rs12905389	rs4098905	20071673	102461162	82389489
16	rs2541696	rs13331261	88165	90274695	90186530
17	rs2396789	rs9897769	8547	81060040	81051493
18	rs12455984	rs12960632	13034	78015180	78002146
19	rs8100066	rs10411093	260912	59097160	58836248
20	rs6139074	rs10460610	63244	62934877	62871633
21	rs28971224	rs10483083	10827533	48100155	37272622
22	rs12157537	rs5771007	16114244	51195728	35081484
Total	2792045802				

Chr: Chromosomes; From SNP: first SNP; To SNP: last SNPs; From location: start chromosomic location; To location: end chromosomic location; Lenght: length of the spanned region. The SNP highlighted in blue was withdrawn from the Reference SNP (rs) cluster on September 2016 due to mapping or clustering errors (Sherry, 2001).

Table A.5 **Sample outliers excluded from BAF segmentation analysis**

SAMPLE	No AI REGION	SumBp/SAMPLE	% AI REGION	MAX Bp
4008	32	2791706788	99.9879	249130276
803	52	2791327191	99.9743	249136839
4610	73	2791026971	99.9635	245610463
1655	89	2790982747	99.9619	220384060
9689	92	2790769031	99.9543	243025474
9104	130	2790549932	99.9464	199752297
4605	425	2790392927	99.9408	59506200
10543	165	2789021333	99.8917	154390319
MLL_10052	227	2789021961	99.8917	104837652
11709	88	2788193664	99.862	174951592
3075	33	2788155651	99.8607	243009823
10780	56	2784558399	99.7318	197623770
9632	372	2783314323	99.6873	98824831
2889	861	2779204191	99.5401	58168924
8396	2050	2767480533	99.1202	53250172
11694	1793	2763726818	98.9857	33126930
11345	1630	2762154763	98.9294	44259447
1692	1594	2761290069	98.8985	42351605
92	2226	2732432452	97.8649	65718106
3717	2147	2701525209	96.7579	39967087
12057	2100	2698551809	96.6514	36505632
11309	4734	2503132860	89.6523	33243631
6439	3630	1078610906	38.6316	7891910
MLL_09977Ra	3369	784428638	28.0951	25068432

Sample: sample IDs; No AI region: total number of AI regions per sample; SumBp/Sample: sum of the length of all AI regions per samples; % AI region: percentage of AI per sample; Max Bp: biggest AI region in each sample. The 19 sample IDs in black were already removed from the analysis because of more than 10% of missingness. The samples highlighted in blue were identified as outliers and removed from the BAF segmentation analysis.

Table A.6 **GWAs results from stages 1 and 2 for all SNPs selected for replication**

Table A.7 **Imputation and analysis of SNPs spanning *TERT***

Table A.8 **Functional annotation for GWAS significant SNPs and their proxies in high LD ($r^2 \geq 0.8$)**

Table A.9 **Functional annotation for *VEGFC* lead SNPs and their proxies in high LD ($r^2 \geq 0.8$)**

Link to view/download the tables https://doi.org/10.5258/SOTON/D2266

Table A.10 **Methylation quantitative trait loci (mQTL) for rs13077541 in blood.**

Trait	CHR	START	END	Effective_Allele	Effect_Size	Effect_Size_Desc	SE	PVAL	FDR	PMID	Sample_Size
cg01132484 (chr3:176916496)	3	176916496	176916496	NA	-0.2388288	beta	0.03152906	1.03E-13	0.00000104	27036880	771
cg01132484 (chr3:176916496)	3	176916496	176916496	NA	-0.276544	beta	0.03328853	4.41E-16	2.67E-09	27036880	764
cg01132484 (chr3:176916496)	3	176916496	176916496	NA	-0.2825467	beta	NA	2.62E-20	2.07E-13	27036880	742
cg01132484 (chr3:176916496)	3	176916496	176916496	NA	-0.2996571	beta	NA	1.87E-23	1.89E-16	27036880	834
cg01132484 (chr3:176916496)	3	176916496	176916496	NA	-0.3387203	beta	NA	2.67E-27	2.96E-20	27036880	837

Appendix B Supplementary Data for Chapter 4

Table B.1 **The 29 likely aUPD events detected in the Schizo-WES02 cohort.**

chr	start	end	mBAF	HetRate	size	No informative SNPs	No all SNPs	Bases per marker	Bases per informative marker	No merged	New size	coverage	Centromere overlap	ggscore	Dawous's score	annotation
1	69270	17264920	0.73	0.36	13211375	229	267	18023.10	20658.79	5	17195650	0.77	0	0.63	62.99	1p
1	865738	18807897	0.69	0.29	17941597	317	355	35874.13	39664.99	4	17942159	1.00	0	0.85	90.34	1p
1	1225959	115258830	0.81	0.65	110878099	129	199	332297.89	544926.43	5	114032871	0.97	0	0.99	81.53	1p
1	12854021	35350573	0.69	0.38	18943245	190	210	54538.36	60735.73	6	22496552	0.84	0	0.67	60.00	1p
1	144220850	248814126	0.69	0.61	102235986	1078	1177	65572.98	69149.21	5	104593276	0.98	0	0.97	644.87	1q
3	118866376	197574936	0.66	0.38	78708561	515	553	142330.13	152832.16	1	78708560	1.00	0	0.97	195.70	3q
4	84230033	190903688	0.72	0.45	104876557	326	346	157849.02	167662.37	2	106673655	0.98	0	0.98	142.63	4q
4	106317429	190903688	0.76	0.41	84584272	296	321	136508.27	148078.71	2	84586259	1.00	0	0.99	121.36	4q
7	64023371	158672619	0.76	0.54	91867464	357	851	103204.51	149005.89	5	94649248	0.97	0	0.99	187.81	7q
7	73097654	158851234	0.82	0.43	83827430	511	572	135647.98	149913.77	3	85753580	0.98	0	1.00	214.79	7q
9	116800	5732483	0.74	0.32	5615684	36	40	140392.10	155991.22	1	5615683	1.00	0	0.66	11.52	9p
9	116800	42368628	0.73	0.41	42122921	202	215	176519.29	187985.18	2	42251828	1.00	0	0.98	82.57	9p
9	117877	33798073	0.81	0.38	33680197	144	156	215898.70	233890.26	1	33680196	1.00	0	1.00	54.72	9p
9	117934	33676094	0.77	0.45	33498291	127	134	116536.23	123717.30	3	33558160	1.00	0	0.95	57.47	9p
9	289557	21350904	0.73	0.32	21061348	109	115	183142.16	193223.38	1	21061347	1.00	0	0.95	34.88	9p
9	43875942	95887320	0.66	0.33	52011379	125	135	385269.47	416091.03	1	52011378	1.00	0.36	0.53	41.25	9q
9	71114312	80932574	0.68	0.26	9818263	98	101	97210.52	100186.36	1	9818262	1.00	0	0.67	25.48	9q
11	193096	48347498	0.75	0.45	45668557	527	577	28505.71	30815.50	6	48154402	0.95	0	0.96	225.74	11p
11	62933774	134244123	0.73	0.42	71310350	554	596	119648.24	128719.04	1	71310349	1.00	0	0.99	232.68	11q
12	119563325	133778796	0.71	0.33	14215472	198	215	66118.47	71795.31	1	14215471	1.00	0	0.89	65.34	12q
13	19751032	115047496	0.61	0.48	92699017	343	356	197277.78	203528.66	2	95296464	0.97	0	0.76	160.15	13q
13	28197436	52971893	0.66	0.36	24774458	153	167	148350.05	161924.56	1	24774457	1.00	0	0.77	55.08	13q
14	30066929	107049080	0.86	0.31	76982152	323	323	238334.84	238334.84	1	76982151	1.00	0	1.00	100.13	14q
14	31354296	107283160	0.72	0.83	75321232	543	543	44380.05	44380.05	4	75928864	0.99	0	0.72	448.43	14q
14	35872926	107283160	0.76	0.78	70983139	578	622	79082.19	82962.93	8	71410234	0.99	0	0.97	448.86	14q
14	91110582	107113968	0.77	0.79	15945703	279	279	26823.58	26823.58	3	16003386	1.00	0	0.61	218.69	14q
17	38634929	81043039	0.63	0.37	42408111	644	679	62456.72	65851.10	1	42408110	1.00	0	0.87	238.28	17q
19	41354606	56520150	0.73	0.36	15165545	23	33	459561.97	659371.52	1	15165544	1.00	0	0.79	8.28	19q
20	29623223	49191228	0.66	0.37	19568006	186	209	93626.82	105204.33	1	19568005	1.00	0	0.71	68.82	20q

Table B.2 **High stringency settings: AI regions identified in the UK biobank exemplar dataset.**

Table B.3 **Low stringency settings: AI regions identified in the UK biobank exemplar dataset.**

Table B.4 **gg score system applied to the UKB-WES50 labelled data**

Table B.5 **gg score system applied to the Schizo-WES02 data**

Link to view/download the tables https://doi.org/10.5258/SOTON/D2266

Bibliography

Adeyemo, A.A., Zaghloul, N.A., Chen, G., Doumatey, A.P., Leitch, C.C., Hostelley, T.L., Nesmith, J.E., Zhou, J., Bentley, A.R., Shriner, D., et al. (2019). ZRANB3 is an African-specific type 2 diabetes locus associated with beta-cell mass and insulin response. Nat. Commun.

Affymetrix (2011). Analysis Guide Axiom™ Genotyping Solution Data Analysis Guide. Analysis 55.

Afyounian, E., Annala, M., and Nykter, M. (2017). Segmentum: a tool for copy number analysis of cancer genomes. BMC Bioinformatics *18*, 215.

Ahmadi, A., Ghaedi, H., Salimian, J., Azimzadeh Jamalkandi, S., and Ghanei, M. (2019). Association between chronic obstructive pulmonary disease and interleukins gene variants: A systematic review and meta-analysis. Cytokine.

Ahola-Olli, A. V., Würtz, P., Havulinna, A.S., Aalto, K., Pitkänen, N., Lehtimäki, T., Kähönen, M., Lyytikäinen, L.P., Raitoharju, E., Seppälä, I., et al. (2017). Genome-wide Association Study Identifies 27 Loci Influencing Concentrations of Circulating Cytokines and Growth Factors. Am. J. Hum. Genet.

Altshuler, D.L., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Collins, F.S., De La Vega, F.M., Donnelly, P., Egholm, M., et al. (2010a). A map of human genome variation from population-scale sequencing. Nature.

Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., Peltonen, L., et al. (2010b). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58.

Altshuler, D.M., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., Flicek, P., Gabriel, S.B., et al. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature.

Anderson, C. a, Pettersson, F.F.H., Clarke, G.G.M., Cardon, L.R., Morris, A.P., and Zondervan, K.T. (2010). Data quality control in genetic case-control association studies. Nat. Protoc. *5*, 1564–1573.

Anderson, K., Lutz, C., Van Delft, F.W., Bateman, C.M., Guo, Y., Colman, S.M., Kempski, H., Moorman, A. V., Titley, I., Swansbury, J., et al. (2011). Genetic variegation of clonal architecture and propagating cells in leukaemia. Nature *469*, 356–361.

Andrew, C.R. (2007). Linkage Disequilibrium and Association Mapping. Methods Mol. Biol.

Apte, R.S., Chen, D.S., and Ferrara, N. (2019). VEGF in Signaling and Disease: Beyond Discovery and Development. Cell.

Arber, D.A., Orazi, A., Hasserjian, R., Thiele, J., Borowitz, M.J., Le Beau, M.M., Bloomfield, C.D., Cazzola, M., and Vardiman, J.W. (2016a). The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. Blood *127*, 2391–2405.

Arber, D.A., Orazi, A., Hasserjian, R., Thiele, J., Borowitz, M.J., Le Beau, M.M., Bloomfield, C.D., Cazzola, M., and Vardiman, J.W. (2016b). The 2016 revision to the World Health Organization (WHO) classification of lymphoid neoplasms.

Arnold, M., Raffler, J., Pfeufer, A., Suhre, K., and Kastenmüller, G. (2015). SNiPA: an interactive, genetic variant-centered annotation browser. Bioinformatics *31*, 1334–1336.

Arock, M., Sotlar, K., Akin, C., Broesby-Olsen, S., Hoermann, G., Escribano, L., Kristensen, T.K., Kluin-Nelemans, H.C., Hermine, O., Dubreuil, P., et al. (2015). KIT mutation analysis in mast cell neoplasms: Recommendations of the European Competence Network on Mastocytosis. Leukemia.

Astle, W.J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A.L., Mead, D., Bouman, H., Riveros-Mckay, F., Kostadima, M.A., et al. (2016). The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell.

Auer, P.L., and Lettre, G. (2015). Rare variant association studies: Considerations, challenges and opportunities. Genome Med.

Auton, A., Brooks, L.D., Durbin, R.M., Abecasis, G.R., Altshuler, D.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., et al. (2015). A global reference for human genetic variation. Nature *526*, 68.

Auwera, G.A. Van der, Carneiro, M.O., Chris Hartl, R.P., Angel, G. del, Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., et al. (2014). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinforma.

Van der Auwera, G.A., and O'Connor, B.D. (2020). Genomics in the Cloud: Using Docker, GATK, and WDL in Terra (O'Reilly Media, Inc.).

Avellino, R., and Delwel, R. (2017). Expression and regulation of C/EBPα in normal myelopoiesis and in malignant transformation. Blood.

Bae, J.M., Shin, S.H., Kwon, H.J., Park, S.Y., Kook, M.C., Kim, Y.W., Cho, N.Y., Kim, N., Kim, T.Y., Kim, D., et al. (2012). ALU and LINE-1 hypomethylations in multistep gastric carcinogenesis and their

prognostic implications. Int. J. Cancer.

Baird, J.H., and Gotlib, J. (2018). Clinical Validation of KIT Inhibition in Advanced Systemic Mastocytosis.

Baker, M. (2011). Making sense of chromatin states. Nat. Methods.

de Bakker, P.I.W., Yelensky, R., Pe'er, I., Gabriel, S.B., Daly, M.J., and Altshuler, D. (2005). Efficiency and power in genetic association studies. Nat. Genet. *37*, 1217–1223.

Bao, E.L., Nandakumar, S.K., Liao, X., Bick, A.G., Karjalainen, J., Tabaka, M., Gan, O.I., Havulinna, A.S., Kiiskinen, T.T.J., Lareau, C.A., et al. (2020). Inherited myeloproliferative neoplasm risk affects haematopoietic stem cells. Nature.

Baptista, R.L.R., dos Santos, A.C.E., Gutiyama, L.M., Solza, C., and Zalcberg, I.R. (2017). Familial Myelodysplastic/Acute Leukemia Syndromes—Myeloid Neoplasms with Germline Predisposition. Front. Oncol.

Baranwal, A., Bagwe, B.R., and M, V. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

Di Bernardo, M.C., Crowther-Swanepoel, D., Broderick, P., Webb, E., Sellick, G., Wild, R., Sullivan, K., Vijayakrishnan, J., Wang, Y., Pittman, A.M., et al. (2008). A genome-wide association study identifies six susceptibility loci for chronic lymphocytic leukemia. Nat. Genet.

Bick, A.G., Weinstock, J.S., Nandakumar, S.K., Fulco, C.P., Bao, E.L., Zekavat, S.M., Szeto, M.D., Liao, X., Leventhal, M.J., Nasser, J., et al. (2020). Inherited causes of clonal haematopoiesis in 97,691 whole genomes. Nature *3*.

Blanco, E., González-Ramírez, M., Alcaine-Colet, A., Aranda, S., and Di Croce, L. (2020). The Bivalent Genome: Characterization, Structure, and Regulation. Trends Genet.

Bolouri, H., Farrar, J.E., Triche, T., Ries, R.E., Lim, E.L., Alonzo, T.A., Ma, Y., Moore, R., Mungall, A.J., Marra, M.A., et al. (2018). The molecular landscape of pediatric acute myeloid leukemia reveals recurrent structural alterations and age-specific mutational interactions. Nat. Med. *24*, 103–112.

Bonasio, R., Tu, S., and Reinberg, D. (2010). Molecular signals of epigenetic states. Science (80-.).

Bonnet, D., and Dick, J.E. (1997). Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. Nat. Med. *3*, 730–737.

Bosch, X. (2004). Spain to establish national genetic database. Lancet *363*, 1044.

Boycott, K.M., Vanstone, M.R., Bulman, D.E., and MacKenzie, A.E. (2013). Rare-disease genetics in the era of next-generation sequencing: discovery to translation. Nat. Rev. Genet. *14*, 681–691.

Boyd, S., and Arber, D. (2011). CHAPTER 18 – Acute myeloid leukemias. In Blood and Bone Marrow Pathology, p.

Boyle, A.P., Hong, E.L., Hariharan, M., Cheng, Y., Schaub, M.A., Kasowski, M., Karczewski, K.J., Park, J., Hitz, B.C., Weng, S., et al. (2012). Annotation of functional variation in personal genomes using RegulomeDB. Genome Res.

Broesby-Olsen, S., Kristensen, T.K., Møller, M.B., Bindslev-Jensen, C., and Vestergaard, H. (2012). Adult-onset systemic mastocytosis in monozygotic twins with KIT D816V and JAK2 V617F mutations. J. Allergy Clin. Immunol. *130*, 806–808.

Buniello, A., Macarthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res.

van de Bunt, M., Manning Fox, J.E., Dai, X., Barrett, A., Grey, C., Li, L., Bennett, A.J., Johnson, P.R., Rajotte, R. V., Gaulton, K.J., et al. (2015). Transcript Expression Data from Human Islets Links Regulatory Signals from Genome-Wide Association Studies for Type 2 Diabetes and Glycemic Traits to Their Downstream Effectors. PLoS Genet.

Burton, P.R., Clayton, D.G., Cardon, L.R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D.P., McCarthy, M.I., Ouwehand, W.H., Samani, N.J., et al. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature.

Bush, W.S., and Moore, J.H. (2012). Chapter 11: Genome-Wide Association Studies. PLoS Comput. Biol. *8*.

Busque, L., Patel, J.P., Figueroa, M.E., Vasanthakumar, A., Provost, S., Hamilou, Z., Mollica, L., Li, J., Viale, A., Heguy, A., et al. (2012). Recurrent somatic TET2 mutations in normal elderly individuals with clonal hematopoiesis. Nat. Genet.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature.

Cai, N., Bigdeli, T.B., Kretzschmar, W., Lei, Y., Liang, J., Song, L., Hu, J., Li, Q., Jin, W., Hu, Z., et al. (2015). Sparse whole-genome sequencing identifies two loci for major depressive disorder. Nature.

Caldas, C. (2012). Cancer sequencing unravels clonal evolution. Nat. Biotechnol. *30*, 408–410.

Calkins, G.N., Boveri, T., Calkins, G.N., and Boveri, T. (1914). Zur Frage der Entstehung maligner Tumoren. Science (80-.). *40*, 64.

Campa, D., Gentiluomo, M., Obazee, O., Ballerini, A., Vodickova, L., Hegyi, P., Soucek, P., Brenner, H., Milanetto, A.C., Landi, S., et al. (2020). Genome-wide association study identifies an early onset pancreatic cancer risk locus. Int. J. Cancer.

Campregher, P.V., Halley, N. da S., Vieira, G.A., Fernandes, J.F., Velloso, E.D.R.P., Ali, S., Mughal, T., Miller, V., Mangueira, C.L.P., Odone, V., et al. (2017). Identification of a novel fusion TBL1XR1–PDGFRB in a patient with acute myeloid leukemia harboring the DEK–NUP214 fusion and clinical response to dasatinib. Leuk. Lymphoma.

Cano-Gamez, E., and Trynka, G. (2020). From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. Front. Genet.

Carithers, L.J., and Moore, H.M. (2015). The Genotype-Tissue Expression (GTEx) Project. Biopreserv. Biobank.

Carlevaro-Fita, J., Lanzós, A., Feuerbach, L., Hong, C., Mas-Ponte, D., Pedersen, J.S., Abascal, F., Amin, S.B., Bader, G.D., Barenboim, J., et al. (2020). Cancer LncRNA Census reveals evidence for deep functional conservation of long noncoding RNAs in tumorigenesis. Commun. Biol.

Carson, A.R., Smith, E.N., Matsui, H., Brækkan, S.K., Jepsen, K., Hansen, J.B., and Frazer, K.A. (2014). Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. BMC Nephrol.

Caughey, G.H. (2016). Mast cell proteases as pharmacological targets. Eur. J. Pharmacol.

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, 7.

Chang, T.Y., Dvorak, C.C., and Loh, M.L. (2014). Bedside to bench in juvenile myelomonocytic leukemia: Insights into leukemogenesis from a rare pediatric leukemia. Blood.

Chanock, S.J., Manolio, T., Boehnke, M., Boerwinkle, E., Hunter, D.J., Thomas, G., Hirschhorn, J.N., Abecasis, G., Altshuler, D., Bailey-Wilson, J.E., et al. (2007). Replicating genotype–phenotype associations. Nature *447*, 655–660.

Chase, A., Leung, W., Tapper, W., Jones, A. V, Knoops, L., Rasi, C., Forsberg, L.A., Guglielmelli, P.,

Zoi, K., Hall, V., et al. (2015). Profound parental bias associated with chromosome 14 acquired uniparental disomy indicates targeting of an imprinted locus. Leukemia *29*, 2069–2074.

