Characterising the genetic architecture of common elastic tissue disorders using UK Biobank

A thesis submitted to the University of Oxford in partial fulfilment of the requirements for the degree of

Master of Science (by Research)

Molecular and Cellular Medicine

Waheed-Ul-Rahman Ahmed
Lady Margaret Hall
NDORMS
Trinity Term 2020

Abstract

Background

Elastic fibres are key extracellular matrix (ECM) components that provide elasticity to tissues throughout the body, allowing them to recover after deformation. Derangement of elastic fibres lead to cutaneous and systemic disorders that show elastic tissue pathology. Common elastic tissue disorders such as varicose veins, haemorrhoids, and hernia have a complex aetiology where genetic predisposition and environmental factors interplay to influence overall phenotypic expression. These elastic diseases are highly prevalent and exert a significant healthcare and socioeconomic burden. However, their genetic basis remains poorly defined, with only a handful of putative genes identified.

Method

To unravel the genetic architecture of varicose veins, haemorrhoids and hernia, genome-wide association study (GWAS) was performed using the UK Biobank resource, a prospective cohort of 500,000 whole genome genotyped individuals, with replication from ~410,000 individuals from 23andMe, Inc. (California) for the varicose veins analysis. Genes and pathways were prioritised using a suite of bioinformatic approaches, and pharmacological targets identified using the Open Targets Platform. A genetic risk score was constructed for each of the three elastic tissue disorders to examine the genetic burden among patients with severe disease. For the hernia analysis, multi-trait and multivariate meta-analysis approaches were deployed to uncover the shared genetic susceptibility to multiple hernia phenotypes.

Results

Performing the largest two-stage GWAS of varicose veins in 810,625 individuals, forty-nine signals at 46 susceptibility loci were discovered, including 29 previously unreported associations. Next, through the first-ever GWA study of haemorrhoids in over 400,000 individuals, 13 signals at 12 novel loci were discovered to associate with haemorrhoids. Lastly, association analysis of inguinal, femoral, umbilical, and hiatus hernia individually yielded 58 signals at 38 loci (34 new) associated with the four hernia phenotypes. When combined in a multi-trait meta-analysis, 12 biologically relevant putative loci were discovered to associate with multiple hernia phenotypes, demonstrating novel and robust evidence of shared susceptibility to hernia. Collectively, a total of 114 independent associated signals at 86 susceptibility loci were discovered to associate with varicose veins, haemorrhoids and hernia, which were mapped to over 400 putative genes, many of which demonstrated evidence of therapeutic tractability and clustered in pathways pertaining to ECM homeostasis. The last and most compelling discovery was that the genetic risk score correlated with disease severity for all three disorders, with patients undergoing surgery having a higher genetic burden than those managed non-surgically.

Conclusion

Prioritised genes and pathways demonstrate significant biological plausibility, and represent promising candidates for further investigation of elastic tissue biology and potential pharmacological targeting. The genetic risk score correlated with disease across all disorders, representing an important proof-of-principle for the future use of genetic risk scoring in personalised medicine approaches to surgical disorders. To this

end, this thesis represents an advance at the nexus of surgical research and complex trait genetics.

Dedicated to my late grandfather

Master Muhammad Yunus Bhatti
(1924-1990)
إِنَّا لِلّهِ وَإِنَّـا إِلَيْهِ رَاجِعُونَ

whose remarkable focus on the importance of education as a teacher in my old country of Kotli, Azad Kashmir, inspired my journey to Oxford

Acknowledgements

الْحَمْدُ لِلَّهِ الَّذِي بِنِعْمَتِهِ تَتِمُّ الصَّالِحَاتُ

All praise is due to Allah (God), by whose honour and Majesty, deeds of virtue are accomplished.

I am deeply grateful to Allah for affording me this extraordinary opportunity to study at the University of Oxford – to apply myself and to learn without bounds, such that I may be able to one day help patients that are devastatingly affected by the diseases that I have been researching.

I am indebted to my Father, Manzoor Ahmad Bhatti, and Mother, Robina Kausar Bhatti for supporting me endlessly through all the difficulties they have had to go through migrating from their family and home in Kashmir, in the hope of better circumstances in England. Their struggles growing up has been instrumental in defining the person I am today, and continue to aspire towards. Most of all, my parents taught me the value of hard work, staying grounded, helping others, and the importance of prayer and gratitude. Coming to Oxford has been a dream beyond my wildest imagination, which I shall never take for granted, and would not have been possible without my parents' sacrifices.

For their inexorable mentorship and guidance over the past year, I am deeply thankful to my supervisors Professor Dominic Furniss, Mr Akira Wiberg, and Professor Krina Zondervan. I learnt a lot under their tutelage, and they believed in me and gave me more time than two MSc students would ever warrant. I have developed into a better scientist, medical student, and human-being because of their generosity, mentorship and wealth of knowledge.

It is easy to follow the path well-trodden, and hence I consider myself deeply fortunate to have been within a minute's walk from Dr Michael Ng and Mr Akira Wiberg. Their extensive experience of GWAS proved invaluable at many important junctures within the MSc. Thank you for your patience whilst answering all my questions, for bouncing ideas, and putting up with all my late-night/ early-morning emails!

Deserving special mention is Mr David Izadi who took a lowly 1st year medical student under his wings and guided him. I am grateful to have met David early in my medical career; his care towards his patients, and honesty and integrity as a surgeon is something I will always aspire towards. In addition, I am privileged to have been mentored by Dr Conor Jones, Mr James Chan, and Dr Zahid Pranjol whose guidance had a profound impact on my formative development at medical school. Lastly and importantly, I am indebted to Mr James Glasbey and Mr Kenneth McLean whose guidance and mentorship over the years within the STARSurg Collaborative has cultivated my leadership and research skills.

I have enjoyed every minute of my time in Oxford and learned more than I knew possible and made countless memories. I am thankful to many great friends for their support this year: Hamez Gacaferi, Conrad Harrison, Kari Shah, Akmal Ali, Paijaan Majid, Temidayo Osunorobi, Osaid Alser, Omar Kouli, Hasan Sukkar, TJ Ashaye, Matthew Saint, Dhaniyyah Hishamuddin, Awais Malik, Muzzaffer Hussain, and Saira Hassan. Thank you for your friendship!

Preface

This thesis is a collection of the research work carried out by myself in the Botnar Research Centre, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, between October 2019 and September 2020, under the supervision of Professor Dominic Furniss, Professor Krina Zondervan, and Mr Akira Wiberg.

Every piece of work that is described in this thesis was conducted solely by me, with the exception of the following:

- Mr Akira Wiberg performed the quality control (QC) of the UK Biobank resource that was used in all chapters, and ran the initial association analysis for varicose veins (Chapter 2).
- Dr Michael Ng performed the case-control matching in Chapter 4 and developed the codes to run the metaUSAT analysis (Chapter 4).

This work presented in Chapter 2 is part of a manuscript titled *'Genome-wide association analysis and replication in 810,625 individuals identifies novel therapeutic targets for varicose veins' (doi: 10.1101/2020.05.14.095653).* Chapters 3 and 4 are both being prepared in manuscript form for submission to a journal.

I state the research described herein is original, though the work of others is cited with appropriate references in-text. No part of this research has previously been submitted for the conferral of a degree outside of the University of Oxford.

Waheed-Ul-Rahman Ahmed

Sep 2020

Abbreviations

AWH	Abdominal Wall Hernia
BMI	Body Mass Index
CADD	Combined Annotation Dependent Depletion
CAG	Constituent Glycosaminoglycan
CAIS	Complete Androgen Insensitivity Syndrome
CEAP	Clinical-Etiology-Anatomy-Pathophysiology
CL	Cutis Laxa
CVD	Chronic Venous Disease
DG-HAL	Doppler-Guided Haemorrhoidal Artery Ligation
DNA	Deoxyribonucleic Acid
EBI	European Bioinformatics Institute
ECM	Extracellular Matrix
EDS	Ehlers-Danlos Syndrome
EFO	Experiment Factor Ontology
eQTL	Expression Quantitative Trait Loci
ERK	Extracellular Signal Regulated Kinase
FDR	False Discovery Rate
FUMA	Functional Mapping And Annotation
GERP	Genomic Evolutionary Rate Profiling
GO	Gene Ontology

GORD	Gastro-Oesophageal Reflux Disease
GWAS	Genome-Wide Association Study
GWGAS	Genome-Wide, Gene Association Analysis
HEIDI	Heterogeneity In Dependent Instruments Test
HGNC	Human Gene Nomenclature Committee
HLA	Human Leukocyte Antigen
HWE	Hardy-Weinberg Equillibrium
IAS	Internal Anal Sphincter
IBD	Identity-By-Descent
ICD	International Classification Of Diseases
IndSigSNPs	Independent Significant Snps
kb	Kilobases
LD	Linkage Disequilibrium
LDS4	Loey-Dietz Syndrome Type 4
LDSC	Linkage Disequilibrium Score
LMM	Linear Mixed Model
Mb	Megabase
metaUSAT	Meta-Analysis Unified Score-Based Association Test
MFS	Marfan Syndrome
MMPs	Matrix Metalloproteinases
MTAG	Multi-Trait Analysis Of GWAS
NHGRI	National Human Genome Research Institute
NHS	National Health Service

OPCS	Office Of Population Censuses And Surveys Classification Of Interventions And Procedures
OR	Odds Ratio
PolyPhen	Polymorphism Phenotyping
PCA	Principal Component Analysis
PXE	Pseudoxanthoma Elasticum
QC	Quality Control
Q-Q	Quantile-Quantile
RBL	Rubber Band Ligation
RCT	Randiomised Controlled Trial
RDB	Regulome Database
RNA	Ribonucleic Acid
RPKM	Read Per Kilobase Per Million
SD	Standard Deviation
SE	Standard Error
SIFT	Sorting Intolerant From Tolerant
SMC	Smooth Muscle Cell
SMR	Summary-Based Mendelian Randomisation
SNP	Single Nucleotide Polymorphism
SVAS	Supravalvular Aortic Stenosis
TGF-β	Transforming Growth Factor-Β
TIMP	Tissue Inhibitor Of Metalloproteinase
vSMC	Vascular Smooth Muscle Cell

WBS	Williams-Beuren Syndrome
wGRS	Weighted Genetic Risk Score
WMS	Weil-Marchesani Syndrome
XGR	Exploring Genomic Relations

Table of Contents

Abstract *2*

Acknowledgements *6*

Preface *7*

Abbreviations *8*

Table of Contents *12*

Chapter 1: Introduction *16*

1.1. Introduction 17
1.1.1. The extracellular matrix 17
1.1.2. Maintenance of the extracellular matrix 20
1.1.3. The elastic tissue 22
1.1.4. Disorders of the elastic tissue 24
1.1.5. Varicose veins as a complex disease model 32
1.1.6. Haemorrhoids as a complex disease model 40
1.1.7. Hernia as a complex disease model 47
1.1.8. Genome-wide association study 51
1.1.9. Scope of thesis 58
1.1.10. Funding 60

1.2. Chapter references 61

Chapter 2: Genome-wide association analysis of varicose veins *79*

2.1. Introduction 80
2.1.1. Rationale and aims 80

2.2. Methods 81
2.2.1. Ethics and consent 81
2.2.2. Study population and phenotype definition 81
2.2.3. Genotyping 84
2.2.4. Quality control 85
2.2.5. Imputation 88
2.2.6. Association analysis 88
2.2.7. Genomic risk loci definition 90
2.2.8. Functional annotation of SNPs 91
2.2.9. Candidate gene mapping 92
2.2.10. Gene set, tissue and pathway analyses 94
2.2.11. SNP-based heritability analysis 95
2.2.12. Genetic correlation analysis 96
2.2.13. Drug-target enrichment analysis 97
2.2.14. Genetic risk score 97
2.2.15. URLs 98

2.3. Results 100
2.3.1. Forty-six replicated varicose veins susceptibility loci 100
2.3.2. In silico annotation 121
2.3.3. Gene mapping 124
2.3.4. Gene set, pathway and tissue-specific enrichment 129
2.3.5 Genetic correlations with varicose veins associated phenotypes 132
2.3.6. Drug target enrichment analysis 134
2.3.7. Genetic risk score for varicose veins 135

2.4. Discussion 137
2.4.1. Summary 137
2.4.2. Angiogenesis 138
2.4.3. Lymphangiogenesis 140
2.4.4. Extracellular matrix regulation 141
2.4.5. Immune response 144
2.4.6. Vascular smooth muscle cell proliferation and migration 146
2.4.7. Apoptosis 147
2.4.8. Genetic risk score 148

2.4.9 Strengths and limitations 149
2.5. Conclusion 151
2.6. Chapter References 152
2.7. Chapter Appendix 162
Chapter 3: Genome-wide association analysis of haemorrhoids 163
3.1. Introduction 164
3.1.1. Rationale and aims 164
3.2. Methods 165
3.2.1. Ethics and consent 165
3.2.2. Study participants 165
3.2.3. Genotyping 167
3.2.4 Quality control 167
3.2.5. Imputation 170
3.2.6. Association analysis 170
3.2.7. Genomic risk loci definition 170
3.2.8. Functional annotation of SNPs 170
3.2.9. Candidate gene mapping 170
3.2.10. Gene set, tissue and pathway analyses 171
3.2.11. SNP-based heritability analysis 171
3.2.12. Genetic correlation analysis 171
3.2.13. Drug-target enrichment analysis 171
3.2.14. Genetic risk score 172
3.2.15. URLs 172
3.3. Results 174
3.3.1. Twelve novel haemorrhoids associated loci 174
3.3.2. In silico annotation 185
3.3.3. Gene mapping 187
3.3.4. Gene set, pathway and tissue-specific enrichment 193
3.3.5 Genetic correlations with haemorrhoids associated phenotypes 198
3.3.6. Drug target enrichment analysis 201
3.3.7. Genetic risk score for haemorrhoids 203
3.4. Discussion 205
3.4.1. Summary 205
3.4.2. Extracellular Matrix Remodelling 206
3.4.3. TGFβ-Signalling Pathway 208
3.4.4. Internal anal sphincter tonicity 210
3.4.5. Haemorrhoids and arterial dilating diseases 212
3.4.6. Haemorrhoids and colorectal cancer 214
3.4.7. Heritability and genetic correlations of haemorrhoids disease 216
3.4.8. Genetic risk score for haemorrhoids correlates with disease severity 218
3.4.9. Strengths and limitations 219
3.5. Conclusion 220
3.6. Chapter References 221
3.7. Chapter Appendix 232
Chapter 4: The shared genetic architecture of hernia phenotypes 233
4.1. Introduction 234
4.1.1. Rationale and aims 234
4.2. Methods 236
4.2.1. Ethics and consent 236
4.2.2. Study participants 236
4.2.3. Genotyping 243
4.2.4. Quality control 243
4.2.5. Imputation 243
4.2.6. Association analyses in BOLT-LMM 243
4.2.7. Genomic risk loci definition for BOLT-LMM studies 244
4.2.8. Functional annotation of BOLT-LMM studies 244
4.2.9. Candidate gene mapping of BOLT-LMM studies 244
4.2.10. Gene set, tissue and pathway analyses of BOLT-LMM studies 244
4.2.11. SNP-based heritability of BOLT-LMM studies 245
4.2.12. Genetic risk score for hernia 245
4.2.13. Multi-trait analysis in MTAG 245
4.2.14. Multivariate meta-analysis in metaUSAT 246
4.2.15. URLs 247

4.3. Results ... 248
4.3.1. Association analysis of individual hernia phenotypes in BOLT-LMM ... 248
4.3.2. In silico annotation of individual hernia loci ... 258
4.3.3. Candidate gene mapping of individual hernia loci ... 260
4.3.4. SNP-based heritability of individual hernia phenotypes ... 261
4.3.5. Genetic risk score for the individual hernia phenotypes ... 262
4.3.6. Combined association analysis of hernia phenotypes in BOLT-LMM ... 264
4.3.7. In silico annotation of combined hernia loci ... 272
4.3.8. Candidate gene mapping of combined hernia loci ... 273
4.3.9. Gene set, pathway and tissue enrichment analysis of combined hernia loci ... 274
4.3.10. SNP-based heritability of combined hernia phenotypes ... 276
4.3.11. Genetic risk score to look for evidence of shared biology between hernia subtypes ... 277
4.3.12. Multi-trait analysis of the individual hernia phenotypes in MTAG to uncover shared genetic biology ... 280
4.3.13. Multivariate meta-analysis of the individual hernia phenotypes in metaUSAT to uncover shared genetic biology ... 293
4.4. Discussion ... 305
4.4.1. Summary ... 305
4.4.2. Six loci with evidence of shared susceptibility towards multiple hernia subtypes ... 306
4.4.3. Six loci discovered through multi-trait analysis that weren’t discovered in the individual analyses ... 311
4.4.4. Genetic risk scoring of hernia severity and multiple hernia risk ... 316
4.4.5. Strengths and limitations ... 317
4.5. Conclusion ... 319
4.6. Chapter References ... 320
4.7. Chapter Appendix ... 331
Chapter 5: Conclusion ... 334
5.1. Conclusion ... 335
5.1.1. Summary ... 335
5.1.2. Implications ... 337
5.1.2. Future plans ... 338
5.2. Chapter references ... 340
5.3. Chapter Appendix ... 341
5.3.1 Academic development and contributions at Oxford ... 341

Chapter 1: Introduction

1.1. Introduction

1.1.1. The extracellular matrix

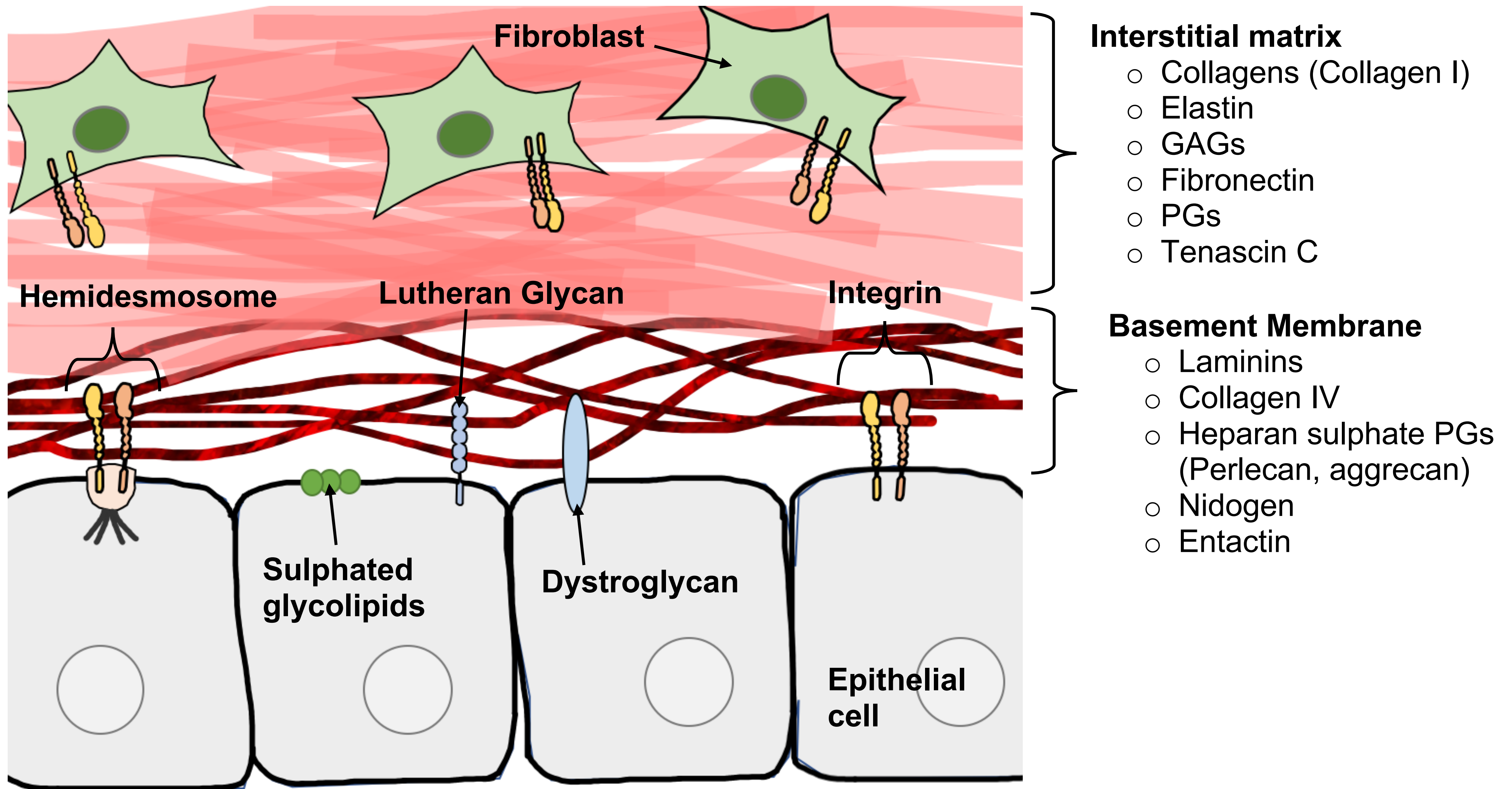
The extracellular matrix (ECM) is a complex three-dimensional network comprising ~1 - 1.5% of the mammalian proteome.[1] The ECM is a highly complex and dynamic organ central to all tissues; it is continuously remodelled to regulate tissue homeostasis.[2] Aside from its mechanical and architectural properties, via cell-ECM and ECM-cell interactions the ECM receives and transmits signals between cells, meaning it plays an important role in fundamental cellular processes such as adhesion, proliferation, migration, differentiation and apoptosis.[3] The ECM can be sub-divided into two fundamental constituents: i) the interstitial connective tissue matrix (encapsulates cells), ii) the basement membrane (separates epithelium from stroma) (**Figure 1.1**).[4] The core matrisome[*] comprises ~300 proteins, including 43 collagen subunits, ~200 glycoproteins and ~36 proteoglycans.[1] Collagens are the most abundant ECM protein, able to fold into different sized trimeric, coiled coil rods that impart structural integrity to the ECM of several tissues, including tendon (fibres), bone and cartilage (organic matrices), basement membrane (laminar sheets), vitreous humour (viscous matrix), and the dermis and capsules of organs (interstitial matrices).[5]

Another central structural component of ECM is elastin, a glycoprotein constituent of the ECM that confers elasticity to tissues and complements the function of collagens. Other extensively studied glycoproteins include fibronectins, laminins, thrombospondins, fibulins and tenascins.[6] Glycoproteins have a myriad of functions

[*] A complete 'parts list' of all the proteins that make up the extracellular matrix (ECM) and proteins that have the ability to contribute to the ECM in different scenarios.

including ECM assembly, and importantly can act as integral membrane proteins (such as the integrins) where they are central to ECM-cell interactions and regulate fundamental cellular processes, often via growth factor signalling.[3] The glycoprotein family of ECM proteins comprise a characteristic repeating oligosaccharide chain attached covalently to amino acid side-chains. Another constituent of the core matrisome is the proteoglycan family which, among other attributes, have important space-filling and lubrication properties due to their constituent glycosaminoglycan (GAG) chains which are strongly negatively charged.[7] Well-known proteoglycans include versican (anti-adhesion molecule in blood vessels and other tissues), perlecan (endothelial barrier function), aggrecan (major proteoglycan in cartilage), and fibromodulin (epidermis).[1]

Figure 1.1. The extracellular matrix (ECM). The ECM consists of two components: an interstitial matrix and a basement membrane. The mammalian matrisome is the complete 'parts list' of ECM proteins that comprise the ECM. Adapted from Bonnans (2014).[4]



1.1.2. Maintenance of the extracellular matrix

The ECM represents a dynamic organ with high turnover; it is continuously degraded, synthesised, re-assembled and modified.[8] This turnover is kept under tight control to maintain normal connective tissue function and homeostasis, particularly during injury. Perturbed ECM remodelling can precipitate morphological changes which can lead to pathology.[4] Disruptions in coordinated ECM synthesis or degradation can result in disorders of ECM accumulation (such as fibrotic disorders or malignancy[9]) or breakdown (such as varicose veins, haemorrhoids and hernia[10]). Remodelling of the structure, composition, and distribution of the ECM involves chemical and enzymatic cross-linking, and cleavage by proteolytic enzymes that are intimately involved in this process.[4] This includes the complex interactions of four groups of proteases:

i) Matrix metalloproteinases (MMPs)

The MMPs are the main enzyme in matrix degradation[4]; they are synergistically able to digest all ECM macromolecules. MMPs are metal-binding proteases that are secreted as latent precursors (*zymogens*) and activated by extracellular activation (most commonly cleavage). The MMP family comprises 23 members, with the majority being secreted and a small proportion being membranous. MMPs consist of three major groups: collagenases, gelatinases, and stromelysins (reviewed in Reynolds (2014)[11]). Their coordinated activity is necessary for tissue homeostasis; during injury, disease, or inflammation, their secretion and activation is heightened. Aside from ECM remodelling, MMPs are involved in the cleavage and activation of ECM-bound or intrinsic growth factors, and therefore have wide-ranging activities.[12]

ii) Adamalysins

The adamalysin family comprise ADAMs (a disintegrin and metalloproteinases), ADAMTS (ADAMs with a thrombospondin motif), and ADAMTSL (ADAMTS-like) proteins.[4] They are membranous or secreted enzymes that function in shedding ECM proteins adjacent to cell membranes, and partake in cell–cell fusion, adhesion and intracellular signalling.[13] A unique feature of adamalysins are their potential adhesion and protease domains, enabling them to partake in a plethora of biological functions, including the formation, remodelling, and degradation of components of the ECM.[14]

iii) Meprins

Meprins are part of the astacin family of membrane-bound and secreted zinc metalloproteinases. They are multi-domain and comprise two subunits - meprin-α and meprin-β.[4] Meprins are key shedding proteins that work at the cell membrane (predominantly meprin-β) or are secreted into peri-cellular space (meprin-α) and work to hydrolyse cell surface and ECM proteins, biologically active peptides, and cytokines.[4] Indeed, meprin synthesis is necessary for the activation of IL1β and IL18 interleukins[15,16] and for the dissemination of leucocytes through the ECM, and therefore has important roles in inflammation.[17,18] Meprins have further roles in the cleavage and maturation of a diverse range of ECM substrates, including collagen IV[4], pro-collagens I[19] and III[20], pro-ADAM-10[21], fibronectin[4], kallikrein[22], nidogen[4]; as well as the maturation of several growth factors (namely EGF and VEGFA) and proteases, including MMPs -3 and -9.[23]

iv) Metalloproteinase inhibitors

Dynamic remodelling and repair of the ECM necessitates careful control of the activity of ECM-degrading metalloproteinases to ensure that excessive and pathological

breakdown of the tissue matrix and cell surface molecules is avoided.[3,4] Metalloproteinase inhibitors therefore play a fundamental role in ECM composition and homeostasis. Four members of the tissue inhibitor of metalloproteinase (TIMP) family have been cloned (TIMPs 1 to 4), which counteract the function of both the matrix- (MMPs) and disintegrin- metalloproteinases (ADAMs and ADAMTSs).[24] TIMPs have a number of sites at which they can form complexes with metalloproteinases, enabling them to selectively and preferentially bind and inhibit specific metalloproteinases. The metalloproteinase:TIMP ratio of tissues therefore governs the overall composition of ECM, and is tightly regulated by cytokines, growth factors, and hormones, many of which are cell-specific and others which are ubiquitous (e.g. TGF-β).[25] Furthermore, enzymes such a LOX, LOXL1 and transglutaminases are involved in the cross-linking of ECM components and support the role of TIMPs in stiffening the ECM.[26]

1.1.3. The elastic tissue

The tissues of several organs need to be robust and extensible to perform their normal physiological roles, such as large arterial vessels during systole, the respiratory tree during inspiration, and the bowel during peristalsis.[27] The elastic properties of these tissues are provided largely by the distribution of elastic fibres throughout their ECM. The vast majority of the mature elastic fibre (~90%) is represented by a central amorphous, insoluble component (elastin), attached to (and interspersed by) a longitudinally aligned 10-15nm thick micro-fibrillar sheet.[28] The tubular microfibril structure consists of several glycoproteins, importantly fibrillin and micro-fibril-associated glycoprotein, which function as an organising scaffold to buttress the homogenous elastin.[27] The elastin precursor, tropoelastin, is secreted into the pericellular space where it is assembled into loose chains by being placed onto the

fibrillar scaffold and cross-linked into elastin fibres.[29] As well as being stabilised by cross-linking, the elastic fibre is promoted and stabilised by proteins such as the fibulins which link elastic fibres to cells.[30]

Alongside elastic properties which enable elastic tissues to recover after deformation, these tissues require tensile reinforcement to resist applied loads and to prevent overstretch.[29] The ECM of different tissues therefore comprises a varying proportion of amorphous elastin and proteoglycans, which impart matrix resilience, and collagens, which provide tensile strength as determined by the mechanical needs of the tissue.[31] Elastic fibres may represent a small component of some tissues (~2-4% of the dry weight of skin), or they can play a more significant role in others, such as in larger arteries, where elastic fibres constitute over 50% of the ECM architecture.[29]

The most abundant collagen, Type I collagen, exists in most mammalian tissues as closely-packed thick fibrils arranged in a superhelix.[12] Reticular collagen fibres are represented by Type III collagens; these fibres are coarse, branched, and segmented. They cross-link into a fine meshwork (reticulin), and support soft tissues such as the liver, bone marrow, and lymphatics.[32] Collagen can also exist in non-fibrous forms in ground substance[12], such as in the amorphous ground substance of hyalin and articular cartilage (Type II collagen) and the basal laminae (Type IV collagen).[32] The ECM fibres of different tissues can therefore be divided into an elastin system and a collagenous system[33], each working in sync to endow overall physiological functionality and character to different tissues.

1.1.4. Disorders of the elastic tissue

Disruption in the tightly controlled balance between elastic and collagen fibre composition and other matrix components can impact tissue homeostasis, and therefore the intrinsic molecular and physiological properties of tissues.[34] Indeed, elastic and collagen fibres represent a great challenge to tissue repair due to their molecular complexity and requirement of chaperone proteins for synthesis.[35] Disorders of elastic tissue are therefore common, and encompass a phenotypic spectrum of disease. Detailed study of genetic disorders that disrupt ECM assembly or homeostasis has enabled an enhanced understanding of the extensive molecular network that underpins normal tissue matrix biology.[36]

Genetic disorders of connective tissue broadly pertain to the assembly of elastin (elastinopathies) or collagen fibres (collagenopathies). The prototypic elastinopathies are the cutis laxa (CL) syndromes, Marfan Syndrome (MFS) and MFS-like syndromes, and pseudoxanthoma elasticum (PXE).[32] CL syndromes are characterised by abnormal elastic fibres causing loose, redundant hypoelastic skin with poor elastic recoil. Several subtypes of inherited CL exist, each with variable severity and clinical features, including skin laxity, hernias, aortic aneurysms, and bladder diverticula, among others (**Table 1.1**).[37] Autosomal dominant forms of MFS result from a mutation in the fibrillin-1 gene (FBN1), which is the core component of the microfibril layer of elastic fibres that is heavily involved in the sequestration and bioavailability of TGF-β peptides.[38] MFS is typified by cardiovascular, musculoskeletal, and ocular manifestations, with aortic aneurysms and dissection being a prominent cause of mortality in these patients.[39] PXE is an autosomal recessive disease caused by loss-of-function mutations of ABCC6, a putative transmembrane receptor.[32] PXE affects

tissues rich in elastic fibres, causing fragmentation and ectopic mineralisation of elastic fibres.[40] The cardinal features of PXE are skin elasticity, ocular neovascularisation, and cardiovascular manifestations, such as early arteriosclerosis and cardiac failure. Other elastinopathies include supravalvular aortic stenosis (SVAS), Williams-Beuren Syndrome (WBS), Weil-Marchesani Syndrome (WMS), congenital contractural arachnodactyly (Beals Syndrome), and geleophysic dysplasia.[41]

Collagenopathies, which alter collagen formation, impact almost all tissues and organ systems of the body and present with a range of phenotypes depending on the disruption of different collagens.[36,42] The most well-known collagenopathies are Ehlers-Danlos Syndromes (EDS), which are a group of collagen and collagen-related disorders showing clinical and genetic heterogeneity.[36] Patients with EDS broadly demonstrate cutaneous, ligamentous and articular abnormalities, as well as abnormalities of the vasculature and internal organs.[36] More specifically, the commonest features of EDS are of skin hyper-extensibility, fragility of tissues, and joint hypermobility (the 13 most common EDS subtypes are listed in **Table 1.2**).[42] Twenty-nine collagen types have been described, with several collagen types known to be associated with disease (**Table 1.3**).[42]

The elastic and connective tissue disorders mentioned thus far are heritable, and are caused largely by the disruption of a single gene (monogenic). Monogenic disorders are often rare, highly-penetrant, syndromic forms of disease; for example, MFS has a prevalence of ~1.5-17.2 per 100,000[43] and EDS has a prevalence of ~4 to 20 per 100,000.[44] Complex disorders, on the other hand, represent the vast majority of

human disease, and consist of both genetic (multiple alleles at distinct loci each with a small effect on phenotype[45]) and non-genetic contributors which together impart overall disease susceptibility. Complex disorders often have a prevalence of *at least* 1 in 3, meaning their disease burden is far greater on a population level and they have a higher societal cost.

Complex disorders of the elastic tissue, like monogenic disease, can impact every organ of the body (See **Figure 1.2**). The study of the genetic contributions to complex elastic tissue disorders is greatly lacking, and this is especially true for “surgical” diseases such as varicose veins, haemorrhoids and hernia, which are the focus of this thesis. Candidate gene studies have largely failed to identify genetic culprits[46,47], and research efforts and funding devoted to these surgical disorders have been limited.[48] In **Sections 1.1.5** through **1.1.7**, we will cover in further detail the background behind the three surgical disorders.

Table 1.1. Cutis laxa sub-types and clinical manifestations. Clinical manifestations are ordered according to how typical they are across all ten cutis laxa sub-types. Table adapted from Berk (2011).[37]

	ADCL	ARCL-IA	ARCL-IB	ARCL-IIA	ARCL-IIB	ARCL-III	XLCL	MACS	URDS	ATS
Genetic defect	*ELN*	*FBLN4*	*FBLN5*	*ATP6V0A2*	*PYCR1*	-	*ATP7A*	*RIN2*	*LTBP4*	*SLC2A10*
Skin laxity	+++	++	+++	+++	+++	+++	+++	+++	+++	+++
Hernia	++	+++	+++	+++	++	++	+++	+	+++	++
Joint laxity		+		+++	+++	++	++	+++	++	+++
Postnatal growth delay		++	+	+++	+++	+++	+		+++	
Prominent ears	+++	+	+++				++		+	+
Scoliosis				++	+	++	++	+++		++
Hypotonia		+		+++		++	++		+++	
Delayed motor development		+	+	+++		+++	+++			
Retrognathia		+++		+++	+++				+++	+++
Emphysema	++	++	+++						+++	
Aortic aneurysm	++	+++		+						++
Arterial tortuosity		+++	+				++			+++
Mental retardation				+++	+++	+++	+++			
Bladder diverticula			+	+			+++		+++	
Patent anterior fontanelle				+++		++	++		++	
IUGR				+++	+++	+++			+	
Congenital hip dislocation		+		+	+++	++				
Hypertelorism		+++							+++	++
Osteoporosis					++		++	+		
Athetoid movements					+	+++				
Corneal opacification					+	+++				
SVAS			+++							
Glycosylation defects				++						
Occipital horns							+++			
Short and broad clavicles							+++			
Macrocephaly								+++		
Alopecia								+++		
Gingival hyperplasia								+++		

ADCL, Autosomal dominant cutis laxa; ARCL, autosomal recessive cutis laxa; ATS, arterial tortuosity syndrome; DBS, De Barsy syndrome; IUGR, intrauterine growth retardation; MACS, macrocephaly-alopecia-cutis laxa-scoliosis syndrome; SVAS, supravalvular aortic stenosis; URDS, Urban-Rifkin-Davis syndrome; XLCL, X-linked cutis laxa; +, rare; ++, not uncommon; +++, Common; blank, not present.

Table 1.2. The different subtypes of Ehlers-Danlos Syndrome (EDS). The 13 EDS subtypes categorised according to the latest latest 2017 international classification of EDS.[49] Table adapted from Meester (2017).[50]

EDS subtype	Gene	Key clinical features
EDS disorders of collagen structure and processing		
Classical	*COL5A/COL5A2*	Skin hyper-elasticity and hypermobile joints
Vascular	*COL3A1*	Skin and vascular fragility, characteristic facies
Athrochalasia	*COL1A/COL1A2*	Severe hypermobile joints, congenital hip dislocation, skin hyper-elasticity
Dermatosparaxis	*ADAMTS2*	Extreme skin fragility, mild hypermobile joints, characteristic facies
Cardiac-valvular	*COL1A2*	Severe cardiac valve abnormalities defects, hypermobile joints, skin hyper-elasticity
EDS disorders of collagen folding and cross-linking		
Kypho-scoliotic	*PLOD1/FKBP14*	Kyphoscoliosis, hypermobile joints, muscular atrophy
EDS disorders of myomatrix structure and function		
Classical-like	*TNXB*	Skin hyper-elasticity, hypermobile joints, skin fragility
Myopathic	*COL12A1*	Muscular atrophy, proximal joint contractures, distal joint hypermobility
EDS disorders of GAG synthesis		
Spondylo-dysplastic	*B4GALT7/B3GALT6*	Short stature, muscle atrophy, bowing of limbs
Musculo-contractural	*CHST14/DSE*	Congenital contractures, characteristic facies, skin fragility.
EDS disorders of the complement pathway		
Periodontal	*C1R/C1S*	Severe periodontitis, lack of attached gingiva, pretibial plaques
EDS disorders of intracellular processes		
Spondylo-dysplastic	*SLC39A13*	Short stature, muscular atrophy, bowing of limbs
Brittle Cornea syndrome	*ZNF469/PRDM5*	Thin cornea, keratoconus, keratoglobus, blue sclerae
EDS disorders that remain unresolved		
Hypermobile	?	Hypermobile joints, skin hyper-elasticity, smooth velvety skin

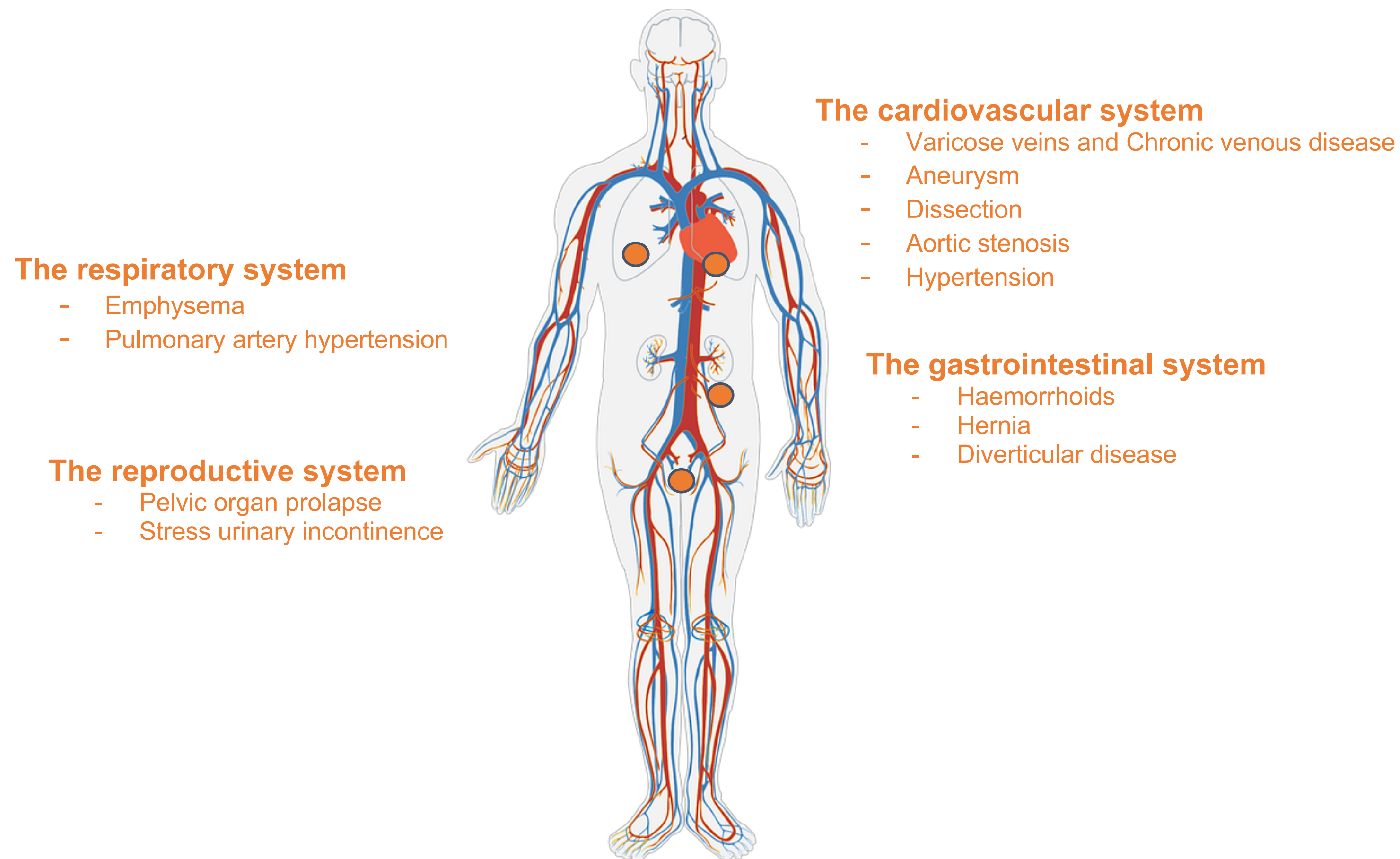
Table 1.3. The collagenopathies associated with disease. Collagen genes are arranged according to type. Table adapted from Jobling (2014).[42]

Collagen gene	Collagen type	Anatomic areas of expression	Associated Disorders
COL1A1	**Type I**	Most connective tissue, skin, tendon, ligament, bone	Caffey disease EDS classical and athrochalasia types OI types I, II, III, IV
COL1A2			EDS athrochalasia and cardiac-valvular types OI Types II, III, IV
COL2A1	**Type II** **Type XI**	Cartilage, nucleus pulposus, vitreous, cornea, inner ear	Achondrogenesis II/hypochondrogenesis Kneist dysplasia Osteoarthritis with mild chondrodysplasia Platyspondylic lethal skeletal dysplasia, Torrence type Sponyloepiphyseal dysplasia congenita Spondyloepimetaphyseal dysplasia, Strudwick type Stickler syndrome type 1, non-syndromic ocular type
COL3A1	**Type III**	Most connective tissue, especially vessels, skin and tendons	EDS vascular type EDS hypermobile type
COL4A1	**Type IV**	Basement membranes	Brain small vessel disease with Axenfeld–Reiger anomaly and hemorrhage Hereditary angiopathy with nephropathy, aneurysm and muscle cramps Porencephaly I
COL4A2			Porencephaly 2
COL4A3			Alport syndrome, AD and AR types
COL4A4			Benign familial hematuria
COL4A5			Alport syndrome, X-linked

COL4A6			Diffuse leiomyomatosis with Alport syndrome
COL5A1	**Type V**	Most connective tissue, especially skin, bone, tendon, cornea, placenta and fetal membranes	EDS classical type EDS brittle cornea syndrome
COL6A1	**Type VI**	Most connective tissue, tendons, contributes to cell matrix adhesion in skeletal muscle	Bethlem myopathy Ullrich congenital muscular dystrophy
COL7A1	**Type VII**	Anchoring fibrils in dermo-epidermal junctions	Epidermolysis bullosa dystrophica, autosomal recessive and dominant types, bart type, inversa type, pruriginosa type, pretibial type
COL9A1	**Type IX**	Cartilage, vitreous, retina, inner ear	MED Type VI Stickler syndrome Type IV, autosomal recessive
COL9A2			MED Type II Stickler syndrome Type V, autosomal recessive
COL9A3			MED Type III Multiple epiphyseal dysplasia with myopathy
COL10A1	**Type X**	Hypertrophic chondrocytes in calcifying cartilage	Metaphyseal chondrodysplasia, Schmid type
COL11A1	**Type XI**	Cartilage, nucleus pulposus, vitreous, cornea, inner ear	Stickler syndrome, Type II, AD Fibrochondrogenesis Marshal syndrome
COL11A2		Cartilage, nucleus pulposus, inner ear	Stickler syndrome, Type III, AD Fibrochondrogenesis Deafness, AD and AR Weissenbacker–Zweymuller syndrome
COL17A1	**Type XVII**	Component of hemidesmosomes	Junctional epidermolysis bullosa, non-Herlitz type
COL18A1	**Type XVIII**	Basement membranes	Knobloch syndrome

AD, autosomal dominant; AR, autosomal recessive EDS, Ehlers-Danlos Syndrome; MED, Multiple epiphyseal dysplasia; OI, Osteogenesis imperfecta;

Figure 1.2. Complex disorders with a primary elastic tissue pathology. A diagram depicting common elastic tissue disorders with elastin and/or collagen dysfunction. Other disorders with secondary elastic tissue involvement have not been included. Vector image from Pixabay – no attribution license.



