Deciphering genetic susceptibility to multiple myeloma

MOLLY WENT

Division of Genetics and Epidemiology
Institute of Cancer Research
SM2 5NG

Submitted for the degree of Doctor of Philosophy in accordance with the regulations of the University of London
2020

Declaration

The work presented in this thesis is entirely my own work, except where clearly stated in the 'Statement of independent work attributable to candidate' on page 9, and has not been submitted for a degree or comparable award to this or any other university or institution.

Abstract

Multiple myeloma (MM) is a malignancy characterised by the clonal expansion of plasma cells primarily from the bone marrow. The two- to four-fold increased risk observed in relatives of MM patients provides support for inherited susceptibility to the disease. Genome-wide association studies (GWAS) have implicated common, low penetrance variants in MM susceptibility, however much of the heritability remains unexplained. To search for novel risk loci, a new GWAS and a meta-analysis with previous GWAS and a replication series, totalling 9,974 MM cases and 247,556 controls of European ancestry was performed. These data provide evidence for six new MM risk loci, bringing the total number to 23. Information from gene expression, regulatory profiling and *in situ* Hi-C data was integrated for the 23 risk loci. Collectively these data implicate disruption of developmental transcriptional regulators as a basis of MM susceptibility, compatible with altered B-cell differentiation, dysregulation of autophagy/apoptosis and cell cycle signalling as key mechanisms. To identify candidate causal genes at GWAS loci and search for novel risk regions, a multi-tissue transcriptome-wide association study (TWAS) was performed by integrating GWAS data with Genotype-Tissue Expression Project (GTEx) data assayed in 48 tissues. 108 genes at 13 independent regions associated with MM risk were identified, all of which were within 1 Mb of known MM GWAS risk variants. Of these, 94 genes, located in eight regions, had not previously been considered as a candidate gene for that locus. Clustering of chronic lymphocytic leukaemia (CLL) and MM is observed in families, suggesting an element of shared inherited susceptibility. To examine this, cross-trait linkage disequilibrium (LD)-score regression of MM and CLL GWAS data sets was performed. A significant genetic correlation between these two B-cell malignancies was shown (Rg= 0.4, P= 0.0046). Furthermore, nine loci pleiotropic to MM and CLL were identified and integration of regulatory and expression data demonstrated that these pleiotropic risk loci were enriched for B-cell regulatory elements, and implicated B-cell developmental genes. No lifestyle or environmental exposures have been consistently linked to an increased risk of MM. Summary data from GWAS of multiple phenotypes can be exploited in a Mendelian randomisation (MR) phenome-wide association study (PheWAS) to search for factors influencing MM risk. An MR-PheWAS was performed analysing 249 phenotypes, proxied by 10,225 genetic variants, and summary GWAS data. Although no significant associations with MM risk were observed among the 249 phenotypes, 28 phenotypes showed evidence suggestive of association, including increased levels of serum vitamin B6 and blood carnitine (P= 1.1×10^{-3}) with greater MM risk, and

increased levels of total cholesterol, blood esterified cholesterol and omega-3 fatty acids (P= 5.4×10^{-4}) with reduced MM risk. Collectively these findings provide insight into genetic and genomic architecture, as well as the aetiology of MM.

Acknowledgements

First, I would like to thank my supervisor Professor Richard Houlston for providing me the opportunity to pursue this PhD, for your support and insight throughout and for expanding my knowledge. I have valued your guidance, both professionally and personally.

Thank you to the Molecular and Population Genetics Team, not only for your support and encouragement in learning, but also for the fun times we have shared and for making my degree an enjoyable and memorable experience. Thank you to the Myeloma Group, with special mention to Sherbs, David, Fabio and Martin for your guidance and lovely conversations. Thank you also to my friends and colleagues across the ICR.

I would like to thank The Institute of Cancer Research, Myeloma UK and Mr Ralph Stockwell for funding my PhD. To collaborators on the projects, thank you all for your help and support.

With huge thanks and lots of love to my parents and my sisters, Lauren and Connie, for your unwavering support throughout my education.

Finally, my thanks go to the patients and relatives, without whom these studies could not have happened.

In memory of Mr Ralph Stockwell, who died in last year.

Publications

Papers published either as a direct result from or through collaborative work during this thesis:

Went M, Sud A, Law PJ, Johnson DC, Weinhold N, Försti A, van Duin M, Mitchell JS, Chen B, Kuiper R, Stephens OW, Bertsch U, Campo C, Einsele H, Gregory WM, Henrion M, Hillengass J, Hoffmann P, Jackson GH, Lenive O, Nickel J, Nöthen MM, da Silva Filho MI, Thomsen H, Walker BA, Broyl A, Davies FE, Langer C, Hansson M, Kaiser M, Sonneveld P, Goldschmidt H, Hemminki K, Nilsson B, Morgan GJ, Houlston RS. Assessing the effect of obesity-related traits on multiple myeloma using a Mendelian randomisation approach. *Blood Cancer Journal*. 2017;7(6):e573-e573.

Li N, Johnson DC, Weinhold N, Kimber S, Dobbins SE, Mitchell JS, Kinnersley B, Sud A, Law PJ, Orlando G, Scales M, Wardell CP, Forsti A, Hoang PH, **Went M**, Holroyd A, Hariri F, Pastinen T, Meissner T, Goldschmidt H, Hemminki K, Morgan GJ, Kaiser M, Houlston RS. Genetic Predisposition to Multiple Myeloma at 5q15 Is Mediated by an ELL2 Enhancer Polymorphism. *Cell Reports*. 2017;20(11):2556-2564.

Went M, Sud A, Försti A, Halvarsson B-M, Weinhold N, Kimber S, van Duin M, Thorleifsson G, Holroyd A, Johnson DC, Li N, Orlando G, Law PJ, Ali M, Chen B, Mitchell JS, Gudbjartsson DF, Kuiper R, Stephens OW, Bertsch U, Broderick P, Campo C, Bandapalli OR, Einsele H, Gregory WA, Gullberg U, Hillengass J, Hoffmann P, Jackson GH, Jöckel K-H, Johnsson E, Kristinsson SY, Mellqvist U-H, Nahi H, Easton D, Pharoah P, Dunning A, Peto J, Canzian F, Swerdlow A, Eeles RA, Kote-Jarai Z, Muir K, Pashayan N, Nickel J, Nöthen MM, Rafnar T, Ross FM, da Silva Filho MI, Thomsen H, Turesson I, Vangsted A, Andersen NF, Waage A, Walker BA, Wihlborg A-K, Broyl A, Davies FE, Thorsteinsdottir U, Langer C, Hansson M, Goldschmidt H, Kaiser M, Sonneveld P, Stefansson K, Morgan GJ, Hemminki K, Nilsson B, Houlston RS. Identification of multiple risk loci and regulatory mechanisms influencing susceptibility to multiple myeloma. *Nature Communications*. 2018;9(1):3707.

Went M, Sud A, Speedy H, Sunter NJ, Försti A, Law PJ, Johnson DC, Mirabella F, Holroyd A, Li N, Orlando G, Weinhold N, van Duin M, Chen B, Mitchell JS, Mansouri L, Juliusson G, Smedby KE, Jayne S, Majid A, Dearden C, Allsup DJ, Bailey JR, Pratt G, Pepper C, Fegan C, Rosenquist R, Kuiper

R, Stephens OW, Bertsch U, Broderick P, Einsele H, Gregory WM, Hillengass J, Hoffmann P, Jackson GH, Jöckel K-H, Nickel J, Nöthen MM, da Silva Filho MI, Thomsen H, Walker BA, Broyl A, Davies FE, Hansson M, Goldschmidt H, Dyer MJS, Kaiser M, Sonneveld P, Morgan GJ, Hemminki K, Nilsson B, Catovsky D, Allan JM, Houlston RS. Genetic correlation between multiple myeloma and chronic lymphocytic leukaemia provides evidence for shared aetiology. *Blood Cancer Journal*. 2018;9(1):1.

Went M, Sud A, Li N, Johnson DC, Mitchell JS, Kaiser M, Houlston RS. Regions of homozygosity as risk factors for multiple myeloma. *Annals of Human Genetics*. 2019;83(4):231-238.

Labreche K, Daniau M, Sud A, Law PJ, Royer-Perron L, Holroyd A, Broderick P, **Went M**, Benazra M, Ahle G, Soubeyran P, Taillandier L, Chinot OL, Casasnovas O, Bay J-O, Jardin F, Oberic L, Fabbro M, Damaj G, Brion A, Mokhtari K, Philippe C, Sanson M, Houillier C, Soussain C, Hoang-Xuan K, Houlston RS, Alentorn A, Network LOC. A genome-wide association study identifies susceptibility loci for primary central nervous system lymphoma at 6p25.3 and 3p22.1: a LOC network study. *Neuro-oncology*. 2019;21(8):1039-1048.

Went M, Kinnersley B, Sud A, Johnson DC, Weinhold N, Försti A, van Duin M, Orlando G, Mitchell JS, Kuiper R, Walker BA, Gregory WM, Hoffmann P, Jackson GH, Nöthen MM, da Silva Filho MI, Thomsen H, Broyl A, Davies FE, Thorsteinsdottir U, Hansson M, Kaiser M, Sonneveld P, Goldschmidt H, Stefansson K, Hemminki K, Nilsson B, Morgan GJ, Houlston RS. Transcriptome-wide association study of multiple myeloma identifies candidate susceptibility genes. *Human Genomics*. 2019;13(1):37-37.

Schmidt AF, Holmes MV, Preiss D, Swerdlow DI, Denaxas S, Fatemifar G, Faraway R, Finan C, Valentine D, Fairhurst-Hunter Z, Hartwig FP, Horta BL, Hypponen E, Power C, Moldovan M, van Iperen E, Hovingh K, Demuth I, Norman K, Steinhagen-Thiessen E, Demuth J, Bertram L, Lill CM, Coassin S, Willeit J, Kiechl S, Willeit K, Mason D, Wright J, Morris R, Wanamethee G, Whincup P, Ben-Shlomo Y, McLachlan S, Price JF, Kivimaki M, Welch C, Sanchez-Galvez A, Marques-Vidal P, Nicolaides A, Panayiotou AG, Onland-Moret NC, van der Schouw YT, Matullo G, Fiorito G, Guarrera S, Sacerdote C, Wareham NJ, Langenberg C, Scott RA, Luan Ja, Bobak M, Malyutina S, Pająk A, Kubinova R, Tamosiunas A, Pikhart H, Grarup N, Pedersen O, Hansen T, Linneberg A, Jess T, Cooper J, Humphries SE, Brilliant M, Kitchner T, Hakonarson H, Carrell DS, McCarty CA, Lester KH, Larson EB, Crosslin DR, de Andrade M, Roden DM, Denny JC, Carty C, Hancock S, Attia

J, Holliday E, Scott R, Schofield P, O'Donnell M, Yusuf S, Chong M, Pare G, van der Harst P, Said MA, Eppinga RN, Verweij N, Snieder H, Lifelines Cohort a, Christen T, Mook-Kanamori DO, Consortium I, Gustafsson S, Lind L, Ingelsson E, Pazoki R, Franco O, Hofman A, Uitterlinden A, Dehghan A, Teumer A, Baumeister S, Dörr M, Lerch MM, Völker U, Völzke H, Ward J, Pell JP, Meade T, Christophersen IE, Maitland-van der Zee AH, Baranova EV, Young R, Ford I, Campbell A, Padmanabhan S, Bots ML, Grobbee DE, Froguel P, Thuillier D, Roussel R, Bonnefond A, Cariou B, Smart M, Bao Y, Kumari M, Mahajan A, Hopewell JC, Seshadri S, ISGC MCot, Dale C, Costa RPE, Ridker PM, Chasman DI, Reiner AP, Ritchie MD, Lange LA, Cornish AJ, Dobbins SE, Hemminki K, Kinnersley B, Sanson M, Labreche K, Simon M, Bondy M, Law P, Speedy H, Allan J, Li N, **Went M**, Weinhold N, Morgan G, Sonneveld P, Nilsson B, Goldschmidt H, Sud A, Engert A, Hansson M, Hemingway H, Asselbergs FW, Patel RS, Keating BJ, Sattar N, Houlston R, Casas JP, Hingorani AD. Phenome-wide association analysis of LDL-cholesterol lowering genetic variants in PCSK9. *BMC Cardiovascular Disorders*. 2019;19(1):240-240.

Pertesi M, **Went M**, Hansson M, Hemminki K, Houlston RS, Nilsson B. Genetic predisposition for multiple myeloma. *Leukemia*. 2020;34(3):697-708.

Went M, Cornish AJ, Law P, Kinnersley B, Van Duin M, Weinhold N, Forsti A, Hansson M, Sonneveld P, Goldschmidt H, Morgan G, Hemminki K, Nilsson B, Kaiser M, Houlston R. Search for multiple myeloma risk factors using Mendelian randomization. *Blood Advances*. 2020; 4 (10): 2172–2179.

Statement of independent work attributable to the candidate

Chapter 1

This chapter is entirely my own work.

Chapter 2

This chapter is entirely my own.

Chapter 3

Bioinformatic work was supervised by Amit Sud (ICR) and Philip Law (ICR). Sample collection, array genotyping and quality control of the four GWAS datasets were carried out as described in their respective publications. David Johnson (ICR) managed and prepared the Medical Research Council (MRC) Myeloma IX and XI Case Study samples, ascertained and collected by Richard Houlston, Martin Kaiser (both ICR), Gareth Morgan (University of Arkansas for Medical Sciences (UAMS)), Faith Davies (UMAS), Walter Gregory (University of Leeds) and Graham Jackson (Royal Victoria Infirmary). Amy Holroyd supervised genotyping, including optimisation in the UK. Asta Försti, Obul R Bandapalli and Chiara Campo (all German Cancer Research Centre) coordinated and performed the German replication genotyping. eQTL analyses were performed by Niels Weinhold (University of Heidelberg and UAMS). Björn Nilsson (Department of Laboratory Medicine, Lund) coordinated the Swedish/Norwegian replication genotyping. *In situ* promoter capture Hi-C experiment was conducted by Scott Kimber (ICR).

Chapter 4

Bioinformatic work was supervised by Amit Sud and Ben Kinnersley.

Chapter 5

This chapter was entirely my own.

Chapter 6

Pipeline to run MR base was written by Alex Cornish.

Chapter 7

This chapter is entirely my own work.

Table of contents

List of abbreviations

A	Adenine
AA	African ancestry
ADME	absorption, distribution, metabolism, and excretion
AID	Activation-induced deaminase
ALSPAC	The Avon Longitudinal Study of Parents and Children
APC	Adenomatous polyposis coli
APEX	The Assessment of Proteasome Inhibition for Extending Remissions
APOBEC	apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like
ARSA	arylsulfatase A
ASB	ankyrin repeat and SOCS box-containing proteins
ASCT	Autologous Stem Cell Transplant
ATP	Adenosine 5'-triphosphate
AVS	Associated variant sets
BCAC	Breast Cancer Association Consortium
BFDP	Bayesian false-discovery probability
BLAST	Basic Local Alignment Search Tool
BLAT	Basic Local Alignment Tool
B-PROOF	The B-vitamins for the Prevention of Osteoporotic Fracture
BR	Broad range
BWA	Burrows-Wheeler Aligner
C	Cytosine
CAGE	Cap analysis gene expression
CENPO	Centromere Protein O
CEU	Utah residents with Northern and Western European ancestry from the CEPH collection
CGEMS	The Cancer Genetic Markers of Susceptibility Study
CHB	Han Chinese in Beijing, China
CI	Confidence interval
CLL	Chronic lymphocytic leukaemia
CRC	Colorectal cancer
CSR	Class switch recombination
CT	Computed tomography
CTD	Cyclophosphamide, thalidomide and dexamethasone
DNA	Deoxyribonucleic acid
DSMM	Deutsche Studiengruppe Multiples Myeloma
DSS	Durie-Salmon Staging System
DTNB	β-dystrobrevin
EBV	Epstein–Barr virus
EDTA	Ethylenediaminetetraacetic acid
EMM	Extramedullary myeloma
EMN	European Myeloma Network
ENCODE	The Encyclopaedia of DNA Elements
eQTL	Expression quantitative trait locus

EUR	European
FA	Fatty acid
FAO	Fatty acid oxidation
FDR	False-discovery rate
FISH	Fluorescence *in situ* hybridisation
FLC	Free light chain
FWER	Family-wise error rate
G	Guanine
GABAA	Gamma-aminobutyric acid type A
GC	Germinal centre
GCTA	Genome-wide complex trait analysis
GELCAPS	The Genetic Lung Cancer Predisposition Study
GEO	Gene expression omnibus
GMMG	German-Speaking Multiple Myeloma Multicenter Study Group
GWAS	Genome-wide association study
H_0	Null hypothesis
H_1	Alternative hypothesis
H3K27Ac	Histone H3 lysine-27 acetylation
H3K27me3	Histone H3 lysine-27 trimethylation
H3K4Me1	Histone H3 lysine-4 monomethylation
H3K4Me3	Histone H3 lysine-4 trimethylation
HEIDI	Heterogeneity in dependent instruments
HICUP	Hi-C User Pipeline
HLA	human leukocyte antigen
HMGXB4	HMG-box containing 4
HMM	Hidden Markov Model
HNR	The Heinz Nixdorf Recall
HOVON	Stichting Hemato-Oncologie voor Volwassenen Nederland
HR	Hazard ratio
HRD	Hyperdiploidy
HS	High sensitivity
HWE	Hardy-Weinberg Equilibrium
I^2	I-squared statistic
IBS	Identity-by-state
ICD	International Classification of Diseases
Ig	Immunoglobulin
IGH	Heavy chain locus
IL	Interleukin
IMWG	International Myeloma Working Group
ISS	International Staging System
IV	Instrumental variable
IVW-FE	Inverse variance weighted-fixed effects
IVW-RE	Inverse variance weighted-random effects
JPT	Japanese in Tokyo, Japan
LD	Linkage disequilibrium
LDAK	Linkage disequilibrium adjusted kinships
MAF	Minor allele frequency

MAFB	V-maf musculoaponeurotic fibrosarcoma oncogene homolog B
MAPK	Mitogen-activated protein kinase
Mb	Megabase
MBE	Mode-based estimate
MGUS	Monoclonal gammopathy of undetermined significance
MM	Multiple myeloma
MPRA	Massively parallel reporter assay
MR	Mendelian randomisation
MRC	Medical Research Council
MRI	Magnetic resonance imaging
MTAP	methylthioadenosine phosphorylase
mTOR	Mechanistic target of rapamycin
MYC	Avian myelocytomatosis viral oncogene homolog
MyIX	Myeloma IX clinical trial
MYNN	Myoneurin
MyXI	Myelomw XI clinical trial
MZ	Marginal-zone
NADH	nicotinamide adenine dinucleotide (NAD) + hydrogen (H)
NADPH	Nicotinamide adenine dinucleotide phosphate
NCBI	The National Centre for Biotechnology Information
NCI	National Cancer Institute
NF-κB	Nuclear factor kappa-light-chain-enhancer of activated B-cells
NIH	National Institutes of Health
NMSG	Nordic Myeloma Study Group
NRAS	Neuroblastoma RAS viral oncogene homolog
NSCCG	The National Study of Colorectal Cancer Genetics
OR	Odds ratio
PAD	Bortezomib, doxorubicin and dexamethasone
PCA	Principal components analysis
PCL	Plasma cell leukaemia
PCR	Polymerase chain reaction
PEMT	Phosphatidylethanolamine N-Methyltransferase
PET-CT	Positron emission tomography–computed tomography
PheWAS	Phenome-wide association study
PLCO	Prostate, Lung, Colon, Ovary Screen Trial
POU5F1	POU domain, class 5, transcription factor 1
PRACTICAL	Prostate Cancer Association Group to Investigate Cancer Associated Alterations in the Genome
PRS	Polygenic risk score
PVE	Proportion of variance explained
QC	Quality control
QPRT	Quinolinate Phosphoribosyltransferase
Q-Q plot	Quantile-quantile plot
r^2	Correlation coefficient
RISS	Revised International Staging System
RNA	Ribonucleic acid
RR	Relative risk
SAP	Shrimp alkaline phosphatase

SD	Standard deviation
SHAPEIT	Segmented haplotype estimation and imputation tool
SHM	Somatic hypermutation
SIFT	Sorting intolerant from tolerant algorithm
SMM	Smouldering multiple myeloma
SMR	Summary data-based Mendelian randomisation
SNAP	SNP annotation and proxy search
SNP	Single nucleotide polymorphism
STARR-seq	self-transcribing active regulatory region sequencing
T	Thymine
TAE	Tris-acetate-EDTA
TE	Tris-EDTA
TERC	Telomerase RNA Component
TERT	Telomerase Reverse Transcriptase
TF	Transcription factor
TIGAR	TP53-induced glycolysis
TNF	Tumor Necrosis Factor
TOM1	Target of myb1 membrane trafficking protein
TP53	Tumour protein 53
TT2	Total Therapy 2
TT3	Total Therapy 3
TWAS	Transcriptome-wide association study
TYMP	Thymidine Phosphorylase
UAMS	University of Arkansas for Medical Sciences
UCSC	The University of California, Santa Cruz
UK	United Kingdom
UKBS	UK National Blood Service
UKGPCS	UK Genetic Prostate Cancer Study
ULK	Unc-51 Like Autophagy Activating Kinase
US	United States
UTR	Untranslated region
VCF	Variant call format
WAC	WW Domain Containing Adaptor With Coiled-Coil
WES	Whole exome sequencing
WGS	Whole genome sequencing
WME	Weighted median estimator
WTCCC	The Wellcome Trust Case Control Consortium
YRI	Yoruba in Ibadan, Nigeria
µg	Microgram
µl	Microlitre
µM	Micromolar
$\chi 2$	Chi-squared

List of figures

List of tables

CHAPTER 1 Introduction

1.1 Overview of multiple myeloma

Multiple myeloma (MM) is a haematological malignancy with an annual incidence of 8.5 individuals per 100,000 in the UK [1-3]. The disease is caused by the clonal expansion of plasma cells infiltrating the bone marrow [4]. Plasma cells result from the terminal differentiation of B-cells and are the mediators of long-term humoral immunity, producing and releasing antibody [5]. Patients typically present with monoclonal immunoglobulin protein, produced by the aberrant plasma cells, in serum and urine [6]. Despite improvements in therapy, MM essentially remains an incurable disease; in patients under the age of 60, 10-year survival is only around 30% [7].

1.1.1 Epidemiology of multiple myeloma

The global burden of MM has significantly increased over the last 30 years [8]. Incidence of the disease is, however highly variable between different countries, with MM being more common in economically developed countries (**Figure 1.1**) though some of the differences in incidence may be due to lack of diagnostic abilities in less economically developed countries compared with more economically developed countries and do not necessarily reflect differences in disease biology [8, 9]. The incidence of MM increases with age with the median age of diagnosis for MM being 70 years [10]. As with other B-cell malignancies MM is more common in men than in women [11-13]. Both MM and its precursor lesion, monoclonal gammopathy of undetermined significance (MGUS), have a higher incidence in those with African than Caucasian ancestry [13-15] (**Figure 1.2**).

Multiple lifestyle and dietary factors have been variously purported to affect the risk of MM or MGUS, including obesity [16-20], diet [21-23], vitamin D [24, 25] and immune dysfunction [26]. Environmental factors proposed to influence MM risk include herbicide [27] or pesticide [28] exposure, occupation as a farmer or firefighter [29, 30], exposure to radiation [31, 32] and exposure to industrial solvent methylene chloride [33]. To date none of these findings have been independently validated and the aetiological basis of MM largely remains unexplained.

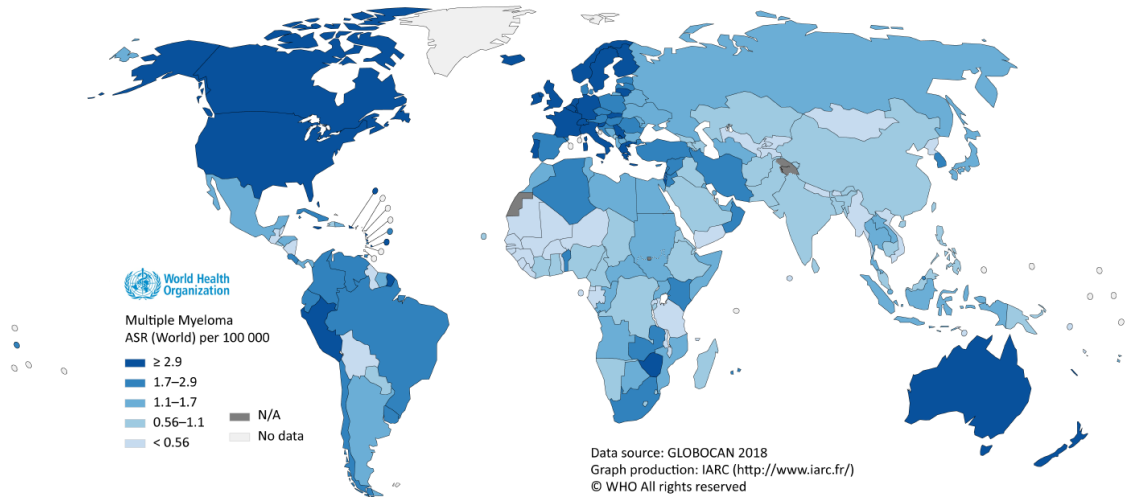


Figure 1.1 World age-standardised incidence rates of multiple myeloma in 2018. GLOBOCAN [34]

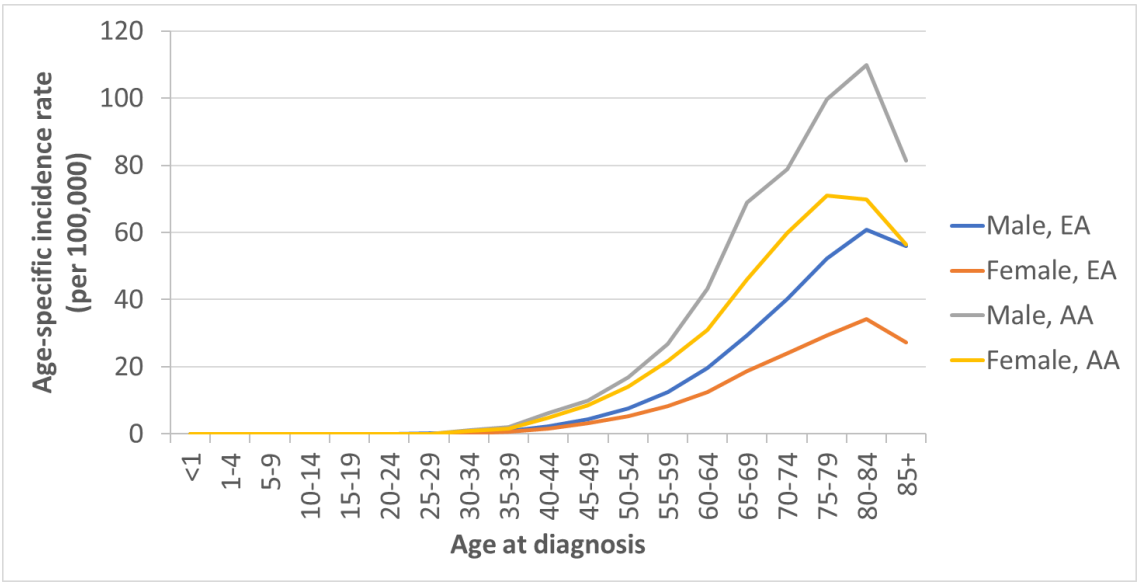


Figure 1.2 Age-specific incidence rate of multiple myeloma per 100,000 by race and sex. Incidence data from the SEER 18 Registries 2007-2011 [35]. EA, European ancestry; AA, African ancestry.

1.1.2 Cellular origin of multiple myeloma

Plasma cells are terminally differentiated cells of B-cell lineage, which develop from haematopoietic stem cells in the bone marrow. Commitment to the B-cell lineage is dependent on transcription factors including PU.1, E2A and paired box protein 5 (PAX5) [36].

Naïve B-cells that exit the bone marrow continue to undergo maturation in the spleen. Here they pass through transitional stages, to form long-lived naïve follicular B-cells with a minority also forming naïve marginal-zone (MZ) B-cells [36]. B1 cells, present in the peritoneal and pleural

cavities of the gut lamina propria, represent another type of mature naïve B-cell [37]. Follicular B, marginal zone B and B1 cells all possess antigen-independent self-renewing ability [5]. B1-cells develop into antibody-secreting cells (ASCs) when challenged with antigens, often from bacterial pathogens or viruses, and form part of the innate immune system [38] (**Figure 1.3**). Similarly, MZ B-cells contribute to the innate immunity by differentiating into ASCs upon exposure to polymeric epitopes of bacteria or viruses [38] (**Figure 1.3**). ASCs developed from B1-cells and MZ B-cells are normally short-lived. Follicular B-cells, as the most abundant mature B-cell subset, can generate ASCs in an early response like B1-cells and MZ B-cells when they encounter foreign antigens. With T-cell involvement follicular B-cells can also undergo a clonal expansion to form a germinal centre (GC) within secondary lymphoid organs (**Figure 1.4**) [5, 36].

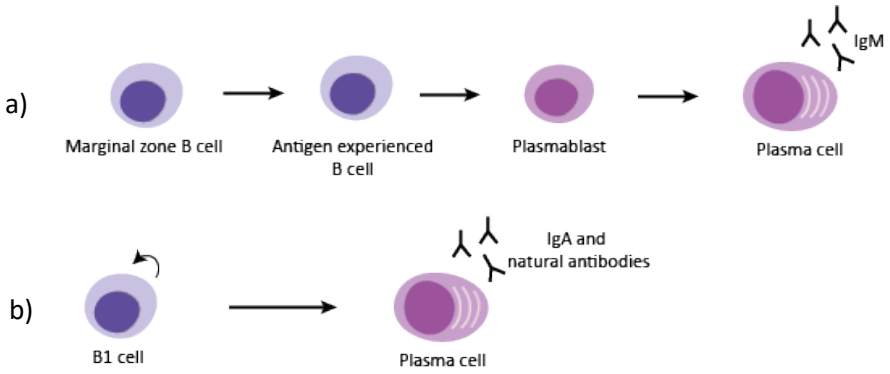


Figure 1.3 Key steps in normal B-cell differentiation formation of a) marginal zone B and b) B1 cells.

The GC is composed of the light and dark zones. Within the dark-zone, activated B-cells undergo somatic hypermutation (SHM) of the immunoglobulin genes to generate diverse antibodies. Dark-zone B-cells differentiate into light-zone B-cells, where those cells expressing a high-affinity antibody either recirculate to the dark zone to undergo further rounds of SHM or differentiate further. During development, B-cells also undergo class switch recombination (CSR). Class switching, the result of CSR, changes the expressed isotype from low-affinity immunoglobulin M (IgM), which characterises an antigen experienced B-cell, to IgG-, IgA-, or IgE-, generating specific antibodies with different functional characteristics [39]. CSR was thought to occur in GCs however recent research revises this previously held assumption, suggesting that CSR can occur prior to SHM in a pre-GC reaction [40]. B-cells that bear high-affinity antibodies of various isotypes, as a result of these selection processes, differentiate into memory B-cells or ASCs, with some plasma cells becoming long-lived and residing in the bone marrow to provide long-lived antibody response [36, 41]. Upon antigen rechallenge, the memory B-cells can differentiate into

plasma cells rapidly or form secondary germinal centre to generate higher-affinity antibodies [42]. The plasma cells associated with MM are by convention termed post-germinal centre B-cells since they have undergone immunoglobulin gene SHM, VDJ recombination (where exons encoding the antigen binding domains are assembled from Variable, Diversity and Joining gene segments), antigen selection and (usually) isotype switch recombination [43].

Plasma cells in the bone marrow can undergo a clonal expansion to form MGUS, the asymptomatic precursor lesion to MM. Progression of MGUS to MM occurs at a rate of 1% per year [44]. Smouldering MM (SMM) is an intermediary state between MGUS and MM, with annual risk of 10% in first five years of progressing to MM, 3% per year in the subsequent five years and 1% per year thereafter [45]. Symptomatic MM is typified by the presence of monoclonal protein (M protein) in the blood or urine produced by the clonally-expanded plasma cells, as well as the associated organ dysfunction [46]. Clonal plasma cells can progress into plasma cell leukaemia (PCL) or extramedullary myeloma (EMM), migrating outside the bone marrow to the peripheral blood. Genetic aberrations, which have been characterised in MM, are considered to disrupt the intrinsic biological pathways of B-cells and plasma cells resulting in the initiation and development of MM [47].

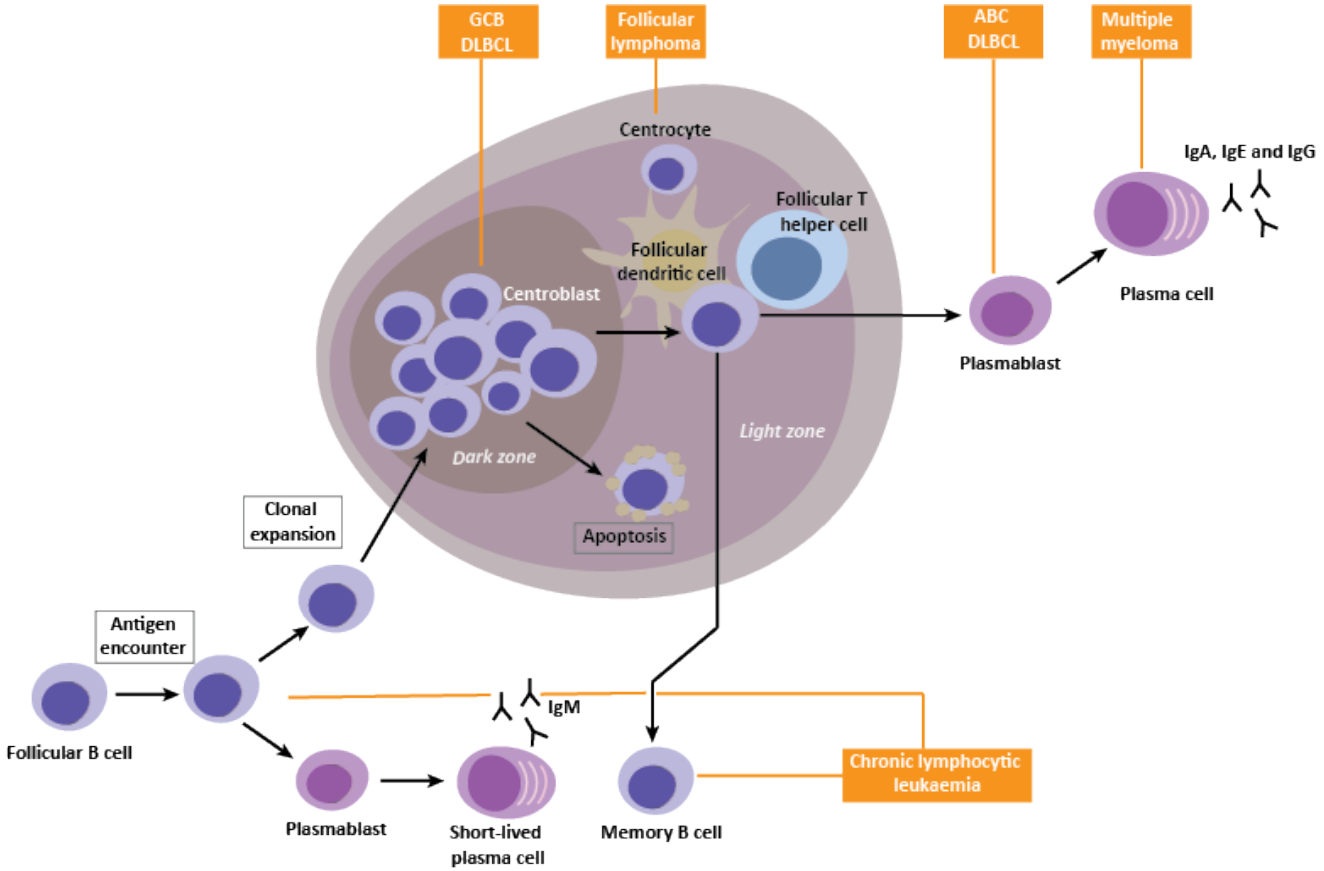


Figure 1.4 Germinal centre reaction. Upon antigen stimulation, mature naïve follicular B-cells undergo clonal expansion in GCs. This is followed by somatic hypermutation, with B-cells bearing the highest affinity antibodies being preferentially selected. B-cells expressing high-antigen-affinity antibodies that have survived the GC reaction differentiate into long-lived memory B-cells, antibody-secreting plasmablasts or plasma cells. Short-lived antibody-secreting plasmablasts and plasma cells can also develop from mature naïve marginal-zone B-cells and B1-cells. B-cell malignancies can arise at various stages of B-cell development. Putative cell of origin of the various B-cell malignancies are indicated, including MM, which is thought to arise from a terminally differentiated plasma cell.

1.1.3 The multiple myeloma genome

Multiple myeloma is a biologically heterogenous disease. Accumulation of genetic abnormalities, including hyperdiploidy (HRD), chromosomal translocations, copy number changes, gene mutations, aberrant methylation and microRNA deregulation [47] characterise the initiation and progression of MM (**Figure 1.5**).

Primary genetic events associated with the development of the MM precursor states are chromosomal translocations (non-HRD) and hyperdiploidy (HRD). HRD is present in 55-60% of MM patients, involving trisomies of odd numbered chromosomes - specifically chromosomes 3, 5, 7, 9, 11, 15, 19 and 21 [48-50]. Non-HRD can be subdivided based on translocations of the IGH locus at 14q32; this process is thought to be the consequence of aberrant CSR during antigen stimulated B-cell proliferation [39]. Normal B-cells undergo CSR to alter the antibody class expressed, mediated by double-strand DNA breaks (DSBs) with the expression of activation-induced deaminase (AID) [43, 51, 52]. Successful recombination results in a B-cell which produces a functional heavy chain in its secreted immunoglobulin. Errors in the process of CSR can result in DNA from another chromosome being translocated and juxtaposed with strong IGH enhancer on chromosome 14 [43]. Recurrent chromosomal translocations observed in MM are summarised in **Table 1.1**.

Primary translocation	Frequency	Translocated gene partner
t(11;14)	15-20%	*CCND1*
t(4;14)	10-15%	*FGFR3, MMSET*
t(6;14)	2-5%	*CCND3*
t(14;16)	5%	*c-MAF*
t(14;20)	1-2%	*MAFB*

Table 1.1 The main primary chromosomal translocations in multiple myeloma.

Secondary chromosomal events in MM include deletion of 1p (30%), 6q (33%), 8p (25%), 12p (15%), 13q (59%), 14q (39%), 16q (35%), 17p (7%), 20 (12%) and 22 (18%) and gain of 1q (40%) [53]. A number of these changes have clinical relevance, with del(17p), which involves loss of *TP53*, and *MYC* translocation being associated with progression from newly diagnosed MM to refractory disease and plasma cell leukaemia [54-57]. Gain of 1q21, which implicates the oncogene *CKS1B*, has been shown to be strongly associated with adverse patient prognosis [2, 58, 59]. In addition to copy number changes, gene mutations in *RAS/MAPK* signalling pathway (e.g. *NRAS* and *KRAS*), aberrant methylation and microRNA deregulation are all features of MM [47].

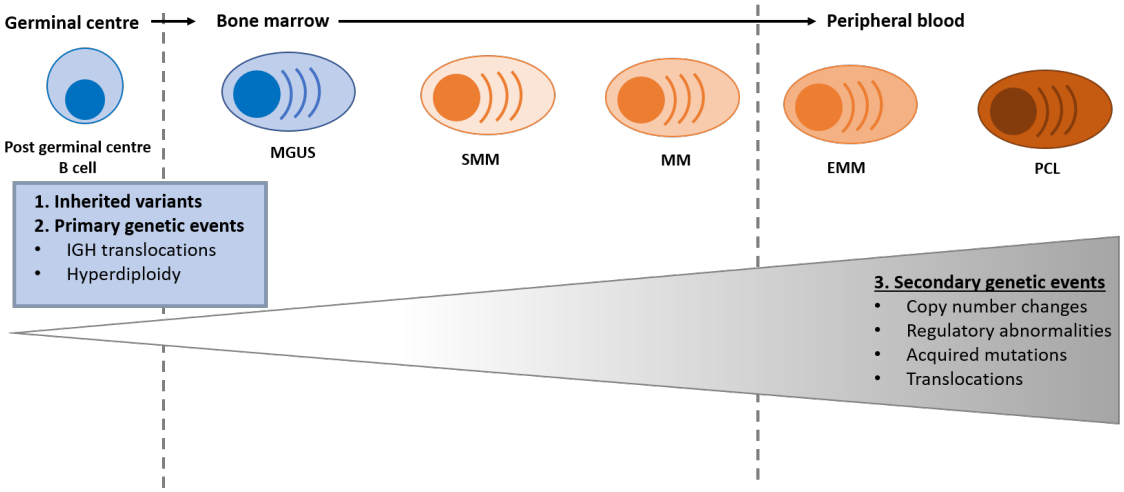


Figure 1.5 Pathogenesis of multiple myeloma. The initial deregulated plasma cell in the bone marrow belongs to MGUS, which develops further genetic abnormalities in the progression to symptomatic MM, EMM/PCL. MGUS, monoclonal gammopathy of undetermined significance; SMM, smouldering multiple myeloma; MM, multiple myeloma; EMM, extramedullary multiple myeloma; PCL, plasma cell leukaemia. Adapted from [47].

1.1.4 Diagnostic classification of multiple myeloma

Diagnostic classification of MM was established by the International Myeloma Working Group (IMWG) (**Table 1.2**) [60]. Serum and urine M proteins (an abnormal immunoglobulin fragment) are measured from patients by electrophoresis and immunofixation. Presence of CRAB symptoms are also evaluated in patients, which assess Calcium levels (hypercalcemia; serum calcium > 2.75 mmol/L), Renal impairment (serum creatinine > 177 µmol/L, or creatinine clearance < 40 ml/min), Anaemia (haemoglobin level < 100 g/L) and Bone lesions (defined as ≥ 1 osteolytic lesions detected on skeletal radiography, computed tomography (CT) or positron emission tomography–computed tomography (PET-CT)) [60].

These diagnostic criteria have recently been revised to reflect changes in available therapy and improvements in accurate biomarker identification, with an aim to identify and treat those individuals at high risk of progression from SMM or MGUS to MM [60]. Specifically, patients with clonal bone marrow plasma cell percentage ≥60%, involved:uninvolved serum free light chain ratio ≥100, or >1 focal lesions on MRI, studies who do not yet show the presence of CRAB features, are now among those eligible for treatment.

Clinical stage	Diagnostic criteria
Monoclonal gammopathy of undetermined significance (MGUS)	• Serum M protein < 30 g/L, and • Clonal plasma cells < 10% in bone marrow, and • Absence of myeloma-related end-organ damage or tissue impairment or CRAB
Asymptomatic/smouldering multiple myeloma (SMM)	• Serum M protein level ≥ 30 g/L or urinary M protein ≥ 500mg per 24 hours, and/or clonal plasma cells 10%-60% in bone marrow, and • Absence of myeloma-defining events (*i.e.* no myeloma-related end-organ damage or tissue impairment or CRAB, involved: uninvolved serum free light chain ratio < 100, no focal lesions identified by magnetic resonance imagining (MRI))
Symptomatic MM	• Clonal plasma cells ≥ 10% in bone marrow or biopsy-proven bony or extramedullary plasmacytoma, and any one of the following: • Clonal plasma cells in bone marrow ≥ 60%, or • Involved: uninvolved serum free light chain ratio ≥ 100 (providing involved FLC ≥ 100mg/L) , or • Evidence of end-organ damage related to myeloma or CRAB, or • > 1 MRI focal lesion

Table 1.2 IMWG diagnostic criteria of multiple myeloma.

1.1.5 Prognostic factors

As a disease, MM is clinically heterogenous with patient survival being affected by host factors including tumour burden (stage), tumour biology (cytogenetic abnormalities), and response to therapy. Traditionally, survival in MM patients has been based on the Durie-Salmon Staging System (DSS) and International Staging System (ISS) [61, 62], but both these systems have limitations [63]. The DSS classifies patients based on tumour burden, but suffers from a lack of reproducibility due to varied interpretation of MM bone disease. While the ISS classification includes measurement of serum albumin and beta2-microglobulin and is generally considered to be more reproducible, these biomarkers can be disproportionately affected by factors that are not disease-specific [63].

Recognising that the molecular subtype and cytogenetic abnormalities in MM have prognostic relevance, a Revised International Staging System (RISS) was created that combines elements of tumour burden (ISS) and disease biology (presence of high-risk cytogenetic abnormalities (**Table 1.3**) or elevated lactate dehydrogenase level) [63]. Gene expression signatures, generated by unsupervised clustering of mRNA expression, are increasingly being used to risk stratify patients [64]. For example, *MAFB* and *c-MAF* overexpression as a consequence of t(14;20) and t(14;16)

respectively, cluster as one subgroup designated “MF”, on the assumption that over-expression of the MAF family results in deregulation of mutual downstream genes in MM [64]. Different molecular subgroups have demonstrated differences in event-free and overall survival, and mutational load have also been linked to a poorer outcome [64, 65]. Risk stratification, which combines the mutational load, molecular subtype classification and gene expression profiling is increasingly being used to define patient treatment, for example, in the Mayo Stratification of Myeloma and Risk-Adapted Therapy (mSMART) [66] and ongoing clinical trials Total Therapy 4 and 5 [67] conducted by the University of Arkansas.

Standard risk	Intermediate risk	High risk
HRD		t(14;16)
t(11;14)	t(4;14)	t(14;20)
t(6;14)		17p deletion

Table 1.3 Cytogenetic risk-stratification of multiple myeloma. Adapted from ref. 59.

1.1.6 Treatment of multiple myeloma

Asymptomatic precursors to MM, including MGUS and SMM, typically do not require treatment. However a subgroup of patients with SMM, who are at high risk of progression to MM, are now being considered for therapy [60]. Treatment for MM generally involves chemotherapy, with radiotherapy, as appropriate, for pathological bone lesions. Patients who are younger (usually < 70 years) without co-morbidities, are typically treated with high-dose therapy, followed by an autologous stem cell transplantation (ASCT), otherwise, chemotherapy only is used [68, 69].

Treatment with bortezomib (Velcade), lenalidomide (Revlimid) and dexamethasone in combination (VRd) is a standard course of therapy for MM, as induction prior to high-dose therapy and stem cell transplantation, as an initial treatment for older and less fit patients, or at relapse [70]. Bortezomib is one of a group of drugs called proteasome inhibitors. Others in this group include carfilozomib and ixazomib, which are also used in treatment of MM. Alternatives to lenalidomide, which is an immunomodulatory agent, are pomalidomide or thalidomide. Combination treatments rely on the synergistic effects of the therapy agents, including targeting the tumour microenvironment. For example, bortezomib targets pathways which are both intrinsic and extrinsic to the plasma cell [71], while dexamethasone targets only intrinsic pathways [72].

DNA damaging agents such as alkylating agent melphalan and cyclophosphamide, and anthracycline agents such as doxorubicin are also treatment options for MM, though stem cell toxins may be avoided in patients who will undergo ASCT [69, 70].

New treatments include repurposed alkylating agents, kinesin spindle protein inhibitors, histone deacetylase inhibitors, and inhibitors of key complexes in MM development and progression, namely cyclin-dependent kinase, IL-6, Bruton's tyrosine kinase, B-cell lymphoma 2, protein kinase B and phosphoinositide 3-kinase pathway components. Novel immunotherapies using monoclonal antibodies (e.g. daratumumab, elotuzumab, indatuximab, SAR650984) are also treatment options currently being investigated in clinical trials for those with relapsed/refractory MM [73]. Many of these treatments benefit from fewer, milder side effects than conventional treatment and thus may be advantageous for long-term management of patients [60].

Currently, MM is essentially an incurable disease and the majority of patients will relapse. Treatment for relapsed patients is informed by the quality and duration of response to previous

drug regimens, timings of relapse and patients health (*e.g.* age, renal function, bone marrow function, and presence of comorbidities). A regimen of formerly administered chemotherapy drugs or novel agents with or without stem cell transplantation is given at relapse.

1.2 Evidence for familial predisposition to multiple myeloma

Evidence for inherited predisposition to MM comes from increased risk of MM seen in relatives of MM patients [74]. The largest study to date investigated familial risk in a range of haematological malignancies, including 25,787 patients diagnosed with a MM. This study quantified familial relative risks (RRs) in 59,413 of the first-degree relatives of MM patients, finding a familial RR of 2.24 (95% CI 1.81-2.75) [74, 75]. This estimate is consistent with earlier studies using the same cancer registry which demonstrated a RR in first-degree relatives of 2.45 and 2.1. Clustering of MM with the precursor condition MGUS is also observed in families, with first-degree relatives of MM patients having a two-fold elevated risk of MGUS (RR 2.1; 95% CI 1.5-3.1). Two further studies have demonstrated a three-fold elevated risk of developing MM and MGUS among relatives of MGUS patients [76, 77]. In addition, there is evidence for clustering of MM with other tumour types including haematological malignancies. An increased risk among relatives of patients with MM has been found for chronic lymphocytic leukaemia (CLL) (RR = 1.33–2.45) [78, 79], Waldenström's macroglobulinemia (RR = 4.0) [79], acute lymphoblastic leukaemia (ALL) (RR = 2.1) [80], and non-Hodgkin lymphoma (NHL) (RR = 1.34–1.35) [78, 81]. Notably, these are all of lymphoid origin.

1.3 Models of inherited genetic predisposition

The two- to three-fold familial risks associated with MM and other cancers are compatible with a range of effect sizes and frequencies of predisposition alleles. Studies have detected two main classes of cancer susceptibility alleles with different levels of risk and prevalence in the general population. First, rare moderate-penetrance variants (risk allele frequency <2%; odds ratios (ORs) >2.0) have been identified through investigation of candidate genes. Second, common low-penetrance alleles (risk allele frequency >5%; ORs <1.5) have been identified by genome-wide association studies (GWAS). The penetrance and frequency spectrum of cancer risk alleles, in general, likely exists on a continuum and the observed dichotomy aforementioned may reflect the methods used to detect risk alleles, rather than the underlying biology [82].

1.3.1 Rare, highly penetrant alleles

Successful identification of cancer susceptibility genes has been dominated by linkage studies of highly selected families. These analyses have led to the identification of most of the currently known high-penetrance susceptibility genes (e.g. *BRAC1* and *BRAC2* associated with breast and ovarian cancer [83, 84], *APC*, *MLH1* and *MSH2* with colorectal cancer (CRC) [85-88]). While mutations of such genes produce highly penetrant phenotypes, these mutations are rare and explain only a minor component of disease susceptibility in these instances.

To date no linkage studies exist in MM high-risk families, though Waller *et al* developed a gene mapping strategy to search for shared genomic segments, using data from 11 MM high-risk families from Utah [89]. This study identified a 1.8-Mb shared segment on 6q16, harbouring nine genes, in one pedigree. Exome sequencing in this region revealed predicted deleterious variants in *USP45* (p.Gln691* and p.Gln621Glu), a gene involved in DNA repair through endonuclease regulation. Additionally, a 1.2 Mb segment at 1p36.11 was identified in two Utah high-risk pedigrees, with coding variants found in *ARID1A* (p.Ser90Gly and p.Met890Val), a gene in the SWI/SNF chromatin remodelling complex. However, these findings have not yet been replicated.

An example of a rare low-penetrance susceptibility allele in MM is provided by the germ-line mutations observed in *CDKN2A* (p16INK4A), a tumour suppressor gene encoding a cell cycle inhibitor. However, this was described in a single family with four individuals affected with melanoma and only one fifth family member affected with MM [90].

1.3.2 Common, low penetrance alleles

To date, high penetrance mutations in more than 72 genes [91] have been associated with susceptibility to cancer, but these account for only a small fraction of the familial risks of the respective cancers, leaving much of the heritability unexplained. It is likely that most of the inherited genetic susceptibility to common cancers result from multiple inherited genetic variants. The “common disease, common variant” hypothesis posits that a substantial proportion of the genetic risk of common diseases can be accounted for by the action of multiple low-penetrance alleles that have a relatively high population frequency [92] (**Figure 1.6**). Although such alleles confer small effects individually, they could contribute significantly to disease susceptibility in the general population. These alleles are highly unlikely to cause multiple cases in families and therefore would have eluded prior detection through linkage studies [93].

1.3.3 Candidate gene association studies

Until more recently, the search for common genetic variants influencing MM have been based on analyses of polymorphisms in pre-selected candidate genes. Hypotheses, which have been examined in candidate gene association studies include the role of cytokines and immune response, DNA repair, folate metabolism, ADME (absorption, distribution, metabolism and excretion), insulin-like growth factors, and apoptosis [94-99]. While some studies report positive associations, these have not been replicated in an independent cohort. These findings are characterised by small case-control studies, whose low power to detect true associations therefore increase the risk of false-positive discoveries and are limited in their ability to appropriately account for population substructures [100, 101]. Prior knowledge of specific disease-related candidate genes, together with the prioritisation of alleles with respect to these genes, form the basis of candidate gene association studies; however current knowledge about the disease aetiology make pre-selection of genes inherently difficult.

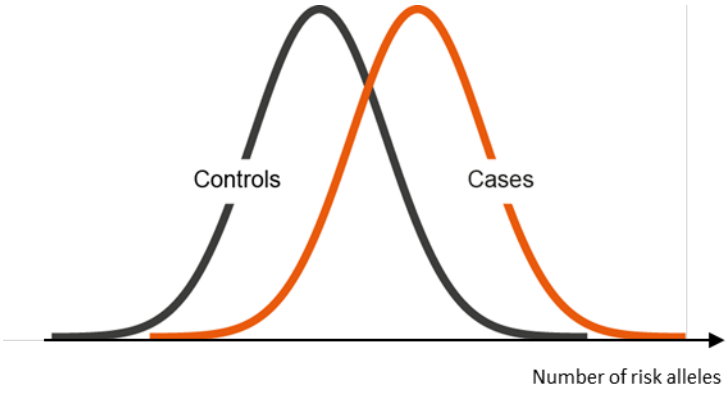


Figure 1.6 Polygenic model of disease susceptibility. The distribution of risk alleles in both cases and controls follows a normal distribution. However, cases have a shift towards a higher number of risk alleles. Figure adapted from [102].

1.4 Genome-wide association studies

The advent of genome-wide association studies (GWAS) has provided a powerful approach in the identification of common, low-penetrance risk alleles for MM and these studies have transformed our understanding of susceptibility to the disease. GWAS generally use single nucleotide polymorphisms (SNPs) as marker variants of investigation [103]. SNPs are common variants in the genome which occur approximately every 300-1,000bp [103]. A SNP marker allele is associated with a disease if one allele is found significantly more frequently in cases than in cancer-free controls. SNPs are inherited in blocks, with SNPs in closer proximity more likely to be inherited together [92]. This non-random association between alleles at loci on the same

chromosome occurs during meiosis and is referred to as linkage disequilibrium (LD) (**Figure 1.7**) [92, 104]. Correlated SNPs co-segregate into a haplotype and this allows certain SNPs across the genome to be selected as 'tag SNPs', which can capture the majority of sequence variation in a given region [105].

The number of SNPs that require genotyping to capture most common variants across the human genome (that is, those with a minor allele frequency >5%) is therefore reduced to around 300,000. Arrays that assess for common genetic variations in the form of SNPs across the entire human genome typically directly genotype 300,000-1,000,000 tagging SNPs. This allows for identification of regions associated with a disease or trait (termed "risk loci") without prior knowledge of genomic location or function. The power of an association study is the likelihood of detecting an effect if there is a true genetic effect present to detect. It is dependent on many factors, including the sample size, the genetic model used, the frequency of the disease allele under study, the effect size of the variant on the trait of interest, and the significance threshold required to declare a true association [106]. In the case of MM especially, GWAS provide a significant advantage over linkage studies as single cases are much more readily available than large extended pedigrees. This allows large enough sample sizes, and therefore increased power, to detect variants with small effects. Furthermore, international collaborations of MM studies can be combined in a meta-analysis of GWAS to greater increase power [107]. An alternative approach is to select cases that are genetically enriched for disease, such as those with a family history or early age of disease onset [108].

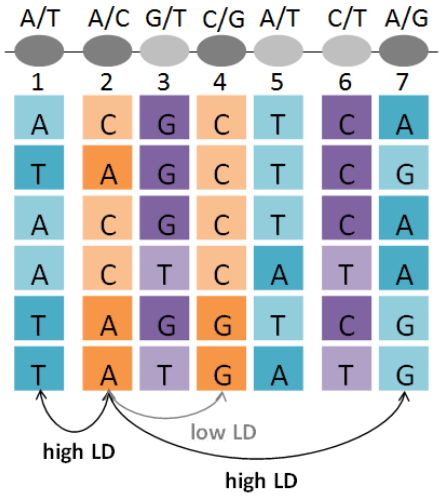


Figure 1.7 Principle of linkage disequilibrium. It is possible to identify genetic variation without genotyping every SNP in a chromosomal region. For example, through genotyping SNP 2 it is possible to infer the genotypes of SNP 1, SNP 4 and SNP 7. SNP 2 therefore can be a 'tag' SNP.

1.4.1 Imputation

Risk SNPs identified through GWAS represent proxies for the association signal and while a minority of GWAS tag SNPs are directly functional, the majority are not themselves necessarily the functional or causative variant at the risk locus. The causative SNP in the association is likely to be correlated with the sentinel tagSNP at the GWAS association peak while not being directly genotyped on a GWAS array. A key step in deciphering the causative SNP at a risk locus is fine-mapping, which is aided by imputation of untyped genotypes (**Figure 1.8**).

Imputation is a computational method that aims to predict the likely genotypes at un-genotyped loci across the genome and makes use of the information provided by haplotypes in a reference panel of sequenced samples such as the 1000 Genomes project [109] and UK10K project [110]. Imputation can boost power of up to 10% over testing only genotyped SNPs and aid in identifying new regions of association at variants that are incompletely tagged by GWAS tagSNPs or at insertion/deletions (indels) that are not fully captured by GWAS arrays [111]. Furthermore, where different genotyping arrays are used in different cohorts, imputation can allow for harmonization of SNPs across the cohorts, so that meta-analysis can be performed. Imputation is limited by the choice of reference panel, the quality and size of which can impact the imputation fidelity.

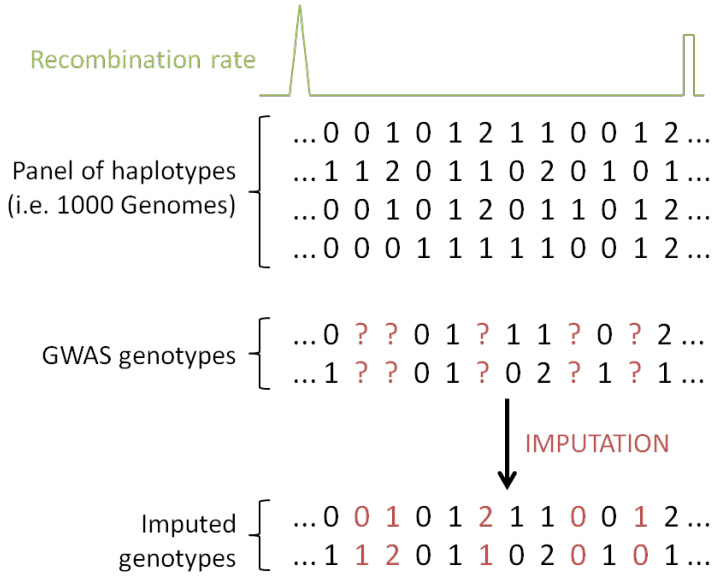


Figure 1.8 Overview of imputation. Imputation utilises a reference panel of phased haplotypes to infer the genotypes at un-typed positions in GWAS datasets, increasing the number of variants that can be tested for an association with disease. Red "?" indicate variants not genotyped on the GWAS array. The most likely haplotype is chosen from the reference panel to predict the genotypes at these variants.

1.4.2 Association studies in multiple myeloma

The first GWAS was carried out by Broderick *et al.* [112] and comprised 1,675 cases from the UK and Germany. The study identified risk loci at 3p22.1 (*ULK4*), 7p15.3 (*CDCA7L*, *DNAH11*), and a promising (5×10^{-6}> *P* >5×10^{-8}) locus at 2p23.3 (*DNMT3A*, *DTNB*). This dataset was extended to 4,692 cases, by Chubb *et al* [113] who identified four risk loci at 3q26.2 (*TERC* and other genes), 6p21.33 (*HLA*), 17p11.2 (*TNFRSF13B*), and 22q13.1 (*CBX7*). An independent GWAS carried out by Swaminathan *et al* [114], comprising 3,031 MM and MGUS cases from a Swedish/Norwegian, Danish, and Icelandic dataset, reported risk loci 5q15 and a promising locus at 22q13. From Weinhold *et al* [115], a subset of 1,655 MM cases from Chubb e*t al,* with linked fluorescence *in situ* hybridisation (FISH) data, was assessed for subtype analysis to identify the risk for developing a specific tumour karyotype in MM. The 11q13.3 locus was identified to be associated with the t(11;14) translocation in which *CCND1* is placed under the control of the immunoglobulin heavy chain enhancer. Mitchell *et al* [116] performed a meta-analysis on 9,866 cases from the data sets from United Kingdom, Germany, and Scandinavia, including two new data sets from the Netherlands and USA. This meta-analysis identified eight new loci at 6p22.3 (*JARID2*), 6q21 (*ATG5*), 7q36.1 (*SMARCD3* and other genes), 8q24.21 (*CCAT1*), 9p21.3 (*CDKN2A*), 10p12.1 (*WAC*), 16q23.1 (*RFWD3*) and 20q13.13 (*PREX1*) and brought the 22q13 (*TOM1*, *HMGXB4*) association from the Scandinavian association study to genome-wide significance (Section 2.3.2).

3p22.1

The 3p22.1 association signal spans *ULK4*, which encodes a serine/threonine-protein kinase. The G-to-A transition at rs1052501 results in Ala542Thr which was predicted by the authors to be tolerated from *in silico* SIFT (sorting intolerant from tolerant) analysis or benign from Polyphen2. Although the exact function of *ULK4* is not known, the Atg1-ULK complex with *ULK1* and *ULK2* regulates mTOR-mediated autophagy, a pathway critical in MM biology [117, 118]. In addition to *ULK4*, the region of LD encompasses *TRAK1*, which regulates the endocytic trafficking of the GABAA receptors [119].

7p15.3

At 7p15.3, *CDCA7L* (Cell division cycle-associated 7-like protein) has been implicated as a candidate gene as it encodes a cell division-associated protein that binds the transcriptional co-activator p75, and is thought to potentiate MYC-mediated transformation events [120-122]. In addition, *CDCA7L* is highly expressed in plasma cells, and Weinhold *et al* [123] and Li *et al* [124]

demonstrated in follow-up studies that the MM risk allele increases *CDCA7L* expression in plasma cells as a result of the lead variant rs4487645 creating a new binding site for the transcription factor IRF4.

2p23.3

DTNB (β-dystrobrevin), a component of the dystrophin-associated protein complex has been implicated at 2p23.3. The LD region also encompasses *DNMT3A* (DNA (cytosine-5)-methyltransferase 3A), a *de novo* DNA methyltransferase lowly expressed in MM. Epigenetic changes are observed in the transition from normal plasma cells, MGUS, MM to relapsed MM, specifically global DNA hypomethylation and gene-specific DNA hypermethylation, suggesting a role of epigenetic deregulation in MM development [125].

3q26.2

rs10936599 at 3q26.2 results in a synonymous mutation in MYNN (myoneurin), a zinc finger protein that is expressed abundantly in muscle. 3q26.2 encompasses *TERC* (telomerase RNA component). Telomerase activity and telomerase-mediated elongation of shorter telomeres is a feature of MM [126]. Sequence variation at *TERC* associates with several other cancer types, including CLL [127], glioma [128-130], CRC [131], and thyroid [132] cancer. Notably, the rs10936599 G risk allele is associated with significantly longer telomeres in CRC patients and has been shown to increase CRC risk [131].

6p21.33

The HLA region contains numerous genes relevant to B- and T-cell development and function, and the MM risk allele seems to represent HLA-DRB5*01. This region spans *PSORS1C1* (psoriasis susceptibility 1 candidate 1) and *POU5F1* (POU domain, class 5, transcription factor 1), which regulates stem cell pluripotency, lineage commitment and tissue-specific gene expression [133]. However there is currently no association between *POU5F1* and MM pathogenesis. 6p21.33 has been shown to be associated with follicular lymphoma [134] and Hodgkin's lymphoma risk [135], defined by variants in the HLA class I and II regions. The MM risk associated with these SNPs was non-significant.

17p11.2

The association at 17p11.2 encompasses *TNFRSF13B* (tumour necrosis factor receptor superfamily member 13B). TNFRSF13B is required for transitional and mature B-cell

development and normal B-cell homeostasis. *TNFRSF13B* −/− mice demonstrate an increase in the number of B-cells in the lymph nodes and spleen, an infiltration of lymphocytes in the liver and kidneys, and increased lymphoma risk [136].

22q13.1

The association at 22q13.1 spans *CBX7* (chromobox homolog 7), which encodes a polycomb group protein. Proteins from this group regulate cell fate determination during normal and pathogenic cell growth and differentiation [137]. *CBX7* mediates transcription repression on *CDKN2A*, transcription of which is required for replicative or oncogene-induced senescence [138]. CBX7 also cooperates with MYC to promote aggressive B-cell lymphomagenesis [139].

5q15

The association at 5q15 spans *ELL2* (elongation factor, RNA polymerase II 2). This gene encodes a key component of the super-elongation complex, which mediates rapid gene induction by suppressing transient pausing of RNA polymerase II [140]. In mature and memory B-cells, which express *ELL2* at a low level, IGH-mRNA is translated to membrane-bound Ig [141]. In plasma cells, *ELL2* is highly expressed and helps RNA polymerase II find a promoter-proximal weak poly(A)-site, allowing IGH-mRNA to be translated to secreted Ig. Both Li *et al* [142] and Ali *et al* [143] have investigated the functional basis of the association at this locus. Li *et al* [142] propose the causal SNP as rs6877329, which forms a chromatin looping interaction with the *ELL2* promoter, with the C allele reducing enhancer activity in MM and conferring lower ELL2 expression in MM patients. Li *et al* [142] also provided data suggesting that the *ELL2* allele preferentially predisposes for the hyperdiploid MM subtype. Ali *et al* [143] identified rs3777189 as a likely candidate causal variant that perturbs a binding site for MAFF/G/K transcription factors, and also found that the ELL2 risk allele increases ribosomal gene expression, proposing this as a possible compensatory reaction. Both studies showed that the *ELL2* MM risk allele reduces *ELL2* expression in CD138+ plasma cells.

22q13

At 22q13, *TOM1* (target of myb1 membrane trafficking protein) was implicated as a candidate gene. *TOM1* encodes an adapter protein required for the maturation of autophagosomes and their fusion with lysosomes, and also displays higher expression in plasma cells relative to other blood cell types [114]. The LD region at 22q13 association also comprises *HMGXB4* (HMG-box containing 4). A dominant mutation in *TOM1* has recently been identified in a family with early-

onset autoimmunity and combined immunodeficiency with decreased levels of immunoglobulins as well as several lymphocyte subsets, including switched memory B-cells [144].

6q21

The 6q21 association marked by rs9372120 maps to intron 6 of *ATG5* (Homo sapiens autophagy related 5). *ATG5* is highly expressed in plasma cells and essential for autophagy and plasma cell survival [145, 146]. Using data from lymphoblastoid cell lines (LCLs), it was shown that the region at 6q21 (rs9372120, *ATG5*) participates in intra-chromosome looping with the transcriptional repressor *PRDM1* (alias *BLIMP1*) [116], which has an established role in plasma cell development and survival [36, 37, 147].

6p22.3

The 6p22.3 (rs34229995) association is 2.2-kb telomeric to the 5' of *JARID2* (jumonji, AT-rich interactive domain 2). JARID2 functions as a transcriptional repressor through recruitment of Polycomb repressive complex 2 and has recently been identified as a regulator of haematopoietic stem cell function [148]. Furthermore, the 6p22.3-p21.31 region is commonly gained in MM tumours [53].

7q36.1

The 7q36.1 (rs7781265) association localizes to intron 2 of *SMARCD3* (swi/snf-related, matrix-associated, actin-dependent regulator of chromatin, subfamily d, member 3). *SMARCD3* recruits BAF chromatin remodelling complexes to specific enhancers.

8q24.21

The 8q24.21 variant rs1948915 maps to *CCAT1* (colon cancer-associated transcript 1). The same region at 8q24.21 harbours multiple independent loci with different tumour specificities, including the B-cell malignancies diffuse B-cell lymphoma [149], Hodgkin's lymphoma [135] and chronic lymphocytic leukaemia [150]. With the exception of CLL, the SNPs underlying these associations have been shown to reside in distinct LD blocks [82].

9p21.3

The 9p21.3 variant rs2811710 maps to intron 1 of *CDKN2A*/*p16INK4A* (cyclin-dependent kinase inhibitor 2A). This region is a susceptibility locus for multiple tumour types including breast and

lung cancer [151], glioma [152] and acute lymphoblastic leukaemia [153]. Furthermore, the 9p21.3 locus interacts with the genomic region containing *MTAP* (methylthioadenosine phosphorylase) and deletion of *MTAP* is common in cancer, being closely linked to homozygous deletion of p16 [154].

10p12.1

The 10p12.1 (rs2790457) association localizes to intron 3 of the gene encoding *WAC* (ww domain-containing adaptor with coiled-coil region), which has been shown to be part of an extended autophagy network [155].

16q23.1

The region at 16q23.1 encompasses *RFWD3*, a gene encoding an E3 ubiquitin ligase that positively regulates p53 stability by forming an RFWD3–MDM2–p53 complex, thereby protecting p53 from degradation by MDM2-mediated polyubiquitination. Variation at 16q23.1 defined with the correlated SNP rs4888262 has previously been shown to influence testicular cancer risk [156].

20q13.13

The 20q13.13 (rs6066835) association mapped to intron 3 of *PREX1* (phosphatidylinositol-3, 4, 5-trisphosphate-dependent Rac exchange factor 1), a Rac guanine exchange factor that coordinates signalling inputs from G protein-coupled receptors and receptor tyrosine kinases. Due to its role in the PI3K/AKT pathway and MEK/ERK signalling, *PREX1* has been proposed as a biomarker and therapeutic target in breast cancer [157].

1.4.3 Perspectives from GWAS

So far, MM GWAS provide evidence to support a polygenic model of MM and have identified 17 risk loci [112-114, 116], with one additional risk locus specific for t(11,14) translocations [115]. These risk SNPs are common (European minor allele frequency [MAF] >0.01) and have modest effect sizes ($1.12 < OR < 1.38$). In addition, the loci encompass genes which are known to be important in plasma cell or cancer biology, for example *TNFRSF13B* at 17p11.2, *TERC* at 3q26.2, *ELL2* at 5q15 and *MYC* at 8q24.21.

It is estimated that the heritability explained by the nine previously identified common MM risk SNPs from GWAS was 2.9%, whereas the heritability explained by all common SNPs was 15.2%

[158]. Comparing the heritability explained by the common variants with that from family studies, a fraction of the heritability may be explained by other genetic variants, such as rare variants [158]. In summary, much of the heritable risk of MM remains unexplained and statistical modelling indicates that further common risk variants remain to be discovered [158].

1.5 Strategies to identify novel myeloma susceptibility loci

1.5.1 Meta-analysis of GWAS

Given that many GWAS exhibit long tails of associations with small effect sizes, much of the underlying genetic architecture of cancer susceptibility may be due to a large number of common susceptibility alleles, which individually account for a very small proportion of the inherited risk. New susceptibility loci could therefore potentially be identified through a new generation of larger GWAS, involving large-scale meta-analysis and replication. Additionally, given that variation at 11q13.3 is driven by the association with the MM subtype t(11;14) [115], it is possible that further studies combining pre-existing and potentially additional GWAS datasets with linked karyotype information will identify further subtype-specific MM risk loci.

1.5.2 Next generation arrays

Another possibility is that low-frequency risk variants (MAF ~0.01) contribute significantly to the familial risk of MM. While current GWAS arrays are designed to capture common risk variants, they do not adequately capture variation at MAF < 0.05 [159, 160]. Using pools of reference haplotypes such as that provided by the 1000 Genomes Project and UK10K Project, whole-genome imputation may extend the frequency range for which associations can be detected from existing datasets [110, 161]. However, it is likely that the discovery of this class of susceptibility allele will be reliant on next-generation SNP arrays, for example, the recently developed genotyping microarray, the OncoArray. This array includes a genome-wide backbone, comprising 230,000 SNPs tagging most common genetic variants, together with dense mapping of known susceptibility regions, rare variants from sequencing experiments, pharmacogenetic markers and cancer related traits [162].

Given low-frequency risk variants are likely to be highly population-specific, they are more difficult to detect by generic array-based technologies. Such considerations, as well as the likelihood that many risk variants have insufficient frequency to be detectable through scans of the general population [163], increasingly suggest that comprehensive characterisation of the contribution of genetic variation to MM risk will rely on sequencing data. This data can be

potentially generated from whole exome sequencing (WES) or whole genome sequencing (WGS) studies of MM individuals in families. This will additionally allow the interrogation of more complex forms of genetic variation to MM risk, such as structural variation and copy number variants (CNVs), which are not as amenable to capture by array-based technologies. Germline high-coverage WES study has been conducted on a general population of 513 MM cases and 1,569 healthy controls [164]. However no protein-coding low-frequency alleles (MAF of 0.01-0.05) were statistically associated with MM risk due to limited power in the study, though a suggestive association with *KIF18A* was found by employing gene burden testing (which collapses information for multiple genetic variants into a single genetic score [165]) . The use of familial cases provides a means of significantly empowering the search for rare disease-causing alleles for cancer. A potentially deleterious missense variant in *EP300* (p.Arg695His) was found using WES in a family with multiple cases of MM and MGUS [166]. As the number of familial MM cases are few, the practicality of adopting this as a means of gene identification for MM and replicating associations is problematic.

1.6 Functional annotation of GWAS risk loci

To date, GWAS have produced risk loci for various diseases. A key task in post-GWAS analysis is to decipher the biological effect these loci confer in disease susceptibility [167]. Consideration of the functional effect of risk SNPs is important in prioritisation of potential causal variants in fine-mapped GWAS association signals, as well as determining the mechanistic effect of the risk locus in disease origin and/or progression. Given the plethora of possibilities by which a variant may functionally act, elucidating the mechanistic basis by which a given SNP exerts its effect on disease risk remains a considerable challenge.

Variants can broadly be classified into those that are coding (*i.e.* directly affect protein function) and non-coding. Only a small number of variants identified from cancer GWAS lie in exons and have been shown to directly impact the amino acid sequence of an expressed protein e.g. *BRCA2* p.Lys3326Ter (rs11571833) and *CHEK2* p.Ile157Thr (rs17879961) for lung cancer [168]. The majority of GWAS loci discovered to date map to non-coding regions of the genome (*e.g.* gene introns, promoters or intergenic regions) and are understood to exert their effect by regulation of gene expression, with a variety of models having been proposed. Variants may affect correct mRNA processing; an example of this is the SNP in the 3' untranslated region (poly(A) tail) of *TP53* (rs78378222) associated with prostate cancer and glioma risk [169, 170] or cause a splice site *e.g.* the rs10069690 variant at 5p15.33 (*TERT*), resulting in decreased telomerase activity

[130, 171, 172]. Notably rs603965, which is associated with t(11,14) CCND1 translocation in MM, has been shown to differentially influence the alternative splicing of *CCND1* mRNA, the 870G allele creating an optimal splice donor site at the exon 4/intron4 boundary resulting in the cyclin D1a transcript. The 870A allele hinders splicing allowing for read-through into intron 4 and production of the variant cyclin D1b transcript, though is not fully penetrant [115]. In addition, variants discovered though GWAS may affect protein translation via binding of microRNAs and altered expression of large intergenic noncoding RNAs. Furthermore, variants may affect gene transcription through disruption of local promoter-transcription complex interactions or potentially long-range enhancer-complex interactions [167, 173]. These proposed mechanisms are summarised in **Figure 1.9**.

1.6.1 Assessing the impact of protein-coding variants on disease predisposition

Coding changes can affect function by altering amino acid sequence (missense changes) and causing protein truncation (creation of premature stop sequence, aberrant splicing of exons) [167]. For protein truncating variants the likely impact on protein function is potentially severe, however, the potential functional impact of missense variants is less clear. *In silico* prediction algorithms, such as SIFT [174] and PolyPhen-2 [175], are available which primarily make use of protein sequence conservation information to decide on the deleteriousness of a query amino acid substitution.

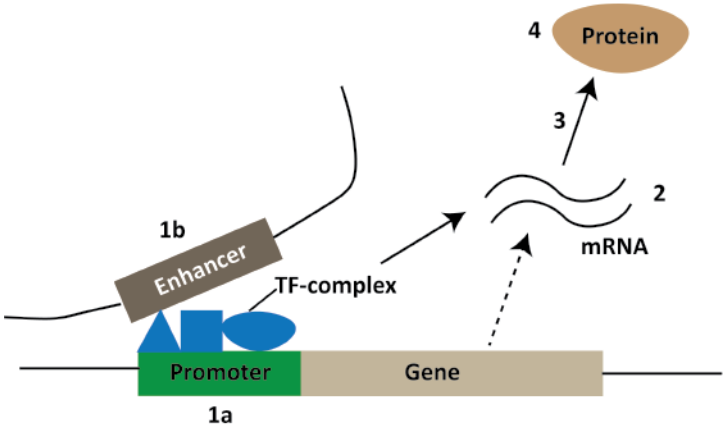


Figure 1.9 Possible basis by which polymorphisms mediate cancer susceptibility. 1a- Affecting gene transcription through disrupting local promoter-transcription complex interactions, 1b- Affecting gene transcription through potentially long-range enhancer-complex interactions, 2- Affecting correct mRNA processing (*e.g.* splicing, poly-adenylation), 3- Affecting protein translation (*e.g.* microRNAs, stop-gain changes) and 4- Affecting protein sequence (amino acid substitution). TF, transcription factor.