Chase, A., Pellagatti, A., Singh, S., Score, J., Tapper, W.J., Lin, F., Hoade, Y., Bryant, C., Trim, N., Yip, B.H., et al. (2019). PRR14L mutations are associated with chromosome 22 acquired uniparental disomy, age-related clonal hematopoiesis and myeloid neoplasia. Leukemia.

Chaudhury, A., Komrokji, R.S., Al Ali, N.H., Zhang, L., Vafaii, P., and Lancet, J.E. (2015). Prognosis and Outcomes in MDS-MPN Unclassifiable: Single Institution Experience of a Rare Disorder. Blood *126*, 1698 LP – 1698.

Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X.C., Stern, D., Winkler, J., Lockhart, D.J., Morris, M.S., and Fodor, S.P.A. (1996). Accessing genetic information with high-density DNA arrays. Science (80-.).

Chen, D., and George, T.I. (2018). Mastocytosis. In Hematopathology: A Volume in the Series: Foundations in Diagnostic Pathology, p.

Chen, L., Ge, B., Casale, F.P., Vasquez, L., Kwan, T., Garrido-Martín, D., Watt, S., Yan, Y., Kundu, K., Ecker, S., et al. (2016). Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. Cell.

Chen, M.H., Raffield, L.M., Mousas, A., Sakaue, S., Huffman, J.E., Moscati, A., Trivedi, B., Jiang, T., Akbari, P., Vuckovic, D., et al. (2020). Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. Cell.

Chial, B.H., Write, P.D., Right, S., and Education, N. (2008). Proto-oncogenes to Oncogenes to Cancer.

Chio, A., Schymick, J.C., Restagno, G., Scholz, S.W., Lombardo, F., Lai, S.L., Mora, G., Fung, H.C., Britton, A., Arepalli, S., et al. (2009). A two-stage genome-wide association study of sporadic amyotrophic lateral sclerosis. Hum. Mol. Genet. *18*, 1524–1532.

Chubb, D., Weinhold, N., Broderick, P., Chen, B., Johnson, D.C., Försti, A., Vijayakrishnan, J., Migliorini, G., Dobbins, S.E., Holroyd, A., et al. (2013). Common variation at 3q26.2, 6p21.33, 17p11.2 and 22q13.1 influences multiple myeloma risk. Nat. Genet.

Church, D.M., Schneider, V.A., Graves, T., Auger, K., Cunningham, F., Bouk, N., Chen, H.C., Agarwala, R., McLaren, W.M., Ritchie, G.R.S., et al. (2011). Modernizing reference genome assemblies. PLoS Biol.

Churpek, J.E., Marquez, R., Neistadt, B., Claussen, K., Lee, M.K., Churpek, M.M., Huo, D., Weiner,

H., Bannerjee, M., Godley, L.A., et al. (2016). Inherited mutations in cancer susceptibility genes are common among survivors of breast cancer who develop therapy-related leukemia. Cancer.

Cingolani, P., Patel, V.M., Coon, M., Nguyen, T., Land, S.J., Ruden, D.M., and Lu, X. (2012). Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. Front. Genet.

Clarke, G.M., Anderson, C. a, Pettersson, F.H., Cardon, L.R., and Andrew, P. (2011). Basic statistical analysis in genetic case-control studies. Nat. Protoc. *6*, 121–133.

Clevers, H. (2011). The cancer stem cell: premises, promises and challenges. Nat. Med. *17*, 313–319.

Coltoff, A., and Mascarenhas, J. (2019). Relevant updates in systemic mastocytosis. Leuk. Res. *81*, 10–18.

Conde-Fernandes, I., Sampaio, R., Moreno, F., Palla-Garcia, J., Teixeira, M. dos A., Freitas, I., Neves, E., Jara-Acevedo, M., Escribano, L., and Lima, M. (2017). Systemic mastocytosis with KIT V560G mutation presenting as recurrent episodes of vascular collapse: response to disodium cromoglycate and disease outcome. Allergy, Asthma Clin. Immunol. *13*, 21.

Crowther-Swanepoel, D., Broderick, P., Di Bernardo, M.C., Dobbins, S.E., Torres, M., Mansouri, M., Ruiz-Ponte, C., Enjuanes, A., Rosenquist, R., Carracedo, A., et al. (2010). Common variants at 2q37.3, 8q24.21, 15q21.3 and 16q24.1 influence chronic lymphocytic leukemia risk. Nat. Genet.

Daley, T., Metcalfe, D.D., and Akin, C. (2001). Association of the Q576R polymorphism in the interleukin-4 receptor α chain with indolent mastocytosis limited to the skin. Blood.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics.

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., et al. (2021). Twelve years of SAMtools and BCFtools. Gigascience *10*.

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods. Nat. Genet.

Davydov, E. V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++.

PLoS Comput. Biol.

Dawoud, A.A.Z., Tapper, W.J., and Cross, N.C.P. (2020). Clonal myelopoiesis in the UK Biobank cohort: ASXL1 mutations are strongly associated with smoking. Leukemia.

DeAngelo, D., Reiter, A., and Radia, D. (2021). CT023 – PATHFINDER: Interim analysis of avapritinib (ava) in patients (pts) with advanced systemic mastocytosis (AdvSM). In Abstract #CT023. Presented at the 2021 American Association for Cancer Research Annual Meeting, April 11, 2021., p.

Debette, S., Visvikis-Siest, S., Chen, M.H., Ndiaye, N.C., Song, C., Destefano, A., Safa, R., Nezhad, M.A., Sawyer, D., Marteau, J.B., et al. (2011). Identification of cis-and trans-acting genetic variants explaining up to half the variation in circulating vascular endothelial growth factor levels. Circ. Res.

Deelen, P., Bonder, M., van der Velde, K., Westra, H.-J., Winder, E., Hendriksen, D., Franke, L., and Swertz, M.A. (2014). Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. BMC Res. Notes *7*, 901.

Depristo, M.A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J.R., Hartl, C., Philippakis, A.A., Del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet.

Devlin, B., and Risch, N. (1995). A Comparison of Linkage Disequilibrium Measures for Fine-Scale Mapping. Genomics.

Ding, L., Ley, T.J., Larson, D.E., Miller, C.A., Koboldt, D.C., Welch, J.S., Ritchey, J.K., Young, M.A., Lamprecht, T., McLellan, M.D., et al. (2012). Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature *481*, 506–510.

Dragu, D.L., Necula, L.G., Bleotu, C., Diaconu, C.C., and Chivu-Economescu, M. (2015). Therapies targeting cancer stem cells: Current trends and future challenges. World J. Stem Cells *7*, 1185–1201.

Duan, Y., Zhang, X., Yang, L., Dong, X., Zheng, Z., Cheng, Y., Chen, H., Lan, B., Li, D., Zhou, J., et al. (2019). Disruptor of telomeric silencing 1-like (DOT1L) is involved in breast cancer metastasis via transcriptional regulation of MALAT1 and ZEB2. J. Genet. Genomics.

Dudbridge, F., and Gusnanto, A. (2008). Estimation of significance thresholds for genomewide association scans. Genet. Epidemiol.

Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S.,

Harrow, J., Kaul, R., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature.

Earp, M.A., and Goode, E.L. (2017). Linkage Disequilibrium. In Encyclopedia of Cancer, p.

Emilsson, V., Ilkov, M., Lamb, J.R., Finkel, N., Gudmundsson, E.F., Pitts, R., Hoover, H., Gudmundsdottir, V., Horman, S.R., Aspelund, T., et al. (2018). Co-regulatory networks of human serum proteins link genetics to disease. Science (80-.).

Engel, E. (1980). A new genetic concept: Uniparental disomy and its potential effect, isodisomy. Am. J. Med. Genet.

Ernst, J., and Kellis, M. (2017). Chromatin-state discovery and genome annotation with ChromHMM. Nat. Protoc.

Ernst, T., Chase, A.J., Score, J., Hidalgo-Curtis, C.E., Bryant, C., Jones, A. V., Waghorn, K., Zoi, K., Ross, F.M., Reiter, A., et al. (2010). Inactivating mutations of the histone methyltransferase gene EZH2 in myeloid disorders. Nat. Genet.

Evangelou, E., Trikalinos, T.A., Salanti, G., and Ioannidis, J.P.A. (2006). Family-based versus unrelated case-control designs for genetic associations. PLoS Genet. *2*, 1147–1155.

Faiyaz-Ul-Haque, M., Al-Dayel, F., Tulba, A., Abalkhail, H., Alhussaini, H., Memon, M., Bazarbashi, S., Amin, T., Satti, M.B., Peltekova, I., et al. (2018). Spectrum of the KIT gene mutations in gastrointestinal stromal tumors in Arab patients. Asian Pacific J. Cancer Prev. *19*, 2905–2910.

Ferrari, R., Hernandez, D.G., Nalls, M.A., Rohrer, J.D., Ramasamy, A., Kwok, J.B.J., Dobson-Stone, C., Brooks William S., B.S., Schofield, P.R., Halliday, G.M., et al. (2014). Frontotemporal dementia and its subtypes: A genome-wide association study. Lancet Neurol. *13*, 686–699.

Ferri, F.J., Pudil, P., Hatef, M., and Kittler, J. (1994). Comparative study of techniques for large-scale feature selection. In Machine Intelligence and Pattern Recognition, p.

Ferrucci, L., Bandinelli, S., Benvenuti, E., Di Iorio, A., Macchi, C., Harris, T.B., and Guralnik, J.M. (2000). Subsystems contributing to the decline in ability to walk: Bridging the gap between epidemiology and geriatric practice in the InCHIANTI study. J. Am. Geriatr. Soc.

Flavahan, W.A., Gaskell, E., and Bernstein, B.E. (2017). Epigenetic plasticity and the hallmarks of cancer. Science (80-.).

Frenzel, L., and Hermine, O. (2013). Mast cells and inflammation. Jt. Bone Spine.

Fritz, A., Percy, C., Jack, A., Shanmugaratnam, K., Sobin, L., Parkin, D.M., and Whelan, S. (2013). International Classification of Diseases for Oncology.

Frost, M.J., Ferrao, P.T., Hughes, T.P., and Ashman, L.K. (2002). Juxtamembrane mutant V560GKit is more sensitive to Imatinib (STI571) compared with wild-type c-Kit whereas the kinase domain mutant D816VKit is resistant. Mol. Cancer Ther.

Fuchsberger, C., Flannick, J., Teslovich, T.M., Mahajan, A., Agarwala, V., Gaulton, K.J., Ma, C., Fontanillas, P., Moutsianas, L., McCarthy, D.J., et al. (2016). The genetic architecture of type 2 diabetes. Nature.

Gallagher, M.D., and Chen-Plotkin, A.S. (2018). The Post-GWAS Era: From Association to Function. Am. J. Hum. Genet.

Ganna, A., Genovese, G., Howrigan, D.P., Byrnes, A., Kurki, M.I., Zekavat, S.M., Whelan, C.W., Kals, M., Nivard, M.G., Bloemendal, A., et al. (2016). Ultra-rare disruptive and damaging mutations influence educational attainment in the general population. Nat. Neurosci.

Gao, H., Gong, N., Ma, Z., Miao, X., Chen, J., Cao, Y., and Zhang, G. (2018). LncRNA ZEB2-AS1 promotes pancreatic cancer cell growth and invasion through regulating the miR-204/HMGB1 axis. Int. J. Biol. Macromol.

Gao, J., Gentzler, R.D., Timms, A.E., Horwitz, M.S., Frankfurt, O., Altman, J.K., and Peterson, L.C. (2014). Heritable GATA2 mutations associated with familial AML-MDS: a case report and review of literature. J. Hematol. Oncol.

Gaulton, K.J., Nammo, T., Pasquali, L., Simon, J.M., Giresi, P.G., Fogarty, M.P., Panhuis, T.M., Mieczkowski, P., Secchi, A., Bosco, D., et al. (2010). A map of open chromatin in human pancreatic islets. Nat. Genet.

Gaunt, T.R., Shihab, H.A., Hemani, G., Min, J.L., Woodward, G., Lyttleton, O., Zheng, J., Duggirala, A., McArdle, W.L., Ho, K., et al. (2016a). Systematic identification of genetic influences on methylation across the human life course. Genome Biol.

Gaunt, T.R., Shihab, H.A., Hemani, G., Min, J.L., Woodward, G., Lyttleton, O., Zheng, J., Duggirala, A., McArdle, W.L., Ho, K., et al. (2016b). Systematic identification of genetic influences on methylation across the human life course. Genome Biol.

Gel, B., and Serra, E. (2017). KaryoploteR: An R/Bioconductor package to plot customizable genomes displaying arbitrary data. Bioinformatics.

Genovese, G., Kähler, A.K., Handsaker, R.E., Lindberg, J., Rose, S.A., Bakhoum, S.F., Chambert, K.,

Mick, E., Neale, B.M., Fromer, M., et al. (2014). Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. N. Engl. J. Med. *371*, 2477–2487.

Geyer, J.T. (2019). Myeloid Neoplasms with Germline Predisposition. Pathobiology.

Di Giacomo, D., La Starza, R., Gorello, P., Pellanera, F., Kalender Atak, Z., De Keersmaecker, K., Pierini, V., Harrison, C.J., Arniani, S., Moretti, M., et al. (2021). 14q32 rearrangements deregulating BCL11B mark a distinct subgroup of T and myeloid immature acute leukemia . Blood.

Gibson, G. (2012). Rare and common variants: Twenty arguments. Nat. Rev. Genet.

Gilreath, J.A., Tchertanov, L., and Deininger, M.W. (2019). Novel approaches to treating advanced systemic mastocytosis. Clin. Pharmacol. Adv. Appl.

Gnanasambandan, K., Magis, A., and Sayeski, P.P. (2010). The constitutive activation of Jak2-V617F is mediated by a π stacking mechanism involving Phenylalanines 595 and 617. Biochemistry.

Gogarten, S.M., Bhangale, T., Conomos, M.P., Laurie, C.A., McHugh, C.P., Painter, I., Zheng, X., Crosslin, D.R., Levine, D., Lumley, T., et al. (2012). GWASTools: An R/Bioconductor package for quality control and analysis of genome-wide association studies. Bioinformatics.

Goode, E.L. (2011). Linkage Disequilibrium. In Encyclopedia of Cancer, (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 2043–2048.

Goossens, S., Wang, J., Tremblay, C.S., De Medts, J., T'sas, S., Nguyen, T., Saw, J., Haigh, K., Curtis, D.J., Van Vlierberghe, P., et al. (2019). ZEB2 and LMO2 drive immature T-cell lymphoblastic leukemia via distinct oncogenic mechanisms. Haematologica.

Gotlib, J., Kluin-Nelemans, H.C., George, T.I., Akin, C., Sotlar, K., Hermine, O., Awan, F.T., Hexner, E., Mauro, M.J., Sternberg, D.W., et al. (2016). Efficacy and Safety of Midostaurin in Advanced Systemic Mastocytosis. N. Engl. J. Med.

Gourvest, M., Brousset, P., and Bousquet, M. (2019). Long noncoding RNAs in acute myeloid leukemia: Functional characterization and clinical relevance. Cancers (Basel).

Graae, A.S., Hollensted, M., Kloppenborg, J.T., Mahendran, Y., Schnurr, T.M., Appel, E.V.R., Rask, J., Nielsen, T.R.H., Johansen, M., Linneberg, A., et al. (2018). An adult-based insulin resistance genetic risk score associates with insulin resistance, metabolic traits and altered fat distribution in Danish children and adolescents who are overweight or obese. Diabetologia.

Grand, F.H., Hidalgo-Curtis, C.E., Ernst, T., Zoi, K., Zoi, C., McGuire, C., Kreil, S., Jones, A., Score, J., Metzgeroth, G., et al. (2009). Frequent CBL mutations associated with 11q acquired uniparental disomy in myeloproliferative neoplasms. Blood *113*, 6182–6192.

Gratten, J., and Visscher, P.M. (2016). Genetic pleiotropy in complex traits and diseases: Implications for genomic medicine. Genome Med.

Greaves, M. (2015). Evolutionary determinants of cancer. Cancer Discov.

Greaves, M., and Maley, C.C. (2012). Clonal evolution in cancer. Nature.

Greiner, G., Sprinzl, B., Górska, A., Ratzinger, F., Gurbisz, M., Witzeneder, N., Schmetterer, K.G., Gisslinger, B., Uyanik, G., Hadzijusufovic, E., et al. (2021). Hereditary α tryptasemia is a valid genetic biomarker for severe mediator-related symptoms in mastocytosis. Blood.

Griffith, M., Spies, N.C., Krysiak, K., McMichael, J.F., Coffman, A.C., Danos, A.M., Ainscough, B.J., Ramirez, C.A., Rieke, D.T., Kujan, L., et al. (2017). CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nat. Genet.

Guan, J., Liu, P., Wang, A., and Wang, B. (2020). Long non-coding RNA ZEB2-AS1 affects cell proliferation and apoptosis via the miR-122-5p/PLK1 axis in acute myeloid leukemia. Int. J. Mol. Med.

Guo, M.H., Nandakumar, S.K., Ulirsch, J.C., Zekavat, S.M., Buenrostro, J.D., Natarajan, P., Salem, R.M., Chiarle, R., Mitt, M., Kals, M., et al. (2017). Comprehensive population-based genome sequencing provides insight into hematopoietic regulatory mechanisms. Proc. Natl. Acad. Sci. U. S. A.

Guo, Y., Hu, Y., Hu, M., He, J., and Li, B. (2018). Long non-coding RNA ZEB2-AS1 promotes proliferation and inhibits apoptosis in human lung cancer cells. Oncol. Lett.

Haferlach, T., Nagata, Y., Grossmann, V., Okuno, Y., Bacher, U., Nagae, G., Schnittger, S., Sanada, M., Kon, A., Alpermann, T., et al. (2014). Landscape of genetic lesions in 944 patients with myelodysplastic syndromes. Leukemia *28*, 241–247.

Hahn, C.N., Chong, C.-E., Carmichael, C.L., Wilkins, E.J., Brautigan, P.J., Li, X.-C., Babic, M., Lin, M., Carmagnac, A., Lee, Y.K., et al. (2011). Heritable GATA2 mutations associated with familial myelodysplastic syndrome and acute myeloid leukemia. Nat. Genet.

Hammerschlag, A.R., Stringer, S., De Leeuw, C.A., Sniekers, S., Taskesen, E., Watanabe, K., Blanken, T.F., Dekker, K., Te Lindert, B.H.W., Wassing, R., et al. (2017). Genome-wide association analysis of insomnia complaints identifies risk genes and genetic overlap with psychiatric and

metabolic traits. Nat. Genet.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: The next generation. Cell *144*, 646–674.

Hansemann, D. (1890). Ueber asymmetrische Zelltheilung in Epithelkrebsen und deren biologische Bedeutung. Virchows Arch. Pathol. Anat. *119*, 299–326.

Harris, J.F., Chambers, a F., Hill, R.P., and Ling, V. (1982). Metastatic variants are generated spontaneously at a high rate in mouse KHT tumor. Proc. Natl. Acad. Sci. U. S. A.

Hartmann, K., Wardelmann, E., Ma, Y., Merkelbach-Bruse, S., Preussner, L.M., Woolery, C., Baldus, S.E., Heinicke, T., Thiele, J., Buettner, R., et al. (2005). Novel germline mutation of KIT associated with familial gastrointestinal stromal tumors and mastocytosis. Gastroenterology *129*, 1042–1046.

Hasan, S.K., Mays, A.N., Ottone, T., Ledda, A., Nasa, G. La, Cattaneo, C., Borlenghi, E., Melillo, L., Montefusco, E., Cervera, J., et al. (2008). Molecular analysis of t(15;17) genomic breakpoints in secondary acute promyelocytic leukemia arising after treatment of multiple sclerosis. Blood.

He, C., Holme, J., and Anthony, J. (2014). SNP genotyping: The KASP assay. Methods Mol. Biol.

Heim, S., and Mitelman, F. (2015). Nonrandom chromosome abnormalities in cancer: An overview. In Cancer Cytogenetics: Fourth Edition, p.

Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat. Genet.

Hidalgo-Curtis, C., Apperley, J.F., Stark, A., Jeng, M., Gotlib, J., Chase, A., Cross, N.C.P., and Grand, F.H. (2010). Fusion of PDGFRB to two distinct loci at 3p21 and a third at 12q13 in imatinib-responsive myeloproliferative neoplasms. Br. J. Haematol.

Higgins, J.P.T., Thompson, S.G., Deeks, J.J., and Altman, D.G. (2003). Measuring inconsistency in meta-analyses. BMJ Br. Med. J. *327*, 557–560.

Hinds, D.A., Barnholt, K.E., Mesa, R.A., Kiefer, A.K., Do, C.B., Eriksson, N., Mountain, J.L., Francke, U., Tung, J.Y., Nguyen, H., et al. (2016). Germ line variants predispose to both JAK2 V617F clonal hematopoiesis and myeloproliferative neoplasms. Blood.

Hirota, S., Isozaki, K., Moriyama, Y., Hashimoto, K., Nishida, T., Ishiguro, S., Kawano, K., Hanada, M., Kurata, A., Takeda, M., et al. (1998). Gain-of-function mutations of c-kit in human

gastrointestinal stromal tumors. Science (80-.).

Hnisz, D., Abraham, B., Lee, T., Lau, A., Saint-Andre, V., Sigova, A., Hoke, H., and Young, R. (2014). Transcriptional super-enhancers connected to cell identity and disease. Cell.