1.1.5. Varicose veins as a complex disease model

Varicose veins are tortuous, dilated (> 3mm diameter), and palpable veins that are a highly prevalent clinical presentation of chronic venous disease (CVD) (**Figure 1.3**). CVD is classified clinically according to the universal CEAP classification system which aids research and reporting around CVD— this standard has recently been revised (**Table 1.4**).[51]

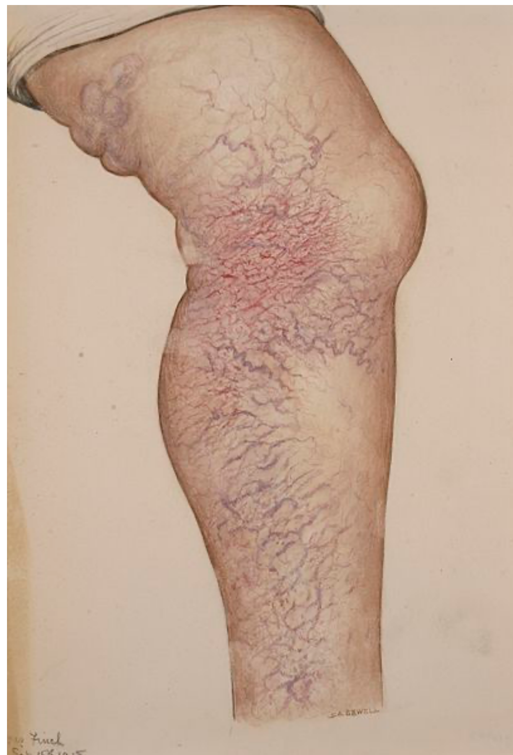


Figure 1.3. Extensive varicose veins affecting the leg. Image from the Wellcome Collection Gallery. St Bartholomew's Hospital Archives & Museum SBHB MU/14/51/4. Photo number: L0061468. Licensed under CC-BY-4.0.

Varicose veins are estimated to affect 25-33% of females and 10-20% of males[52,53], and temporal trend data demonstrate a growing disease prevalence.[54] The Bonn Vein

Study demonstrated that chronic venous insufficiency leads to significant skin changes in around 3-14% of cases[55], including oedema, lipodermatosclerosis, skin ulcers, and infrequently amputation.[56] Each year in the US, *at least* 20,556 patients receive a new diagnosis of venous ulceration[57], with chronic venous insufficiency responsible for ~72% of these.[58] Despite established treatment protocols, 25-50% of venous ulcers of the leg continue unhealed after six months of therapy[59], and the annual cost of managing venous leg ulcers in the USA is ~$14.9 billion.[60] Each year around 2% of healthcare expenditure in the UK is directly attributable to the management of varicose veins and its complications, which represents ~10-30% of the workload of district nurses.[61] The cost of venous-related illness on society at large is extensive: each year 4.5 million work days in the US *alone* are lost to venous-related illness.[62] Varicose veins therefore lead to significant economic and societal health costs.

Presently there are no medical treatments for varicose veins, with management restricted to surgical intervention (Marsden *et al*.[63] summarise NICE Guidance around the diagnosis and management of varicose veins). Endovenous ablative techniques are the first-line treatment for symptomatic varicose veins[63]; however, recurrence of varicosities following surgery is as high as 20%, meaning it is no better than open truncal stripping surgery.[64] Thus, for a significant subset of patients, varicose veins are complicated by ongoing sequelae and significantly impact patient-reported quality of life[57], requiring repeated intervention.

The aetiopathology of varicose veins is multifactorial (**Figure 1.4**).[65] A widely recognised theory is that varicose veins develop from valvular insufficiency which causes haemodynamic backflow and stasis, resulting in venous hypertension; this

precipitates vessel wall alterations as well as inflammation and activation of the venous wall.[66,67] However, the exact sequence of pathological events is not well understood and shows significant heterogeneity between patients[66,67], suggesting varicose veins is a complex disorder with several concomitant susceptibilities which lead to overall pathology. Risk factors for varicose veins established from large epidemiological studies include older age, female sex, obesity, height, orthostatic professions, a history of DVT, and a positive family history.[68,69]

Up to 85% of varicose veins patients report a positive family history[70], and among offspring with one affected parent the familial standardised incidence ratio is 2.39[71], with a narrow-sense heritability of ~17%.[72] Moreover, genetic factors contribute to 30-40% of femoral vein capacity and elasticity (a predisposing phenotypic contributor to varicose veins).[73] This suggests that inherent venous characteristics, and thus varicose vein pathology, are at least *in part* under genetic control.[74] Moreover, several candidate gene and linkage studies have identified putative genes implicated in varicose veins biology, including *FOXC2* (Forkhead Box C2), *THBD* (Thrombomodulin), *DMN* (Desmulin), and *MTHFR* (methylenetetrahydrofolate reductase) genes. Through an analysis of cDNA libraries comparing varicose veins with normal veins, Lee *et al.*[75] identified differential expression in several ECM proteins, including tubulin, lumican, and versican.

Two previous genome-wide association studies (GWAS) of varicose veins have been described in the literature. Ellinghaus *et al.*[76] identified two genome-wide significant susceptibility loci at *EFEMP1* and *KCNH8,* in their discovery cohort of 323 cases and 4,619 controls and independent replication in 1,946 cases and 3,146 control subjects.

More recently, Fukaya *et al*.[77] performed a GWAS in 9,577 cases and 327,959 control subjects from the UK Biobank, identifying 30 additional signals (27 loci) associated with varicose veins. This study, however, has two main limitations. Firstly, cases were defined only by the International Classification of Diseases (ICD) 9th or 10th edition codes, meaning that thousands of cases defined by operative intervention codes were misclassified as controls. Secondly, the genetic associations discovered were not tested in an independent replication cohort.[77]

Table 1.4. Updated 2020 CEAP classification system and reporting standard for chronic venous disease. The CEAP (Clinical-Etiology-Anatomy-Pathophysiology) classification standard to describe patients with chronic venous disorders (CVD). The system is based on the clinical manifestations of CVD and a current understanding of its aetiology, pathophysiology, and anatomical involvement. This table is based on the latest revised 2020 CEAP classification published by the American Venous Forum.[51]

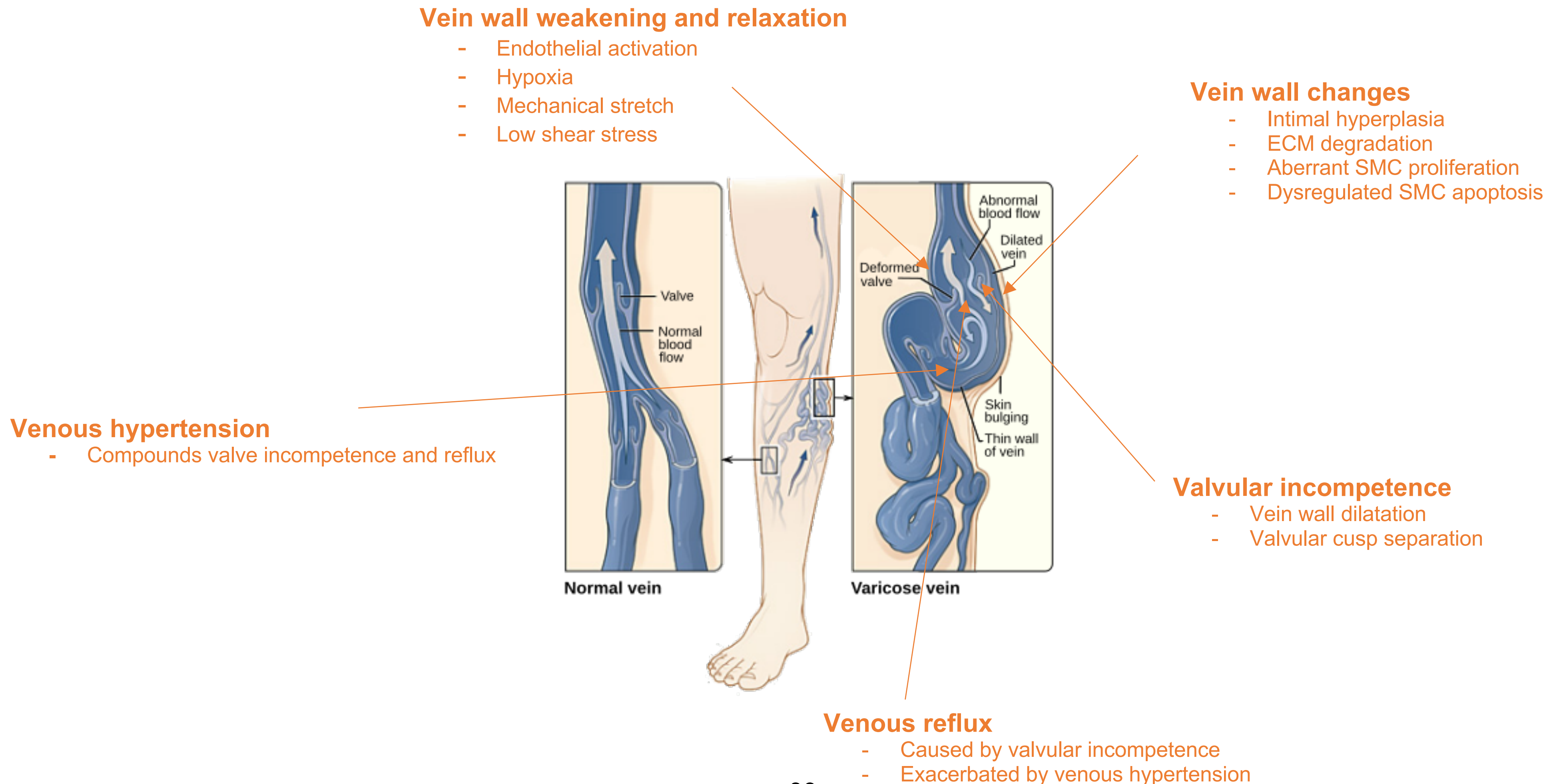
CEAP Classification 2020

Clinical classification	
$\mathbf{C_0}$	No visible or palpable signs of venous disease
$\mathbf{C^1}$	Telangiectasias or reticular veins
$\mathbf{C^2}$	Varicose veins
$\mathbf{C^{2r}}$	Recurrent varicose veins
$\mathbf{C_3}$	Oedema
$\mathbf{C_4}$	Changes in skin and subcutaneous tissue secondary to CVD
$\mathbf{C_{4a}}$	Pigmentation or eczema
$\mathbf{C_{4b}}$	Lipodermatosclerosis or atrophie blanche
$\mathbf{C_{4c}}$	Corona phlebectatica
$\mathbf{C_5}$	Healed
$\mathbf{C_6}$	Active venous ulcer
$\mathbf{C_{6r}}$	Recurrent active venous ulcer
(A)etiologic classification	
$\mathbf{E_p}$	Primary
$\mathbf{E_s}$	Secondary
$\mathbf{E_{si}}$	Secondary – intravenous

E_{se}	Secondary – extravenous
E_c	Congenital
E_n	No cause identified
Anatomic classification	
A_s	**Superficial** Tel Telangiectasia Ret Reticular veins GSVa Great saphenous vein above knee GSVb Great saphenous vein below knee SSV Small saphenous vein AASV Anterior accessory saphenous vein NSV Nonsaphenous vein
A_d	**Deep** IVC Inferior vena cava CIV Common iliac vein IIV Internal iliac vein EIV External iliac vein PELV Pelvic veins CFV Common femoral vein DFV Deep femoral vein FV Femoral vein POPV Popliteal vein TIBV Crural (tibial) vein PRV Peroneal vein ATV Anterior tibial vein PTV Posterior tibial vein MUSV Muscular veins GAV Gastrocnemius vein SOV Soleal vein

A_p	**Perforator** TPV Thigh perforator vein CPV Calf perforator vein
A_n	**No venous anatomic location identified**
Patho-physiologic classification	
P_r	Reflux
P_o	Obstruction
$P_{r,o}$	Reflux and obstruction
P_n	No pathophysiology identified

Figure 1.4. Pathophysiology of varicose veins. The pathophysiology of varicose veins is highly complex and the order of pathological events is a topic of debate. Outlined below are the most well-defined contributors to varicose veins pathobiology, alongside predisposing factors such as age, sex, weight, height, and genetics.[67] Image courtesy of the National Health Lung and Blood Institute (NIH)®, Sep 2017. Image modified by *Jmarchn*, made available by Wikimedia commons, under a CC-BY-SA-3.0 license.



1.1.6. Haemorrhoids as a complex disease model

Haemorrhoids (*piles*) result from the enlargement of the haemorrhoidal veins and distal displacement of the anal cushions (**Figure 1.5**).[78] Haemorrhoidal veins are normal structures that play an important role in faecal continence[79]; at rest they are full of blood, acting as a plug and contribute 15-20% of resting anal pressure.[80]

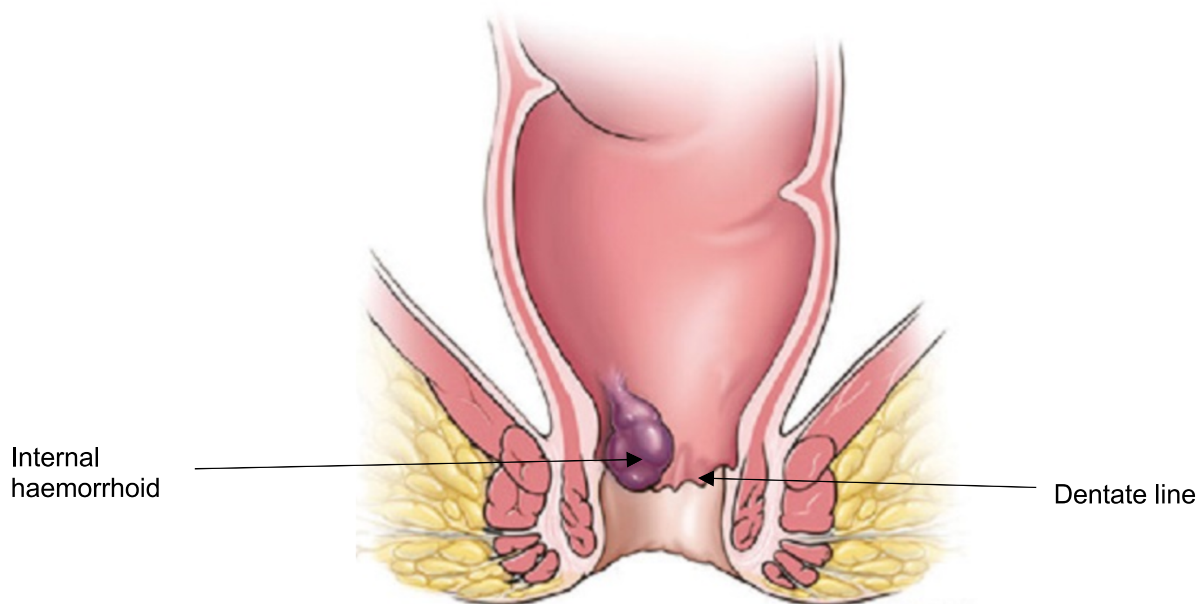


Figure 1.5. An internal haemorrhoid. The schematic depicts a Grade I internal haemorrhoid. Internal haemorrhoids are those originating above the dentate line and external haemorrhoids are those originating below. Image adapted from ConsultQD (Cleveland Clinic).

Haemorrhoids are the third most common outpatient gastrointestinal diagnosis[81], reported incidentally in 40%[82] of Americans after screening colonoscopies and 86% of proctoscopies in rectal surgery clinics[83], with an estimated worldwide prevalence

ranging from 4.4% to 36.4%.[84–86] Haemorrhoids disease[*] occurs in ~10-30% of haemorrhoids patients[87], representing 3.3 million ambulatory care visits annually in the USA.[82] Crosland and Jones found only 41% of patients in general practice reporting rectal bleeding had sought medical advice for it[88], suggesting the actual incidence of symptomatic haemorrhoids is likely much higher due to the associated anxiety and personal and cultural stigma.[89] Haemorrhoids can also co-exist with serious pathology, or can often be mis-diagnosed in its place.[90] General practice prescriptions for topical treatments for haemorrhoids have been found to increase in the year prior to a rectal cancer diagnosis.[91] It is therefore important that correct diagnoses for haemorrhoids are made and that risk of more serious proximal pathology is ruled out.[92]

Internal haemorrhoids severity is commonly graded according to the Goligher Classification System **(Table 1.5)**.[93] Management of haemorrhoids disease is directed by the severity of symptoms and degree of prolapse (**Figure 1.6**)[92], and is predicted to increase 23% over the next 20 years.[94] Conservative approaches for early stage haemorrhoids can include fibre supplementation[95], lifestyle modification and topical preparations, with only limited data existing for their long-term benefit. Surgical intervention is the mainstay of treatment for haemorrhoids, with over 35,000 surgical interventions performed for haemorrhoids in England in 2018-19.[96] Surgical interventions include outpatient procedures, principally rubber band ligation (RBL) which is associated with bleeding[97], pain, and high-recurrence rates (particularly in haemorrhoids with significant prolapse) which can be as high as 49%.[98] Radical

[*] Haemorrhoids disease is the symptomatic presentation of pathological haemorrhoids (i.e. dilated haemorrhoidal veins). Most patients with haemorrhoids report no symptoms and therefore do not have haemorrhoids disease.

surgery requiring general anaesthesia is the definitive treatment for high-grade haemorrhoids that do not respond to outpatient treatment, and is particularly effective in larger prolapsed haemorrhoids[99]; however it is associated with short and long-term complications.[100,101] Thus, for a significant proportion of patients, haemorrhoids significantly impact their quality of life.

The aetiology of haemorrhoids has been a source of debate and discussion over the centuries.[95] Haemorrhoids is thought to have a complex multi-factorial aetiology, with four theories proposed (**Figure 1.7)**: the varicose vein theory[*] (which is the least accepted of all four[102]), the internal anal sphincter (IAS) hypertonicity theory[†], the (anal cushion) vascular hyperplasia theory[‡] (both of which have been postulated to increase resting pressure in the anal canal), the vascular theory[§], and the more widely accepted sliding anal lining theory[**] (thought to be caused by disruption of the stromal component of the anal canal and rectal redundancy).[102,103] It is likely that several of these factors are at play in a complex manner, each imparting a heightened risk of haemorrhoids disease. Risk factors for haemorrhoids[79] include middle-age (peak prevalence of 45-65, following a Gaussian distribution[84]), male sex, prolonged straining (due to constipation or diarrhoea), hard stool, a low-fibre diet, pregnancy, obesity, high socioeconomic status and associated conditions[79] such as hernia[104,105], varicose veins[106,107], genitourinary prolapse[108,109], diverticular disease.[110,111], and other diseases of connective tissue.[104,112] Several studies have pointed towards a

---

[*] Varicose vein theory – that haemorrhoids are varicosities, however this is now obsolete as haemorrhoids and anorectal varices are not the same.
[†] IAS hypertonicity theory – that haemorrhoids are caused by an increased resting anal pressure due to IAS hypertonicity.
[‡] Vascular hyperplasia theory – that haemorrhoids cushions are corpus cavernosum recti and maintain anal continence.
[§] Vascular theory – that haemorrhoidal veins occur due to hypertension in the portal venous system.
[**] Sliding anal-lining theory – that haemorrhoids are caused by a displacement of the anal lining mucosa of the anal cushions due to ECM fragmentation

positive family history as a contributor to haemorrhoids aetiopathology; however, no studies have as yet performed heritability estimates for haemorrhoids.[79,113–115] The existing literature has largely overlooked the role of genetic susceptibility in haemorrhoids[116], whilst the genetic contributions of other anorectal and associated disorders have been studied *in detail.*[117,118]

.

Table 1.5. **Goligher classification for internal haemorrhoids**. Internal haemorrhoids severity is commonly classified according to a four-grade system which describes the prominence of the haemorrhoidal veins and the degree of prolapse. Adapted from Cataldo (2005).[119]

Grade	Physical findings
I	Engorged and hyperaemic haemorrhoidal veins – no prolapse
II	Haemorrhoidal cushions prolapse into anal canal during defecation – spontaneous reduction
III	Haemorrhoidal cushion prolapse into anal during defecation, but remain prolapsed – manual reduction
IV	Entirely exteriorised haemorrhoids, permanently prolapsed – manual reduction not effective

Figure 1.6. Evidence-based treatment algorithm for Grade I-IV haemorrhoids disease. Adapted from the recent European Society of Coloproctology guideline for haemorrhoidal disease.[92] RBL, Rubber Band Ligation; DG-HAL, Doppler-Guided Haemorrhoidal Artery Ligation.

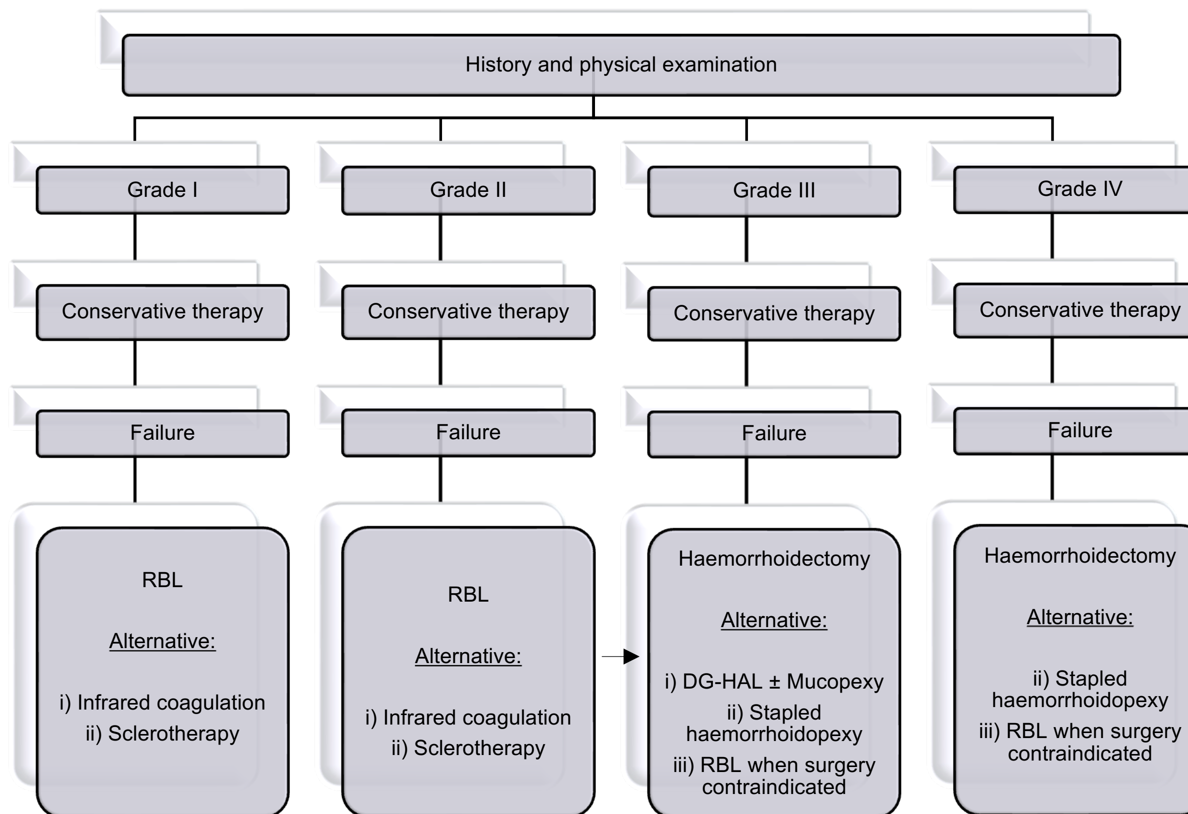
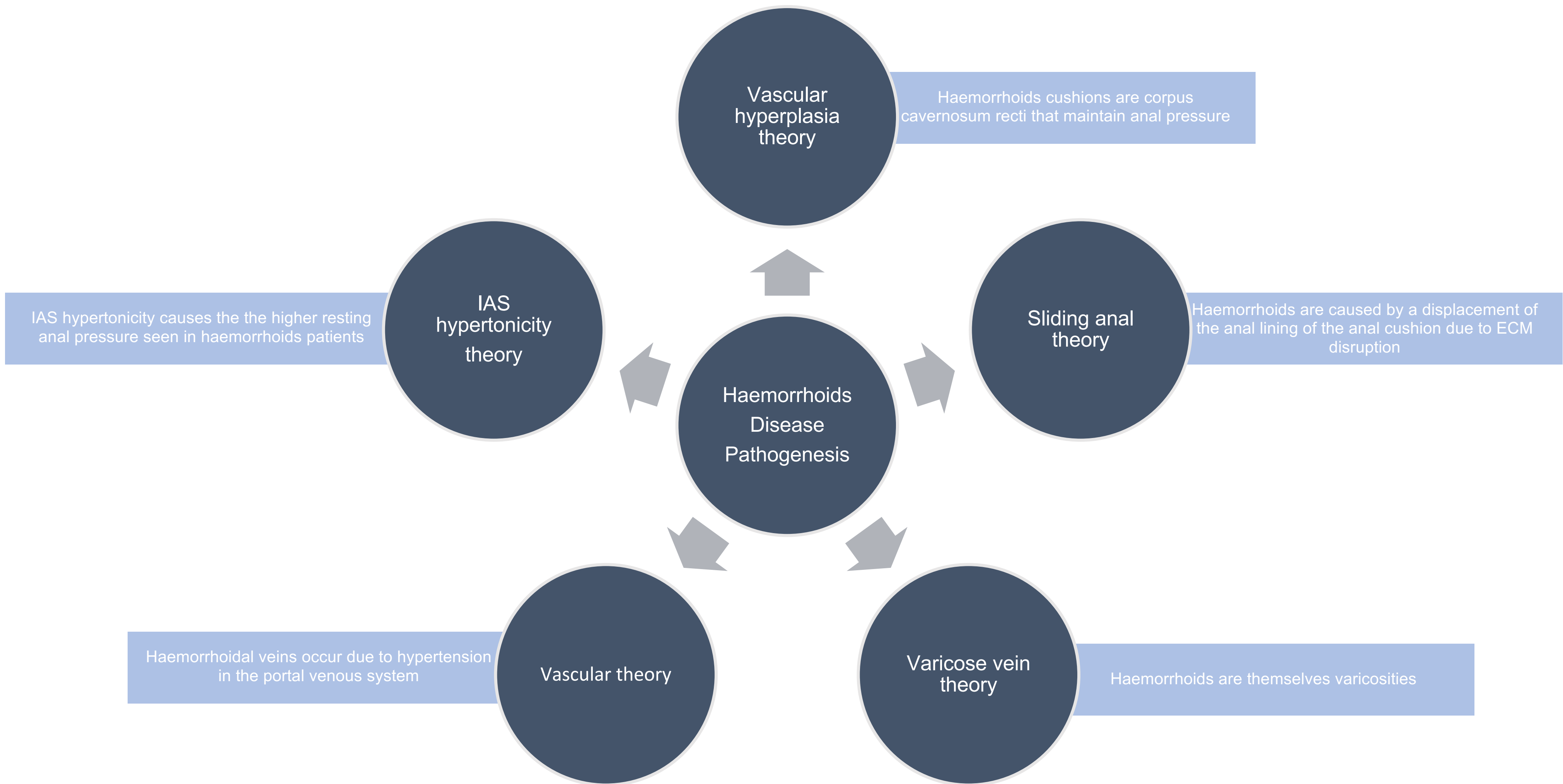


Figure 1.7. The pathogenesis of haemorrhoids disease. The five commonly held theories around the pathogenesis of haemorrhoids are provided below, alongside an explanation putting into context each theory.



1.1.7. Hernia as a complex disease model

Hernias are a group of conditions characterised by the abnormal protrusion of an organ from the anatomic cavity in which it resides. Abdominal wall hernia (AWH) represent the majority of hernia types (**Figure 1.8**), though hernia can occur outside the abdominal wall: principally hiatus hernia, which involves protrusion of the abdominal contents through the diaphragm into the mediastinum. The prevalence of AWH is ~100 to 300 per 100,000 of the population, and each year, at least 20 million AWHs are repaired worldwide[120,121]. Inguinal hernia represents around three-quarters of all hernias[121], with a lifetime risk of 27% in males and 3% in females.[122] In England, over 135,000 surgical procedures were performed in 2019 for AWH[123], making it one of the most commonly performed elective operations.[124]

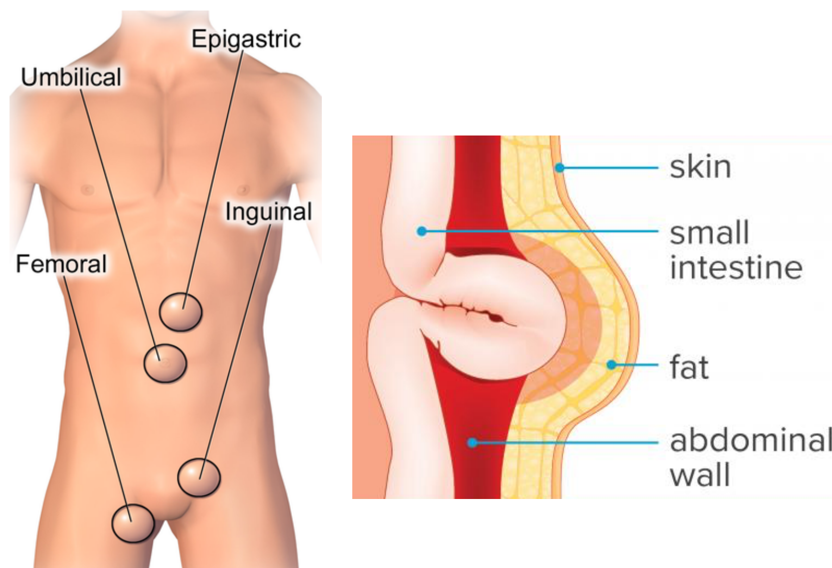


Figure 1.8. Common abdominal wall hernias. An image depicting common sites of hernia. *Image on the left made available by Wikimedia Commons under a CC-BY-SA-4.0 license. OTRS ticket #2019081910003902. Image on the right adapted from Medial News Today.*

Up to a third of abdominal wall and hiatus hernia patients are symptomatic.[125] The main clinical manifestation of hiatus hernia is gastro-oesophageal reflux disease (GORD), and the cornerstone of medical therapy is to alleviate symptoms by inhibiting gastric secretions.[126] Symptomatic AWH can present with swelling, a feeling of heaviness in the abdomen, burning sensations, sharp pain, and discomfort on coughing, activity and defaecation, with sudden onset severe pain and an irreducible lump suggestive of strangulation (a surgical emergency).[127] Surgery is the definitive treatment for symptomatic AWH; however, it is associated with complications including chronic postoperative pain, seroma or haematoma, infection, and failure of surgical repair[128,129]— with the reoperation rate for inguinal hernia recurrence being 12.3% at five years and 23.1% at 13 years.[130] Recurrence is lower and return to work is faster among prosthetic mesh repairs compared to sutured repairs[131,132], however mesh-related complications are common– including contraction, erosion and infection.[132] In the case of femoral hernia, diagnostic difficulties result in these hernias often being misdiagnosed. Up to 35-40% of these patients are not diagnosed until there is strangulation or bowel obstruction requiring emergency hernia repair (which is associated with high mortality rates).[133–135] Surgical management of AWHs can therefore be challenging, with high recurrence rates and risks of complication that can significantly reduce patients' quality of life. This emphasises a growing need to improve our understanding of hernia aetiopathology (**Figure 1.9**), which may guide new therapeutic avenues to improve patient outcomes.

Alongside the shared risk factors of age, sex, and body mass index, family history is a major contributor to hernia pathology. A positive family history is associated with an eight-fold increased risk of groin hernia[136,137], and is implicated in an increased

susceptibility to contralateral and recurrent inguinal hernia.[138–140] Among patients with a sibling with a surgically treated AWH, Zöller *et al.*[141], demonstrated the concordant standardised incidence ratio of inguinal hernia (1.97), femoral hernia (3.40), umbilical hernia (3.61), incisional hernia (2.24) and epigastric hernia (5.57). The characteristic presence of hernia in several genetic syndromes, including MFS[47], EDS[47] and CL syndromes (**Table 1.1**)[31], provide further evidence for an underlying genetic cause. Zöller *et al.*[141] also found several discordant risks to be over 2, suggesting a strong shared familial susceptibility among all five AWHs, with the greatest familial susceptibility shared between femoral and inguinal hernia. These results are supported by several clinical studies which identify a common co-occurrence of multiple hernia subtypes.[142,143] Indeed, patients with a first-degree relative with inguinal hernia are more at risk of femoral, umbilical, incisional and epigastric hernias.[144]

It is therefore likely that identifiable genetic risk factors may selectively predispose to distinct hernia pathology and even multiple hernia risk. However, at present no studies have characterised the genetic basis of non-congenital hernia and only a handful of gene studies have identified molecular candidates.[145–151] Only one GWAS of inguinal hernia (identifying four loci) has been performed[152], and a GWAS of GORD in 6,750 individuals did not identify any genome-wide significant loci.[153]

Figure 1.9. Pathophysiology of abdominal wall hernia. Summarised below are contributors of abdominal wall hernia pathology as highlighted by Abrahamson.[154] Image made available by Wikimedia Commons, under a CC-0-1.0 license.

Integrity of the fascial component

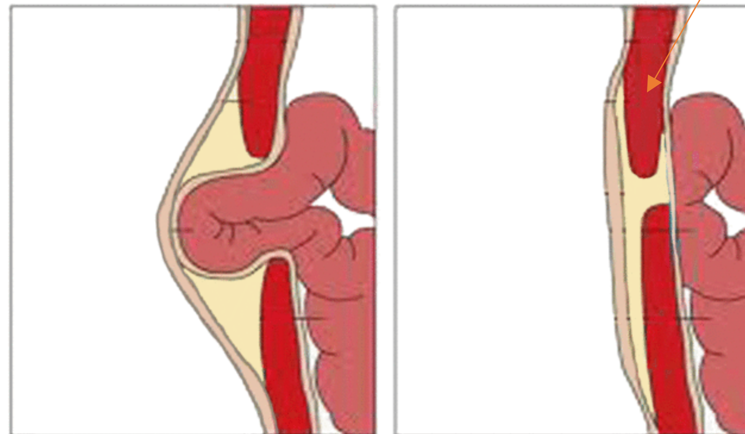
- Previous surgery
- Heavy lifting injury
- Patent processes vaginalis
- Weak fascia transversalis
- Collagen composition of fascia
- Faulty shutter mechanism

Intrinsic weakness in the abdominal wall

- Previous surgery
- Ageing

Increased intra-abdominal pressure

- Obesity
- Pregnancy
- Coughing
- Straining (constipation)
- Heavy lifting / straining



1.1.8. Genome-wide association study

The study of the influence of genetics in complex disorders relies on approaches beyond familial linkage studies. Genome-wide association studies (GWAS) are the current best way to study the polygenic architecture of these disorders (a field known as complex trait genetics). GWAS seeks to identify the correlation between genetic variation in regions of the genome (in particular common single-nucleotide polymorphisms (SNPs)[*]) and disease status. The first landmark GWA study, performed by Klein *et al*.[155] in 2005, investigated associations for age-related macular degeneration in a total cohort of 96 cases and 60 controls of white origin, identifying two significant SNP variants in a region of *CFH*, a regulator of innate immunity.[156] Fifteen years on, considerable strides in rapid, low-budget SNP array and high throughput sequencing technology[157], alongside sizable advances in data storage capacities have enabled association analysis across a much larger numbers of genetic variants, and across increasingly larger population sizes — facilitating the largest association analysis to date in 2,370,390 individuals.[158] As of 26th August 2020, a total of 4,681 association studies for 5,506 unique diseases and traits have been catalogued by the National Human Genome Research Institute–European Bioinformatics Institute (NHGRI-EBI) GWAS Catalog.[159]j GWAS is central to this thesis, hence a brief overview of its principles is warranted.