1.6.2 Assessing the impact of non-coding variants on disease predisposition

As the majority of variants identified through GWAS localise to non-coding regions, there have been major efforts which aim to inform the regulatory mechanisms perturbed at cancer risk loci. Knowledge of cell-type-specific transcription factor binding, histone mark characterisation, expression quantitative trait loci, methylation quantitative trait loci and chromatin conformation, can all be used to annotate GWAS loci. Indeed, large consortia such as ENCODE [176], NIH Roadmap Epigenomics [177, 178] and Blueprint [179] have catalogued such experiments which map regulatory regions. Programs such as ChromHMM, which uses a multivariate Hidden Markov Model that explicitly models the presence or absence of chromatin marks in different cell types [180] have enabled understanding of higher-order structures governing disease susceptibility. In parallel to the increasing availability of regulatory data from a range of cell types statistical methods which integrate gene expression with GWAS datasets such as Summary-data-based Mendelian Randomization (SMR) [181] and transcriptome-wide association studies (TWAS), aid in identification of candidate causal genes. Furthermore, since TWAS aggregates the effects of multiple variants into a single testing unit, and facilitates prioritisation of genes at known risk regions for functional validation, it potentially also affords increased study power to identify new risk regions.

1.7 Further applications of GWAS

As well as informing cancer biology, GWAS can assist in identifying individuals at increased risk of cancer, aid in drug discovery and repositioning (via understanding of the genes and pathways identified) and can elucidate aetiological risk factors for cancer.

1.7.1 Polygenic risk scores

Polygenic risk scores (PRS) are the weighted sum of the number of risk alleles carried by an individual. While previous clinical applications have focussed on rare, highly penetrant mutations which confer increased risk (e.g. *BRCA1* and *TP53* in breast cancer [182]), there is increasing evidence that PRS of complex disease can be used to identify individuals at higher risk and have been considered as an aid in stratified screening [183]. This has been demonstrated for colorectal cancer (CRC) as well as breast and prostate cancers [184] [185, 186].

1.7.2 Informing therapy

Knowledge of germline genetic variation is demonstrating increased potential to inform treatment. For example, GWAS has been used to identify individuals at risk of treatment related

toxicity [187, 188]. Furthermore, in the case of MM, germline genetic variation has been linked to survival [189], indicating that inherited genotypes could provide prognostic information in the context of this disease.

1.7.3 Mendelian randomisation

Mendelian randomisation (MR) analysis uses genetic markers (termed instrumental variables [IVs]) known to be associated with a potential risk factor in the assessment of that risk factor's effect on another trait or disease [190, 191]. The availability of large GWAS data sets has established robust IVs in the form of genetic risk variants and MR offers the ability to identify non-genetic risk factors using these IVs [192]. For example, increased body mass index (BMI) has recently been implicated as a risk factor for CRC using MR [192]. Such studies have also identified chemopreventive agents [193] and performed safety analysis of therapies [194]. MR provides an advantage over conventional observational studies as genetic variants are randomly assigned at conception so they are not influenced by reverse causation and can provide unconfounded estimates of disease risk.

1.8 Study aims and scope of enquiry

The inherited predisposition to MM is currently understood to involve multiple low-penetrance risk SNPs, however a large proportion of the genetic risk to MM currently remains unaccounted for.

The work detailed in this thesis aims to demonstrate further insight into genetic predisposition to MM, making use of currently available technologies and analytical methods. It is anticipated that research into the genetic basis of this plasma cell malignancy will lead to increased insight into MM biology and potentially identify novel therapeutic strategies.

Specifically:

- Chapter 3 details a new GWAS and meta-analysis with previously existing datasets, performed to identify new risk loci. Estimation of the contribution of common variation to the narrow-sense heritability of MM is calculated and PRS are constructed using GWAS datasets. New and established risk loci are functionally annotated using regulatory data, including information from ChIP, to identify enhancer histone marks, and Hi-C, to identify long range interactions.
- Chapter 4 describes a TWAS to identify candidate causal genes for MM.

- Chapter 5 investigates genetic correlation between CLL and MM, identifies pleiotropic risk loci between the two haematological malignancies and characterises these regions using regulatory data from B-cells, in order to gain insight in the cellular aetiology of MM.
- Chapter 6 uses MR to investigate potential risk factors for MM in a phenome-wide association study.

CHAPTER 2 Materials and methods

2.1 Subjects

Datasets used in this thesis are detailed within this section. The diagnosis of MM (International Classification of Diseases, 10th Revision (ICD-10) C90.0) was established in accordance with World Health Organization guidelines [195]. All samples from patients for genotyping were obtained before treatment or at presentation. Collection of patient samples and associated clinicopathological information was undertaken with written informed consent and relevant ethical review board approval at respective study centres in accordance with the tenets of the Declaration of Helsinki.

2.1.1 Genome-wide association study datasets

UK OncoArray GWAS

Cases

Post-QC, the OncoArray GWAS series comprised 878 cases ascertained through the UK Myeloma XI trial [196]. The Myeloma-XI (MyXI) [196] Phase III clinical trial was set up in 2007 and recruited 4,400 patients. The trial was designed to test different combinatorial drug schemes in patients involving lenalidomide, cyclophosphamide, dexamethasone, carfilzomib and vorinostat. All cases were UK residents and had self-reported European ancestry. Samples were collected from patients aged 18 years or older and newly diagnosed as having symptomatic MM or non-secretory MM. Samples were subject to SNP and sample quality control (QC) as described in Section 2.3.1.

Controls

The OncoArray GWAS control series comprised (i) 2,976 cancer-free men (age <65 years) recruited by the PRACTICAL Consortium [197], the UK Genetic Prostate Cancer Study (UKGPCS), and (ii) 4,446 cancer-free women from across the UK recruited via the Breast Cancer Association Consortium (BCAC) [198]. The UKGPCS established in 1993, is a nationwide study of inherited risk to prostate cancer, which aims to find genetic changes which are associated with prostate cancer risk. Prostate Cancer Association Group to Investigate Cancer Associated Alterations in the Genome (PRACTICAL) is a collaboration of researchers investigating the inherited risk of prostate cancer. Formed in April 2005, BCAC is an international multidisciplinary consortium which investigates the inherited risk of breast cancer.

Dutch GWAS

Cases

Post-QC, the Dutch GWAS consisted of a total of 555 cases, recruited from three clinical trials: HOVON 65/GMMG-HD4 (restricted to Dutch cases; n = 158), HOVON 87/NMSG18 (n = 292) and HOVON 95/EMN02 (n = 105) [199-201]. DNA was extracted from EDTA-venous blood samples. The multi-centre HOVON Foundation, from which Dutch GWAS samples were recruited, was set up in 1985 and has run several clinical trials of MM [199-201]. The HOVON Foundation has collaborated with the GMMG Study Group, Nordic Myeloma Study Group (NMSG) and the European Myeloma Network (EMN) for multiple clinical trials.

Controls

The Dutch GWAS controls consisted of 2,669 individuals recruited from the B-vitamins for the Prevention of Osteoporotic Fractures (B-PROOF) Dataset [202]. The B-PROOF study is a multi-centre study initiated in 2008 in The Netherlands [202] to investigate bone fracture risk. The study population involved 2,919 individuals aged 65 years or older with homocysteine concentration above a defined threshold. Controls for this GWAS were from the placebo-controlled group of the B-PROOF study.

German GWAS

Cases

Post-QC, the German GWAS comprised 1,508 cases recruited by the German-Speaking Multiple Myeloma Multicenter Study Group (GMMG). DNA was prepared from EDTA-venous blood or CD138-negative bone marrow cells (< 1% tumour contamination). Patients were recruited from GMMG-HD3 (2001-2005, n=550), GMMG-HD4 (2005-2011, n=399), and GMMG-HD5 (2010-2016, n=604) trials [199, 200, 203, 204]. In GMMG-HD3/4 patients were aged 18-65, diagnosed of stage II/III MM according to the Salmon and Durie criteria [204]. In GMMG-HD5 patients were aged 18-70 and diagnosed with MM requiring systemic therapy.

Controls

The German GWAS comprised 2,107 healthy individuals as controls, who were enrolled in the Heinz Nixdorf Recall (HNR) study [205]. The HNR was a population-based, prospective cohort study which investigated the ability of a coronary artery calcification scoring system in predicting

the risk of developing major cardiovascular events. It was initiated in the late 1990s and recruited 4,814 German residents, aged 45- 75 years, between 2000 and 2003 [205].

Icelandic GWAS

Cases

The Icelandic GWAS comprised 480 MM cases from the Icelandic Cancer Registry [206]. The Icelandic Cancer Registry was established in 1955 with the aim of registering all cancers diagnosed in Iceland, collecting histological, cytological, hematological and/or autopsy data on registered patients [206, 207].

Controls

The Icelandic GWAS comprised 212,164 controls ascertained from different research projects at deCODE Genetics [207].

Swedish GWAS

Cases

Post-QC, the Swedish/Norwegian GWAS consisted of 1,714 cases from the Swedish National Myeloma Biobank (Skåne University Hospital, Lund, Sweden) and the Norwegian Biobank for Myeloma (Trondheim, Norway) [208].

Controls

Post-QC the Swedish GWAS controls comprised genotype data on 10,391 individuals, obtained from previously published studies of schizophrenia [209] and TWINGENE [210]. The schizophrenia GWAS in the Swedish population was conducted in 2013, recruiting 6,243 individuals as controls. The control individuals had not been hospitalised for schizophrenia or bipolar disorder, with Scandinavian parents and aged over 18 years [211]. The TWINGENE Study was conducted from 2004 to 2008 in the Swedish population on twins born between 1911 and 1958 [210]. Samples were obtained from the control arm of the schizophrenia GWAS and one individual from each twin pair was used from the TWINGENE study.

UK GWAS

Cases

Post-QC, the UK GWAS comprised 2,282 cases ascertained through the UK Medical Research Council (MRC) UK Myeloma IX (MyIX) [212, 213] and UK Myeloma XI (MyXI) trials [196]. The

MyIX [212, 214, 215] Phase III clinical trial was set up between 2003 and 2014 and recruited 1,970 patients through over 120 centres in the UK. Patients were randomised for intensive or non-intensive therapy, followed by stem cell transplantation with or without thalidomide maintenance therapy. All cases were UK residents and had self-reported European ancestry. Samples were collected from patients aged 18 years or older and newly diagnosed as having symptomatic MM or non-secretory MM.

Controls

Post-QC, 5,197 controls were used from publicly accessible genotype data generated by the Wellcome Trust Case Control Consortium (WTCCC). Specifically, controls for the UK GWAS were selected from the 1958 Birth Cohort [216] (also known as the National Child Development Study) and the UK National Blood Service (UKBS) [217].

USA GWAS

Cases

Post-QC, the US GWAS comprised 780 incident cases. These were recruited from Total Therapy clinical trials UAMS-TT2 (1998-2004), TT3 (2004-2014), TT3b (2006-2017) and TT4 (2008-2017), which were coordinated by the University of Arkansas for Medical Science (UAMS) Myeloma Institute [11]. Total Therapy trials combine chemotherapy in patients with autologous stem cell transplantation, followed by maintenance strategy. Germline DNA was extracted from leukapheresis products of patients aged 18-75.

Controls

Post-QC the USA GWAS controls comprised data on 1,857 healthy individuals from the Cancer Genetic Markers of Susceptibility Study (CGEMS) [218]. The NCI CGEMS study comprised several population-based studies and was used to investigate common genetic variations in breast and prostate cancers [218].

Chapter 3 details a meta-analysis of the OncoArray GWAS with the above GWAS conducted in the UK, Germany, Sweden/Norway, the US, the Netherlands and Iceland, which had been previously published in their entirety with strict QC procedures [112-114, 116]. Post-QC case and control sample numbers and genotyping array for the GWAS datasets are detailed in **Table 2.1**.

	Numbers		Genotyping arrays	
	Cases	Controls	Cases	Controls
UK	2,282	5,197	Illumina Human OmniExpress-12 v1.0	Illumina Human 1-2M-Duo Custom v1.0
Germany	1,508	2,107	Illumina Human OmniExpress-12 v1.0	Illumina Human Omni1-Quad v1.0; Illumina Human OmniExpress-12 v1.0
Sweden/ Norway	1,714	10,391	Illumina Human OmniExpress-Exome	TWINGENE: Illumina 317K; Illumina Human OmniExpress 700K; Schizophrenia GWAS: Illumina Human OmniExpress; Affymetrix 5.0; Affymetrix 6.0
US	780	1,857	Illumina Human OmniExpress-12 v1.0; Illumina HumanOmni1-Quad	Illumina Sentrix HumanHap550
Netherlands	555	2,669	Illumina Human OmniExpress-12 v1.0	Illumina Human OmniEpress Exome-8 v1.1
Iceland	480	212,164	Illumina microarrays	Illumina microarrays
Oncoarray	878	7,083	Illumina Infinium OncoArray-500K	Illumina Infinium OncoArray-500K

Table 2.1 Details of datasets used in GWAS. Post QC quality control sample number and genotyping arrays for the UK, German, Sweden/Norway, US, Netherlands and Iceland GWAS datasets. The OncoArray GWAS, analysed in Chapter 3 is highlighted in grey.

2.1.2 Replication datasets

Informed consent was obtained from all study participants and each study was carried out with ethical review board approval. Cases for replication of promising ($5\times10^{-6} > P > 5\times10^{-8}$) genetic associations with MM were obtained from the Germany, Sweden and Denmark. Replication genotyping was performed using allele-specific PCR KASPar chemistry (LGC, Hertfordshire, UK). Call rates for SNP genotypes were > 95% in each of the replication series. To ensure the quality of genotyping in all assays, at least two negative controls and duplicate samples (showing a concordance of > 99%) were genotyped at each centre. The fidelity of imputation was assessed by directly sequencing a set of 147 randomly selected samples from the UK OncoArray case series. Imputation was found to be robust; concordance was > 90%. Genotyping and sequencing primers are detailed in **Appendix 1**.

Germany

The German replication series comprised 911 cases collected by the German Myeloma Study Group (Deutsche Studiengruppe Multiples Myeloma (DSMM)), GMMG, University Clinic, Heidelberg and University Clinic, Ulm. Controls comprised 1,477 healthy German blood donors recruited between 2004 and 2007 by the Institute of Transfusion Medicine and Immunology, University of Mannheim, Germany.

Sweden

The Swedish replication series comprised 534 MM cases from the Swedish National Myeloma Biobank. As controls 2,382 Swedish blood donors were analysed.

Denmark

The Danish replication series comprised 332 MM cases from the University Hospital of Copenhagen. As controls 2,229 individuals from Denmark and Skåne County, Sweden (the southernmost part of Sweden adjacent to Denmark) were analysed.

2.1.3 Chronic Lymphocytic Leukaemia (CLL) datasets

In Chapter 5, genetic correlation between MM and CLL was investigated. For this, data from three previously reported CLL GWAS [127, 150, 219, 220] were used. All these studies were based on individuals of European ancestry and comprised: CLL UK1 (505 cases and 2,698 controls), CLL UK2 (1,236 cases and 2,501 controls) and CLL US (2,174 cases and 2,682 controls). The diagnosis of CLL (ICD-10-CM C91.10, ICD-O M9823/3 and 9670/3) was established in accordance with the International Workshop on Chronic Lymphocytic Leukaemia guidelines [221]. After application to NCBI database of Genotypes and Phenotypes (dbGaP), these samples were downloaded and were subject to SNP and sample QC as described in Section 2.3.1

2.1.4 Datasets for expression and survival analysis

Gene expression array data used in Chapter 3 were obtained pre-treatment from CD138+ bone marrow plasma cell samples from MyIX trial generated using Affymetrix Human Genome U133 2.0 Plus Array data. This data is publicly available from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) with accession number GSE21349. Expression data from newly diagnosed MM patients were also obtained from UAMS TT2/3 trials (Section 2.1.1**Error! Reference source not found.**) with expression data generated using Affymetrix GeneChip® Human Full Length arrays (GSE2658, n=559; GSE31161, n=1,038) [222-225], and the HOVON65/GMMG-HD4 trial (Section 2.1.1) with expression data from Affymetrix GeneChip® Human Genome U133 plus 2.0 arrays (GSE19784, n=328) [226]. Expression array data was obtained from relapsed MM plasma cell samples in the Assessment of Proteasome Inhibition for Extending Remissions (APEX) trial generated by Affymetrix GeneChip® Human Genome U133 A/B arrays (GSE9782, n=528) [227]. Expression data from CD138+ bone marrow plasm cells from previously untreated MM patients were obtained from GMMG trials (Section 2.1.1) with Affymetrix GeneChip Human Genome U133 Plus 2.0 array (E-MTAB-372, n=280; E-

MTAB-2299, n=665) [228, 229]. Overall survival analyses in Chapter 3 were obtained from each of the patient datasets listed above.

A summary of the datasets used can be found in **Table 2.2**. Briefly, expression quantitative trait locus (eQTL) analyses were carried out using Affymetrix Human Genome U133 2.0 Plus Array data for CD138+ plasma cells from 183 MRC Myeloma IX trial patients, 658 Heidelberg GMMG patients and 608 US UAMS patients (Section 2.1.1).

Dataset accession number	Clinical trial	Sample size	Type of MM cases	Analysis
GSE21349	MyIX	491	Newly diagnosed	Expression profiling, clinical outcome
EGAS00001001147	MyXI	463	Newly diagnosed	Whole exome sequencing
EGAS00001001, EGAD00001001021	MyIX, MyXI	513	Newly diagnosed	Whole exome sequencing
GSE2658	TT2/TT3	559	Newly diagnosed	Expression profiling, clinical outcome
GSE31161	TT2/TT3	1,038	Newly diagnosed & Relapsed	Expression profiling, clinical outcome (Newly diagnosed patients)
GSE9782	APEX	528	Relapsed	Expression profiling, clinical outcome
GSE19784	HOVON65/ GMMG-HD4	328	Newly diagnosed	Expression profiling, clinical outcome
E-MTAB-372	GMMG-HD3/ GMMG-HD4/ GMMG-HD5	280	Newly diagnosed	Expression profiling, clinical outcome
E-MTAB-2299	GMMG-HD3/ GMMG-HD4/ GMMG-HD5	665	Newly diagnosed	Expression profiling, clinical outcome

Table 2.2 Clinical datasets used in this study. Patients in this study are of HapMap Utah residents of Western and Northern European ancestry.

2.2 Molecular Methods

2.2.1 DNA extraction

Genomic DNA was extracted from EDTA-venous blood, as per standard protocol, unless otherwise stated in Section 2.1 [230].

2.2.2 DNA quantification

2.2.2.1 Picogreen

Extracted DNA was quantified using Picogreen dsDNA quantitation (Invitrogen Molecular Probes, Paisley, UK). First, stock DNA was diluted 1:225 in TE buffer, before 5µl of the diluted DNA was mixed with 95µl Picogreen dsDNA quantitation reagent dye solution (1:200 TE; pH 7.5, Invitrogen Molecular Probes, Paisley, UK). DNA samples were scanned using Labsystems Ascent Fluoroscan (Life Sciences International, Basingstoke, UK) and concentrations calculated using Ascent Software v2.6 (Life Sciences International, Basingstoke, UK).

2.2.2.2 Qubit

DNA concentrations were measured using Qubit Fluorometric dsDNA Quantitation (Q33216, ThermoFisher Scientific, Waltham, USA) according to manufacturer's guidelines. The broad range (BR) and high sensitivity (HS) assay kits were both used depending on the sensitivity requirement for measuring DNA concentrations (BR: 100pg/µl–1µg/µl; HS: 10pg/µl–100ng/µl). The technology relies on dsDNA-selective fluorescent dye to quantitate nucleic acids in solution.

2.2.3 Genotyping

2.2.3.1 Array SNP microarrays

GWAS datasets were genotyped on SNP arrays as detailed in **Table 2.1Error! Reference source not found.**. Prior to genotyping, the concentration of DNA samples were quantified, normalised and plated in 96-well plates at a final concentration of 50ng/µl. The principles of a BeadChip microarray can be illustrated by the Illumina Infinium II assay. Briefly, genomic DNA (≈750ng) is amplified with no allelic partiality. The amplified DNA is enzymatically fragmented and precipitated in alcohol, before resuspension. The DNA is then hybridised onto BeadChip arrays covalently linked to locus-specific 50-mer oligonucleotides. Allele detection occurs in a two-step process. First, a primer hybridises to the complementary region of DNA with the primer terminal 3' end directly adjacent to the SNP to be identified, forming a duplex. Following this, the primer oligonucleotide is enzymatically extended by one base, with the incorporated nucleotide base being covalently linked to a fluorescent tag (**Figure 2.1**). The intensity of this florescence is

detected by the Illumina BeadArray Reader, analysed using Illumina's software and sample genotypes are called automatically.

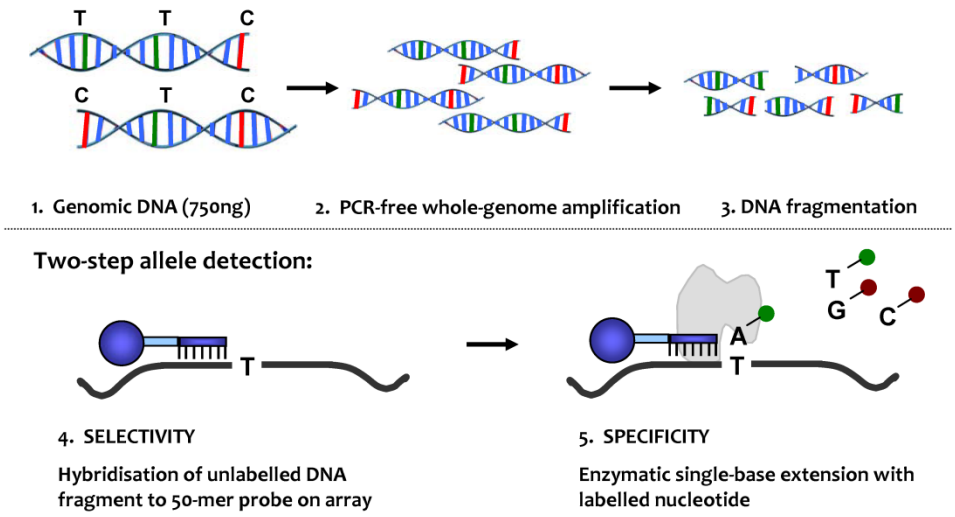


Figure 2.1 The Illumina Infinium II genotyping assay [231].

2.2.3.2 KASPar genotyping

Competitive allele-specific PCR (KASPar) was used to genotype replication samples. KASPar uses fluorescence resonance energy transfer chemistry to quantify the presence of a genotype at a locus within genomic DNA. Allele-specific primers are designed to incorporate a tail that is alike to either a VIC- or FAM-labelled oligonucleotide. Allele-specific primers are added to genomic DNA and KASPar mix, containing polymerase and VIC- and FAM- labelled oligonucleotide. In the first PCR step, genomic DNA is denatured and allele specific primers anneal along with the common primer before DNA is amplified across the target region. In the second round of PCR, the complement of the allele-specific tail sequence is generated. In subsequent rounds of PCR, the levels of allele-specific tail increase and VIC- or FAM-labelled oligonucleotide binds to its complement. This releases the fluorescent dye from its quencher, such that the presence of either VIC or FAM fluorescent signal acts as an indication of the specific allele present at a locus (**Figure 2.2**).

Primer design

Polymerase chain reaction (PCR) oligonucleotide primers for KASPar genotyping were designed using Primer Picker (KBiosciences, Hertfordshire, UK). To confirm specificity, primer sequences were searched against the human genome using BLAST and were subject to *in silico* PCR using UCSC Genome Browser [232]. Oligonucleotides were obtained from Invitrogen (Paisley, UK) and resuspended in dH_2O (1µg/µl single stranded DNA). Details of oligonucleotide primers used are

provided in **Appendix 1**. To ensure sequences did not contain any SNPs present in the European population, target regions were surveyed using SNPmasker.

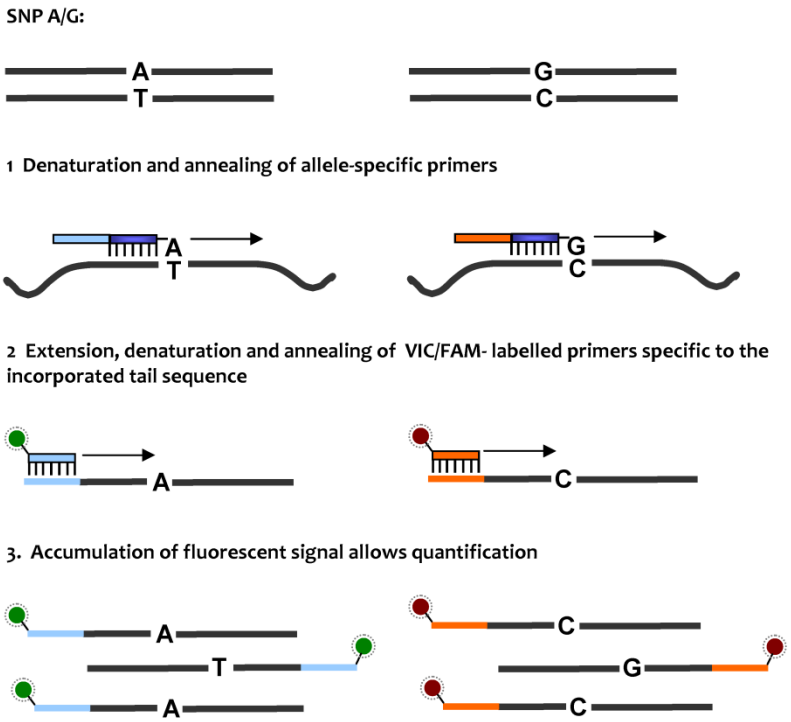


Figure 2.2 The KASPar SNP genotyping system. Two competitive allele-specific tailed forward primers (dark blue) incorporate a nucleotide sequence that is complementary to one of two fluorescently labelled primers (cyan and orange). Accumulation of the fluorescent signal enables quantification and scoring of the alleles.

Amplification by polymerase chain reaction

For every SNP genotyped an assay mix, consisting of 12µM of each allele-specific primer and 30µM of the common primer, was prepared. In each well of a 384-well plate, 1µl of the 2x reaction mix (containing KTaq polymerase, ROX passive reference, VIC- and FAM-labelled primers), 0.055µl assay mix, 0.032µl 50mM MgCl2, and 0.9µl of dH2O was added to 2µl of DNA at a concentration of 2.5ng/µl. Cycling conditions were:

1× cycle	- denaturation for 15 minutes at 94°C
20× cycles	- denaturation for 10 seconds at 94°C
	- annealing for 5 seconds at 57°C
	- extension for 10 seconds at 72°C
22× cycles	- denaturation for 10 seconds at 94°C
	- annealing for 20 seconds at 57°C
	- extension for 40 seconds at 72°C

2.2.4 Polymerase chain reaction

Polymerase chain reaction (PCR) oligonucleotides for sequencing were designed using Primer 3 [233] or manually. Predicted primers were subjected to BLAST searches and *in silico* PCR of the human genome using the online UCSC Genome Browser [232] to confirm specificity. Primer characteristics were checked *in silico* with the IDT OligoAnalyser [234]. The Gibbs free energy (ΔG) of primer homo- and hetero-dimers was thresholded to > -10kcal/mol, and/or the difference in melting temperature (T_m) of the primer set to < 2-5 °C apart. Target regions were surveyed using SNPmasker (http://bioinfo.ut.ee/snpmasker/) to ensure primer sequences did not include SNPs present in the European population (*e.g.* MAF > 0.05). Oligonucleotides were obtained from Sigma (Poole, UK) and re-suspended in TE buffer to a stock concentration of 100μM and in distilled water to a working solution of 10μM. All primers used in this PCR are listed in **Appendix 1**.

2.2.4.1 Standard PCR protocol

A total of 10-100ng genomic DNA was used for amplification by PCR prior to sequencing reaction with the ThermoPrime Taq DNA Polymerase kit (AB0301B; ThermoFisher Scientific, Waltham, USA). DNA was pipetted into a single well on a 96-well microtitre plate (I1402-9800; Star Lab, Milton Keynes, UK) containing a reaction mastermix with, at final concentration, 1 × reaction buffer IV, 0.2mM each deoxynucleotide triphosphate (dNTP), 0.5μM each of forward and reverse primer, 1.5mM $MgCl_2$ and 0.625U Thermoprime Plus DNA polymerase. The microtitre plate was covered with an adhesive lid (AB-0580; ThermoFisher Scientific, Waltham, USA) and the plate was transferred to a thermocycler (Thermo-Hybaid, Middlesex, UK). The heated lid option was selected for all programs to prevent evaporation of products.

The optimum annealing temperature for all primer pairs used was determined by PCR with control human placental DNA (D-3035; Sigma-Aldrich, Poole, UK) over a range of temperatures. "Touchdown" temperatures were used where the first annealing temperature was set to 68°C, 60°C or 55°C. The annealing temperature was then reduced by 1°C every cycle until a temperature of 50°C was achieved (45°C if the initial temperature was 55°C). This was done to increase binding specificity of the primer set to the DNA template. This was followed by amplification at a lower temperature at 60°C.

Typically, PCR conditions were:

1 × cycle - denaturation for 5 mins at 94°C

18 × cycles - denaturation for 30 sec at 94°C
- annealing for 1 min at 68°C (-1°C/cycle)
- extension for 1 min/kb at 72°C

17 × cycles - denaturation for 30 sec at 94°C
- annealing for 1 min at 60°C
- extension for 1 min/kb at 72°C

1 × cycle - extension for 5 mins at 72°C

Primers that failed to amplify satisfactorily were evaluated over a range of $MgCl_2$ concentrations (0.5mM - 2.0mM), and in the presence of 1.0M - 1.7M aqueous Betaine (B0300; Sigma-Aldrich, Poole, UK).

2.2.4.2 Agarose gel electrophoresis

UltraPure™ Agarose (16500500; Thermo Fisher Scientific, Waltham, USA) was dissolved in 1× Tris-acetate-EDTA (TAE) buffer (ICR Laboratory Support Services) in a microwave oven. To make 1% agarose gel, 1 agarose was dissolved in 100ml TAE buffer. 2µl 10mg/ml solution of Ethidium Bromide (E1510; Sigma-Aldrich, Poole, UK) was added per 100ml gel. 1× loading buffer (B7024S; New England Biolabs, Beverly, MA) was added to each sample (*e.g.* PCR end reaction) prior to gel loading. 100bp ladder (N3231S) or 1kb DNA ladder (N3232S; New England Biolabs, Beverly, MA) were used as size standards. Electrophoresis was carried out in 1x TAE with a PowerPac™ Basic Power Supply (Bio-Rad, South San Francisco, USA) at 100V for 45 minutes or until all the products had migrated a sufficient distance to resolve the bands. After electrophoresis the gel was visualised by transillumination under ultra-violet light at 340nm and a photographic record was made using a Gel Doc-It Imagining Ultraviolet Trans-Illuminator system (BioImaging Systems, Cambridge, UK).

2.2.5 Sanger sequencing

2.2.5.1 Generation and preparation of sequencing template

Target DNA was amplified with PCR reaction (Section 2.2.4.1). After amplification, 5µl of each product was visualised by agarose gel electrophoresis to confirm successful amplification (Section 2.2.4.2). Removal of unincorporated primer oligonucleotides and dNTPs was performed using ExoSAP-IT PCR cleanup reagent (Exonuclease I, Shrimp Alkaline Phosphatase; 78200; USB, Ohio, USA). A combination of two hydrolytic enzymes, the Exonuclease 1 removes unincorporated single-stranded primers and shrimp alkaline phosphatase (SAP) removes

unincorporated nucleotides. One unit of ExoSAP enzyme was added to 5µl PCR product. The restriction digestion conditions were:

1× cycle - digestion for 30 mins at 37°C
1× cycle - ExoSAP inactivation for 15 mins at 80°C

2.2.6 Cycle sequencing reaction

Direct DNA sequencing was undertaken using fluorescent "terminator dyes" attached to each of the four dideoxynucleotides (ddNTPs). When these ddNTPs are incorporated into a PCR product during amplification, they prevent further DNA chain elongation [235]. By chance, at least one terminator nucleotide will be incorporated at each base position in a target sequence during amplification, resulting in a population of PCR products differing in size by one base pair. This mixture can then be separated by electrophoresis and the products excited by a laser. As the dye attached to each nucleotide fluoresces at a different wavelength, the end nucleotide of each PCR product can be identified and the sequence determined.

Sequencing reactions were performed in a total reaction volume of 10µl consisting of 0.5µl BigDye version 3.1 (containing BigDye terminator fluorescently-labeled ddNTPs, dNTPs, AmpliTaq® DNA polymerase, $MgCl_2$ and reaction buffer), 2µl BigDye version 3.1 5 × Sequencing buffer (Applied Biosystems, Foster City, CA, USA), 0.5µl 5µM sequencing primer, 1.5µl exosapped PCR product and 5.5µl RNAse/DNase-free deionised water (dH_2O), placed in a 96-well microtitre plate and covered with an adhesive lid. DNA was amplified using the following cycling conditions:

1 x cycle - denaturation for 5 minutes at 96°C
25 x cycles - denaturation for 30 seconds at 96°C
- annealing for 15 seconds at 50°C
- extension for 1 minute at 60°C

2.2.6.1 Clean-up of sequencing reaction

The sequencing products were purified by adding 1.2µl 125 mM EDTA (pH 8.0) (ICR Laboratory Support Services), 1.2µl 3 M sodium acetate, and 30µl 100% ethanol and incubating at room temperature for 15 minutes to precipitate the amplified DNA. The reaction mixture was centrifuged at 2600×g for 30 mins at 4°C, to pellet the DNA, and the supernatant removed by

centrifuging upside-down on tissue paper at 180×g for 1 min at 4 °C. Residual unincorporated primers and nucleotides were removed by the addition of 42µl 70% ethanol to each sample followed by centrifugation at 1,650×g for 15 mins at 4°C. The supernatant was removed by centrifuging upside-down on tissue paper as before, followed by incubating for 15 mins at 37°C to dry the DNA pellet. Samples were re-suspended in 12µl formamide Hi-Dye (Applied Biosystems, Foster City, CA, USA) and analysed using the ABI 3730 Automated Fluorescent DNA Sequencer (Applied Biosystems, Foster City, CA, USA). Resulting chromatograms were analysed using the Sequencher™ version 4.8/ build 3767 software package (Gene Codes, Ann Arbor, USA).

2.2.7 *In situ* promoter capture Hi-C

In situ promoter capture Hi-C (CHi-C) was previously conducted in the MM cell line KMS11 to determine the 3D architecture of the genome. Briefly, *in situ* Hi-C libraries were prepared as previously described [236] and as depicted in **Figure 2.3**. To increase cell lysis efficiency, three aliquots of 8 million cells were fixed separately in 1% v/v formaldehyde for 10 mins, each aliquot lysed in 15 ml Hi-C lysis buffer (10mM Tris-HCl, pH8.0, 10mM NaCl,, 0.2% Igepal, 1× protease inhibitor) and incubated on ice for 1 hour before being combined. Cross-linked DNA was digested by restriction enzyme HindIII (R0104L, New England Biolabs, Ipswich, US). Digested chromatin ends were filled and marked with biotin-14-dATP (19524016, Thermo Fisher Scientific, Waltham, US). The resulting blunt end fragments were ligated at 16°C in the nucleus with T4 DNA ligase (M0202L, New England Biolabs, Ipswich, US). DNA purified after crosslinking was reversed by Proteinase K (P8102, New England Biolabs, Ipswich, US) treatment. DNA was sheared by sonication (E220, Covaris, Massachusetts, USA) and approximately 200bp - 650bp fragments were selected. Biotin tagged DNA was pulled down with streptavidin beads and ligated with Illumina paired end adapters. Six cycles of PCR were performed to amplify libraries before capture. Promoter capture was based on 32,313 biotinylated 120-mer RNA baits (5190-4396, Agilent, Santa Clara, USA) targeting both ends of HindIII restriction fragments that overlap Ensembl promoters of protein-coding, non-coding, antisense, snRNA, miRNA and snoRNA transcripts. After library enrichment, a post-capture PCR step was carried out using 5 amplification cycles. Hi-C libraries were sequenced using Illumina HiSeq 2000 technology. Reads were aligned to the GRCh37 build using bowtie2 v2.2.6 and identification of valid di-tags was performed using HICUP v0.5.9 [237]. To declare significant contacts, HiCUP output was processed using CHiCAGO v1.1.8 [238] (Section 2.4.3). Data from three independent biological replicates were combined to obtain a definitive set of contacts. A CHiCAGO score > 5 was taken to indicate a significant contact [238]. The interaction of MM risk SNPs with *APOBEC* genes in

Chapter 4 was plotted on WashU Epigenome Browser [239]. Promoter capture Hi-C on GM12878 used in Chapter 4 was obtained from publicly available EMBL-EBI: E-MTAB-2323 [240].

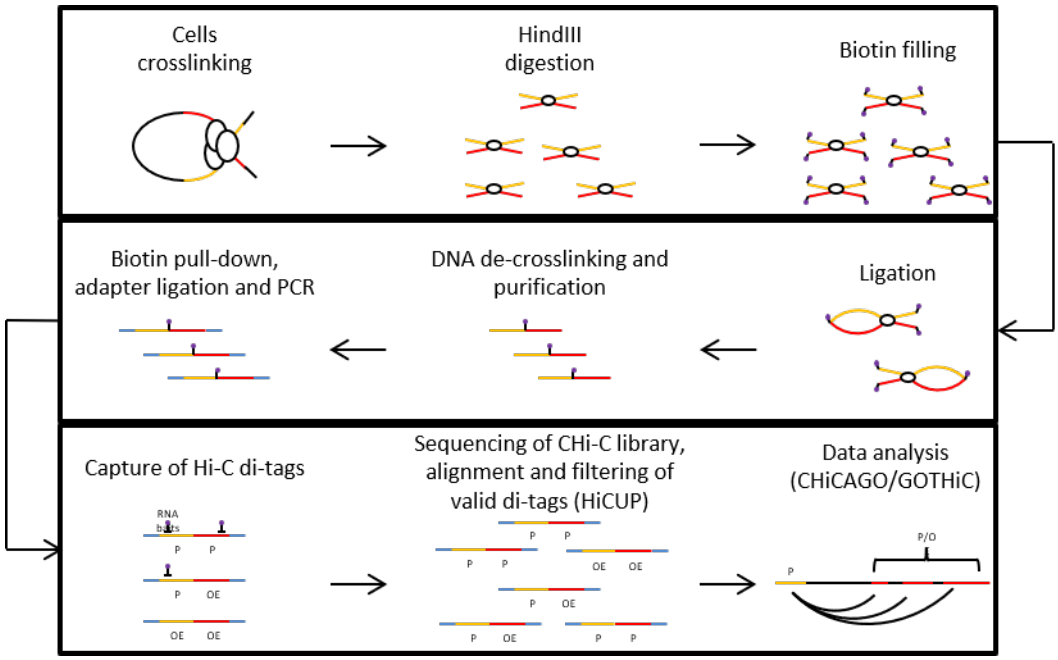


Figure 2.3 General work-flow for CHi-C library generation and analysis. Main steps for the generation of CHi-C libraries, illustrating the comprehensive of the Hi-C protocol and the capture of selected di-tags using custom RNA baits (adapted from Orlando *et al* [241]).

2.3 Statistical analyses

Statistical analyses were primarily performed using PLINK v1.9 [242], R v3.1.3 [243] and custom perl scripts.

2.3.1 Quality control in association studies

2.3.1.1 Software

PLINK

PLINK is a whole-genome analysis tool which can perform a range of large scale analyses on genomic data [242]. Specifically the following quality control steps in the GWAS analysis pipeline were performed using PLINK:

- Test for deviation of genotype frequencies from Hardy-Weinberg Equilibrium (HWE)
- Verifying that genetic sex as estimated from X-heterozygosity is consistent with phenotypic sex
- Calculating sample and SNP missing rates
- Tests for significant differences in missing data between cases and controls
- Calculation of pairwise identity by state

R

R v3.1.3 [243] is a publicly available software environment for statistical computing and graphics [243]. Installation of packages allows a large number of statistical and bioinformatics techniques to be performed.

2.3.1.2 Identity-by-state analysis

Using PLINK, genotype data was analysed to search for duplicates and closely related individuals within and between datasets. Identity-by-state (IBS) values were calculated based on at least 2,000 shared markers for each pair of individuals. For any pair with distance > 0.2 (a threshold estimated to remove all first-degree relatives) the control from a case-control pair was removed; otherwise, the individual with the lower call rate was removed from further analysis.

2.3.1.3 Quantile-quantile plots

Quantile-quantile (Q-Q) plots were used to assess the adequacy of case-control matching and the possibility of differential genotyping of cases and controls by comparing the distribution of observed test statistics against that of values expected under a null hypothesis. A Q-Q plot is a probability plot comparing two probability distributions by plotting their quantiles against each

other. The highest observed value is plotted against the highest expected value. Deviation from the line *y=x* implies a consistent difference between cases and controls, suggesting bias by a confounder. The comparatively few variants with much higher observed than expected values are assumed to represent true associations. The inflation factor λ was calculated by dividing the median of the test statistics by the median expected values from a χ2 distribution with 1 degree of freedom (df). The inflation factor λ was based on the 90% least-significant SNPs [244]. Since λ scales with sample size, λ_{1000} can be calculated for an equivalent study of 1,000 cases and 1,000 controls by rescaling λ [245].

2.3.1.4 Principal components analysis

Principal components analysis (PCA) was used to identify and remove individuals with large scale differences in ancestry (*i.e.* not of European ancestry). The *smartpca* package, part of EIGENSOFT v4.2 [246], was used to perform PCA [247, 248]. For each dataset, SNP data was merged with 60 European (CEU), 60 Nigerian (YRI), 90 Japanese from Tokyo (JPT) and 90 Han Chinese (CHB) individuals from the Phase II HapMap project. Due to large genetic differences between these three ancestral groups, a plot of the first two principal components was sufficient to identify any individual not within the main CEU cluster and exclude from further analyses.

2.3.1.5 Linkage disequilibrium-based SNP pruning

Certain analyses, such as PCA require a set of uncorrelated SNPs. These were estimated in PLINK using the --indep flag.

2.3.1.6 Hardy-Weinberg equilibrium

The Hardy-Weinberg principle states that the allele and genotype frequencies in a population will remain constant from generation to generation in the absence of evolutionary influences [249]. At a single locus with two alleles denoted A and B with frequencies $f(A)=p$ and $f(B)=q$, respectively, expected genotype frequencies are $f(AA)=p^2$, $f(BB)=q^2$ and $f(AB)=2pq$ for the AA homozygote, BB homozygote and AB heterozygote respectively. The sum of all genotype frequencies must equal 1 (*i.e.* $p^2 + 2pq + q^2 = 1$). If a genetic locus satisfies this equation it is said to be in Hardy-Weinberg equilibrium (HWE). The χ^2-test was used to assess genotype frequencies in controls for evidence of departure from HWE [249], which may indicate population stratification. $P < 1 \times 10^{-5}$ was considered to be out of HWE.

2.3.2 Assessing statistical significance

The *P*-value is defined as the probability of obtaining a value that is at least as extreme as that of the actual sample by chance. If the *P*-value is smaller than a pre-set threshold then the null hypothesis of no association is rejected and the result is considered significant. For a single test $P < 0.05$ is deemed significant in order to control the family wise error rate (FWER; the probability of making even one type I error) at 0.05. However, in GWAS where many SNPs are being tested simultaneously keeping the threshold for significance at 0.05 would lead a large number of false positives (if the null hypothesis is correct for 1,000,000 SNPs tested, then 5%, that is 50,000 SNPs, are expected to have $P < 0.05$ by chance).

To minimise type I error and keep the FWER at 0.05, a Bonferroni correction of the *P*-value can be applied. The corrected *P*-value is given by the equation $P = \alpha/n$, where α equates to the generally accepted level of significance (0.05) and n to the number of polymorphisms genotyped. This is likely an overcorrection when multiple tests can be correlated. Simulations generating an infinitely dense set of polymorphisms identified a *P*-value cut off of 5×10^{-8} as appropriate in genome-wide studies [250-252].

In the GWAS conducted in Chapter 3, the threshold for statistical significance in GWAS was taken to be $P < 5 \times 10^{-8}$. Additional analyses were explicitly corrected according to the number of tests carried out unless stated otherwise.

2.3.2.1 Bayesian false-discovery probability

A Bayesian approach to the false discovery rate, in which the probability of the null hypothesis (H_0) is true, given the observed data, can be used to interpret findings from a GWAS. Calculation of the Bayesian false-discovery probability (BFDP) requires the *P*-value, the study power (Section 2.3.3) and an estimate of the prior for a given SNP. This BFDP statistic is complementary to the *P*-value, which tests the probability of the data, given the null hypothesis (H_0) is true. For borderline associations identified in the GWAS performed in Chapter 3, the BFDP was calculated using the methodology of Wakefield *et al* [253]. The BFDP was calculated based on a plausible OR of 1.2 (based on the 95th percentile of meta-analysis OR values) and a prior probability of association of 0.0001.

2.3.3 Calculation of study power

Study power is defined as the probability of rejecting the null hypothesis (H_0) of no association when the alternative hypothesis (H_1) is true [106]. When the probability of rejecting H_0 when it is true (type 1 error) is α, and the failure to reject H_0 when it is false (type 2 error) is β, statistical power is defined as 1-β. Whilst α can be controlled by the study investigator, β is subject to factors outside the investigators control such as the effect size and frequency of the associated variant, the level of correlation between the typed marker and the true causal variants and the underlying disease model. Study power can be maximised by increasing the number of samples studied and using cases genetically enriched for disease susceptibility (*i.e.* early onset and with family history of disease).

2.3.4 Estimating linkage disequilibrium

During meiotic recombination, stretches of DNA are non-independently co-inherited. SNPs at different sites in the genome are not randomly inherited; they are strongly correlated and likely to co-segregate together in a haplotype. This non-random association of alleles, termed linkage disequilibrium (LD), allows certain SNPs to act as proxies, or tag SNPs, for correlated SNPs. This reduces the number of SNPs that need to be genotyped to capture most common variants (that is, those with a minor allele frequency > 5%) to around 300,000.

The most common metrics of LD are *D'* and the correlation coefficient (r^2). *D'* varies between 0 and 1 with a value of 1 corresponding to complete LD. Values less than one indicate disrupted LD and have no clear statistical interpretation particularly as *D'* is strongly inflated in small sample sizes and only measures recombination history. Therefore, intermediate values should not be used to measure the extent of LD. The more stable r^2 is the preferred measure of the extent of LD as it summarises both the recombination and the mutational history of the markers [254, 255]. The r^2 statistic is equal to *D'* divided by the product of the allele frequencies at the two loci with perfect LD indicated by $r^2 = 1$.

2.3.5 Haploview

Haploview v4.2 is a Java software package used to compute LD and haplotype blocks [256]. It uses primary genotype data and data derived from publicly available databases. In addition to LD statistics and haplotype blocks, Haploview also generates marker quality statistics, population haplotype frequencies and single marker association statistics. It can also be used for haplotype association analysis.

2.3.5.1 SNAP

SNP Annotation and Proxy (SNAP) search is an annotation tool used to find proxy SNPs based on LD, physical distance and/or presence on commercial genotyping platforms [257]. SNAP implements Haploview 4.0 to calculate pairwise r^2 and D' measures of linkage disequilibrium based on data from the 1000 Genomes pilot project. This allows the user to generate plots of regional LD and query pair-wise LD metrics. Furthermore, the tool includes annotation information from multiple commercial arrays, so can be used to check for SNPid aliases across dbSNP builds. Plots can be created using publicly available R code.

2.3.5.2 The International HapMap project

The International HapMap project [258] provides a haplotype map of the human genome [259]. By cataloguing the common genetic variants in the human genome, the resource contains high-density SNP genotype data from individuals across four different populations (Caucasian, Chinese, Japanese, and African).

2.3.5.3 VCFtools

VCFtools is a package of programs for working with VCF files [260]. It can be used to compute r^2 and D' metrics when supplied with phased haplotype data (*e.g*. from the 1000 Genomes project or UK10K project).

2.3.6 Imputation

Imputation is a method used in GWAS to extrapolate genetic data from a densely characterised reference panel to a sparsely typed sample set [161, 261]. This method predicts (or 'imputes') the genotypes at untyped variants in each individual using a reference panel of known haplotypes sequenced or genotyped at a dense set of variants. Imputation methods identify stretches of shared haplotype in the reference dataset and study dataset and fill in missing genotypes in the study dataset by copying alleles observed in the matching haplotype of the reference dataset.

2.3.6.1 Imputation reference panels

Accuracy of imputation is critically dependent on the reference panel used to infer missing genotypes, while identification of risk loci is dependent on the extent to which disease-causing

variation is catalogued within this reference panel. The GWAS performed in Chapter 3 utilised a merged reference panel, combining data from the 1000 Genomes and UK10K projects.

1000 Genomes project

Haplotype data from 1,092 individuals from Africa (n=246), Asia (n=286), Europe (n=379) and the Americas (n=181) produced in phase one of the project are available as a reference panel for imputation [262]. Only data from those of European ancestry were used in the GWAS in Chapter 3.

UK10K project

Phased haplotype data from 3,781 UK individuals was additionally available from the UK10K project for use as an imputation reference panel [110].

2.3.6.2 SHAPEIT

Generally, genotyping data is unphased; it is not directly observed which of the two parental chromosomes, or haplotypes, a particular allele falls on. Hence which alleles are co-localised on the same chromosome is unknown. This information is essential for imputation algorithms which look for haplotypes which are shared between a study dataset and the reference data. Segmented HAPlotype Estimation and Imputation Tool (SHAPEIT) is a fast and accurate method of estimating haplotypes from genotype data based on a Hidden Markov Model (HMM) [263]. This method can be used to create a set of phased haplotypes from genotype data which can subsequently be used as a reference panel for imputation. GWAS data can also be pre-phased, greatly increasing imputation speed.

2.3.6.3 IMPUTE 4

IMPUTE 4 is a statistical program for imputing unobserved genotypes in SNP association studies utilising a reference panel of known haplotypes such as the 1000 Genomes project [264]. It uses an approximate population genetics model that gives more weight to genotypes that are consistent with the local patterns of LD. Use is made of information from all markers in LD with an untyped SNP in a way that decreases with genetic distance from the SNP being imputed. Marginal probabilities of each possible genotype for each unknown genotype under study are output, allowing for uncertainties in prediction to be taken into account in association testing. In addition, associated 'information' metrics are produced for each SNP reflecting the certainty of imputation, where a quality score of 1 corresponds to a near perfectly imputed SNPs [261].

To guard against poorly imputed SNPs, only SNPs with an info score > 0.8 were retained. IMPUTE 4 implements the haploid imputation options included in IMPUTE 2, but is much faster and more memory efficient.

2.3.7 SNPTEST

SNPTEST v2.5 is a program for the analysis of single SNP associations in GWAS [265]. It takes in the marginal probability output from IMPUTE and allows for conditioning on covariates. SNPs can be tested for association assuming an additive, dominant, recessive, general or heterozygote model. As with IMPUTE, an information score is output to reflect confidence in the imputation.

2.3.8 META

META is a program for meta-analysis of GWAS studies which reads in the output from SNPTEST [266]. Along with *P*-values of association, for each SNP META outputs Cochran's *Q* and I^2 statistics of heterogeneity between datasets. A threshold imputation information score can be specified whereby for each SNP, only studies passing this threshold will be included in the *P*-value estimate. Three different methods can be used to combine *P*-values; an inverse variance method under a fixed effects model, an inverse variance under a random effects model or a z-statistics combination method based on a fixed effects model.

Meta-analyses were undertaken to obtain pooled estimates using the Mantel-Haenszel method to combine raw data. Joint ORs and 95% confidence intervals (CIs) were calculated using META v1.7 assuming an inverse variance weighted, fixed-effects model, and tests of the significance of the pooled effect sizes were calculated using a standard normal distribution. Cochran's Q statistic to test for heterogeneity and the I^2 statistic (where $I^2 = ((Q - df)/Q)*100$ and *df* is number of studies -1) to quantify the proportion of the total variation due to heterogeneity were calculated. I^2 values of ≥ 75% can be considered characteristic of large heterogeneity [267, 268].

2.3.9 Association analyses

2.3.9.1 Conditional analyses

Conditional analyses were carried out to rule out the existence of multiple statistical signals at each risk locus. Association statistics were calculated for all SNPs conditioning on the top SNP in each locus showing genome-wide significance. The genome-wide complex trait analysis (GCTA) program [269] was used, employing the conditional and joint genome-wide association analysis tool with summary statistics.

2.3.9.2 Subtype analysis

FISH and ploidy classification of UK and German samples had previously been conducted by UK MyIX and MyXI trials and the German GMMG trial (Section 2.1.1) [270-272]. The XL *IGH* Break Apart probe (MetaSystems, Altlussheim, Germany) was used to detect any *IGH* translocation in the German samples. Association between SNP genotype and MM risk under a variety of molecular subtypes was tested under a logistic regression in case-only and case-control analyses.

2.3.9.3 Age and sex association analysis

In Chapter 3, association between sex and genotype for the top SNP at each of the genome-wide significant regions was conducted with logistic regression, and with linear model for the association between age at diagnosis and genotype. All individuals in five of the six sample sets were used (UK n = 2,282, German n = 1,508, US n = 780, Sweden/Norway n = 1,714, Netherlands n = 555).

2.4 Bioinformatic analysis

2.4.1 Databases

2.4.1.1 University of California, Santa Cruz (UCSC) genome

The University of California, Santa Cruz (UCSC) genome browser (http://genome.ucsc.edu/) is a virtual map of the human genome, annotated with known genes, transcripts, polymorphic variation, repeated sequences, conservation, structural variation and experimental data from external databases such as ENCODE (Section 2.4.1.3). These features are mapped against their physical positions in the genome. Various bioinformatics tools are contained within the website and were utilised as follows:

- *Genome Browser* tool was used to query specific regions of DNA and visualise genes, introns, regulatory elements and other features of the genomic location.
- *BLAT* tool was used to assess the binding accuracy of primers designed for PCR by finding possible spurious binding sites with > 95% similarity to the sequence of interest.
- *LiftOver* tool was used to convert genome coordinates between different genome assemblies. Specifically, early GWAS SNPs may be mapped to NCBI Build 36 (hg18) whereas sequencing reads are mapped to the more recent Build 37 (hg19).

- *Table Browser* tool was used to download data associated with specific tracks in the genome browser.

2.4.1.2 National Centre for Biotechnology Information

The National Centre for Biotechnology Information (NCBI) web server (http://www.ncbi.nlm.nih.gov/) hosts a multitude of databases and bioinformatics tools [273]. Specific tools used in this work are:

- PubMed for literature searches and citations.
- Basic Local Alignment Search Tool (BLAST) for nucleotide based searches.
- RefSeq to obtain reference sequences of chromosomes, genomic contigs, mRNAs and proteins. These data can also be queried in UCSC.
- dbSNP database of short genetic variations to query specific SNPs for position, allele and frequency information.

2.4.1.3 The Encyclopedia of DNA Elements

The encyclopedia of DNA elements (ENCODE) [274, 275] aims to build a comprehensive list of functional elements in the human genome, including elements that act at the protein and RNA level, as well as DNA regulatory elements. The ENCODE project integrates genome-wide experimental data for over 100 different cell types. Data includes: chromatin structure (*e.g*. HiC), open chromatic prediction (*e.g*. DNase hypersensitivity), histone modifications and transcription factor binding prediction (ChIP-seq) and RNA transcription (RNA-seq and CAGE). All data is publicly available for download and can be viewed in the UCSC genome browser (Section 2.4.1.1). From this data the functionality of specific genomic regions can be inferred which is critical in fine-mapping studies and prioritisation of sequence variants.

2.4.1.4 1000 Genomes project

The 1000 Genomes Project aims to provide a comprehensive catalogue of human genetic variation with frequencies > 1% through sequencing large numbers of individuals at 4x coverage [109]. Combining data from all individuals will then allow for accurate imputation of variants not directly covered in this low coverage sequencing. Data from the pilot phase, phase one and phase three of the project have been made publicly available. It is currently the largest publicly available resource for genome-wide variant frequency data across different populations

worldwide. Variant data from individuals of European descent in the 1000 Genomes project were used, as part of a reference panel for imputation (Chapter 3).

2.4.1.5 UK10K project

The UK10K project aims to sequence 10,000 phenotyped people at 6x coverage in order to better understand the link between low-frequency and rare genetic changes and human disease [110]. The 10,000 individuals are split into three cohorts; the Twins UK and ALSPAC cohorts comprise 1,854 and 1,927 whole-genome sequenced individuals respectively and a further 6,000 individuals with extreme health problems (neurodevelopment, obesity and rare diseases) are to be exome sequenced. It is currently the largest publicly available resource for variant frequency data in the UK population. Variant data from the Twins UK and ALSPAC cohorts (3,781 individuals) were used for haplotype data, as part of a reference panel for imputation (Chapter 3).

2.4.1.6 The International HapMap project

The international HapMap project aims to catalogue all common genetic variants in the human genome across different populations [259]. The resource allows the retrieval of high-density SNP genotype data from large numbers of individuals that are representative of different populations (Caucasian, Chinese, Japanese and African).

2.4.1.7 Ensembl genome browser

The Ensembl genome browser [276] is a genome annotation database supported by the European Bioinformatics Institute. Along with the Ensembl Biomart, it is of particular use for retrieval of gene information including genomic organisation of exons, introns and known regulatory domains, known transcripts, proteins, homologues and recorded variation within the gene sequence [277, 278].

2.4.1.8 WashU Epigenome Browser

WashU Epigenome Browser allows visualisation of chromatin interaction data from a variety of chromatin conformation capture experiments [239]. In Chapter 4, WashU Epigenome browser was used to display chromatin looping interactions.

2.4.1.9 Blueprint

Blueprint is an epigenomic project specifically focusing on hematopoiesis [179]. It aims to generate epigenomic maps of a wide variety of cell types from blood, including both primary human cells from healthy individuals and blood-based diseases. A range of epigenomic and regulatory data is available including RNA-seq, DNase-seq, ChIP and histone modifications [179]. Blueprint data is available from other data portals including the Ensembl Biomart and UCSC browsers.

2.4.2 Expression quantitative trait Locus (eQTL) analysis

eQTL analysis aims to find associations between SNP genotypes and expression levels of genes in *cis*, by assessing mRNA levels for samples for which SNP genotype data are also available. In this thesis the following datasets were used for eQTL analysis:

eQTL in MM patients

eQTL analyses were performed using Affymetrix Human Genome U133 2.0 Plus Array data for plasma cells from 183 MRC Myeloma IX trial patients, 658 Heidelberg patients and 608 US patients (Section 2.1.4). German, UK and US data was separately pre-processed and analysed using a Bayesian approach to probabilistic estimation of expression residuals to infer broad variance components, thus accounting for hidden determinants influencing global expression such as copy number, translocation status and batch effects [279]. The association between genotype of the sentinel variant and gene expression of genes within 500 kb either side was evaluated by linear regression. Data from each study cohort was pooled under a fixed-effects model, controlling for FDR, and calling significant associations with a FDR ≤0.05.

Genotype-Tissue Expression Consortium (GTEx)

The GTEx (Genotype-Tissue Expression) Consortium is a tissue biobank for gene-expression levels across individuals for diverse tissues of the human body, with a broad sampling of normal, non-diseased human tissues from postmortem donors [280]. The GTEx project includes publicly available genotype, gene expression, histological and clinical data for 491 human donors across 48 tissues. This enables the study of tissue-specific gene expression and the identification of genetic associations with gene expression levels (expression quantitative trait loci, or eQTLs) across many tissues, including both local (*cis*-eQTLs) and distal (*trans*-eQTLs) effects. For association analysis of predicted gene expression with MM risk in Chapter 4, SNP weights and

their respective covariance in 48 tissues from 80 to 491 individuals were obtained from predict.db, which is based on GTEx version 7 [280] eQTL data.

2.4.3 Hi-C analysis

Recent studies making use of techniques such as Hi-C have led to increased insights into three-dimensional genome structure [236]. More recently, "topologically associating domains" (TADs) have been identified as megabase-sized local chromatin interaction domains, which are stable across different cell types and conserved across mammalian species [281]. Defining these interaction domains can aid in interpretation of DNA sequence function.

Hi-C data was used to map the candidate causal SNPs to chromosomal TADs and identify patterns of relevant, local chromatin interactions using a range of cell lines. Hi-C data for the KMS11 cell line, used in Chapter 3, was generated in-house by Dr Scott Kimber. GM12878 Hi-C data used in Chapter 4 was publicly available [240]. Valid Hi-C pairs were generated aligning raw reads to the reference genome using Burrows-Wheeler alignment (BWA) [282], matching pairs of reads and filtering for biases. *Bona fide* Hi-C ditags were allocated to a contact matrix, with a predefined, uniform resolution of 5 kb. Experimental bias was corrected using the matrix balancing approach [283]. TADs were inferred from the contact matrix by means of the arrowhead algorithm for domain detection [236]

2.4.4 Heritability estimation

Linkage disequilibrium adjusted kinships (LDAK) [284] was used to estimate the polygenic variance (*i.e.* heritability) ascribable to all genotyped and imputed GWAS SNPs from summary statistic data. SNP-specific expected heritability, adjusted for LD, MAF and genotype certainty was calculated from the UK10K [110] and 1000 Genomes [109] data. Samples were excluded with a call rate <0.99 or if individuals were closely related or of divergent ancestry from CEU. Individual SNPs were excluded if they showed deviation from HWE with $P < 1 \times 10^{-5}$, an individual SNP genotype yield <95%, MAF <1%, SNP imputation score <0.99 and the absence of the SNP in the GWAS summary statistic data. This resulted in a total 1,254,459 SNPs which were used to estimate the heritability of MM.

To estimate the sample size required for a given proportion of the GWAS heritability, a likelihood-based approach was implemented to model the effect-size distribution using association statistics, from the MM meta-analysis, and LD information, obtained from

individuals of European ancestry in the 1000 Genomes Project Phase 3 [109]. LD values were based on an r^2 threshold of 0.1 and a window size of 1 MB. The goodness of fit of the observed distribution of *P*-values against the expected from a two-component model (single normal distribution) and a three-component model (mixture of two normal distributions) were assessed [285]. The percentage of GWAS heritability explained for a projected sample size was determined using this model and is based on power calculations for the discovery of genome-wide significant SNPs. The genetic variance explained was calculated as the proportion of total GWAS heritability explained by SNPs reaching genome-wide significance at a given sample size. The 95% confidence intervals were determined using 10,000 simulations.

2.4.5 Summary-data-based Mendelian Randomisation (SMR)

The relationship between SNP genotype and gene expression was carried out using SMR analysis as per Zhu *et al* [181]. Briefly, if b_{xy} is the effect size of *x* (gene expression) on *y* (slope of *y* regressed on the genetic value of *x*), b_{zx} is the effect of *z* on *x* and b_{zy} be the effect of *z* on *y*, b_{xy} (b_{zy}/b_{zx}) is the effect of *x* on *y*. To distinguish pleiotropy from linkage where the top associated *cis*-eQTL is in LD with two causal variants, one affecting gene expression and the other affecting a trait, heterogeneity in dependent instruments (HEIDI) was tested for using multiple SNPs in each *cis*-eQTL region. Under the hypothesis of pleiotropy, b_{xy} values for SNPs in LD with the causal variant should be identical. For each probe that passed significance threshold for the SMR test, the heterogeneity in the b_{xy} values estimated for multiple SNPs in the *cis*-eQTL region was tested using HEIDI.

GWAS summary statistics files were based on the meta-analysis performed in Chapter 3. A threshold for the SMR test of $P_{SMR} < 1 \times 10^{-3}$ was set for the analysis performed in Chapter 3, corresponding to a Bonferroni correction for 45 tests, *i.e.* 45 probes which demonstrated an association in the SMR test. For the study of genetic correlation between MM and CLL in Chapter 5, a threshold for the SMR test of $P_{SMR} < 2.5 \times 10^{-5}$ was set. For all genes passing this threshold, we generated plots of the eQTL and GWAS associations at the locus, as well as plots of GWAS and eQTL effect sizes (*i.e.* input for the HEIDI heterogeneity test). HEIDI test *P*-values < 0.05 were considered as reflective of heterogeneity. This threshold is, however, conservative for gene discovery because it retains fewer genes than when correcting for multiple testing.

2.4.6 Transcriptome imputation

MetaXcan is a suite of tools to perform integrative gene mapping studies. Amongst these, PrediXcan is a gene-level association approach that tests the effects of gene expression levels on phenotypes [286]. PrediXcan imputes transcriptome levels with models trained in measured transcriptome datasets. These predicted expression levels are then correlated with the phenotype in a gene association test. Associations between predicted gene expression and MM risk were examined using PrediXcan, which combines GWAS and eQTL data, accounting for LD-confounded associations. Briefly, genes likely to be disease-causing were prioritised using S-PrediXcan which uses GWAS summary statistics and pre-specified weights to predict gene expression, given co-variances of SNPs. SNP weights and their respective covariance in 48 tissues from 80 to 491 individuals were obtained from predict.db which is based on GTEx version 7 eQTL data [280]. A full list of the sample count by tissue can be found at the GTEx portal [287]. To combine S-PrediXcan data across the different tissues taking into account tissue-tissue correlations, S-MultiXcan [288] was used.

To determine if associations between genetically predicted gene expression and MM risk were influenced by variants previously identified by GWAS, conditional analyses were performed adjusting for sentinel GWAS risk SNPs using GCTA-COJO [269]. Adjusted output files were provided as the input GWAS summary statistics for S-PrediXcan analyses as above. To account for multiple comparisons, a Bonferroni-corrected *P*-value threshold of 1.96×10^{-6} (*i.e.* 0.05/25,520 genes) was considered as being statistically significant.

2.4.7 Transcription factor and histone mark enrichment analysis

To examine enrichment in specific TF binding across risk loci, the method of Cowper-Sal lari *et al* [289] was adapted. Briefly, for each risk locus, a region of strong LD (defined as $r^2 > 0.8$ and D' > 0.8) was determined, and these SNPs were considered the associated variant set (AVS). Publicly available data on TF ChIP-seq uniform peak data was obtained from ENCODE [274] for the GM12878 cell line, including data for 82 TF and 11 histone marks. In addition, ChIP-seq peak data for six histone marks from KMS11 cell line were generated in-house, and naïve B-cell ChIP-seq data was downloaded from Blueprint Epigenome Project [179]. For each mark, the overlap of the SNPs in the AVS and the binding sites was assessed to generate a mapping tally. A null distribution was produced by randomly selecting SNPs with the same characteristics as the risk-associated SNPs, and the null mapping tally calculated. This process was repeated 10,000 times, and *P*-values were calculated as the proportion of permutations, where null mapping tally was

greater or equal to the AVS mapping tally. An enrichment score was calculated by normalising the tallies to the median of the null distribution. Thus, the enrichment score is the number of standard deviations of the AVS mapping tally from the median of the null distribution tallies.

2.4.8 Estimation of genetic correlation using LD score regression

To investigate genetic correlation between MM and CLL (Chapter 5), cross-trait LD score regression by Bulik-Sullivan *et al* [290] was implemented. This method is an extension of single trait LD score regression; it estimates genetic correlation using only GWAS summary statistics and is not biased by sample overlap. Summary statistics from the CLL and MM GWAS meta-analysis were used and filters as recommended by the authors were applied. Specifically, filtering SNPs to INFO > 0.9, MAF > 0.01, and harmonizing to Hap-Map3 SNPs with 1000 Genomes EUR MAF > 0.05, removing indels and structural variants, removing strand-ambiguous SNPs and removing SNPs where alleles did not match those in 1000 Genomes. This was performed by running the munge-sumstats.pr script included with ldsc. The script, ldsc.py, part of the ldsc package was run, excluding the HLA region. Heritability estimates are reported on the observed scale. There is no distinction between observed and liability scale genetic correlation for case/control traits.

2.4.8.1 Stratified LD score regression

A variation of LD score regression, namely stratified LD score regression, can be used to partition heritability according to different genomic categories. For both MM and CLL, stratified LD score regression was applied across the baseline model used in Finucane *et al* [291]. Enrichment of functional categories was plotted for each disease and is defined as the proportion of heritability divided by the total heritability. Additional flanking regions around each functional category, which authors designed to allow observation of enrichment of SNP heritability in intermediary regions, were excluded from plots.

2.4.9 Chromatin state annotation

ChromHMM is a software for learning and characterizing chromatin states [180]. Multiple genomic datasets (*e.g.* ChIP-seq, histone marks) are integrated into a hidden Markov model that models the presence or absence of each chromatin mark to demarcate the genome into a defined number of states corresponding to different biological functions (*e.g*. active promoter, strong enhancer or repetitive). This inference of regulatory elements aids in interpretation of SNP effect. Variant sets (*i.e.* sentinel risk SNP and correlated SNPs, $r^2 > 0.8$) were annotated for

putative functional effect based upon histone mark ChIP-seq data for H3K27ac, H3K4me1, H3K27me3, H3K9me3, H3K36me3 and H3K27me3 from KMS11 cell lines, generated in-house and naïve B-cells from Blueprint Epigenome Project [179] using ChromHMM. The software package was used to infer chromatin states by integrating information on these histone modifications, training the model on three MM cell lines; KMS11, MM1S and JJN3. Genome-wide signal tracks were binarised (including input controls for ChIP-seq data), and a set of learned models were generated using ChromHMM software. A 12-state model was suitable for interpretation, and biological meaning was assigned to the states based on chromatin marks that use putative rules as previously described [176, 292, 293].

2.4.10 Cell-type-specific analyses

In Chapter 5, chromatin mark overlap enrichment for genome-wide significant loci in different cell types was performed using the methodology of Trynka *et al* [294]. This approach scores GWAS SNPs based on proximity to chromatin mark and fold-enrichment of respective chromatin mark, assessing significance using a tissue-specific permutation method. ChIP-seq data for H3K4me3 from primary blood cells and CLL samples was downloaded from Blueprint Epigenome project [179]. In addition, 4 MM cell lines, KMS11, JJN3, MM1-S, and L363, were included.

2.4.11 Annotation of regulatory elements

For the integrated functional annotation of risk loci in Chapter 3, variant sets (*i.e*. all SNPs in LD $r^2 > 0.8$ with the sentinel SNP) were annotated with: (i) presence of a Hi-C contact linking to a gene promoter, (ii) presence of an association from SMR analysis, (iii) presence of a regulatory ChromHMM state, (iv) evidence of transcription factor binding and (v) presence of a nonsynonymous coding change. Candidate causal genes were then assigned to MM risk loci using the target genes implicated in annotation tracks (i), (ii), (iii) and (iv). If the data supported multiple gene candidates, the gene with the highest number of individual functional data points was considered as the candidate. Where multiple genes have the same number of data points, all genes are listed. Direct non-synonymous coding variants were allocated additional weighting. Competing mechanisms for the same gene (*e.g.* both coding and promoter variants) were permitted.

2.4.12 Mendelian randomisation analyses

Mendelian randomisation (MR) is an analytical method that exploits genetic variants as instrumental variables (IVs), to infer the causal relevance of an exposure to an outcome, such as

a disease [191]. Because the genetic variants are randomly assigned at conception they are not influenced by reverse causation and in the absence of pleiotropy (*i.e.* genetic variants being associated with the disease through alternative pathways) they can provide unconfounded estimates of disease risk [191]. An agnostic strategy to identify causal relationships has recently been proposed, integrating the phenome-wide association study (PheWAS) and MR methodology, termed MR-PheWAS [295].

2.4.12.1 Genetic instruments for phenotypes

Two-sample MR was conducted using the TwoSampleMR R package [296]. Genetic instruments for each of the traits investigated were SNPs identified from recent meta-analyses, the largest studies published to date, or those curated by MR-Base. For each SNP, the chromosome position, the effect estimate expressed in standard deviations (SDs) of the trait per-allele and the corresponding standard errors (SEs) were recovered. SNPs were only considered as potential instruments if they were associated with each trait at $P < 5 \times 10^{-8}$ in GWAS of European populations and had a minor allele frequency > 0.01. To avoid co-linearity between SNPs for each trait, correlated SNPs within each trait were excluded (LD threshold, $r^2 \geq 0.01$). Only SNPs with the strongest effect on the trait were considered. The proportion of variance explained (PVE) by the associated SNPs were computed from the association statistics. Traits were only considered if the power to identify an OR_{SD} of 0.67 or 1.50 was > 80%. As analysis of binary traits (such as disease status) with binary outcomes in two-sample MR frameworks can result in inaccurate causal estimates, only continuous traits were considered [297].

2.4.12.2 Estimation of study power

The power of MR to demonstrate a causal effect depends on the percentage of risk factor variance explained by the genetic variants used as instruments [107]. The study power was estimated, stipulating an alpha of 0.05, for each risk factor *a priori* across a range of effect sizes.

2.4.12.3 Mendelian randomization analysis

The MR methodology assumes that genetic variants, used as instruments for a risk factor, are associated with the risk factor and not with confounders or alternative causal relationships [191]. Additionally, associations must be linear and unaffected by interactions. For each SNP, causal effects were estimated for MM as an odds ratio per one SD unit increase in the putative risk factor (OR_{SD}), with 95% confidence intervals (CIs), using the Wald ratio. For traits with multiple SNPs as IVs, causal effects were estimated under inverse variance weighted random-

effects (IVW-RE) and inverse variance weighted fixed-effects (IVW-FE) models. To assess the robustness of the findings, weighted median estimates (WME) [298] and mode-based estimates (MBE) [299] were obtained for results which were suggestively significant and had > 2 SNPs included in the analysis. Pleiotropy exists when a single genetic variant influences multiple phenotypes [300]. Horizontal pleiotropy refers to a situation where the genetic instrument influences the disease outcome via a different pathway which is not under investigation. Where pleiotropic effects are balanced and there exists no systematic bias across a set of genetic instruments, MR estimates remain valid. If horizontal pleiotropy is unbalanced (directional) it may result in a biased MR estimate [301]. Directional pleiotropy was therefore assessed using MR-Egger regression [302]. A consistent effect across these four complementary methods (IVW, MBE,WME and MR-Egger), which make different assumptions about horizontal pleiotropy, is less likely to be a false positive [303]. The potential impact of outlying and pleiotropic SNPs on causal estimates were examined by adopting a leave-one-out strategy, under the IVW-RE model [296]. This method performs the MR analysis, but leaves out each SNP in turn to identify whether a single SNP is driving the association. Heterogeneity (I^2) was calculated from Cochran's *Q*-value. To account for multiple correction testing, a Bonferroni-corrected *P*-value was considered as being statistically significant, while a $P < 0.05$ was considered to be suggestive evidence of a causal association. Statistical analyses were performed using R version 3.4.0 and MR-Base [296].

2.4.12.4 Availability of data

Genetic instruments can be obtained through MR-Base [296] or from published work. **Appendix 19** lists the phenotypes which were investigated the MR analysis and details the publications they were obtained from.

2.4.13 Primer design

These programs and tools were used for designing primers in Chapters 3 and 4.

2.4.13.1 Primer 3

Primer 3 [233] is a web-based program for designing PCR oligonucleotide primers allowing user parameter specification. It is essential in design of primers that do not misprime to the human genome.

2.4.13.2 KBioSciences Primer Picker

KBioSciences Primer Picker [304] is a web-based program provided by the manufacturer for designing KASPar SNP genotyping primers.

CHAPTER 3 Identification of risk loci for multiple myeloma

3.1 Overview and rationale

Although no lifestyle or environmental exposures have been consistently linked to an increased risk of MM, the two- to four-fold increased risk observed in relatives of MM patients provides support for inherited genetic predisposition [78]. Understanding of MM susceptibility has recently been informed by genome-wide association studies (GWAS), which have so far identified 17 independent risk loci for MM [112-114, 116], with an additional locus being subtype-specific for t(11;14) translocation MM [115]. Much of the heritable risk of MM, however, remains unexplained and statistical modelling indicates that further common risk variants remain to be discovered [158].

To gain a more comprehensive insight into MM aetiology, a new GWAS followed by a meta-analysis with existing GWAS and replication genotyping (totalling 9,974 cases and 247,556 controls) was performed. Six new MM susceptibility loci were identified as well as refined risk estimates for the previously reported loci. In addition, the possible gene regulatory mechanisms underlying the associations seen at all 23 GWAS risk loci was investigated by analysing *in situ* promoter Capture Hi-C (CHi-C) in MM cells to characterise chromatin interactions between predisposition single-nucleotide polymorphism (SNPs) and target genes, integrating these data with chromatin immunoprecipitation-sequencing (ChIP-seq) data generated in house and a range of publicly available genomics data. Finally, the contribution of both new and previously discovered loci to the heritable risk of MM was quantified and a likelihood-based approach to estimate sample sizes required to explain 80% of the heritability was implemented.

3.2 Study design

A new GWAS was performed and meta-analysed with previous GWAS datasets. Replication genotyping of promising associations was performed and imputed genotypes were validated. SNPs achieving genome-wide significance after replication genotyping were then functionally annotated using CHi-C, gene expression, ChIP-seq and a range of publicly available genomics data (**Figure 3.1**).

