Detection, causes and consequences of sex chromosome mosaicism

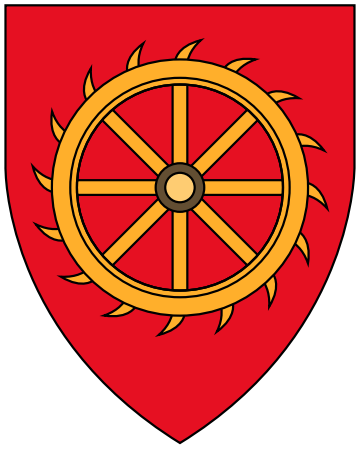
This thesis is submitted for the degree of Doctor of Philosophy

Yajie Zhao

MRC Epidemiology Unit

& St Catharine's College

August 2022



Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the Degree Committee of Clinical Medicine and Clinical Veterinary Medicine.

Abstract

Detection, causes and consequences of sex chromosome mosaicism

Yajie Zhao

Sex chromosome mosaicism, including male mosaic loss of chromosome Y (LOY) and female mosaic loss of chromosome X (LOX), is the most common form of clonal haematopoiesis (CH) that can be defined as the age-related clonal expansion of blood cells with somatic mutations. With the decreased cost of sequencing, the development of new bioinformatics methods and the emergence of large cohorts with both genotype and phenotype data, there has been much progress in the detection, causes and consequences of sex chromosome mosaicism especially LOY. On the contrary, the studies of LOX are still very limited.

The most recent genome-wide association study (GWAS) to investigate the genetic determinants of LOY in 205,011 males identified 156 independent signals and highlighted a key role for genes involved in cell-cycle regulation and DNA damage response. Population studies typically determine LOY using genotype intensities derived from genotype array data, the accuracy of which varies by the number of Y chromosome probes on the array and are technically noisy. Inaccurate estimation of LOY reduces the power to identify genetic and phenotypic associations with LOY. To overcome these constraints, I developed a robust estimator of LOY , which was derived from several orthogonal approaches using both whole exome sequence and genotype array data. The same method was also implemented for LOX.

In chapter 2, in genetic data derived from 205,604 UK Biobank males and 243,765 females, the new method improved the accuracy of LOY/LOX estimation as measured by the strength of association between LOY/LOX and age, smoking status, and polygenic risk of LOY/LOX derived from previous GWAS. In chapter 3, I then used this revised and validated LOY instrument to conduct a new GWAS of LOY status in the UK Biobank. Beyond the previously

identified 156 signals, I identified 22 novel LOY-associated loci. I leveraged the shared genetic architecture between LOY and other related traits to improve the power to identify variants associated with the risk of myeloproliferative neoplasm (MPN). Based on available MPN GWAS summary statistics, I identified 13 novel loci reaching genome-wide significance, including locus near *PARP1*, which encodes an established target of cancer therapy.
The method used to detect somatic LOY/LOX can also be used to identify congenital sex chromosome abnormalities. In chapter 4, the detection method and characterisation of male sex chromosome abnormalities were reported, as a similar study on female sex chromosome abnormalities in UK Biobank had already been published.

The whole exome sequence also provided the chance to explore the effect of rare non-synonymous variants, which are rarely captured by GWAS arrays. In chapter 5, the first exome-wide association study (ExWAS) for LOY was conducted on over 80,000 men from UK Biobank. As well as *CHEK2*, which had been identified on a previous GWAS on LOY, a novel gene, *GIGYF1*, was identified, in which loss-of-function variants increased the risk of LOY and Type 2 diabetes (T2D) by 6-fold. This finding illuminated the potential link between LOY and metabolism. In chapter 6, the first ExWAS of LOX and an extended ExWAS of LOY were performed on over 450,000 samples from UK Biobank. For LOY, the power increase was observed for the two identified genes, *CHEK2* and *GIGYF1*. In addition, loss of function variants in three clonal haematopoiesis of indeterminate potential (CHIP) genes, *DNMT3A*, *TET2* and *ASXL1*, were negatively associated with LOY. For LOX, rare damaging variants in *FBXO10* were identified to increase the risk of LOX.

In summary, this thesis shows that the accuracy of estimating LOY/LOX was improved by combining multiple approaches using both GWAS array and whole exome sequence data. Using both GWAS and ExWAS approaches, the thesis further improved the understanding of the genetic causes of sex chromosome mosaicism. This innovative approach improved the power to detect novel mechanisms that regulate clonal mosaicism in blood and can be used to enhance the identification of novel genes associated with the risk of related cancers.

Acknowledgement

At the end of this splendid journey and the start of the new adventure, I want to thank my best supervisors Prof John Perry and Prof Ken Ong and my kindest mentors Dr Hana Lango Allen and Dr Eugene Gardner. Without their unconditional and endless guidance, support and help during these three years, I could not make any achievements I have made.

Many thanks to my examiners, Prof Alison Dunning and Prof Vijay Sankaran. Their insightful comments on this thesis provided many potential directions for my future research. It's also my great privilege that my work could be recognised by them.

I want also to thank my parents and my girlfriend, Miss Mingyan Xin, whose emotional and financial support has been the powerhouse in the journey of my PhD. Many thanks to the care and support from Prof Yi Rao and Mrs Jianjin Dong when I was in China during the pandemic time.

Many thanks to my fantastic colleagues in our group and MRC Epidemiology Unit, every time I need help and support from them, they will always offer the best they can. Many thanks to my collaborators from all over the world, I am looking forward to working with you on more exciting projects in the future. Many thanks to all participants from UK Biobank, without their generous decision to participate in this great cohort, there is no possibility to finish any studies described in this thesis.

As a non-native speaker, for me, writing such a long thesis was a great challenge. Special thanks to my colleague Raina Jia and my roommate Rosie Good, who helped me carefully proofread this thesis.

Special thanks to my cutest cat in the world, Peach.

Finally, I want to thank all who I love and who love me.

In Science We Trust.

Table of Contents

Tables and figures

List of Abbreviations

ACAT-V	set-based Aggregated Cauchy Association Test
AD	individual genotype call-level Allelic Depths for the ref and alt alleles in the order listed
AF-LOY/LOX	Affected cell Fractions of LOY/LOX estimated from MoChA
AML	Acute Myeloid Leukaemia
AQ	variant site-level Allele Quality score reflecting evidence for each alternate allele, Phred scale
AWS	Amazon Web Services
BAF	B-allele frequency
BAM	Binary Alignment Map
BED	Browser Extensible Data
BMD	Bone Mineral Density
BMI	Body Mass Index
CADD	Combined Annotation Dependent Depletion
CH	clonal hematopoiesis
CHIP	Clonal hematopoiesis of indeterminate potential; CHIP
Coloc-ABF	Bayesian colocalisation analysis using Bayes Factors
COPD	Chronic Obstructive Pulmonary Disease
DDR	DNA Damage Response
DMG	Damaging variants, the combination of High Confidence protein-truncating variant and Missense variants with CADD scores ≥ 25
DP	individual genotype call-level approximate read DePth
eQTL	expression Quantitative Trait Loci
ExWAS	Exome Wide gene burden Association Study
FDR	False Discovery Rate
GC	Genomic control
GLM	Generalised Linear Model
GQ	individual genotype call-level Genotype Quality, Phred scale
GRCh37	Genome Reference Consortium Human Build 37
GRCh38	Genome Reference Consortium Human Build 38
GRM	Genetic Relationship Matrix
GRS	Genetic Risk Score
GTEx	Genotype-Tissue Expression
GWAS	Genome Wide Association Study
HC_PTV	High Confidence Protein-Truncating Variant
HPC	High performance computing
HRC	Haplotype Reference Consortium
HSCs	Hematopoietic Stem Cells
HSPCs	Haematopoietic Stem and Progenitor Cells

HTS	High-Throughput Sequencing
InDels	Insertion or Deletions
IV	Instrumental Variable
KEGG	Kyoto Encyclopedia of Genes and Genomes
KO	Knock Out
KS	Klinefelter Syndrome
LDSC	Linkage Disequilibrium Score Regression
LDSC-SEG	LD SCore regression applied to Specifically Expressed Genes
LOFTEE	Loss-Of-Function Transcript Effect Estimator
LOX	mosaic loss of chromosome X
LOY	mosaic loss of chromosome Y
LRR	log2-transformed R ratio
LTL	Leukocyte Telomere Length
MAF	Minor Allele Frequency
MAGMA	Multi-marker Analysis of GenoMic Annotation
MANE	Matched Annotation from NCBI and EMBL-EBI
mCAs	mosaic Chromosomal Alterations
MDS	MyeloDysplastic Syndromes
MISS_CADD25	MISSense variant with CADD score ≥ 25
mLRR-Y	median or mean of log R ratio values of the probes on MSY
MoChA	MOsaic CHromosomal Alterations
MoChA-LOX	LOX status estimated from MoChA
MPN	MyeloProliferative Neoplasms
MR	Mendelian Randomisation
MR-EGGER	Mendelian randomization-Egger test
MR-IVW	Inverse-Variance Weighted MR
mRNA	Messenger RNA
MR-PWM	Penalized Weighted Median MR
MR-WM	Weighted Median MR
MSY	Male-Specific Regions
MTAG	Multi-Trait Analysis of GWAS
NMR	Nuclear Magnetic Resonance
OR	Odds Ratio
PAR	Pseudo-Autosomal Regions
PAR-LOY	LOY status estimated from MoChA
PC	Principal Component
PheWAS	Phenome-Wide Association Study
POLYPHEN	Polymorphism Phenotyping
PoPS	Polygenic Priority Score
PTVs	Protein-Truncating Variants
QC	Quality Control

qPCR	quantitative Polymerase Chain Reaction
QUAL	variant site-level QUALity core, Phred scale
RAP	UK Biobank Research Analysis Platform
REVEL	Rare Exome Variant Ensemble Learner
RR	Relative Risk
SCAVENGE	Single Cell Analysis of Variant Enrichment through Network propagation of GEnomic data
scDRS	single-cell Disease-Relevance Score
SIFT	Sorting Intolerant From Tolerant
SKAT	Sequence Kernel Association Test
SMR-HEIDI	Summary-data-based Mendelian randomisation (SMR) and HEterogeneity In Dependent Instruments (HEIDI) tests
SNP	Single-nucleotide polymorphism
SNVs	Single-Nucleotide Variants
STAAR	variant-set test for association using annotation information
SVs	Structural Variations
T2D	Type 2 Diabetes
TOPMed	The Trans-Omics for Precision Medicine program
VAF	Variant Allele Fraction
VCF	Variant Call Format
VEP	Ensembl Variant Effect Predictor
WDL	Workflow Description Language
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
WP	WikiPathways
XDR	X-degenerate regions

Chapter 1 Introduction

1.1 Sex chromosomes

Most males and females carry 22 pairs of autosomal chromosomes and 1 pair of sex chromosomes. Same as other mammals, the pair of human sex chromosomes is male heterogamety (males XY, females XX)[1]. However, inborn sex chromosome abnormalities could occur as various forms, such as male 47,XXY (carrying one extra X chromosome) - also known as the Klinefelter syndrome, 47, XYY (carrying one extra Y chromosome), 48, XXYY (carrying one extra pair XY chromosome), female 45, XO (losing one X chromosome) - also known as the Turner syndrome, 47, XXX (carrying one extra X chromosome) and various forms of sex chromosome mosaicism[2]. When compared to people with normal sex chromosomes, carriers of abnormal karyotypes can be affected by various health consequences ranging from reproductivity to intelligence[2]. However, the understanding of the mechanisms underlying these health burdens is still very limited.

There have been many studies illustrating that human sex chromosomes differ from autosomes in many aspects of genome biology, such as organization, gene content and gene expression[1]. The sex chromosome pair shows an extreme imbalance in size and function, which is different from the autosome pairs. The X chromosome is large (165 Mb) and gene-rich (about 1000 genes with diverse general and specialised functions), but the Y chromosome is small (~60 Mb) and heterochromatic (most transcribed units[3] are pseudogenes or amplified copies)[4]. From the evolutionary perspective, the Y chromosome can be seen as the degraded X chromosome, as 20 of the 27 genes on the male-specific regions of the human Y chromosome evolved from the genes on the X chromosome[4,5]. The most important function of the sex chromosomes is sex determination, especially for the Y chromosome. Over 30 years ago, the gene in the Y chromosome, *SRY*, which encodes the testis-determining factor was first identified by Sinclair et al[6]. This illuminated an essential genetic pathway which determines the sex of a new-born[5]. Unlike other chromosomes, not just *SRY*, the functions of many other genes on the Y chromosome converge to affect the sex or fertility of males[4,7]. Because of its characteristics, much of the Y chromosome has been thought to be “functional wasteland” which will completely disappear from human

genome[8,9]. However, this view has been challenged because several ubiquitously expressed Y-chromosome genes have been identified and multiple association between the Y chromosome, immune system and complex polygenic traits including coronary artery disease were discovered[8].

Females, unlike males, inherit one X chromosome from each parent. During the early developmental stage, one copy of the X chromosomes partially becomes transcriptionally inactive[10]. The inactivation process is random and irreversible and then transferred to the daughter cells[10]. One of the mechanisms to explain this is to compensate for the gene dosage imbalance of sex chromosomes of 46,XX females and 46,XY males[11]. However, with increasing age, the expected 1:1 ratio of the inactivated maternal and paternal X chromosome copies shows skewness, and this skewness was discovered in different tissues[12,13]. The detectable skewness of X chromosome inactivation in white blood cells might be an indicator of the depletion of hematopoietic stem cells (HSCs), selective pressure of white blood cells, or clonal hematopoiesis.

1.2 Clonal haematopoiesis

The age-related accumulations of postzygotic DNA mutations resulting in tissue genetic heterogeneity are known as somatic mosaicism, which arises because of errors in the repair or replication of damaged DNA[14]. The concept of genome mosaicism dates back more than a century ago, when several scientists speculated that cancer was a somatic mosaic by genetic alterations[14]. 40 years since then, two scientists proposed that cancer and ageing were consequences of the accumulation of *de novo* somatic mutations. After another two decades, this concept was proved to be correct for cancer[14]. Since then, most of the studies on somatic mosaicism have focused on cancer. However, the roles of somatic mosaicism in rare diseases and complex traits, especially ageing and ageing-related diseases, are still unclear.

Somatic mosaicism in ageing has been identified since the 1950s[14–16]. However, somatic mosaicism in normal tissue has been difficult to study as the mutation events are random and rare, and the fractions of affected cells are so low that it is hard to be detected[14,15,17].

In this thesis, I mainly focused on the age-related accumulated somatic mutations in blood, which is called clonal haematopoiesis (CH) and can be detected from SNP-array used for genome wide association studies (GWAS) or high-throughput sequencing (HTS) data[18]. The types of clonal haematopoiesis include the gain, loss or recombination resulting in the loss of heterozygosity of large pieces of one or more chromosomes (mCAs), mosaic loss of sex chromosomes (LOY/LOX) and the somatic mutations occurring in myeloid-associated genes at ≥ 2% variant allele fraction (VAF) called clonal haematopoiesis of indeterminate potential (CHIP)[18].

Over the past ten years, there has been substantial progress in identifying mCAs in large cohorts, which was benefited from the decreasing cost of sequencing and the emergence of large cohorts with genomics data[18]. In 2012, Laurie et al.[19] detected large structural mosaic events of autosomal chromosomes based on SNP array data from over 50,000 individuals recruited for GWAS by exploring log R ratio (LRR) and B-allele frequency (BAF) data **(Figure 1-1)**. They identified 514 large structural mosaic events in 404 of 50,222 participants and found the frequency of clonal mosaicism for large chromosomal anomalies in blood sample is low (<0.5%) from birth until 50 years of age, after which it rapidly rises to 2–3% in the elderly[19].

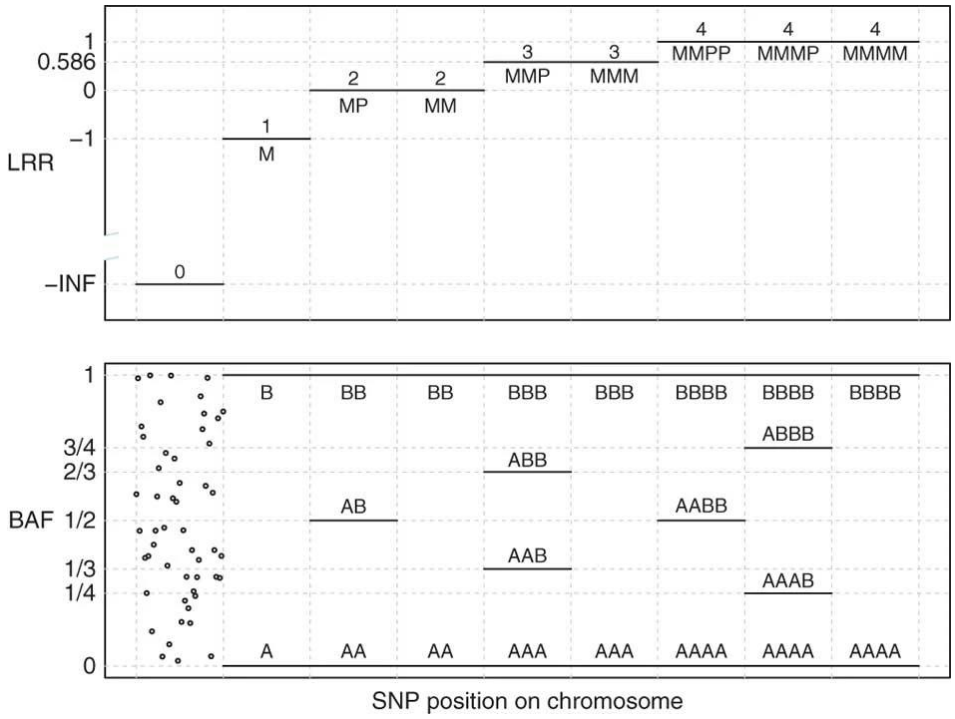


Figure 1-1 Illustration of log R ratio (LRR) and B-allele frequency (BAF) (Cited from Laurie et al. [19])

In the same issue of *Nature Genetics*, Jacobs et al.[20] reported similar results as Laurie et al. They identified 514 mosaic events of autosomal chromosomes based on 31,717 cancer cases and 26,136 cancer cancer-free controls[20]. In 2015, Machiela et al. detected 1,315 large structural mosaic events of autosomal chromosomes in 925 of 127,179 individuals[21]. Most DNA materials used in these studies were extracted from blood and the sensitivity of detection largely depended on the fraction of blood cell with mosaic events (> 5%) and sample size[19–21]. The phase information was also not used in the studies mentioned above. Therefore, many more mosaic events may be missed.

Using statistical phasing, Vattathil and Scheet increased the mosaic detection sensitivity (>1%) and identified 1,141 large structural mosaic events of autosomal chromosomes in 901 of 31,110 participants[22]. An important breakthrough in detecting large structural mosaic events was made by Loh et al in 2018[17]. They exploited the information provided by long-range phasing and hugely increased the detection sensitivity (>0.1%). With their novel phase-based computational method based on SNP-array data, 8,342 chromosomal structural mosaic events ranging from 50 kb to 249 Mb in blood-derived DNA were identified in 151,202 UK Biobank participants[17]. By applying their new method on the whole UK Biobank cohort, they further identified 19,632 autosomal mosaic chromosomal alterations in 482,789 participants[23]. Their new method shed a light on this area and has been a cornerstone for the downstream analysis of specific chromosomal structural mosaic events, because of its high and accurate detection rate compared to the previous methods based on SNP array data.

Although the studies using SNP array data have made significant progress in detecting clonal haematopoiesis events especially for mCAs, there is still a great potential to explore clonal haematopoiesis events using genotype data generated by HTS. The advancement of HTS, including whole exome sequencing (WES) and whole genome sequencing (WGS) technology, and its continuing decline in cost, promise to transform population-based genetics studies. HTS data enables the identification of other types of genetic variation that are often overlooked, such as copy number changes and mosaic alterations present in only a fraction of cells. Detection of somatic variation or chromosomal gains/losses can be accomplished by calculating the proportion of sequenced reads.

One application of HTS data is to detect CHIP mutations. Unlike other types of clonal haematopoiesis, CHIP is hard to detect from SNP-array data as many CHIP mutations are rare and cannot be captured or imputed from SNP-array. Recently, there have been two large studies on CHIP. Using the high coverage WGS of blood DNA in 97,691 participants of The Trans-Omics for Precision Medicine (TOPMed) program, Bick et al. identified 4,938 CHIP mutations in 4,229 participants[24]. They found that over 75% of identified CHIP mutations were in one of three genes, *DNMT3A*, *TET2* and *ASXL1*[24]. Kar et al. used the blood WES data of 200,453 participants from UK Biobank to detect CHIP mutations[25]. They identified 11,697 mutations in 10,924 participants. Again, *DNMT3A*, *TET2* and *ASXL1* were the most mutated CHIP genes[25]. From these two studies, several loci were identified to be significantly associated with CHIP including loci near *TERT*, *TET2*, *KPNA4-TRIM59*, *PARP1*, *ATM*, *CHEK2*, *CD164*, and *SETBP1*[24,25].

1.3 Mosaic loss of chromosome Y

The studies described above introduced the progress in autosomal mCA and CHIP events. However, the most common type of chromosomal structural mosaic event is mosaic loss of chromosome Y (LOY) in circulating white blood cells. LOY is defined as a lower-than-expected abundance of DNA from the Y chromosome with a certain threshold of detection[14,16,26]. LOY has been discovered for over 50 years from the earliest cytogenetic analyses, and it has a high prevalence among the ageing men[15,16]. Previous studies have shown significant associations between LOY and smoking status[27]. There is numerous epidemiological evidence demonstrating that LOY associates with several types of cancer, autoimmune conditions, age-related macular degeneration, cardiovascular disease, Alzheimer's disease, type 2 diabetes, obesity, and all-cause mortality[16]. However, the causes of LOY and the mechanisms behind these associations are still unclear.

Like other large structural mosaic events, LOY can be detected using SNP-array genotyping for GWAS as the arrays include a varying number of probes on chromosome Y [28]. Human chromosome Y has a special structure and can be divided into two different parts: pseudo autosomal regions (PAR) and male-specific regions (MSY) consisting of X-transposed, X-degenerate (XDR) and Ampliconic regions[3,29] **(Figure 1-2)**. Since the MSY region does not participate in recombination, the degree of LOY can be estimated by calculating the median or mean of log R ratio (mLRR) values of the probes on MSY (chrY: 2,694,521-59,034,049, hg19/GRCh37; 6,671,498–22,919,969, hg18/Build 36)[28,30,31]. The mLRR value close to zero is the normal state and the more negative value means that the larger proportion of cells may have loss of chromosome Y.

Previous studies on LOY have set different thresholds for mLRR-Y values to distinguish mLOY events from controls. By analysing the peripheral blood DNA from 1,153 elderly men in an age window of 70.7-83.6 years, Forsberg et al. found at least 8.2% of these participants with LOY in their blood cells[31]. After applying the same approach but choosing a different threshold, Zhou et al. also identified LOY events for 8,679 cancer cases and 5,110 cancer-free controls and found that 7% of the men had LOY and the prevalence of LOY increased with age, reaching 18.7% in men aged over 80 years old[30]. It is obvious that the different

arbitrary thresholds explain the difference in LOY prevalence among the older males in these studies.

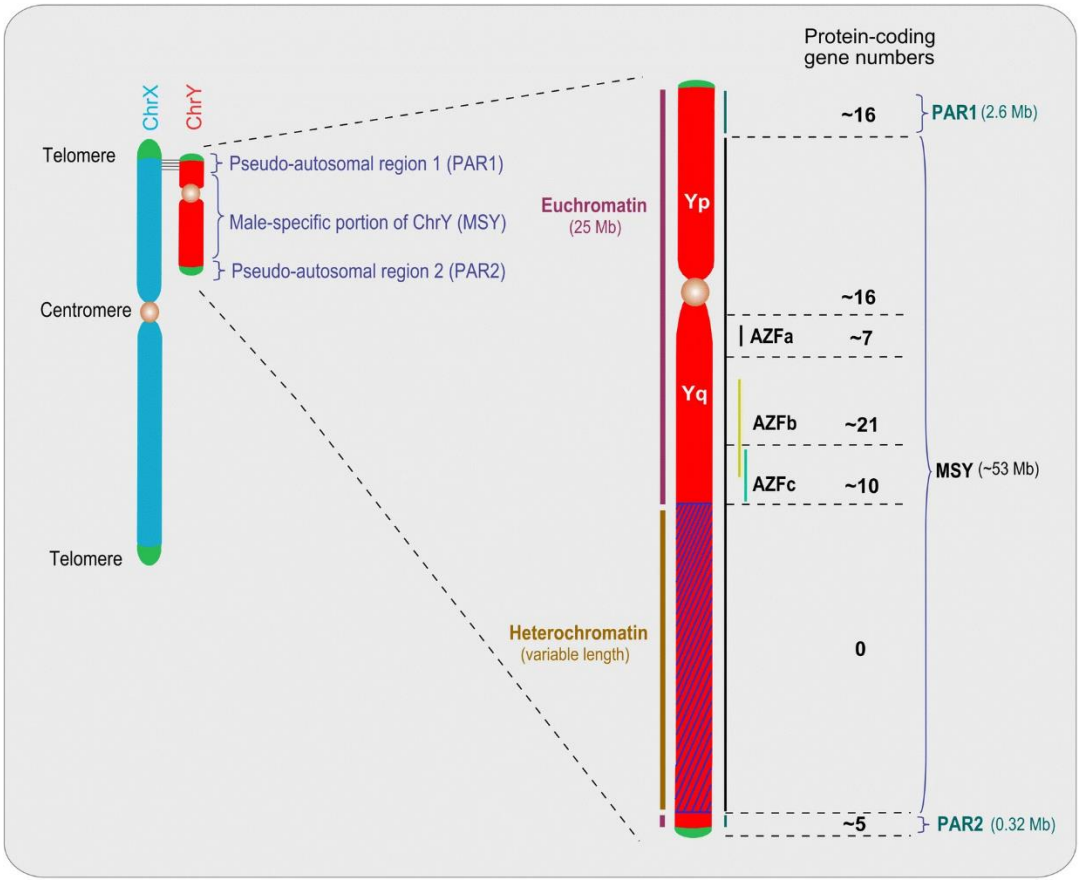


Figure 1-2 Structure of human chromosome Y (Cited from Guo et al. [29]).

Detecting large structural mosaic events such as LOY is just the first step, the key questions needing to be answered are the mechanisms behind these events and their contributions to the developmental traits and diseases. If the large structural mosaic events can be regarded as one type of "genetic phenotypes", it will be worth to conduct GWAS to understand the genetic contribution of common variants to them and then uncover the potential biological pathways. For LOY, Zhou et al. discovered the first common variant susceptibility locus (rs2887399, OR=1.55, 95% CI=1.36-1.78; P=1.37×10^{-10}) relating to mLOY by conducting GWAS, which mapped at 14q32.13 to *TCL1A* (encoding T cell leukaemia/lymphoma 1A) that functions as a co-activator of the cell survival kinase AKT and is often over- expressed in haematological malignancies of T and B cells[30].

Using genotype-array-intensity data and sequence reads from 85,542 male UK Biobank participants, Wright et al. identified 19 genomic regions ($P < 5\times10^{-8}$) associated with LOY, which was estimated by the mean of log R ratio directly through implementing GWAS[32]. These identified loci included genes that covered several aspects of cell proliferation and cell cycle regulation, including DNA synthesis (*NPAT*), DNA damage response (*ATM*), mitosis (*PMF1*, *CENPN* and *MAD1L1*) and apoptosis (*TP53*) [32].

In 2019, Thompson et al. published the largest research on LOY and found that about 20% of male participants aged from 40 to 70 years old in UK Biobank (N=205,011) had detectable LOY, using the phase-based computational method (PAR-LOY) developed by Loh et al[16,17]. In this method, they used allele-specific genotyping intensities in PAR which is the chromosome Y region affected by recombination[16]. Then, they performed a GWAS for dichotomous LOY status estimated by PAR-LOY and identified 156 statistically independent signals ($P<5\times10^{-8}$). Among the identified signals, 19 were previously reported , which were involved in cell-cycle regulation and cancer susceptibility, as well as acting as the somatic drivers of tumour growth and targets of cancer therapy[16]. In order to compare these two LOY estimation methods (mLRR-Y and PAR-LOY), they performed a GWAS for continuous LOY variables estimated by mLRR-Y on the same study samples and found that only 61 of the 156 loci reached genome-wide significance, which showed that PAR-LOY may have much more power for detecting LOY than mLRR-Y[16].

Table 1-1 Characteristics of previous population-scale studies on LOY

LOY Study	Year	LOY measures	Sample size	No. of LOY cases	No. of Leading Signals
Forsberg et al. [31]	2014	mLRR-Y (< -0.139)	1,141	93	NA
Dumanski et al.[27]	2015	mLRR-Y (< -0.139)	5,738	900	NA
Zhou et al. [30]	2016	mLRR-Y (< -0.15)	13,729	970	1
Wright et al. [32]	2017	mLRR-Y	85,542	NA	19
Terao et al. [33]	2019	mLRR-Y	95,380	NA	50
Thompson et al. [16]	2019	PAR-LOY	205,011	41,791	156

Most current LOY studies use white Europeans as their study group, but there may exist different genetic architectures for different ancestry groups. Loftfield et al. observed less LOY in men of African ancestry (0.4%) compared to men of European ancestry (1.8%, P=0.003) in UKBB[34]. Before the study of Thompson et al.[16], Terao et al.[33] conducted a GWAS

on 95,380 Japanese males and identified 50 independent genetic markers in 46 loci at the genome-wide significant level, 35 of which were unreported. Among these signals, 15 of them replicated the significant signals from Wright et al.[32]

Therefore, conducting GWAS on non-European groups may identify novel ancestry-specific signals, which can reflect the differences in the genetic architecture of LOY between ancestry groups.

1.4 Mosaic loss of chromosome X

Same as LOY in males, mosaic loss of chromosome X (LOX) is the most common mosaic event in females[17], which might be affected by the X chromosome inactivation and can increase the risk of leukaemia[35–37]. On the contrary, male LOX is extremely rare[38]. Compared with the genetic and epidemiological studies on LOY, studies on LOX are still lacking. Although there have been over 150 signals reported to be associated with LOY that reveal the underlying mechanisms[16], there has been no large-scale GWAS of LOX since now.

There is reported moderate genetic correlation (r_g =0.3, P=3.98×10^{-4}) between LOY and LOX[39], suggesting some shared mechanisms between these two types of sex chromosome mosaicism. But the distinct pattern of LOX means that other mechanisms underlying LOX may also exist. As reported by Loh et al., two common variants (Xp11.1 near *DXZ1* and Xq23 near *DXA4*) on chromosome X that weakly elevate the risk of LOX but strongly impact which X chromosome is lost in the expanded clone in heterozygous females[17]. Additionally, the loci near *HLA* and *SP140L* were associated with a higher risk of LOX [17].

1.5 Exploring sex chromosome mosaicism using high-throughput sequencing data from large cohorts

Same as other studies on somatic mosaicism, the comprehensive exploration of LOY/LOX estimates derived from HTS data and comparison with LOY/LOX estimates based on SNP array data are still lacking until now. Meanwhile, GWAS have identified hundreds of signals associated with sex chromosome mosaicism, especially for LOY[16,30,32,33].

As an intrinsic limitation of GWAS, most identified LOY-associated signals are common variants with a minor allele frequency (MAF)>5% in the non-coding regions and only have a

tiny effect size[40]. The potential causal genes and regions still need to be targeted through several downstream analytical pipelines from the identified signals, which makes it difficult to conduct follow-up wet-lab validation work for the targeted genes or regions through these pipelines.

From the evolutionary theory, protein-truncating or missense variants with deleterious effects are likely to be rare due to natural selection[41,42]. Compared to common non-coding variants, protein-truncating or missense variants in some genes and regions may play a much greater role, but they are often poorly included in SNP arrays designed for GWAS[43] and are hard to be imputed from the reference sequence. The classical single variant association tests for GWAS may not be suitable for these low frequency (0.5%<MAF<5%) and rare (MAF<0.5%) variants as the statistical power would be very low unless the sample sizes or effect sizes are very large[40].

Due to the rarity of rare nonsynonymous variants, the cumulative effects of a group of variants with similar predictive functions in the same gene can be explored by implementing gene-based aggregation tests. When multiple variants in the group are associated with a given disease or trait, the statistical power can be increased[40]. The current widely used approaches for the gene-based aggregation tests are mainly based on two hypotheses. The first hypothesis is that a large proportion of the variants are causal, and their effects are in the same direction. Based on this hypothesis, the rare variants are simply collapsed into genetic scores to gain more power, which is called burden tests. However, this approach might be less powerful if only a small fraction of variants are causal or if the effects of causal variants have different directions. The variance-component tests (i.e., SKAT (the sequence kernel association test)[44]) are proposed to address this limitation by testing the variance of the genetic effects. As the prior knowledge about the real association is lacking for most tests, the omnibus test can also be performed by combining burden and variance-component tests, which can better account for the proportion of causal variants and the existence of the variants with the opposite effects[40]. However, if one of the hypotheses strongly holds, the omnibus test might be less powerful than the test based on that hypothesis[40]. Recently, more methods (i.e., STAAR (variant-set test for association using annotation information)[45]) have been developed to set a flexible weighting scheme

according to the MAF of variants and the quantitative annotation scores (i.e., CADD[46], REVEL[47], SIFT[48], POLYPHEN[49] etc.) (Figure 6)

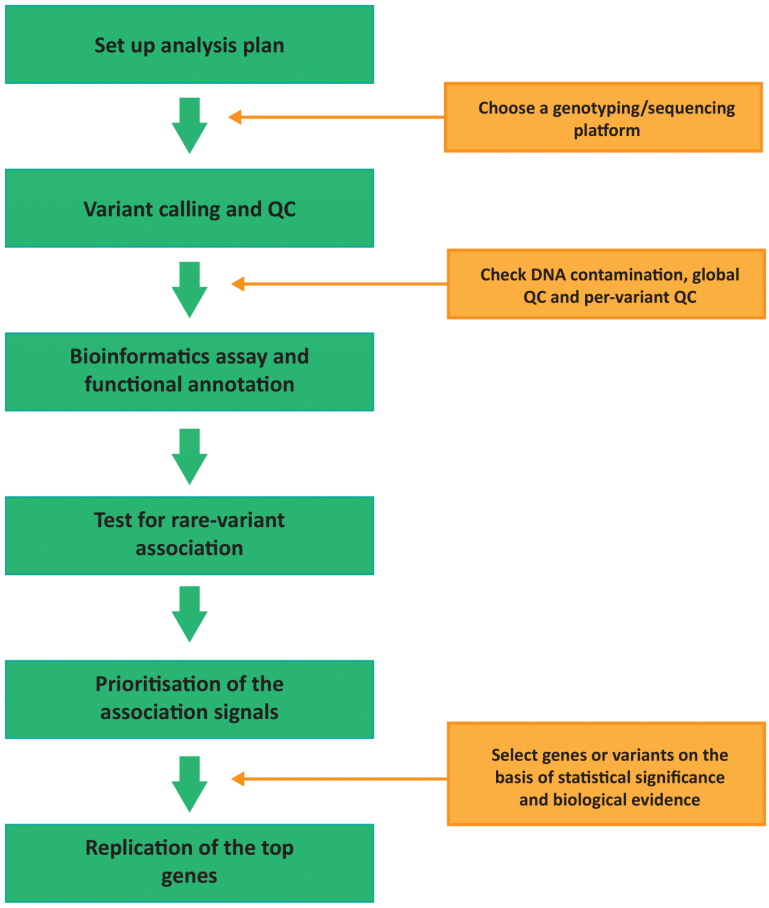


Figure 1-3 Data-processing and analysis flow chart for Sequencing-Based Association Studies (cited and redrawn from Lee et al.[40])

However, the absence of large cohorts with HTS data has been one major obstacle for these studies to estimate LOY/LOX and explore the cumulative effects of rare variants. This obstacle can be resolved now because the large cohorts such as UK Biobank made their HTS data publicly available. UK Biobank released its WES data for all 454,834 participants through 3 batches from Mar. 2019 to Oct. 2021 and WGS data for over 200,000 participants in Nov. 2021. Along with the SNP-array data, UK Biobank is ideal to explore the LOY/LOX estimation from different sequencing method and identify the genes which have direct effects on LOY/LOX by using the HTS data. With the abundant health-related data including

medical history, environment exposure and health outcome etc. of the UK Biobank, more associations between LOY and health outcomes might be revealed[50].

1.6 Structure and aim of this thesis

This thesis aims to explore the detection, causes and consequences of sex chromosome mosaicism by combining the data from both SNP-array and WES of half a million participants and implementing advanced statistical methods.

Chapter 2 describes the methods used to estimate LOY/LOY from SNP-array and WES data and the systematic comparison among different LOY/LOX metrics, including two combined LOY/LOX calls incorporating LOY/LOX metrics estimated from both SNP-array and WES data.

Chapter 3 describes a new GWAS for LOY, exploiting the new combined LOY metric estimated from SNP-array and WES data and the exploration of whether the summary statistics of LOY and LTL can be used to boost the power of detecting the signals of MPN.

Chapter 4 describes the detection and characteristics of the males with abnormal karyotypes in the UK Biobank.

Chapter 5 describes the first exome-wide association study of LOY on over 80,000 males in the UK Biobank and the comprehensive investigation of the novel LOY-associated gene *GIGYF1*.

Chapter 6 describes exome-wide association studies of LOY and LOX based on over 450,000 participants from the UK Biobank.

Chapter 2 Development and statistical power evaluation of sex chromosome mosaicism measures

2.1 Contributions

This chapter describes the analyses that I conducted for the development and statistical power evaluation of sex chromosome mosaicism measures. I estimated LOY and LOX from SNP-array data for all UK Biobank participants and the Y and X chromosome dosages from WES data for over 200,000 UK Biobank participants. I then calculated the new combined LOY and LOX calls, checked correlations among the different LOY and LOX measures and performed systematic comparisons of their statistical power. Dr Hana Lango Allen designed and tested the pipeline to estimate the average read depth of the autosomes and sex chromosomes. Prof Po-Ru Loh (Harvard University) proposed the formulas to combine the different sex chromosome mosaicism measures. Dr Eugene J. Gardner implemented the read depth pipeline in UKBB RAP platform and provided valuable advice on Figure 2-3. Prof John R.B. Perry and Ken K. Ong supervised the analyses and provided guidance.

2.2 Abstract

Sex chromosome altering events, including mosaic loss of the Y (LOY) and X (LOX) chromosomes, accumulate during the ageing process. However, the causes and consequences of these events are unclear. Previous studies have used SNP-array intensity data to estimate LOY and LOX in thousands of human individuals in order to identify the genetic underpinnings of LOY/LOX. Estimates of an individual's LOY/LOX status are then used to perform Genome-Wide Association Studies (GWAS) and Phenome-Wide Association Studies (PheWAS) to identify genetic loci or phenotypes, respectively, that increase an individual's risk of mosaic sex chromosomes loss. Yet, sex chromosome mosaicism estimated from high-throughput sequencing data (e.g., Whole-Exome [WES] and Whole-Genome Sequencing [WGS]) and the systematic evaluation of these measures is lacking. In this chapter, I proposed a new approach to the estimation of sex chromosome mosaicism from WES. I calculated the combined estimates incorporating several independent measures, including continuous and binary LOY/LOX estimation from SNP-array intensity and the Y/X dosage estimated from WES for LOY and LOX respectively. Through systematic comparisons between LOY measures, my new 3-way combined metric showed an improvement from previous approaches. Although the results pertaining to LOX were more difficult to interpret compared with LOY, the WES based dosage and 3-way combined metric still showed some advantages compared to SNP-array based approaches especially a stronger association with age, the dominant risk factor of all CH events. These results provide evidence for selecting the measures of sex chromosome mosaicism with the highest statistical power for epidemiological studies, SNP-based genome-wide association, and gene-based burden tests.

2.3 Introduction

Most studies in human genetics have mainly focused on inherited variation transmitted through the germline. Many have concentrated on identifying disease-causing variants in monogenic diseases and investigating the genetic contribution to complex phenotypes through the use of genome-wide association and sequencing-based studies[15]. However, as part of the ageing process and owing to errors in DNA repair or replication of damaged DNA as cells divide[14], many different mutations, such as Single-Nucleotide Variants (SNVs), insertion or deletions (InDels), and large chromosomal structural variations (SVs) can arise and accumulate over time in the somatic cells of the human body. Consequently, cells originating from a single fertilised zygote may differ in their genetic makeup. The age-related accumulation of postzygotic DNA mutations results in tissue genetic heterogeneity in adult humans and is known as somatic mosaicism[14].

Although somatic mosaicism can arise in all tissues of the human body[14], the most well-studied is somatic mosaicism in the blood[18]. More specifically known as clonal hematopoiesis (CH), CH can be defined as an age-related expansion of mutated hematopoietic clones, of which several distinct subtypes exist: coding mutations in approximately 20 genes recurrently found as mutated in CH[51] (Clonal hematopoiesis of indeterminate potential; CHIP), gains/losses and copy-neutral loss of chromosomal segments (mosaic Chromosomal Alterations; mCAs), and loss of whole sex chromosomes[39]. Of these subtypes, the most common form of CH is the mosaic loss of chromosome Y (LOY) in circulating white blood cells. LOY was discovered over 50 years ago from cytogenetic analyses and has a high prevalence among aging men[16,31].

LOY has several clinical definitions, the most common one classifies LOY as a lower-than-expected abundance of DNA from the Y chromosome within a certain threshold of detection[14,16,26]. According to the largest study on LOY (N=205,011 males), around 20% of European males between 40-70 years of age have some evidence of LOY[16]. Other studies have also shown that LOY is associated with a wide range of behaviours and comorbidities such as age, smoking status, cancer, autoimmune conditions, type 2 diabetes, and all-cause mortality[16,26,28,30,31,52]. Due to limitations in sample size and methodology, somatic

mosaicism has been difficult to identify; mutation events are random, rare, and the fraction of affected cells is often below the detection limit of current technologies[14,15,17].

Nonetheless, in the past ten years, there has been significant progress in identifying somatic mosaicism events due to the decreasing cost of high-throughput sequencing and the emergence of large cohorts with SNP array data[18] the identification of common genetic variation. The standard approach calculates the $\log_2$-transformed *R* ratio (LRR), which represents the difference between the fluorescence signal intensity of the 'A' and 'B' alleles assayed for each SNP on the genotyping array[28]. Because the male-specific regions (MSY) portion of the Y chromosome does not participate in recombination with the X chromosome, the degree of LOY can be estimated by calculating the median or mean of LRR values (mLRR-Y) of all probes in the MSY region[28,30,31] . Recently, Thompson et al.[16] proposed a new approach to identify dichotomous LOY events (PAR-LOY). Using PAR-LOY, they identified 156 leading LOY-associated signals by performing GWAS for over 200,000 male participants in UK Biobank[16]. This approach uses the phase-based computational method developed by Loh et al[17] and was based on allele-specific genotyping intensities in the sex chromosome pseudo autosomal regions (PAR) rather than MSY[16]. In order to compare these two LOY estimation methods (mLRR-Y and PAR-LOY), they also performed a GWAS for continuous mLRR-Y values on the same study samples and found that only 61 of the 156 loci reached genome-wide significance, which indicated that PAR-LOY may have much more power for detecting LOY than mLRR-Y[16].

LOX can also be estimated using these above-mentioned approaches. Different from LOY, female mosaic loss of the chromosome X (LOX) is also age-related but much less common. Previous research on LOX identified 124 LOX events>2 Mb in size in only 0.25% (97 out of 38,303) women using SNP-array intensity data[35]. Compared with LOY, large-scale studies on LOX are still lacking.

Most of these current implemented approaches to detect sex chromosome mosaicism events have relied on SNP-array probes on the sex chromosomes, so the accuracy of the estimations depends on the number of probes, which can vary greatly between arrays especially for LOY. For most genotype arrays, there are only hundreds to thousands of probes on the Y chromosome. Because mLRR-Y is the mean or median of the LRR values of

these probes, the number of probes can largely affect the accuracy of the estimation. By contrast, data from WES and WGS provides the sequence read depth of the exomes and genomes of Y chromosome. Therefore, WES and WGS have great potential for exploring somatic mosaicism, including sex chromosomes mosaicism events. From WES and WGS data, the detection of somatic variation or chromosomal gains/losses can be accomplished by calculating the proportion of sequenced reads. It is unknown whether the combination of different sex chromosome mosaicism measures can improve the detection power.

In this chapter, I develop a new method based on read depth from WES data to estimate sex chromosome mosaicism events. I first introduced the method for the combined calls of the independent sex chromosome mosaicism variables. Then I conducted a systematic comparison among widely used sex chromosome mosaicism estimations including mLRR-Y/X, dichotomous mLRR-Y/X, PAR-LOY/MoChA-LOX, sex chromosome mosaicism variables (Y/X dosages) estimated from WES, dichotomous Y/X dosages and the combined calls .