Modern SNP arrays genotype around ~0.5 to 1 million variants, capturing ~1% of variation in the human genome. GWAS leverages the principle of linkage disequilibrium (LD) at the population level – the non-random association between

[*] The most common form of genetic variation resulting from a change (polymorphism) of a single nucleotide. The term SNPs, variants, and markers are used synonymously in the context of GWAS.

genetic markers at different loci – meaning SNPs that are adjacent to each other are generally inherited together during recombination events.[160] Coordinated international research consortiums such as the 1000 Genomes Project[161] and the HapMap Project[162] have been successful in cataloguing SNPs that capture common genomic variation across different populations and their structural relation to one another. This development has been pivotal in facilitating imputation of ungenotyped SNPs from those that have been typed, meaning common variants in an individual can be predicted from lower-coverage SNP arrays.

The full-breadth of steps involved in a GWAS study are detailed by Hardy and Singleton[163]. Briefly, a GWA study comprises a case-control comparison for each typed or imputed variant between a population of cases[*] and a population of controls[†]. Statistical analysis is performed to examine the frequency of each allele in the two populations to look for significant association (See **Figure 1.10**). To account for multiple-testing and to control false positive rate, a genome-wide statistical significance threshold is set a $P < 5\times10^{-8}$ (0.05/ 1,000,000). GWAS is therefore rooted in an experimental, hypothesis-free and gene-agnostic methodology, identifying common variants that confer modest, incremental risk towards a disease phenotype at multiple loci (median odds ratio (OR) ~1.33 per SNP) (**Figure 1.11**).

Significant GWAS-associated variants, in themselves, are a location marker for disease, and are not indicative of a particular gene candidate. Indeed, for two-thirds of associated loci, the prioritised gene is not the nearest gene to the most associated

[*] individuals that have the trait or disease in question
[†] individuals that do not have a particular trait or disease

SNP at a locus.[164] Therefore functional annotation and mapping of loci is an important step to unravelling the genetic architecture of a disease and to prioritise actionable variants that can support functional validation experiments and therapeutic targeting.[165]

Several limitations are inherent to GWAS methodology– the most striking is that the vast majority of associated variants reside in non-coding or intergenic regions[166], and therefore for a significant proportion of GWAS hits, it is unclear how these loci play a role in disease biology. Another important limitation is that for complex disorders, a significant proportion of heritability are ascribed to the rare genetic variation which are not ascribed to loci captured within GWAS analyses, meaning there is significant missing heritability for complex traits that is difficult to attribute.[167] Moreover, to date, 88.9% of GWAS participants across all studies have been from European ancestry[168], meaning the transferability of GWAS results to non-European populations is restricted.[169] This is highlighted by the significantly lower predictive capabilities of polygenic risk scores (derived from European population) when examined in non-Europeans.[170] Moreover, significantly more genetic variation is present in non-European populations, which are unfortunately under-studied, meaning that current efforts are deprived of variants that are common in non-Europeans. Clearly there is much progress to be made in this regard, with efforts such as the GWAS Diversity Monitor[168] (gwasdiversitymonitor.com) and programmes such as the African Genome Variation Project, GemomeAsia 100K and H3Africa, representing important steps towards greater diversity in GWAS studies.

Despite its many limitations, the access to large summary statistics datasets and international collaborations has grown exponentially in the last few years which has enabled the genetic community to make great strides in understanding complex disease genetics through GWAS. Perhaps the most striking GWAS successes have been for auto-immune, metabolic and psychiatric disease.[164] Type II Diabetes Mellitus (T2DM) for instance is the most studied trait of all GWA studies (147 studies, 3% of all GWA studies)[159], leading to the identification of 2486 individual SNP associations.[159] This includes the identification of a loss-of-function mutation in *SLC30A8* (encodes a zinc transport ion channel expressed in beta cells) which is protective for TD2M, and has driven pharmaceutical efforts to develop ZnT-8 antagonists which are currently being investigated.[171] Other benefits of GWAS, include the utility of GWAS-derived variants as genetic predictors (polygenic risk scoring)[172], which have enabled the development of promising risk scoring tools for diseases such as glaucoma[173], and which may foreseeably be implemented on a population scale as more individuals are genotyped (for example, by direct-to-consumer companies such as 23andMe) and as genetic databases acquire greater population sizes to improve their predictive capabilities.

The very first GWA studies in small cohorts of < 1000 individuals often failed to capture the full extent of heritability of complex disorders, however we are now in an era of powerful association studies of up to 1 million individuals. The development of large population-based initiatives over the past five years, have been necessary to advance GWAS research efforts. The UK Biobank is a prospective cohort study of ~500,000 individuals of white British ancestry that have had whole genome genotyping and linkage to their electronic medical records and represents the principal data source

used in this thesis to uncover the genetic architecture of three common surgical disorders.[174]

Figure 1.10. Method for GWA Study designs. Calculation example illustrating the methodology behind GWA studies. The genotype counts for SNP 1 are taken from the 9p21 SNP as identified in the Wellcome Trust Case Control Consortium Study (2007) of seven common diseases.[175] The figure shows the G allele of SNP1 is significantly (P <5e-8) over-represented among cases compared to controls and may be a marker for disease. Adapted and made available by Wikimedia Commons under a CC-BY-3.0 license.

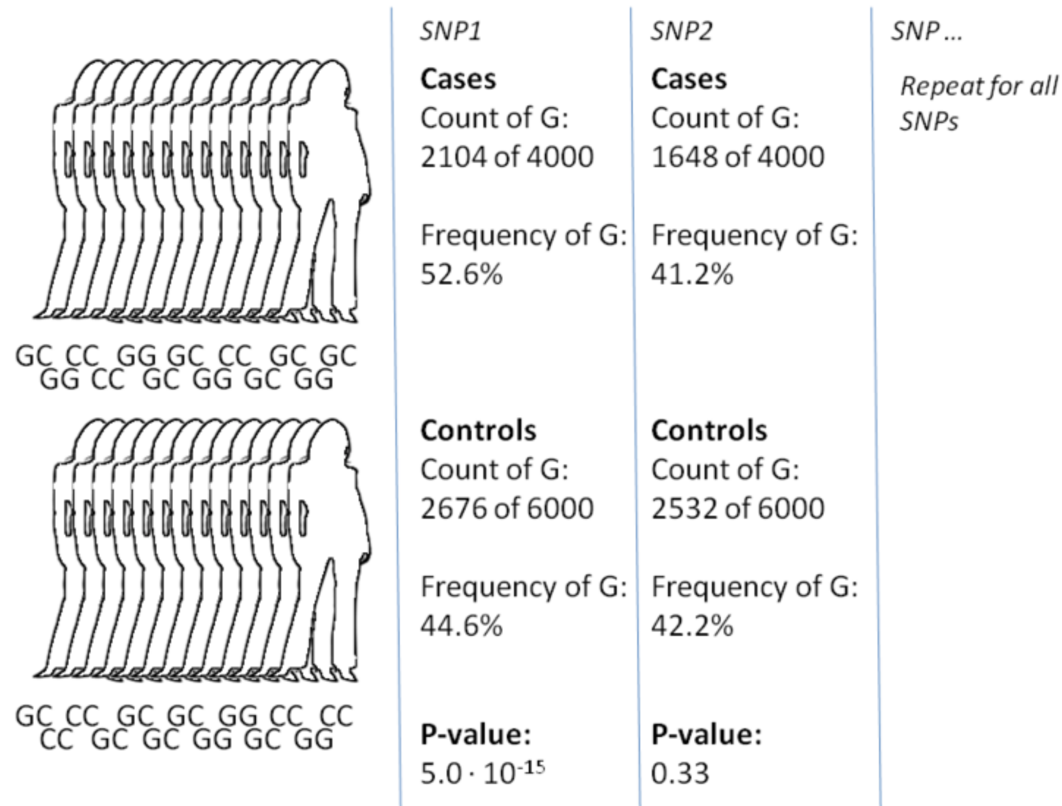
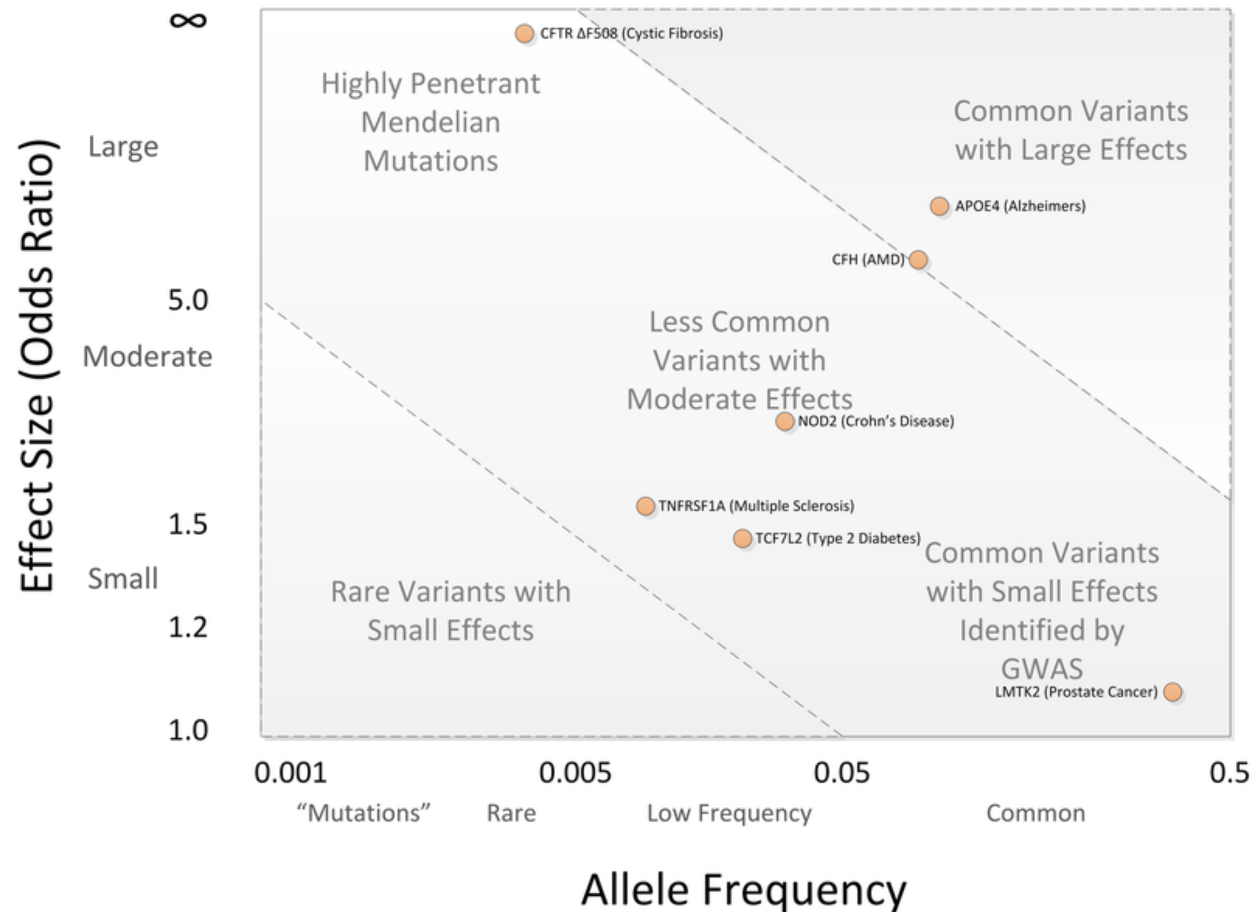


Figure 1.11. Spectrum of disease allelic effects identified by GWAS. A GWAS is positioned to identify common variants each with small effect sizes (OR < 1.5), each contributing to overall disease risk. Adapted from Bush (2012)[176] and made available by Wikimedia Commons under a CC-BY-2.5 license.



1.1.9. Scope of thesis

My thesis is titled 'Characterising the genetic architecture of complex elastic tissue disorders in UK Biobank'. It is important to define 'genetic architecture' and 'complex elastic tissue disorders':

i) *Genetic architecture* – pertains to the "characteristics of genetic variation that are responsible for heritable phenotypic variability."[177] It comprises the full extent of genetic variants that affect a trait, the frequency of these variants in the population, the effect size of each variant, and the variants' interaction with each other and the environment.[177]

ii) *Complex elastic tissue disorders* – it is important to note that there is no recognised classification of 'elastic tissue disorders', and the term is often used synonymously with connective tissue disorders. However, to distinguish the two, elastic tissue disorders herein refers to acquired diseases pertaining to significant involvement of elastic fibre dysfunction (disruption of elastin content) in their pathobiology and the resultant loss of resilience and elasticity of tissues.[31] The complex elastic tissue disorders being investigated are three: varicose veins, haemorrhoids, and hernia.

To this end, the chapters that follow have these broad aims:

i) **Chapter 2** – map the genetic architecture of varicose veins through a two-stage GWAS and prioritise candidate genes that are important in its pathobiology. Additionally, to perform bioinformatic functional analyses and construct a genetic risk score for the prognostication of varicose veins.

ii) **Chapter 3** – map the genetic architecture of haemorrhoids through the first-ever GWAS, and prioritise candidate genes that are important in its pathobiology. Additionally, to perform bioinformatic functional analyses and construct a genetic risk score for the prognostication of haemorrhoids.

iii) **Chapter 4** – map the genetic architecture of four hernia sub-types individually (inguinal, femoral, umbilical and hiatus hernia) and through multi-trait meta-analysis approaches, elucidate the shared genetic biology between multiple hernia subtypes. To construct a genetic risk score for the prognostication of hernia individually and for multiple hernia risk.

1.1.10. Funding

The author is recipient to scholarships from the Aziz Foundation, Wolfson Foundation and Royal College Surgeons of England. The author also receives stipendiary support from the National Institute of Health Research (NIHR) Oxford Biomedical Research Centre (Musculoskeletal theme). Previously, the author was funded by summer vacation studentships from the British Association of Plastic, Reconstructive and Aesthetic Surgeons (BAPRAS) and the Royal College of Surgeons of Edinburgh for work pertaining to **Chapter 2** of this thesis. The mentioned funding bodies have at no stage had any influence on the study and results reported herein.

1.2. Chapter references

1. Hynes, R. O. & Naba, A. Overview of the matrisome-An inventory of extracellular matrix constituents and functions. *Cold Spring Harb. Perspect. Biol.* **4**, (2012).
2. Hynes, R. O. The extracellular matrix: Not just pretty fibrils. *Science* **326**, 1216–1219 (2009).
3. Frantz, C., Stewart, K. M. & Weaver, V. M. The extracellular matrix at a glance. *J. Cell Sci.* **123**, 4195–4200 (2010).
4. Bonnans, C., Chou, J. & Werb, Z. Remodelling the extracellular matrix in development and disease. *Nature Reviews Molecular Cell Biology* **15**, 786–801 (2014).
5. Miller, R. T. Mechanical properties of basement membrane in health and disease. *Matrix Biology* **57–58**, 366–373 (2017).
6. Naba, A. *et al.* The matrisome: In silico definition and in vivo characterization by proteomics of normal and tumor extracellular matrices. *Mol. Cell. Proteomics* **11**, (2012).
7. Yue, B. Biology of the extracellular matrix: An overview. *Journal of Glaucoma* **23**, S20–S23 (2014).
8. Lu, P., Takai, K., Weaver, V. M. & Werb, Z. Extracellular Matrix degradation and remodeling in development and disease. *Cold Spring Harb. Perspect. Biol.* **3**, (2011).
9. Cox, T. R. & Erler, J. T. Remodeling and homeostasis of the extracellular matrix: Implications for fibrotic diseases and cancer. *DMM Disease Models and Mechanisms* **4**, 165–178 (2011).
10. Raffetto, J. D. Pathophysiology of Chronic Venous Disease and Venous Ulcers. *Surgical Clinics of North America* **98**, 337–347 (2018).

11. Reynolds, J. J., Hembry, R. M. & Meikle, M. C. Connective tissue degradation in health and periodontal disease and the roles of matrix metalloproteinases and their natural inhibitors. *Advances in dental research* **8**, 312–319 (1994).

12. Ricard-Blum, S. The Collagen Family. *Cold Spring Harb. Perspect. Biol.* **3**, 1–19 (2011).

13. Mead, T. J. & Apte, S. S. ADAMTS proteins in human disorders. *Matrix Biology* **71**–**72**, 225–239 (2018).

14. Kelwick, R., Desanlis, I., Wheeler, G. N. & Edwards, D. R. The ADAMTS (A Disintegrin and Metalloproteinase with Thrombospondin motifs) family. *Genome Biol.* **16**, 113 (2015).

15. Herzog, C. *et al.* Meprin A and meprin α generate biologically functional IL-1β from pro-IL-1β. *Biochem. Biophys. Res. Commun.* **379**, 904–908 (2009).

16. Banerjee, S. & Bond, J. S. Prointerleukin-18 is activated by meprin β in vitro and in vivo in intestinal inflammation. *J. Biol. Chem.* **283**, 31371–31377 (2008).

17. Sun, Q., Jin, H. J. & Bond, J. S. Disruption of the meprin α and β genes in mice alters homeostasis of monocytes and natural killer cells. *Exp. Hematol.* **37**, 346–356 (2009).

18. Crisman, J. M., Zhang, B., Norman, L. P. & Bond, J. S. Deletion of the Mouse Meprin β Metalloprotease Gene Diminishes the Ability of Leukocytes to Disseminate through Extracellular Matrix. *J. Immunol.* **172**, 4510–4519 (2004).

19. Broder, C. *et al.* Metalloproteases meprin α and meprin β are C- and N-procollagen proteinases important for collagen assembly and tensile strength. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 14219–14224 (2013).

20. Kronenberg, D. *et al.* Processing of procollagen III by meprins: New players in extracellular matrix assembly. *J. Invest. Dermatol.* **130**, 2727–2735 (2010).

21. Jefferson, T. *et al.* The substrate degradome of meprin metalloproteases reveals an unexpected proteolytic link between meprin β and ADAM10. *Cell. Mol. Life Sci.* **70**, 309–333 (2013).

22. Ohler, A., Debela, M., Wagner, S., Magdolen, V. & Becker-Pauly, C. Analyzing the protease web in skin: Meprin metalloproteases are activated specifically by KLK4, 5 and 8 vice versa leading to processing of proKLK7 thereby triggering its activation. in *Biological Chemistry* **391**, 455–460 (Biol Chem, 2010).

23. Geurts, N. *et al.* Meprins process matrix metalloproteinase-9 (MMP-9)/gelatinase B and enhance the activation kinetics by MMP-3. *FEBS Lett.* **586**, 4264–4269 (2012).

24. Nagase, H., Visse, R. & Murphy, G. Structure and function of matrix metalloproteinases and TIMPs. *Cardiovascular Research* **69**, 562–573 (2006).

25. Arpino, V., Brock, M. & Gill, S. E. The role of TIMPs in regulation of extracellular matrix proteolysis. *Matrix Biology* **44**–**46**, 247–254 (2015).

26. Lucero, H. A. & Kagan, H. M. Lysyl oxidase: An oxidative enzyme and effector of cell function. *Cellular and Molecular Life Sciences* **63**, 2304–2316 (2006).

27. Midwood, K. S. & Schwarzbauer, J. E. Elastic Fibers : Building Bridges Between Cells and Their Matrix. *Curr. Biol.* **12**, 279–281 (2002).

28. Mecham, R. P. Elastin in lung development and disease pathogenesis. *Matrix Biol.* **73**, 6–20 (2018).

29. Rosenbloom, J., Abrams, W. R. & Mecham, R. Extracellular matrix 4: The elastic fiber. *FASEB J.* **7**, 1208–1218 (1993).

30. Yanagisawa, H. *et al.* Fibulin-5 is an elastin-binding protein essential for elastic fibre development in vivo. *Nature* **415**, 168–171 (2002).

31. Kielty, C. M. Elastic fibres in health and disease. *Expert Rev. Mol. Med.* **8**, 1–23

(2006).

32. Vanakker, O., Callewaert, B., Malfait, F. & Coucke, P. The Genetics of Soft Connective Tissue Disorders. *Annu. Rev. Genomics Hum. Genet.* **16**, 229–255 (2015).

33. Montes, G. S. Structural biology of the fibres of the collagenous and elastic systems. *Cell Biol. Int.* **20**, 15–27 (1996).

34. Milewicz, D. M., Urban, Z. & Boyd, C. *Genetic disorders of the elastic fiber system*. *Matrix Biology* **19**, (2000).

35. Shifren, A. & Mecham, R. P. The stumbling block in lung repair of emphysema: Elastic fiber assembly. in *Proceedings of the American Thoracic Society* **3**, 428–433 (2006).

36. Harrison, P. & Wordsworth, P. Metabolic bone disease and inherited disorders of bone and connective tissue. in *The Rheumatology Handbook* 247–298 (Imperial College Press, 2011). doi:10.1142/9781848163218_0005

37. Berk, D. R., Bentley, D. D., Bayliss, S. J., Lind, A. & Urban, Z. Cutis laxa: A review. *J. Am. Acad. Dermatol.* **66**, 842.e1-842.e17 (2012).

38. Benke, K. *et al.* The role of transforming growth factor-beta in Marfan syndrome. *Cardiol. J.* **20**, 227–234 (2013).

39. Landis, B. J., Veldtman, G. R. & Ware, S. M. Genotype-phenotype correlations in Marfan syndrome. **103**, (2017).

40. Christiano, A. M. & Uitto, J. Molecular pathology of the elastic fibers. in *Journal of Investigative Dermatology* **103**, S53–S57 (1994).

41. Baldwin, A. K., Simpson, A., Steer, R., Cain, S. A. & Kielty, C. M. Elastic fibres in health and disease. *Expert Rev. Mol. Med.* **15**, (2013).

42. Jobling, R. *et al.* The collagenopathies: Review of clinical phenotypes and

molecular correlations. *Curr. Rheumatol. Rep.* **16**, (2014).

43. Kodolitsch, Y. Von *et al.* Perspectives on the revised ghent criteria for the diagnosis of marfan syndrome. *Appl. Clin. Genet.* **8**, 137–155 (2015).
44. Germain, D. P. Ehlers-Danlos syndrome type IV. *Orphanet Journal of Rare Diseases* **2**, 32 (2007).
45. Mitteroecker, P., Cheverud, J. M. & Pavlicev, M. Multivariate analysis of genotype–phenotype association. *Genetics* **202**, 1345–1363 (2016).
46. Grant, Y., Onida, S. & Davies, A. Genetics in chronic venous disease. *Phlebol. J. Venous Dis.* **32**, 3–5 (2017).
47. Barnett, C., Langer, J. C., Hinek, A., Bradley, T. J. & Chitayat, D. Looking past the lump: genetic aspects of inguinal hernia in children. *Journal of Pediatric Surgery* **44**, 1423–1431 (2009).
48. Mann, M., Tendulkar, A., Birger, N., Howard, C. & Ratcliffe, M. B. National Institutes of Health Funding for Surgical Research. *Ann. Surg.* **247**, 217–221 (2008).
49. Malfait, F. *et al.* The 2017 international classification of the Ehlers–Danlos syndromes. *Am. J. Med. Genet. Part C Semin. Med. Genet.* **175**, 8–26 (2017).
50. Meester, J. A. N. *et al.* Differences in manifestations of Marfan syndrome, Ehlers-Danlos syndrome, and Loeys-Dietz syndrome. *Ann. Cardiothorac. Surg.* **6**, 582–594 (2017).
51. Lurie, F. *et al.* The 2020 update of the CEAP classification system and reporting standards. *J. Vasc. Surg. Venous Lymphat. Disord.* **8**, 342–352 (2020).
52. Beebe-Dimmer, J. L., Pfeifer, J. R., Engle, J. S. & Schottenfeld, D. The epidemiology of chronic venous insufficiency and varicose veins. *Ann. Epidemiol.* **15**, 175–184 (2005).

53. Evans, C. J., Fowkes, F. G. R., Ruckley, C. V. & Lee, A. J. Prevalence of varicose veins and chronic venous insufficiency in men and women in the general population: Edinburgh Vein Study. *J. Epidemiol. Community Health* **53**, 149–153 (1999).

54. Heit, J. A. *et al.* Trends in the incidence of venous stasis syndrome and venous ulcer: A 25-year population-based study. *J. Vasc. Surg.* **33**, 1022–1027 (2001).

55. Rabe, E. *et al.* Bonn Vein Study by the German Society of Phlebology: Epidemiological study to investigate the prevalence and severity of chronic venous disorders in the urban and rural residential populations. *Phlebologie* **32**, 1–14 (2003).

56. Anwar, M. A. *et al.* A review of familial, genetic, and congenital aspects of primary varicose vein disease. *Circ. Cardiovasc. Genet.* **5**, 460–466 (2012).

57. Smith, J. J., Garratt, A. M., Guest, M., Greenhalgh, R. M. & Davies, A. H. Evaluating and improving health-related quality of life in patients with varicose veins. *J. Vasc. Surg.* **30**, 710–719 (1999).

58. Bergqvist, D., Lindholm, C. & Nelzen, O. Chronic leg ulcers: The impact of venous disease. *J. Vasc. Surg.* **29**, 752–755 (1999).

59. Singer, A. J., Tassiopoulos, A. & Kirsner, R. S. Evaluation and Management of Lower-Extremity Ulcers. *N. Engl. J. Med.* **377**, 1559–1567 (2017).

60. Rice, J. B. *et al.* Burden of venous leg ulcers in the United States. *J. Med. Econ.* **17**, 347–356 (2014).

61. Laing, W. Chronic Venous Diseases of the Leg. *Office of Health Economics: London* (1992). Available at: https://www.ohe.org/publications/chronic-venous-diseases-leg. (Accessed: 1st January 2020)

62. Stanley, J. C. *et al.* Vascular surgery in the United States: Workforce issues:

Report of the Society for Vascular Surgery and the International Society for Cardiovascular Surgery, North American Chapter, Committee on Workforce Issues. *J. Vasc. Surg.* **23**, 172–181 (1996).

63. Marsden, G., Perry, M., Kelley, K. & Davies, A. H. Diagnosis and management of varicose veins in the legs: Summary of NICE guidance. *BMJ (Online)* **347**, (2013).
64. O'Donnell, T. F., Balk, E. M., Dermody, M., Tangney, E. & Iafrati, M. D. Recurrence of varicose veins after endovenous ablation of the great saphenous vein in randomized trials. *Journal of Vascular Surgery: Venous and Lymphatic Disorders* **4**, 97–105 (2016).
65. Segiet, O. A. *et al.* Biomolecular mechanisms in varicose veins development. *Annals of Vascular Surgery* **29**, 377–384 (2015).
66. Raffetto, J. D. & Khalil, R. A. Mechanisms of varicose vein formation: Valve dysfunction and wall dilation. *Phlebology* **23**, 85–89 (2008).
67. Lim, C. S. & Davies, A. H. Pathogenesis of primary varicose veins. *Br. J. Surg.* **96**, 1231–1242 (2009).
68. Lee, A. J., Evans, C. J., Allan, P. L., Ruckley, C. V & Fowkes, F. G. R. Lifestyle factors and the risk of varicose veins: Edinburg Vein Study. *J. Clin. Epidemiol.* **56**, 171–179 (2003).
69. Sisto, T. *et al.* Prevalence and risk factors of varicose veins in lower extremities: Mini-Finland health survey. *Eur. J. Surgery, Acta Chir.* **161**, 405–414 (1995).
70. Scott TE, LaMorte WW, Gorin DR, M. J. Risk factors for chronic venous insufficiency: a dual case-control study. *J. Vasc. Surg.* **22**, 622–8 (1995).
71. Zöller, B., Ji, J., Sundquist, J. & Sundquist, K. Family history and risk of hospital treatment for varicose veins in Sweden. *Br. J. Surg.* **99**, 948–953 (2012).

72. Fiebig, A. *et al.* Heritability of chronic venous disease. *Hum. Genet.* **127**, 669–674 (2010).

73. Molnár, A. Á. *et al.* Heritability of venous biomechanics. *Arterioscler. Thromb. Vasc. Biol.* **33**, 152–157 (2013).

74. Brinsuk, M., Tank, J., Luft, F. C., Busjahn, A. & Jordan, J. Heritability of Venous Function in Humans. *Arterioscler. Thromb. Vasc. Biol.* **24**, 207–211 (2004).

75. Lee, S. *et al.* Gene Expression Profiles in Varicose Veins Using Complementary DNA Microarray. *Dermatologic Surg.* **31**, 391–395 (2006).

76. Ellinghaus, E. *et al.* Genome-wide association analysis for chronic venous disease identifies EFEMP1 and KCNH8 as susceptibility loci. *Sci. Rep.* **7**, 1–9 (2017).

77. Fukaya, E. et al. Clinical and Genetic Determinants of Varicose Veins. *Circulation* 1–12 (2018). doi:10.1161/CIRCULATIONAHA.118.035584

78. Henry, MM; Swash, M. *Coloproctology and the pelvic floor: Pathophysiology and management*. *Butterworths* (Butterworths, 1985).

79. Loder, P. B., Nicholls, R. J. & Phillips, A. K. S. *Haemorrhoids: pathology, pathophysiology and aetiology*. *British Journal of Surgery* **81**, (1994).

80. Banov, L., Knoepp, L. F., Erdman, L. H. & Alia, R. T. Management of hemorrhoidal disease. *J. S. C. Med. Assoc.* **81**, 398–401 (1985).

81. Sandler, R. S. & Peery, A. F. Rethinking What We Know About Hemorrhoids. *Clinical Gastroenterology and Hepatology* **17**, 8–15 (2019).

82. Everhart, J. E. & Ruhl, C. E. Burden of Digestive Diseases in the United States Part II: Lower Gastrointestinal Diseases. *Gastroenterology* **136**, 741–754 (2009).

83. Haas, P. A., Haas, G. P., Schmaltz, S. & Fox, T. A. The prevalence of

hemorrhoids. *Dis. Colon Rectum* **26**, 435–439 (1983).

84. Johanson, J. F. & Sonnenberg, A. The prevalence of hemorrhoids and chronic constipation. An epidemiologic study. *Gastroenterology* **98**, 380–386 (1990).

85. ACHESON, R. M. Haemorrhoids in the adult male; a small epidemiological study. *Guys. Hosp. Rep.* **109**, 184–195 (1960).

86. Gazet, J. C., Redding, W. & Rickett, J. W. The prevalence of haemorrhoids. A preliminary survey. *Proc. R. Soc. Med.* **63 Suppl**, 78–80 (1970).

87. Abramowitz, L. *et al.* The European Journal of General Practice The prevalence of proctological symptoms amongst patients who see general practitioners in France The prevalence of proctological symptoms amongst patients who see general practitioners in France. (2014). doi:10.3109/13814788.2014.899578

88. Crosland, A. & Jones, R. Rectal bleeding: Prevalence and consultation behaviour. *BMJ* **311**, 486 (1995).

89. Shi, Y. *et al.* Factors influencing patient delay in individuals with haemorrhoids: A study based on theory of planned behavior and common sense model. *J. Adv. Nurs.* **75**, 1018–1028 (2019).

90. Ristvedt, S. L. *et al.* Delayed treatment for rectal cancer. *Dis. Colon Rectum* **48**, 1736–1741 (2005).

91. Hansen, P. L., Hjertholm, P. & Vedsted, P. Increased diagnostic activity in general practice during the year preceding colorectal cancer diagnosis. *Int. J. Cancer* **137**, 615–624 (2015).

92. Tol, R. R. *et al.* European Society of ColoProctology: guideline for haemorrhoidal disease. *Color. Dis.* **22**, 650–662 (2020).

93. Lohsiriwat, V. Approach to Hemorrhoids. doi:10.1007/s11894-013-0332-6

94. Etzioni, D. A., Beart, R. W., Madoff, R. D. & Ault, G. T. Impact of the Aging

Population on the Demand for Colorectal Procedures. *Dis. Colon Rectum* **52**, 583–590 (2009).

95. Altomare, D. F. & Giuratrabocchetta, S. Conservative and surgical treatment of haemorrhoids. *Nature Reviews Gastroenterology and Hepatology* **10**, 513–521 (2013).

96. Hospital Admitted Patient Care Activity 2018-19 - NHS Digital. Available at: https://digital.nhs.uk/data-and-information/publications/statistical/hospital-admitted-patient-care-activity/2018-19. (Accessed: 4th August 2020)

97. Iyer, V. S., Shrier, I. & Gordon, P. H. Long-term outcome of rubber band ligation for symptomatic primary and recurrent internal hemorrhoids. *Dis. Colon Rectum* **47**, 1364–1370 (2004).

98. Brown, S. R. *et al.* Haemorrhoidal artery ligation versus rubber band ligation for the management of symptomatic second-degree and third-degree haemorrhoids (HubBLe): a multicentre, open-label, randomised controlled trial. *Lancet* **388**, 356–364 (2016).

99. Shanmugam V, Thaha MA, Rabindranath KS, Campbell KL, Steele RJ, L. M. Rubber band ligation versus excisional haemorrhoidectomy for haemorrhoids. *Cochrane Database Syst Rev* **3**, (2005).

100. Bleday R, Pena JP, Rothenberger DA, Goldberg SM, B. J. Symptomatic hemorrhoids: Current incidence and complications. *Dis. Colon Rectum* **35**, 477–481 (1992).

101. Sardinha, T. C. & Corman, M. L. Hemorrhoids. *Surgical Clinics of North America* **82**, 1153–1167 (2002).

102. Margetis, N. Pathophysiology of internal hemorrhoids. *Ann. Gastroenterol.* **32**, 264–272 (2019).

103. Pata, F. *et al.* Anatomy, Physiology and Pathophysiology of Haemorrhoids. *Rev. Recent Clin. Trials* **15**, (2020).

104. Salnikova, L. E., Khadzhieva, M. B. & Kolobkov, D. S. Biological findings from the PheWAS catalog: focus on connective tissue-related disorders (pelvic floor dysfunction, abdominal hernia, varicose veins and hemorrhoids). *Hum. Genet.* **135**, 779–795 (2016).

105. Abramson, J. H., Gofin, J., Hopp, C., Makler, A. & Epstein, L. M. The epidemiology of inguinal hernia. A survey in western Jerusalem. *J. Epidemiol. Community Health* **32**, 59–67 (1978).

106. Ekici, U., Kartal, A. & Ferhatoglu, M. F. Association Between Hemorrhoids and Lower Extremity Chronic Venous Insufficiency. *Cureus* **11**, (2019).

107. Burkitt, D. P. Varicose Veins Deep Vein Thrombosis, and Haemorrhoids: Epidemiology and Suggested Aetiology. *Br. Med. J.* **2**, 556 (1972).

108. Heslop, J. PILES AND RECTOCELES. *ANZ J. Surg.* **57**, 935–938 (1987).

109. Miedel, A., Tegerstedt, G., Mæhle-Schmidt, M., Nyrén, O. & Hammarström, M. Nonobstetric Risk Factors for Symptomatic Pelvic Organ Prolapse. *Obstet. Gynecol.* **113**, 1089–1097 (2009).

110. Humphreys, D. M. Diverticular disease: Three studies: Part I—Relation to other disorders and fibre intake. *Br. Med. J.* **1**, 424–425 (1976).

111. S, R. J. vd L. G. D. Diverticular disease. Pathology and clinical aspects based on 368 autopsy cases. *Zentralbl Chir* **116**, 991–8 (1991).

112. Willis, S., Junge, K., Ebrahimi, R., Prescher, A. & Schumpelick, V. Haemorrhoids - a collagen disease? *Color. Dis.* **12**, 1249–1253 (2010).

113. Bruch, H. P. & Roblick, U. J. Pathophysiologie des Hämorrhoidalleidens. *Chirurg* **72**, 656–659 (2001).

114. Khan, N. M. M. Injection sclerotherapy versus electrocoagulation in the management outcome of early haemorrhoids - PubMed. *J Pak Med Assoc* **56**, 579–582 (2006).

115. Yildiz, T., Aydin, D. B., Ilce, Z., Yucak, A. & Karaaslan, E. External hemorrhoidal disease in child and teenage: Clinical presentations and risk factors. *Pakistan J. Med. Sci.* **35**, 696–700 (2019).

116. Kaidar-Person, O., Person, B. & Wexner, S. D. Hemorrhoidal Disease: A Comprehensive Review. *Journal of the American College of Surgeons* **204**, 102–117 (2007).

117. Bogaert, J. & Prenen, H. Molecular genetics of colorectal cancer. *Annals of Gastroenterology* **27**, 9–14 (2014).

118. Cleynen, I. *et al.* Inherited determinants of Crohn's disease and ulcerative colitis phenotypes: A genetic association study. *Lancet* **387**, 156–167 (2016).

119. Cataldo, P. *et al.* Practice Parameters for the Management of Hemorrhoids (Revised). *Dis. Colon Rectum* **48**, 189–194 (2005).

120. Bay-Nielsen, M. *et al.* Quality assessment of 26 304 herniorrhaphies in Denmark: A prospective nationwide study. *Lancet* **358**, 1124–1128 (2001).

121. Kingsnorth, A. N. & LeBlanc, K. A. *Management of Abdominal Hernias*. (Springer London, 2013). doi:10.1007/978-1-84882-877-3

122. Primatesta, P. & Goldacre, M. J. Inguinal hernia repair: Incidence of elective and emergency surgery, readmission and mortality. *Int. J. Epidemiol.* **25**, 835–839 (1996).

123. NHS Digital. NHS Hospital Admitted Patient Care Activity, 2018-19: Procedures and Interventions. Available at: https://digital.nhs.uk/data-and-information/publications/statistical/hospital-admitted-patient-care-activity/2018-

19. (Accessed: 10th September 2020)

124. Abbott, T. E. F. *et al.* Frequency of surgical treatment and related hospital procedures in the UK: A national ecological study using hospital episode statistics. *Br. J. Anaesth.* **119**, 249–257 (2017).

125. Hair, A., Paterson, C., Wright, D., Baxter, J. N. & O'Dwyer, P. J. What effect does the duration of an inguinal hernia have on patient symptoms? *J. Am. Coll. Surg.* **193**, 125–129 (2001).

126. Roman, S. & Kahrilas, P. J. The diagnosis and management of hiatus hernia. *BMJ (Online)* **349**, (2014).

127. Jenkins, J. T. & O'Dwyer, P. J. Inguinal hernias. *BMJ* **336**, 269–272 (2008).

128. Simons, M. P. *et al.* European Hernia Society guidelines on the treatment of inguinal hernia in adult patients. *Hernia* **13**, 343–403 (2009).

129. Poelman, M. M. *et al.* EAES Consensus Development Conference on endoscopic repair of groin hernias. *Surgical Endoscopy* **27**, 3505–3519 (2013).

130. Flum, D. R., Horvath, K. & Koepsell, T. Have outcomes of incisional hernia repair improved with time? A population-based analysis. *Ann. Surg.* **237**, 129–135 (2003).

131. Scott, N. *et al.* Open Mesh versus non-Mesh for groin hernia repair. *Cochrane Database Syst. Rev.* (2001). doi:10.1002/14651858.cd002197

132. Smart, N. J., Marshall, M. & Daniels, I. R. Biological meshes: A review of their use in abdominal wall hernia repairs. *Surgeon* **10**, 159–171 (2012).

133. Dahlstrand, U., Wollert, S., Nordin, P., Sandblom, G. & Gunnarsson, U. Emergency femoral hernia repair: A study based on a national register. *Ann. Surg.* **249**, 672–676 (2009).

134. Koch, A., Edwards, A., Haapaniemi, S., Nordin, P. & Kald, A. Prospective

evaluation of 6895 groin hernia repairs in women. *Br. J. Surg.* **92**, 1553–1558 (2005).

135. Nilsson, H., Stylianidis, G., Haapamäki, M., Nilsson, E. & Nordin, P. Mortality after groin hernia surgery. *Ann. Surg.* **245**, 656–660 (2007).

136. Lau, H., Fang, C., Yuen, W. K. & Patil, N. G. Risk factors for inguinal hernia in adult males: A case-control study. *Surgery* **141**, 262–266 (2007).

137. Burcharth, J., Pommergaard, H. C. & Rosenberg, J. The inheritance of groin hernia: A systematic review. *Hernia* **17**, 183–189 (2013).

138. Ikeda, H. *et al.* Risk of contralateral manifestation in children with unilateral inguinal hernia: Should hernia in children be treated contralaterally? *J. Pediatr. Surg.* **35**, 1746–1748 (2000).

139. Jansen, P. L., Klinge, U., Jansen, M. & Junge, K. Risk factors for early recurrence after inguinal hernia repair. *BMC Surg.* **9**, 18 (2009).

140. Maddox, M. M. & Smith, D. P. A long-term prospective analysis of pediatric unilateral inguinal hernias: Should laparoscopy or anything else influence the management of the contralateral side? *J. Pediatr. Urol.* **4**, 141–145 (2008).

141. Zöller, B., Ji, J., Sundquist, J. & Sundquist, K. Shared and nonshared familial susceptibility to surgically treated inguinal hernia, femoral hernia, incisional hernia, epigastric hernia, and umbilical hernia. *J. Am. Coll. Surg.* **217**, (2013).

142. Glassow, F. Femoral hernia. Review of 2,105 repairs in a 17 year period. *Am. J. Surg.* **150**, 353–356 (1985).

143. De Luca, L. *et al. Relationship Between Hiatal Hernia and Inguinal Hernia*. *Digestive Diseases and Sciences* **49**, (2004).