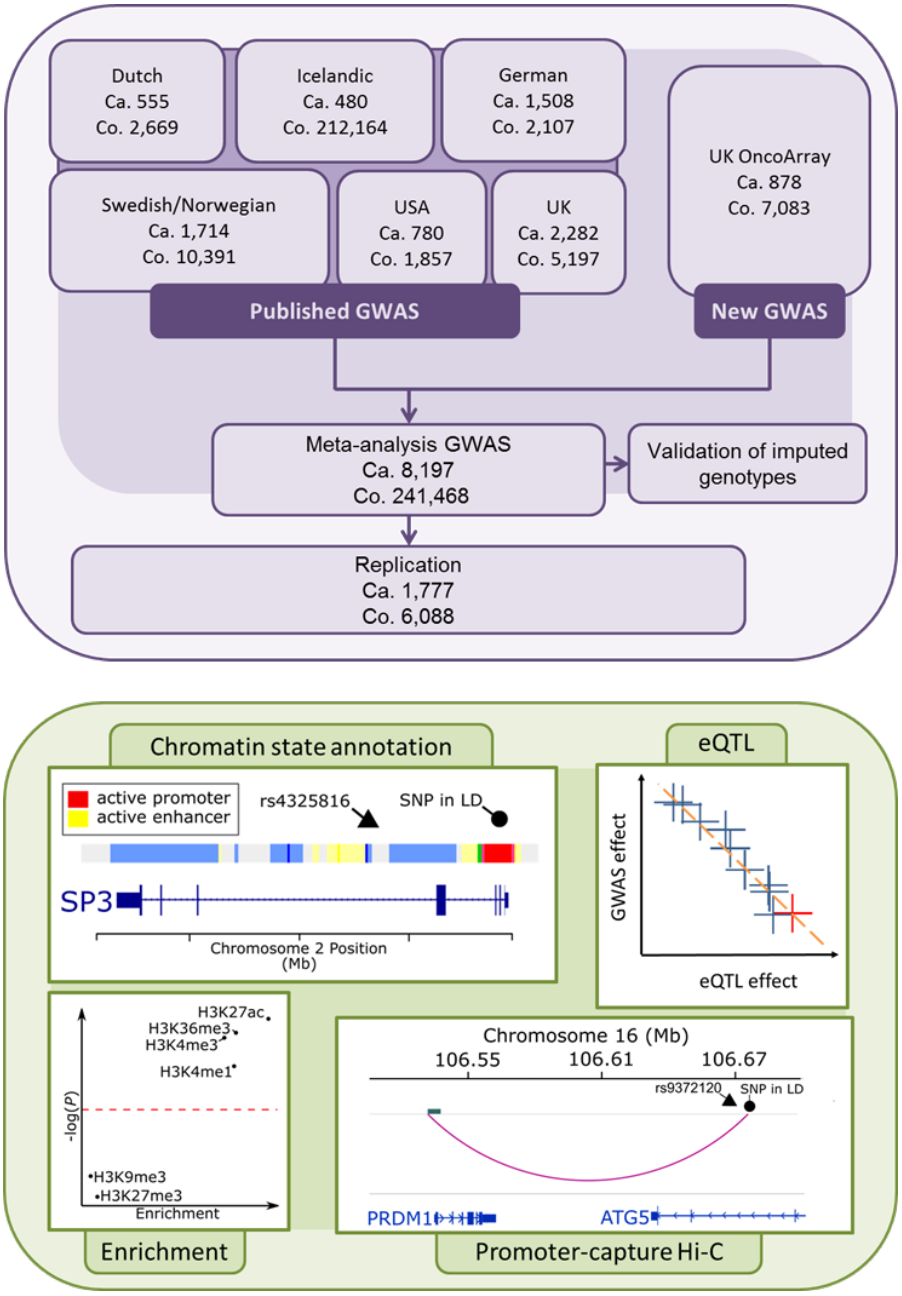


Figure 3.1 Overview of study design. GWAS study design. Details of the new and existing GWAS samples, including recruitment centres or trials and quality control, are provided in **Table 3.1** and **Table 3.2**. Trials or centres from which replication samples were recruited are detailed in Section 2.1. Ca.: cases, Co.: controls, eQTL: expression quantitative trait loci, SNP: single-nucleotide polymorphism, LD: linkage disequilibrium.

3.2.1 Genome-wide association studies

Details of sample numbers, ascertainment and quality control (QC) for the new OncoArray study and previous UK, US, German, Icelandic, Swedish/Norwegian and Dutch MM studies are detailed in **Table 3.1** and **Table 3.2.**

		Trial/ Recruitment Centre	Pre-QC	Sex discrepancy	Call rate fail	Heterozygosity rate	Related Individuals	Non-European Ancestry	Post-QC
UK	**Cases**	UK RC My IX, UK MRC MyXI	2,329	10	1	NA	2	34	2,282
	Controls	1958 Birth Cohort, National Blood Service	5,199	0	0	NA	2	0	5,197
Sweden/Norway	**Cases**	Swedish National Myeloma Biobank, Norwegian Biobank for Myeloma							1,714
	Controls	TWINGENE							10,391
Germany	**Cases**	GMMG-HD3, GMMG-HD4, GMMG-HD5	1,512	1	0	NA	0	3	1,508
	Controls	Heinz Nixdorf Recall	2,107	0	0	NA	0	0	2,107
Netherlands	**Cases**	HOVON65/GMMG-HD4, HOVON95/EMN02, HOVON87/NMSG18	608	0	2	7	0	44	555
	Controls	B-PROOF	2,669	0	0	0	0	0	2,669
USA	**Cases**	Total Therapy II, Total Therapy III, Total Therapy 3B, Total Therapy 4	1,076	0	0	9	1	286	780
	Controls	Cancer Genetic Markers of Susceptibility	2,234	0	4	2	0	369	1,857
Iceland	**Cases**	Icelandic Cancer Registry							480
	Controls	deCODE							212,164
OncoArray	**Cases**	UK MRC MyXI	931	6	1	5	3	44	878
	Controls	PRACTICAL, BCAC	7,519	8	1	7	68	364	7,083

Table 3.1 Details of the quality control filters applied to each GWAS. For the OncoArray dataset, highlighted in grey, samples were excluded due to call rate (<95% or failed genotyping), ancestry (principle components analysis or other samples reported to be not of white, European descent), relatedness (any individuals found to be duplicated or related within or between data sets through IBS) or sex discrepancy. Dutch, German, UK, USA, Sweden/Norway and Iceland studies have been previously reported in their entirety with comprehensive details on QC [112-114, 116]. MyIX, Myeloma IX; MyXI, Myeloma XI; B-PROOF, B-vitamins for the prevention of osteoporotic fractures; UKGPCS, UK Genetic Prostate Cancer Study; BCAC, Breast Cancer Association Consortium.

	UK	Sweden/Norway	Germany	Netherlands	USA	Iceland	OncoArray
Pre-QC	409,429		401,405	646,124	296,998		459,068
Call rate fail	997		113	6,523	4		6,851
HWE fail	7		0	18,104	171		12
MAF < 0.01	3		1	0	9,151		73,239
Post-QC	408,422		401,291	621,497	287,672		378,966
Imputed (filtered)	8,517,071	7,182,761	8,282,831	8,628,799	8,085,846	10,291,845	3,874,958

Table 3.2 Details of the quality control filters applied to each GWAS. For the OncoArray cohort, highlighted in grey, genotyped SNPs with a call rate <95% were excluded as were those with a MAF <0.01 or showing significant deviation from Hardy-Weinberg equilibrium (*i.e*. P < 10-5). Imputed SNPs with information score <0.8 and MAF <0.01 were excluded. Dutch, German, UK, USA, Sweden/Norway and Iceland studies have been previously reported in their entirety with comprehensive details on QC [112-114, 116].

3.2.2 Replication genotyping

Details of the series for replication of SNPs taken forward (German, Sweden/Norway and Denmark) are detailed in **Table 3.3**. Genotyping was performed using competitive allele-specific PCR KASPar chemistry as detailed in Section 2.2.3.2 for the nine promising risk loci. Call rates for SNP genotypes were > 95% in each of the replication series. To ensure the quality of genotyping in all assays, at least two negative controls and duplicate samples (showing a concordance of > 99%) were genotyped at each centre.

		Samples	Trial/ Recruitment Centre
Germany	**Cases**	911	German Myeloma Study Group
	Controls	1,477	Institute of Transfusion Medicine and Immunology, University of Mannheim, Germany
Swedish	**Cases**	534	Swedish National Myeloma Biobank
	Controls	2,382	Swedish Blood Donors
Denmark	**Cases**	332	University Hospital of Copenhagen
	Controls	2,229	Individuals from Denmark and Skane County

Table 3.3 Details of the replication sample recruitment.

3.2.3 Imputation concordance assessment

The fidelity of imputation at three imputed risk loci was assessed by directly sequencing a set of 147 randomly selected samples from the UK OncoArray case series. Targeted sequencing of imputed SNPs was performed by Sanger sequencing with primer sequences detailed in **Appendix 1**.

3.2.4 Statistical and bioinformatics analyses

Imputation and association testing were carried out as detailed in Section 2.3.6 and Section 2.3.7. Expression quantitative trait loci analyses on GWAS-identified SNPs were detailed in Section 2.4.5. Biological inferences were made based on the annotation of GWAS-identified SNPs as detailed in Section 2.4.11.

3.3 Results

3.3.1 Association analysis

A new GWAS using the OncoArray platform [162] (878 MM cases and 7,083 controls from the UK) was conducted, followed by a meta-analysis with six published MM GWAS data sets [112-114, 116, 305] (totalling 7,319 cases and 234,385 controls). Standard QC [306] measures, as described in Section 2.3.1, were applied to the OncoArray dataset; individuals with low call rate (< 95%) were excluded and those found to have non-European ancestry on the basis of HapMap version 2 CEU, JPT, CHB and YRI population reference data (**Appendix 2**). For first-degree relative pairs, the control or the individual with the lower call rate was excluded. To increase genomic resolution, array SNP genotypes were imputed to > 10 million SNPs. Quantile–quantile (Q–Q) plots for SNPs with minor allele frequency (MAF) > 1% after imputation did not show evidence of substantive over-dispersion for the OncoArray GWAS ($\lambda = 1.03$, $\lambda_{1000} = 1.02$, **Appendix 3**). Meta-analysis was undertaken using an inverse-variance approach under a fixed-effects model to derive odds ratios (ORs) for each SNP with MAF > 1%. Finally, validation of nine SNPs associated at $P < 1 \times 10^{-6}$ in the meta-analysis, which did not map to known MM risk loci and displayed a consistent OR across all GWAS data sets was sought by genotyping an additional 1,777 cases and 6,088 controls from three independent series (Germany, Denmark and Sweden). Targeted sequencing of imputed SNPs was performed and imputation was found to be robust; concordance was > 90% (**Appendix 4**). After meta-analysis of the new and pre-existing GWAS data sets and replication series, genome-wide significant associations (*i.e.* $P < 5 \times 10^{-8}$) [93] were identified for six new loci at 2q31.1, 5q23.2, 7q22.3, 7q31.33, 16p11.2 and 19p13.11 (**Appendix 5**, **Table 3.4** and **Figure 3.2**). Additionally, borderline associations were identified at two loci with *P*-values of 5.93×10^{-8} (6p25.3) and 9.90×10^{-8} (7q21.11). Bayesian false-discovery probabilities (BFDP) were calculated for these promising loci to assess the noteworthiness of the observed association [253] ; they had BFDP of 4% (6p25.3) and 6% (7q21.11). Conditional analysis of GWAS data showed no evidence for additional independent signals at the loci (**Appendix 6**). Finally, there was no evidence to support the existence of the putative risk locus at 2p12.3 (rs1214346), previously proposed by Erickson *et al* [307] (GWAS meta-analysis *P*-value = 0.32).

					OncoArray		Previous data		Replication		Combined meta		
SNP	Chr.	Pos. (b37)	Risk Allele	RAF	OR	P_{trend}	OR	P_{trend}	OR	P_{trend}	OR	P_{meta}	I^2
rs7577599	2	25613146	T	0.81	1.22	2.63×10^{-3}	1.24	1.24×10^{-16}	-	-	1.23	1.29×10^{-18}	0
rs4325816	**2**	**174808899**	**T**	**0.77**	**1.16**	**1.23×10^{-2}**	**1.11**	**1.30×10^{-5}**	**1.16**	**3.00×10^{-3}**	**1.12**	**7.37×10^{-9}**	**9**
rs6599192	3	41992408	G	0.16	1.24	1.35×10^{-3}	1.26	8.75×10^{-18}	-	-	1.26	4.96×10^{-20}	0
rs10936600	3	169514585	A	0.75	1.18	5.12×10^{-3}	1.20	5.94×10^{-15}	-	-	1.20	1.20×10^{-16}	0
rs1423269	5	95255724	A	0.75	1.09	0.125	1.17	1.57×10^{-11}	-	-	1.16	8.30×10^{-12}	23
rs6595443	**5**	**122743325**	**T**	**0.43**	**1.14**	**9.87×10^{-3}**	**1.10**	**4.69×10^{-6}**	**1.10**	**0.022**	**1.11**	**1.20×10^{-8}**	**0**
rs34229995	6	15244018	G	0.02	1.05	0.781	1.40	1.76×10^{-8}	-	-	1.36	5.60×10^{-8}	0
rs3132535	6	31116526	A	0.29	1.26	2.67×10^{-5}	1.20	2.97×10^{-17}	-	-	1.21	6.00×10^{-21}	0
rs9372120	6	106667535	G	0.21	1.18	7.74×10^{-3}	1.20	8.72×10^{-14}	-	-	1.19	2.40×10^{-15}	0
rs4487645	7	21938240	C	0.65	1.23	1.06×10^{-4}	1.24	5.30×10^{-25}	-	-	1.24	2.80×10^{-28}	0
rs17507636	**7**	**106291118**	**C**	**0.74**	**1.12**	**5.71×10^{-2}**	**1.12**	**5.54×10^{-7}**	**1.10**	**0.036**	**1.12**	**9.20×10^{-9}**	**50**
rs58618031	**7**	**124583896**	**T**	**0.72**	**1.17**	**7.61×10^{-3}**	**1.11**	**4.70×10^{-6}**	**1.10**	**0.061**	**1.12**	**2.73×10^{-8}**	**0**

					OncoArray		Previous data		Replication		Combined meta		
SNP	Chr.	Pos. (b37)	Risk Allele	RAF	OR	P_{trend}	OR	P_{trend}	OR	P_{trend}	OR	P_{meta}	I^2
rs7781265	7	150950940	A	0.12	1.33	3.23×10^{-4}	1.20	1.82×10^{-7}	-	-	1.22	4.82×10^{-10}	49
rs1948915	8	128222421	C	0.32	1.19	1.68×10^{-3}	1.14	3.14×10^{-10}	-	-	1.15	2.53×10^{-12}	26
rs2811710	9	21991923	C	0.63	1.13	1.76×10^{-2}	1.14	6.50×10^{-10}	-	-	1.14	3.64×10^{-11}	0
rs2790457	10	28856819	G	0.73	1.09	0.124	1.12	8.44×10^{-7}	-	-	1.11	2.66×10^{-6}	0
rs13338946	**16**	**30700858**	**C**	**0.26**	**1.17**	$\mathbf{7.90\times10^{-3}}$	**1.12**	$\mathbf{2.22\times10^{-7}}$	**1.26**	$\mathbf{2.5\times10^{-7}}$	**1.15**	$\mathbf{1.02\times10^{-13}}$	**26**
rs7193541	16	74664743	T	0.58	1.14	9.01×10^{-3}	1.12	1.14×10^{-8}	-	-	1.12	3.68×10^{-10}	34
rs34562254	17	16842991	A	0.10	1.32	7.63×10^{-4}	1.30	3.63×10^{-17}	-	-	1.30	1.18×10^{-19}	29
rs11086029	**19**	**16438661**	**T**	**0.24**	**1.26**	$\mathbf{1.02\times10^{-4}}$	**1.12**	$\mathbf{1.69\times10^{-6}}$	**1.15**	$\mathbf{5.00\times10^{-3}}$	**1.14**	$\mathbf{6.79\times10^{-11}}$	**42**
rs6066835	20	47355009	C	0.08	1.13	0.162	1.24	1.16×10^{-9}	-	-	1.23	6.58×10^{-10}	38
rs138747	22	35700488	A	0.66	-	-	1.21	2.58×10^{-8}	-	-	1.21	2.58×10^{-8}	0
rs139402	22	39546145	C	0.44	1.11	4.15×10^{-2}	1.23	4.98×10^{-26}	-	-	1.22	3.84×10^{-26}	56

Table 3.4 Summary of genotyping results for all genome-wide MM risk SNPs. New loci discovered through this study are emboldened. RAF, risk allele frequency; P_{trend}, *P*-value for trend, via logistic regression; P_{meta}, *P*-value for fixed effects meta-analysis; I^2, heterogeneity index (0–100). RAF are based on the UK cohort control series, with the exception of rs138747, which is sourced from the 1000 Genomes Project[55]

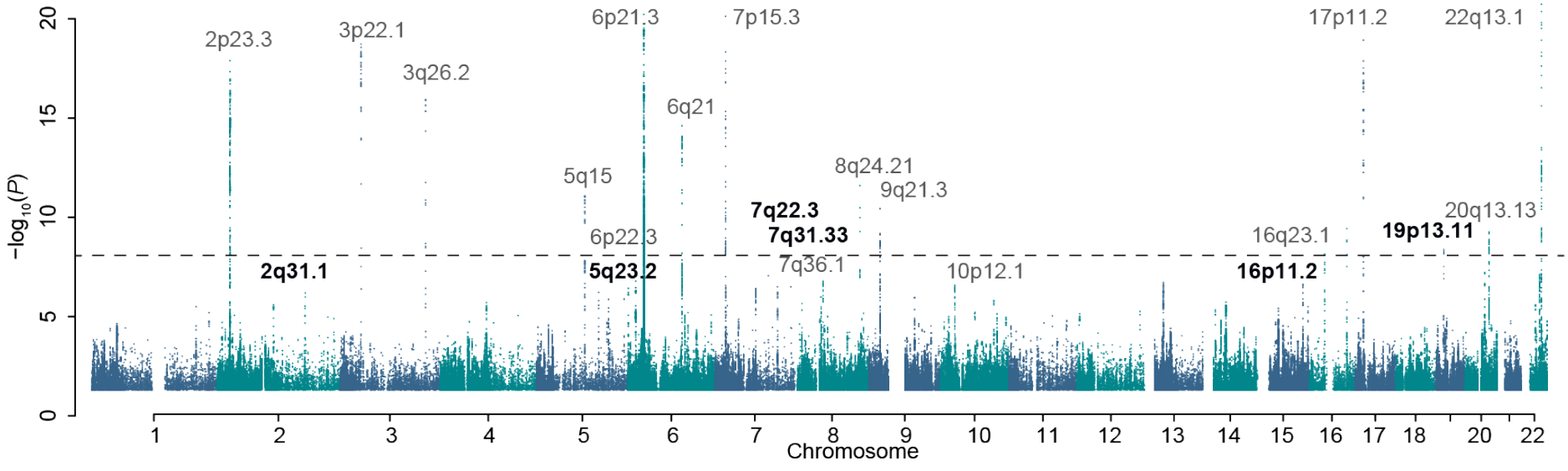


Figure 3.2 Manhattan plot of association signals. Genomic location of MM risk alleles identified in genome-wide association studies. Risk loci identified in the latest meta-analysis including the new OncoArray GWAS are emboldened. Dashed line represents threshold for genome-wide significance ($P < 5\times10^{-8}$).

3.3.2 Contribution of risk SNPs to heritability

Linkage disequilibrium adjusted kinships (LDAK) calculates heritability of a trait considering factors such as allele frequency and linkage disequilibrium, which influence heritability estimates [308]. Using this method, the heritability of MM ascribable to all common variation was calculated to be 15.6% (±4.7); collectively the previously identified and new risk loci account for 15.7% of the heritability (13.6% and 2.1%, respectively). To assess the collective impact of all identified risk SNPs, polygenic risk scores (PRS) considering the combined effect of all risk SNPs modelled under a log-normal relative risk distribution were constructed [186]. Using this approach, an individual in the top 1% of genetic risk has a threefold increased risk of MM when compared to an individual with median genetic risk (**Figure 3.3**). An enrichment of risk variants among familial MM compared with both sporadic MM cases and population-based controls was observed, comparable to that expected in the absence of a strong monogenic predisposition (respective *P*-values 0.027 and 1.60×10^{-5}; **Appendix 10**). Undoubtedly, the identification of further risk loci through the analysis of larger GWAS are likely to improve the performance of any PRS model. To estimate the sample size required to explain a greater proportion of the GWAS heritability, a likelihood-based approach using association statistics in combination with LD information was implemented to model the effect-size distribution [285, 309]. The effect-size distributions for susceptibility SNPs were best modelled using the three-component model (mixture of two normal distributions). Under this model, to identify SNPs explaining 80% of the GWAS heritability is likely to require sample sizes in excess of 50,000 (**Appendix 11**).

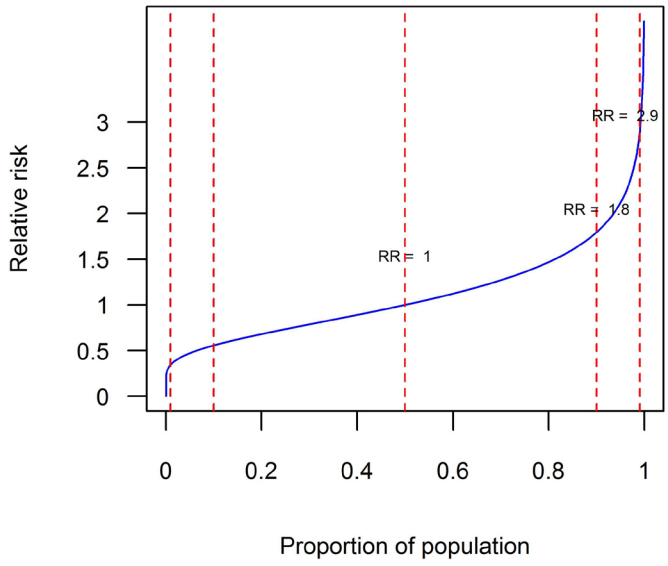


Figure 3.3 Population distribution of polygenic risk score (PRS). Ordered by relative risk (RR) (compared with population median risk). PRS is based on the 23 risk SNPs. Vertical red lines (left to right) correspond to 1%, 10%, 50%, 90%, and 99% centile, respectively.

3.3.3 Functional annotation and biological inference of risk loci

To the extent that they have been studied, many GWAS risk SNPs localise to non-coding regions and influence gene regulation [167]. To investigate the functional role of previously reported and new MM risk SNPs, a global analysis of SNP associations using ChIP-seq data generated on the MM cell line KMS11 and publicly accessible naïve B-cell Blueprint Epigenome Project data [179, 310] was performed. There was evidence of enrichment of MM SNPs in regions of active chromatin, as indicated by the presence of H3K27ac, H3K4Me3 and H3K4Me1 marks (**Figure 3.4**). An enrichment of relevant B-cell transcription factor-(TF) binding sites was also observed using ENCODE GM12878 lymphoblastoid cell line (LCL) data (**Figure 3.5**). Collectively these data support the tenet that the MM predisposition loci influence risk through effects on *cis*-regulatory networks involved in transcriptional initiation and enhancement.

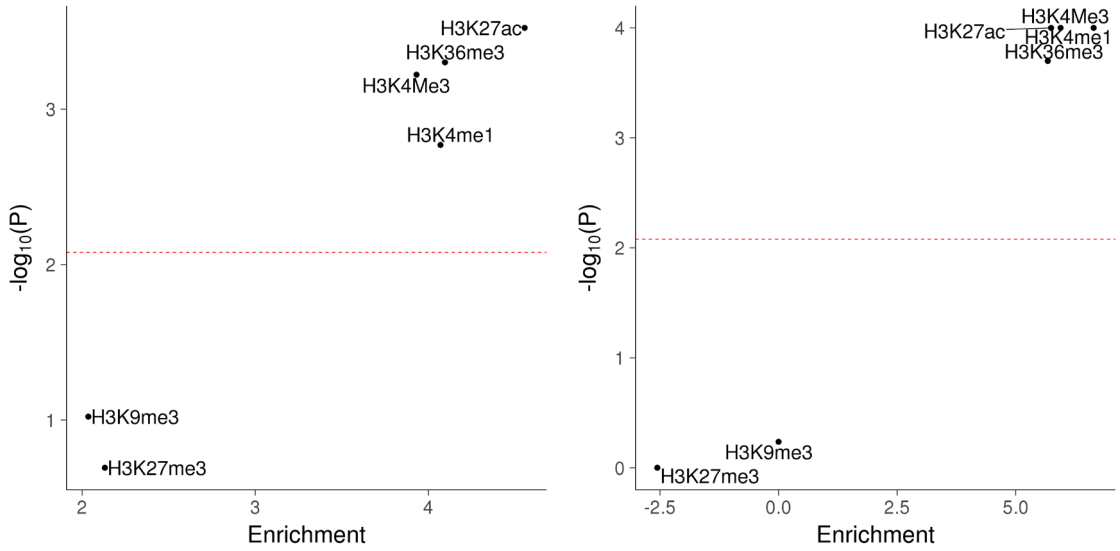


Figure 3.4 Enrichment of histone marks. The overrepresentation of histone marks in (left) naïve B and (right) KMS11 cells at the location of new and known MM risk SNPs demonstrates that risk SNPs are enriched in regions of open chromatin. The red line denotes the Bonferroni corrected *P*-value threshold. Note axes are on different scales.

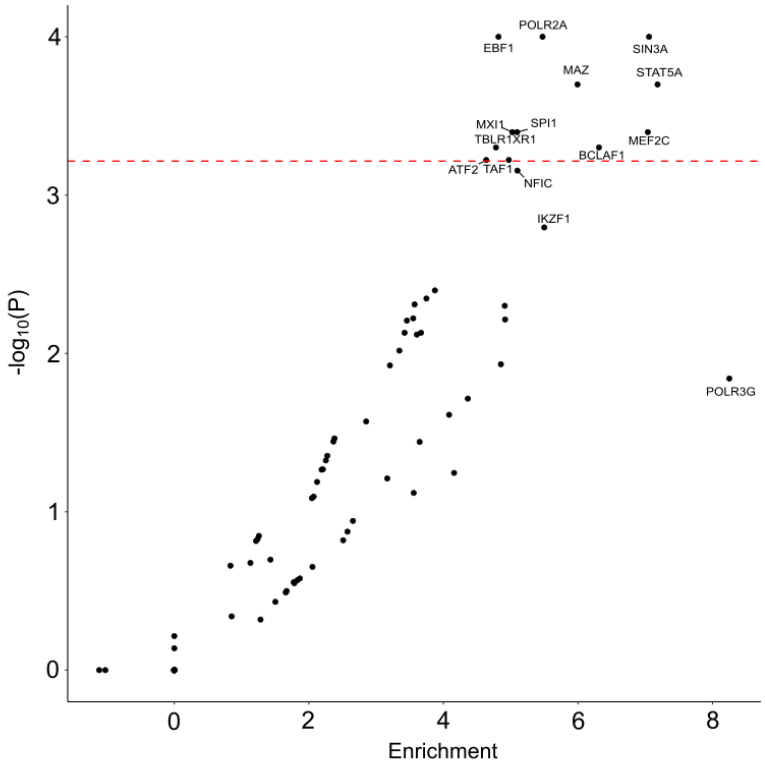


Figure 3.5 Enrichment of transcription factor binding sites. The overrepresentation of transcription factor (TF) binding sites in GM12878 cells at the location of new and known MM risk SNPs demonstrates that risk SNPs are enriched in regions of B-cell relevant TF binding. The red line denotes the Bonferroni corrected *P*-value threshold.

Since genomic spatial proximity and chromatin looping interactions are key to the regulation of gene expression, physical interactions at respective genomic regions in KMS11 and naïve B-cells were interrogated using CHi-C data [173]. To gain insight into the possible biological mechanisms for associations, an expression quantitative trait locus (eQTL) analysis was performed using mRNA expression data on CD138-purified MM plasma cells; specifically, Summary data-based Mendelian Randomization (SMR) analysis [181] was implemented to test for pleiotropy between GWAS signal and *cis*-eQTL for genes within 1 Mb of the sentinel SNP to identify a causal relationship (**Appendix 12**). Risk loci with variants mapping to binding motifs of B-cell-specific TFs were annotated. Finally, direct promoter variants and non-synonymous coding mutations were catalogued for genes within risk loci (**Table 3.5**).

Although preliminary and requiring functional validation, this analysis delineates four potential candidate disease mechanisms across the 23 MM risk loci. Firstly, four of the risk loci contain candidate genes linked to regulation of cell cycle and genomic instability, as evidenced by CHi-C looping interactions in KMS11 cells to *MTAP* (at 9p21.3) and eQTL effects for *CEP120* (at 5q23.2) (**Figure 3.6** and **Figure 3.7**). *CEP120* is required for microtubule assembly and elongation with

overexpression of *CEP120* leading to uncontrolled centriole elongation [311]. rs58618031 (7q31.33) maps 5' of POT1, the protection of telomeres 1 gene. *POT1* is part of the shelterin complex that functions to protect telomeres and maintain chromosomal stability [312, 313]. While mutated *POT1* is not a feature of MM, it is commonly observed in B-cell chronic lymphocytic leukaemia [127, 314, 315]. The looping interaction from the rs58618031 annotated enhancer element implicates *ASB15*. Members of the ASB family feature, as protein components of the ubiquitin–proteasome system, are intriguingly being investigated as a potential therapeutic target in MM [316-318].

Second, candidate genes encoding proteins involved in chromatin remodelling were implicated at three of the MM risk loci, supported by promoter variants at 2q31.1, 7q36.1 and 22q13.1. The new locus at 2q31.1 implicates *SP3*, encoding a TF, which through promoter interaction, has a well-established role in B-cell development influencing the expression of germinal centre genes, including activation-induced cytidine deaminase (*AID)* [319, 320].

Third, the central role of *IRF4-MYC*-mediated apoptosis/autophagy in MM oncogenesis is supported by variation at five loci, including eQTL effects *WAC* (at 10p12.1) (**Figure 3.6** and **Figure 3.7**) and Hi-C looping interactions (at 8q24.21 and 16q23.1). The 7p15.3 association ascribable to rs4487645 has been documented to influence expression of c-*MYC*-interacting *CDCA7L* through differential IRF4 binding [124]. Similarly, the long-range interaction between *CCAT1* (colon cancer-associated transcript 1) and *MYC* provides an attractive biological basis for the 8q24.21 association, given the notable role of *MYC* in MM [321, 322]. It is noteworthy that the promising risk locus at 6p25.3 contains *IRF4*. At the new locus 19p13.11, the missense variant (NP_057354.1:p.Leu104Pro) and the correlated promoter SNP rs11086029 implicates *KLF2* in MM biology. Demethylation by KDM3A histone demethylase sustains *KLF2* expression and influences IRF4-dependent MM cell survival [323]. The new 16p11.2 risk locus contains a number of genes including Proline-Rich Protein 14 (*PRR14*), which is implicated in PI3-kinase/Akt/mTOR signalling, a therapeutic target in myelomatous plasma cells [324].

Fourth, loci related to B-cell and plasma cell differentiation and function are supported by variation at three loci, including eQTL effects (*ELL2* at 5q15) [124] and Hi-C looping interactions (at 6q21). As previously inferred from GM12878 cell line data, the region at 6q21 (rs9372120, *ATG5*) participates in intra-chromosome looping with the B-cell transcriptional repressor *PRDM1* (alias *BLIMP1*) [116]. Additionally, SNP rs34562254 at 17p11.2 is responsible for the amino acid

substitution (NP_036584.1:p.Pro251Leu) in *TNFRSF13B*, a key regulator of normal B-cell homeostasis, which has an established role in MM biology [325-330].

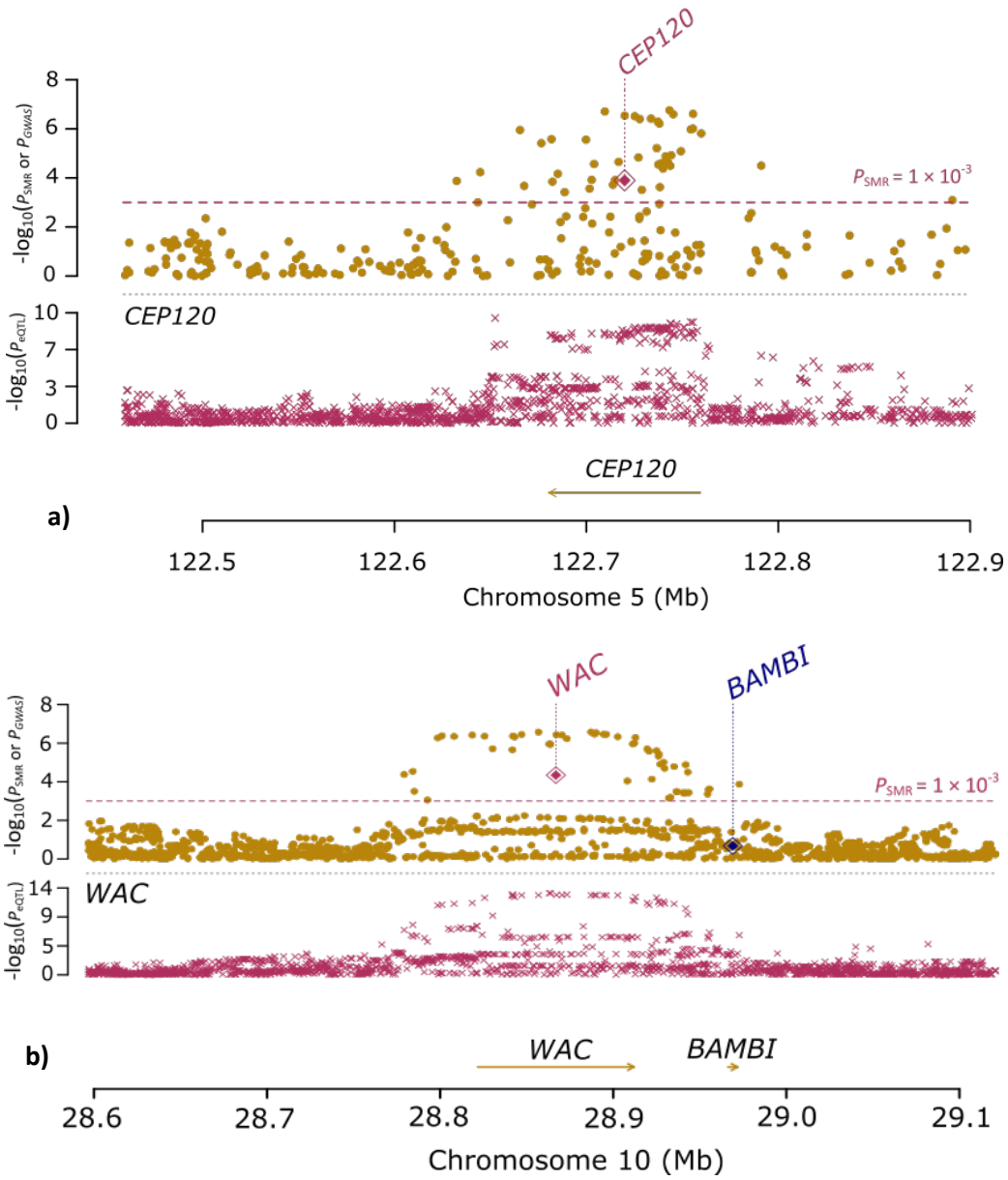


Figure 3.6 Summary data-based Mendelian Randomization (SMR) analysis locus plot. a) 5q23.2 and b) 10p12.1. Upper panel - brown dots represent *P*-values for SNPs from the GWAS meta-analysis, diamonds represent *P*-values for probes from the SMR test; lower panel – crosses represent eQTL *P*-values of SNPs from MM plasma cells from 183 MRC MyIX trial patients (GEO: GSE21349) and 658 Heidelberg GMMG patients (EMBL-EBI: E-MTAB-2299), with genes passing the SMR (*i.e.* $P_{SMR} < 0.001$) and HEIDI (*i.e.* $P_{HEIDI} > 0.05$) tests highlighted in red. Probeset ID refers to Affymetrix U133 2.0 Plus Array custom chip definition file (CDF v.17) mapping to Entrez genes.

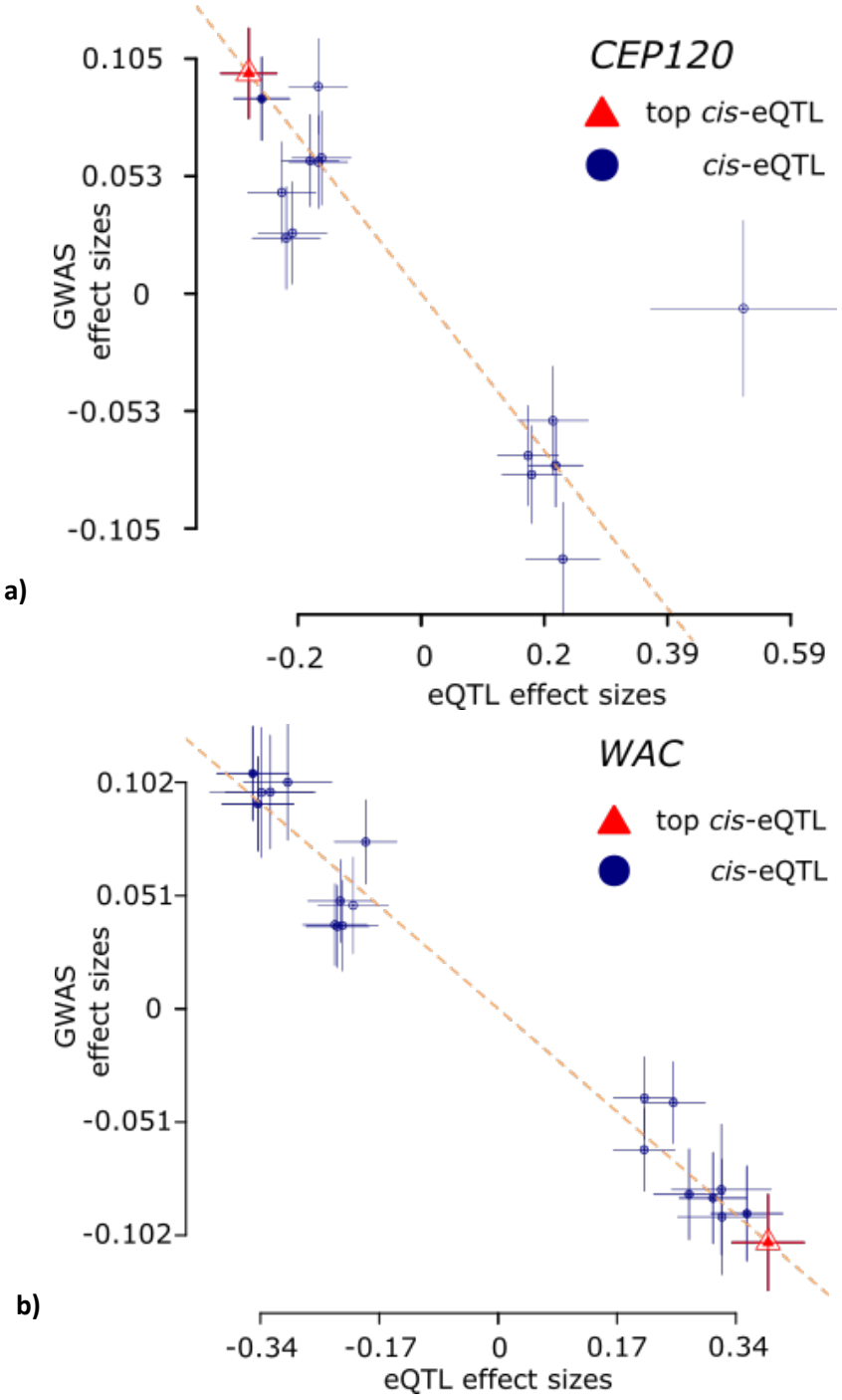


Figure 3.7 Summary data-based Mendelian Randomization analysis effect plot. (a) 5q23.2 and (b) 10p12.1. Blue dots represent effect sizes of SNPs from the GWAS meta-analysis against those from the eQTL study of MM plasma cells from 183 MRC MyIX trial patients (GEO: GSE21349) and 658 Heidelberg GMMG patients (EMBL-EBI: E-MTAB-2299). The top *cis*-eQTL is highlighted by a red diamond. Error bars are the standard errors of the SNP effects. An estimate of b_{xy} at the top *cis*-eQTL is represented by the orange dotted line.

								Functional evidence					
SNP	**Locus**	**bp(b37)**	**Genes in LD block**	**Coding variant**	**Promoter variant**	**Promoter/ enhancer chromatin states**	**TF binding[1]**	**Hi-C contact(s) in KMS11 cells**	**Hi-C contact(s) in naïve B cells**	**eQTL**	**Functional study**	**Candidate causal gene(s)**	**Candidate disease mechanism**
rs7577599	2p23.3	25613146	*DTNB*										
rs4325816	**2q31.1**	**174808899**	***SP3***		***SP3***	**active promoter, transcribed enhancer weakly acetylated, intermediate enhancer**	**BATF, CTCF, MAZ, NFIC, RAD21, YY1**					***SP3***	**Chromatin remodelling**
rs6599192	3p22.1	41992408	*ULK4*										
rs10936600	3q26.2	169514585	*ACTRT3 MYNN LRRC34*	*LRRC34*		active promoter, distal promoter	ATF2, EBF1, MAZ, MXI1, POL2RA, SIN3A, STAT5A, TAF1, TBLR1XR1 +21	*GPR160, SEC62-AS1*	*GPR160, LRRC31, MYNN, PDCD10, SERPINI1, SEC62, SAMD7, SEC62-AS1, SKIL, PHC3, PDCD10*			*LRRC34*	
rs1423269	5q15	95255724	*ELL2*	*ELL2*		intermediate enhancer, active enhancer, distal promoter	ATF2, BCLAF1, EBF1, IKZF1, MAZ, MEF2C, MXI1, SPI1, STAT5A, TBLR1XR1 +23	*VPS13C*				*ELL2*	B-cell development
rs6595443	**5q23.2**	**122743325**	***CEP120***	***CEP120***		**transcribed enhancer weakly acetylated**	**SPI1**	***SNX2, SNX24***	***SNX2, SNX24***	**CEP120**		***CEP120***	**cell cycle/ genomic stability**

								Functional evidence					
SNP	**Locus**	**bp(b37)**	**Genes in LD block**	**Coding variant**	**Promoter variant**	**Promoter/ enhancer chromatin states**	**TF binding[1]**	**Hi-C contact(s) in KMS11 cells**	**Hi-C contact(s) in naïve B cells**	**eQTL**	**Functional study**	**Candidate causal gene(s)**	**Candidate disease mechanism**
rs34229995	6p22.3	15244018	*JARID2*			intermediate enhancer, active promoter, distal promoter	FOXM1, IKZF1, MEF2A, NFIC, RELA, RUNX3, SPI1, YY1, ZNF143						
rs3132535	6p21.3	31116526	*PSORS1C1 CCHCR1*										
rs9372120	6q21	106667535	*ATG5*			intermediate enhancer, active enhancer		*PREP*	*PRDM1, PREP*			*PRDM1*	B-cell development
rs4487645	7p15.3	21938240	*DNAH11 CDCA7L*				IRF4, MYC, POLR2A, POU2F2, RUNX3, SPI1, TAF1, WRNIP1				CDCA7L	*CDCA7L*	apoptosis/ autophagy
rs17507636	**7q22.3**	**106291118**	***CCDC71L***										
rs58618031	**7q31.33**	**124583896**	***POT1***			**distal promoter, active enhancer, intermediate enhancer**	**NFIC**		***ASB15, IQUB, WASL***				**cell cycle/ genomic stability**
rs7781265	7q36.1	150950940	*ABCF2 CHPF2 SMARCD3*		*ABCF2, CHPF2*	active promoter, poised promoter	EBF1, EZH2, POLR2A, SIN3A, TAF1, YY1	*ASIC3, ABCF2, ATG9B*				*ABCF2*	chromatin remodelling

								Functional evidence					
SNP	**Locus**	**bp(b37)**	**Genes in LD block**	**Coding variant**	**Promoter variant**	**Promoter/ enhancer chromatin states**	**TF binding[1]**	**Hi-C contact(s) in KMS11 cells**	**Hi-C contact(s) in naïve B cells**	**eQTL**	**Functional study**	**Candidate causal gene(s)**	**Candidate disease mechanism**
rs1948915	8q24.21	128222421					ATF2, BCLAF1, EBF1, MAZ, MXI1, POL2RA, SIN3A, SPI1, STAT5A +18		*CASC11, MYC*			*MYC*	apoptosis/ autophagy
rs2811710	9p21.3	21991923	*CDKN2A, MTAP, CDKN2B-AS1*	*CDKN2A*	*CDKN2A, CDKN2B-AS1*	active promoter		*MTAP*	*MTAP*			*CDKN2A, MTAP*	cell cycle/ genomic stability
rs2790457	10p12.1	28856819	*WAC*			intermediate enhancer	CTCF	*LYZL1*	*MASTL, YME1L1*	WAC		*WAC*	apoptosis/ autophagy
rs13338946	**16p11.2**	**30700858**	***PRR14 FBRS SRCAP***	***PRR14***	***FBRS***	**active promoter, distal promoter**	**EBF1, MAZ, MXI1, POL2RA, SIN3A, SPI1, TAF1 +11**	***DCTPP1, DOC2A, FBXL19, GDPD3, ITGAL, MYLPF, PPP4C, SEPHS2, SEPT1, TBC1D10B, ZNF48, ZNF771***	***FBRS, PRR14, DCTPP1, MYLPF, TBC1D10B, SEPHS2***			***PRR14***	**apoptosis/ autophagy**
rs7193541	16q23.1	74664743	*RFWD3 GLG1*	*RFWD3*	*RFWD3*	active promoter	PML, TBP	*GLG1, NPIPL2*	*GLG1, HSPE1P, CFDP1, PSMD7, RFWD3, GABARAPL2*			*RFWD3*	

								Functional evidence					
SNP	**Locus**	**bp(b37)**	**Genes in LD block**	**Coding variant**	**Promoter variant**	**Promoter/ enhancer chromatin states**	**TF binding**[1]	**Hi-C contact(s) in KMS11 cells**	**Hi-C contact(s) in naïve B cells**	**eQTL**	**Functional study**	**Candidate causal gene(s)**	**Candidate disease mechanism**
rs34562254	17p11.2	16842991	*TNFRSF13B*	*TNFRSF13B*		intermediate enhancer, distal promoter, active enhancer	CTCF, POL2RA, STAT5A					*TNFRSF13B*	B-cell development
rs11086029	**19p13.11**	**16438661**	***KLF2***	***KLF2***	***KLF2***	**poised promoter**	**CTCF, EGR1, IKZF1, NFYB, POLR2A, RFX5, SIN3A, SPI1**					***KLF2***	**apoptosis/ autophagy**
rs6066835	20q13.13	47355009	*PREX1*			poised promoter	ATF2, EBF1, IKZF1, MEF2C, POL2RA, SPI1, TBLR1XR1 +9		*ARFGEF2*				
rs138747	22q13.1	35700488	*HMGXB4 TOM1*	*HMGXB4*	*TOM1, HMGXB4*	active promoter, transcribed enhancer weakly acetylated, intermediate enhancer, distal promoter, active enhancer, transcribed weak enhancer weakly acetylated	BCLAF1, EBF1, MAZ, POL2RA, STAT5A +46	*CRYBB1, HMOX1, APOL3, TOM1, LARGE, HMGXB4*	*FBXO7, HMGXB4, RASD2, MB*				
rs139402	22q13.1	39546145	*CBX7*		*CBX7*	distal promoter, intermediate enhancer, active promoter, poised promoter	BCLAF1, CHD2, CTCF, EBF1, MAZ, NFYB, POLR2A, RELA, RFX5, TBP, TAF1, ZNF143		*APOBEC3B-AS1, RPL3*			*CBX7*	chromatin remodelling

Table 3.5 Summary of functional annotation of the 23 risk loci. Newly identified risk loci are emboldened. Chromatin states were determined using ChromHMM. Heat maps were used to assign states based on previously described rules and these are shown in **Appendix 13**). Where > 10 TF were implicated at a locus, only those that overlap with TF which demonstrated enrichment in GM12878 are shown here. A full list of TFs localising to loci are detailed in **Appendix 14**.

3.4 Discussion

The meta-analysis of a new GWAS series in conjunction with previously published MM data sets performed in this chapter has identified six novel risk loci. Together, the new and previously reported loci explain an estimated 16% of the SNP heritability for MM in European populations. Ancestral differences in the risk of developing MM are well recognised, with a greater prevalence of MM in African Americans as compared with those with European ancestry [331]. It is plausible that the effects of MM risk SNPs may differ between Europeans and non-Europeans and hence contribute to differences in prevalence rates. Thus far, there has only been limited evaluation of this possibility with no evidence for significant differences [332].

Integration of CHi-C data with ChIP-seq chromatin profiling from MM and LCLs and naïve B-cells and eQTL analysis, from patient expression data, has allowed preliminary insight into the biological basis of MM susceptibility. The analysis within this chapter suggests a model of MM susceptibility based on transcriptional dysregulation consistent with altered B-cell differentiation, where dysregulation of autophagy/apoptosis and cell cycle signalling feature as recurrently modulated pathways. Specifically, the findings here implicate mTOR-related genes *ULK4*, *ATG5* and *WAC*, and by virtue of the role of *IRF4-MYC* related autophagy, *CDCA7L*, *DNMT3A*, *CBX7* and *KLF2* in MM development (**Table 3.5**). Further investigations are necessary to decipher the functional basis of risk SNPs, nevertheless this analysis highlights mTOR signalling and the ubiquitin–proteasome pathway, targets of approved drugs in MM. As a corollary of this, genes elucidated via the functional annotation of GWAS that discovered MM risk loci may represent promising therapeutic targets for myeloma drug discovery. Finally, estimation of sample sizes required to identify a larger proportion of the heritable risk of MM attributable to common variation underscores the need for further international collaborative analyses.

CHAPTER 4 Transcriptome-wide association study of multiple myeloma

4.1 Overview and rationale

Consistent with findings from many different cancer GWAS, bar a few notable exceptions, the functional variants and target susceptibility genes at the MM risk regions are yet to be identified. Knowledge of the causal genes responsible for defining disease predisposition is important in furthering understanding of MM tumourigenesis and has the potential to inform the development of novel therapeutic strategies [82]. While most GWAS risk variants map to non-coding regions of the genome, they are enriched for variants correlated with gene expression levels [167, 333]. Exploiting this characteristic, the integration of GWAS signals with expression quantitative trait loci (eQTLs) has implicated *ELL2* and *CDCA7L* as the risk genes likely to be responsible for the 5q15 and 7p15.3 MM associations, respectively [123, 124, 142, 143]. The high frequency of eQTLs coupled with linkage disequilibrium (LD) across regions can, however, make disentangling the risk genes from spurious co-localisation at the same region problematic.

Transcriptome-wide association studies (TWAS) have been proposed as a strategy to identify risk genes underlying complex traits [286]. This approach imputes genetic data from GWAS using reference sets of weights generated from eQTL data, before correlating this genetic component of gene expression with the phenotype of interest. Since TWAS aggregates the effects of multiple variants into a single testing unit, and facilitates prioritisation of genes at known risk regions for functional validation, it potentially also affords increased study power to identify new risk regions.

While MM is caused by the clonal expansion of malignant plasma cells, if a TWAS is to be based on expression data from a single cell, deciding on the most appropriate source is inherently problematic [334]. Utilising eQTL data from tumours is complicated by copy number alterations and tumours essentially represent terminal stage in disease progression. Moreover, the effect of any risk allele may be acting at the level of the tumour micro-environment [335]. Studies have shown that eQTLs strongly enriched in GWAS signals are not necessarily specific to the eQTL discovery tissue [333]. Taking advantage of this principle allows a multi-tissue TWAS to be conducted integrating expression across multiple tissues, thereby leveraging information on shared eQTLs for candidate gene discovery [336].

Within this chapter a multi-tissue TWAS is performed to prioritise candidate causal genes at known risk regions for MM and search for new risk regions. Specifically, gene expression data from 48 tissue panels measured in 8,756 individuals is analysed in conjunction with summary association statistics on 7,319 MM cases and 234,385 controls of European descent. 108 genes at 13 loci associated with MM risk are identified and additional evidence of a potential role for a number of genes dysregulated in MM tumourigenesis is provided.

4.2 Study design

GWAS data was integrated with Genotype-Tissue Expression Project (GTEx) data assayed from lymphocyte cell lines and whole blood, to predict gene expression. At the 22q13.1 locus looping interactions and histone modifications in the lymphoblastoid cell line (LCL) GM12878, were interrogated for evidence of gene regulation.

4.2.1 Genome-wide association study datasets

MM genotyping data was derived from the meta-analysis of seven GWAS datasets totalling 7,319 cases and 234,385 controls of European descent detailed in Chapter 3.

4.2.2 Expression data

SNP weights, used to impute expression levels for the whole transcriptome, and their respective covariance in 48 tissues from 80 to 491 individuals were obtained from predict.db [286], which is based on GTEx version 7 eQTL data [280]. A full list of the sample count by tissue can be found at the GTEx Portal [287].

4.2.3 Association analysis of predicted gene expression with myeloma risk

Associations between predicted gene expression and MM risk were examined using MetaXcan [286], which combines GWAS and eQTL data, accounting for LD-confounded associations. Briefly, genes likely to be disease-causing were prioritised using S-PrediXcan [286] which uses GWAS summary statistics and pre-specified weights to predict gene expression, given co-variances of SNPs. To combine S-PrediXcan data across the different tissues considering tissue-tissue correlations, S-MultiXcan was used [336].

To determine if associations between genetically predicted gene expression and MM risk were influenced by variants previously identified by GWAS, conditional analyses were performed

adjusting for sentinel GWAS risk SNPs [337]. To account for multiple comparisons, a Bonferroni-corrected *P*-value threshold of 1.96×10^{-6} (*i.e*. 0.05/25,520 genes) was considered as being statistically significant.

4.2.4 Regulatory annotation

To map risk SNPs to interactions involving promoter contacts and identify genes involved in MM susceptibility at the 22q13.1 locus, previously published promoter capture Hi-C data on the GM12878 cell line was analysed as a model B-cell [240]. This data was downloaded from the ArrayExpress database, accession code E-MTAB-2323.

4.2.5 Statistical power for association tests

The methodology of Wu *et al* [338] was used to estimate the power of the TWAS to identify associations using a simulation analysis. An estimate of the population prevalence of MM was obtained from Cancer Research UK [3]. The statistical power was calculated at $P < 1.96 \times 10^{-6}$, corresponding to the TWAS genome-wide significance level, according to various *cis*-heritability (h^2) thresholds that are assumed to be equivalent to gene expression prediction models (R^2). The results are based on 1,000 replicates.

4.3 Results

The association between predicted gene expression levels and MM risk was examined using MetaXcan with summary statistics for GWAS SNPs in 7,319 MM cases and 234,385 controls. MetaXcan is a statistical method which leverages substantial sharing of eQTLs across tissue and improves the ability to identify potential target genes [288]. In total, the expression levels of 25,520 genes across 48 tissues were tested for an association with MM risk. Quantile-quantile plots of TWAS association statistics did not show evidence of systematic inflation (**Figure 4.1**Error! Reference source not found.). **Figure 4.2** shows Manhattan plots for respective GWAS and TWAS associations.

Applying a Bonferroni threshold, 108 genes at 13 independent regions were identified as being associated with risk of MM (**Appendix 15**). All identified genes except those localising to the HLA region on chromosome 6p21 were within 1 Mb of previously reported MM risk SNPs. For all loci, except those in the HLA region, association signals were abrogated after adjusting for the top risk SNP, consistent with variation in expression of the identified gene being functionally related to the MM risk association. The complex LD patterns within the HLA region make deconvolution

of significant results within the region difficult [339, 340]; therefore, the principal focus was confined to 31 genes at 12 loci outside 6p21, which are shown in **Table 4.1**.

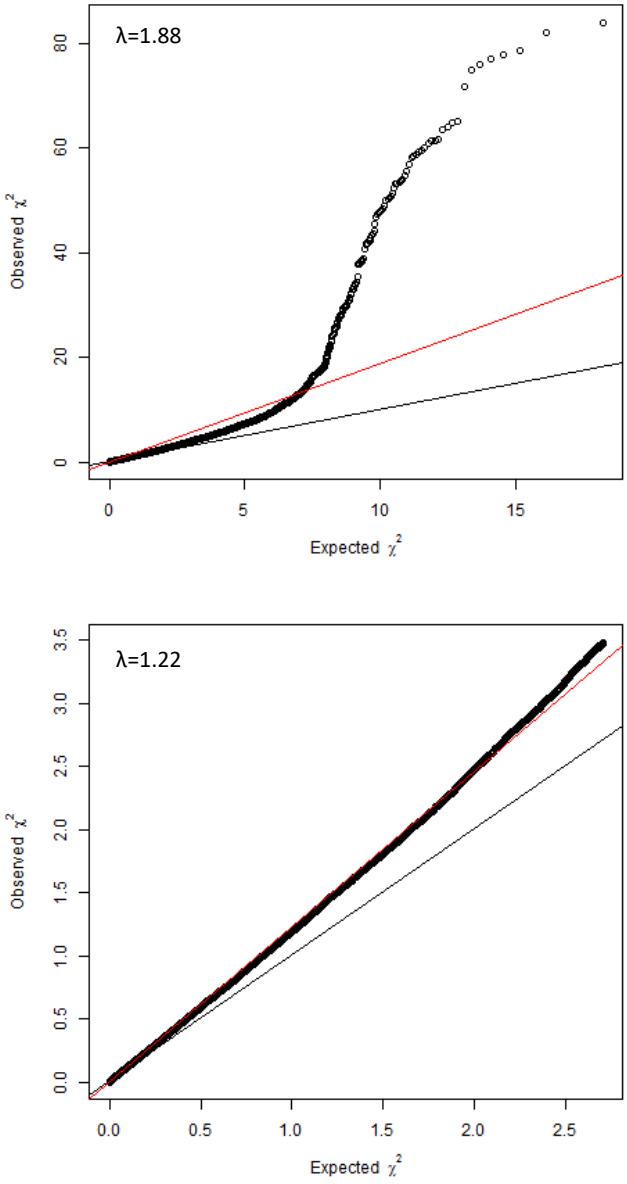


Figure 4.1 Quantile-Quantile plots of GWAS and TWAS. Shown are of $-\log_{10}$ (*P*-value) associations (top) TWAS for MM; (bottom) TWAS for MM (lower 90% of associations).

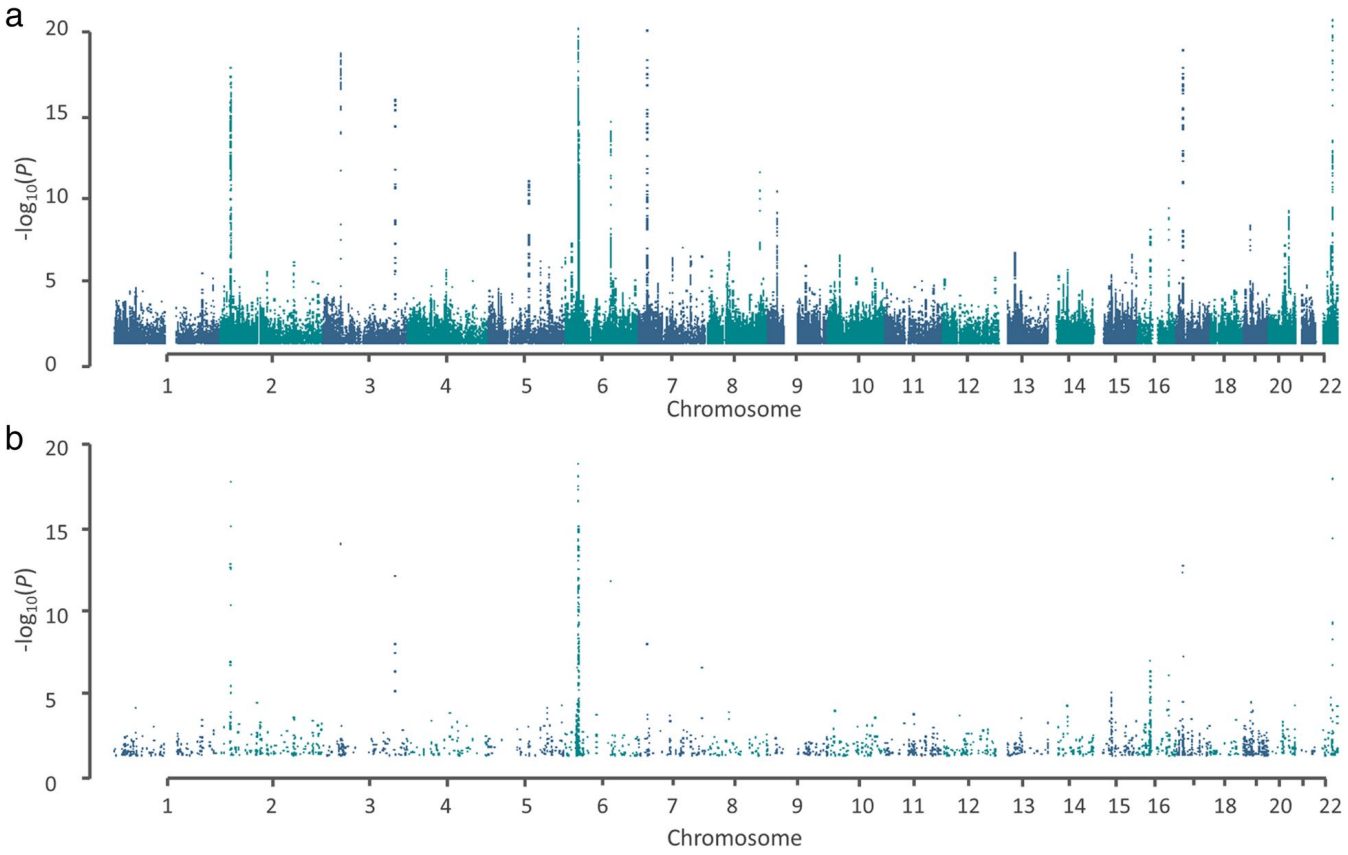


Figure 4.2 Manhattan plots of association signals. Manhattan plots of gene genomic co-ordinates against $-\log_{10}$ (*P*-value) of GWAS and TWAS association statistics. a- GWAS association statistics. b- TWAS association statistics.

.

Locus	Gene	*P*-value	N/N_{indep}	Z-score min	Z-score max	Z -score mean	Z-score s.d.	SNP adjusting for	*P*-value after SNP adjustment
2p23.3	*KIF3C*	1.65×10^{-18}	6/6	-9.4	4.35	-1.19	4.5	rs7577599	1.4×10^{-9}
2p23.3	*EPT1*	8.37×10^{-16}	9/9	-1.76	6	1.3	2.72	rs7577599	2.1×10^{-5}
2p23.3	*CENPO*	1.48×10^{-13}	12/8	-6.6	2.22	-0.05	2.57	rs7577599	6.1×10^{-8}
2p23.3	*DNMT3A*	2.44×10^{-13}	8/8	-2.89	7.96	1.94	3.07	rs7577599	0.01
2p23.3	*AC010150.1*	2.90×10^{-13}	4/4	-0.88	7.89	1.61	4.2	rs7577599	8.9×10^{-10}
2p23.3	*PTGES3P2*	4.46×10^{-11}	7/5	-4.23	2.03	-2.46	2.08	rs7577599	1.1×10^{-4}
2p23.3	*DTNB*	1.16×10^{-7}	11/10	-3.88	5.78	0.36	2.38	rs7577599	3.1×10^{-3}
2p23.3	*DNAJC27*	1.74×10^{-7}	8/8	-0.74	4.52	1.95	1.58	rs7577599	0.11
3p22.1	*ULK4*	9.01×10^{-15}	43/6	0.9	8.89	6.6	2.24	rs6599192	0.85
3q26.2	*MYNN*	7.84×10^{-13}	6/6	-7.91	1.58	-1.66	3.32	rs10936600	0.17
3q26.2	*LRRIQ4*	9.63×10^{-9}	3/2	-5.94	-0.88	-4.25	2.92	rs10936600	0.03
3q26.2	*LRRC34*	3.35×10^{-8}	21/2	3.97	6.47	5.12	0.66	rs10936600	0.82
3q26.2	*ACTRT3*	4.28×10^{-7}	4/4	-0.94	5.8	1.56	2.94	rs10936600	0.48
6q21	*ATG5*	1.55×10^{-12}	4/4	0.93	5.89	3.72	2.41	rs9372120	0.07
7p15.3	*CDCA7L*	9.61×10^{-9}	8/8	-3.11	4.61	1.12	2.42	rs75341503	0.23
7q36.1	*CHPF2*	2.53×10^{-7}	6/6	-2.01	2.13	0.4	1.49	rs7781265	0.06

Locus	Gene	*P*-value	N/N_{indep}	Z-score min	Z-score max	Z -score mean	Z-score s.d.	SNP adjusting for	*P*-value after SNP adjustment
16p11.2	*QPRT*	1.01×10^{-7}	17/8	-2.73	3.04	-0.59	1.63	rs13338946	0.15
16p11.2	*RNF40*	4.02×10^{-7}	24/3	0.05	5.68	4.67	1.48	rs13338946	0.89
16p11.2	*PRR14*	4.28×10^{-7}	2/2	-5.38	-0.2	-2.79	3.66	rs13338946	0.34
16p11.2	*C16orf93*	8.07×10^{-7}	13/5	-5.74	-0.34	-4.59	1.73	rs13338946	0.24
16p11.2	*RP11-2C24.5*	1.54×10^{-6}	5/5	-5.64	4.43	-0.58	3.8	rs13338946	0.73
16p11.2	*PRSS53*	1.71×10^{-6}	16/8	-5.19	3.68	-1.04	2.71	rs13338946	0.79
16q23.1	*RFWD3*	7.71×10^{-7}	34/7	-3.41	6.35	2.51	3.26	rs7193541	0.47
17p11.2	*TBC1D27*	1.95×10^{-13}	6/6	-1.91	4.19	0.51	2.16	rs34562254	0.89
17p11.2	*USP32P1*	4.88×10^{-13}	3/3	-7.29	2.8	-1.36	5.27	rs34562254	0.01
17p11.2	*PEMT*	5.65×10^{-8}	14/7	-1.74	5.43	1.36	1.93	rs34562254	0.01
22q13.1	*APOBEC3C*	1.10×10^{-18}	21/8	-8.93	0.24	-4.09	2.21	rs139402	0.13
22q13.1	*APOBEC3H*	4.28×10^{-15}	7/5	-5.45	7.92	-0.95	4.38	rs139402	0.76
22q13.1	*FAM83F*	4.65×10^{-10}	11/8	-4.25	2.56	-0.48	2.01	rs139402	1.1×10^{-4}
22q13.1	*APOBEC3D*	6.2×10^{-10}	29/7	-8.38	-0.85	-4.15	1.56	rs139402	0.04
22q13.1	*APOBEC3F*	5.15×10^{-9}	5/4	-6.34	6.15	1.09	5.07	rs139402	0.13
22q13.1	*APOBEC3G*	1.81×10^{-7}	43/2	0.36	6.57	4.94	1.17	rs139402	0.17

Table 4.1 Genes significantly associated with risk of multiple myeloma. Excludes associations found in the HLA region. s.d., standard deviation. Detailed are the S-MultiXcan *P*-values for association between gene expression MM, and the corresponding Z-scores quantifying this relationship (*e.g*. a positive score indicates increased gene expression increases risk). N and N_{indep} indicate the total number of single-tissue results used for S-MultiXcan analysis and the number of independent components after singular value decomposition, respectively.

For many loci, this TWAS finds support for the involvement of a number of genes that have previously been implicated in defining MM [112-114, 116, 305]. Specifically, single-gene associations were identified at 3p22.1 (*ULK4*), 6q21 (*ATG5*), 7p15.3 (*CDCA7L*), 7q36.1 (*CHPF2*) and 16q23.1 (*RFWD3*). However, at a number of regions, this analysis identified multiple significant genes, notably, 2p23.3 (*KIF3C*, *EPT1*, *CENPO*, *DTNB*, *DNM3TA*, *PTGES3P2*, *DNAJC27*), 3q26.2 (*MYNN*, *LRRC34*, *LRRIQ4*, *ACTRT3*), 16p11.2 (*QPRT*, *RNF40*, *PRR14*, *C16orf93*, *RP11-2C24.5*, *PRSS53*) and 17p11.2 (*TBC1D27*, *USP32P1*, *PEMT*). A complete list of novel genes identified at known GWAS risk loci is provided in **Table 4.2**.

Interestingly, several of the *APOBEC* genes were identified at 22q13.1. These genes localise within a distinct LD block adjacent to the one to which the sentinel GWAS risk SNPs maps (**Figure 4.3).** To gain insight into the potential for genome-wide significant SNPs in 22q13.1 in to influence regulation via a *cis*-regulatory enhancer, looping interaction and histone modifications in GM12878 were mapped across this region. GM12878, a cell line with negligible genetic and phenotypic abnormalities, was chosen as a model for early B-cell differentiation [341]. There was evidence of enhancer marks and looping interactions from SNPs in 22q13.1 to *APOBEC* genes (**Figure 4.3**), highlighting active chromatin and spatial proximity present in this region, necessary to mediate gene expression [173]. No significant genes were identified at 12 reported MM risk regions (2q31.1, 5q15, 5q23.2, 6p22.3, 7q22.3, 7q31.33, 8q24.21, 9p21.3, 10p12.1, 17p11.2, 19p13.1, 20q13.1).

4.3.1 Biological inference

These findings provide further support for a number of the genes previously implicated by GWAS whose expression influences the risk of developing MM, including *CDCA7L* at 7p15.3, which has been functionally validated. At 7p15.3, rs4487645 resides in an enhancer of c-Myc-interacting *CDCA7L* and increases IRF4 binding, affecting MM proliferation [124]. Furthermore, *ULK4* at 3p22.1, *ATG5* at 6q21 and *RFWD3* at 16q23 have been identified here and implicated previously. Additionally, this TWAS implicates new genes at known risk regions, notably *APOBEC3C*, *APOBEC3D*, *APOBEC3F*, *APOBEC3G* and *APOBEC3H* at 22q13.1 as playing a role in defining MM predisposition. Aberrant *APOBEC* cytidine deaminase activity has been shown to correlate with an increased mutational burden and is a recognised feature of MM, caused by triggering DNA mutation through dC deamination [342-344]. Furthermore, *KIF3C*, identified at 2p23.3, is a gene which regulates microtubule dynamics and has been previously implicated in breast cancer [345, 346]. Also at 2p23.3, this analysis identified *CENPO*, a gene involved in cell cycle progression via

regulation of kinetochore assembly [347]. At 16p11.2, *RNF40* is a promising candidate for MM susceptibility due to its role in double-strand break repair during homologous recombination (HR) and class switch recombination [348, 349]. This gene has also been implicated in colorectal cancer [350]. A further candidate at this locus, *QPRT* has been demonstrated to confer resistance to chemotherapy and radiotherapy when studied in glioma and leukaemia [351, 352]. As such, genes identified within this TWAS build upon previously suggested candidate disease mechanisms which may confer MM predisposition [305], including anti-apoptotic effects, roles in DNA double-strand break repair and cell cycle regulation. Furthermore, many of the genes identified have been previously investigated in vitro for their roles in cancer and this adds further support as plausible candidate genes for MM predisposition.

6p21.33, which encodes much of the major histocompatibility complex, is an especially gene rich region. As well as the class I HLA-A and class II genes HLA-DQA1 and HLA-DRB1/5, multiple genes localise to the region including *TCF19* which encodes the cell cycle progression and proliferation transcription factor 19 [353, 354]. Complex LD patterns within this region make deconvolution of significant results within the region inherently problematic [339]. Additional work is required to reveal the contribution of genes in this region to MM development.

A number of previously reported MM risk regions were not implicated in this TWAS. At some regions such as 5q15, the high tissue specificity associated with the causal gene *ELL2* [142] may not be best modelled herein. At other loci, it is less obvious why an association was not detected. Speculatively, models at earlier developmental stages may yield greater insights at these loci, especially if they are influencing differentiation along B-cell lineages. Additionally, other mechanistic effects may explain the functional basis of such loci, including methylation and splicing.

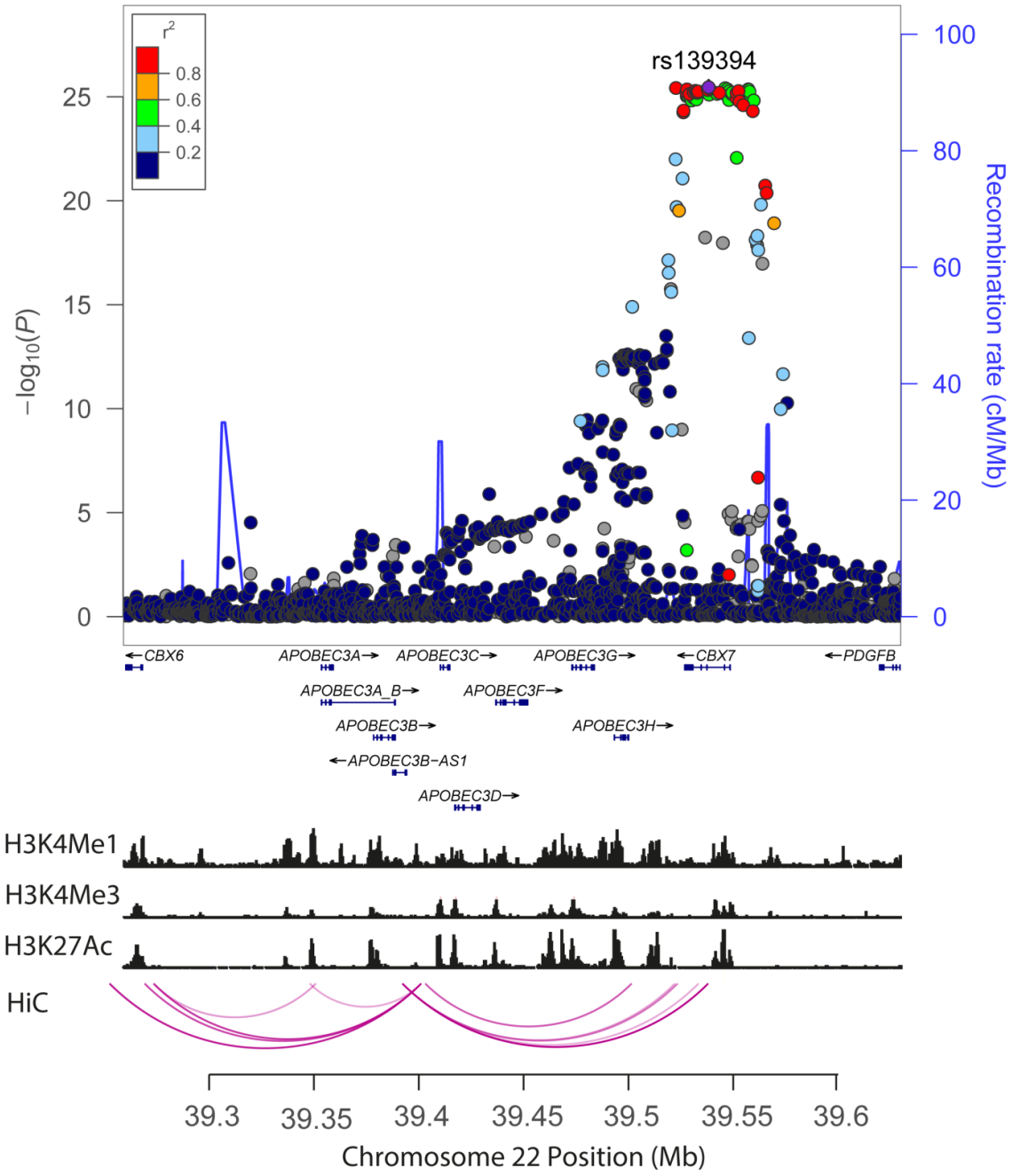


Figure 4.3 Regional plot of association at 22q13. Regional plot of association results at 22q13 in MM alongside recombination rates and histone marks in GM12878. Plot shows discovery association results of both genotyped and imputed SNPs in the GWAS samples and recombination rates. $-\log_{10}$ (*P*-values) (y axes) of the SNPs are shown according to their chromosomal positions (x axes). The colour of each symbol reflects the extent of LD with the top genotyped SNP. Genetic recombination rates, estimated using HapMap samples from Utah residents of western and northern European ancestry (CEU), are shown with a blue line. Physical positions are based on NCBI build 37 of the human genome. Also shown are the relative positions of GENCODE v19 genes mapping to the region of association. Below the association plot are the relative positions of GENCODE v19 genes mapping to the region of association and the histone marks and chromatin loops for LCL, GM12878.