Holle, R., Happich, M., Löwel, H., Wichmann, H.E., and MONICA/KORA Study Group (2005). KORA--a research platform for population based health research. Gesundheitswes. (Bundesverband Der Ärzte Des Öffentlichen Gesundheitsdienstes *67 Suppl 1*, S19-25.

Holstege, H., Pfeiffer, W., Sie, D., Hulsman, M., Nicholas, T.J., Lee, C.C., Ross, T., Lin, J., Miller, M.A., Ylstra, B., et al. (2014). Somatic mutations found in the healthy blood compartment of a 115-yr-old woman demonstrate oligoclonal hematopoiesis. Genome Res.

Hong, E.P., and Park, J.W. (2012). Sample size and statistical power calculation in genetic association studies. Genomics Inform. *10*, 117–122.

Horn, S., Figl, A., Rachakonda, P.S., Fischer, C., Sucker, A., Gast, A., Kadel, S., Moll, I., Nagore, E., Hemminki, K., et al. (2013). TERT promoter mutations in familial and sporadic melanoma. Science (80-.).

Van Hout, C. V., Tachmazidou, I., Backman, J.D., Hoffman, J.D., Liu, D., Pandey, A.K., Gonzaga-Jauregui, C., Khalid, S., Ye, B., Banerjee, N., et al. (2020). Exome sequencing and characterization of 49,960 individuals in the UK Biobank. Nature.

Howie, B.N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet.

Huang, F.W., Hodis, E., Xu, M.J., Kryukov, G. V., Chin, L., and Garraway, L.A. (2013). Highly recurrent TERT promoter mutations in human melanoma. Science (80-.).

Huarte, M. (2015). The emerging role of lncRNAs in cancer. Nat. Med.

Hung, R.J., Spitz, M.R., Houlston, R.S., Schwartz, A.G., Field, J.K., Ying, J., Li, Y., Han, Y., Ji, X., Chen, W., et al. (2019). Lung Cancer Risk in Never-Smokers of European Descent is Associated With Genetic Variation in the 5p15.33 TERT-CLPTM1Ll Region. J. Thorac. Oncol.

Hunter, D.J., Kraft, P., Jacobs, K.B., Cox, D.G., Yeager, M., Hankinson, S.E., Wacholder, S., Wang, Z., Welch, R., Hutchinson, A., et al. (2007). A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat. Genet.

Hur, K., Cejas, P., Feliu, J., Moreno-Rubio, J., Burgos, E., Boland, C.R., and Goel, A. (2014). Hypomethylation of long interspersed nuclear element-1 (LINE-1) leads to activation of protooncogenes in human colorectal cancer metastasis. Gut.

Husemoen, L.L.N., Thomsen, T.F., Fenger, M., Jørgensen, H.L., and Jørgensen, T. (2003). Contribution of thermolabile methylenetetrahydrofolate reductase variant to total plasma homocysteine levels in healthy men and women. Inter99 (2). Genet. Epidemiol.

Iglesias, A.I., van der Lee, S.J., Bonnemaijer, P.W.M., Höhn, R., Nag, A., Gharahkhani, P., Khawaja, A.P., Broer, L., Foster, P.J., Hammond, C.J., et al. (2017). Haplotype reference consortium panel: Practical implications of imputations with large reference panels. Hum. Mutat.

Igolkina, A.A., Zinkevich, A., Karandasheva, K.O., Popov, A.A., Selifanova, M. V., Nikolaeva, D., Tkachev, V., Penzar, D., Nikitin, D.M., and Buzdin, A. (2019). H3K4me3, H3K9ac, H3K27ac, H3K27me3 and H3K9me3 Histone Tags Suggest Distinct Regulatory Evolution of Open and Condensed Chromatin Landmarks. Cells.

Illumina (2010). Interpreting Infinium Assay Data for Whole-Genome Structural Variation. Analysis 0–9.

Irvine, A.D., and McLean, W.H.I. (2006). Breaking the (un)sound barrier: Filaggrin is a major gene for atopic dermatitis. J. Invest. Dermatol.

Ishikawa, Y., Kawashima, N., Atsuta, Y., Sugiura, I., Sawa, M., Dobashi, N., Yokoyama, H., Doki, N., Tomita, A., Kiguchi, T., et al. (2020). Prospective evaluation of prognostic impact of KIT mutations on acute myeloid leukemia with RUNX1-RUNX1T1 and CBFB-MYH11. Blood Adv.

Jacobs, K.B., Yeager, M., Zhou, W., Wacholder, S., Wang, Z., Rodriguez-Santiago, B., Hutchinson, A., Deng, X., Liu, C., Horner, M.J., et al. (2012). Detectable clonal mosaicism and its relationship to aging and cancer. Nat. Genet.

Jagadeesan, M., Khetan, V., and Mallipatna, A. (2016). Genetic perspective of retinoblastoma: From present to future. Indian J. Ophthalmol.

Jaiswal, S., and Ebert, B.L. (2019). Clonal hematopoiesis in human aging and disease. Science (80-.).

Jaiswal, S., Fontanillas, P., Flannick, J., Manning, A., Grauman, P. V., Mar, B.G., Lindsley, R.C., Mermel, C.H., Burtt, N., Chavez, A., et al. (2014). Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. N. Engl. J. Med. *371*, 2488–2498.

Jaiswal, S., Natarajan, P., Silver, A.J., Gibson, C.J., Bick, A.G., Shvartz, E., McConkey, M., Gupta, N., Gabriel, S., Ardissino, D., et al. (2017). Clonal Hematopoiesis and risk of atherosclerotic cardiovascular disease. N. Engl. J. Med.

James, G., Witten, D., Hastie, T., and Tibishirani, R. (2013). An Introduction to Statistical Learning with Applications in R.

Javierre, B.M., Sewitz, S., Cairns, J., Wingett, S.W., Várnai, C., Thiecke, M.J., Freire-Pritchett, P., Spivakov, M., Fraser, P., Burren, O.S., et al. (2016). Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell.

Jawhar, M., Schwaab, J., Schnittger, S., Sotlar, K., Horny, H.P., Metzgeroth, G., Müller, N., Schneider, S., Naumann, N., Walz, C., et al. (2015). Molecular profiling of myeloid progenitor cells in multi-mutated advanced systemic mastocytosis identifies KIT D816V as a distinct and late event. Leukemia.

Jayakumar, R., and Xie, S. (2018). 165 A Rare Case of Systemic Mastocytosis With Associated Clonal Hematological Non-Mast Cell Lineage Disease. Am. J. Clin. Pathol.

Jeltsch, M., Kaipainen, A., Joukov, V., Meng, X., Lakso, M., Rauvala, H., Swartz, M., Fukumura, D., Jain, R.K., and Alitalo, K. (1997). Hyperplasia of lymphatic vessels in VEGF-C transgenic mice. Science (80-.).

Johansson, B., and Harrison, C.J. (2015). Acute myeloid leukemia. In Cancer Cytogenetics: Fourth Edition, p.

Jones, S.R. (2003). An introduction to power and sample size estimation. Emerg. Med. J. *20*, 453–458.

Jones, A. V., Chase, A., Silver, R.T., Oscier, D., Zoi, K., Wang, Y.L., Cario, H., Pahl, H.L., Collins, A., Reiter, A., et al. (2009). JAK2 haplotype is a major risk factor for the development of myeloproliferative neoplasms. Nat. Genet.

Jones, A. V., Campbell, P.J., Beer, P.A., Schnittger, S., Vannucchi, A.M., Zoi, K., Percy, M.J., McMullin, M.F., Scott, L.M., Tapper, W., et al. (2010). The JAK2 46/1 haplotype predisposes to MPL-mutated myeloproliferative neoplasms. Blood.

Jørgensen, T., Borch-John sen, K., Thomsen, T.F., Ibsen, H., Glümer, C., and Pisinger, C. (2003). A randomized non-pharmacological intervention study for prevention of ischaemic heart disease: Baseline results Inter99 (1). Eur. J. Prev. Cardiol.

Julià, A., Domènech, E., Ricart, E., Tortosa, R., García-Sánchez, V., Gisbert, J.P., Nos Mateu, P., Gutiérrez, A., Gomollón, F., Mendoza, J.L., et al. (2013). A genome-wide association study on a southern European population identifies a new Crohn's disease susceptibility locus at RBX1-EP300. Gut *62*, 1440 LP – 1445.

Kanai, M., Akiyama, M., Takahashi, A., Matoba, N., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K., et al. (2018). Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. Nat. Genet.

Kang, Z.J., Liu, Y.F., Xu, L.Z., Long, Z.J., Huang, D., Yang, Y., Liu, B., Feng, J.X., Pan, Y.J., Yan, J.S., et al. (2016). The philadelphia chromosome in leukemogenesis. Chin. J. Cancer *35*.

Keene, O.N. (1995). The log transformation is special. Stat. Med.

Khera, A. V., Chaffin, M., Aragam, K.G., Haas, M.E., Roselli, C., Choi, S.H., Natarajan, P., Lander, E.S., Lubitz, S.A., Ellinor, P.T., et al. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat. Genet.

Kheradpour, P., and Kellis, M. (2014). Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. Nucleic Acids Res.

Khoury, M.J., and Yang, Q. (1998). The future of genetic studies of complex human diseases. Epidemiology.

Khoury, P., and Lyons, J.J. (2019). Mast cell activation in the context of elevated basal serum tryptase: genetics and presentations. Curr. Allergy Asthma Rep.

Kilpeläinen, T.O., Bentley, A.R., Noordam, R., Sung, Y.J., Schwander, K., Winkler, T.W., Jakupović, H., Chasman, D.I., Manning, A., Ntalla, I., et al. (2019). Multi-ancestry study of blood lipid levels identifies four loci interacting with physical activity. Nat. Commun.

Kilpivaara, O., Mukherjee, S., Schram, A.M., Wadleigh, M., Mullally, A., Ebert, B.L., Bass, A., Marubayashi, S., Heguy, A., Garcia-Manero, G., et al. (2009). A germline JAK2 SNP is associated with predisposition to the development of JAK2(V617F)-positive myeloproliferative neoplasms. Nat. Genet.

Kim, S.Y., Im, K., Park, S.N., Kwon, J., Kim, J.A., and Lee, D.S. (2015). CALR, JAK2, and MPL mutation profiles in patients with four different subtypes of myeloproliferative neoplasms: Primary myelofibrosis, essential thrombocythemia, polycythemia vera, and myeloproliferative neoplasm, unclassifiable. Am. J. Clin. Pathol. *143*, 635–644.

Kircher, M., Witten, D.M., Jain, P., O'roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet.

Kitamura, Y., Matsuda, H., and Hatanaka, K. (1979). Clonal nature of mast-cell clusters formed in W/Wv mice after bone marrow transplantation. Nature.

Kitamura, Y., Oboki, K., and Ito, A. (2007). Development of mast cells. Proc. Jpn. Acad. Ser. B. Phys. Biol. Sci.

Klampfl, T., Gisslinger, H., Harutyunyan, A.S., Nivarthi, H., Rumi, E., Milosevic, J.D., Them, N.C.C., Berg, T., Gisslinger, B., Pietra, D., et al. (2013). Somatic Mutations of Calreticulin in Myeloproliferative Neoplasms. N. Engl. J. Med.

Knudson, a G. (2001). Two genetic hits (more or less) to cancer. Nat. Rev. Cancer *1*, 157–162.

Koboldt, D.C., Ding, L., Mardis, E.R., and Wilson, R.K. (2010). Challenges of sequencing human genomes. Brief. Bioinform.

KORA_A rsID Conversion File, I.S. KORA A: Illumina Human Omni 2.5, Loci Name to rsID Conversion File.

KORA_B rsID Conversion File, I.S. KORA B: Illumina Omni Express, Loci Name to rsID Conversion File.

Korn, C., and Méndez-Ferrer, S. (2017). Myeloid malignancies and the microenvironment. Blood *129*, 811–822.

Kouri, N., Ross, O.A., Dombroski, B., Younkin, C.S., Serie, D.J., Soto-Ortolaza, A., Baker, M., Finch, N.C.A., Yoon, H., Kim, J., et al. (2015). Genome-wide association study of corticobasal degeneration identifies risk variants shared with progressive supranuclear palsy. Nat. Commun. *6*, 7247.

Kraft, P., Zeggini, E., and Ioannidis, J.P.A. (2009). Replication in genome-wide association studies. Stat. Sci.

Kralovics, R., Guan, Y., and Prchal, J.T. (2002). Acquired uniparental disomy of chromosome 9p is a frequent stem cell defect in polycythemia vera. Exp. Hematol. *30*, 229–236.

Kristensen, T., Vestergaard, H., and Møller, M.B. (2011). Improved detection of the KIT D816V mutation in patients with systemic mastocytosis using a quantitative and highly sensitive real-time qPCR assay. J. Mol. Diagnostics *13*, 180–188.

Ku, C.S., Loy, E.Y., Pawitan, Y., and Chia, K.S. (2010). The pursuit of genome-wide association studies: where are we now? J. Hum. Genet. *55*, 195–206.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.

L'Abbate, A., Tolomeo, D., Cifola, I., Severgnini, M., Turchiano, A., Augello, B., Squeo, G., D'Addabbo, P., Traversa, D., Daniele, G., et al. (2018). MYC-containing amplicons in acute myeloid leukemia: genomic structures, evolution, and transcriptional consequences. Leukemia.

LaFramboise, T. (2009). Single nucleotide polymorphism arrays: A decade of biological, computational and technological advances. Nucleic Acids Res. *37*, 4181–4193.

Laine, E., de Beauchêne, I.C., Perahia, D., Auclair, C., and Tchertanov, L. (2011). Mutation D816V alters the internal structure and dynamics of C-Kit receptor cytoplasmic region: Implications for dimerization and activation mechanisms. PLoS Comput. Biol.

Lambert, S.A., Abraham, G., and Inouye, M. (2019). Towards clinical utility of polygenic risk scores. Hum. Mol. Genet.

Lancho, O., and Herranz, D. (2018). The MYC Enhancer-ome: Long-Range Transcriptional Regulation of MYC in Cancer. Trends in Cancer.

Landau, D.A., Carter, S.L., Getz, G., and Wu, C.J. (2014). Clonal evolution in hematological malignancies and therapeutic implications. Leukemia *28*, 34–43.

Landrum, M.J., Chitipiralla, S., Brown, G.R., Chen, C., Gu, B., Hart, J., Hoffman, D., Jang, W., Kaur, K., Liu, C., et al. (2020). ClinVar: Improvements to accessing data. Nucleic Acids Res.

Lange, M., Gleń, J., Zabłotna, M., Nedoszytko, B., Sokołowska-Wojdyło, M., Rębała, K., Ługowska-Umer, H., Niedoszytko, M., Górska, A., Sikorska, M., et al. (2017). Interleukin-31 polymorphisms and serum IL-31 level in patients with mastocytosis: Correlation with clinical presentation and pruritus. Acta Derm. Venereol.

Langemeijer, S.M.C., Kuiper, R.P., Berends, M., Knops, R., Aslanyan, M.G., Massop, M., Stevens-Linders, E., Van Hoogen, P., Van Kessel, A.G., Raymakers, R.A.P., et al. (2009). Acquired mutations in TET2 are common in myelodysplastic syndromes. Nat. Genet.

Laurie, C.C., Laurie, C.A., Rice, K., Doheny, K.F., Zelnick, L.R., McHugh, C.P., Ling, H., Hetrick, K.N., Pugh, E.W., Amos, C., et al. (2012). Detectable clonal mosaicism from birth to old age and its relationship to cancer. Nat. Genet. *44*, 642–650.

Ledergerber, C., and Dessimoz, C. (2011). Base-calling for next-generation sequencing platforms. Brief. Bioinform.

Lee, W.C., Reuben, A., Hu, X., McGranahan, N., Chen, R., Jalali, A., Negrao, M. V., Hubert, S.M., Tang, C., Wu, C.C., et al. (2020). Multiomics profiling of primary lung cancers and distant

metastases reveals immunosuppression as a common characteristic of tumor cells with metastatic plasticity. Genome Biol.

de Leeuw, C.A., Mooij, J.M., Heskes, T., and Posthuma, D. (2015). MAGMA: Generalized Gene-Set Analysis of GWAS Data. PLoS Comput. Biol.

Lettre, G., Sankaran, V.G., Bezerra, M.A.C., Araújo, A.S., Uda, M., Sanna, S., Cao, A., Schlessinger, D., Costa, F.F., Hirschhorn, J.N., et al. (2008). DNA polymorphisms at the BCL11A, HBS1L-MYB, and β-globin loci associate with fetal hemoglobin levels and pain crises in sickle cell disease. Proc. Natl. Acad. Sci. U. S. A.

Lewontin, R.C. (1964). The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. Genetics.

Li, H. (2011). Tabix: Fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics.

Li, B., Carey, M., and Workman, J.L. (2007). The Role of Chromatin during Transcription. Cell.

Li, W., Qi, Y., Cui, X., Huo, Q., Zhu, L., Zhang, A., Tan, M., Hong, Q., Yang, Y., Zhang, H., et al. (2018). Characteristic of HPV Integration in the Genome and Transcriptome of Cervical Cancer Tissues. Biomed Res. Int.

Lieb, R. (2013). Population-based study. In Encyclopedia of Behavioral Medicine, pp. 1507–1508.

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M.F., Parker, B.J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., et al. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. Nature.

Liu, Q.Y., Kong, L.F., Xu, Z.G., Li, Z., and Xue, H.Z. (2020). Mutation of the KIT Gene, excluding Exon 11, in Gastrointestinal Stromal Tumors. Biomed. Environ. Sci. *33*, 369–373.

Liu, Y., Cheng, Z., Pang, Y., Cui, L., Qian, T., Quan, L., Zhao, H., Shi, J., Ke, X., and Fu, L. (2019). Role of microRNAs, circRNAs and long noncoding RNAs in acute myeloid leukemia. J. Hematol. Oncol.

Lloyd-Jones, L.R., Holloway, A., McRae, A., Yang, J., Small, K., Zhao, J., Zeng, B., Bakshi, A., Metspalu, A., Dermitzakis, M., et al. (2017). The Genetic Architecture of Gene Expression in Peripheral Blood. Am. J. Hum. Genet.

Lyons, J.J., Yu, X., Hughes, J.D., Le, Q.T., Jamil, A., Bai, Y., Ho, N., Zhao, M., Liu, Y., O'Connell, M.P.,

et al. (2016). Elevated basal serum tryptase identifies a multisystem disorder associated with increased TPSAB1 copy number. Nat. Genet.

Lyons, J.J., Stotz, S.C., Chovanec, J., Liu, Y., Lewis, K.L., Nelson, C., DiMaggio, T., Jones, N., Stone, K.D., Sung, H., et al. (2018). A common haplotype containing functional CACNA1H variants is frequently coinherited with increased TPSAB1 copy number. Genet. Med.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. *45*, D896–D901.

Mailman, M.D., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., Hao, L., Kiang, A., Paschall, J., Phan, L., et al. (2007). The NCBI dbGaP database of genotypes and phenotypes. Nat. Genet.

Makishima, H., and Maciejewski, J.P. (2011). Pathogenesis and consequences of uniparental disomy in cancer. Clin. Cancer Res. *17*, 3913–3923.

Mallo, M., del Rey, M., Ibáñez, M., Calasanz, M.J., Arenillas, L., Larráyoz, M.J., Pedro, C., Jerez, A., Maciejewski, J., Costa, D., et al. (2013). Response to lenalidomide in myelodysplastic syndromes with del(5q): Influence of cytogenetics and mutations. Br. J. Haematol. *162*, 74–86.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.

Manthri, S., Costello, P.N., and Krishnan, K. (2020). Chronic mast cell leukaemia with exon 9 KIT mutation A502_Y503dup: a rare imatinib responsive variant. BMJ Case Rep.

Marcella, S., Petraroli, A., Braile, M., Parente, R., Ferrara, A.L., Galdiero, M.R., Modestino, L., Cristinziano, L., Rossi, F.W., Varricchi, G., et al. (2021). Vascular endothelial growth factors and angiopoietins as new players in mastocytosis. Clin. Exp. Med.

Marchese, F.P., Raimondi, I., and Huarte, M. (2017). The multidimensional mechanisms of long noncoding RNA function. Genome Biol.

Marees, A.T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., and Derks, E.M. (2018). A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. Int. J. Methods Psychiatr. Res.

Marioni, R.E., Harris, S.E., Zhang, Q., McRae, A.F., Hagenaars, S.P., Hill, W.D., Davies, G., Ritchie,

C.W., Gale, C.R., Starr, J.M., et al. (2018). GWAS on family history of Alzheimer's disease. Transl. Psychiatry.

Martincorena, I., Raine, K.M., Gerstung, M., Dawson, K.J., Haase, K., Van Loo, P., Davies, H., Stratton, M.R., and Campbell, P.J. (2017). Universal Patterns of Selection in Cancer and Somatic Tissues. Cell.

Martyn, G.E., Wienert, B., Yang, L., Shah, M., Norton, L.J., Burdach, J., Kurita, R., Nakamura, Y., Pearson, R.C.M., Funnell, A.P.W., et al. (2018). Natural regulatory mutations elevate the fetal globin gene via disruption of BCL11A or ZBTB7A binding. Nat. Genet.