144. Liem, M. S. L., Van Der Graaf, Y., Beemer, F. A. & Van Vroonhoven, T. J. M. V. Increased risk for inguinal hernia in patients with Ehlers-Danlos syndrome.

Surgery **122**, 114–115 (1997).

145. Sezer, S. *et al.* Association of collagen type I alpha 1 gene polymorphism with inguinal hernia. *Hernia* **18**, 507–512 (2014).

146. Han, Q. *et al.* Functional sequence variants within the SIRT1 gene promoter in indirect inguinal hernia. *Gene* **546**, 1–5 (2014).

147. Zhang, Y. *et al.* Genetic analysis of the TBX2 gene promoter in indirect inguinal hernia. *Hernia* **18**, 513–517 (2014).

148. Zhang, Y. *et al.* Genetic analysis of the TBX1 gene promoter in indirect inguinal hernia. *Gene* **535**, 290–293 (2014).

149. Antoniou, G. A. *et al.* Assessment of insertion/deletion polymorphism of the angiotensin-converting enzyme gene in abdominal aortic aneurysm and inguinal hernia. *Vascular* **21**, 1–5 (2013).

150. Wang, D., Han, Y., Xu, X., Chen, J. & Chen, Y. Matrix Metalloproteinases (MMP-2) and Tissue Inhibitors of Metalloproteinases (TIMP-2) in Patients with Inguinal Hernias. *World J. Surg.* 1–8 (2020). doi:10.1007/s00268-020-05674-0

151. Mao, Y. *et al.* A network analysis revealed the essential and common downstream proteins related to inguinal hernia. *PLoS One* **15**, e0226885 (2020).

152. Jorgenson, E. *et al.* A genome-wide association study identifies four novel susceptibility loci underlying inguinal hernia. *Nat. Commun.* **6**, (2015).

153. Bonfiglio, F. *et al.* A meta-analysis of reflux genome-wide association studies in 6750 Northern Europeans from the general population. *Neurogastroenterol. Motil.* **29**, (2017).

154. Abrahamson, J. *Groin Hernia Surgery Etiology And Pathophysiology Of Primary And Recurrent Groin Hernia Formation*.

155. Klein, R. J. *et al.* Complement factor H polymorphism in age-related macular

degeneration. *Science (80-.).* **308**, 385–389 (2005).

156. Rodríguez De Córdoba, S., Esparza-Gordillo, J., Goicoechea De Jorge, E., Lopez-Trascasa, M. & Sánchez-Corral, P. The human complement factor H: Functional roles, genetic variations and disease associations. *Molecular Immunology* **41**, 355–367 (2004).

157. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).

158. Baselmans, B. M. L. *et al.* Multivariate genome-wide analyses of the well-being spectrum. *Nat. Genet.* **51**, 445–451 (2019).

159. D., W. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006

160. Slatkin, M. Linkage disequilibrium - Understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics* **9**, 477–485 (2008).

161. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

162. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).

163. Hardy, J. & Singleton, A. Genomewide association studies and human disease. *N. Engl. J. Med.* **360**, 1759–1768 (2009).

164. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).

165. Breen, G. *et al.* Translating genome-wide association findings into new therapeutics for psychiatry. *Nature Neuroscience* **19**, 1392–1396 (2016).

166. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science (80-.).* **337**, 1190–1195 (2012).

167. Mayhew, A. J. & Meyre, D. Assessing the Heritability of Complex Traits in Humans: Methodological Challenges and Opportunities. *Curr. Genomics* **18**, (2017).

168. Mills, M. C. & Rahal, C. The GWAS Diversity Monitor tracks diversity by disease in real time. *Nature Genetics* **52**, 242–243 (2020).

169. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).

170. Duncan, L. *et al.* Analysis of polygenic risk score usage and performance in diverse human populations. *Nat. Commun.* **10**, 1–9 (2019).

171. Flannick, J. *et al.* Loss-of-function mutations in SLC30A8 protect against type 2 diabetes. *Nat. Genet.* **46**, 357–363 (2014).

172. Evans, D. M., Visscher, P. M. & Wray, N. R. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Hum. Mol. Genet.* **18**, 3525–3531 (2009).

173. Craig, J. E. *et al.* Multitrait analysis of glaucoma identifies new risk loci and enables polygenic prediction of disease susceptibility and progression. *Nat. Genet.* **52**, 160–166 (2020).

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

175. Burton, P. R. *et al.* Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).

176. Bush, W. S. & Moore, J. H. Chapter 11: Genome-Wide Association Studies. *PLoS Comput. Biol.* **8**, e1002822 (2012).

177. Timpson, N. J., Greenwood, C. M. T., Soranzo, N., Lawson, D. J. & Richards, J. B. Genetic architecture: The shape of the genetic contribution to human traits

and disease. *Nature Reviews Genetics* **19**, 110–124 (2018).

Chapter 2: Genome-wide association analysis of varicose veins

---

2.1. Introduction

2.1.1. Rationale and aims

As discussed in **Chapter 1**, varicose veins represent a common disorder with a high socioeconomic burden and associated patient morbidity.[1] There are currently no medical treatments for varicose veins, and endovenous surgical approaches to manage symptomatic varicose veins are associated with high recurrence.[2] The majority of varicose veins patients report a positive family history[3], thus demonstrating the importance of genetic susceptibility to varicose veins. However to date, relatively few genes have been implicated through candidate gene studies and GWAS approaches.[4–7] The aim of this chapter is to advance our understanding of the genetic architecture of varicose veins and discover clinically relevant biological pathways and potential molecular targets associated with varicose veins. This will be achieved by undertaking the largest two-stage GWAS of varicose veins using the UK Biobank cohort, with independent replication in a cohort from 23andMe, Inc.

2.2. Methods

2.2.1. Ethics and consent

UK Biobank received research ethics approval from the North West Research Ethics Committee (MREC) (11/NW/0382), in accordance with the Declaration of Helsinki. Informed consent was obtained to collate and disseminate genotype data for all study participants for the purposes of medical research. The work presented in this thesis was performed under UK Biobank study ID 10948. All 23andMe participants were from the userbase of 23andMe, Inc. (Sunnyvale, California) — a direct-to-consumer personal genomics company. Genotyping of participants was performed by the 23andMe Personal Genome Service. All 23andMe research participants included in this chapter completed an online questionnaire and provided informed consent for their genotype data to be used for research purposes, under a protocol approved by the external AAHRPP-accredited IRB, Ethical and Independent Review Services (E&I Review).

2.2.2. Study population and phenotype definition

UK Biobank is a population-level resource: a multicentre prospective cohort study of 488,377 UK participants aged 40-69 years recruited from 2006 to 2010.[8] Participants were invited to the study centre to provide written consent, complete a touch-screen questionnaire, and take part in a computer-based interview; physical and functional measures were also obtained, alongside genetic samples for whole genome genotyping.[9] Participants' genotype data was linked with their electronic medical

records to permit deep phenotyping.[9] The characteristics of the full UK Biobank cohort are described in detail elsewhere.[10]

In the discovery analysis, varicose vein cases were identified from the UK Biobank data showcase (ukbiobank.ac.uk) using diagnostic, operative and self-report codes. Cases for varicose veins were defined if participants had at least one of the following four codes (specific codes are in parentheses and **Table 2.1**)

1. Primary and/or secondary ICD-10 codes for varicose veins (I83)
2. Primary and/or secondary OPCS code for varicose vein surgery: (L84-L88)
3. Self-reported operation code for varicose vein surgery (1479)
4. Self-reported non-cancer illness code for varicose veins (1494)

In summary, 27,165 participants from the UK Biobank cohort had at least one of the above codes and were classified as varicose vein cases. Participants without any of these codes were designated as controls.

Table 2.1. Codes used for varicose veins case definition in UK Biobank. The total number of participants with each of the diagnostic codes is shown below. A total of 27,165 participants possessed at least one of the diagnostic codes for varicose veins.

Source of Data	UK Biobank Data Field	Code	Description	N
Primary ICD-10	41202	I83.0	Varicose veins of lower extremities with ulcer	12195
		I83.1	Varicose veins of lower extremities with inflammation	
		I83.2	Varicose veins of lower extremities with both ulcer and inflammation	
		I83.9	Varicose veins of lower extremities without ulcer or inflammation	
Secondary ICD-10	41204	As above	As above	1168
Primary OPCS	41200	L84	Combined operations on varicose vein of leg	12528
		L85	Ligation of varicose vein of leg	
		L86	Injection into varicose vein of leg	
		L87	Other operations on varicose vein of leg	
		L88	Transluminal operations on varicose vein of leg	
Secondary OPCS	41210	As above	As above	8116
Non-cancer illness (self-report)	20002	1494	Varicose veins	2266
Operation (self-report)	20004	1479	Varicose vein surgery	20115
Total (excluding overlaps)				**27165**

Following quality control (outlined in **Section 2.2.4. Quality control**), the final discovery analysis consisted of 22,473 cases and 379,183 controls.

In the 23andMe replication cohort, participants provided answers to the varicose veins-related question, *'Do you have varicose veins on your legs?' (Yes/Not Sure/No)*. Self-reported varicose veins cases were defined if they answered, *'yes'* to the above question, while controls were identified as those that answered *'No'.* Using this approach, in the final replication analysis, a total of 113,041 self-reported varicose veins cases and 295,928 control participants were included.

2.2.3. Genotyping

Genome-wide genotyping data was made available for 488,377 participants in the UK Biobank cohort.[9] The initial 49,950 participants were genotyped on the Affymetrix UK BiLEVE Axiom array (807,411 genotyped variants), with the second batch of 438,427 participants from the cohort genotyped on the Affymetrix UK Biobank Axiom array (825,927 genotyped variants). The two arrays were almost identical, sharing over 95% marker content. The present study is based on the third release of the UK Biobank cohort (July 2017), which contained the complete set of genotypes for the 488,377 participants (805,426 directly genotyped variants).

The 23andMe independent replication cohort was genotyped using one of four custom arrays (v1/v2, v3, v4, v5). Illumina HumanHap550+ BeadChip was used for v1/v2 (1,680 cases, 4,882 controls) and the Illumina OmniExpress+ BeadChip was used for v3 (21,342 cases, 56,448 controls). For v4 a fully customised array (58,883 cases, 148,637 controls) was used, and for v5, the Illumina Infinium Global Screening Array

was implemented (31,136 cases, 85,961 controls). Successive arrays contained significant overlap between all previous array chips.

2.2.4. Quality control

Quality control (QC) for the UK Biobank discovery cohort used a combination of UK Biobank's own QC and additional layers of more stringent QC performed locally by the Furniss Group. The full protocol has been described in detail elsewhere (**Figure 2.1**).[11] Briefly, all SNPs with a call rate < 90% were removed. This was followed by sample-level QC — participants were excluded if: (i) they demonstrated heterozygosity > 3 S.D. from the mean (calculated using UK Biobank's PCA-adjusted heterozygosity values, Data Field 22004); (ii) there was disparity between genetically inferred sex (Data Field 22001) and self-reported sex (Data Field 31) or participants with aneuploidy of sex chromosomes (Data Field 22019); and (iii) had a call rate < 98%. Further, all participants who were not white British in ancestry (based on principal component analysis (PCA) and self-reported ethnicity (Data Field 22006)) were excluded.

Following PCA, 86,693 participants were excluded from the discovery GWAS analysis. Using a linear mixed model implemented in BOLT-LMM[12] enabled the inclusion of related participants.

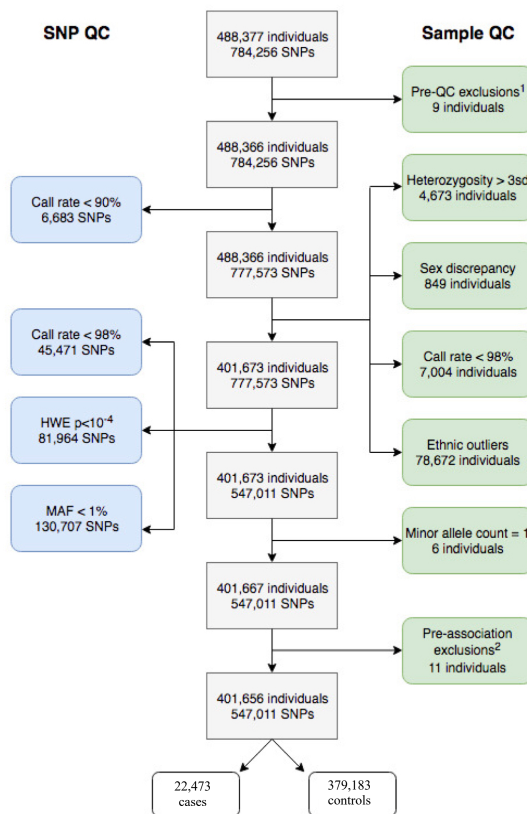
Next, SNP-level QC was performed, following which 230,562 SNPs were excluded based on: (i) a call rate < 98%, (ii) Hardy-Weinberg Equilibrium (HWE) $P < 1\times10^{-4}$, (iii) minor allele frequency (MAF) < 0.01. Six participants were further excluded because they were visual outliers when autosomal heterozygosity was plotted against call rate.

The post-QC discovery GWAS therefore consisted of 401,667 participants and 547,011 directly genotyped variants (**Figure 2.1**).

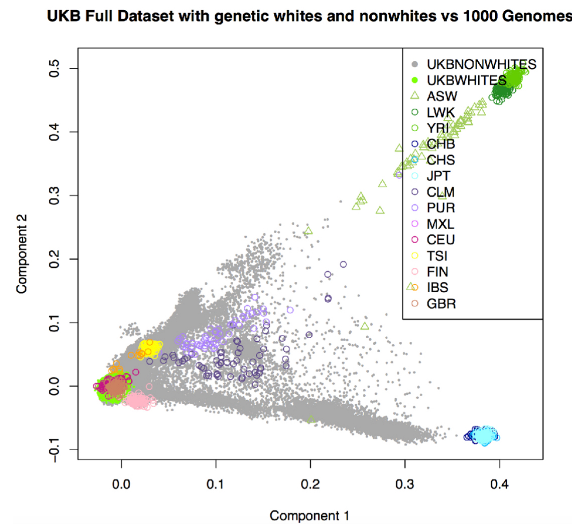
In the 23andMe replication analysis, samples were restricted to participants from European ancestry determined through an analysis of local ancestry.[13] A maximal set of unrelated participants was chosen for each analysis using a segmental identity-by-descent (IBD) estimation algorithm. Participants were defined as related if they shared more than 700 cM IBD, including regions where the two participants share either one or both genomic segments IBD. When selecting participants for case/ control phenotype analyses, the selection process was designed to always maximise case sample size by preferentially retaining cases over controls. Specifically, if both an individual case and an individual control were found to be related, then the case was retained in the analysis. Variant QC was implemented independently to genotyped and imputed GWAS results. The SNPs failing QC were flagged based on multiple criteria, namely HWE P-value, call rate, imputation R-square and test statistics of batch effects.

Figure 2.1. Overview of Quality Control. **A**) Flowchart summarising the quality control (QC) protocol. Excluded SNPs are in blue panels on the left and excluded participants are in green panels on the right. [1]Pre-QC exclusions: 3 participants with invalid IDs and sex, and 8 participants who have withdrawn from UK Biobank were excluded prior to QC. [2]Pre-association exclusions: 11 participants who were not present in UK Biobank's sample file accompanying the BGEN files were excluded prior to association. **B)** Principal Component Analysis (PCA) for demonstration of ethnicity of UK Biobank participants. The UK Biobank cohort was merged with publicly available data from the 1000 Genomes Project and PCA was performed using flashpca. Participants identified by UK Biobank as having white British ancestry are coloured in lime green, and the remaining UK Biobank participants are in grey. In this graph of principal component 1 vs principal component 2, a near-perfect overlap can be seen between the UK Biobank "white British" participants and both GBR (British in England and Scotland - light brown) and CEU (Utah residents with Northern and Western European ancestry - magenta) participants from the 1000 Genomes Project. Figure modified from Wiberg, *et al.*[11]

a



b



2.2.5. Imputation

UK Biobank autosome phasing was performed centrally using a 1000 Genomes Consortium[14] Phase 3 reference panel in SHAPEIT3[15], and is described in detail elsewhere.[9] The ~800,000 directly genotyped variants in the full UK Biobank cohort were imputed to 92,693,895 autosomal markers using a combination reference panel from the Haplotype Reference Consortium[9], the 1000 Genomes Project and the UK10K Project.[14,16] After the quality control checks (detailed in **2.2.4. Quality control**), low-quality (imputation INFO Score < 0.9) and rare (MAF < 0.01) variants from the imputation dataset were excluded, leaving ~9 million variants for the final discovery association analysis (discussed further in **2.2.6. Association analysis**).

In the 23andMe replication analysis, out-of-sample modified versions of the Beagle graph-based haplotype phasing algorithm[17] and Eagle2[18] algorithm were implemented to phase samples. The samples were imputed against a single unified imputation reference panel combining the 1000 Genomes Phase 3 haplotypes[14] with the UK10K project imputation reference panel[16] using Minimac3.[19]

2.2.6. Association analysis

In the UK Biobank discovery analysis, genome-wide association testing was performed across a total of 8,944,547 SNPs (547,011 directly genotyped (MAF ≥ 0 .01) and 8,397,536 imputed SNPs using a linear mixed non-infinitesimal model implemented in BOLT-LMM v2.323.[12] The reference human genome assembly used was GRCh37 (hg19) and linkage disequilibrium scores were obtained from participants of European-ancestry extracted from the BOLT-LMM package.[12] To

account for residual population structure, adjustments were included in the model for the covariates: genetic sex and genotyping platform. Association testing was implemented by linear regression assuming an additive allelic effect using imputed allelic dosages. BOLT-LMM[12] implements a linear regression and hence outputs a beta regression coefficient (with an associated standard error) for case-control outcomes - odds ratios for each variant were computed using the following equation:

$$\ln(OR) = \frac{\beta}{\mu(1-\mu)}$$

Where β is the beta regression coefficient calculated by BOLT LMM[12], and μ is the fraction of the cases in the sample (μ = 0.0593).

Conditional regression analysis was performed in BOLT-LMM for the top signal at each of the genome-wide significant loci from the association analysis (except the MHC region due to the high density of genes and high linkage of variants).[12] The conditional regression model regressed on genetic sex and genotyping platform, with the genotypic dosage of the top signal as calculated by QCTOOL 2.0 as a third covariate (see **2.2.15. URLs**). If any genome-wide significant signals remained at that locus, this process was repeated iteratively (by adding the genotypic dosage of the top residual signal as a further covariate), until no genome-wide significant signal remained.

In the 23andMe replication analysis, summary statistics were generated though logistic regression assuming an additive model for allelic effects. Association analysis

was performed by regressing on age, sex, the first five principal components, and the genotyping platform.

The lead 118 independent variants from the discovery analysis were tested for association with varicose veins in 23andMe. A Bonferroni-corrected $P < 4.24 \times 10^{-4}$ (0.05/118) was defined as the significant threshold for replication. Replication data for 108 of 118 variants were available in the 23andMe summary statistics: nine variants did not meet the SNP quality control within 23andMe, and one variant (10:79677281_CA_C) was not identified. A fixed-effects meta-analysis of each of the top independent variants was performed in Genome-Wide Association Meta-Analysis (GWAMA) software.[20]

2.2.7. Genomic risk loci definition

Risk loci interpretation of the genetic associations was performed within the platform FUMA GWAS v1.3.3 (*Functional Mapping and Annotation of Genome-Wide Association Studies,* see **2.1.15 URLs**).[21] Independent significant SNPs (IndSigSNPs) were first established by identifying all genome-wide significant SNPs ($P < 5 \times 10^{-8}$) and those that were independent from each other at $r^2 < 0.6$. These IndSigSNPs were then represented by lead SNPs, a proportion of IndSigSNPs that were in linkage equilibrium with each other at $r^2 < 0.1$. Lead SNPs were manually selected in FUMA as those independent signals emerging from the conditional regression analysis (Refer to **2.2.6. Association analysis**), and subsequent positional and eQTL gene mapping was based on these variants (see **2.2.9. Candidate gene mapping**). Independent genomic risk loci were identified as physical genomic regions around lead SNPs (Lead SNPs that were < 250kb apart were merged), with genome risk loci borders

established by identifying variants in linkage disequilibrium ($r^2 \geq 0.6$) of one of the IndSigSNPs. A single independent genomic risk locus was defined as the genomic region housing all variants (defined as candidate variants) in linkage with each IndSigSNP. Novel risk loci were those risk loci that had not been previously related to varicose veins in the NHGRI-EBI catalogue of published genome-wide association studies.[22]

2.2.8. Functional annotation of SNPs

Functional mapping and annotation of the associated SNPs was performed in FUMA *SNP2GENE* v1.3.6[21], an online platform that collates external databases to provide comprehensive annotation information for functional interpretation (**2.1.15 URLs**). For all analyses, the summary statistics from the UK Biobank discovery cohort were used with default settings. All candidate genes in each defined genomic risk locus in linkage disequilibrium with an IndSigSNP ($r^2 \geq 0.6$), a $P < 5\times10^{-2}$, and MAF > 0.01 were included in the FUMA *SNP2GENE* annotation. FUMA *SNP2GENE* uses each candidate variant's genomic location and effect/non-effect allele to collate functional annotation data from established genetic annotation databases[21], including ANNOVAR[23], RegulomeDB[24], CADD[25], and 15-core chromatin state categories.[26] ANNOVAR was queried to identify the gene location and function of candidate variants.[23] Combined Annotation Dependent Depletion (CADD) is a tool for scoring the deleteriousness of candidate variants and their likelihood to affect protein structure or function.[25] A CADD score ≥ 12.37 is the suggested threshold for a predicted pathogenic variant[27], with a scaled CADD score > 20 indicative of a variant in the top 1% of deleterious variants in the human genome, and a score > 30 indicating a variant in the top 0.1% of deleterious variants.[25] RegulomeDB is a resource that uses eQTL

data and chromatin marks to highlight DNA features and regulatory elements in non-coding genomic regions.[24] RegulomeDB scores are categorical and range from 1a to 7, with a score of 1a indicative of a variant with the highest possibility of affecting transcription factor binding and linked to expression of a gene target (i.e. EQTL + TF Binding + matched TF motif + matched DNase Footprint + DNase peak).[24] Chromatin state annotations refer to the accessibility of genomic regions to transcriptional effects. The 15-core chromatin state model (NIH Roadmap Epigenomes Consortium[26]), provides a reference epigenomic landscape of the human genome based on a core-set of 5 chromatin marks assayed across 127 epigenomes (H3K4me3, H3K4me1, H3K36me3, H3K27me3 and H3K9me3).[26] The 15-states capture key interactions between chromatin marks, with a score of 1 to 7 indicative of a region of open chromatin, and a score of 1 denoting a region with the highest accessibility (i.e. an active transcription start site).[26] Exonic SNPs were investigated further using gnomAD and Ensembl genome browsers to uncover non-synonymous functionality (Refer to **2.2.15. URLs**).[28]

2.2.9. Candidate gene mapping

To map candidate variants identified in the GWAS to genes, four gene mapping approaches were implemented - positional mapping[21], eQTL mapping[21], MAGMA genome-wide, gene association analysis (GWGAS)[29], and summary-based mendelian randomisation (SMR)[30]:

i) Positional mapping - the candidate variants at each locus were mapped to protein-coding genes that lay within a genomic window of 10kb on either side of the variant.

ii) eQTL mapping was used to map candidate variants within each locus to a gene if it had a genome-wide significant eQTL association ($P < 5\times10^{-8}$, FDR < 0.05) for that gene[*] in tibial artery tissue from the GTEx repository[21], based on having a cis-eQTL within 1MB of the gene. (Tibial artery tissue was chosen for this analysis as it was thought to be the closest surrogate for lower limb venous tissue out of all 53 tissue types present in GTEx).

iii) A MAGMA v1.07[29] genome-wide gene association test was implemented in FUMA *SNP2GENE*.[21] This mapped at least one SNP from the GWAS individually to 18,733 protein-coding genes obtained from Ensembl build 85.[28] To map variants to these genes across the genome, a strict Bonferroni correction was implemented to account for multiple-testing ($P < 2.67\times10^{-6}$ (i.e. 0.05/18733)). A Quantile-Quantile plot for the GWGAS was generated and genes whose P-values reached genome-wide significance were additionally labelled in a Manhattan plot.

iv) SMR was used to identify genes with expression levels associated with varicose veins due to pleiotropy.[30] The association between a gene's expression in tibial artery tissue (eQTL data taken from GTEx V7 tibial artery[21]) and varicose veins was analysed using the top-associated eQTL for each gene as a genetic instrument. SMR significant genes ($P_{SMR} < 0.05$/ number of probes) are those providing evidence of pleiotropy[†], but also of co-localisation[‡].[30] To examine for heterogeneity in SMR estimates and to untangle pleiotropy from co-localisation, a HEterogeneity In Dependent Instrument (HEIDI) test was implemented[30], with SNPs passing the HEIDI test (a $P_{HEIDI} < 0.05$/

---

[*]in other words, allelic variation at the SNP is associated with altered gene expression levels

[†]i.e. the expression of a gene and that of a trait are influenced by the same causal variant at the gene locus

[‡]possibility that SNPs controlling gene expression are in LD with those associated with the traits

number of P_{SMR}-significant Probes) associated with varicose veins through pleiotropy (rather than co-localisation); and therefore identifying genes whose expression levels mediate the association between SNPs and varicose veins.

2.2.10. Gene set, tissue and pathway analyses

Gene-set analysis was performed in MAGMA v1.07[29] (implemented in FUMA *SNP2GENE*[21]), with the full distribution of SNP P-values from the GWGAS analysis (described in **Section 2.2.9**) that lay positionally within the start and end points of a protein-coding gene (predefined distances were set to 0kb on both sides). Using competitive testing for gene set enrichment, 15,496 gene sets obtained from MSigDB v8.0[31] were tested (5500 curated gene sets and 9996 GO terms). Curated gene sets were derived from nine data sources[31], including KEGG, REACTOME and BioCarta and GO terms made up of three categories: biological processes, cellular components and molecular function. To account for multiple testing, enrichment across the gene sets was corrected to account for the number of gene sets tested ($P < 3.23\times10^{-6}$ (0.05/ 15496)). Enrichment of the overlap between GWAS variants and those reported in previous GWAS within the NIH GWAS Catalog were also examined[22], with enrichment P-values for the proportion of overlap in the genes determined.

MAGMA tissue expression analysis[29] was performed to test the relationship between GWAS associations (the full distribution of SNP p-values were used in the gene-property analysis) and highly expressed genes from individual tissues in the GTEx v8 30 general tissue types collection and 54 specific tissue types collection separately.[21] Gene property analysis was implemented using the averaged expression of genes in each tissue type as a covariate[29], and gene expression values depicted as log2-

transformed average RPKM per tissue type after winsorized at 50 based on the GTEx RNA-seq data.[21]

The above described gene set and tissue expression analyses were then repeated within FUMA *GENE2FUNC* v1.3.5d[21], to specifically hone in on the functionality of *only* the genes prioritised directly from the four candidate gene mapping approaches (i.e. a credible set of genes)[*] (described in **Section 2.2.9.**). Gene set enrichment analyses of the gene sets within MSigDB v8.0[31] were tested, and gene property and tissue enrichment analyses within GTEx consortium tissue was also performed distinctly for the prioritised varicose veins associated genes.[21]

Using eXploring Genomic Relations (XGR) software[32], pathway enrichment analysis of the prioritised genes was performed to highlight canonical pathways that were enriched. A hypergeometric distribution test was performed and adjusted FDR < 0.05 used to highlight prioritised gene sets. No restriction on overlap between the input genes was in place.

2.2.11. SNP-based heritability analysis

Using Linkage Disequilibrium Score (LDSC) regression[33], the LD intercept and mean chi-squared test score for the varicose veins GWAS was calculated, with the attenuation score calculated using the equation: (LDSC intercept - 1)/ (mean χ^2 - 1). LDSC was used to produce a SNP-based heritability estimate for varicose veins in UK

[*] i.e. genes *specifically* prioritised through one of the four described gene mapping approaches and *not* only from the GWGAS test performed in MAGMA (which is what FUMA *SNP2GENE* does) using the full distribution of P-values of variants throughout the genome to identify enriched candidate genes.

Biobank and 23andMe (h^2_g).[34] This approach derived the heritability for varicose veins by regressing each variants association statistic onto its LD Score*. h^2_g is a measure of genetic variance defined as the proportion of phenotypic variance explained by all or selected SNPs on a genotyping array. For the LDSC calculations[33], we harmonised the varicose veins GWAS summary statistics to include 1,170,823 variants that were well-imputed in the HapMap 3 panel and LD pruned ($r^2 < 0.1$) with long range LD regions removed to avoid capturing excess variance of LD regions. A two-step estimator cut-off of 30 was set to remove SNPs with large effect sizes.[33]

2.2.12. Genetic correlation analysis

Using the varicose veins summary statistics, we performed a genetic correlation analysis in the LD Hub database v1.9.3.[34] LD Hub is a centralised database of summary-level GWAS results for 832 diseases/traits from several publicly-available consortia.[34] 176 preselected traits across nine trait categories from the LDHub database were tested for correlation with varicose veins: metabolites, glycaemic traits, autoimmune diseases, anthropometric traits, smoking behaviour, lipids, cardiometabolic traits, reproductive traits and haematological traits. Trait categories were pre-defined based on associations in the literature. Genetic correlations (r_g) between the traits were defined by regression on each variant's Z-score product from the two phenotypes, against its LDSC.[33] To account for multiple testing, a Bonferroni correction of $P < 5.56\times10^{-3}$ (0.05/9) was applied.

---

* i.e. 'The sum of LD r^2 measured with all other SNPs'.

2.2.13. Drug-target enrichment analysis

Genes prioritised through the gene-mapping approaches were queried in the Open Targets Platform.[35] The Open Targets Platform is a comprehensive data integration resource for access to and visualisation of potential therapeutics targets with associated disorders.[35] Drug targets may be proteins, protein complexes or RNA molecules as identified by the Human Gene Nomenclature Committee (HGNC), with integration from Ensembl (protein-coding genes)[28]. Relationships between gene targets and diseases are collated by mapping to Experiment Factor Ontology (EFO) terms. [35] Open Targets Platform determines the tractability of proteins encoded by the prioritised genes to therapeutic targeting, or whether they are under investigation in clinical trials (data extracted from clinicaltrials.gov). The Platform summarises the available evidence for target-disease associations using a plethora of information for target and disease. Fisher's exact test was used to determine the overlap of varicose veins prioritised genes with pharmacologically active drug targets in several diseases, with a nominal $P < 5\times10^{-2}$ indicative of significance.[35]

2.2.14. Genetic risk score

Genetic risk score profiles for the UK Biobank cohort were calculate via a weighted genetic risk score (wGRS), based on the top independent variants at each replicated risk locus. The wGRS was compared between between six groups of participants from the GWAS: i) all cases vs all controls; ii) surgical cases vs non-surgical cases; iii) cases with ulceration vs cases with no ulceration. Surgical cases were defined as those with OPCS (*Office of Population Censuses and Surveys Classification of Interventions and Procedures*) or self-reported operative codes. Ulceration cases were

those that had a primary and/or secondary ICD-10 codes for varicose veins with ulceration (I83.0 and I83.2) (detailed in **2.2.4. Quality control**). The following formula was implemented[36]:

$$wGRS = \sum_{i=1}^{n} W_i X_i$$

where i is the lead SNP at each genomic risk locus, n is the total number of lead SNPs in the GWAS (n = 49), W_i is the weighting for each of the SNPs (the natural logarithm of the odds ratio for each effect allele), and X_i is the number of effect alleles each individual possesses for each SNP. Each subject's risk allele was used to compute a SNP dosage (QCTOOL v2). wGRS calculations and unpaired t-testing between the different subgroups was performed in R v3.3.1 (see URLs).

2.2.15. URLs

ANNOVAR, www.annovar.openbioinformatics.org/en/latest/; BOLT-LMM, www.data.broadinstitute.org/alkesgroup/BOLT-LMM/; CADD, cadd.gs.washington.edu/; Ensembl, www.ensembl.org/index.html; flashpca, github.com/gabraham/flashpca/; FUMA, www.fuma.ctglab.nl/; GERP, www.mendel.stanford.edu/SidowLab/downloads/gerp/; GnomAD, www.gnomad.broadinstitute.org/; GTEx Portal, www.gtexportal.org/home/; GWAMA, www.genomics.ut.ee/en/tools/gwama; Human Genome Variation Society (HGVS), www.varnomen.hgvs.org/; HRC, www.haplotype-reference-consortium.org/ ; LD Hub, www.ldsc.broadinstitute.org/ldhub/; LD Link, www.ldlink.nci.nih.gov/; MAGMA, www.ctg.cncr.nl/software/magma; Open Targets Platform, www.targetvalidation.org/;

PLINK, www.pngu.mgh.harvard.edu/~purcell/plink/; Polyphen-2, www.genetics.bwh.harvard.edu/pph2; QCTOOL,www.well.ox.ac.uk/~gav/qctool_v2/#overview; R, www.r-project.org; RegulomeDB, www.regulomedb.org/; SHAPEIT3, jmarchini.org/shapeit3/; SIFT, www.sift.bii.a-star.edu.sg/; UK Biobank, www.ukbiobank.ac.uk/; XGR, www.galahad.well.ox.ac.uk:3040; 1000 Genomes Project, www.1000genomes.org; 23andMe, https://research.23andme.com/

2.3. Results

2.3.1. Forty-six replicated varicose veins susceptibility loci

The overall two-stage association analysis workflow is provided in **Figure 2.2**. The discovery cohort consisted of 22,473 cases and 379,183 controls of white British ancestry from the UK Biobank dataset. Association testing yielded genome-wide significant associations at 109 risk loci (12,391 variants). A further nine independent signals at eight loci were identified through conditional regression analysis. The λ_{GC} demonstrated inflation (1.25), with the LDSC regression intercept (1.06) and attenuation ratio (0.13) in keeping with the expectations of polygenicity and large sample size (**Figure 2.3-A**).[33]

The 118 lead independent signals at the 109 risk loci were tested in the 23andMe association analysis consisting of 113,041 self-reported varicose vein cases and 295,928 controls. Here again, the LDSC intercept demonstrated moderate inflation ($\lambda = 1.13$, S.E. = 0.01). Forty-nine of 118 variants demonstrated significant association with varicose veins at a Bonferroni-corrected threshold of $P < 4.24\times10^{-4}$ (**Table 2.2**). Thus, in total 49 independent significant associations at 46 risk loci were identified (**Figure 2.3-B**; regional Locus Zoom plots for all 49 signals are presented in **Figure 2.4).** Across both cohorts, allelic effects were concordant at all 49 replicated variants, with minimal evidence of heterogeneity between the two association studies at all loci (Q-statistic > 0.05). Eighteen of the 45 risk loci were previously reported, and 28 are novel (Table 1). Sixty-nine variants (at 63 risk loci) did not replicate, and are therefore not included in subsequent post-association analyses. These can be found in **Appendix Table 2.1**.

Figure 2.2. Varicose veins GWA study design and analysis workflow. A two-stage GWAS conducted in UK Biobank, with replication of the lead independent variants variants within the 23andMe replication cohort. Of the 118 tested variants, data on 117 variants were available for replication in the 23andMe Cohort, of which 108 passed QC within the replication cohort (see **2.2.6. Association Analysis**). Forty-nine independent variants at 46 loci met the Bonferroni-corrected threshold in the replication cohort, and subsequently were interrogated further in multiple analyses.

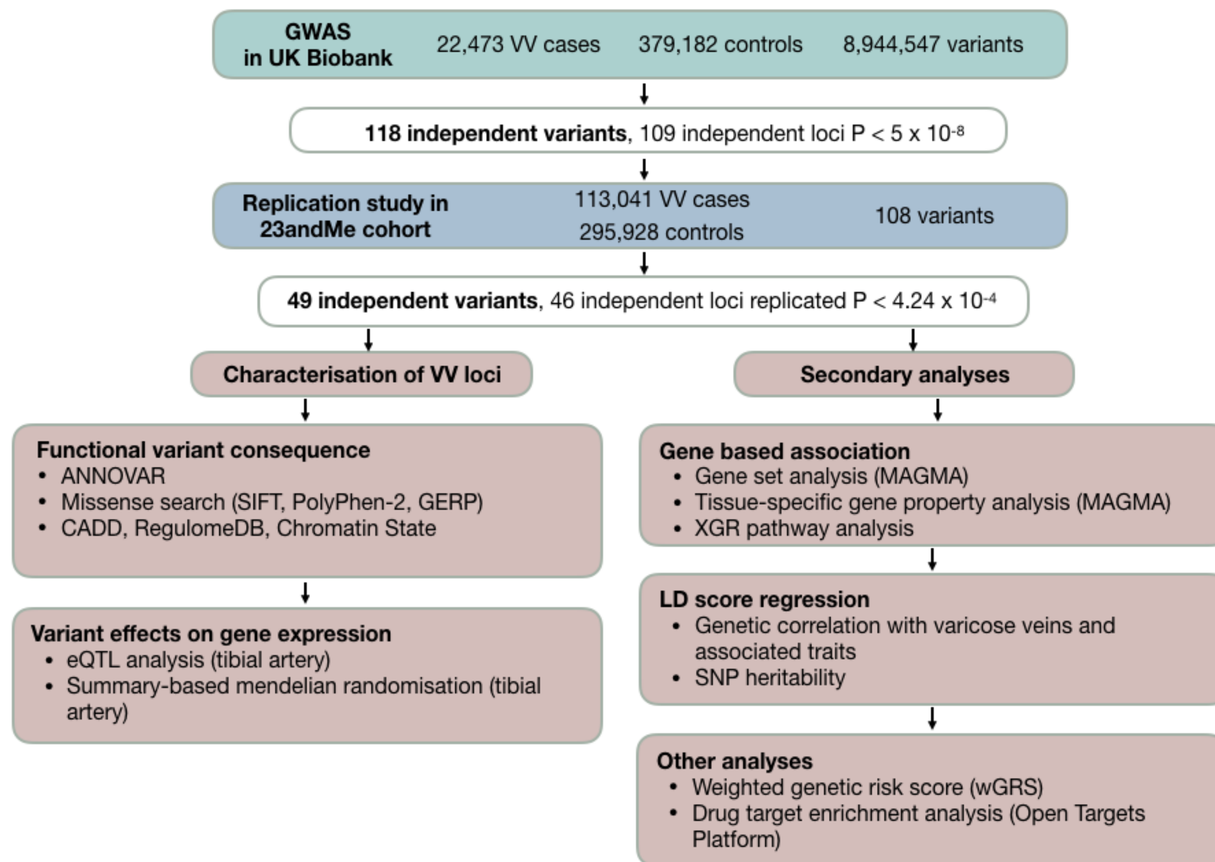


Table 2.2. Forty-nine significant variants at 46 susceptibility loci associated with varicose veins in a two-stage GWAS of 135,514 cases and 675,111 controls from the UK Biobank and 23andMe, Inc.