SNP	Locus	bp(b37)	Newly implicated genes	Previously identified genes	Study
rs7577599	2p23.3	25,613,146	*KIF3C, EPT1, CENPO, DNMT3A, AC010150.1, PTGES3P2, DNAJC27*	*DTNB*	Broderick *et al* 2011
rs4325816	2q31.1	174,808,899		*SP3*	Went *et al* 2018
rs6599192	3p22.1	41,992,408		*ULK4*	Broderick *et al* 2011
rs10936600	3q26.2	169,514,585	*LRRIQ4*	*TERC, ACTRT3, MYNN, LRRC34, GPR160, LRRC31, MYNN, PDCD10, SERPINI1, SEC62, SAMD7, SEC62-AS1, SKIL, PHC3, PDCD10*	Chubb *et al* 2013, Went *et al* 2018
rs1423269	5q15	95,255,724		*ELL2, VPS13C*	Swaminathan *et al*, Li *et al* 2017
rs6595443	5q23.2	122,743,325		*CEP120, SNX2, SNX24*	Went *et al* 2018
rs34229995	6p22.3	15,244,018		*JARID2*	Mitchell *et al* 2016
rs3132535	6p21.3	31,116,526		*PSORS1C1, CCHCR1, CDSN, TCF19, POU5F1*	Chubb *et al* 2013, Went *et al* 2018
rs9372120	6q21	106,667,535		*ATG5, PREP, PRDM1*	Micthell *et al* 2016, Went *et al* 2018
rs4487645	7p15.3	21,938,240		*DNAH11 CDCA7L*	Broderick *et al* 2011, Li *et al* 2016
rs17507636	7q22.3	106,291,118		*CCDC71L*	Went *et al* 2018

SNP	Locus	bp(b37)	Newly implicated genes	Previously identified genes	Study
rs58618031	7q31.33	124,583,896		*POT1, ASB15, IQUB, WASL*	Went *et al* 2018
rs7781265	7q36.1	150,950,940		*ABCF2 CHPF2 SMARCD3 ASIC3 ATG98*	Mitchell *et al* 2016
rs1948915	8q24.21	128,222,421		*CASC11, MYC*	Mitchell *et al* 2016, Went *et al* 2018
rs2811710	9p21.3	21,991,923		*CDKN2A, MTAP, CDKN2B-AS1*	Mitchell *et al* 2016
rs2790457	10p12.1	28,856,819		*WAC, LYZL1, MASTL, YME1L1*	Mitchell *et al* 2016, Went *et al* 2018
rs13338946	16p11.2	30,700,858	*QPRT, RNF40, RP11-2C24.5, C16orf93*	*PRR14, FBRS, SRCAP, DCTPP1, DOC2A, FBXL19, GDPD3, ITGAL, MYLPF, PPP4C, SEPHS2, SEPT1, TBC1D10B, ZNF48, ZNF771*	Went *et al* 2018
rs7193541	16q23.1	74,664,743		*RFWD3, GLG1, HSPE1P, CFDP1, PSMD7, GABARAPL2, NPIPL2*	Mitchell *et al* 2016
rs34562254	17p11.2	16,842,991	*PEMT, USP32P1, TBC1D27*	*TNFRSF13B*	Chubb *et al* 2013
rs11086029	19p13.11	16,438,661	N/A	*KLF2*	Went *et al* 2018
rs6066835	20q13.13	47,355,009	N/A	*PREX1, ARFGEF2*	Mitchell *et al* 2016, Went *et al* 2018
rs138747	22q13.1	35,700,488	N/A	*CRYBB1, HMOX1, APOL3, TOM1, LARGE, FBXO7, HMGXB4, RASD2, MB*	Swaminathan *et al* 2015, Went *et al* 2018
rs139402	22q13.1	39,546,145	*APOBEC3C, APOBEC3D, APOBEC3F, APOBEC3G, APOBEC3H, FAM83F*	*CBX7, APOBEC3B-AS1, RPL3*	Chubb *et al* 2013, Went *et al* 2018

Table 4.2 New and previously implicated genes at each genome wide significant MM locus [112-114, 116, 355].

4.4 Discussion

Within this chapter a large TWAS involving 7,319 MM cases of European ancestry has been performed, identifying genetically predicted expression levels in 108 genes associated with MM risk. Of these, there were 94 genes located in eight regions that, although mapping within 1 Mb of a MM risk locus, had not previously been considered as a candidate gene for that locus.

The increasing appreciation that regulation of gene expression forms the mechanistic basis of many GWAS risk regions makes the TWAS an attractive approach to identify causal genes. Traditionally, studies have only tended to consider an eQTL and risk SNP to overlap if they are in linkage at a specified threshold. This is however, conservative as multiple local SNPs may independently contribute to risk. Furthermore, stipulating genome-wide significance thresholds for the GWAS signal (*i.e.* $P < 5 \times 10^{-8}$) and linkage strength (*i.e.* LD > 0.5) between pairs of SNPs for evidence of expression influencing risk, constrains study power. The TWAS approach is essentially agnostic as it jointly considers all SNPs in the region, regardless of reported GWAS association strength. There are, however, limitations to TWAS. Firstly, TWAS is based on fitting predictive linear models of gene expression based on local genotype data, followed by prediction into large cohorts and subsequent association testing; therefore, it does not capture total expression which includes environmental and technical components [356]. Secondly, TWAS will also lose power if gene expression is a nonlinear function of local SNPs, or when *trans* (or distal) regulation is a major determinant of expression levels.

All conclusions from this TWAS come with several caveats. While TWAS associations are consistent with models of gene expression level influencing MM risk, there is the possibility of confounding factors. Imputed gene expression levels are generated from weighted linear combinations of SNPs, and many of which may tag non-regulatory mechanisms driving risk and result in inflated association statistics. Inevitably, despite addressing LD, since genes with eQTLs are common, associations may be the result of chance co-localization between eQTLs and MM risk.

The ability to identify gene expression significantly associated with MM risk in this TWAS may be affected by tissue specificity. On the basis of the power calculation, this TWAS analysis had only 80% power to detect an odds ratio of ~ 1.1 for MM risk per one standard deviation increase (or decrease) in the expression level of a gene whose *cis*-heritability is 60% in EBV-transformed lymphocytes (**Figure 4.5**), which has been chosen as a proxy for plasma cells. In light of abundant shared *cis*-regulation of expression across tissues, by combining data,

it would be expected that any model could yield greater power as the number of tissues increases in which a variant is functional. Hence, the aim was to robustly capture genetically regulated gene expression using a large sample size.

In summary, work within this chapter highlights the value of integrating expression with GWAS to prioritise candidate causal genes. A number of identified genes have plausible roles in MM tumourigenesis (*e.g*. *APOBEC*, *RNF40*) or have been previously implicated in other malignancies (*e.g*. *QPRT*). The genes identified in this TWAS can be explored for follow-up and validation to further understand their role in MM biology.

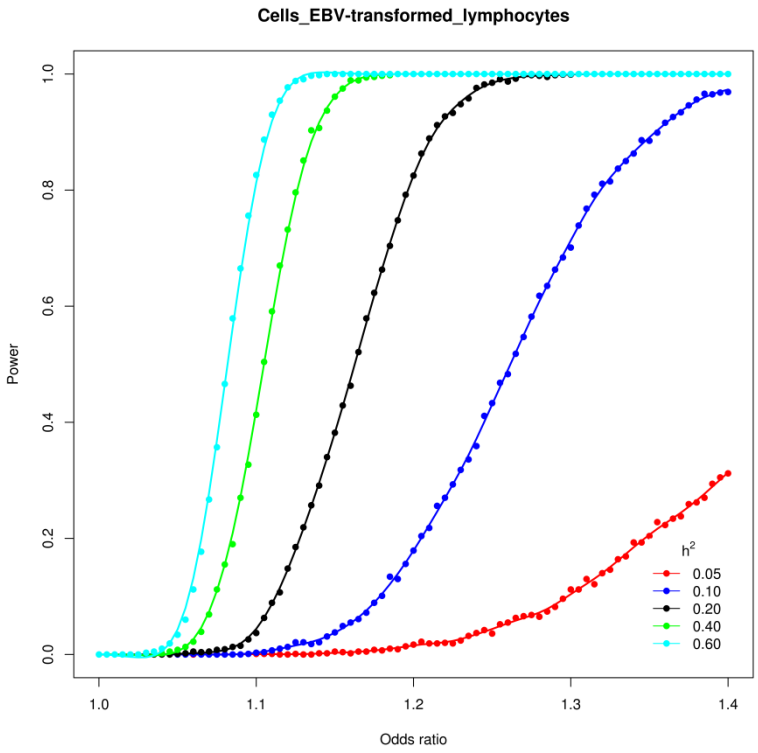


Figure 4.4 Power of TWAS based on 147 samples of EBV-transformed lymphocytes. Simulation analysis based on 7,319 cases and 234,385 controls. Gene expression was generated from the distribution of gene expression levels from EBV-transformed lymphocyte tissue (n=147). Statistical power was calculated at $P < 1.96\times10^{-6}$ (the significance threshold used in the main TWAS analysis) according to various *cis*-heritability (h^2) thresholds which are assumed to be equivalent to gene expression prediction models (R^2). Power calculations were per 1 s.d. change in gene expression based on 1,000 replicates.

CHAPTER 5 Co-heritability of multiple myeloma and chronic lymphocytic leukaemia.

5.1 Overview and rationale

Chronic lymphocytic leukaemia (CLL) and multiple myeloma (MM) are both B-cell malignancies, which arise from the clonal expansion of progenitor cells at different stages of B-cell maturity [357-359]. Epidemiological observations on familial cancer risks across the different B-cell malignancies suggest an element of shared inherited susceptibility, especially between CLL and MM [75].

Genome-wide association studies (GWAS) have transformed understanding of genetic susceptibility to the B-cell malignancies, identifying 45 CLL [127, 150, 219, 220] and 17 MM risk loci [112-114, 116]. Furthermore, statistical modelling of GWAS data indicates that common genetic variation is likely to account for 34% of CLL and 15% of MM heritability [158, 219]. There is the possibility that part of the shared heritable basis to both MM and CLL is likely to be enshrined in the same common risk variants.

Linkage disequilibrium (LD) score regression is a method which exploits the feature of a test statistic for a given single nucleotide polymorphism (SNP), whereby that test statistic will incorporate the effects of correlated SNPs [360]. Conventional LD score regression regresses trait $\chi 2$ statistics against the LD score for a given SNP, with the coefficient of the regression line providing an estimate of trait heritability. This method can be modified by instead regressing the product of SNP Z-scores from two traits against the SNP LD score, with the slope providing an estimate of genetic covariance between the two traits [290]. The method can be applied to summary statistics, is not biased by sample overlap, and does not require multiple traits to be measured for each individual.

Within this chapter, application of LD score regression to MM and CLL GWAS data demonstrates a positive genetic correlation between CLL and MM. There is evidence of shared genetic susceptibility at 10 known risk loci and integration of promoter capture Hi-C (CHi-C) data, ChIP-seq and gene expression data provide insight into the shared biological basis of CLL and MM.

5.2 Study design

Cross-trait linkage disequilibrium (LD)-score regression of multiple myeloma (MM) and chronic lymphocytic leukaemia (CLL) genome-wide association study (GWAS) data sets was performed, totalling 11,734 cases and 29,468 controls (**Figure 5.1**Error! Reference source not found.). Integration of eQTL, CHi-C and ChIP-seq data was performed at pleiotropic risk loci.

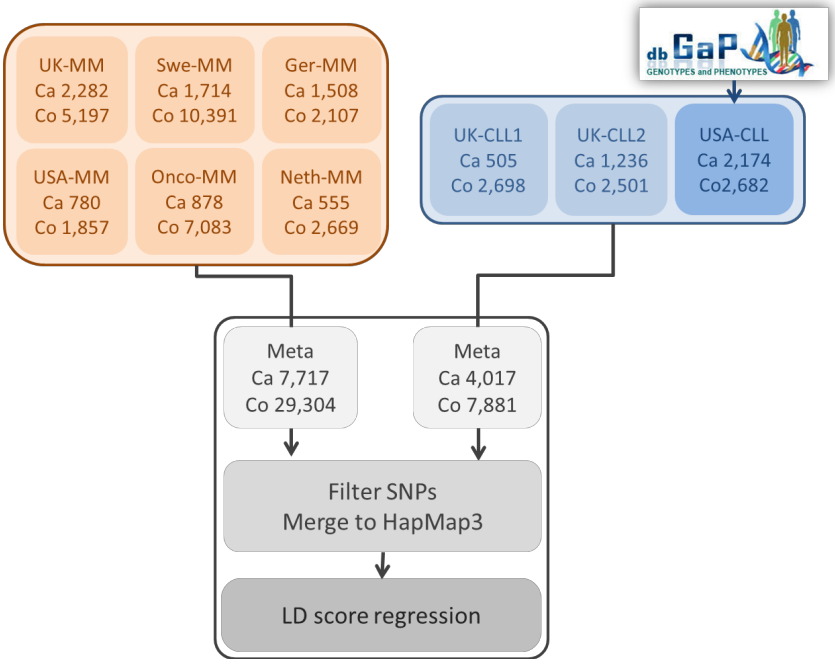


Figure 5.1 Overview of study design. Schematic outlining the processing of data sets used in the genetic correlation. Ca; cases, Co; controls

5.2.1 Multiple myeloma and chronic lymphocytic leukaemia datasets

Summary level data from the MM GWAS in Chapter 3 were used for LD score regression. Data from three previously reported CLL GWAS [127, 219, 220] were used to generate a meta-analysis of CLL datasets, following standard quality control measures [306]. The details of the MM datasets are in **Table 3.1** and **Table 3.2**. Details of the CLL dataset are in **Table 5.1** and **Table 5.2**. The summary level data from this meta-analysis was then used in the LD score regression analysis.

5.2.2 LD score regression

To investigate genetic correlation between MM and CLL, cross-trait LD score regression by Bulik-Sullivan *et al* [290] was used. Using summary statistics from the GWAS meta-analyses, filters were implemented as recommended by the authors. Specifically, filtering SNPs to INFO >0.9, MAF >0.01, and harmonising to Hap Map3 SNPs with 1000 Genomes EUR MAF >0.05, removing indels and structural variants, removing strand-ambiguous SNPs and removing SNPs where

alleles did not match those in 1000 Genomes. Heritability estimates were reported on the observed scale. There is no distinction between observed and liability scale genetic correlation for case/control traits [290].

	UK 1		UK 2		US	
	Cases	**Controls**	**Cases**	**Controls**	**Cases**	**Controls**
Pre-QC	517	2,698	1,403	2,501	2,178	2,685
Sex discrepancy					1	3
Call rate fail					0	0
Heterozygosity rate					0	0
Related Individuals					3	0
Non-European Ancestry					0	0
Post-QC	505	2,698	1,236	2,501	2,174	2,682

Table 5.1 Details of the quality control filters applied to each CLL GWAS. Samples were excluded due to call rate (< 95% or failed genotyping), ancestry (principle components analysis or other samples reported to be not of white, European descent), relatedness (any individuals found to be duplicated or related within or between data sets through IBS) or sex discrepancy. These studies have been previously reported in their entirety with comprehensive details on QC.

CLL	**UK1**	**UK2**	**US**
Pre-QC			727,545
Call rate fail			2,388
HWE fail/MAF < 0.01			81,128
Post-QC	301,786	630,366	644,029
Imputed (filtered)			8,899,686

Table 5.2 Details of the quality control filters applied to each CLL GWAS. For the OncoArray genotyped SNPs with a call rate < 95% were excluded as were those with a MAF < 0.01 or showing significant deviation from Hardy-Weinberg equilibrium (*i.e.* $P < 10^{-5}$). Imputed SNPs with information score < 0.8 and MAF < 0.01 were excluded.

5.2.3 Partitioned heritability

Stratified LD score regression can be used to partition heritability according to different genomic categories [291]. The enrichment of functional categories- defined as proportion heritability divided by the total heritability- was plotted for MM and CLL as per the method described in Section 2.4.8.1.

5.2.4 Shared risk loci

To identify pleiotropic risk loci, that is genetic loci that influence two traits, SNPs previously reported to be associated with each disease at genome-wide significance ($P < 5 \times 10^{-8}$), as well as highly correlated variants ($r^2 > 0.8$) were identified at the 45 and 23 known risk loci for CLL and MM, respectively. Within these correlated variant sets at each locus, many of the CLL susceptibility loci that were associated with MM at region-wide significance after Bonferroni correction for multiple testing (*i.e*. $P_{adj} < 0.05/45$) were identified. The process was then repeated, examining MM susceptibility SNPs in CLL, applying a significance level of $P_{adj} < 0.05/23$.

5.2.5 Variant set enrichment

The method of Cowper-Sal lari *et al* [289] was implemented to investigate enrichment of specific histone mark binding; this is described in detail in Section 2.4.7. For this publicly available ChIP-seq data for six histone marks from naïve B-cells was downloaded from Blueprint Epigenome Project [179].

5.2.6 Cell-type-specific analyses

The chromatin mark overlap enrichment for genome-wide significant loci in different cell types was investigated using the methodology of Trynka *et al* [294]. Briefly, this approach scores GWAS SNPs based on proximity to chromatin mark and fold-enrichment of respective chromatin mark, assessing significance using a tissue-specific permutation method. ChIP-seq data for H3K4me3 from primary blood cells and CLL samples was downloaded from Blueprint Epigenome project [179]. In addition, four MM cell lines, KMS11, JJN3, MM1-S and L363, were included in the analysis.

5.2.7 eQTL

eQTL analyses were performed using publicly available whole-blood data downloaded from GTeX [280]. The relationship between SNP genotype and gene expression was carried out using

Summary-data-based Mendelian Randomization (SMR) analysis as per Zhu *et al* [181], which is described in Section 2.4.5.

5.3 Results

5.3.1 Genetic correlation and heritability

Cross trait LD-score regression was performed using summary statistics from GWAS meta-analyses based on 7,717 MM cases and 21,587 controls, and 4,017 CLL cases and 7,881 controls. Details of the MM GWAS QC are in **Table 3.1** and **Table 3.2**. Detail of CLL QC are in **Table 5.1** and **Table 5.2**. In addition to standard GWAS QC, additional filters as per Bulik-Sullivan *et al* [290, 360] were implemented resulting in 1,055,728 harmonized SNPs between the two data sets. Heritability estimates from cross-trait LD score regression of 9.2 (±1.8%) and 22 (±5.9%) were comparable with previous estimates for MM and CLL. LD-score regression revealed a significant positive genetic correlation between MM and CLL with an R_g value of 0.44 ($P = 4.6 \times 10^{-3}$).

5.3.2 Identification of pleiotropic risk loci

Of the 45 CLL risk loci, four were associated with MM ($P_{adj} < 0.0011$) while, of 23 MM risk loci, five were significantly associated in CLL ($P_{adj} < 0.0022$) (**Table 5.3**, **Figure 5.2**). Correlated SNPs ($r^2 > 0.8$) at 3q26.2 are associated with both CLL and MM at genome-wide significance, bringing the total number of pleiotropic loci to 10.

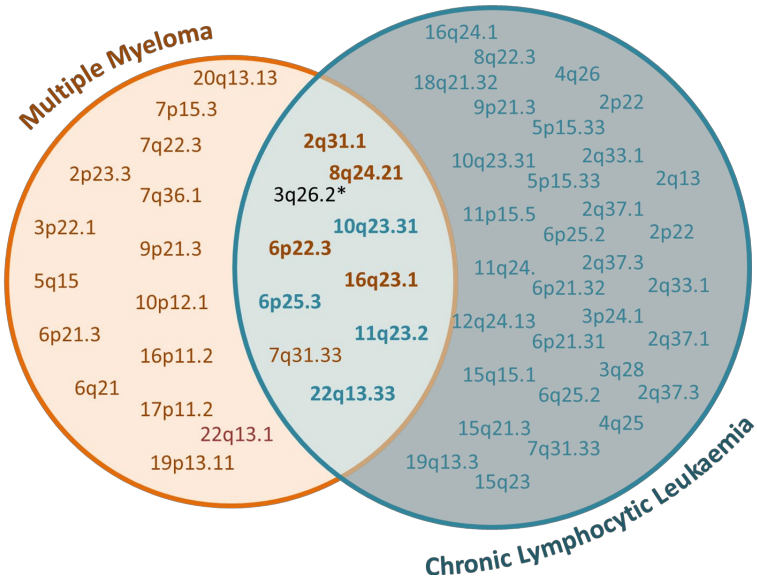


Figure 5.2 Overlap of loci in multiple myeloma and chronic lymphocytic leukaemia. *correlated variants at 3q26.2 had been previously published as genome wide significant in each data set prior to this analysis.

Locus	Discovery GWAS	Sentinel variant	Correlated variant	Position (hg19)	Risk allele		Odds Ratio		*P*-value	
					CLL	MM	CLL	MM	CLL	MM
2q31.1	MM	rs4325816		174,808,899	T	T	1.11	1.12	2.0×10^{-3}	6.4×10^{-7}
			rs72919402	174,750,200	T	-	1.13	-	4.6×10^{-4}	-
3q26.2	MM & CLL	rs1317082		169,497,585	A	A	1.2	1.19	7.1×10^{-8}	2.2×10^{-16}
			rs3821383	169,489,946	A	A	1.2	1.18	4.2×10^{-8}	4.5×10^{-15}
6p25.3	CLL	rs872071		411,064	G	G	1.37	1.1	2.8×10^{-27}	7.5×10^{-7}
			rs1050976	408,079	T	T	1.37	1.1	1.9×10^{-27}	3.7×10^{-7}
6p22.3	MM	rs34229995		15,244,018	G	G	1.37	1.36	8.5×10^{-3}	5.6×10^{-8}
			rs13197919	15,282,334	T	T	1.35	1.32	1.3×10^{-3}	3.42×10^{-7}
7q31.33	MM	rs58618031		124,583,896	T	T	1.15	1.11	3.2×10^{-5}	1.7×10^{-7}
			rs59294613	124,554,267	C	-	1.16	-	4.4×10^{-6}	-
8q24.21	MM	rs1948915		128,222,421	C	C	1.17	1.15	7.6×10^{-7}	2.5×10^{-12}
			-	-	-	-	-	-	-	-
10q23.31	CLL	rs6586163		90,752,018	A	A	1.28	1.06	1.1×10^{-16}	1.8×10^{-3}
			rs7082101	90,741,615	-	C	-	1.06	-	8.2×10^{-4}
11q23.2	CLL	rs11601504		113,526,853	C	C	1.2	1.09	2.3×10^{-5}	8.5×10^{-4}
			-	-	-	-	-	-	-	-
16q23.1	MM	rs7193541		74,664,743	T	T	1.12	1.12	1.0×10^{-4}	3.7×10^{-10}
	CLL		-	-	-	-	-	-	-	-
22q13.33		rs140522		50,971,266	T	T	1.17	1.08	3.7×10^{-7}	1.2×10^{-4}
			-	-	-	-	-	-	-	-

Table 5.3 Risk loci demonstrating association of alleles at respective loci in both CLL and MM.

5.3.3 Biological inference

Trynka *et al* [294] have recently shown that chromatin marks highlighting active regulatory regions overlap with phenotype-associated variants in a cell-type-specific manner. As H3K4me3 was shown to be the most phenotypically cell-type-specific chromatin mark, cell-type specificity of the 10 pleiotropic risk loci was examined by analysing H3K4me3 chromatin marks in normal haematopoietic cells and CLL patient samples from Blueprint. Additionally, *de novo* data on the KMS11, MM1S, JJN3 and L363 MM cell lines were included. Cell types showing the strongest enrichment of risk SNPs at H3K4me3 marks included naïve B-cells and CD38- B-cells. Notably, variants at 2q31.1, 6p25.3, 8q24.21, 16q23.1 and 22q13.33 were enriched for H3K4me3 in naïve B-cells (**Figure 5.3**).

Most GWAS signals map to non-coding regions of the genome [167] and influence gene expression through chromatin looping interactions [240, 241]. Application of partitioned heritability analysis, stratifying across 28 genomic categories, demonstrated enrichment of CLL and MM heritability in functional elements of the genome, in particular FANTOM5 enhancers (CLL and MM) transcription start sites and 5' untranslated region and coding regions (MM) (**Appendix 16**). Furthermore, there was significant enrichment of SNPs in the shared loci within regions of active chromatin, as indicated by the presence of H3K27ac and H3K4me3 marks in naïve B-cells, supporting the principle that SNPs in shared loci influence risk through regulatory effects (**Figure 5.3**Error! Reference source not found.). To identify target genes, CHi-C data on naïve B-cells from Blueprint [179] was used. Finally, to gain insight into the possible biological mechanisms for associations, eQTL analysis was performed using mRNA expression data on blood from GTEx. This involved application of SMR [181] to test for pleiotropy between GWAS signal and *cis*-eQTL for genes to identify a causal relationship (**Appendix 17** and **Appendix 18**). Broadly, this analysis of the shared loci groups them into those that act on B-cell regulation and differentiation, and those that underpin the distinctive biology of cancer; specifically, loci relating to genome instability, angiogenesis and dysregulated apoptosis. These are summarised in **Table 5.4**.

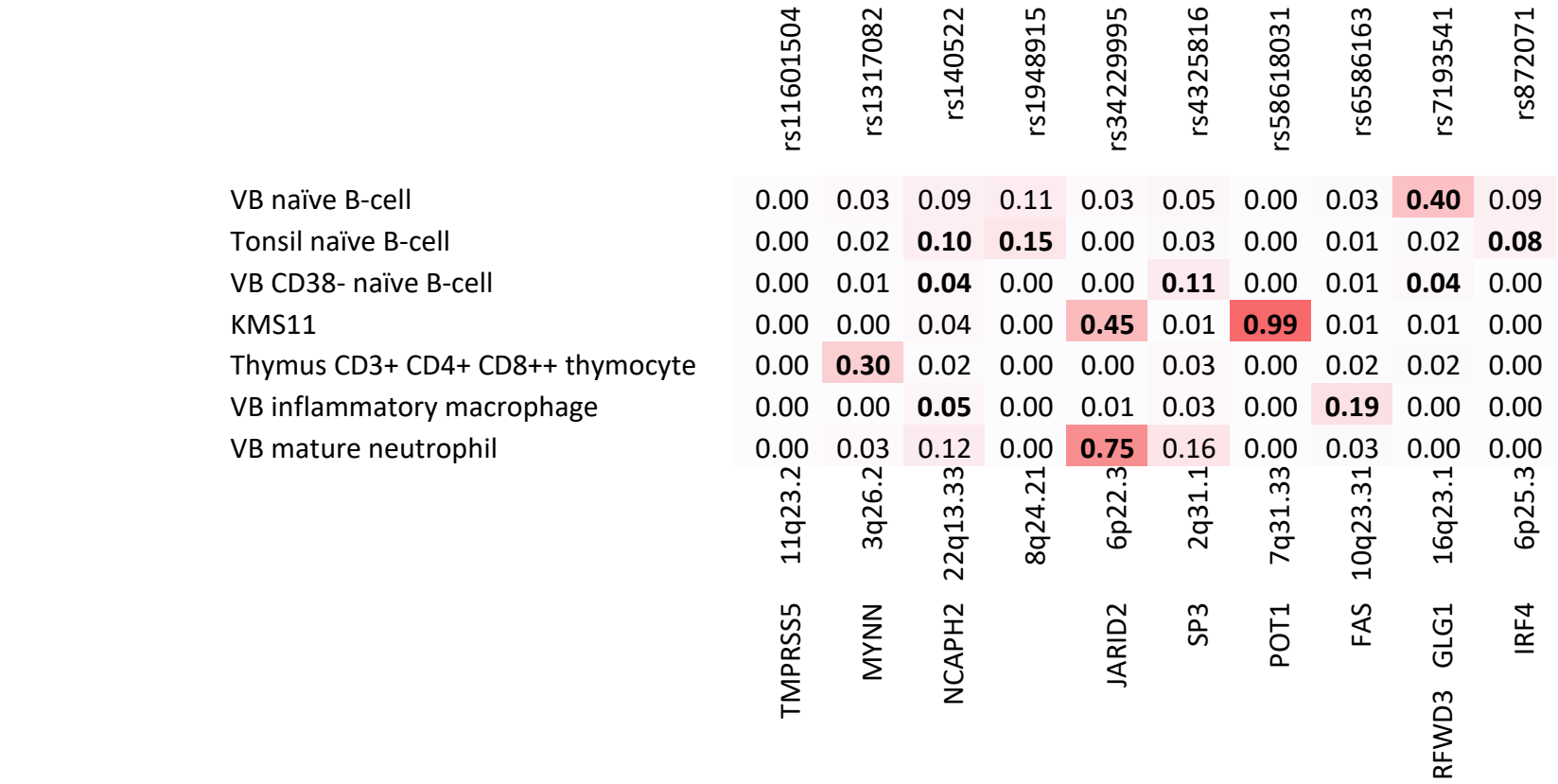


Figure 5.3 Tissue specific H3K4me3 mark enrichment for shared loci. Bold denotes SNPs significantly enriched.

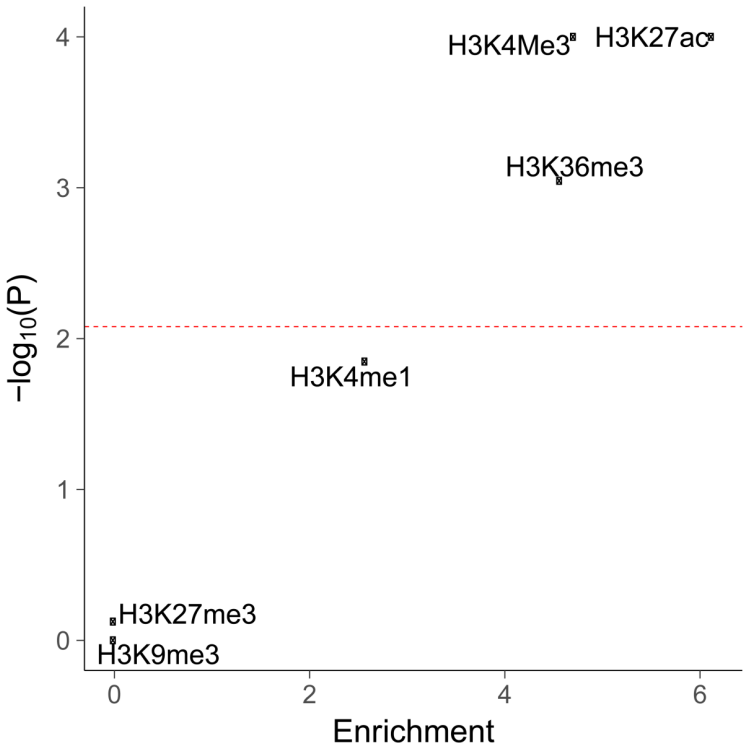


Figure 5.4 The overrepresentation of histone marks from naïve B-cells at the location of shared CLL and MM risk loci. The red line denotes the Bonferroni corrected *P*-value threshold.

Locus	Chr	rsID	Position (hg19)	Functional Evidence				Candidate causal gene(s)
				Proximal genes	Naïve B HiC	eQTL	Naïve B Histone Marks present	
2q31.1	2	rs4325816	174,808,899	***SP3***	***SP3 (promoter)***		H3K27ac+ H3K4me3	*SP3*
					RP11-394I13.2			
				ACTRT3	*PDCD10*		H3K27ac+ H3K4me3	*SEC62*
				MYNN LRRC34	*SERPINI1*		H3K27ac+ H3K4me1	*TERC*
				TERC	*RP11-379K17.4*			
					SEC62			
					SEC62-AS1			
					GPR160			
3q26.2	3	rs1317082	169,497,585		*RNU4-38P*			
					PHC3			
					RNU6-315P			
					NA			
					SKIL			
					MYNN			
					SAMD7			
6p25.3	6	rs872071	411,064	*SERPINB6*	*DUSP22*		H3K27ac (weak)+ H3K4me1	
					RP3-416J7.5			

Locus	Chromosome	rsID	Position (hg19)	Functional Evidence				
				Proximal genes	Naïve B HiC	eQTL	Naïve B Histone Marks present	Candidate causal gene(s)
6p22.3	6	rs34229995	15,244,018	*JARID2*			H3K27ac (weak)+ H3K4me1	
				POT1	*IQUB*			
				IQUB	*ASB15*			
				ASB15	*RP11-390E23.6*			
7q31.33	7	rs58618031	124,583,896	*WASL*	*WASL*			
				ACTRT3	*ACTRT3*			
				RNU6-11P	*RP11-816J6.3*			
					RNU6-11P			
8q24.21	8	rs1948915	128,222,421				H3K27ac + H3K4me1	
				ACTA	*ACTA2 (promoter)*	*ACTA*		*FAS*
10q23.31	10	rs6586163	90,752,018	*FAS*	*FAS (promoter)*	*FAS*	H3K27ac (weak)+ H3K4me3	*ACTA2*
					CH25H			
11q23.2	11	rs11601504	113,526,853					

Locus	Chromosome	rsID	Position (hg19)	Functional Evidence				
				Proximal genes	Naïve B HiC	eQTL	Naïve B Histone Marks present	Candidate causal gene(s)
16q23.1	16	rs7193541	74,664,743	***RFWD3*** *GLG1*	*GLG1* *RNU6-237P* *NPIPB15* *AC009120.4* *PSMD7* *GABARAPL2* *TERF2IP* *KARS* *CFDP1* ***RFWD3 (promoter)***	***RFWD3***	H3K27ac+ H3K4me3	*RFWD3*
22q13.33	22	rs140522	50,971,266	*TYMP*	*RP11-144N1.1* *HSPE1P7* *CTA-384D8.36* *NCAPH2* *ODF3B* *SCO2* *TYMP* *LMF2* *NCAPH2* *SYCE3* *ARSA*	*TYMP* *ODF3B*		*TYMP*

Table 5.4 Functional evidence at each of the shared loci.

Of the shared loci, three were related to B-cell regulation. This included a composite of evidence at 10q23.31, from looping interaction in naïve B-cells and correlation in GWAS effect size and expression, which provide evidence for two candidate genes *ACTA2* (**Appendix 17** and **Appendix 18**), encoding smooth muscle (α)-2 actin, a protein involved in cell movement and contraction of muscles [361] and *FAS*, a member of the TNF-receptor superfamily. FAS, has a central role in regulating the immune response through apoptosis of B-cells [362, 363]. At 2q31.1, looping interactions implicated transcription factor SP3, which has been shown to influence expression of germinal centre genes [319, 320]. Variants at 6p25.3 reside in the 3'-UTR of *IRF4*, which has an established role in B-cell regulation and MM oncogenesis [124, 323, 364].

Three of the 10 loci contain genes with roles in maintenance of genomic stability. Specifically, evidence from expression and CHi-C data implicated *RFWD3* at 16q23.1 (**Appendix 17** and **Appendix 18**). This gene encodes an E3 ubiquitin-protein ligase, which has been shown to promote progression to late stage homologous recombination through ubiquitination and timely removal of RAD51 and RPA at sites of DNA damage and is necessary for replication fork restart [365, 366]. Variants in this locus demonstrated enrichment of H3K4me3 marks in two samples of naïve B-cells, which represents a plausible cell of disease origin. rs58618031 (7q31.33) maps 5' of *POT1*, which is part of the shelterin complex and functions to maintain chromosomal stability [312, 313]. Variant rs1317082 at 3q26.2 is located proximal to *TERC*, a gene which has been shown to influence telomere length [367]. Additionally, we observed looping interactions to a number of genes at 3q26.2 including *SEC62*, which has been proposed as a cancer biomarker [367-370]. Intriguingly, variants at 3q26.2 this locus have been implicated in colorectal [131], thyroid [132] and bladder cancer [371].

Several genes were implicated at 22q13.33 by looping interactions for *SCO2*, *LMF2*, *ODF3B*, *TYMP*/*ECGF1*, *NCAPH2*, *SYCE3* and *ARSA*, with *TYMP*/*ECGF1* and *SCO2* demonstrating evidence of correlation in GWAS and eQTL effect size, albeit not significant after multiple testing ($P_{SMR} = 2.38 \times 10^{-4}$ and 3.19×10^{-4}). Variants within this locus were enriched in H3K4me3 chromatin marks in both CD38- B-cells and inflammatory macrophages. *TYMP* (alias *ECGF1*) encodes thymidine phosphorylase, which is often overexpressed in tumours and has been linked to angiogenesis [372, 373]. A detailed study on this gene has implicated *TYMP* in the development of lytic bone lesions in MM, via a mechanism involving activation of PI3K/Akt signalling and increased *DNMT3A* expression resulting in hypermethylation of *RUNX2*, osterix, and *IRF8* [374]. Furthermore, *SCO2* (synthesis of cytochrome c oxidase), also mapping to this

locus, has been implicated in the development of breast cancer [375, 376], gastric cancer [377] and leukaemia [378], through glucose metabolism reprogramming [379], a hallmark of cancer [380]. Tumour suppressor, p53, regulates metabolic pathways, p53-transactivated TP53-induced glycolysis (TIGAR), and regulation of apoptosis in part through *SCO2* [376-378]. Finally, whereas these data were indifferent to decipher 8q24.21, this locus has also been shown to harbour risk SNPs for other cancers, which localize within distinct LD blocks and likely reflect tissue specificity.

5.4 Discussion

Principally, work within this chapter has identified a significant genetic correlation between MM and CLL and has discovered 10 risk loci shared between them, supporting epidemiological data demonstrating elevated familial risks between these B-cell malignancies [75]. Applying a working hypothesis that the loci may act in pleiotropic fashion, relevant cells representing a common tissue of disease origin were selected; namely naïve B-cells. While requiring biological validation, integration of data from CHi-C, chromatin mark enrichment and eQTL at shared loci has provided insight into how these loci may confer susceptibility to both CLL and MM. The shared loci identified could be grouped into those containing genes related to B-cell regulation and differentiation and those containing genes involved in angiogenesis, genome stability and apoptosis, supporting the tenet that these alleles can influence aetiology of either disease. With the expansion of GWAS of the B-cell malignancies, more detailed characterisation of common underlying risk alleles and affected pathways can inform the biology of B-cell oncogenesis.

CHAPTER 6 Search for multiple myeloma risk factors using Mendelian randomisation

6.1 Overview and rationale

The global burden of MM has substantially increased in the last 30 years, but its incidence is highly variable between different countries. Although MM is more common in high sociodemographic index countries, the temporal increase in disease incidence has been higher in middle and low-middle sociodemographic index countries [8]. This data suggests, albeit indirectly, that lifestyle factors influence the risk of developing MM.

Identifying aetiological risk factors for MM has the potential to inform prevention and intervention strategies to reduce disease burden. Numerous factors have been reported to affect the risk of either MM or its precursor monoclonal gammopathy of unknown significance (MGUS), including obesity [16-18, 20], diet [21-23], vitamin D [381] [24, 25], immune dysfunction [26] and radiation exposure [31, 32]. Aside from obesity, studies have either been inconsistent, found non-significant results or not been independently validated.

These observational epidemiological studies are, however, prone to reverse causation, unmeasured confounding and recall bias, which can preclude causal inferences [382]. Furthermore, the studies that have been conducted to date have had a limited scope of enquiry. Specifically, examining factors with established associations for other cancers or for which information can be readily collected.

Mendelian randomisation (MR) is an analytical method that exploits genetic variants as instrumental variables (IVs), to infer the causal relevance of an exposure to an outcome, such as a disease [191]. Because the genetic variants are randomly assigned at conception they are not influenced by reverse causation and in the absence of pleiotropy (*i.e.* genetic variants being associated with the disease through alternative pathways) they can provide unconfounded estimates of disease risk (**Figure 6.1**) [191]. So far, the application of MR to study MM aetiology has been confined to examining the relationship between obesity [383] and immunoglobulin [384] levels to MM risk. An agnostic strategy to identify causal relationships has recently been proposed, termed MR-PheWAS [295], which integrates a phenome-wide association study (PheWAS) and MR methodology.

To gain insight into the aetiological basis of MM, work within this chapter implements an MR-PheWAS to test purported associations and to search for novel causal relationships. Specifically, 249 phenotypes, proxied by 10,225 genetic variants, were analysed using summary genetic data from the genome-wide association study (GWAS) of MM in Chapter 3 comprising 7,717 case and 29,304 control subjects.

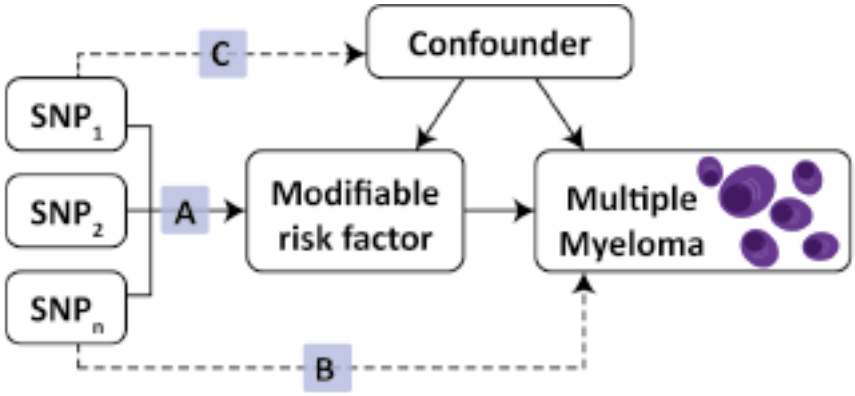


Figure 6.1 Principles of Mendelian randomisation. The assumptions that need to be satisfied to derive unbiased causal effect estimates. Dashed lines represent direct causal and potential pleiotropic effects that would violate Mendelian randomisation assumptions. A: genetic variants used as instrumental variables are only associated with the modifiable risk factor; B: genetic variants only influence the risk of developing MM through the modifiable risk factor; C: genetic variants are not associated with any measured or unmeasured confounders. SNP: single nucleotide polymorphism.

6.2 Study design

6.2.1 Genetic instruments for phenotypes

Two sample MR was conducted using the TwoSampleMR R package [296]. Full details of the quality control steps applied to select genetic instruments for analysis can be found in Section 2.4.12. Briefly, genetic instruments for each of the traits investigated were single nucleotide polymorphisms (SNPs) identified from recent meta-analyses, the largest studies published to date, or those curated by MR-Base [296] (**Appendix 19**). For each SNP, the chromosome position, the effect estimate expressed in standard deviations (SDs) of the trait per-allele and the corresponding standard errors (SEs) were recovered (**Appendix 20**). Only continuous traits were considered, as analysis of binary traits (such as disease status) with binary outcomes in two-sample MR frameworks can result in inaccurate causal estimates [297].

6.2.2 Multiple myeloma data

The association of each genetic instrument with MM risk was examined using summary statistics from the MM GWAS in Chapter 3, excluding the Icelandic dataset. This meta-analysis related > 3 million genetic variants to 7,717 MM cases and 29,304 controls of European descent. As some potentially modifiable reproductive risk factors are female-specific, where sex data was available, MM association statistics were computed using 2,190 female cases and 9,060 female controls.

6.2.3 Estimation of study power

The power of MR to demonstrate a causal effect depends on the percentage of risk factor variance explained by the genetic variants used as instruments. Stipulating an alpha of 0.05, the study power was estimated for each risk factor *a priori* across a range of effect sizes [107].

6.2.4 Mendelian randomisation analysis

MR methodology assumes that genetic variants, used as instruments for a risk factor, are associated with the risk factor and not with confounders or alternative causal relationships (**Figure 6.1**). Additionally, associations must be linear and unaffected by interactions [190]. For each SNP, causal effects were estimated for MM as an odds ratio per one SD unit increase in the putative risk factor (OR_{SD}), with 95% confidence intervals (CIs), using the Wald ratio. For traits with multiple SNPs as IVs, causal effects were estimated under inverse variance weighted random-effects (IVW-RE) and inverse variance weighted fixed-effects (IVW-FE) models. To assess the robustness of our findings, we also obtained weighted median estimates (WME) [298] and mode-based estimates (MBE) [299] for results which were suggestively significant and had >2 SNPs included in the analysis. Pleiotropy exists when a single genetic variant influences multiple phenotypes [300]. Horizontal pleiotropy refers to a situation where the genetic instrument influences disease outcome via a different pathway which is not under investigation. Where pleiotropic effects are balanced and there exists no systematic bias across a set of genetic instruments, MR estimates remain valid. If horizontal pleiotropy is unbalanced (directional) it may result in a biased MR estimate [300, 301]. Directional pleiotropy was therefore assessed using MR-Egger regression [302]. A consistent effect across these four complementary methods (IVW, MBE, WME and MR-Egger), which make different assumptions about horizontal pleiotropy, is less likely to be a false positive [303]. The potential impact of outlying and pleiotropic SNPs on causal estimates was examined adopting a leave-one-out strategy, under the IVW-RE model [296]. This method performs the MR analysis, but leaves out each SNP in turn

to identify whether a single SNP is driving the association. Heterogeneity observed within each trait (I^2) was calculated from Cochran's Q-value.

To account for multiple testing, a Bonferroni-corrected *P*-value of 2×10^{-4} (*i.e.* 0.05/249 putative risk factors) was considered as being statistically significant. A $P > 2\times 10^{-4}$ but < 0.05 was considered to be suggestive evidence of a causal association. Statistical analyses were performed using R version 3.4.0 and MR-Base [296].

6.3 Results

The median PVE by variants used as IVs for each of the 249 phenotypes examined as potential risk factors for MM was 5.45% (0.61 - 60.43%). The power of this study to demonstrate a causal association for MM is tabulated for each exposure in **Appendix 19**.

The strength of the association between each of the 249 phenotypes studied and risk of MM under IVW-RE models is shown in **Figure 6.2**; with corresponding tabulated data in **Appendix 21** and **Appendix 22**. None of the traits showed a statistically significant association with risk of MM, while 28 phenotypes showed suggestive evidence of association ($P < 0.05$) with risk of MM (**Figure 6.3**).

6.3.1 Fatty acids and metabolism

Genetically predicted increased levels of alpha-linolenic acid and decreased levels of docosapentaenoic acid, both omega-3 fatty acids (FAs), showed a suggestive association with MM risk (Wald ratio: OR_{SD}= 1.20, 95% CI: 1.04-1.38, *P*= 0.011 and IVW-RE: OR_{SD}=0.90, 95% CI: 0.81-0.99, *P*= 0.037 respectively). Overall, genetically predicted higher levels of omega-3 FAs were associated with a decreased risk of MM (IVW-RE: OR_{SD}= 0.74, 95% CI: 0.62-0.88, $P=5.4 \times 10^{-4}$); causal effect estimates being similar under WME and MBE approaches (**Appendix 23**). In the omega-6 FA class, decreased levels of adrenic acid, arachidonic acid and gamma-linolenic and increased levels of dihomo-gamma-linoleic acid and linoleic acid were associated with increased risk of MM (**Figure 6.3**). While FAs within the class were individually significant, overall the omega-6 FAs as a class were not suggestively associated with increased risk of MM. Similarly, while higher levels of oleic acid were suggestively associated with increased MM risk, overall omega-7 and omega-9 FA classes were not significant. FA metabolism involves sequential enzymatic conversions and genes involved in FA processing form parts of numerous FA pathways. As a result, SNPs influencing the metabolism of one FA are often associated with circulating concentrations of multiple FAs [385]. Leave-one-out analysis showed rs174547 was

a major driver of association across multiple FAs, although omega-3 FAs as a class remained significant after excluding this SNP from the analysis (*P*= 0.020, **Appendix 24**). When applying WME and MBE approaches, causal effect estimates for omega-3 FAs remained significant.

Increased levels of genetically predicted blood carnitine showed a suggestive association with increased risk of MM (OR_{SD} = 1.13, 95% CI: 1.05-1.22, *P* = 1.1×10-3). MR Egger analysis did not show evidence of bias in causal estimates ($P_{intercept}$ > 0.05, **Appendix 25**) and leave-one-out analysis demonstrated no single SNP was driving the association (**Appendix 24**). Although altered levels of a number of acyl carnitine esters were also suggestively significant for MM risk, including *cis*-4-decenoyl carnitine, decanoylcarnitine, hexanoylcarnitine, hydroxyisovaleroyl carnitine, isovalerylcarnitine, octanoylcarnitine, propionylcarnitine and stearoylcarnitine, these acyl cartnitines follow similar biosynthetic pathways and their levels may influenced by the same underlying SNPs [386].

6.3.2 Telomere length

While genetically increased telomere length was associated with MM risk (IVW-RE: OR_{SD} = 2.33, 95% CI: 1.20-4.52, *P* = 0.013), there was marked heterogeneity between the seven SNPs used as IVs (I^2 = 86%). The association was primarily driven by the 3q26 *TERC* SNP (rs10936599) and after exclusion of this SNP the association was non-significant (*P* = 0.161) (**Appendix 24**, **Figure 6.4**Error! Reference source not found.). This SNP has previously been associated with MM [113].

6.3.3 Diet, lifestyle and other factors

Among the dietary factors considered, an increased level of serum vitamin B6 was suggestively associated with increased risk of MM (OR_{SD} = 1.26, 95% CI: 1.01-1.58 *P* = 0.041), while vitamin D, which has been suggested as a risk factor for MM [25], was not associated with MM risk in this study (*P* = 0.54).

In keeping with previous findings from meta-analysis of prospective studies which have demonstrated an association between obesity and increased risk of MM (relative risk = 1.21 95% CI: 1.08-1.35) [20], increased BMI was associated with increased MM risk, albeit non-significant (OR: 1.10, 95% CI: 0.99-1.22, *P*=0.082). All other obesity-related traits including whole body water mass, basal metabolic rate, weight, impedance of whole body, body mass index, whole body fat mass, body fat percentage, trunk fat percentage, waist circumference, birth weight, hip

circumference, waist-to-hip ratio and birth weight of first child demonstrated non-significant associations (**Appendix 21**).

Furthermore, this analysis showed non-significant relationships between IL-6 polymorphisms and IL-6 receptors with MM risk (**Appendix 21**).

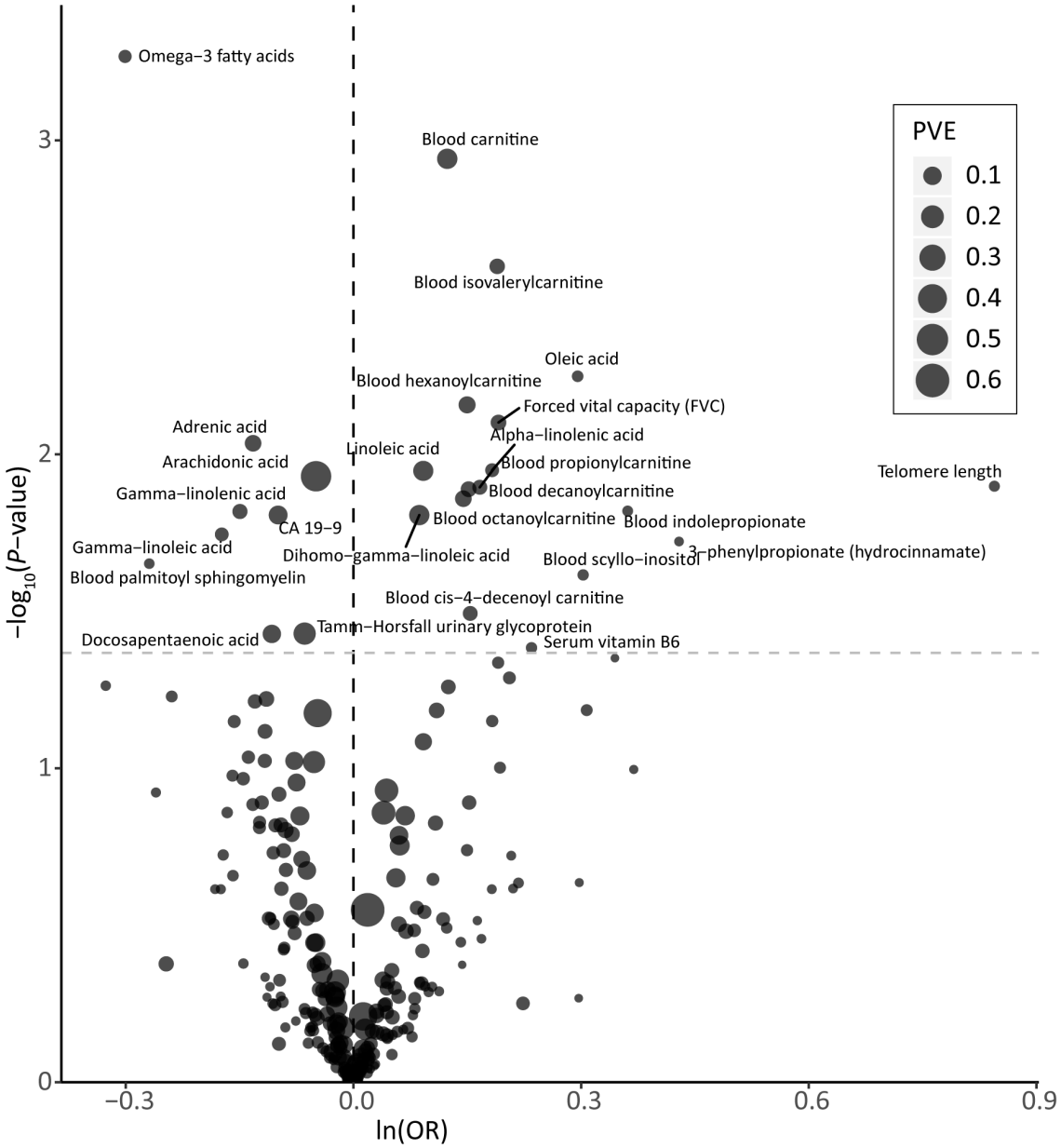


Figure 6.2 Volcano plot of odds ratio of the association between 249 phenotypes with risk of MM. Odds ratio per standard deviation from random-effects inverse variance weighted or Wald ratio Mendelian randomisation analysis of 256 phenotypes with risk of MM. Dashed grey line corresponds to *P* = 0.05. PVE, proportion of variance explained.

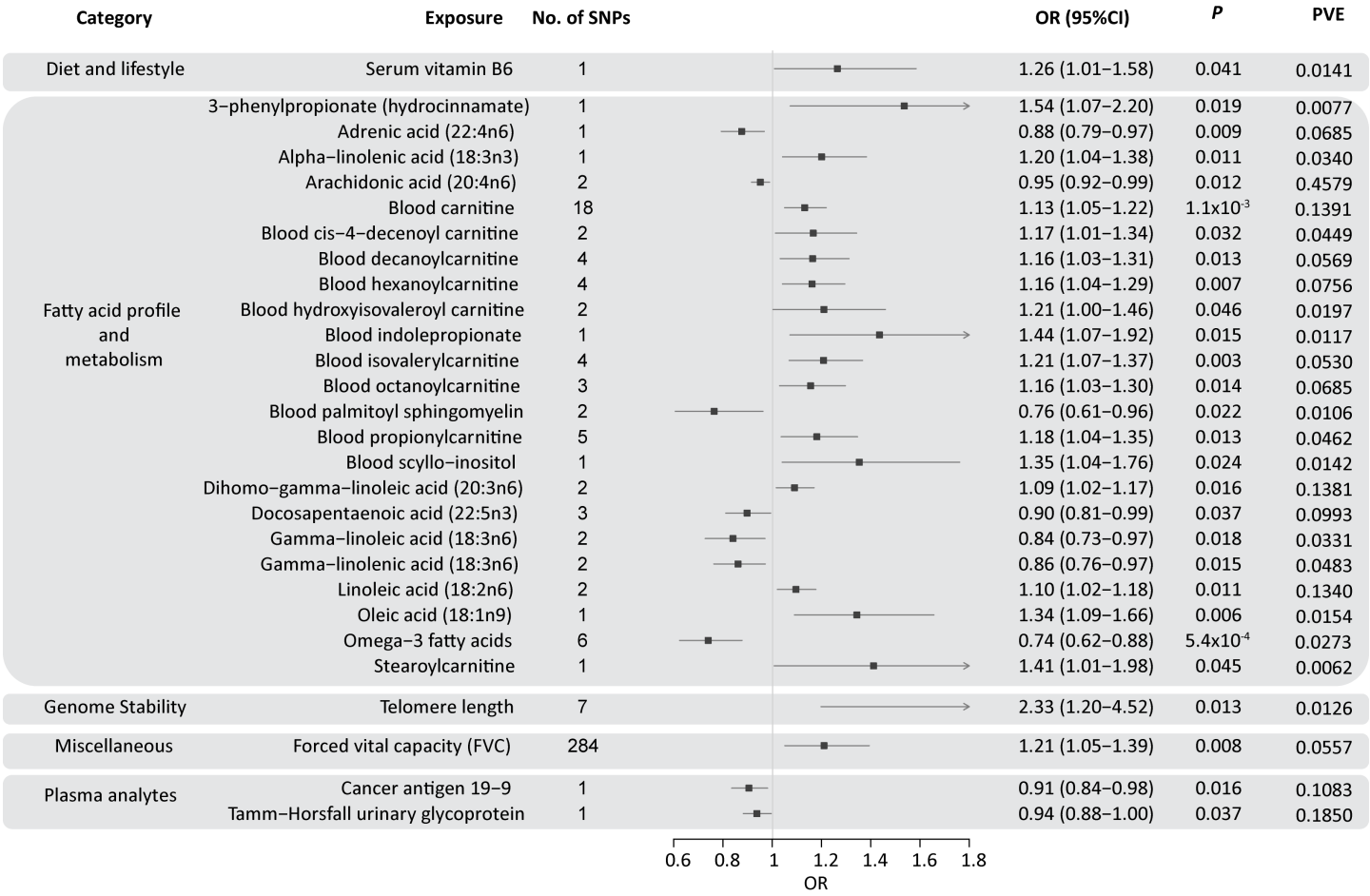


Figure 6.3 Forest plot of 28 phenotypes suggestively associated with risk of MM. Confidence intervals indicated by line width. Vertical line denotes the null value (OR_{SD} = 1).

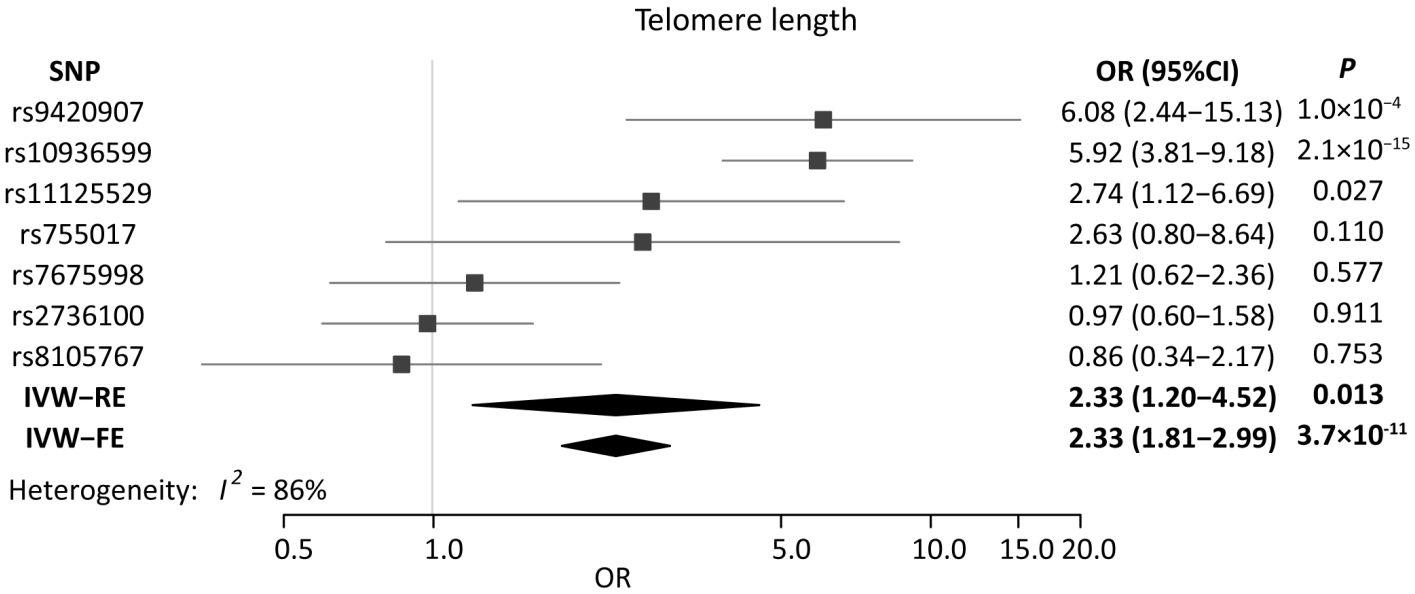


Figure 6.4 Forest plot showing the effect of alleles associated with longer telomere length on MM risk. Diamonds represents overall causal effects estimated using fixed- and random-effects inverse variance weighted models (IVW-FE and IVW-RE, respectively). Confidence intervals indicated by diamond width. Vertical line denotes the null value (OR_{SD} = 1).

6.4 Discussion

Despite its comparative rarity, MM is one of the cancers of unmet need given the significant morbidity and mortality associated with it. Incidence of MM limits the power of a conventional cohort study to demonstrate a causal association. As a consequence of this, little is known about the aetiological basis of MM, which is a barrier to developing strategies to reduce disease burden [387]. This contrasts markedly to the success of cohort and case control studies of the common cancers such as breast [388], lung [389] and colorectal [390-392] which have identified major determinants of risk.

MR can circumvent many limitations of a conventional observational study and the methodology is therefore increasingly being used to examine the impact of interventions on disease risk. The value of MR has been greatly enhanced by the wealth of GWAS data now available on multiple traits, which provide SNPs that can be used as IVs. These data enable testing of the relationship between multiple traits and MM risk in a hypothesis-free manner by performing a MR-PheWAS. Notably, among the 249 exposures analysed using IVW-RE the majority of the suggestively significant results were related to fatty acid transport and fatty acid oxidation (FAO) pathways, including carnitine, acyl carnitines and omega-3 fatty acids. Briefly, fatty acids are transported into the mitochondria where they are oxidised with concomitant production of nicotinamide adenine dinucleotide (NADH), nicotinamide adenine dinucleotide phosphate (NADPH), flavin adenine dinucleotide (FADH2), and ATP for energy to sustain cellular metabolism. During this process, carnitine, fatty acids and acyl-CoA are utilised [393]. The metabolic requirement of plasma cells to perform antibody production and how this alters when the cells become malignant, as in MM, is relatively unknown, though studies have shown that B-cells are metabolically flexible to support the production and secretion of antibodies [394]. This metabolic reprogramming may be mediated by cells in the bone marrow microenvironment, such as bone marrow adipocytes, which store triglycerides and convert them to fatty acids. The constituents of the microenvironment may shift myeloma cells from aerobic glycolysis to utilise readily available fatty acids and produce more energy by FAO. As such, targeting FAO in MM is an area of interest for therapeutic investigation [393]. Such metabolic reprogramming, itself a hallmark of cancer [380], is still not fully understood, and thus this analysis provides support for aberrant fatty acid and blood carnitine levels influencing MM risk using genetic markers as IVs.
Longer telomere length has been associated with risk of MM [126] and other cancers, including glioma [128]. The relationship between telomere length and cancer risk is a long-standing question in cancer epidemiology, though it has been proposed that a predisposition to longer

telomeres may permit cells to escape growth arrest and so undergo malignant transformation [129]. In this analysis, we found that longer telomere length was nominally associated with increased risk of MM. This was predominantly driven by a SNP in the *TERC* gene, with other variants showing only limited support for an association (**Figure 6.4**Error! Reference source not found., **Appendix 24**).

Additionally, this analysis found no evidence for association between traits that have previously been considered as potential risk factors for MM, including vitamin D [25] and IL-6 polymorphisms [395, 396]. However, albeit non-significant, this study provides some supporting evidence for the reported association between obesity and risk of MM [18-20, 397]. Intriguingly observational studies, have demonstrated an increased risk of transformation from MGUS to MM in overweight and obese individuals [398, 399], suggesting obesity-related pathways being determinants of tumour progression rather than affecting early phase of neoplasic development.

This analysis has been able to leverage a greater number of SNPs as IVs, thereby increasing study power; for 202 of the exposures, we had at least 80% power to demonstrate an OR_{SD} of 1.33 stipulating a *P*-value of 0.05. However, there is a possibility that the null results we observed were simply a consequence of limited study power if the true effect of these phenotypes is marginal. Furthermore, the causal effects estimated by MR-Egger were non-significant for many phenotypes, although this may be the result of reduced power of this test to detect causal effects compared to other MR methodologies [302].

The strength of this MR study is the exploitation of large GWAS datasets to examine the relationship between multiple phenotypes and risk of MM thereby increasing study power and enabling demonstration of effects of small magnitude. A central assumption in MR is that the variants used as IVs are associated with the exposure being investigated. To ensure this was the case, only SNPs associated with exposure traits at genome-wide significance ($P < 5 \times 10^{-8}$) from GWAS were used. Furthermore, only the data from individuals of European descent were used to limit bias from population stratification. This analysis does however have limitations. Firstly, it is limited to studying phenotypes with genetic instruments available. Secondly, correcting for multiple testing inevitably means the potential for false-negatives is not unsubstantial. Thirdly, though only traits for which there was >80% study power at OR_{SD}=1.50 were considered, for a

large number of traits there was still limited power to demonstrate causal associations of small effect.

In conclusion, the work within this chapter has provided further insight into the landscape of MM aetiology and shed light on factors for which the evidence from conventional epidemiological studies has been mixed. The advent of larger meta-analyses of MM GWAS datasets and exposures offers the prospect of using MR-based strategies to search for possible causal associations with smaller effect sizes.

CHAPTER 7 Discussion

7.1 Identification of novel susceptibility loci for myeloma

To date genome-wide association studies (GWAS) have informed much of our understanding of inherited susceptibility to MM. While previous studies have identified 17 MM risk loci [112-114, 116], the majority of the heritability remains unexplained [158].

With an aim to gain further insight into inherited susceptibility to MM and identify new risk loci, a new GWAS was performed and meta-analysed with previous GWAS and a replication series, totalling 9,974 MM cases and 247,556 controls of European ancestry. Six new loci were identified; 2q31.1, 5q23.2, 7q22.3, 7q31.33, 16p11.2 and 19p13.11. Previously identified loci accounted for 13.6% of the GWAS heritability, while the additional loci discovered account for 2.1%. Collectively the discovered loci account for 15.7% of the heritability of MM. Construction of polygenic risk scores (PRS) considering the combined effect of all risk single nucleotide polymorphisms (SNPs) found that an individual in the top 1% of genetic risk has a threefold increased risk of MM when compared to an individual with median genetic risk.

7.2 Functional annotation and biological inference of myeloma risk loci

The risk loci identified from MM GWAS were shown to map to genomic regions of cell-type-specific active chromatin, as indicated by the presence of H3K27ac, H3K4me3 and H3K4me1 histone marks. This supports the idea that risk loci likely influence risk via subtle regulatory effects on gene expression. To gain insight into the possible biological and functional mechanisms underlying all 23 MM risk loci, data from patient expression quantitative trait locus (eQTL) analysis, B-cell-specific transcription factors (TFs) and histone marks, and promoter capture Hi-C (CHi-C) data was integrated to prioritise candidate genes at each locus. This comprehensive analysis of all MM risk loci used novel CHi-C and ChIP-seq data from MM cell lines to inform the prioritisation of a candidate gene at each locus. While some of the identified genes at newly discovered loci may be plausible for candidate gene studies given the current knowledge of MM biology (e.g. *KLF2* at 19p13.11), it is unlikely that others could have been anticipated. For example, *CEP120* at 5q23.2 would have been unlikely *a priori* to be considered a candidate based on the existing knowledge of its function. Importantly, variation at 5q23.2 is now associated with *CEP120* expression in MM patients, highlighting the significance of the agnostic approach of GWAS.

The genes identified could be broadly grouped into four potential candidate disease mechanisms; those related to B-cell development and function; those related to cell cycle and genomic instability; those related to apoptosis/autophagy; and those related involved in chromatin remodelling. Biological investigation of the functional mechanism behind all the risk loci is likely to generate profound insight into MM biology and pathogenesis.

The transcriptome-wide association study (TWAS) performed in Chapter 4 provided evidence for further genes underlying GWAS associations, as well as support for genes discussed in Chapter 3. In contrast to the Summary-data-based Mendelian Randomization (SMR) analysis performed in chapter 3, this analysis leveraged data from multiple non-tumour tissues, making it complementary to the expression data analysis in Chapter 3, which used patient expression data. While MM is a malignancy of plasma cells, there is increasing evidence of the role of the microenvironment in progression of MM precursor lesion monoclonal gammopathy of undetermined significance (MGUS) to MM and in sustaining MM. Furthermore, malignant transformation to form MM may occur at an early stage of B-cell development; a proposition potentially supported by the genetic correlation with related B-cell malignancy chronic lymphocytic leukaemia (CLL) and identification and annotation of pleiotropic risk loci between these diseases reported in Chapter 5. As such a TWAS may not be best represented by patient plasma cell expression data. Notably, this TWAS found evidence for *APOBEC3C*, *APOBEC3D*, *APOBEC3F*, *APOBEC3G* and *APOBEC3H* at 22q13.1 as playing a role in defining MM predisposition. *APOBEC* cytidine deaminase activity is a recognised feature of MM, caused by triggering DNA mutation [343, 344].

7.3 Genetic correlation between B-cell malignancies

Studies prior to this thesis had provided evidence for shared susceptibility to MM and CLL [75]. Both these are malignancies of B-cell origin, however prior studies were based on observational familial relationships, so may not distinguish between environmental and genetic factors influencing disease risk. Application of linkage disequilibrium (LD) score regression to examine a correlation between MM and CLL found a positive genetic correlation between these malignancies, suggesting that this shared susceptibility does have an inherited genetic basis. Identification and annotation of pleiotropic risk loci within this chapter demonstrated enrichment of these loci in naïve B-cell and CD38- B-cells and may provide evidence that inherited predisposition to both malignancies may be happening at an early B-cell stage. Notably, genes identified by annotation of these risk loci were related to B-cell development providing further support. Exploration of genetic correlation with additional B-cell malignancies

(indicated in **Figure 1.4**) would provide further insight into the shared aetiological basis of lymphoid malignancies, however these analyses are currently limited by the small sample sizes. The methodology typically requires samples sizes in the thousands and the low incidence of these malignancies precludes studies into genetic correlation using LD score regression.

7.4 Investigating aetiological risk factors for myeloma

In addition to LD score regression, application of Mendelian randomisation (MR) using GWAS datasets can investigate the aetiological basis of disease [190, 191, 400]. Work in Chapter 6 performed a phenome-wide association study using MR (MR-PheWAS), examining the relationship between 249 exposures and MM risk using instrumental variables (IVs) constructed from GWAS datasets. Although no significant associations with MM risk were observed among the 249 phenotypes, 28 phenotypes showed evidence suggestive of association, including decreased blood carnitine (P=1.1×10^{-3}) and increased levels of omega-3 fatty acids (P=5.4×10^{-4}) with reduced MM risk. Few previously suggested modifiable risk factors showed evidence of association with MM; for example obesity and vitamin D. It is also notable that the most significantly associated risk factors were related to metabolism. Plasma and B-cells of earlier lineage have a unique metabolic flexibility to support the production and secretion of antibodies. Such metabolic reprogramming is itself a hallmark of cancer [380] and, in combination with previous studies in the area of MM metabolic pathways, this work may provide support for targeting of the pathways around fatty acid oxidation and carnitine metabolism as areas of therapeutic interest in MM. This work also highlights the value in leveraging data from a wide range of GWAS datasets to inform cancer biology.

7.5 Future studies in genetic predisposition to myeloma

Collectively the previously identified and new risk loci account for 15.7% of the GWAS heritability (13.6% and 2.1%, respectively), so much of the heritability of MM remains unexplained. Future GWAS and meta-analysis with larger datasets may uncover more common risk loci of low effect size. It is possible that as MM GWAS sample sizes increase, further variants associated with MM karyotypes may be identified, as for MM risk locus 11q13.3 which has been associated with t(11;14) [115], and the association of 5q15 risk locus, driven by HRD MM [142]. Primary translocation and HRD occur at the initiating stages of MM development therefore discovery of subtype associated loci may provide insight into the aetiology of MM initiation and subtype determination. Furthermore, application of TWAS and genetic correlation to GWAS of MM subtypes may provide insight into biology underlying specific karyotypes. Currently, small sample sizes limit power and preclude application of such techniques for subtype associations.

It is possible that rare (MAF < 0.01) germline variants contribute to the 'missing heritability' for MM however GWAS are underpowered to identify these variants [93].

Whole-genome sequencing of MM germline samples may reveal novel loci previously undiscovered through GWAS and imputation, as well as explore the possible existence of high-penetrance non-coding susceptibility alleles. Indeed, there has been a shift in recent years to analyses of whole genome and whole exome sequencing, however analysis of large-scale whole-genome sequencing data brings challenges [401]. Currently the rate of generation of high-throughput sequencing data is exceeding the pace at which it can be analysed with care and accuracy. Whole-exome sequencing is more cost effective than whole-genome sequencing, however, the non-coding region of the genome harbours most of the genetic variants involved in disease predisposition [167]. Custom capture of regulatory regions is an attractive alternative, however this requires prior knowledge of potential regions of interest and negates the agnostic nature which is an asset of GWAS.

Deconvolution of the functional mechanisms behind MM risk loci will be invaluable for the clinical utility of GWAS, especially in the case of MM where the aetiology of the disease is still largely unknown. Laboratory techniques employed in these studies, such as luciferase reporter assays, are low-throughput, timely and costly. Recently there have been advances in functional assays, including massively parallel reporter assays (MPRA) [402-406] and Self-Transcribing Active Regulatory Region sequencing (STARR-seq) [402, 407]. Both of these assays aim to identify regulatory regions using a high throughout sequencing-based methodology, however the methods have different library sources. In STARR-seq a source genome is fragmented and ligated before recombination into a vector, while in MPRA, a library is designed *in silico* and is synthesised as a pool of oligos which can be inserted in to a vector. As such, MPRA may be the desirable assay for prioritisation of genetic variants in LD at an identified locus.

7.6 Overall conclusion

In summary this thesis has studied inherited susceptibility to MM, providing quantitative and qualitative information about germline genetic contribution to disease risk and biology and highlighting the polygenic nature of the disease. In Chapter 3, a GWAS was performed, identifying six new loci. Variation in common SNPs was demonstrated to explain approximately 15.6% of MM risk and polygenic risk scores were constructed to identify the increased risk of MM in those with the highest genetic risk. Furthermore, a global analysis of all the new and established MM risk loci was performed, which delineated four potential candidate disease

mechanisms across the 23 MM risk loci, by integrating regulatory data on histone marks, TFs, CHi-C and patient gene expression. A TWAS, performed in Chapter 4, provided further evidence which consolidated many of the previously implicated genes and provided new potential genes dysregulated in MM, including those from the *APOBEC* family of genes. A genetic correlation between CLL and MM detailed in Chapter 5, and subsequent annotation of identified pleiotropic loci, highlighted that common pathways may be involved in malignant transformation of progenitor B-cells to either disease. Exploration of tissues outside of the myeloma and plasma cells, for example functional studies in germinal centre models, co-cultured plasma cells and the tumour microenvironment, may be essential to fully appreciate the role and context of GWAS loci in disease risk. A strength of GWAS is the reproducibility of risk loci; indeed, regions of the genome associated with MM risk are repeatedly validated with each subsequent GWAS and meta-analysis performed. This reflects the strength of the study design, which aims to reduce 'winners curse', by considering factors such as replication cohort sample size and population stratification. Risk loci identified via GWAS therefore provide a robust, reproducible origin from which a proposed mechanism of disease origin or progression can be functionally annotated, beginning from a genetic association at the level of DNA, working towards understanding of aberrant regulation of gene pathways and malignant cell transformation. The collective findings from this thesis suggest future efforts in genetic predisposition to MM are likely to involve further GWAS, whole-exome sequencing and integration of MPRA with regulatory datasets to functionally annotate risk loci and validate mechanisms of disease biology.

References

1. International Myeloma Working, G., *Criteria for the classification of monoclonal gammopathies, multiple myeloma and related disorders: a report of the International Myeloma Working Group.* British journal of haematology, 2003. **121**(5): p. 749-757.
2. Hanamura, I., et al., *Frequent gain of chromosome band 1q21 in plasma-cell dyscrasias detected by fluorescence in situ hybridization: incidence increases from MGUS to relapsed myeloma and is related to prognosis and disease progression following tandem stem-cell transplantation.* Blood, 2006. **108**(5): p. 1724-1732.
3. UK, C.R. [cited 2020 April 2020]; Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/myeloma/incidence.
4. Dutta, A.K., et al., *Subclonal evolution in disease progression from MGUS/SMM to multiple myeloma is characterised by clonal stability.* Leukemia, 2019. **33**(2): p. 457-468.
5. Nutt, S.L., et al., *The generation of antibody-secreting plasma cells.* Nat Rev Immunol, 2015. **15**(3): p. 160-71.
6. Kumar, S.K., et al., *Multiple myeloma.* Nature reviews. Disease primers, 2017. **3**: p. 17046-17046.
7. Brenner, H., A. Gondos, and D. Pulte, *Recent major improvement in long-term survival of younger patients with multiple myeloma.* Blood, 2008. **111**(5): p. 2521-2526.
8. Cowan, A.J., et al., *Global Burden of Multiple Myeloma: A Systematic Analysis for the Global Burden of Disease Study 2016.* JAMA oncology, 2018. **4**(9): p. 1221-1227.
9. Becker, N., *Epidemiology of multiple myeloma.* Recent results in cancer research. Fortschritte der Krebsforschung. Progres dans les recherches sur le cancer, 2011. **183**: p. 25-35.
10. Smith, A., et al., *Guidelines on the diagnosis and management of multiple myeloma 2005.* British journal of haematology, 2006. **132**(4): p. 410-451.
11. Morgan, G.J., F.E. Davies, and M. Linet, *Myeloma aetiology and epidemiology.* Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie, 2002. **56**(5): p. 223-234.
12. Renshaw, C., et al., *Trends in the incidence and survival of multiple myeloma in South East England 1985-2004.* BMC cancer, 2010. **10**: p. 74-74.
13. Reis, L., et al., *SEER Cancer statistics review, 1973-1997.* Bethesda, MD: National Cancer Institute, 2000.
14. Waxman, A.J., et al., *Racial disparities in incidence and outcome in multiple myeloma: a population-based study.* Blood, 2010. **116**(25): p. 5501-5506.
15. Greenberg, A.J., C.M. Vachon, and S.V. Rajkumar, *Disparities in the prevalence, pathogenesis and progression of monoclonal gammopathy of undetermined significance and multiple myeloma between blacks and whites.* Leukemia, 2012. **26**(4): p. 609-614.
16. Carson, K.R., M.L. Bates, and M.H. Tomasson, *The skinny on obesity and plasma cell myeloma: a review of the literature.* Bone Marrow Transplant, 2014. **49**(8): p. 1009-15.
17. De Pergola, G. and F. Silvestris, *Obesity as a major risk factor for cancer.* J Obes, 2013. **2013**: p. 291546.
18. Teras, L.R., et al., *Body size and multiple myeloma mortality: a pooled analysis of 20 prospective studies.* Br J Haematol, 2014. **166**(5): p. 667-76.
19. Birmann, B.M., et al., *Body mass index, physical activity, and risk of multiple myeloma.* Cancer Epidemiol Biomarkers Prev, 2007. **16**(7): p. 1474-8.