2.4 Methods

2.4.1 Study populations

In this chapter, all analyses were conducted using the data from UK Biobank, which contains around 500,000 participants aged from 40 to 70 across England, Wales, and Scotland. For each participant, a list of phenotypic and health-related data was collected, including physical measurements, lifestyle indicators, blood and urine biomarkers, imaging, and routine health record data[50]. UK Biobank has approval from the North West Multi-centre Research Ethics Committee (REC reference 21/NW/0157) and informed consent was provided by each participant.

Two SNP genotyping arrays were used to genotype the 488,377 participants: UK Biobank Lung Exome Variant Evaluation (UK BiLEVE study, N=49,950) and Affymetrix Axiom UK Biobank array (UK Biobank Axiom (Affymetrix), N=438,427). There were 807,411 and 825,927 SNP probes respectively, with 95% overlap between arrays.

UK Biobank released WES data for over 450,000 participants in three batches: the first batch of 49,981 samples was available in Mar 2019, the second batch containing 200,602 samples was released in Oct 2020 and the third batch containing all 454,834 samples was released in Oct 2021.

2.4.2 Estimation of intensities across the entire Y and X chromosome from SNP-array data

To estimate the intensities across the entire X or Y chromosome from SNP-array data, I downloaded the genotyping fluorescence signal intensity (LRR) and quality control (QC) information for all SNPs on the X and Y chromosome from the UK Biobank data showcase (https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=22431 and https://biobank.ndph.ox.ac.uk/showcase/refer.cgi?id=1955). SNPs which (i) were located within PARs, (ii) did not have a calculable LRR on both arrays, (iii) did not pass QC in all 106 batches, or (iv) were flagged as failing QC by UK Biobank were excluded.

Following this exclusion process, 16,599 SNPs on the X chromosome and 579 SNPs on the Y chromosome remained. The median LRR across all remaining SNPs on the X and Y chromosome was calculated to generate the values for mLRR-X and mLRR-Y, respectively.

These values represent the median fluorescence signal intensities across the entire X or Y chromosome relative to all autosomal SNP signals.

2.4.3 Dichotomous mosaic status of Y and X chromosome from SNP-array data

The MoChA-LOX and PAR-LOY data in this study was acquired from previous studies conducted by Loh et al.[17,23] and Thompson et al.[16] For the dichotomous mosaic status of the X and the Y chromosome, it is necessary to know the fractions of cells without the X and the Y chromosomes. The affected cell fractions (AF-LOY) of LOY and LOX were also calculated from the same pipeline used for estimating LOY and LOX status. They were calculated from BAF (B-Allele Frequency) values. The formula to calculate them can be divided in to two parts: *muDiff=2×0.01×BAF* and *AF=2×muDiff/(1+muDiff)*.

2.4.4 Estimation of Y and X chromosome dosages from whole exome sequence data

The pipeline used to estimate the X and the Y chromosome dosages based on WES data involved two steps. Firstly, I processed WES data for each sample to generate their average coverages of the autosomal chromosomes, the X chromosome, and the X-degenerate regions (XDR) of the Y chromosome. More accurate average coverage of Y chromosome can be estimated through just focusing on XDR, because the other male-specific regions on the Y chromosome are ~99% identical to the homologous region on the X chromosome or all highly repetitive and include palindromes (inverted repeats)[53]. I then extracted the regions of XDRs from the target regions of UK Biobank WES capture experiment according to their GRCh38 coordinates. The target region of the autosomal chromosomes was generated by excluding the target regions on the X and Y chromosomes. The target region of the X chromosome was generated by excluding the target regions on PARs.

I applied Samtools (version:1.9)[54] to convert the CRAM files of each sample to BAM files based on the GRCh38 reference sequence (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/GRCh38_reference_genome/GRCh38_full_analysis_set_plus_decoy_hla.fa). To calculate the average coverages of the autosomal chromosomes, the X chromosome, and the XDR of Y chromosome, I converted the BED (Browser Extensible Data) files of the new target regions to Picard Interval Lists using Picard (version: 2.21.6-SNAPSHOT) function *BedToIntervalList*, based on the same

reference sequence. With these Picard Interval Lists, the BAM file of each participant was inputted to calculate the average coverages of the autosomal chromosomes, the X chromosome, and the XDR of Y chromosome through Picard function *CollectHsMetrics*. It took about 40 minutes for each sample to generate its average coverages of these three regions using one CPU. The second step was to calculate the relative read depth of XDRs and X chromosome, which were defined as the average coverage of XDRs and X chromosome divided by the average coverage of autosomal chromosomes. For males the average Y chromosome read depth should distribute around 0.5, as there are two copies of autosomal chromosomes but only one copy of the Y chromosome. Male Y dosage is therefore defined as the relative read depth multiplied by 2, to provide a proxy of the Y chromosome copy number. Female X dosage is the relative read depth of X chromosome, which is a direct proxy of the X chromosome copy number.

For the first and second batch of WES data, I downloaded all the individual level WES data and stored in the HPC (High performance computing) system of the University of Cambridge, and all the above-mentioned analyses were conducted by submitting parallel jobs to the Slurm job allocation system. For the third batch, all the individual level WES data was stored and called in an online cloud system and all the analyses were transferred to the UK Biobank Research Analysis Platform (RAP). The new WDL (Workflow Description Language) pipeline was designed to conduct the analyses on AWS (Amazon Web Services). For the same sample from the first batch, the dosage of chrX and chrY have exactly the same values estimated in the HPC system and UK Biobank RAP.

2.4.5 Combining the LOY and LOX estimates to build a new LOY and LOX estimate

The main goal of this chapter is to explore whether the combination of these LOY and LOX measures estimated by the three different methods can improve the power of detecting LOY/X.

For males, two LOY combined models using PAR-LOY as a baseline were proposed:

i) *LOY Combined call (2-way)=PAR-LOY (the binary LOY status estimated by PAR-LOY)+3×AF-LOY (the estimated fraction of cells without chromosome Y)-3×mLRR-Y (cropped to the range [0,2])*

ii) *LOY Combined call (3-way)=PAR-LOY+2×AF-LOY-2×mLRR-Y-4×(Y dosage-1) (cropped to the range [0,2]).*

For a sample with LOY, it should have PAR-LOY value equal to 1, AF-LOY value greater than 0, mLRR-Y value less than 0 and Y dosage value less than 1. PAR-LOY and mLRR-Y are two independent estimates of LOY because they are derived from different regions of chromosome Y. For Y dosage, as it comes from WES data, it is independent of the LOY estimates from the array data. The value of Y dosage minus 1 represents the degree of LOY. These combined calls can fully utilise LOY estimates from these methods to reduce the number of false negative and false positive cases introduced by single LOY estimate. For example, if a sample is identified as without LOY by PAR-LOY, but has a positive value of AF-LOY, a negative value of mLRR-Y and a value of Y dosage less than 1, then the combined call can correct the LOY estimation from PAR-LOY and provide the degree of LOY.

The same approach was also applied to find LOX measures with the largest statistical power. Two LOX combined models using MoChA-LOX as the baseline were also proposed:

i) *LOX Combined call (2-way)=MoChA-LOX+3×AF-LOX-3×mLRR-Y (cropped to the range [0,2])*
ii) *LOX Combined call (3-way)=MoChA-LOX+2×AF-LOX-2×mLRR-X-4×(X dosage-2) (cropped to the range [0,2]).*

As only LOY for 46,XY and LOX for 46,XX were analysed, for all these measures, the abnormal karyotypes carriers for both males and females were excluded (see chapter 4). Then, I performed association tests in R among the pairs of mLRR-Y/X, Y/X dosage, PAR-LOY/MoChA-LOX, LOY/X Combined call (2-way), and LOY/X Combined call (3-way) for males and females, respectively.

2.4.6 Assessment of statistical power of different LOY and LOX estimates

Statistical analysis was performed using R software (version 3.6.2). As the mLRR-Y and Y dosage are continuous variables, I tested a wide range of dichotomous variables by creating thresholds for the mLRR-Y and the Y dosage data, respectively. By defining each centile from 1% to 50% of the mLRR-Y and the Y dosage data as the threshold, I created 50 case/control

comparisons according to samples' mLRR-Y and Y dosage values separately. For example, participants whose Y dosage value was less than the 1st centile of Y dosage data were assigned to the case group with LOY phenotype, and the remaining participants were assigned to the control group. I had different types of LOY including mLRR-Y, Y dosage, PAR-LOY, two combined calls and the 100 dichotomous variables for assessing their statistical power.

Previous studies had shown that the LOY phenotype such as mLRR-Y and PAR-LOY had strong correlations with age and 'ever smoking' status. In order to identify which LOY estimate maximally reflected the association with age and 'ever smoking' status, I extracted the data of age and 'ever smoking' status from UK Biobank. Linear regression was performed with genotype chip, exome batch and PC1-10 as covariates. For the linear regression against 'ever smoking' status, age was also introduced as a covariate. I then calculated the absolute values of Z scores (beta/SE) from the summary statistics of linear regression and compared for all LOY estimates. To date, with the development and application of GWAS, several variants on autosomal chromosomes which had a significant association with LOY had been identified. Wright et al.[32] identified 19 leading SNPs that are associated with LOY estimated from mLRR-Y based on 85,542 males in UK Biobank. Thompson et al.[16] identified 156 leading LOY-associated signals. I calculated the Genetic Risk Scores (GRS) of the 19 SNPs and the 156 SNPs for the UK Biobank male participants separately using PLINK. The same analysis was performed for these two GRS as for 'ever smoking' status.

The same steps were also conducted for LOX.

2.5 Results

2.5.1 Summary of Sex chromosome dosage estimation from WES data and comparison with mLRR-X and mLRR-Y from genotyping array

I estimated the LOY and LOX measures for each participant in UK Biobank from their SNP-array and WES data **(Table 2-1)**.

Table 2-1 Different approaches used for estimating LOY/X

LOY/X estimation methods	Description	Data	Regions on chrY/X
mLRR-Y/X (continuous)	compute the median or mean of log R ratio values of the probes on male-specific regions of the Y chromosome/non pseudo autosomal regions of the X chromosome	SNP-array	MSY/nonPAR
PAR-LOY/MoChA-LOX (dichotomous)	explore allelic imbalance on pseudo autosomal regions of the Y chromosome/non pseudo autosomal regions of the X chromosome using phase-based computational method (MoChA)	SNP-array	PAR/whole X
Y/X dosage (continuous)	calculate the relative depth of the X-degenerate regions of the Y chromosome/non pseudo autosomal regions of the X chromosome	WES	XDR/nonPAR
LOY/X Combined call (2-way, continuous)	PAR-LOY/MoChA-LOX + 3 × AF-LOY/X-3×mLRR-Y/X (cropped to the range [0,2])	SNP-array	MSY+PAR/nonPAR
LOY/X Combined call (3-way, continuous)	PAR-LOY/ MoChA-LOX+2×AF-LOY/X-2 × mLRR-Y/X-4×(Y/X dosage-1) (cropped to the range [0,2])	SNP-array and WES	MSY+PAR+XDR/nonPAR

Compared with mLRR values estimated from genotyping array, the X and Y dosage estimation from WES data provided clearer separation between males and females and normal and abnormal karyotypes (**Figure 2-1)**. In the 205,604 males, the Y dosage ranged from 0.119 to 2.380 with a mean (SD) 1.013 (0.085), and the X dosage ranged from 0.904 to 2.035 with a mean (SD) 0.983 (0.027). In the 243,765 females, the X dosage ranged from 0.958 to 3.011 with a mean (SD) 1.957 (0.038).

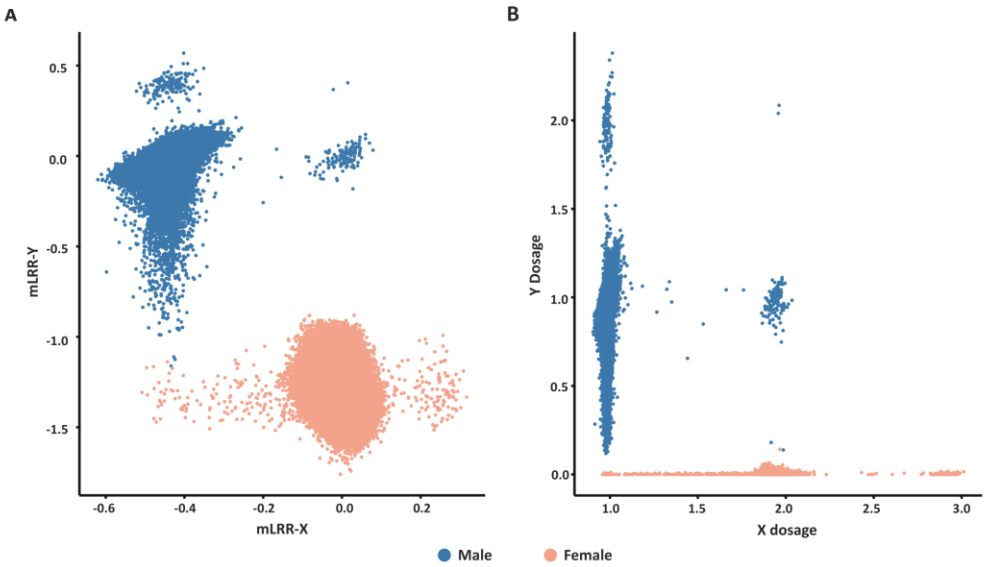


Figure 2-1 (A) mLRR-X and mLRR-Y for each of 205,604 males and 243,765 females. **(**B) X dosage and Y dosage for the same samples.

There was a low to moderate positive association between sex chromosome dosages and mLRR values of sex chromosome. For males, the Pearson's correlation coefficient between Y dosage and mLRR-Y was 0.612 and the correlation coefficient between X dosage and mLRR-X was 0.297. For females, the correlation coefficient between X dosage and mLRR-X was 0.247.

2.5.2 Correlation between different LOY and LOX measures

The samples identified as LOY by PAR-LOY among 46, XY samples showed significant mean difference in mLRR-Y and Y-dosage (**Table 2-2, Figure 2-2A, B, C)**.

Table 2-2 Summary statistics of mLRR-Y/X and Y/X dosage for LOY/X cases and controls identified by PAR-LOY/MoChA-LOX

	mLRR-Y		**Y dosage**				**mLRR-X**		**X dosage**	
PAR-LOY	Case	Control	Case	Control		**MoChA-LOX**	Case	Control	Case	Control
Mean	-0.046	0.009	0.940	1.03		**Mean**	-0.002	-0.001	1.939	1.958
Min	-1.161	-1.112	0.140	0.119		**Min**	-0.169	-0.168	1.350	1.677
Max	0.140	0.213	1.284	1.456		**Max**	0.135	0.145	2.089	2.23
SD	0.091	0.039	0.119	0.056		**SD**	0.031	0.031	0.046	0.026
P_{ANOVA}	$<1\times10^{-250}$		$<1\times10^{-250}$			P_{ANOVA}	1.45×10^{-5}		$<1\times10^{-250}$	

The samples identified as LOX by MoChA-LOX among 46, XX samples also showed mean differences from controls in mLRR-X and X-dosage. But from the P_{ANOVA}, the difference in mLRR-X was not as significant as in X dosage (**Table 2-2, Figure 2-2D, E, F)**.

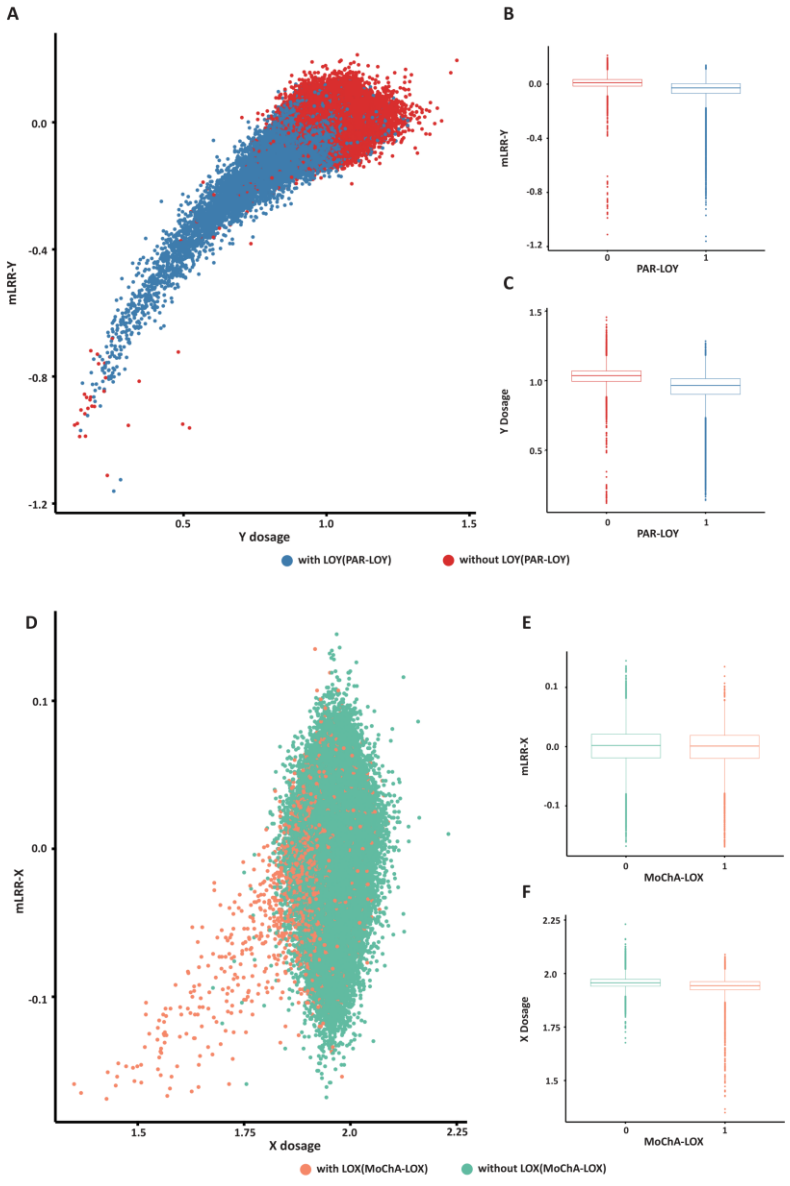


Figure 2-2 (**A)** Y dosage plotted against mLRR-Y for 46,XY males (N=205,321). (**B)** Difference between LOY cases and controls identified by PAR-LOY in mLRR-Y. (**C)** Difference between LOY cases and controls identified by PAR-LOY in Y dosage. (**D)** X dosage plotted against mLRR-X for 46,XX females (N=243,520). (**E)** Difference between LOX cases and controls identified by MoChA-LOX in mLRR-X. (**F)** Difference between LOX cases and controls identified by MoChA-LOX in X dosage.

All of the LOY measures were significantly associated with each other ($P<1\times10^{-250}$), but Pearson's correlation coefficients varied from 0.382 to 0.971. PAR-LOY was set as a baseline measure for two combined calls, so it had correlation coefficients equalling 0.952 and 0.926 with LOY Combined call (2-way) and LOY Combined call (3-way). These two combined calls also strongly correlated with each other (r=0.971). The smallest correlation coefficient was between mLRR-Y and PAR-LOY (r=-0.382) (**Table 2-3A)**.

Similar to the observations with LOY, all LOX measures were significantly associated with each other ($P<1.45\times10^{-5}$), however, the correlation coefficients were weak. Specifically, the correlation coefficient between mLRR-X and X dosage was only r=0.089 and the correlation coefficient between mLRR-X and MoChA-LOX was only r=-0.009 (**Table 2-3B)**.

Table 2-3 (A) Pairwise correlations between each LOY measures. The upper diagonal gives the T-statistic derived using cor.test ( ) in R, and the lower diagonal gives correlation coefficient. **(B)** Pairwise correlations between each LOX measures. The upper diagonal gives the T-statistic derived using cor.test ( ) in R, and the lower diagonal gives correlation coefficient.

A

T-statistic/correlation coefficient	mLRR-Y	Y dosage	PAR-LOY	LOY (2-way)	LOY (3-way)
mLRR-Y	NA	335.537	-187.255	-320.331	-294.24
Y dosage	0.595	NA	-223.018	-313.438	-408.439
PAR-LOY	-0.382	-0.442	NA	1412.486	1107.534
LOY Combined call (2-way)	-0.577	-0.569	0.952	NA	1835.657
LOY Combined call (3-way)	-0.545	-0.67	0.926	0.971	NA

B

T-statistic/correlation coefficient	mLRR-X	X dosage	MoChA-LOX	LOX (2-way)	LOX (3-way)
mLRR-X	NA	43.868	-4.337	-122.258	-124.17
X dosage	0.089	NA	-75.831	-96.69	-279.573
MoChA-LOX	-0.009	-0.151	NA	1711.724	1046.932
LOX Combined call (2-way)	-0.24	-0.192	0.961	NA	1286.411
LOX Combined call (3-way)	-0.244	-0.493	0.905	0.934	NA

2.5.3 Statistical power of different LOY and LOX measures

I have restricted the analyses to the samples with all the available LOY or LOX measures to perform a systematic comparison by running regression against the benchmarking traits including age and smoking status and two GRS of LOY. For LOY, the LOY Combined call (3-

way) method, consisting of all three independent LOY measures, outperformed all the other LOY measures for all traits and GRS we tested according to its Z scores. LOY Combined call (2-way) also provided a better statistical power than the LOY measures used in the previous studies and the new LOY measure estimated from WES. Although PAR-LOY is a binary LOY measure, it still generated larger Z scores than the two continuous LOY measures. The Y dosage estimated from WES had a better performance than mLRR-Y from genotyping array. For all of the dichotomous LOY measures derived from two continuous measures: mLRR-Y and Y dosage, their Z scores were less than the Z scores of the continuous mLRR-Y and Y dosage (**Figure 2-3A)**.

The rank of LOX measures was comparatively complicated. LOX Combined call (3-way) outperformed the other LOX measures for associations with age and smoking status. However, the WES X dosage generated the largest Z scores for associations with GRS. The mLRR-X did not reach a significant value for smoking status (P=0.11). If there was the same mechanism behind the LOX and the LOY, the LOX measures should be significantly associated with the LOY GRS. The results indicated a significant association of the LOX measures with the GRS generated from the 19 SNPs identified by mLRR-Y ($P<1.40\times10^{-7}$) and 156 SNPs identified by PAR-LOY ($P<2.81\times10^{-3}$), the WES X dosage outperformed all other LOX measures. As for dichotomous LOX measures derived from mLRR-X and X dosage, under some settings, they had larger Z scores than other LOX measures (**Figure 2-3B)**.

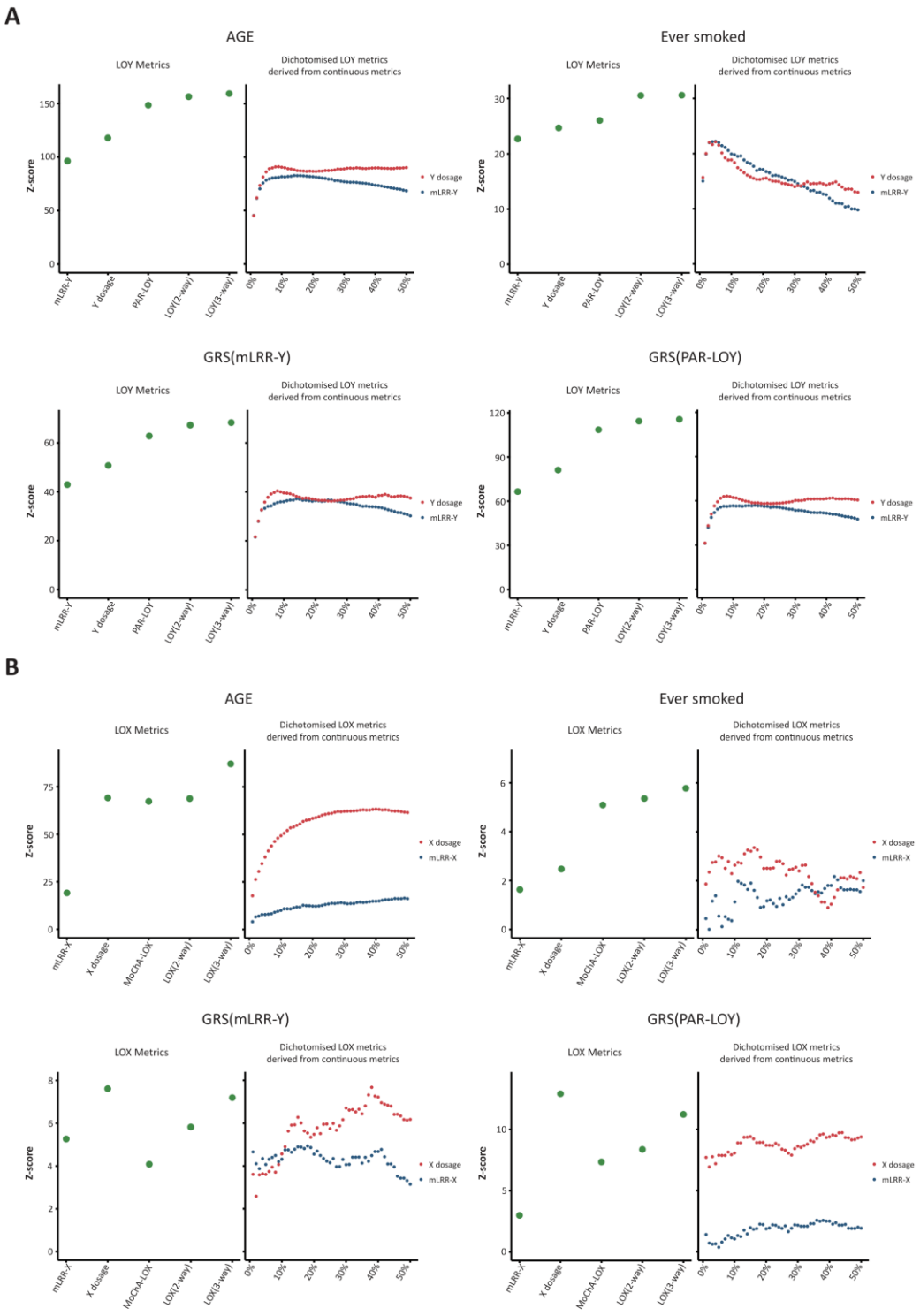


Figure 2-3 (A) Z scores derived from the linear regression of LOY metrics against age, ever smoking status, GRS (mLRR-Y), and GRS (PAR-LOY) respectively (From top left to down right). (**B)** Z scores derived from the linear regression of LOX metrics against age, ever smoking status, GRS (mLRR-Y), and GRS (PAR-LOY) respectively (From top left to down right).

2.6 Discussion

Everyone may have sex chromosome mosaicism, just to different degrees. Who we detect as having LOY or LOX is a function of our detection methods and the proportion of cells impacted. Compared to current sex chromosome mosaicism measures, although performing a PCR test may provide a more accurate estimation but is expensive and time-consuming, particularly in studies with large sample sizes. All of the sex chromosome mosaicism measures mentioned in this chapter can be seen as proxies of the abundance of the X and Y chromosome genetical material from SNP-array and WES data. In this chapter, I aimed to find the measures that provide the largest statistical power to undertake downstream analyses, including the investigation of their causes and consequences.

For LOY, the findings clearly show that different measures can hugely influence the statistical power of downstream analyses. The success of the 3-way combined call showed that the accuracy of LOY measure can be improved by maximally using the data from SNP genotyping arrays and WES. The 2-way combined call based on the independent LOY metrics estimated from SNP-array data was only slightly inferior to the 3-way combined calls. As there were just a few cohorts with WES or WGS data, this 2-way combined call should be the first choice for measuring LOY for the cohorts only having SNP genotyping array data. One advantage of PAR-LOY over mLRR-Y is that PAR-LOY allows for the detection of LOY events even in a smaller proportion of cells. Because mLRR-Y values can vary largely due to technical artefacts, the estimation from mLRR-Y was noisier. In contrast, the Y dosage estimation provides a cleaner continuous LOY variable. However, it should be noted that some samples with very small mLRR-Y and Y dosage values were not identified as LOY events by PAR-LOY, which might be because that the algorithm of PAR-LOY categorised these samples as 45, XO. In these cases, the combined call can accurately rescue them as LOY events.

For LOX, the combined calls improved power. In contrast, the results showed that mLRR-X may not be suitable for measuring LOX, as its estimation contained lots of technical noises. On the contrary, WES X dosage showed potential to be a LOX measure, because it outperformed 2-way combined call and MoChA-LOX for age and outperformed all the other LOX measures for associations with the two LOY genetic risk scores.

The results from this chapter provide a reliable reference for choosing sex chromosome mosaicism variables. Using the variables with stronger power to conduct traditional epidemiological studies and genetic association studies can better improve the understanding of the causes and consequences of sex chromosome mosaicism. There are some limitations of this study. Firstly, I compared the sex chromosome mosaicism measures solely based on statistical tests, and all measures were derived from genotyping chip and WES. The results would be more robust if qPCR for a small set of samples could be performed as a control to validate the sex chromosome mosaicism estimations. Secondly, these LOY measures may have more complex relationships with each other. Therefore, more sophisticated methods, such as advanced machine learning techniques will be required to maximise the power of combined calls. Finally, the results for LOX were not as clear as LOY, which might be due to its low prevalence and the use of genetic benchmark traits based on LOY. Since LOX has not been extensively studied as LOY, more studies will be needed to generate benchmark traits for LOX. Finally, the sex chromosome mosaicism measures were restricted to the DNA materials from blood. Future studies will be required to validate these measures or develop new methods to detect sex chromosome mosaicism events.

In conclusion, by combining the LOY/LOX metrics estimated from both SNP-array and WES data, the new proposed new Combined 3-way calls for LOY/X showed the most significant association with age, which is the primary risk factor for both LOY and LOX.

Chapter 3 Genetic analysis of Y chromosome mosaicism and its mechanistic link to myeloproliferative neoplasms

3.1 Contributions

This chapter details the work I conducted to re-analyse existing GWAS data of LOY using an improved measure of LOY calling and using the summary statistics of LOY to boost the power to detect the loci associated with MPN. I performed the GWAS analysis for the LOY Combined Call (3-way), identified the novel leading signals and compared with the previous GWAS studies that used PAR-LOY. I performed LDSC, MR, colocalisation and MTAG analyses for LOY, LTL and MPN. Dr Katherine Kentistou developed the pipeline to identify the independent significant signals from GWAS summary statistics and highlight the likely functional genes. She also shared her script with me to perform the colocalisation analysis. Dr Felix R Day developed the MR script and shared it with me to conduct MR analysis. Prof John R.B. Perry and Ken K. Ong supervised the analyses and provided guidance.

3.2 Abstract

Previous GWAS analyses have produced many mechanistic insights into LOY, however the LOY calls used in these studies were mainly based on SNP-array data. The studies still lacked that systematic exploration of whether combining the Y dosage estimated from next-generation sequencing data can improve the power to detect more LOY-related signals. Illustrated in chapter 2, the LOY Combined Call (3-way) that combined both LOY measures estimated from SNP-array and whole-exome sequence data outperformed other LOY measures. In this chapter, I implemented in the same GWAS pipeline used by the previous GWAS analysis of PAR-LOY for 204,770 male participants in the UK Biobank. There were 22 novel LOY signals identified.

Previous analysis showed the mechanistic link between LOY, leukocyte telomere length (LTL) and blood cancers including myeloproliferative neoplasms (MPN). For MPN, the largest GWAS study identified only 17 MPN risk loci on over 1 million samples. Because it is hard to increase the sample size of MPN cases, instead, through combining the GWAS summary statistics of LOY, LTL and MPN, the shared underlying mechanisms were further revealed and 13 novel MPN signals were identified. Collectively, these results highlighted that the use of more accurate LOY estimation can improve the statistical power to detect LOY-related signals. Moreover, systematic evaluation of the impact of LOY-related signals on health outcomes can increase the power to detect loci that influence LOY-related health outcomes.

3.3 Introduction

Like other types of clonal mosaicism, there has been fruitful progress in understanding the causes and consequences of LOY over the past 10 years, largely due to the decreased cost of sequencing technology, the innovation of novel bioinformatic tools and the large cohorts consisting of hundreds of thousands of individuals[18]. The largest-scale study of LOY to date exploited the genotype data from SNP-based genotype array of over 200,000 males from UK Biobank by developing a new detection method and validated the findings of other large cohorts. This study substantially improved the understanding of the mechanisms behind LOY by conducting GWAS analysis[16]. However, the accurate estimation of LOY from SNP-array data largely depended on the number of probes on the Y chromosome, which varied from hundreds to thousands depending on the chips used for sequencing. The limited number of probes introduced technical noises in the estimation of LOY[16]. Therefore, whether more Y dosage information could be generated from next-generation sequencing data remained uninvestigated. The newly released whole-exome sequencing data of 450,000 samples from UK Biobank provided the chance to investigate this question. In chapter 2, LOY Combined Call (3-way) combined by the different LOY metrics estimated from both SNP-array and WES were observed to have improved statistical power. In this study, this variable was used to conduct the GWAS and its downstream analyses in the same manner as previous analyses conducted for PAR-LOY.

Previous analyses on LOY demonstrated that LOY plays a role as a biomarker of the defective DNA damage response and cell cycle regulation, which suggests that the genetic analysis of LOY represents an exclusive opportunity to identify variants associated with DNA damage response and cell cycle regulation that influence the risk of multiple types of cancer. This can be supported by the observations that the LOY-predisposing alleles carriers have a higher risk of breast, prostate, testicular, brain and renal cell cancers[16]. Arising from similar origins in haematopoietic stem and progenitor cells (HSPCs)[55–57], myeloid malignancies, such as myeloproliferative neoplasms (MPN), myelodysplastic syndromes (MDS) and acute myeloid leukaemia (AML)[18] become the primary diseases of interest due to the shared mechanisms with LOY. Based on nearly 1 million people, the largest scale GWAS

of MPN only identified 17 leading loci. The very low incidence (less than 1/100,000)[58] of MPN meant that it would be very difficult to get more cases to expend the scale of GWAS.

Given that it has a shared genetic architecture, to test the assumption that LOY can be used to detect more MPN risk loci and reveal any shared mechanisms, I also investigated leukocyte telomere length (LTL), which is associated with MPN, and performed several analyses, including LDSC[59], MR[60], Co-localisation[61] and MTAG[62] analyses.

In this chapter, the newly proposed LOY call not only increased the power of the known signals but also identified 22 novel LOY signals. Taken together, these results demonstrate the power of this novel approach. Furthermore, by integrating different methods based on GWAS summary statistics, 13 MPN risk loci were identified and the shared mechanisms of LOY, LTL and MPN were illustrated.

In summary, the improvement of LOY metric identified novel signals compared to the previous method using the same sample. Additionally, the analyses on MPN significantly increased the power to detect loci that influence MPN via the same mechanism that causes LOY.

3.4 Methods

3.4.1 Genetic association testing in the UK Biobank for newly proposed LOY metric

The study implemented the same pipeline to conduct GWAS for the new LOY measure as the previous analysis using PAR-LOY[16].

The linear mixed model implemented in BOLT-LMM (version: 2.3.2)[63] was performed to conduct the common variants genetics association testing, which can account for the cryptic population structure and relatedness. The genetic data with bgen format from V3[50] release of UK Biobank containing the complete set of Haplotype Reference Consortium (HRC) and 1K Genomes imputed variants were used as genotype input. The k-means clustering method was applied to the first four genetic principal components[8] to cluster the participants. Only participants identified as having "White European" ancestry were included in this study.

Participants with inconsistent ancestry identification answers in the questionnaire were excluded. Additionally, participants who were with abnormal sex chromosome karyotypes, failed to pass the criteria of quality control, withdrew their consent were also removed from the analysis.

Different from the previous GWAS on LOY, I used the novel LOY Combined Call (3-way) as the LOY metric. Because not all participants in UK Biobank had WES data, I imputed the LOY Combined Call (2-way) values for the samples without WES data. Therefore, the analysis included both genotype and phenotype data for 204,770 male participants. For BOLT-LMM, the genetic relationship matrix was included as a covariate, which was calculated from the genotyped variants which were: on autosomes with minor allele frequency (MAF)>1%, included in both genotyping arrays, and passed QC in all 106 batches[16]. The other covariates included genotyping chips, age at first check, and the first 10 genetic principal components. All the common variants genetic association testing analyses were conducted in the High-Performance Computing (HPC) system maintained by University of Cambridge.

A clumping algorithm was used to select the signals with a $P<5\times10^{-8}$, an imputation quality score (INFO)>0.5, and a MAF>0.1% at a 1-Mb window[16]. If the genome-wide significant leading signals shared any correlation with each other due to the long-range linkage disequilibrium ($r^2>0.05$), these signals were excluded from further analysis. The

approximate conditional analyses implemented in GCTA (Genome-wide Complex Trait Analysis)[64] were conducted to augment the leading loci. Signals which were uncorrelated (r^2<0.05) with previous identified leading signals and with $P<5\times10^{-8}$ before and after conditional analysis were also identified as leading signals.

Then, all leading signals were extracted from the published summary statistics of PAR-LOY, and all the signals with $P<5 \times 10^{-8}$ for LOY Combined Call (3-way) but not for PAR-LOY were considered potential novel signals. Extra conditional analyses were performed for all potential novel signals. If there were any signals correlated with one potential novel signal (r^2>0.05) or within a 1-Mb window and with $P<5\times10^{-8}$ or for PAR-LOY, this signal was not considered a novel signal for LOY. All the remaining signals were considered novel signals associated with LOY. These SNPs were mapped to the nearest protein coding gene within a 1-Mb window, according to its GRCh37 coordinates.

3.4.2 Exploring the mechanistic link between LOY, LTL and MPN

Previous epidemiological studies made it clear that LOY is associated with blood cancer. Myeloproliferative neoplasms are a group of blood cancers that are characterized by the excessive production of mature myeloid cells[57]. This study conducted further systematic analyses to explore the relationship among LOY and MPN. Additionally, the telomere length of leukocyte was also included in this study, as the identified 17 MPN risk loci include two telomere length associated genes *MECOM* and *TERT*[57]. The measured LTL GWAS summary statistics for the 472,174 UK Biobank participants was taken from Codd et al.[65] For MPN, the summary statistics of largest scale GWAS for MPN containing 2,949 cases of MPN and 835,554 controls was taken from Bao et al.[57]

3.4.2.1 LD Score regression

Linkage disequilibrium score regression (LDSC)[59] was used to calculate the pairwise genetic correlations between LOY, LTL and MPN. The basic European LD score reference panel was used, which was provided by the developers of LDSC and downloaded from their website (https://data.broadinstitute.org/alkesgroup/LDSCORE/eur_w_ld_chr.tar.bz2). The LD scores were generated from 1000 Genomes and contained 1,217,312 SNPs. For the GWAS summary statistics used in LDSC, the SNPs with INFO≤0.9 and MAF≤0.01 were removed.

3.4.2.2 Mendelian randomisation

I conducted the Bi-directional Mendelian Randomisation (MR) analyses between LOY, LTL and MPN. MR analyses are often used to estimate the directional association between exposure and outcome traits, using signals from GWAS summary statistics as instrumental variables (IVs). The advantage of MR is that it can mimic the biological link between exposure and outcome traits[66]. The IV for MR analyses must satisfy three assumptions: 1) it is associated with the risk factor, 2) it is not associated with any confounder of the risk factor–outcome association, 3) it is conditionally independent of the outcome given the risk factor and confounders[67,68].

I chose the leading signals identified from GWAS of LOY Combined Call (3-way) as the IVs for LOY. The independent sentinel variants associated with LTL were used as the IVs for LTL[65]. For MPN, due to the limited number of its associated signals, I used both signals that reached genome-wide significance ($P<5\times10^{-8}$) or suggestive significance ($P<1\times10^{-6}$)[57] as IVs. As LOY is specific to males, I removed the IVs on the X chromosome from all MR analyses. If any signals were missing in the outcome summary statistics, I collected proxies for these signals using GCTA with European UK Biobank individuals as reference (within 1 MB of reported signals and $r^2>0.4$). I chose the proxy of each missing signal with the largest r^2 value as the replacement IV, which was contained in both GWAS summary statistics of the exposure and outcome.

I extracted the summary statistics of IVs for both exposure and outcome phenotype from the original GWAS summary statistics files. Then, I aligned the IVs of exposure phenotype to increasing allele and the IVs of outcome phenotype were subsequently realigned accordingly. Steiger filtering as implemented in R package 'TwoSampleMR' [69] was used to filter out IVs which may lead to reverse causality, due to them having a more significant association with the outcome than the exposure. Next, the IVs that were identified as outliers according to Rücker's Q' statistic[70] were further excluded. The remaining IVs were used to conduct the MR analysis.

The MR inverse-variance weighted (MR-IVW) model was used as the primary model to conduct MR[60]. Compared with other MR methods, the MR-IVW can provide high statistical power. I checked the sensitivity of MR models based on the degree of heterogeneity (I^2

statistics and Cochran's Q-derived P-value), horizontal pleiotropy (MR-Egger $P_{intercept} < 0.05$), and funnel and dosage plots. To account for potential horizontal pleiotropy and heterogeneity, three additional MR tests were performed: MR-EGGER[71], weighted median (MR-WM)[72], and penalised weighted median (MR-PWM).

3.4.2.3 Co-localisation analysis

Bayesian testing was used to assess whether two association signals were consistent with a shared causal variant, looking for colocalisation between pairs of both LTL and MPN, and LOY and MPN, and using their summary statistics and the leading signals through implementation of R package 'coloc' (Version: 5.1.0)[61]. SNPs with h4.pp (the posterior probability that both traits are associated and share a single causal variant)≥0.75 were defined as co-localised causal variants for both traits.

3.4.2.4 Multi-trait analysis of GWAS

GWAS summary statistics for LOY, LTL and MPN were used to conduct a meta-analysis by implementing the multi-trait analysis of GWAS (MTAG)[62,73].

Based on the summary statistics from GWAS of multiple correlated traits, MTAG can enhance the statistical power to identify genetic associations for each trait included in the analysis. There are several advantages of MTAG: 1. It works on an arbitrary number of traits' GWAS summary statistics and doesn't require individual-level data; 2. The summary statistics aren't needed to generate from independent study cohorts: MATG can account for the sample overlap between GWAS summary statistics by conducting bivariate LDSC[59]; 3. MTAG can provide the effect size estimation of all SNPs for all traits analysed; 4. Even when many traits are included, MTAG is computationally quick because every step has a closed-form solution[62].

For an individual SNP, the MTAG results were calculated in three steps: 1. Estimation of the variance-covariance matrix of GWAS estimation error using a sequence of LD score regression; 2. Estimation of the variance-covariance matrix of the SNP effects using the method of moments; 3. Finally, for each SNP, these estimates were substituted into the equation described by Turley et al[62].

I performed the MTAG analysis using the Python command line tool. Prior to the analysis, I excluded the variants with MAF<0.01 from the summary statistics of all three traits. A potential problem for MTAG is that SNPs can be totally flat for one trait but not flat for another trait, but this can cause the MTAG's effect size estimations of these SNPs for the first trait to shift away from 0. Then, this causes the false positive rate (FDR) to increase.[62,73] Therefore, I estimated the max FDR for each trait by invoking "—fdr" when running MTAG in command line.

The same clumping algorithm used for LOY was applied on the summary statistics of MPN generated from MTAG. For all leading signals, I extracted their summary statistics from the original GWAS summary statistics. In total, 35 independent leading signals were identified. I then applied the Bonferroni correction for the identified signals. I further excluded the signals with P>0.05/35=0.00143 in the original GWAS to avoid the issues mentioned above, as GWAS for both LOY and LTL identified many more leading signals than MPN, which thus increased the FDR for MPN.

3.5 Results

3.5.1 Identifying novel leading signals for LOY

This study identified 20,025 Genome-wide significant ($P<5\times10^{-8}$) SNPs and 173 independent signals **(Figure 3-2)**. The most significant associated SNP (rs2887399, $P=8\times10^{-164}$) was in the first identified LOY-related gene, *TCL1A*[30]. Among these SNPs all the 156 signals[16] identified by previous GWAS of PAR-LOY passed genome-wide significance, with the exception of three signals with p-value ($P<7.1\times10^{-7}$) . For all 156 previously identified signals using PAR-LOY, there was an average 11% increase in test statistic (χ^2 value) in my new GWAS using LOY Combined Call (3-way) **(Figure 3-1)**. Compared with the PAR-LOY GWAS, the genomic inflation factor lambda GC (the ratio of expected to observed median test statistics) of this study remained at 1.2, but the overall mean χ^2 value increased from 1.47 to 1.54. There was no evidence of signal inflation due to population structure, as the LDSC intercept was 1.01. This evidence illustrates that the GWAS of LOY Combined Call (3-way) has stronger statistical power than the PAR-LOY GWAS. After clumping and excluding the previously identified signals, 22 novel LOY-associated leading signals remained, with p-values ranging from 7.1×10^{-10} to 4.8×10^{-8} **(Table 3-1)**. These signals were also nominally significantly associated with PAR-LOY with p-values from 6.8×10^{-8} to 1.9×10^{-5}.