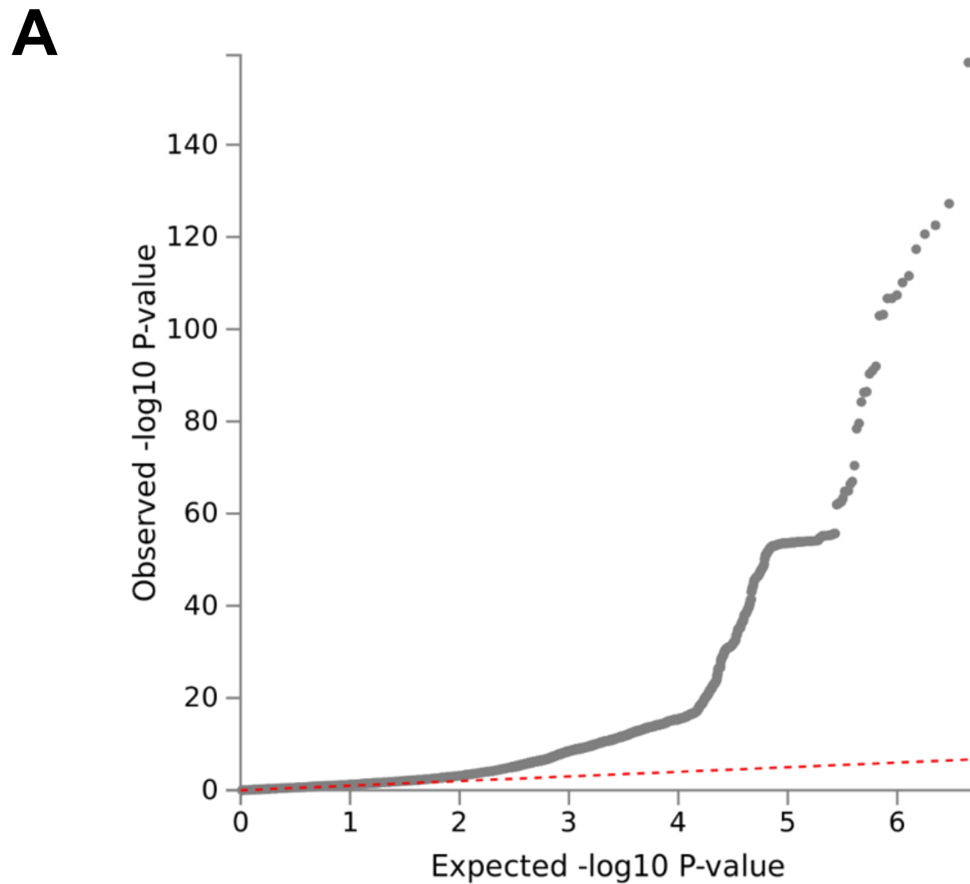
SNP					Discovery GWAS in UK Biobank				Replication GWAS in 23andMe				Meta-Analysis		
Chr	**SNP**	**Position[*]**	**EA[†]**	**NEA[‡]**	**EAF[§]**	**INFO[‖]**	**OR[#]**	**P-Value**	**EAF[§]**	**INFO[‖]**	**OR[#]**	**P-Value**	**OR[#]**	**P-Value**	**Candidate genes**
1	rs11121615	10825577	C	T	0.31	0.984	1.32 (1.30-1.35)	1.40×10^{-158}	0.32	0.944	1.57 (1.48-1.67)	2.72×10^{-50}	1.35 (1.32-1.38)	6.95×10^{-201}	*CASZ1*
1	**rs7518191**	**115603413**	**T**	**C**	**0.30**	**0.986**	**1.07 (1.05-1.09)**	$\mathbf{2.30\times10^{-10}}$	**0.31**	**0.998**	**1.14 (1.08-1.21)**	$\mathbf{5.72\times10^{-6}}$	**1.08 (1.06-1.10)**	$\mathbf{6.90\times10^{-14}}$	***TSPAN2***
1	**rs17712208[**]**	**214150445**	**A**	**T**	**0.04**	**G**	**1.13 (1.07-1.19)**	$\mathbf{3.00\times10^{-6}}$	**0.03**	**0.687**	**1.50 (1.25-1.81)**	$\mathbf{1.71\times10^{-5}}$	**1.15 (1.10-1.21)**	$\mathbf{1.68\times10^{-8}}$	***PROX1***
1	**rs340875**	**214158986**	**G**	**C**	**0.48**	**0.993**	**1.07 (1.05-1.09)**	$\mathbf{2.30\times10^{-11}}$	**0.50**	**0.993**	**1.27 (1.20-1.34)**	$\mathbf{3.22\times10^{-18}}$	**1.09 (1.07-1.11)**	$\mathbf{4.22\times10^{-20}}$	***PROX1***
1	**rs2820464**	**219693220**	**A**	**G**	**0.34**	**0.997**	**1.07 (1.04-1.09)**	$\mathbf{3.60\times10^{-10}}$	**0.32**	**0.997**	**1.16 (1.09-1.22)**	$\mathbf{6.91\times10^{-7}}$	**1.07 (1.05-1.10)**	$\mathbf{4.48\times10^{-14}}$	**-**
2	rs9967884	30488183	A	G	0.77	0.995	1.10 (1.07-1.12)	1.60×10^{-15}	0.78	0.986	1.20 (1.12-1.28)	4.99×10^{-8}	1.11 (1.08-1.13)	1.40×10^{-20}	*LBH*
2	rs3791679	56096892	A	G	0.77	G	1.07 (1.04-1.09)	1.90×10^{-8}	0.76	G	1.22 (1.15-1.30)	4.42×10^{-10}	1.08 (1.06-1.11)	1.59×10^{-13}	*EFEMP1*
2	rs2861819	68489221	C	G	0.66	0.996	1.17 (1.15-1.20)	2.10×10^{-56}	0.65	0.925	1.40 (1.32-1.49)	1.04×10^{-29}	1.20 (1.17-1.22)	2.65×10^{-77}	*PPP3R1*
2	**rs4849044**	**112898933**	**C**	**T**	**0.51**	**0.991**	**1.07 (1.05-1.09)**	$\mathbf{2.00\times10^{-11}}$	**0.50**	**0.990**	**1.12 (1.06-1.18)**	$\mathbf{5.78\times10^{-5}}$	**1.07 (1.05-1.09)**	$\mathbf{1.98\times10^{-14}}$	***FBLN7***
2	**rs17819430**	**118886398**	**C**	**A**	**0.96**	**0.991**	**1.15 (1.10-1.21)**	$\mathbf{5.00\times10^{-9}}$	**0.96**	**0.979**	**1.40 (1.22-1.61)**	$\mathbf{1.63\times10^{-6}}$	**1.18 (1.13-1.23)**	$\mathbf{1.51\times10^{-12}}$	***INSIG2***
2	**rs55889669**	**173194747**	**A**	**G**	**0.19**	**0.994**	**1.08 (1.05-1.11)**	$\mathbf{3.60\times10^{-10}}$	**0.19**	**0.977**	**1.20 (1.12-1.29)**	$\mathbf{2.62\times10^{-7}}$	**1.09 (1.07-1.12)**	$\mathbf{2.89\times10^{-14}}$	**-**
3	**rs844176**	**14827080**	**G**	**A**	**0.89**	**0.974**	**1.11 (1.07-1.14)**	$\mathbf{1.60\times10^{-10}}$	**0.89**	**0.966**	**1.18 (1.08-1.29)**	$\mathbf{1.81\times10^{-4}}$	**1.11 (1.08-1.15)**	$\mathbf{3.39\times10^{-13}}$	***FGD5***
3	rs2713575	128294355	G	A	0.50	0.980	1.11 (1.09-1.13)	7.30×10^{-28}	0.50	0.999	1.21 (1.15-1.28)	5.02×10^{-12}	1.12 (1.10-1.14)	1.82×10^{-36}	*GATA2*
3	**rs9877579**	**188058716**	**C**	**T**	**0.26**	**0.986**	**1.07 (1.05-1.10)**	$\mathbf{7.90\times10^{-11}}$	**0.26**	**0.983**	**1.14 (1.07-1.21)**	$\mathbf{3.21\times10^{-5}}$	**1.08 (1.06-1.10)**	$\mathbf{5.84\times10^{-14}}$	***LPP***
4	rs28558138	26818080	G	C	0.58	0.979	1.14 (1.12-1.16)	2.00×10^{-40}	0.58	0.890	1.21 (1.14-1.28)	8.71×10^{-11}	1.15 (1.13-1.17)	9.35×10^{-49}	*TBC1D19*
4	**rs56155140**	**57824451**	**A**	**G**	**0.19**	**G**	**1.09 (1.07-1.12)**	$\mathbf{6.50\times10^{-13}}$	**0.19**	**0.996**	**1.21 (1.13-1.29)**	$\mathbf{5.17\times10^{-8}}$	**1.11 (1.08-1.13)**	$\mathbf{8.30\times10^{-18}}$	***IGFBP7***
4	**rs1471251**	**87976359**	**A**	**T**	**0.60**	**0.993**	**1.06 (1.04-1.08)**	$\mathbf{4.50\times10^{-8}}$	**0.61**	**0.987**	**1.12 (1.06-1.18)**	$\mathbf{5.31\times10^{-5}}$	**1.06 (1.04-1.08)**	$\mathbf{8.33\times10^{-11}}$	***AFF1***

4	**rs34154818**	**89726823**	**A**	**AT**	**0.55**	**0.980**	**1.05 (1.03-1.08)**	**4.90×10^{-8}**	**0.53**	**0.951**	**1.14 (1.08-1.21)**	**2.25×10^{-6}**	**1.06 (1.04-1.08)**	**2.13×10^{-11}**	***FAM13A***
4	**rs10007409**	**120142306**	**C**	**T**	**0.69**	**0.999**	**1.06 (1.04-1.08)**	**3.20×10^{-8}**	**0.68**	**0.990**	**1.21(1.14-1.28)**	**1.12×10^{-10}**	**1.07 (1.05-1.10)**	**2.09×10^{-13}**	***USP53***
4	**rs11728719**	**186696172**	**A**	**C**	**0.76**	**0.978**	**1.08 (1.05-1.10)**	**3.00×10^{-11}**	**0.77**	**0.958**	**1.15 (1.08-1.23)**	**1.63×10^{-5}**	**1.09 (1.06-1.11)**	**1.65×10^{-14}**	***SORBS2***
5	**rs57253948**	**38754162**	**G**	**A**	**0.06**	**G**	**1.12 (1.07-1.16)**	**3.70×10^{-8}**	**0.06**	**0.989**	**1.32 (1.18-1.47)**	**1.15×10^{-6}**	**1.14 (1.09-1.18)**	**1.05×10^{-11}**	**-**
5	rs3749748	127350549	T	C	0.25	0.992	1.16 (1.13-1.18)	5.60×10^{-39}	0.24	0.964	1.42 (1.33-1.51)	3.69×10^{-27}	1.18 (1.16-1.21)	1.06×10^{-56}	*FBN2, SLC12A2*
5	rs11135046	158230013	G	T	0.46	0.995	1.12 (1.10-1.14)	1.60×10^{-32}	0.45	0.992	1.16 (1.10-1.23)	5.81×10^{-8}	1.13 (1.11-1.15)	1.33×10^{-38}	*EBF1*
6	rs7773004	26267755	A	G	0.51	0.998	1.09 (1.07-1.11)	6.00×10^{-20}	0.50	0.984	1.18 (1.12-1.25)	1.38×10^{-9}	1.10 (1.08-1.12)	2.33×10^{-26}	*HFE*
6	**rs11967262**	**43760327**	**C**	**G**	**0.51**	**0.996**	**1.06 (1.04-1.08)**	**5.20×10^{-9}**	**0.51**	**0.990**	**1.34 (1.27-1.42)**	**8.13×10^{-27}**	**1.09 (1.07-1.11)**	**1.45×10^{-19}**	***VEGFA***
6	rs1936800	127436064	C	T	0.47	G	1.06 (1.04-1.08)	2.60×10^{-9}	0.49	0.986	1.20 (1.13-1.26)	7.26×10^{-11}	1.07 (1.05-1.09)	8.29×10^{-15}	*RSPO3*
8	**rs34022079**	**6648676**	**C**	**T**	**0.64**	**0.975**	**1.07 (1.05-1.09)**	**8.20×10^{-11}**	**0.63**	**0.888**	**1.62 (1.52-1.71)**	**4.59×10^{-58}**	**1.11 (1.09-1.14)**	**1.61×10^{-29}**	**-**
8	rs10504825	87567848	C	A	0.41	G	1.07 (1.05-1.09)	7.20×10^{-13}	0.41	0.997	1.14 (1.08-1.21)	1.55×10^{-6}	1.08 (1.06-1.10)	6.51×10^{-17}	*CPNE3*
9	**rs78216177**	**232148**	**C**	**G**	**0.14**	**0.992**	**1.10 (1.07-1.13)**	**3.70×10^{-11}**	**0.14**	**0.988**	**1.16 (1.08-1.25)**	**1.28×10^{-4}**	**1.10 (1.08-1.13)**	**5.80×10^{-14}**	***DOCK8***
9	**rs753085**	**117045447**	**G**	**A**	**0.73**	**0.994**	**1.06 (1.04-1.09)**	**1.50×10^{-8}**	**0.73**	**0.995**	**1.14 (1.07-1.21)**	**4.11×10^{-5}**	**1.07 (1.05-1.09)**	**2.17×10^{-11}**	***COL27A1***
9	rs10817762	118161597	A	C	0.52	0.994	1.08 (1.06-1.10)	1.80×10^{-16}	0.52	0.994	1.14 (1.08-1.20)	2.05×10^{-6}	1.09 (1.07-1.11)	1.02×10^{-20}	*TNC*
10	**rs61863928**	**64449549**	**G**	**T**	**0.62**	**0.955**	**1.06 (1.04-1.08)**	**3.00×10^{-8}**	**0.64**	**0.887**	**1.37 (1.29-1.45)**	**4.22×10^{-25}**	**1.09 (1.07-1.11)**	**1.51×10^{-17}**	**-**
11	**rs79465012**	**128258136**	**C**	**T**	**0.93**	**G**	**1.13 (1.09-1.17)**	**9.70×10^{-11}**	**0.93**	**0.842**	**1.28 (1.14-1.44)**	**2.46×10^{-5}**	**1.14 (1.10-1.18)**	**1.08×10^{-13}**	**-**
12	**rs7308356**	**50539611**	**G**	**A**	**0.63**	**0.997**	**1.08 (1.06-1.11)**	**4.10×10^{-16}**	**0.62**	**0.997**	**1.14 (1.08-1.21)**	**2.93×10^{-6}**	**1.09 (1.07-1.11)**	**3.02×10^{-20}**	***CERS5***
12	**rs1054852**	**124496316**	**G**	**A**	**0.38**	**0.904**	**1.06 (1.04-1.08)**	**1.60×10^{-8}**	**0.37**	**0.927**	**1.29 (1.21-1.36)**	**1.44×10^{-17}**	**1.08 (1.06-1.11)**	**2.87×10^{-16}**	***DNAH10OS***
13	**rs41286076**	**73634859**	**T**	**C**	**0.26**	**0.998**	**1.08 (1.05-1.10)**	**1.10×10^{-11}**	**0.25**	**0.977**	**1.14 (1.07-1.21)**	**5.64×10^{-5}**	**1.08 (1.06-1.11)**	**1.06×10^{-14}**	***KLF5***
14	**rs72683923**	**50735947**	**C**	**T**	**0.02**	**G**	**1.22 (1.14-1.30)**	**1.40×10^{-8}**	**0.02**	**0.721**	**2.38 (1.90-2.99)**	**1.06×10^{-13}**	**1.28 (1.20-1.37)**	**3.62×10^{-14}**	***CDKL1***
15	rs11852492	96167544	T	C	0.84	0.999	1.12 (1.09-1.15)	8.90×10^{-18}	0.83	0.994	1.20 (1.11-1.29)	1.49×10^{-6}	1.13 (1.10-1.15)	3.30×10^{-22}	-
16	**rs11076178**	**57146402**	**T**	**C**	**0.11**	**0.986**	**1.09 (1.06-1.12)**	**2.10×10^{-8}**	**0.11**	**0.905**	**1.18 (1.08-1.28)**	**3.16×10^{-4}**	**1.10 (1.07-1.13)**	**1.05×10^{-10}**	***CPNE2***

16	rs111350029**	88796770	G	GGGA GGC	0.14	0.910	1.24 (1.20-1.27)	1.90×10^{-48}	0.14	0.810	1.57 (1.44-1.71)	7.82×10^{-25}	1.27 (1.23-1.30)	7.39×10^{-66}	*PIEZO1, GALNS*
16	rs11646394**	88812279	C	A	0.87	0.995	1.19 (1.16-1.23)	2.30×10^{-34}	0.88	0.924	1.47 (1.35-1.61)	5.00×10^{-18}	1.22 (1.18-1.25)	3.62×10^{-46}	*PIEZO1, GALNS*
16	rs2002833	88842117	G	A	0.33	0.988	1.19 (1.17-1.22)	1.10×10^{-65}	0.31	0.988	1.43 (1.35-1.52)	8.55×10^{-34}	1.22 (1.19-1.24)	2.47×10^{-90}	*PIEZO1, GALNS*
17	**rs6503321**	**2096580**	**A**	**G**	**0.38**	**0.999**	**1.06 (1.04-1.08)**	$\mathbf{3.80\times10^{-8}}$	**0.37**	**0.990**	**1.14 (1.08-1.21)**	$\mathbf{3.28\times10^{-6}}$	**1.07 (1.05-1.08)**	$\mathbf{1.77\times10^{-11}}$	***SMG6***
17	rs638538	68216128	A	C	0.27	0.979	1.11 (1.09-1.14)	6.90×10^{-23}	0.28	0.994	1.19 (1.12-1.26)	1.79×10^{-8}	1.12 (1.10-1.14)	5.85×10^{-29}	*KCNJ2*
17	rs9895127	70029808	T	C	0.43	0.987	1.10 (1.08-1.12)	6.40×10^{-22}	0.44	0.989	1.16 (1.10-1.22)	1.48×10^{-7}	1.11 (1.09-1.13)	2.88×10^{-27}	*AC007461.1*
19	**rs12609241**†**	**16360926**	**G**	**A**	**0.75**	**0.990**	**1.07 (1.05-1.10)**	$\mathbf{3.00\times10^{-10}}$	**0.75**	**0.950**	**1.33 (1.25-1.42)**	$\mathbf{1.84\times10^{-18}}$	**1.10 (1.08-1.12)**	$\mathbf{1.32\times10^{-18}}$	***KLF2***
20	rs3787184	50157837	A	G	0.83	0.978	1.16 (1.13-1.19)	3.10×10^{-32}	0.82	0.949	1.17 (1.09-1.26)	1.32×10^{-5}	1.16 (1.14-1.19)	2.51×10^{-36}	*NFATC2*
20	**rs76602912**	**57459868**	**T**	**C**	**0.98**	**G**	**1.25 (1.18-1.33)**	$\mathbf{7.50\times10^{-13}}$	**0.97**	**0.857**	**1.52 (1.26-1.83)**	$\mathbf{1.41\times10^{-5}}$	**1.28 (1.20-1.36)**	$\mathbf{3.56\times10^{-16}}$	***GNAS***
20	rs6062619	62683002	A	G	0.73	0.952	1.10 (1.08-1.12)	1.90×10^{-17}	0.73	0.824	1.27 (1.19-1.36)	4.97×10^{-13}	1.11 (1.09-1.14)	5.48×10^{-25}	*SOX18*

*Based on NCBI Genome Build 37 (hg19). †The effect allele. ‡The alternate (non-effect) allele. §The effect allele frequency in the study population. ||The imputation quality score; G= genotyped SNP. #Odds ratio (95% confidence intervals). OR > 1 indicative of increased risk with effect allele. **denotes four residual significant signals following conditional regression analysis at the lead SNP. **†At this locus, 19p13.11, the lead SNP in the UK Biobank cohort was rs451367 ($P_{discovery} = 2.10\times10^{-10}$), however, this did not replicate in 23andMe ($P_{replication} = 0.47$; Appendix Table 2.1) - the independent residual signal, rs12609241 shown here, did however replicate. Bold variants represent loci not previously reported.

Figure 2.3. Results of genome-wide association study in varicose veins. A) Quantile-Quantile plot of observed vs. expected P-values for the association analysis for varicose veins. B) Manhattan plot showing genome-wide P-values plotted against position on each of the autosomes. The dark blue, light blue, and green dots refer to the discovery cohort in UK Biobank, with the red dots corresponding to the forty-nine variants from the 23andMe cohort at each replicated locus. The dark blue peaks correspond to the 46 loci that replicated in the 23andMe cohort at a Bonferroni-corrected threshold of $P < 4.24\times10^{-4}$. Candidate genes at each locus are named above each signal, with newly discovered genetic loci in blue, and previously described loci in black.



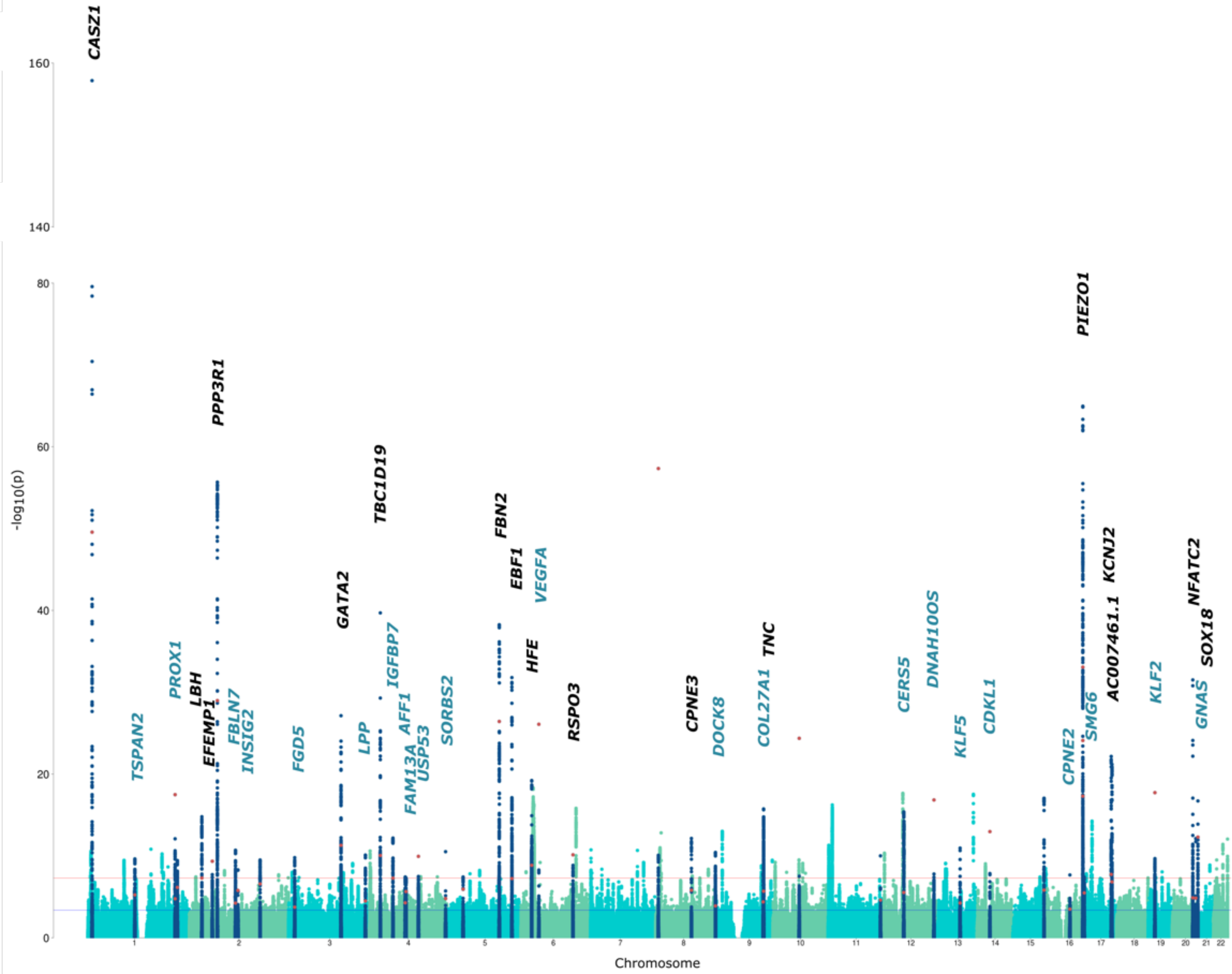
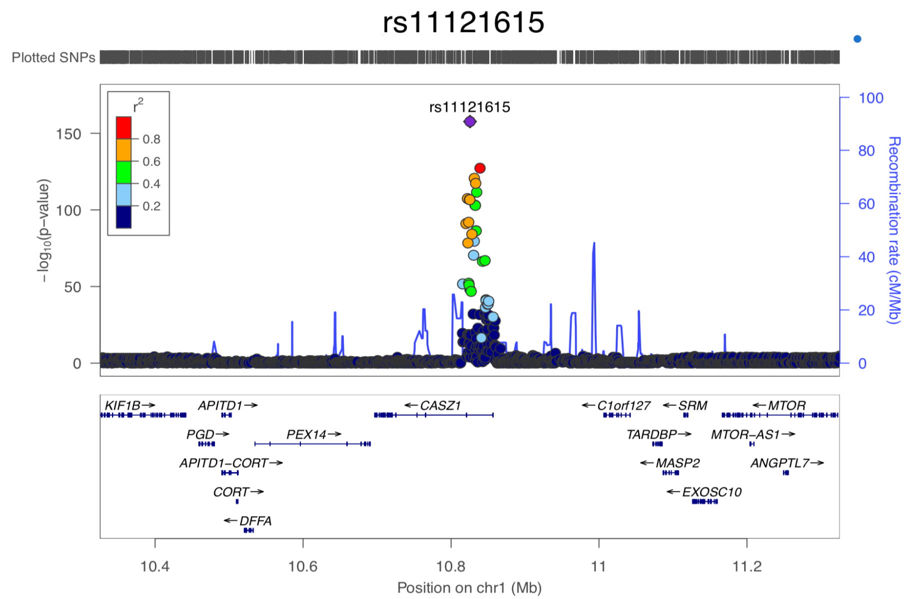
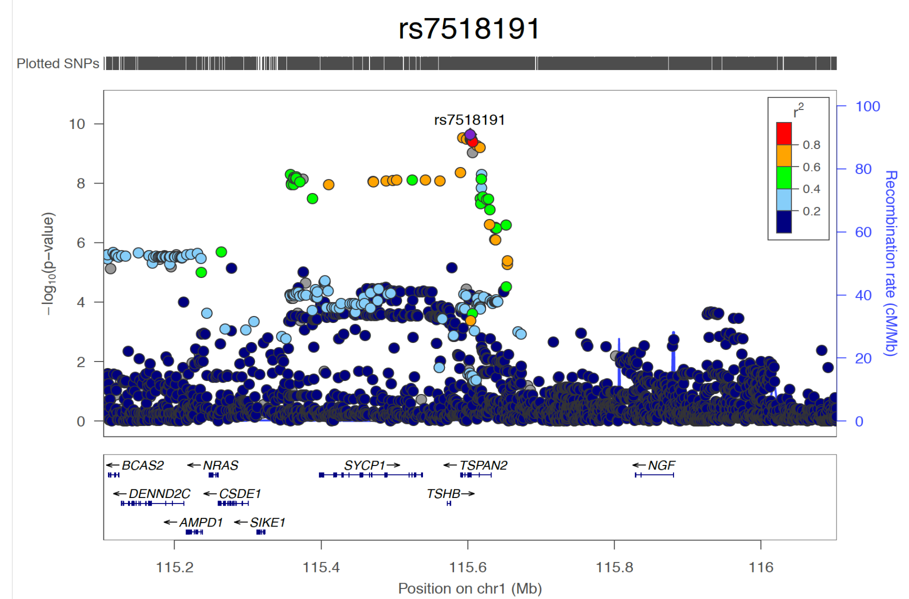
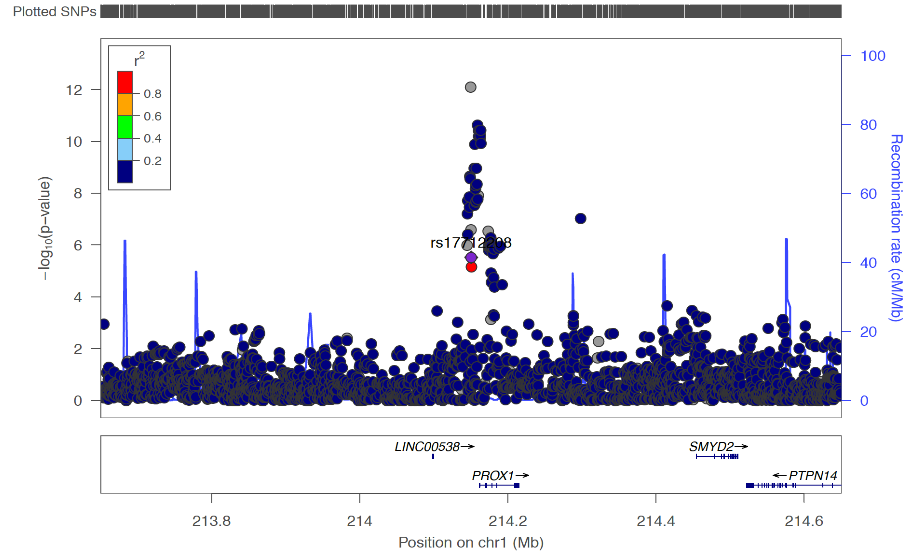
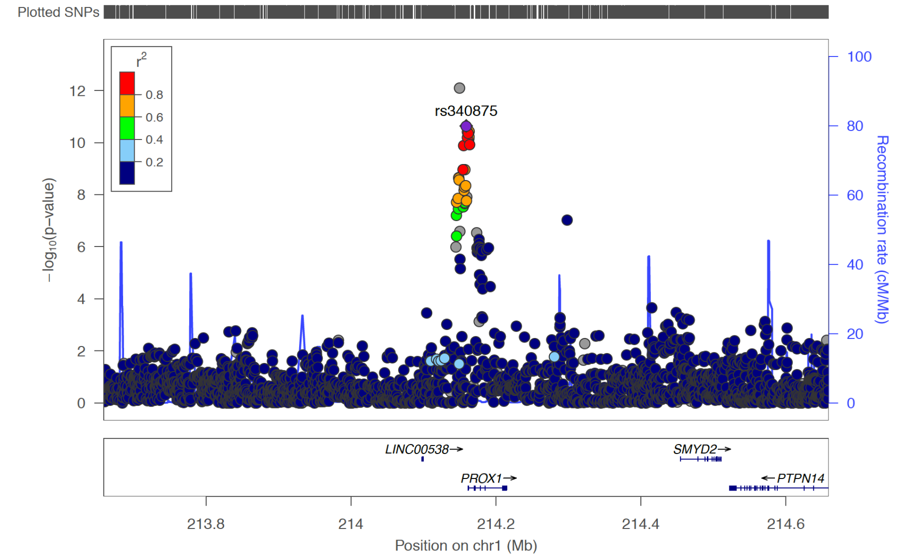

B
-log10(p)
Chromosome
CASZ1
PPP3R1
TBC1D19
FBN2
EBF1
VEGFA
GATA2
PIEZO1
KCNJ2
NFATC2
AC007461.1
SOX18
TNC
DNAH10OS
CERS5
PROX1
LBH
EFEMP1
FBLN7
INSIG2
TSPAN2
FGD5
LPP
IGFBP7
AFF1
FAM13A
USP53
SORBS2
HFE
RSPO3
CPNE3
DOCK8
COL27A1
KLF5
CDKL1
CPNE2
SMG6
KLF2
GNAS

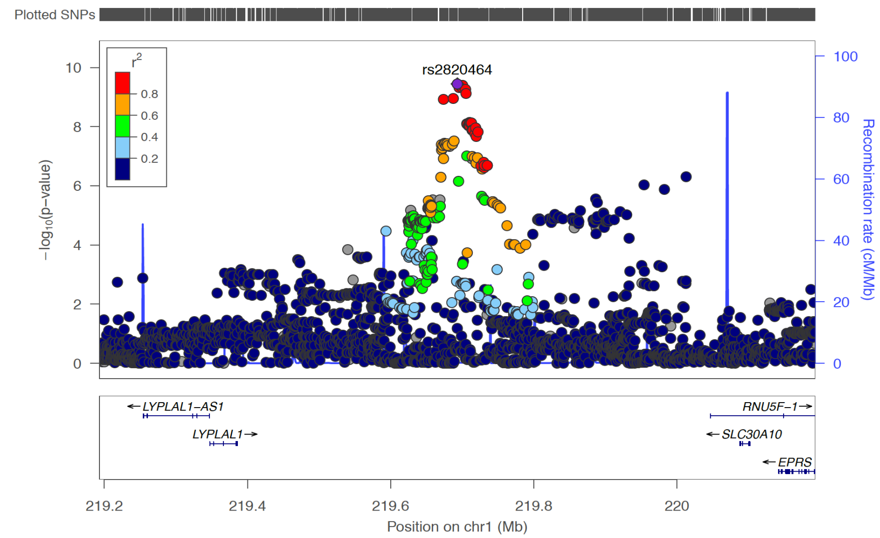
Figure 2.4. Regional Locus Zoom plots of all varicose veins associated signals. LocusZoom plots of the 49 independent genome-wide significant variants at the 46 replicated varicose veins associated susceptibility loci. Plots are ordered by chromosome number and genomic position. SNP position is shown on the x-axis, and strength of association on the y-axis ($-\log_{10}$ P-value). The linkage disequilibrium (LD) relationship between the lead SNP and the surrounding SNPs is indicated by the r^2 legend. In the lower panel of each sub-figure, genes within 500kb on either side of the index SNP are shown. The position on each chromosome is depicted in relation to Human Genome build hg19 (GRCh37). Note: variants rs34154818 and rs111350029 did not exist in the 1000 genomes reference panel and therefore r^2 values could not be generated

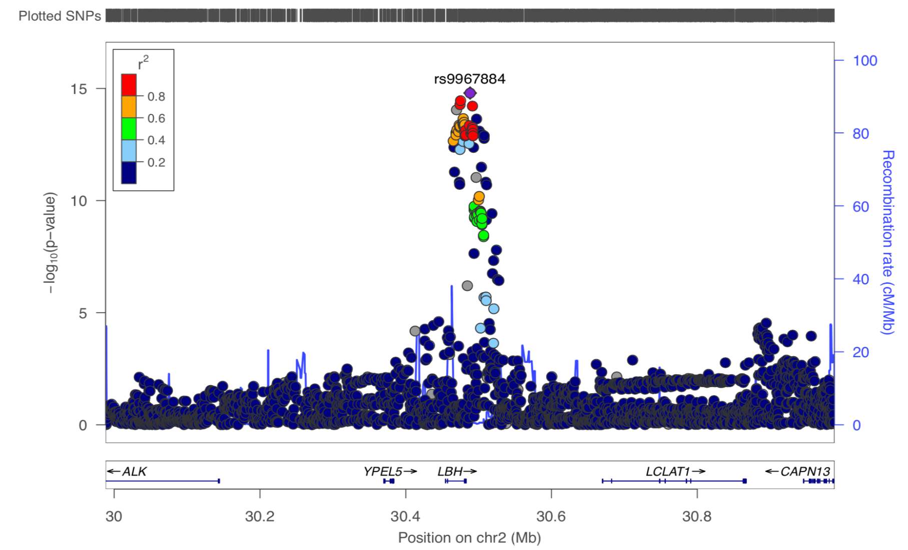


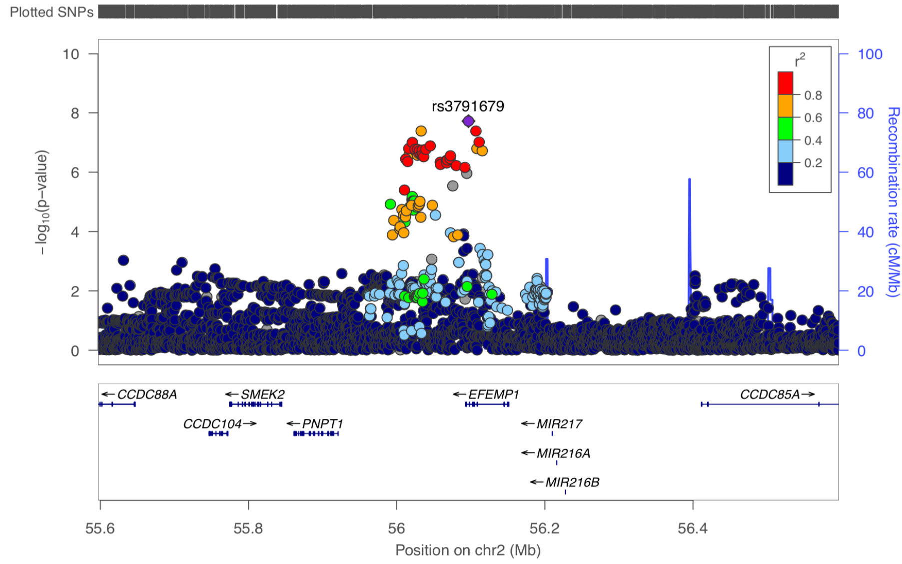



rs17712208
Plotted SNPs
r2
0.8
0.6
0.4
0.2
-log10(p-value)
Recombination rate (cM/Mb)
LINC00538
PROX1
SMYD2
PTPN14
213.8
214
214.2
214.4
214.6
Position on chr1 (Mb)

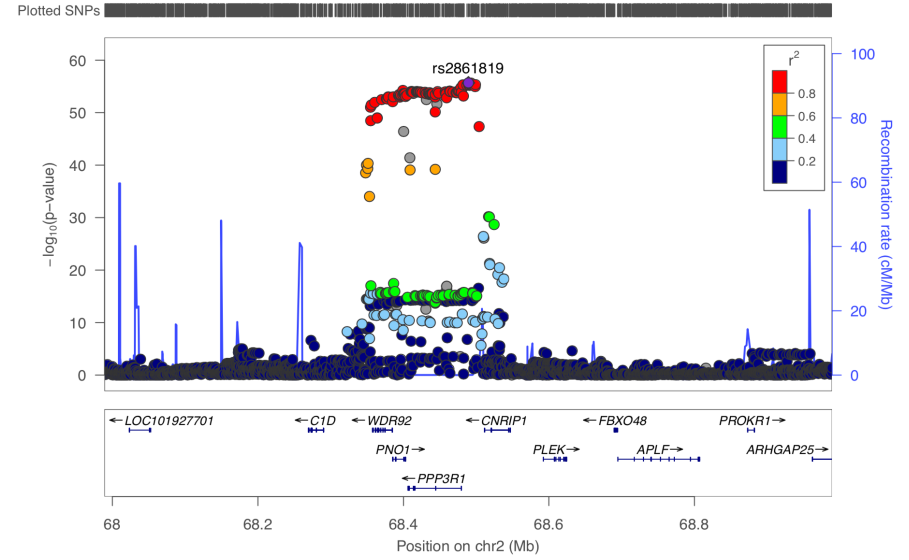

rs340875
Plotted SNPs
r2
0.8
0.6
0.4
0.2
-log10(p-value)
Recombination rate (cM/Mb)
LINC00538
PROX1
SMYD2
PTPN14
213.8
214
214.2
214.4
214.6
Position on chr1 (Mb)


rs2820464
Plotted SNPs
r2
0.8
0.6
0.4
0.2
-log10(p-value)
Recombination rate (cM/Mb)
LYPLAL1-AS1
LYPLAL1
RNU5F-1
SLC30A10
EPRS
219.2
219.4
219.6
219.8
220
Position on chr1 (Mb)

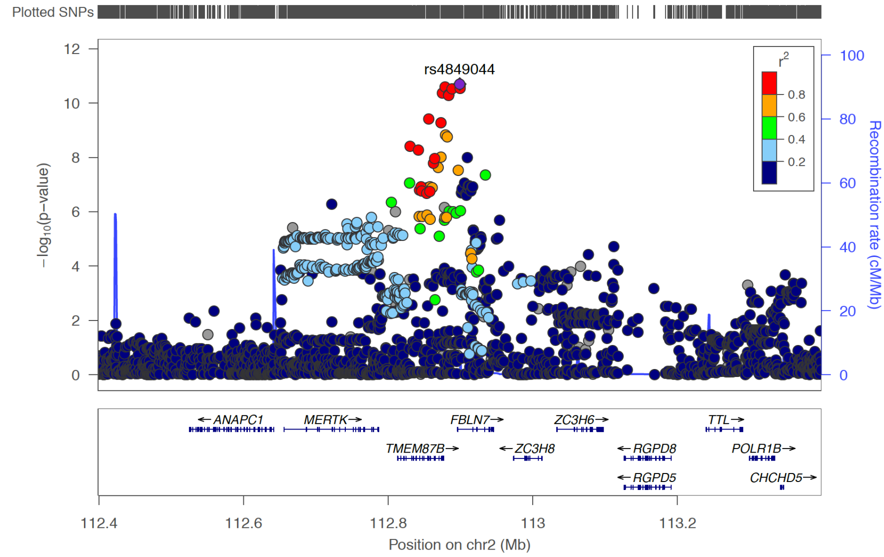

rs9967884
Plotted SNPs
r2
0.8
0.6
0.4
0.2
-log10(p-value)
Recombination rate (cM/Mb)
ALK
YPEL5
LBH
LCLAT1
CAPN13
30
30.2
30.4
30.6
30.8
Position on chr2 (Mb)