Massé, A., Vainchenker, W., Dupont, S., Alberdi, A., Delhommeau, F., Fontenay, M., Robert, F., Lécluse, Y., Plo, I., Viguié, F., et al. (2009). Mutation in TET2 in Myeloid Cancers . N. Engl. J. Med.

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nat. Genet.

McClay, J.L., Shabalin, A.A., Dozmorov, M.G., Adkins, D.E., Kumar, G., Nerella, S., Clark, S.L., Bergen, S.E., Hultman, C.M., Magnusson, P.K.E., et al. (2015). High density methylation QTL analysis in human blood via next-generation sequencing of the methylated genomic DNA fraction. Genome Biol.

McGranahan, N., and Swanton, C. (2017). Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. Cell.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res.

Meggendorfer, M., Jeromin, S., Haferlach, C., Kern, W., and Haferlach, T. (2018). The mutational landscape of 18 investigated genes clearly separates four subtypes of myelodysplastic/myeloproliferative neoplasms. Haematologica.

de Melo Campos, P., Machado-Neto, J.A., Scopim-Ribeiro, R., Visconte, V., Tabarroki, A., Duarte, A.S.S., Barra, F.F.C., Vassalo, J., Rogers, H.J., Lorand-Metze, I., et al. (2014). Familial systemic mastocytosis with germline KIT K509I mutation is sensitive to treatment with imatinib, dasatinib and PKC412. Leuk. Res. *38*, 1245–1251.

Michaud, J., Wu, F., Osato, M., Cottles, G.M., Yanagida, M., Asou, N., Shigesada, K., Ito, Y., Benson, K.F., Raskind, W.H., et al. (2002). In vitro analyses of known and novel RUNX1/AML1 mutations in dominant familial platelet disorder with predisposition to acute myelogenous leukemia:

Implications for mechanisms of pathogenesis. Blood.

Mital, A., Piskorz, A., Lewandowski, K., Wasag, B., Limon, J., and Hellmann, A. (2011). A case of mast cell leukaemia with exon 9 KIT mutation and good response to imatinib. Eur. J. Haematol.

Mobuchon, L., Battistella, A., Bardel, C., Scelo, G., Renoud, A., Houy, A., Cassoux, N., Milder, M., Cancel-Tassin, G., Cussenot, O., et al. (2017). A GWAS in uveal melanoma identifies risk polymorphisms in the CLPTM1L locus. Npj Genomic Med. *2*, 5.

Mohamedali, A.M., Smith, A.E., Gaken, J., Lea, N.C., Mian, S.A., Westwood, N.B., Strupp, C., Gattermann, N., Germing, U., and Mufti, G.J. (2009). Novel TET2 mutations associated with UPD4q24 in myelodysplastic syndrome. J. Clin. Oncol. *27*, 4002–4006.

Molderings, G.J., Haenisch, B., Bogdanow, M., Fimmers, R., and Nöthen, M.M. (2013). Familial Occurrence of Systemic Mast Cell Activation Disease. PLoS One *8*.

Mullighan, C.G., Phillips, L.A., Su, X., Ma, J., Miller, C.B., Shurtleff, S.A., and Downing, J.R. (2008). Genomic Analysis of the Clonal Origins of Relapsed Acute Lymphoblastic Leukemia. Science (80-.).

Murakami, N., Okuno, Y., Yoshida, K., Shiraishi, Y., Nagae, G., Suzuki, K., Narita, A., Sakaguchi, H., Kawashima, N., Wang, X., et al. (2018). Integrated molecular profiling of juvenile myelomonocytic leukemia. Blood.

Nadaf, J., Majewski, J., and Fahiminiya, S. (2015). ExomeAI: Detection of recurrent Allelic Imbalance in tumors using whole Exome sequencing data. Bioinformatics *31*, 429–431.

Naeim, F., Nagesh Rao, P., Song, S.X., and Phan, R.T. (2018). Acute Myeloid Leukemias With Recurrent Genetic Abnormalities. In Atlas of Hematopathology, p.

Nalls, M. a, Pankratz, N., Lill, C.M., Do, C.B., Hernandez, D.G., Saad, M., DeStefano, A.L., Kara, E., Bras, J., Sharma, M., et al. (2014). Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. Nat. Genet. *056*, 1–7.

Nandakumar, S.K., Liao, X., and Sankaran, V.G. (2020). In The Blood: Connecting Variant to Function In Human Hematopoiesis. Trends Genet.

Nedoszytko, B., Niedoszytko, M., Lange, M., Van Doormaal, J., Gleń, J., Zabłotna, M., Renke, J., Vales, A., Buljubasic, F., Jassem, E., et al. (2009). Interleukin-13 promoter gene polymorphism -1112C/T is associated with the systemic form of mastocytosis. Allergy Eur. J. Allergy Clin. Immunol.

Nedoszytko, B., Lange, M., Renke, J., Niedoszytko, M., Zabłotna, M., Gleń, J., and Nowicki, R. (2018). The Possible Role of Gene Variant Coding Nonfunctional Toll-Like Receptor 2 in the Pathogenesis of Mastocytosis. Int. Arch. Allergy Immunol.

Nedoszytko, B., Sobalska-Kwapis, M., Strapagiel, D., Lange, M., Górska, A., Oude Elberink, J.N.G., van Doormaal, J., Słomka, M., Kalinowski, L., Gruchała-Niedoszytko, M., et al. (2020). Results from a genome-wide association study (GWAS) in mastocytosis reveal new gene polymorphisms associated with who subgroups. Int. J. Mol. Sci.

Neve, R.M., Lane, H.A., and Hynes, N.E. (2001). The role of overexpressed HER2 in transformation. Ann. Oncol.

Nguyen, B., Williams, A.B., Young, D.J., Ma, H., Li, L., Levis, M., Brown, P., and Small, D. (2017). FLT3 activating mutations display differential sensitivity to multiple tyrosine kinase inhibitors. Oncotarget.

Nica, A.C., and Dermitzakis, E.T. (2013). Expression quantitative trait loci: Present and future. Philos. Trans. R. Soc. B Biol. Sci.

Nielsen, R., Paul, J.S., Albrechtsen, A., and Song, Y.S. (2011). Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet.

NIH National Cancer Institute CANCER CLASSIFICATION.

Nikoloski, G., Langemeijer, S.M.C., Kuiper, R.P., Knops, R., Massop, M., Tönnissen, E.R.L.T.M., van der Heijden, A., Scheele, T.N., Vandenberghe, P., de Witte, T., et al. (2010). Somatic mutations of the histone methyltransferase gene EZH2 in myelodysplastic syndromes. Nat. Genet. *42*, 665–667.

Nowell, P.C. (1976). The clonal evolution of tumor cell populations. Science.

O'Keefe, C., McDevitt, M.A., and Maciejewski, J.P. (2010). Copy neutral loss of heterozygosity: A novel chromosomal lesion in myeloid malignancies. Blood *115*, 2731–2739.

O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., Bodily, P., Tian, L., Hakonarson, H., Johnson, W.E., et al. (2013). Low concordance of multiple variant-calling pipelines: Practical implications for exome and genome sequencing. Genome Med.

Obrochta, E., and Godley, L.A. (2018). Identifying patients with genetic predisposition to acute myeloid leukemia. Best Pract. Res. Clin. Haematol.

Olcaydu, D., Harutyunyan, A., Jäger, R., Berg, T., Gisslinger, B., Pabinger, I., Gisslinger, H., and Kralovics, R. (2009). A common JAK2 haplotype confers susceptibility to myeloproliferative neoplasms. Nat. Genet.

Oracki, S.A., Walker, J.A., Hibbs, M.L., Corcoran, L.M., and Tarlinton, D.M. (2010). Plasma cell development and survival. Immunol. Rev. *237*, 140–159.

Osumi, T., Tsujimoto, S. ichi, Tamura, M., Uchiyama, M., Nakabayashi, K., Okamura, K., Yoshida, M., Tomizawa, D., Watanabe, A., Takahashi, H., et al. (2018). Recurrent RARB translocations in acute promyelocytic leukemia lacking RARA translocation. Cancer Res.

Papaemmanuil, E., Hosking, F.J., Vijayakrishnan, J., Price, A., Olver, B., Sheridan, E., Kinsey, S.E., Lightfoot, T., Roman, E., Irving, J.A.E., et al. (2009). Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. Nat. Genet.

Papaemmanuil, E., Gerstung, M., Malcovati, L., Tauro, S., Gundem, G., Van Loo, P., Yoon, C.J., Ellis, P., Wedge, D.C., Pellagatti, A., et al. (2013). Clinical and biological implications of driver mutations in myelodysplastic syndromes. Blood *122*, 3616–3627.

Patel, U., Luthra, R., Medeiros, L.J., and Patel, K.P. (2017). Diagnostic, Prognostic, and Predictive Utility of Recurrent Somatic Mutations in Myeloid Neoplasms. Clin. Lymphoma, Myeloma Leuk. *17*, S62–S74.

Payne, V., and Kam, P.C.A. (2004). Mast cell tryptase: A review of its physiology and clinical significance. Anaesthesia.

Pe'er, I., Yelensky, R., Altshuler, D., and Daly, M.J. (2008). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genet. Epidemiol.

Pearson, T. a, and Manolio, T. a (2008). How to interpret a genome-wide association study. JAMA.

Peiris, M.N., Li, F., and Donoghue, D.J. (2019). BCR: A promiscuous fusion partner in hematopoietic disorders. Oncotarget.

Peprah, E., Xu, H., Tekola-Ayele, F., and Royal, C.D. (2015). Genome-wide association studies in Africans and African Americans: Expanding the framework of the genomics of human traits and disease. Public Health Genomics.

Perrotti, D., Cesi, V., Trotta, R., Guerzoni, C., Santilli, G., Campbell, K., Iervolino, A., Condorelli, F., Gambacorti-Passerini, C., Caligiuri, M.A., et al. (2002). BCR-ABL suppresses C/EBPα expression through inhibitory action of hnRNP E2. Nat. Genet. *30*, 48–58.

Peters, A.H.F.M., Kubicek, S., Mechtler, K., O'Sullivan, R.J., Derijck, A.A.H.A., Perez-Burgos, L., Kohlmaier, A., Opravil, S., Tachibana, M., Shinkai, Y., et al. (2003). Partitioning and Plasticity of Repressive Histone Methylation States in Mammalian Chromatin. Mol. Cell.

Pharoah, P.D.P., Tsai, Y.-Y., Ramus, S.J., Phelan, C.M., Goode, E.L., Lawrenson, K., Buckley, M., Fridley, B.L., Tyrer, J.P., Shen, H., et al. (2013). GWAS meta-analysis and replication identifies three new susceptibility loci for ovarian cancer. Nat. Genet. *45*, 362–370, 370e1-2.

Pisinger, C., Vestbo, J., Borch-Johnsen, K., and Jørgensen, T. (2005). Smoking cessation intervention in a large randomised population-based study. The Inter99 study. Prev. Med. (Baltim).

Popejoy, A.B., and Fullerton, S.M. (2016). Genomics is failing on diversity. Nature.

Powers, D.M.W. (2007). Evaluation: from precision, recall and f-factor. Tech. Rep. SEI-07-001.

Pruim, R.J., Welch, R.P., Sanna, S., Teslovich, T.M., Chines, P.S., Gliedt, T.P., Boehnke, M., Abecasis, G.R., Willer, C.J., and Frishman, D. (2011). LocusZoom: Regional visualization of genome-wide association scan results. In Bioinformatics, pp. 2336–2337.

Purcell, S., Cherny, S.S., and Sham, P.C. (2003). Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. Bioinformatics *19*, 149–150.

Purcell, S.M., Moran, J.L., Fromer, M., Ruderfer, D., Solovieff, N., Roussos, P., O'Dushlaine, C., Chambert, K., Bergen, S.E., Kähler, A., et al. (2014). A polygenic burden of rare disruptive mutations in schizophrenia. Nature.

Qi, X., Zhang, D.H., Wu, N., Xiao, J.H., Wang, X., and Ma, W. (2015). ceRNA in cancer: Possible functions and clinical implications. J. Med. Genet.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics.

Radomska, H.S., Bassères, D.S., Zheng, R., Zhang, P., Dayaram, T., Yamamoto, Y., Sternberg, D.W., Lokker, N., Giese, N.A., Bohlander, S.K., et al. (2006). Block of C/EBPα function by phosphorylation in acute myeloid leukemia with FLT3 activating mutations . J. Exp. Med. *203*, 371–381.

Rafnar, T., Sulem, P., Stacey, S.N., Geller, F., Gudmundsson, J., Sigurdsson, A., Jakobsdottir, M., Helgadottir, H., Thorlacius, S., Aben, K.K.H., et al. (2009). Sequence variants at the TERT-CLPTM1L locus associate with many cancer types. Nat. Genet.

Raghavan, M., Smith, L.L., Lillington, D.M., Chaplin, T., Kakkas, I., Molloy, G., Chelala, C., Cazier, J.B., Cavenagh, J.D., Fitzgibbon, J., et al. (2008). Segmental uniparental disomy is a commonly acquired genetic event in relapsed acute myeloid leukemia. Blood *112*, 814–821.

Rathinasamy, B., and Velmurugan, B.K. (2018). Role of lncRNAs in the cancer development and progression and their regulation by various phytochemicals. Biomed. Pharmacother.

Rausz, E., Szilágyi, A., Nedoszytko, B., Lange, M., Niedoszytko, M., Lautner-Csorba, O., Falus, A., Aladzsity, I., Kokai, M., Valent, P., et al. (2013). Comparative analysis of IL6 and IL6 receptor gene polymorphisms in mastocytosis. Br. J. Haematol.

Rayner, N.W., and Mccarthy, M.I. (2011). Development and Use Of a Pipeline to Generate Strand and Position Information for Common Genotyping Chips. In Presented at the Annual Meeting of the American Society of Human Genetics, Montreal, Canada, p. 3.

Regier, A.A., Farjoun, Y., Larson, D.E., Krasheninina, O., Kang, H.M., Howrigan, D.P., Chen, B.J., Kher, M., Banks, E., Ames, D.C., et al. (2018). Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. Nat. Commun.

Reinig, E.F., and He, R. (2017). Myelodysplastic/myeloproliferative neoplasm with ring sideroblasts and thrombocytosis with co-mutated JAK2 and SF3B1. Blood *129*, 656.

Remacle, J.E., Kraft, H., Lerchner, W., Wuytens, G., Collart, C., Verschueren, K., Smith, J.C., and Huylebroeck, D. (1999). New mode of DNA binding of multi-zinc finger transcription factors: δEF1 family members bind with two hands to two target sites. EMBO J.

Van Riel, B., and Rosenbauer, F. (2014). Epigenetic control of hematopoiesis: The PU.1 chromatin connection. Biol. Chem.

Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J.L., Kähler, A.K., Akterin, S., Bergen, S.E., Collins, A.L., Crowley, J.J., Fromer, M., et al. (2013). Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nat. Genet.

Robins, C., Liu, Y., Fan, W., Duong, D.M., Meigs, J., Harerimana, N. V., Gerasimov, E.S., Dammer, E.B., Cutler, D.J., Beach, T.G., et al. (2021). Genetic control of the human brain proteome. Am. J. Hum. Genet.

Robinson, W.P. (2000). Mechanisms leading to uniparental disomy and their clinical consequences. BioEssays.

Roman-Gomez, J., Jimenez-Velasco, A., Agirre, X., Cervantes, F., Sanchez, J., Garate, L., Barrios, M., Castillejo, J.A., Navarro, G., Colomer, D., et al. (2005). Promoter hypomethylation of the LINE-1 retrotransposable elements activates sense/antisense transcription and marks the progression of chronic myeloid leukemia. Oncogene.

Rumi, E., Pietra, D., Guglielmelli, P., Bordoni, R., Casetti, I., Milanesi, C., Sant'Antonio, E., Ferretti,

V., Pancrazzi, A., Rotunno, G., et al. (2013). Acquired copy-neutral loss of heterozygosity of chromosome 1p as a molecular event associated with marrow fibrosis in MPL-mutated myeloproliferative neoplasms. Blood.

Ryland, G.L., Doyle, M.A., Goode, D., Boyle, S.E., Choong, D.Y.H., Rowley, S.M., Li, J., Bowtell, D.D., Tothill, R.W., Campbell, I.G., et al. (2015). Loss of heterozygosity: what is it good for? BMC Med. Genomics *8*, 45.

Safaee, S., Fardi, M., Hemmat, N., Khosravi, N., Derakhshani, A., Silvestris, N., and Baradaran, B. (2021). Silencing ZEB2 Induces Apoptosis and Reduces Viability in Glioblastoma Cell Lines. Molecules.

Saleh, R., Wedeh, G., Herrmann, H., Bibi, S., Cerny-Reiterer, S., Sadovnik, I., Blatt, K., Hadzijusufovic, E., Jeanningros, S., Blanc, C., et al. (2014). A new human mast cell line expressing a functional IgE receptor converts to tumorigenic growth by KIT D816V transfection. Blood.

Salmena, L., Poliseno, L., Tay, Y., Kats, L., and Pandolfi, P.P. (2011). A ceRNA hypothesis: The rosetta stone of a hidden RNA language? Cell.

San Lucas, F.A., Sivakumar, S., Vattathil, S., Fowler, J., Vilar, E., and Scheet, P. (2016). Rapid and powerful detection of subtle allelic imbalance from exome sequencing data with hapLOHseq. Bioinformatics.

Sankaran, V.G., Menne, T.F., Xu, J., Akie, T.E., Lettre, G., Van Handel, B., Mikkola, H.K.A., Hirschhorn, J.N., Cantor, A.B., and Orkin, S.H. (2008). Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor BCL11A. Science (80-.).

Sankaran, V.G., Xu, J., Ragoczy, T., Ippolito, G.C., Walkley, C.R., Maika, S.D., Fujiwara, Y., Ito, M., Groudine, M., Bender, M.A., et al. (2009). Developmental and species-divergent globin switching are driven by BCL11A. Nature.

Schneider, V.A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.C., Kitts, P.A., Murphy, T.D., Pruitt, K.D., Thibaud-Nissen, F., Albracht, D., et al. (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res.

Schwartz, L.B., Lewis, R.A., and Austen, K.F. (1981). Tryptase from human pulmonary mast cells. Purification and characterization. J. Biol. Chem.

Schwartz, L.B., Sakai, K., Bradford, T.R., Ren, S., Zweiman, B., Worobec, A.S., and Metcalfe, D.D. (1995). The α form of human tryptase is the predominant type present in blood at baseline in normal subjects and is elevated in those with systemic mastocytosis. J. Clin. Invest.

Score, J., and Cross, N.C.P. (2012). Acquired Uniparental Disomy in Myeloproliferative Neoplasms. Hematol. Oncol. Clin. North Am. *26*, 981–991.

Semaan, A., Bernard, V., Lee, J.J., Wong, J.W., Huang, J., Swartzlander, D.B., Stephens, B.M., Monberg, M.E., Weston, B.R., Bhutani, M.S., et al. (2021). Defining the Comprehensive Genomic Landscapes of Pancreatic Ductal Adenocarcinoma Using Real-World Endoscopic Aspiration Samples. Clin. Cancer Res.

Sherry, S.T. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308–311.

Shi, X., Li, J., Ma, L., Wen, L., Wang, Q., Yao, H., Ruan, C., Wu, D., Zhang, X., and Chen, S. (2019). Overexpression of zeb2-as1 lncrna is associated with poor clinical outcomes in acute myeloid leukemia. Oncol. Lett.

Siegmund, K.D., Marjoram, P., Woo, Y.J., Tavaré, S., and Shibata, D. (2009). Inferring clonal expansion and cancer stem cell dynamics from DNA methylation patterns in colorectal cancers. Proc. Natl. Acad. Sci. U. S. A.

Sirugo, G., Williams, S.M., and Tishkoff, S.A. (2019). The Missing Diversity in Human Genetic Studies. Cell.

Sivakumar, S., Lucas, F.A.S., Yasminka A Jakubek, Ozcan, Z., Fowler, J., and Scheet, P. (2021). Pan cancer patterns of allelic imbalance from chromosomal alterations in 33 tumor types. Genetics.

Skoda, R.C., Duek, A., and Grisouard, J. (2015). Pathogenesis of myeloproliferative neoplasms. Exp. Hematol. *43*, 599–608.

Smith, J.G., and Newton-Cheh, C. (2009). Genome-wide association study in humans. Methods Mol Biol *573*, 231–258.

Smith, M.L., Cavenagh, J.D., Lister, T.A., and Fitzgibbon, J. (2004). Mutation of CEBPA in familial acute myeloid leukemia. N. Engl. J. Med.

Solovieff, N., Cotsapas, C., Lee, P.H., Purcell, S.M., and Smoller, J.W. (2013). Pleiotropy in complex traits: Challenges and strategies. Nat. Rev. Genet.

Soussi, T., and Wiman, K.G. (2015). TP53: An oncogene in disguise. Cell Death Differ.

Spain, S.L., and Barrett, J.C. (2015). Strategies for fine-mapping complex traits. Hum. Mol. Genet.

Staaf, J., Lindgren, D., Vallon-Christersson, J., Isaksson, A., Göransson, H., Juliusson, G., Rosenquist, R., Höglund, M., Borg, A., and Ringnér, M. (2008). Segmentation-based detection of

allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays. Genome Biol. *9*, R136.

Van Steensel, B. (2011). Chromatin: Constructing the big picture. EMBO J.