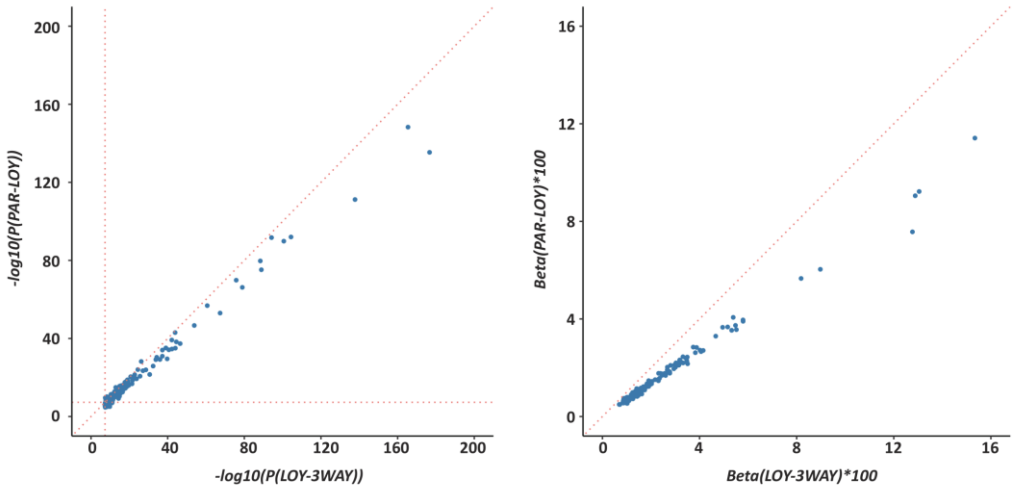


Figure 3-1 The comparison of estimated -$\log_{10}$(p-value) (left) and effect size (right) of 173 independent signals from the GWAS of LOY Combined Call (3-way) (x-axis) and PAR-LOY (y-axis).

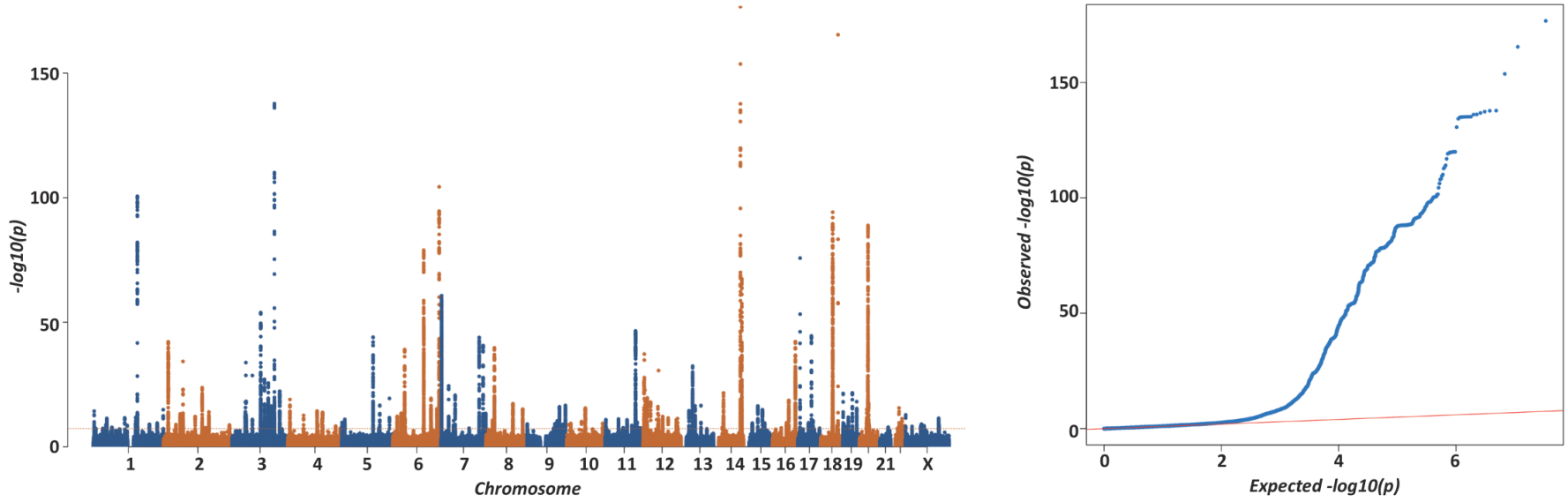


Figure 3-2 Manhattan plot and quantile–quantile (Q-Q) plot illustrating the results of the GWAS of LOY Combined Call (3-way) in 204,770 male participants in UK Biobank. Orange dotted line indicates genome-wide significance level ($P<5\times10^{-8}$).

Table 3-1 Novel signals identified for LOY based on LOY Combined Call (3-way). Their summary statistics for PAR-LOY are also shown. All alleles are aligned to LOY-increasing alleles. A1/A0: effect/non-effect allele, A1FREQ: allele frequency of effect allele, BETA: estimated effect size of the effect allele, SE: standard error, *P*: P-values from BOLT-LMM models.

LOY (3WAY)								PAR-LOY (Thompson et al.[16], 2019)			
SNP	**CHR**	**BP**	**A1/A0**	**A1FREQ**	**BETA**	**SE**	***P***	**BETA**	**SE**	***P***	**Nearest Gene**
rs6427752	1	198795389	C/T	0.483	0.010	0.002	9.10E-10	0.006	0.001	2.50E-06	*PTPRC*
rs11211005	1	44937451	G/A	0.785	0.012	0.002	1.30E-09	0.008	0.001	1.30E-07	*RNF220*
rs34985293	1	65654059	CAG/C	0.318	0.010	0.002	1.50E-08	0.006	0.001	2.50E-06	*AK4*
1:33418349_CAA_C	1	33418349	C/CAA	0.291	0.009	0.002	4.40E-08	0.006	0.001	1.50E-06	*RNF19B*
rs11121242	1	8906301	G/A	0.513	0.009	0.002	4.80E-08	0.006	0.001	3.60E-06	*ENO1*
rs77552263	2	43786818	G/A	0.922	0.017	0.003	1.20E-09	0.011	0.002	6.50E-07	*THADA*
rs113823725	5	131563501	C/G	0.539	0.008	0.002	3.20E-08	0.006	0.001	6.70E-07	*P4HA2*
rs1124275	6	158622125	A/G	0.861	0.013	0.002	9.00E-09	0.008	0.002	1.50E-06	*GTF2H5*
6:37938688_ACAAAC_A	6	37938688	A/ACAAAC	0.091	0.014	0.003	4.50E-08	0.009	0.002	2.50E-06	*ZFAND3*
rs79516659	9	93942899	C/T	0.932	0.016	0.003	3.80E-08	0.009	0.002	1.90E-05	*AUH*
rs7116797	11	116707338	G/A	0.893	0.014	0.003	2.10E-08	0.009	0.002	8.00E-07	*APOA1*
rs73031459	11	124636645	A/G	0.031	0.028	0.005	3.10E-08	0.018	0.004	1.40E-06	*MSANTD2*
rs663503	11	128587411	T/C	0.641	0.009	0.002	4.60E-08	0.006	0.001	3.30E-06	*FLI1*
rs2159599	12	710441	G/A	0.257	0.010	0.002	1.70E-08	0.006	0.001	8.70E-07	*NINJ2*
rs61976859	14	31583512	T/C	0.102	0.015	0.003	1.90E-08	0.009	0.002	1.50E-05	*HECTD1*
rs62057094	16	31128004	G/T	0.892	0.014	0.003	2.70E-08	0.009	0.002	1.40E-06	*KAT8*
rs11077394	17	76693711	G/A	0.554	0.009	0.002	3.70E-08	0.006	0.001	1.60E-06	*CYTH1*
rs62131484	19	4012097	G/A	0.895	0.016	0.003	7.10E-10	0.010	0.002	7.80E-08	*PIAS4*
rs11701821	21	38892075	T/C	0.347	0.010	0.002	1.30E-08	0.006	0.001	1.90E-06	*DYRK1A*
rs5996675	22	24618331	A/G	0.790	0.011	0.002	3.40E-08	0.007	0.001	5.50E-07	*GGT5*
rs769428149	23	66944119	C/CA	0.859	0.010	0.002	3.50E-09	0.006	0.001	8.30E-07	*AR*
23:24064168_CAT_C	23	24064168	CAT/C	0.413	0.007	0.001	7.00E-09	0.005	0.001	6.50E-08	*EIF2S3*

3.5.2 Mechanistic links between LOY, LTL and MPN

To check the genetic correlations between LOY, LTL and MPN, I conducted the LDSC using their GWAS summary statistics. The p-value for all three LDSC was not significant, but LOY and MPN showed a positive trend (r_g=0.34). In contrast, the genetic correlation coefficients for LTL and LOY and for LTL and MPN were negligible, -0.05 and 0.02 respectively **(Table 3-2)**.

Table 3-2 The test statistics for the pair-wise LD Score regression among LOY, LTL and MPN. p1: trait 1, p2: trait 2, rg: genetic correlation, se: standard error of rg, z: Z-score, p: p-value

p1	p2	rg	se	z	*p*
LOY (3WAY)	**MPN**	0.34	0.21	1.62	0.11
LOY (3WAY)	**LTL**	-0.05	0.05	-0.98	0.33
MPN	**LTL**	0.02	0.10	0.20	0.84

I conducted MR analyses to check whether there were directional associations among LOY, MPN and LTL. From MR results after applying both radial and Steiger filters, higher LOY was associated with increased risk of MPN ($beta_{MR\text{-}IVW}$=0.98, $SE_{MR\text{-}IVW=}$0.18, $P_{MR\text{-}IVW}=1.6\times10^{-7}$) but MPN was not correspondingly associated with risk of LOY ($P_{MR\text{-}IVW}$>0.05). In addition, longer LTL was associated with higher risk of MPN ($beta_{MR\text{-}IVW}$=0.71, $SE_{MR\text{-}IVW}$=0.13, $P_{MR\text{-}IVW}=2\times10^{-7}$) but, conversely, MPN was associated with shorter LTL ($beta_{MR\text{-}IVW}$=-0.01, $SE_{MR\text{-}IVW}$=0.002, $P_{MR\text{-}IVW}=5\times10^{-4}$). On the other hand, longer LTL was associated with less LOY ($beta_{MR\text{-}IVW}$=-0.06, $SE_{MR\text{-}IVW}$=0.01, $P_{MR\text{-}IVW}=5.8\times10^{-11}$) but there was no significant effect of LOY on LTL ($P_{MR\text{-}IVW}$>0.05). Other than the primary model MR-IVW, all other sensitivity MR models generated consistent estimates **(Table 3-3, Figure 3-4)**.

Table 3-3 Bi-directional pair-wise Mendelian randomisation results among LOY, LTL and MPN with Steiger and Radial filters. n_IVs: number of instrumental variables, betaIVW: effect size estimated from MR-IVW model, sebetaIVW: standard error of effect size of MR-IVW model, pIVW: p-value of MR-IVW model, CochQp: Cochran's Q-derived P-value, Isq: I^2 statistics, , betaEGGER: effect size estimated from MR-EGGER model, sebetaEGGER: standard error of effect size of MR-EGGER model, pEGGER: p-value of MR-EGGER model, interEGGER: intercept estimated from MR-EGGER model, seinterEGGER: standard error of the intercept of MR-EGGER model, pinterEGGER: p-value of the intercept of MR-EGGER, betaWM: effect size estimated from MR-WM model, sebetaWM: standard error of effect size of MR-WM model, pWM: p-value of MR-WM model, betaPWM: effect size estimated from MR-PWM model, sebetaPWM: standard error of effect size of MR-PWM model, pPWM: p-value of MR-PWM model

Exposure	Outcome	n_IVs	betaIVW	sebetaIVW	pIVW	CochQp	Isq	betaEGGER	sebetaEGGER	pEGGER
LOY (3WAY)	**MPN**	114	0.98	0.18	1.6E-07	0.34	4.65	1.04	0.34	3.1E-03
LOY (3WAY)	**LTL**	91	-0.01	0.02	4.5E-01	0.34	5.38	-0.01	0.04	7.7E-01
MPN	**LOY (3WAY)**	8	0.00	0.00	4.2E-01	0.53	0.00	0.00	0.01	9.9E-01
MPN	**LTL**	12	-0.01	0.00	5.0E-04	0.42	2.02	-0.01	0.00	1.2E-02
LTL	**LOY (3WAY)**	84	-0.06	0.01	5.8E-11	0.45	1.30	-0.07	0.01	3.4E-06
LTL	**MPN**	97	0.71	0.13	2.0E-07	0.53	0.00	1.29	0.24	4.2E-07

Exposure	Outcome	interEGGER	seinterEGGER	pinterEGGER	betaWM	sebetaWM	pWM	betaPWM	sebetaPWM	pPWM
LOY (3WAY)	**MPN**	-0.001	0.01	8.5E-01	0.89	0.28	1.7E-03	0.89	0.29	2.6E-03
LOY (3WAY)	**LTL**	0.000	0.00	9.6E-01	-0.04	0.03	1.6E-01	-0.04	0.02	1.5E-01
MPN	**LOY (3WAY)**	0.001	0.00	6.8E-01	0.00	0.01	7.0E-01	0.00	0.01	6.8E-01
MPN	**LTL**	0.000	0.00	8.9E-01	-0.01	0.00	2.0E-03	-0.01	0.00	2.0E-03
LTL	**LOY (3WAY)**	0.000	0.00	4.7E-01	-0.07	0.01	1.1E-07	-0.07	0.01	1.9E-07
LTL	**MPN**	-0.018	0.01	3.1E-03	0.97	0.22	3.3E-05	0.89	0.21	6.4E-05

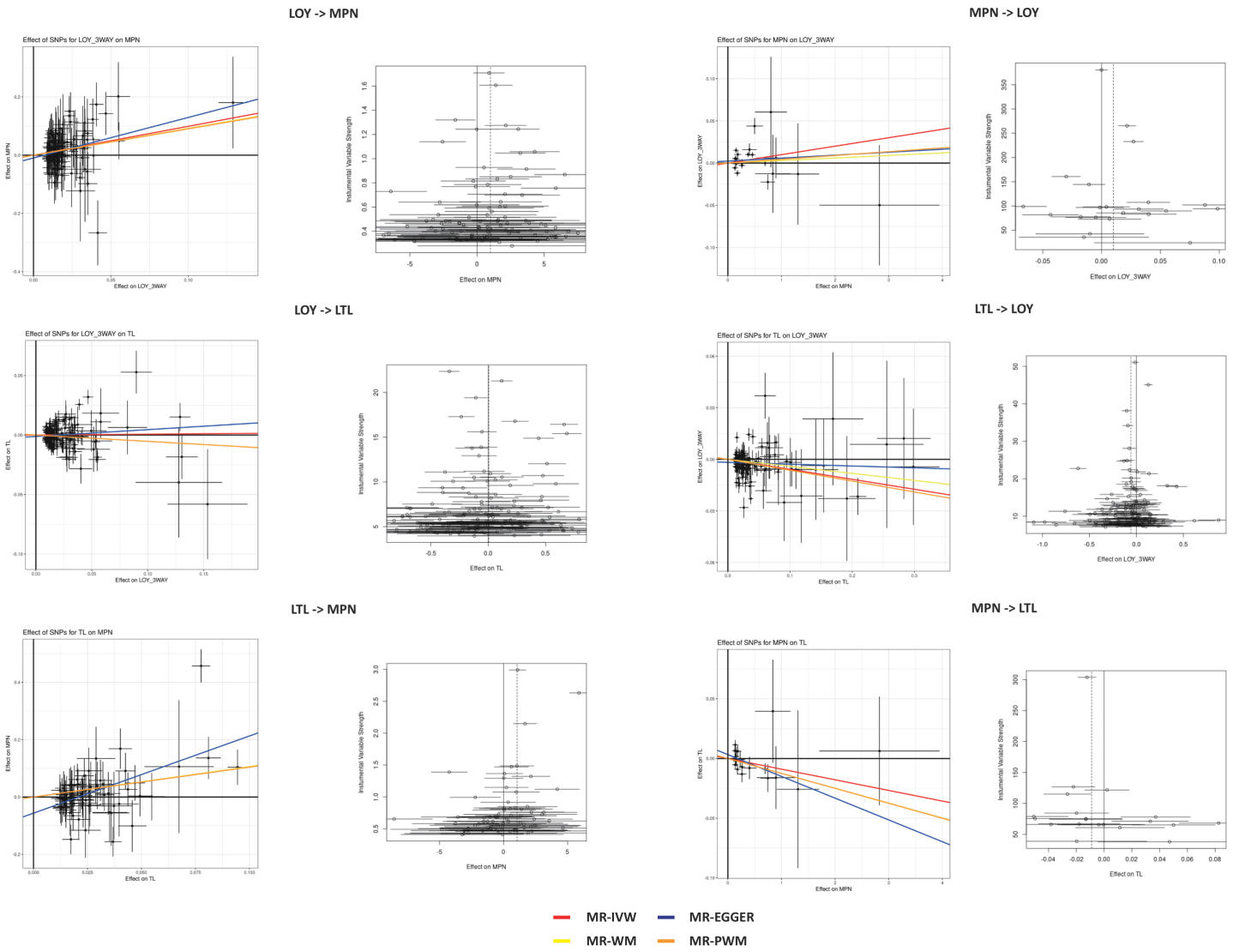


Figure 3-3 Scatter and funnel plots for the MR analyses.

Based on co-localisation analysis, 12 leading SNPs for LOY co-localised with MPN. These mapped to identified MPN genes including: *NPAT*, *MECOM*, *TET2*, *TERT*, *ATM* and *GFI1B* and other genes at sub genome-wide significance for MPN ($P<3.5\times10^{-5}$) including *PARP1*, *DLK1*, *TP53*, *NREP*, *MAD1L1* and *RBPMS*. In addition, 5 leading SNPs for LTL co-localised with MPN. These mapped to *TERT*, *NFE2*, *PARP1* and *ATM*. Of note, leading SNPs at *TERT*, *PARP1* and *ATM* colocalised with all 3 traits, LOY, LTL and MPN **(Table 3-4, Table 3-5)**.

Table 3-4 Co-localised SNPs between LOY and MPN and their nearest mapped genes. h0.pp: posterior probability of the hypothesis that no association with either trait, h1.pp: posterior probability of the hypothesis that association with trait 1, not with trait 2, h2.pp: posterior probability of the hypothesis that association with trait 2, not with trait 1, h3.pp: posterior probability of the hypothesis that association with trait 1 and trait 2, two independent SNPs, h4.pp: posterior probability of the hypothesis that association with trait 1 and trait 2, one shared SNP, coloc_SNP_a0_a1: the shared SNP associated with both trait 1 and trait 2.

	co-localisation test statistics						LOY (3WAY)								MPN			
LOY (3WAY) signal	**h0.pp**	**h1.pp**	**h2.pp**	**h3.pp**	**h4.pp**	**coloc_SNP_a0_a1**	**SNP**	**CHR**	**BP**	**A1/A0**	**A1FREQ**	**BETA**	**SE**	***P***	**BETA**	**SE**	***P***	**Nearest Gene**
rs138994074	2.5E-06	1.3E-02	7.3E-06	3.6E-02	0.95	1_226537388_T_C	rs4653728	1	226537388	C/T	0.85	0.014	0.002	8.5E-10	0.18	0.04	1.8E-06	*PARP1*
rs7129527	3.8E-44	3.6E-05	1.7E-40	1.6E-01	0.84	11_108105593_A_G	rs228595	11	108105593	A/G	0.41	0.023	0.002	2.1E-46	0.16	0.03	1.9E-09	*ATM*
rs56116444	2.5E-38	7.6E-03	2.2E-38	5.7E-03	0.99	5_111061847_G_T	rs56116444	5	111061847	G/T	0.07	0.041	0.003	9.9E-45	-0.27	0.06	2.8E-06	*NREP*
rs2280548	4.9E-55	2.0E-02	1.5E-54	6.0E-02	0.92	7_1976457_T_C	rs1801368	7	1976457	T/C	0.40	0.027	0.002	2.5E-61	0.13	0.03	4.0E-06	*MAD1L1*
rs1824914	4.0E-20	4.4E-04	1.1E-17	1.2E-01	0.88	3_168838408_A_G	rs2293661	3	168838408	G/A	0.45	0.016	0.002	4.8E-23	0.14	0.03	5.1E-08	*MECOM*
4:105864529_ACT_A	1.3E-23	3.0E-22	5.3E-04	1.1E-02	0.99	4_105806108_T_A	rs144317085	4	105806108	A/T	0.97	0.025	0.004	5.9E-09	-0.72	0.07	7.2E-26	*TET2*
rs72698720	4.5E-63	1.9E-02	6.2E-63	2.6E-02	0.96	14_101178555_A_G	rs72698718	14	101178555	G/A	0.87	0.041	0.002	6.3E-68	0.18	0.04	4.5E-06	*DLK1*
rs13167280	2.0E-51	2.7E-48	3.0E-05	4.0E-02	0.96	5_1285974_A_C	rs7705526	5	1285974	A/C	0.33	0.010	0.002	3.9E-09	0.46	0.03	4.8E-54	*TERT*
rs2853677	2.0E-51	2.7E-48	3.0E-05	4.0E-02	0.96	5_1285974_A_C	rs7705526	5	1285974	A/C	0.33	0.010	0.002	3.9E-09	0.46	0.03	4.8E-54	*TERT*
rs2979469	1.9E-31	4.7E-02	4.9E-32	1.1E-02	0.94	8_30285091_C_G	rs2979469	8	30285091	C/G	0.74	0.023	0.002	1.8E-40	0.14	0.03	8.8E-06	*RBPMS*
rs78378222	6.0E-38	5.3E-02	5.5E-39	3.9E-03	0.94	17_7578671_T_C	rs35850753	17	7578671	T/C	0.02	0.079	0.006	6.3E-47	0.37	0.09	3.5E-05	*TP53*
rs621940	5.1E-15	4.0E-05	2.5E-12	1.9E-02	0.98	9_135870130_G_C	rs621940	9	135870130	C/G	0.84	0.018	0.002	2.4E-17	-0.20	0.04	5.8E-09	*GFI1B*

Table 3-5 Co-localised SNPs between LTL and MPN and their nearest mapped genes.

	co-localisation test statistics						LTL								MPN			
LTL signal	**h0.pp**	**h1.pp**	**h2.pp**	**h3.pp**	**h4.pp**	**coloc_SNP_a0_a1**	**SNP**	**CHR**	**BP**	**A1/A0**	**A1FREQ**	**BETA**	**SE**	***P***	**BETA**	**SE**	***P***	**Nearest Gene**
rs61748181	0.0E+00	6.6E-50	8.6E-278	1.4E-13	1.00	5_1285974_A_C	rs7705526	5	1285974	A/C	0.327	0.078	0.002	2.4E-282	0.46	0.03	4.8E-54	*TERT*
rs7705526	0.0E+00	6.6E-50	8.6E-278	1.4E-13	1.00	5_1285974_A_C	rs7705526	5	1285974	A/C	0.327	0.078	0.002	2.4E-282	0.46	0.03	4.8E-54	*TERT*
rs79977579	1.1E-11	6.0E-03	4.7E-12	1.7E-03	0.99	12_54698408_A_G	rs79755767	12	54698408	A/G	0.096	0.028	0.003	4.8E-16	0.21	0.04	1.5E-06	*NFE2*
rs932002	3.0E-42	1.7E-02	8.9E-42	4.9E-02	0.93	1_226577306_T_C	rs932002	1	226577306	C/T	0.849	0.040	0.003	7.3E-47	0.17	0.04	2.6E-06	*PARP1*
rs611646	4.0E-71	2.4E-05	1.8E-67	1.1E-01	0.89	11_108141701_T_G	rs651030	11	108141701	G/T	0.592	0.037	0.002	6.0E-73	-0.16	0.03	2.0E-09	*ATM*

Finally, I performed MTAG analysis to boost the power of MPN GWAS to identify more MPN-associated signals. According to the output of MTAG, the max FDR of MPN was 0.1, which meant that some of the leading MPN signals were driven by LOY and LTL. The same clumping algorithm was applied to the GWAS summary statistics of MPN generated from MTAG. 35 independent MPN signals were identified after clumping and conditional & joint association analysis. However, 12 of these signals showed weak associations ($P>1.4\times10^{-3}$) in the original MPN GWAS and these were dropped to avoid reporting false positive signals. Among the remaining 23 MPN signals, 10 were identified by the original MPN GWAS analysis as having a genome-significant *P*-value. Finally, 13 novel signals for MPN were identified from MTAG analysis, which included co-localised SNPs near *PARP1*, *MAD1L1*, *DLK1* and *RBPMS* and *TP53*, in which missense mutations are the most common mutations in human cancers **(Figure 3-4, Table 3-6)**.

To reveal the biological functions of these MPN leading signals, enrichment analysis was conducted based on the list of nearest genes mapped by these signals. There were 37 pathways from different databases with evidence of enrichment. The results from both KEGG and WP showed associations with the apoptosis pathway. Besides, the associated pathways from KEGG included many cancer-related pathways including human T-cell leukaemia virus 1 infection, chronic myeloid leukaemia, small cell lung cancer, viral carcinogenesis, and pancreatic cancer, which were driven by the known cancer related genes *PARP1*, *NFKBIA*, *TP53*, *BCL2L1* and *MAD1L1*.

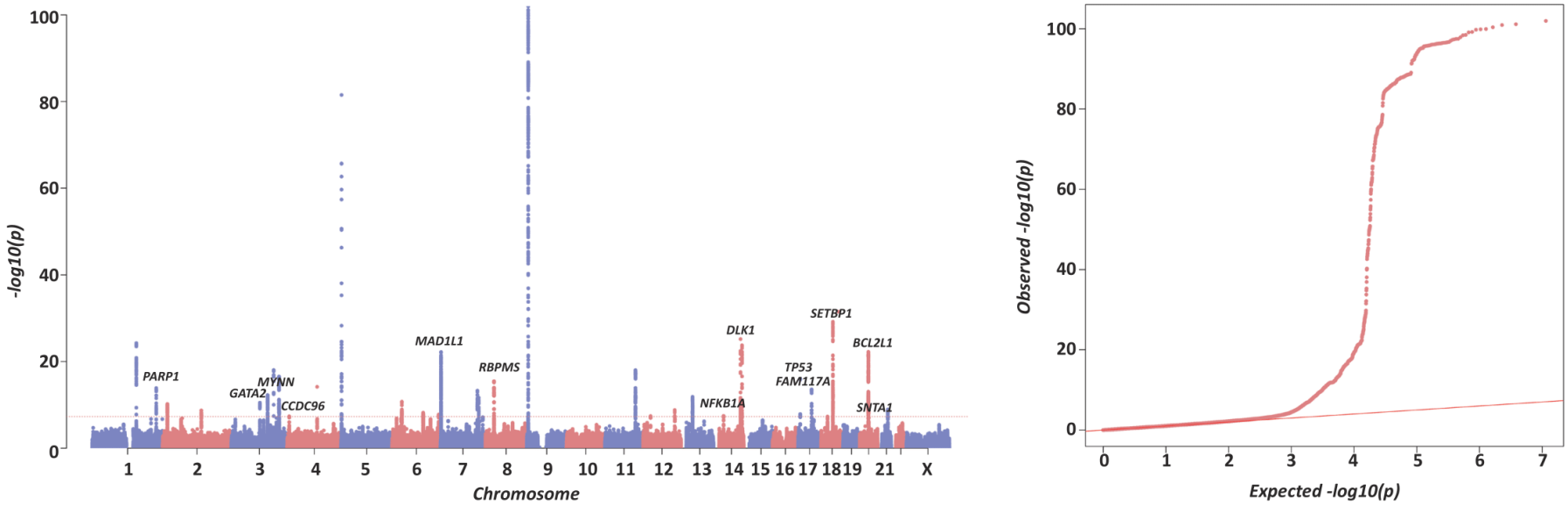


Figure 3-4 Manhattan plot and quantile–quantile (Q-Q) plot illustrating the summary statistics for MPN from MTAG. Pink dot line indicates genome-wide significance level ($P<5\times10^{-8}$).

Table 3-6 Novel signals identified for MPN from MTAG and the comparison with summary statistics from the original MPN GWAS.

MPN (MTAG)								MPN (Bao et al.[57], 2020)			
SNP	**CHR**	**BP**	**A1/A0**	**A1FREQ**	**BETA**	**SE**	***P***	**BETA**	**SE**	***P***	**Nearest Gene**
rs76887998	1	226539353	T/C	0.83	0.014	0.002	1.36E-14	0.176	0.037	1.71E-06	*PARP1*
rs2712431	3	128316890	A/C	0.69	0.011	0.002	5.54E-13	0.144	0.029	6.72E-07	*GATA2*
rs1920123	3	169506173	C/T	0.73	0.009	0.002	3.70E-09	0.100	0.030	9.13E-04	*MYNN*
rs3951348	4	7040389	T/C	0.78	0.009	0.002	4.53E-08	0.129	0.033	9.07E-05	*CCDC96*
rs4719366	7	1977906	A/G	0.37	0.014	0.001	6.57E-23	0.129	0.027	2.65E-06	*MAD1L1*
rs2979488	8	30280630	G/A	0.74	0.013	0.002	3.39E-16	0.135	0.030	9.48E-06	*RBPMS*
rs72698718	14	101178555	G/A	0.85	0.020	0.002	1.84E-24	0.177	0.039	4.48E-06	*DLK1*
rs2233406	14	35874799	G/A	0.72	0.008	0.002	3.61E-08	0.124	0.030	3.46E-05	*NFKBIA*
rs35850753	17	7578671	T/C	0.02	0.041	0.005	1.11E-16	0.367	0.089	3.48E-05	*TP53*
rs118035610	17	47804307	C/T	0.94	0.022	0.003	2.66E-14	0.202	0.060	7.70E-04	*FAM117A*
rs11082395	18	42072716	T/C	0.13	0.023	0.002	6.89E-30	0.150	0.038	8.09E-05	*SETBP1*
rs291688	20	32009341	A/C	0.10	0.014	0.002	2.11E-09	0.155	0.043	3.14E-04	*SNTA1*
rs6089050	20	30312945	C/T	0.78	0.017	0.002	5.91E-23	0.123	0.037	7.77E-04	*BCL2L1*

3.6 Discussion

Using the newly proposed LOY call incorporating whole-exome sequence read depth information in UK Biobank male participants, this study identified 20 more novel LOY signals, which proved the power of the strategy. However, compared with the 2.5-fold increase in the χ^2 association statistics from mLRR-Y to PAR-LOY[16], only an 11% increment was observed and most of the signals already had sub genome-wide significant p-values in PAR-LOY models, which means that the method only provides a small boost in power. Importantly, this study only analysed the white-European males in UK Biobank. With the growing number of other large cohorts with more diverse participants around the world, it is important to extend the analysis by including more samples from different ancestries, in order to increase the detection power. The trans-ethnic meta-analysis approach could also increase the ability to fine-map the causal variants, due to reduced linkage disequilibrium windows.

As a relatively high prevalent genetic biomarker indicating the status of defective DNA damage response and cell cycle regulation, evidence on the functional consequences of LOY is still very limited, but previous epidemiological studies already identified the associations between LOY and a wide range of health outcomes. This study aimed to examine whether LOY can improve knowledge about these health outcomes. By using the summary statistics for LOY, LTL and MPN and conducting LDSC, bi-directional MR, colocalisation and MTAG analyses, the shared mechanism behind these three traits were revealed. From LDSC analyses, no significant genetic correlation was observed between LOY, LTL and MPN. However, the MR results demonstrated significant positive associations between LOY, LTL and MPN. This apparent inconsistency might be because there are directional differences and heterogeneous relationships between these 3 traits. Additionally, both colocalisation and MTAG analyses identified *PARP1*, whose inhibitor was approved to use as a treatment of *BRCA1/2* deficient cancers[74]. From the summary statistics of MPN generated by MTAG, more MPN associated signals were identified, which illustrated the power of using LOY as an DNA damage indicator.

In the future, this approach could be implemented in other LOY-associated health outcomes beyond MPN in order to provide mechanistic insights into the links between LOY, DNA damage response, and these health outcomes. They will also identify novel targets of disease susceptibility which will likely not be identified through the conventional approaches. Although these *in-silico* analyses provided reliable evidence about the mechanisms of LOY and its related health outcomes, more experimental work will be needed to illuminate the mechanisms and then help to develop any potential therapeutic targets.

Chapter 4 Detection and characterisation of male sex chromosome abnormalities in the UK Biobank study

Published in:

Genetic Medicine

Yajie Zhao, Eugene J. Gardner, Marcus A. Tuke, Huairen Zhang, Maik Pietzner, Mine Koprulu, Raina Y. Jia, Katherine S. Ruth, Andrew R. Wood, Robin N. Beaumont, Jessica Tyrrell, Samuel E. Jones, Hana Lango Allen, Felix R. Day, Claudia Langenberg, Timothy M. Frayling, Michael N. Weedon, John R.B. Perry, Ken K. Ong*, Anna Murray*

**These authors jointly supervised this work*

Online ahead of print 9 June, 2022.

doi:10.1016/j.gim.2022.05.011

4.1 Contributions

This chapter describes the work I performed on male sex chromosome abnormalities in the UK Biobank study. I identified the males with sex chromosome abnormalities from their mLRR-Y and mLRR-X values and then validated them using WES data. I extracted and processed the phenotypes from UK Biobank and performed the regression analyses using both BOLT and GLM pipelines. I drafted this chapter with the help of Prof Ken K. Ong and Dr Eugene J. Gardner. Prof Claudia Langenberg, Dr Maik Pietzner and Mine Koprulu conducted the PheWAS analysis of this paper. Raina Y. Jia performed the analyses for NMR metabolic biomarkers. Dr Hana Lango Allen designed the pipeline to estimate read depth from WES data. Dr Felix R. Day prepared some phenotypes used in this chapter. Dr Eugene J. Gardner helped to draft the method part and provided many useful advice on the analyses. The other colleagues from University of Exeter ran the same analyses for most of traits mentioned in this chapter independently, which replicated the results generated from this chapter. Prof John R.B. Perry, Prof Ken K. Ong and Prof Anna Murray supervised the work conducted in this chapter. All the contributors of this chapter reviewed and checked writing of this chapter.

4.2 Abstract

To systematically ascertain male sex chromosome abnormalities, 47,XXY (Klinefelter syndrome) and 47,XYY, and characterise their risks of adverse health outcomes. I analysed genotyping array or exome sequence data in 207,067 men of European ancestry aged 40-70 from the UK Biobank and related these to extensive routine health record data. Only 49/213 (23%) of men whom we identified with Klinefelter syndrome and only 1/143 (0.7%) with 47,XYY had a diagnosis of abnormal karyotype on their medical records or self-report. We observed expected associations for Klinefelter syndrome with reproductive dysfunction (late puberty: risk ratio, RR=2.7; childlessness: RR=4.2; testosterone concentration -3.8 nmol/L, all $P<2\times10^{-8}$), whereas XYY men appeared to have normal reproductive function. Despite this difference, we identified several higher disease risks shared across both Klinefelter syndrome and 47,XYY, including Type 2 diabetes (RR=3.0 and 2.6, respectively), venous thrombosis (RR=6.4 and 7.4), pulmonary embolism (RR=3.3 and 3.7), and chronic obstructive pulmonary disease (RR=4.4 and 4.6; all $P<7\times10^{-6}$). Klinefelter syndrome and 47,XYY were mostly unrecognised but conferred substantially higher risks of metabolic, vascular and respiratory diseases, which were only partially explained by higher levels of BMI, deprivation and smoking.

4.3 Introduction

The most common sex chromosome aneuploidies in men are 47,XXY (Klinefelter syndrome) and 47,XYY, with population prevalence estimates of 100 per 100,000 and 18-100 per 100,000,[75,76] respectively. Men with Klinefelter syndrome typically present during adolescence with delayed puberty or as adults with infertility. Other recognised features include tall adult stature, high body fat percentage,[77] poor muscle tone, low bone mineral density, and increased risks of neurocognitive disability, psychoses, and disorders of personality.[78] Klinefelter syndrome has also been associated with higher risks of Type 2 diabetes and venous thromboembolism.[79,80] By contrast, 47,XYY is less well characterised as many of these individuals may not present to health services and thus are unaware of their karyotype. Reported features associated with 47,XYY may therefore be affected by sampling bias. These include tall stature, scoliosis, learning difficulties,[81] poor muscle tone,[82] increased central fat, and increased risks of seizures, asthma, and emotional and behavioural problems (e.g., autism and attention deficit disorder).[83] While infertility has been reported in some men with XYY, most studies report normal sexual development and fertility.[84]

Previous studies identified men with Klinefelter syndrome or 47,XYY from medical records, and therefore case ascertainment was based on recognition of their typical phenotypic features. Therefore, the reported penetrance of these features may have been biased and the full spectrum of clinical features overlooked. A more robust alternative approach is to identify such individuals from large population-based studies using systematic measurements to produce unbiased estimates of the effects of sex chromosome aneuploidy on unselected diseases. We recently used this approach to show that mosaic X chromosome aneuploidy in women (mosaic Turner's syndrome, 45,X) conferred a lower penetrance of infertility than had been reported by earlier clinic-based studies.[85]

Here, I analysed single-nucleotide polymorphism (SNP) array genotype data in 207,067 men of European ancestry aged 40-70 from the UK Biobank. I identified 213 men with sex chromosome aneuploidy indicative of Klinefelter syndrome and 143 men with 47,XYY, and related these karyotypes to extensive study data and medical records to understand the

penetrance of male sex chromosome aneuploidy on typical reproductive outcomes and its wider clinical impacts.

4.4 Methods

4.4.1 Study population

The UK Biobank is a large prospective cohort which recruited approximately 500,000 participants aged 40 to 70 years across the island of Great Britain. A broad range of phenotypic and health-related information was collected from each participant, including physical measurements, lifestyle indicators, biomarkers in blood and urine, imaging and routine health record data.[50] UK Biobank has approval from the North West Multi-centre Research Ethics Committee (REC reference 21/NW/0157) and informed consent was provided by each participant.

In the UK Biobank, 488,377 participants had DNA samples assayed using one of two genotyping arrays: UK Biobank Lung Exome Variant Evaluation (UK BiLEVE study, N=49,950) and Affymetrix Axiom UK Biobank array (UK Biobank Axiom (Affymetrix), N=438,427). These two arrays tested 807,411 and 825,927 SNPs respectively, with 95% overlap between arrays.[50] We restricted our analysis to men of 'white European' genetic ancestry as classified by the approach previously described by Thompson et al.[16] Briefly, this approach uses k-means-clustering to group individuals by the first four genetic principal components. Additionally, I excluded individuals who were classified as White European by our k-means approach but self-identified as being of ancestry other than White European.[16] I further excluded individuals whose samples failed UK Biobank genotyping quality control parameters and those who withdrew consent. Accordingly, 207,067 men were included in all association testing analyses. I was unable to incorporate non-European individuals when modelling the relationship between abnormal karyotypes and phenotypes outlined in this manuscript. I identified only 16 non-white European males with abnormal karyotypes: 9 with 47,XXY and 7 with 47,XYY.

4.4.2 Identification of male sex chromosome aneuploidy heterozygotes from SNP array data

To identify men with sex chromosome aneuploidy, I downloaded genotyping fluorescence signal intensity (log2ratios, LRR) and quality control (QC) information for all SNPs on the X

(chrX) and Y chromosomes (chrY) from the UK Biobank data showcase (https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=22431 and https://biobank.ndph.ox.ac.uk/showcase/refer.cgi?id=1955). I excluded SNPs that: i. were located within Pseudo-Autosomal Regions (PAR),[86] ii. Did not have a calculable LRR on both arrays, iii. Did not pass QC in all 106 batches, or iv. Were flagged as failing QC by UK Biobank. After these steps, 16,599 chrX SNPs and 579 chrY SNPs remained. I then calculated the median LRR across all remaining SNPs on chrX and chrY to generate the values mLRR-X and mLRR-Y, respectively. These values represent the median fluorescence signal intensities across the entire X or Y chromosome.[32] Using the thresholds described by Bycroft et al.,[50] men with [-1≤mLRR-Y<0.23 and mLRR-X>-0.2] were categorised as having 47,XXY (Klinefelter syndrome) and men with [mLRR-Y≥0.23 and mLRR-X<-0.2] as 47,XYY (men with [mLRR-Y≥0.23 and mLRR-X>-0.2] were categorised as 48,XXYY and were not included in further analyses).

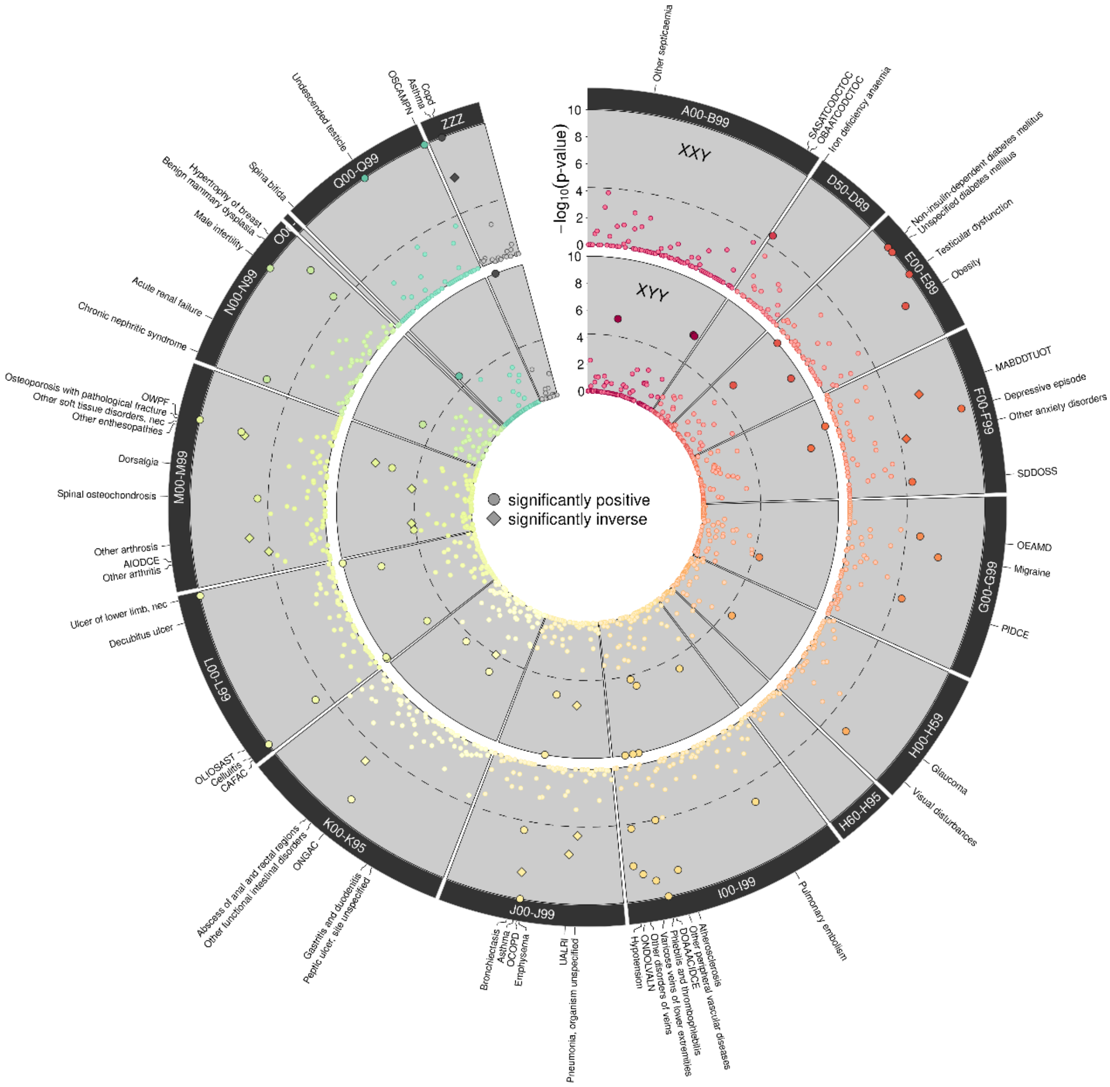
4.4.3 Confirmation of male sex chromosome aneuploidy heterozygotes from exome sequencing data

To confirm sex chromosome aneuploidy status using an orthogonal approach, we used exome sequencing data available for 83,104 white European men in UK Biobank.[87,88] To estimate sex chromosome dosage, I calculated the average read depth of three target regions: i. non-PAR regions on chrX, ii. X-degenerate regions (XDRs) on chrY as defined by Skov et al.,[53] and iii. Autosomes. First, we used *samtools (*version: 1.9) to convert the provided CRAM files for each participant to Binary Alignment Map (BAM) files based on the GRCh38 reference sequence (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/GRCh38_reference_genome/GRCh38_full_analysis_set_plus_decoy_hla.fa). UK Biobank provided the GRCh38 coordinates of the targeted regions for its exome sequencing design with a BED (Browser Extensible Data) file (https://biobank.ndph.ox.ac.uk/showcase/ukb/auxdata/xgen_plus_spikein.GRCh38.bed). I created three subsets of this BED file by extracting the overlap between the target regions and non-PAR regions on chrX, XDRs on chrY, and autosomes according to their GRCh38 coordinates. Then, they were converted to Picard Interval Lists using the Picard (version:

2.21.6-SNAPSHOT) function *BedToIntervalList*, based on the same reference sequence. Using these Picard Interval Lists, the BAM file of each participant was inputted to calculate the average coverages of non-PAR regions on chrX, XDRs on chrY, and autosomes using the Picard function *CollectHsMetrics*. The relative read depth of non-PAR regions on chrX and XDRs on chrY were defined as the average coverage in each of these regions divided by the average coverage across the autosomes. The relative read depth of non-PAR regions on chrX and XDRs on chrY multiplied by 2 were used as a proxy of chrX dosage and chrY dosage, respectively. Men with [X dosage>1.2] were categorised as having 47,XXY (Klinefelter syndrome) and men with [Y dosage>1.5] as 47,XYY. Men with [X dosage>1.2 and Y dosage>1.5] were categorised as 48,XXYY.