20. Wallin, A. and S.C. Larsson, *Body mass index and risk of multiple myeloma: a meta-analysis of prospective studies.* Eur J Cancer, 2011. **47**(11): p. 1606-15.
21. Thordardottir, M., et al., *Dietary intake is associated with risk of multiple myeloma and its precursor disease.* PloS one, 2018. **13**(11): p. e0206047-e0206047.
22. Fritschi, L., et al., *Dietary fish intake and risk of leukaemia, multiple myeloma, and non-Hodgkin lymphoma.* Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology, 2004. **13**(4): p. 532-537.
23. Brown, L.M., et al., *Diet and nutrition as risk factors for multiple myeloma among blacks and whites in the United States.* Cancer causes & control : CCC, 2001. **12**(2): p. 117-125.
24. Gascoyne, D.M., et al., *Vitamin D Receptor Expression in Plasmablastic Lymphoma and Myeloma Cells Confers Susceptibility to Vitamin D.* Endocrinology, 2017. **158**(3): p. 503-515.
25. Burwick, N., *Vitamin D and plasma cell dyscrasias: reviewing the significance.* Annals of hematology, 2017. **96**(8): p. 1271-1277.
26. Lindqvist, E.K., et al., *Personal and family history of immune-related conditions increase the risk of plasma cell disorders: a population-based study.* Blood, 2011. **118**(24): p. 6284-6291.
27. Landgren, O., et al., *Agent Orange Exposure and Monoclonal Gammopathy of Undetermined Significance: An Operation Ranch Hand Veteran Cohort Study.* JAMA oncology, 2015. **1**(8): p. 1061-1068.
28. Merhi, M., et al., *Occupational exposure to pesticides and risk of hematopoietic cancers: meta-analysis of case-control studies.* Cancer causes & control : CCC, 2007. **18**(10): p. 1209-1226.
29. Khuder, S.A. and A.B. Mutgi, *Meta-analyses of multiple myeloma and farming.* American journal of industrial medicine, 1997. **32**(5): p. 510-516.
30. LeMasters, G.K., et al., *Cancer risk among firefighters: a review and meta-analysis of 32 studies.* Journal of occupational and environmental medicine, 2006. **48**(11): p. 1189-1202.
31. Hsu, W.-L., et al., *The incidence of leukemia, lymphoma and multiple myeloma among atomic bomb survivors: 1950-2001.* Radiation research, 2013. **179**(3): p. 361-382.
32. Preston, D.L., et al., *Cancer incidence in atomic bomb survivors. Part III: Leukemia, lymphoma and multiple myeloma, 1950-1987.* Radiation research, 1994. **137**(2s): p. S68-S97.
33. Liu, T., et al., *Occupational exposure to methylene chloride and risk of cancer: a meta-analysis.* Cancer causes & control : CCC, 2013. **24**(12): p. 2037-2049.
34. Ferlay, J., et al., *Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods.* Int J Cancer, 2019. **144**(8): p. 1941-1953.
35. 1975-2017, S.C.S.R. [cited 2020 April 2020]; Surveillance, Epidemiology, and End Results program]. Available from: https://seer.cancer.gov/csr/1975_2017/.
36. Shapiro-Shelef, M. and K. Calame, *Regulation of plasma-cell development.* Nature Reviews Immunology, 2005. **5**(3): p. 230-242.
37. Fairfax, K.A., et al., *Plasma cell development: from B-cell subsets to long-term survival niches.* Seminars in immunology, 2008. **20**(1): p. 49-58.
38. Martin, F., A.M. Oliver, and J.F. Kearney, *Marginal zone and B1 B cells unite in the early response against T-independent blood-borne particulate antigens.* Immunity, 2001. **14**(5): p. 617-629.
39. González, D., et al., *Immunoglobulin gene rearrangements and the pathogenesis of multiple myeloma.* Blood, 2007. **110**(9): p. 3112-3121.
40. Roco, J.A., et al., *Class-Switch Recombination Occurs Infrequently in Germinal Centers.* Immunity, 2019. **51**(2): p. 337-350.e7.

41. Slifka, M.K., M. Matloubian, and R. Ahmed, *Bone marrow is a major site of long-term antibody production after acute viral infection.* Journal of virology, 1995. **69**(3): p. 1895-1902.
42. Tangye, S.G., et al., *Intrinsic differences in the proliferation of naive and memory human B cells as a mechanism for enhanced secondary immune responses.* Journal of immunology (Baltimore, Md. : 1950), 2003. **170**(2): p. 686-694.
43. Fenton, J.A.L., et al., *Isotype class switching and the pathogenesis of multiple myeloma.* Hematological oncology, 2002. **20**(2): p. 75-85.
44. Kyle, R.A., et al., *A long-term study of prognosis in monoclonal gammopathy of undetermined significance.* The New England journal of medicine, 2002. **346**(8): p. 564-569.
45. Kyle, R.A., et al., *Clinical course and prognosis of smoldering (asymptomatic) multiple myeloma.* The New England journal of medicine, 2007. **356**(25): p. 2582-2590.
46. Kyle, R.A. and S.V. Rajkumar, *Multiple myeloma.* The New England journal of medicine, 2004. **351**(18): p. 1860-1873.
47. Morgan, G.J., B.A. Walker, and F.E. Davies, *The genetic architecture of multiple myeloma.* Nature reviews. Cancer, 2012. **12**(5): p. 335-348.
48. Gould, J., et al., *Plasma cell karyotype in multiple myeloma.* Blood, 1988. **71**(2): p. 453-456.
49. Sawyer, J.R., et al., *Cytogenetic findings in 200 patients with multiple myeloma.* Cancer genetics and cytogenetics, 1995. **82**(1): p. 41-49.
50. Smadja, N.V., et al., *Chromosomal analysis in multiple myeloma: cytogenetic evidence of two different diseases.* Leukemia, 1998. **12**(6): p. 960-969.
51. Keim, C., et al., *Regulation of AID, the B-cell genome mutator.* Genes & development, 2013. **27**(1): p. 1-17.
52. Nutt, S.L., et al., *The genetic network controlling plasma cell differentiation.* Seminars in immunology, 2011. **23**(5): p. 341-349.
53. Walker, B.A., et al., *A compendium of myeloma-associated chromosomal copy number abnormalities and their prognostic value.* Blood, 2010. **116**(15): p. e56-65.
54. Chng, W.J., et al., *Clinical significance of TP53 mutation in myeloma.* Leukemia, 2007. **21**(3): p. 582-584.
55. Fonseca, R., et al., *Clinical and biologic implications of recurrent genomic aberrations in myeloma.* Blood, 2003. **101**(11): p. 4569-4575.
56. Avet-Loiseau, H., et al., *Genetic abnormalities and survival in multiple myeloma: the experience of the Intergroupe Francophone du Myélome.* Blood, 2007. **109**(8): p. 3489-3495.
57. Drach, J., et al., *Presence of a p53 gene deletion in patients with multiple myeloma predicts for short survival after conventional-dose chemotherapy.* Blood, 1998. **92**(3): p. 802-809.
58. An, G., et al., *Chromosome 1q21 gains confer inferior outcomes in multiple myeloma treated with bortezomib but copy number variation and percentage of plasma cells involved have no additional prognostic value.* Haematologica, 2014. **99**(2): p. 353-9.
59. Weinhold, N., et al., *Concomitant gain of 1q21 and MYC translocation define a poor prognostic subgroup of hyperdiploid multiple myeloma.* Haematologica, 2016. **101**(3): p. e116-9.
60. Rajkumar, S.V., et al., *International Myeloma Working Group updated criteria for the diagnosis of multiple myeloma.* The Lancet. Oncology, 2014. **15**(12): p. e538-e548.
61. Smith, D. and K. Yong, *Advances in understanding prognosis in myeloma.* British Journal of Haematology, 2016. **175**(3): p. 367-380.
62. Brigle, K. and B. Rogers, *Pathobiology and Diagnosis of Multiple Myeloma.* Semin Oncol Nurs, 2017. **33**(3): p. 225-236.

63. Palumbo, A., et al., *Revised International Staging System for Multiple Myeloma: A Report From International Myeloma Working Group.* Journal of clinical oncology : official journal of the American Society of Clinical Oncology, 2015. **33**(26): p. 2863-2869.
64. Zhan, F., et al., *The molecular classification of multiple myeloma.* Blood, 2006. **108**(6): p. 2020-2028.
65. Walker, B.A., et al., *Mutational Spectrum, Copy Number Changes, and Outcome: Results of a Sequencing Study of Patients With Newly Diagnosed Myeloma.* Journal of clinical oncology : official journal of the American Society of Clinical Oncology, 2015. **33**(33): p. 3911-3920.
66. Mikhael, J.R., et al., *Management of newly diagnosed symptomatic multiple myeloma: updated Mayo Stratification of Myeloma and Risk-Adapted Therapy (mSMART) consensus guidelines 2013.* Mayo Clinic proceedings, 2013. **88**(4): p. 360-376.
67. Usmani, S.Z., et al., *Second malignancies in total therapy 2 and 3 for newly diagnosed multiple myeloma: influence of thalidomide and lenalidomide during maintenance.* Blood, 2012. **120**(8): p. 1597-1600.
68. Bird, J.M., et al., *Guidelines for the diagnosis and management of multiple myeloma 2011.* British journal of haematology, 2011. **154**(1): p. 32-75.
69. Smith, D. and K. Yong, *Multiple myeloma.* BMJ (Clinical research ed.), 2013. **346**: p. f3863-f3863.
70. Gerecke, C., et al., *The Diagnosis and Treatment of Multiple Myeloma.* Deutsches Arzteblatt international, 2016. **113**(27-28): p. 470-476.
71. Accardi, F., et al., *Mechanism of Action of Bortezomib and the New Proteasome Inhibitors on Myeloma Cells and the Bone Microenvironment: Impact on Myeloma-Induced Alterations of Bone Remodeling.* BioMed research international, 2015. **2015**: p. 172458-172458.
72. Sharma, S. and A. Lichtenstein, *Dexamethasone-induced apoptotic mechanisms in myeloma cells investigated by analysis of mutant glucocorticoid receptors.* Blood, 2008. **112**(4): p. 1338-1345.
73. Naymagon, L. and M. Abdul-Hay, *Novel agents in the treatment of multiple myeloma: a review about the future.* Journal of hematology & oncology, 2016. **9**(1): p. 52-52.
74. Sud, A., et al., *The landscape of familial risk of hematological malignancies: an analysis of 153,115 cases.* (Under review).
75. Sud, A., et al., *Analysis of 153 115 patients with hematological malignancies refines the spectrum of familial risk.* Blood, 2019. **134**(12): p. 960-969.
76. Landgren, O., et al., *Risk of plasma cell and lymphoproliferative disorders among 14621 first-degree relatives of 4458 patients with monoclonal gammopathy of undetermined significance in Sweden.* Blood, 2009. **114**(4): p. 791-795.
77. Vachon, C.M., et al., *Increased risk of monoclonal gammopathy in first-degree relatives of patients with multiple myeloma or monoclonal gammopathy of undetermined significance.* Blood, 2009. **114**(4): p. 785-790.
78. Altieri, A., et al., *Familial risks and temporal incidence trends of multiple myeloma.* Eur J Cancer, 2006. **42**(11): p. 1661-70.
79. Landgren, O., et al., *Risk of plasma cell and lymphoproliferative disorders among 14621 first-degree relatives of 4458 patients with monoclonal gammopathy of undetermined significance in Sweden.* Blood, 2009. **114**(4): p. 791-5.
80. Kristinsson, S.Y., et al., *Risk of lymphoproliferative disorders among first-degree relatives of lymphoplasmacytic lymphoma/Waldenström macroglobulinemia patients: a population-based study in Sweden.* Blood, 2008. **112**(8): p. 3052-3056.
81. Schinasi, L.H., et al., *Multiple myeloma and family history of lymphohaematopoietic cancers: Results from the International Multiple Myeloma Consortium.* British Journal of Haematology, 2016. **175**(1): p. 87-101.

82. Sud, A., B. Kinnersley, and R.S. Houlston, *Genome-wide association studies of cancer: current insights and future perspectives.* Nat Rev Cancer, 2017. **17**(11): p. 692-704.
83. Miki, Y., et al., *A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1.* Science (New York, N.Y.), 1994. **266**(5182): p. 66-71.
84. Wooster, R., et al., *Identification of the breast cancer susceptibility gene BRCA2.* Nature, 1995. **378**(6559): p. 789-792.
85. Kinzler, K.W., et al., *Identification of FAP locus genes from chromosome 5q21.* Science, 1991. **253**(5020): p. 661.
86. Lindblom, A., et al., *Genetic mapping of a second locus predisposing to hereditary non–polyposis colon cancer.* Nature Genetics, 1993. **5**(3): p. 279-282.
87. Fishel, R., et al., *The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer.* Cell, 1993. **75**(5): p. 1027-1038.
88. Moslein, G., et al., *Microsatellite instability and mutation analysis of hMSH2 and hMLH1 in patients with sporadic, familial and hereditary colorectal cancer.* Human molecular genetics, 1996. **5**(9): p. 1245-1252.
89. Waller, R.G., et al., *Novel pedigree analysis implicates DNA repair and chromatin remodeling in multiple myeloma risk.* PLoS genetics, 2018. **14**(2): p. e1007111-e1007111.
90. Dilworth, D., et al., *Germline CDKN2A mutation implicated in predisposition to multiple myeloma.* Blood, 2000. **95**(5): p. 1869-1871.
91. Ballinger, M.L., et al., *Monogenic and polygenic determinants of sarcoma risk: an international genetic study.* The Lancet. Oncology, 2016. **17**(9): p. 1261-1271.
92. Reich, D.E. and E.S. Lander, *On the allelic spectrum of human disease.* Trends Genet, 2001. **17**(9): p. 502-10.
93. Risch, N. and K. Merikangas, *The future of genetic studies of complex human diseases.* Science, 1996. **273**(5281): p. 1516-7.
94. Hayden, P.J., et al., *Variation in DNA repair genes XRCC3, XRCC4, XRCC5 and susceptibility to myeloma.* Human molecular genetics, 2007. **16**(24): p. 3117-3127.
95. Pratt, G., et al., *A polymorphism in the 3' UTR of IRF4 linked to susceptibility and pathogenesis in chronic lymphocytic leukaemia and Hodgkin lymphoma has limited impact in multiple myeloma.* British journal of haematology, 2010. **150**(3): p. 371-373.
96. Roddam, P.L., et al., *Genetic variants of NHEJ DNA ligase IV can affect the risk of developing multiple myeloma, a tumour characterised by aberrant class switch recombination.* Journal of medical genetics, 2002. **39**(12): p. 900-905.
97. Davies, F.E., et al., *High-producer haplotypes of tumor necrosis factor alpha and lymphotoxin alpha are associated with an increased risk of myeloma and have an improved progression-free survival after treatment.* Journal of clinical oncology : official journal of the American Society of Clinical Oncology, 2000. **18**(15): p. 2843-2851.
98. Ma, L.-M., L.-H. Ruan, and H.-P. Yang, *Meta-analysis of the association of MTHFR polymorphisms with multiple myeloma risk.* Scientific reports, 2015. **5**: p. 10735-10735.
99. Hosgood, H.D., 3rd, et al., *Genetic variation in cell cycle and apoptosis related genes and multiple myeloma risk.* Leukemia research, 2009. **33**(12): p. 1609-1614.
100. Patnala, R., J. Clements, and J. Batra, *Candidate gene association studies: a comprehensive guide to useful in silico tools.* BMC genetics, 2013. **14**: p. 39-39.
101. Tabor, H.K., N.J. Risch, and R.M. Myers, *Candidate-gene approaches for studying complex genetic traits: practical considerations.* Nature reviews. Genetics, 2002. **3**(5): p. 391-397.
102. Whiffin, N. and R.S. Houlston, *Architecture of inherited susceptibility to colorectal cancer: a voyage of discovery.* Genes (Basel), 2014. **5**(2): p. 270-84.

103. Botstein, D. and N. Risch, *Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease.* Nature genetics, 2003. **33 Suppl**: p. 228-237.
104. Pritchard, J.K. and M. Przeworski, *Linkage disequilibrium in humans: models and data.* American journal of human genetics, 2001. **69**(1): p. 1-14.
105. Hirschhorn, J.N. and M.J. Daly, *Genome-wide association studies for common diseases and complex traits.* Nature reviews. Genetics, 2005. **6**(2): p. 95-108.
106. Sham, P.C. and S.M. Purcell, *Statistical power and significance testing in large-scale genetic studies.* Nature reviews. Genetics, 2014. **15**(5): p. 335-346.
107. Brion, M.J., K. Shakhbazov, and P.M. Visscher, *Calculating statistical power in Mendelian randomization studies.* Int J Epidemiol, 2013. **42**(5): p. 1497-501.
108. Houlston, R.S. and J. Peto, *The future of association studies of common cancers.* Human Genetics, 2003. **112**(4): p. 434-435.
109. Genomes Project, C., et al., *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. **491**(7422): p. 56-65.
110. Huang, J., et al., *Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel.* Nat Commun, 2015. **6**: p. 8111.
111. Marchini, J. and B. Howie, *Genotype imputation for genome-wide association studies.* Nature reviews. Genetics, 2010. **11**(7): p. 499-511.
112. Broderick, P., et al., *Common variation at 3p22.1 and 7p15.3 influences multiple myeloma risk.* Nat Genet, 2011. **44**(1): p. 58-61.
113. Chubb, D., et al., *Common variation at 3q26.2, 6p21.33, 17p11.2 and 22q13.1 influences multiple myeloma risk.* Nat Genet, 2013. **45**(10): p. 1221-1225.
114. Swaminathan, B., et al., *Variants in ELL2 influencing immunoglobulin levels associate with multiple myeloma.* Nat Commun, 2015. **6**: p. 7213.
115. Weinhold, N., et al., *The CCND1 c.870G>A polymorphism is a risk factor for t(11;14)(q13;q32) multiple myeloma.* Nat Genet, 2013. **45**(5): p. 522-525.
116. Mitchell, J.S., et al., *Genome-wide association study identifies multiple susceptibility loci for multiple myeloma.* Nat Commun, 2016. **7**: p. 12050.
117. Guglielmelli, T., et al., *mTOR pathway activation in multiple myeloma cell lines and primary tumour cells: pomalidomide enhances cytoplasmic-nuclear shuttling of mTOR protein.* Oncoscience, 2015. **2**(4): p. 382-394.
118. Jung, C.H., et al., *mTOR regulation of autophagy.* FEBS letters, 2010. **584**(7): p. 1287-1295.
119. Gilbert, S.L., et al., *Trak1 mutation disrupts GABA(A) receptor homeostasis in hypertonic mice.* Nature genetics, 2006. **38**(2): p. 245-250.
120. Maertens, G.N., P. Cherepanov, and A. Engelman, *Transcriptional co-activator p75 binds and tethers the Myc-interacting protein JPO2 to chromatin.* Journal of cell science, 2006. **119**(Pt 12): p. 2563-2571.
121. Ou, X.-M., K. Chen, and J.C. Shih, *Monoamine oxidase A and repressor R1 are involved in apoptotic signaling pathway.* Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(29): p. 10923-10928.
122. Zhou, L., et al., *Silencing of thrombospondin-1 is critical for myc-induced metastatic phenotypes in medulloblastoma.* Cancer research, 2010. **70**(20): p. 8199-8210.
123. Weinhold, N., et al., *The 7p15.3 (rs4487645) association for multiple myeloma shows strong allele-specific regulation of the MYC-interacting gene CDCA7L in malignant plasma cells.* Haematologica, 2015. **100**(3): p. e110.
124. Li, N., et al., *Multiple myeloma risk variant at 7p15.3 creates an IRF4-binding site and interferes with CDCA7L expression.* Nat Commun, 2016. **7**: p. 13656.

125. Heuck, C.J., et al., *Myeloma is characterized by stage-specific alterations in DNA methylation that occur early during myelomagenesis.* Journal of immunology (Baltimore, Md. : 1950), 2013. **190**(6): p. 2966-2975.
126. Campa, D., et al., *Risk of multiple myeloma is associated with polymorphisms within telomerase genes and telomere length.* International journal of cancer, 2015. **136**(5): p. E351-E358.
127. Speedy, H.E., et al., *A genome-wide association study identifies multiple susceptibility loci for chronic lymphocytic leukemia.* Nat Genet, 2014. **46**(1): p. 56-60.
128. Andersson, U., et al., *The association between longer relative leukocyte telomere length and risk of glioma is independent of the potentially confounding factors allergy, BMI, and smoking.* Cancer causes & control : CCC, 2019. **30**(2): p. 177-185.
129. Walsh, K.M., et al., *Telomere maintenance and the etiology of adult glioma.* Neuro-oncology, 2015. **17**(11): p. 1445-1452.
130. Walsh, K.M., et al., *Variants near TERT and TERC influencing telomere length are associated with high-grade glioma risk.* Nature genetics, 2014. **46**(7): p. 731-735.
131. Houlston, R.S., et al., *Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33.* Nat Genet, 2010. **42**(11): p. 973-7.
132. Gudmundsson, J., et al., *A genome-wide association study yields five novel thyroid cancer risk loci.* 2017. **8**: p. 14517.
133. Wu, G. and H.R. Schöler, *Role of Oct4 in the early embryo development.* Cell regeneration (London, England), 2014. **3**(1): p. 7-7.
134. Skibola, C.F., et al., *Genetic variants at 6p21.33 are associated with susceptibility to follicular lymphoma.* Nature genetics, 2009. **41**(8): p. 873-875.
135. Enciso-Mora, V., et al., *A genome-wide association study of Hodgkin's lymphoma identifies new susceptibility loci at 2p16.1 (REL), 8q24.21 and 10p14 (GATA3).* Nature genetics, 2010. **42**(12): p. 1126-1130.
136. Gross, J.A., et al., *TACI-Ig neutralizes molecules critical for B cell development and autoimmune disease. impaired B cell maturation in mice lacking BLyS.* Immunity, 2001. **15**(2): p. 289-302.
137. Gil, J., D. Bernard, and G. Peters, *Role of polycomb group proteins in stem cell self-renewal and cancer.* DNA and cell biology, 2005. **24**(2): p. 117-125.
138. Aguilo, F., M.-M. Zhou, and M.J. Walsh, *Long noncoding RNA, polycomb, and the ghosts haunting INK4b-ARF-INK4a expression.* Cancer research, 2011. **71**(16): p. 5365-5369.
139. Scott, C.L., et al., *Role of the chromobox protein CBX7 in lymphomagenesis.* Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(13): p. 5389-5394.
140. Martincic, K., et al., *Transcription elongation factor ELL2 directs immunoglobulin secretion in plasma cells by stimulating altered RNA processing.* Nature immunology, 2009. **10**(10): p. 1102-1109.
141. Milcarek, C., et al., *The eleven-nineteen lysine-rich leukemia gene (ELL2) influences the histone H3 protein modifications accompanying the shift to secretory immunoglobulin heavy chain mRNA production.* The Journal of biological chemistry, 2011. **286**(39): p. 33795-33803.
142. Li, N., et al., *Genetic Predisposition to Multiple Myeloma at 5q15 Is Mediated by an ELL2 Enhancer Polymorphism.* Cell Rep, 2017. **20**(11): p. 2556-2564.
143. Ali, M., et al., *The multiple myeloma risk allele at 5q15 lowers ELL2 expression and increases ribosomal gene expression.* Nature Communications, 2018. **9**(1): p. 1649.
144. Keskitalo, S., et al., *Dominant TOM1 mutation associated with combined immunodeficiency and autoimmune disease.* npj Genomic Medicine, 2019. **4**(1): p. 14.

145. Conway, K.L., et al., *ATG5 regulates plasma cell differentiation.* Autophagy, 2013. **9**(4): p. 528-537.
146. Cenci, S., *Autophagy, a new determinant of plasma cell differentiation and antibody responses.* Molecular immunology, 2014. **62**(2): p. 289-295.
147. Buckland, J., *BLIMP1, BCL6 and B-cell fate.* Nature Reviews Immunology, 2002. **2**(9): p. 629-629.
148. Kinkel, S.A., et al., *Jarid2 regulates hematopoietic stem cell function by acting with polycomb repressive complex 2.* Blood, 2015. **125**(12): p. 1890-1900.
149. Cerhan, J.R., et al., *Genome-wide association study identifies multiple susceptibility loci for diffuse large B cell lymphoma.* Nature genetics, 2014. **46**(11): p. 1233-1238.
150. Crowther-Swanepoel, D., et al., *Common variants at 2q37.3, 8q24.21, 15q21.3 and 16q24.1 influence chronic lymphocytic leukemia risk.* Nat Genet, 2010. **42**(2): p. 132-6.
151. Timofeeva, M.N., et al., *Influence of common genetic variation on lung cancer risk: meta-analysis of 14 900 cases and 29 485 controls.* Human molecular genetics, 2012. **21**(22): p. 4980-4995.
152. Dahlin, A.M., et al., *Genetic Variants in the 9p21.3 Locus Associated with Glioma Risk in Children, Adolescents, and Young Adults: A Case-Control Study.* Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology, 2019. **28**(7): p. 1252-1258.
153. Vijayakrishnan, J., et al., *Identification of four novel associations for B-cell acute lymphoblastic leukaemia risk.* Nature Communications, 2019. **10**(1): p. 5348.
154. Marjon, K., et al., *MTAP Deletions in Cancer Create Vulnerability to Targeting of the MAT2A/PRMT5/RIOK1 Axis.* Cell Rep, 2016. **15**(3): p. 574-587.
155. Joachim, J., et al., *Coiling up with SCOC and WAC: two new regulators of starvation-induced autophagy.* Autophagy, 2012. **8**(9): p. 1397-400.
156. Litchfield, K., et al., *Identification of 19 new risk loci and potential regulatory mechanisms influencing susceptibility to testicular germ cell tumor.* Nature genetics, 2017. **49**(7): p. 1133-1140.
157. Liu, H.-J., et al., *PtdIns(3,4,5)P3-dependent Rac Exchanger 1 (PREX1) Rac-Guanine Nucleotide Exchange Factor (GEF) Activity Promotes Breast Cancer Cell Proliferation and Tumor Growth via Activation of Extracellular Signal-regulated Kinase 1/2 (ERK1/2) Signaling.* The Journal of biological chemistry, 2016. **291**(33): p. 17258-17270.
158. Mitchell, J.S., et al., *Implementation of genome-wide complex trait analysis to quantify the heritability in multiple myeloma.* Sci Rep, 2015. **5**: p. 12473.
159. Li, M., C. Li, and W. Guan, *Evaluation of coverage variation of SNP chips for genome-wide association studies.* European journal of human genetics : EJHG, 2008. **16**(5): p. 635-643.
160. Kent, J.W., Jr., *Rare variants, common markers: synthetic association and beyond.* Genetic epidemiology, 2011. **35 Suppl 1**(Suppl 1): p. S80-S84.
161. Howie, B., J. Marchini, and M. Stephens, *Genotype imputation with thousands of genomes.* G3 (Bethesda, Md.), 2011. **1**(6): p. 457-470.
162. Amos, C.I., et al., *The OncoArray Consortium: A Network for Understanding the Genetic Architecture of Common Cancers.* Cancer Epidemiol Biomarkers Prev, 2017. **26**(1): p. 126-135.
163. Babron, M.-C., et al., *Rare and low frequency variant stratification in the UK population: description and impact on association tests.* PloS one, 2012. **7**(10): p. e46519-e46519.
164. Scales, M., et al., *Search for rare protein altering variants influencing susceptibility to multiple myeloma.* Oncotarget; Vol 8, No 22, 2017.
165. Lee, S., et al., *Rare-variant association analysis: study designs and statistical tests.* American journal of human genetics, 2014. **95**(1): p. 5-23.

166. Bolli, N., et al., *Next-generation sequencing of a family with a high penetrance of monoclonal gammopathies for the identification of candidate risk alleles.* Cancer, 2017. **123**(19): p. 3701-3708.
167. Freedman, M.L., et al., *Principles for the post-GWAS functional characterization of cancer risk loci.* Nat Genet, 2011. **43**(6): p. 513-8.
168. Wang, Y., et al., *Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer.* Nature Genetics, 2014. **46**(7): p. 736-741.
169. Wang, Y., et al., *A novel TP53 variant (rs78378222 A > C) in the polyadenylation signal is associated with increased cancer susceptibility: evidence from a meta-analysis.* Oncotarget, 2016. **7**(22): p. 32854-32865.
170. Enciso-Mora, V., et al., *Low penetrance susceptibility to glioma is caused by the TP53 variant rs78378222.* British Journal of Cancer, 2013. **108**(10): p. 2178-2185.
171. Bojesen, S.E., et al., *Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer.* Nature genetics, 2013. **45**(4): p. 371-384e3842.
172. He, G., et al., *TERT rs10069690 polymorphism and cancers risk: A meta-analysis.* Molecular genetics & genomic medicine, 2019. **7**(10): p. e00903-e00903.
173. Risca, V.I. and W.J. Greenleaf, *Unraveling the 3D genome: genomics tools for multiscale exploration.* Trends Genet, 2015. **31**(7): p. 357-72.
174. Ng, P.C. and S. Henikoff, *SIFT: Predicting amino acid changes that affect protein function.* Nucleic acids research, 2003. **31**(13): p. 3812-3814.
175. Adzhubei, I., D.M. Jordan, and S.R. Sunyaev, *Predicting functional effect of human missense mutations using PolyPhen-2.* Current protocols in human genetics, 2013. **Chapter 7**: p. Unit7.20-Unit7.20.
176. Hoffman, M.M., et al., *Integrative annotation of chromatin elements from ENCODE data.* Nucleic Acids Res, 2013. **41**(2): p. 827-41.
177. Bernstein, B.E., et al., *The NIH Roadmap Epigenomics Mapping Consortium.* Nature biotechnology, 2010. **28**(10): p. 1045-1048.
178. Satterlee, J.S., et al., *Community resources and technologies developed through the NIH Roadmap Epigenomics Program.* Methods in molecular biology (Clifton, N.J.), 2015. **1238**: p. 27-49.
179. Fernandez, J.M., et al., *The BLUEPRINT Data Analysis Portal.* Cell Syst, 2016. **3**(5): p. 491-495.e5.
180. Ernst, J. and M. Kellis, *ChromHMM: automating chromatin-state discovery and characterization.* Nat Methods, 2012. **9**(3): p. 215-6.
181. Zhu, Z., et al., *Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets.* Nat Genet, 2016. **48**(5): p. 481-7.
182. Easton, D.F., et al., *Gene-Panel Sequencing and the Prediction of Breast-Cancer Risk.* New England Journal of Medicine, 2015. **372**(23): p. 2243-2257.
183. Halvarsson, B.M., et al., *Direct evidence for a polygenic etiology in familial multiple myeloma.* Blood Advances, 2017. **1**(10): p. 619-623.
184. Khera, A.V., et al., *Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations.* Nature Genetics, 2018. **50**(9): p. 1219-1224.
185. Pashayan, N., et al., *Polygenic susceptibility to prostate and breast cancer: implications for personalised screening.* British journal of cancer, 2011. **104**(10): p. 1656-1663.
186. Frampton, M.J., et al., *Implications of polygenic risk for personalised colorectal cancer screening.* Ann Oncol, 2016. **27**(3): p. 429-34.
187. Aminkeng, F., et al., *A coding variant in RARG confers susceptibility to anthracycline-induced cardiotoxicity in childhood cancer.* Nature genetics, 2015. **47**(9): p. 1079-1084.

188. Fachal, L., et al., *A three-stage genome-wide association study identifies a susceptibility locus for late radiotherapy toxicity at 2q24.1.* Nature Genetics, 2014. **46**(8): p. 891-894.
189. Johnson, D.C., et al., *Genome-wide association study identifies variation at 6q25.1 associated with survival in multiple myeloma.* Nat Commun, 2016. **7**: p. 10290.
190. Lawlor, D.A., et al., *Mendelian randomization: using genes as instruments for making causal inferences in epidemiology.* Statistics in medicine, 2008. **27**(8): p. 1133-1163.
191. Davey Smith, G. and G. Hemani, *Mendelian randomization: genetic anchors for causal inference in epidemiological studies.* Human molecular genetics, 2014. **23**(R1): p. R89-R98.
192. Jarvis, D., et al., *Mendelian randomisation analysis strongly implicates adiposity with risk of developing colorectal cancer.* British journal of cancer, 2016. **115**(2): p. 266-272.
193. Rodriguez-Broadbent, H., et al., *Mendelian randomisation implicates hyperlipidaemia as a risk factor for colorectal cancer.* International journal of cancer, 2017. **140**(12): p. 2701-2708.
194. Interleukin-6 Receptor Mendelian Randomisation Analysis, C., et al., *The interleukin-6 receptor as a target for prevention of coronary heart disease: a mendelian randomisation analysis.* Lancet (London, England), 2012. **379**(9822): p. 1214-1224.
195. Organisation, W.H., *International Statistical Classification of Diseases and Related Health Problems 10th Revision*. 2010 ed. Vol. 2. 2011.
196. Jackson, G.H., et al., *Lenalidomide maintenance versus observation for patients with newly diagnosed multiple myeloma (Myeloma XI): a multicentre, open-label, randomised, phase 3 trial.* The Lancet. Oncology, 2019. **20**(1): p. 57-73.
197. Schumacher, F.R., et al., *Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci.* Nature Genetics, 2018. **50**(7): p. 928-936.
198. Michailidou, K., et al., *Association analysis identifies 65 new breast cancer risk loci.* Nature, 2017. **551**(7678): p. 92-94.
199. Goldschmidt, H., et al., *Joint HOVON-50/GMMG-HD3 randomized trial on the effect of thalidomide as part of a high-dose therapy regimen and as maintenance treatment for newly diagnosed myeloma patients.* Ann Hematol, 2003. **82**(10): p. 654-9.
200. Scheid, C., et al., *Bortezomib before and after autologous stem cell transplantation overcomes the negative prognostic impact of renal impairment in newly diagnosed multiple myeloma: a subgroup analysis from the HOVON-65/GMMG-HD4 trial.* Haematologica, 2014. **99**(1): p. 148-154.
201. Sonneveld, P., et al., *Bortezomib induction and maintenance treatment in patients with newly diagnosed multiple myeloma: results of the randomized phase III HOVON-65/ GMMG-HD4 trial.* J Clin Oncol, 2012. **30**(24): p. 2946-55.
202. van Wijngaarden, J.P., et al., *Rationale and design of the B-PROOF study, a randomized controlled trial on the effect of supplemental intake of vitamin B12 and folic acid on fracture incidence.* BMC geriatrics, 2011. **11**: p. 80-80.
203. Merz, M., et al., *Subcutaneous versus intravenous bortezomib in two different induction therapies for newly diagnosed multiple myeloma: an interim analysis from the prospective GMMG-MM5 trial.* Haematologica, 2015. **100**(7): p. 964-9.
204. Sonneveld, P., et al., *Bortezomib induction and maintenance treatment in patients with newly diagnosed multiple myeloma: results of the randomized phase III HOVON-65/ GMMG-HD4 trial.* Journal of clinical oncology : official journal of the American Society of Clinical Oncology, 2012. **30**(24): p. 2946-2955.
205. Erbel, R., et al., *[The Heinz Nixdorf Recall study].* Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz, 2012. **55**(6-7): p. 809-15.
206. Sigurdardottir, L.G., et al., *Data quality at the Icelandic Cancer Registry: comparability, validity, timeliness and completeness.* Acta oncologica (Stockholm, Sweden), 2012. **51**(7): p. 880-889.

207. Gulcher, J. and K. Stefansson, *Population genomics: laying the groundwork for genetic disease modeling and targeting.* Clinical chemistry and laboratory medicine, 1998. **36**(8): p. 523-527.
208. Biobank, N.M. [cited 2018 August 2018]; Available from: https://bbmri.no/norwegian-research-biobank-multiple-myeloma/introduction.
209. Ripke, S., et al., *Genome-wide association analysis identifies 13 new risk loci for schizophrenia.* Nature genetics, 2013. **45**(10): p. 1150-1159.
210. Magnusson, P.K.E., et al., *The Swedish Twin Registry: establishment of a biobank and other recent developments.* Twin research and human genetics : the official journal of the International Society for Twin Studies, 2013. **16**(1): p. 317-329.
211. Ripke, S., et al., *Genome-wide association analysis identifies 13 new risk loci for schizophrenia.* Nat Genet, 2013. **45**(10): p. 1150-9.
212. Morgan, G.J., et al., *Long-term follow-up of MRC Myeloma IX trial: Survival outcomes with bisphosphonate and thalidomide treatment.* Clinical cancer research : an official journal of the American Association for Cancer Research, 2013. **19**(21): p. 6030-6038.
213. Morgan, G.J., et al., *First-line treatment with zoledronic acid as compared with clodronic acid in multiple myeloma (MRC Myeloma IX): a randomised controlled trial.* Lancet (London, England), 2010. **376**(9757): p. 1989-1999.
214. Morgan, G.J., et al., *Cyclophosphamide, thalidomide, and dexamethasone as induction therapy for newly diagnosed multiple myeloma patients destined for autologous stem-cell transplantation: MRC Myeloma IX randomized trial results.* Haematologica, 2012. **97**(3): p. 442-450.
215. Morgan, G.J., et al., *First-line treatment with zoledronic acid as compared with clodronic acid in multiple myeloma (MRC Myeloma IX): a randomised controlled trial.* Lancet, 2010. **376**(9757): p. 744-51.
216. Power, C. and J. Elliott, *Cohort profile: 1958 British birth cohort (National Child Development Study).* International journal of epidemiology, 2006. **35**(1): p. 34-41.
217. Wellcome Trust Case Control, C., *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.* Nature, 2007. **447**(7145): p. 661-678.
218. *Identifying Cancer Genetic Markers of Susceptibility Using HighThroughput SNP Arrays.* Cancer Biology & Therapy, 2007. **6**(5): p. 638-639.
219. Law, P.J., et al., *Genome-wide association analysis implicates dysregulation of immunity genes in chronic lymphocytic leukaemia.* 2017. **8**: p. 14175.
220. Berndt, S.I., et al., *Genome-wide association study identifies multiple risk loci for chronic lymphocytic leukemia.* (1546-1718 (Electronic)).
221. Hallek, M., et al., *Guidelines for the diagnosis and treatment of chronic lymphocytic leukemia: a report from the International Workshop on Chronic Lymphocytic Leukemia updating the National Cancer Institute-Working Group 1996 guidelines.* Blood, 2008. **111**(12): p. 5446-56.
222. Zhan, F., et al., *The molecular classification of multiple myeloma.* Blood, 2006. **108**(6): p. 2020-8.
223. Hanamura, I., et al., *Prognostic value of cyclin D2 mRNA expression in newly diagnosed multiple myeloma treated with high-dose chemotherapy and tandem autologous stem cell transplantations.* Leukemia, 2006. **20**(7): p. 1288-90.
224. Zhan, F., et al., *Gene-expression signature of benign monoclonal gammopathy evident in multiple myeloma is linked to good prognosis.* Blood, 2007. **109**(4): p. 1692-700.
225. Chen, L., et al., *Identification of early growth response protein 1 (EGR-1) as a novel target for JUN-induced apoptosis in multiple myeloma.* Blood, 2010. **115**(1): p. 61-70.
226. Broyl, A., et al., *Gene expression profiling for molecular classification of multiple myeloma in newly diagnosed patients.* Blood, 2010. **116**(14): p. 2543-53.

227. Mulligan, G., et al., *Gene expression profiling and correlation with outcome in clinical trials of the proteasome inhibitor bortezomib.* Blood, 2007. **109**(8): p. 3177-88.
228. Moreaux, J., et al., *Development of gene expression-based score to predict sensitivity of multiple myeloma cells to DNA methylation inhibitors.* Mol Cancer Ther, 2012. **11**(12): p. 2685-92.
229. Weinhold, N., et al., *The 7p15.3 (rs4487645) association for multiple myeloma shows strong allele-specific regulation of the MYC-interacting gene CDCA7L in malignant plasma cells.* Haematologica, 2015. **100**(3): p. e110-3.
230. Miller, S.A., D.D. Dykes, and H.F. Polesky, *A simple salting out procedure for extracting DNA from human nucleated cells.* Nucleic acids research, 1988. **16**(3): p. 1215-1215.
231. Illumina. [cited 2020 April 2020]; Available from: https://www.tst-web.illumina.com/content/illumina-marketing/amr/en/technology/beadarray-technology/infinium-hd-assay.html.
232. Kent, W.J., et al., *The human genome browser at UCSC.* Genome Res, 2002. **12**(6): p. 996-1006.
233. Untergasser, A., et al., *Primer3--new capabilities and interfaces.* Nucleic Acids Res, 2012. **40**(15): p. e115.
234. Technologies, I.D. [cited 2016 November 2016]; Available from: www.idtdna.com/calc/analyzer.
235. Prober, J.M., et al., *A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides.* Science, 1987. **238**(4825): p. 336-41.
236. Rao, S.S., et al., *A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.* Cell, 2014. **159**(7): p. 1665-80.
237. Wingett, S., et al., *HiCUP: pipeline for mapping and processing Hi-C data.* F1000Res, 2015. **4**: p. 1310.
238. Cairns, J., et al., *CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data.* Genome Biol, 2016. **17**(1): p. 127.
239. Zhou, X., et al., *Exploring long-range genome interactions using the WashU Epigenome Browser.* Nature methods, 2013. **10**(5): p. 375-376.
240. Mifsud, B., et al., *Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C.* Nature Genetics, 2015. **47**: p. 598.
241. Orlando, G., B. Kinnersley, and R.S. Houlston, *Capture Hi-C Library Generation and Analysis to Detect Chromatin Interactions.* Curr Protoc Hum Genet, 2018: p. e63.
242. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* American journal of human genetics, 2007. **81**(3): p. 559-575.
243. Team, R.C., *R: A language and environment for statistical computing.* 2018.
244. Clayton, D.G., et al., *Population structure, differential bias and genomic control in a large-scale, case-control association study.* Nat Genet, 2005. **37**(11): p. 1243-6.
245. de Bakker, P.I.W., et al., *Practical aspects of imputation-driven meta-analysis of genome-wide association studies.* Human molecular genetics, 2008. **17**(R2): p. R122-R128.
246. Patterson, N., A.L. Price, and D. Reich, *Population structure and eigenanalysis.* PLoS Genet, 2006. **2**(12): p. e190.
247. Patterson, N., A.L. Price, and D. Reich, *Population structure and eigenanalysis.* PLoS genetics, 2006. **2**(12): p. e190-e190.
248. Price, A.L., et al., *Principal components analysis corrects for stratification in genome-wide association studies.* Nature genetics, 2006. **38**(8): p. 904-909.
249. Emigh, T.H., *A comparison of tests for Hardy-Weinberg equilibrium.* Biometrics, 1980. **36**(4): p. 627-642.
250. Dudbridge, F. and A. Gusnanto, *Estimation of significance thresholds for genomewide association scans.* Genetic epidemiology, 2008. **32**(3): p. 227-234.

251. Pe'er, I., et al., *Estimation of the multiple testing burden for genomewide association studies of nearly all common variants.* Genetic epidemiology, 2008. **32**(4): p. 381-385.
252. Hoggart, C.J., et al., *Genome-wide significance for dense SNP and resequencing data.* Genetic epidemiology, 2008. **32**(2): p. 179-185.
253. Wakefield, J., *A Bayesian measure of the probability of false discovery in genetic epidemiology studies.* Am J Hum Genet, 2007. **81**(2): p. 208-27.
254. Devlin, B. and N. Risch, *A comparison of linkage disequilibrium measures for fine-scale mapping.* Genomics, 1995. **29**(2): p. 311-22.
255. Jorde, L.B., *Linkage disequilibrium and the search for complex disease genes.* Genome Res, 2000. **10**(10): p. 1435-44.
256. Barrett, J.C., et al., *Haploview: analysis and visualization of LD and haplotype maps.* Bioinformatics (Oxford, England), 2005. **21**(2): p. 263-265.
257. Li, S., et al., *Snap: an integrated SNP annotation platform.* Nucleic acids research, 2007. **35**(Database issue): p. D707-D710.
258. Thorisson, G.A., et al., *The International HapMap Project Web site.* Genome Res, 2005. **15**(11): p. 1592-3.
259. International HapMap, C., *The International HapMap Project.* Nature, 2003. **426**(6968): p. 789-796.
260. Danecek, P., et al., *The variant call format and VCFtools.* Bioinformatics (Oxford, England), 2011. **27**(15): p. 2156-2158.
261. Marchini, J. and B. Howie, *Genotype imputation for genome-wide association studies.* Nat Rev Genet, 2010. **11**(7): p. 499-511.
262. Abecasis, G.R., et al., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.
263. Delaneau, O., J. Marchini, and J.-F. Zagury, *A linear complexity phasing method for thousands of genomes.* Nature methods, 2011. **9**(2): p. 179-181.
264. Bycroft, C., et al., *Genome-wide genetic data on ~500,000 UK Biobank participants.* bioRxiv, 2017: p. 166298.
265. Marchini, J., et al., *A new multipoint method for genome-wide association studies by imputation of genotypes.* Nat Genet, 2007. **39**(7): p. 906-13.
266. Liu, J.Z., et al., *Meta-analysis and imputation refines the association of 15q25 with smoking quantity.* Nat Genet, 2010. **42**(5): p. 436-40.
267. Higgins, J.P.T., et al., *Measuring inconsistency in meta-analyses.* BMJ (Clinical research ed.), 2003. **327**(7414): p. 557-560.
268. Higgins, J.P. and S.G. Thompson, *Quantifying heterogeneity in a meta-analysis.* Stat Med, 2002. **21**(11): p. 1539-58.
269. Yang, J., et al., *Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits.* Nature genetics, 2012. **44**(4): p. 369-S3.
270. Neben, K., et al., *Combining information regarding chromosomal aberrations t(4;14) and del(17p13) with the International Staging System classification allows stratification of myeloma patients undergoing autologous stem cell transplantation.* Haematologica, 2010. **95**(7): p. 1150-7.
271. Chiecchio, L., et al., *Deletion of chromosome 13 detected by conventional cytogenetics is a critical prognostic factor in myeloma.* Leukemia, 2006. **20**(9): p. 1610-7.
272. Boyle, E.M., et al., *A molecular diagnostic approach able to detect the recurrent genetic prognostic factors typical of presenting myeloma.* Genes Chromosomes Cancer, 2015. **54**(2): p. 91-8.
273. Coordinators, N.R., *Database resources of the National Center for Biotechnology Information.* Nucleic acids research, 2018. **46**(D1): p. D8-D13.
274. de Souza, N., *The ENCODE project.* Nat Methods, 2012. **9**(11): p. 1046.

275. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
276. Yates, A.D., et al., *Ensembl 2020.* Nucleic Acids Research, 2019. **48**(D1): p. D682-D688.
277. Cunningham, F., et al., *Ensembl 2015.* Nucleic acids research, 2015. **43**(Database issue): p. D662-D669.
278. Flicek, P., et al., *Ensembl 2014.* Nucleic acids research, 2014. **42**(Database issue): p. D749-D755.
279. Stegle, O., et al., *Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses.* Nat Protoc, 2012. **7**(3): p. 500-7.
280. *The Genotype-Tissue Expression (GTEx) project.* Nat Genet, 2013. **45**(6): p. 580-5.
281. Dixon, J.R., et al., *Topological domains in mammalian genomes identified by analysis of chromatin interactions.* Nature, 2012. **485**(7398): p. 376-380.
282. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics (Oxford, England), 2009. **25**(14): p. 1754-1760.
283. Knight, P.A. and D. Ruiz, *A fast algorithm for matrix balancing.* IMA Journal of Numerical Analysis, 2012. **33**(3): p. 1029-1047.
284. Speed, D., et al., *Improved heritability estimation from genome-wide SNPs.* American journal of human genetics, 2012. **91**(6): p. 1011-1021.
285. Chatterjee, N., *Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits and implications for the future.* BioRxiv, 2017.
286. Gamazon, E.R.A.-O.h.o.o., et al., *A gene-based association method for mapping traits using reference transcriptome data.* (1546-1718 (Electronic)).
287. GTEx. *Dataset Summary of Analysis Samples*. [cited 2019 January 2019]; Available from: https://gtexportal.org/home/tissueSummaryPage.
288. Barbeira, A.N., et al., *Integrating Predicted Transcriptome From Multiple Tissues Improves Association Detection.* bioRxiv, 2018: p. 292649.
289. Cowper-Sal lari, R., et al., *Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression.* Nat Genet, 2012. **44**(11): p. 1191-8.
290. Bulik-Sullivan, B., et al., *An atlas of genetic correlations across human diseases and traits.* 2015. **47**(11): p. 1236-41.
291. Finucane, H.K., et al., *Partitioning heritability by functional annotation using genome-wide association summary statistics.* 2015. **47**(11): p. 1228-35.
292. Fiziev, P., et al., *Systematic Epigenomic Analysis Reveals Chromatin States Associated with Melanoma Progression.* Cell Rep, 2017. **19**(4): p. 875-889.
293. Schoenfelder, S., et al., *Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome.* Nat Genet, 2015. **47**(10): p. 1179-1186.
294. Trynka, G., et al., *Chromatin marks identify critical cell types for fine mapping complex trait variants.* Nat Genet, 2013. **45**(2): p. 124-30.
295. Millard, L.A.C., et al., *MR-PheWAS: hypothesis prioritization among potential causal effects of body mass index on many outcomes, using Mendelian randomization.* Scientific reports, 2015. **5**: p. 16645-16645.
296. Hemani, G., et al., *The MR-Base platform supports systematic causal inference across the human phenome.* eLife, 2018. **7**: p. e34408.
297. Disney-Hogg, L., et al., *Impact of atopy on risk of glioma: a Mendelian randomisation study.* BMC medicine, 2018. **16**(1): p. 42-42.
298. Bowden, J., et al., *Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator.* Genetic epidemiology, 2016. **40**(4): p. 304-314.

299. Hartwig, F.P., G. Davey Smith, and J. Bowden, *Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption.* International journal of epidemiology, 2017. **46**(6): p. 1985-1998.
300. Hemani, G., J. Bowden, and G. Davey Smith, *Evaluating the potential role of pleiotropy in Mendelian randomization studies.* Human molecular genetics, 2018. **27**(R2): p. R195-R208.
301. Fan, Q., et al., *HDL-cholesterol levels and risk of age-related macular degeneration: a multiethnic genetic study using Mendelian randomization.* International Journal of Epidemiology, 2017. **46**(6): p. 1891-1902.
302. Bowden, J., G. Davey Smith, and S. Burgess, *Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression.* Int J Epidemiol, 2015. **44**(2): p. 512-25.
303. Wootton, R.E., et al., *Evaluation of the causal effects between subjective wellbeing and cardiometabolic health: mendelian randomisation study.* BMJ, 2018. **362**: p. k3788.
304. [cited 2016 October 2016]; Available from: http://www.kbioscience.co.uk.
305. Went, M., et al., *Identification of multiple risk loci and regulatory mechanisms influencing susceptibility to multiple myeloma.* Nature Communications, 2018. **9**(1): p. 3707.
306. Anderson, C.A., et al., *Data quality control in genetic case-control association studies.* Nat Protoc, 2010. **5**(9): p. 1564-73.
307. Erickson, S.W., et al., *Genome-wide scan identifies variant in 2q12.3 associated with risk for multiple myeloma.* Blood, 2014. **124**(12): p. 2001-3.
308. Speed, D., et al., *Reevaluation of SNP heritability in complex human traits.* Nat Genet, 2017. **49**(7): p. 986-992.
309. Chatterjee, N., et al., *Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies.* Nat Genet, 2013. **45**(4): p. 400-5, 405e1-3.
310. Javierre, B.M., et al., *Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters.* Cell, 2016. **167**(5): p. 1369-1384.e19.
311. Comartin, D., et al., *CEP120 and SPICE1 cooperate with CPAP in centriole elongation.* Curr Biol, 2013. **23**(14): p. 1360-6.
312. Pinzaru, A.M., et al., *Telomere Replication Stress Induced by POT1 Inactivation Accelerates Tumorigenesis.* Cell Rep, 2016. **15**(10): p. 2170-84.
313. Rice, C., et al., *Structural and functional analysis of the human POT1-TPP1 telomeric complex.* Nat Commun, 2017. **8**: p. 14928.
314. Chang, S., *Cancer chromosomes going to POT1.* Nat Genet, 2013. **45**(5): p. 473-5.
315. Speedy, H.E., et al., *Germ line mutations in shelterin complex genes are associated with familial chronic lymphocytic leukemia.* Blood, 2016. **128**(19): p. 2319-2326.
316. McDaneld, T.G., K. Hannon, and D.E. Moody, *Ankyrin repeat and SOCS box protein 15 regulates protein synthesis in skeletal muscle.* Am J Physiol Regul Integr Comp Physiol, 2006. **290**(6): p. R1672-82.
317. McDaneld, T.G. and D.M. Spurlock, *Ankyrin repeat and suppressor of cytokine signaling (SOCS) box-containing protein (ASB) 15 alters differentiation of mouse C2C12 myoblasts and phosphorylation of mitogen-activated protein kinase and Akt.* J Anim Sci, 2008. **86**(11): p. 2897-902.
318. Kile, B.T., et al., *The SOCS box: a tale of destruction and degradation.* Trends Biochem Sci, 2002. **27**(5): p. 235-41.
319. Park, S.R., et al., *HoxC4 binds to the promoter of the cytidine deaminase AID gene to induce AID expression, class-switch DNA recombination and somatic hypermutation.* Nat Immunol, 2009. **10**(5): p. 540-50.

320. Steinke, J.W., et al., *Identification of an Sp factor-dependent promoter in GCET, a gene expressed at high levels in germinal center B cells.* Mol Immunol, 2004. **41**(12): p. 1145-53.
321. Holien, T., et al., *Addiction to c-MYC in multiple myeloma.* Blood, 2012. **120**(12): p. 2450-3.
322. Kuehl, W.M. and P.L. Bergsagel, *MYC addiction: a potential therapeutic target in MM.* Blood, 2012. **120**(12): p. 2351-2.
323. Ohguchi, H., et al., *The KDM3A-KLF2-IRF4 axis maintains myeloma cell survival.* 2016. **7**: p. 10258.
324. Yang, M., et al., *PRR14 is a novel activator of the PI3K pathway promoting lung carcinogenesis.* Oncogene, 2016. **35**(42): p. 5527-5538.
325. Ju, S., et al., *Correlation of expression levels of BLyS and its receptors with multiple myeloma.* Clin Biochem, 2009. **42**(4-5): p. 387-99.
326. Mackay, F. and P. Schneider, *TACI, an enigmatic BAFF/APRIL receptor, with new unappreciated biochemical and biological properties.* Cytokine Growth Factor Rev, 2008. **19**(3-4): p. 263-76.
327. Moreaux, J., et al., *The level of TACI gene expression in myeloma cells is associated with a signature of microenvironment dependence versus a plasmablastic signature.* Blood, 2005. **106**(3): p. 1021-30.
328. Moreaux, J., et al., *TACI expression is associated with a mature bone marrow plasma cell signature and C-MAF overexpression in human myeloma cell lines.* Haematologica, 2007. **92**(6): p. 803-11.
329. Moreaux, J., et al., *APRIL and TACI interact with syndecan-1 on the surface of multiple myeloma cells to form an essential survival loop.* Eur J Haematol, 2009. **83**(2): p. 119-29.
330. Novak, A.J., et al., *Expression of BCMA, TACI, and BAFF-R in multiple myeloma: a mechanism for growth and survival.* Blood, 2004. **103**(2): p. 689-94.
331. Landgren, O. and B.M. Weiss, *Patterns of monoclonal gammopathy of undetermined significance and multiple myeloma in various ethnic/racial groups: support for genetic factors in pathogenesis.* Leukemia, 2009. **23**(10): p. 1691-7.
332. Rand, K.A., et al., *A Meta-analysis of Multiple Myeloma Risk Regions in African and European Ancestry Populations Identifies Putatively Functional Loci.* Cancer Epidemiol Biomarkers Prev, 2016. **25**(12): p. 1609-1618.
333. Ip, H.F., et al., *Characterizing the Relation Between Expression QTLs and Complex Traits: Exploring the Role of Tissue Specificity.* (1573-3297 (Electronic)).
334. Landgren, O. and S.V. Rajkumar, *New Developments in Diagnosis, Prognosis, and Assessment of Response in Multiple Myeloma.* (1078-0432 (Print)).
335. Sud, A., et al., *Genome-wide association study implicates immune dysfunction in the development of Hodgkin lymphoma.* Blood, 2018. **132**(19): p. 2040.
336. Barbeira, A.N.A.-O.h.o.o., et al., *Integrating predicted transcriptome from multiple tissues improves association detection.* (1553-7404 (Electronic)).
337. Yang, J., et al., *GCTA: a tool for genome-wide complex trait analysis.* Am J Hum Genet, 2011. **88**(1): p. 76-82.
338. Wu, L., et al., *A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer.* (1546-1718 (Electronic)).
339. Moutsianas, L. and J. Gutierrez-Achury, *Genetic Association in the HLA Region.* Methods Mol Biol, 2018. **1793**: p. 111-134.
340. Beksac, M., et al., *HLA polymorphism and risk of multiple myeloma.* Leukemia, 2016. **30**: p. 2260.
341. Hussain, T. and R. Mulherkar, *Lymphoblastoid Cell lines: a Continuous in Vitro Source of Cells to Study Carcinogen Sensitivity and DNA Repair.* Int J Mol Cell Med, 2012. **1**(2): p. 75-87.

342. Bolli, N., et al., *Genomic patterns of progression in smoldering multiple myeloma.* 2018. **9**(1): p. 3363.
343. Maura, F., et al., *Biological and prognostic impact of APOBEC-induced mutations in the spectrum of plasma cell dyscrasias and multiple myeloma cell lines.* Leukemia, 2018. **32**(4): p. 1044-1048.
344. Walker, B.A., et al., *APOBEC family mutational signatures are associated with poor prognosis translocations in multiple myeloma.* 2015. **6**: p. 6997.
345. Guzik-Lendrum, S., I. Rayment, and S.P. Gilbert, *Homodimeric Kinesin-2 KIF3CC Promotes Microtubule Dynamics.* (1542-0086 (Electronic)).
346. Wang, C., et al., *Suppression of motor protein KIF3C expression inhibits tumor growth and metastasis in breast cancer by inhibiting TGF-beta signaling.* (1872-7980 (Electronic)).
347. Eskat, A., et al., *Step-wise assembly, maturation and dynamic behavior of the human CENP-P/O/R/Q/U kinetochore sub-complex.* (1932-6203 (Electronic)).
348. So, C.C., S. Ramachandran, and A. Martin, *E3 Ubiquitin Ligases RNF20 and RNF40 Are Required for Double-Stranded Break (DSB) Repair: Evidence for Monoubiquitination of Histone H2B Lysine 120 as a Novel Axis of DSB Signaling and Repair.* Mol Cell Biol, 2019. **39**(8).
349. Shiloh, Y., et al., *RNF20-RNF40: A ubiquitin-driven link between gene expression and the DNA damage response.* FEBS Lett, 2011. **585**(18): p. 2795-802.
350. Schneider, D., et al., *The E3 ubiquitin ligase RNF40 suppresses apoptosis in colorectal cancer cells.* Clin Epigenetics, 2019. **11**(1): p. 98.
351. Sahm, F., et al., *The endogenous tryptophan metabolite and NAD+ precursor quinolinic acid confers resistance of gliomas to oxidative stress.* Cancer Res, 2013. **73**(11): p. 3225-34.
352. Ullmark, T., et al., *Anti-apoptotic quinolinate phosphoribosyltransferase (QPRT) is a target gene of Wilms' tumor gene 1 (WT1) protein in leukemic cells.* Biochem Biophys Res Commun, 2017. **482**(4): p. 802-807.
353. Sen, S., et al., *Transcription factor 19 interacts with histone 3 lysine 4 trimethylation and controls gluconeogenesis via the nucleosome-remodeling-deacetylase complex.* J Biol Chem, 2017. **292**(50): p. 20362-20378.
354. Krautkramer, K.A., et al., *Tcf19 is a novel islet factor necessary for proliferation and survival in the INS-1 beta-cell line.* Am J Physiol Endocrinol Metab, 2013. **305**(5): p. E600-10.
355. Went, M. and A. Sud, *Identification of multiple risk loci and regulatory mechanisms influencing susceptibility to multiple myeloma.* 2018. **9**(1): p. 3707.
356. Wainberg, M., et al., *Opportunities and challenges for transcriptome-wide association studies.* Nature Genetics, 2019. **51**(4): p. 592-599.
357. Barlogie, B. and R.P. Gale, *Multiple myeloma and chronic lymphocytic leukemia: parallels and contrasts.* Am J Med, 1992. **93**(4): p. 443-50.
358. Shaffer, A.L., A. Rosenwald, and L.M. Staudt, *Lymphoid malignancies: the dark side of B-cell differentiation.* Nat Rev Immunol, 2002. **2**(12): p. 920-32.
359. Kuppers, R., *Mechanisms of B-cell lymphoma pathogenesis.* Nat Rev Cancer, 2005. **5**(4): p. 251-62.
360. Bulik-Sullivan, B.K., et al., *LD Score regression distinguishes confounding from polygenicity in genome-wide association studies.* 2015. **47**(3): p. 291-5.
361. Guo, D.C., et al., *Mutations in smooth muscle alpha-actin (ACTA2) lead to thoracic aortic aneurysms and dissections.* Nat Genet, 2007. **39**(12): p. 1488-93.
362. Figgett, W.A., et al., *The TACI receptor regulates T-cell-independent marginal zone B cell responses through innate activation-induced cell death.* Immunity, 2013. **39**(3): p. 573-83.

363. Akagi, T., T. Yoshino, and E. Kondo, *The Fas antigen and Fas-mediated apoptosis in B-cell differentiation.* Leuk Lymphoma, 1998. **28**(5-6): p. 483-9.
364. Zhang, S., et al., *IRF4 promotes cell proliferation by JNK pathway in multiple myeloma.* Med Oncol, 2013. **30**(2): p. 594.
365. Inano, S., et al., *RFWD3-Mediated Ubiquitination Promotes Timely Removal of Both RPA and RAD51 from DNA Damage Sites to Facilitate Homologous Recombination.* Mol Cell, 2017. **66**(5): p. 622-634.e8.
366. Elia, A.E., et al., *RFWD3-Dependent Ubiquitination of RPA Regulates Repair at Stalled Replication Forks.* Mol Cell, 2015. **60**(2): p. 280-93.
367. Codd, V., et al., *Identification of seven loci affecting mean telomere length and their association with disease.* Nature Genetics, 2013. **45**: p. 422.
368. Linxweiler, M., B. Schick, and R. Zimmermann, *Let's talk about Secs: Sec61, Sec62 and Sec63 in signal transduction, oncology and personalized medicine.* Signal Transduct Target Ther, 2017. **2**: p. 17002.
369. Bergmann, T.J., et al., *Role of SEC62 in ER maintenance: A link with ER stress tolerance in SEC62-overexpressing tumors?* Mol Cell Oncol, 2017. **4**(2): p. e1264351.
370. Jung, V., et al., *Genomic and expression analysis of the 3q25-q26 amplification unit reveals TLOC1/SEC62 as a probable target gene in prostate cancer.* Mol Cancer Res, 2006. **4**(3): p. 169-76.
371. Figueroa, J.D., et al., *Genome-wide association study identifies multiple loci associated with bladder cancer risk.* Hum Mol Genet, 2014. **23**(5): p. 1387-98.
372. Bijnsdorp, I.V., et al., *Thymidine phosphorylase in cancer cells stimulates human endothelial cell migration and invasion by the secretion of angiogenic factors.* Br J Cancer, 2011. **104**(7): p. 1185-92.
373. Deves, C., et al., *The kinetic mechanism of Human Thymidine Phosphorylase - a molecular target for cancer drug development.* Mol Biosyst, 2014. **10**(3): p. 592-604.
374. Liu, H., et al., *Thymidine phosphorylase exerts complex effects on bone resorption and formation in myeloma.* Sci Transl Med, 2016. **8**(353): p. 353ra113.
375. Liu, F., et al., *The oncoprotein HBXIP promotes glucose metabolism reprogramming via downregulating SCO2 and PDHA1 in breast cancer.* Oncotarget, 2015. **6**(29): p. 27199-213.
376. Won, K.Y., et al., *Regulatory role of p53 in cancer metabolism via SCO2 and TIGAR in human breast cancer.* Hum Pathol, 2012. **43**(2): p. 221-8.
377. Kim, S.H., et al., *Distinctive interrelation of p53 with SCO2, COX, and TIGAR in human gastric cancer.* Pathol Res Pract, 2016. **212**(10): p. 904-910.
378. Papadopoulou, L.C., et al., *Imatinib inhibits the expression of SCO2 and FRATAXIN genes that encode mitochondrial proteins in human Bcr-Abl(+) leukemia cells.* Blood Cells Mol Dis, 2014. **53**(1-2): p. 84-90.
379. Nath, A. and C. Chan, *Genetic alterations in fatty acid transport and metabolism genes are associated with metastatic progression and poor prognosis of human cancers.* Sci Rep, 2016. **6**: p. 18669.
380. Hanahan, D. and R.A. Weinberg, *The hallmarks of cancer.* Cell, 2000. **100**(1): p. 57-70.
381. Ng, A.C., et al., *Impact of vitamin D deficiency on the clinical presentation and prognosis of patients with newly diagnosed multiple myeloma.* Am J Hematol, 2009. **84**(7): p. 397-400.
382. Yarmolinsky, J., et al., *Causal Inference in Cancer Epidemiology: What Is the Role of Mendelian Randomization?* Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology, 2018. **27**(9): p. 995-1010.
383. Went, M., et al., *Assessing the effect of obesity-related traits on multiple myeloma using a Mendelian randomisation approach.* Blood cancer journal, 2017. **7**(6): p. e573-e573.

384. Chattopadhyay, S., et al., *Eight novel loci implicate shared genetic etiology in multiple myeloma, AL amyloidosis, and monoclonal gammopathy of unknown significance.* Leukemia, 2019.
385. Wu, J.H.Y., et al., *Genome-wide association study identifies novel loci associated with concentrations of four plasma phospholipid fatty acids in the de novo lipogenesis pathway: results from the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium.* Circulation. Cardiovascular genetics, 2013. **6**(2): p. 171-183.
386. Kim, H.I., et al., *Fine Mapping and Functional Analysis Reveal a Role of SLC22A1 in Acylcarnitine Transport.* American journal of human genetics, 2017. **101**(4): p. 489-502.
387. Alexander, D.D., et al., *Multiple myeloma: a review of the epidemiologic literature.* International journal of cancer, 2007. **120 Suppl 12**: p. 40-61.
388. Akram, M., et al., *Awareness and current knowledge of breast cancer.* Biological research, 2017. **50**(1): p. 33-33.
389. Barta, J.A., C.A. Powell, and J.P. Wisnivesky, *Global Epidemiology of Lung Cancer.* Annals of global health, 2019. **85**(1): p. 8.
390. Roncucci, L. and F. Mariani, *Prevention of colorectal cancer: How many tools do we have in our basket?* European journal of internal medicine, 2015. **26**(10): p. 752-756.
391. Fund, W.C.R. and A.I.f.C. Research, *Food, nutrition, physical activity, and the prevention of cancer: a global perspective*. Vol. 1. 2007: Amer Inst for Cancer Research.
392. Aran, V., et al., *Colorectal Cancer: Epidemiology, Disease Mechanisms and Interventions to Reduce Onset and Mortality.* Clinical colorectal cancer, 2016. **15**(3): p. 195-203.
393. Masarwi, M., et al., *Multiple Myeloma and Fatty Acid Metabolism.* JBMR plus, 2019. **3**(3): p. e10173-e10173.
394. Caro-Maldonado, A., et al., *Metabolic reprogramming is required for antibody production that is suppressed in anergic but exaggerated in chronically BAFF-exposed B cells.* Journal of immunology (Baltimore, Md. : 1950), 2014. **192**(8): p. 3626-3636.
395. Ziakas, P.D., et al., *Interleukin-6 polymorphisms and hematologic malignancy: a re-appraisal of evidence from genetic association studies.* Biomarkers : biochemical indicators of exposure, response, and susceptibility to chemicals, 2013. **18**(7): p. 625-631.
396. Li, Y., et al., *Association of IL-6 Promoter and Receptor Polymorphisms with Multiple Myeloma Risk: A Systematic Review and Meta-Analysis.* Genetic testing and molecular biomarkers, 2016. **20**(10): p. 587-596.
397. Alexander, D.D., et al., *Multiple myeloma: a review of the epidemiologic literature.* Int J Cancer, 2007. **120 Suppl 12**: p. 40-61.
398. Chang, S.-H., et al., *Obesity and the Transformation of Monoclonal Gammopathy of Undetermined Significance to Multiple Myeloma: A Population-Based Cohort Study.* Journal of the National Cancer Institute, 2016. **109**(5): p. djw264.
399. Thordardottir, M., et al., *Obesity and risk of monoclonal gammopathy of undetermined significance and progression to multiple myeloma: a population-based study.* Blood advances, 2017. **1**(24): p. 2186-2192.
400. Smith, G.D. and S. Ebrahim, *'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?* International journal of epidemiology, 2003. **32**(1): p. 1-22.
401. Batley, J. and D. Edwards, *Genome sequence data: management, storage, and visualization.* BioTechniques, 2009. **46**(5): p. 333-336.
402. Klein, J., et al., *A systematic evaluation of the design, orientation, and sequence context dependencies of massively parallel reporter assays.* bioRxiv, 2019: p. 576405.

403. Kheradpour, P., et al., *Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay.* Genome research, 2013. **23**(5): p. 800-811.
404. Melnikov, A., et al., *Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay.* Nature biotechnology, 2012. **30**(3): p. 271-277.
405. Tewhey, R., et al., *Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay.* Cell, 2016. **165**(6): p. 1519-1529.
406. Ulirsch, J.C., et al., *Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits.* Cell, 2016. **165**(6): p. 1530-1545.
407. Arnold, C.D., et al., *Genome-wide quantitative enhancer activity maps identified by STARR-seq.* Science (New York, N.Y.), 2013. **339**(6123): p. 1074-1077.
408. Neben, K., et al., *Combining information regarding chromosomal aberrations t(4;14) and del(17p13) with the International Staging System classification allows stratification of myeloma patients undergoing autologous stem cell transplantation.* Haematologica, 2010. **95**(7): p. 1150-1157.

Appendix 1

RSID	Sequencing Primer Sequence		Sequencing Direction	Sequencing Additives
rs6595443	Forward	AAGGAGTCAATTCTGCAAAAAG	Reverse	1M Betaine
	Reverse	TGCTGTTGTTGTTTGAAGTGG		
rs58618031	Forward	TGATAGTCATTTCTCACAAGAGCTG	Forward	1M Betaine
	Reverse	TCTCTGTCAAAATGAAACTTACCTTC		
rs11629542	Forward	CCAACCTCCTCATTGTAGGG	Forward	1M Betaine
	Reverse	AGCAAGAAACAAAGCACAGG		

Details of sequencing primers.

rsID		KASP Primer Sequence	Conditions
rs4325816	KASP Primer A1	GAAGGTGACCAAGTTCATGCTAACCTAGGTTGCTGGGAGAATGAT	Std42
	KASP Primer A2	GAAGGTCGGAGTCAACGGATTCCTAGGTTGCTGGGAGAATGAC	
	KASP Common Primer	CATGTGACGTTGTTTTCATAAATCTCATAA	
rs6595443	KASP Primer A1	GAAGGTGACCAAGTTCATGCTCCATTTCTGATAGTGTGTGTTAAAGTCT	Std42plus5
	KASP Primer A2	GAAGGTCGGAGTCAACGGATTCCATTTCTGATAGTGTGTGTTAAAGTCA	
	KASP Common Primer	GTGAATGCACCTAACAGAGTATCAAAATA	
rs1050976	KASP Primer A1	GAAGGTGACCAAGTTCATGCTAAGTGATGTGTTTACATTTACTGAAATGC	Std42plus5
	KASP Primer A2	GAAGGTCGGAGTCAACGGATTCAAGTGATGTGTTTACATTTACTGAAATGT	
	KASP Common Primer	TTTTCTCTGTCTTCCAGCAAGACCTAAT	
rs17507636	KASP Primer A1	GAAGGTGACCAAGTTCATGCTTTCACTGTAGCCATCTGTATCCC	Std42plus5
	KASP Primer A2	GAAGGTCGGAGTCAACGGATTCTTTCACTGTAGCCATCTGTATCCT	
	KASP Common Primer	CCTGCTTCTTTAATTATGTATAGGGTAGAA	
rs17501560	KASP Primer A1	GAAGGTGACCAAGTTCATGCTCAAGATACAACAGGTGAGACCCAA	Std42plus5
	KASP Primer A2	GAAGGTCGGAGTCAACGGATTAAGATACAACAGGTGAGACCCAG	
	KASP Common Primer	TGTCCTTAATAGTTTAGTCTCCAAAATCAT	
rs58618031	KASP Primer A1	GAAGGTGACCAAGTTCATGCTAGGAGGCCTCAGGAAACTTACG	Std42plus15
	KASP Primer A2	GAAGGTCGGAGTCAACGGATTAAGGAGGCCTCAGGAAACTTACA	
	KASP Common Primer	CTGACATTTTCCCCACTGGCATTCAT	
rs11629542	KASP Primer A1	GAAGGTGACCAAGTTCATGCTAAGTACGTGCCTAAAAGATGGACAC	Std42plus10
	KASP Primer A2	GAAGGTCGGAGTCAACGGATTAAGTACGTGCCTAAAAGATGGACAG	
	KASP Common Primer	GCCATGTCTGGGGCACTATTTCTAA	
rs13338946	KASP Primer A1	GAAGGTGACCAAGTTCATGCTCGAGACTCTATCTCAATAAATGAATAAAATG	Std42
	KASP Primer A2	GAAGGTCGGAGTCAACGGATTGCGAGACTCTATCTCAATAAATGAATAAAATA	
	KASP Common Primer	CACCCCACTTCATTTTTTCATAACACGTA	
rs11086029	KASP Primer A1	GAAGGTGACCAAGTTCATGCTGTGGCCTCCTCTACGTTGAAAAAAAA	Std42plus5
	KASP Primer A2	GAAGGTCGGAGTCAACGGATTGTGGCCTCCTCTACGTTGAAAAAAAT	
	KASP Common Primer	GGCTTCCAGGAAGAGGTAAGTAGTT	

Details of genotyping primers

KASPAR genotyping conditions

Std42

- Hot Start: 94ºC for 15 minutes
- Stage 1: 20 cycles
    - o 94ºC for 10 seconds
    - o 57ºC for 5 seconds
    - o 72ºC for 10 seconds
- Stage 2: 22 cycles
    - o 94ºC for 10 seconds
    - o 57ºC for 20 seconds
    - o 72ºC for 40 seconds

Std42plus5

- Hot Start: 94ºC for 15 minutes
- Stage 1: 20 cycles
    - o 94ºC for 10 seconds
    - o 57ºC for 5 seconds
    - o 72ºC for 10 seconds
- Stage 2: 22 cycles
    - o 94ºC for 10 seconds
    - o 57ºC for 20 seconds
    - o 72ºC for 40 seconds
- Stage 3: 5 cycles
    - o 94ºC for 10 seconds
    - o 57ºC for 1 minute

Std42plus10

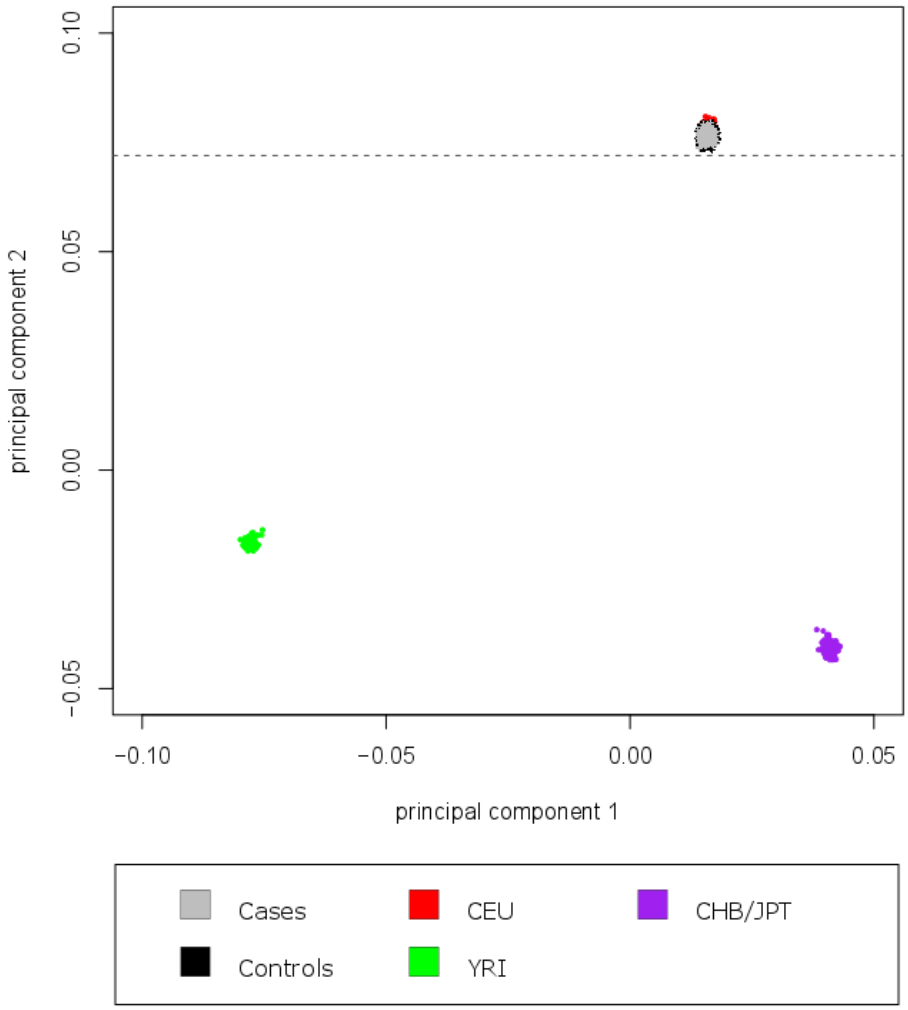
- Hot Start: 94ºC for 15 minutes
- Stage 1: 20 cycles
    - o 94ºC for 10 seconds
    - o 57ºC for 5 seconds
    - o 72ºC for 10 seconds
- Stage 2: 22 cycles
    - o 94ºC for 10 seconds
    - o 57ºC for 20 seconds
    - o 72ºC for 40 seconds

- Stage 3: 10 cycles
    - 94ºC for 10 seconds
    - 57ºC for 1 minute

Std42plus15

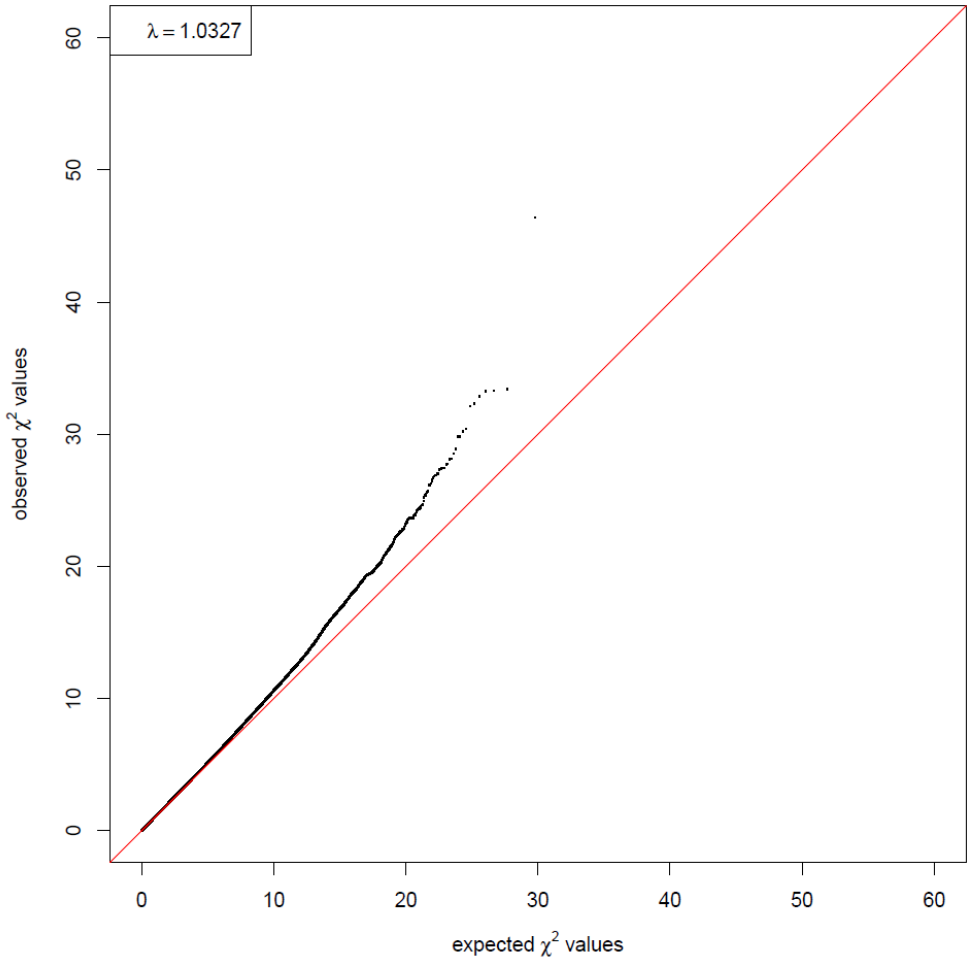
- Hot Start: 94ºC for 15 minutes
- Stage 1: 20 cycles
    - 94ºC for 10 seconds
    - 57ºC for 5 seconds
    - 72ºC for 10 seconds
- Stage 2: 22 cycles
    - 94ºC for 10 seconds
    - 57ºC for 20 seconds
    - 72ºC for 40 seconds
- Stage 3: 15 cycles
    - 94ºC for 10 seconds
    - 57ºC for 1 minute

Appendix 2



Principal components analysis plot for the OncoArray cohort. Post removal of cases of non-European ancestry. The first two principal components of the analysis are plotted. Cases and controls outside of the intervals $0.0155 \leq x \leq 0.019$, and $0.0735 \leq y \leq 0.079$ were excluded in order to remove individuals of non-European ancestry (grey dotted line shows the lower threshold of the second principal component). HapMap CEU individuals are plotted in red; CHB/JPT individuals are plotted in purple; YRI individuals are plotted in green. Cases are plotted in grey, controls plotted in black.