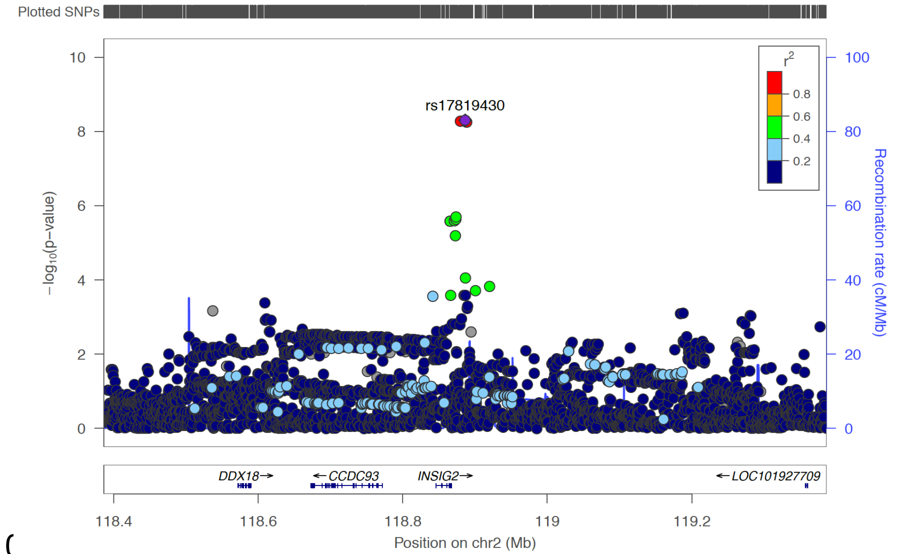
rs3791679
Plotted SNPs
rs3791679
−log10(p−value)
Recombination rate (cM/Mb)
r2
0.8
0.6
0.4
0.2
CCDC88A
SMEK2
EFEMP1
CCDC85A
CCDC104
PNPT1
MIR217
MIR216A
MIR216B
55.6
55.8
56
56.2
56.4
Position on chr2 (Mb)



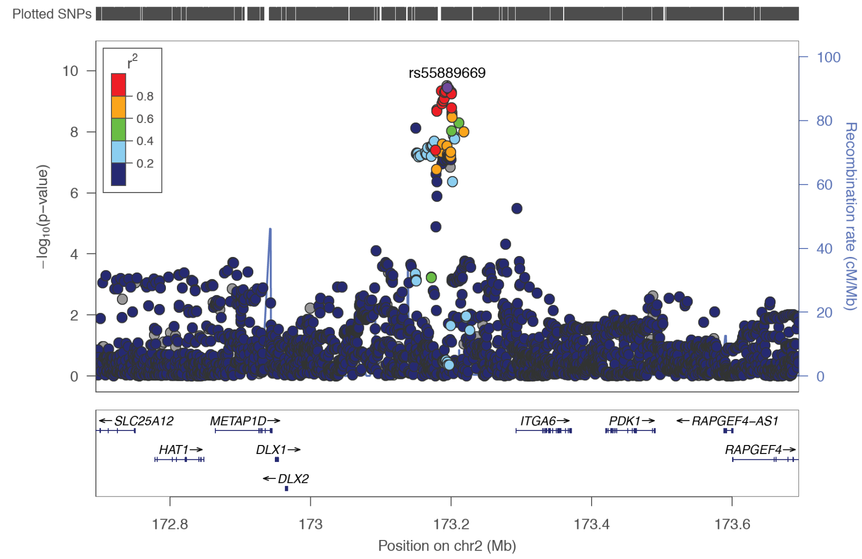
rs2861819
Plotted SNPs
rs2861819
−log10(p−value)
Recombination rate (cM/Mb)
r2
0.8
0.6
0.4
0.2
LOC101927701
C1D
WDR92
CNRIP1
FBXO48
PROKR1
PNO1
PLEK
APLF
ARHGAP25
PPP3R1
68
68.2
68.4
68.6
68.8
Position on chr2 (Mb)

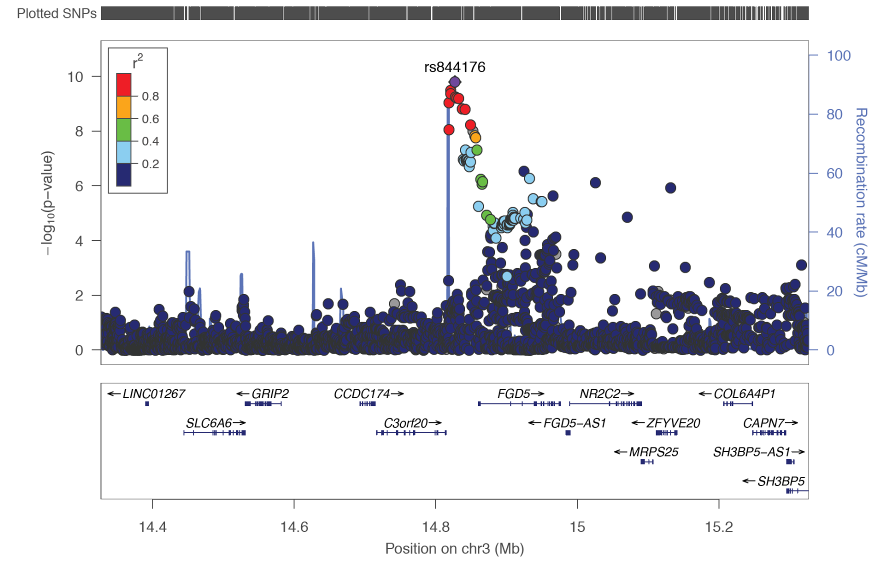


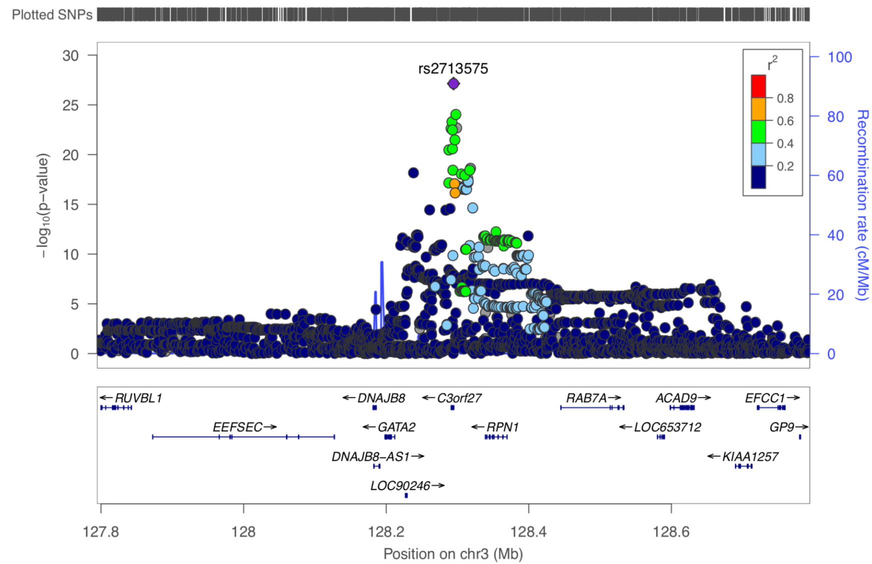
rs4849044
Plotted SNPs
rs4849044
−log10(p−value)
Recombination rate (cM/Mb)
r2
0.8
0.6
0.4
0.2
ANAPC1
MERTK
FBLN7
ZC3H6
TTL
TMEM87B
ZC3H8
RGPD8
POLR1B
RGPD5
CHCHD5
112.4
112.6
112.8
113
113.2
Position on chr2 (Mb)

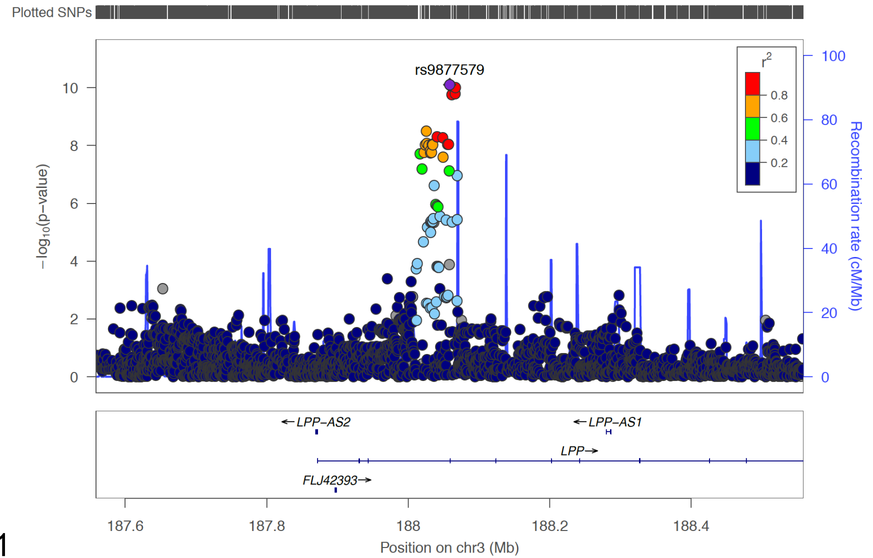


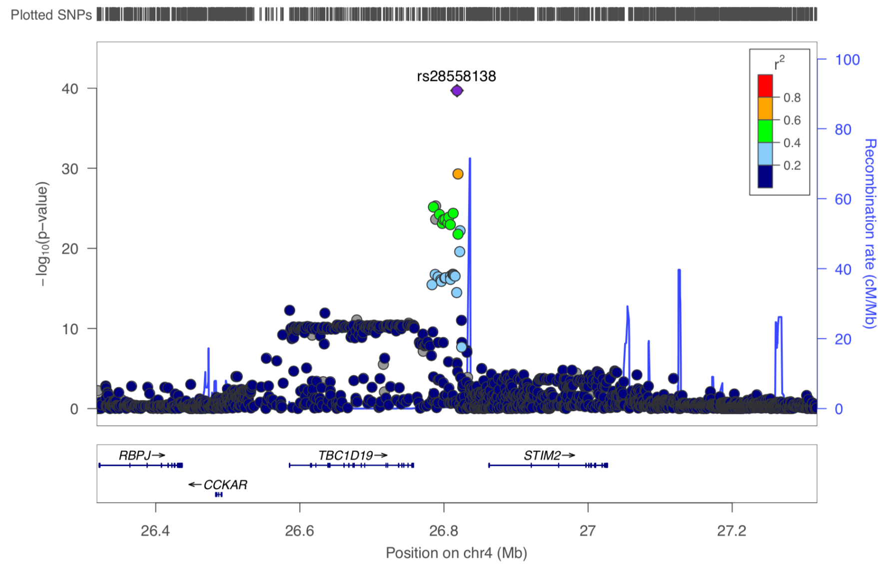
rs17819430
Plotted SNPs
rs17819430
−log10(p−value)
Recombination rate (cM/Mb)
r2
0.8
0.6
0.4
0.2
DDX18
CCDC93
INSIG2
LOC101927709
118.4
118.6
118.8
119
119.2
Position on chr2 (Mb)

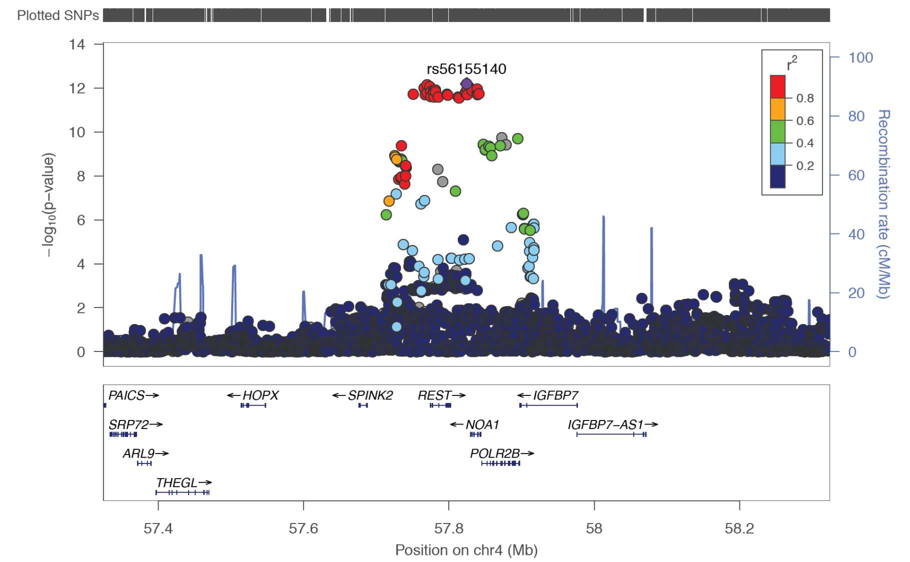

rs55889669
Plotted SNPs
rs55889669
Recombination rate (cM/Mb)
Position on chr2 (Mb)
SLC25A12
METAP1D
HAT1
DLX1
DLX2
ITGA6
PDK1
RAPGEF4-AS1
RAPGEF4

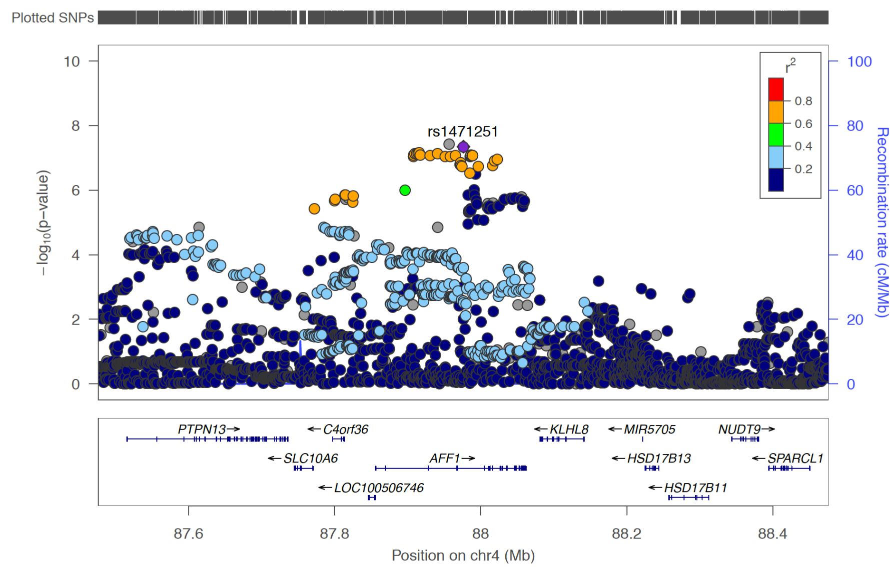

rs844176
Plotted SNPs
rs844176
Recombination rate (cM/Mb)
Position on chr3 (Mb)
LINC01267
GRIP2
CCDC174
FGD5
NR2C2
COL6A4P1
SLC6A6
C3orf20
FGD5-AS1
ZFYVE20
CAPN7
MRPS25
SH3BP5-AS1
SH3BP5

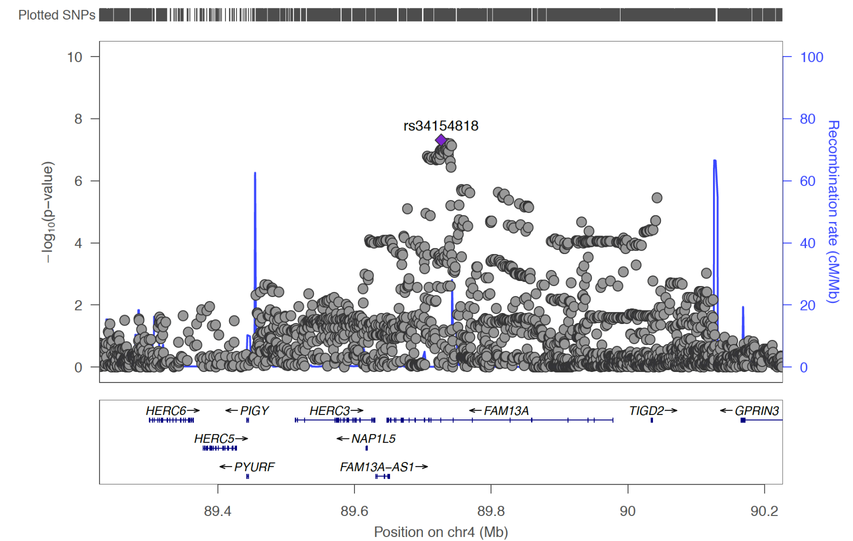

rs2713575
Plotted SNPs
rs2713575
Recombination rate (cM/Mb)
Position on chr3 (Mb)
RUVBL1
DNAJB8
C3orf27
RAB7A
ACAD9
EFCC1
EEFSEC
GATA2
RPN1
LOC653712
GP9
DNAJB8-AS1
KIAA1257
LOC90246

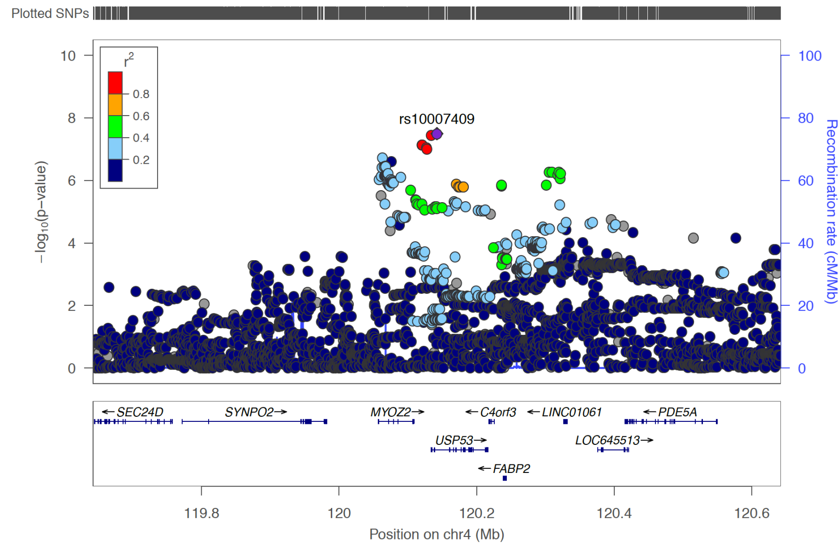

rs9877579
Plotted SNPs
rs9877579
Recombination rate (cM/Mb)
Position on chr3 (Mb)
LPP-AS2
LPP-AS1
LPP
FLJ42393


rs28558138
Plotted SNPs
rs28558138
r2
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
RBPJ→
CCKAR
TBC1D19→
STIM2→
26.4
26.6
26.8
27
27.2
Position on chr4 (Mb)

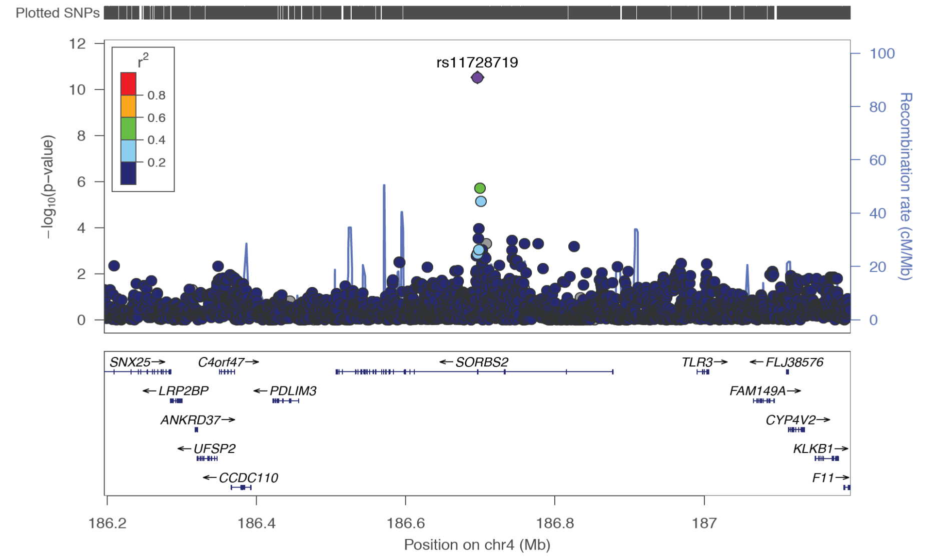

rs56155140
Plotted SNPs
rs56155140
r2
−log10(p−value)
Recombination rate (cM/Mb)
PAICS→
←HOPX
←SPINK2
REST→
←IGFBP7
SRP72→
←NOA1
IGFBP7−AS1→
ARL9→
POLR2B→
THEGL→
57.4
57.6
57.8
58
58.2
Position on chr4 (Mb)


rs1471251
Plotted SNPs
rs1471251
r2
−log10(p−value)
Recombination rate (cM/Mb)
PTPN13→
←C4orf36
←KLHL8
←MIR5705
NUDT9→
←SLC10A6
AFF1→
←HSD17B13
←SPARCL1
←LOC100506746
←HSD17B11
87.6
87.8
88
88.2
88.4
Position on chr4 (Mb)

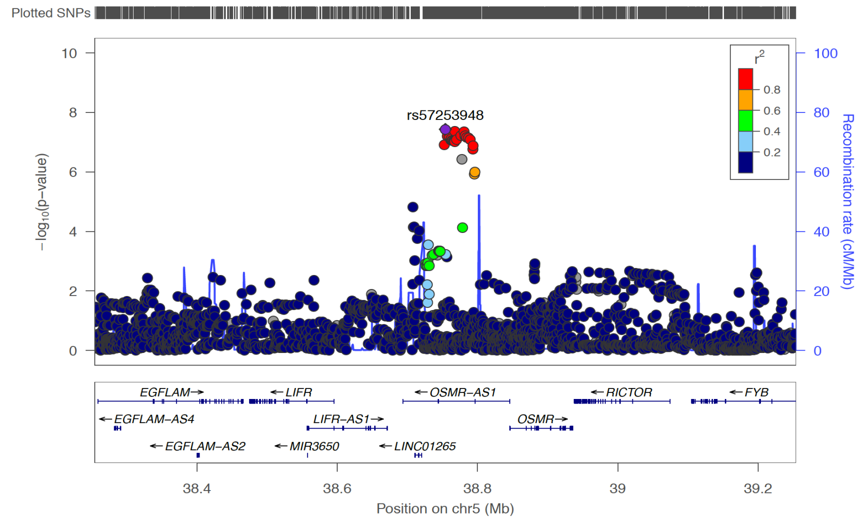

rs34154818
Plotted SNPs
rs34154818
−log10(p−value)
Recombination rate (cM/Mb)
HERC6→
←PIGY
HERC3→
←FAM13A
TIGD2→
←GPRIN3
HERC5→
←NAP1L5
←PYURF
FAM13A−AS1→
89.4
89.6
89.8
90
90.2
Position on chr4 (Mb)



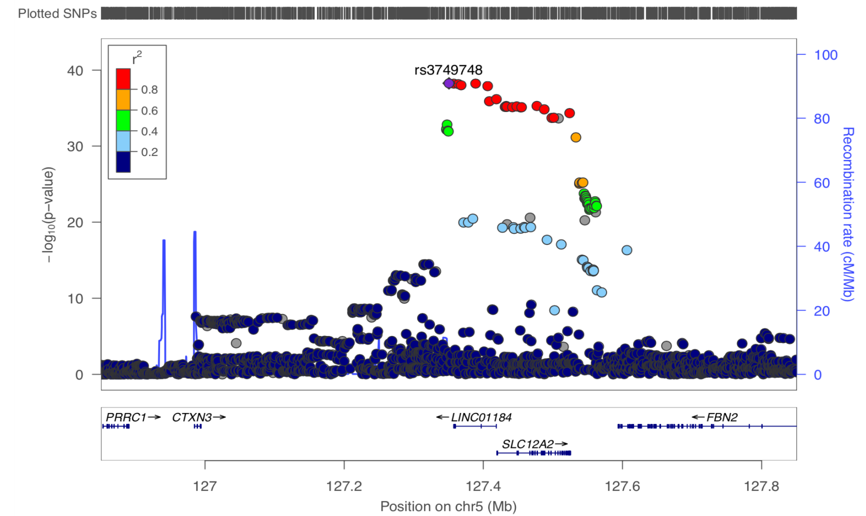
rs10007409
Plotted SNPs
−log10(p−value)
Recombination rate (cM/Mb)
rs10007409
SEC24D
SYNPO2
MYOZ2
C4orf3
LINC01061
PDE5A
USP53
LOC645513
FABP2
Position on chr4 (Mb)



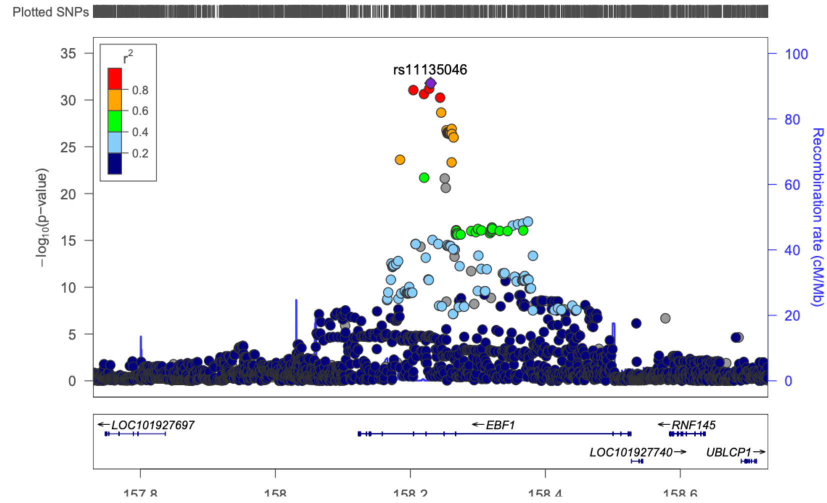
rs11728719
Plotted SNPs
−log10(p−value)
Recombination rate (cM/Mb)
rs11728719
SNX25
C4orf47
SORBS2
TLR3
FLJ38576
LRP2BP
PDLIM3
FAM149A
ANKRD37
CYP4V2
UFSP2
KLKB1
CCDC110
F11
Position on chr4 (Mb)

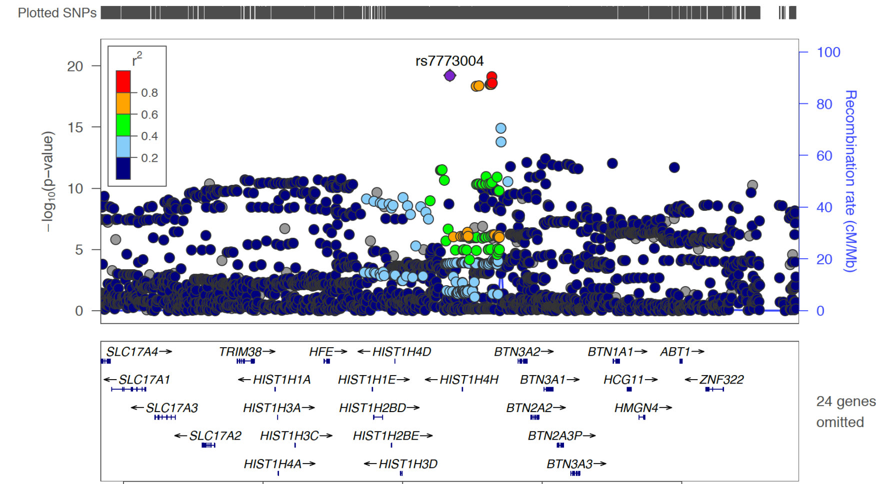


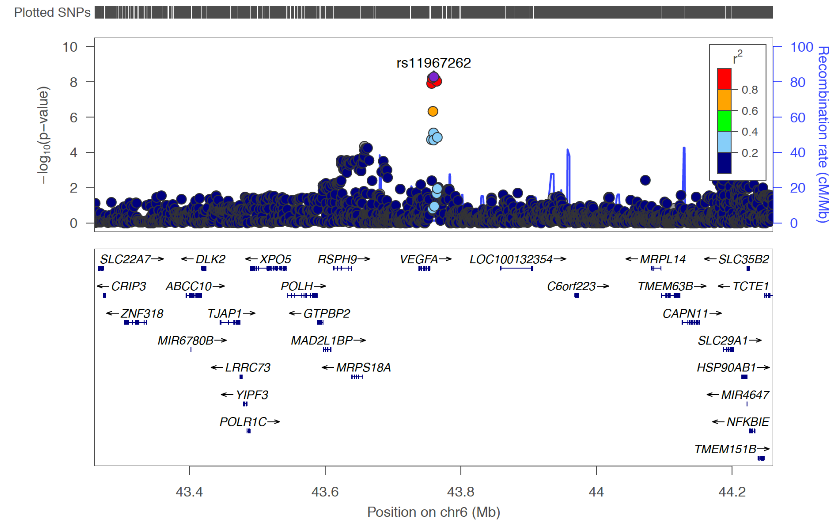
rs57253948
Plotted SNPs
−log10(p−value)
Recombination rate (cM/Mb)
rs57253948
EGFLAM
LIFR
OSMR−AS1
RICTOR
FYB
EGFLAM−AS4
LIFR−AS1
OSMR
EGFLAM−AS2
MIR3650
LINC01265
Position on chr5 (Mb)

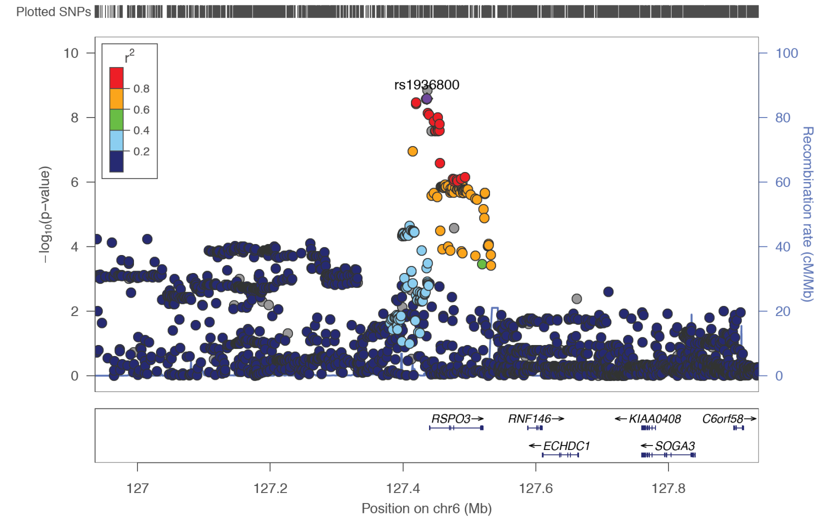


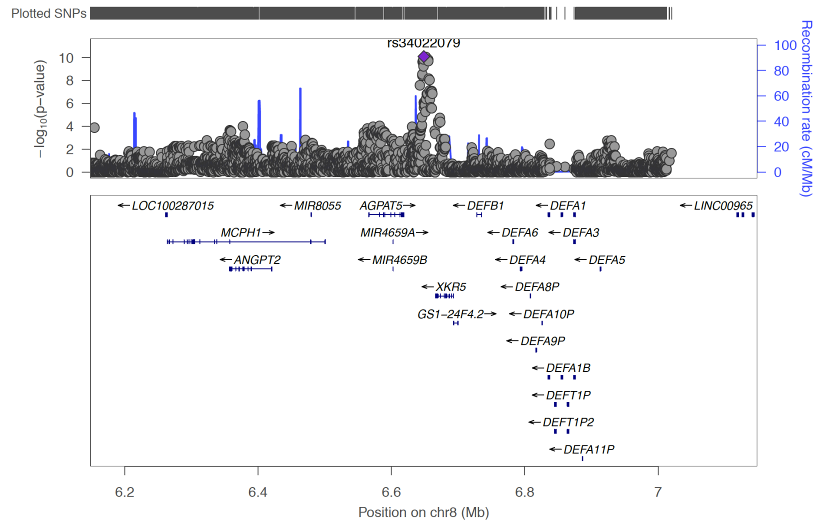
rs3749748
Plotted SNPs
−log10(p−value)
Recombination rate (cM/Mb)
rs3749748
PRRC1
CTXN3
LINC01184
FBN2
SLC12A2
Position on chr5 (Mb)

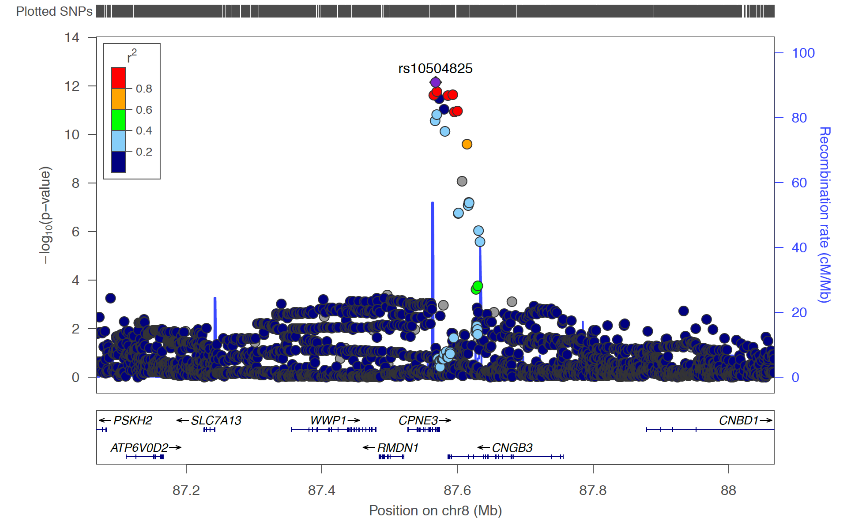

rs11135046
Plotted SNPs
rs11135046
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
←LOC101927697
←EBF1
←RNF145
LOC101927740→
UBLCP1→

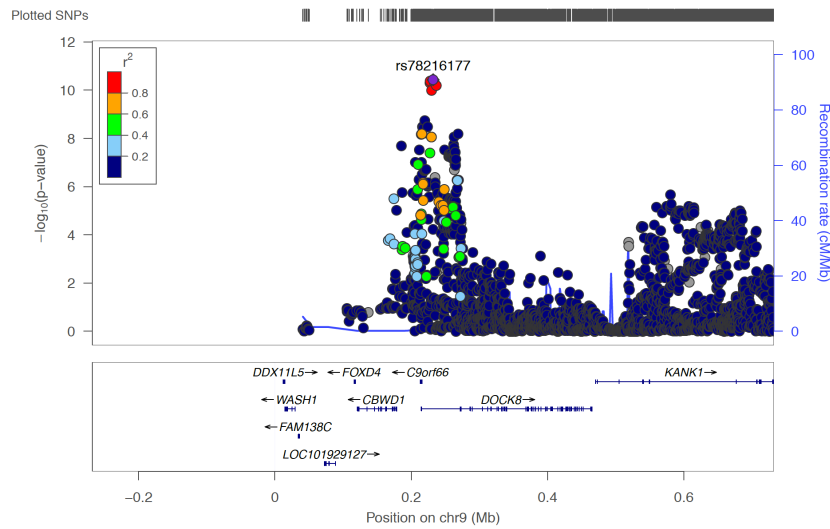

rs7773004
Plotted SNPs
rs7773004
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
SLC17A4→
TRIM38→
HFE→
←HIST1H4D
BTN3A2→
BTN1A1→
ABT1→
←SLC17A1
←HIST1H1A
HIST1H1E→
←HIST1H4H
BTN3A1→
HCG11→
←ZNF322
←SLC17A3
HIST1H3A→
HIST1H2BD→
BTN2A2→
HMGN4→
←SLC17A2
HIST1H3C→
HIST1H2BE→
BTN2A3P→
HIST1H4A→
←HIST1H3D
BTN3A3→
24 genes omitted

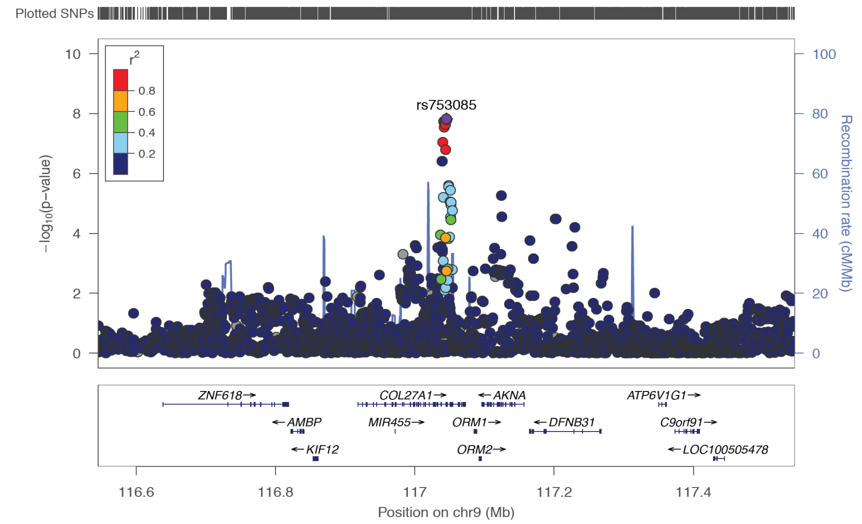

rs11967262
Plotted SNPs
rs11967262
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
SLC22A7→
←DLK2
←XPO5
RSPH9→
VEGFA→
LOC100132354→
←MRPL14
←SLC35B2
←CRIP3
ABCC10→
POLH→
C6orf223→
TMEM63B→
←TCTE1
←ZNF318
TJAP1→
←GTPBP2
CAPN11→
MIR6780B→
MAD2L1BP→
SLC29A1→
←LRRC73
←MRPS18A
HSP90AB1→
←YIPF3
←MIR4647
POLR1C→
←NFKBIE
TMEM151B→
Position on chr6 (Mb)

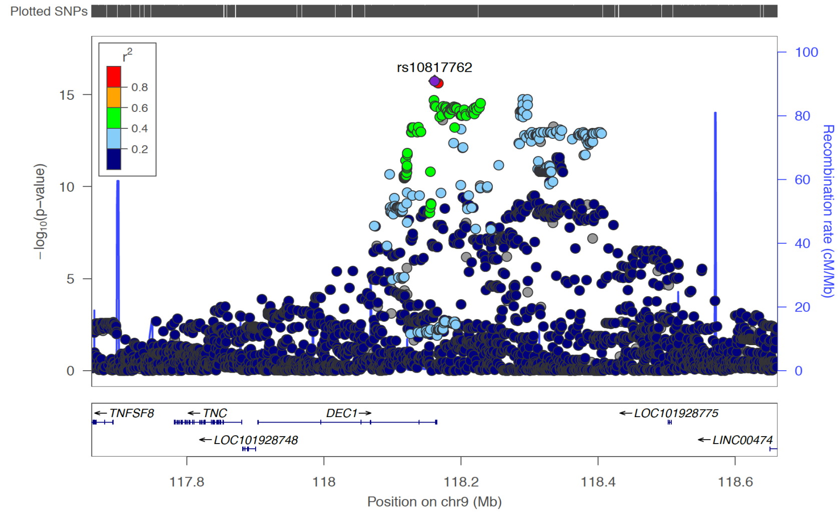

rs1936800
Plotted SNPs
rs1936800
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
RSPO3→
RNF146→
←KIAA0408
C6orf58→
←ECHDC1
←SOGA3
Position on chr6 (Mb)

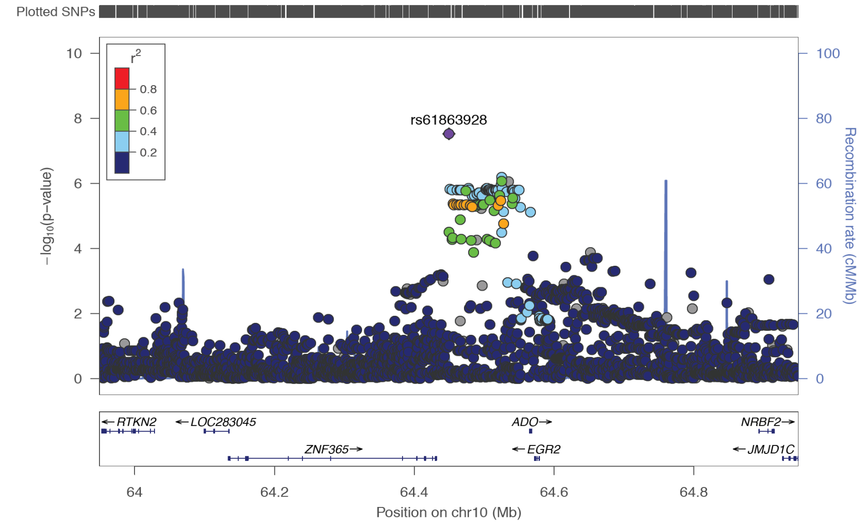

rs34022079
Plotted SNPs
−log10(p−value)
Recombination rate (cM/Mb)
Position on chr8 (Mb)
LOC100287015
MIR8055
AGPAT5
DEFB1
DEFA1
LINC00965
MCPH1
MIR4659A
DEFA6
DEFA3
ANGPT2
MIR4659B
DEFA4
DEFA5
XKR5
DEFA8P
GS1−24F4.2
DEFA10P
DEFA9P
DEFA1B
DEFT1P
DEFT1P2
DEFA11P