4.4.4 Disease association testing

To test for the disease burden associated with male sex chromosome aneuploidies, we performed logistic regression models with Klinefelter syndrome or 47,XYY (coded '1') compared to the normal male karyotype 46,XY (coded '0') as the exposure. Outcomes comprised 875 ICD-10 coded diseases amalgamated from death registries, hospital episode statistics, primary care records (in a subset, N=94,959), and self-reported conditions (from the 'first occurrence of disease' dataset released by UK Biobank). The dataset contains further 19 case definitions from dedicated working groups that used multiple sources for case identifications, such as for chronic obstructive pulmonary disease or end stage renal disease. For each participant, all events from all sources were mapped to an ICD-10 code and the date of the first disease occurrence from any source was taken as the event date. From this dataset, we filtered out likely erroneous disease events if the disease occurrence date: (i) was unknown or missing, ii) matched or preceded the date or year of birth, (iii) occurred after the dataset release date. We performed logistic regression models in R (version: 3.6.0) among unrelated men of 'white European' genetic ancestry (maximum N=162,322) and adjusted for age at study baseline, test centre, and the first 10 genetically derived principal components. Resulting odds ratios (OR) were converted to risk ratios (RR) using the formula described by Zhang and Yu.[89] We applied a stringent Bonferroni corrected p-value threshold of $P<0.05/875=5.7\times10^{-5}$ to define statistical significance **(Supplementary Table 4-1 (Appendix B), Figure 4-1)**.



Other septicaemia
A00-B99
XXY
XYY
−log10(p-value)
10
8
6
4
2
0
SASATCODCTOC
OBAATCODCTOC
Iron deficiency anaemia
D50-D89
Non-insulin-dependent diabetes mellitus
Unspecified diabetes mellitus
Testicular dysfunction
Obesity
E00-E89
MABDDTUOT
Depressive episode
Other anxiety disorders
F00-F99
SDDOSS
OEAMD
Migraine
G00-G99
PIDCE
H00-H59
Glaucoma
Visual disturbances
H60-H95
Pulmonary embolism
I00-I99
Atherosclerosis
Other peripheral vascular diseases
DOAAACIDCE
Phlebitis and thrombophlebitis
Varicose veins of lower extremities
Other disorders of veins
ONDOLVALN
Hypotension
UALRI
Pneumonia, organism unspecified
J00-J99
Bronchiectasis
Asthma
OCOPD
Emphysema
Gastritis and duodenitis
Peptic ulcer, site unspecified
K00-K95
ONGAC
Abscess of anal and rectal regions
Other functional intestinal disorders
OLIOSAST
Cellulitis
CAFAC
L00-L99
Ulcer of lower limb, nec
Decubitus ulcer
Other arthrosis
AIODCE
Other arthritis
Spinal osteochondrosis
Dorsalgia
Osteoporosis with pathological fracture
OWPF
Other soft tissue disorders, nec
Other enthesopathies
M00-M99
Chronic nephritic syndrome
Acute renal failure
N00-N99
Male infertility
Benign mammary dysplasia
Hypertrophy of breast
Spina bifida
Undescended testicle
Q00-Q99
OSCAMPN
Asthma
Copd
ZZZ
significantly positive
significantly inverse

Figure 4-1 Circos plot summarizing phenome-wide disease association tests for KS and 47,XYY compared to 46,XY. Each segment represents each ICD-10 chapter in lexicographical order. P-values (on a negative logarithmic scale) were from logistic regression models for KS (outer circle) and XYY (inner circle) with each of 875 ICD-10 coded disease outcomes, adjusted for age and ten principal genetic components. Outcomes reaching the multiple testing corrected statistical significance threshold ($P<0.05/875=5.7\times10^{-5}$; dashed line) are indicated by large circles (for positive associations) and diamonds (for negative associations). SASATCODCTOC=streptococcus and staphylococcus as the cause of diseases classified to other chapters; OBAATCODCTOC=other bacterial agents as the cause of diseases classified to other chapters; MABDDTUOT=mental and behavioural disorders due to use of tobacco; SDDOSS=specific developmental disorders of scholastic skills; OEAMD=other extrapyramidal and movement disorders; PIDCE=polyneuropathy in diseases classified elsewhere; DOAAACIDCE=disorders of arteries, arterioles and capillaries in diseases classified elsewhere; ONDOLVALN=other non-infective disorders of lymphatic vessels and lymph nodes; UALRI=unspecified acute lower respiratory infection; OCOPD=other chronic obstructive pulmonary disease; ONGAC=other non-infective gastro-enteritis and colitis; CAFAC=cutaneous abscess, furuncle and carbuncle; OLIOSAST=other local infections of skin and subcutaneous tissue; AIODCE=arthropathies in other diseases classified elsewhere; OWPF=osteoporosis without pathological fracture; OSCAMPN=other sex chromosome abnormalities, male phenotype, NEC; NEC=not elsewhere classified

4.4.5 Study phenotype association testing

To test the association of male sex chromosome aneuploidy status against selected anthropometric, reproductive, metabolic, cardiovascular, learning/memory, and behavioural study-measured traits **(Supplementary Table 4-2 (Appendix B))**, I used a linear mixed model implemented in BOLT-LMM (version: 2.3.2)[63]. The outcome 'childlessness' was derived from the response 'zero' to the question "How many children have you fathered?" among men aged 55 and older. The Townsend Deprivation index was used as an indicator of socio-economic status, based on participants' home postcodes. The two binary exposure variables described above were converted to BGEN file format using PLINK2 (version:2.00-alpha) and inputted to BOLT-LMM via the *bgenFile* flag. A genetic relationship matrix (GRM) was generated based on all autosomal variants which had minor allele frequency (MAF)>1%, passed QC in all 106 batches and were present on both genotyping arrays. Genotyping chip, age at baseline, and the first 10 genetically derived principal components were included as covariates. For binary outcomes, we also performed logistic regression and calculated the RR from the OR as described above.

4.4.6 NMR metabolic biomarkers association testing

We analysed 168 circulating metabolic traits measured by proton nuclear magnetic resonance (NMR) spectroscopy (Nightingale Health Plc.) in non-fasting plasma samples in UK Biobank men with 46,XY (N=49,806), Klinefelter syndrome (N=48) or 47,XYY (N=38). For each metabolic traits we first performed adjustment for technical variations using the R package *ukbnmr* (https://www.medrxiv.org/content/10.1101/2021.09.24.21264079v2.), then performed inverse rank normalisation, and then further adjusted for sex, age at the first study visit, BMI and the first 10 genetically derived principal components. Associations between abnormal karyotype and each metabolic trait were tested in separate linear regression models **(Supplementary Figure 4-1 (Appendix B))**.

4.5 Results

4.5.1 Prevalence of Male Sex Chromosome Aneuploidy in a Population Scale Biobank

Using genotyping array data, of 207,067 men of European ancestry, we identified 213 men with 47,XXY (Klinefelter syndrome (KS), prevalence 103/100,000) and 143 with 47,XYY (69/100,000; Figure 2A). Of these cases who also had exome sequencing data, we observed 100% confirmation of aneuploidy status (62/62 men with KS and 54/54 men with 47,XYY) **(Figure 4-2B,C)**.

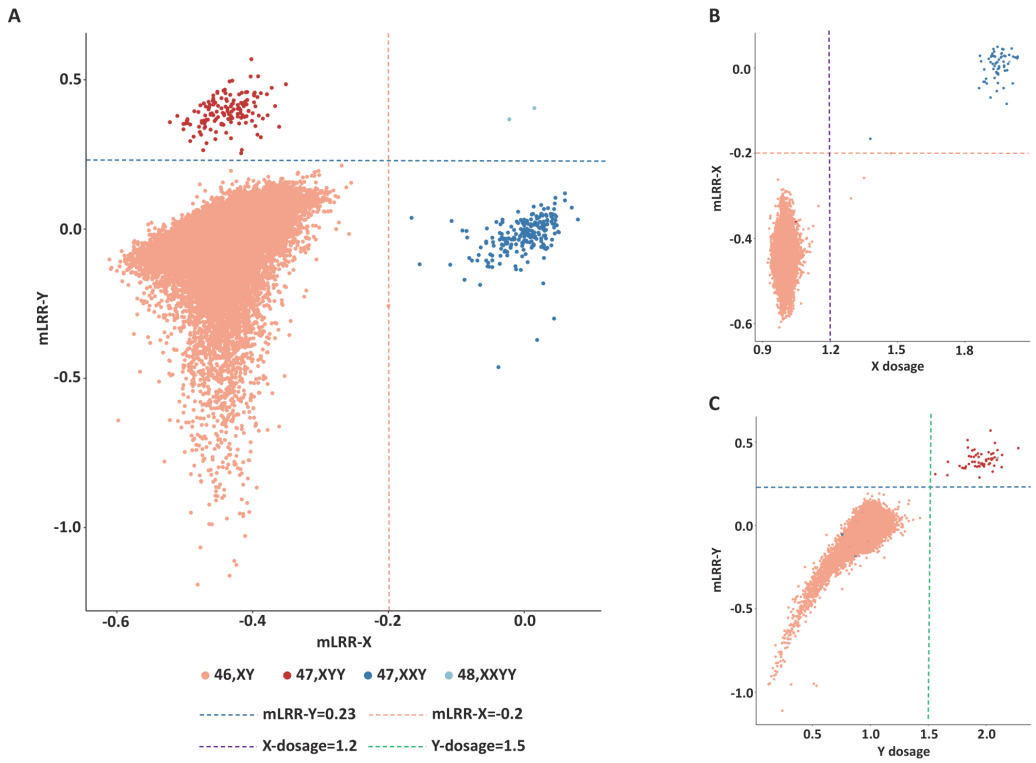


Figure 4-2 (A) Median array genotype intensity on the X (mLRR-X) and Y (mLRR-Y) chromosomes for each of N=207,067 men, including 213 with 47,XXY (Klinefelter syndrome), 143 with 47,XYY and 2 with 48,XXYY. **(B)** X dosage estimated from exome sequencing plotted against mLRR-X (N=83,104). **(C)** Y dosage estimated from exome sequencing plotted against mLRR-Y (N=83,104).

Only 49/213 (23.0%) of men with KS and 1/143 (0.7%) with 47,XYY had a diagnosis of sex chromosome abnormality on routine medical records or self-reported data (ICD10: Q98 Other sex chromosome abnormalities, male phenotype, not elsewhere classified). Similar

proportions were found in the subsample of men who had primary care data: only 24/89 (27.0%) with KS and 1/76 (1.3%) with 47,XYY had known sex chromosome abnormality. Conversely, of the men with a diagnosis of sex chromosome abnormality on their health record, by our analysis we classified four as 46,XX (mLRR-Y<-1) and a further eight as having a normal male karyotype.

4.5.2 Quantification of typical features of 47,XXY and 47,XYY

Compared to men with a normal karyotype (46,XY), men with KS and 47,XYY had taller adult height, by 2.7 cm (P=7×10^{-15}) and 7.9 cm (P=8×10^{-77}), respectively, and were more likely to have 'taller than average' childhood height (RR=1.3, P=0.01 and RR=1.7, P=5×10^{-10}, respectively). Men with KS and 47,XYY were more likely to be childless (RR=4.2, P=4×10^{-117} and RR=2.4, P=2×10^{-17}, respectively), but only those with KS (not 47,XYY) were more likely to report later than average puberty timing (RR=2.7, P=2×10^{-8}) **(Table 4-1)**.

In addition, men with KS and 47,XYY were less likely to have a university or college degree (RR=0.39, P=5×10^{-8} and RR=0.50, P=9×10^{-5}, respectively) and had lower fluid intelligence test scores (beta=-2.1, P=3×10^{-15} and -1.6, P=5×10^{-8}, respectively), were more likely to be smokers (RR=1.1, P= 0.28 and RR=1.2, P=8×10^{-3}, respectively), report depressive episodes (RR=2.4, P=6×10^{-10} and RR=2.7, P=3×10^{-9}) and live in areas with higher deprivation index (beta=1.6, P=1×10^{-15} and 1.6, P=1×10^{-10}, respectively). Men with KS and 47,XYY were also more likely to live without a partner (RR=2.2, P=4×10^{-20} and RR=2.1, P=1×10^{-12}) and report loneliness and isolation (RR=2.2, P=4×10^{-13} and RR=2.5, P=3×10^{-13}) and poor overall health (RR=4.2, P=2×10^{-26} and RR=3.8, P=2×10^{-14}) **(Table 4-1)**.

Table 4-1 Typical features of Klinefelter syndrome and 47,XYY compared to men with normal (46,XY) karyotypes

	46,XY	KS	47,XYY	KS vs. 46,XY			47,XYY vs. 46,XY		
Continuous traits	**Mean (SD)**	**Mean (SD)**	**Mean (SD)**	**Beta**	**95% CI**	***P***	**Beta**	**95% CI**	***P***
Height (cm)	175.9 (6.8)	178.7 (7.6)	184.4 (7.6)	2.7	2.0 - 2.7	7×10^{-15}	7.9	7.0 - 7.9	8×10^{-77}
Fluid intelligence test scores	6.3 (2.2)	4.2 (1.9)	4.7 (1.9)	-2.2	-2.7 to -2.2	3×10^{-15}	-1.6	-2.1 to -1.6	5×10^{-8}
Townsend deprivation index	-1.4 (3.0)	0.3 (3.5)	0.2 (3.5)	1.6	1.2 - 1.6	1×10^{-15}	1.6	1.1 - 1.6	1×10^{-10}
Binary traits	**%**	**%**	**%**	**RR**	**95% CI**	***P***	**RR**	**95% CI**	***P***
Childless	21.2%	87.6%	51.8%	4.1	3.9 - 4.3	4×10^{-117}	2.4	2.0 - 2.8	2×10^{-17}
'Taller than average' childhood height	31.7%	39.9%	55.6%	1.3	1.0 - 1.5	1×10^{-2}	1.7	1.5 - 2.2	5×10^{-10}
Late puberty timing	6.2%	17.4%	7.3%	2.7	1.8 - 3.8	2×10^{-8}	1.1	0.5 - 2.0	8×10^{-1}
University or college degree	40.4%	16.3%	20.2%	0.39	0.3 - 0.6	5×10^{-8}	0.50	0.3 - 0.7	9×10^{-5}
Depressive episodes	9.1%	22.1%	24.5%	2.4	1.9 - 3.0	6×10^{-10}	2.7	2.0 - 3.5	3×10^{-9}
Ever smoked	51.6%	54.5%	61.0%	1.1	0.95 - 1.2	3×10^{-1}	1.2	1.1 - 1.4	8×10^{-3}
Lives alone	17.2%	37.4%	35.3%	2.1	1.8 - 2.5	3×10^{-14}	2.1	1.6 - 2.5	2×10^{-8}
Lives without a partner	22.7%	50.0%	48.6%	2.2	1.9 - 2.5	4×10^{-20}	2.1	1.8 - 2.5	1×10^{-12}
Loneliness, isolation	14.8%	33.3%	37.5%	2.2	1.8 - 2.6	4×10^{-13}	2.5	2.0 - 3.0	3×10^{-13}
Poor overall health	5.0%	21.1%	19.1%	4.2	3.2 - 5.3	2×10^{-26}	3.8	2.6 - 5.2	2×10^{-14}
Long-standing illness or infirmity	35.7%	65.7%	63.8%	1.9	1.7 - 2.1	1×10^{-19}	1.8	1.6 - 2.1	4×10^{-13}

KS, Klinefelter syndrome

Beta, Regression coefficient from linear regression models

RR, Relative risk from logistic regression models

4.5.3 Anthropometric features of men with 47,XXY and 47,XYY

Compared to 46,XY men, those with KS and 47,XYY had higher BMI (beta=1.8 Kg/m^2, $P=8\times10^{-11}$ and 2.2, $P=5\times10^{-11}$, respectively), higher percent total body fat (beta=4.8%, $P=8\times10^{-40}$ and 2.2%, $P=3\times10^{-7}$) and weaker hand grip strength (beta=-7.1 Kg, $P=1\times10^{-36}$ and -2.6, $P=3\times10^{-4}$) (Table 2). Other features differed between the groups. Even after accounting for their taller adult heights, men with 47,XYY (but not KS) had higher total fat-free mass (beta=3.8 Kg, $P=4\times10^{-15}$). Furthermore, men with 47,XYY had slightly higher bone mineral density (BMD) (beta=0.04 g/cm^2, $P=2\times10^{-2}$), whereas men with KS had lower BMD (-0.05, $P=5\times10^{-6}$) and higher likelihood of osteoporosis with pathological fracture (RR=10.8, $P=2\times10^{-7}$) **(Table 4-2)**.

Table 4-2 Anthropometric characteristics of Klinefelter syndrome and 47,XYY compared to men with normal (46,XY) karyotypes

	46,XY	KS	47,XYY	KS vs. 46,XY			47,XYY vs. 46,XY		
Continuous traits	**Mean (SD)**	**Mean (SD)**	**Mean (SD)**	**Beta**	**95% CI**	***P***	**Beta**	**95% CI**	***P***
BMI (Kg/m2)	27.9 (4.2)	29.7 (5.7)	30.2 (5.7)	1.8	1.2 - 2.3	8×10^{-11}	2.2	1.6 - 2.9	5×10^{-11}
Body fat percentage (%)	29.5 (5.4)	34.3 (6.0)	31.6 (5.7)	4.8	4.1 - 5.6	9×10^{-40}	2.2	1.4 - 3.1	4×10^{-7}
Hand grip strength (Kg)	41.9 (8.9)	35.2 (8.9)	39.9 (9.7)	-7.1	-8.2 to -6.0	1×10^{-36}	-2.6	-3.9 to -1.2	3×10^{-4}
Fat-free mass, adj. height (Kg)	63.9 (7.7)	66.1 (9.4)	73.5 (9.2)	0.2	-0.6 to 0.9	6×10^{-1}	3.8	2.8 - 4.7	4×10^{-15}
Bone mineral density (g/cm2)	0.6 (0.2)	0.5 (0.2)	0.6 (0.2)	-0.05	-0.08 to -0.03	5×10^{-6}	0.04	0.01 - 0.08	2×10^{-2}
Binary traits	**%**	**%**	**%**	**RR**	**95% CI**	***P***	**RR**	**95% CI**	***P***
Osteoporosis with pathological fracture	0.3%	2.4%	0.7%	10.8	4.5 - 25.3	2×10^{-7}	3.1	0.4 - 20.6	3×10^{-1}
Osteoporosis without pathological fracture	1.6%	8.9%	0.7%	6.2	4.0 - 8.9	5×10^{-15}	0.5	0.1 - 3.4	5×10^{-1}

KS, Klinefelter syndrome

Beta, Regression coefficient from linear regression models

RR, Relative risk from logistic regression models

4.5.4 Hormonal, metabolic and vascular features of men with 47,XXY and 47,XYY

Compared to 46,XY men, those with KS (but not 47,XYY) had lower plasma total testosterone concentration (beta=-3.8 nmol/L, $P=2\times10^{-50}$). Both KS and 47,XYY men had lower plasma IGF-1 concentrations (beta=-1.7 nmol/L, $P=1\times10^{-6}$ and -2.3, $P=3\times10^{-8}$, respectively).

We also identified several adverse metabolic and vascular health outcomes shared across both KS and 47,XYY, including higher risks for Type 2 diabetes (RR=3.0, $P=2\times10^{-20}$ and RR=2.6, $P=3\times10^{-10}$, respectively), albuminuria (RR=1.9, $P=5\text{x}10^{-4}$ and RR=2.4, $P=5\text{x}10^{-6}$), venous thrombosis (RR=6.4, $P=3\text{x}10^{-23}$ and RR=7.4, $P=7\text{x}10^{-22}$), pulmonary embolism (RR=3.3, $P=2\text{x}10^{-6}$ and RR=3.7, $P=7\text{x}10^{-6}$) and atherosclerosis (RR=3.1, $P=6\text{x}10^{-3}$ and RR=5.5, $P=8\text{x}10^{-6}$). These disease associations were only partially attenuated after adjustments for BMI, household deprivation and smoking **(Table 4-3)**. Exploration of red blood cell and platelet traits showed lower haematocrit and haemoglobin concentrations in KS (but not 47,XYY) men, but no obvious explanation for higher thrombosis risk was found **(Supplementary Table 4-2 (Appendix B))**.

Men with KS and 47,XYY had lower levels of HDL cholesterol (beta=-0.11 mmol/L, $P=2\times10^{-7}$ and -0.17, $P=3\times10^{-11}$, respectively) and men with 47,XYY (but not KS) had higher triglycerides (beta=0.32 mmol/L, $P=1\times10^{-3}$). In the subgroup with NMR metabolic data, we observed that KS (n=48) and XYY (n=38) men had lower levels of most HDL-related traits. Furthermore, 47,XYY (but not KS) men showed lower levels across all lipid classes, apart from triglycerides **(Supplementary Figure 4-1 (Appendix B))**.

4.5.5 Respiratory features of men with 47,XXY and 47,XYY

Compared to 46,XY men, those with KS and 47,XYY had lower forced expiratory volume (beta=-0.68 L, $P=1\times10^{-46}$ and -0.26, $P=2\times10^{-5}$, respectively) peak expiratory flow (beta=-125 L/min, $P=2\times10^{-38}$ and -68, $P=4\times10^{-8}$) and vital capacity (beta=-0.93 L, $P=2\times10^{-50}$ and -0.25, $P=3\times10^{-3}$), and higher risks for chronic obstructive pulmonary disease (RR=4.4, $P=5\times10^{-18}$ and RR=4.6, $P=2\times10^{-13}$) and asthma (RR=2.0, $P=9\times10^{-9}$ and RR=1.7, $P=4\times10^{-3}$). Again, these

disease associations were only partially attenuated after adjustments for BMI, household deprivation and smoking **(Table 4-3)**.

Table 4-3 Hormonal, metabolic, vascular and respiratory characteristics of men with Klinefelter syndrome and 47,XYY compared to men with normal (46,XY) karyotypes

	46,XY	KS	47,XYY	KS vs. 46,XY						47,XYY vs. 46,XY					
				Baseline model			adj. BMI, TDI and smoking			Baseline model			adj. BMI, TDI and smoking		
Continuous traits	**Mean (SD)**	**Mean (SD)**	**Mean (SD)**	**Beta**	**95% CI**	***P***	**Beta**	**95% CI**	***P***	**Beta**	**95% CI**	***P***	**Beta**	**95% CI**	***P***
Testosterone (nmol/L)	12.0 (3.7)	8.2 (6.0)	11.5 (4.7)	-3.8	-4.3 to -3.3	2×10^{-50}	-3.3	-3.8 to -2.8	2×10^{-41}	-0.5	-1.1 to 0.1	9×10^{-2}	0.1	-0.5 to 0.7	8×10^{-1}
SHBG (nmol/L)	39.9 (16.8)	41.1 (22.7)	40.3 (20.4)	2.3	0.1 - 4.4	2×10^{-2}	4.2	2.1 - 6.2	3×10^{-5}	1.3	-1.4 to 4.0	3×10^{-1}	3.7	1.1 - 6.2	2×10^{-3}
IGF-1 (nmol/L)	21.9 (5.5)	20.4 (6.5)	20.0 (5.7)	-1.7	-2.3 to -1.0	1×10^{-6}	-1.2	-1.9 to -0.5	7×10^{-4}	-2.3	-3.1 to -1.4	3×10^{-8}	-1.8	-2.6 to -0.9	2×10^{-5}
HDL cholesterol (mmol/L)	1.3 (0.3)	1.2 (0.3)	1.1 (0.3)	-0.11	-0.15 to -0.07	2×10^{-7}	-0.06	-0.10 to -0.02	5×10^{-3}	-0.17	-0.22 to -0.12	3×10^{-11}	-0.10	-0.15 to -0.06	1×10^{-5}
LDL cholesterol (mmol/L)	3.5 (0.9)	3.4 (0.9)	3.4 (0.9)	-0.10	-0.22 to 0.01	6×10^{-2}	-0.06	-0.18 to 0.05	3×10^{-1}	-0.15	-0.29 to -0.01	2×10^{-2}	-0.11	-0.25 to 0.03	9×10^{-2}
Triglycerides (mmol/L)	2.0 (1.2)	2.1 (1.0)	2.3 (1.4)	0.10	-0.05 to 0.25	2×10^{-1}	-0.01	-0.16 to 0.13	8×10^{-1}	0.32	0.14 - 0.51	1×10^{-3}	0.13	-0.05 to 0.31	3×10^{-1}
Forced expiratory volume (L)	3.3 (0.8)	2.7 (0.8)	3.2 (0.9)	-0.7	-0.8 to -0.6	1×10^{-46}	-0.6	-0.7 to -0.5	1×10^{-39}	-0.3	-0.4 to -0.1	2×10^{-5}	-0.2	-0.3 to -0.1	5×10^{-3}
Peak expiratory flow (L/min)	469 (138)	348 (132)	408 (132)	-125	-144 to -106	2×10^{-38}	-117	-136 to -98	3×10^{-34}	-68	-92 to -44	4×10^{-8}	-58	-81 to -34	4×10^{-6}
Vital capacity (L)	4.5 (1.0)	3.5 (0.9)	4.3 (1.0)	-0.9	-1.1 to -0.8	2×10^{-50}	-0.8	-0.9 to -0.7	4×10^{-41}	-0.2	-0.4 to -0.1	3×10^{-3}	-0.1	-0.03 to 0.03	2×10^{-1}
Binary traits	**%**	**%**	**%**	**RR**	**95% CI**	***P***	**RR**	**95% CI**	***P***	**RR**	**95% CI**	***P***	**RR**	**95% CI**	***P***
Type 2 diabetes	10.3%	28.6%	25.0%	3.0	2.4 - 3.6	2×10^{-20}	2.3	1.7 - 2.9	5×10^{-12}	2.6	1.9 - 3.4	3×10^{-10}	1.8	1.2 - 2.5	3×10^{-5}
Albuminuria	13.9%	26.6%	30.3%	1.9	1.3 - 2.6	5×10^{-4}	1.7	1.2 - 2.4	4×10^{-3}	2.4	1.7 - 3.2	5×10^{-6}	2.1	1.4 - 2.9	1×10^{-4}
Venous thrombosis	1.8%	10.9%	12.6%	6.4	4.2 - 9.1	3×10^{-23}	5.4	3.5 - 7.8	2×10^{-22}	7.4	4.6 - 10.9	7×10^{-22}	5.6	3.4 - 8.6	2×10^{-18}
Pulmonary embolism	2.5%	4.9%	7.5%	3.3	2.0 - 5.2	2×10^{-6}	2.8	1.7 - 4.5	5×10^{-5}	3.7	2.1 - 6.2	7×10^{-6}	2.8	1.5 - 4.9	9×10^{-4}
Atherosclerosis	1.0%	2.8%	4.9%	3.1	1.4 - 6.8	6×10^{-3}	2.2	0.9 - 5.2	9×10^{-2}	5.5	2.6 - 11.2	8×10^{-6}	4.4	2.1 - 9.1	1×10^{-4}
COPD	4.2%	16.4%	16.8%	4.4	3.2 - 6.0	5×10^{-18}	3.7	2.6 - 5.1	2×10^{-12}	4.6	3.1 - 6.5	2×10^{-13}	3.3	2.1 - 5.0	2×10^{-7}
Asthma	11.8%	23.5%	19.6%	2.0	1.6 - 2.6	1×10^{-7}	1.9	1.5 - 2.5	1×10^{-6}	1.7	1.2 - 2.3	4×10^{-3}	1.6	1.1 - 2.2	1×10^{-2}

KS, Klinefelter syndrome

Beta, Regression coefficient from linear regression models

RR, Relative risk from logistic regression models

TDI, Townsend Deprivation Index

4.6 Discussion

Using systematic case ascertainment in a large, unselected population of men of European ancestry aged 40-70, we report the prevalence of KS (103/100,000) and 47,XYY (69/100,000). Notably, only a small minority of these men had a diagnosis of sex chromosome abnormality on their medical records or by self-report (23% of KS and 0.7% of 47,XYY) and yet these conditions conferred substantially increased risks for multiple, potentially preventable diseases.

The under-diagnosis of KS and 47,XYY has been previously indirectly quantified in other settings, based on the differences in their clinical prevalence compared to estimates from population-based cytogenetic surveys in new-born infants. Such studies estimated that only between 7% (in UK) and 57% (in Australia) of expected KS cases were diagnosed based on clinical presentation. For XYY, only between 3% (in UK) and 18% (in Denmark) of expected cases were diagnosed.[76]

Our prevalence estimates in an adult study population are somewhat lower than those reported by those new-born infants (KS: 152/100,000; and 47,XYY: 98/100,000 males).[76] While this could be interpreted as indicating higher mortality rates, it is recognised that UK Biobank comprises a more educated and healthier sample compared to the general population, likely due to 'healthy volunteer' bias.[90] Similarly, the prevalence of other adverse genetic conditions is reportedly lower in UK Biobank than in other more representative studies.[91]

Previous studies have highlighted higher disease risks in men with KS. Bojesen et al.[92] identified 832 men with KS from hospital records in Denmark and reported higher risks for venous thrombosis (hazard ratio=5.3, 95% CI 3.3–8.5), pulmonary embolism (3.6, 1.9–6.7), chronic obstructive pulmonary disease (3.9, 2.5–6.1), Type 2 diabetes (3.7, 2.1–6.4), and atherosclerosis (4.5, 2.8–7.1). Swerdlow et al.[80] accessed data on 3,518 UK patients with KS diagnosed since 1959 and followed to mid-2003 and reported higher mortality from Diabetes mellitus (standardized mortality ratio: 5.8, 95% CI, 3.4–9.3), pulmonary embolism (5.7, 2.5–11.3) and chronic lower respiratory disease (2.1, 1.4–3.0). Zöller et al.[79] reported a higher risk for venous thromboembolism (incidence rate ratio, IRR=6.4, 95% CI 5.1-7.9) in

1,085 men diagnosed with KS between 1969 and 2010 in Sweden. Our findings confirm these strong disease associations and also the reported higher risks of psychiatric illness and osteoporosis.

By contrast, few studies have reported on disease risks in men with 47,XYY. Berglund et al. identified 255 men with 47,XYY from hospital records in Denmark and reported higher risks for venous thrombosis (IRR=10.2, 95% CI 4.6–22.6), and chronic obstructive pulmonary disease (IRR=5.8, 2.4–15.1).[93] Our findings confirm those observations and extend the list of diseases strongly associated with 47,XYY to also include Type 2 diabetes, pulmonary embolism and atherosclerosis.

We observed some notable differences between KS and 47,XYY. KS is a well-recognised cause of reproductive dysfunction, and this was reflected in our data by their later age at puberty, lower testosterone levels and high risk of being childless. Reproductive dysfunction likely also contributes to their tall stature (due to later pubertal growth completion), lower bone density and muscle strength and greater adiposity. By contrast, men with 47,XYY appeared to have normal reproductive function, with no alteration in their puberty timing or testosterone levels, and a more modestly higher risk of being childless which could be explained by their similarly higher chance of living without a partner.

Hence, despite these marked differences in reproductive function, it is unclear why both KS and 47,XYY should show striking similarities in conferring substantially higher risks for many diseases in common – Type 2 diabetes, atherosclerosis, venous thrombosis, pulmonary embolism, and COPD, which persisted after adjustments for several lifestyle behavioural-related traits (BMI, smoking, deprivation). Higher risks of Type 2 diabetes, atherosclerosis and microalbuminuria, with lower HDL cholesterol and higher adiposity, together indicate higher insulin resistance in both KS and 47,XYY men. Both conditions confer a triple dose of the pseudo-autosomal region, which contains the growth-related *SHOX* gene, which likely partially contributes to their tall stature, and there are case reports of insulin resistance in other conditions characterised by *SHOX* excess.[94,95] However, the underlying mechanisms are yet unknown.

Similarly, it is unclear why risks for venous thrombosis and pulmonary embolism are raised in both KS and 47,XYY to a similar substantial degree, around 6-7 fold higher risk for venous thrombosis. This is a similar or even higher than that conferred by Factor V Leiden, a genetic variant carried by around 5% of white and of European descent.[96] Hence, it might be considered to add sex chromosome aneuploidy to the screening for genetic causes of thrombophilia. Furthermore, as KS and 47,XYY confer higher risks for multiple potentially preventable diseases, future studies should explore the potential benefits of wider testing.

Strengths of our study include the use of systematic case ascertainment and 100% confirmation of GWAS array-based categorisation using exome sequencing data in a large sub-sample. Furthermore, our numbers of individuals with a male sex chromosome aneuploidy are similar to those reported by UK Biobank, which found 355 such individuals (99.2% of our 358 cases).[50] GWAS array genotyping and exome sequencing are increasingly performed in clinical settings, however male sex chromosome aneuploidy status is not routinely derived. In addition, the wide range of traits and diseases available in UK Biobank allowed us to systematically quantify the disease and phenotypic impacts of male sex chromosome aneuploidy.

Limitations include the 'healthy volunteer' bias of the UK Biobank sample and the yet incomplete linkage to primary care health data. Hence it is likely that the true disease risks associated with KS and 47,XYY analyses are even higher than the substantial estimates that we observed.

In conclusion, our findings show that male sex chromosome aneuploidy can be reliably detected using GWAS or exome sequencing data. Klinefelter syndrome and 47,XYY were mostly unrecognised but conferred substantially higher risks of diverse potentially preventable diseases, including metabolic and vascular diseases, which were only partially explained by higher levels of BMI, deprivation, and smoking. Future studies should consider the utility of deriving male sex chromosome aneuploidy status when genetic testing is undertaken for existing clinical indications, e.g., for thrombosis risk. Furthermore, our findings add significantly to ongoing debates regarding the potential benefits of wider population genetic screening.

Chapter 5 *GIGYF1* loss of function is associated with clonal mosaicism and adverse metabolic health

Published in:

Nature Communications

Yajie Zhao, Stasa Stankovic, Mine Koprulu, Eleanor Wheeler, Felix R Day, Hana Lango Allen, Nicola D Kerrison, Maik Pietzner, Po-Ru Loh, Nicholas J Wareham, Claudia Langenberg, Ken K Ong, and John R.B. Perry

July 7, 2021

doi: 10.1038/s41467-021-24504-y

5.1 Contributions

This chapter describes the work I performed on the first exome-wide gene burden testing for LOY. I processed the WES data for over 200,000 participants from UK Biobank by implementing several QC procedures under the guidance of Dr Hana Lango Allen. I set up and tested the analysis pipeline of STAAR and conducted the leave-one-out analysis. Prof John R.B. Perry set up the analysis pipeline of BOLT-LMM and conducted the sensitivity analyses. Stasa Stankovic and Mine Koprulu annotated the variants using the VEP. Dr Eleanor Wheeler performed the multi-tissue eQTL associations in GTEx. Dr Felix R Day prepared the phenotypes and checked principal components and geographical distribution of *GIGYF1* loss of function carriers. Nicola D Kerrison checked the details of the *GIGYF1* loss of function carriers with T2D. Prof Po-Ru Loh proposed the formula of PAR-LOYq. Dr Maik Pietzner, Prof Nicholas J Wareham, Prof Claudia Langenberg and Prof Ken K Ong provided valuable advice on the analyses and writing.

5.2 Abstract

Mosaic loss of chromosome Y (LOY) in leukocytes is the most common form of clonal mosaicism, caused by dysregulation in cell-cycle and DNA damage response pathways. Previous genetic studies have focussed on identifying common variants associated with LOY, which we now extend to rarer, protein-coding, variation using exome sequences from 82,277 male UK Biobank participants. Loss of function of two genes – *CHEK2* and *GIGYF1* - reached exome-wide significance. Rare alleles in *GIGYF1* have not previously been implicated in any complex trait, but here loss-of-function carriers exhibit six-fold increased susceptibility to LOY (OR=5.99 [3.04-11.81], $P=1.3\times10^{-10}$). These same alleles were also associated with adverse metabolic health, including higher susceptibility to Type 2 Diabetes (OR=6.10 [3.51-10.61], $P=1.8\times10^{-12}$), 4kg higher fat mass ($P=1.3\times10^{-4}$), 2.32 nmol/L lower serum IGF1 levels ($P=1.5\times10^{-4}$) and 4.5kg reduced handgrip strength ($P=4.7\times10^{-7}$). These associations were mirrored by a common variant nearby associated with expression of *GIGYF1*. Our observations are consistent with *GIGYF1* enhancing insulin and IGF-1 receptor signaling, highlighting a potential direct connection between clonal mosaicism and metabolic health.

5.3 Introduction

Mosaic loss of the Y chromosome in leukocytes (LOY) is the most common form of clonal mosaicism, first noted over fifty years ago[97,98]. It has been associated with risk of a number of complex diseases and traits, however the biological mechanisms underpinning these observations are unclear. Like other forms of clonal mosaicism, LOY is strongly associated with age, reflecting greater opportunity for mitotic errors in haemopoietic stem cell division and subsequent clonal expansion to occur. Predisposition to LOY also has a heritable component and to date, over 150 associated common genetic variants have been identified[16,30,32,33]. These loci have implicated genes involved in cell-cycle fidelity and DNA damage response (DDR), supporting the idea that LOY is a readily detectable manifestation of subtle defects in these processes[16,32]. We have hypothesized that predisposition to genomic instability that is shared across multiple cell types, including leukocytes, may explain the observational associations between LOY and other health outcomes[16]. This concept is most apparent for *CHEK2* loss of function, which both promotes LOY in men and extends reproductive life in women through the shared mechanism of inhibiting DNA damage sensing and apoptosis. Identifying novel genetic determinants associated with LOY has the potential, therefore, to not only increase our knowledge of clonal haematopoiesis, but also to identify loci that underlie susceptibility to other complex traits through shared biological mechanisms. We previously demonstrated this with Type 2 Diabetes (T2D), where overlap with LOY highlights loci which likely impact cellular homeostasis in metabolic tissues. For example, alleles in *CCND2* both increase the risk of T2D and LOY[16], with this gene encoding the major D-type cyclin that is expressed in pancreatic β-cells and is essential for adult β cell growth[99].

To date, genetic studies for LOY have focussed on genotype-array imputed common genetic variation, which largely misses the contributions of rarer, often more deleterious, alleles[16,30,32]. To address this, we perform the first exome-sequence GWAS for LOY, assessing the role of rare protein-coding variation. We extend and confirm previous observations supporting the role of *CHEK2* and identify a novel association with *GIGYF1* loss of function, highlighting an intriguing link between LOY and metabolic health.

5.4 Methods

5.4.1 Phenotype definitions

Until now, there were two established LOY estimation methods based on SNP-array data: (1) the median or mean of log R ratio (mLRR-Y) genotyping intensity values of the probes on the male-specific regions of chromosome Y (MSY); and (2) the phase-based computational method that estimates allelic imbalance using only the pseudo autosomal regions (PAR-LOY) detailed by previously[16]. The mLRR-Y and PAR-LOY are independent approaches as they are estimated from non-overlapping regions of the Y chromosome. Although there is considerable correlation in the LOY estimates produced by these two methods, we sought to combine the independent information considered by the two approaches to gain increased power for genetic association analyses. We combined PAR-LOY and mLRR-Y with an additional measure, the estimated fraction of cells with LOY (AF-LOY) which was estimated when generating PAR-LOY[16]. Our new combined call of LOY (PAR-LOYq) is defined as PAR-LOY+ 3×AF-LOY– 3×mLRR-Y (cropped to the range [0,2]). The intuition behind this formula is to augment the binary PAR-LOY variable by up-weighting individuals who have a larger LOY cell fraction (as estimated by AF-LOY and mLRR-Y), which may be more strongly associated with risk alleles.

We compared the performance of the three LOY estimates derived from the genotyping array data using the full set of male UKBB participants[16]. We performed association testing with age and ever smoking status, which are two established risk factors for LOY[27,28], and the 156 previously reported LOY-associated loci[16]. For both age and smoking status, PAR-LOYq outperformed than two established LOY estimation methods using the same sample; the t-test statistic of PAR-LOYq for age increased 65.4% and 5.2% respectively and the t-test statistic of PAR-LOYq for ever smoking status increased 44.9% and 11.1% respectively. Improvement of PAR-LOYq over PAR-LOY was also evaluated for the 156 previously identified variants by assessing the median improvement in chisq statistic.

Participants were classified as cases of type 2 diabetes (T2D) according to the previously published UKBB probable T2D algorithm[100] based on baseline self-reported diabetes or medications, in addition to evidence from electronic health records (Hospital Episode Statistics or Death Registration) consistent with T2D (International Statistical Classification

of Diseases and Related Health Problems Tenth Revision code E11). Any possible or probable type 1 diabetes cases were excluded. Controls were participants without evidence of T2D as defined above. The GWAS on random glucose and HbA1C – using BOLT-LMM pipeline described below – was performed after excluding individuals with our defined T2D criteria. The T2D test statistic for the common variant was taken from the DIAMANTE consortium GWAS meta-analysis[101]. All other phenotypes used in this study were available from UK Biobank and any applied transformations described in the relevant results tables.

5.4.2 UK Biobank exome-sequence data processing and QC

We downloaded VCF and PLINK format files for whole exome sequencing (WES) data of 200,643 UK Biobank participants, which were made available in October 2020. The overview of this 200K WES release is described at https://biobank.ndph.ox.ac.uk/ukb/label.cgi?id=170. Details of sequence data processing (read alignment, variant calling etc.) are described in papers of Szustakowski et al.[87] and Yun et al.[88]

I merged individual VCF files into a single VCF file of each chromosome using BCFtools v1.9[102]. I converted each chromosome file losslessly to a GDS (Genomic Data Structure) format file (an RData object) using the seqVCF2GDS () function from the R package SeqArray v1.30.0[103]. I used SeqArray package and GDS data object to extract the dosage matrix and perform additional variant and genotype level filtering below. Such genotype data processing is faster than using a flat text VCF file because GDS is implemented using an optimized C++ library and a high-level R interface is provided by the platform-independent R package gdsfmt[103,104].

I used SeqArray package to calculate and extract the QC metrics. Firstly, I identified and flagged 7,913,671 on-target variants (those defined by the xgen_plus_spikein.GRCh38.bed file genomic coordinates) among the total of 15,916,398 called variants on autosomes and chromosome X. The UKBB released VCF file has a number of QC metrics which can be used for variant site and individual genotype filtering: QUAL (variant site-level quality score, Phred scale); AQ (variant site-level allele quality score reflecting evidence for each alternate allele, Phred scale); DP (individual genotype call-level approximate read depth (reads with MQ=255 or with bad mates had already been filtered out)); AD (individual genotype call-

level allelic depths for the ref and alt alleles in the order listed); GQ (individual genotype call-level Genotype Quality, Phred scale). I additionally calculated the site-level genotype missingness (the number of samples at each site without genotype call).

After generating the summary statistics of QUAL and AQ metrics, I noted that the released UKBB 200K WES data already had some QC filters applied. The values of QUAL and AQ ranged from 20 (error rate=1%) to 99 (error rate<0.0001%) with mean 44.5 (error rate<0.01%). For all chromosomes the distributions of the values of QUAL and AQ are nearly the same. I decided not to apply additional stricter filters on these two site-level metrics. I calculated summary statistics (minimum, maximum, mean, and 1st, 2nd and 3rd quartile) for DP and GQ for each variant based on all 200,643 samples for autosomes and 110,438 female samples for the X chromosome. I recorded the number of samples with GQ<20 at each variant. I calculated allelic balance for each heterozygous genotype calls at on-target bi-allelic sites (ABratio), defined as the number of alternate allele's reads (provided in the AD field) divided by the total depth which equals to the sum of read depths of reference allele and alternative allele. I then generated the same per-site summary statistics as above for ABratio. I defined and excluded a heterozygous genotype call as imbalanced if ABratio≤0.25 or ABratio≥0.8.