Stelzer, G., Dalah, I., Stein, T.I., Satanower, Y., Rosen, N., Nativ, N., Oz-Levi, D., Olender, T., Belinky, F., Bahir, I., et al. (2011). In-silico human genomics with GeneCards. Hum. Genomics *5*.

Strachan, T., and Read, A. (2011). Human molecular genetics.

Stransky, N., Cerami, E., Schalm, S., Kim, J.L., and Lengauer, C. (2014). The landscape of kinase fusions in cancer. Nat. Commun. *5*.

Stratton, M.R., Campbell, P.J., and Futreal, P.A. (2009). The cancer genome. Nature *458*, 719–724.

Straver, R., Weiss, M.M., Waisfisz, Q., Sistermans, E.A., and Reinders, M.J.T. (2017). WISExome: A within-sample comparison approach to detect copy number variations in whole exome sequencing data. Eur. J. Hum. Genet.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Med.

Sun, B.B., Maranville, J.C., Peters, J.E., Stacey, D., Staley, J.R., Blackshaw, J., Burgess, S., Jiang, T., Paige, E., Surendran, P., et al. (2018). Genomic atlas of the human plasma proteome. Nature.

Swerdlow, S.H., Campo, E., Harris, N.L., Jaffe, E.S., Pileri, S.A., Stein, H., Thiele, J., and Vardiman, J.. (2008). WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues. Lyon, France.

Swerdlow, S.H., Campo, E., Pileri, S.A., Lee Harris, N., Stein, H., Siebert, R., Advani, R., Ghielmini, M., Salles, G.A., Zelenetz, A.D., et al. (2016). The 2016 revision of the World Health Organization classification of lymphoid neoplasms. Blood *127*, 2375–2390.

Swets, J.A. (1988). Measuring the accuracy of diagnostic systems. Sci. Sci.

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. Nat. Rev. Genet.

Tanaka, J., Oshima, T., Hori, K., Tomita, T., Kim, Y., Watari, J., Oh, K., Hirota, S., Matsumoto, T., and Miwa, H. (2010). Small gastrointestinal stromal tumor of the stomach showing rapid growth and early metastasis to the liver. Dig. Endosc.

Tanaka, T., Shen, J., Abecasis, G.R., Kisialiou, A., Ordovas, J.M., Guralnik, J.M., Singleton, A., Bandinelli, S., Cherubini, A., Arnett, D., et al. (2009). Genome-wide association study of plasma

polyunsaturated fatty acids in the InCHIANTI study. PLoS Genet.

Tapper, W., Jones, A. V, Kralovics, R., Harutyunyan, A.S., Zoi, K., Leung, W., Godfrey, A.L., Guglielmelli, P., Callaway, A., Ward, D., et al. (2015). Genetic variation at MECOM, TERT, JAK2 and HBS1L-MYB predisposes to myeloproliferative neoplasms. Nat. Commun. *6*, 6691.

Tefferi, A., and Vardiman, J.W. (2008). Classification and diagnosis of myeloproliferative neoplasms: the 2008 World Health Organization criteria and point-of-care diagnostic algorithms. Leukemia *22*, 14–22.

Teo, Y.Y., Inouye, M., Small, K.S., Gwilliam, R., Deloukas, P., Kwiatkowski, D.P., and Clark, T.G. (2007). A genotype calling algorithm for the Illumina BeadArray platform. Bioinformatics *23*, 2741–2746.

Tesi, B., Davidsson, J., Voss, M., Rahikkala, E., Holmes, T.D., Chiang, S.C.C., Komulainen-Ebrahim, J., Gorcenco, S., Nilsson, A.R., Ripperger, T., et al. (2017). Gain-of-function SAMD9L mutations cause a syndrome of cytopenia, immunodeficiency, MDS, and neurological symptoms. Blood.

Thurner, M., van de Bunt, M., Torres, J.M., Mahajan, A., Nylander, V., Bennett, A.J., Gaulton, K.J., Barrett, A., Burrows, C., Bell, C.G., et al. (2018). Integration of human pancreatic islet genomic data refines regulatory mechanisms at type 2 diabetes susceptibility loci. Elife.

Tiedt, R., Tichelli, A., Skoda, R.C., Kralovics, R., Buser, A.S., Cazzola, M., Teo, S.-S., Passamonti, F., and Passweg, J.R. (2005). A Gain-of-Function Mutation of JAK2 in Myeloproliferative Disorders . N. Engl. J. Med.

Torabi, K., Erola, P., Alvarez-Mora, M.I., Díaz-Gay, M., Ferrer, Q., Castells, A., Castellví-Bel, S., Milà, M., Lozano, J.J., Miró, R., et al. (2019). Quantitative analysis of somatically acquired and constitutive uniparental disomy in gastrointestinal cancers. Int. J. Cancer.

Torres, J., Abdalla, M., Payne, A., Fernandez-Tajes, J., Thurner, M., Nylander, V., Gloyn, A., Mahajan, A., and McCarthy, M. (2020). A multi-omic integrative scheme characterizes tissues of action at loci associated with type 2 diabetes. Am. J. Hum. Genet. *107*, 1011–1028.

van der Touw, W., Chen, H.M., Pan, P.Y., and Chen, S.H. (2017). LILRB receptor-mediated regulation of myeloid cell maturation and function. Cancer Immunol. Immunother.

Trerotola, M., Relli, V., Simeone, P., and Alberti, S. (2015). Epigenetic inheritance and the missing heritability. Hum. Genomics.

Tuna, M., and Amos, C.I. (2010). Uniparental Disomy in Cancer - A New Tool in Molecular Cancer.

In Encyclopedia of Life Sciences, p.

Tuna, M., Knuutila, S., and Mills, G.B. (2009). Uniparental disomy in cancer. Trends Mol. Med.

Tuna, M., Smid, M., Martens, J.W.M., and Foekens, J.A. (2012). Prognostic value of acquired uniparental disomy (aUPD) in primary breast cancer. Breast Cancer Res. Treat.

Turner, S.D. (2018). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. J. Open Source Softw.

Turner, S., Armstrong, L.L., Bradford, Y., Carlsony, C.S., Crawford, D.C., Crenshaw, A.T., de Andrade, M., Doheny, K.F., Haines, J.L., Hayes, G., et al. (2011). Quality control procedures for genome-wide association studies. Curr. Protoc. Hum. Genet.

Uda, M., Galanello, R., Sanna, S., Lettre, G., Sankaran, V.G., Chen, W., Usala, G., Busonero, F., Maschio, A., Albai, G., et al. (2008). Genome-wide association study shows BCL11A associated with persistent fetal hemoglobin and amelioration of the phenotype of β-thalassemia. Proc. Natl. Acad. Sci. U. S. A.

Ulirsch, J.C., Lareau, C.A., Bao, E.L., Ludwig, L.S., Guo, M.H., Benner, C., Satpathy, A.T., Kartha, V.K., Salem, R.M., Hirschhorn, J.N., et al. (2019). Interrogation of human hematopoiesis at single-cell and single-variant resolution. Nat. Genet.

Ustun, C., Arock, M., Kluin-Nelemans, H.C., Reiter, A., Sperr, W.R., George, T., Horny, H.P., Hartmann, K., Sotlar, K., Damaj, G., et al. (2016). Advanced systemic mastocytosis: From molecular and genetic progress to clinical practice. Haematologica *101*, 1133–1143.

Valent, P., Akin, C., and Metcalfe, D.D. (2017a). Mastocytosis: 2016 updated WHO classification and novel emerging treatment concepts. Blood.

Valent, P., Akin, C., Hartmann, K., Nilsson, G., Reiter, A., Hermine, O., Sotlar, K., Sperr, W.R., Escribano, L., George, T.I., et al. (2017b). Advances in the classification and treatment of mastocytosis: Current status and outlook toward the future. Cancer Res. *77*, 1261–1270.

Vattathil, S., and Scheet, P. (2013). Haplotype-based profiling of subtle allelic imbalance with SNP arrays. Genome Res.

Verma, A. (2012). Risk Alleles for Multiple Sclerosis Identified by a Genomewide Study. Yearb. Neurol. Neurosurg. *2008*, 114–115.

Viñuela, A., Varshney, A., van de Bunt, M., Prasad, R.B., Asplund, O., Bennett, A., Boehnke, M., Brown, A.A., Erdos, M.R., Fadista, J., et al. (2020). Genetic variant effects on gene expression in human pancreatic islets and their implications for T2D. Nat. Commun.

Visscher, P.M., Hill, W.G., and Wray, N.R. (2008). Heritability in the genomics era - Concepts and misconceptions. Nat. Rev. Genet.

Visscher, P.M., Brown, M.A., McCarthy, M.I., and Yang, J. (2012). Five years of GWAS discovery. Am. J. Hum. Genet. *90*, 7–24.

Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. Am. J. Hum. Genet.

Volders, P.J., Anckaert, J., Verheggen, K., Nuytens, J., Martens, L., Mestdagh, P., and Vandesompele, J. (2019). Lncipedia 5: Towards a reference set of human long non-coding rnas. Nucleic Acids Res.

Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M.J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Kasela, S., et al. (2018a). Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. BioRxiv.

Võsa, U., Claringbould, A., Westra, H.J., Bonder, M.J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Kasela, S., et al. (2018b). Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. BioRxiv 1–57.

Vuckovic, D., Bao, E.L., Akbari, P., Lareau, C.A., Mousas, A., Jiang, T., Chen, M.H., Raffield, L.M., Tardaguila, M., Huffman, J.E., et al. (2020). The Polygenic and Monogenic Basis of Blood Traits and Diseases. Cell.

Walsh, C.S., Ogawa, S., Scoles, D.R., Miller, C.W., Kawamata, N., Narod, S.A., Koeffle, H.P., and Karlan, B.Y. (2008). Genome-wide loss of heterozygosity and uniparental disomy in BRCA1/2-associated ovarian carcinomas. Clin. Cancer Res.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res.

Wang, L., Wheeler, D.A., and Prchal, J.T. (2016). Acquired uniparental disomy of chromosome 9p in hematologic malignancies. Exp. Hematol. *44*, 644–652.

Wang, S.A., Hasserjian, R.P., Fox, P.S., Rogers, H.J., Geyer, J.T., Chabot-Richards, D., Weinzierl, E., Hatem, J., Jaso, J., Kanagal-Shamanna, R., et al. (2014a). Atypical chronic myeloid leukemia is clinically distinct from unclassifiable myelodysplastic/myeloproliferative neoplasms. Blood.

Wang, Z., Zhu, B., Zhang, M., Parikh, H., Jia, J., Chung, C.C., Sampson, J.N., Hoskins, J.W., Hutchinson, A., Burdette, L., et al. (2014b). Imputation and subset-based association analysis

across different cancer types identifies multiple independent risk loci in the TERT-CLPTM1L region on chromosome 5p15.33. Hum. Mol. Genet.

Ward, L.D., and Kellis, M. (2016). HaploReg v4: Systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. Nucleic Acids Res.

Wasag, B., Niedoszytko, M., Piskorz, A., Lange, M., Renke, J., Jassem, E., Biernat, W., Debiec-Rychter, M., and Limon, J. (2011). Novel, activating KIT-N822I mutation in familial cutaneous mastocytosis. Exp. Hematol. *39*.

Watanabe, K., Taskesen, E., Van Bochoven, A., and Posthuma, D. (2017). Functional mapping and annotation of genetic associations with FUMA. Nat. Commun.

Westra, H.J., Peters, M.J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M.W., Fairfax, B.P., Schramm, K., Powell, J.E., et al. (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat. Genet.

Wigginton, J.E., Cutler, D.J., and Abecasis, G.R. (2005). A Note on Exact Tests of Hardy-Weinberg Equilibrium. Am. J. Hum. Genet.

Wray, N.R. (2005). Allele Frequencies and the r2 Measure of Linkage Disequilibrium: Impact on Design and Interpretation of Association Studies. Twin Res. Hum. Genet.

Wu, X., Yan, T., Wang, Z., Wu, X., Cao, G., and Zhang, C. (2017). LncRNA ZEB2-AS1 promotes bladder cancer cell proliferation and inhibits apoptosis by regulating miR-27b. Biomed. Pharmacother.

Wu, Y.H., Graff, R.E., Passarelli, M.N., Hoffman, J.D., Ziv, E., Hoffmann, T.J., and Witte, J.S. (2018). Identification of pleiotropic cancer susceptibility variants from genome-wide association studies reveals functional characteristics. Cancer Epidemiol. Biomarkers Prev.

Xie, M., Lu, C., Wang, J., McLellan, M.D., Johnson, K.J., Wendl, M.C., McMichael, J.F., Schmidt, H.K., Yellapantula, V., Miller, C.A., et al. (2014). Age-related mutations associated with clonal hematopoietic expansion and malignancies. Nat. Med.

Xu, C., Cui, H., Li, H., Wu, Y., An, H., and Guo, C. (2019). Long non-coding RNA ZEB2-AS1 expression is associated with disease progression and predicts outcome in gastric cancer patients. J. B.U.ON.

Yang, H., and Wang, K. (2015). Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. Nat. Protoc.

Yang, M., Huang, W., Sun, Y., Liang, H., Chen, M., Wu, X., Wang, X., Zhang, L., Cheng, X., Fan, Y., et al. (2019). Prognosis and modulation mechanisms of COMMD6 in human tumours based on

expression profiling and comprehensive bioinformatics analysis. Br. J. Cancer.

Yao, J., Chen, J., Li, L.Y., and Wu, M. (2020). Epigenetic plasticity of enhancers in cancer. Transcription.

Yeboah, M., Papagregoriou, C., Des, C.J., Claude Chan, H.T., Hu, G., McPartlan, J.S., Schiött, T., Mattson, U., Ian Mockridge, C., Tornberg, U.C., et al. (2020). LILRB3 (ILT5) is a myeloid cell checkpoint that elicits profound immunomodulation. JCI Insight.

Young, A.L., Challen, G.A., Birmann, B.M., and Druley, T.E. (2016). Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. Nat. Commun.

Zakrzewski, F., de Back, W., Weigert, M., Wenke, T., Zeugner, S., Mantey, R., Sperling, C., Friedrich, K., Roeder, I., Aust, D., et al. (2019). Automated detection of the HER2 gene amplification status in Fluorescence in situ hybridization images for the diagnostics of cancer tissues. Sci. Rep.

Zanotti, R., Simioni, L., Garcia-Montero, A.C., Perbellini, O., Bonadonna, P., Caruso, B., Jara-Acevedo, M., Bonifacio, M., and De Matteis, G. (2013). Somatic D816V KIT mutation in a case of adult-onset familial mastocytosis. J. Allergy Clin. Immunol. *131*, 605–607.

Zeggini, E., and Ioannidis, J.P.A. (2009). Meta-analysis in genome-wide association studies. Pharmacogenomics *10*, 191–201.

Zhang, G., Li, H., Sun, R., Li, P., Yang, Z., Liu, Y., Wang, Z., Yang, Y., and Yin, C. (2019). Long non-coding RNA ZEB2-AS1 promotes the proliferation, metastasis and epithelial mesenchymal transition in triple-negative breast cancer by epigenetically activating ZEB2. J. Cell. Mol. Med.

Zheng, Z., Huang, D., Wang, J., Zhao, K., Zhou, Y., Guo, Z., Zhai, S., Xu, H., Cui, H., Yao, H., et al. (2020). QTLbase: an integrative resource for quantitative trait loci across multiple human molecular phenotypes. Nucleic Acids Res.

Zink, F., Stacey, S.N., Norddahl, G.L., Frigge, M.L., Magnusson, O.T., Jonsdottir, I., Thorgeirsson, T.E., Sigurdsson, A., Gudjonsson, S.A., Gudmundsson, J., et al. (2017). Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. Blood.

Zondervan, K.T., and Cardon, L.R. (2007). Designing candidate gene and genome-wide case–control association studies. Nat. Protoc. *2*, 2492–2501.

Zuvich, R.L., Armstrong, L.L., Bielinski, S.J., Bradford, Y., Carlson, C.S., Crawford, D.C., Crenshaw, A.T., de Andrade, M., Doheny, K.F., Haines, J.L., et al. (2011). Pitfalls of merging GWAS data:

Lessons learned in the eMERGE network and quality control procedures to maintain high data quality. Genet. Epidemiol. *35*, 887–898.

ARTICLE

Genome-wide association study identifies novel susceptibility loci for *KIT* D816V positive mastocytosis

Gabriella Galatà,[1] Andrés C. García-Montero,[2,3] Thomas Kristensen,[4,5] Ahmed A.Z. Dawoud,[1] Javier I. Muñoz-González,[2,3] Manja Meggendorfer,[6] Paola Guglielmelli,[7] Yvette Hoade,[1] Ivan Alvarez-Twose,[8] Christian Gieger,[9,10,11,12] Konstantin Strauch,[9,13,14] Luigi Ferrucci,[15] Toshiko Tanaka,[15] Stefania Bandinelli,[16] Theresia M. Schnurr,[17] Torsten Haferlach,[6] Sigurd Broesby-Olsen,[5,18,19] Hanne Vestergaard,[5,20] Michael Boe Møller,[4,5] Carsten Bindslev-Jensen,[5,18,19] Alessandro M. Vannucchi,[7] Alberto Orfao,[2,3] Deepti Radia,[21] Andreas Reiter,[22] Andrew J. Chase,[1,23] Nicholas C.P. Cross,[1,23,24,*] and William J. Tapper[1,24]

Summary

Mastocytosis is a rare myeloid neoplasm characterized by uncontrolled expansion of mast cells, driven in >80% of affected individuals by acquisition of the *KIT* D816V mutation. To explore the hypothesis that inherited variation predisposes to mastocytosis, we performed a two-stage genome-wide association study, analyzing 1,035 individuals with *KIT* D816V positive disease and 17,960 healthy control individuals from five European populations. After quality control, we tested 592,007 SNPs at stage 1 and 75 SNPs at stage 2 for association by using logistic regression and performed a fixed effects meta-analysis to combine evidence across the two stages. From the meta-analysis, we identified three intergenic SNPs associated with mastocytosis that achieved genome-wide significance without heterogeneity between cohorts: rs4616402 ($p_{meta} = 1.37 \times 10^{-15}$, OR = 1.52), rs4662380 ($p_{meta} = 2.11 \times 10^{-12}$, OR = 1.46), and rs13077541 ($p_{meta} = 2.10 \times 10^{-9}$, OR = 1.33). Expression quantitative trait analyses demonstrated that rs4616402 is associated with the expression of *CEBPA* ($p_{eQTL} = 2.3 \times 10^{-14}$), a gene encoding a transcription factor known to play a critical role in myelopoiesis. The role of the other two SNPs is less clear: rs4662380 is associated with expression of the long non-coding RNA gene *TEX41* ($p_{eQTL} = 2.55 \times 10^{-11}$), whereas rs13077541 is associated with the expression of *TBL1XR1,* which encodes transducin (β)-like 1 X-linked receptor 1 ($p_{eQTL} = 5.70 \times 10^{-8}$). In individuals with available data and non-advanced disease, rs4616402 was associated with age at presentation (p = 0.009; beta = 4.41; n = 422). Additional focused analysis identified suggestive associations between mastocytosis and genetic variation at *TERT*, *TPSAB1/TPSB2*, and *IL13*. These findings demonstrate that multiple germline variants predispose to *KIT* D816V positive mastocytosis and provide novel avenues for functional investigation.