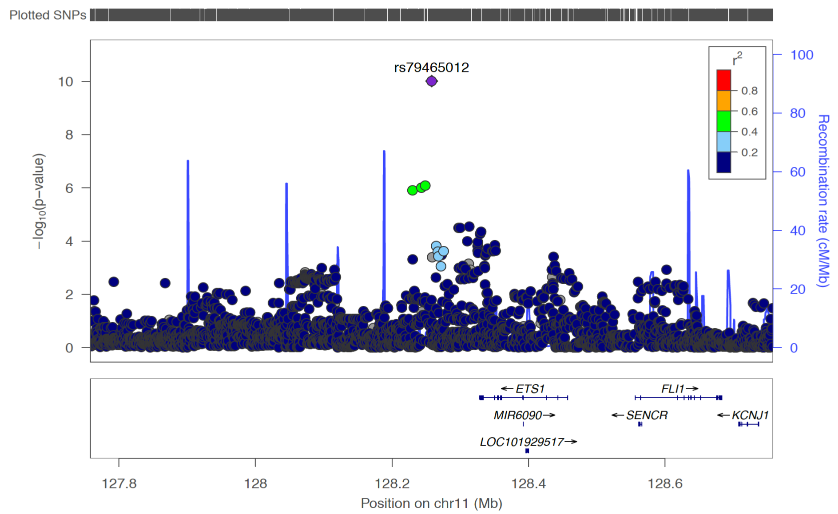

rs10504825
Plotted SNPs
r^2
−log10(p−value)
Recombination rate (cM/Mb)
Position on chr8 (Mb)
PSKH2
SLC7A13
WWP1
CPNE3
CNBD1
ATP6V0D2
RMDN1
CNGB3

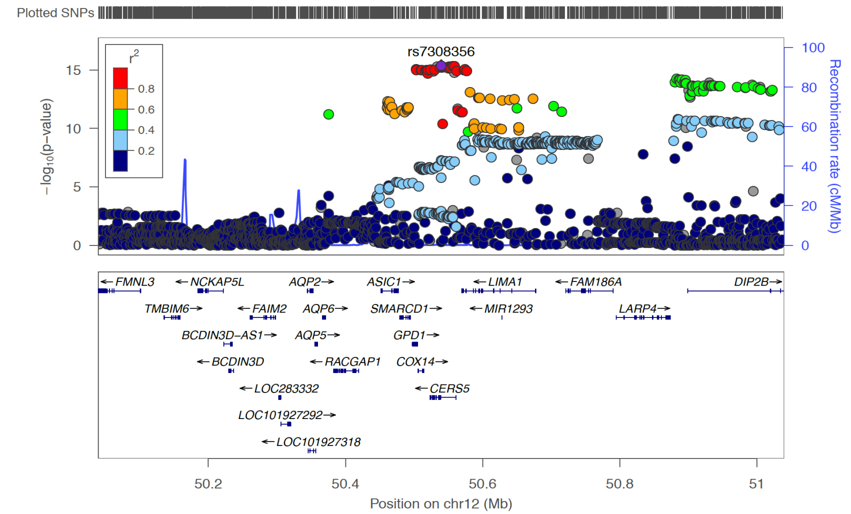

rs78216177
Plotted SNPs
r^2
−log10(p−value)
Recombination rate (cM/Mb)
Position on chr9 (Mb)
DDX11L5
FOXD4
C9orf66
KANK1
WASH1
CBWD1
DOCK8
FAM138C
LOC101929127

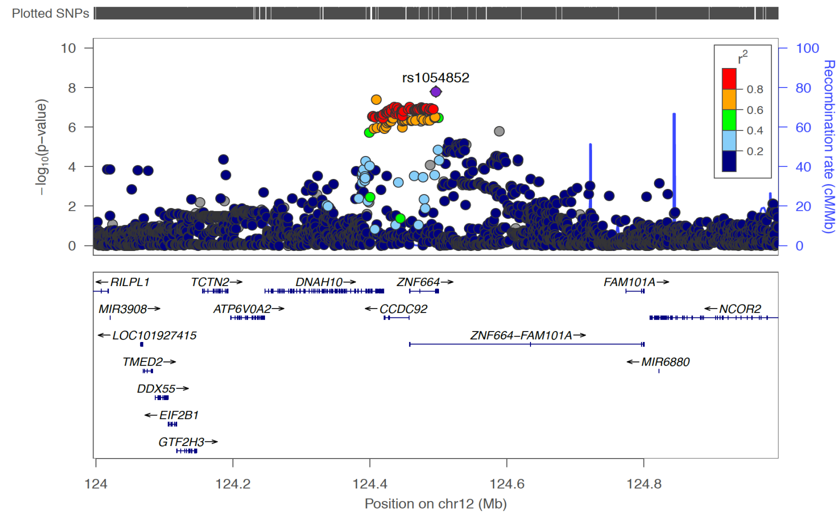

rs753085
Plotted SNPs
r^2
−log10(p−value)
Recombination rate (cM/Mb)
Position on chr9 (Mb)
ZNF618
COL27A1
AKNA
ATP6V1G1
AMBP
MIR455
ORM1
DFNB31
C9orf91
KIF12
ORM2
LOC100505478

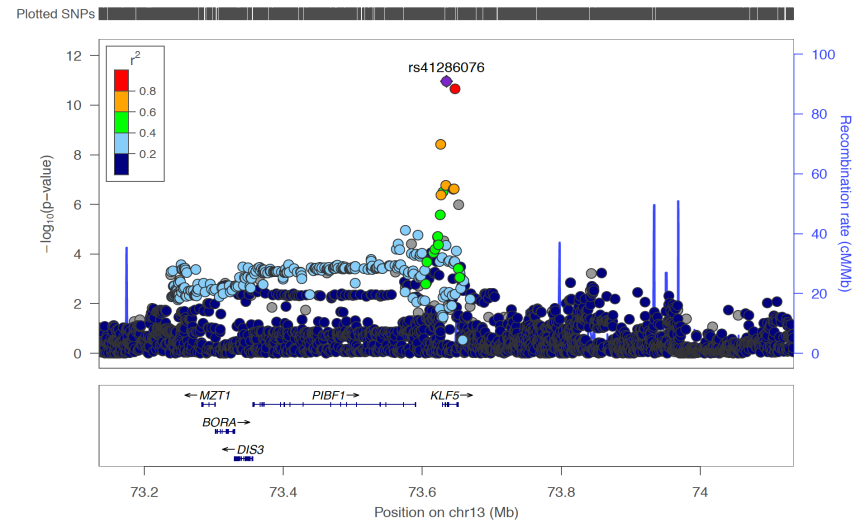

rs10817762
Plotted SNPs
rs10817762
−log10(p−value)
Recombination rate (cM/Mb)
TNFSF8
TNC
DEC1
LOC101928775
LOC101928748
LINC00474
Position on chr9 (Mb)

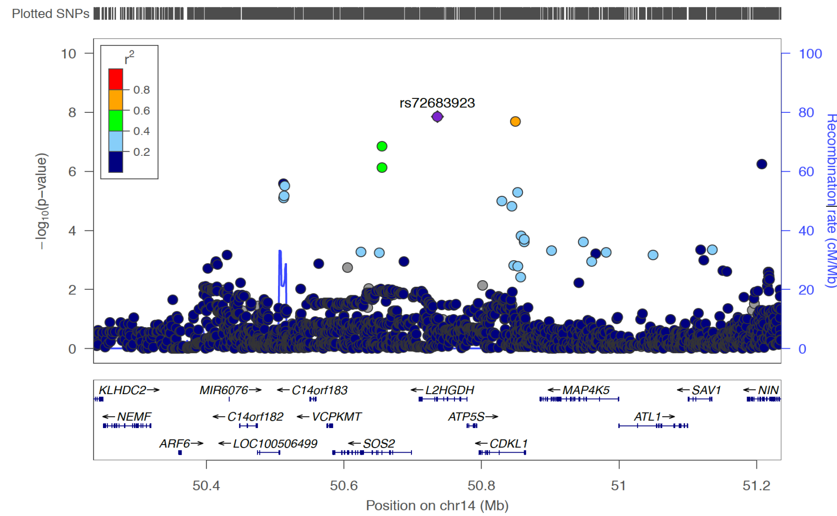

rs61863928
Plotted SNPs
rs61863928
−log10(p−value)
Recombination rate (cM/Mb)
RTKN2
LOC283045
ADO
NRBF2
ZNF365
EGR2
JMJD1C
Position on chr10 (Mb)

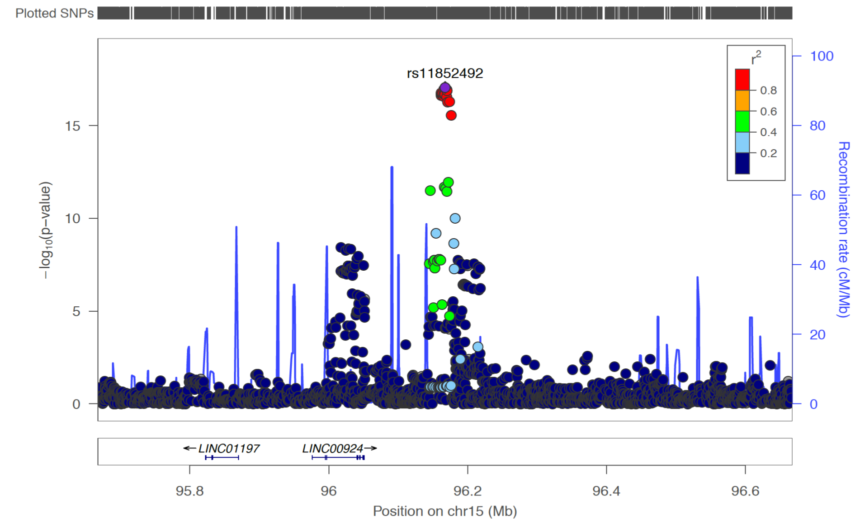

rs79465012
Plotted SNPs
rs79465012
−log10(p−value)
Recombination rate (cM/Mb)
ETS1
FLI1
MIR6090
SENCR
KCNJ1
LOC101929517
Position on chr11 (Mb)

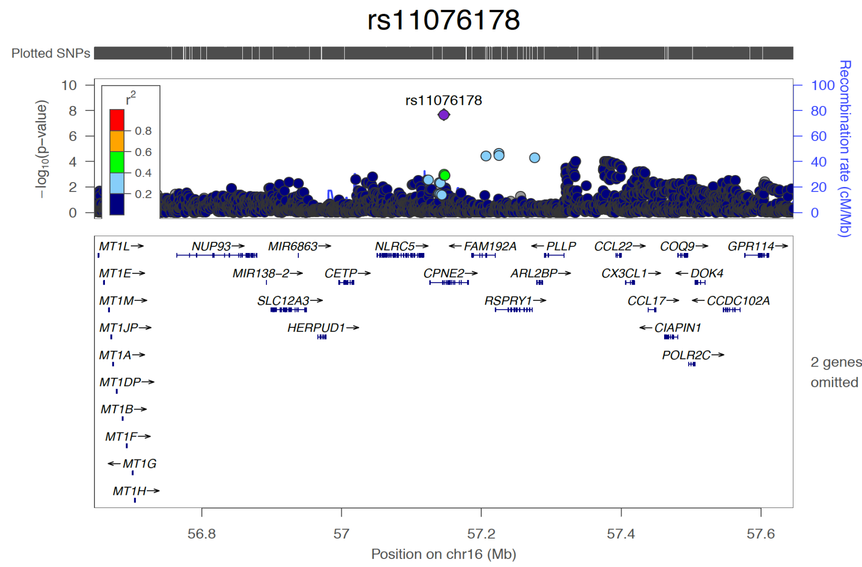

rs7308356
Plotted SNPs
rs7308356
−log10(p−value)
Recombination rate (cM/Mb)
FMNL3
NCKAP5L
AQP2
ASIC1
LIMA1
FAM186A
DIP2B
TMBIM6
FAIM2
AQP6
SMARCD1
MIR1293
LARP4
BCDIN3D−AS1
AQP5
GPD1
BCDIN3D
RACGAP1
COX14
LOC283332
CERS5
LOC101927292
LOC101927318
Position on chr12 (Mb)


rs1054852
Plotted SNPs
rs1054852
−log10(p−value)
Recombination rate (cM/Mb)
RILPL1
TCTN2
DNAH10
ZNF664
FAM101A
MIR3908
ATP6V0A2
CCDC92
NCOR2
LOC101927415
ZNF664−FAM101A
TMED2
MIR6880
DDX55
EIF2B1
GTF2H3
Position on chr12 (Mb)

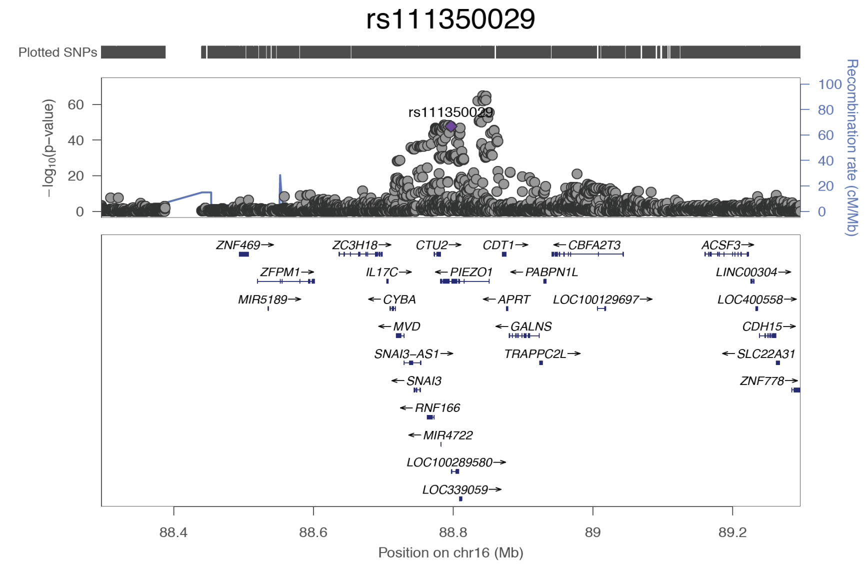

rs41286076
Plotted SNPs
rs41286076
−log10(p−value)
Recombination rate (cM/Mb)
MZT1
PIBF1
KLF5
BORA
DIS3
Position on chr13 (Mb)


rs72683923
Plotted SNPs
rs72683923
−log10(p−value)
Recombination rate (cM/Mb)
KLHDC2
MIR6076
C14orf183
L2HGDH
MAP4K5
SAV1
NIN
NEMF
C14orf182
VCPKMT
ATP5S
ATL1
ARF6
LOC100506499
SOS2
CDKL1
Position on chr14 (Mb)

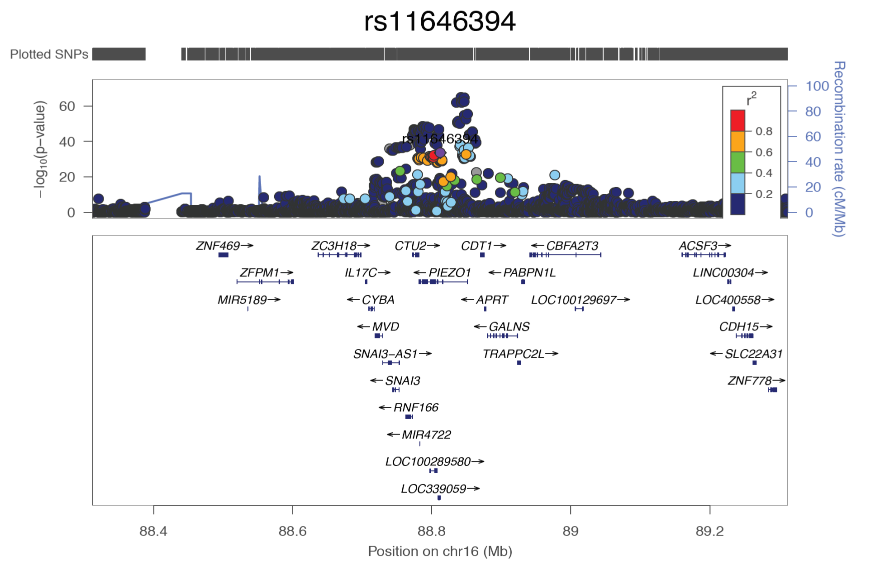

rs11852492
Plotted SNPs
rs11852492
−log10(p−value)
Recombination rate (cM/Mb)
LINC01197
LINC00924
Position on chr15 (Mb)



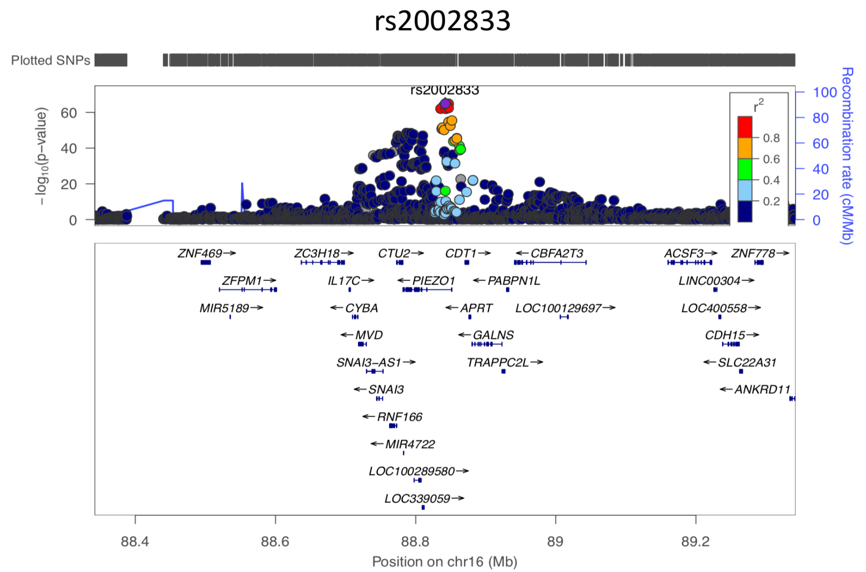
rs11076178
Plotted SNPs
−log10(p−value)
Recombination rate (cM/Mb)
r2
0.8
0.6
0.4
0.2
rs11076178
MT1L→
NUP93→
MIR6863→
NLRC5→
←FAM192A
←PLLP
CCL22→
COQ9→
GPR114→
MT1E→
MIR138−2→
CETP→
CPNE2→
ARL2BP→
CX3CL1→
←DOK4
MT1M→
SLC12A3→
RSPRY1→
CCL17→
←CCDC102A
MT1JP→
HERPUD1→
←CIAPIN1
MT1A→
POLR2C→
MT1DP→
MT1B→
MT1F→
←MT1G
MT1H→
2 genes omitted
56.8
57
57.2
57.4
57.6
Position on chr16 (Mb)



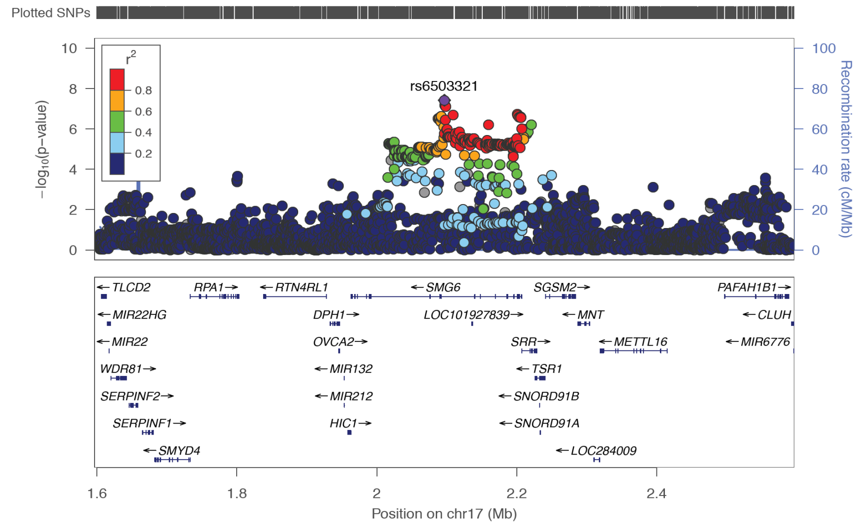
rs111350029
Plotted SNPs
−log10(p−value)
Recombination rate (cM/Mb)
rs111350029
ZNF469→
ZC3H18→
CTU2→
CDT1→
←CBFA2T3
ACSF3→
ZFPM1→
IL17C→
←PIEZO1
←PABPN1L
LINC00304→
MIR5189→
←CYBA
←APRT
LOC100129697→
LOC400558→
←MVD
←GALNS
CDH15→
SNAI3−AS1→
TRAPPC2L→
←SLC22A31
←SNAI3
ZNF778→
←RNF166
←MIR4722
LOC100289580→
LOC339059→
88.4
88.6
88.8
89
89.2
Position on chr16 (Mb)

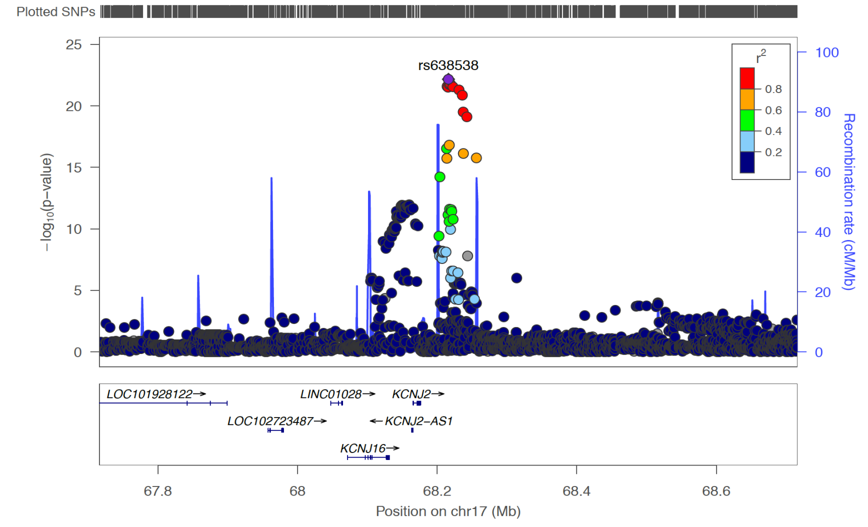


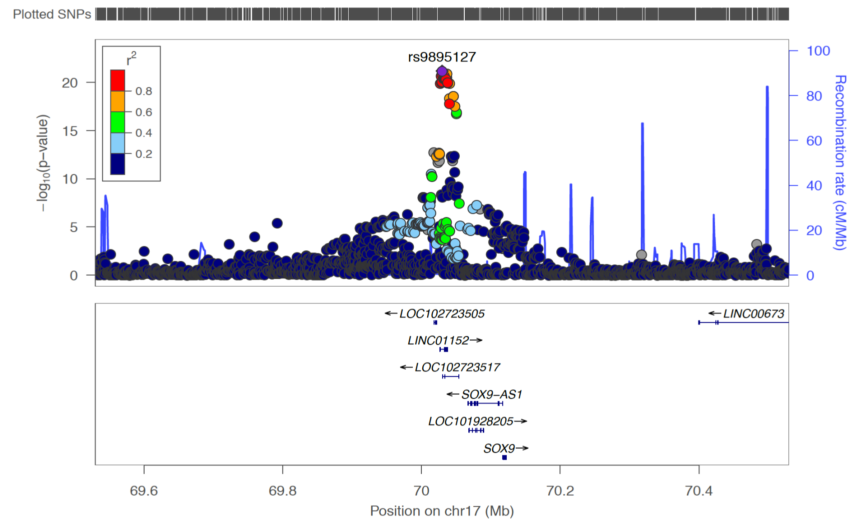
rs11646394
Plotted SNPs
−log10(p−value)
Recombination rate (cM/Mb)
rs11646394
r2
0.8
0.6
0.4
0.2
ZNF469→
ZC3H18→
CTU2→
CDT1→
←CBFA2T3
ACSF3→
ZFPM1→
IL17C→
←PIEZO1
←PABPN1L
LINC00304→
MIR5189→
←CYBA
←APRT
LOC100129697→
LOC400558→
←MVD
←GALNS
CDH15→
SNAI3−AS1→
TRAPPC2L→
←SLC22A31
←SNAI3
ZNF778→
←RNF166
←MIR4722
LOC100289580→
LOC339059→
88.4
88.6
88.8
89
89.2
Position on chr16 (Mb)

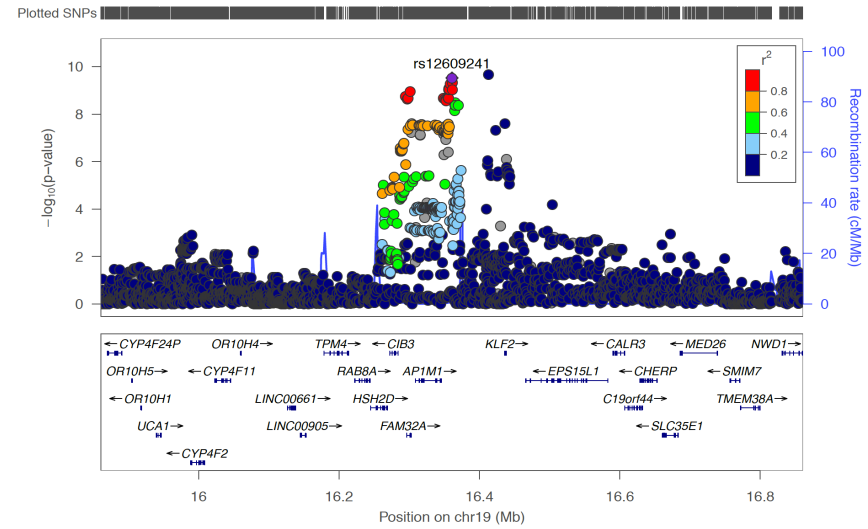


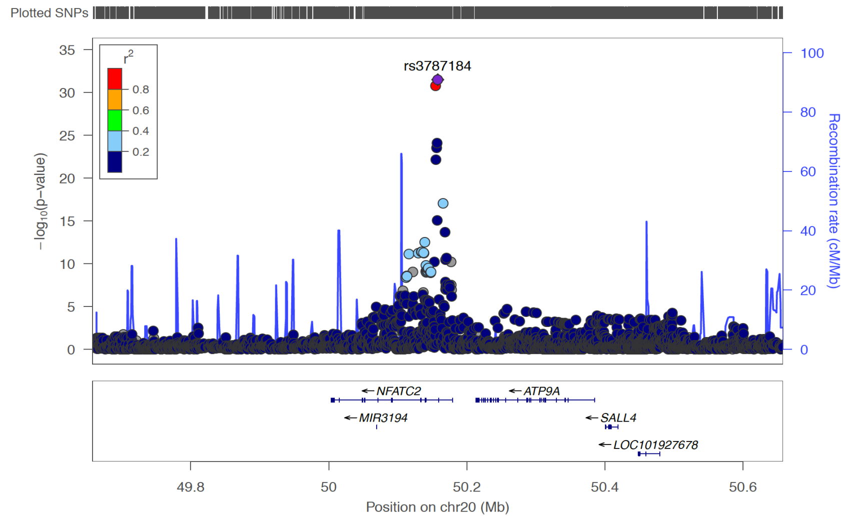
rs2002833
Plotted SNPs
−log10(p−value)
Recombination rate (cM/Mb)
rs2002833
r2
0.8
0.6
0.4
0.2
ZNF469→
ZC3H18→
CTU2→
CDT1→
←CBFA2T3
ACSF3→
ZNF778→
ZFPM1→
IL17C→
←PIEZO1
←PABPN1L
LINC00304→
MIR5189→
←CYBA
←APRT
LOC100129697→
LOC400558→
←MVD
←GALNS
CDH15→
SNAI3−AS1→
TRAPPC2L→
←SLC22A31
←SNAI3
←ANKRD11
←RNF166
←MIR4722
LOC100289580→
LOC339059→
88.4
88.6
88.8
89
89.2
Position on chr16 (Mb)

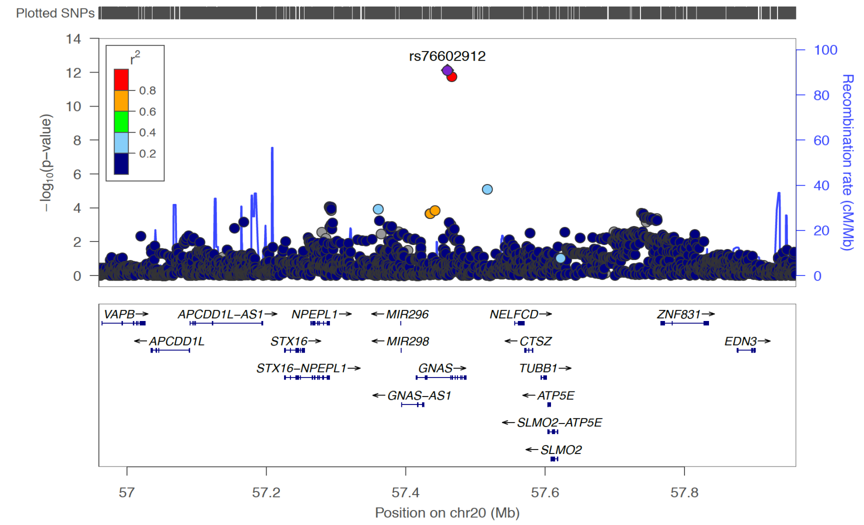

rs6503321
Plotted SNPs
rs6503321
$-\log_{10}$(p-value)
Recombination rate (cM/Mb)
r^2
0.8
0.6
0.4
0.2
TLCD2
RPA1
RTN4RL1
SMG6
SGSM2
PAFAH1B1
MIR22HG
DPH1
LOC101927839
MNT
CLUH
MIR22
OVCA2
SRR
METTL16
MIR6776
WDR81
MIR132
TSR1
SERPINF2
MIR212
SNORD91B
SERPINF1
HIC1
SNORD91A
SMYD4
LOC284009
Position on chr17 (Mb)

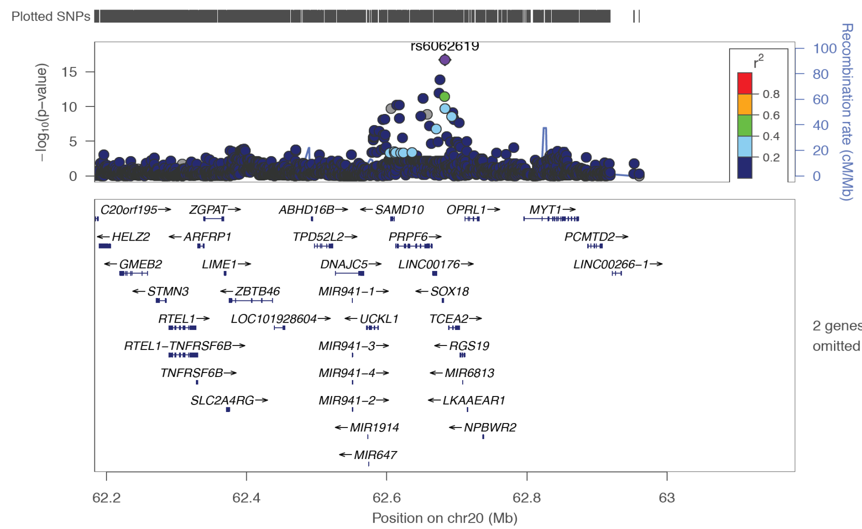

rs638538
Plotted SNPs
rs638538
$-\log_{10}$(p-value)
Recombination rate (cM/Mb)
r^2
0.8
0.6
0.4
0.2
LOC101928122
LINC01028
KCNJ2
LOC102723487
KCNJ2-AS1
KCNJ16
Position on chr17 (Mb)


rs9895127
Plotted SNPs
rs9895127
$-\log_{10}$(p-value)
Recombination rate (cM/Mb)
r^2
0.8
0.6
0.4
0.2
LOC102723505
LINC00673
LINC01152
LOC102723517
SOX9-AS1
LOC101928205
SOX9
Position on chr17 (Mb)


rs12609241
Plotted SNPs
rs12609241
$-\log_{10}$(p-value)
Recombination rate (cM/Mb)
r^2
0.8
0.6
0.4
0.2
CYP4F24P
OR10H4
TPM4
CIB3
KLF2
CALR3
MED26
NWD1
OR10H5
CYP4F11
RAB8A
AP1M1
EPS15L1
CHERP
SMIM7
OR10H1
LINC00661
HSH2D
C19orf44
TMEM38A
UCA1
LINC00905
FAM32A
SLC35E1
CYP4F2
Position on chr19 (Mb)


rs3787184
Plotted SNPs
rs3787184
−log10(p−value)
Recombination rate (cM/Mb)
NFATC2
ATP9A
MIR3194
SALL4
LOC101927678
Position on chr20 (Mb)


rs76602912
Plotted SNPs
rs76602912
−log10(p−value)
Recombination rate (cM/Mb)
VAPB
APCDD1L−AS1
NPEPL1
MIR296
NELFCD
ZNF831
APCDD1L
STX16
MIR298
CTSZ
EDN3
STX16−NPEPL1
GNAS
TUBB1
GNAS−AS1
ATP5E
SLMO2−ATP5E
SLMO2
Position on chr20 (Mb)


rs6062619
Plotted SNPs
rs6062619
−log10(p−value)
Recombination rate (cM/Mb)
C20orf195
ZGPAT
ABHD16B
SAMD10
OPRL1
MYT1
HELZ2
ARFRP1
TPD52L2
PRPF6
PCMTD2
GMEB2
LIME1
DNAJC5
LINC00176
LINC00266−1
STMN3
ZBTB46
MIR941−1
SOX18
RTEL1
LOC101928604
UCKL1
TCEA2
RTEL1−TNFRSF6B
MIR941−3
RGS19
TNFRSF6B
MIR941−4
MIR6813
SLC2A4RG
MIR941−2
LKAAEAR1
MIR1914
NPBWR2
MIR647
2 genes omitted
Position on chr20 (Mb)

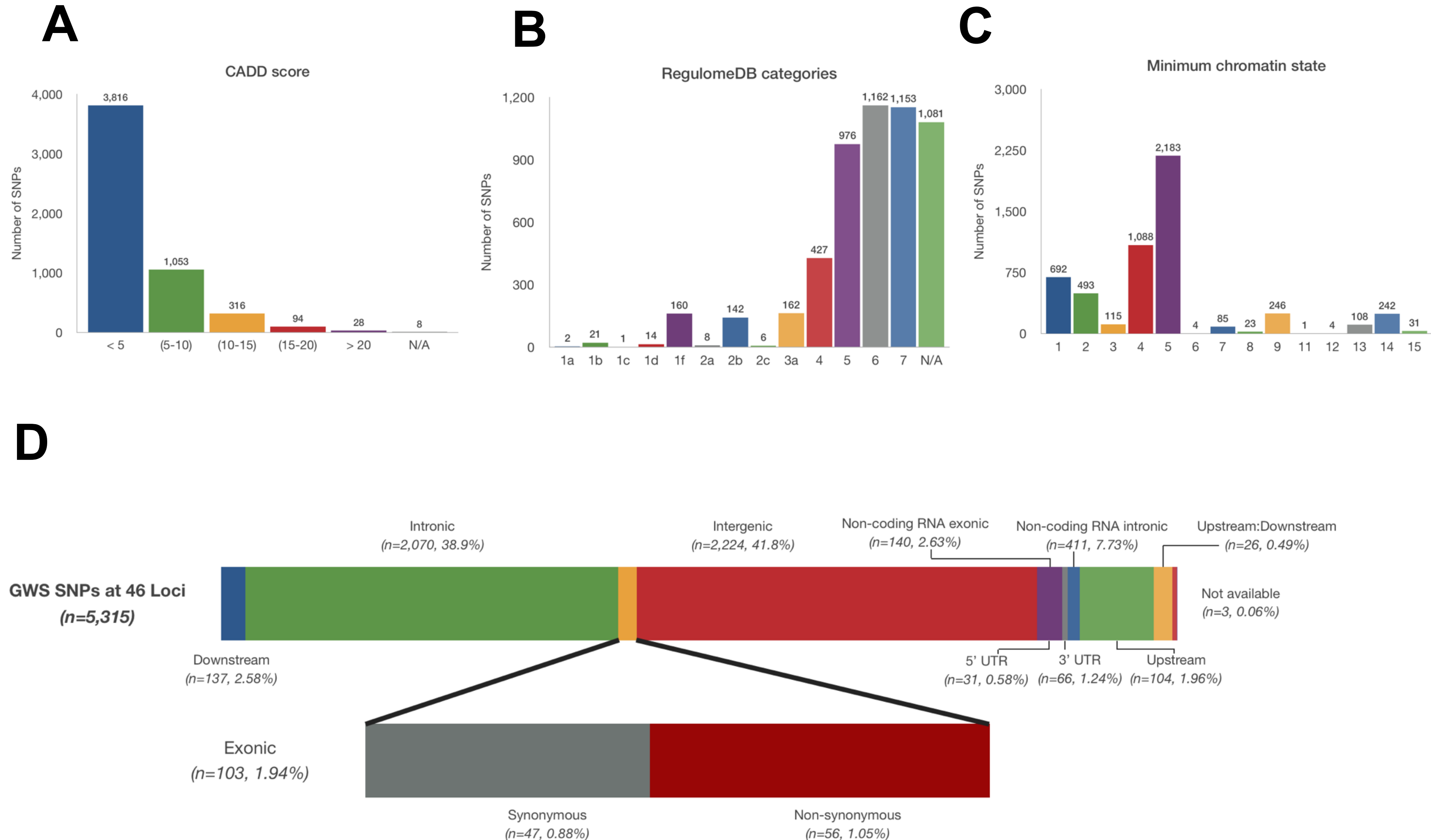
2.3.2. In silico annotation

FUMA[21] was used to interrogate and annotate SNPs at the susceptibility loci, identifying 5,315 genome-wide significant candidate SNPs from the discovery cohort associated with varicose veins at 45 of the 46 replicated loci (**Figure 2.5**). ~2% of candidate variants (n = 103) were exonic, of which 47 were synonymous and 56 were non-synonymous (52 missenses, two stop gains, one splice site variant, and one frameshift variant). Four missense variants were non-synonymous and predicted to alter protein structure and/or function and demonstrated modest linkage with the index replicated variant at three loci ($r^2 \geq 0.22$ and $D' \geq 0.75$) – 12q13.12 (rs7308356), 16q24.3 (rs2002833) and 17q24.3 (rs9895127) (**Appendix Table 2.2**). This includes rs12303082-T/G ($P = 2.0\times10^{-9}$, OR = 1.06, $r^2_{index} = 0.27$, $D'_{index} = 0.92$) which resides within a highly conserved region (Genomic evolutionary rate profiling (GERP) score: 0.83) of *FAM186A*, causing a p.Lys187Gln amino acid substitution which is predicted to be deleterious and alter protein function (SIFT[37]: 0.01, PolyPhen-2[38]: 0.91). Another missense variant, rs7184427-A/G ($P = 9.1\times10^{-40}$, OR = 1.19, $r^2_{index} = 0.22$, $D'_{index} = 0.75$), results in a predicted deleterious p.Val250Ala substitution within *PIEZO1* (SIFT[37]: 0.00). *PIEZO1* encodes a mechanosensitive cation channel involved in the detection of vascular shear stress, and was previously associated with varicose veins and lymphoedema.[5,39]

Of the intronic and intergenic variants (n = 4294, 80.8%) highlighted by ANNOVAR[23] (**Figure 2.5**), 3,735 (87.0%) lie in open chromatin regions. 163 of the intronic and intergenic variants show evidence of functionality with a CADD Score[25] ≥ 12.37 (**Appendix Table 2.3**), of which 17 also demonstrate regulatory potential with an RDB[24] score of 2b or less (*likely to affect binding*) and eight with an RDB score of 1f

or less (*likely to affect binding and linked to expression of a gene target* (i.e. an eQTL)) (**Appendix Table 2.3**).

Figure 2.5. Functional annotation of the genome-wide significant variants at the forty-six varicose veins-associated loci. Functional consequences of the SNPs on genes were obtained by performing ANNOVAR gene-based annotation using Ensembl genes (build 85) in FUMA. A) CADD scores, B) RegulomeDB scores and C) 15-core chromatin state were annotated to all 5,315 SNPs in 1000G phase 3 by FUMA through matching chromosome, position, reference, and alternative alleles. D) Positional classification of the 5,315 SNPs.



2.3.3. Gene mapping

204 putative genes were mapped to 38 replicated loci based on genomic proximity at these loci.[21] Eighty genes were prioritised based on their association with variants that are known to alter expression of these genes (eQTLs) within tibial artery tissue from the GTEx consortium v8.0[40] ($P_{eqtl} < 5×10^{-8}$). Of these, 30 were *not* positionally mapped (i.e. they reside outside the 10kb positional proximity window). A genome-wide, gene-based association study (GWGAS) implemented in MAGMA v1.07[29] prioritised 248 protein-coding genes significantly associated with varicose veins at a Bonferroni-corrected P-value < $2.67×10^{-6}$; of which, 117 lay within the confines of the replicated loci (**Figure 2.6**; **Figure 2.7**; **Appendix Table 2.4**).