In our sensitivity analysis, I applied three variant-level filters to exclude variants at potentially poorly performing sites: filter 1:>5% missingness (samples without genotype calls); filter 2: the maximum of the read depth of genotype calls (DP) across samples<10; and filter 3:>20% genotype calls with GQ<20. After applying these three filters, 1,161,679 (7.3%) of the initial 15,916,398 variants, and 96,640 (1.2%) of the 7,913,671 on-target variants were excluded. For the variants included in our variant-set analysis, I also generated the same QC metrics restricted only to rare allele carriers. Ultimately all of these metrics were used to filter out variants in sensitivity analyses that were initially performed using the default QC parameters applied to the UKBB released dataset.

5.4.3 Variant annotation and definition of gene burden sets

We annotated variants released in UK Biobank (UKBB) 200K whole exome sequencing (WES) VCF files using the Ensembl Variant Effect Predictor tool release 99 based on build hg38[105]. For each uploaded variant, the default VEP features include consequence and impact of the

variant, overlapping gene, position at cDNA and protein level and amino acid change, if applicable. In addition to the default features, the following plugins from VEP were used: (i) SIFT[106], which predicts whether an amino acid substitution affects protein function based on sequence homology and the physical properties of the amino acid, (ii) Polyphen-2[49], which predicts possible impact of an amino acid substitution on the structure and function of a protein, (iii) CADD[46] which provides deleteriousness prediction scores for all variants based on diverse genomic features and (iv) LOFTEE[107] which provides loss of function prediction for variants. The variants were annotated for every available overlapping transcript in Ensembl. We used the most severe variant definition for each variant-gene pair, which provides the annotation of the variant for the transcript it has the most severe consequence on.

We defined loss of function variants as those with 'high impact' prediction by VEP. This includes frameshift variants, transcript ablating or transcript amplifying variants, splice acceptor or splice donor variants, stop lost, start gained or stop gained variants. 'Moderate impact' variants include missense variants, inframe deletion or insertions, missense variants and protein altering variants.

5.4.4 Gene association testing

Gene burden scores were created by collapsing all annotated rare alleles together to define a binary call denoting whether an individual carries none versus one or more rare alleles at a given gene. Reported effect estimates therefore represent the trait difference between carriers and non-carriers of these alleles. These dummy variables were then transformed into BGEN file format genotype call files for association testing using a linear mixed model implemented in BOLT-LMM[63] to account for cryptic population structure and relatedness. Only autosomal genetic variants that were common (minor allele frequency (MAF)>1%), passed quality control in all 106 batches and were present on both genotyping arrays were included in the genetic relationship matrix (GRM). Genotyping chip, age at baseline and ten genetically derived principal components were included as covariates. Samples were excluded from analysis if they failed UK Biobank quality control parameters, were of non-European ancestry or if the participant withdrew consent from the study.

5.4.5 Secondary association testing

I applied STAAR (variant-Set Test for Association using Annotation infoRmation)[45] as a secondary analytical approach for associated genes. STAAR is a general framework for performing a rare variants association study at scale, suitable for whole exome or genome population-level datasets such as UKBB. STAAR accounts for population structure and relatedness, by fitting linear and logistic mixed models for quantitative and dichotomous traits. It takes as input individual data frames for genotypes, phenotypes, covariates including age, age^2, sex, chip, PC1-PC10 were generated from the SNP array data and (sparse) GRM.

I used the basic function of STAAR (with CADD-score weighting additionally performed in a sensitivity analysis) and set the thresholds of MAF≤0.5% and ≥2 rare variants count in a gene. The output of STAAR provides p-values for a number of different rare variant set burden tests including SKAT (sequence kernel association test), Burden test and ACAT-V (set-based aggregated Cauchy association test). Additionally, STAAR provides an omnibus test result by using the combined Cauchy association test to aggregate the association across the different tests.

To ensure that the individual gene-level result is not disproportionally affected by a single variant of considerably larger effect and that the others are part of the same variant set, I performed a drop-one-out analysis using STAAR for our target gene.

Effect estimates for dichotomous traits were estimated by using logistic regression performed in R (3.3.3). Where these are reported they include the P-value obtained from the linear mixed-model generated by BOLT-LMM.

5.5 Results

Previous studies have quantified LOY using either a quantitative measure derived from the mean log2-transformed R ratio of signal intensity (mLRR-Y)[30], or more recently a more-powered dichotomous measure (PAR-LOY) using allele-specific genotyping intensities in the sex chromosome pseudo-autosomal region (PAR)[16]. We note that both measures are proxies for the abundance of Y chromosome genetic material in the measured biological samples, derived from intensity data which contains a lot of experimental 'noise'. As these measures are independent – one relies on PAR genotypes only whilst the other excludes them – we hypothesised that an aggregate of the two would further help improve the signal to noise ratio of these measures and therefore increase statistical power to detect genetic associations. We name this combined quantitative measure PAR-LOYq and estimated it in the same UK Biobank participants who were previously studied for PAR-LOY (N=205,011 men). As expected, PAR-LOYq calls provided a more powerful measure for discovery analysis, with a median 10.6% increase in chi-square association statistic for the 156 LOY loci previously identified by PAR-LOY **(Supplementary Table 5-1 (Appendix C))**[16].

To identify genes associated with LOY, we performed gene burden analyses for PAR-LOYq in 82,277 male UK Biobank participants with exome sequence data. Two models were tested exome-wide, by collapsing together rare (MAF<0.5%) loss of function or moderate-impact variants in each individual gene. The association of burden test in two genes, *CHEK2* and *GIGYF1*, were statistically significant exome-wide (P<1.6x10^{-6}) across these analyses **(Figure 5-1)**. Loss of function variants in *CHEK2* (N=543 carriers, effect=0.23 SD increase in PAR-LOYq between rare allele carriers vs non-carriers, P=3.4x10^{-9}) have previously been implicated with LOY as the most common frameshift variant (1100delC, MAF~0.2%) is well captured by GWAS imputation and directly genotyped on the UKBB array. This single variant accounted for 76% of loss of function carriers and the *CHEK2* association was nominally significant when it was excluded (P=0.02, effect=0.18 SD). An independent burden test of rare moderate-impact alleles in *CHEK2* (not including 1100delC and other loss of function alleles) was also associated with PAR-LOYq (**Supplementary Table 5-2 (Appendix C),** N=1057 carriers, effect=0.11 SDs, P=1.7×10^{-4}).

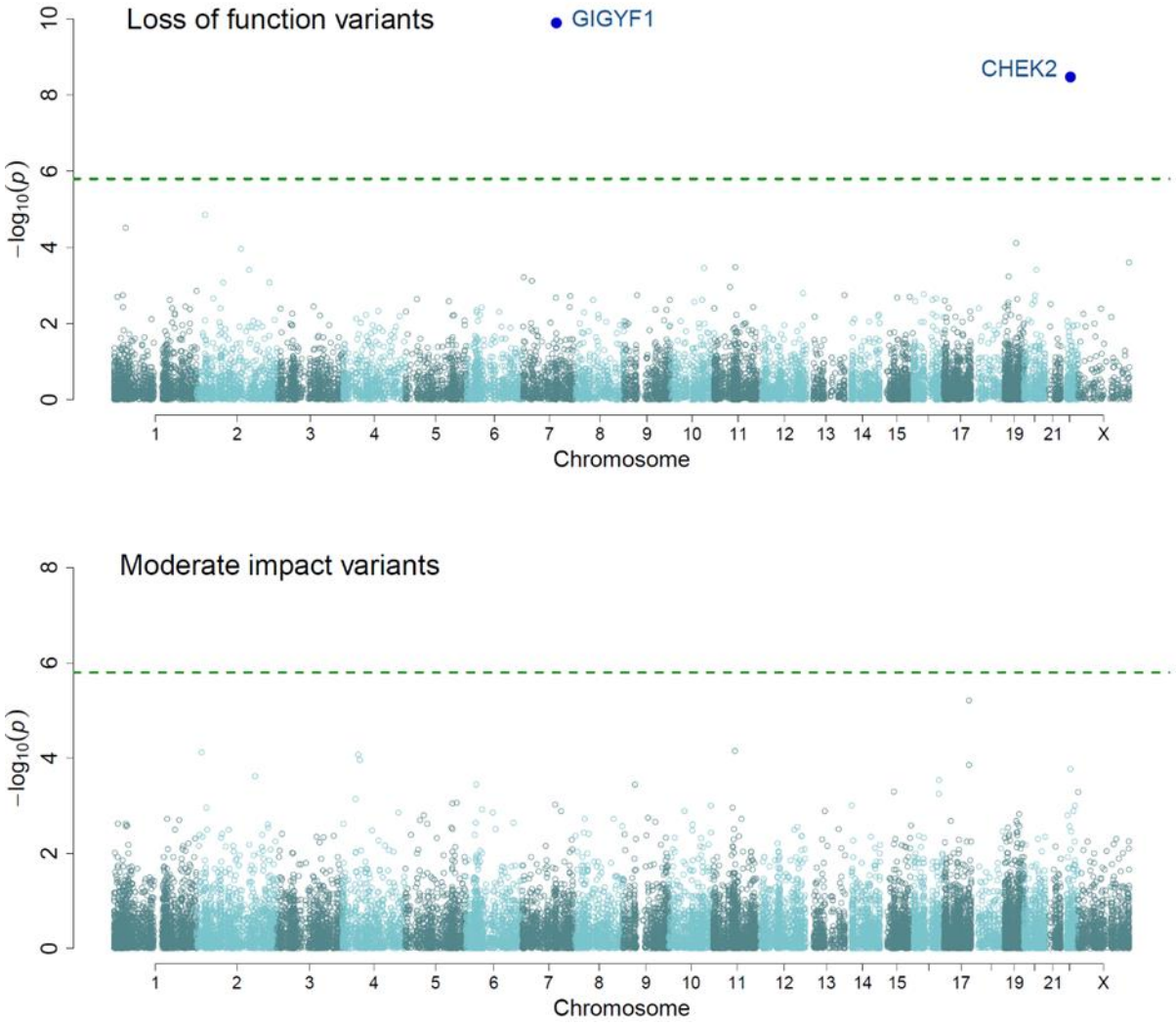


Figure 5-1 Manhattan plot for exome-wide gene burden test statistics. Dashed line denotes the multiple test adjusted P-value threshold (P<1.6×10^{-6}).

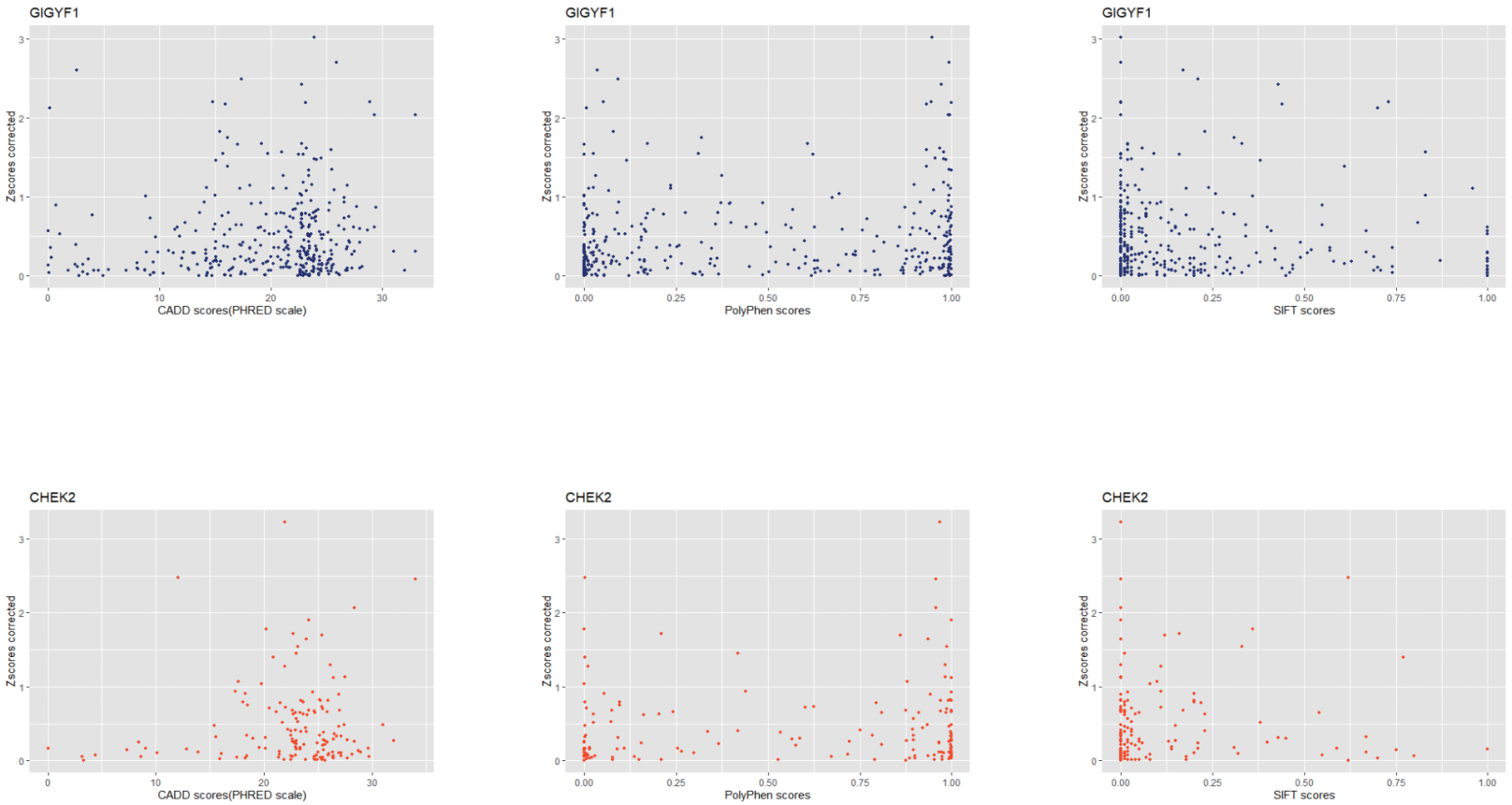


Figure 5-2 Relationship between bioinformatically predicted function and LOY association for *GIGYF1* and *CHEK2* moderate-impact variants. Y-axis shows the PAR-LOYq association of each variant assessed by absolute Z-score divided by minor allele count.

GIGYF1 loss of function (N=40 carriers) was associated with a 0.93 (0.64-1.21, P=1.3×10^{-10}) standard deviation higher PAR-LOYq. This burden signal combined the effects of 27 rare variants: a single base insertion frameshift with 10 carriers, 4 doubletons and 22 singleton rare alleles. No individual variant was more significant than the overall *GIGYF1* test result, which remained significant in a leave-one-out analysis of each variant **(Table 5-1)**. Rare moderate-impact alleles were not associated with LOY in aggregate (P=0.70), however several individual moderate-impact variants exhibited nominally significant associations **(Supplementary Table 5-3 (Appendix C))**. We note that missense alleles will likely represent a heterogeneous collection of loss of function, gain of function and benign effects. As with *CHEK2*, bioinformatic filters were a poor predictor of which missense variants in *GIGYF1* were associated with LOY **(Figure 5-2)**. Furthermore, running genome-wide burden analyses in STAAR, weighting each variant by its CADD score did not identify additional genes **(Supplementary Table 5-4 (Appendix C))**.

Table 5-1 Leave-one-out gene burden association analyses for *GIGYF1*

		STAAR Association Test P-value						
Variant	**MAC**	**STAAR_O**	**SKAT_1_25**	**SKAT_1_1**	**Burden_1_25**	**Burden_1_1**	**ACAT-V_1_25**	**ACAT-V_1_1**
ALL	40	1.73E-10	2.05E-05	2.07E-05	1.15E-10	1.15E-10	1.15E-10	1.15E-10
7:100687545:D:1	10	1.10E-07	7.56E-06	7.58E-06	7.40E-08	7.40E-08	7.40E-08	7.40E-08
7:100687546:I:1	2	1.02E-09	2.70E-05	2.71E-05	6.82E-10	6.82E-10	6.82E-10	6.82E-10
7:100687532:G:A	2	2.40E-11	1.88E-05	1.89E-05	1.60E-11	1.60E-11	1.60E-11	1.60E-11
7:100683017:G:A	2	2.47E-11	1.88E-05	1.89E-05	1.65E-11	1.65E-11	1.65E-11	1.65E-11
7:100687357:G:A	2	3.03E-11	1.85E-05	1.87E-05	2.02E-11	2.02E-11	2.02E-11	2.02E-11
7:100684236:C:T	1	6.19E-11	1.96E-05	1.97E-05	4.13E-11	4.13E-11	4.13E-11	4.13E-11
7:100688238:T:C	1	1.31E-10	1.95E-05	1.96E-05	8.75E-11	8.75E-11	8.75E-11	8.75E-11
7:100686356:G:A	1	8.04E-11	1.94E-05	1.96E-05	5.36E-11	5.36E-11	5.36E-11	5.36E-11
7:100683366:G:A	1	9.52E-11	1.94E-05	1.95E-05	6.34E-11	6.35E-11	6.34E-11	6.35E-11
7:100686365:G:A	1	5.28E-11	1.98E-05	1.99E-05	3.52E-11	3.52E-11	3.52E-11	3.52E-11
7:100685129:G:A	1	9.52E-10	2.48E-05	2.50E-05	6.35E-10	6.34E-10	6.35E-10	6.34E-10
7:100687297:C:T	1	6.44E-10	2.29E-05	2.31E-05	4.29E-10	4.29E-10	4.29E-10	4.29E-10
7:100682749:I:1	1	2.11E-10	1.99E-05	2.01E-05	1.40E-10	1.40E-10	1.40E-10	1.40E-10
7:100687045:I:1	1	1.95E-09	2.99E-05	3.01E-05	1.30E-09	1.30E-09	1.30E-09	1.30E-09
7:100683122:D:2	1	5.80E-11	1.97E-05	1.98E-05	3.87E-11	3.87E-11	3.87E-11	3.87E-11
7:100683112:I:20	1	9.49E-11	1.94E-05	1.95E-05	6.33E-11	6.33E-11	6.33E-11	6.33E-11
7:100682484:T:C	1	1.14E-10	1.94E-05	1.96E-05	7.60E-11	7.60E-11	7.60E-11	7.60E-11
7:100683231:C:T	1	1.21E-10	1.94E-05	1.96E-05	8.05E-11	8.05E-11	8.05E-11	8.05E-11
7:100686817:G:A	1	4.60E-10	2.17E-05	2.18E-05	3.07E-10	3.07E-10	3.07E-10	3.07E-10
7:100682700:D:2	1	6.64E-10	2.30E-05	2.32E-05	4.43E-10	4.43E-10	4.43E-10	4.43E-10
7:100682387:D:1	1	6.75E-10	2.31E-05	2.33E-05	4.50E-10	4.50E-10	4.50E-10	4.50E-10
7:100683374:D:5	1	1.17E-09	2.60E-05	2.62E-05	7.82E-10	7.81E-10	7.82E-10	7.81E-10
7:100687408:T:C	1	1.27E-09	2.66E-05	2.67E-05	8.46E-10	8.45E-10	8.46E-10	8.45E-10
7:100686749:C:T	1	1.79E-09	2.92E-05	2.94E-05	1.19E-09	1.19E-09	1.19E-09	1.19E-09
7:100682120:D:1	1	2.06E-09	3.05E-05	3.07E-05	1.38E-09	1.37E-09	1.38E-09	1.37E-09
7:100687323:G:A	1	3.27E-09	3.56E-05	3.58E-05	2.18E-09	2.18E-09	2.18E-09	2.18E-09
7:100682198:I:2	1	3.97E-09	3.83E-05	3.85E-05	2.65E-09	2.65E-09	2.65E-09	2.65E-09
7:100682071:C:T	0	1.73E-10	2.05E-05	2.07E-05	1.15E-10	1.15E-10	1.15E-10	1.15E-10
7:100688225:D:2	0	1.73E-10	2.05E-05	2.07E-05	1.15E-10	1.15E-10	1.15E-10	1.15E-10
7:100685054:G:A	0	1.73E-10	2.05E-05	2.07E-05	1.15E-10	1.15E-10	1.15E-10	1.15E-10
7:100686182:D:2	0	1.73E-10	2.05E-05	2.07E-05	1.15E-10	1.15E-10	1.15E-10	1.15E-10
7:100681994:C:T	0	1.73E-10	2.05E-05	2.07E-05	1.15E-10	1.15E-10	1.15E-10	1.15E-10
7:100683218:C:A	0	1.73E-10	2.05E-05	2.07E-05	1.15E-10	1.15E-10	1.15E-10	1.15E-10
7:100683303:C:T	0	1.73E-10	2.05E-05	2.07E-05	1.15E-10	1.15E-10	1.15E-10	1.15E-10
7:100683548:A:C	0	1.73E-10	2.05E-05	2.07E-05	1.15E-10	1.15E-10	1.15E-10	1.15E-10
7:100683585:D:2	0	1.73E-10	2.05E-05	2.07E-05	1.15E-10	1.15E-10	1.15E-10	1.15E-10
7:100684338:C:G	0	1.73E-10	2.05E-05	2.07E-05	1.15E-10	1.15E-10	1.15E-10	1.15E-10
7:100684339:T:G	0	1.73E-10	2.05E-05	2.07E-05	1.15E-10	1.15E-10	1.15E-10	1.15E-10

We next performed several sensitivity analyses to further explore the genetic architecture of this *GIGYF1*-LOY association. Firstly, we observed consistent effects using the two previous LOY traits, with a 6-fold (OR=5.99 [3.04-11.81], $P=6\times10^{-7}$) higher risk of a PAR-LOY dichotomous call and a -0.038 (~0.81 SD, $P=8.8\times10^{-9}$) reduction in mLRR-Y. Secondly, in a sensitivity analysis, PAR-LOYq association results were highly consistent when excluding multi-allelic sites ($P=8.4\times10^{-9}$) or indels ($P=9.9\times10^{-3}$) and when restricting to high-confidence loss of function variants defined by LOFTEE ($P=4.1\times10^{-13}$)[107]. Sequencing quality control parameters for each individual variant appeared robust. Thirdly, we reproduced the same association signal using a second analytical pipeline implemented in STAAR ($P=1.73\times10^{-10}$)[45]. Finally, we showed that *GIGYF1* loss of function was not associated with any genetically derived principal component and carriers were geographically dispersed across the UK **(Figure 5-3, 5-4)**.

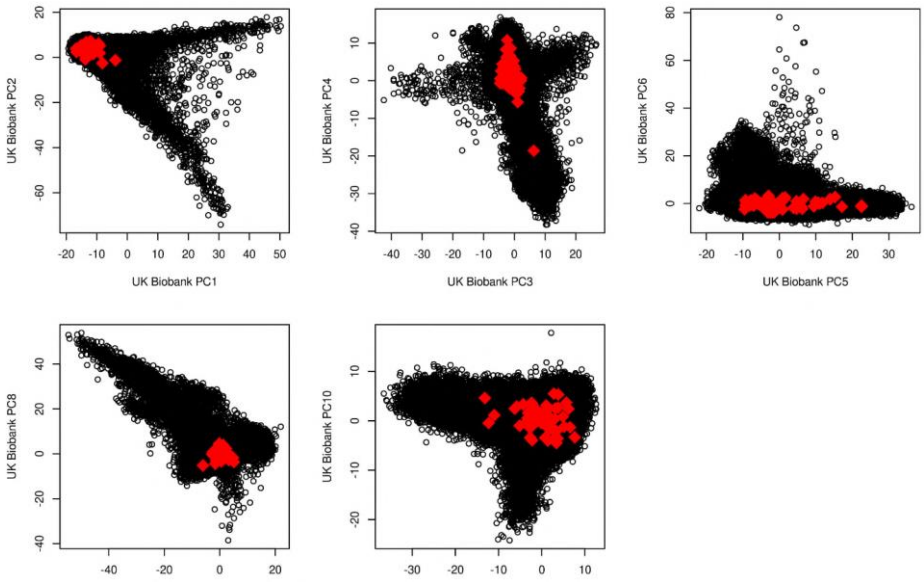


Figure 5-3 Impact of *GIGYF1* loss of function carriers on genetically-defined principal components. *GIGYF1* carriers are highlighted in red, all other analysed samples in black. Analysis performed in maximum available sample-size (N=184,972).

Figure 5-4 Geographical distribution of *GIGYF1* loss of function carriers by location of birth. GIGYF1 carriers are highlighted in red, all other analysed samples in black. Analysis performed in maximum available sample-size (N=184,972).

GIGYF1 is named after its known binding to growth factor receptor-bound protein 10 (GRB10) and interacts with both the insulin and *IGF1* receptors[108]. We therefore postulated that loss of function alleles may also impact on metabolic health, and, therefore, repeated the *GIGYF1* association analyses across 17 metabolic-health related traits in men and women **(Table 5-2)**.*GIGYF1* loss of function (N=64 carriers) was associated with higher susceptibility to type 2 diabetes (OR=6.10 [3.51-10.61], P=1.8×10^{-12}) and higher acute and longer-term average levels of glycaemia in non-diabetic individuals (random glucose P=2.6x10^{-5} and HbA1c P=6.6×10^{-7}). Of the 64 carriers, 19 (30%) had T2D, compared to 7.1% in the population of UK Biobank in whom sequence data is available. Carrier status was also associated with a 1.85 kg/m^2 higher body mass index (P=5.3×10^{-4}), 4 kg higher fat mass (P=1.3×10^{-4}), 1.85 kg higher lean mass (P=5.2×10^{-3}), 0.04 higher waist-to-hip ratio (P=1.8x10^{-6}), -0.01 lower sitting to standing height ratio (P=4.3×10^{-7}), 4.5 kg lower grip strength (P=4.7×10^{-7}) and 2.32 nmol/L lower serum IGF1 levels (P=1.5×10^{-4}). The T2D association was largely unattenuated by adjustment for BMI (OR 5.07 [2.78-9.27] P=8.9×10^{-11}) and the clinical characteristics of the rare allele carriers with T2D did not provide any evidence of phenotype distinct from typical T2D **(Supplementary Table 5-5 (Appendix C))**. Notably *GIGYF1* loss of function was not associated with birthweight, puberty timing, childhood body size or adult height (P>0.05).

Table 5-2 The association of *GIGYF1* loss of function on metabolic traits

Trait	Units	N	MAC	BETA	SE	P_BOLT_LMM
T2D	binary (estimate on linear scale)	184972	64 (19 cases)	0.22	0.03	1.80E-12
T2D (adj BMI)	binary (estimate on linear scale)	184344	64 (19 cases)	0.19	0.03	8.90E-11
Sitting / standing height ratio	ratio	184111	64	-0.01	0.00	4.30E-07
Handgrip strength	Kilograms	184493	64	-4.50	0.88	4.70E-07
HbA1C (exlcuding T2D)	rank normalised	163960	45	0.55	0.12	6.60E-07
Waist to Hip ratio	ratio	184594	64	0.04	0.01	1.80E-06
Glucose (excluding t2d)	rank normalised	150285	42	0.55	0.14	2.60E-05
Total fat mass	grams	178001	60	4051	1104	1.30E-04
IGF1	nmol/L	175034	63	-2.32	0.64	1.50E-04
BMI	kg/m2	184305	64	1.85	0.56	5.30E-04
body fat %	percentage	178149	60	2.33	0.73	8.60E-04
Total lean mass	grams	178147	60	1853	681	5.20E-03
Childhood height (age 10)	categorical	181949	61	0.10	0.08	2.30E-01
Voice breaking	categorical	77054	37	-0.05	0.05	3.30E-01
birth weight	Standard Deviations	90770	28	0.18	0.19	3.80E-01
height	centimetres	184536	64	0.51	0.66	5.80E-01
Childhood body size (age 10)	categorical	181868	61	0.01	0.08	7.70E-01
menarche	years	98880	22	-0.05	0.33	8.10E-01

We next examined whether common genetic variation in GIGYF1 was also associated with LOY and metabolic health parameters. We observed that an intergenic variant (rs221781, MAF=11%, **Table 5-3 and Figure 5-5**) ~25kb from *GIGYF1* was significantly associated with higher glucose (P=4.80×10^{-15}) and HbA1c (P=3.40×10^{-10}). This same allele was associated with increased risk of T2D (OR adj BMI=1.06 (1.04-1.09), P=8.50×10^{-8}) and LOY (P=3.00×10^{-6}), but with lower circulating LDL (P=3.40×10^{-10}) and HDL (P=1.90×10^{-18}) levels. The variant was not associated with BMI (P=0.09). The lead signal for T2D (rs221781) is also the lead conditionally independent eQTL for *GIGYF1* across a number of GTEx tissues including subcutaneous adipose **(Figure 5-6)**, in which we observed that increased expression of *GIGYF1* was associated with lower risk of T2D. The lead eQTL for *GIGYF1* is rs221792 in cultured fibroblasts (P=1.3×10^{-32}) which is in high LD (r^2=0.71, D'=1) with rs221781. The association of common *GIGYF1* variants with T2D was also confirmed in Million Veteran Program data, in which we found a previously reported lead SNP for T2D was in high LD with rs221781 (rs534043, r^2=1, P=8.03×10^{-10}) with a consistent direction of effect[109].

Table 5-3 Common variant associations on metabolic health at the *GIGYF1* locus.

Trait		SNP*	CHR	BP (b37)	Effect Allele	Other Allele	EAF	BETA	SE	P value	OR (95%CI)
T2D (DIAMANTE T2DadjBMI, Neff=157384)	Lead T2D SNP	rs221781	7	100295908	A	G	0.11	-0.06	0.01	8.50E-08	0.94 (0.92,0.96)
T2D (DIAMANTE T2DadjBMI, Neff=157384)	Lead eQTL SNP	rs221792	7	100278657	A	G	0.16	-0.03	0.01	9.50E-04	0.97 (0.95,0.99)
T2D (DIAMANTE T2Dunadjusted, Neff=231,420)	Lead T2D SNP	rs221781	7	100295908	A	G	0.11	-0.05	0.01	3.10E-06	0.95 (0.93,0.97)
T2D (DIAMANTE T2Dunadjusted, Neff=231,420)	Lead eQTL SNP	rs221792	7	100278657	A	G	0.15	-0.01	0.01	1.20E-01	0.99 (0.97,1.00)
PoRu_Males_LOY_combined_Imputed (N=205,011)	Lead T2D SNP	rs221781	7	100295908	A	G	0.11	-0.01	0.00	1.30E-05	-
PoRu_Males_LOY_combined_Imputed (N=205,011)	Lead eQTL SNP	rs221792	7	100278657	A	G	0.15	-0.01	0.00	4.20E-03	-
PoRu_Males_mosaicY_likelyLOY_Imputed (N=205,011)	Lead T2D SNP	rs221781	7	100295908	A	G	0.11	-0.01	0.00	3.00E-06	-
PoRu_Males_mosaicY_likelyLOY_Imputed (N=205,011)	Lead eQTL SNP	rs221792	7	100278657	A	G	0.15	-0.01	0.00	1.70E-03	-
Invn HbA1c (UKBB, N=431,089)	Lead T2D SNP	rs221781	7	100295908	A	G	0.11	-0.02	0.00	**3.40E-10**	-
Invn HbA1c (UKBB, N=431,089)	Lead eQTL SNP	rs221792	7	100278657	A	G	0.15	-0.01	0.00	7.30E-04	-
Invn Glucose (UKBB, N=394,144)	Lead T2D SNP	rs221781	7	100295908	A	G	0.11	-0.03	0.00	**4.80E-15**	-
Invn Glucose (UKBB, N=394,144)	Lead eQTL SNP	rs221792	7	100278657	A	G	0.15	-0.02	0.00	1.70E-11	-
Invn BMI (UKBB, N=450,669)	Lead T2D SNP	rs221781	7	100295908	A	G	0.11	0.01	0.00	9.30E-02	-
Invn BMI (UKBB, N=450,669)	Lead eQTL SNP	rs221792	7	100278657	A	G	0.15	0.01	0.00	4.30E-04	-
BOLT_UKBBGWAS_invn_hdl_nostatin-imputed.txt.gz (N=430,960)	Lead T2D SNP	rs221781	7	100295908	A	G	0.11	0.03	0.00	**1.90E-18**	-
BOLT_UKBBGWAS_invn_hdl_nostatin-imputed.txt.gz (N=430,960)	Lead eQTL SNP	rs221792	7	100278657	A	G	0.15	0.01	0.00	2.30E-06	-
BOLT_UKBBGWAS_invn_ldl_adjLipids-imputed.txt.gz (N=430,160)	Lead T2D SNP	rs221781	7	100295908	A	G	0.11	0.03	0.00	**3.40E-10**	-
BOLT_UKBBGWAS_invn_ldl_adjLipids-imputed.txt.gz (N=430,160)	Lead eQTL SNP	rs221792	7	100278657	A	G	0.15	0.01	0.00	2.00E-04	-
LD between rs221781 and rs221792: r2=0.71, D'=1 (in phase alleles are AA/GG)											

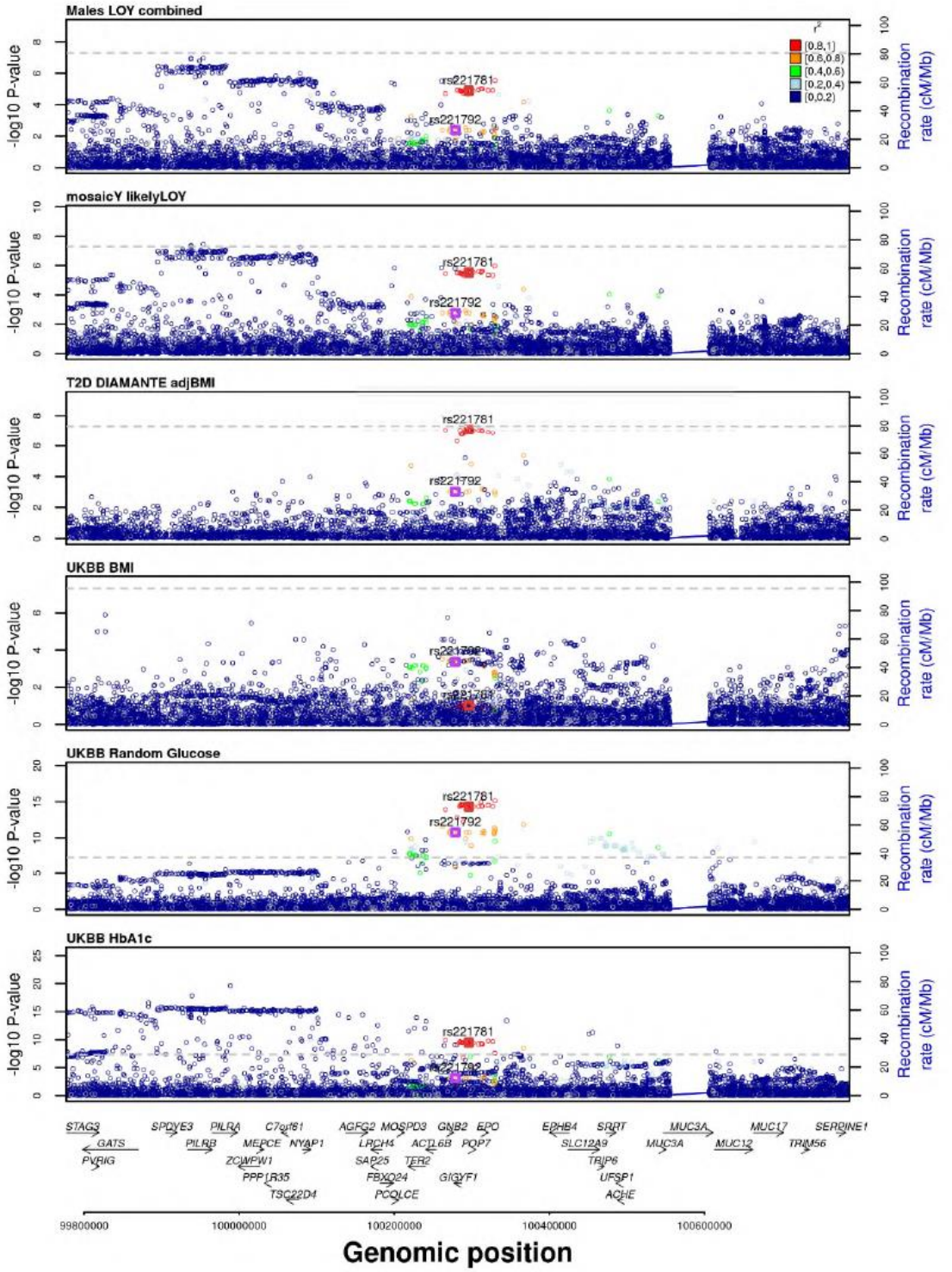


Figure 5-5 Regional association of common variants with Type 2 Diabetes, LOY and related traits in the region around *GIGYF1* (+/- 500kb). Highlighted variants are the lead variant associated with T2D r2221781 (red) and lead eQTL for *GIGYF1* rs221792 (purple).

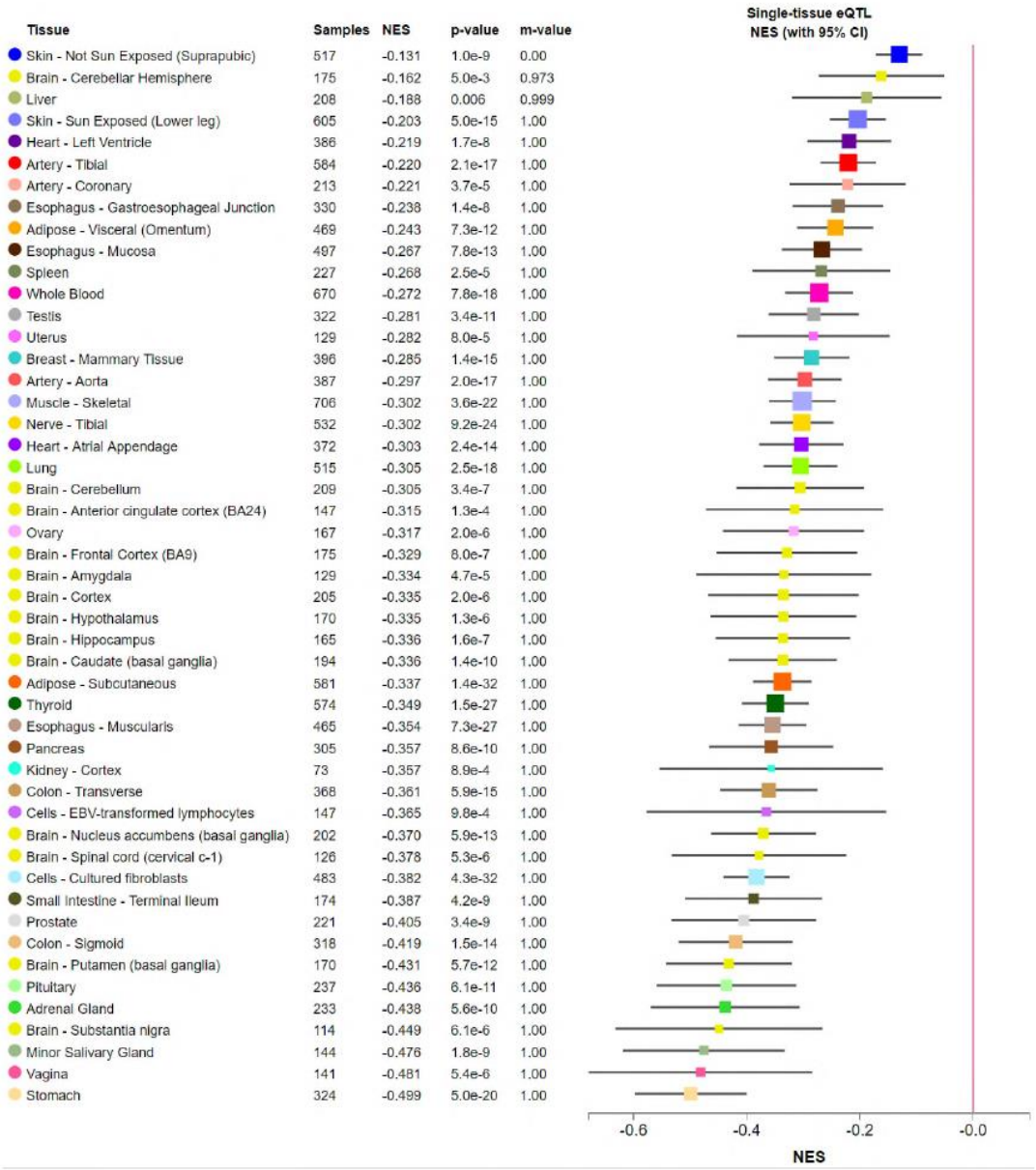


Figure 5-6 Multi-tissue eQTL associations in GTEx for common variant rs221781. The solid pink line represents the null effect. Each square represents the beta estimate from a linear regression model of the variant against mRNA transcript abundance. Test statistic is a two-sided P-value and no correction for multiple testing has been made.

5.6 Discussion

In summary, this exome-wide approach identified rare loss of function alleles in *GIGYF1* exhibiting an effect on LOY ~5 times larger than any genetic variants previously identified by GWAS. Similarly, these alleles confer effect sizes on a number of metabolic outcomes far larger than those previously identified by imputed GWAS and other smaller sequencing studies. For example, rare variants in *PDX1*, *CCND2, SLC30A8* and *PAM* are associated with double the odds of T2D[110–112], whereas *GIGYF1* loss of function is associated with a six-fold increased risk (OR=5.96[3.43-10.38]). The majority of common variants associated with T2D confer much more modest effects (OR<1.5)[101].

GIGYF1 encodes a member of the gyf family of adaptor proteins. It binds growth factor receptor bound 10 (*GRB10*), which is another adaptor protein that binds activated insulin receptors and insulin-like growth factor-1 (IGF-1) receptors to negatively regulate receptor signaling, metabolic responses and IGF1-induced mitogenesis[108,113,114]. Transfection of cells with GRB10-binding fragments of *GIGYF1* leads to greater activation of both the insulin receptor and the IGF-1 receptor[115]. Our findings relating loss of function variants in *GIGYF1* to metabolic and anthropometric outcomes are broadly consistent with the notion that in wild-type carriers *GIGYF1* enhances insulin and IGF-1 receptor signaling, leading to greater handgrip strength (relative to loss of function carriers), sitting height and circulating IGF-1 levels (due to increased insulin signaling), and lower % body fat, WHR, HbA1c, glucose levels and susceptibility to T2D. We previously highlighted the potential role of IGF signalling in promoting chromosomal instability and the cellular accumulation of DNA damage and reported that genetically higher IGF-1 levels are related to greater LOY[116]. It may therefore appear paradoxical that here we find that loss of function in *GIGYF1 (*putatively leading to decreased IGF-1 signalling) should be associated with increased rather than decreased LOY. We hypothesize that *GIGYF1* might enhance DDR mechanisms to protect DNA integrity in the face of IGF-1-mediated tissue growth and differentiation. *GIGYF1* and the related protein *GIGYF2* are implicated in translational repression[117] and translation-coupled mRNA decay[118], which suggests that they may have biological roles beyond insulin and IGF-1 receptor signaling. Although *GIGYF1* is broadly expressed[119], the lack of associations in our data with some established IGF-1-related traits, such as birthweight and adult height, might

reflect tissue or developmental specificity in its effects. We anticipate that future experimental work will shed light on these questions to better understand the links between clonal mosaicism and metabolic health.

Chapter 6 Identification of rare non-synonymous variants affecting sex chromosome mosaicism

6.1 Contributions

This chapter describes the work I conducted to perform exome-wide gene burden testing for LOY and LOX based on the 450,000 participants from UK Biobank. In this chapter, I ran the burden testing and sensitivity analyses for both LOY and LOX. I also performed the PheWAS analysis for identified genes whose non-synonymous variants affect LOY or LOX. Additionally, I compared the effects on LOY and LOX for the identified genes. Dr Eugene J. Gardner designed, tested, and implemented the QC, annotation, and burden testing pipelines for the WES data on the UKBB RAP. Dr Eugene J. Gardner also performed the LOY and CHIP genes association check. Prof John R.B. Perry and Ken K. Ong supervised and provided guidance on the analyses.

6.2 Abstract

Mosaic Y chromosome loss (LOY) in circulating leukocytes is the most common form of clonal mosaicism. Previous studies of common germline genetic variants have identified over 150 LOY associated genomic loci, and the exome-wide gene burden analysis described in chapter 5 identified two genes, *GIGYF1* and *CHEK2*, that had a large effect on the risk of LOY, which was estimated from SNP array data in over 82,277 men recruited by the UK Biobank.

In this chapter, the exome-wide gene burden analysis was re-conducted with an extended sample size of 190,573 men and using an improved measure of LOY based on SNP array data combined with Y dosage information estimated from whole-exome sequencing data, which was described in chapter 2. In addition, rare genetic determinants of LOX were sought using a similar approach in 226,125 UK Biobank women.