Introduction

Mastocytosis (MIM: 154800) is an uncommon myeloid neoplasm characterized by expansion and accumulation of clonal mast cells in one or more organ systems, including bone marrow, skin, liver, spleen, and gastrointestinal tract. The extent of organ infiltration and organ damage serves as the basis for classification as cutaneous mastocytosis (CM) or systemic mastocytosis (SM).[1] CM is typically found in children, while most adults with mastocytosis have SM with involvement of the bone marrow. Six main subtypes of SM are recognized: indolent SM (ISM) and smoldering systemic mastocytosis (SMM) are relatively benign forms that usually have a stable clinical course over many years. In contrast, SM with an associated hematologic neoplasm (SM-AHN), aggressive SM (ASM), and mast cell

[1]School of Medicine, University of Southampton, Southampton SO17 1BJ, UK; [2]Institute of Biomedical Research of Salamanca, Salamanca 37007, Spain; [3]Servicio de Citometría, Departamento de Medicina, CIBERONC, and Instituto de Biología Molecular y Celular del Cáncer, CSIC/Universidad de Salamanca, Salamanca 37007, Spain; [4]Department of Pathology, Odense University Hospital, 5000 Odense, Denmark; [5]Mastocytosis Centre Odense University Hospital, 5000 Odense, Denmark; [6]Munich Leukemia Laboratory, 81377 Munich, Germany; [7]Centro di Ricerca e Innovazione per le Malattie Mieloproliferative, Azienda Ospedaliera Universitaria Careggi, Dipartimento di Medicina Sperimentale e Clinica, Università Degli Studi di Firenze, 50134 Firenze, Italy; [8]Instituto de Mastocitosis de Castilla La Mancha, Hospital Virgen del Valle, 45071 Toledo, Spain; [9]Institute of Genetic Epidemiology, Helmholtz Zentrum München – German Research Center for Environmental Health, 85764 Neuherberg, Germany; [10]German Centre for Cardiovascular Research Partner Site Munich Heart Alliance, 80802 Munich, Germany; [11]Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, Germany Research Center for Environmental Health, 85764 Neuherberg, Germany; [12]German Center for Diabetes Research, 85764 Neuherberg, Germany; [13]Chair of Genetic Epidemiology, IBE, Faculty of Medicine, LMU Munich, 80539 Munich, Germany; [14]Institute of Medical Biostatistics, Epidemiology and Informatics, University Medical Center, Johannes Gutenberg University, 55131 Mainz, Germany; [15]Longitudinal study section, Translation Gerontology Branch, National Institute on Aging, Baltimore, MD 21224, USA; [16]Geriatric Unit, Azienda USL Toscana centro, 50137 Firenze, Italy; [17]Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen, Denmark; [18]Department of Dermatology and Allergy Centre, Odense University Hospital, 5000 Odense, Denmark; [19]Odense Research Center for Anaphylaxis, Odense University Hospital, 5000 Odense, Denmark; [20]Department of Hematology, Odense University Hospital, 5000 Odense, Denmark; [21]Department of Clinical Haematology, Guy's and St Thomas' NHS Hospitals, London SE1 9RT, UK; [22]University Hospital Mannheim, Heidelberg University, 68167 Mannheim, Germany; [23]Wessex Regional Genetics Laboratory, Salisbury NHS Foundation Trust, Salisbury SP2 8BJ, UK
[24]These authors contributed equally
*Correspondence: ncpc@soton.ac.uk
https://doi.org/10.1016/j.ajhg.2020.12.007.

leukemia (MCL), collectively known as advanced SM (advSM), are associated with a poor prognosis.[2] ISM is the most common of the six subtypes, accounting for 80% of SM-affected individuals.[3]

Approximately 80%–90% of adult SM-affected individuals across all subtypes test positive for the somatic mutation *KIT* c.2447A>T (p.Asp816Val), usually referred to as *KIT* D816V. Due to the nature of the disease, the mutant allele frequency is often very low, particularly in peripheral blood samples, and sensitive methods are needed for its detection.[4] *KIT* D816V mutation burden, serum tryptase, and β2-microglobulin levels correlate with disease burden and severity,[5–8] and for advSM, additional somatic mutations in *SRSF2*, *ASXL1*, and *RUNX1* indicate an adverse prognosis.[9–11]

Mastocytosis is usually a sporadic disorder, but familial forms have been described, often in association with inherited, weakly activating *KIT* mutations.[12,13] Very occasionally, familial clustering of *KIT* D816V has been observed, but in all affected individuals, this mutation is somatically acquired[14] and, as a strongly activating variant, *KIT* D816V is believed to be incompatible with normal embryonic development and thus not transmissible through the germline. Other lines of evidence suggest the possibility of a broader role for genetic variation in mastocytosis. The presence of germline variants in genes known to be somatically mutated in myeloid disorders was one of several factors related to adverse clinical outcome in SM.[11] Studies of mast cell activation disease (MCAD), a disorder that overlaps with SM, indicate a substantial excess of symptoms in first-degree relatives of affected individuals, which might suggest a common genetic susceptibility.[15,16] Several constitutional genetic variants have been associated with the development of different mastocytosis phenotypes in relatively small candidate gene studies[17–21] and a recent single-stage genome-wide association study (GWAS) of 234 affected individuals.[22] Finally, it has been clearly established that constitutional genetic variation at several loci predispose to other myeloproliferative neoplasms (MPN).[23,24]

To determine whether common genetic variation plays a role in predisposition to mastocytosis, we have performed a robust two-stage GWAS focusing on affected individuals that tested positive for *KIT* D816V regardless of clinical subtype to help ensure a genetically homogeneous cohort. We anticipate that the identification of validated genetic markers associated with mastocytosis will provide novel lines of investigation to understand this complex disorder.

Material and methods

Discovery and replication cohorts

Prior to quality control (QC), the stage 1 discovery individuals consisted of 479 *KIT* D816V positive mastocytosis-affected individuals recruited from the UK (n = 329) and Germany (n = 150). These affected individuals were compared with healthy control individuals from the UK Wellcome Trust Case Control Consortium (WTCCC2, n = 5,200)[25] and the German Cooperative Health Research in the Region of Augsburg study (KORA, n = 4,397), respectively.[26] At stage 2, 666 independent *KIT* D816V positive replication individuals were recruited from Spain (n = 399), Denmark (n = 185), and Italy (n = 82) and compared to published population controls from the Spanish National DNA Bank (SNDNAB, n = 1,062),[27,28] a Danish study of ischemic heart disease (Inter99, n = 6,184),[29,30] and the Italian Invecchiare in Chianti study (InCHIANTI, n = 1,210).[31,32] Participants provided informed consent for sampling according to the Declaration of Helsinki. The number of samples that were recruited and used for analysis after QC in the discovery and replication stages is shown in Table S1. An overview of the two-stage study design and sample numbers is shown in Figure S1. All mastocytosis-affected individuals were adults diagnosed via standard procedures. Further details on the five cohorts are provided in the Supplemental methods.[2,4]

Genotyping

DNA was extracted from peripheral blood or bone marrow. The stage 1 affected individuals were genotyped for 960,919 SNPs via Infinium OmniExpress exome chips (version 8_1.4_A1) and the Genome Studio software (GSGT version 1.9.4) at the Clinical Research Facility in Edinburgh. These data are available on request from ArrayExpress (accession number E-MTAB-9358). The stage-2 affected individuals were genotyped for 92 SNPs via custom designed Kompetitive Allele Specific PCR (KASP) at LGC.[33] Genotypic data for the control cohorts were obtained from published studies. In WTCCC2, genotypes were called with Illumina 1.2M Duo chips and Illumina's program to call SNPs with a posterior probability >0.95.[34] KORA control individualss were genotyped for 2,443,177 SNPs via the Illumina human Omni chip (version 2.5-4v1_B) in KORA_A (a subset of follow-up F3 of the population-based survey KORA S3) and 730,372 SNPs with Illumina human Omni express chips (version 12v1_H) in KORA_B (an independent subset of KORA S3/F3). Control individuals from SNDNAB, Inter99, and InCHIANTI were genotyped with Illumina Global Screening arrays, Illumina HumanOmniExpress-24 (versions 1.0A and 1.1A), and Illumina Infinium HumanHap 550K SNP arrays, which include 18, 90, and 45 of the SNPs selected for replication, respectively. Genotypes for the remaining SNPs were determined by imputation.

Quality control

Standard GWAS QC measures[35] were applied to the genotypic data with Plink prior to analysis.[36] These measures included genotype missingness (per sample and per SNP), minor allele frequency (MAF), Hardy Weinberg equilibrium (HWE), heterozygosity (Figure S2), sex inference, cryptic relatedness, strand orientation, and population stratification with multidimensional scaling (MDS) (Figure S3). Since the affected individuals and control individuals were genotyped separately, SNPs were excluded if they had modest deviation from HWE in control individuals (p value < 0.001) or extreme deviation in affected individuals (p value $\leq 1 \times 10^{-10}$), which most likely reflects poor genotyping rather than disease association.[37] The number of SNPs and samples removed by these QC measures is shown in Table S1. QC and imputation of the stage 2 control individuals has previously been described.[28–32] Full details regarding the QC and imputation procedures are given in the Supplemental methods.

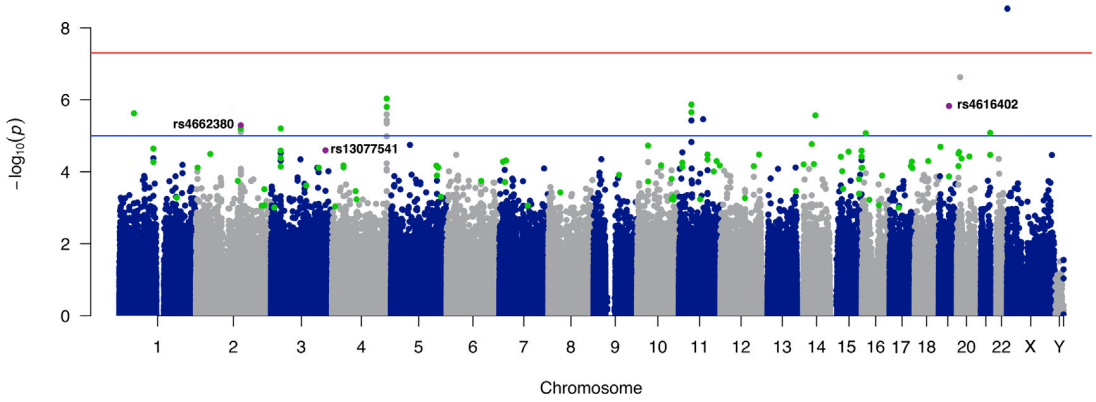


Figure 1. Genome-wide association of *KIT* D816V positive mastocytosis
Manhattan plot showing results from the stage 1 meta-analysis of the UK and German cohorts for all 24 chromosomes. Results are plotted for 592,007 SNPs tested as −log10 of the meta-analysis p values on the y axis against genomic location on the x axis. One SNP was identified with genome-wide significance (p value $< 5 \times 10^{-8}$), indicated by the red line, and a further 18 SNPs were identified with suggestive p values ($<1 \times 10^{-5}$), indicated by the blue line. SNPs selected for replication are highlighted in green, and the three SNPs that reached genome-wide significance after meta-analysis of stages 1 and 2 are highlighted in purple.

Imputation

Imputation of the discovery cohorts was used to increase SNP density and enable fine mapping around significant loci. SNPs were imputed with the Sanger imputation server,[38] which used EAGLE2 for pre-phasing into the Haplotype Reference Consortium (HRC release 1.1) and positional Burrows-Wheeler transform (PBWT) for imputation. Imputed genotypes were quality controlled by exclusion of SNPs with info score <0.80, posterior genotype probabilities less than 0.99, MAF less than 1%, greater than 10% missing genotypes, or extreme deviation from HWE (p value $\leq 1 \times 10^{-10}$).

Statistical analysis

SNPs were tested for association via binary logistic regression in Plink. We carried out a fixed effects inverse variance-weighted meta-analysis by using Plink to combine evidence from the stage 1 cohorts (UK and Germany) and to determine the final effect sizes and significance levels by combining evidence across stages 1 and 2. Heterogeneity between studies was estimated with the χ^2-based Cochran's Q statistic and the I^2 statistic, which describes the percentage of variation across studies that is due to heterogeneity rather than chance. To examine the effectiveness of the QC measures and assess evidence for any systematic biases, we used the qqnorm and qqplot procedures in R to construct quantile-quantile (QQ) plots for the stage 1 analysis of the UK and German cohorts and the stage 1 meta-analysis (Figure S4). Samples with evidence of non-Caucasian ancestry were excluded rather than adjusting the association analysis for population stratification. To examine the effect of this decision, we retained the ancestry outliers and repeated the stage 1 analyses with adjustment for the first two principal components from the MDS analysis (Figure S5 and Table S2).

We visualized and interpreted the results from the stage 1 meta-analysis by using the qqman package[39] in R to create a Manhattan plot (Figure 1) and the FUMA software to generate regional plots.[40] Results from the final meta-analysis of stages 1 and 2 were displayed in a forest plot with Stata (Figure 2).

The power to detect SNPs associated with SM was estimated with the genetic power calculator[41] under a multiplicative genetic risk model and a type 1 error rate of 5×10^{-8} (Figure S6). We used a range of genotype relative risks (1.1–2.0) and risk allele frequencies (MAF 0.05–0.4) to estimate power assuming a disease prevalence of 1 in 100,000[42] and unselected control individuals.

Selection of SNPs for replication

To minimize false positives and the potential for overlooking signals with compelling functional evidence but modest significance, we used the following method to select SNPs for follow-up at stage 2. First, we used a clumping procedure in Plink to generate a shortlist of index SNPs ($p < 0.001$) with support from correlated SNPs (SNPs $r2 > 0.5$, within 500 kb and $p < 0.01$) based on the stage 1 meta-analysis. From this shortlist, 92 index SNPs were selected for replication, and priority, but not exclusivity, was given to SNPs that were either located in or flanked by a gene with functional relevance according to annotation from GeneAlacart.[43] Relevant functions were signal transduction components, hematopoiesis, myeloid leukemia, and myeloproliferative or mast cell conditions from GeneAlacart.[43] A total of 44 SNPs were selected with functional relevance. We then infilled the number of selected SNPs to 82 by selecting the most significant remaining index SNPs. We selected an additional 10 SNPs were selected as backups and to add support to the most promising signals in terms of either their biological relevance, individual significance, or level of support from correlated SNPs.

Identification of chromosomal abnormalities

We identified regions of acquired uniparental disomy (aUPD) and copy number gains or losses in the stage 1 SM-affected individuals by using B allele frequency (BAF) segmentation[44] followed by post processing to select likely somatic events as described[45] and manual review of all BAF plots (Figure S7). See Supplemental methods for further details.

Functional annotation of variants

We explored the biological relevance of regions containing genome-wide significant SNPs by using HaploReg (version 4.1)[46] to annotate the lead SNP and its proxies ($r^2 \geq 0.8$) with respect to histone modification, sequence conservation by using genomic evolutionary rate profiling (GERP),[47] estimated pathogenicity by using combined annotation-dependent depletion (CADD)

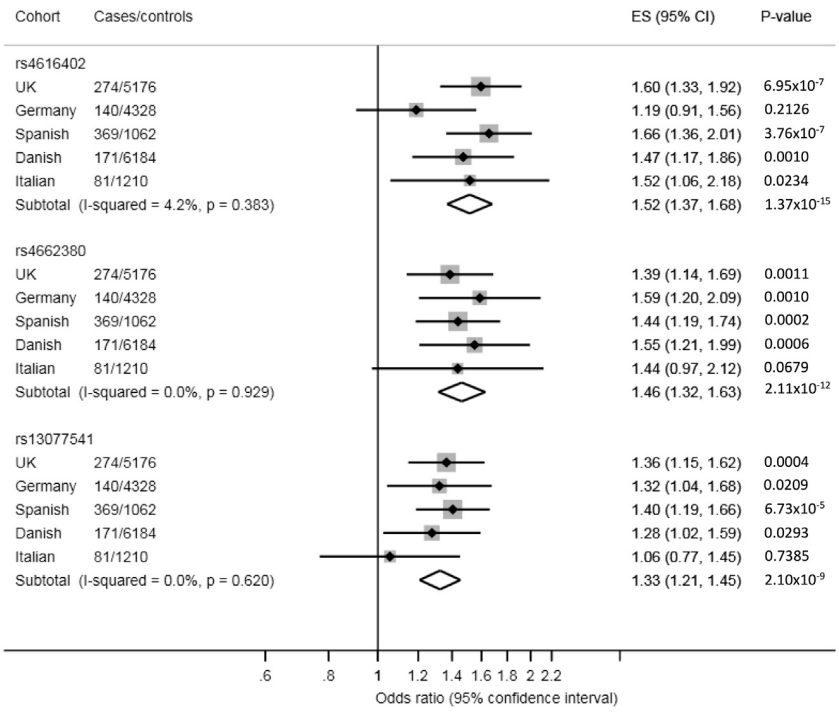


Figure 2. Forest plots and meta-analysis for three SNPs reaching genome-wide significance

Forest plots for each SNP associated with SM at a genome-wide level of significance. Odds ratios (OR = ES) and 95% confidence intervals (CIs) are displayed on the x axis. Results are shown for each cohort (UK, German, Spanish, Danish, and Italian) and the combined analysis. The SNP subtotals and diamond show the final OR and CI for a fixed effects meta-analysis of all five cohorts and uses I^2 to assess heterogeneity in effect sizes between cohorts.

scores,[48] predicted effect on protein binding by using RegulomeDB[49] scores (SNPs scoring ≤3 are likely to affect binding), and previous associations with clinical phenotypes by using the NHGRI-EBI GWAS catalog.[50] Additionally, candidate regions were annotated against a 15-state chromatin model[51] in primary hematopoietic stem cells (E035) and a myeloid leukemia cell line (K562). This model categorizes non-coding DNA into active or repressed states that are respectively enriched and depleted for phenotype-associated SNPs.[52] To gain further functional insight, we performed expression and methylation quantitative trait loci (eQTL and mQTL, respectively) analyses on the lead SNP and its proxies ($r^2 \geq 0.8$) by using GTEx v8[53] and QTLbase.[54] Finally, we used LNCipedia[55] and the Cancer LncRNA Census (CLC)[56] to investigate the function of long non-coding RNA (lncRNA).

Association with clinical features

Diagnostic and phenotypic variables for initial diagnosis (advanced, ASM, SM-AHN, MCL; non-advanced, all other subtypes), the presence or absence of skin lesions (yes or no), gender, baseline serum tryptase (ng/mL), and age were available for most of the Spanish (n = 369) and Italian (n = 81) individuals but not for other cohorts. Three categorical variables (initial diagnosis, skin lesions, and sex) were tested for association with allelic counts for the three significant SNPs via Fisher's exact test. Continuous variables (tryptase and age) were tested via linear regression following Kolmogorov-Smirnov checks for normal distribution and normalization of tryptase levels via quantile transformation. We used a fixed effects inverse variance-weighted meta-analysis to combine evidence from the two cohorts.

Results

Discovery stage

After QC of the stage 1 data, 592,007 SNPs were tested for association with *KIT* D816V positive mastocytosis via binary logistic regression in the UK (274 affected individuals versus 5,176 control individuals) and German cohorts (140 affected individuals versus 4,328 control individuals) (Table S1). Summary statistics from these analyses, which are available from LocusZoom, were combined with a fixed effects meta-analysis.[57] The QQ plots for each analysis and their low genomic inflation factors ($\lambda \leq 1.038$) demonstrate a close agreement with the null hypothesis until the tail of the distribution where SNPs with p values less than 10^{-4} become more significant than expected by chance alone (Figure S4). Consequently, systematic biases such as the separate genotyping of our affected individuals and control individuals, residual population stratification, or clonal somatic changes are unlikely to account for the significance of these SNPs. A Manhattan plot summarizing the results of the stage 1 meta-analysis is shown in Figure 1. A total of 18 SNPs were identified with suggestive p values ($p \leq 1 \times 10^{-5}$).

Replication and final meta-analysis

According to the number of samples that passed QC and using a multiplicative disease model, we estimated the stage 1 analysis to have 80% power to detect common SNPs (MAF = 0.4) with a relative risk (RR) of 1.56 and rare SNPs (MAF = 0.1) with an RR of 1.82 (Figure S6A). Because of the potential to overlook SNPs with smaller effect sizes, we used a set of selection criteria rather than significance alone (see Material and methods) to identify 92 SNPs for replication. These SNPs were selected to have support from correlated SNPs and were either the most significant (n = 38), surpassed a moderate significance threshold ($p < 0.001$) and were located in or flanked by a functionally relevant gene (n = 44), or were selected as backups for the most promising signals (n = 10). One SNP, rs7884433, achieved genome-wide significance in the stage 1 analysis, but it was not selected for replication because it lacked support from any of the SNPs in strong linkage disequilibrium (LD) and is thus likely to be a technical artifact.

Table 1. Summary of the most significant SNPs from meta-analysis of stages 1 and 2

SNP	Chr	Location (hg19)	Alleles	RAF	Gene	p_{meta}	OR (CI)	I^2
rs4616402	19q13	33,753,555	A/G	0.240	*SLC7A10-CEBPA*	1.37×10^{-15}	1.52 (1.37–1.68)	4.2
rs4662380	2q22	145,316,407	C/T	0.189	*LINC01412*	2.11×10^{-12}	1.46 (1.32–1.63)	0
rs13077541	3q26	176,925,740	G/A	0.464	*TBL1XR1-LINC00501*	2.10×10^{-9}	1.33 (1.21–1.45)	0

SNP, rs identifier from dbSNP; alleles, risk associated/non-risk associated allele; RAF, risk allele frequency in Europeans from 1000 genomes; p_{meta}, fixed effects meta-analysis of stages 1 and 2; OR, odds ratio; CI, 95% confidence interval; I^2, heterogeneity index (0–100).

Of the 92 SNPs selected, 75 were successfully genotyped in 666 *KIT* D816V mastocytosis-affected individuals from Spain, Denmark, and Italy. Additional control individuals (n = 8,456) from the same populations that had previously been genotyped were used for comparison. After QC, 621 affected individuals and all the control individuals remained for analysis. All SNPs passed QC in affected individuals, although 19 were excluded from the Spanish control individuals because of per SNP missingness (≥10%) following imputation. Samples were tested for association with SM as three separate cohorts via binary logistic regression. We determined the final significance levels and effect sizes by using a fixed effects inverse variance-weighted meta-analysis to combine evidence from stages 1 and 2. This meta-analysis identified three intergenic SNPs with genome-wide significance: rs4616402 ($p_{meta} = 1.37 \times 10^{-15}$), rs4662380 ($p_{meta} = 2.11 \times 10^{-12}$), and rs13077541 ($p_{meta} = 2.10 \times 10^{-9}$) (Table 1). Results for the three SNPs reaching genome-wide significance are summarized in a forest plot that shows that each SNP is significant in four of the five cohorts tested and that there is evidence for the same trend in the remaining population (Figure 2). Cochran's Q test and I^2 statistics showed that for each SNP there was no evidence of heterogeneity between cohorts. Results from the meta-analysis of stages 1 and 2 for all SNPs tested are shown in Table S3.

To investigate the possibility of residual population stratification, we repeated the stage 1 analyses without removing 26 samples with evidence of outlying ancestry (Table S1) and adjusting the association analysis by using the first two principal components from MDS. The top three SNPs retained genome-wide significance, and rs4662380 and rs13077541 became slightly more significant (Table S2), which suggests an absence of residual population stratification in the original analysis.