Appendix 3



Quantile-Quantile (Q-Q) plot. Observed and expected χ^2 values of association between SNP genotype and risk of multiple myeloma after imputation for the OncoArray cohort. λ=1.0327, λ_{1000}=1.0209**.** The red line represents the null hypothesis of no true association. *Q-Q* plots for the UK, Sweden/Norway, Germany, Iceland, USA and Netherlands sets have been previously reported.

Appendix 4

	UK Cases			
rsID	AA	Aa	aa	r^2
rs58618031	7/7	58/59	83/83	0.99
rs11629542	28/31	59/63	51/53	0.91
rs6595443	53/54	74/78	41/41	0.97

Concordance between directly sequenced and imputed genotype. Shown are SNPs which were genome-wide significant after replication. These comprised 147 randomly selected samples from the Oncoarray case series. AA, major homozygote; Aa, heterozygote; aa, minor homozygote. r^2 indicates Pearson product-moment correlation coefficient between imputed and sequenced genotype.

Appendix 5

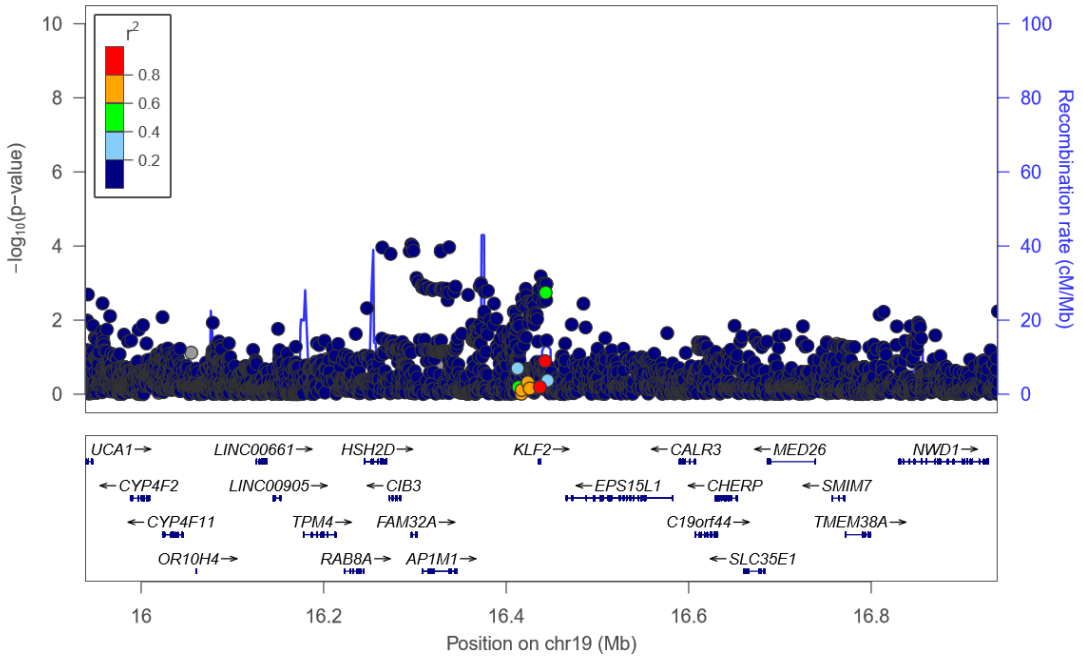
Table containing details of the replication of top association signals (overleaf).

				German Replication				Swedish Replication				Danish Replication			
SNP	Chr.	Pos. (b37)	Risk Allele	Cases RAF	Controls RAF	OR	*P*-value	Cases RAF	Controls RAF	OR	*P*-value	Cases RAF	Controls RAF	OR	*P*-value
rs4325816	2	174808899	T	0.79	0.76	1.19	0.014	0.78	0.77	1.06	0.461	0.83	0.79	1.25	0.037
rs6595443	5	122743325	T	0.47	0.45	1.08	0.202	0.45	0.42	1.15	0.038	0.44	0.43	1.04	0.639
rs17507636	7	106291118	C	-	-	-	-	0.79	0.76	1.19	0.036	0.74	0.75	0.93	0.485
rs58618031	7	124583896	T	-	-	-	-	0.75	0.72	1.18	0.032	0.72	0.71	1.03	0.761
rs13338946	16	30700858	C	0.32	0.28	1.24	0.001	0.29	0.27	1.13	0.112	0.37	0.28	1.51	9.0×10^{-6}
rs11086029	19	16438661	T	0.23	0.21	1.18	0.022	0.24	0.22	1.12	0.149	0.26	0.24	1.11	0.293
rs1050976	6	408079	T	0.49	0.47	1.07	0.268	0.47	0.45	1.10	0.188	0.48	0.46	1.08	0.371
rs11629542	15	90098754	G	0.44	0.46	0.92	0.205	0.54	0.54	1.01	0.910	0.57	0.54	1.16	0.091
rs17501560	7	81415783	A	0.81	0.81	1.00	0.954	0.83	0.80	1.22	0.016	0.82	0.81	1.08	0.461

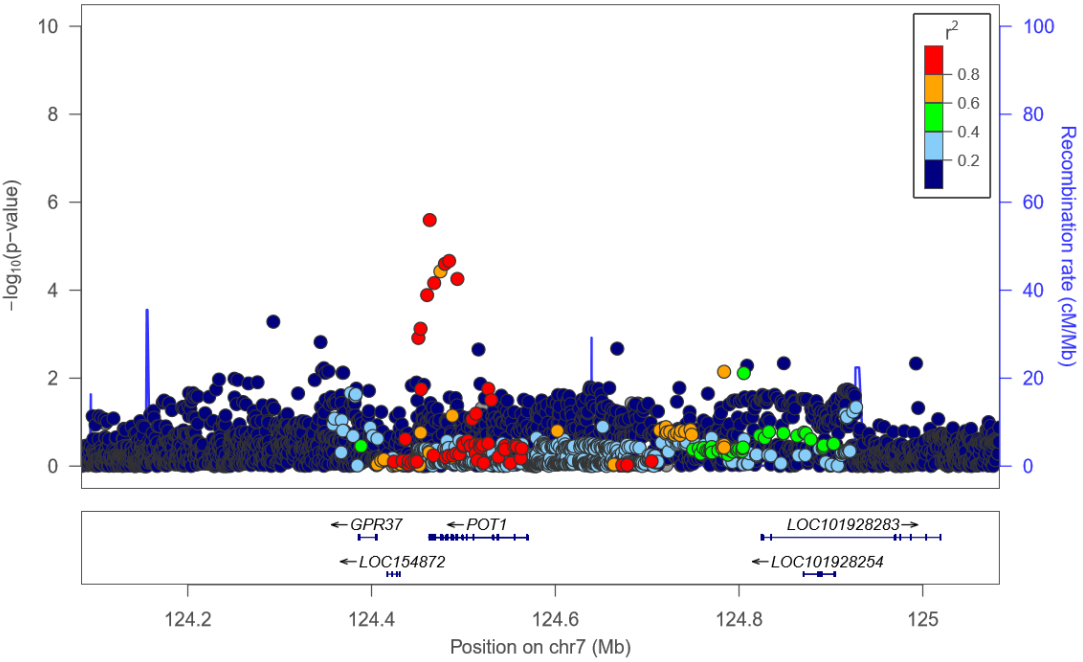
Replication of top association signals. Shown are SNPs which were taken forward for replication genotyping. Cases RAF, risk allele frequency of replication cases; Control RAF, risk allele frequency of replication controls. *P*-values are shown for each replication series (logistic regression). rs17507636 had been previously replicated in the German cohort, with association values [116]; cases RAF: 0.760, controls RAF: 0.735, OR: 1.15, *P*-value: 0.06. A meta-analysis of this with discovery cohorts and replication series was performed using R version 3.3.1 (R Development Core Team, Vienna, Austria).

Appendix 6

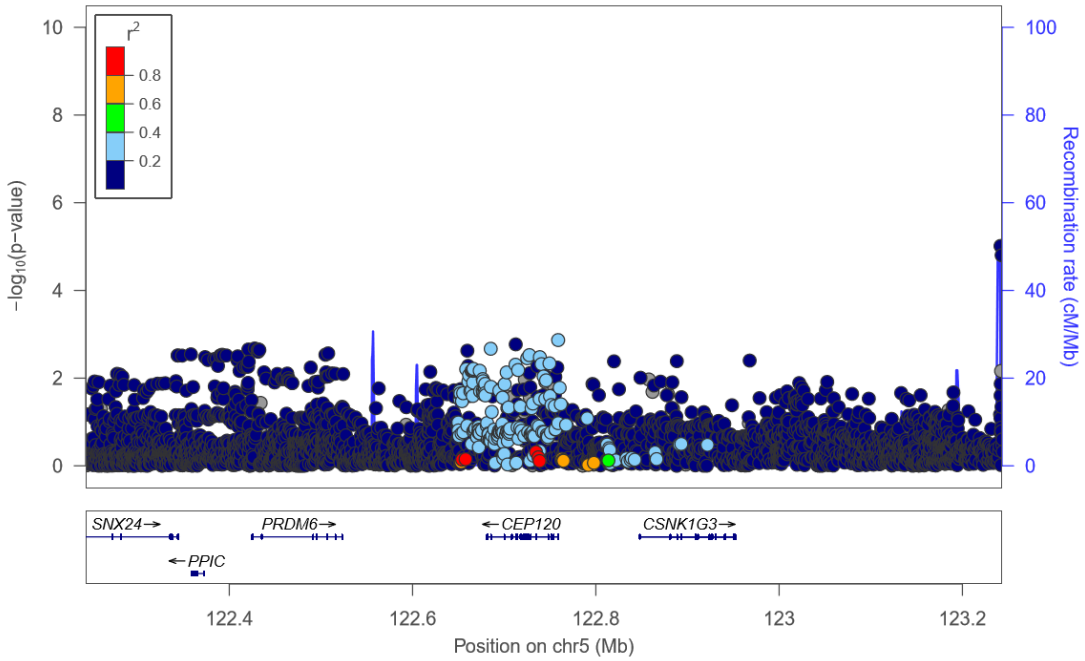
Plots of six newly discovered loci after conditioning on sentinel SNP at each locus.



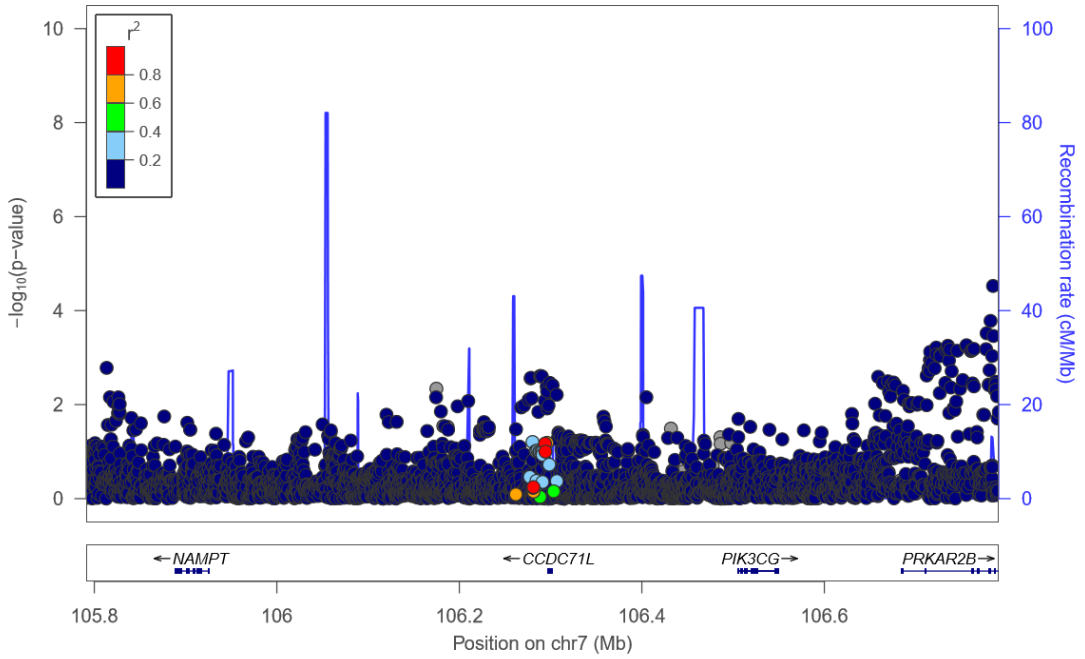
rs11086029



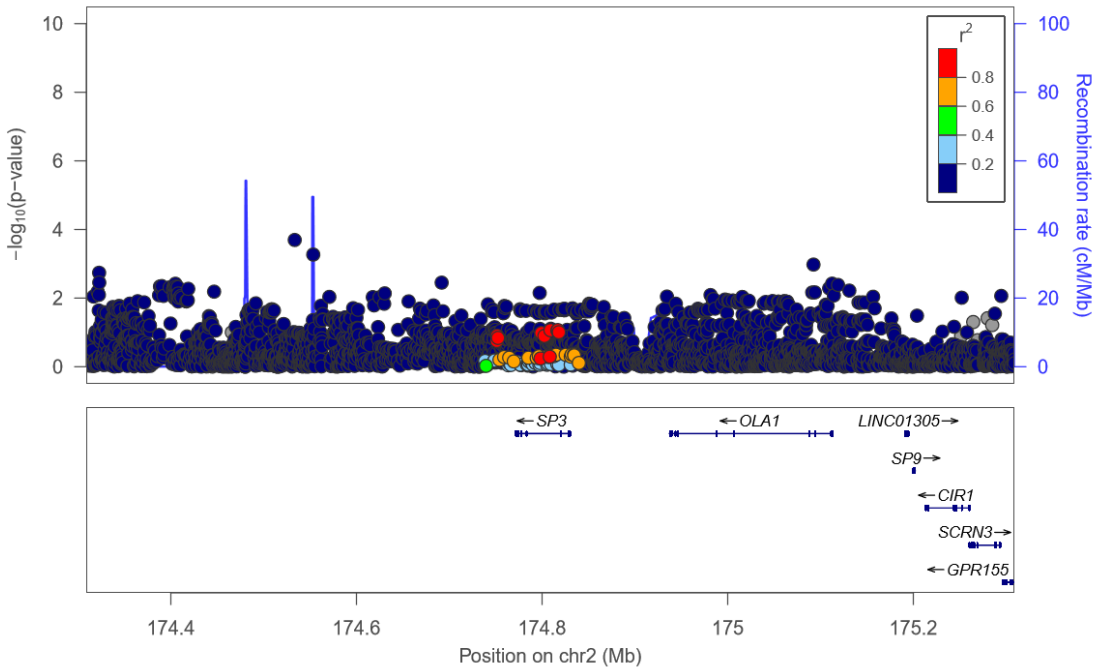
rs58618031



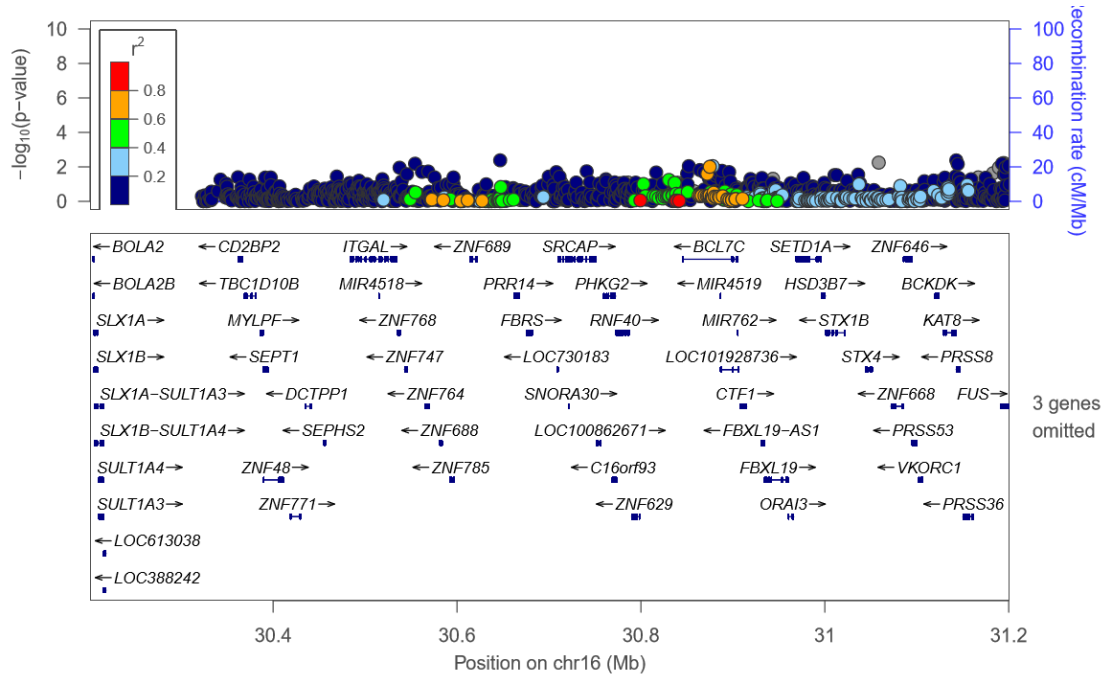
rs6595443



rs17507636



rs4325816



rs13338946

Appendix 7

Tables showing association between SNP genotype and age at diagnosis, and between SNP genotype and sex (overleaf).

	USA		Germany		OncoArray		UK		Meta				
RSID	**Beta**	***P*-value**	**Beta**	***P*-value**	**Beta**	***P*-value**	**Beta**	***P*-value**	**OR**	**95% CIs**	***P*-value**	**I^2**	**P_{HET}**
rs34229995	-0.43	0.15	-0.08	0.69	-0.23	0.50	-0.05	0.79	0.87	0.69 - 1.10	0.23	0	0.73
rs9372120	0.16	0.20	-0.11	0.22	0.02	0.87	-0.01	0.85	0.99	0.90 - 1.08	0.84	6	0.36
rs7781265	-0.06	0.69	0.02	0.84	0.15	0.28	-0.11	0.22	0.98	0.87 - 1.10	0.72	0	0.44
rs1948915	0.02	0.86	0.05	0.53	0.07	0.49	0.04	0.58	1.04	0.96 - 1.13	0.30	0	0.99
rs2811710	0.13	0.24	0.03	0.74	-0.02	0.83	-0.03	0.67	1.01	0.93 - 1.10	0.80	0	0.65
rs2790457	0.04	0.75	0.06	0.45	-0.02	0.83	0.12	0.08	1.07	0.98 - 1.17	0.12	0	0.73
rs7193541	0.05	0.67	0.01	0.88	0.08	0.44	0.02	0.74	1.03	0.95 - 1.12	0.44	0	0.95
rs6066835	-0.60	0.00	-0.24	0.06	0.30	0.07	0.07	0.49	0.94	0.83 - 1.07	0.36	82	0.00
rs7577599	-0.04	0.79	-0.10	0.33	0.22	0.12	0.01	0.94	1.00	0.90 - 1.12	0.96	14	0.32
rs6599192	0.12	0.37	0.01	0.93	-0.13	0.28	-0.03	0.67	0.98	0.89 - 1.08	0.73	0	0.56
rs4487645	0.09	0.46	-0.03	0.79	0.04	0.73	0.07	0.29	1.05	0.96 - 1.16	0.27	0	0.89
rs34562254	0.19	0.24	-0.07	0.51	-0.01	0.93	0.05	0.60	1.02	0.91 - 1.14	0.72	0	0.58

Relationship between SNP genotype and sex (continued on next page).

	USA		Germany		OncoArray		UK		Meta				
RSID	**Beta**	***P*-value**	**Beta**	***P*-value**	**Beta**	***P*-value**	**Beta**	***P*-value**	**OR**	**95% CIs**	***P*-value**	**I^2**	**P_{HET}**
rs3132535	0.03	0.82	0.11	0.17	-0.12	0.25	0.03	0.61	1.03	0.95 - 1.12	0.49	0	0.39
rs1423269	0.25	0.04	0.00	0.99	0.02	0.87	0.10	0.15	1.09	0.99 - 1.19	0.07	5	0.37
rs139402	-0.06	0.56	-0.09	0.24	-0.11	0.27	0.01	0.83	0.96	0.89 - 1.03	0.25	0	0.64
rs138747	-0.34	0.21	0.22	0.17	-0.46	0.04	-0.04	0.80	0.94	0.79 - 1.13	0.51	59	0.06
rs10936600	0.13	0.30	0.02	0.85	-0.10	0.40	-0.04	0.55	0.99	0.90 - 1.09	0.81	0	0.54
rs11086029	-0.11	0.40	0.15	0.21	-0.06	0.16	-0.10	0.41	0.96	0.88 - 1.05	0.37	27	0.25
rs13338946	0.08	0.51	0.07	0.31	0.07	0.49	-0.08	0.28	1.06	0.97 - 1.15	0.20	0	0.69
rs17507636	0.06	0.14	-0.04	0.50	0.09	0.75	-0.20	0.22	1.03	0.94 - 1.12	0.59	25	0.26
rs4325816	-0.02	0.91	0.01	0.80	-0.04	0.96	-0.01	0.56	0.97	0.89 - 1.07	0.60	0	0.99
rs58618031	0.05	0.19	0.01	0.56	-0.06	0.96	-0.16	0.48	0.97	0.89 - 1.07	0.56	0	0.53
rs6595443	0.13	0.36	0.07	0.09	0.09	0.47	-0.10	0.12	1.08	1.00 - 1.16	0.07	8	0.35

Relationship between SNP genotype and sex. Continued from previous page. Analysis based on beta values calculated from logistic regression on the discovery phase data sets from UK (2,282 cases), Oncoarray (878 cases), German (1,508 cases) and USA (780 cases) series. The meta-analysis was conducted using a fixed-effects model. This assumes that the underlying effect across all studies is the same. To test for potential heterogeneity, Cochran's Q-statistic was calculated such that $P_{HET} > 0.05$ implied the presence of non-significant heterogeneity. The heterogeneity index, I^2 (0-100), was also measured; this quantifies the proportion of the total variation due to heterogeneity.

	USA		Germany		OncoArray		UK		Meta				
RSID	Beta	*P*-value	Beta	*P*-value	Beta	*P*-value	Beta	*P*-value	OR	95% CIs	*P*-value	I^2	P_{HET}
rs34229995	0.93	0.45	0.07	0.94	-1.34	0.43	0.44	0.67	1.23	0.40 - 3.79	0.71	0	0.74
rs9372120	-1.18	0.03	0.15	0.74	0.31	0.60	0.64	0.08	1.15	0.74 - 1.80	0.53	61	0.05
rs7781265	0.02	0.97	-0.41	0.43	0.45	0.53	-0.09	0.84	0.92	0.53 - 1.61	0.77	0	0.80
rs1948915	0.28	0.56	0.36	0.35	-0.22	0.68	-0.43	0.19	0.95	0.64 - 1.43	0.82	0.5	0.39
rs2811710	-0.02	0.97	-0.40	0.32	-0.01	0.98	-0.20	0.56	0.83	0.55 - 1.25	0.36	0	0.92
rs2790457	0.51	0.36	-0.24	0.57	-0.14	0.80	-0.36	0.31	0.86	0.55 - 1.33	0.50	0	0.62
rs7193541	-0.55	0.25	-0.22	0.56	-0.62	0.22	-0.02	0.94	0.77	0.53 - 1.14	0.19	0	0.69
rs6066835	-0.71	0.35	0.14	0.82	-1.45	0.08	-0.07	0.88	0.71	0.38 - 1.34	0.29	0	0.40
rs7577599	-0.50	0.46	0.22	0.63	0.87	0.21	0.13	0.75	1.19	0.71 - 1.97	0.51	0	0.56
rs6599192	-0.68	0.26	-0.20	0.63	-0.52	0.40	0.04	0.92	0.79	0.50 - 1.26	0.33	0	0.73
rs4487645	-0.68	0.19	0.33	0.56	-0.46	0.40	0.18	0.58	0.93	0.60 - 1.45	0.75	0	0.39
rs34562254	-1.00	0.16	0.35	0.45	0.35	0.62	-0.14	0.76	0.98	0.57 - 1.66	0.93	0	0.40

(Continued on next page)

	USA		Germany		OncoArray		UK		Meta				
RSID	**Beta**	***P*-value**	**Beta**	***P*-value**	**Beta**	***P*-value**	**Beta**	***P*-value**	**OR**	**95% CIs**	***P*-value**	**I^2**	**P_{HET}**
rs3132535	0.03	0.95	-0.28	0.40	-0.65	0.22	0.02	0.94	0.84	0.57 - 1.24	0.38	0	0.70
rs1423269	0.50	0.36	-0.54	0.16	-1.06	0.06	0.23	0.53	0.85	0.55 - 1.30	0.45	52	0.10
rs139402	-0.32	0.50	0.68	0.03	0.89	0.07	-0.30	0.32	1.22	0.85 - 1.75	0.29	63	0.04
rs138747	1.33	0.24	1.03	0.15	0.86	0.45	0.71	0.34	2.56	1.10 - 5.95	0.03	0	0.97
rs10936600	0.49	0.40	-0.48	0.22	-1.55	0.01	-0.53	0.17	0.61	1.00 - 2.47	0.03	51	0.11
rs11086029	-1.10	0.04	-0.27	0.48	-1.36	0.11	-0.17	0.61	0.64	0.41 - 0.99	0.04	15	0.32
rs13338946	-1.07	0.04	-0.30	0.41	0.25	0.65	0.50	0.14	0.95	0.64 - 1.43	0.81	59	0.06
rs17507636	0.60	0.30	0.59	0.17	0.54	0.36	0.12	0.73	1.48	0.95 - 2.31	0.09	0	0.80
rs4325816	0.08	0.90	-0.36	0.37	-0.01	0.99	-0.15	0.69	0.85	0.54 - 1.33	0.48	0	0.93
rs58618031	0.22	0.69	-0.76	0.07	-0.83	0.15	-0.16	0.67	0.68	0.44 - 1.07	0.10	0	0.39
rs6595443	0.65	0.16	-0.22	0.51	-0.43	0.38	-0.26	0.39	0.89	0.62 - 1.27	0.51	15	0.32

Relationship between SNP genotype and age at diagnosis. (Continued from previous page). Analysis based on beta values calculated from linear regression on the discovery phase data sets from UK (2,282 cases), Oncoarray (878 cases), German (1,508 cases) and USA (780 cases) cohorts. The meta-analysis was conducted using a fixed-effects model. This assumes that the underlying effect across all studies is the same. To test for potential heterogeneity, Cochran's Q-statistic was calculated such that $P_{HET} > 0.05$ implied the presence of non-significant heterogeneity. The heterogeneity index, I^2 (0-100), was also measured; this quantifies the proportion of the total variation due to heterogeneity.

Appendix 8

	German		UK		OncoArray		Meta	
RSID	**Beta**	***P*-value**	**Beta**	***P*-value**	**Beta**	***P*-value**	**Beta**	***P*-value**
rs2790457	0.12	0.41	-0.18	0.18	0.44	0.14	0.002	0.99
rs13338946	-0.17	0.23	0.03	0.78	0.28	0.34	-0.03	0.76
rs7193541	0.13	0.29	-0.07	0.53	-0.17	0.52	0.002	0.98
rs34562254	-0.02	0.92	-0.26	0.12	-0.40	0.29	-0.17	0.14
rs11086029	-0.14	0.33	-0.11	0.39	0.11	0.70	-0.10	0.28
rs6066835	-0.13	0.54	-0.35	0.06	0.90	0.07	-0.16	0.22
rs138747	0.42	0.13	-0.18	0.52	1.00	0.10	0.20	0.28
rs139402	0.08	0.52	-0.07	0.56	-0.25	0.33	-0.02	0.78
rs4325816	0.02	0.90	-0.08	0.58	0.15	0.62	-0.01	0.88
rs6599192	0.26	0.11	-0.10	0.47	0.23	0.48	0.07	0.48
rs10936600	-0.03	0.83	0.11	0.46	0.03	0.91	0.04	0.69
rs1423269	0.12	0.44	0.05	0.70	-0.11	0.69	0.06	0.54
rs6595443	-0.30	0.02	-0.13	0.24	-0.07	0.78	-0.19	0.02
rs34229995	0.52	0.11	-0.39	0.29	-1.15	0.26	0.05	0.83
rs3132535	0.02	0.87	0.01	0.91	-0.11	0.69	0.005	0.95
rs9372120	-0.11	0.47	0.04	0.76	0.37	0.22	0.02	0.87
rs4487645	0.28	0.05	0.05	0.67	0.76	0.01	0.21	0.02
rs17507636	-0.003	0.98	0.02	0.86	-0.20	0.54	-0.01	0.95
rs58618031	-0.03	0.84	0.16	0.20	0.02	0.95	0.07	0.41
rs7781265	0.21	0.25	-0.02	0.90	0.01	0.99	0.08	0.49
rs1948915	0.03	0.83	0.15	0.20	0.02	0.94	0.09	0.30
rs2811710	0.17	0.22	-0.02	0.89	-0.16	0.56	0.04	0.63
rs7577599	0.29	0.10	-0.03	0.86	0.04	0.91	0.10	0.34

Relationship between SNP genotype and t(4;14) subtype. German cases: 142, UK cases: 170, Oncoarray cases: 33, Meta: 345. Case-only analysis; Beta values obtained from logistic regression. FISH and ploidy classification of UK and German samples were determined as previously described [270, 271].

	German		UK		OncoArray		Meta	
RSID	**Beta**	***P*-value**	**Beta**	***P*-value**	**Beta**	***P*-value**	**Beta**	***P*-value**
rs2790457	-0.09	0.43	0.32	0.004	-0.06	0.82	0.10	0.19
rs13338946	-0.05	0.64	0.36	0.001	-0.24	0.34	0.11	0.12
rs7193541	0.02	0.82	-0.02	0.85	0.48	0.03	0.04	0.50
rs34562254	-0.05	0.68	-0.14	0.35	-0.46	0.15	-0.12	0.19
rs11086029	-0.04	0.74	0.12	0.27	-0.11	0.66	0.03	0.70
rs6066835	-0.13	0.41	0.17	0.30	-0.44	0.29	-0.02	0.88
rs138747	0.14	0.52	-0.42	0.06	-0.36	0.50	-0.15	0.33
rs139402	-0.17	0.07	-0.09	0.37	-0.27	0.23	-0.14	0.03
rs4325816	0.004	0.97	0.04	0.75	-0.13	0.62	0.01	0.93
rs6599192	0.13	0.27	-0.03	0.83	0.13	0.65	0.06	0.46
rs10936600	-0.14	0.22	0.01	0.94	-0.26	0.32	-0.09	0.27
rs1423269	-0.003	0.98	0.13	0.26	0.12	0.61	0.07	0.37
rs6595443	-0.04	0.71	-0.09	0.34	0.21	0.36	-0.04	0.53
rs34229995	-0.23	0.35	0.03	0.92	-0.44	0.62	-0.15	0.44
rs3132535	-0.05	0.60	0.11	0.31	0.27	0.24	0.04	0.53
rs9372120	0.09	0.43	0.19	0.10	0.10	0.68	0.14	0.08
rs4487645	0.13	0.23	-0.14	0.18	-0.06	0.80	-0.01	0.87
rs17507636	0.03	0.81	-0.13	0.25	-0.05	0.85	-0.05	0.52
rs58618031	0.02	0.83	-0.05	0.66	-0.19	0.46	-0.03	0.71
rs7781265	-0.08	0.54	-0.38	0.01	-0.30	0.34	-0.22	0.02
rs1948915	-0.20	0.05	0.08	0.42	0.60	0.01	-0.003	0.96
rs2811710	0.02	0.89	-0.12	0.26	-0.08	0.71	-0.05	0.44
rs7577599	-0.11	0.42	-0.09	0.49	0.04	0.89	-0.09	0.33

Relationship between SNP genotype and t(11;14) subtype. German cases: 277, UK cases: 231, Oncoarray cases: 47, Meta: 555. Case-only analysis; Beta values obtained from logistic regression. FISH and ploidy classification of UK and German samples were determined as previously described [270, 271].

	German		UK		OncoArray		Meta	
RSID	**Beta**	***P*-value**	**Beta**	***P*-value**	**Beta**	***P*-value**	**Beta**	***P*-value**
rs2790457	0.28	0.37	0.34	0.31	1.06	0.07	0.41	0.05
rs13338946	0.20	0.49	-0.06	0.86	-0.24	0.68	0.05	0.82
rs7193541	-0.03	0.92	-0.08	0.80	-0.05	0.93	-0.05	0.80
rs34562254	-0.24	0.52	0.56	0.20	-0.07	0.92	0.08	0.76
rs11086029	-0.18	0.57	-0.24	0.45	0.43	0.44	-0.12	0.56
rs6066835	-0.49	0.28	-0.90	0.06	0.66	0.49	-0.54	0.09
rs138747	-0.06	0.93	-0.63	0.33	0.00	1.00	-0.29	0.49
rs139402	0.24	0.36	-0.62	0.04	-0.06	0.91	-0.13	0.48
rs4325816	0.45	0.17	0.01	0.98	0.10	0.87	0.23	0.31
rs6599192	-0.42	0.22	0.19	0.61	0.08	0.90	-0.11	0.64
rs10936600	-0.12	0.71	0.04	0.91	-0.54	0.36	-0.12	0.59
rs1423269	0.61	0.06	-0.12	0.72	1.12	0.04	0.41	0.06
rs6595443	-0.45	0.10	0.08	0.78	0.14	0.79	-0.16	0.40
rs34229995	0.38	0.59	1.67	0.08	-1.06	0.59	0.70	0.20
rs3132535	-0.09	0.75	0.15	0.63	0.13	0.81	0.03	0.88
rs9372120	-0.18	0.59	0.31	0.36	-0.95	0.10	-0.08	0.71
rs4487645	0.52	0.08	0.08	0.80	-0.24	0.67	0.24	0.24
rs17507636	0.29	0.35	0.05	0.87	-0.34	0.59	0.12	0.57
rs58618031	0.10	0.75	-0.04	0.90	-0.68	0.23	-0.06	0.77
rs7781265	0.20	0.59	-0.21	0.63	0.92	0.21	0.15	0.58
rs1948915	0.05	0.86	-0.14	0.66	0.07	0.89	-0.02	0.91
rs2811710	-0.24	0.42	0.19	0.54	0.43	0.40	0.03	0.87
rs7577599	0.34	0.36	-0.39	0.33	0.70	0.36	0.07	0.77

Relationship between SNP genotype and t(14;16) subtype. German cases: 29, UK cases: 24, Oncoarray cases: 8, Meta: 61 Case-only analysis; Beta values obtained from logistic regression. FISH and ploidy classification of UK and German samples were determined as previously described [271, 408].

	German		UK		OncoArray		Meta	
RSID	**Beta**	***P*-value**	**Beta**	***P*-value**	**Beta**	***P*-value**	**Beta**	***P*-value**
rs2790457	0.03	0.71	0.14	0.06	-	-	0.10	0.09
rs13338946	-0.04	0.59	-0.05	0.50	-0.02	0.91	-0.04	0.39
rs7193541	0.08	0.28	0.04	0.57	-	-	0.06	0.25
rs34562254	0.01	0.92	0.15	0.12	0.51	0.04	0.12	0.09
rs11086029	0.11	0.22	-0.01	0.92	-0.10	0.60	0.03	0.59
rs6066835	0.06	0.60	0.37	0.001	-0.40	0.21	0.20	0.01
rs138747	-0.32	0.04	-0.05	0.74	-0.28	0.49	-0.19	0.08
rs139402	0.07	0.33	0.18	0.01	0.22	0.19	0.14	0.003
rs4325816	-0.05	0.61	-0.06	0.43	-	-	-0.06	0.35
rs6599192	-0.15	0.13	-0.05	0.55	-	-	-0.09	0.15
rs10936600	0.04	0.63	0.01	0.86	-	-	0.03	0.64
rs1423269	-0.07	0.40	-0.13	0.08	-	-	-0.11	0.06
rs6595443	-0.05	0.50	-0.002	0.98	-0.15	0.39	-0.03	0.49
rs34229995	-0.15	0.43	0.03	0.89	-	-	-0.07	0.63
rs3132535	0.09	0.23	0.04	0.54	-	-	0.07	0.21
rs9372120	-0.04	0.68	-0.11	0.14	-0.14	0.48	-0.09	0.12
rs4487645	-0.08	0.33	-0.03	0.62	-	-	-0.05	0.32
rs17507636	-0.01	0.91	-0.001	0.99	0.08	0.69	0.001	0.98
rs58618031	-0.09	0.27	0.03	0.68	0.20	0.30	-0.01	0.92
rs7781265	0.01	0.94	-0.01	0.95	-	-	0.0001	1.00
rs1948915	0.08	0.30	-0.04	0.52	-0.34	0.06	-0.02	0.74
rs2811710	0.001	0.99	-0.02	0.78	-	-	-0.01	0.84
rs7577599	-0.061	0.56	0.08	0.35			0.02	0.74

Relationship between SNP genotype and hyperdiploid subtype. German cases: 661, UK cases: 702, Oncoarray cases: 257, Meta: 1,620. Case-only analysis; Beta values obtained from logistic regression. FISH and ploidy classification of UK and German samples were determined as previously described [270, 271].

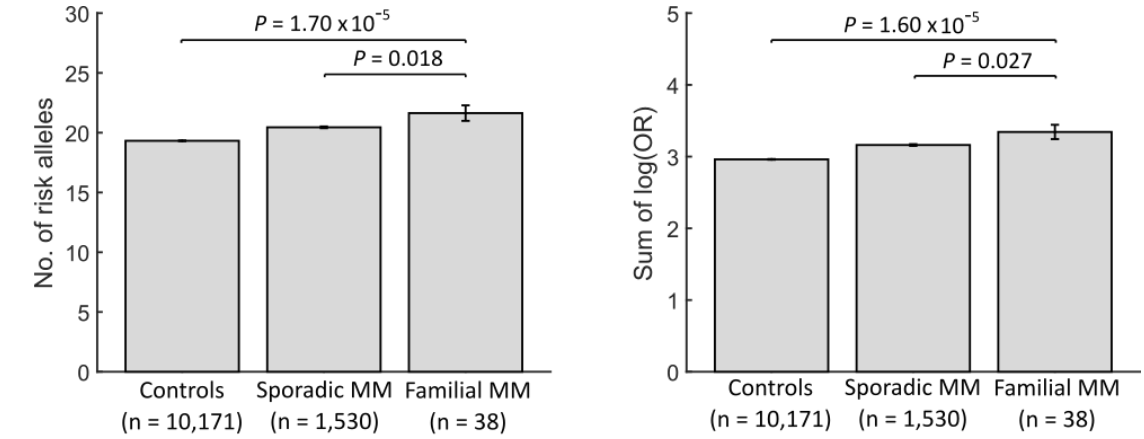
Appendix 9

Relationship between genome-wide significant SNPs genotype and patient overall survival (overleaf).

		Germ-GMMG		UK-MyIX		UK-MyXI		USA-UAMS		Meta	
SNP	**Risk Allele**	**HR**	***P*-value**	**HR**	***P*-value**	**HR**	***P*-value**	**HR**	***P*-value**	**HR**	***P*-value**
rs4325816	T	1.06	0.69	0.88	0.12	0.86	0.27	0.85	0.11	0.89	0.03
rs6599192	G	1.27	0.14	1.01	0.95	1.08	0.54	0.98	0.86	1.04	0.50
rs10936600	A	1.06	0.69	1.00	0.95	1.10	0.45	0.98	0.83	1.02	0.70
rs1423269	A	1.07	0.61	1.05	0.49	0.97	0.80	0.97	0.78	1.02	0.69
rs6595443	T	1.14	0.27	0.98	0.78	0.92	0.39	1.03	0.66	1.00	0.91
rs34229995	G	1.12	0.71	0.68	0.03	1.24	0.56	0.99	0.96	0.88	0.30
rs3132535	A	0.87	0.27	0.91	0.18	0.89	0.27	1.07	0.40	0.95	0.21
rs9372120	G	0.99	0.92	1.22	0.01	0.84	0.12	1.05	0.63	1.06	0.21
rs4487645	C	1.04	0.77	0.94	0.39	1.10	0.38	0.99	0.90	0.99	0.89
rs17507636	C	1.06	0.70	1.03	0.73	1.10	0.42	1.04	0.67	1.05	0.36
rs58618031	T	1.00	0.99	1.01	0.89	1.14	0.23	1.17	0.08	1.07	0.14
rs7781265	A	0.82	0.20	1.20	0.07	0.90	0.46	-	-	1.02	0.77
rs1948915	C	0.96	0.73	1.02	0.83	0.96	0.68	1.04	0.59	1.01	0.89
rs2790457	G	0.88	0.40	0.91	0.21	0.92	0.48	0.94	0.55	0.92	0.08
rs13338946	C	0.76	0.03	1.02	0.75	1.07	0.54	0.91	0.25	0.96	0.32
rs7193541	T	1.01	0.94	0.94	0.34	1.05	0.61	1.06	0.43	1.00	0.96
rs34562254	A	0.91	0.51	1.07	0.46	1.40	0.05	1.13	0.33	1.09	0.16
rs11086029	T	0.98	0.89	0.82	0.01	1.06	0.57	1.05	0.62	0.94	0.17
rs6066835	C	0.92	0.63	1.03	0.77	1.13	0.48	0.90	0.41	0.99	0.87
rs139402	C	0.92	0.49	1.05	0.49	1.06	0.60	1.03	0.68	1.03	0.51

Relationship between genome-wide significant SNPs genotype and patient overall survival. Data from: 1,165 cases from the UK MRC Myeloma-IX trial (UK-MyIX); 877 MM cases from the UK MRC Myeloma-XI trial (UK-MyXI); 511 of the patients recruited to the German-GWAS (GER-GMMG); 703 MM cases in the UAMS Myeloma Institute for Research and Therapy GWAS (US-UAMS). *P*-values calculated from Cox regression analysis. Data for SNPs rs2811710, rs7577599 and rs138747, or a correlated SNP ($r^2 > 0.6$) to use as proxy, were not present in the survival analysis.

Appendix 10



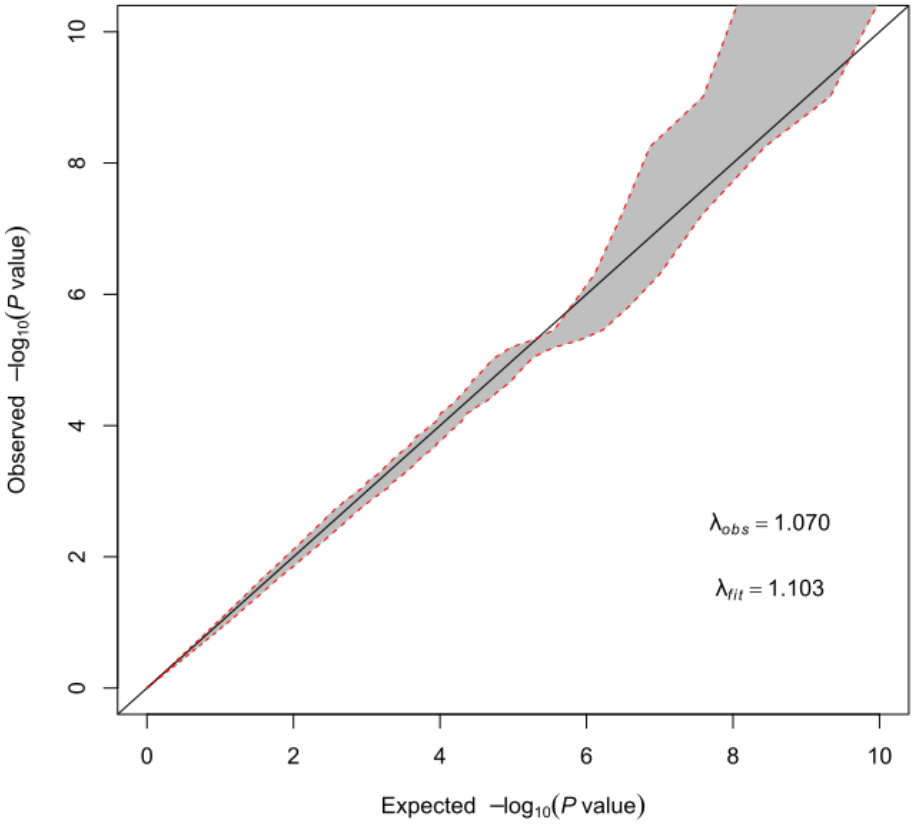
Median:	19.2	20.4	21.6	Median:	2.96	3.15	3.33
Mean:	19.3	20.4	21.6	Mean:	2.96	3.16	3.34

Polygenic risk scores (PRS) for familial MM, sporadic MM and population-controls. A higher risk allele burden is seen in the familial MM compared with both sporadic MM and controls (difference in PRS score tested by one-sided Student's t-test). (a) Based on number of risk alleles carried; (b) Calculated as the sum log-transformed odds ratios. The observed 1.08-fold enrichment of PRS in familial over sporadic cases is entirely compatible the expected familial risk attributable to the 23 risk SNPs of 1.10 given by:

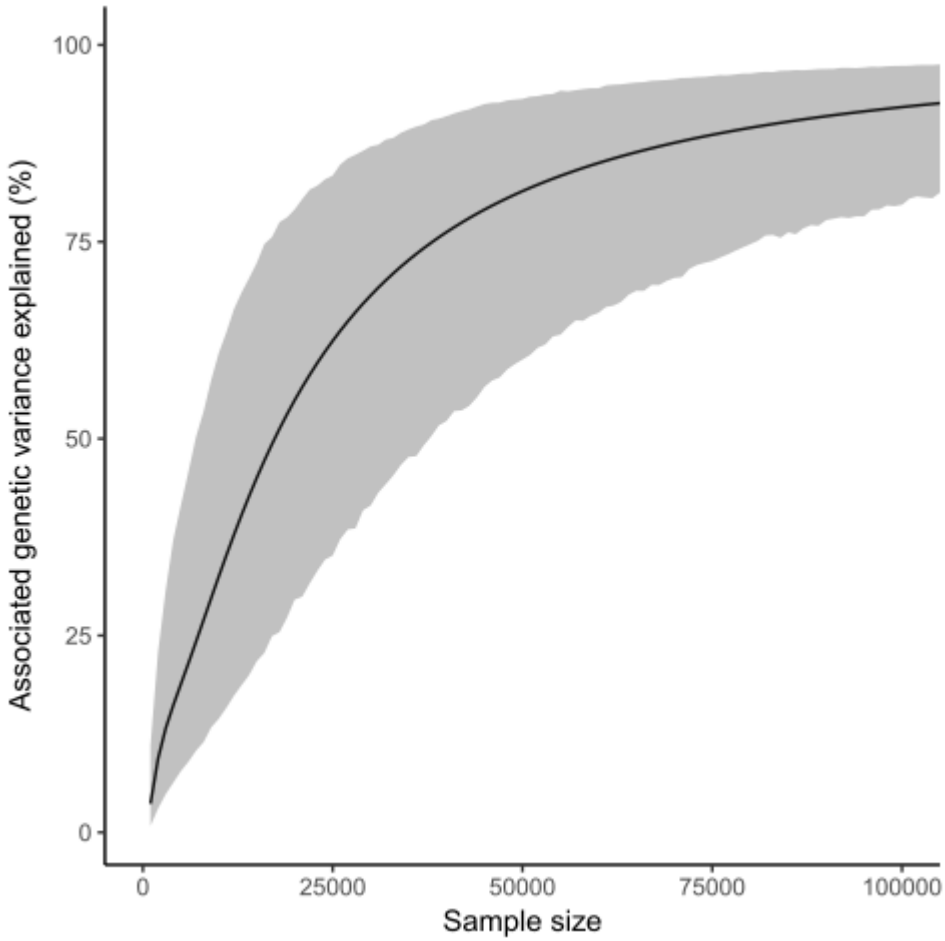
$$\prod_{i}^{n=23} \frac{p_i r_i^2 + q_i}{p_i r_i + q_i^2}$$

where p_i is the frequency of the risk allele for locus i, $q_i = 1 - p_i$, and r_i is the estimated per-allele OR.

Appendix 11



***Q-Q* plot comparing observed distributions of association statistics against those expected under a three-component model.** Grey shaded area represents the 80% confidence interval.



Projected percentage of GWAS heritability explained for a given sample size. Results were obtained using a three-component model to estimate distribution of effect sizes. Grey shaded area represents the 95% confidence interval of the heritability estimate.

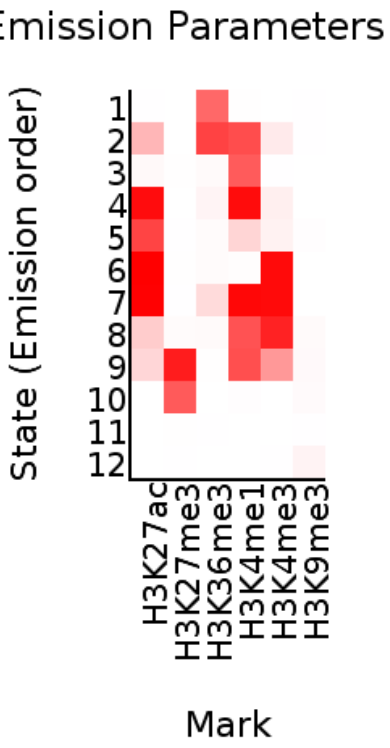
Appendix 12

rsID	Locus	Probe chromosome	Gene	P_{SMR}	P_{HEIDI}
rs6595443	5q23.2	5	CEP120	1.27×10^{-4}	6.60×10^{-2}
rs2807754	10p21.1	10	WAC	4.53×10^{-5}	6.28×10^{-1}
rs1423269	5q15	5	ELL2	7.08×10^{-7}	5.58×10^{-3}
rs4487645	7p15.3	7	CDCA7L	8.37×10^{-15}	1.08×10^{-2}
rs6090899	20q13.13	20	PREX1	4.01×10^{-4}	5.46×10^{-3}

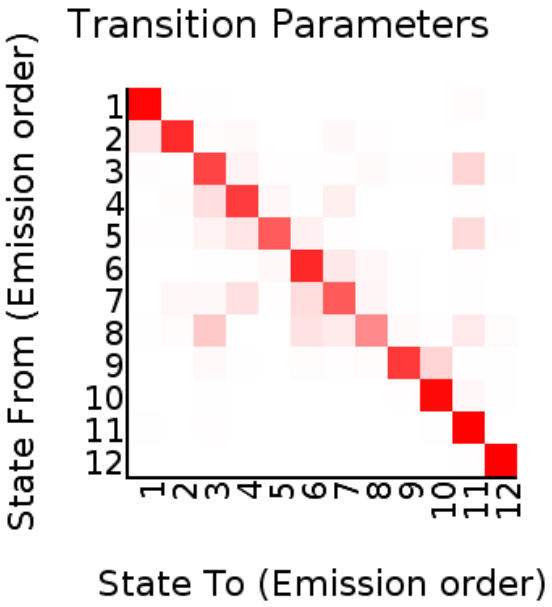
Summary of results from SMR analysis. A threshold for the SMR test of $P_{SMR} < 1\times10^{-3}$ corresponding to a Bonferroni correction for 45 tests was set. For all genes passing this threshold plots of the eQTL and GWAS associations at the locus were generated, as well as plots of GWAS and eQTL effect sizes (*i.e*. corresponding to input for the HEIDI heterogeneity test). HEIDI test *P*-values < 0.05 were considered as being reflective of heterogeneity. This threshold is conservative for gene discovery because it retains fewer genes than when correcting for multiple testing. Probes which passed the HEIDI threshold are highlighted in grey.

Appendix 13

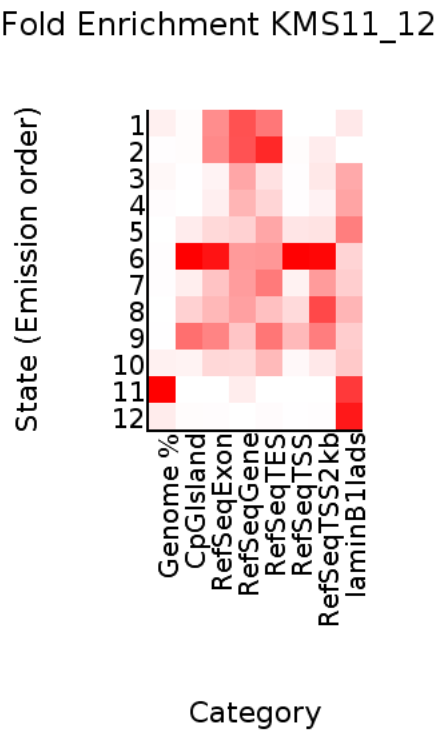
a



b



c



Heat maps outputted by ChromHMM pipeline show a- emission parameters, b- transition parameters and c- state functional enrichments for the KMS11 MM cell line. Columns in **(c)** are labelled as follows: Genome % indicates the relative percentage of the genome represented by each state and relative fold enrichment for RefSeq transcription start sites (TSS); CpG Islands; 2000 base pair intervals around the TSS; exons; genes; transcript end sites (TES); evolutionary conservation; and nuclear lamina associated regions, respectively. Heat maps shown were used to assign states based on previously described rules [176, 292, 293]. The ChromHMM model was learned across 3 MM cell lines; JJN3, KMS11 and MM1S.

Appendix 14

Locus	Lead SNP	Transcription Factor			
3q26.2	rs10936600	ATF2	BATF	CEBPB	CHD1
		POLR2A	POLR3G	POU2F2	RUNX3
		CTCF	EBF1	ELF1	ELK1
		RXRA	SIN3A	STAT5A	TAF1
		MAZ	MTA3	MXI1	NFIC
		MAX	PML	WRNIP1	YY1
		EP300	ETS1	FOXM1	IRF4
		TBL1XR1	TBP		
5q15	rs1423269	ATF2	BATF	BCL11A	BCL3
		MXI1	NFATC1	NFIC	PML
		BCLAF1	BHLHE40	CEBPB	CHD2
		POU2F2	RELA	RUNX3	SP1
		EBF1	EP300	FOXM1	IKZF1
		SPI1	STAT3	STAT5A	TBL1XR1
		IRF4	JUND	MAZ	MEF2A
		TBP	TCF12	TCF3	MEF2C
		MTA3			
8q24	rs1948915	ATF2	BCL3	BCLAF1	CEBPB
		RELA	RUNX3	SIN3A	SPI1
		CTCF	EBF1	EP300	FOXM1
		STAT3	STAT5A	TBL1XR1	YY1
		JUND	MAZ	MEF2A	MEF2C
		MTA3	MXI1	NFIC	PML
		POLR2A	RAD21	SMC3	
16p11	rs13338946	BCL3	CHD1	CHD2	CTCF
		EBF1	MAZ	MXI1	NFIC
		POLR2A	RELA	RUNX3	SIN3A
		SP1	SPI1	TAF1	TCF12
		WRNIP	YY1		
20q13	rs6066835	ATF2	BCL11A	EBF1	ELF1
		EP300	FOXM1	IKZF1	MEF2A
		MEF2C	NFIC	POLR2A	RUNX3
		SPI1	TBL1XR1	USF1	WRNIP1
22q13	rs138747	ATF2	ATF3	BCL3	BHLHE40
		MAZ	MEF2C	MTA3	MXI1
		CEBPB	CHD1	CHD2	EBF1
		NFATC1	NFE2	NFIC	NFYA
		EGR1	ELF1	ELK1	EP300
		NFYB	NR2C2	PAX5	PBX3
		FOS	FOXM1	GABPA	IKZF1
		PML	POLR2A	POU2F2	RELA
		IRF4	MAX	TAF1	TBL1XR1
		RUNX3	SIN3A	TBP	TCF12
		TBP	TCF12	TCF3	USF1
		USF2	WRNIP1	YY1	ZEB1
		SP1	SPI1	SRF	STAT5A
		ZNF143			

Full lists of TF binding at selected loci. TF ChIP-seq (161 factors) with Factorbook Motifs for GM12878 were downloaded from ENCODE.

Appendix 15

Genes significantly associated with risk of multiple myeloma (overleaf). Includes associations seen in the HLA region (6p21.32-33,6p22.1). s.d., standard deviation. Detailed are the S-MultiXcan *P*-value for association between gene expression and the corresponding Z-scores quantifying this relationship (*e.g.* a positive score indicates increased gene expression increases risk). N and N_{indep} indicate the total number of single-tissue results used for S-MultiXcan analysis and the number of independent components after singular value decomposition, respectively.

Locus	Gene	*P*-value	N/N_{indep}	Z-score min	Z-score max	Z -score mean	Z-score s.d.	SNP adjusting for	*P*-value after SNP adjustment
16p11.2	QPRT	1.01×10^{-7}	17/8	-2.73	3.04	-0.59	1.63	rs13338946	0.15
16p11.2	RNF40	4.02×10^{-7}	24/3	0.05	5.68	4.67	1.48	rs13338946	0.89
16p11.2	PRR14	4.28×10^{-7}	2/2	-5.38	-0.20	-2.79	3.66	rs13338946	0.34
16p11.2	C16orf93	8.07×10^{-7}	13/5	-5.74	-0.34	-4.59	1.73	rs13338946	0.24
16p11.2	RP11-2C24.5	1.54×10^{-6}	5/5	-5.64	4.43	-0.58	3.80	rs13338946	0.73
16p11.2	PRSS53	1.71×10^{-6}	16/8	-5.19	3.68	-1.04	2.71	rs13338946	0.79
16q23.1	RFWD3	7.71×10^{-7}	34/7	-3.41	6.35	2.51	3.26	rs7193541	0.47
17p11.2	TBC1D27	1.95×10^{-13}	6/6	-1.91	4.19	0.51	2.16	rs34562254	0.89
17p11.2	USP32P1	4.88×10^{-13}	3/3	-7.29	2.80	-1.36	5.27	rs34562254	0.01
17p11.2	PEMT	5.65×10^{-8}	14/7	-1.74	5.43	1.36	1.93	rs34562254	0.01
22q13.1	APOBEC3C	1.10×10^{-18}	21/8	-8.93	0.24	-4.09	2.21	rs139402	0.13
22q13.1	APOBEC3H	4.28×10^{-15}	7/5	-5.45	7.92	-0.95	4.38	rs139402	0.76
22q13.1	FAM83F	4.65×10^{-10}	11/8	-4.25	2.56	-0.48	2.01	rs139402	1.1×10^{-4}
22q13.1	APOBEC3D	6.2×10^{-10}	29/7	-8.38	-0.85	-4.15	1.56	rs139402	0.04

Locus	Gene	*P*-value	N/N_{indep}	Z-score min	Z-score max	Z -score mean	Z-score s.d.	SNP adjusting for	*P*-value after SNP adjustment
22q13.1	APOBEC3F	5.15×10^{-9}	5/4	-6.34	6.15	1.09	5.07	rs139402	0.13
22q13.1	APOBEC3G	1.81×10^{-7}	43/2	0.36	6.57	4.94	1.17	rs139402	0.17
2p23.3	KIF3C	1.65×10^{-18}	6/6	-9.40	4.35	-1.19	4.50	rs7577599	1.4×10^{-9}
2p23.3	EPT1	8.37×10^{-16}	9/9	-1.76	6.00	1.30	2.72	rs7577599	2.1×10^{-5}
2p23.3	CENPO	1.48×10^{-13}	12/8	-6.60	2.22	-0.05	2.57	rs7577599	6.1×10^{-8}
2p23.3	DNMT3A	2.44×10^{-13}	8/8	-2.89	7.96	1.94	3.07	rs7577599	0.01
2p23.3	AC010150.1	2.90×10^{-13}	4/4	-0.88	7.89	1.61	4.20	rs7577599	8.9×10^{-10}
2p23.3	PTGES3P2	4.46×10^{-11}	7/5	-4.23	2.03	-2.46	2.08	rs7577599	1.1×10^{-4}
2p23.3	DTNB	1.16×10^{-7}	11/10	-3.88	5.78	0.36	2.38	rs7577599	3.1×10^{-3}
2p23.3	DNAJC27	1.74×10^{-7}	8/8	-0.74	4.52	1.95	1.58	rs7577599	0.11
3p22.1	ULK4	9.01×10^{-15}	43/6	0.90	8.89	6.60	2.24	rs6599192	0.85
3q26.2	MYNN	7.84×10^{-13}	6/6	-7.91	1.58	-1.66	3.32	rs10936600	0.17
3q26.2	LRRIQ4	9.63×10^{-9}	3/2	-5.94	-0.88	-4.25	2.92	rs10936600	0.03
3q26.2	LRRC34	3.35×10^{-8}	21/2	3.97	6.47	5.12	0.66	rs10936600	0.82
3q26.2	ACTRT3	4.28×10^{-7}	4/4	-0.94	5.80	1.56	2.94	rs10936600	0.48

Locus	Gene	*P*-value	N/N_{indep}	Z-score min	Z-score max	Z -score mean	Z-score s.d.	SNP adjusting for	*P*-value after SNP adjustment
6p21.32	HLA-DRB9	4.71×10^{-12}	26/21	-6.00	0.30	-3.44	1.86		
6p21.32	PRRT1	3.97×10^{-11}	13/8	-6.55	0.44	-3.82	1.94	rs3132535	7.1×10^{-4}
6p21.32	HLA-DRB6	5.21×10^{-10}	47/6	-6.66	0.54	-5.01	1.43		
6p21.32	HLA-DQB1	6.88×10^{-9}	47/3	2.08	5.91	4.68	0.96		
6p21.32	PPT2	2.03×10^{-8}	4/3	-4.89	1.19	-2.52	2.83		
6p21.32	AGER	2.78×10^{-8}	21/7	-6.14	1.37	-3.28	1.62		
6p21.32	HLA-DRB1	2.99×10^{-8}	43/14	-3.93	6.42	3.31	2.15		
6p21.32	HLA-DRB5	5.29×10^{-8}	47/2	2.86	6.55	5.29	0.95		
6p21.32	RPL32P1	1.17×10^{-7}	25/25	-1.74	5.91	0.73	1.75		
6p21.32	EGFL8	1.34×10^{-7}	4/4	-4.56	2.27	-1.39	2.86		
6p21.32	HLA-DQA1	4.02×10^{-7}	37/10	-0.01	5.65	3.24	1.63		
6p21.33	TCF19	5.36×10^{-20}	36/6	-7.00	4.61	-3.65	2.70	rs3132535	0.02
6p21.33	HCG27	1.37×10^{-19}	44/5	-8.39	-2.45	-5.97	1.41	rs3132535	0.03
6p21.33	VARS2	7.77×10^{-19}	45/24	-6.98	4.55	-0.13	2.74	rs3132535	0.66
6p21.33	NRM	3.05×10^{-18}	21/19	-4.59	6.59	1.28	3.09	rs3132535	0.28

Locus	Gene	*P*-value	N/N_{indep}	Z-score min	Z-score max	Z -score mean	Z-score s.d.	SNP adjusting for	*P*-value after SNP adjustment
6p21.33	SFTA2	4.90×10^{-18}	7/7	-7.61	2.81	-1.96	3.83	rs3132535	0.17
6p21.33	ABCF1	2.40×10^{-17}	18/17	-8.55	1.70	-2.07	3.14	rs3132535	0.60
6p21.33	XXbac-BPG248L24.12	7.79×10^{-16}	34/11	-7.66	5.48	-3.17	3.30	rs3132535	7.5×10^{-5}
6p21.33	LY6G6C	1.20×10^{-15}	4/4	-1.07	7.32	2.24	3.76	rs3132535	5.5×10^{-4}
6p21.33	ABHD16A	1.72×10^{-15}	9/9	-6.83	4.82	-0.21	3.37	rs3132535	0.12
6p21.33	PPP1R18	4.66×10^{-15}	14/13	-2.51	4.86	1.92	2.24	rs3132535	0.40
6p21.33	GTF2H4	5.89×10^{-15}	12/12	-4.08	8.38	1.53	3.88	rs3132535	0.28
6p21.33	GPANK1	1.49×10^{-14}	14/11	-5.02	3.96	-0.53	2.59	rs3132535	0.03
6p21.33	FLOT1	1.87×10^{-14}	24/16	-9.23	4.66	-1.30	3.37	rs3132535	0.07
6p21.33	C6orf25	2.20×10^{-14}	5/5	-2.18	6.53	2.27	3.70	rs3132535	1.2×10^{-3}
6p21.33	XXbac-BPG299F13.14	2.32×10^{-14}	11/10	-7.81	0.75	-4.30	2.68	rs3132535	3.1×10^{-3}
6p21.33	DDX39B	4.83×10^{-14}	8/6	-5.69	2.95	-1.01	3.09	rs3132535	0.03
6p21.33	CCHCR1	8.99×10^{-14}	40/5	-5.56	8.24	-2.01	3.41	rs3132535	0.15
6p21.33	HLA-B	2.94×10^{-13}	33/9	-8.18	3.60	-4.80	2.47	rs3132535	0.06
6p21.33	DDAH2	3.09×10^{-13}	13/6	-1.45	7.41	3.26	2.21	rs3132535	0.23
6p21.33	HCG20	1.03×10^{-12}	27/23	-3.24	6.23	1.35	2.70	rs3132535	0.10

Locus	Gene	*P*-value	N/N_{indep}	Z-score min	Z-score max	Z -score mean	Z-score s.d.	SNP adjusting for	*P*-value after SNP adjustment
6p21.33	XXbac-BPG181B23.7	1.13×10^{-12}	47/9	-1.51	7.36	3.32	2.19	rs3132535	0.37
6p21.33	PSORS1C2	1.41×10^{-12}	32/5	-3.72	5.60	-1.24	1.67	rs3132535	0.06
6p21.33	CYP21A2	2.88×10^{-12}	29/11	-8.41	-0.30	-3.90	1.58	rs3132535	0.10
6p21.33	DDR1	3.51×10^{-12}	12/11	-4.95	6.01	0.05	3.42	rs3132535	0.36
6p21.33	IER3	3.65×10^{-12}	10/10	-6.45	3.57	-0.31	3.11	rs3132535	0.15
6p21.33	CLIC1	3.86×10^{-12}	9/7	-0.59	6.46	3.72	1.98	rs3132535	0.15
6p21.33	WASF5P	6.42×10^{-12}	6/6	0.80	4.71	2.83	1.53	rs3132535	0.75
6p21.33	POU5F1	7.81×10^{-12}	41/6	-5.12	5.04	0.05	1.69	rs3132535	0.21
6p21.33	AIF1	1.55×10^{-11}	6/6	-6.65	-2.66	-4.06	1.48	rs3132535	0.01
6p21.33	PSORS1C1	3.05×10^{-11}	41/4	-2.32	6.28	0.82	1.95	rs3132535	0.01
6p21.33	LST1	5.78×10^{-11}	10/10	-7.04	2.06	-2.19	2.83	rs3132535	0.09
6p21.33	MSH5	8.08×10^{-11}	8/8	-4.41	4.20	-0.79	2.54	rs3132535	0.09
6p21.33	DHX16	8.37×10^{-11}	13/13	-7.67	4.73	-0.39	3.30	rs3132535	0.22
6p21.33	HSPA1B	9.81×10^{-11}	5/5	-0.36	6.24	2.53	2.79	rs3132535	5.3×10^{-4}
6p21.33	VARS	1.05×10^{-10}	2/2	1.26	6.77	4.02	3.90	rs3132535	0.73

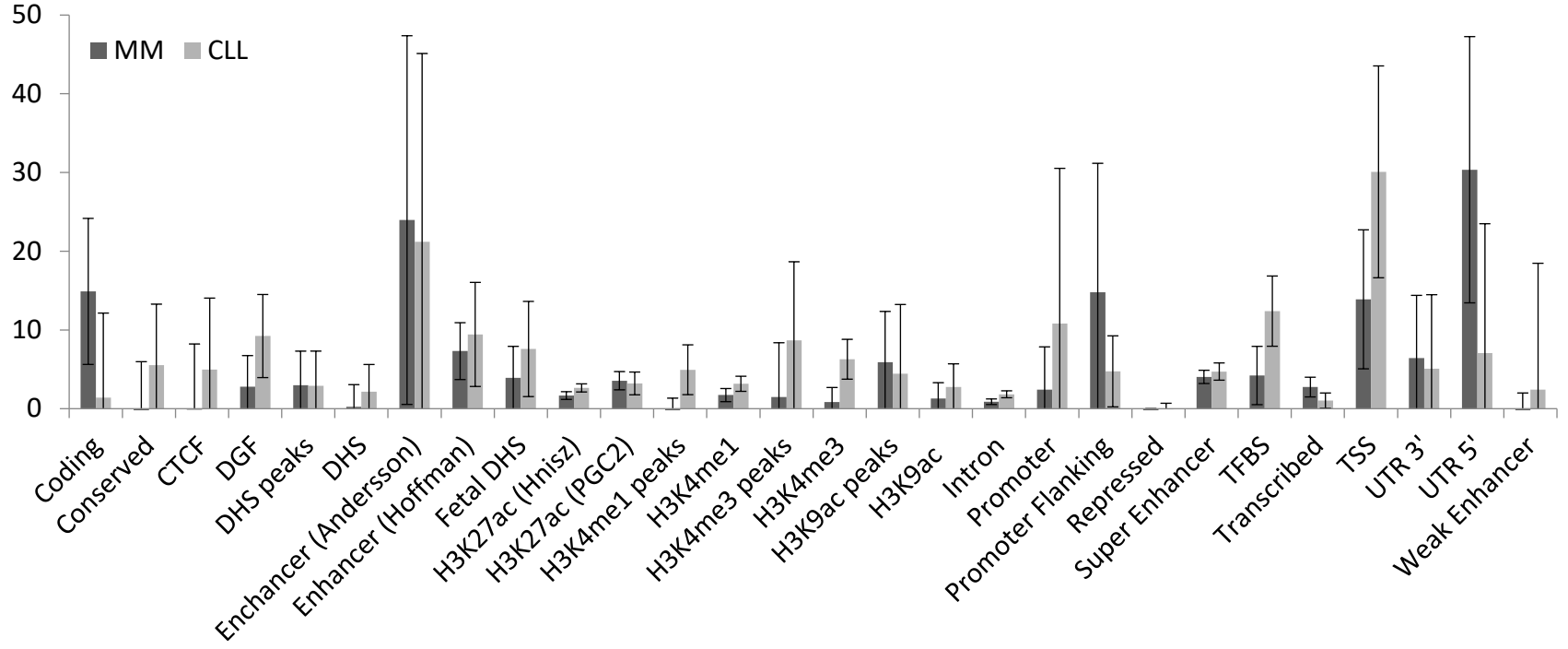
Locus	Gene	*P*-value	N/N_{indep}	Z-score min	Z-score max	Z -score mean	Z-score s.d.	SNP adjusting for	*P*-value after SNP adjustment
6p21.33	TNXA	1.13×10^{-10}	29/7	-3.14	3.85	0.18	2.04	rs3132535	1.3×10^{-3}
6p21.33	C2	4.60×10^{-10}	7/7	-7.02	-0.65	-3.01	2.39	rs3132535	0.44
6p21.33	MICB	6.50×10^{-10}	41/4	-5.08	4.27	-1.30	1.78	rs3132535	0.23
6p21.33	PRR3	6.54×10^{-10}	4/4	-5.48	1.69	-2.18	3.15	rs3132535	0.30
6p21.33	XXbac-BPG27H4.8	7.77×10^{-10}	3/3	-4.29	2.96	-1.33	3.80	rs3132535	0.12
6p21.33	MDC1	8.07×10^{-10}	8/8	-3.05	5.09	1.38	3.18	rs3132535	0.17
6p21.33	HLA-E	2.68×10^{-9}	11/11	-6.59	2.04	-2.49	2.53	rs3132535	0.57
6p21.33	EHMT2	4.69×10^{-9}	5/5	-6.17	1.85	-1.45	3.91	rs3132535	0.29
6p21.33	C4B	6.27×10^{-9}	31/14	-5.44	3.77	-1.22	2.31	rs3132535	0.01
6p21.33	VWA7	8.93×10^{-9}	7/4	-2.58	4.65	2.67	2.50	rs3132535	0.04
6p21.33	NELFE	1.40×10^{-8}	16/6	-3.76	5.43	-0.60	2.29	rs3132535	1.1×10^{-3}
6p21.33	PRRC2A	4.44×10^{-8}	8/7	-5.91	3.22	-1.22	2.54	rs3132535	0.05
6p21.33	XXbac-BPG181B23.6	4.49×10^{-8}	1/1	5.47	5.47	5.47	NA	rs3132535	0.04
6p21.33	BAG6	5.07×10^{-8}	31/10	-4.29	3.78	-0.71	1.88	rs3132535	0.01
6p21.33	C4A	5.73×10^{-8}	44/3	0.68	5.46	3.95	1.06	rs3132535	0.12

Locus	Gene	*P*-value	N/N_{indep}	Z-score min	Z-score max	Z -score mean	Z-score s.d.	SNP adjusting for	*P*-value after SNP adjustment
6p21.33	HCG22	6.02×10^{-8}	32/2	-5.47	7.75	2.48	4.68	rs3132535	0.01
6p21.33	HSPA1L	8.13×10^{-8}	4/4	-0.18	5.19	2.60	2.41	rs3132535	0.14
6p21.33	SAPCD1	8.27×10^{-8}	4/4	-5.91	0.75	-1.58	3.06	rs3132535	0.91
6p21.33	LTA	1.50×10^{-7}	6/6	-5.34	4.68	-0.01	3.35	rs3132535	0.37
6p21.33	C6orf15	3.77×10^{-7}	4/4	-5.59	0.09	-1.98	2.55	rs3132535	0.39
6p21.33	CFB	4.66×10^{-7}	4/4	-4.49	2.59	-1.02	2.89	rs3132535	0.01
6p21.33	FKBPL	8.80×10^{-7}	6/5	-4.37	3.40	0.82	3.16	rs3132535	1.4×10^{-4}
6p21.33	ATF6B	1.00×10^{-6}	28/2	-4.48	1.22	-1.14	1.16	rs3132535	0.03
6p22.1	HCG17	4.98×10^{-15}	11/10	-4.32	8.63	1.82	3.46	rs34229995	0.50
6p22.1	TRIM39	1.24×10^{-14}	2/2	-7.88	-0.61	-4.25	5.14	rs34229995	0.39
6p22.1	HLA-L	5.14×10^{-12}	41/3	-6.62	0.34	-4.23	1.63	rs34229995	2.9×10^{-3}
6p22.1	RPP21	1.87×10^{-10}	5/5	-6.58	1.52	-0.95	3.24	rs34229995	0.31
6p22.1	PAIP1P1	1.63×10^{-8}	7/7	-3.77	2.01	-0.75	2.13	rs34229995	0.05
6p22.1	ZNRD1	5.76×10^{-8}	22/14	-3.20	3.30	1.07	1.53		
6p22.1	ZSCAN9	2.53×10^{-7}	19/9	-4.02	1.96	-1.13	1.73		

Locus	Gene	*P*-value	N/N_{indep}	Z-score min	Z-score max	Z -score mean	Z-score s.d.	SNP adjusting for	*P*-value after SNP adjustment
6q21	ATG5	1.55×10^{-12}	4/4	0.93	5.89	3.72	2.41	rs9372120	0.07
7p15.3	CDCA7L	9.61×10^{-9}	8/8	-3.11	4.61	1.12	2.42	rs75341503	0.23
7q36.1	CHPF2	2.53×10^{-7}	6/6	-2.01	2.13	0.40	1.49	rs7781265	0.06

Appendix 16

Partitioned heritability analysis showing results for 28 functional categories (overleaf).