My results, replicated the previously reported associations between LOY and *GIGYF1* and *CHEK2* with increased statistical confidence. Additionally, rare variants in three known clonal haematopoiesis of indeterminate potential (CHIP) genes (*ASXL1*, *TET2* and *DNMT3A*) were associated with a decreased risk of LOY. Among females, carriers of rare variants in *FBXO10*, which may play a role in apoptosis, showed increased mosaic LOX. These findings revealed the relationship between different CHIP events. More in-depth experimental work will be needed to investigate the mechanisms behind these associations.

6.3 Introduction

Identified over fifty years ago, mosaic Loss of the Y chromosome (LOY) in leukocytes is the most common form of clonal mosaicism[97,98,120]. In a recent study based on UK Biobank, one in five male participants aged from 40 to 70 have LOY[16]. Numerous epidemiological studies have demonstrated the association between LOY and health conditions including a wide range of cancers[31,36,121,122], type 2 diabetes[34,36], neurodegenerative diseases[26,123], obesity[34,36], and all-cause mortality[31]. But the causes and consequences of LOY remain limited. The emergence of large cohorts with SNP-array data provides the opportunity to investigate the genetic predisposition to LOY by conducting large-scale GWAS studies in hundreds of thousands of individuals[16,30,32,33].

Since the first identification of LOY-associated loci near *TCL1A*[30], there have been over 150 SNPs identified that predispose individuals to mosaic LOY[16]. The implicated genes mapped by these SNPs involve many aspects of cell-cycle regulation and DNA damage response (DDR) pathway[16,32], which could explain the mechanisms behind LOY. In chapter 5, I described an exome-wide, rare variant gene burden test for LOY in over 80,000 males recruited by the UK Biobank[124]. In this chapter, I showed that rare LOF variants in *GIGYF1* result in a 6-fold increase in risk of LOY

In contrast to LOY, the number of epidemiological and genetic studies on mosaic loss of the X chromosome (LOX) in females is comparatively limited, largely due to LOX being significantly less prevalent than LOY[35]; approximately 8% of women aged over 65 have detectable somatic mosaic X chromosome loss[17,18].

Both LOY and LOX are the special common type of clonal haematopoiesis, the other types of clonal haematopoiesis include mosaic chromosomal alterations (mCAs) and clonal haematopoiesis of indeterminate potential (CHIP) defined as clonal haematopoiesis arise through the age-related acquisition of somatic mutations in the myeloid-associated genes[18,24]. Previously GWAS studies identified several shared genetic risk variants between LOY and the other types of clonal haematopoiesis

including variants mapped to *TET2* shared between LOY and CHIP, *TCL1A* shared between LOY, CHIP and *DNMT3A*-CH, *CHEK2* and *ATM* shared between LOY, CHIP and *JAK2*-CH, *TERT* shared between LOY, CHIP and mCAs, and *HLA*, *DLK1* and *MAD1L1* shared between LOY and mCAs[18]. In this chapter, both the loss of function with MAF< 0.1% and missense variants with CADD scores>25 and MAF<0.1% of *CHEK2* were identified significantly associated with LOY and nominally associated with LOX, which further confirmed the role of *CHEK2* on clonal haematopoiesis. Additionally, the loss of function variants with MAF<0.1% of three typical clonal haematopoiesis of indeterminate potential genes including *ASXL1*, *TET2* and *DNMT3A* were identified to significantly decrease the risk of LOY, which may reveal the relationships between LOY and CHIP.

In this chapter, I used the combined measures of sex chromosome mosaicism presented in chapter 2 from an extended collection of 416,698 individuals recruited by the UK Biobank (190,573 males and 226,125 females), to explore the effects of rare nonsynonymous variants on mosaic sex chromosome loss.

6.4 Methods

6.4.1 LOY and LOX measures for analysis

The LOY and LOX measures used in this chapter were described in chapter 2. Compared with other LOY and LOX measures, the LOY/X Combined Call (3-way) exploited the sex chromosome dosage information generated from both SNP array and whole-exome sequence data. To compare the statistical power of these LOY and LOX measures, in the benchmark test against age, the primary risk factor for LOY and LOX, both variables showed the most significant association. The formula used to calculate LOY/X Combined Call (3-way) was:

LOY/X Combined Call (3-way)=PAR-LOY/MoChA-LOX+2×AF-LOY/X-2×mLRR-Y/X - 4×(Y/X dosage-1/2) (cropped to the range [0,2])

6.4.2 Study population

The analyses described in this chapter were conducted on the samples from UK Biobank. The samples for the gene-burden testing were restricted to "white European" as defined in previous chapters. The samples with abnormal karyotypes, including male 47, XXY, 47, XYY, 48, XXYY and female 45, XO and 47, XXX were excluded from the analysis as defined in chapter 4. Individual samples with an excess of heterozygosity or autosomal variant missingness≥5% on released SNP-array data, or not contained in the subset of phased samples as defined in Bycroft et al.[50] were also set as missing.

6.4.3 Defining rare variant masks

Before conducting the rare variants gene-burden testing, several quality control and data format conversion procedures were performed on both participants and genotypes[125]. The whole-exome-sequencing data in the VCF format for 454,787 participants was extracted via the UK Biobank Research Access Platform (UK Biobank RAP, https://ukbiobank.dnanexus.com/)[126].

For the gVCF files of genotype data, BCFtools[54] was implemented to split and left-normalise multi-allelic sites. A missingness-based approach was used to filter variants. Based on this approach, genotype calls in the following categories were set as missing (i.e.,". /."):

1. Single nucleotide variants (SNVs) with sequencing depth (DP) less than 7, Genotype Quality (GQ) less than 20, and with an allelic balance p. value<0.001

2. Small insertions and deletions (InDels) with DP less than 10 and GQ less than 20

After setting all genotype calls in these categories as missing, variants with greater than 50% missing genotype calls were excluded from further analysis.

To annotate the remaining variants on the autosomal and X chromosomes, the ENSEMBL Variant Effect Predictor (VEP) (version:104)[105] was used with the "everything" flag and LOFTEE plugin. The predicted consequence of each variant was estimated by comparison to a single MANE (version:0.97) or VEP canonical ENSEMBL transcript and the most damaging consequence as defined by VEP defaults. The variants with high confidence (HC, as defined by LOFTEE[107]) stop gained, splice donor/acceptor, and frameshift consequences were grouped as protein-truncating variants (PTVs). CADD (version:1.6) was used to calculate the Combined Annotation Dependent Depletion (CADD) scores for all variants[127].

6.4.4 Rare variants gene-burden testing

Using BOLT-LMM (version:2.3.5)[63], I performed rare variants gene-burden testing using provided UK Biobank WES data . After dropping individuals with missing data, 190,573 males with LOY Combined Call (3-way) and 226,125 females with LOX Combined Call (3-way) data remained for association testing. To compute the GRM required for BOLT-LMM, I inputted genotyping data for variants with allele counts greater than 100. Additionally, I also performed single marker tests for all variants genotyped from WES passing QC as defined above. A set of dummy genotypes were

generated, which represented participant carrier status per-gene for PTVs, missense variants with CADD scores ≥ 25 (MISS_CADD25) and damaging variants that are the combination of the PTV with high confidence and the missense variants with CADD scores ≥ 25 (HC_PTV + MISS_CADD25) respectively. The Minor Allele Frequency (MAF) threshold for these rare non-synonymous variants was set as 0.1%. For each gene, carriers with non-synonymous variants were set as heterozygous ("0/1") and the non-carriers were set as homozygous reference ("0/0"). All models were controlled for age, age^2, WES batch, sex, and the first ten genetic ancestry principal components (PCs) as described in Bycroft et al.[50]

I further excluded genes with less than 50 carriers, resulting in the final inclusion of 8,975, 14,682 and 16,064 genes with PTV, MISS_CADD25 and DMG variant masks, respectively for LOY, and 9,858, 15,144 and 16,493 genes with PTV, MISS_CADD25 and DMG masks, respectively for LOX. After Bonferroni multiple testing correction, the exome-wide significant threshold for a statistically significant association was set as $0.05/39{,}721=1.26\times10^{-6}$ and $0.05/41{,}495=1.20\times10^{-6}$ for LOY and LOX, respectively.

For significant genes, I conducted a leave-one-out analysis using a generalised linear model (GLM) with the `glm` function in R to identify gene associations driven by a single variant. As BOLT does not provide accurate odds ratio (OR) estimates for binary traits, ORs were extracted from GLMs for PAR-LOY and MoChA-LOX in R.

As an orthogonal approach for gene burden testing, I applied STAAR[45] using identical dummy genotype and covariates as BOLT-LMM. Gene burden p-values from STAAR output were extracted and compared with BOLT p-values.

6.4.5 Associations between the CHIP loss of function variants and LOY

To investigate whether rare variants within these genes might have arisen somatically, we queried per-genotype allelic depth (the number of reads from sequencing supporting each of the alleles of that site) information from quality-

controlled and annotated variant call format (VCF) files generated for rare variant burden testing.

Allelic depth (i.e., the number of sequencing reads supporting the alternative and reference alleles) was extracted for all carriers of PTV variants with MAF<0.1% for genes associated with LOY. Variant Allele Fraction (VAF) for each genotype call was calculated for each genotype using the following formula:

$$AD\_alt / (AD\_alt + AD\_ref)$$

A VAF of 0.5 indicates that the balance of reads supporting the alternative and reference alleles is the same, and thus consistent with heterozygous germline inheritance. Significant departures from this ratio may indicate somatic events.

All variants of these genes were annotated based on whether they were known, or likely CHIP driver mutations based on the criteria proposed by Bick et al.[24]. For each gene, the association tests between PTV carrier status and LOY were performed under six different settings by excluding the individuals carrying the variants with the following characteristics:

1. Frameshift InDels with a binomial test p. value for allele balance < 0.001 (i.e., filtering InDels identically to SNVs, see section 11.4.3).
2. Any variant with VAF<0.25 or>0.75.
3. Any variant with VAF<0.4 or>0.6.
4. Any variant with VAF>0.35.
5. A variant explicitly listed in Supplementary Table 3 from Bick et al.[24]
6. As above in (5), but also matching the criteria in Supplementary Table 2 from Bick et al.[24]

All these association tests were conducted separately for each gene using a linear regression model with the same covariates as used in the rare variant burden tests described above.

6.4.6 PheWAS analysis

To explore the wider health consequences of rare variants, we assessed several phenotypes representing a wide range of health conditions including cancers, metabolic traits, reproductive traits, basic anthropometric measures, blood biomarkers and behaviours. All these analyses were performed independently by using the pipeline described above. For each identified LOY and LOX related gene, I extracted test statistics for each trait. The exome-wide significant multiple-testing threshold was set the same as the threshold used for discovering LOY and LOX related genes.

6.5 Results

6.5.1 Non-synonymous variants affecting LOY

I performed an exome-wide gene-burden analysis for LOY in male participants from UK Biobank. As in the results presented as part of Chapter 9, My analysis confirmed that carriers of loss of function variants within *GIGYF1* and *CHEK2* were more likely to have mosaic LOY. My analysis also identified three novel genes including *ASXL1*, *TET2* and *DNMT3A*, which are all known CHIP genes[24,25] **(Figure 6-1, Table 6-1)**. The carriers with loss of function variants with MAF<0.1% in *GIGYF1* still showed the most significant combined association with higher risk of having LOY (N=81 carriers, beta=0.451, SE=0.056, P=9.20×10^{-16}). For *CHEK2*, not only the carriers with loss of function variants with MAF < 0.1% had more risk to have LOY (N=325 carriers, beta=0.143, SE=0.028, P=3.50×10^{-7}), but also the carriers with missense variants with CADD scores>25 and MAF<0.1% (N=811 carriers, beta=0.096, SE=0.018, P=6.30×10^{-8}) variants and the carriers with damaging variants with MAF<0.1% (N=1136 carriers, beta=0.110, SE=0.015, P=2.80×10^{-13}).

Different from *GIGYF1* and *CHEK2*, the carriers with loss of function variants with MAF<0.1% within three known CHIP genes are associated with a decreased risk of having LOY. Additionally, the carriers with damaging variants with MAF<0.1% in *DNMT3A* also significantly decreased the risk to have LOY. The carriers with loss of function variants with MAF<0.1% in *ASXL1*, which is one of the most frequently mutated genes in all subtypes of myeloid malignancies[128], were associated with decreased risk of LOY (N=212 carriers, beta=-0.264, SE=0.035, P=3×10^{-14}). For *TET2*, which can drive tumorigenesis in several blood cancers as well as in solid cancers[129], the carriers of loss of function variants with MAF< 0.1% in *TET2* had a decreased risk of LOY (N=192 carriers, beta=-0.261, SE=0.036, P=8.4×10^{-13}). *DNMT3A* is frequently mutated in a large variety of immature and mature hematologic neoplasms[130]. The carriers of both loss of function variants with MAF<0.1% (N=89 carriers, beta=-0.278, SE=0.053, P=2.1×10^{-7}) and damaging variants with MAF<0.1% (N=270 carriers, beta=-

0.173, SE=0.031, P=1.9×10^{-8}) in *DNMT3A* were associated with decreased risk of having LOY.

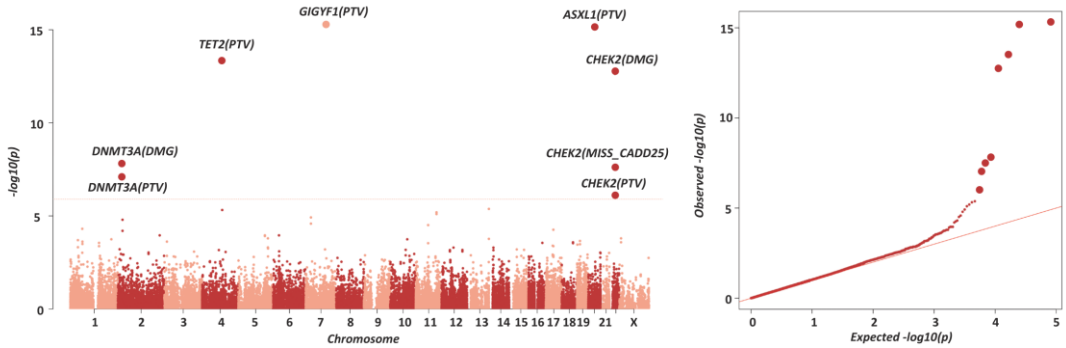


Figure 6-1 Manhattan and Quantile-Quantile (Q-Q) plots for rare variants gene-burden test statistics for LOY. The dashed blue denotes the exome-wide significance threshold (P<1.26×10^{-6}). The Genomic Inflation Factor (λ) is 1.05 and the sample size is 190,573.

Several sensitivity analyses were performed for these significant associations. None of these gene burden associations were driven by single variants, as shown by leave-one-out analyses **(Table 6-2)**. All these gene burden associations still reached nominal significance (P<4.38×10^{-5}) after dropping the most significant single variant. To estimate the relative risk of having LOY, logistic regressions were conducted for the binary PAR-LOY measure **(Table 6-1)**. The loss of function variants with MAF<0.1% carriers of *GIGYF1* conferred a 5-fold (OR=4.9, 95% C.I.=3 to 8, P=1.51×10^{-10}) higher risk to have LOY. The carriers with loss of function (OR=1.8, 95% C.I.=1.4 to 2.3, P=1.18×10^{-5}), missense variants with CADD scores>25 (OR=1.6, 95% C.I.=1.3 to 1.9, P=9.76×10^{-8}) and damaging (OR=1.6, 95% C.I.= 1.4 to 1.9, P=6.81×10^{-12}) variants with MAF<0.1% of *CHEK2* had a 2-fold increased risk to have LOY.

In contrast, the loss of function variants with MAF<0.1% carriers of *TET2* had a 6-fold (OR=0.16, 95% C.I.= 0.09 to 0.29, P=1.51×10^{-10}) lower risk to have LOY. Both the loss of function variants with MAF<0.1% carriers of *ASXL1 (*OR=0.32, 95% C.I.= 0.21 to 0.49, P=6.53×10^{-8}) and *DNMT3A* (OR=0.3, 95% C.I.= 0.16 to 0.59, P=4.44×10^{-4}) had 3-fold lower risks to have LOY. The damaging variants with MAF<0.1% of *DNMT3A* had 2-fold lower risk (OR=0.44, 95% C.I.= 0.31 to 0.62, P=3.22×10^{-6}). The orthogonal

approach, STAAR, generated consistent significant p-values for all identified genes **(Table 6-1)**. There was no gene identified from synonymous variants mask.

Table 6-1 Exome-wide significant gene burden associations with LOY and LOX

		BOLT_LMM								Logistic Regression		STAAR			
PHENOTYPE	**SYMBOL**	**FREQ**	**BETA**	**SE**	**CHISQ**	***P***	**MASK**	**MAF**	**AC**	**OR**	***P***	***P (STAAR-O)***	***P (SKAT)***	***P (Burden)***	***P (ACAT)***
LOY (3WAY)	***GIGYF1***	2.13E-04	0.45	0.06	64.59	9.20E-16	PTV	MAF_01	81	4.92[3.02, 8]	1.51E-10	1.33E-15	6.08E-05	4.38E-16	7.67E-13
	CHEK2	2.98E-03	0.11	0.02	53.35	2.80E-13	DMG	MAF_01	1136	1.63[1.42, 1.88]	6.81E-12	1.59E-13	2.43E-06	5.33E-14	1.19E-03
	CHEK2	2.13E-03	0.10	0.02	29.26	6.30E-08	MISS_CADD25	MAF_01	811	1.57[1.33, 1.86]	9.76E-08	8.59E-08	1.73E-04	2.87E-08	1.91E-03
	CHEK2	8.53E-04	0.14	0.03	25.95	3.50E-07	PTV	MAF_01	325	1.78[1.37, 2.3]	1.18E-05	3.81E-07	7.19E-04	1.28E-07	1.05E-03
	ASXL1	5.56E-04	-0.26	0.03	57.72	3.00E-14	PTV	MAF_01	212	0.32[0.21, 0.49]	6.53E-08	4.85E-12	6.05E-09	1.60E-12	1.16E-08
	TET2	5.04E-04	-0.26	0.04	51.19	8.40E-13	PTV	MAF_01	192	0.16[0.09, 0.29]	9.27E-10	2.21E-11	1.38E-01	1.47E-11	1.47E-11
	DNMT3A	7.08E-04	-0.17	0.03	31.61	1.90E-08	DMG	MAF_01	270	0.44[0.31, 0.62]	3.22E-06	7.52E-08	9.90E-04	2.51E-08	1.53E-05
	DNMT3A	2.34E-04	-0.28	0.05	26.98	2.10E-07	PTV	MAF_01	89	0.3[0.16, 0.59]	4.44E-04	3.44E-07	7.94E-01	2.29E-07	2.29E-07
LOX (3WAY)	***FBXO10***	1.28E-03	0.06	0.01	27.21	1.80E-07	MISS_CADD25	MAF_01	581	1.97[1.53, 2.54]	1.35E-07	6.81E-07	2.87E-06	2.48E-07	5.38E-05
	FBXO10	1.44E-03	0.05	0.01	24.44	7.70E-07	DMG	MAF_01	650	2.06[1.59,2.68]	6.34E-08	1.99E-06	3.29E-06	8.39E-07	5.71E-05

Table 6-2 Test statistics after dropping the variant with most significant effect on the burden test

PHENOTYPE	**SYMBOL**	**MASK**	**VAR dropped**	**BETA**	**SE**	**T-statistics**	***P***
LOY (3WAY)	***GIGYF1***	PTV	7:100687545:CA:C	0.50	0.07	7.51	5.73E-14
	CHEK2	PTV	22:28695238:TA:T	0.16	0.04	4.09	4.38E-05
	CHEK2	MISS_CADD25	22:28725338:T:C	0.09	0.02	4.43	9.36E-06
	CHEK2	DMG	22:28725338:T:C	0.11	0.02	6.57	4.91E-11
	ASXL1	PTV	20:32434638:A:AG	-0.21	0.05	-4.40	1.06E-05
	TET2	PTV	4:105234885:TC:T	-0.26	0.04	-6.95	3.78E-12
	DNMT3A	PTV	2:25247049:A:C	-0.27	0.06	-4.92	8.60E-07
	DNMT3A	DMG	2:25234374:G:A	-0.16	0.03	-4.99	5.98E-07
LOX (3WAY)	***FBXO10***	DMG	9:37518378:G:A	0.05	0.01	4.25	2.16E-05
	FBXO10	MISS_CADD25	9:37518378:G:A	0.05	0.01	4.45	8.45E-06

6.5.2 Somatic and germline mutations in CHIP genes show similar effects on LOY

Clonal Haematopoiesis of Indeterminate Potential (CHIP) can be defined as the accumulation of somatic mutations over time that lead to clonal expansion in regenerating haematopoietic stem cell populations[24]. Rare variant gene-burden testing identified 3 known CHIP genes associated with LOY[24,25]: *ASXL1* (N=212 carriers), *DNMT3A* (N=89), and *TET2* (N=192). Yet it was unclear if these associations were due to reverse-causality; somatic genome instability, consequent to LOY, could cause or occur in parallel with somatic mutations arising within CHIP genes. As part of WES variant quality control (section 6.4.3), a filter on variant allele frequency (VAF) had applied to heterozygous alleles for all variants. This filtering approach will exclude variants with strong VAF imbalance but, due to the characteristics of CHIP genes, it was further explored whether the associations between the CHIP genes and LOY were driven by variants with more marginal VAF scores, which could be indicative of somatic mutation. Compared with the LOY associated variants in *GIGYF1* whose VAF distributions showed a mean close to 0.5, the VAF distribution of the three CHIP genes was skewed to the left (i.e.,<< 0.5), which indicates that some identified variants might be of somatic origin **(Figure 6-2)**; variants that arose only in a single white blood stem-cell population will exist in a lower fraction of cells in the entire blood cell population and thus appear to be at lower VAF. Therefore, different VAF and variants category filters were tested to distinguish the effects of germline and somatic mutations (**Figure 6-3**). Although the p-values varied between the different variant filters (likely due to differences in numbers of tested variants), the effect size of CHIP genes kept consistent, which indicated that both somatic (VAF< 0.35) and germline mutations of CHIP genes can decrease the risk of LOY **(Figure 6-3).** As a positive control, we also performed the same analysis for *GIGYF1,* which is associated with LOY but not involved in CHIP. *GIGYF1* also showed highly consistent positive effects on LOY, except for when variants were restricted to potential somatic mutations. This is because most variants on *GIGYF1* are likely to be of germline origin, and thus the number of carriers of potential somatic mutations is

very small. Therefore, the *P* value became non-significant, but the effect size was still consistent.

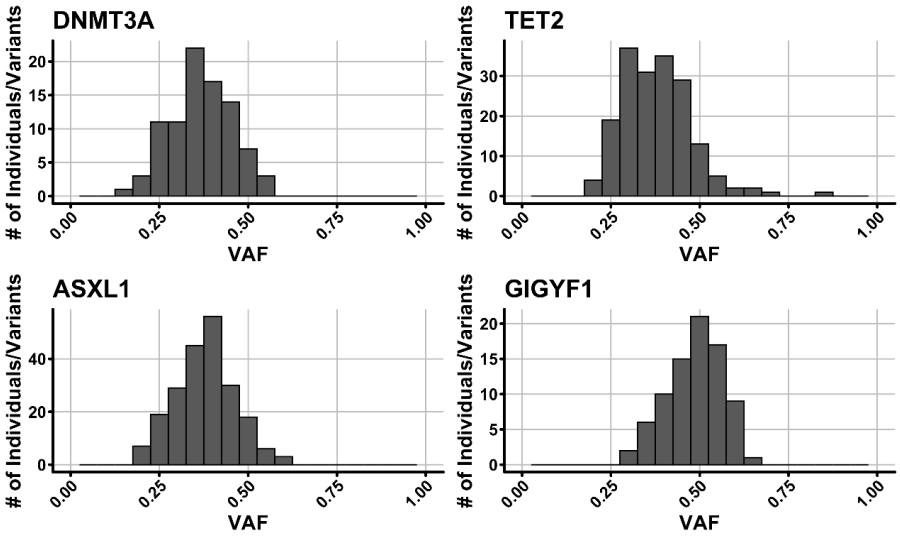


Figure 6-2 Variant Allele Frequency (VAF) histograms for the four genes *DNMT3A*, *TET2*, *ASXL1* and *GIGYF1*. Bars indicate total number of genotypes in 0.05 VAF bins.

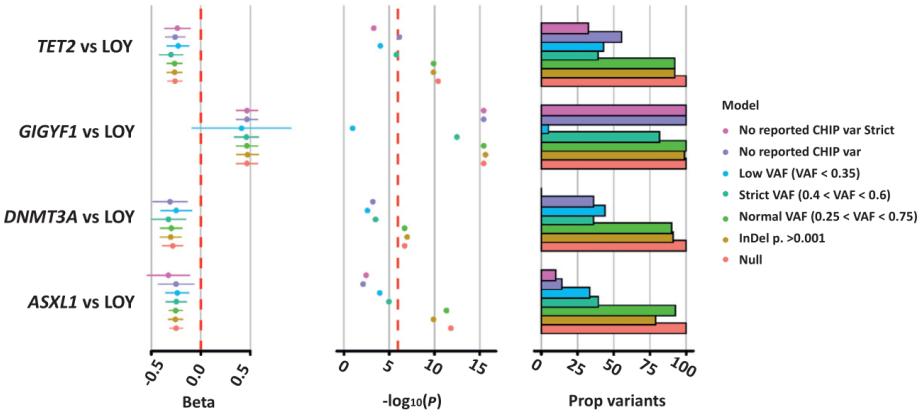


Figure 6-3 Plotted are beta ± 95% CI (left), -log10 (p. value) (middle) and proportion of variants remaining after filtering (right) for each gene/model combination. ”No CHIP Vars” and ”No CHIP Vars Strict” indicate models excluding known CHIP variants or known CHIP variants and variants identified by a broader set of criteria presented in Bick et al.[24] , respectively. No variants remained for *DNMT3A* after performing filtering according to criteria outlined in Bick et al.[24], thus beta and p. value estimates are not presented for this model. Also plotted are unfiltered models for all four genes for comparative purposes (Null).

6.5.3 PheWAS of LOY-related genes

To further explore the associations between the identified LOY-associated genes and other health related traits, the phenome-wide association study (PheWAS) was conducted **(Table 6-3)**. Apart from *GIGYF1*, the damaging variants with MAF< 0.1% of *DNMT3A* and *CHEK2* and the loss of function variants with MAF<0.1% of *ASXL1*, *TET2* and *DNMT3A* were significantly associated with increased risks of blood cancers. This was consistent with previous observations[128–130]. Moreover, the loss of function variants with MAF<0.1% of *ASXL1* and *TET2* were associated with shorter leukocyte telomere length. For *CHEK2*, the reported association with later age at natural menopause was also identified in this analysis[131] - its loss of function and damaging variants with MAF<0.1% were also associated with increased combined risk of a female hormone-sensitive cancer (breast cancer, ovarian cancer, or uterine cancer) and its loss of function variants with MAF<0.1% were associated with an increased risk of male prostate cancer. The previously reported significant associations between the loss of function variants with MAF<0.1% of *GIGYF1* and metabolic-health related traits were replicated in this analysis in a more than doubled sample size.

Table 6-3 Significant associations identified for LOY and LOX associated genes from the PheWAS analysis

PHENOTYPE	SYMBOL	A1FREQ	BETA	SE	CHISQ	*P*	MASK	MAF	AC
Adjusted leukocyte telomere length	*ASXL1*	3.96E-04	-0.07	0.01	88.01	6.50E-21	PTV	MAF_01	324
Cancers(any type, male only)	*ASXL1*	5.56E-04	0.15	0.03	27.60	1.50E-07	PTV	MAF_01	214
Cancers(excluding skin cancers)	*ASXL1*	3.91E-04	0.14	0.02	44.74	2.30E-11	PTV	MAF_01	329
Cancers(excluding skin cancers, male only)	*ASXL1*	5.56E-04	0.17	0.03	46.46	9.30E-12	PTV	MAF_01	214
Blood cancers	*ASXL1*	3.91E-04	0.10	0.01	185.48	3.10E-42	PTV	MAF_01	329
Blood cancers(female only)	*ASXL1*	2.52E-04	0.07	0.01	38.82	4.70E-10	PTV	MAF_01	115
Blood cancers(male only)	*ASXL1*	5.56E-04	0.11	0.01	131.59	1.80E-30	PTV	MAF_01	214
Cancers(any type)	*CHEK2*	2.20E-03	0.05	0.01	25.21	5.10E-07	MISS_CADD25	MAF_01	1856
Cancers(any type)	*CHEK2*	3.02E-03	0.05	0.01	43.51	4.20E-11	DMG	MAF_01	2542
Cancers(any type, male only)	*CHEK2*	2.98E-03	0.06	0.01	24.89	6.10E-07	DMG	MAF_01	1149
Cancers(excluding blood cancers)	*CHEK2*	3.02E-03	0.05	0.01	32.57	1.20E-08	DMG	MAF_01	2542
Cancers(excluding skin cancer)	*CHEK2*	2.20E-03	0.05	0.01	36.28	1.70E-09	MISS_CADD25	MAF_01	1856
Cancers(excluding skin cancer)	*CHEK2*	3.02E-03	0.06	0.01	70.32	5.00E-17	DMG	MAF_01	2542
Cancers(excluding skin cancer)	*CHEK2*	8.15E-04	0.09	0.01	38.61	5.20E-10	PTV	MAF_01	686
Cancers(excluding skin cancers, female only)	*CHEK2*	3.05E-03	0.06	0.01	35.20	3.00E-09	DMG	MAF_01	1393
Cancers(excluding skin cancers, male only)	*CHEK2*	2.98E-03	0.07	0.01	35.43	2.60E-09	DMG	MAF_01	1149
Blood cancer	*CHEK2*	3.02E-03	0.02	0.00	34.29	4.80E-09	DMG	MAF_01	2542
Blood cancer(male only)	*CHEK2*	2.98E-03	0.02	0.00	28.63	8.70E-08	DMG	MAF_01	1149
Hormone sensitive cancer(female only)	*CHEK2*	3.05E-03	0.05	0.01	44.38	2.70E-11	DMG	MAF_01	1393
Hormone sensitive cancer(female only)	*CHEK2*	7.79E-04	0.10	0.02	35.78	2.20E-09	PTV	MAF_01	356
Hormone sensitive cancer(male only)	*CHEK2*	2.98E-03	0.04	0.01	29.90	4.60E-08	DMG	MAF_01	1149
Age at menopause	*CHEK2*	2.01E-03	1.25	0.21	35.47	2.60E-09	MISS_CADD25	MAF_01	431
Age at menopause	*CHEK2*	2.70E-03	1.58	0.18	75.64	3.40E-18	DMG	MAF_01	578
Age at menopause	*CHEK2*	6.87E-04	2.53	0.36	49.40	2.10E-12	PTV	MAF_01	147
Blood cancers	*DNMT3A*	7.27E-04	0.04	0.01	46.09	1.10E-11	DMG	MAF_01	612
Blood cancers	*DNMT3A*	2.56E-04	0.06	0.01	50.61	1.10E-12	PTV	MAF_01	216
Blood cancers(female only)	*DNMT3A*	7.48E-04	0.03	0.01	24.31	8.20E-07	DMG	MAF_01	342
Blood cancers(male only)	*DNMT3A*	2.31E-04	0.10	0.02	40.39	2.10E-10	PTV	MAF_01	89
glucose	*GIGYF1*	1.64E-04	0.62	0.07	70.48	4.60E-17	PTV	MAF_01	120
glucose(excluding T2D cases)	*GIGYF1*	1.28E-04	0.40	0.07	32.60	1.10E-08	PTV	MAF_01	87
HbA1c	*GIGYF1*	1.64E-04	4.41	0.41	118.07	1.70E-27	PTV	MAF_01	131
HbA1c(excluding T2D cases)	*GIGYF1*	1.31E-04	2.61	0.36	52.31	4.70E-13	PTV	MAF_01	97
HDL	*GIGYF1*	1.68E-04	-0.15	0.03	28.31	1.00E-07	PTV	MAF_01	124
LDL	*GIGYF1*	1.64E-04	-0.48	0.07	50.94	9.50E-13	PTV	MAF_01	132
T2D	*GIGYF1*	1.59E-04	0.18	0.02	66.06	4.40E-16	PTV	MAF_01	133
T2D(female only)	*GIGYF1*	1.17E-04	0.16	0.03	26.33	2.90E-07	PTV	MAF_01	53
T2D(male only)	*GIGYF1*	2.09E-04	0.20	0.03	36.02	2.00E-09	PTV	MAF_01	80
waist-to-hip ratio adjusted for BMI(male only)	*GIGYF1*	2.14E-04	0.03	0.01	23.72	1.10E-06	PTV	MAF_01	82
Adjusted leukocyte telomere length	*TET2*	4.49E-04	-0.05	0.01	62.52	2.60E-15	PTV	MAF_01	367
Cancers(any type)	*TET2*	4.56E-04	0.12	0.02	31.00	2.60E-08	PTV	MAF_01	384
Cancers(any type, male only)	*TET2*	5.01E-04	0.15	0.03	24.41	7.80E-07	PTV	MAF_01	193
Cancers(excluding skin cancers)	*TET2*	4.56E-04	0.11	0.02	36.92	1.20E-09	PTV	MAF_01	384
Cancers(excluding skin cancers, male only)	*TET2*	5.01E-04	0.17	0.03	39.29	3.70E-10	PTV	MAF_01	193
Blood cancers	*TET2*	4.56E-04	0.12	0.01	343.62	1.00E-76	PTV	MAF_01	384
Blood cancers(female only)	*TET2*	4.18E-04	0.10	0.01	138.49	5.70E-32	PTV	MAF_01	191
Blood cancers(male only)	*TET2*	5.01E-04	0.15	0.01	198.18	5.20E-45	PTV	MAF_01	193

6.5.4 Damaging variants in *FBXO10* associated with LOX

To identify genes, whose non-synonymous variants were associated with LOX, the exome-wide rare variants genes burden tests were performed for 226,125 females with the LOX Combined Call (3-way) **(Figure 6-4, Table 6-1)**. Only one gene, *FBXO10* (F-Box Protein 10), reached statistically significant ($P<1.24\times10^{-6}$) . Both the carriers with missense variants with CADD scores>25 (n=581 carriers) and the carriers damaging variants (n=650 carriers) with MAF<0.1% of *FBXO10* were associated with an increased risk of LOX with $P=1.8\times10^{-7}$ (beta=0.059, SE=0.011) and 7.7×10^{-7} (beta=0.052, SE=0.01), respectively. To estimate the relative risk of having LOX, I conducted a logistic regression for dichotomous LOX-status. The carriers of missense variants with CADD scores> 25 and damaging variants with MAF<0.1% of *FBXO10* had 2-fold increased risk of having LOX (OR=2.1, 95% C.I.= 1.6 to 2.7, $P=1.35\times10^{-7}$ and OR=2.0, 95% C.I.= 1.5 to 2.5, $P=6.34\times10^{-8}$).

In sensitivity analyses, to exclude the possibility of one single variant driving the association, after dropping any variant, the association between *FBXO10* and LOX remained significant ($P<8.45\times10^{-6}$ for missense variants with CADD scores> 25 and MAF<0.1%, and $P<2.16\times10^{-5}$ for damaging variants with MAF<0.1%) (Table 6-1). By implementing the orthogonal analytical pipeline STAAR, the association between *FBXO10* and LOX was confirmed ($P<2.48\times10^{-7}$ for missense variants with CADD scores> 25 and MAF< 0.1%, and $P<8.39\times10^{-7}$ for damaging variants with MAF< 0.1%) **(Table 6-1)**. There was no gene identified from synonymous variants mask.

FBXO10 is located on 9p13.2, which is the substrate-recognition component of SCF (SKP1-CUL1-F-box protein)-type E3 ubiquitin ligase complex. The SCF(*FBXO10*) complex mediates ubiquitination and degradation of the antiapoptotic protein, *BCL2 (BCL2 Apoptosis Regulator*), thereby playing a role in apoptosis by controlling the stability of *BCL2*[132], which is a LOY associated gene from previous GWAS study[16]. There was no association between *FBXO10* and diseases or other traits in PheWAS. The most significant nominal association was with shorter height of the carriers

(P=2.4×10^{-6}, beta=-0.71, SE=0.15, for missense variants with CADD scores>25 and MAF<0.1%, and P=1.7×10^{-6}, beta=-0.68, SE=0.14, for damaging variants with MAF<0.1%).

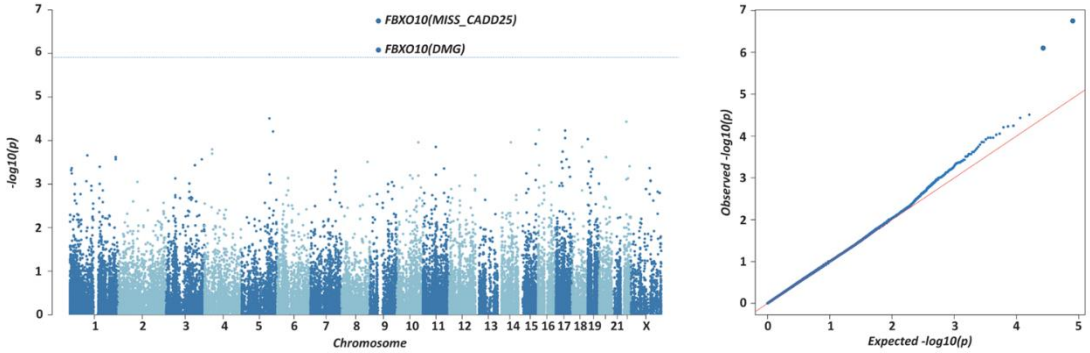


Figure 6-4 Manhattan and Quantile-Quantile (Q-Q) plots for rare variants gene-burden test statistics. The dashed blue denotes the exome-wide significance threshold (P<1.20×10^{-6}). The Genomic Inflation Factor is 1.002 and the sample size is 226,125.

6.5.5 Comparison between the effect on LOY and LOX for the identified genes

In order to explore whether the LOY-associated genes can affect LOX and vice versa, the summary statistics of these gene settings were extracted **(Table 6-4)**. *ASXL1*, *CHEK2* and *TET2* were nominally associated (P<0.05) with an increased risk of LOX. In contrast, *FBXO10* was not associated with LOY. It should be noted that there were 100 more male carriers of *ASXL1* than females (212 vs. 113; test of equal proportions $P=2.5\times10^{-12}$). This sex imbalance might reflect that *ASXL1* may mutate much more frequently in males. The same pattern was also observed on the loss of function variants of *GIGYF1* (81 vs. 52; $P=6.2\times10^{-4}$).

Table 6-4 Test statistics of LOY and LOX for all LOY and LOX associated variant masks

			LOX (3WAY)				LOY (3WAY)			
SYMBOL	MASK	MAF	**AC**	**BETA**	**SE**	***P***	**AC**	**BETA**	**SE**	***P***
ASXL1	PTV	MAF_01	113	0.064	0.026	1.30E-02	212	-0.264	0.035	3.00E-14
CHEK2	MISS_CADD25	MAF_01	1028	0.024	0.009	4.80E-03	811	0.096	0.018	6.30E-08
CHEK2	DMG	MAF_01	1381	0.030	0.007	3.90E-05	1136	0.110	0.015	2.80E-13
CHEK2	PTV	MAF_01	353	0.048	0.015	9.10E-04	325	0.143	0.028	3.50E-07
DNMT3A	DMG	MAF_01	336	-0.021	0.015	0.16	270	-0.173	0.031	1.90E-08
DNMT3A	PTV	MAF_01	125	-0.010	0.024	0.69	89	-0.278	0.054	2.10E-07
GIGYF1	PTV	MAF_01	52	0.054	0.038	0.15	81	0.451	0.056	9.20E-16
TET2	PTV	MAF_01	190	0.051	0.020	1.00E-02	192	-0.261	0.037	8.40E-13
FBXO10	MISS_CADD25	MAF_01	581	0.059	0.011	1.80E-07	469	-0.030	0.023	0.20
FBXO10	DMG	MAF_01	650	0.053	0.011	7.70E-07	519	-0.025	0.022	0.26

6.6 Discussion

In summary, this study substantially increased the power to identify genes in which loss-of-function or missense variants may have a direct effect on LOY or LOX, by increasing the sample size and using improved LOY and LOX measures with stronger statistical power. The p-value for *GIGYF1* loss of function variants with MAF<0.1% was boosted after increasing the sample size (from $P=1.3\times10^{-10}$ to 5.4×10^{-14} using PAR-LOYq), and still further by using the LOY 3-way Combined Call (to $P=9.2\times10^{-16}$). Furthermore, the significant associations between *GIGYF1* loss of function variants with MAF<0.1% and the adverse effects on metabolic health were further confirmed, which had also been replicated by other independent groups[133–135]. As the current knowledge about *GIGYF1* was very limited, more function validations are needed to reveal the mechanism behind the strong association between LOY and T2D.

In chapter 5, only the loss of function variants of *CHEK2* were associated with LOY, which was driven by the most common single frameshift variant (1100delC, MAF~0.2%) in *CHEK2*. After excluding this variant, the p-value for *CHEK2* decreased to 0.02. In this study, both loss of function and missense variants with CADD score>25 were significantly associated with LOY without including the aforementioned frameshift variant. This confirmed the effect of *CHEK2* on LOY. Additionally, *CHEK2* was also associated at nominal significance with an increased risk of LOX. Moreover, *CHEK2* was associated with later menopause, increased risk of having hormone-sensitive cancers, and blood cancers. The serine/threonine-protein kinase coded by *CHEK2* is necessary for checkpoint-mediated cell cycle arrest, activation of DNA repair and apoptosis in response to the presence of DNA double-strand breaks[136,137]. All the evidence about *CHEK2* indicated that defects of DNA damage response may be a driver of both LOY and LOX. Therefore, LOY for males and LOX for females may represent a valuable biomarker of DDR which can be measured at a population scale. If LOY and LOX can be treated as a biomarker to measure DDR, further longitudinal studies might be conducted to explore how LOY and LOX were progressed and their association with the development of cancers.

This study firstly identified that three typical CHIP genes, *ASXL1*, *DNMT3A* and *TET2,* were negatively associated with LOY. The hypothesis of the underlying mechanisms might be that

LOY is deleterious for clonal expansion if other driver mutations are present. More comprehensive studies are needed to understand the relationship between these two types of CH.

Finally, the knowledge about the only LOX-associated gene *FBXO10* was also limited. From the known biological pathway that this gene participates in, *FBXO10* might play a role in apoptosis by controlling the stability of *BCL2*, which supports the hypothesis that defects of cell-cycle regulation and DDR might also be a driver for LOX. More in-depth studies on *FBXO10* will be needed to further illustrate the roles it played in LOX.

In this study, due to the limited number of non-Europeans in the UK Biobank, the analyses were restricted to Europeans. However, previous studies showed that the prevalence of LOY was different among genetic ancestry groups[34] and the variant allele frequency might also vary. Therefore, the multi-ancestry analysis is required for future analysis. Moreover, the participants in UK Biobank have a better average health condition than the general population, which may underestimate the prevalence of LOY and LOX and the true genetic effect sizes. Sample size is key in conducting rare variant gene burden testing. After more than doubling the sample size, more genes were identified for LOY. Due to the low prevalence of LOX and possibly less heritable than LOY, more samples may be required to identify signals for LOX.

Chapter 7 Conclusions

7.1 Overview of the thesis

This thesis systematically investigated the detection, causes and consequences of sex chromosome mosaicism by fully exploiting the sequence and phenotype data from large cohorts such as UK Biobank. The thesis can be divided into two major parts: exploration of sex chromosome mosaicism and detection and characteristics of male sex chromosome aneuploidies. Chapters 2, 3, 5, 6 focused on the first part and chapter 4 focused on the second part.

For the sex chromosome mosaicism, this thesis mainly focused on the LOY and extended the same analyses to LOX. In the last ten years, there have been lots of genetic and epidemiological studies on LOY. On the contrary, the studies on LOX are very limited. Therefore, based on the current studies on LOY, I can easily evaluate the results for LOY but cannot find reliable references for LOX. However, the analytical pipeline can be robustly implemented on the analyses on LOX, which may provide some preliminary insights on the mechanisms that contribute to LOX. The whole thesis primarily tried to improve the understanding of sex chromosome mosaicism. However, the individuals with chromosome aneuploidies can also be identified when I used our methods to estimate the level of sex chromosome mosaicism. As the study on females with sex chromosome aneuploidies based on the same study participants had previously been published[85], in this thesis, the males with sex chromosome aneuploidies were studied.

In chapter 2, I showed that the LOY/ LOX calls can be further enhanced by incorporating the exome sequence read depth information of the sex chromosomes. The new LOY/LOX calls were calculated by combining the calls from the SNP-array and the exome sequence to get the continuous call capturing the presence and degree of LOY and LOX. Additionally, the new LOY/LOX calls also were compared with previous widely used LOY/LOX calls. The improved statistical power of the new LOY call was shown by benchmarking the strength of associations against age, smoking, and known genetic determinants. However, the power improvement of the LOX call was not as clear as the LOY call.