Functional annotation and candidate gene mapping

To explore the functional relevance of the regions associated with mastocytosis, we used HaploReg and RegulomeDB to determine whether the risk SNP or its proxies ($r^2 \geq 0.8$) were located in regions with potential regulatory functions based on chromatin modification, DNA methylation, and alteration of transcription factor (TF)-binding motifs (Table S4). To gain further functional insight, we performed eQTL and mQTL analyses on the lead SNP and its proxies by using GTEx v8[53] and QTLbase.[54] Finally, we repeated the stage 1 meta-analysis by using imputation to enable fine mapping around the lead SNPs and to generate association results for proxies, which had not been directly genotyped.

The most significant SNP, rs4616402, confers a 1.52-fold increased risk of developing mastocytosis and is situated in an intergenic region on chromosome 19 between a solute carrier gene (*SLC7A10*, 36.8 kb downstream) and a gene encoding a transcription factor (*CEBPA*, 37.2 kb downstream) that coordinates proliferation and differentiation of myeloid progenitor cells (Figure 3A). Using QTLbase, we found that rs4616402 is strongly associated with the expression of *CEBPA* in whole blood according to data from three previous eQTL studies ($p_{eQTL} = 2.30 \times 10^{-14}$; $p_{eQTL} = 2.96 \times 10^{-11}$; $p_{eQTL} = 9.20 \times 10^{-9}$).[58–60] There is no evidence that *SLC7A10* has a role in carcinogenesis, including myeloid malignancies, and no additional SNPs were identified in strong LD with rs4616402. However, there is weak evidence that rs4616402 may have functional consequences according to the RegulomeDB score (score = 4). The chromatin surrounding rs4616402 is characterized as an enhancer (7_Enh) in primary hematopoietic stem cells because of an enrichment of the H3K4me1 signature. Additionally, the risk allele is predicted to alter three TF-binding motifs (Arnt_1, Gm397, and Hmx_1, Table S4).

The second most significant SNP, rs4662380, increases the risk of developing mastocytosis by 1.46-fold and is located in the first intron of a lincRNA gene (*LINC01412*) (Figure 3B). Twelve additional SNPs in LINC01412 were identified in strong LD with the lead. Three of these proxies are located in chromatin enhancers (7_Enh: rs6722387, rs16823865, and rs13413446) in primary hematopoietic stem cells, and one is located in a flanking active transcription start site (2_TssAFlnk: rs16823855) in K562 (Table S4). The RegulomeDB scores indicate that two of the proxies, rs4662227 (score = 2c) and rs13413446 (score = 3a), are likely to affect TF binding, while the remaining SNPs are estimated to have weak evidence for functional consequences. However, using the GWAS catalog,[50] we found that one of the remaining proxies, rs16823866, was strongly associated with white blood cell counts in two previous studies ($p = 4 \times 10^{-18}$ and $p = 6 \times 10^{-11}$).[62,63] Finally, using QTLbase, we found that the lead SNP ($p_{eQTL} = 2.55 \times 10^{-11}$) and four proxies, including rs16823866 ($p_{eQTL} = 2.55 \times 10^{-11}$), were strongly associated with the expression of the nearby gene *TEX41* in neutrophils.[64]

The final SNP, rs13077541, is associated with a 1.33-fold increase in risk of developing mastocytosis and is located

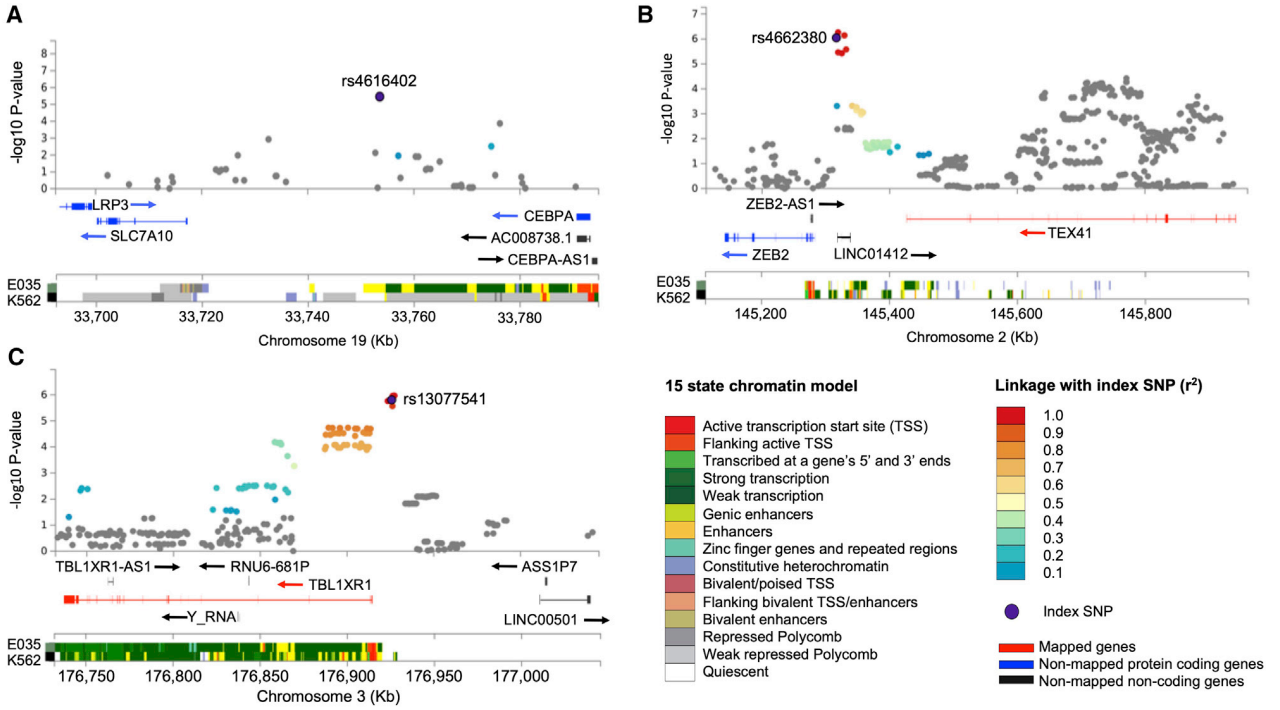


Figure 3. Regional plots of the imputed stage 1 meta-analysis for SNPs reaching genome-wide significance in the final meta-analysis
(A–C) Results from the imputed stage 1 meta-analysis in a region surrounding three SNPs (rs4616402 [A], rs4662380 [B], and rs13077541 [C]) that predispose to SM and reached genome-wide significance in the final meta-analysis. In each plot, the leading SNP is indicated by a purple circle and the color of other SNPs represent the strength of linkage disequilibrium (r^2) with the lead SNP. Protein-coding genes and RNA genes are shown in the track below with arrows to indicate the direction of transcription and wider lines representing the location of exons. The lower panel displays the 15-state chromatin track (chromHMM) in primary hematopoietic stem cells (E035) and K562 with data from the NIH Roadmap Epigenomics Consortium.[61] Physical positions are relative to build 37 (hg19) of the human genome.

in an intergenic region of chromosome 3 between transducin (β)-like 1 X-linked receptor 1 (*TBL1XR1*, 10.6 kb upstream) and another lincRNA gene (*LINC00501*, 86.5 kb upstream) (Figure 3C). Fifty-three additional SNPs were identified in strong LD with the lead, including 27 intronic SNPs in *TBL1XR1* (Table S4). Eleven of these proxies are located in active chromatin regions, including three in an active transcription start site (1_TssA: rs12493005, rs12486557, and rs34302523) and two in a 5′ transcribed region (3_TxFlnk: rs35072945 and rs34311793) in K562. The RegulomeDB scores indicate that five of the proxies are likely to affect binding (score2a–c: rs6790639, rs34302523, rs6772872, rs7616138, and rs1920131). Of these, rs6790639 is particularly relevant because the PU.1 TF, which is encoded by the Spi-1 proto-oncogene (*SPI1*), has been shown to bind to this region in K562 via ChIP sequencing.[65] PU.1, together with other TFs, regulates the expression of genes involved myelopoiesis.[66] Using QTLbase, we found that the lead SNP ($p_{eQTL} = 5.70 \times 10^{-8}$) and one of the proxies, rs16823866 ($p_{eQTL} = 9.52 \times 10^{-9}$), were strongly associated with the expression of *TBL1XR1* in CD4+ naive T cells.[64]

Association with clinical features

To determine whether variants that predispose to the development of mastocytosis relate to particular clinical features, we used Fisher's exact tests and linear regression to correlate allelic counts for the three significant SNPs with clinical phenotypes in the Spanish and Italian cohorts (Table 2), the only affected individuals for which clinical information was available. A significant association that remained significant after correction for multiple testing was identified between rs4616402 and age at presentation (n = 422; p = 0.009; beta = 4.41) in individuals with non-advanced disease. No association with age was seen in the much smaller group of individuals (n = 26) with advanced disease, a subgroup for which additional mutations may be a confounding factor. In affected individuals, the age of onset was estimated to increase by 4.41 years per risk allele. No associations were seen with baseline tryptase levels, gender, skin lesions, or disease phenotype.

Association with *TPSAB1* and *TPSB2*

Increased copy number variation at *TPSAB1*, the gene at 16p13 encoding α-tryptase, is associated with elevated serum tryptase levels in hereditary α-tryptasemia.[67] Our analysis did not include direct copy number analysis of this gene; however, a recent study linked *TPSAB1* duplications with three SNPs, including rs58124832.[68] This SNP was genotyped at stage 1 and met our criteria for analysis at stage 2, yielding a suggestive overall association with SM ($p_{meta} = 9.03 \times 10^{-6}$). The Cochran's Q test and I^2 statistics showed no evidence of heterogeneity between

Table 2. Association between the most significant SNPs and clinical phenotypes in the Spanish and Italian cohorts

Phenotype	Number of affected individuals	rs4662380		rs13077541		rs4616402	
		p value	Effect size (CI)	p value	Effect size (CI)	p value	Effect size (CI)
Initial diagnosis (indolent/advanced)	422/26	0.175	0.58 (0.26–1.27)	0.646	0.88 (0.50–1.54)	0.238	0.60 (0.25–1.40)
Sex (F/M)	235/214	0.266	1.18 (0.88–1.60)	0.384	1.12 (0.86–1.46)	0.904	1.03 (0.65–1.61)
Skin lesions (+/−)	275/122	0.638	1.08 (0.77–1.51)	0.151	0.81 (0.60–1.08)	0.406	1.23 (0.75–2.00)
Age at diagnosis	422	0.668	0.55 (−1.97–3.07)	0.625	0.67 (−2.02–3.35)	0.009	4.41 (1.09–7.73)
Tryptase	417	0.452	−0.08 (−0.29–0.13)	0.136	−0.17 (−0.39–0.05)	0.249	0.17 (−0.12–0.45)

Categorical phenotypes: initial diagnosis (422 indolent versus 26 advanced mastocytosis-affected individuals), sex (235 female versus 214 male individuals), and skin lesions (275 individuals with skin phenotype versus 122 individuals without skin phenotype); p value, fixed effects meta-analysis of Italian and Spanish Fisher's exact test; effect size, odds ratio; CI, 95% confidence interval. Continuous phenotypes: age at diagnosis and tryptase levels tested in individuals with non-advanced phenotype; p value, linear regression; effect size, regression coefficient beta; CI, 95% confidence interval.

cohorts; however, the association was significant in only three cohorts ($p_{German} = 0.0058$, $p_{UK} = 0.0042$, and $p_{Spanish} = 0.05$). The eQTL analysis showed that rs58124832 is strongly associated with the expression of *TPSAB1* ($p_{eQTL} < 1.9 \times 10^{-58}$) and *TPSB2* (tryptase-β2; $p_{eQTL} = 1.96 \times 10^{-75}$) in blood.

Association with *TERT*

Several *TERT* SNPs have been identified as risk factors for the development of hematological malignancies, including MPN, as well as some solid tumors. Our stage 1 analysis included rs2853677, which has been linked to both MPN and *JAK2* V617F associated clonal hematopoiesis.[24] This SNP marginally failed to meet our criteria for analysis at stage 2; however, the stage 1 meta-analysis for directly genotyped UK and German affected individuals showed $p_{meta} = 0.0011$, suggesting the possibility of an association. To examine this in more detail, we imputed genotypes for 64 additional SNPs spanning *TERT* and tested their association with SM. As shown in Table S5, seven SNPs achieved p values < 0.001. The strongest of these was for rs7726159 ($p_{meta} = 8 \times 10^{-5}$), an established risk SNP for multiple cancer types.[69] We identified one secondary association at *TERT* for rs2853677, which remained significant after conditioning on rs7726159 ($p_{conditional} = 0.035$). No associations were seen with other SNPs that predispose to MPN[70] or clonal hematopoiesis of indeterminate potential[71] in our stage 1 data (Table S6).

Associations with other genetic factors

To the best of our knowledge, 14 SNPs have been associated with the development or phenotype of human mastocytosis in published studies.[17–22] Of these, 11 were directly genotyped or could be imputed from our stage 1 data (Table S7), but only one of these was significant: rs1800925 in the promoter region of *IL13* at 5q31 ($p_{imputed} = 0.008$). This SNP has been linked to the development of adult SM and serum interleukin-13 levels[18] and inflammatory disorders such as chronic obstructive pulmonary disease.[72]

Discussion

Despite being characterized by a common somatic oncogenic driver mutation, mastocytosis is a complex disorder with a broad range of clinical phenotypes and outcomes. In this study, we have identified constitutional genotype as an additional factor contributing to the heterogeneity of mastocytosis. The use of a molecular definition for affected individuals rather than clinically defined subtypes and careful ethnicity matching of affected individuals and control individuals aimed to reduce the chance of heterogeneity both in the primary and replication cohorts. Thus, with a relatively modest cohort size for a GWAS, we were able to identify and validate three novel SNPs that achieved genome-wide significance and additional suggestive associations at *TERT*, *TPSAB1/TPSB2*, and *IL13* that merit further investigation. Notably, apart from rs1800925 (*IL13*), we did not confirm any of the previously published associations derived from candidate gene studies and a recent GWAS that did not include a replication cohort (Table S6). In addition, we found no evidence that genetic variation at *KIT* is associated with acquisition of *KIT* D816V, unlike the finding in MPN that the *JAK2* haplotype strongly influences the probability of acquiring *JAK2* V617F.[73]

Theoretically, common genetic variation may influence mastocytosis by distinct mechanisms, for example by promoting or favoring the outgrowth of a *KIT* D816V positive clone that arose by random mutation (fertile ground hypothesis); by increasing the probability that a *KIT* D816V mutation arises in a stem cell (hypermutability hypothesis); or by promoting the development of signs or symptoms in an individual with a *KIT* D816V positive clone, thus increasing the chance of clinical investigation (phenotypic hypothesis). We considered the possibility that clonal somatic changes might affect the analysis; however, we found that mastocytosis genomes are relatively simple in that only a small proportion of affected individuals showed likely somatic copy number changes or acquired uniparental disomy (Figure S7). Furthermore,

apart from isolated affected individuals, the genomic regions with somatic changes did not include the risk factors we identified.

Of the three significant SNPs identified in this study, the strongest association was seen for rs4616402 at 19q13. Interestingly, this SNP was significantly associated with age of diagnosis in individuals with non-advanced disease. This SNP is located in a candidate enhancer, and the risk allele is linked to reduced expression of *CEBPA*,[60] located 37.3 kb upstream. Another 19q13 SNP, rs78744187, has previously been linked to basophil counts and shown to modulate the activity of a *CEBPA* enhancer;[74] however, this variant is not in LD with rs4616402 ($r^2 = 0.22$). *CEBPA* is an intronless gene that encodes a leucine zipper TF that binds to the CCAAT motif in the promoter of its target genes. It is expressed in myeloid progenitor cells, and several studies have defined its critical role in myelopoiesis and malignant transformation of myeloid cells.[75] Of particular relevance, high C/EBPα expression inhibits the production of mast cells from mast/basophil common progenitors, whereas low C/EBPα expression inhibits the production of basophils.[71] Although the consequence of reduced *CEBPA* levels in the context of *KIT* D816V remains to be defined, reduced *CEBPA* expression associated with rs4616402 may be relevant to the fertile ground and phenotypic hypothesis defined above by creating an environment that favors the production of mast cells. It is striking that *CEBPA* or its product, C/EBPα, is targeted by two other oncogenic tyrosine kinases: BCR-ABL1 downregulates *CEBPA* by a post-transcriptional mechanism[76] and oncogenic FLT3 mutants disrupt C/EBPα function by ERK1/2-mediated phosphorylation.[77] Furthermore, low *CEBPA* expression is commonly seen in acute myeloid leukemia, although the underlying mechanism is unclear.[75] Detailed functional studies are needed to clarify the relationship between *KIT* D816V-driven clonal outgrowth and *CEBPA* expression.

The second most significant SNP, rs4662380, is located at 2q22 within the lincRNA *LINC01412* and associated with higher expression of the nearby gene *TEX41*. Both are of unknown function, but because of the possibility of long range interactions between GWAS signals and target genes, it is unclear whether either are directly relevant to SM. *ZEB2* is another nearby gene that has been linked to both myeloid and lymphoid leukemias,[78,79] but we found no association between rs4662380 and *ZEB2* expression. Interestingly, rs16823866, a SNP strongly linked to rs4662380, was associated with elevated white blood cells and, specifically, basophils in three independent population studies.[62,63,80] Although the underlying mechanism is unclear, this may be relevant to the phenotypic hypothesis in that affected individuals with abnormal blood counts may be more likely to be investigated clinically. The final SNP, rs13077541, is linked to expression of *TBL1XR1*. This gene has been reported as a fusion partner of *PDGFRB, ROS1*, *RARA*, and *RARB* in myeloid malignancies,[81–83] but its significance in relation to SM remains to be established.

Data and code availability

Genotyping data are available at ArrayExpress (https://www.ebi.ac.uk/arrayexpress/; accession number E-MTAB-9358). GWAS summary statistics are available at LocusZoom (http://locuszoom.org/ under "Mastocytosis GWAS").

Supplemental Information

Supplemental Information can be found online at https://doi.org/10.1016/j.ajhg.2020.12.007.

Acknowledgments

A full list of the investigators who contributed to the generation of the WTCCC data is available from the WTCCC website, funding for which was provided by The Wellcome Trust under award 07611. The KORA study was initiated and financed by the Helmholtz Zentrum München—German Research Center for Environmental Health, which is funded by the German Federal Ministry of Education and Research (BMBF) and by the State of Bavaria. Furthermore, KORA research was supported within the Munich Center of Health Sciences (MC-Health), Ludwig-Maximilians-Universität, as part of LMUinnovativ. The KORA Study Group consists of A. Peters (speaker), J. Heinrich, R. Holle, R. Leidl, C. Meisinger, K. Strauch, and their co-workers, who are responsible for the design and conduct of the KORA studies. We gratefully acknowledge the contribution of all members of field staff conducting the KORA study, and we are grateful to all study participants of KORA for their invaluable contributions to this study. The InCHIANTI study baseline (1998–2000) was supported as a "targeted project" (ICS110.1/RF97.71) by the Italian Ministry of Health and in part by the U.S. National Institute on Aging (263 MD 9164 and 263 MD 821336). The Spanish mastocytosis cohort and the Spanish National DNA Bank were supported by grants from the Instituto de Salud Carlos III and FEDER (PI16/00642 and PT17/0015/0044). The Italian cohort of mastocytosis-affected individuals was funded by the Associazione Italiana per la Ricerca sul cancro, Mynerva project, 21267. The Novo Nordisk Foundation Center for Basic Metabolic Research is an independent Research Center at the University of Copenhagen partially funded by an unrestricted donation from the Novo Nordisk Foundation. A.A.Z.D. was supported by a Lady Tata International Award; G.G., N.C.P.C., Y.H., A.J.C., and W.J.T. were supported by Blood Cancer UK (13002 and 18007).