Partitioned heritability analysis showing results for 28 functional categories. The full baseline model as per Finucane *et al* [291], was used in this analysis, excluding category flanking regions from the plot.

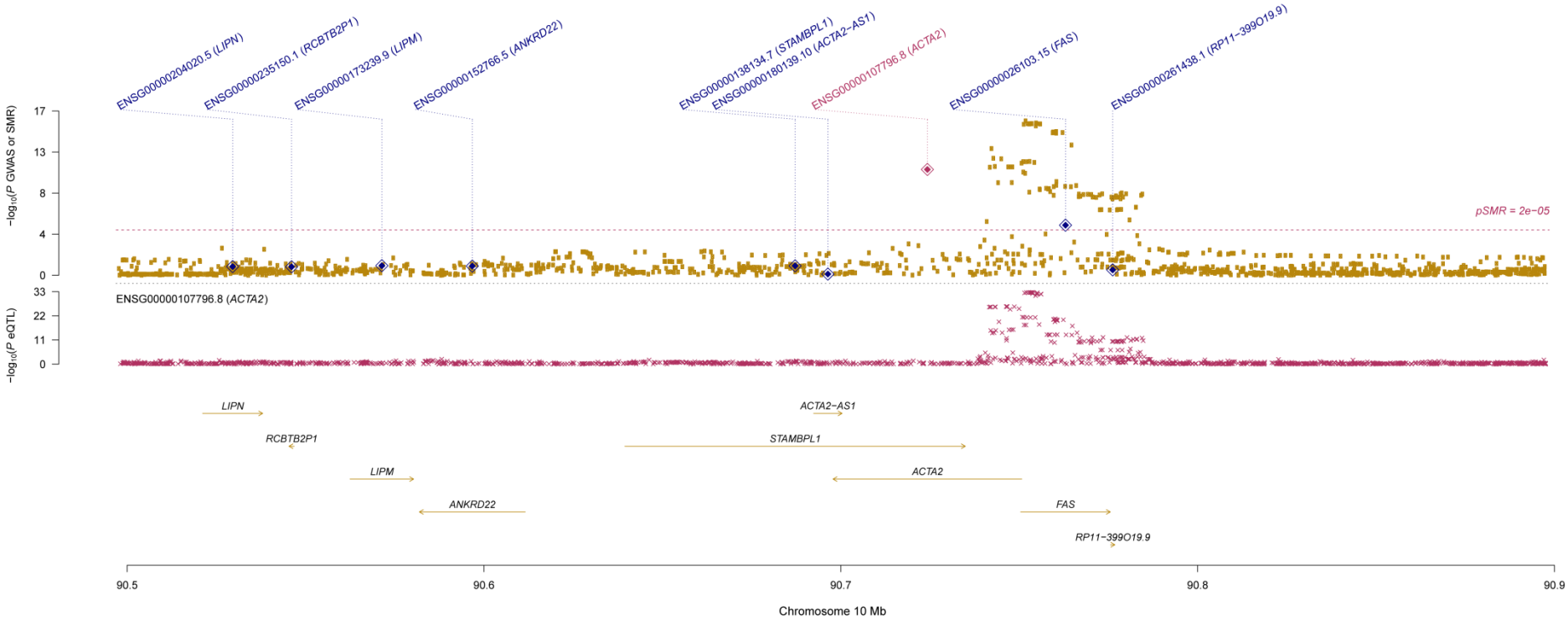
Appendix 17

				CLL		MM	
Locus	**Pos**	**SNP**	**Gene**	P_{SMR}	P_{HEIDI}	P_{SMR}	P_{HEIDI}
10q23.31	90,752,018	rs6586163	*ACTA2*	1×10^{-11}	0.2	3×10^{-3}	0.5
			FAS	6×10^{-6}	2×10^{-5}	2×10^{-3}	0.02
16q23.1	74,664,743	rs7193541	*RFWD3*	9×10^{-3}	0.4	1×10^{-6}	0.005
22q13.33	50,971,266	rs140522	*SCO2*	1×10^{-4}	5×10^{-6}	3×10^{-4}	2×10^{-4}
			TYMP	7×10^{-5}	0.03	2×10^{-4}	0.2
			ODF3B	5×10^{-5}	0.1	0.01	0.3

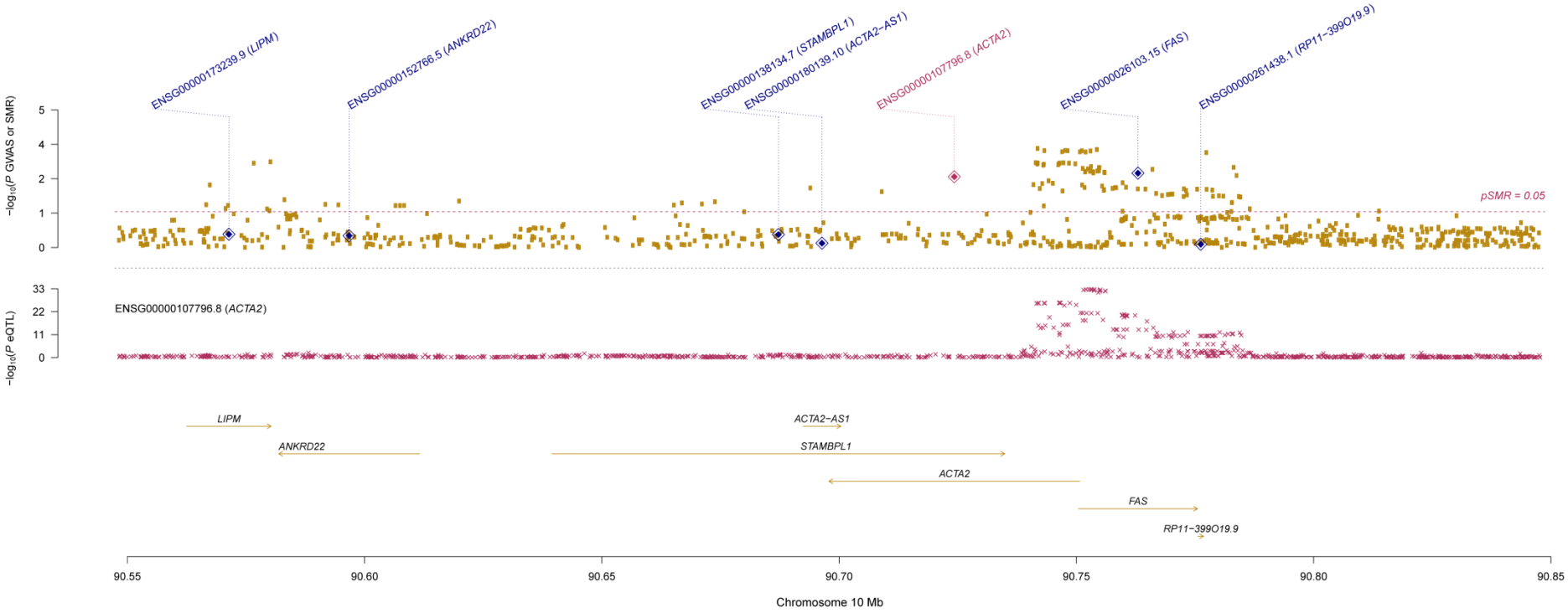
Summary of results from SMR analysis. A threshold for the SMR test of $P_{SMR} < 2.5\times10^{-5}$ was set corresponding to a Bonferroni correction for 2,000 probes. For all genes passing this threshold plots of the eQTL and GWAS associations at the locus were generated, as well as plots of GWAS and eQTL effect sizes (*i.e.* corresponding to input for the HEIDI heterogeneity test). HEIDI test *P*-values < 0.05 were considered as being reflective of heterogeneity. This threshold is conservative for gene discovery because it retains fewer genes than when correcting for multiple testing.

Appendix 18

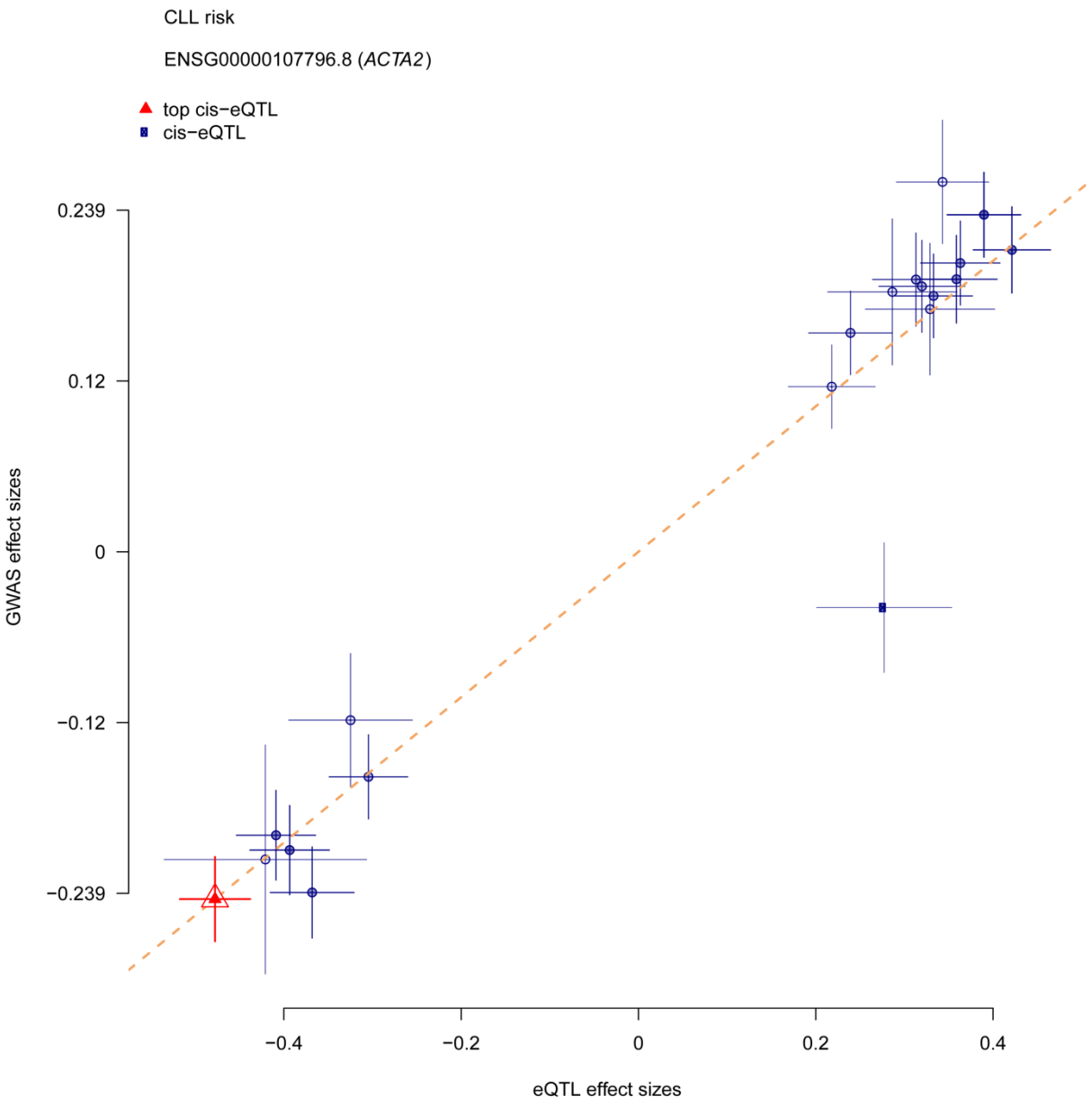
Summary data-based Mendelian Randomization (SMR) analysis plots (overleaf).



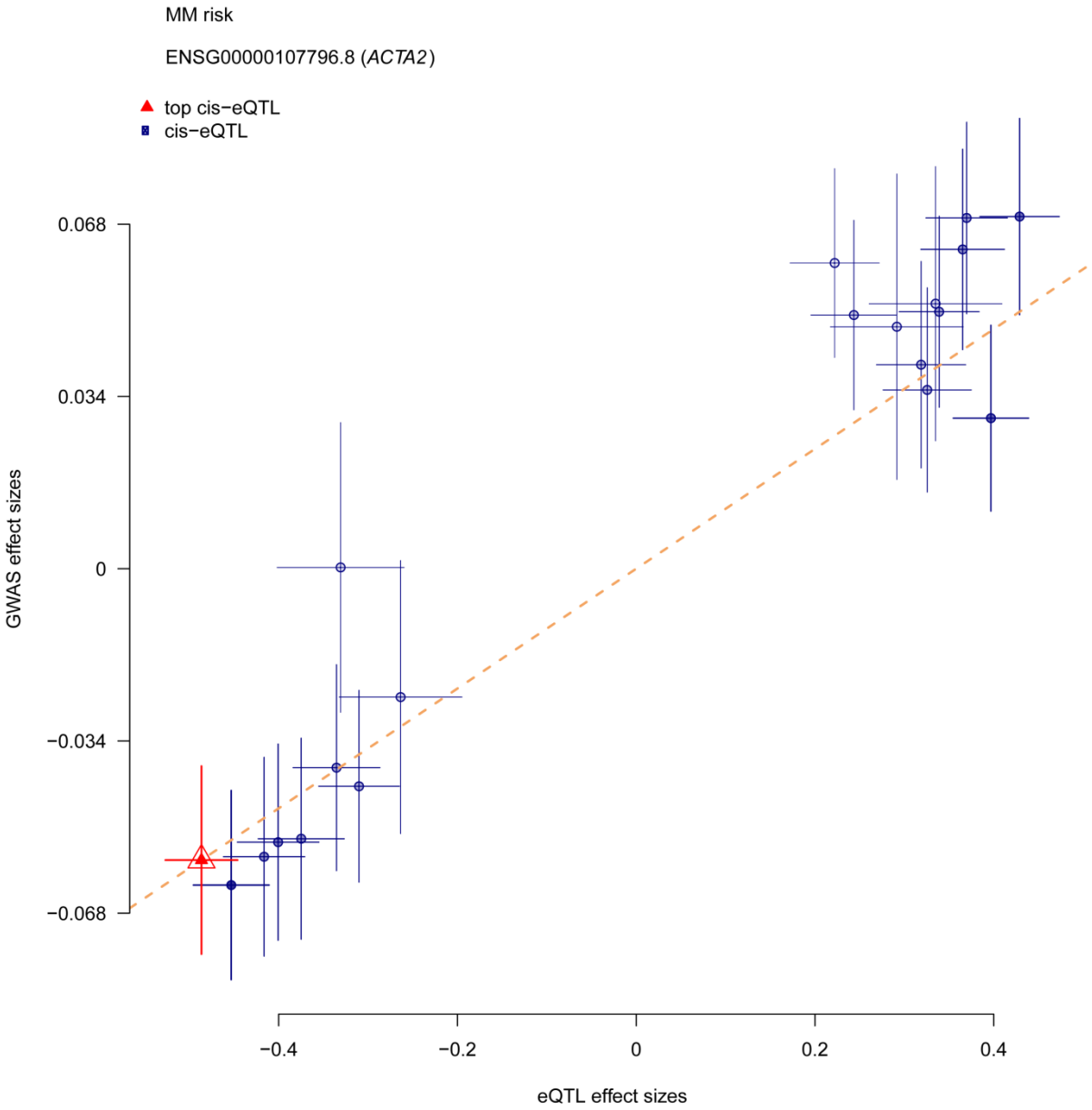
Summary data-based Mendelian Randomization (SMR) analysis locus plot for CLL. Upper panel - brown dots represent *P*-values for SNPs from the GWAS meta-analysis, diamonds represent *P*-values for probes from the SMR test; lower panel – crosses represent eQTL *P*-values of SNPs from whole blood with genes passing the SMR (*i.e.* $P_{SMR} < 0.001$) and HEIDI (*i.e.* $P_{HEIDI} > 0.05$) tests highlighted in red.



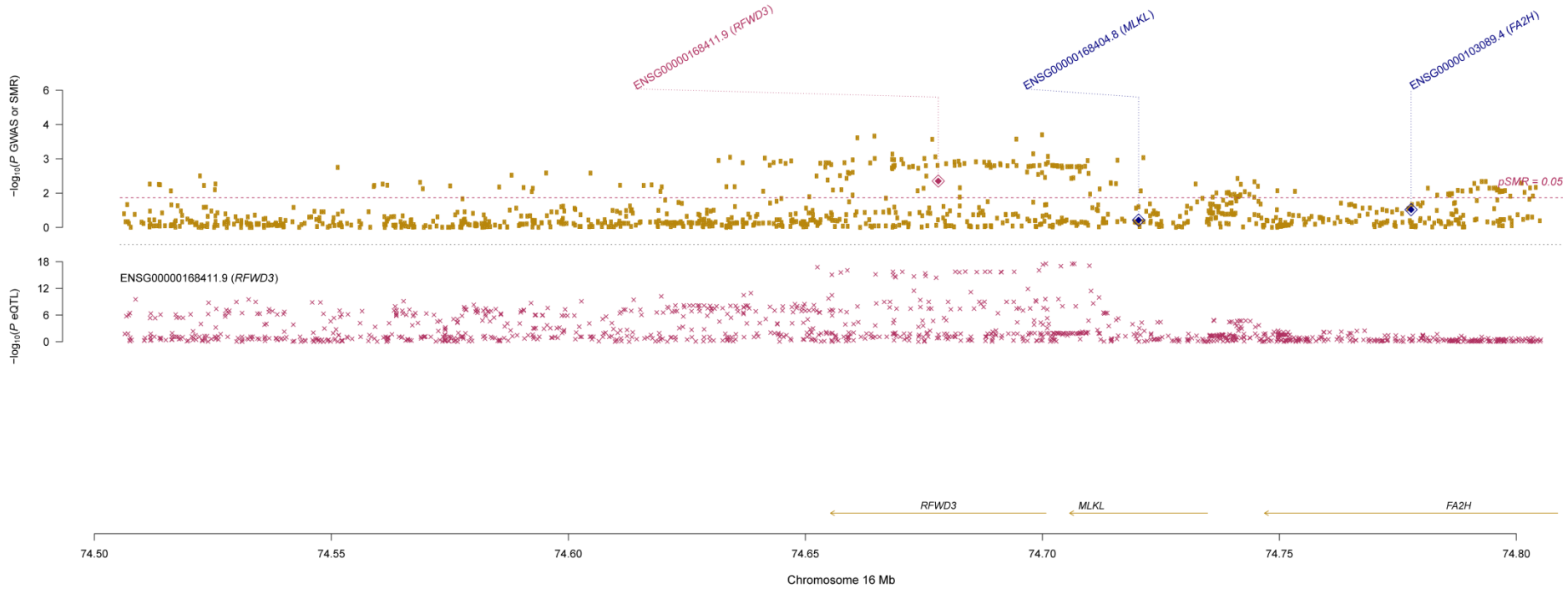
Summary data-based Mendelian Randomization (SMR) analysis locus plot for MM. Upper panel - brown dots represent *P*-values for SNPs from the GWAS meta-analysis, diamonds represent *P*-values for probes from the SMR test; lower panel – crosses represent eQTL *P*-values of SNPs from whole blood with genes passing the SMR (*i.e.* $P_{SMR} < 0.001$) and HEIDI (*i.e.* $P_{HEIDI} > 0.05$) tests highlighted in red.



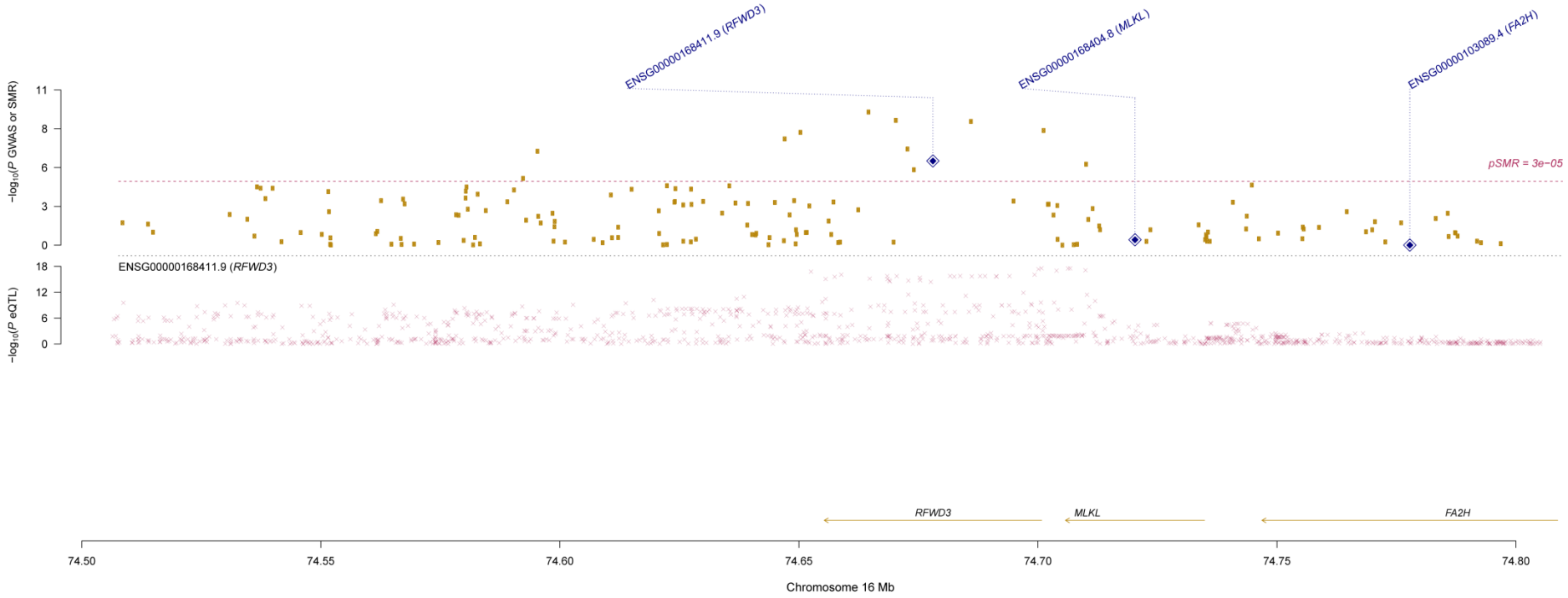
Summary data-based Mendelian Randomization analysis effect plot for CLL. Blue dots represent effect sizes of SNPs from the GWAS meta-analysis against those from the eQTL study of whole blood. The top *cis*-eQTL is highlighted by a red diamond. Error bars are the standard errors of the SNP effects. An estimate of b_{xy} at the top *cis*-eQTL is represented by the orange dotted line.



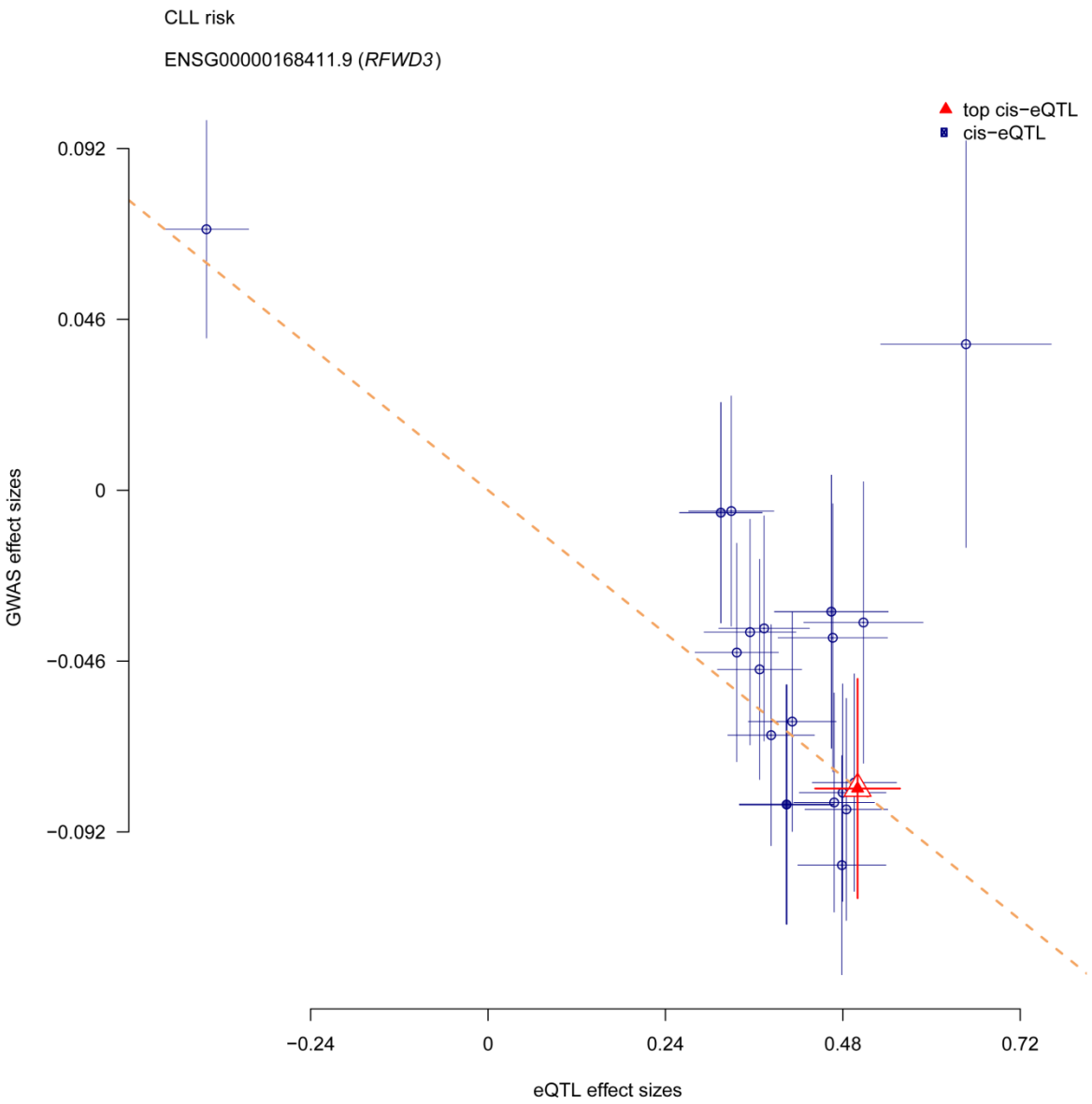
Summary data-based Mendelian Randomization analysis effect plot for MM. Blue dots represent effect sizes of SNPs from the GWAS meta-analysis against those from the eQTL study of whole blood. The top *cis*-eQTL is highlighted by a red diamond. Error bars are the standard errors of the SNP effects. An estimate of b_{xy} at the top *cis*-eQTL is represented by the orange dotted line.



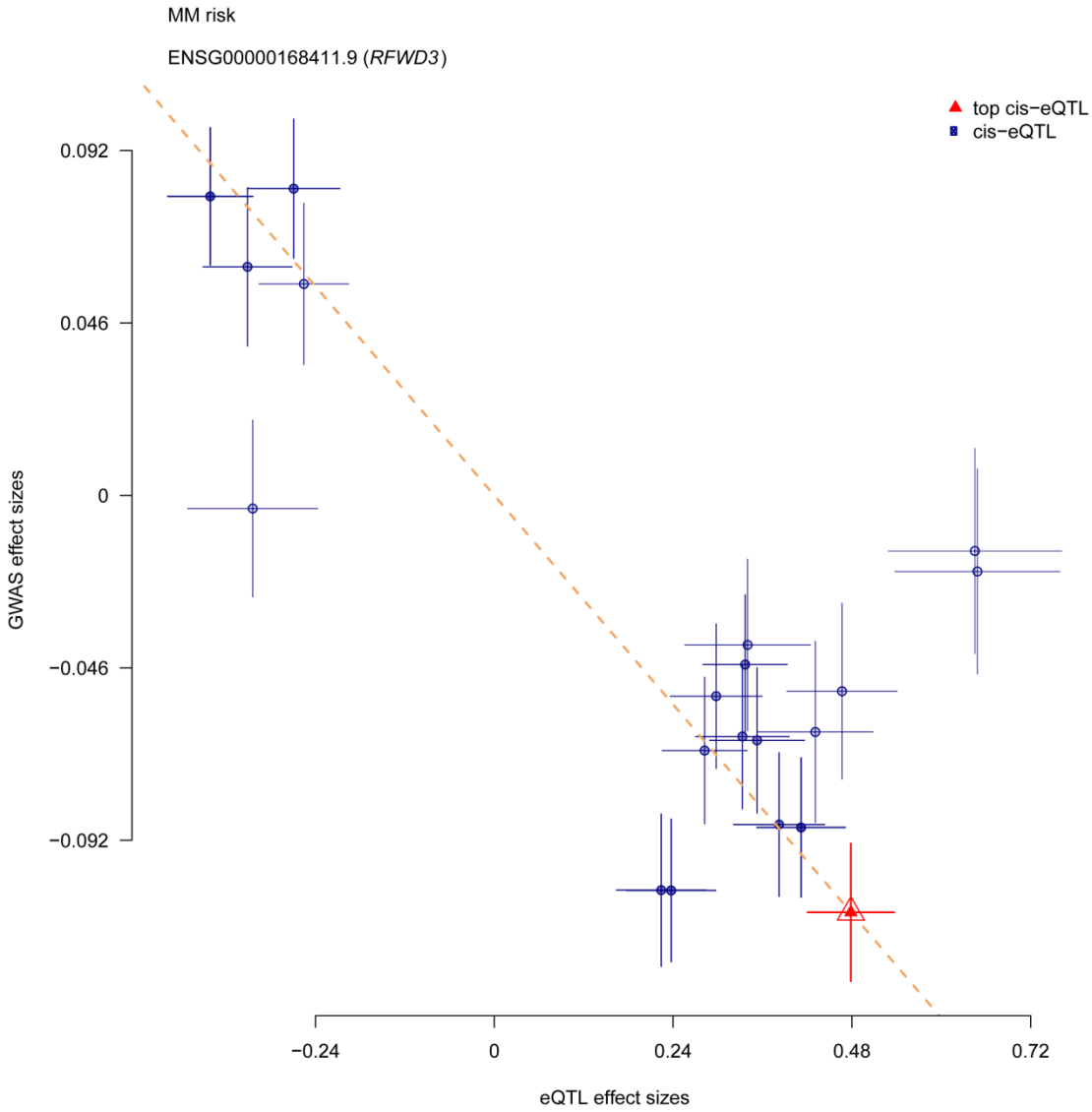
Summary data-based Mendelian Randomization (SMR) analysis locus plot for CLL. Upper panel - brown dots represent *P*-values for SNPs from the GWAS meta-analysis, diamonds represent *P*-values for probes from the SMR test; lower panel – crosses represent eQTL *P*-values of SNPs from whole blood with genes passing the SMR (*i.e.* $P_{SMR} < 0.001$) and HEIDI (*i.e.* $P_{HEIDI} > 0.05$) tests highlighted in red.



Summary data-based Mendelian Randomization (SMR) analysis locus plot for MM. Upper panel - brown dots represent *P*-values for SNPs from the GWAS meta-analysis, diamonds represent *P*-values for probes from the SMR test; lower panel – crosses represent eQTL *P*-values of SNPs from whole blood with genes passing the SMR (*i.e*. $P_{SMR} < 0.001$) and HEIDI (*i.e*. $P_{HEIDI} > 0.05$) tests highlighted in red.



Summary data-based Mendelian Randomization analysis effect plot for CLL. Blue dots represent effect sizes of SNPs from the GWAS meta-analysis against those from the eQTL study of whole blood. The top *cis*-eQTL is highlighted by a red diamond. Error bars are the standard errors of the SNP effects. An estimate of b_{xy} at the top *cis*-eQTL is represented by the orange dotted line.



Summary data-based Mendelian Randomization analysis effect plot for MM. Blue dots represent effect sizes of SNPs from the GWAS meta-analysis against those from the eQTL study of whole blood. The top *cis*-eQTL is highlighted by a red diamond. Error bars are the standard errors of the SNP effects. An estimate of b_{xy} at the top *cis*-eQTL is represented by the orange dotted line.

Appendix 19

Modifiable risk factors considered in MR analysis (tables overleaf).
*Indicates power estimated using MM summary statistics from females only. HDL, high density lipoprotein; LDL, low density lipoprotein; LOY, loss of Y chromosome

Trait	PubMed ID/ DOI	No. SNPs	PVE by SNPs	Power to identify OR_{SD} of 0.91 or 1.10	0.83 or 1.20	0.75 or 1.33	0.67 or 1.50	F-statistic
Plasma IL-6 sRa	29875488	1	0.6043	1.00	1.00	1.00	1.00	5039
Arachidonic acid (20:4n6)	24823311	2	0.4579	1.00	1.00	1.00	1.00	3643
Height	30124842	2481	0.3798	1.00	1.00	1.00	1.00	174
Interleukin-6 receptor	10.1038/srep18092	1	0.3700	1.00	1.00	1.00	1.00	479
Angiotensin-converting enzyme	10.1038/srep18092	2	0.2282	0.96	1.00	1.00	1.00	120
Apolipoprotein A-IV	10.1038/srep18092	1	0.2274	0.95	1.00	1.00	1.00	240
E-selectin	10.1038/srep18092	1	0.2221	0.95	1.00	1.00	1.00	233
Blood butyrylcarnitine	24816252	9	0.2128	0.94	1.00	1.00	1.00	220
Fetuin-A	10.1038/srep18092	1	0.1923	0.92	1.00	1.00	1.00	194
Tamm-Horsfall urinary glycoprotein	10.1038/srep18092	1	0.1850	0.91	1.00	1.00	1.00	185
Blood N-acetylornithine	24816252	4	0.1837	0.91	1.00	1.00	1.00	402
Blood glycoproteins	27005778	28	0.1790	0.90	1.00	1.00	1.00	146
Apolipoprotein H	10.1038/srep18092	3	0.1656	0.88	1.00	1.00	1.00	54
Heel bone mineral density (BMD)	10.1038/s41586-018-0579-z	409	0.1653	0.88	1.00	1.00	1.00	94
Factor VII	10.1038/srep18092	2	0.1578	0.86	1.00	1.00	1.00	76
Angiotensinogen	10.1038/srep18092	1	0.1533	0.85	1.00	1.00	1.00	148
Interleukin-16	10.1038/srep18092	1	0.1447	0.83	1.00	1.00	1.00	138
Circulating fetuin-A	28379451	1	0.1433	0.83	1.00	1.00	1.00	1332
Blood carnitine	24816252	18	0.1391	0.82	1.00	1.00	1.00	66
Dihomo-gamma-linoleic acid (20:3n6)	24823311	2	0.1381	0.81	1.00	1.00	1.00	691
Linoleic acid (18:2n6)	24823311	2	0.1340	0.80	1.00	1.00	1.00	668
Whole body water mass	10.1038/s41586-018-0579-z	735	0.1316	0.79	1.00	1.00	1.00	68
Chemokine CC-4	10.1038/srep18092	1	0.1268	0.78	1.00	1.00	1.00	119

Trait	PubMed ID/ DOI	No. SNPs	PVE by SNPs	Power to identify OR_{SD} of				F-statistic
				0.91 or 1.10	0.83 or 1.20	0.75 or 1.33	0.67 or 1.50	
Blood biliverdin	24816252	3	0.1228	0.77	1.00	1.00	1.00	291
Basal metabolic rate	10.1038/s41586-018-0579-z	693	0.1215	0.76	1.00	1.00	1.00	66
Apolipoprotein E	10.1038/srep18092	1	0.1209	0.76	1.00	1.00	1.00	112
Carcinoembryonic antigen	10.1038/srep18092	2	0.1159	0.74	1.00	1.00	1.00	53
Dihomo-gamma-linolenic acid (20:3n6)	24823311	2	0.1150	0.74	1.00	1.00	1.00	560
Blood glycine	27005778	6	0.1110	0.72	1.00	1.00	1.00	390
Blood estrone 3-sulfate	24816252	1	0.1089	0.72	1.00	1.00	1.00	31
Cancer antigen 19-9	10.1038/srep18092	1	0.1083	0.71	1.00	1.00	1.00	99
Blood succinylcarnitine	24816252	7	0.1059	0.70	1.00	1.00	1.00	111
Blood glutaroyl carnitine	24816252	9	0.1049	0.70	1.00	1.00	1.00	94
Blood bilirubin (Z,Z)	24816252	2	0.1037	0.69	1.00	1.00	1.00	370
LDL	24097068	102	0.1031	0.69	1.00	1.00	1.00	177
Blood 2-aminooctanoic acid	24816252	3	0.1029	0.69	1.00	1.00	1.00	272
Myeloid progenitor inhibitory factor 1	10.1038/srep18092	2	0.1023	0.69	1.00	1.00	1.00	46
Docosapentaenoic acid (22:5n3)	21829377	3	0.0993	0.68	1.00	1.00	1.00	326
Total cholesterol	24097068	123	0.0973	0.67	1.00	1.00	1.00	164
CD 40 antigen	10.1038/srep18092	1	0.0963	0.66	1.00	1.00	1.00	87
Blood apolipoprotein B	27005778	27	0.0932	0.65	1.00	1.00	1.00	79
Weight	10.1038/s41586-018-0579-z	576	0.0914	0.64	1.00	1.00	1.00	59
Blood tryptophan	24816252	19	0.0906	0.64	0.99	1.00	1.00	38
Impedance of whole body	10.1038/s41586-018-0579-z	564	0.0902	0.63	0.99	1.00	1.00	58
Blood bradykinin, des-arg(9)	24816252	3	0.0892	0.63	0.99	1.00	1.00	96
Blood androsterone sulfate	24816252	4	0.0875	0.62	0.99	1.00	1.00	176
HDL	24097068	124	0.0833	0.60	0.99	1.00	1.00	131

Trait	PubMed ID/ DOI	No. SNPs	PVE by SNPs	Power to identify OR_{SD} of 0.91 or 1.10	0.83 or 1.20	0.75 or 1.33	0.67 or 1.50	F-statistic
Body mass index	30124842	964	0.0786	0.57	0.99	1.00	1.00	70
Blood hexanoylcarnitine	24816252	4	0.0756	0.56	0.99	1.00	1.00	150
Blood proline	24816252	4	0.0727	0.54	0.98	1.00	1.00	144
Blood tetradecanedioate	24816252	3	0.0716	0.54	0.98	1.00	1.00	145
Macrophage inflammatory protein-1 alpha	10.1038/srep18092	1	0.0711	0.53	0.98	1.00	1.00	62
Blood alpha-glutamyltyrosine	24816252	3	0.0705	0.53	0.98	1.00	1.00	38
Adrenic acid (22:4n6)	24823311	1	0.0685	0.52	0.98	1.00	1.00	635
Blood octanoylcarnitine	24816252	3	0.0685	0.52	0.98	1.00	1.00	180
Serotransferrin	10.1038/srep18092	1	0.0679	0.51	0.98	1.00	1.00	59
Blood HDL diameter	27005778	10	0.0675	0.51	0.97	1.00	1.00	116
Haptoglobin	10.1038/srep18092	1	0.0635	0.49	0.97	1.00	1.00	55
Whole body fat mass	10.1038/s41586-018-0579-z	415	0.0632	0.49	0.97	1.00	1.00	54
Triglycerides	24097068	70	0.0619	0.48	0.96	1.00	1.00	166
Blood 5-oxoproline	24816252	1	0.0616	0.48	0.96	1.00	1.00	482
Fasting proinsulin	20081858	8	0.0613	0.48	0.96	1.00	1.00	87
Total triglycerides	24097068	34	0.0612	0.47	0.96	1.00	1.00	180
Monocyte chemotactic protein 2	10.1038/srep18092	1	0.0607	0.47	0.96	1.00	1.00	53
Blood urate	24816252	2	0.0582	0.46	0.95	1.00	1.00	228
Blood decanoylcarnitine	24816252	4	0.0569	0.45	0.95	1.00	1.00	110
Blood 5-alpha-pregnan-3beta,20alpha-disulfate	24816252	4	0.0567	0.45	0.95	1.00	1.00	38
Blood leucine	24816252	11	0.0565	0.44	0.95	1.00	1.00	40
Forced vital capacity (FVC)	10.1038/s41586-018-0579-z	284	0.0557	0.44	0.94	1.00	1.00	53
Blood hexadecanedioate	24816252	3	0.0555	0.44	0.94	1.00	1.00	126
Omega-6 fatty acids	27005778	13	0.0553	0.44	0.94	1.00	1.00	61

Trait	PubMed ID/ DOI	No. SNPs	PVE by SNPs	Power to identify OR_{SD} of				F-statistic
				0.91 or 1.10	0.83 or 1.20	0.75 or 1.33	0.67 or 1.50	
Alpha-1-antitrypsin	10.1038/srep18092	1	0.0552	0.44	0.94	1.00	1.00	48
Blood N-(2-furoyl)glycine	24816252	1	0.0552	0.44	0.94	1.00	1.00	30
Blood phenylalanylserine	24816252	2	0.0540	0.43	0.94	1.00	1.00	69
Blood 10-undecenoate (11:1n1)	24816252	3	0.0539	0.43	0.94	1.00	1.00	140
Blood isovalerylcarnitine	24816252	4	0.0530	0.42	0.93	1.00	1.00	103
Body fat percentage	10.1038/s41586-018-0579-z	365	0.0528	0.42	0.93	1.00	1.00	50
Blood epiandrosterone sulfate	24816252	2	0.0527	0.42	0.93	1.00	1.00	204
Blood apolipoprotein A-I	27005778	12	0.0515	0.41	0.93	1.00	1.00	83
Age at menopause*	10.1038/s41586-018-0579-z	48	0.0508	0.15	0.45	0.84	0.99	77
Tenascin-C	10.1038/srep18092	1	0.0500	0.40	0.92	1.00	1.00	43
Glutathione S-transferase alpha	10.1038/srep18092	1	0.0491	0.40	0.92	1.00	1.00	42
Matrix metalloproteinase-7	10.1038/srep18092	1	0.0489	0.40	0.91	1.00	1.00	42
Trunk fat percentage	10.1038/s41586-018-0579-z	334	0.0489	0.40	0.91	1.00	1.00	51
Blood isobutyrylcarnitine	24816252	3	0.0483	0.39	0.91	1.00	1.00	125
Gamma-linolenic acid (18:3n6)	24823311	2	0.0483	0.39	0.91	1.00	1.00	219
Blood N-[3-(2-Oxopyrrolidin-1-yl)propyl]acetamide	24816252	5	0.0482	0.39	0.91	1.00	1.00	75
Serum vitamin B12	23754956	9	0.0474	0.39	0.91	1.00	1.00	252
Age at menarche*	10.1038/s41586-018-0579-z	73	0.0322	0.11	0.31	0.65	0.94	83
Blood leucylalanine	24816252	2	0.0472	0.38	0.90	1.00	1.00	67
Blood phosphatidylcholine and other cholines	27005778	10	0.0471	0.38	0.90	1.00	1.00	67
Blood kynurenine	24816252	4	0.0468	0.38	0.90	1.00	1.00	90
Waist circumference	10.1038/s41586-018-0579-z	316	0.0464	0.38	0.90	1.00	1.00	52
Neuronal cell adhesion molecule	10.1038/srep18092	1	0.0463	0.38	0.90	1.00	1.00	40
Blood propionylcarnitine	24816252	5	0.0462	0.38	0.90	1.00	1.00	71

Trait	PubMed ID/ DOI	No. SNPs	PVE by SNPs	Power to identify OR_{SD} of				F-statistic
				0.91 or 1.10	0.83 or 1.20	0.75 or 1.33	0.67 or 1.50	
Total fatty acids	27005778	12	0.0462	0.38	0.90	1.00	1.00	54
Blood N-acetylglycine	24816252	3	0.0462	0.38	0.90	1.00	1.00	108
Fibroblast growth factor 4	10.1038/srep18092	1	0.0462	0.38	0.90	1.00	1.00	40
Blood copper	23720494	2	0.0460	0.38	0.90	1.00	1.00	63
Total phosphoglycerides	27005778	10	0.0459	0.38	0.90	1.00	1.00	65
Blood zinc	23720494	2	0.0459	0.38	0.90	1.00	1.00	63
Blood *cis*-4-decenoyl carnitine	24816252	2	0.0449	0.37	0.89	1.00	1.00	170
Eicosapentaenoic acid (20:5n3)	21829377	5	0.0448	0.37	0.89	1.00	1.00	83
CD5	10.1038/srep18092	1	0.0436	0.36	0.88	1.00	1.00	37
Blood mannose	24816252	1	0.0436	0.36	0.88	1.00	1.00	334
VLDL diameter	27005778	11	0.0432	0.36	0.88	1.00	1.00	79
Blood 4-acetamidobutanoate	24816252	2	0.0424	0.35	0.87	1.00	1.00	145
Sortilin	10.1038/srep18092	1	0.0419	0.35	0.87	1.00	1.00	36
Blood 5alpha-androstan-3beta,17beta-diol disulfate	24816252	4	0.0415	0.35	0.87	1.00	1.00	75
Blood betaine	24816252	5	0.0402	0.34	0.85	1.00	1.00	62
B lymphocyte chemoattractant	10.1038/srep18092	1	0.0396	0.33	0.85	1.00	1.00	34
Trefoil factor 3	10.1038/srep18092	1	0.0390	0.33	0.84	1.00	1.00	33
Blood N-acetylcarnosine	24816252	3	0.0388	0.33	0.84	1.00	1.00	84
Leptin	10.1038/srep18092	1	0.0385	0.32	0.84	1.00	1.00	33
Epithelial-derived neutrophil-activating	10.1038/srep18092	1	0.0384	0.32	0.84	1.00	1.00	33
Macrophage inflammatory protein-1 beta	10.1038/srep18092	1	0.0382	0.32	0.84	1.00	1.00	32
Interleukin-13	10.1038/srep18092	1	0.0382	0.32	0.84	1.00	1.00	32
Cystatin-C	10.1038/srep18092	1	0.0381	0.32	0.84	1.00	1.00	32
Receptor for advanced glycosylation end	10.1038/srep18092	1	0.0375	0.32	0.83	1.00	1.00	32

Trait	PubMed ID/ DOI	No. SNPs	PVE by SNPs	Power to identify OR_{SD} of 0.91 or 1.10	0.83 or 1.20	0.75 or 1.33	0.67 or 1.50	F-statistic
Growth-regulated alpha protein	10.1038/srep18092	1	0.0369	0.31	0.82	1.00	1.00	31
Angiopoietin-2	10.1038/srep18092	1	0.0369	0.31	0.82	1.00	1.00	31
Thymus-expressed chemokine	10.1038/srep18092	1	0.0368	0.31	0.82	1.00	1.00	31
Blood sphingomyelins	27005778	9	0.0366	0.31	0.82	1.00	1.00	57
Blood asparagine	24816252	2	0.0365	0.31	0.82	1.00	1.00	105
Macrophage colony-stimulating factor 1	10.1038/srep18092	1	0.0364	0.31	0.82	1.00	1.00	31
Interleukin-18	10.1038/srep18092	1	0.0361	0.31	0.82	1.00	1.00	31
Circulating C-reactive protein	21300955	14	0.0359	0.31	0.81	1.00	1.00	220
Fasting glucose	22581228	23	0.0358	0.31	0.81	1.00	1.00	94
Thrombopoietin	10.1038/srep18092	1	0.0356	0.30	0.81	1.00	1.00	30
Blood 12-hydroxyeicosatetraenoate (12-HETE)	24816252	1	0.0356	0.30	0.81	1.00	1.00	100
Blood serine	24816252	3	0.0354	0.30	0.81	1.00	1.00	90
Vascular cell adhesion molecule-1	10.1038/srep18092	1	0.0353	0.30	0.81	1.00	1.00	30
Interleukin-8	10.1038/srep18092	1	0.0352	0.30	0.81	1.00	1.00	30
Blood 1,5-anhydroglucitol (1,5-AG)	24816252	3	0.0352	0.30	0.81	1.00	1.00	89
Plasma progesterone	26014426	2	0.0348	0.30	0.80	1.00	1.00	52
Plasma progesterone*	26014426	2	0.0348	0.12	0.33	0.69	0.95	52
Platelet count	22139419	39	0.0346	0.30	0.80	0.99	1.00	61
Alpha-linolenic acid (18:3n3)	21829377	1	0.0340	0.29	0.79	0.99	1.00	312
Gamma-linoleic acid (18:3n6)	24823311	2	0.0331	0.29	0.78	0.99	1.00	148
Pulse rate	10.1038/s41586-018-0579-z	59	0.0327	0.28	0.78	0.99	1.00	63
Blood citrate	24816252	6	0.0307	0.27	0.75	0.99	1.00	39
Omega-9 and saturated fatty acids	27005778	7	0.0301	0.26	0.74	0.99	1.00	60
Glycoprotein acetyls	27005778	10	0.0298	0.26	0.74	0.99	1.00	59

Trait	PubMed ID/ DOI	No. SNPs	PVE by SNPs	Power to identify OR_{SD} of				F-statistic
				0.91 or 1.10	0.83 or 1.20	0.75 or 1.33	0.67 or 1.50	
Mono-unsaturated fatty acids	27005778	7	0.0286	0.25	0.72	0.98	1.00	57
Blood alpha-hydroxyisovalerate	24816252	3	0.0285	0.25	0.72	0.98	1.00	71
Circulating carotenoids	19185284	1	0.0277	0.25	0.71	0.98	1.00	106
Omega-3 fatty acids	27005778	6	0.0273	0.24	0.70	0.98	1.00	63
Circulating 25-hydroxyvitamin D	29343764	5	0.0265	0.24	0.69	0.98	1.00	431
Blood alpha-ketoglutarate	24816252	1	0.0263	0.24	0.68	0.98	1.00	30
Omega-7 and -9 and saturated fatty acids	27005778	6	0.0262	0.24	0.68	0.98	1.00	61
Monounsaturated fatty acids	27005778	6	0.0257	0.23	0.67	0.97	1.00	59
Serum calcium	24068962	7	0.0253	0.23	0.67	0.97	1.00	226
Birth weight	10.1038/s41586-018-0579-z	93	0.0247	0.23	0.66	0.97	1.00	53
Corrected insulin response	24699409	3	0.0246	0.22	0.65	0.97	1.00	45
Blood cysteine-glutathione disulfide	24816252	1	0.0243	0.22	0.65	0.97	1.00	49
Palmitoleic acid (16:1n7)	23362303	5	0.0236	0.22	0.64	0.96	1.00	43
Hip circumference	25673412	89	0.0235	0.22	0.64	0.96	1.00	39
LDL diameter	27005778	5	0.0234	0.22	0.63	0.96	1.00	92
Blood gamma-glutamyltyrosine	24816252	5	0.0227	0.21	0.62	0.96	1.00	33
Blood 2-hydroxyisobutyrate	24816252	3	0.0220	0.21	0.61	0.95	1.00	46
Stearic acid (18:0)	23362303	3	0.0220	0.21	0.61	0.95	1.00	67
Blood aspartylphenylalanine	24816252	1	0.0211	0.20	0.59	0.94	1.00	83
Blood selenium	23720494	1	0.0205	0.19	0.58	0.94	1.00	114
Blood acetylcarnitine	24816252	2	0.0200	0.19	0.57	0.93	1.00	75
Blood glutamine	27005778	6	0.0199	0.19	0.57	0.93	1.00	75
Blood hydroxyisovaleroyl carnitine	24816252	2	0.0197	0.19	0.56	0.93	1.00	53
Blood methylcysteine	24816252	1	0.0195	0.19	0.56	0.92	1.00	55

Trait	PubMed ID/ DOI	No. SNPs	PVE by SNPs	Power to identify OR_{SD} of				F-statistic
				0.91 or 1.10	0.83 or 1.20	0.75 or 1.33	0.67 or 1.50	
Blood gamma-glutamylglutamine	24816252	3	0.0194	0.19	0.55	0.92	1.00	48
Blood 3-methyl-2-oxovalerate	24816252	3	0.0193	0.19	0.55	0.92	1.00	48
Blood citrulline	24816252	4	0.0188	0.18	0.54	0.92	1.00	35
Blood inosine	24816252	1	0.0186	0.18	0.54	0.91	1.00	50
HbA1C levels	20858683	11	0.0183	0.18	0.53	0.91	1.00	79
Circulating adiponectin	22479202	10	0.0178	0.18	0.52	0.90	1.00	65
Blood dihomo-linolenate (20:3n3 or n6)	24816252	2	0.0178	0.18	0.52	0.90	1.00	67
Waist-to-hip ratio	25673412	35	0.0176	0.17	0.51	0.90	1.00	58
Blood tryptophan betaine	24816252	1	0.0175	0.17	0.51	0.90	1.00	125
Blood tyrosine	24816252	3	0.0175	0.17	0.51	0.90	1.00	44
Blood homocitrulline	24816252	1	0.0173	0.17	0.51	0.89	1.00	69
Blood uridine	24816252	3	0.0172	0.17	0.51	0.89	1.00	43
Blood octadecanedioate	24816252	2	0.0171	0.17	0.50	0.89	1.00	60
Fluid intelligence score	10.1038/s41586-018-0579-z	50	0.0170	0.17	0.50	0.89	1.00	38
Blood 1-palmitoylglycerophosphoethanolamine	24816252	2	0.0168	0.17	0.50	0.88	1.00	63
Putamen volume	25607358	4	0.0163	0.17	0.49	0.87	1.00	55
Blood stearate (18:0)	24816252	2	0.0162	0.16	0.48	0.87	1.00	61
Serum IgE	22075330	3	0.0161	0.16	0.48	0.87	1.00	80
Blood 1-linoleoylglycerol (1-monolinolein)	24816252	1	0.0159	0.16	0.48	0.87	1.00	45
Oleic acid (18:1n9)	23362303	1	0.0154	0.16	0.46	0.86	0.99	141
Blood phenyllactate (PLA)	24816252	1	0.0154	0.16	0.46	0.85	0.99	89
Blood O-sulfo-L-tyrosine	24816252	2	0.0152	0.16	0.46	0.85	0.99	57
Birth weight of first child	10.1038/s41586-018-0579-z	45	0.0150	0.16	0.45	0.85	0.99	49
Blood tetradecadienoate	24816252	1	0.0149	0.15	0.45	0.84	0.99	111

Trait	PubMed ID/ DOI	No. SNPs	PVE by SNPs	Power to identify OR_{SD} of				F-statistic
				0.91 or 1.10	0.83 or 1.20	0.75 or 1.33	0.67 or 1.50	
Blood 2-aminobutyrate	24816252	2	0.0148	0.15	0.45	0.84	0.99	55
Blood albumin	27005778	4	0.0146	0.15	0.44	0.84	0.99	70
Plasma IGF-I	29875488	1	0.0145	0.15	0.44	0.83	0.99	49
Blood dehydroisoandrosterone sulfate (DHEA-S)	24816252	2	0.0144	0.15	0.44	0.83	0.99	54
Blood alanine	27005778	6	0.0144	0.15	0.44	0.83	0.99	55
Morning/evening person (chronotype)	10.1038/s41586-018-0579-z	99	0.0143	0.15	0.44	0.83	0.99	44
Blood scyllo-inositol	24816252	1	0.0142	0.15	0.43	0.82	0.99	88
Blood valine	27005778	5	0.0142	0.15	0.43	0.82	0.99	71
Blood myo-inositol	24816252	2	0.0141	0.15	0.43	0.82	0.99	53
Serum vitamin B6	19303062	1	0.0141	0.15	0.43	0.82	0.99	27
Blood creatine	24816252	1	0.0140	0.15	0.43	0.82	0.99	105
Blood N2,N2-dimethylguanosine	24816252	2	0.0140	0.15	0.43	0.82	0.99	35
Blood histidine	27005778	5	0.0135	0.14	0.42	0.81	0.99	53
Insulin disposition index	24699409	1	0.0127	0.14	0.40	0.78	0.98	68
Telomere length	23535734	7	0.0126	0.14	0.39	0.78	0.98	69
Blood indolepropionate	24816252	1	0.0117	0.13	0.37	0.75	0.97	87
Iron status	25352340	3	0.0115	0.13	0.37	0.74	0.97	190
Blood levulinate (4-oxovalerate)	24816252	2	0.0113	0.13	0.36	0.74	0.97	37
Blood creatinine	27005778	6	0.0113	0.13	0.36	0.74	0.97	47
Blood glucose	27005778	3	0.0110	0.13	0.35	0.72	0.97	92
Neuroticism score	10.1038/s41586-018-0579-z	78	0.0108	0.13	0.35	0.71	0.96	38
Plasma estradiol	26014426	1	0.0107	0.12	0.34	0.71	0.96	31
Blood palmitoyl sphingomyelin	24816252	2	0.0106	0.12	0.34	0.71	0.96	39
Phenylalanine	27005778	4	0.0100	0.12	0.32	0.68	0.95	54

Trait	PubMed ID/ DOI	No. SNPs	PVE by SNPs	Power to identify OR_{SD} of 0.91 or 1.10	0.83 or 1.20	0.75 or 1.33	0.67 or 1.50	F-statistic
Indoleacetate	24816252	2	0.0098	0.12	0.32	0.67	0.95	36
Years of schooling	27225129	74	0.0098	0.12	0.32	0.67	0.95	39
4-methyl-2-oxopentanoate	24816252	2	0.0097	0.12	0.32	0.67	0.94	36
Lysine	24816252	1	0.0095	0.12	0.31	0.66	0.94	71
Hypoxanthine	24816252	1	0.0093	0.11	0.31	0.65	0.94	65
1,6-anhydroglucose	24816252	1	0.0090	0.11	0.30	0.63	0.93	31
Gamma-glutamylphenylalanine	24816252	2	0.0088	0.11	0.29	0.63	0.92	33
Acetylphosphate	24816252	2	0.0087	0.11	0.29	0.62	0.92	32
Heptanoate (7:0)	24816252	2	0.0087	0.11	0.29	0.62	0.92	32
Stearidonate (18:4n3)	24816252	1	0.0087	0.11	0.29	0.62	0.92	64
Myristoleate (14:1n5)	24816252	1	0.0086	0.11	0.29	0.62	0.92	64
Caprylate (8:0)	24816252	2	0.0084	0.11	0.28	0.60	0.91	31
Laurate (12:0)	24816252	2	0.0083	0.11	0.28	0.60	0.91	31
HOMA-B	20081858	4	0.0081	0.11	0.27	0.59	0.90	74
Margarate (17:0)	24816252	1	0.0080	0.11	0.27	0.58	0.90	59
Nap during day	10.1038/s41586-018-0579-z	58	0.0079	0.10	0.27	0.58	0.89	47
Palmitoleate (16:1n7)	24816252	1	0.0079	0.10	0.27	0.58	0.89	58
3-phenylpropionate (hydrocinnamate)	24816252	1	0.0077	0.10	0.26	0.57	0.89	45
Time spent watching television (TV)	10.1038/s41586-018-0579-z	65	0.0077	0.10	0.26	0.57	0.88	38
Laurylcarnitine	24816252	1	0.0076	0.10	0.26	0.56	0.88	38
3,4-dihydroxybutyrate	24816252	1	0.0072	0.10	0.25	0.54	0.86	48
Indolelactate	24816252	1	0.0072	0.10	0.25	0.54	0.86	50
2hr glucose	22885924	7	0.0071	0.10	0.24	0.54	0.86	44
Docosahexaenoic acid (22:6n3)	21829377	1	0.0071	0.10	0.24	0.53	0.86	63

Trait	PubMed ID/ DOI	No. SNPs	PVE by SNPs	Power to identify OR_{SD} of				F-statistic
				0.91 or 1.10	0.83 or 1.20	0.75 or 1.33	0.67 or 1.50	
3-(4-hydroxyphenyl)lactate	24816252	1	0.0070	0.10	0.24	0.53	0.85	52
Serum vitamin A1	21878437	2	0.0070	0.10	0.24	0.53	0.85	35
1-oleoylglycerol (1-monoolein)	24816252	1	0.0067	0.10	0.23	0.51	0.84	37
1-stearoylglycerophosphoethanolamine	24816252	1	0.0067	0.10	0.23	0.51	0.84	46
LOY	10.1038/s41586-019-1765-3	92	0.0066	0.10	0.23	0.51	0.83	15
N1-methyl-3-pyridone-4-carboxamide	24816252	1	0.0066	0.10	0.23	0.51	0.83	48
Serum vitamin E	21729881	3	0.0065	0.09	0.23	0.50	0.83	11
Fasting insulin	22885924	14	0.0065	0.09	0.23	0.50	0.83	51
Hippocampus volume	25607358	2	0.0062	0.09	0.22	0.48	0.81	41
Stearoylcarnitine	24816252	1	0.0062	0.09	0.22	0.48	0.81	42
Dodecanedioate	24816252	1	0.0061	0.09	0.22	0.48	0.80	32

Appendix 21

Causal estimates from each Mendelian randomization approach for each trait and MM risk (tables overleaf).
* Indicates P < 0.05, ** Causal effects estimated using MM summary statistics from females only, † indicates significant heterogeneity. HDL, high density lipoprotein; LDL, low density lipoprotein; LOY, loss of Y chromosome.

Category	Exposure	No. of SNPs	IVW-RE					Wald Ratio					I^2	
			P-value		Odds ratio (95% CIs)			Odds ratio (95% CIs)						
Developmental and growth factors	Height	2481	0.617		1.01	0.96	- 1.07	NA	NA	-	NA		21.71	
Developmental and growth factors	Putamen volume	4	0.928		0.99	0.74	- 1.31	NA	NA	-	NA		46.08	
Developmental and growth factors	Plasma IGF-I	1	0.687		NA	NA	- NA	0.94	0.71	-	1.25		NA	
Developmental and growth factors	Hippocampus volume	2	0.540		1.35	0.52	- 3.48	NA	NA	-	NA		84.22	†
Diet and lifestyle	Heel bone mineral density (BMD) T-score	409	0.578		0.98	0.91	- 1.06	NA	NA	-	NA		18.06	
Diet and lifestyle	Serum vitamin B12	9	0.701		1.04	0.85	- 1.27	NA	NA	-	NA		57.33	
Diet and lifestyle	Blood copper	2	0.935		0.99	0.80	- 1.22	NA	NA	-	NA		48.42	
Diet and lifestyle	Blood zinc	2	0.984		1.00	0.87	- 1.15	NA	NA	-	NA		0.00	
Diet and lifestyle	Fasting glucose	23	0.287		1.10	0.92	- 1.30	NA	NA	-	NA		31.25	
Diet and lifestyle	Pulse rate	59	0.920		0.99	0.82	- 1.20	NA	NA	-	NA		34.08	
Diet and lifestyle	Circulating 25-hydroxyvitamin D	5	0.541		1.08	0.84	- 1.40	NA	NA	-	NA		65.39	
Diet and lifestyle	Serum calcium	7	0.130		0.88	0.74	- 1.04	NA	NA	-	NA		0.00	
Diet and lifestyle	Blood selenium	1	0.373		NA	NA	- NA	0.91	0.75	-	1.12		NA	
Diet and lifestyle	Serum vitamin B6	1	0.04130	*	NA	NA	- NA	1.26	1.01	-	1.58		NA	
Diet and lifestyle	Iron status	3	0.232		1.24	0.87	- 1.78	NA	NA	-	NA		47.90	
Diet and lifestyle	Blood creatinine	6	0.977		1.01	0.70	- 1.44	NA	NA	-	NA		47.92	
Diet and lifestyle	Serum vitamin A1	2	0.306		1.18	0.86	- 1.61	NA	NA	-	NA		2.90	
Diet and lifestyle	Serum vitamin E	3	0.536		0.89	0.62	- 1.28	NA	NA	-	NA		0.00	
Diet and lifestyle	Fasting insulin	14	0.231		1.35	0.83	- 2.19	NA	NA	-	NA		52.12	

Category	Exposure	No. of SNPs	IVW-RE					Wald Ratio				I^2
			P-value		Odds ratio (95% CIs)			Odds ratio (95% CIs)				
Fatty acid profile and metabolism	3-phenylpropionate (hydrocinnamate)	1	0.01895	*	NA	NA	NA	1.54	1.07	-	2.20	NA
Fatty acid profile and metabolism	Blood butyrylcarnitine	9	0.887		1.01	0.93	1.08	NA	NA	-	NA	22.86
Fatty acid profile and metabolism	Blood N-acetylornithine	4	0.096		0.95	0.89	1.01	NA	NA	-	NA	0.00
Fatty acid profile and metabolism	Blood glycoproteins	28	0.517		0.98	0.91	1.05	NA	NA	-	NA	0.00
Fatty acid profile and metabolism	Adrenic acid (22:4n6)	1	0.00922	*	NA	NA	NA	0.88	0.79	-	0.97	NA
Fatty acid profile and metabolism	Alpha-linolenic acid (18:3n3)	1	0.01123	*	NA	NA	NA	1.20	1.04	-	1.38	NA
Fatty acid profile and metabolism	Blood biliverdin	3	0.531		0.98	0.90	1.05	NA	NA	-	NA	0.00
Fatty acid profile and metabolism	Dihomo-gamma-linolenic acid (20:3n6)	2	0.223		1.06	0.97	1.16	NA	NA	-	NA	0.00
Fatty acid profile and metabolism	Blood glycine	6	0.142		0.93	0.85	1.02	NA	NA	-	NA	0.00
Fatty acid profile and metabolism	Blood estrone 3-sulfate	1	0.359		NA	NA	NA	0.95	0.86	-	1.06	NA
Fatty acid profile and metabolism	Arachidonic acid (20:4n6)	2	0.01174	*	0.95	0.92	0.99	NA	NA	-	NA	0.00
Fatty acid profile and metabolism	Blood succinylcarnitine	7	0.412		0.96	0.87	1.06	NA	NA	-	NA	28.78
Fatty acid profile and metabolism	Blood glutaroyl carnitine	9	0.163		1.06	0.98	1.16	NA	NA	-	NA	0.00
Fatty acid profile and metabolism	Blood bilirubin (Z,Z)	2	0.641		0.98	0.90	1.07	NA	NA	-	NA	0.00
Fatty acid profile and metabolism	Blood 2-aminooctanoic acid	3	0.760		0.99	0.91	1.07	NA	NA	-	NA	0.00
Fatty acid profile and metabolism	Blood carnitine	18	0.00114	*	1.13	1.05	1.22	NA	NA	-	NA	1.32
Fatty acid profile and metabolism	Blood *cis*-4-decenoyl carnitine	2	0.03211	*	1.17	1.01	1.34	NA	NA	-	NA	0.00
Fatty acid profile and metabolism	Blood apolipoprotein B	27	0.095		0.93	0.84	1.01	NA	NA	-	NA	0.00
Fatty acid profile and metabolism	Blood tryptophan	19	0.111		0.93	0.85	1.02	NA	NA	-	NA	0.00
Fatty acid profile and metabolism	Blood bradykinin, des-arg(9)	3	0.265		0.93	0.82	1.06	NA	NA	-	NA	44.46
Fatty acid profile and metabolism	Blood androsterone sulfate	4	0.828		1.01	0.92	1.11	NA	NA	-	NA	0.00
Fatty acid profile and metabolism	Blood decanoylcarnitine	4	0.01291	*	1.16	1.03	1.31	NA	NA	-	NA	0.00
Fatty acid profile and metabolism	Blood proline	4	0.508		0.97	0.87	1.07	NA	NA	-	NA	0.00

Category	Exposure	No. of SNPs	IVW-RE					Wald Ratio					I^2
			P-value		Odds ratio (95% CIs)			Odds ratio (95% CIs)					
Fatty acid profile and metabolism	Blood tetradecanedioate	3	0.195		0.93	0.84	- 1.04	NA	NA	-	NA		6.61
Fatty acid profile and metabolism	Blood alpha-glutamyltyrosine	3	0.472		1.04	0.94	- 1.16	NA	NA	-	NA		0.00
Fatty acid profile and metabolism	Blood hexanoylcarnitine	4	0.00695	*	1.16	1.04	- 1.29	NA	NA	-	NA		0.00
Fatty acid profile and metabolism	Blood HDL diameter	10	0.302		0.92	0.79	- 1.08	NA	NA	-	NA		54.57
Fatty acid profile and metabolism	Blood 5-oxoproline	1	0.749		NA	NA	- NA	0.98	0.89	-	1.09		NA
Fatty acid profile and metabolism	Blood hydroxyisovaleroyl carnitine	2	0.04608	*	1.21	1.00	- 1.46	NA	NA	-	NA		0.00
Fatty acid profile and metabolism	Blood urate	2	0.835		0.98	0.82	- 1.17	NA	NA	-	NA		59.83
Fatty acid profile and metabolism	Blood indolepropionate	1	0.01514	*	NA	NA	- NA	1.44	1.07	-	1.92		NA
Fatty acid profile and metabolism	Blood 5-alpha-pregnan-3beta,20alpha-disulfate	4	0.760		0.98	0.86	- 1.12	NA	NA	-	NA		0.00
Fatty acid profile and metabolism	Blood leucine	11	0.314		1.06	0.95	- 1.19	NA	NA	-	NA		0.00
Fatty acid profile and metabolism	Blood isovalerylcarnitine	4	0.00252	*	1.21	1.07	- 1.37	NA	NA	-	NA		0.00
Fatty acid profile and metabolism	Blood hexadecanedioate	3	0.162		0.92	0.82	- 1.03	NA	NA	-	NA		0.00
Fatty acid profile and metabolism	Blood N-(2-furoyl)glycine	1	0.065		NA	NA	- NA	1.12	0.99	-	1.25		NA
Fatty acid profile and metabolism	Blood phenylalanylserine	2	0.860		0.99	0.88	- 1.11	NA	NA	-	NA		0.00
Fatty acid profile and metabolism	Blood 10-undecenoate (11:1n1)	3	0.607		0.97	0.85	- 1.10	NA	NA	-	NA		0.00
Fatty acid profile and metabolism	Blood octanoylcarnitine	3	0.01384	*	1.16	1.03	- 1.30	NA	NA	-	NA		5.70
Fatty acid profile and metabolism	Blood epiandrosterone sulfate	2	0.613		1.03	0.92	- 1.16	NA	NA	-	NA		0.00
Fatty acid profile and metabolism	Blood palmitoyl sphingomyelin	2	0.02229	*	0.76	0.61	- 0.96	NA	NA	-	NA		0.00
Fatty acid profile and metabolism	Blood isobutyrylcarnitine	3	0.899		1.01	0.89	- 1.14	NA	NA	-	NA		0.00
Fatty acid profile and metabolism	Blood propionylcarnitine	5	0.01273	*	1.18	1.04	- 1.35	NA	NA	-	NA		11.55
Fatty acid profile and metabolism	Blood N-[3-(2-Oxopyrrolidin-1-yl)propyl]acetamide	5	0.542		0.96	0.86	- 1.09	NA	NA	-	NA		0.00
Fatty acid profile and metabolism	Blood leucylalanine	2	0.511		0.96	0.85	- 1.08	NA	NA	-	NA		0.00
Fatty acid profile and metabolism	Blood phosphatidylcholine and other cholines	10	0.506		0.96	0.84	- 1.09	NA	NA	-	NA		0.00

Category	Exposure	No. of SNPs	IVW-RE					Wald Ratio				I^2	
			P-value		Odds ratio (95% CIs)			Odds ratio (95% CIs)					
Fatty acid profile and metabolism	Blood kynurenine	4	0.688		1.03	0.91	- 1.16	NA	NA	-	NA	0.00	
Fatty acid profile and metabolism	Blood scyllo-inositol	1	0.02421	*	NA	NA	- NA	1.35	1.04	-	1.76	NA	
Fatty acid profile and metabolism	Blood N-acetylglycine	3	0.625		0.95	0.79	- 1.15	NA	NA	-	NA	46.75	
Fatty acid profile and metabolism	Total phosphoglycerides	10	0.183		0.91	0.80	- 1.04	NA	NA	-	NA	0.00	
Fatty acid profile and metabolism	Eicosapentaenoic acid (20:5n3)	5	0.076		0.89	0.78	- 1.01	NA	NA	-	NA	0.00	
Fatty acid profile and metabolism	Blood mannose	1	0.055		NA	NA	- NA	1.13	1.00	-	1.29	NA	
Fatty acid profile and metabolism	VLDL diameter	11	0.533		1.06	0.88	- 1.28	NA	NA	-	NA	50.80	
Fatty acid profile and metabolism	Blood 4-acetamidobutanoate	2	0.621		1.05	0.86	- 1.29	NA	NA	-	NA	60.30	
Fatty acid profile and metabolism	Blood 5alpha-androstan-3beta,17beta-diol disulfate	4	0.650		0.97	0.85	- 1.11	NA	NA	-	NA	0.00	
Fatty acid profile and metabolism	Blood betaine	5	0.382		1.10	0.89	- 1.34	NA	NA	-	NA	50.22	
Fatty acid profile and metabolism	Blood N-acetylcarnosine	3	0.709		1.05	0.83	- 1.32	NA	NA	-	NA	63.64	
Fatty acid profile and metabolism	Blood sphingomyelins	9	0.129		0.89	0.76	- 1.04	NA	NA	-	NA	0.00	
Fatty acid profile and metabolism	Blood asparagine	2	0.610		0.95	0.78	- 1.15	NA	NA	-	NA	47.10	
Fatty acid profile and metabolism	Blood 12-hydroxyeicosatetraenoate (12-HETE)	1	0.152		NA	NA	- NA	0.90	0.78	-	1.04	NA	
Fatty acid profile and metabolism	Blood serine	3	0.335		0.93	0.79	- 1.08	NA	NA	-	NA	0.00	
Fatty acid profile and metabolism	Blood 1,5-anhydroglucitol (1,5-AG)	3	0.484		1.09	0.85	- 1.41	NA	NA	-	NA	59.71	
Fatty acid profile and metabolism	Dihomo-gamma-linoleic acid (20:3n6)	2	0.01561	*	1.09	1.02	- 1.17	NA	NA	-	NA	0.00	
Fatty acid profile and metabolism	Blood citrate	6	0.328		1.08	0.92	- 1.27	NA	NA	-	NA	0.00	
Fatty acid profile and metabolism	Omega-9 and saturated fatty acids	7	0.186		0.90	0.77	- 1.05	NA	NA	-	NA	0.00	
Fatty acid profile and metabolism	Glycoprotein acetyls	10	0.108		0.86	0.72	- 1.03	NA	NA	-	NA	0.00	
Fatty acid profile and metabolism	Mono-unsaturated fatty acids	7	0.092		0.87	0.74	- 1.02	NA	NA	-	NA	0.00	
Fatty acid profile and metabolism	Blood alpha-hydroxyisovalerate	3	0.978		1.00	0.72	- 1.38	NA	NA	-	NA	75.83	†
Fatty acid profile and metabolism	Docosapentaenoic acid (22:5n3)	3	0.03732	*	0.90	0.81	- 0.99	NA	NA	-	NA	32.62	

Category	Exposure	No. of SNPs	IVW-RE					Wald Ratio					I^2
			P-value	Odds ratio (95% CIs)				Odds ratio (95% CIs)					
Fatty acid profile and metabolism	Blood alpha-ketoglutarate	1	0.154	NA	NA	-	NA	0.88	0.74	-	1.05	NA
Fatty acid profile and metabolism	Omega-7 and -9 and saturated fatty acids	6	0.148	0.88	0.75	-	1.05	NA	NA	-	NA	0.00
Fatty acid profile and metabolism	Monounsaturated fatty acids	6	0.071	0.85	0.72	-	1.01	NA	NA	-	NA	0.00
Fatty acid profile and metabolism	Blood cysteine-glutathione disulfide	1	0.226	NA	NA	-	NA	1.11	0.94	-	1.32	NA
Fatty acid profile and metabolism	Palmitoleic acid (16:1n7)	5	0.673	1.07	0.77	-	1.50	NA	NA	-	NA	71.56
Fatty acid profile and metabolism	LDL diameter	5	0.473	0.91	0.70	-	1.18	NA	NA	-	NA	53.00
Fatty acid profile and metabolism	Blood gamma-glutamyltyrosine	5	0.999	1.00	0.75	-	1.33	NA	NA	-	NA	54.46
Fatty acid profile and metabolism	Blood 2-hydroxyisobutyrate	3	0.747	0.95	0.72	-	1.27	NA	NA	-	NA	55.02
Fatty acid profile and metabolism	Stearic acid (18:0)	3	0.567	0.90	0.63	-	1.28	NA	NA	-	NA	74.02
Fatty acid profile and metabolism	Blood aspartylphenylalanine	1	0.899	NA	NA	-	NA	0.99	0.82	-	1.19	NA
Fatty acid profile and metabolism	Blood acetylcarnitine	2	0.071	1.20	0.98	-	1.46	NA	NA	-	NA	4.95
Fatty acid profile and metabolism	Blood glutamine	6	0.555	0.91	0.67	-	1.24	NA	NA	-	NA	62.51
Fatty acid profile and metabolism	Blood methylcysteine	1	0.182	NA	NA	-	NA	1.16	0.93	-	1.45	NA
Fatty acid profile and metabolism	Blood gamma-glutamylglutamine	3	0.895	0.98	0.70	-	1.36	NA	NA	-	NA	66.82
Fatty acid profile and metabolism	Blood 3-methyl-2-oxovalerate	3	0.798	0.97	0.74	-	1.27	NA	NA	-	NA	50.68
Fatty acid profile and metabolism	Blood citrulline	4	0.602	0.94	0.74	-	1.19	NA	NA	-	NA	29.10
Fatty acid profile and metabolism	Blood inosine	1	0.599	NA	NA	-	NA	0.95	0.78	-	1.16	NA
Fatty acid profile and metabolism	Blood dihomo-linolenate (20:3n3 or n6)	2	0.100	1.21	0.96	-	1.53	NA	NA	-	NA	0.00
Fatty acid profile and metabolism	Blood tryptophan betaine	1	0.584	NA	NA	-	NA	0.94	0.74	-	1.18	NA
Fatty acid profile and metabolism	Blood tyrosine	3	0.690	1.06	0.80	-	1.41	NA	NA	-	NA	46.31
Fatty acid profile and metabolism	Blood homocitrulline	1	0.834	NA	NA	-	NA	1.02	0.83	-	1.25	NA
Fatty acid profile and metabolism	Blood uridine	3	0.065	1.36	0.98	-	1.89	NA	NA	-	NA	62.15
Fatty acid profile and metabolism	Blood octadecanedioate	2	0.378	0.91	0.74	-	1.12	NA	NA	-	NA	0.00