In chapter 3, I used the new LOY call to conduct the GWAS and compared the results with the previous published LOY GWAS results. From the results, the statistical power was slightly boosted, and 22 new LOY-associated signals were identified. By implementing the statistical methods and software including LDSC, MR, colocalisation and MTAG, I illustrated the directional correlation among LOY, LTL and MPN, identified the colocalised signals and boosted the signal detection power for MPN. These results increased the understanding of the mechanistic links between LOY and MPN. The analytical pipelines can be extended to other health outcomes in the future.

In chapter 4, the males with sex chromosome aneuploidies were identified from the UK Biobank by exploiting the sequence intensity data on the non-PARs of X and Y chromosomes and validated by calculating the relative read depth of the non-PARs of X and Y chromosomes. From the results, about 1/500 males in UK Biobank have an extra sex chromosome and just a few of them know their sex chromosome aneuploidies condition. By conducting the association tests against a wide range of health conditions, it was shown that both men with one extra X and Y chromosome had increased risks of several health conditions including T2D and COPD.

In chapter 5, the first exome-wide association test analysis was conducted for LOY using both the basic burden test and novel omnibus test methods for rare non-synonymous variants based on over 80,000 male UK Biobank participants with WES data. These analyses not only identified the known LOY-associated gene, *CHEK2* but also a novel LOY-associated gene, *GIGYF1*. Based on the potential biological function of *GIGYF1*, the significant associations between the LOF variants of *GIGYF1* and metabolic-related traits were revealed by conducting the gene-burden test, which linked the underlying mechanisms between LOY and metabolism.

In chapter 6, ExWAS was conducted for LOY and LOX using the LOY/LOX calls that combined the LOY/LOX calls from both SNP-array and WES for over 450,000 UK Biobank participants. For LOY, both *CHEK2* and *GIGYF1* identified in chapter 5 were replicated. Some CHIP genes including *TET2*, *ASXL1* and *DNMT3A* showed a negative association with LOY. After further checking the allele frequency of the variants for each carrier of these genes to distinguish the germline and somatic mutation, it was revealed that both germline and somatic

mutations of these CHIP genes had a negative association with LOY. In addition, the first LOX-associated gene, *FBXO10*, was identified from ExWAS.

Overall, this thesis provided a reference for future studies on LOY and LOX in choosing the best variable. By fully using the analytical pipelines for common variants and rare non-synonymous variants, the thesis revealed the full spectrum of variants affecting the sex chromosome mosaicism, which improved the understanding of the genetic predisposition to LOY and LOX. At the same time, our pilot study on MPN using LOY as a tool illuminated the future path to other health outcomes which have been observationally associated with LOY. Additionally, the study on male sex chromosome aneuploidies had clear clinical implications, which can improve the awareness of male sex chromosome aneuploidies and encourage routine tests for them due to the relatively high prevalence and multiple adverse health consequences. The analytical pipelines on genetic data mentioned in this thesis were very robust and can be easily implemented on other phenotypes, which can help to gain more mechanistic insights of genetic causes.

However, there were still some limitations of this thesis. Firstly, the LOY and LOX variables developed and used in this thesis were derived from the sequence data of bulk DNA from blood samples, but no qPCR test was conducted to validate the calling. Secondly, as the thesis only used the samples from UK Biobank, only "White European" samples were analysed because the majority of the population in UK Biobank is white British. Finally, as the thesis mainly focused on the detection, causes and consequences of LOY and LOX from genetic prospects, a deeper investigation of the specific biological mechanisms of both LOY and LOX was still lacking.

7.2 Future Avenues

7.2.1 Multi-ancestry analyses on LOY and LOX

Like most current genetic analyses[138], the genetic analysis on LOY also suffered from the issue of lack of diversity. Restricting the samples to European also affected the ability to identify more ancestry-specific loci and further affect the understanding of the underlying mechanism. For LOX, large-scale genetic analysis is still lacking. In the future, my colleagues, collaborators, and I are planning to collaborate with other large population studies with genotype data to conduct the multi-ancestry meta-analysis for both LOY and LOX. The target population included in this study will contain White European, African American, Latino, East Asian, and South Asian, which can represent most of the population in the world. Except for White European, most of them are underrepresented in current genetic studies. The increased sample size can not only enhance the power to detect the LOY/LOX associated loci but provide a valuable chance to explore the difference in the pattern of LOY/LOX among different ancestral groups. Additionally, the multi-ancestry meta-analysis can also increase the power to fine map the causal variants due to the reduced linkage disequilibrium windows.

7.2.2 Identification LOY and LOX associated genes using proteomic and single-cell sequencing data

The recent progress on proteomic technologies has made measuring thousands of blood circulating proteins for thousands of participants in the large population studies possible. In MRC Epidemiology Unit, we have used SomaLogic platform to measure 4775 plasma proteins in 10,708 samples from Fenland study[139]. Recently, Sun et al. described the characteristics of proteomic study from UK Biobank, which measured 1,463 proteins for 54,306 participants using Olink Explore 1536 platform[140]. The proteomics data from UK Biobank will be available for my study in Q4, 2022. By combining the data from Fenland, UK Biobank, and other studies such as deCODE[141], more novel and reproducible findings might be identified. I will systematically investigate the associations between individual plasma protein and LOY/LOX by conducting the linear regression model adjusting for age, smoking status, and other study related covariates. After getting the results from individual study,

my colleagues and I will also conduct a fix-effects meta-analysis for the proteins measured in different study and report the significant protein-LOY/LOX association after Bonferroni correction. The LOY-associated proteins can improve our understanding of underlying mechanisms of LOY and illuminate the mechanistic link between LOY/LOX and other health outcomes. Additionally, the protein data can also be used to check whether the identified LOY/LOX-associated can be mapped to the putatively causal genes.

The most recent LOY study investigated the single-cell RNA sequencing data in 86,160 cells from 19 men aged from 64 to 89 and found that about 16% of cells lacked the Y chromosome[16]. This study further showed that the LOY cells were overexpressed T-cell leukaemia/lymphoma protein 1A relative (*Tcl1a*) to the normal cells without LOY. Notably, the germline variation in *TCL1A* is the first identified LOY-associated loci[30], which was confirmed by the following studying on LOY. These observations suggested that there might be a bi-directional relationship between LOY and *Tcl1a* function[16]. In my future study, my colleagues, collaborators, and I will detect the normal expression of 20 protein-coding genes on the male-specific regions of the Y chromosome, which are normally expressed in leukocytes. Therefore, single cells lacking the Y chromosome can be identified by the absence of the sequence reads from these genes. By comparing the gene expressions between the cells lacking and not lacking the Y chromosome, the dysregulations of autosomally expressed genes will be identified. For the single cell data, I hope to work with Sarah Teichmann (Wellcome Sanger Institute), who is the co-founder and principal leader of the Human Cell Atlas international consortium. These analyses will provide unique insight into the cellular consequences of LOY, shedding further light on how to genetic perturbation of white blood cells might directly impact the disease risk. The same analyses can also be easily to extend to LOX.

Additionally, with the GWAS summary statistics and single-cell sequencing data, I am also planning to implement novel methods (SCAVENGE[142] and scDRS[143]) which can help to explore the pathology and cellular origin of LOY and LOX.

7.2.3 Using a systems biology approach to systematically link LOY and LOX associated variants to genes

The important and necessary step for the GWAS studies is to identify the genes that might be the causal risk factors and illustrate the mechanisms. With the list of significant LOY and LOX associated variants and the summary statistics generated from above-mentioned multi-ancestry meta-analysis, my colleagues and I will implement the pipeline based on the systems biology approach to link these variants to their potential causal effector genes.

Previous studies showed that the closest gene of a given leading signals are often causal gene[144,145], but it's imperfect to be a predictor of causality. The novel pipeline developed by our group can fully exploit the information generated from several current widely used variants to genes software and approaches and aggregate them to reveal more biological mechanisms of LOY/LOX.

By integrating the publicly available expression-QTL and protein-QTL data, the pipeline firstly will assess whether the leading signals directly impact the transcription and/or translation of target genes. The tissue enrichment analysis via LD score regression applied to specifically expressed genes (LDSC-SEG)[146] and Cell type-specific analyses (https://github.com/bulik/ldsc/wiki/Cell-type-specific-analyses) can be conducted. The tissues with $p<0.05$ will be used, alongside the data from GTEx tissue fixed-effects meta-analysis (V7)[147], eQTLGen[148] and Brain-eMeta[149]. Co-localisation will be performed using the summary-data-based Mendelian randomisation (SMR) and heterogeneity in dependent instruments (HEIDI) tests (SMR-HEIDI)[150] and the fully Bayesian colocalisation analysis using Bayes Factors (Coloc-ABF)[61] to avoid coincidental overlap of signals due to extend patterns of LD. By incorporating the protein quantitative trait loci (pQTL) data from the Fenland study[139], the above analyses will be conducted as well.

The activity-by-contact (ABC)[151] model has created an enhancer-gene map across 131 human cell types and tissues. For each identified LOY/LOX leading signal, the lists of proxy signals with high LD ($r2>0.8$) of that signal will be calculated. With the data generated from the ABC model, individual genes will be scored if the leading signal or proxy signals fall with one of these identified enhancer regions.

If there were coding variants among the list of proxy signals, they will be annotated by SIFT[152] and POLYPHEN[49]. The gene-level coding variant Multi-marker Analysis of GenoMic Annotation (MAGMA)[153] analysis and Polygenic Priority Score (PoPS)[154] to estimate the effect on the LOY/LOX of the given genes will be performed as well.

After obtaining the list of genes from each of these analyses, the scores based on different categories will be assigned to each gene and will be summed up for each gene. According to the ranks based on the summed scores, the high confidence causal genes will be prioritised, which might unlock the biological insights that can be learned from GWAS summary statistics.

7.2.4 *GIGYF1* function follow-up

In chapter 5, it was observed that the loss of function variants of *GIGYF1* can 6-fold increase the risk of LOY and T2D. Not just LOY and T2D, the loss of function variants of *GIGYF1* were significantly associated with several metabolic traits including BMI, waist-hip ratio, HbA1c, glucose, grip strength and body fat mass. In chapter 6, these associations were replicated, and more associations were found including LDL and HDL. The evidence together illustrated the important but long-time neglected biological functions of *GIGYF1*. In order the decipher the underlying mechanisms of *GIGYF1*, I am planning to collaborate with wet lab researchers to build the *GIGYF1* KO mice models and replicate the phenotypes observed from human genetic studies. Then, comprehensive molecular and cellular experiments will be conducted to illuminate how *GIGYF1* can affect such a wide range of phenotypes, which may reveal the new biology and then guide to the novel therapeutic targets.

References

1. Ellegren, H. Sex-chromosome evolution: recent progress and the influence of male and female heterogamety. *Nat Rev Genet* **12**, 157–166 (2011).

2. Berglund, A., Stochholm, K. & Gravholt, C. H. The epidemiology of sex chromosome abnormalities. *Am J Med Genet C Semin Med Genet* **184**, 202–215 (2020).

3. Skaletsky, H. *et al.* The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**, 825–837 (2003).

4. Graves, J. A. M. Sex chromosome specialization and degeneration in mammals. *Cell* **124**, 901–914 (2006).

5. Graves, J. A. M. In retrospect: Twenty-five years of the sex-determining gene. *Nature* **528**, 343–344 (2015).

6. Sinclair, A. H. *et al.* A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. *Nature* **346**, 240–244 (1990).

7. Lahn, B. T. & Page, D. C. Functional coherence of the human Y chromosome. *Science* **278**, 675–680 (1997).

8. Maan, A. A. *et al.* The Y chromosome: a blueprint for men's health? *European Journal of Human Genetics 2017 25:11* **25**, 1181–1188 (2017).

9. Quintana-Murci, L. & Fellous, M. The human Y chromosome: the biological role of a "functional wasteland." *Journal of Biomedicine and Biotechnology* **1**, 18 (2001).

10. Lyon, M. F. Gene action in the X-chromosome of the mouse (Mus musculus L.). *Nature* **190**, 372–373 (1961).

11. Lucchesi, J. C., Kelly, W. G. & Panning, B. Chromatin remodeling in dosage compensation. *Annu Rev Genet* **39**, 615–651 (2005).

12. Gale, R. E., Wheadon, H., Boulos, P. & Linch, D. C. Tissue Specificity of X-Chromosome Inactivation Patterns. *Blood* **83**, 2899–2905 (1994).

13. Nonrandom X-Inactivation Patterns in Normal Females: Lyonization Ratios Vary With Age. *Blood* **88**, 59–65 (1996).

14. Vijg, J. & Dong, X. Pathogenic Mechanisms of Somatic Mutation and Genome Mosaicism in Aging. *Cell* **182**, 12–23 (2020).

15. Forsberg, L. A., Gisselsson, D. & Dumanski, J. P. Mosaicism in health and disease-clones picking up speed. *Nature Reviews Genetics* **18**, 128–142 (2017).

16. Thompson, D. J. *et al.* Genetic predisposition to mosaic Y chromosome loss in blood. *Nature* **575**, 652–657 (2019).

17. Loh, P. R. *et al.* Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature* **559**, 350–355 (2018).

18. Silver, A. J., Bick, A. G. & Savona, M. R. Germline risk of clonal haematopoiesis. *Nat Rev Genet* **22**, 603–617 (2021).

19. Laurie, C. C. *et al.* Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nature Genetics* **44**, 642–650 (2012).

20. Jacobs, K. B. *et al.* Detectable clonal mosaicism and its relationship to aging and cancer. *Nature Genetics* **44**, 651–658 (2012).

21. Machiela, M. J. *et al.* Characterization of large structural genetic mosaicism in human autosomes. *American Journal of Human Genetics* **96**, 487–497 (2015).

22. Vattathil, S. & Scheet, P. Extensive Hidden Genomic Mosaicism Revealed in Normal Tissue. *American Journal of Human Genetics* **98**, 571–578 (2016).

23. Loh, P. R., Genovese, G. & McCarroll, S. A. Monogenic and polygenic inheritance become instruments for clonal selection. *Nature* **584**, 136–141 (2020).

24. Bick, A. G. *et al.* Inherited causes of clonal haematopoiesis in 97,691 whole genomes. *Nature* **586**, 763–768 (2020).

25. Kar, S. P. *et al.* Genome-wide analyses of 200,453 individuals yield new insights into the causes and consequences of clonal hematopoiesis. *Nature Genetics 2022* 1–12 (2022) doi:10.1038/s41588-022-01121-z.

26. Dumanski, J. P. *et al.* Mosaic Loss of Chromosome y in Blood Is Associated with Alzheimer Disease. *American Journal of Human Genetics* **98**, 1208–1219 (2016).

27. Dumanski, J. P. *et al.* Smoking is associated with mosaic loss of chromosome Y. *Science (1979)* **347**, 81–83 (2015).

28. Forsberg, L. A. Loss of chromosome Y (LOY) in blood cells is associated with increased risk for disease and mortality in aging men. *Human Genetics* **136**, 657–663 (2017).

29. Guo, X. *et al.* Mosaic loss of human Y chromosome: what, how and why. *Human Genetics* (2020) doi:10.1007/s00439-020-02114-w.

30. Zhou, W. *et al.* Mosaic loss of chromosome Y is associated with common variation near TCL1A. *Nature Genetics* **48**, 563–568 (2016).

31. Forsberg, L. A. *et al.* Mosaic loss of chromosome y in peripheral blood is associated with shorter survival and higher risk of cancer. *Nature Genetics* **46**, 624–628 (2014).

32. Wright, D. J. *et al.* Genetic variants associated with mosaic Y chromosome loss highlight cell cycle genes and overlap with cancer susceptibility. *Nature Genetics* **49**, 674–679 (2017).

33. Terao, C. *et al.* GWAS of mosaic loss of chromosome Y highlights genetic effects on blood cell differentiation. *Nat Commun* **10**, (2019).

34. Loftfield, E. *et al.* Predictors of mosaic chromosome Y loss and associations with mortality in the UK Biobank. *Scientific Reports* **8**, 1–10 (2018).

35. MacHiela, M. J. *et al.* Female chromosome X mosaicism is age-related and preferentially affects the inactivated X chromosome. *Nat Commun* **7**, (2016).

36. Lin, S. H. *et al.* Incident disease associations with mosaic chromosomal alterations on autosomes, X and Y chromosomes: insights from a phenome-wide association study in the UK Biobank. *Cell Biosci* **11**, (2021).

37. Zekavat, S. M. *et al.* Hematopoietic mosaic chromosomal alterations increase the risk for diverse types of infection. *Nat Med* **27**, 1012–1024 (2021).

38. Zhou, W. *et al.* Detectable chromosome X mosaicism in males is rarely tolerated in peripheral leukocytes. *Scientific Reports 2021 11:1* **11**, 1–5 (2021).

39. Brown, D. W. *et al.* Shared and distinct genetic etiologies for different types of clonal hematopoiesis. (2022).

40. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: Study designs and statistical tests. *American Journal of Human Genetics* **95**, 5–23 (2014).

41. Gibson, G. Rare and common variants: twenty arguments. *Nat Rev Genet* **13**, 135–145 (2012).

42. Kryukov, G. v., Shpunt, A., Stamatoyannopoulos, J. A. & Sunyaev, S. R. Power of deep, all-exon resequencing for discovery of human trait genes. *Proc Natl Acad Sci U S A* **106**, 3871–3876 (2009).

43. Liu, J. Z., Chen, C., Tsai, E. A., Whelan, C. D. & Sexton, D. The burden of rare protein-truncating genetic variants on human lifespan. 1–18 (2020).

44. Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* **89**, 82–93 (2011).

45. Li, X. *et al.* Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nature Genetics* vol. 35 57 Preprint at https://doi.org/10.1038/s41588-020-0676-4 (2020).

46. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* **47**, D886–D894 (2019).

47. Ioannidis, N. M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet* **99**, 877–885 (2016).

48. Ng, P. C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res* **11**, 863–874 (2001).

49. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.]* **0 7**, Unit7.20 (2013).

50. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

51. Challen, G. A. & Goodell, M. A. Clonal hematopoiesis: mechanisms driving dominance of stem cell clones. *Blood* **136**, 1590–1598 (2020).

52. Dumanski, J. P., Sundström, J. & Forsberg, L. A. Loss of Chromosome y in Leukocytes and Major Cardiovascular Events. *Circulation: Cardiovascular Genetics* **10**, 1–2 (2017).

53. Skov, L. & Schierup, M. H. Analysis of 62 hybrid assembled human Y chromosomes exposes rapid structural changes and high rates of gene conversion. *PLOS Genetics* **13**, e1006834 (2017).

54. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, (2021).

55. Steensma, D. P. *et al.* Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. *Blood* **126**, 9–16 (2015).

56. Sperling, A. S., Gibson, C. J. & Ebert, B. L. The genetics of myelodysplastic syndrome: from clonal haematopoiesis to secondary leukaemia. *Nat Rev Cancer* **17**, 5–19 (2017).

57. Bao, E. L. *et al.* Inherited myeloproliferative neoplasm risk affects haematopoietic stem cells. *Nature 2020 586:7831* **586**, 769–775 (2020).

58. Titmarsh, G. J. *et al.* How common are myeloproliferative neoplasms? A systematic review and meta-analysis. *Am J Hematol* **89**, 581–587 (2014).

59. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nature Genetics 2015 47:11* **47**, 1236–1241 (2015).

60. Slob, E. A. W. & Burgess, S. A comparison of robust Mendelian randomization methods using summary data. *Genet Epidemiol* **44**, 313–329 (2020).

61. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet* **10**, (2014).

62. Turley, P. *et al.* Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat Genet* **50**, 229–237 (2018).

63. Loh, P. R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics 2015 47:3* **47**, 284–290 (2015).

64. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76–82 (2011).

65. Codd, V. *et al.* Polygenic basis and biomedical consequences of telomere length variation. *Nat Genet* **53**, 1425–1433 (2021).

66. Smith, G. D. & Ebrahim, S. "Mendelian randomization": can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* **32**, 1–22 (2003).

67. Didelez, V. & Sheehan, N. Mendelian randomization as an instrumental variable approach to causal inference. *Stat Methods Med Res* **16**, 309–330 (2007).

68. Burgess, S. & Thompson, S. G. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *Am J Epidemiol* **181**, 251–260 (2015).

69. Hemani, G., Tilling, K. & Davey Smith, G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLoS Genet* **13**, (2017).

70. Bowden, J. *et al.* Improving the accuracy of two-sample summary-data Mendelian randomization: moving beyond the NOME assumption. *Int J Epidemiol* **48**, 728–742 (2019).

71. Burgess, S. & Thompson, S. G. Interpreting findings from Mendelian randomization using the MR-Egger method. *Eur J Epidemiol* **32**, 377–389 (2017).

72. Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet Epidemiol* **40**, 304–314 (2016).

73. Hollis, B. *et al.* Genomic analysis of male puberty timing highlights shared genetic basis with hair colour and lifespan. *Nat Commun* **11**, (2020).

74. Rajman, L., Chwalek, K. & Sinclair, D. A. Therapeutic Potential of NAD-Boosting Molecules: The In Vivo Evidence. *Cell Metab* **27**, 529–547 (2018).

75. Ratcliffe, S. Long term outcome in children of sex chromosome abnormalities. *Archives of Disease in Childhood* **80**, 192–195 (1999).

76. Berglund, A. *et al.* Changes in the cohort composition of turner syndrome and severe non-diagnosis of Klinefelter, 47,XXX and 47,XYY syndrome: A nationwide cohort study 11 Medical and Health Sciences 1117 Public Health and Health Services. *Orphanet Journal of Rare Diseases* **14**, 1–9 (2019).

77. Chang, S. *et al.* Anthropometry in Klinefelter syndrome - Multifactorial influences due to CAG length, testosterone treatment and possibly intrauterine hypogonadism. *Journal of Clinical Endocrinology and Metabolism* **100**, E508–E517 (2015).

78. Skakkebæk, A. *et al.* The role of genes, intelligence, personality, and social engagement in cognitive performance in Klinefelter syndrome. *Brain and Behavior* **7**, 1–11 (2017).

79. Zöller, B., Ji, J., Sundquist, J. & Sundquist, K. High Risk of Venous Thromboembolism in Klinefelter Syndrome. *J Am Heart Assoc* **5**, 1–6 (2016).

80. Swerdlow, A. J., Higgins, C. D., Schoemaker, M. J., Wright, A. F. & Jacobs, P. A. Mortality in patients with Klinefelter syndrome in britain: A cohort study. *Journal of Clinical Endocrinology and Metabolism* **90**, 6516–6522 (2005).

81. Leggett, V., Jacobs, P., Nation, K., Scerif, G. & Bishop, D. V. M. Neurocognitive outcomes of individuals with a sex chromosome trisomy: XXX, XYY, or XXY: A systematic review. *Developmental Medicine and Child Neurology* **52**, 119–129 (2010).

82. Asano, A. *et al.* Myotonic dystrophy associated with 47 XYY syndrome. *Psychiatry and Clinical Neurosciences* **54**, 113–116 (2000).

83. Tartaglia, N. R. *et al.* Autism spectrum disorder in males with sex chromosome aneuploidy. *Journal of Developmental & Behavioral Pediatrics* **38**, 197–207 (2017).

84. Borjian Boroujeni, P. *et al.* Clinical aspects of infertile 47,XYY patients: a retrospective study. *Human Fertility* **22**, 88–93 (2019).

85. Tuke, M. A. *et al.* Mosaic Turner syndrome shows reduced penetrance in an adult population study. *Genetics in Medicine* **21**, 877–886 (2019).

86. Flaquer, A., Rappold, G. A., Wienker, T. F. & Fischer, C. The human pseudoautosomal regions: A review for genetic epidemiologists. *European Journal of Human Genetics* **16**, 771–779 (2008).

87. Szustakowski, J. D. *et al.* Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. *Nature Genetics* **53**, 942–948 (2021).

88. Yun, T. *et al.* Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics* **36**, 5582–5589 (2020).

89. Zhang, J. & Yu, K. F. What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *J Am Med Assoc* **280**, 1690–1691 (1998).

90. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants with Those of the General Population. *American Journal of Epidemiology* **186**, 1026–1034 (2017).

91. Wade, K. H. *et al.* Loss-of-function mutations in the melanocortin 4 receptor in a UK birth cohort. *Nature Medicine* **27**, 1088–1096 (2021).

92. Bojesen, A., Juul, S., Birkebæk, N. H. & Gravholt, C. H. Morbidity in Klinefelter syndrome: A Danish register study based on hospital discharge diagnoses. *Journal of Clinical Endocrinology and Metabolism* **91**, 1254–1260 (2006).

93. Berglund, A., Stochholm, K. & Gravholt, C. H. Morbidity in 47,XYY syndrome: a nationwide epidemiological study of hospital diagnoses and medication use. *Genetics in Medicine* **22**, 1542–1551 (2020).

94. Kanaka-Gantenbein, C. *et al.* Tall stature, insulin resistance, and disturbed behavior in a girl with the triple X syndrome harboring three SHOX genes: Offspring of a father with mosaic Klinefelter syndrome but with two maternal X chromosomes. *Horm Res* **61**, 205–210 (2004).

95. Lanktree, M. B., Fantus, I. G. & Hegele, R. A. Triple X syndrome in a patient with partial lipodystrophy discovered using a high-density oligonucleotide microarray: A case report. *Journal of Medical Case Reports* **3**, 1–5 (2009).

96. Van Langevelde, K., Flinterman, L. E., Vlieg, A. V. H., Rosendaal, F. R. & Cannegieter, S. C. Broadening the factor V Leiden paradox: Pulmonary embolism and deep-vein thrombosis as 2 sides of the spectrum. *Blood* **120**, 933–946 (2012).

97. Jacobs, P. A., Brunton, M., Brown, W. M. C., Doll, R. & Goldstein, H. Change of human chromosome count distribution with age: evidence for a sex differences. *Nature* **197**, 1080–1081 (1963).

98. Jacobs, P. A., Court Brown, W. M. & Doll, R. Distribution of human chromosome counts in relation to age. *Nature* **191**, 1178–1180 (1961).

99. He, L. M. *et al.* Cyclin D2 protein stability is regulated in pancreatic beta-cells. *Mol Endocrinol* **23**, 1865–1875 (2009).

100. Eastwood, S. V *et al.* Algorithms for the Capture and Adjudication of Prevalent and Incident Diabetes in UK Biobank. *PLoS One* **11**, e0162388 (2016).

101. Mahajan, A. *et al.* Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat Genet* **50**, 1505–1513 (2018).

102. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

103. Zheng, X. *et al.* SeqArray-a storage-efficient high-performance data format for WGS variant calls. *Bioinformatics* **33**, 2251–2257 (2017).

104. Zheng, X. *et al.* A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–8 (2012).

105. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, (2016).

106. Sim, N.-L. *et al.* SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res* **40**, W452-7 (2012).

107. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature 2020 581:7809* **581**, 434–443 (2020).

108. Giovannone, B. *et al.* Two novel proteins that are linked to insulin-like growth factor (IGF-I) receptors by the Grb10 adapter and modulate IGF-I signaling. *J Biol Chem* **278**, 31564–73 (2003).

109. Vujkovic, M. *et al.* Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. *Nat Genet* **52**, 680–691 (2020).

110. Steinthorsdottir, V. *et al.* Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nat Genet* **46**, 294–8 (2014).

111. Huyghe, J. R. *et al.* Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nat Genet* **45**, 197–201 (2013).

112. Flannick, J. *et al.* Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature* **570**, 71–76 (2019).

113. Dufresne, A. M. & Smith, R. J. The adapter protein GRB10 is an endogenous negative regulator of insulin-like growth factor signaling. *Endocrinology* **146**, 4399–409 (2005).

114. Holt, L. J. & Siddle, K. Grb10 and Grb14: enigmatic regulators of insulin action--and more? *Biochem J* **388**, 393–406 (2005).

115. Preston, E., Butler, K. & Haas, N. Does magnetic resonance imaging compromise integrity of the blood-brain barrier? *Neurosci Lett* **101**, 46–50 (1989).

116. Stankovic, S. *et al.* Elucidating the genetic architecture underlying IGF1 levels and its impact on genomic instability and cancer risk. *Wellcome Open Research* **6**, 20 (2021).

117. Peter, D. *et al.* GIGYF1/2 proteins use auxiliary sequences to selectively bind to 4EHP and repress target mRNA expression. *Genes Dev* **31**, 1147–1161 (2017).

118. Weber, R. *et al.* 4EHP and GIGYF1/2 Mediate Translation-Coupled Messenger RNA Decay. *Cell Rep* **33**, 108262 (2020).

119. GTEx Consortium *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

120. Forsberg, L. A. *et al.* Mosaic loss of chromosome Y in leukocytes matters. *Nat Genet* **51**, 4–7 (2019).

121. Loftfield, E. *et al.* Mosaic Y Loss Is Moderately Associated with Solid Tumor Risk. *Cancer Res* **79**, 461–466 (2019).

122. Machiela, M. J. *et al.* Mosaic chromosome Y loss and testicular germ cell tumor risk. *J Hum Genet* **62**, 637–640 (2017).

123. Dumanski, J. P. *et al.* Immune cells lacking Y chromosome show dysregulation of autosomal gene expression. *Cell Mol Life Sci* **78**, 4019–4033 (2021).

124. Zhao, Y. *et al.* GIGYF1 loss of function is associated with clonal mosaicism and adverse metabolic health. *Nat Commun* **12**, (2021).

125. Gardner, E. J. *et al.* Damaging missense variants in IGF1R implicate a role for IGF-1 resistance in the aetiology of type 2 diabetes. *medRxiv* 2022.03.26.22272972 (2022) doi:10.1101/2022.03.26.22272972.

126. Backman, J. D. *et al.* Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628–634 (2021).

127. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310–315 (2014).

128. Asada, S., Fujino, T., Goyama, S. & Kitamura, T. The role of ASXL1 in hematopoiesis and myeloid malignancies. *Cell Mol Life Sci* **76**, (2019).

129. Jiang, S. Tet2 at the interface between cancer and immunity. *Commun Biol* **3**, (2020).

130. Brunetti, L., Gundry, M. C. & Goodell, M. A. DNMT3A in Leukemia. *Cold Spring Harb Perspect Med* **7**, (2017).

131. Ruth, K. S. *et al.* Genetic insights into biological mechanisms governing human ovarian ageing. *Nature* **596**, 393–397 (2021).

132. Chiorazzi, M. *et al.* Related F-box proteins control cell death in Caenorhabditis elegans and human lymphoma. *Proc Natl Acad Sci U S A* **110**, 3943–3948 (2013).

133. Jurgens, S. J. *et al.* Analysis of rare genetic variation underlying cardiometabolic diseases and traits among 200,000 individuals in the UK Biobank. *Nat Genet* **54**, (2022).

134. Curtis, D. Analysis of rare coding variants in 200,000 exome-sequenced subjects reveals novel genetic risk factors for type 2 diabetes. *Diabetes Metab Res Rev* **38**, (2022).

135. Deaton, A. M. *et al.* Gene-level analysis of rare variants in 379,066 whole exome sequences identifies an association of GIGYF1 loss of function with type 2 diabetes. *Sci Rep* **11**, 21565 (2021).

136. Waterman, D. P., Haber, J. E. & Smolka, M. B. Checkpoint Responses to DNA Double-Strand Breaks. *Annu Rev Biochem* **89**, 103–133 (2020).

137. Matsuoka, S., Huang, M. & Elledge, S. J. Linkage of ATM to cell cycle regulation by the Chk2 protein kinase. *Science* **282**, 1893–1897 (1998).

138. Loos, R. J. F. 15 years of genome-wide association studies and no signs of slowing down. *Nature Communications 2020 11:1* **11**, 1–3 (2020).

139. Pietzner, M. *et al.* Mapping the proteo-genomic convergence of human diseases. *Science* **374**, (2021).

140. Sun, B. B. *et al.* Genetic regulation of the human plasma proteome in 54,306 UK Biobank participants. *bioRxiv* **20**, 2022.06.17.496443 (2022).

141. Ferkingstad, E. *et al.* Large-scale integration of the plasma proteome with genetics and disease. *Nature Genetics 2021 53:12* **53**, 1712–1721 (2021).

142. Yu, F. *et al.* Variant to function mapping at single-cell resolution through network propagation. *Nature Biotechnology 2022* 1–10 (2022) doi:10.1038/s41587-022-01341-y.

143. Zhang, M. J. *et al.* Polygenic enrichment distinguishes disease associations of individual cells in single-cell RNA-seq data. *bioRxiv* 2021.09.24.461597 (2022) doi:10.1101/2021.09.24.461597.

144. Stacey, D. *et al.* ProGeM: a framework for the prioritization of candidate causal genes at molecular quantitative trait loci. *Nucleic Acids Res* **47**, (2019).

145. Aragam, K. G. *et al.* Discovery and systematic characterization of risk variants and genes for coronary artery disease in over a million participants. *medRxiv* **28**, 2021.05.24.21257377 (2021).

146. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat Genet* **50**, 621–629 (2018).

147. Ardlie, K. G. *et al.* Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).

148. Võsa, U. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nature Genetics 2021 53:9* **53**, 1300–1310 (2021).

149. Qi, T. *et al.* Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. *Nat Commun* **9**, (2018).

150. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* **48**, 481–487 (2016).

151. Nasser, J. *et al.* Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–243 (2021).

152. Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M. & Ng, P. C. SIFT missense predictions for genomes. *Nat Protoc* **11**, 1–9 (2016).

153. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput Biol* **11**, (2015).

154. Weeks, E. M. *et al.* Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. *medRxiv* **23**, 2020.09.08.20190561 (2020).

Appendix A

List of publications (Equal 1st Authors - *)

Liu A*, Genovese G*, **Zhao Y***, et al. Genetic investigation of mosaic loss of the X chromosome in peripheral leukocytes of 918,085 women identifies germline predisposition and strong signals of haplotype selection. In Preparation (2022).

Stankovic S*, Shekari S*, Huang QQ*, Gardner EJ*, Owens NDL*, (5 other authors), **Zhao Y**, et al. Genetic susceptibility to earlier ovarian ageing increases de novo mutation rate in offspring. *medRxiv.* Published online June 23, 2022:2022.06.23.22276698. doi:10.1101/2022.06.23.22276698

Zhao Y, Gardner EJ, Tuke MA, et al. Detection and characterization of male sex chromosome abnormalities in the UK Biobank study. *Genetics in Medicine.* Published online June 9, 2022. doi:10.1016/J.GIM.2022.05.011

Brown DW*, Cato LD*, **Zhao Y***, et al. Shared and distinct genetic etiologies for different types of clonal hematopoiesis. *bioRxiv.* Published online March 14, 2022:2022.03.14.483644. doi:10.1101/2022.03.14.483644

Koprulu M*, **Zhao Y***, Wheeler E, et al. Identification of Rare Loss-of-Function Genetic Variation Regulating Body Fat Distribution. *The Journal of Clinical Endocrinology & Metabolism.* 2022;107(4):1065-1077. doi:10.1210/CLINEM/DGAB877

Zhao Y, Stankovic S, Koprulu M, et al. GIGYF1 loss of function is associated with clonal mosaicism and adverse metabolic health. *Nature Communications.* 2021;12(1). doi:10.1038/S41467-021-24504-Y

Stankovic S, Day FR, **Zhao Y**, et al. Elucidating the genetic architecture underlying IGF1 levels and its impact on genomic instability and cancer risk. *Wellcome Open Research* 2021 6:20. 2021;6:20. doi:10.12688/wellcomeopenres.16417.1

Appendix B

Supplementary Tables and Figures of Chapter 4

This file can be found attached to the electronic version of this thesis and can also be download from the following links.

Supplementary Table 4-1:

Description: Summary of phenome-wide disease association tests for KS and 47,XYY compared to 46,XY with each of 875 ICD-10 coded disease outcomes, from logistic regression models adjusted for age and ten principal genetic components. Outcomes reaching the multiple testing corrected statistical significance threshold (P<0.05/875=5.7x10^{-5}) are indicated in bold. (https://ars.els-cdn.com/content/image/1-s2.0-S1098360022007778-mmc2.xlsx)

Supplementary Table 4-2:

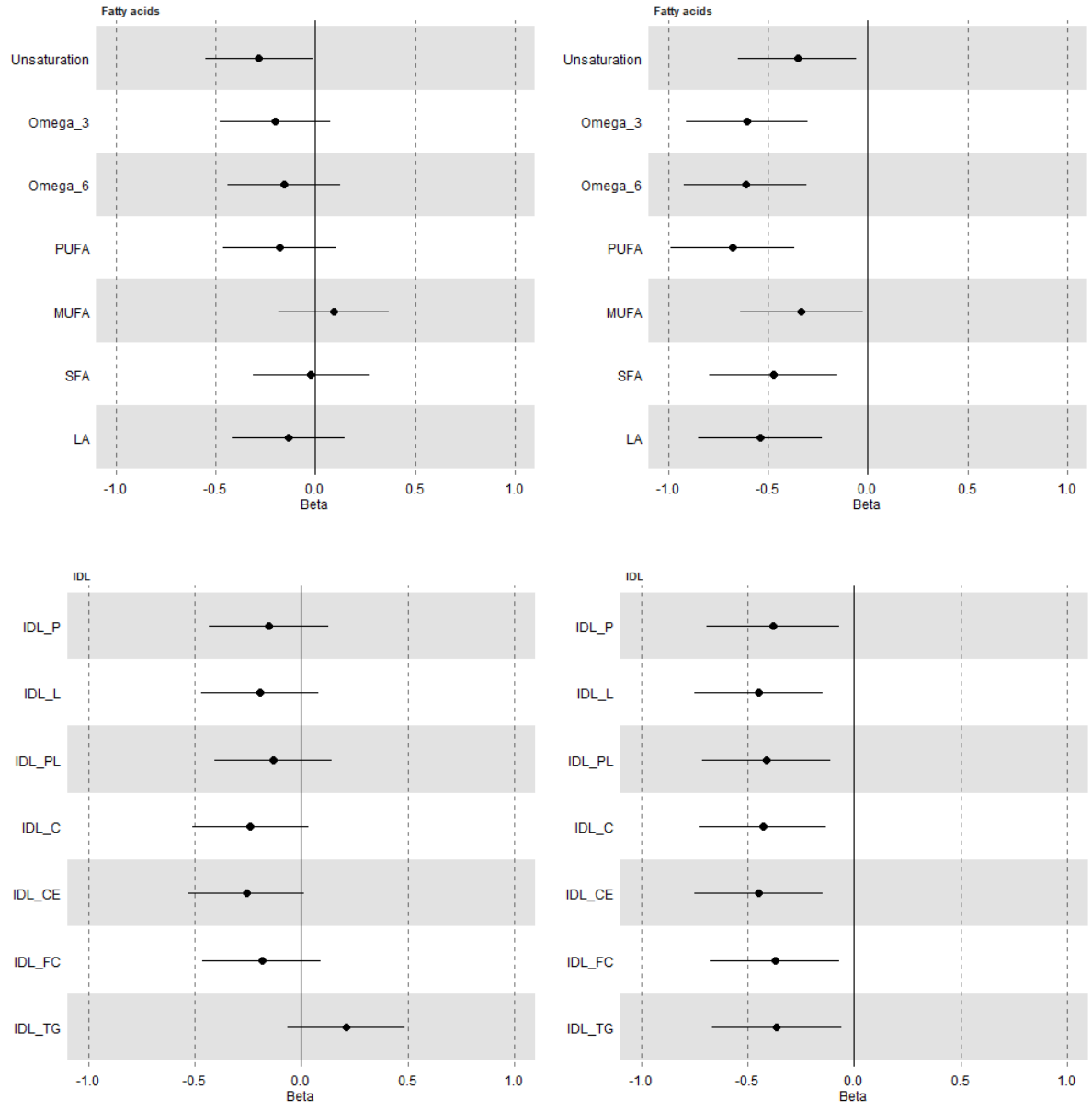
Description: Summary of association tests for KS and 47,XYY compared to 46,XY with each of 18 red blood cell or platelet traits, from linear regression models adjusted for age and ten principal genetic components. Outcomes reaching the multiple testing corrected statistical significance threshold (P<0.05/18=2.8x10^{-3}) are highlighted.