Declaration of interests

The authors declare no competing interests

Received: October 9, 2020
Accepted: December 7, 2020
Published: January 8, 2021

Web resources

OMIM, https://www.omim.org/entry/154800
Wellcome Trust Case Control Consortium, https://www.wtccc.org.uk/

References

1. Arber, D.A., Orazi, A., Hasserjian, R., Thiele, J., Borowitz, M.J., Le Beau, M.M., Bloomfield, C.D., Cazzola, M., and Vardiman, J.W. (2016). The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. Blood *127*, 2391–2405.
2. Valent, P., Akin, C., and Metcalfe, D.D. (2017). Mastocytosis: 2016 updated WHO classification and novel emerging treatment concepts. Blood *129*, 1420–1427.
3. Cohen, S.S., Skovbo, S., Vestergaard, H., Kristensen, T., Møller, M., Bindslev-Jensen, C., Fryzek, J.P., and Broesby-Olsen, S. (2014). Epidemiology of systemic mastocytosis in Denmark. Br. J. Haematol. *166*, 521–528.
4. Arock, M., Sotlar, K., Akin, C., Broesby-Olsen, S., Hoermann, G., Escribano, L., Kristensen, T.K., Kluin-Nelemans, H.C., Hermine, O., Dubreuil, P., et al. (2015). KIT mutation analysis in mast cell neoplasms: recommendations of the European Competence Network on Mastocytosis. Leukemia *29*, 1223–1232.
5. Sperr, W.R., Kundi, M., Alvarez-Twose, I., van Anrooij, B., Oude Elberink, J.N.G., Gorska, A., Niedoszytko, M., Gleixner, K.V., Hadzijusufovic, E., Zanotti, R., et al. (2019). International prognostic scoring system for mastocytosis (IPSM): a retrospective cohort study. Lancet Haematol. *6*, e638–e649.
6. Erben, P., Schwaab, J., Metzgeroth, G., Horny, H.P., Jawhar, M., Sotlar, K., Fabarius, A., Teichmann, M., Schneider, S., Ernst, T., et al. (2014). The KIT D816V expressed allele burden for diagnosis and disease monitoring of systemic mastocytosis. Ann. Hematol. *93*, 81–88.
7. Hoermann, G., Gleixner, K.V., Dinu, G.E., Kundi, M., Greiner, G., Wimazal, F., Hadzijusufovic, E., Mitterbauer, G., Mannhalter, C., Valent, P., and Sperr, W.R. (2014). The KIT D816V allele burden predicts survival in patients with mastocytosis and correlates with the WHO type of the disease. Allergy *69*, 810–813.
8. Muñoz-González, J.I., Álvarez-Twose, I., Jara-Acevedo, M., Henriques, A., Viñas, E., Prieto, C., Sánchez-Muñoz, L., Caldas, C., Mayado, A., Matito, A., et al. (2019). Frequency and prognostic impact of *KIT* and other genetic variants in indolent systemic mastocytosis. Blood *134*, 456–468.
9. Jawhar, M., Schwaab, J., Álvarez-Twose, I., Shoumariyeh, K., Naumann, N., Lübke, J., Perkins, C., Muñoz-González, J.I., Meggendorfer, M., Kennedy, V., et al. (2019). MARS: Mutation-Adjusted Risk Score for Advanced Systemic Mastocytosis. J. Clin. Oncol. *37*, 2846–2856.
10. Jawhar, M., Schwaab, J., Schnittger, S., Meggendorfer, M., Pfirrmann, M., Sotlar, K., Horny, H.P., Metzgeroth, G., Kluger, S., Naumann, N., et al. (2016). Additional mutations in SRSF2, ASXL1 and/or RUNX1 identify a high-risk group of patients with KIT D816V(+) advanced systemic mastocytosis. Leukemia *30*, 136–143.
11. Muñoz-González, J.I., Jara-Acevedo, M., Alvarez-Twose, I., Merker, J.D., Teodosio, C., Hou, Y., Henriques, A., Roskin, K.M., Sanchez-Muñoz, L., Tsai, A.G., et al. (2018). Impact of somatic and germline mutations on the outcome of systemic mastocytosis. Blood Adv. *2*, 2814–2828.
12. Zhang, L.Y., Smith, M.L., Schultheis, B., Fitzgibbon, J., Lister, T.A., Melo, J.V., Cross, N.C., and Cavenagh, J.D. (2006). A novel K509I mutation of KIT identified in familial mastocytosis-in vitro and in vivo responsiveness to imatinib therapy. Leuk. Res. *30*, 373–378.
13. Wasag, B., Niedoszytko, M., Piskorz, A., Lange, M., Renke, J., Jassem, E., Biernat, W., Debiec-Rychter, M., and Limon, J. (2011). Novel, activating KIT-N822I mutation in familial cutaneous mastocytosis. Exp. Hematol. *39*, 859–65.e2.
14. Zanotti, R., Simioni, L., Garcia-Montero, A.C., Perbellini, O., Bonadonna, P., Caruso, B., Jara-Acevedo, M., Bonifacio, M., and De Matteis, G. (2013). Somatic D816V KIT mutation in a case of adult-onset familial mastocytosis. J. Allergy Clin. Immunol. *131*, 605–607.
15. Molderings, G.J., Haenisch, B., Bogdanow, M., Fimmers, R., and Nöthen, M.M. (2013). Familial occurrence of systemic mast cell activation disease. PLoS ONE *8*, e76241.
16. Haenisch, B., Nöthen, M.M., and Molderings, G.J. (2012). Systemic mast cell activation disease: the role of molecular genetic alterations in pathogenesis, heritability and diagnostics. Immunology *137*, 197–205.
17. Daley, T., Metcalfe, D.D., and Akin, C. (2001). Association of the Q576R polymorphism in the interleukin-4 receptor alpha chain with indolent mastocytosis limited to the skin. Blood *98*, 880–882.
18. Nedoszytko, B., Niedoszytko, M., Lange, M., van Doormaal, J., Gleń, J., Zabłotna, M., Renke, J., Vales, A., Buljubasic, F., Jassem, E., et al. (2009). Interleukin-13 promoter gene polymorphism -1112C/T is associated with the systemic form of mastocytosis. Allergy *64*, 287–294.
19. Rausz, E., Szilágyi, A., Nedoszytko, B., Lange, M., Niedoszytko, M., Lautner-Csorba, O., Falus, A., Aladzsity, I., Kokai, M., Valent, P., et al. (2013). Comparative analysis of IL6 and IL6 receptor gene polymorphisms in mastocytosis. Br. J. Haematol. *160*, 216–219.
20. Lange, M., Gleń, J., Zabłotna, M., Nedoszytko, B., Sokołowska-Wojdyło, M., Rębała, K., Ługowska-Umer, H., Niedoszytko, M., Górska, A., Sikorska, M., et al. (2017). Interleukin-31 Polymorphisms and Serum IL-31 Level in Patients with Mastocytosis: Correlation with Clinical Presen-tation and Pruritus. Acta Derm. Venereol. *97*, 47–53.
21. Nedoszytko, B., Lange, M., Renke, J., Niedoszytko, M., Zabłotna, M., Gleń, J., and Nowicki, R. (2018). The Possible Role of Gene Variant Coding Nonfunctional Toll-Like Receptor 2 in the Pathogenesis of Mastocytosis. Int. Arch. Allergy Immunol. *177*, 80–86.
22. Nedoszytko, B., Sobalska-Kwapis, M., Strapagiel, D., Lange, M., Górska, A., Elberink, J.N.G.O., van Doormaal, J., Słomka, M., Kalinowski, L., Gruchała-Niedoszytko, M., et al. (2020). Results from a Genome-Wide Association Study (GWAS) in Mastocytosis Reveal New Gene Polymorphisms Associated with WHO Subgroups. Int. J. Mol. Sci. *21*, 5506.
23. Tapper, W., Jones, A.V., Kralovics, R., Harutyunyan, A.S., Zoi, K., Leung, W., Godfrey, A.L., Guglielmelli, P., Callaway, A., Ward, D., et al. (2015). Genetic variation at MECOM, TERT, JAK2 and HBS1L-MYB predisposes to myeloproliferative neoplasms. Nat. Commun. *6*, 6691.
24. Hinds, D.A., Barnholt, K.E., Mesa, R.A., Kiefer, A.K., Do, C.B., Eriksson, N., Mountain, J.L., Francke, U., Tung, J.Y., Nguyen, H.M., et al. (2016). Germ line variants predispose to both JAK2 V617F clonal hematopoiesis and myeloproliferative neoplasms. Blood *128*, 1121–1128.
25. International Parkinson's Disease Genomics Consortium (IPDGC); and Wellcome Trust Case Control Consortium 2 (WTCCC2) (2011). A two-stage meta-analysis identifies several new loci for Parkinson's disease. PLoS Genet. *7*, e1002142.
26. Wichmann, H.E., Gieger, C., Illig, T.; and MONICA/KORA Study Group (2005). KORA-gen–resource for population genetics, controls and a broad spectrum of disease phenotypes. Gesundheitswesen *67* (*Suppl 1*), S26–S30.

27. Bosch, X. (2004). Spain to establish national genetic database. Lancet *363*, 1044.
28. Julià, A., Domènech, E., Ricart, E., Tortosa, R., García-Sánchez, V., Gisbert, J.P., Nos Mateu, P., Gutiérrez, A., Gomollón, F., Mendoza, J.L., et al. (2013). A genome-wide association study on a southern European population identifies a new Crohn's disease susceptibility locus at RBX1-EP300. Gut *62*, 1440–1445.
29. Jørgensen, T., Borch-Johnsen, K., Thomsen, T.F., Ibsen, H., Glümer, C., and Pisinger, C. (2003). A randomized non-pharmacological intervention study for prevention of ischaemic heart disease: baseline results Inter99. Eur. J. Cardiovasc. Prev. Rehabil. *10*, 377–386.
30. Pisinger, C., Vestbo, J., Borch-Johnsen, K., and Jørgensen, T. (2005). Smoking cessation intervention in a large randomised population-based study. The Inter99 study. Prev. Med. *40*, 285–292.
31. Ferrucci, L., Bandinelli, S., Benvenuti, E., Di Iorio, A., Macchi, C., Harris, T.B., and Guralnik, J.M. (2000). Subsystems contributing to the decline in ability to walk: bridging the gap between epidemiology and geriatric practice in the InCHIANTI study. J. Am. Geriatr. Soc. *48*, 1618–1625.
32. Tanaka, T., Shen, J., Abecasis, G.R., Kisialiou, A., Ordovas, J.M., Guralnik, J.M., Singleton, A., Bandinelli, S., Cherubini, A., Arnett, D., et al. (2009). Genome-wide association study of plasma polyunsaturated fatty acids in the InCHIANTI Study. PLoS Genet. *5*, e1000338.
33. He, C., Holme, J., and Anthony, J. (2014). SNP genotyping: the KASP assay. Methods Mol. Biol. *1145*, 75–86.
34. Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661–678.
35. Anderson, C.A., Pettersson, F.H., Clarke, G.M., Cardon, L.R., Morris, A.P., and Zondervan, K.T. (2010). Data quality control in genetic case-control association studies. Nat. Protoc. *5*, 1564–1573.
36. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
37. Marees, A.T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., and Derks, E.M. (2018). A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. Int. J. Methods Psychiatr. Res. *27*, e1608.
38. McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., et al.; Haplotype Reference Consortium (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nat. Genet. *48*, 1279–1283.
39. Turner, S.D. (2014). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. bioRxiv. https://doi.org/10.1101/005165.
40. Watanabe, K., Taskesen, E., van Bochoven, A., and Posthuma, D. (2017). Functional mapping and annotation of genetic associations with FUMA. Nat. Commun. *8*, 1826.
41. Purcell, S., Cherny, S.S., and Sham, P.C. (2003). Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. Bioinformatics *19*, 149–150.
42. Coltoff, A., and Mascarenhas, J. (2019). Relevant updates in systemic mastocytosis. Leuk. Res. *81*, 10–18.
43. Stelzer, G., Dalah, I., Stein, T.I., Satanower, Y., Rosen, N., Nativ, N., Oz-Levi, D., Olender, T., Belinky, F., Bahir, I., et al. (2011). In-silico human genomics with GeneCards. Hum. Genomics *5*, 709–717.
44. Staaf, J., Lindgren, D., Vallon-Christersson, J., Isaksson, A., Göransson, H., Juliusson, G., Rosenquist, R., Höglund, M., Borg, A., and Ringnér, M. (2008). Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays. Genome Biol. *9*, R136.
45. Dawoud, A.A.Z., Tapper, W.J., and Cross, N.C.P. (2020). Clonal myelopoiesis in the UK Biobank cohort: ASXL1 mutations are strongly associated with smoking. Leukemia *34*, 2660–2672.
46. Ward, L.D., and Kellis, M. (2016). HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. Nucleic Acids Res. *44* (D1), D877–D881.
47. Cooper, G.M., Stone, E.A., Asimenos, G., Green, E.D., Batzoglou, S., Sidow, A.; and NISC Comparative Sequencing Program (2005). Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. *15*, 901–913.
48. Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. *46*, 310–315.
49. Boyle, A.P., Hong, E.L., Hariharan, M., Cheng, Y., Schaub, M.A., Kasowski, M., Karczewski, K.J., Park, J., Hitz, B.C., Weng, S., et al. (2012). Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. *22*, 1790–1797.
50. Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. *47* (D1), D1005–D1012.
51. Ernst, J., and Kellis, M. (2017). Chromatin-state discovery and genome annotation with ChromHMM. Nat. Protoc. *12*, 2478–2492.
52. Hoffman, M.M., Ernst, J., Wilder, S.P., Kundaje, A., Harris, R.S., Libbrecht, M., Giardine, B., Ellenbogen, P.M., Bilmes, J.A., Birney, E., et al. (2013). Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res. *41*, 827–841.
53. Carithers, L.J., and Moore, H.M. (2015). The Genotype-Tissue Expression (GTEx) Project. Biopreserv. Biobank. *13*, 307–308.
54. Zheng, Z., Huang, D., Wang, J., Zhao, K., Zhou, Y., Guo, Z., Zhai, S., Xu, H., Cui, H., Yao, H., et al. (2020). QTLbase: an integrative resource for quantitative trait loci across multiple human molecular phenotypes. Nucleic Acids Res. *48* (D1), D983–D991.
55. Volders, P.J., Anckaert, J., Verheggen, K., Nuytens, J., Martens, L., Mestdagh, P., and Vandesompele, J. (2019). LNCipedia 5: towards a reference set of human long non-coding RNAs. Nucleic Acids Res. *47* (D1), D135–D139.
56. Carlevaro-Fita, J., Lanzós, A., Feuerbach, L., Hong, C., Mas-Ponte, D., Pedersen, J.S., Johnson, R.; PCAWG Drivers and Functional Interpretation Group; and PCAWG Consortium (2020). Cancer LncRNA Census reveals evidence for deep functional conservation of long noncoding RNAs in tumorigenesis. Commun. Biol. *3*, 56.
57. Pruim, R.J., Welch, R.P., Sanna, S., Teslovich, T.M., Chines, P.S., Gliedt, T.P., Boehnke, M., Abecasis, G.R., and Willer, C.J. (2010). LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics *26*, 2336–2337.
58. Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M.J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Kasela, S.,

et al. (2018). Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. bioRxiv. https://doi.org/10.1101/447367.
59. Westra, H.J., Peters, M.J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M.W., Fairfax, B.P., Schramm, K., Powell, J.E., et al. (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat. Genet. *45*, 1238–1243.
60. Lloyd-Jones, L.R., Holloway, A., McRae, A., Yang, J., Small, K., Zhao, J., Zeng, B., Bakshi, A., Metspalu, A., Dermitzakis, M., et al. (2017). The Genetic Architecture of Gene Expression in Peripheral Blood. Am. J. Hum. Genet. *100*, 228–237.
61. Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al.; Roadmap Epigenomics Consortium (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.
62. Kanai, M., Akiyama, M., Takahashi, A., Matoba, N., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K., et al. (2018). Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. Nat. Genet. *50*, 390–400.
63. Astle, W.J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A.L., Mead, D., Bouman, H., Riveros-Mckay, F., Kostadima, M.A., et al. (2016). The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell *167*, 1415–1429.e19.
64. Chen, L., Ge, B., Casale, F.P., Vasquez, L., Kwan, T., Garrido-Martín, D., Watt, S., Yan, Y., Kundu, K., Ecker, S., et al. (2016). Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. Cell *167*, 1398–1414.e24.
65. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.
66. van Riel, B., and Rosenbauer, F. (2014). Epigenetic control of hematopoiesis: the PU.1 chromatin connection. Biol. Chem. *395*, 1265–1274.
67. Lyons, J.J., Yu, X., Hughes, J.D., Le, Q.T., Jamil, A., Bai, Y., Ho, N., Zhao, M., Liu, Y., O'Connell, M.P., et al. (2016). Elevated basal serum tryptase identifies a multisystem disorder associated with increased TPSAB1 copy number. Nat. Genet. *48*, 1564–1569.
68. Lyons, J.J., Stotz, S.C., Chovanec, J., Liu, Y., Lewis, K.L., Nelson, C., DiMaggio, T., Jones, N., Stone, K.D., Sung, H., et al. (2018). A common haplotype containing functional CACNA1H variants is frequently coinherited with increased TPSAB1 copy number. Genet. Med. *20*, 503–512.
69. Wang, Z., Zhu, B., Zhang, M., Parikh, H., Jia, J., Chung, C.C., Sampson, J.N., Hoskins, J.W., Hutchinson, A., Burdette, L., et al. (2014). Imputation and subset-based association analysis across different cancer types identifies multiple independent risk loci in the TERT-CLPTM1L region on chromosome 5p15.33. Hum. Mol. Genet. *23*, 6616–6633.
70. Bao, E.L., Nandakumar, S.K., Liao, X., Bick, A.G., Karjalainen, J., Tabaka, M., Gan, O.I., Havulinna, A.S., Kiiskinen, T.T.J., Lareau, C.A., et al.; FinnGen; and 23andMe Research Team (2020). Inherited myeloproliferative neoplasm risk affects haematopoietic stem cells. Nature *586*, 769–775.
71. Bick, A.G., Weinstock, J.S., Nandakumar, S.K., Fulco, C.P., Bao, E.L., Zekavat, S.M., Szeto, M.D., Liao, X., Leventhal, M.J., Nasser, J., et al.; NHLBI Trans-Omics for Precision Medicine Consortium (2020). Inherited causes of clonal haematopoiesis in 97,691 whole genomes. Nature *586*, 763–768.
72. Ahmadi, A., Ghaedi, H., Salimian, J., Azimzadeh Jamalkandi, S., and Ghanei, M. (2019). Association between chronic obstructive pulmonary disease and interleukins gene variants: A systematic review and meta-analysis. Cytokine *117*, 65–71.
73. Jones, A.V., Chase, A., Silver, R.T., Oscier, D., Zoi, K., Wang, Y.L., Cario, H., Pahl, H.L., Collins, A., Reiter, A., et al. (2009). JAK2 haplotype is a major risk factor for the development of myeloproliferative neoplasms. Nat. Genet. *41*, 446–449.
74. Guo, M.H., Nandakumar, S.K., Ulirsch, J.C., Zekavat, S.M., Buenrostro, J.D., Natarajan, P., Salem, R.M., Chiarle, R., Mitt, M., Kals, M., et al. (2017). Comprehensive population-based genome sequencing provides insight into hematopoietic regulatory mechanisms. Proc. Natl. Acad. Sci. USA *114*, E327–E336.
75. Avellino, R., and Delwel, R. (2017). Expression and regulation of C/EBPα in normal myelopoiesis and in malignant transformation. Blood *129*, 2083–2091.
76. Perrotti, D., Cesi, V., Trotta, R., Guerzoni, C., Santilli, G., Campbell, K., Iervolino, A., Condorelli, F., Gambacorti-Passerini, C., Caligiuri, M.A., and Calabretta, B. (2002). BCR-ABL suppresses C/EBPalpha expression through inhibitory action of hnRNP E2. Nat. Genet. *30*, 48–58.
77. Radomska, H.S., Bassères, D.S., Zheng, R., Zhang, P., Dayaram, T., Yamamoto, Y., Sternberg, D.W., Lokker, N., Giese, N.A., Bohlander, S.K., et al. (2006). Block of C/EBP alpha function by phosphorylation in acute myeloid leukemia with FLT3 activating mutations. J. Exp. Med. *203*, 371–381.
78. Bolouri, H., Farrar, J.E., Triche, T., Jr., Ries, R.E., Lim, E.L., Alonzo, T.A., Ma, Y., Moore, R., Mungall, A.J., Marra, M.A., et al. (2018). The molecular landscape of pediatric acute myeloid leukemia reveals recurrent structural alterations and age-specific mutational interactions. Nat. Med. *24*, 103–112.
79. Goossens, S., Wang, J., Tremblay, C.S., De Medts, J., T'Sas, S., Nguyen, T., Saw, J., Haigh, K., Curtis, D.J., Van Vlierberghe, P., et al. (2019). ZEB2 and LMO2 drive immature T-cell lymphoblastic leukemia via distinct oncogenic mechanisms. Haematologica *104*, 1608–1616.
80. Vuckovic, D., Bao, E.L., Akbari, P., Lareau, C.A., Mousas, A., Jiang, T., Chen, M.H., Raffield, L.M., Tardaguila, M., Huffman, J.E., et al.; VA Million Veteran Program (2020). The Polygenic and Monogenic Basis of Blood Traits and Diseases. Cell *182*, 1214–1231.e11.
81. Murakami, N., Okuno, Y., Yoshida, K., Shiraishi, Y., Nagae, G., Suzuki, K., Narita, A., Sakaguchi, H., Kawashima, N., Wang, X., et al. (2018). Integrated molecular profiling of juvenile myelomonocytic leukemia. Blood *131*, 1576–1586.
82. Osumi, T., Tsujimoto, S.I., Tamura, M., Uchiyama, M., Nakabayashi, K., Okamura, K., Yoshida, M., Tomizawa, D., Watanabe, A., Takahashi, H., et al. (2018). Recurrent *RARB* Translocations in Acute Promyelocytic Leukemia Lacking *RARA* Translocation. Cancer Res. *78*, 4452–4458.
83. Campregher, P.V., Halley, N.D.S., Vieira, G.A., Fernandes, J.F., Velloso, E.D.R.P., Ali, S., Mughal, T., Miller, V., Mangueira, C.L.P., Odone, V., and Hamerschlak, N. (2017). Identification of a novel fusion TBL1XR1-PDGFRB in a patient with acute myeloid leukemia harboring the DEK-NUP214 fusion and clinical response to dasatinib. Leuk. Lymphoma *58*, 2969–2972.