Category	Exposure	No. of SNPs	IVW-RE					Wald Ratio				I^2	
			P-value		Odds ratio (95% CIs)			Odds ratio (95% CIs)					
Fatty acid profile and metabolism	Blood 1-palmitoylglycerophosphoethanolamine	2	0.314		0.90	0.73	1.10	NA	NA	NA		0.00	
Fatty acid profile and metabolism	Blood stearate (18:0)	2	0.932		1.02	0.67	1.54	NA	NA	NA		73.80	
Fatty acid profile and metabolism	Blood 1-linoleoylglycerol (1-monolinolein)	1	0.685		NA	NA	NA	0.95	0.73	1.22		NA	
Fatty acid profile and metabolism	Gamma-linoleic acid (18:3n6)	2	0.01798	*	0.84	0.73	0.97	NA	NA	NA		0.00	
Fatty acid profile and metabolism	Blood phenyllactate (PLA)	1	0.220		NA	NA	NA	0.85	0.66	1.10		NA	
Fatty acid profile and metabolism	Blood O-sulfo-L-tyrosine	2	0.300		0.90	0.73	1.10	NA	NA	NA		0.00	
Fatty acid profile and metabolism	Blood tetradecadienoate	1	0.985		NA	NA	NA	1.00	0.81	1.25		NA	
Fatty acid profile and metabolism	Blood 2-aminobutyrate	2	0.138		0.85	0.68	1.06	NA	NA	NA		0.00	
Fatty acid profile and metabolism	Blood albumin	4	0.816		1.05	0.69	1.61	NA	NA	NA		0.00	
Fatty acid profile and metabolism	Blood dehydroisoandrosterone sulfate (DHEA-S)	2	0.832		0.98	0.77	1.23	NA	NA	NA		0.00	
Fatty acid profile and metabolism	Blood alanine	6	0.837		0.97	0.72	1.30	NA	NA	NA		37.26	
Fatty acid profile and metabolism	Gamma-linolenic acid (18:3n6)	2	0.01520	*	0.86	0.76	0.97	NA	NA	NA		0.00	
Fatty acid profile and metabolism	Blood valine	5	0.583		1.08	0.81	1.45	NA	NA	NA		35.14	
Fatty acid profile and metabolism	Blood myo-inositol	2	0.716		1.08	0.71	1.64	NA	NA	NA		70.38	
Fatty acid profile and metabolism	Linoleic acid (18:2n6)	2	0.01128	*	1.10	1.02	1.18	NA	NA	NA		0.00	
Fatty acid profile and metabolism	Blood creatine	1	0.189		NA	NA	NA	0.84	0.65	1.09		NA	
Fatty acid profile and metabolism	Oleic acid (18:1n9)	1	0.00564	*	NA	NA	NA	1.34	1.09	1.66		NA	
Fatty acid profile and metabolism	Blood N2,N2-dimethylguanosine	2	0.481		1.09	0.86	1.39	NA	NA	NA		0.00	
Fatty acid profile and metabolism	Blood histidine	5	0.751		0.94	0.65	1.36	NA	NA	NA		47.24	
Fatty acid profile and metabolism	Omega-3 fatty acids	6	0.00054	*	0.74	0.62	0.88	NA	NA	NA		0.00	
Fatty acid profile and metabolism	Blood levulinate (4-oxovalerate)	2	0.993		1.00	0.55	1.80	NA	NA	NA		81.76	†
Fatty acid profile and metabolism	Blood glucose	3	0.670		0.94	0.73	1.23	NA	NA	NA		0.00	
Fatty acid profile and metabolism	Stearoylcarnitine	1	0.04456	*	NA	NA	NA	1.41	1.01	1.98		NA	

Category	Exposure	No. of SNPs	IVW-RE				Wald Ratio				I^2
			P-value	Odds ratio (95% CIs)			Odds ratio (95% CIs)				
Fatty acid profile and metabolism	Phenylalanine	4	0.419	0.87	0.61	1.23	NA	NA	NA		25.72
Fatty acid profile and metabolism	Indoleacetate	2	0.611	1.08	0.80	1.46	NA	NA	NA		0.00
Fatty acid profile and metabolism	4-methyl-2-oxopentanoate	2	0.707	1.05	0.80	1.38	NA	NA	NA		0.00
Fatty acid profile and metabolism	Lysine	1	0.849	NA	NA	NA	0.97	0.74	1.28		NA
Fatty acid profile and metabolism	Hypoxanthine	1	0.495	NA	NA	NA	1.10	0.84	1.44		NA
Fatty acid profile and metabolism	1,6-anhydroglucose	1	0.900	NA	NA	NA	1.02	0.76	1.36		NA
Fatty acid profile and metabolism	Gamma-glutamylphenylalanine	2	0.516	1.10	0.82	1.49	NA	NA	NA		0.00
Fatty acid profile and metabolism	Acetylphosphate	2	0.496	1.11	0.82	1.50	NA	NA	NA		12.05
Fatty acid profile and metabolism	Heptanoate (7:0)	2	0.669	0.91	0.61	1.38	NA	NA	NA		52.15
Fatty acid profile and metabolism	Stearidonate (18:4n3)	1	0.119	NA	NA	NA	0.77	0.56	1.07		NA
Fatty acid profile and metabolism	Myristoleate (14:1n5)	1	0.243	NA	NA	NA	0.84	0.63	1.13		NA
Fatty acid profile and metabolism	Caprylate (8:0)	2	0.513	1.12	0.80	1.57	NA	NA	NA		0.00
Fatty acid profile and metabolism	Laurate (12:0)	2	0.680	1.07	0.78	1.46	NA	NA	NA		0.00
Fatty acid profile and metabolism	Margarate (17:0)	1	0.243	NA	NA	NA	1.20	0.88	1.63		NA
Fatty acid profile and metabolism	Palmitoleate (16:1n7)	1	0.243	NA	NA	NA	0.83	0.61	1.13		NA
Fatty acid profile and metabolism	Laurylcarnitine	1	0.349	NA	NA	NA	1.18	0.83	1.68		NA
Fatty acid profile and metabolism	3,4-dihydroxybutyrate	1	0.638	NA	NA	NA	0.93	0.68	1.27		NA
Fatty acid profile and metabolism	Indolelactate	1	0.496	NA	NA	NA	0.90	0.65	1.23		NA
Fatty acid profile and metabolism	Docosahexaenoic acid (22:6n3)	1	0.984	NA	NA	NA	1.00	0.73	1.38		NA
Fatty acid profile and metabolism	3-(4-hydroxyphenyl)lactate	1	0.876	NA	NA	NA	1.03	0.71	1.48		NA
Fatty acid profile and metabolism	1-oleoylglycerol (1-monoolein)	1	0.886	NA	NA	NA	1.03	0.70	1.50		NA
Fatty acid profile and metabolism	1-stearoylglycerophosphoethanolamine	1	0.984	NA	NA	NA	1.00	0.73	1.39		NA
Fatty acid profile and metabolism	N1-methyl-3-pyridone-4-carboxamide	1	0.908	NA	NA	NA	1.02	0.70	1.49		NA
Fatty acid profile and metabolism	Dodecanedioate	1	0.423	NA	NA	NA	1.15	0.81	1.64		NA

Category	Exposure	No. of SNPs	IVW-RE					Wald Ratio					I^2	
			P-value		Odds ratio (95% CIs)			Odds ratio (95% CIs)						
Genome Stability	Telomere length	7	0.01263	*	2.33	1.20	- 4.52	NA	NA	-	NA		85.78	†
Genome Stability	LOY	92	0.101		1.45	0.93	- 2.25	NA	NA	-	NA		38.27	
Inflammatory factors	Plasma IL-6 sRa	1	0.282		NA	NA	- NA	1.02	0.98	-	1.05		NA	
Inflammatory factors	Circulating C-reactive protein	14	0.599		1.05	0.89	- 1.23	NA	NA	-	NA		17.36	
Lipids and lipid transport	Circulating fetuin-A	1	0.868		NA	NA	- NA	1.01	0.94	-	1.08		NA	
Lipids and lipid transport	LDL	102	0.289		0.95	0.86	- 1.04	NA	NA	-	NA		27.15	
Lipids and lipid transport	Total cholesterol	123	0.359		0.95	0.85	- 1.06	NA	NA	-	NA		42.55	
Lipids and lipid transport	HDL	124	0.992		1.00	0.89	- 1.12	NA	NA	-	NA		25.63	
Lipids and lipid transport	Triglycerides	70	0.762		0.98	0.86	- 1.12	NA	NA	-	NA		22.74	
lipids and lipid transport	Total triglycerides	34	0.419		0.95	0.85	- 1.07	NA	NA	-	NA		0.00	
Lipids and lipid transport	Omega-6 fatty acids	13	0.060		0.89	0.79	- 1.00	NA	NA	-	NA		0.00	
Lipids and lipid transport	Blood apolipoprotein A-I	12	0.424		0.95	0.84	- 1.08	NA	NA	-	NA		0.00	
Lipids and lipid transport	Total fatty acids	12	0.151		0.91	0.80	- 1.04	NA	NA	-	NA		0.00	
Lipids and lipid transport	Circulating adiponectin	10	0.853		1.02	0.83	- 1.26	NA	NA	-	NA		4.52	

Category	Exposure	No. of SNPs	IVW-RE					Wald Ratio				I^2
			P-value		Odds ratio (95% CIs)			Odds ratio (95% CIs)				
Miscellaneous	Fasting proinsulin	8	0.924		0.99	0.89	- 1.11	NA	NA	-	NA	0.00
Miscellaneous	Platelet count	39	0.302		1.13	0.90	- 1.41	NA	NA	-	NA	48.68
Miscellaneous	Corrected insulin response	3	0.823		1.02	0.86	- 1.21	NA	NA	-	NA	0.00
Miscellaneous	Forced vital capacity (FVC)	284	0.00791	*	1.21	1.05	- 1.39	NA	NA	-	NA	29.48
Miscellaneous	HbA1C levels	11	0.106		0.85	0.70	- 1.03	NA	NA	-	NA	0.00
Miscellaneous	Fluid intelligence score	50	0.059		0.79	0.61	- 1.01	NA	NA	-	NA	23.21
Miscellaneous	Serum IgE	3	0.936		0.99	0.81	- 1.21	NA	NA	-	NA	0.00
Miscellaneous	Morning/evening person (chronotype)	99	0.725		1.05	0.82	- 1.34	NA	NA	-	NA	6.59
Miscellaneous	Insulin disposition index	1	0.664		NA	NA	- NA	0.95	0.74	-	1.21	NA
Miscellaneous	Neuroticism score	78	0.358		1.15	0.85	- 1.56	NA	NA	-	NA	21.49
Miscellaneous	Years of schooling	74	0.055		0.72	0.52	- 1.01	NA	NA	-	NA	24.95
Miscellaneous	HOMA-B	4	0.533		0.91	0.67	- 1.23	NA	NA	-	NA	0.00
Miscellaneous	Nap during day	58	0.190		1.23	0.90	- 1.68	NA	NA	-	NA	0.00
Miscellaneous	Time spent watching television (TV)	65	0.242		1.23	0.87	- 1.75	NA	NA	-	NA	17.56
Miscellaneous	2hr glucose	7	0.463		0.89	0.65	- 1.21	NA	NA	-	NA	18.82

Category	Exposure	No. of SNPs	IVW-RE				Wald Ratio				I^2	
			P-value	Odds ratio (95% CIs)			Odds ratio (95% CIs)					
Plasma analytes	Interleukin-6 receptor	1	0.067	NA	NA	NA	0.95	0.91	-	1.00	NA	
Plasma analytes	Angiotensin-converting enzyme	2	0.138	1.04	0.99	1.10	NA	NA	-	NA	0.00	
Plasma analytes	Apolipoprotein A-IV	1	0.671	NA	NA	NA	0.99	0.92	-	1.05	NA	
Plasma analytes	E-selectin	1	0.118	NA	NA	NA	1.04	0.99	-	1.10	NA	
Plasma analytes	Fetuin-A	1	0.475	NA	NA	NA	0.98	0.93	-	1.04	NA	
Plasma analytes	Apolipoprotein H	3	0.678	1.02	0.94	1.09	NA	NA	-	NA	60.85	
Plasma analytes	Factor VII	2	0.451	0.96	0.86	1.07	NA	NA	-	NA	63.29	
Plasma analytes	Angiotensinogen	1	0.973	NA	NA	NA	1.00	0.92	-	1.08	NA	
Plasma analytes	Interleukin-16	1	0.972	NA	NA	NA	1.00	0.93	-	1.07	NA	
Plasma analytes	Chemokine CC-4	1	0.537	NA	NA	NA	0.98	0.90	-	1.05	NA	
Plasma analytes	Apolipoprotein E	1	0.653	NA	NA	NA	0.98	0.89	-	1.08	NA	
Plasma analytes	Carcinoembryonic antigen	2	0.804	0.97	0.79	1.20	NA	NA	-	NA	87.58	†
Plasma analytes	Myeloid progenitor inhibitory factor 1	2	0.212	0.94	0.85	1.04	NA	NA	-	NA	0.00	
Plasma analytes	CD 40 antigen	1	0.781	NA	NA	NA	1.01	0.93	-	1.10	NA	
Plasma analytes	Macrophage inflammatory protein-1 alpha	1	0.968	NA	NA	NA	1.00	0.90	-	1.12	NA	
Plasma analytes	Serotransferrin	1	0.832	NA	NA	NA	0.99	0.88	-	1.11	NA	
Plasma analytes	Haptoglobin	1	0.157	NA	NA	NA	0.91	0.81	-	1.04	NA	
Plasma analytes	Monocyte chemotactic protein 2	1	0.877	NA	NA	NA	1.01	0.91	-	1.12	NA	
Plasma analytes	Alpha-1-antitrypsin	1	0.594	NA	NA	NA	1.03	0.92	-	1.15	NA	
Plasma analytes	Tenascin-C	1	0.300	NA	NA	NA	0.94	0.84	-	1.06	NA	
Plasma analytes	Glutathione S-transferase alpha	1	0.796	NA	NA	NA	1.02	0.90	-	1.15	NA	
Plasma analytes	Matrix metalloproteinase-7	1	0.441	NA	NA	NA	1.05	0.93	-	1.20	NA	

Category	Exposure	No. of SNPs	IVW-RE						Wald Ratio				I^2
			P-value		Odds ratio (95% CIs)				Odds ratio (95% CIs)				
Plasma analytes	Neuronal cell adhesion molecule	1	0.121		NA	NA	-	NA	0.91	0.80	-	1.03	NA
Plasma analytes	Fibroblast growth factor 4	1	0.693		NA	NA	-	NA	1.03	0.89	-	1.20	NA
Plasma analytes	Cancer antigen 19-9	1	0.01560	*	NA	NA	-	NA	0.91	0.84	-	0.98	NA
Plasma analytes	CD5	1	0.478		NA	NA	-	NA	1.05	0.92	-	1.19	NA
Plasma analytes	Sortilin	1	0.778		NA	NA	-	NA	1.02	0.89	-	1.16	NA
Plasma analytes	B lymphocyte chemoattractant	1	0.129		NA	NA	-	NA	1.16	0.96	-	1.42	NA
Plasma analytes	Trefoil factor 3	1	0.061		NA	NA	-	NA	0.88	0.77	-	1.01	NA
Plasma analytes	Leptin	1	0.242		NA	NA	-	NA	0.91	0.78	-	1.07	NA
Plasma analytes	Epithelial-derived neutrophil-activating	1	0.565		NA	NA	-	NA	1.04	0.90	-	1.21	NA
Plasma analytes	Macrophage inflammatory protein-1 beta	1	0.953		NA	NA	-	NA	1.00	0.86	-	1.16	NA
Plasma analytes	Interleukin-13	1	0.722		NA	NA	-	NA	0.98	0.86	-	1.11	NA
Plasma analytes	Cystatin-C	1	0.931		NA	NA	-	NA	0.99	0.87	-	1.13	NA
Plasma analytes	Receptor for advanced glycosylation end	1	0.211		NA	NA	-	NA	0.91	0.80	-	1.05	NA
Plasma analytes	Growth-regulated alpha protein	1	0.278		NA	NA	-	NA	1.09	0.93	-	1.26	NA
Plasma analytes	Angiopoietin-2	1	0.095		NA	NA	-	NA	0.89	0.78	-	1.02	NA
Plasma analytes	Thymus-expressed chemokine	1	0.777		NA	NA	-	NA	0.98	0.87	-	1.11	NA
Plasma analytes	Macrophage colony-stimulating factor 1	1	0.503		NA	NA	-	NA	1.04	0.92	-	1.19	NA
Plasma analytes	Interleukin-18	1	0.308		NA	NA	-	NA	0.92	0.79	-	1.08	NA
Plasma analytes	Thrombopoietin	1	0.568		NA	NA	-	NA	1.04	0.91	-	1.20	NA
Plasma analytes	Vascular cell adhesion molecule-1	1	0.501		NA	NA	-	NA	1.06	0.90	-	1.24	NA
Plasma analytes	Interleukin-8	1	0.755		NA	NA	-	NA	1.02	0.89	-	1.18	NA
Plasma analytes	Tamm-Horsfall urinary glycoprotein	1	0.03723	*	NA	NA	-	NA	0.94	0.88	-	1.00	NA

			IVW-RE				Wald Ratio			
Category	Exposure	No. of SNPs	*P*-value	Odds ratio (95% CIs)			Odds ratio (95% CIs)			I^2
Obesity	Whole body water mass	735	0.176	1.06	0.97	1.16	NA	NA	NA	22.59
Obesity	Basal metabolic rate	693	0.142	1.07	0.98	1.17	NA	NA	NA	21.49
Obesity	Weight	576	0.930	1.00	0.90	1.12	NA	NA	NA	22.19
Obesity	Impedance of whole body	564	0.685	0.98	0.88	1.09	NA	NA	NA	26.21
Obesity	Body mass index	964	0.082	1.10	0.99	1.22	NA	NA	NA	9.96
Obesity	Whole body fat mass	415	0.940	1.00	0.87	1.13	NA	NA	NA	26.97
Obesity	Body fat percentage	365	0.330	1.07	0.93	1.23	NA	NA	NA	23.84
Obesity	Trunk fat percentage	334	0.149	1.11	0.96	1.29	NA	NA	NA	25.76
Obesity	Waist circumference	316	0.848	1.02	0.87	1.18	NA	NA	NA	27.55
Obesity	Birth weight	93	0.052	1.23	1.00	1.51	NA	NA	NA	28.60
Obesity	Hip circumference	89	0.813	1.03	0.82	1.28	NA	NA	NA	36.65
Obesity	Waist-to-hip ratio	35	0.779	0.96	0.73	1.27	NA	NA	NA	45.58
Obesity	Birth weight of first child	45	0.322	1.13	0.89	1.44	NA	NA	NA	13.05
Sex hormones and reproduction	Age at menopause**	48	0.419	0.78	0.43	1.42	NA	NA	NA	28.44
Sex hormones and reproduction	Age at menarche**	73	0.561	1.25	0.59	2.65	NA	NA	NA	30.61
Sex hormones and reproduction	Plasma progesterone	2	0.301	0.89	0.72	1.11	NA	NA	NA	39.80
Sex hormones and reproduction	Plasma progesterone**	2	0.754	0.91	0.49	1.68	NA	NA	NA	0.00
Sex hormones and reproduction	Circulating carotenoids	1	0.925	NA	NA	NA	1.01	0.86	1.18	NA
Sex hormones and resproduction	Plasma estradiol	1	0.562	NA	NA	NA	0.90	0.63	1.29	NA

Appendix 22

Causal estimates from each Mendelian randomization approach for each trait and MM risk (tables overleaf).

* Indicates P < 0.05**.** ** Causal effects estimated using MM summary statistics from females only. HDL, high density lipoprotein; LDL, low density lipoprotein; LOY, loss of Y chromosome

Exposure	IVW-RE *P*-value		IVW-RE Odds ratio (95% CIs)				IVW-FE Odds ratio (95% CIs)				Wald Ratio Odds ratio (95% CIs)				Causal estimate from MR-Egger slope Odds ratio (95% CIs)			
Plasma IL-6 sRa	0.282		NA	NA	-	NA	NA	NA	-	NA	1.02	0.98	-	1.05	NA	NA	-	NA
3-phenylpropionate (hydrocinnamate)	0.01895	*	NA	NA	-	NA	NA	NA	-	NA	1.54	1.07	-	2.20	NA	NA	-	NA
Height	0.617		1.01	0.96	-	1.07	1.01	0.97	-	1.06	NA	NA	-	NA	1.06	0.96	-	1.18
Interleukin-6 receptor	0.067		NA	NA	-	NA	NA	NA	-	NA	0.95	0.91	-	1.00	NA	NA	-	NA
Angiotensin-converting enzyme	0.138		1.04	0.99	-	1.10	1.04	0.99	-	1.10	NA	NA	-	NA	NA	NA	-	NA
Apolipoprotein A-IV	0.671		NA	NA	-	NA	NA	NA	-	NA	0.99	0.92	-	1.05	NA	NA	-	NA
E-selectin	0.118		NA	NA	-	NA	NA	NA	-	NA	1.04	0.99	-	1.10	NA	NA	-	NA
Blood butyrylcarnitine	0.887		1.01	0.93	-	1.08	1.01	0.94	-	1.07	NA	NA	-	NA	0.95	0.82	-	1.09
Fetuin-A	0.475		NA	NA	-	NA	NA	NA	-	NA	0.98	0.93	-	1.04	NA	NA	-	NA
Blood N-acetylornithine	0.096		0.95	0.89	-	1.01	0.95	0.89	-	1.01	NA	NA	-	NA	0.98	0.86	-	1.11
Blood glycoproteins	0.517		0.98	0.91	-	1.05	0.98	0.91	-	1.05	NA	NA	-	NA	0.93	0.83	-	1.05
Apolipoprotein H	0.678		1.02	0.94	-	1.09	1.02	0.97	-	1.06	NA	NA	-	NA	2.22	0.54	-	9.06
Heel bone mineral density (BMD) T-score	0.578		0.98	0.91	-	1.06	0.98	0.91	-	1.05	NA	NA	-	NA	NA	NA	-	NA
Factor VII	0.451		0.96	0.86	-	1.07	0.96	0.90	-	1.02	NA	NA	-	NA	NA	NA	-	NA
Angiotensinogen	0.973		NA	NA	-	NA	NA	NA	-	NA	1.00	0.92	-	1.08	NA	NA	-	NA

Exposure	IVW-RE *P*-value		IVW-RE Odds ratio (95% CIs)				IVW-FE Odds ratio (95% CIs)				Wald Ratio Odds ratio (95% CIs)				Causal estimate from MR-Egger slope Odds ratio (95% CIs)			
Interleukin-16	0.972		NA	NA	-	NA	NA	NA	-	NA	1.00	0.93	-	1.07	NA	NA	-	NA
Circulating fetuin-A	0.868		NA	NA	-	NA	NA	NA	-	NA	1.01	0.94	-	1.08	NA	NA	-	NA
Adrenic acid (22:4n6)	0.00922	*	NA	NA	-	NA	NA	NA	-	NA	0.88	0.79	-	0.97	NA	NA	-	NA
Alpha-linolenic acid (18:3n3)	0.01123	*	NA	NA	-	NA	NA	NA	-	NA	1.20	1.04	-	1.38	NA	NA	-	NA
Whole body water mass	0.176		1.06	0.97	-	1.16	1.06	0.98	-	1.15	NA	NA	-	NA	NA	NA	-	NA
Chemokine CC-4	0.537		NA	NA	-	NA	NA	NA	-	NA	0.98	0.90	-	1.05	NA	NA	-	NA
Blood biliverdin	0.531		0.98	0.90	-	1.05	0.98	0.90	-	1.05	NA	NA	-	NA	0.98	0.86	-	1.12
Basal metabolic rate	0.142		1.07	0.98	-	1.17	1.07	0.99	-	1.16	NA	NA	-	NA	0.94	0.73	-	1.20
Apolipoprotein E	0.653		NA	NA	-	NA	NA	NA	-	NA	0.98	0.89	-	1.08	NA	NA	-	NA
Carcinoembryonic antigen	0.804		0.97	0.79	-	1.20	0.97	0.90	-	1.05	NA	NA	-	NA	NA	NA	-	NA
Dihomo-gamma-linolenic acid (20:3n6)	0.223		1.06	0.97	-	1.16	1.06	0.97	-	1.16	NA	NA	-	NA	NA	NA	-	NA
Blood glycine	0.142		0.93	0.85	-	1.02	0.93	0.85	-	1.02	NA	NA	-	NA	0.96	0.83	-	1.11
Blood estrone 3-sulfate	0.359		NA	NA	-	NA	NA	NA	-	NA	0.95	0.86	-	1.06	NA	NA	-	NA
Arachidonic acid (20:4n6)	0.01174	*	0.95	0.92	-	0.99	0.95	0.92	-	0.99	NA	NA	-	NA	NA	NA	-	NA
Blood succinylcarnitine	0.412		0.96	0.87	-	1.06	0.96	0.88	-	1.04	NA	NA	-	NA	0.99	0.78	-	1.25

Exposure	IVW-RE			IVW-FE			Wald Ratio			Causal estimate from MR-Egger slope		
	P-value	Odds ratio (95% CIs)		Odds ratio (95% CIs)			Odds ratio (95% CIs)			Odds ratio (95% CIs)		
Blood glutaroyl carnitine	0.163	1.06	0.98 - 1.16	1.06	0.98 - 1.16		NA	NA - NA		0.78	0.53 - 1.15	
Blood bilirubin (Z,Z)	0.641	0.98	0.90 - 1.07	0.98	0.90 - 1.07		NA	NA - NA		NA	NA - NA	
LDL	0.289	0.95	0.86 - 1.04	0.95	0.88 - 1.03		NA	NA - NA		0.90	0.78 - 1.04	
Blood 2-aminooctanoic acid	0.760	0.99	0.91 - 1.07	0.99	0.91 - 1.07		NA	NA - NA		0.94	0.82 - 1.08	
Myeloid progenitor inhibitory factor 1	0.212	0.94	0.85 - 1.04	0.94	0.85 - 1.04		NA	NA - NA		NA	NA - NA	
Blood carnitine	0.00114 *	1.13	1.05 - 1.22	1.13	1.05 - 1.22		NA	NA - NA		1.06	0.88 - 1.27	
Total cholesterol	0.359	0.95	0.85 - 1.06	0.95	0.87 - 1.03		NA	NA - NA		0.85	0.71 - 1.01	
CD 40 antigen	0.781	NA	NA - NA	NA	NA - NA		1.01	0.93 - 1.10		NA	NA - NA	
Blood *cis*-4-decenoyl carnitine	0.03211 *	1.17	1.01 - 1.34	1.17	1.01 - 1.34		NA	NA - NA		NA	NA - NA	
Blood apolipoprotein B	0.095	0.93	0.84 - 1.01	0.93	0.84 - 1.01		NA	NA - NA		0.86	0.73 - 1.02	
Weight	0.930	1.00	0.90 - 1.12	1.00	0.92 - 1.10		NA	NA - NA		1.28	0.31 - 5.37	
Blood tryptophan	0.111	0.93	0.85 - 1.02	0.93	0.85 - 1.02		NA	NA - NA		1.79	0.56 - 5.73	
Impedance of whole body	0.685	0.98	0.88 - 1.09	0.98	0.89 - 1.07		NA	NA - NA		1.10	0.79 - 1.53	
Blood bradykinin, des-arg(9)	0.265	0.93	0.82 - 1.06	0.93	0.85 - 1.02		NA	NA - NA		0.71	0.50 - 1.00	
Blood androsterone sulfate	0.828	1.01	0.92 - 1.11	1.01	0.92 - 1.11		NA	NA - NA		1.02	0.90 - 1.16	

Exposure	IVW-RE				IVW-FE				Wald Ratio				Causal estimate from MR-Egger slope			
	P-value		Odds ratio (95% CIs)				Odds ratio (95% CIs)				Odds ratio (95% CIs)				Odds ratio (95% CIs)			
HDL	0.992		1.00	0.89	-	1.12	1.00	0.91	-	1.10	NA	NA	-	NA	0.94	0.76	-	1.15
Body mass index	0.082		1.10	0.99	-	1.22	1.10	0.99	-	1.21	NA	NA	-	NA	1.16	0.85	-	1.58
Blood decanoylcarnitine	0.01291	*	1.16	1.03	-	1.31	1.16	1.03	-	1.31	NA	NA	-	NA	NA	NA	-	NA
Blood proline	0.508		0.97	0.87	-	1.07	0.97	0.87	-	1.07	NA	NA	-	NA	0.93	0.79	-	1.11
Blood tetradecanedioate	0.195		0.93	0.84	-	1.04	0.93	0.85	-	1.03	NA	NA	-	NA	1.05	0.73	-	1.52
Macrophage inflammatory protein-1 alpha	0.968		NA	NA	-	NA	NA	NA	-	NA	1.00	0.90	-	1.12	NA	NA	-	NA
Blood alpha-glutamyltyrosine	0.472		1.04	0.94	-	1.16	1.04	0.94	-	1.16	NA	NA	-	NA	1.05	0.78	-	1.42
Blood hexanoylcarnitine	0.00695	*	1.16	1.04	-	1.29	1.16	1.04	-	1.29	NA	NA	-	NA	1.03	0.83	-	1.27
Serotransferrin	0.832		NA	NA	-	NA	NA	NA	-	NA	0.99	0.88	-	1.11	NA	NA	-	NA
Blood HDL diameter	0.302		0.92	0.79	-	1.08	0.92	0.83	-	1.02	NA	NA	-	NA	1.06	0.66	-	1.71
Haptoglobin	0.157		NA	NA	-	NA	NA	NA	-	NA	0.91	0.81	-	1.04	NA	NA	-	NA
Whole body fat mass	0.940		1.00	0.87	-	1.13	1.00	0.89	-	1.11	NA	NA	-	NA	NA	NA	-	NA
Triglycerides	0.762		0.98	0.86	-	1.12	0.98	0.87	-	1.10	NA	NA	-	NA	0.85	0.69	-	1.05
Blood 5-oxoproline	0.749		NA	NA	-	NA	NA	NA	-	NA	0.98	0.89	-	1.09	NA	NA	-	NA
Fasting proinsulin	0.924		0.99	0.89	-	1.11	0.99	0.89	-	1.11	NA	NA	-	NA	1.05	0.79	-	1.39

Exposure	IVW-RE			IVW-FE			Wald Ratio			Causal estimate from MR-Egger slope		
	P-value		Odds ratio (95% CIs)	Odds ratio (95% CIs)			Odds ratio (95% CIs)			Odds ratio (95% CIs)		
Total triglycerides	0.419		0.95 (0.85 - 1.07)	0.95 (0.85 - 1.07)			NA (NA - NA)			0.92 (0.76 - 1.11)		
Monocyte chemotactic protein 2	0.877		NA (NA - NA)	NA (NA - NA)			1.01 (0.91 - 1.12)			NA (NA - NA)		
Blood hydroxyisovaleroyl carnitine	0.04608	*	1.21 (1.00 - 1.46)	1.21 (1.00 - 1.46)			NA (NA - NA)			NA (NA - NA)		
Blood urate	0.835		0.98 (0.82 - 1.17)	0.98 (0.88 - 1.10)			NA (NA - NA)			NA (NA - NA)		
Blood indolepropionate	0.01514	*	NA (NA - NA)	NA (NA - NA)			1.44 (1.07 - 1.92)			NA (NA - NA)		
Blood 5-alpha-pregnan-3beta,20alpha-disulfate	0.760		0.98 (0.86 - 1.12)	0.98 (0.86 - 1.12)			NA (NA - NA)			1.46 (0.59 - 3.60)		
Blood leucine	0.314		1.06 (0.95 - 1.19)	1.06 (0.95 - 1.19)			NA (NA - NA)			1.04 (0.51 - 2.12)		
Blood isovalerylcarnitine	0.00252	*	1.21 (1.07 - 1.37)	1.21 (1.07 - 1.37)			NA (NA - NA)			1.87 (0.55 - 6.35)		
Blood hexadecanedioate	0.162		0.92 (0.82 - 1.03)	0.92 (0.82 - 1.03)			NA (NA - NA)			1.02 (0.65 - 1.60)		
Omega-6 fatty acids	0.060		0.89 (0.79 - 1.00)	0.89 (0.79 - 1.00)			NA (NA - NA)			0.78 (0.57 - 1.07)		
Alpha-1-antitrypsin	0.594		NA (NA - NA)	NA (NA - NA)			1.03 (0.92 - 1.15)			NA (NA - NA)		
Blood N-(2-furoyl)glycine	0.065		NA (NA - NA)	NA (NA - NA)			1.12 (0.99 - 1.25)			NA (NA - NA)		
Blood phenylalanylserine	0.860		0.99 (0.88 - 1.11)	0.99 (0.88 - 1.11)			NA (NA - NA)			NA (NA - NA)		
Blood 10-undecenoate (11:1n1)	0.607		0.97 (0.85 - 1.10)	0.97 (0.85 - 1.10)			NA (NA - NA)			1.03 (0.73 - 1.44)		
Blood octanoylcarnitine	0.01384	*	1.16 (1.03 - 1.30)	1.16 (1.03 - 1.29)			NA (NA - NA)			0.96 (0.72 - 1.28)		

Exposure	*P*-value		IVW-RE Odds ratio (95% CIs)				IVW-FE Odds ratio (95% CIs)				Wald Ratio Odds ratio (95% CIs)				Causal estimate from MR-Egger slope Odds ratio (95% CIs)			
Body fat percentage	0.330		1.07	0.93	-	1.23	1.07	0.95	-	1.21	NA	NA	-	NA	0.98	0.59	-	1.62
Blood epiandrosterone sulfate	0.613		1.03	0.92	-	1.16	1.03	0.92	-	1.16	NA	NA	-	NA	NA	NA	-	NA
Blood apolipoprotein A-I	0.424		0.95	0.84	-	1.08	0.95	0.84	-	1.08	NA	NA	-	NA	1.21	0.83	-	1.77
Age at menopause**	0.419		0.78	0.43	-	1.42	0.78	0.47	-	1.30	NA	NA	-	NA	0.53	0.12	-	2.44
Tenascin-C	0.300		NA	NA	-	NA	NA	NA	-	NA	0.94	0.84	-	1.06	NA	NA	-	NA
Blood palmitoyl sphingomyelin	0.02229	*	0.76	0.61	-	0.96	0.76	0.61	-	0.96	NA	NA	-	NA	NA	NA	-	NA
Glutathione S-transferase alpha	0.796		NA	NA	-	NA	NA	NA	-	NA	1.02	0.90	-	1.15	NA	NA	-	NA
Matrix metalloproteinase-7	0.441		NA	NA	-	NA	NA	NA	-	NA	1.05	0.93	-	1.20	NA	NA	-	NA
Trunk fat percentage	0.149		1.11	0.96	-	1.29	1.11	0.98	-	1.26	NA	NA	-	NA	NA	NA	-	NA
Blood isobutyrylcarnitine	0.899		1.01	0.89	-	1.14	1.01	0.89	-	1.14	NA	NA	-	NA	1.06	0.78	-	1.44
Blood propionylcarnitine	0.01273	*	1.18	1.04	-	1.35	1.18	1.04	-	1.34	NA	NA	-	NA	0.95	0.68	-	1.33
Blood N-[3-(2-Oxopyrrolidin-1-yl)propyl]acetamide	0.542		0.96	0.86	-	1.09	0.96	0.86	-	1.09	NA	NA	-	NA	1.06	0.69	-	1.62
Serum vitamin B12	0.701		1.04	0.85	-	1.27	1.04	0.91	-	1.19	NA	NA	-	NA	0.88	0.56	-	1.38
Age at menarche**	0.561		1.25	0.59	-	2.65	1.25	0.67	-	2.34	NA	NA	-	NA	3.22	0.87	-	11.90
Blood leucylalanine	0.511		0.96	0.85	-	1.08	0.96	0.85	-	1.08	NA	NA	-	NA	NA	NA	-	NA

Exposure	*P*-value		IVW-RE Odds ratio (95% CIs)				IVW-FE Odds ratio (95% CIs)				Wald Ratio Odds ratio (95% CIs)				Causal estimate from MR-Egger slope Odds ratio (95% CIs)			
Blood phosphatidylcholine and other cholines	0.506		0.96	0.84	-	1.09	0.96	0.84	-	1.09	NA	NA	-	NA	0.96	0.64	-	1.44
Blood kynurenine	0.688		1.03	0.91	-	1.16	1.03	0.91	-	1.16	NA	NA	-	NA	1.04	0.82	-	1.33
Waist circumference	0.848		1.02	0.87	-	1.18	1.02	0.89	-	1.16	NA	NA	-	NA	1.01	0.62	-	1.65
Neuronal cell adhesion molecule	0.121		NA	NA	-	NA	NA	NA	-	NA	0.91	0.80	-	1.03	NA	NA	-	NA
Blood scyllo-inositol	0.02421	*	NA	NA	-	NA	NA	NA	-	NA	1.35	1.04	-	1.76	NA	NA	-	NA
Total fatty acids	0.151		0.91	0.80	-	1.04	0.91	0.80	-	1.04	NA	NA	-	NA	0.82	0.45	-	1.46
Blood N-acetylglycine	0.625		0.95	0.79	-	1.15	0.95	0.83	-	1.10	NA	NA	-	NA	0.79	0.56	-	1.12
Fibroblast growth factor 4	0.693		NA	NA	-	NA	NA	NA	-	NA	1.03	0.89	-	1.20	NA	NA	-	NA
Blood copper	0.935		0.99	0.80	-	1.22	0.99	0.85	-	1.15	NA	NA	-	NA	NA	NA	-	NA
Total phosphoglycerides	0.183		0.91	0.80	-	1.04	0.91	0.80	-	1.04	NA	NA	-	NA	1.02	0.64	-	1.63
Blood zinc	0.984		1.00	0.87	-	1.15	1.00	0.87	-	1.15	NA	NA	-	NA	NA	NA	-	NA
Cancer antigen 19-9	0.01560	*	NA	NA	-	NA	NA	NA	-	NA	0.91	0.84	-	0.98	NA	NA	-	NA
Eicosapentaenoic acid (20:5n3)	0.076		0.89	0.78	-	1.01	0.89	0.78	-	1.01	NA	NA	-	NA	0.82	0.59	-	1.13
CD5	0.478		NA	NA	-	NA	NA	NA	-	NA	1.05	0.92	-	1.19	NA	NA	-	NA
Blood mannose	0.055		NA	NA	-	NA	NA	NA	-	NA	1.13	1.00	-	1.29	NA	NA	-	NA

Exposure	*P*-value	IVW-RE Odds ratio (95% CIs)	IVW-FE Odds ratio (95% CIs)	Wald Ratio Odds ratio (95% CIs)	Causal estimate from MR-Egger slope Odds ratio (95% CIs)
VLDL diameter	0.533	1.06 0.88 - 1.28	1.06 0.93 - 1.21	NA NA - NA	1.05 0.48 - 2.30
Blood 4-acetamidobutanoate	0.621	1.05 0.86 - 1.29	1.05 0.93 - 1.20	NA NA - NA	NA NA - NA
Sortilin	0.778	NA NA - NA	NA NA - NA	1.02 0.89 - 1.16	NA NA - NA
Blood 5alpha-androstan-3beta,17beta-diol disulfate	0.650	0.97 0.85 - 1.11	0.97 0.85 - 1.11	NA NA - NA	1.01 0.77 - 1.33
Blood betaine	0.382	1.10 0.89 - 1.34	1.10 0.95 - 1.26	NA NA - NA	1.76 0.90 - 3.45
B lymphocyte chemoattractant	0.129	NA NA - NA	NA NA - NA	1.16 0.96 - 1.42	NA NA - NA
Trefoil factor 3	0.061	NA NA - NA	NA NA - NA	0.88 0.77 - 1.01	NA NA - NA
Blood N-acetylcarnosine	0.709	1.05 0.83 - 1.32	1.05 0.91 - 1.20	NA NA - NA	1.20 0.12 - 12.40
Leptin	0.242	NA NA - NA	NA NA - NA	0.91 0.78 - 1.07	NA NA - NA
Epithelial-derived neutrophil-activating	0.565	NA NA - NA	NA NA - NA	1.04 0.90 - 1.21	NA NA - NA
Macrophage inflammatory protein-1 beta	0.953	NA NA - NA	NA NA - NA	1.00 0.86 - 1.16	NA NA - NA
Interleukin-13	0.722	NA NA - NA	NA NA - NA	0.98 0.86 - 1.11	NA NA - NA
Cystatin-C	0.931	NA NA - NA	NA NA - NA	0.99 0.87 - 1.13	NA NA - NA
Receptor for advanced glycosylation end	0.211	NA NA - NA	NA NA - NA	0.91 0.80 - 1.05	NA NA - NA
Growth-regulated alpha protein	0.278	NA NA - NA	NA NA - NA	1.09 0.93 - 1.26	NA NA - NA

Exposure	*P*-value	IVW-RE Odds ratio (95% CIs)	IVW-FE Odds ratio (95% CIs)	Wald Ratio Odds ratio (95% CIs)	Causal estimate from MR-Egger slope Odds ratio (95% CIs)
Angiopoietin-2	0.095	NA NA - NA	NA NA - NA	0.89 0.78 - 1.02	NA NA - NA
Thymus-expressed chemokine	0.777	NA NA - NA	NA NA - NA	0.98 0.87 - 1.11	NA NA - NA
Blood sphingomyelins	0.129	0.89 0.76 - 1.04	0.89 0.76 - 1.04	NA NA - NA	0.67 0.48 - 0.94
Blood asparagine	0.610	0.95 0.78 - 1.15	0.95 0.83 - 1.09	NA NA - NA	NA NA - NA
Macrophage colony-stimulating factor 1	0.503	NA NA - NA	NA NA - NA	1.04 0.92 - 1.19	NA NA - NA
Interleukin-18	0.308	NA NA - NA	NA NA - NA	0.92 0.79 - 1.08	NA NA - NA
Circulating C-reactive protein	0.599	1.05 0.89 - 1.23	1.05 0.90 - 1.21	NA NA - NA	1.20 0.88 - 1.66
Fasting glucose	0.287	1.10 0.92 - 1.30	1.10 0.95 - 1.27	NA NA - NA	1.21 0.82 - 1.80
Thrombopoietin	0.568	NA NA - NA	NA NA - NA	1.04 0.91 - 1.20	NA NA - NA
Blood 12-hydroxyeicosatetraenoate (12-HETE)	0.152	NA NA - NA	NA NA - NA	0.90 0.78 - 1.04	NA NA - NA
Blood serine	0.335	0.93 0.79 - 1.08	0.93 0.79 - 1.08	NA NA - NA	0.53 0.12 - 2.29
Vascular cell adhesion molecule-1	0.501	NA NA - NA	NA NA - NA	1.06 0.90 - 1.24	NA NA - NA
Interleukin-8	0.755	NA NA - NA	NA NA - NA	1.02 0.89 - 1.18	NA NA - NA
Blood 1,5-anhydroglucitol (1,5-AG)	0.484	1.09 0.85 - 1.41	1.09 0.93 - 1.28	NA NA - NA	1.58 0.83 - 3.00
Plasma progesterone	0.301	0.89 0.72 - 1.11	0.89 0.76 - 1.05	NA NA - NA	NA NA - NA

Exposure	*P*-value		IVW-RE Odds ratio (95% CIs)				IVW-FE Odds ratio (95% CIs)				Wald Ratio Odds ratio (95% CIs)				Causal estimate from MR-Egger slope Odds ratio (95% CIs)			
Plasma progesterone**	0.754		0.91	0.49	-	1.68	0.91	0.49	-	1.68	NA	NA	-	NA	NA	NA	-	NA
Platelet count	0.302		1.13	0.90	-	1.41	1.13	0.96	-	1.32	NA	NA	-	NA	1.04	0.50	-	2.17
Dihomo-gamma-linoleic acid (20:3n6)	0.01561	*	1.09	1.02	-	1.17	1.09	1.02	-	1.17	NA	NA	-	NA	NA	NA	-	NA
Pulse rate	0.920		0.99	0.82	-	1.20	0.99	0.85	-	1.15	NA	NA	-	NA	0.95	0.58	-	1.57
Blood citrate	0.328		1.08	0.92	-	1.27	1.08	0.92	-	1.27	NA	NA	-	NA	1.18	0.74	-	1.88
Omega-9 and saturated fatty acids	0.186		0.90	0.77	-	1.05	0.90	0.77	-	1.05	NA	NA	-	NA	0.81	0.43	-	1.51
Glycoprotein acetyls	0.108		0.86	0.72	-	1.03	0.86	0.72	-	1.03	NA	NA	-	NA	0.91	0.50	-	1.64
Mono-unsaturated fatty acids	0.092		0.87	0.74	-	1.02	0.87	0.74	-	1.02	NA	NA	-	NA	0.75	0.39	-	1.44
Blood alpha-hydroxyisovalerate	0.978		1.00	0.72	-	1.38	1.00	0.85	-	1.17	NA	NA	-	NA	0.54	0.22	-	1.32
Circulating carotenoids	0.925		NA	NA	-	NA	NA	NA	-	NA	1.01	0.86	-	1.18	NA	NA	-	NA
Docosapentaenoic acid (22:5n3)	0.03732	*	0.90	0.81	-	0.99	0.90	0.83	-	0.98	NA	NA	-	NA	NA	NA	-	NA
Circulating 25-hydroxyvitamin D	0.541		1.08	0.84	-	1.40	1.08	0.93	-	1.26	NA	NA	-	NA	0.89	0.57	-	1.38
Blood alpha-ketoglutarate	0.154		NA	NA	-	NA	NA	NA	-	NA	0.88	0.74	-	1.05	NA	NA	-	NA
Omega-7 and -9 and saturated fatty acids	0.148		0.88	0.75	-	1.05	0.88	0.75	-	1.05	NA	NA	-	NA	0.86	0.44	-	1.68
Monounsaturated fatty acids	0.071		0.85	0.72	-	1.01	0.85	0.72	-	1.01	NA	NA	-	NA	0.92	0.32	-	2.65

Exposure	IVW-RE		IVW-FE		Wald Ratio		Causal estimate from MR-Egger slope	
	P-value	Odds ratio (95% CIs)	Odds ratio (95% CIs)		Odds ratio (95% CIs)		Odds ratio (95% CIs)	
Serum calcium	0.130	0.88 (0.74 - 1.04)	0.88 (0.74 - 1.04)		NA (NA - NA)		0.75 (0.57 - 1.00)	
Birth weight	0.052	1.23 (1.00 - 1.51)	1.23 (1.03 - 1.46)		NA (NA - NA)		1.50 (0.75 - 3.00)	
Corrected insulin response	0.823	1.02 (0.86 - 1.21)	1.02 (0.86 - 1.21)		NA (NA - NA)		NA (NA - NA)	
Blood cysteine-glutathione disulfide	0.226	NA (NA - NA)	NA (NA - NA)		1.11 (0.94 - 1.32)		NA (NA - NA)	
Palmitoleic acid (16:1n7)	0.673	1.07 (0.77 - 1.50)	1.07 (0.90 - 1.28)		NA (NA - NA)		1.53 (0.35 - 6.66)	
Hip circumference	0.813	1.03 (0.82 - 1.28)	1.03 (0.86 - 1.22)		NA (NA - NA)		1.31 (0.88 - 1.96)	
LDL diameter	0.473	0.91 (0.70 - 1.18)	0.91 (0.76 - 1.09)		NA (NA - NA)		0.86 (0.41 - 1.82)	
Blood gamma-glutamyltyrosine	0.999	1.00 (0.75 - 1.33)	1.00 (0.83 - 1.21)		NA (NA - NA)		4.94 (0.89 - 27.44)	
Blood 2-hydroxyisobutyrate	0.747	0.95 (0.72 - 1.27)	0.95 (0.79 - 1.15)		NA (NA - NA)		0.75 (0.28 - 1.98)	
Stearic acid (18:0)	0.567	0.90 (0.63 - 1.28)	0.90 (0.75 - 1.08)		NA (NA - NA)		0.01 (0.00 - 0.27)	
Blood aspartylphenylalanine	0.899	NA (NA - NA)	NA (NA - NA)		0.99 (0.82 - 1.19)		NA (NA - NA)	
Blood selenium	0.373	NA (NA - NA)	NA (NA - NA)		0.91 (0.75 - 1.12)		NA (NA - NA)	
Blood acetylcarnitine	0.071	1.20 (0.98 - 1.46)	1.20 (0.99 - 1.46)		NA (NA - NA)		NA (NA - NA)	
Blood glutamine	0.555	0.91 (0.67 - 1.24)	0.91 (0.75 - 1.10)		NA (NA - NA)		0.79 (0.42 - 1.49)	
Forced vital capacity (FVC)	0.00791 *	1.21 (1.05 - 1.39)	1.21 (1.08 - 1.36)		NA (NA - NA)		NA (NA - NA)	

Exposure	IVW-RE				IVW-FE				Wald Ratio				Causal estimate from MR-Egger slope			
	P-value	Odds ratio (95% CIs)				Odds ratio (95% CIs)				Odds ratio (95% CIs)				Odds ratio (95% CIs)			
Blood methylcysteine	0.182	NA	NA	-	NA	NA	NA	-	NA	1.16	0.93	-	1.45	NA	NA	-	NA
Blood gamma-glutamylglutamine	0.895	0.98	0.70	-	1.36	0.98	0.81	-	1.18	NA	NA	-	NA	0.45	0.23	-	0.86
Blood 3-methyl-2-oxovalerate	0.798	0.97	0.74	-	1.27	0.97	0.80	-	1.17	NA	NA	-	NA	2.07	0.23	-	19.00
Blood citrulline	0.602	0.94	0.74	-	1.19	0.94	0.77	-	1.15	NA	NA	-	NA	2.38	0.05	-	110.92
Blood inosine	0.599	NA	NA	-	NA	NA	NA	-	NA	0.95	0.78	-	1.16	NA	NA	-	NA
HbA1C levels	0.106	0.85	0.70	-	1.03	0.85	0.70	-	1.03	NA	NA	-	NA	NA	NA	-	NA
Circulating adiponectin	0.853	1.02	0.83	-	1.26	1.02	0.83	-	1.25	NA	NA	-	NA	1.25	0.69	-	2.29
Blood dihomo-linolenate (20:3n3 or n6)	0.100	1.21	0.96	-	1.53	1.21	0.96	-	1.53	NA	NA	-	NA	NA	NA	-	NA
Waist-to-hip ratio	0.779	0.96	0.73	-	1.27	0.96	0.78	-	1.18	NA	NA	-	NA	1.50	0.53	-	4.27
Blood tryptophan betaine	0.584	NA	NA	-	NA	NA	NA	-	NA	0.94	0.74	-	1.18	NA	NA	-	NA
Blood tyrosine	0.690	1.06	0.80	-	1.41	1.06	0.86	-	1.30	NA	NA	-	NA	7.26	0.79	-	66.68
Blood homocitrulline	0.834	NA	NA	-	NA	NA	NA	-	NA	1.02	0.83	-	1.25	NA	NA	-	NA
Blood uridine	0.065	1.36	0.98	-	1.89	1.36	1.11	-	1.66	NA	NA	-	NA	3.25	1.50	-	7.06
Blood octadecanedioate	0.378	0.91	0.74	-	1.12	0.91	0.74	-	1.12	NA	NA	-	NA	NA	NA	-	NA
Fluid intelligence score	0.059	0.79	0.61	-	1.01	0.79	0.63	-	0.98	NA	NA	-	NA	NA	NA	-	NA

Exposure	IVW-RE			IVW-FE			Wald Ratio			Causal estimate from MR-Egger slope		
	P-value	Odds ratio (95% CIs)		Odds ratio (95% CIs)			Odds ratio (95% CIs)			Odds ratio (95% CIs)		
Blood 1-palmitoylglycerophosphoethanolamine	0.314	0.90	0.73 - 1.10	0.90	0.73 - 1.10		NA	NA - NA		NA	NA - NA	
Putamen volume	0.928	0.99	0.74 - 1.31	0.99	0.80 - 1.22		NA	NA - NA		NA	NA - NA	
Blood stearate (18:0)	0.932	1.02	0.67 - 1.54	1.02	0.82 - 1.26		NA	NA - NA		NA	NA - NA	
Serum IgE	0.936	0.99	0.81 - 1.21	0.99	0.81 - 1.21		NA	NA - NA		0.50	0.16 - 1.50	
Blood 1-linoleoylglycerol (1-monolinolein)	0.685	NA	NA - NA	NA	NA - NA		0.95	0.73 - 1.22		NA	NA - NA	
Gamma-linoleic acid (18:3n6)	0.01798 *	0.84	0.73 - 0.97	0.84	0.73 - 0.97		NA	NA - NA		NA	NA - NA	
Blood phenyllactate (PLA)	0.220	NA	NA - NA	NA	NA - NA		0.85	0.66 - 1.10		NA	NA - NA	
Blood O-sulfo-L-tyrosine	0.300	0.90	0.73 - 1.10	0.90	0.73 - 1.10		NA	NA - NA		NA	NA - NA	
Birth weight of first child	0.322	1.13	0.89 - 1.44	1.13	0.90 - 1.42		NA	NA - NA		0.96	0.39 - 2.38	
Blood tetradecadienoate	0.985	NA	NA - NA	NA	NA - NA		1.00	0.81 - 1.25		NA	NA - NA	
Blood 2-aminobutyrate	0.138	0.85	0.68 - 1.06	0.85	0.68 - 1.06		NA	NA - NA		NA	NA - NA	
Blood albumin	0.816	1.05	0.69 - 1.61	1.05	0.69 - 1.61		NA	NA - NA		0.61	0.09 - 4.24	
Plasma IGF-I	0.687	NA	NA - NA	NA	NA - NA		0.94	0.71 - 1.25		NA	NA - NA	
Blood dehydroisoandrosterone sulfate (DHEA-S)	0.832	0.98	0.77 - 1.23	0.98	0.77 - 1.23		NA	NA - NA		NA	NA - NA	
Blood alanine	0.837	0.97	0.72 - 1.30	0.97	0.77 - 1.23		NA	NA - NA		0.45	0.13 - 1.62	

Exposure	IVW-RE *P*-value		IVW-RE Odds ratio (95% CIs)	IVW-FE Odds ratio (95% CIs)	Wald Ratio Odds ratio (95% CIs)	Causal estimate from MR-Egger slope Odds ratio (95% CIs)
Morning/evening person (chronotype)	0.725		1.05 0.82 - 1.34	1.05 0.82 - 1.33	NA NA - NA	NA NA - NA
Gamma-linolenic acid (18:3n6)	0.01520	*	0.86 0.76 - 0.97	0.86 0.76 - 0.97	NA NA - NA	NA NA - NA
Blood valine	0.583		1.08 0.81 - 1.45	1.08 0.86 - 1.37	NA NA - NA	3.24 1.05 - 10.04
Blood myo-inositol	0.716		1.08 0.71 - 1.64	1.08 0.86 - 1.35	NA NA - NA	NA NA - NA
Linoleic acid (18:2n6)	0.01128	*	1.10 1.02 - 1.18	1.10 1.02 - 1.18	NA NA - NA	NA NA - NA
Blood creatine	0.189		NA NA - NA	NA NA - NA	0.84 0.65 - 1.09	NA NA - NA
Oleic acid (18:1n9)	0.00564	*	NA NA - NA	NA NA - NA	1.34 1.09 - 1.66	NA NA - NA
Blood N2,N2-dimethylguanosine	0.481		1.09 0.86 - 1.39	1.09 0.86 - 1.39	NA NA - NA	NA NA - NA
Blood histidine	0.751		0.94 0.65 - 1.36	0.94 0.72 - 1.23	NA NA - NA	3.35 0.54 - 20.71
Insulin disposition index	0.664		NA NA - NA	NA NA - NA	0.95 0.74 - 1.21	NA NA - NA
Omega-3 fatty acids	0.00054	*	0.74 0.62 - 0.88	0.74 0.62 - 0.88	NA NA - NA	0.59 0.31 - 1.11
Serum vitamin B6	0.04130	*	NA NA - NA	NA NA - NA	1.26 1.01 - 1.58	NA NA - NA
Iron status	0.232		1.24 0.87 - 1.78	1.24 0.96 - 1.61	NA NA - NA	1.57 0.99 - 2.50
Blood levulinate (4-oxovalerate)	0.993		1.00 0.55 - 1.80	1.00 0.78 - 1.28	NA NA - NA	NA NA - NA
Blood creatinine	0.977		1.01 0.70 - 1.44	1.01 0.78 - 1.30	NA NA - NA	0.56 0.06 - 4.80

Exposure	*P*-value		IVW-RE Odds ratio (95% CIs)				IVW-FE Odds ratio (95% CIs)				Wald Ratio Odds ratio (95% CIs)				Causal estimate from MR-Egger slope Odds ratio (95% CIs)			
Blood glucose	0.670		0.94	0.73	-	1.23	0.94	0.73	-	1.23	NA	NA	-	NA	0.97	0.30	-	3.08
Neuroticism score	0.358		1.15	0.85	-	1.56	1.15	0.88	-	1.50	NA	NA	-	NA	NA	NA	-	NA
Plasma estradiol	0.562		NA	NA	-	NA	NA	NA	-	NA	0.90	0.63	-	1.29	NA	NA	-	NA
Stearoylcarnitine	0.04456	*	NA	NA	-	NA	NA	NA	-	NA	1.41	1.01	-	1.98	NA	NA	-	NA
Phenylalanine	0.419		0.87	0.61	-	1.23	0.87	0.64	-	1.17	NA	NA	-	NA	0.33	0.03	-	3.48
Indoleacetate	0.611		1.08	0.80	-	1.46	1.08	0.80	-	1.46	NA	NA	-	NA	NA	NA	-	NA
Years of schooling	0.055		0.72	0.52	-	1.01	0.72	0.54	-	0.96	NA	NA	-	NA	NA	NA	-	NA
4-methyl-2-oxopentanoate	0.707		1.05	0.80	-	1.38	1.05	0.80	-	1.38	NA	NA	-	NA	NA	NA	-	NA
Lysine	0.849		NA	NA	-	NA	NA	NA	-	NA	0.97	0.74	-	1.28	NA	NA	-	NA
Hypoxanthine	0.495		NA	NA	-	NA	NA	NA	-	NA	1.10	0.84	-	1.44	NA	NA	-	NA
1,6-anhydroglucose	0.900		NA	NA	-	NA	NA	NA	-	NA	1.02	0.76	-	1.36	NA	NA	-	NA
Gamma-glutamylphenylalanine	0.516		1.10	0.82	-	1.49	1.10	0.82	-	1.49	NA	NA	-	NA	NA	NA	-	NA
Acetylphosphate	0.496		1.11	0.82	-	1.50	1.11	0.84	-	1.47	NA	NA	-	NA	NA	NA	-	NA
Heptanoate (7:0)	0.669		0.91	0.61	-	1.38	0.91	0.69	-	1.21	NA	NA	-	NA	NA	NA	-	NA
Stearidonate (18:4n3)	0.119		NA	NA	-	NA	NA	NA	-	NA	0.77	0.56	-	1.07	NA	NA	-	NA

Exposure	*P*-value		IVW-RE Odds ratio (95% CIs)				IVW-FE Odds ratio (95% CIs)				Wald Ratio Odds ratio (95% CIs)				Causal estimate from MR-Egger slope Odds ratio (95% CIs)			
Myristoleate (14:1n5)	0.243		NA	NA	-	NA	NA	NA	-	NA	0.84	0.63	-	1.13	NA	NA	-	NA
Caprylate (8:0)	0.513		1.12	0.80	-	1.57	1.12	0.80	-	1.57	NA	NA	-	NA	NA	NA	-	NA
Laurate (12:0)	0.680		1.07	0.78	-	1.46	1.07	0.78	-	1.46	NA	NA	-	NA	NA	NA	-	NA
HOMA-B	0.533		0.91	0.67	-	1.23	0.91	0.67	-	1.23	NA	NA	-	NA	1.67	0.49	-	5.63
Margarate (17:0)	0.243		NA	NA	-	NA	NA	NA	-	NA	1.20	0.88	-	1.63	NA	NA	-	NA
Nap during day	0.190		1.23	0.90	-	1.68	1.23	0.90	-	1.68	NA	NA	-	NA	NA	NA	-	NA
Palmitoleate (16:1n7)	0.243		NA	NA	-	NA	NA	NA	-	NA	0.83	0.61	-	1.13	NA	NA	-	NA
Tamm-Horsfall urinary glycoprotein	0.03723	*	NA	NA	-	NA	NA	NA	-	NA	0.94	0.88	-	1.00	NA	NA	-	NA
Time spent watching television (TV)	0.242		1.23	0.87	-	1.75	1.23	0.90	-	1.70	NA	NA	-	NA	NA	NA	-	NA
Telomere length	0.01263	*	2.33	1.20	-	4.52	2.33	1.81	-	2.99	NA	NA	-	NA	NA	NA	-	NA
Laurylcarnitine	0.349		NA	NA	-	NA	NA	NA	-	NA	1.18	0.83	-	1.68	NA	NA	-	NA
3,4-dihydroxybutyrate	0.638		NA	NA	-	NA	NA	NA	-	NA	0.93	0.68	-	1.27	NA	NA	-	NA
Indolelactate	0.496		NA	NA	-	NA	NA	NA	-	NA	0.90	0.65	-	1.23	NA	NA	-	NA
2hr glucose	0.463		0.89	0.65	-	1.21	0.89	0.67	-	1.18	NA	NA	-	NA	NA	NA	-	NA
Docosahexaenoic acid (22:6n3)	0.984		NA	NA	-	NA	NA	NA	-	NA	1.00	0.73	-	1.38	NA	NA	-	NA

Exposure	IVW-RE			IVW-FE			Wald Ratio			Causal estimate from MR-Egger slope		
	P-value	Odds ratio (95% CIs)			Odds ratio (95% CIs)			Odds ratio (95% CIs)			Odds ratio (95% CIs)		
3-(4-hydroxyphenyl)lactate	0.876	NA	NA -	NA	NA	NA -	NA	1.03	0.71 -	1.48	NA	NA -	NA
Serum vitamin A1	0.306	1.18	0.86 -	1.61	1.18	0.87 -	1.60	NA	NA -	NA	NA	NA -	NA
1-oleoylglycerol (1-monoolein)	0.886	NA	NA -	NA	NA	NA -	NA	1.03	0.70 -	1.50	NA	NA -	NA
1-stearoylglycerophosphoethanolamine	0.984	NA	NA -	NA	NA	NA -	NA	1.00	0.73 -	1.39	NA	NA -	NA
LOY	0.101	1.45	0.93 -	2.25	1.45	1.02 -	2.05	NA	NA -	NA	NA	NA -	NA
N1-methyl-3-pyridone-4-carboxamide	0.908	NA	NA -	NA	NA	NA -	NA	1.02	0.70 -	1.49	NA	NA -	NA
Serum vitamin E	0.536	0.89	0.62 -	1.28	0.89	0.62 -	1.28	NA	NA -	NA	1.23	0.10 -	14.93
Fasting insulin	0.231	1.35	0.83 -	2.19	1.35	0.96 -	1.89	NA	NA -	NA	1.30	0.13 -	13.06
Hippocampus volume	0.540	1.35	0.52 -	3.48	1.35	0.92 -	1.96	NA	NA -	NA	NA	NA -	NA
Dodecanedioate	0.423	NA	NA -	NA	NA	NA -	NA	1.15	0.81 -	1.64	NA	NA -	NA

Appendix 23

Causal estimates from weighted median estimates and mode-based estimates for suggestively significant traits and MM risk (tables overleaf). * Indicates P < 0.05. HDL, high density lipoprotein; LDL, low density lipoprotein; LOY, loss of Y chromosome

		Weighted Median				Weighted Mode			
Exposure	*P*-value	Odds ratio (95% CIs)		*P*-value		Odds ratio (95% CIs)		*P*-value	
Blood carnitine	0.00114	1.10	1.00 - 1.22	0.0567		1.09	0.95 - 1.25	0.2276	
Blood decanoylcarnitine	0.01291	1.15	1.01 - 1.31	0.0412	*	1.14	1.00 - 1.31	0.1496	
Blood hexanoylcarnitine	0.00695	1.13	1.01 - 1.27	0.0371	*	1.12	0.99 - 1.27	0.1736	
Blood isovalerylcarnitine	0.00252	1.14	0.98 - 1.32	0.0928		1.15	0.98 - 1.37	0.1941	
Blood octanoylcarnitine	0.01384	1.13	1.01 - 1.27	0.0345	*	1.13	0.99 - 1.28	0.2062	
Blood propionylcarnitine	0.01273	1.10	0.94 - 1.29	0.2424		1.09	0.92 - 1.29	0.3729	
Docosapentaenoic acid (22:5n3)	0.03732	0.90	0.83 - 0.99	0.0270	*	0.90	0.82 - 0.99	0.1577	
Forced vital capacity (FVC)	0.00791	1.23	1.03 - 1.47	0.0214	*	1.28	0.83 - 1.97	0.2667	
Omega-3 fatty acids	0.00054	0.71	0.58 - 0.88	0.0013	*	0.71	0.57 - 0.88	0.0270	*
Telomere length	0.01263	1.55	0.95 - 2.52	0.0771		1.06	0.69 - 1.64	0.7959	

Appendix 25

Results of MR-Egger analysis of potential bias in causal estimates (tables overleaf). * Effects estimated using MM summary statistics from females only. HDL, high density lipoprotein; LDL, low density lipoprotein; LOY, loss of Y chromosome

			MR-Egger intercept		
Category	Exposure	No. of SNPs	SE	*P*-value	Intercept
Developmental and growth factors	Height	2481	0.001	0.303	-0.001
Developmental and growth factors	Putamen volume	4	0.046	0.205	0.086
Diet and lifestyle	Heel bone mineral density (BMD) T-score	409	0.003	0.646	-0.001
Diet and lifestyle	Serum vitamin B12	9	0.031	0.445	0.025
Diet and lifestyle	Fasting glucose	23	0.014	0.589	-0.008
Diet and lifestyle	Pulse rate	59	0.010	0.869	0.002
Diet and lifestyle	Circulating 25-hydroxyvitamin D	5	0.029	0.364	0.030
Diet and lifestyle	Serum calcium	7	0.016	0.239	0.021
Diet and lifestyle	Iron status	3	0.029	0.411	-0.039
Diet and lifestyle	Blood creatinine	6	0.080	0.615	0.043
Diet and lifestyle	Serum vitamin E	3	0.108	0.841	-0.027
Diet and lifestyle	Fasting insulin	14	0.041	0.977	0.001
Genome Stability	Telomere length	7	0.092	0.384	-0.088
Genome Stability	LOY	92	0.008	0.655	-0.004
Inflammatory factors	Circulating C-reactive protein	14	0.014	0.330	-0.014

			MR-Egger intercept		
Category	Exposure	No. of SNPs	SE	*P*-value	Intercept
Fatty acid profile and metabolism	Blood butyrylcarnitine	9	0.018	0.380	0.017
Fatty acid profile and metabolism	Blood N-acetylornithine	4	0.028	0.687	-0.013
Fatty acid profile and metabolism	Blood glycoproteins	28	0.009	0.326	0.009
Fatty acid profile and metabolism	Blood carnitine	18	0.012	0.449	0.009
Fatty acid profile and metabolism	Blood biliverdin	3	0.022	0.954	-0.002
Fatty acid profile and metabolism	Blood glycine	6	0.018	0.598	-0.011
Fatty acid profile and metabolism	Blood succinylcarnitine	7	0.023	0.795	-0.006
Fatty acid profile and metabolism	Blood glutaroyl carnitine	9	0.036	0.153	0.058
Fatty acid profile and metabolism	Blood 2-aminooctanoic acid	3	0.024	0.569	0.019
Fatty acid profile and metabolism	Docosapentaenoic acid (22:5n3)	3	0.038	0.820	-0.011
Fatty acid profile and metabolism	Blood apolipoprotein B	27	0.010	0.319	0.010
Fatty acid profile and metabolism	Blood tryptophan	19	0.059	0.283	-0.065
Fatty acid profile and metabolism	Blood bradykinin, des-arg(9)	3	0.045	0.353	0.072
Fatty acid profile and metabolism	Blood androsterone sulfate	4	0.017	0.824	-0.004
Fatty acid profile and metabolism	Blood hexanoylcarnitine	4	0.022	0.324	0.029
Fatty acid profile and metabolism	Blood proline	4	0.023	0.665	0.012
Fatty acid profile and metabolism	Blood tetradecanedioate	3	0.052	0.629	-0.034
Fatty acid profile and metabolism	Blood alpha-glutamyltyrosine	3	0.038	0.942	-0.004
Fatty acid profile and metabolism	Blood octanoylcarnitine	3	0.033	0.399	0.046
Fatty acid profile and metabolism	Blood HDL diameter	10	0.037	0.560	-0.022
Fatty acid profile and metabolism	Blood decanoylcarnitine	4	0.029	0.613	0.017
Fatty acid profile and metabolism	Blood 5-alpha-pregnan-3beta,20alpha-disulfate	4	0.102	0.474	-0.089

			MR-Egger intercept		
Category	Exposure	No. of SNPs	SE	*P*-value	Intercept
Fatty acid profile and metabolism	Blood leucine	11	0.038	0.946	0.003
Fatty acid profile and metabolism	Blood hexadecanedioate	3	0.055	0.714	-0.027
Fatty acid profile and metabolism	Blood 10-undecenoate (11:1n1)	3	0.052	0.758	-0.021
Fatty acid profile and metabolism	Blood isovalerylcarnitine	4	0.129	0.553	-0.091
Fatty acid profile and metabolism	Blood isobutyrylcarnitine	3	0.031	0.775	-0.011
Fatty acid profile and metabolism	Blood N-[3-(2-Oxopyrrolidin-1-yl)propyl]acetamide	5	0.040	0.673	-0.019
Fatty acid profile and metabolism	Blood phosphatidylcholine and other cholines	10	0.026	0.985	-0.001
Fatty acid profile and metabolism	Blood kynurenine	4	0.023	0.893	-0.004
Fatty acid profile and metabolism	Blood propionylcarnitine	5	0.026	0.266	0.036
Fatty acid profile and metabolism	Blood N-acetylglycine	3	0.032	0.443	0.038
Fatty acid profile and metabolism	Total phosphoglycerides	10	0.027	0.639	-0.013
Fatty acid profile and metabolism	Eicosapentaenoic acid (20:5n3)	5	0.031	0.607	0.018
Fatty acid profile and metabolism	VLDL diameter	11	0.052	0.987	0.001
Fatty acid profile and metabolism	Blood 5alpha-androstan-3beta,17beta-diol disulfate	4	0.029	0.759	-0.010
Fatty acid profile and metabolism	Blood betaine	5	0.047	0.244	-0.068
Fatty acid profile and metabolism	Blood N-acetylcarnosine	3	0.199	0.924	-0.024
Fatty acid profile and metabolism	Blood sphingomyelins	9	0.021	0.110	0.039
Fatty acid profile and metabolism	Blood serine	3	0.123	0.593	0.092
Fatty acid profile and metabolism	Blood 1,5-anhydroglucitol (1,5-AG)	3	0.056	0.443	-0.067
Fatty acid profile and metabolism	Blood citrate	6	0.030	0.732	-0.011
Fatty acid profile and metabolism	Omega-9 and saturated fatty acids	7	0.036	0.738	0.013
Fatty acid profile and metabolism	Glycoprotein acetyls	10	0.037	0.873	-0.006

			MR-Egger intercept		
Category	Exposure	No. of SNPs	SE	*P*-value	Intercept
Fatty acid profile and metabolism	Mono-unsaturated fatty acids	7	0.040	0.661	0.019
Fatty acid profile and metabolism	Blood alpha-hydroxyisovalerate	3	0.065	0.393	0.091
Fatty acid profile and metabolism	Omega-3 fatty acids	6	0.039	0.503	0.029
Fatty acid profile and metabolism	Omega-7 and -9 and saturated fatty acids	6	0.040	0.930	0.004
Fatty acid profile and metabolism	Monounsaturated fatty acids	6	0.072	0.901	-0.010
Fatty acid profile and metabolism	Palmitoleic acid (16:1n7)	5	0.089	0.658	-0.044
Fatty acid profile and metabolism	LDL diameter	5	0.041	0.893	0.006
Fatty acid profile and metabolism	Blood gamma-glutamyltyrosine	5	0.092	0.163	-0.169
Fatty acid profile and metabolism	Blood 2-hydroxyisobutyrate	3	0.067	0.689	0.035
Fatty acid profile and metabolism	Stearic acid (18:0)	3	0.208	0.221	0.575
Fatty acid profile and metabolism	Blood glutamine	6	0.032	0.631	0.016
Fatty acid profile and metabolism	Blood gamma-glutamylglutamine	3	0.046	0.247	0.114
Fatty acid profile and metabolism	Blood 3-methyl-2-oxovalerate	3	0.129	0.619	-0.088
Fatty acid profile and metabolism	Blood citrulline	4	0.201	0.681	-0.096
Fatty acid profile and metabolism	Blood tyrosine	3	0.139	0.337	-0.237
Fatty acid profile and metabolism	Blood uridine	3	0.046	0.263	-0.106
Fatty acid profile and metabolism	Blood albumin	4	0.161	0.627	0.091
Fatty acid profile and metabolism	Blood alanine	6	0.055	0.294	0.066
Fatty acid profile and metabolism	Blood valine	5	0.051	0.148	-0.100
Fatty acid profile and metabolism	Blood histidine	5	0.080	0.259	-0.111
Fatty acid profile and metabolism	Blood glucose	3	0.058	0.977	-0.002
Fatty acid profile and metabolism	Phenylalanine	4	0.095	0.500	0.077

			MR-Egger intercept		
Category	Exposure	No. of SNPs	SE	*P*-value	Intercept
Lipids and lipid transport	LDL	102	0.004	0.343	0.004
Lipids and lipid transport	Total cholesterol	123	0.005	0.093	0.008
Lipids and lipid transport	HDL	124	0.005	0.465	0.003
Lipids and lipid transport	Triglycerides	70	0.005	0.096	0.009
lipids and lipid transport	Total triglycerides	34	0.006	0.613	0.003
Lipids and lipid transport	Omega-6 fatty acids	13	0.020	0.394	0.018
Lipids and lipid transport	Blood apolipoprotein A-I	12	0.026	0.215	-0.034
Lipids and lipid transport	Total fatty acids	12	0.033	0.719	0.012
Lipids and lipid transport	Circulating adiponectin	10	0.020	0.496	-0.014
Miscellaneous	Fasting proinsulin	8	0.021	0.708	-0.008
Miscellaneous	Forced vital capacity (FVC)	284	0.005	0.077	-0.009
Miscellaneous	Platelet count	39	0.016	0.832	0.003
Miscellaneous	Corrected insulin response	3	0.074	0.608	0.052
Miscellaneous	HbA1C levels	11	0.017	0.379	0.016
Miscellaneous	Fluid intelligence score	50	0.018	0.687	-0.007
Miscellaneous	Serum IgE	3	0.084	0.431	0.104
Miscellaneous	Morning/evening person (chronotype)	99	0.008	0.337	0.008
Miscellaneous	Neuroticism score	78	0.016	0.630	0.008
Miscellaneous	Years of schooling	74	0.013	0.474	-0.009
Miscellaneous	HOMA-B	4	0.043	0.418	-0.043
Miscellaneous	Nap during day	58	0.011	0.923	-0.001
Miscellaneous	Time spent watching television (TV)	65	0.015	0.831	-0.003
Miscellaneous	2hr glucose	7	0.034	0.862	0.006

			MR-Egger intercept		
Category	Exposure	No. of SNPs	SE	*P*-value	Intercept
Obesity	Whole body water mass	735	0.003	0.379	0.002
Obesity	Basal metabolic rate	693	0.003	0.248	0.003
Obesity	Weight	576	0.003	0.583	0.002
Obesity	Impedance of whole body	564	0.003	0.458	-0.003
Obesity	Body mass index	964	0.002	0.701	-0.001
Obesity	Whole body fat mass	415	0.004	0.746	-0.001
Obesity	Body fat percentage	365	0.005	0.707	0.002
Obesity	Trunk fat percentage	334	0.005	0.203	0.007
Obesity	Waist circumference	316	0.005	0.984	0.000
Obesity	Birth weight	93	0.009	0.548	-0.006
Obesity	Hip circumference	89	0.011	0.252	0.013
Obesity	Waist-to-hip ratio	35	0.019	0.392	-0.016
Obesity	Birth weight of first child	45	0.014	0.711	0.005
Plasma analytes	Apolipoprotein H	3	0.534	0.472	-0.583
Sex hormones and reproduction	Age at menopause*	48	0.040	0.590	0.022
Sex hormones and reproduction	Age at menarche*	73	0.022	0.090	-0.038