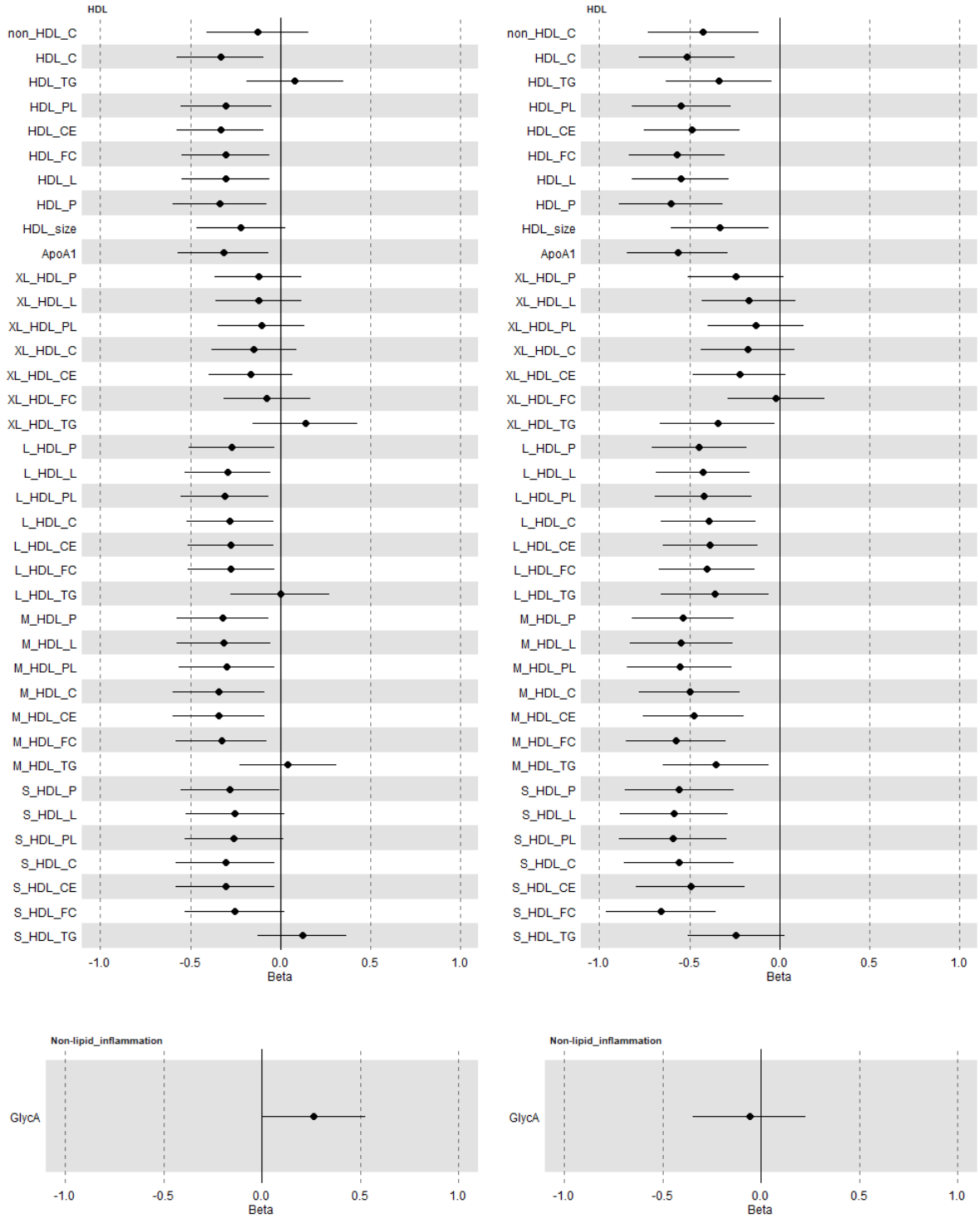
		linear regression					
trait	**karyotype**	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**	**2.5% C.I.**	**97.5% C.I.**
Red blood cells							
Haematocrit percentage	KS	-1.510	0.213	-7.084	1.41E-12	-1.928	-1.093
Haematocrit percentage	XYY	0.155	0.251	0.619	5.36E-01	-0.336	0.647
Red blood cell (erythrocyte) count	KS	-0.151	0.026	-5.766	8.13E-09	-0.202	-0.100
Red blood cell (erythrocyte) count	XYY	0.062	0.031	2.015	4.39E-02	0.002	0.122
Mean corpuscular volume	KS	-0.233	0.301	-0.774	4.39E-01	-0.822	0.357
Mean corpuscular volume	XYY	-0.852	0.354	-2.408	1.60E-02	-1.546	-0.158
Mean sphered cell volume	KS	0.231	0.372	0.621	5.34E-01	-0.498	0.961
Mean sphered cell volume	XYY	-0.048	0.438	-0.109	9.13E-01	-0.907	0.811
Red blood cell (erythrocyte) distribution width	KS	0.679	0.061	11.160	6.51E-29	0.560	0.798
Red blood cell (erythrocyte) distribution width	XYY	0.716	0.072	10.011	1.38E-23	0.576	0.857
Early RBCs							
Reticulocyte count	KS	0.000	0.003	0.074	9.41E-01	-0.006	0.006
Reticulocyte count	XYY	0.001	0.003	0.360	7.19E-01	-0.006	0.008
Reticulocyte percentage	KS	0.054	0.063	0.844	3.99E-01	-0.071	0.178
Reticulocyte percentage	XYY	0.010	0.075	0.137	8.91E-01	-0.136	0.157
Immature reticulocyte fraction	KS	0.017	0.004	4.024	5.71E-05	0.009	0.026
Immature reticulocyte fraction	XYY	0.016	0.005	3.102	1.92E-03	0.006	0.025
Mean reticulocyte volume	KS	1.510	0.547	2.759	5.81E-03	0.437	2.583
Mean reticulocyte volume	XYY	2.097	0.645	3.252	1.14E-03	0.833	3.360
Nucleated red blood cell count	KS	0.000	0.002	-0.199	8.42E-01	-0.005	0.004
Nucleated red blood cell count	XYY	-0.002	0.003	-0.638	5.24E-01	-0.007	0.003
Nucleated red blood cell percentage	KS	-0.005	0.027	-0.173	8.63E-01	-0.058	0.049
Nucleated red blood cell percentage	XYY	-0.026	0.032	-0.799	4.24E-01	-0.089	0.037
Haemoglobin							
Haemoglobin concentration	KS	-0.638	0.072	-8.831	1.04E-18	-0.780	-0.496
Haemoglobin concentration	XYY	-0.025	0.085	-0.296	7.67E-01	-0.192	0.141
Mean corpuscular haemoglobin	KS	-0.342	0.124	-2.752	5.92E-03	-0.586	-0.098
Mean corpuscular haemoglobin	XYY	-0.466	0.146	-3.185	1.45E-03	-0.752	-0.179
Mean corpuscular haemoglobin concentration	KS	-0.279	0.074	-3.766	1.66E-04	-0.425	-0.134
Mean corpuscular haemoglobin concentration	XYY	-0.183	0.087	-2.098	3.59E-02	-0.354	-0.012
Platelets							
Platelet count	KS	-9.640	3.957	-2.436	1.48E-02	-17.395	-1.884
Platelet count	XYY	-10.360	4.656	-2.225	2.61E-02	-19.485	-1.235
Platelet crit	KS	-0.006	0.003	-1.845	6.50E-02	-0.012	0.000
Platelet crit	XYY	-0.012	0.004	-3.194	1.40E-03	-0.019	-0.005
Mean platelet (thrombocyte) volume	KS	0.122	0.076	1.605	1.08E-01	-0.027	0.271
Mean platelet (thrombocyte) volume	XYY	-0.085	0.089	-0.957	3.39E-01	-0.261	0.090
Platelet distribution width	KS	0.067	0.037	1.793	7.30E-02	-0.006	0.140
Platelet distribution width	XYY	0.001	0.044	0.018	9.86E-01	-0.085	0.087

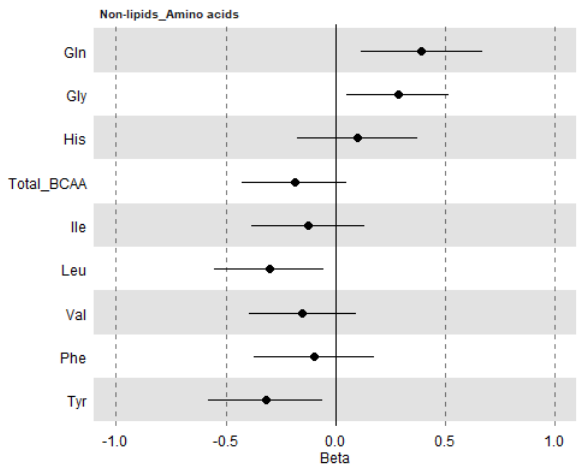
		Summary(heterozygotes)								Summary(Non-heterozygotes)							
trait	**karyotype**	**no. carriers**	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	**s.d.**	**no. non-carriers**	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	**s.d.**
Red blood cells																	
Haematocrit percentage	KS	198	30.5	39.7	41.7	41.8	44.0	58.4	3.5	201035	0.1	41.5	43.3	43.3	45.2	71.1	3.0
Haematocrit percentage	XYY	143	31.5	41.7	43.4	43.5	45.2	55.8	3.4	201035	0.1	41.5	43.3	43.3	45.2	71.1	3.0
Red blood cell (erythrocyte) count	KS	198	3.3	4.3	4.6	4.6	4.9	6.0	0.4	201035	0.0	4.5	4.7	4.7	5.0	7.8	0.4
Red blood cell (erythrocyte) count	XYY	143	3.6	4.5	4.8	4.8	5.1	5.8	0.4	201035	0.0	4.5	4.7	4.7	5.0	7.8	0.4
Mean corpuscular volume	KS	198	77.3	88.5	91.0	91.3	93.9	107.8	4.8	201034	53.5	89.0	91.6	91.6	94.2	160.3	4.3
Mean corpuscular volume	XYY	143	78.6	87.9	90.7	90.7	92.7	112.0	4.7	201034	53.5	89.0	91.6	91.6	94.2	160.3	4.3
Mean sphered cell volume	KS	197	68.2	79.6	82.3	82.9	86.2	97.6	5.4	197819	44.5	79.3	82.5	82.8	86.0	175.6	5.3
Mean sphered cell volume	XYY	142	70.9	79.5	82.5	82.6	85.5	109.6	5.4	197819	44.5	79.3	82.5	82.8	86.0	175.6	5.3
Red blood cell (erythrocyte) distribution width	KS	198	12.3	13.4	13.9	14.1	14.4	22.7	1.1	201034	11.1	12.9	13.3	13.4	13.8	38.3	0.9
Red blood cell (erythrocyte) distribution width	XYY	143	12.2	13.6	14.0	14.1	14.5	19.3	0.9	201034	11.1	12.9	13.3	13.4	13.8	38.3	0.9
Early RBCs																	
Reticulocyte count	KS	197	0.0	0.1	0.1	0.1	0.1	0.3	0.0	197819	0.0	0.0	0.1	0.1	0.1	2.4	0.0
Reticulocyte count	XYY	142	0.0	0.0	0.1	0.1	0.1	0.1	0.0	197819	0.0	0.0	0.1	0.1	0.1	2.4	0.0
Reticulocyte percentage	KS	197	0.4	1.1	1.4	1.4	1.7	6.6	0.6	197818	0.0	1.0	1.3	1.4	1.7	66.7	0.9
Reticulocyte percentage	XYY	142	0.4	1.0	1.3	1.4	1.7	2.7	0.5	197818	0.0	1.0	1.3	1.4	1.7	66.7	0.9
Immature reticulocyte fraction	KS	197	0.0	0.3	0.3	0.3	0.4	0.5	0.1	197819	0.0	0.3	0.3	0.3	0.3	0.8	0.1
Immature reticulocyte fraction	XYY	142	0.1	0.3	0.3	0.3	0.3	0.5	0.1	197819	0.0	0.3	0.3	0.3	0.3	0.8	0.1
Mean reticulocyte volume	KS	197	73.7	102.2	106.7	107.6	113.3	126.0	8.3	197819	46.9	101.8	106.2	106.3	110.9	204.5	7.8
Mean reticulocyte volume	XYY	142	83.4	103.5	108.3	108.2	113.3	132.9	7.8	197819	46.9	101.8	106.2	106.3	110.9	204.5	7.8
Nucleated red blood cell count	KS	198	0.0	0.0	0.0	0.0	0.0	0.1	0.0	200659	0.0	0.0	0.0	0.0	0.0	6.9	0.0
Nucleated red blood cell count	XYY	143	0.0	0.0	0.0	0.0	0.0	0.0	0.0	200659	0.0	0.0	0.0	0.0	0.0	6.9	0.0
Nucleated red blood cell percentage	KS	198	0.0	0.0	0.0	0.0	0.0	2.0	0.2	200657	0.0	0.0	0.0	0.0	0.0	41.2	0.4
Nucleated red blood cell percentage	XYY	143	0.0	0.0	0.0	0.0	0.0	0.0	0.0	200657	0.0	0.0	0.0	0.0	0.0	41.2	0.4
Haemoglobin																	
Haemoglobin concentration	KS	198	10.2	13.7	14.4	14.4	15.2	19.8	1.2	201034	0.1	14.4	15.0	15.0	15.7	20.5	1.0
Haemoglobin concentration	XYY	143	10.4	14.3	15.0	15.0	15.6	18.2	1.1	201034	0.1	14.4	15.0	15.0	15.7	20.5	1.0
Mean corpuscular haemoglobin	KS	198	22.5	30.4	31.3	31.4	32.4	36.8	1.7	201034	0.0	30.8	31.7	31.8	32.7	93.9	1.8
Mean corpuscular haemoglobin	XYY	143	25.6	30.3	31.2	31.3	32.2	37.7	1.7	201034	0.0	30.8	31.7	31.8	32.7	93.9	1.8
Mean corpuscular haemoglobin concentration	KS	198	29.1	33.7	34.3	34.4	35.0	36.9	1.0	201032	19.4	34.0	34.6	34.7	35.3	83.0	1.0
Mean corpuscular haemoglobin concentration	XYY	143	32.0	33.8	34.4	34.5	35.2	36.9	1.0	201032	19.4	34.0	34.6	34.7	35.3	83.0	1.0
Platelets																	
Platelet count	KS	198	62.0	193.4	225.5	229.8	262.7	425.3	54.1	201034	0.3	202.0	234.0	238.4	269.4	1488.0	55.9
Platelet count	XYY	143	106.5	190.6	224.0	228.9	261.9	445.4	58.3	201034	0.3	202.0	234.0	238.4	269.4	1488.0	55.9
Platelet crit	KS	198	0.1	0.2	0.2	0.2	0.2	0.4	0.0	201032	0.0	0.2	0.2	0.2	0.2	1.2	0.0
Platelet crit	XYY	143	0.1	0.2	0.2	0.2	0.2	0.4	0.0	201032	0.0	0.2	0.2	0.2	0.2	1.2	0.0
Mean platelet (thrombocyte) volume	KS	198	6.9	8.7	9.3	9.4	10.1	14.4	1.1	201032	5.7	8.5	9.2	9.3	9.9	15.8	1.1
Mean platelet (thrombocyte) volume	XYY	143	6.9	8.5	9.1	9.2	9.8	12.5	1.1	201032	5.7	8.5	9.2	9.3	9.9	15.8	1.1
Platelet distribution width	KS	198	15.4	16.2	16.6	16.6	16.9	18.7	0.6	201032	13.9	16.2	16.5	16.6	16.9	20.0	0.5
Platelet distribution width	XYY	143	15.6	16.2	16.5	16.6	16.9	17.9	0.5	201032	13.9	16.2	16.5	16.6	16.9	20.0	0.5

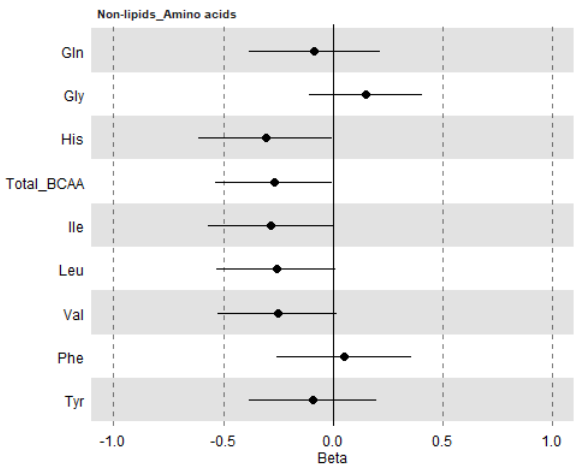
Supplementary Figure 4-1:

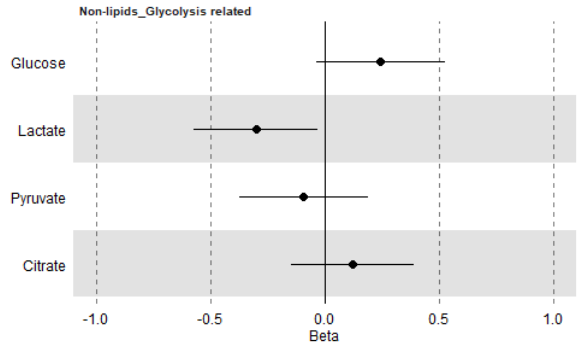
Description: Forest plots showing mean differences (and 95% CI) in plasma metabolic traits measured by nuclear magnetic resonance (NMR) spectroscopy in men with Klinefelter syndrome (47,XXY, KS, left) and 47,XYY (right) compared to men with normal (46,XY) karyotypes.

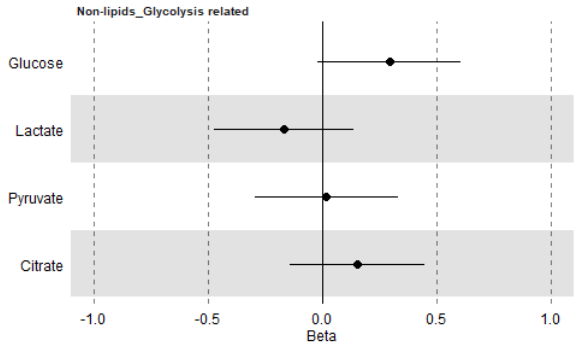


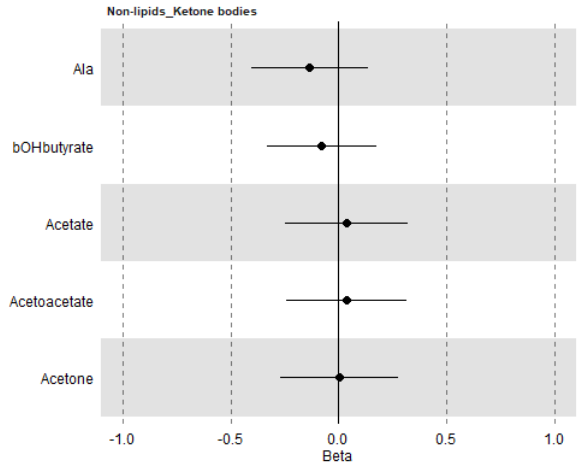

HDL
non_HDL_C
HDL_C
HDL_TG
HDL_PL
HDL_CE
HDL_FC
HDL_L
HDL_P
HDL_size
ApoA1
XL_HDL_P
XL_HDL_L
XL_HDL_PL
XL_HDL_C
XL_HDL_CE
XL_HDL_FC
XL_HDL_TG
L_HDL_P
L_HDL_L
L_HDL_PL
L_HDL_C
L_HDL_CE
L_HDL_FC
L_HDL_TG
M_HDL_P
M_HDL_L
M_HDL_PL
M_HDL_C
M_HDL_CE
M_HDL_FC
M_HDL_TG
S_HDL_P
S_HDL_L
S_HDL_PL
S_HDL_C
S_HDL_CE
S_HDL_FC
S_HDL_TG
-1.0
-0.5
0.0
0.5
1.0
Beta
HDL
non_HDL_C
HDL_C
HDL_TG
HDL_PL
HDL_CE
HDL_FC
HDL_L
HDL_P
HDL_size
ApoA1
XL_HDL_P
XL_HDL_L
XL_HDL_PL
XL_HDL_C
XL_HDL_CE
XL_HDL_FC
XL_HDL_TG
L_HDL_P
L_HDL_L
L_HDL_PL
L_HDL_C
L_HDL_CE
L_HDL_FC
L_HDL_TG
M_HDL_P
M_HDL_L
M_HDL_PL
M_HDL_C
M_HDL_CE
M_HDL_FC
M_HDL_TG
S_HDL_P
S_HDL_L
S_HDL_PL
S_HDL_C
S_HDL_CE
S_HDL_FC
S_HDL_TG
-1.0
-0.5
0.0
0.5
1.0
Beta
Non-lipid_inflammation
GlycA
-1.0
-0.5
0.0
0.5
1.0
Beta
Non-lipid_inflammation
GlycA
-1.0
-0.5
0.0
0.5
1.0
Beta

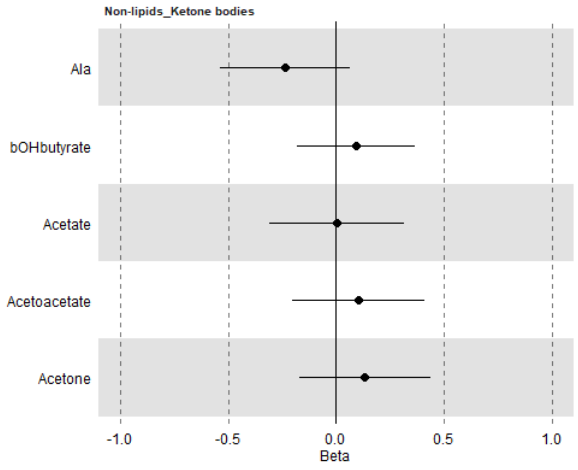

Non-lipids_Amino acids
Gln
Gly
His
Total_BCAA
Ile
Leu
Val
Phe
Tyr
-1.0
-0.5
0.0
0.5
1.0
Beta

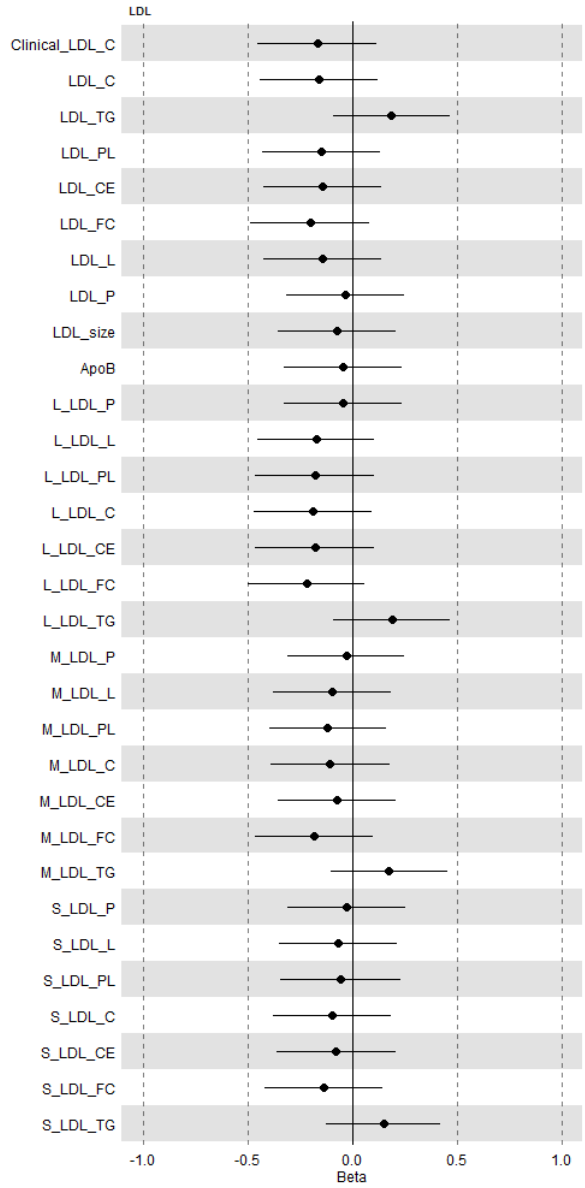

Non-lipids_Amino acids
Gln
Gly
His
Total_BCAA
Ile
Leu
Val
Phe
Tyr
-1.0
-0.5
0.0
0.5
1.0
Beta

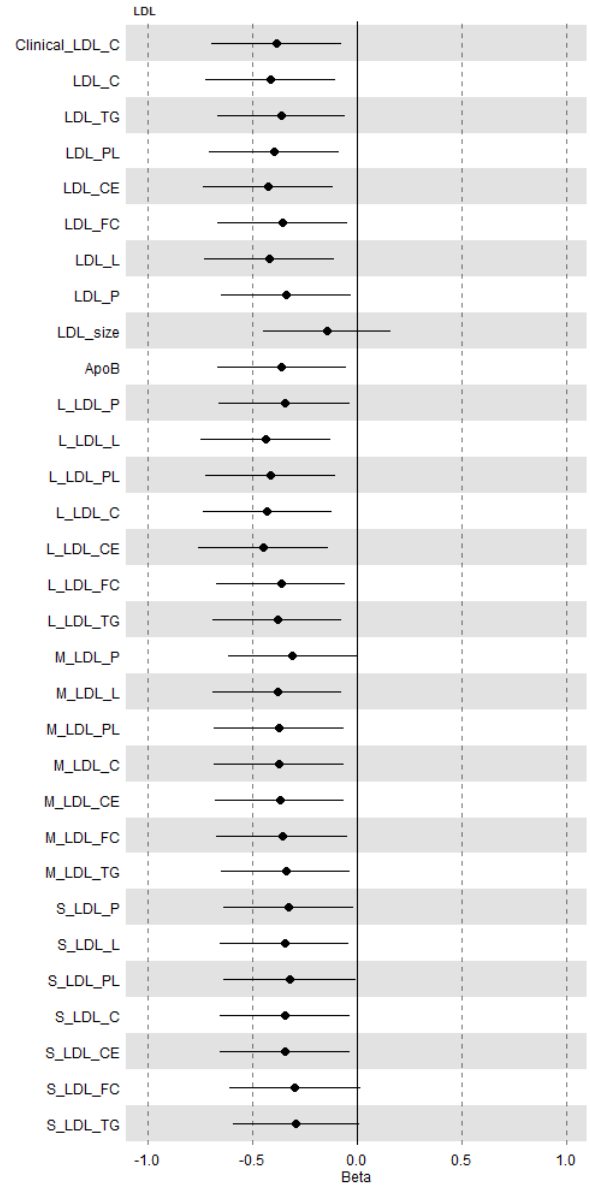

Non-lipids_Glycolysis related
Glucose
Lactate
Pyruvate
Citrate
-1.0
-0.5
0.0
0.5
1.0
Beta

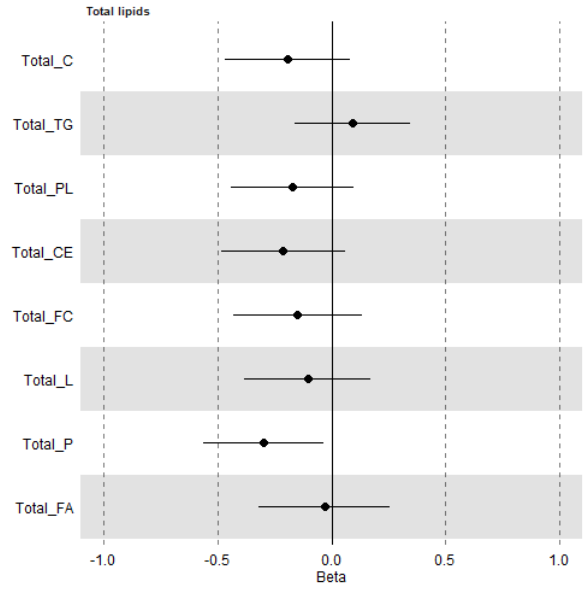

Non-lipids_Glycolysis related
Glucose
Lactate
Pyruvate
Citrate
-1.0
-0.5
0.0
0.5
1.0
Beta

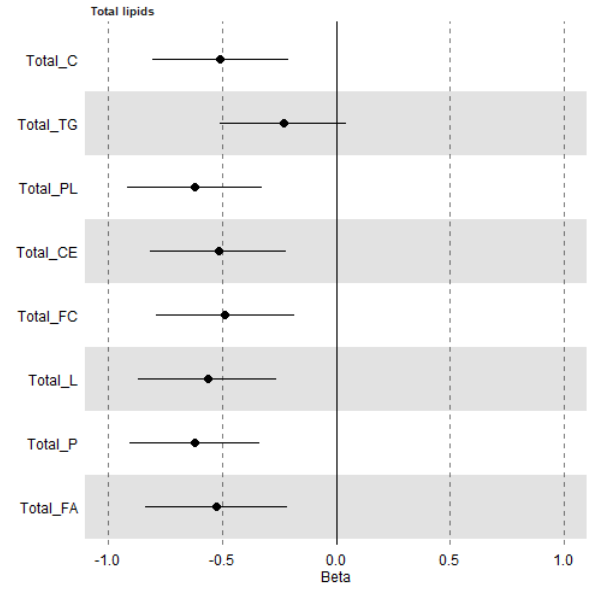

Non-lipids_Ketone bodies
Ala
bOHbutyrate
Acetate
Acetoacetate
Acetone
-1.0
-0.5
0.0
0.5
1.0
Beta

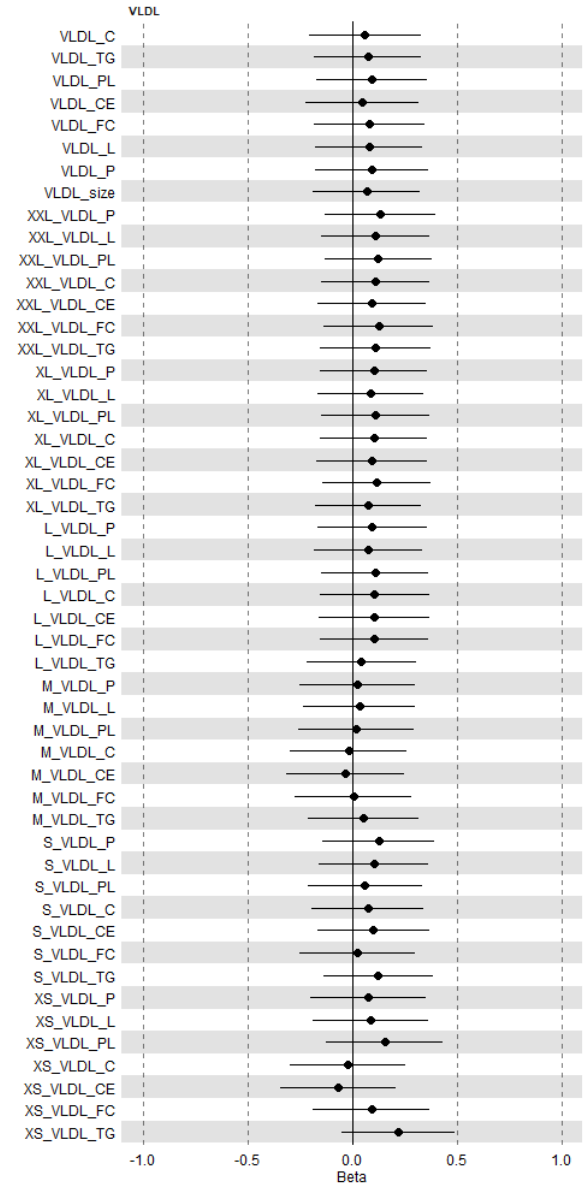

Non-lipids_Ketone bodies
Ala
bOHbutyrate
Acetate
Acetoacetate
Acetone
-1.0
-0.5
0.0
0.5
1.0
Beta

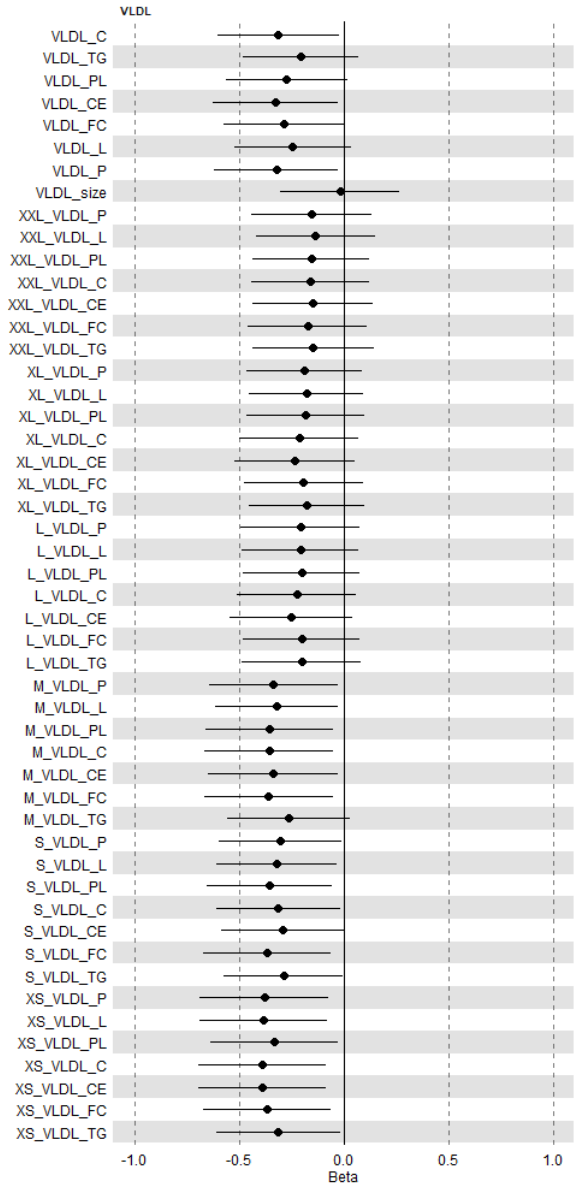

LDL
Clinical_LDL_C
LDL_C
LDL_TG
LDL_PL
LDL_CE
LDL_FC
LDL_L
LDL_P
LDL_size
ApoB
L_LDL_P
L_LDL_L
L_LDL_PL
L_LDL_C
L_LDL_CE
L_LDL_FC
L_LDL_TG
M_LDL_P
M_LDL_L
M_LDL_PL
M_LDL_C
M_LDL_CE
M_LDL_FC
M_LDL_TG
S_LDL_P
S_LDL_L
S_LDL_PL
S_LDL_C
S_LDL_CE
S_LDL_FC
S_LDL_TG
-1.0
-0.5
0.0
0.5
1.0
Beta

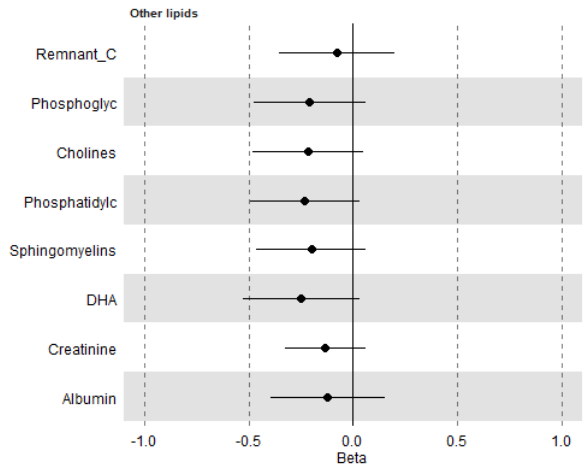

LDL
Clinical_LDL_C
LDL_C
LDL_TG
LDL_PL
LDL_CE
LDL_FC
LDL_L
LDL_P
LDL_size
ApoB
L_LDL_P
L_LDL_L
L_LDL_PL
L_LDL_C
L_LDL_CE
L_LDL_FC
L_LDL_TG
M_LDL_P
M_LDL_L
M_LDL_PL
M_LDL_C
M_LDL_CE
M_LDL_FC
M_LDL_TG
S_LDL_P
S_LDL_L
S_LDL_PL
S_LDL_C
S_LDL_CE
S_LDL_FC
S_LDL_TG
-1.0
-0.5
0.0
0.5
1.0
Beta

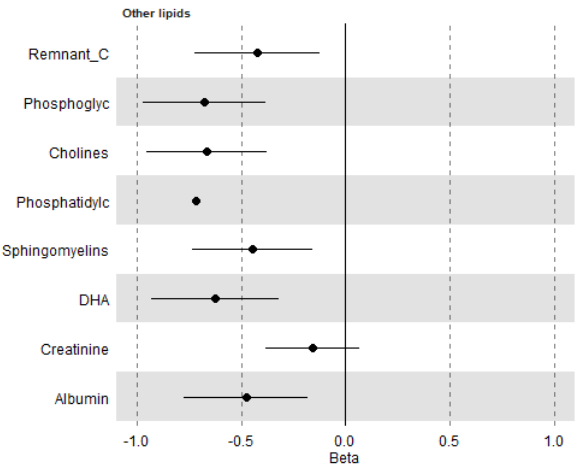

Total lipids
Total_C
Total_TG
Total_PL
Total_CE
Total_FC
Total_L
Total_P
Total_FA
-1.0
-0.5
0.0
0.5
1.0
Beta


Total lipids
Total_C
Total_TG
Total_PL
Total_CE
Total_FC
Total_L
Total_P
Total_FA
-1.0
-0.5
0.0
0.5
1.0
Beta


VLDL
VLDL_C
VLDL_TG
VLDL_PL
VLDL_CE
VLDL_FC
VLDL_L
VLDL_P
VLDL_size
XXL_VLDL_P
XXL_VLDL_L
XXL_VLDL_PL
XXL_VLDL_C
XXL_VLDL_CE
XXL_VLDL_FC
XXL_VLDL_TG
XL_VLDL_P
XL_VLDL_L
XL_VLDL_PL
XL_VLDL_C
XL_VLDL_CE
XL_VLDL_FC
XL_VLDL_TG
L_VLDL_P
L_VLDL_L
L_VLDL_PL
L_VLDL_C
L_VLDL_CE
L_VLDL_FC
L_VLDL_TG
M_VLDL_P
M_VLDL_L
M_VLDL_PL
M_VLDL_C
M_VLDL_CE
M_VLDL_FC
M_VLDL_TG
S_VLDL_P
S_VLDL_L
S_VLDL_PL
S_VLDL_C
S_VLDL_CE
S_VLDL_FC
S_VLDL_TG
XS_VLDL_P
XS_VLDL_L
XS_VLDL_PL
XS_VLDL_C
XS_VLDL_CE
XS_VLDL_FC
XS_VLDL_TG
-1.0
-0.5
0.0
0.5
1.0
Beta


VLDL
VLDL_C
VLDL_TG
VLDL_PL
VLDL_CE
VLDL_FC
VLDL_L
VLDL_P
VLDL_size
XXL_VLDL_P
XXL_VLDL_L
XXL_VLDL_PL
XXL_VLDL_C
XXL_VLDL_CE
XXL_VLDL_FC
XXL_VLDL_TG
XL_VLDL_P
XL_VLDL_L
XL_VLDL_PL
XL_VLDL_C
XL_VLDL_CE
XL_VLDL_FC
XL_VLDL_TG
L_VLDL_P
L_VLDL_L
L_VLDL_PL
L_VLDL_C
L_VLDL_CE
L_VLDL_FC
L_VLDL_TG
M_VLDL_P
M_VLDL_L
M_VLDL_PL
M_VLDL_C
M_VLDL_CE
M_VLDL_FC
M_VLDL_TG
S_VLDL_P
S_VLDL_L
S_VLDL_PL
S_VLDL_C
S_VLDL_CE
S_VLDL_FC
S_VLDL_TG
XS_VLDL_P
XS_VLDL_L
XS_VLDL_PL
XS_VLDL_C
XS_VLDL_CE
XS_VLDL_FC
XS_VLDL_TG
-1.0
-0.5
0.0
0.5
1.0
Beta


Other lipids
Remnant_C
Phosphoglyc
Cholines
Phosphatidylc
Sphingomyelins
DHA
Creatinine
Albumin
-1.0
-0.5
0.0
0.5
1.0
Beta


Other lipids
Remnant_C
Phosphoglyc
Cholines
Phosphatidylc
Sphingomyelins
DHA
Creatinine
Albumin
-1.0
-0.5
0.0
0.5
1.0
Beta

Appendix C

Supplementary Tables of Chapter 5

These files can be found attached to the electronic version of this thesis and can also be download from the following links.

Supplementary Table 5-1:

Description: A comparison of PAR-LOY vs PAR-LOYq association statistics for the previously reported LOY signals (https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-021-24504-y/MediaObjects/41467_2021_24504_MOESM3_ESM.xlsx)

Supplementary Table 5-2:

Description: Exome-wide gene-burden association test statistics for LOY (https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-021-24504-y/MediaObjects/41467_2021_24504_MOESM4_ESM.xlsx)

Supplementary Table 5-3:

Description: Moderate and high impact coding variants identified in *GIGYF1* (https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-021-24504-y/MediaObjects/41467_2021_24504_MOESM5_ESM.xlsx)

Supplementary Table 5-4:

Description: The impact of CADD variant weighting using STAAR for LOY gene burden testing. (https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-021-24504-y/MediaObjects/41467_2021_24504_MOESM7_ESM.xlsx)

Supplementary Table 5-5:

Description: Phenotypic characteristics of *GIGYF1* loss of function carriers.

Variant carried by individual	Number of rare alleles	sex	T2D	T2D age (group) of onset	BMI category (baseline)	HbA1C category (at baseline(mmol/mol))
chr7_100684236_C_T	1	Male	1	30-39	30-35	>48
chr7_100686356_G_A	1	Male	1	40-49	30-35	>48
chr7_100682120_CA_C	1	Male	1	60-69	<25	>48
chr7_100682387_TG_T	1	Male	1	50-59	25-30	42-48
chr7_100683218_C_A	1	Female	1	60-69	25-30	>48
chr7_100683374_TTCTCC_T	1	Male	1	40-49	25-30	<42
chr7_100685054_G_A	1	Female	1	40-49	30-35	<42
chr7_100686033_TG_T	1	Female	1	70-79	35+	<42
chr7_100686749_C_T	1	Male	1	50-59	<25	42-48
chr7_100687045_C_CT	1	Male	1	50-59	25-30	42-48
chr7_100687297_C_T	1	Male	1	50-59	35+	>48
chr7_100687323_G_A	1	Male	1	60-69	35+	42-48
chr7_100687357_G_A	1	Female	1	60-69	35+	>48
chr7_100687545_CA_C	1	Male	1	40-49	30-35	>48
chr7_100687545_CA_C	1	Male	1	50-59	35+	42-48
chr7_100687545_CA_C	1	Male	1	50-59	30-35	>48
chr7_100687545_CA_C	1	Male	1	40-49	30-35	>48
chr7_100687545_CA_C	1	Female	1	60-69	30-35	>48
chr7_100687546_A_AG	1	Male	1	50-59	35+	42-48
chr7_100688238_T_C	1	Male	1	50-59	30-35	<42
chr7_100684338_C_G, chr7_100684339_T_G	2	Female	1	70-79	30-35	42-48
chr7_100681994_C_T	1	Female	0	-	25-30	<42
chr7_100682071_C_T	1	Female	0	-	25-30	<42
chr7_100682071_C_T	1	Female	0	-	30-35	<42
chr7_100682198_G_GGA	1	Male	0	-	25-30	<42
chr7_100682484_T_C	1	Male	0	-	25-30	<42
chr7_100682700_CTT_C	1	Male	0	-	<25	<42
chr7_100682700_CTT_C	1	Female	0	-	<25	<42
chr7_100682749_A_AG	1	Male	0	-	25-30	<42
chr7_100682749_A_AG	1	Female	0	-	<25	<42
chr7_100683017_G_A	1	Male	0	-	25-30	<42
chr7_100683017_G_A	1	Male	0	-	25-30	>48
chr7_100683017_G_A	1	Female	0	-	<25	<42
chr7_100683017_G_A	1	Female	0	-	30-35	<42
chr7_100683112_A_ATGGACAG CCCCTGCTTGGCC	1	Male	0	-	<25	<42
chr7_100683122_CCT_C	1	Male	0	-	25-30	<42
chr7_100683231_C_T	1	Male	0	-	30-35	<42
chr7_100683303_C_T	1	Female	0	-	25-30	<42
chr7_100683366_G_A	1	Male	0	-	30-35	<42
chr7_100683548_A_C	1	Female	0	-	35+	<42
chr7_100683585_CCT_C	1	Female	0	-	<25	<42
chr7_100685129_G_A	1	Male	0	-	35+	<42
chr7_100686182_TGA_T	1	Male	0	-	25-30	<42

chr7_100686365_G_A	1	Male	0	-	25-30	<42
chr7_100686817_G_A	1	Male	0	-	30-35	<42
chr7_100687357_G_A	1	Male	0	-	25-30	<42
chr7_100687357_G_A	1	Male	0	-	25-30	<42
chr7_100687357_G_A	1	Female	0	-	30-35	<42
chr7_100687357_G_A	1	Female	0	-	25-30	<42
chr7_100687408_T_C	1	Male	0	-	30-35	42-48
chr7_100687532_G_A	1	Male	0	-	25-30	<42
chr7_100687532_G_A	1	Male	0	-	<25	<42
chr7_100687545_CA_C	1	Male	0	-	25-30	<42
chr7_100687545_CA_C	1	Female	0	-	<25	<42
chr7_100687545_CA_C	1	Female	0	-	35+	>48
chr7_100687545_CA_C	1	Male	0	-	25-30	<42
chr7_100687545_CA_C	1	Female	0	-	30-35	<42
chr7_100687545_CA_C	1	Male	0	-	<25	<42
chr7_100687545_CA_C	1	Male	0	-	<25	<42
chr7_100687545_CA_C	1	Male	0	-	30-35	<42
chr7_100687545_CA_C	1	Female	0	-	<25	<42
chr7_100687545_CA_C	1	Male	0	-	25-30	<42
chr7_100687546_A_AG	1	Male	0	-	30-35	<42
chr7_100688225_GTC_G	1	Female	0	-	25-30	42-48
chr7_100687545_CA_C	1	Male	0	-	25-30	<42

Variant carried by individual	Self-reported insulin use	Self-reported use of biguanide drugs	Self-reported use of other oral antidiabetic drugs	Insulin within 1 year	Family history of diabetes
chr7_100684236_C_T	1	-	-	1	0
chr7_100686356_G_A	1	-	-	1	0
chr7_100682120_CA_C	-	1	1	0	0
chr7_100682387_TG_T	-	-	-	-	0
chr7_100683218_C_A	-	1	-	0	0
chr7_100683374_TTCTCC_T	-	1	-	0	1
chr7_100685054_G_A	-	-	-	0	1
chr7_100686033_TG_T	-	-	-	-	1
chr7_100686749_C_T	-	1	1	0	0
chr7_100687045_C_CT	-	-	1	0	1
chr7_100687297_C_T	-	-	-	0	1
chr7_100687323_G_A	-	-	-	-	0
chr7_100687357_G_A	-	1	-	0	0
chr7_100687545_CA_C	-	1	1	0	1
chr7_100687545_CA_C	-	1	1	0	0
chr7_100687545_CA_C	-	-	-	-	0
chr7_100687545_CA_C	1	1	1	0	0
chr7_100687545_CA_C	-	1	-	0	0
chr7_100687546_A_AG	-	1	1	0	0
chr7_100688238_T_C	-	-	-	-	1
chr7_100684338_C_G, chr7_100684339_T_G	-	-	-	-	0

chr7_100681994_C_T	-	-	-	-	1
chr7_100682071_C_T	-	-	-	-	0
chr7_100682071_C_T	-	-	-	-	0
chr7_100682198_G_GGA	-	-	-	-	0
chr7_100682484_T_C	-	-	-	-	0
chr7_100682700_CTT_C	-	-	-	-	0
chr7_100682700_CTT_C	-	-	-	-	0
chr7_100682749_A_AG	-	-	-	-	0
chr7_100682749_A_AG	-	-	-	-	1
chr7_100683017_G_A	-	-	-	-	0
chr7_100683017_G_A	-	-	-	-	0
chr7_100683017_G_A	-	-	-	-	1
chr7_100683017_G_A	-	-	-	-	0
chr7_100683112_A_ATGGACAGCCCCTGCTTGGCC	-	-	-	-	0
chr7_100683122_CCT_C	-	-	-	-	0
chr7_100683231_C_T	-	-	-	-	0
chr7_100683303_C_T	-	-	-	-	0
chr7_100683366_G_A	-	-	-	-	0
chr7_100683548_A_C	-	-	-	-	0
chr7_100683585_CCT_C	-	-	-	-	0
chr7_100685129_G_A	-	-	-	-	0
chr7_100686182_TGA_T	-	-	-	-	0
chr7_100686365_G_A	-	-	-	-	0
chr7_100686817_G_A	-	-	-	-	1
chr7_100687357_G_A	-	-	-	-	0
chr7_100687357_G_A	-	-	-	-	1
chr7_100687357_G_A	-	-	-	-	1
chr7_100687357_G_A	-	-	-	-	1
chr7_100687408_T_C	-	-	-	-	0
chr7_100687532_G_A	-	-	-	-	0
chr7_100687532_G_A	-	-	-	-	0
chr7_100687545_CA_C	-	-	-	-	0
chr7_100687545_CA_C	-	-	-	-	0
chr7_100687545_CA_C	-	-	-	-	0
chr7_100687545_CA_C	-	-	-	-	0
chr7_100687545_CA_C	-	-	-	-	0
chr7_100687545_CA_C	-	-	-	-	0
chr7_100687545_CA_C	-	-	-	-	1
chr7_100687545_CA_C	-	-	-	-	0
chr7_100687545_CA_C	-	-	-	-	0
chr7_100687545_CA_C	-	-	-	-	0
chr7_100687546_A_AG	-	-	-	-	0
chr7_100688225_GTC_G	-	-	-	-	0
chr7_100687545_CA_C	-	-	-	-	0