Summary-based mendelian randomisation (SMR) analysis[30] was performed using eQTL data from GTEx v8[40] tibial artery tissue as an instrumental variable to test association between gene expression levels and varicose veins. SMR testing was performed across 4,946 probes with a cis-eQTL $P < 5×10^{-8}$; with the threshold for significance set at $P_{SMR} < 1.01×10^{-5}$. Forty-four putative genes passed this stringent correction and were subsequently tested via a HEIDI analysis to exclude associations with varicose veins through linkage disequilibrium or co-localisation. Twenty-seven SMR-significant genes passed the HEIDI test ($P_{HEIDI} \geq 1.12×10^{-3}$ (0.05/44)) (**Appendix Table 2.5**), 14 of which lay within the varicose veins susceptibility loci and therefore associated with varicose veins through pleiotropy.

In summary, 237 unique genes were mapped to 39 of 46 varicose vein susceptibility loci by at least one of the four mapping strategies, 61 of which were novel putative genes (**Table 2.3**). Substantial overlap was found across the mapping strategies, with

the majority of genes (54.9%, n = 130) being mapped by two or more approaches. Thirty-six genes were prioritised by three mapping approaches, and six genes (*ATF1, AP1M1, DNAH10OS, FBLN7, LBH, WDR92*) were mapped to the varicose veins susceptibility loci by all four mapping approaches (**Figure 2.8**).

Figure 2.6. MAGMA gene-based association study quantile-quantile plot. Quantile-Quantile (Q-Q) plot for the genome-wide, gene-based association test computed by MAGMA v1.07.[29]

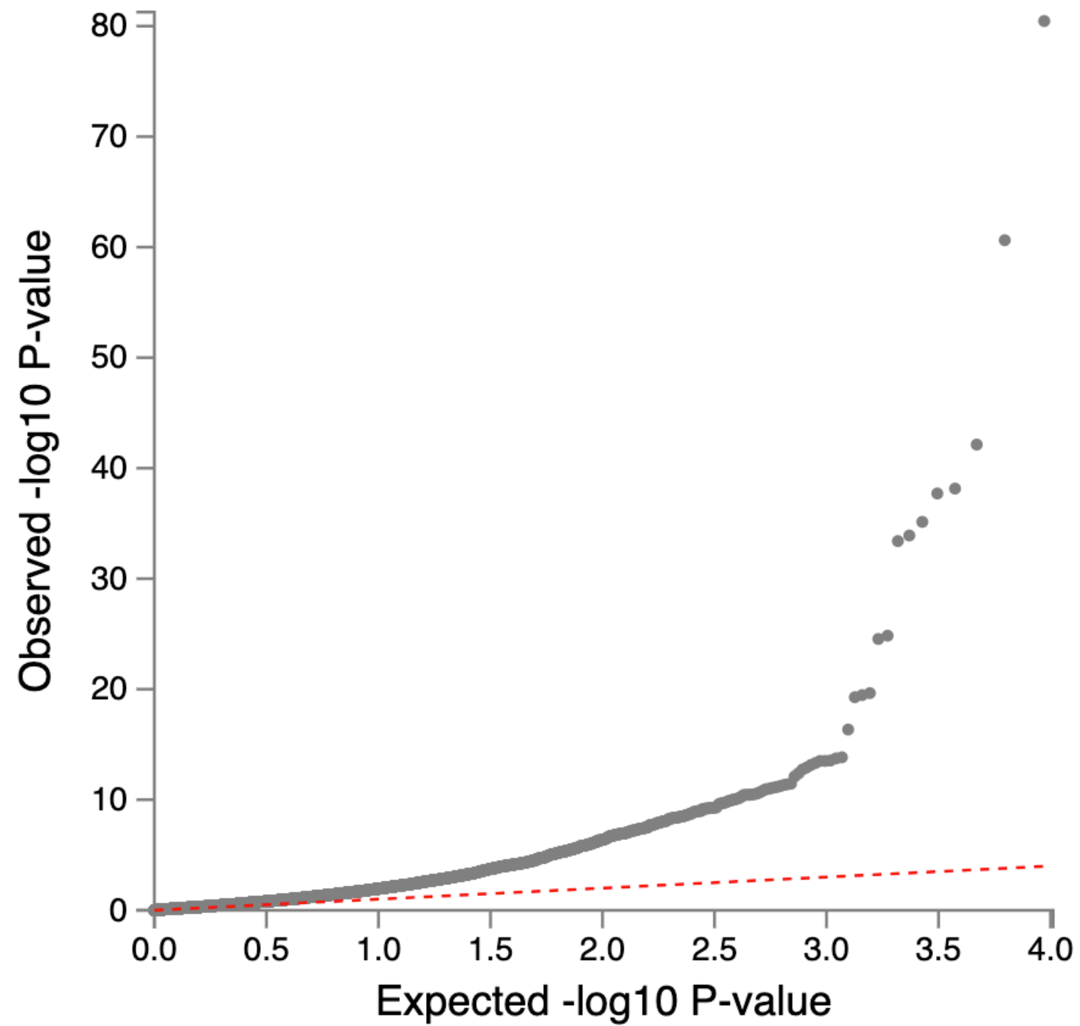


Figure 2.7. MAGMA genome-wide, gene-based association study Manhattan plot. Manhattan plot for the MAGMA GWGAS for varicose veins. The dotted red line indicates the threshold for genome-wide significance ($P < 2.68 \times 10^{-6}$). 248 genes reached genome-wide significance in this analysis, with the top-ten enriched MAGMA genes are annotated in the Manhattan plot.

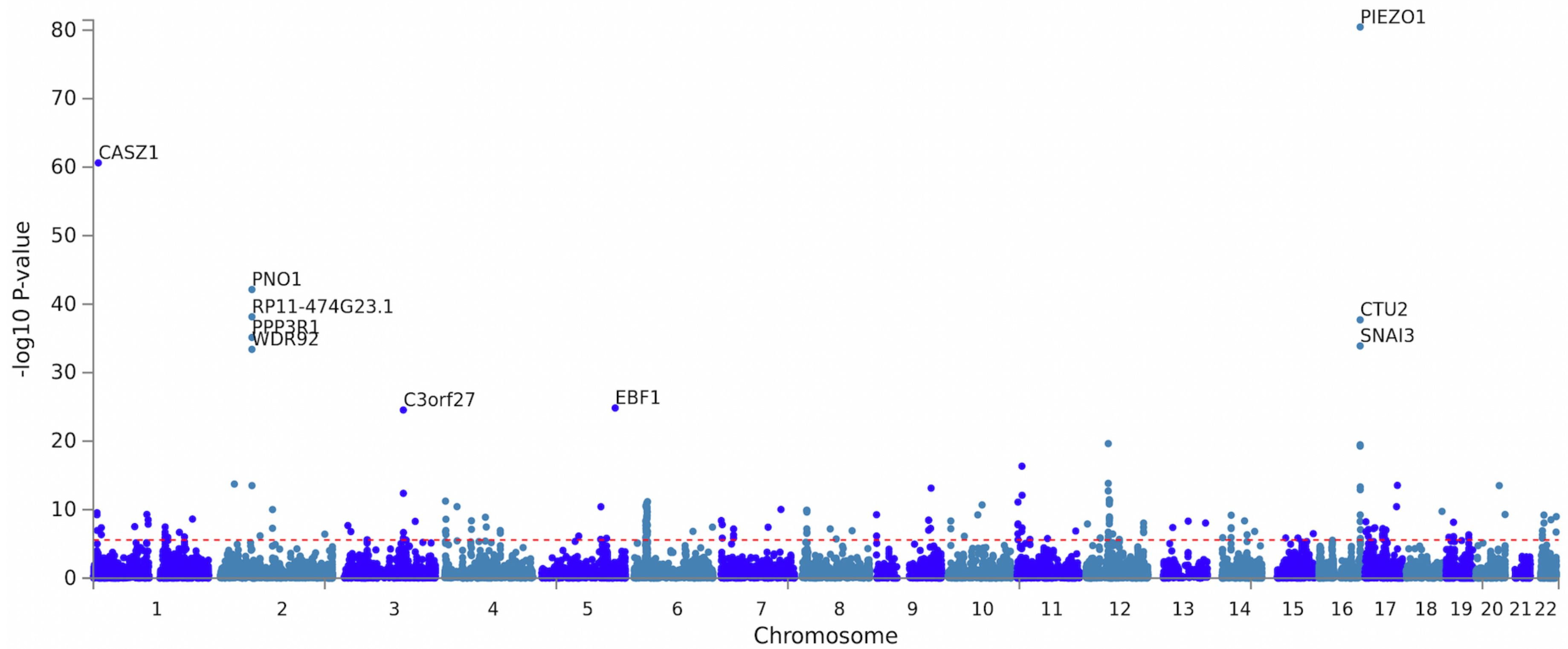
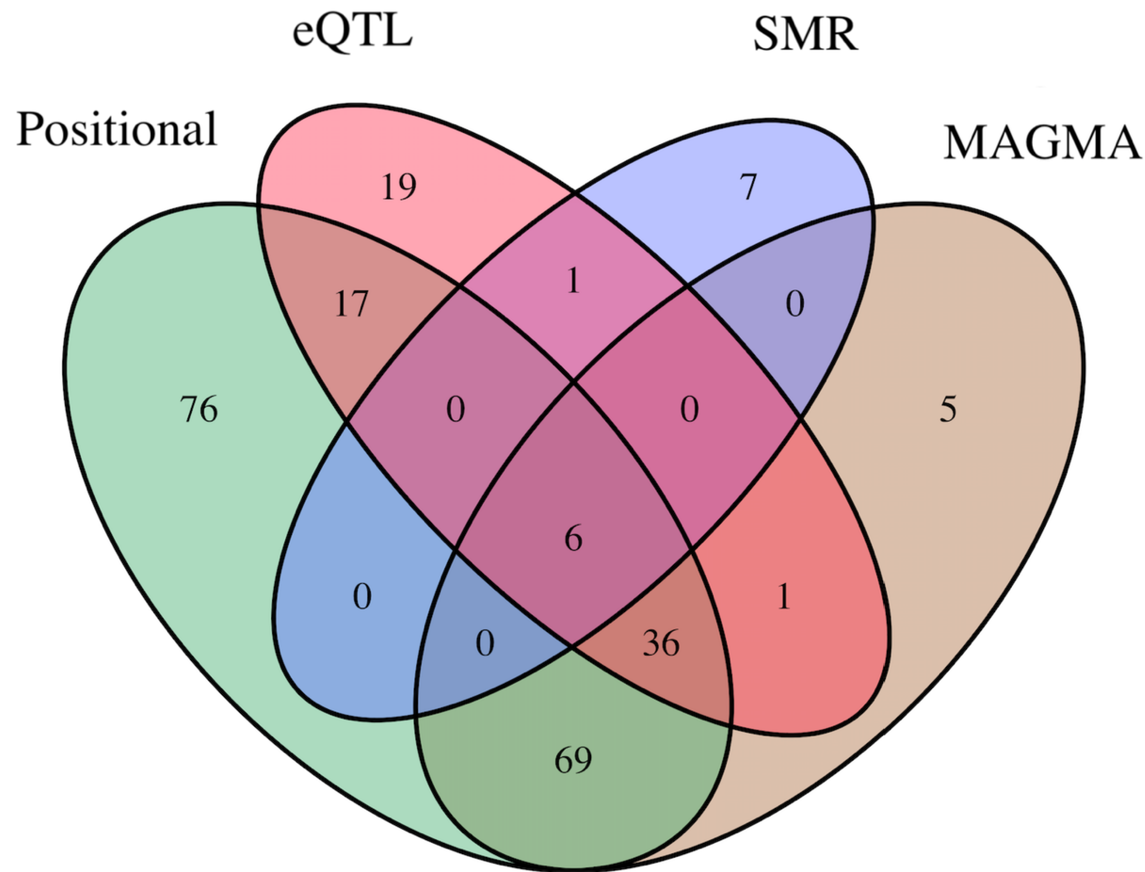


Figure 2.8. Venn diagram for the 237 genes prioritised at the varicose veins associated loci. 237 unique genes were mapped to thirty-nine of the 46 associated loci by one or more gene mapping strategies. 204 genes were mapped via positional mapping, 80 genes were mapped via eQTL mapping, 14 genes were mapped by SMR, and 117 genes were mapped by MAGMA. Overlap between the four different mapping strategies is shown in the Venn diagram. See **Appendix Table 2.6** for a complete list of all prioritised genes.



2.3.4. Gene set, pathway and tissue-specific enrichment

To delineate gene sets and enriched pathways at which the 237 prioritised genes converged, gene-set analysis was conducted in MAGMA v1.07.[29] Four Gene Ontology (GO) gene sets were significantly over-represented in the summary statistics: Cardiovascular Development ($P = 1.56\times10^{-8}$, n = 666); Tube Morphogenesis ($P = 9.35\times10^{-8}$, n =778); Blood Vessel Morphogenesis ($P = 9.39\times10^{-7}$, n = 555); and Tube Development ($P = 1.68\times10^{-6}$, n = 956) (**Appendix Table 2.7**). MAGMA[29] tissue-specific gene property analysis of 54 specific tissue types from the GTEx v8.0 consortium[40] demonstrated significant gene expression in all three vascular tissue types present in GTEx — Coronary Artery ($P = 6.23\times10^{-7}$, 2nd most enriched), Tibial Artery ($P = 1.05\times10^{-6}$, 3rd most enriched) and Aorta ($P = 3.92\times10^{-5}$, 8th most enriched) (**Figure 2.9-A**). MAGMA tissue enrichment within GTEx 30 general tissue types[40] also demonstrated blood vessel tissue to be highly enriched ($P = 3.8\times10^{-4}$, 3rd most enriched) (**Figure 2.9-B**). Next, performing enrichment analysis of the 237 prioritised genes within eXploring Genomic Relations (XGR),[32] six canonical pathways were significantly over-represented. This included genes in pathways pertaining to extracellular matrix biology, the VEGF and VEGFR signalling network, and intracellular Ca^{2+} signalling in the T-Cell Receptor (TCR) Pathway (**Table 2.4**).

Figure 2.9. MAGMA tissue expression analysis. MAGMA Tissue Expression Analysis of varicose veins GWAS-summary data, implemented in FUMA in A) 54 specific and B) 30 general tissue types. This analysis tests the relationship between highly expressed genes in a specific tissue and the genetic associations from the GWAS. Gene-property analysis is performed using average expression of genes per tissue type as a gene covariate. Gene expression values are log2 transformed average RPKM (Read Per Kilobase Per Million) per tissue type after winsorization at 50, and are based on GTEx v8 RNA-Seq data across 54 specific tissue types and 30 general tissue types. The dotted line indicates the Bonferroni-corrected α level, and the tissues that meet this significance threshold are highlighted in red.

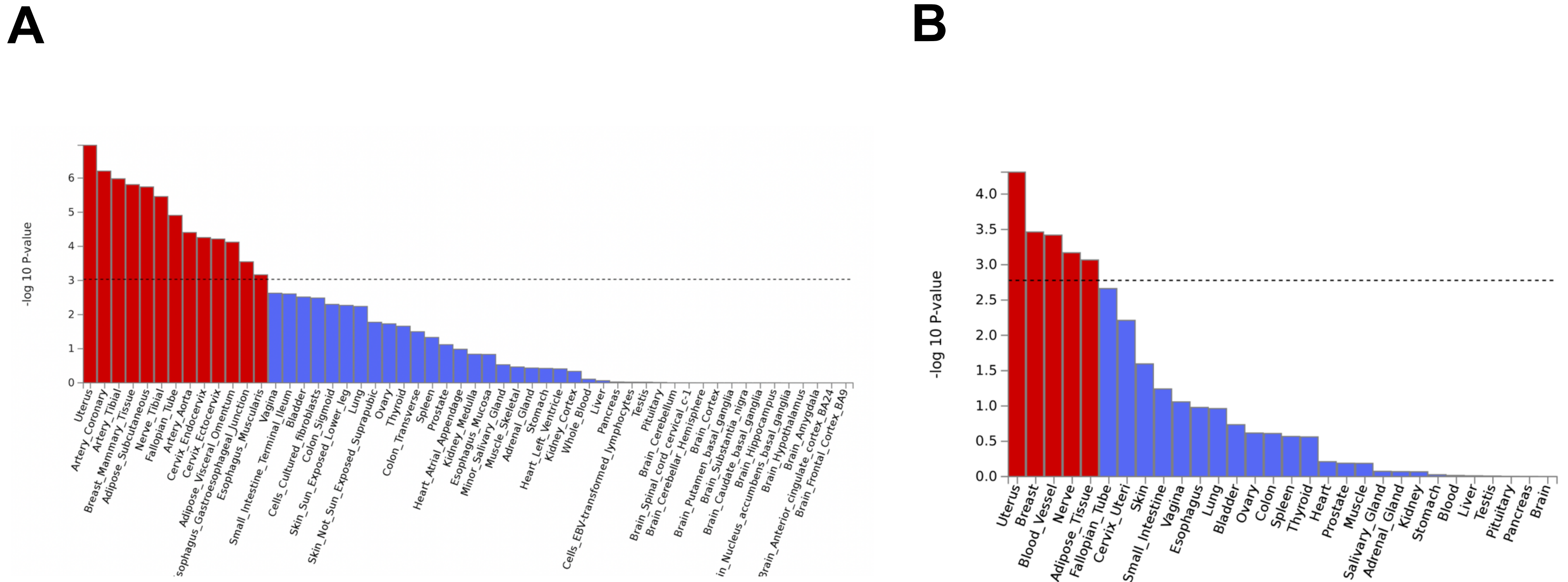


Table 2.4. Gene-based enrichment analysis for varicose veins associated genes in eXploring Genomic Relations (XGR).

Biological Process	Z-Score	P-Value	FDR	Number of overlapped genes	Genes
Alpha9 beta1 integrin signalling events	3.85	0.0012	0.032	*2*	*TNC, VEGFA*
Genes encoding structural ECM glycoproteins	3.39	0.0012	0.032	*6*	*EFEMP1, FBLN7, FBN2, IGFBP7, RSPO3, TNC*
Calcium signalling in the CD4+ TCR pathway	3.5	0.0019	0.032	*2*	*NFATC2, PPP3R1*
Ensemble of genes encoding core extracellular matrix including ECM glycoproteins, collagens and proteoglycans	3.11	0.0019	0.032	*7*	*COL27A1, EFEMP1, FBLN7, FBN2, IGFBP7, RSPO3, TNC*
Non-canonical WNT signalling pathway	3.29	0.0025	0.035	*2*	*MAPK10, NFATC2*
VEGF and VEGFR signalling network	3.11	0.0032	0.036	*1*	*VEGFA*

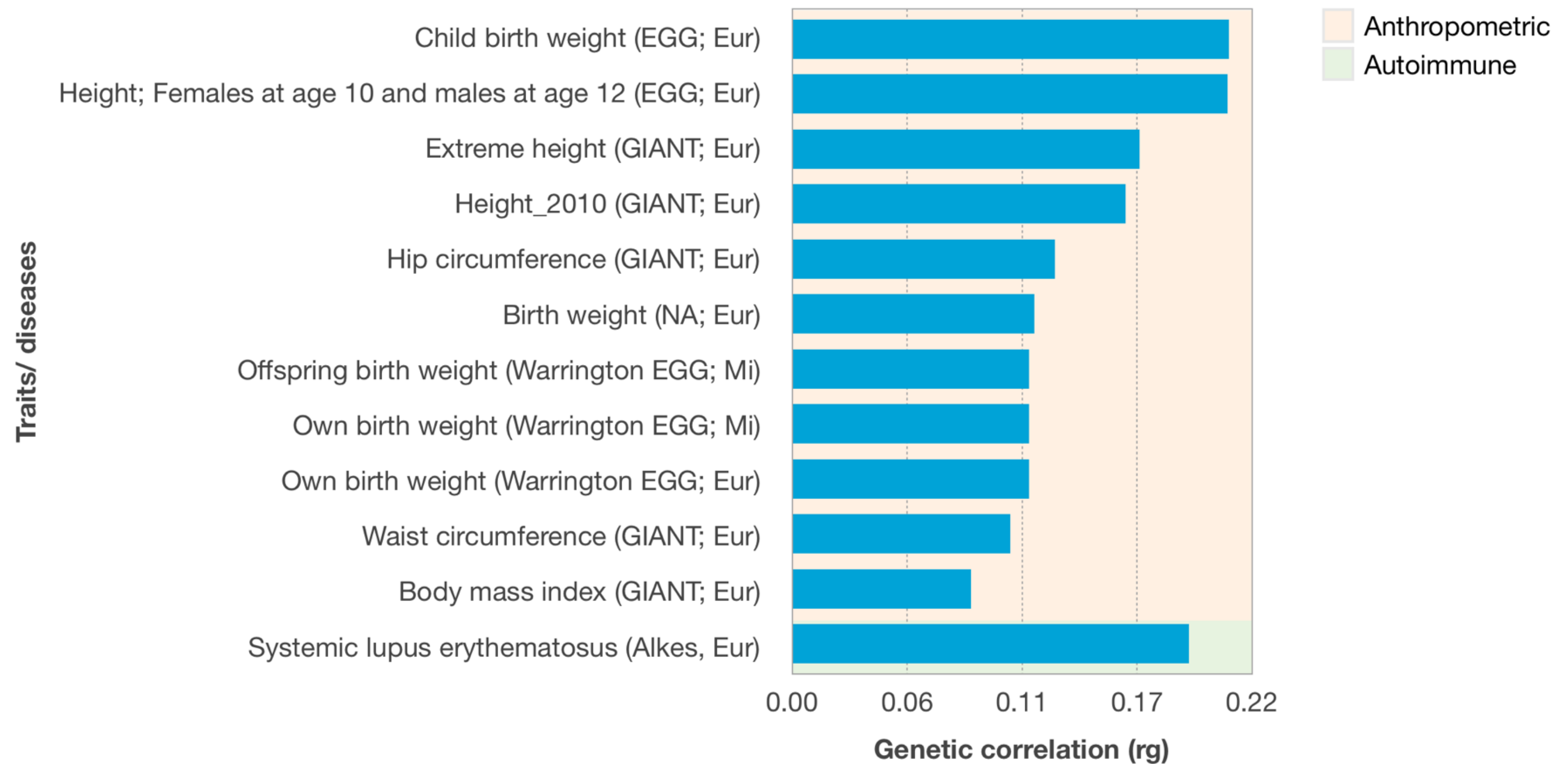
2.3.5 Genetic correlations with varicose veins associated phenotypes

The total contribution of common SNP variants to the varicose veins phenotype was calculated using LDSC Regression.[33] Using LD Scores from ~1.2 million variants found in European populations, the SNP-based heritability* (h^2g) for varicose veins in the discovery population was estimated to be 5.03% (S.E. = 0.30%). The h^2_g was found to be near-identical in the independent 23andMe cohort (h^2_g = 5.40%, S.E. = 0.30%).

Correlation testing between the polygenic architecture of varicose veins and 171 publicly available traits (nine trait categories) from the LDHub database[34] were calculated using LDSC regression.[33] Of the nine trait categories tested, two categories (anthropometric and autoimmune) contained traits that met the Bonferroni-corrected significance threshold (**Appendix Table 2.8**). All twelve significant traits positively correlated with varicose veins (**Figure 2.10**), eleven of which belonged to the anthropometric category (pertaining to height and weight phenotypes). In the autoimmune trait category, systemic lupus erythematosus was discovered to be correlated with varicose veins, demonstrating ~19% genetic overlap ($P = 4.2\times10^{-3}$, r_g = 0.195). Of note, Morris *et al.* previously found variant rs17321999-C at 2p23.1 to be associated with systemic lupus erythematosus ($P = 2.22\times10^{-16}$, OR = 0.83)[41], which is also significantly associated with varicose veins in the discovery analysis ($P = 3.20\times10^{-14}$, OR = 1.09), and is in high linkage with the lead SNP at locus 2p23.1, rs9967884 (r^2 = 0.87).

* i.e. the total phenotypic variance between the population attributable to the SNPs identified in this GWAS

Figure 2.10. Genetic correlation between varicose veins and other phenotypes from the LDHub Database. All twelve significant traits ($P < 5.56\times10^{-3}$) are positively correlated with varicose veins. The consortium and ethnicity for each study is provided in parentheses. *EGG, Early Growth Genetics Consortium; GIANT, Genetic Investigation of ANthropometric Traits Consortium ; Alkes, Alkes Group (Harvard T.H Chan School of Public Health); NA, Not Applicable. Eur, European ethnicity; Mi, Mixed ethnicity.*



2.3.6. Drug target enrichment analysis

The Open Targets Platform investigated the potential for therapeutic targeting of the protein products of 200 of the 237 prioritised genes.[35] Forty-two therapeutic pathways reached nominal significance ($P < 5\times10^{-2}$), with the Butyrophillin family interactions pathway demonstrating top enrichment ($P = 2.6\times10^{-7}$, six drug targets), followed by the Calcineurin-NFAT pathway ($P = 9.4\times10^{-4}$, two targets), and the Transcriptional regulation by RUNX1 pathway (highest number of enriched targets (n = 14), $P = 1.6\times10^{-3}$) (**Appendix Table 2.9**). Pharmacological tractability data for 105 gene targets were available, with 65 predicted to be tractable to antibody targeting to a high confidence, and 26 genes predicted to be tractable to small molecule targeting (**Appendix Table 2.10**). Eight genes had the highest level of evidence as pharmacologically active targets, with known pharmaceutical interactions (*CDK10, COL27A1, GABBR1, KCNJ2, MAPK10, OPRL1, TNC and VEGFA*) (**Appendix Table 2.11**). This includes *VEGFA*, which is a target for several antibody, protein and oligosaccharide agents which are currently under phase two, three and four clinical trials (clinicaltrials.gov) for several ocular vascular disorders.

2.3.7. Genetic risk score for varicose veins

The 49 independent significant signals associated with varicose veins were used to calculate a weighted genetic risk score (wGRS) for all 401,667 participants in UK Biobank. As expected, the wGRS for varicose veins cases (5.179) was higher than in controls (4.986; $P = 9.90 \times 10^{-324}$). Moreover, varicose veins cases that had undergone surgery had a higher wGRS (5.185) compared to varicose veins cases that had not undergone a surgical procedure to manage their disease (5.076; $P = 2.46 \times 10^{-13}$). Varicose veins cases without ulceration (5.184) were found to have a higher wGRS than cases with ulceration (5.092; $P = 3.65 \times 10^{-5}$) (**Table 2.5**).

Table 2.5. Weighted genetic risk score for varicose veins in the UK Biobank cohort

Group	VVs cases	Controls	P-value[†]	VVs cases with operation code	VVs cases without operation code	P-value[§]	VVs cases with ulceration code	VVs cases without ulceration code	P-value[±]
N	22,473	379,183		21,407	1,066		596	21877	
Mean wGRS* (standard deviation)	5.179 (0.454)	4.986 (0.451)	9.90×10^{-324}	5.185 (0.453)	5.076 (0.468)	2.46×10^{-13}	5.092 (0.543)	5.184 (0.543)	3.65×10^{-5}

*VVs, Varicose veins. *wGRS: weighted genetic risk score. [†]Unpaired two-tailed t-test between VVs cases and controls. [§]Unpaired two-tailed t-test between VVs cases with an operation code and VVs cases without an operation code. ±Unpaired two-tailed t-test between VVs cases with an ulceration code and VVs cases without an ulceration code.*

2.4. Discussion

2.4.1. Summary

Varicose veins are a major national health burden, resulting in significant patient morbidity and large healthcare costs, with treatment restricted to surgical intervention and compounded by high recurrence. There is a growing need to understand the biology of varicose veins in order to guide new therapeutic approaches. The present study represents hitherto the largest and most comprehensive association analysis of varicose veins, involving 135,514 cases and 675,111 controls. The association study identified twenty-eight novel risk loci (29 novel signals), and independently replicated 18 of 29 previously reported (but non-replicated) loci (20 of 32 known signals). The suite of *in silico* analyses used provide robust evidence of functional variants in varicose veins-associated genomic regions. Moreover, strong enrichment for genes expressed in the extracellular matrix, immune cell signalling and circulatory system development was demonstrated in the pathway analyses. Of note, a new genetic correlation between the polygenic architecture of varicose veins and systemic lupus erythematosus was discovered. Several prioritised genes show potential for pharmacological targeting, and are currently under active study in other disease models. Finally, the weighted genetic risk score correlated with a more severe prognosis – a fundamental step in facilitating personalised medicine approaches to varicose veins.

The 46 risk loci contain genes that cluster into five functional categories: angiogenesis and lymphangiogenesis, smooth muscle cell biology, extracellular matrix regulation, the immune response, and apoptosis (**Appendix Table 2.12**).

2.4.2. Angiogenesis

Angiogenesis plays a central role in the development and maintenance of healthy veins. Disruption in normal angiogenic processes can lead to varicose veins[42], potentially because of a failure to develop properly-formed venous walls and valves, or to repair defects following vascular stress injury.

The analysis yielded a total of nine presumed angiogenesis-related varicose veins susceptibility loci, most interestingly rs11967262 at 6p21.1 ($P_{meta} = 1.45\times10^{-19}$, OR = 1.09), which resides in an intergenic region ~7kb upstream of vascular endothelial growth factor A (*VEGFA*). VEGFA is a critical regulator of angiogenesis, and is fundamental to conserving the integrity and functionality of the vessel wall.[43] Binding to its receptor VEGR2, VEGFA functions as a selective endothelial mitogen, inducing endothelial cell proliferation, migration and differentiation. Hypoxia and mechanical stretch are key inducers of VEGFR2 expression via activation of the HIF pathway, which itself has been implicated in varicose veins.[44] VEGFA and VEGFR2 expression are considerably heightened in varicose vein walls compared to normal vein, especially in varicose veins complicated by thrombophlebitis.[45] VEGFA also functions as a potent vascular permeability factor[46], which is posited to lead to varicose vein progression and complications via fenestration of the endothelium.[47] The increased microvascular permeability may therefore be an important feature in the progression of varicose veins to chronic venous insufficiency (oedema and venous ulceration).[47] Plasma levels of VEGFA have also been shown to be significantly amplified in varicose veins patients.[48] VEGFA may therefore be implicated in the pathogenesis of varicose veins via its vasodilatory effects, which are thought to decrease vessel tone and lead

to stasis, as well as oxygen free radical release, causing weakness of the vessel wall.[47] VEGFA intriguingly also promotes inflammation via expression of vascular and intracellular adhesion molecules on endothelial cells, connecting angiogenesis to immune dysfunction.[49]

Furthermore, rs2713575 (P_{meta} = 1.82×10^{-36}, OR = 1.12) at 3q21.3 lies in an intronic region near *GATA2*. *GATA2* is a haematopoietic transcription factor necessary for vascular integrity that acts downstream of VEGF, and has been demonstrated to regulate VEGF-induced angiogenesis and lymphangiogenesis.[50] The VEGF axis may therefore be a promising candidate for therapeutic targeting in the treatment of varicose veins. To this end, several anti-VEGFA agents were enriched in our drug-target enrichment analysis, and are currently being investigated in randomised trials for the treatment of several proliferative retinopathies and retinal vein occlusion.

2.4.3. Lymphangiogenesis

Positional and MAGMA mapping highlighted genes at four loci relating to lymphangiogenesis. Indeed, the lymphatic system develops from veins, and its function is intimately linked to the venous circulation, draining extracellular fluid back into circulation. It is therefore feasible that similar genetic defects may result in either lymphoedema, varicose veins, or a combination of the two conditions. rs340875, (P_{meta} = 4.22×10^{-20}, OR = 1.09) is ~2kb upstream of *PROX1*, a master inducer gene necessary for the development of lymphatic vasculature.[51] PROX1 knock-out mice are deficient of lymphatic vasculature.[52] Of significance, during developmental lymphangiogenesis, *PROX1* has been shown to be necessary for the formation of lymphovenous valves.[53,54] This suggests that PROX1 dysfunction may predispose to varicose vein development by causing defects in venous valves. Furthermore, PROX1 is co-expressed and functions alongside the transcription factor FOXC2 in lymphatic valve forming cells at the earliest stage of lymphatic development.[55] Mutations in *FOXC2* cause hereditary lymphoedema-distichiasis, a disease which is characterised by varicose veins and peripheral lymphoedema – highlighting genetic overlap between the two disorders.[56] Previous twin studies have demonstrated linkage of the FOXC2 region with varicose veins[57,58], however the present study did not yield evidence of association between varicose veins and *FOXC2*.

2.4.4. Extracellular matrix regulation

Varicose veins show deposition of extracellular matrix (ECM) in the perivascular space – possibly a compensatory mechanism to buttress an already weakened wall.[59] Moreover, intimal hypertrophy and an increased vessel diameter (characteristic features of ECM disruption) are also seen in varicose veins.[60] This luminal dilatation in varicose veins, alongside valve ring enlargement, is thought to compromise the ability of venous valves to co-apt, compounding venous reflux and leading to stasis and venous hypertension.[4] Of note, a marked imbalance of the structural ECM proteins, collagen and elastin is seen in varicose veins, with a preponderance of collagen compared to normal veins.[61] Redundancy in the connective tissue components of the valves or venous wall may therefore predispose to varicose veins pathology.

The prioritised genes significantly overlapped with canonical pathways relating to ECM components, including Collagen Type XXVII Alpha 1 Chain (*COL27A1*), EGF-containing Fibulin-like Extracellular Matrix Protein 1 (*EFEMP1)* and Fibulin-7 (*FBLN7*).

rs753085 ($P = 2.17\times10^{-11}$, OR = 1.07) resides within an intronic region of Collagen Type XXVII Alpha 1 Chain (*COL27A1*). COL27A1 is a fibrillar collagen in the extracellular matrices of several tissues, including blood vessels.[62] Diminished expression of *COL27A1* has been demonstrated in varicose vein samples.[63] The drug enrichment analysis demonstrated the potential candidacy of COL27A1 as a therapeutic target for varicose vein prevention or treatment, with multiple pharmaceutical agents currently being investigated in several clinical trials.

Another lead signal, rs4849044, lies in an intronic region of *FBLN7* at 2q13 (P_{meta} = 1.98×10^{-14}, OR = 1.07). Fibulins are secreted glycoproteins that stabilise the ECM and are expressed in matrices, elastic fibres and basement membranes.[64] FBLN7 is a cell adhesion molecule that interacts with extracellular matrix proteins, and is highly expressed in blood vessels. The C-terminal fragment of FBLN7 (FBLN7-C) demonstrates anti-angiogenic activity, binding to venous endothelium and disrupting tube formation and vessel sprouting.[65] Of note, *FBLN7* was mapped to the 2q13 locus by all four mapping strategies, and contains eQTL variants which associate *FBLN7* to varicose veins through pleiotropy in the SMR analysis, demonstrating its candidacy as a potential functional player in varicose veins biology.

My study replicated the previously reported association between rs3791679 and varicose veins (P = 1.59×10^{-13}, OR = 1.08), mapped to *EFEMP1*.[66] *EFEMP1* encodes another member of the fibulin family, fibulin-3[67], which is highly expressed in venous endothelia. Fibulin-3 antagonises vascular development through its effect on decreasing expression of the matrix metalloproteinases, MMP2 and MMP3, and increasing expression of MMP inhibitors (TIMPs) in endothelial tissue.[68] To this end, varicose veins have been found to demonstrate a characteristic reduction in expression of MMP2, and heightened expression of TIMP1 and MMP1 protein levels within the saphenofemoral junction.[69] Altered expression of these enzymes may therefore precipitate inherent weakness in the vein wall, predisposing patients to varicose veins. rs3791679 is a notable polymorphism at this locus, with 45 associations across 12 traits, including carpal tunnel syndrome, joint hypermobility, and several anthropometric measures of BMI and height (which have been previously

associated with varicose veins in epidemiological studies).[11,70,71] Consistently, the LDSC genetic correlation analysis demonstrated striking genetic overlap between varicose veins and height and weight phenotypes. Moreover, the drug enrichment analysis identified fibulin-3 to be tractable to antibody targeting with high confidence, and metformin has been previously demonstrated to perturb fibulin-3 levels through inhibition at the transcriptional level.[72] Fibulin-3 therefore necessitates further study as a potential therapeutic target for varicose veins.

2.4.5. Immune response

Heightened expression of inflammatory mediators has been observed in varicose veins compared to normal veins.[4] Specifically, varicose vein walls show increased mast cells, monocytes and macrophages compared to normal veins.[4] Chronic inflammation in the vein wall has therefore been postulated to be a key feature of varicose vein biology.[60]

The association analysis defined five inflammation-associated risk loci, in particular rs78216177 ($P_{meta} = 5.80\times10^{-14}$, OR = 1.10), which lies in an intron of *DOCK8*. *DOCK8* plays a significant role in the innate and adaptive immune systems, with *DOCK8* deletion strongly associated with Hyper-IgE syndrome, a type of primary immunodeficiency that affects multiple systems including the vasculature.[73] Indeed, vascular abnormalities in hyper-IgE syndrome include arterial dilating pathology, aneurysmal changes, and abnormalities in great vessels. These occur in a different vascular territory to varicose veins, and are thought to have overlapping pathological features.

Using publicly available GWAS data, a substantial genetic overlap between varicose veins and systemic lupus erythematosus was discovered. Lead variant rs1471251 ($P = 8.33\times10^{-11}$, OR = 1.06) is an eQTL of *AFF1,* and has been associated with systemic lupus erythematosus.[74] Supporting this shared polygenic architecture, the C allele of SNP variant rs17321999, which associates significantly with varicose veins, also increases the risk of systemic lupus.[41]

Canonical pathway analysis within XGR demonstrated enrichment for *Intracellular calcium signalling in the CD4+ T-Cell Receptor (TCR) pathway* (P = 1.9×10^{-3}, Z = 3.5), specifically highlighting Nuclear Factor of Activated T-Cells, Cytoplasmic, Calcineurin-Dependent 2 (*NFATC2*) and Protein Phosphatase 3 Regulatory Subunit B, Alpha (*PPP3R1*) which are closely involved in this pathway. Two significant signals were discovered at: i) 20q13.2 – an intronic region of *NFATC2* (rs3787184 (P_{meta} = 2.51×10^{-36}, OR = 1.16)) and ii) 2p14 – an intergenic region ~19kb upstream of *PPP3R1* (rs2861819 (P_{meta} = 2.65×10^{-77}, OR = 1.20)). *PPP3R1* encodes Calcineurin subunit B type 1, a Ca^{2+} influx-induced serine/threonine-specific phosphatase, which, alongside NFAT transcription factors, regulates the activation of native T-Cells.[75] Varicose veins are characterised by clustering and infiltration of T lymphocytes[59], which are distributed proximate to the venous valve agger[*].[76] Therefore, altered calcium signalling in T-Cells through aberrant *PPP3R1* and *NFATC2* signalling may be involved in the valvular pathology depicted in varicose veins.

[*] a fibroelastic structure located at the base of venous valves where media meets adventitia

2.4.6. Vascular smooth muscle cell proliferation and migration

Varicose vein walls demonstrate a pathologically altered phenotype defined by vein wall remodelling, consisting of vascular smooth muscle cell (vSMC) hypertrophy, proliferation and migration into intima.[77–79] VSMCs in varicose veins are disarranged and undergo de-differentiation from a contractile to a synthetic phenotype. These changes impair the normal contractile function of SMCs in varicose vein tissue.[78,79] My study implicates for the first time in GWAS, genes that might be intimately involved in this process.

Six loci related to vSMC proliferation and migration were identified, most notably rs7518191 at 1p13.2 (P_{meta} = 6.90×10^{-14}, OR = 1.07) which lies in an intron of Tetraspanin 2 (*TSPAN2*), and is an eQTL for *TSPAN2* in several tissues, including tibial artery (P_{eQTL} = 1.3×10^{-6}). Tetraspanins are expressed at cell surfaces where they function in cell adhesion, cell migration, proliferation and differentiation.[80] *TSPAN2* is selectively enriched in vSMCs within blood vessels; its expression is closely associated with vSMC differentiation.[80] However, *TSPAN2* expression is inhibited when the vSMC undergoes phenotypic modulation in diseased human vessel, which could perhaps lead to the vSMC de-differentiation and migration observed in varicose veins.[81,82]

2.4.7. Apoptosis

Varicose veins demonstrate a significantly reduced expression of components in the intrinsic apoptotic pathway, specifically of bax and Caspase 9.[59] Furthermore, reduction in *Cyclin-D1* and over-expression of *BCL2* has also been demonstrated in the media and intima of varicose veins compared to normal tissue.[83] De-differentiation of vSMC away from a contractile phenotype in varicose veins is thought to be caused by, or at least exacerbated by, disruption in apoptosis.[83,84] Apoptosis may therefore be a contributory factor in the pathogenesis of varicose veins.

The GWAS identified two novel apoptosis-related loci. rs7308356 ($P_{meta} = 3.02 \times 10^{-20}$, OR = 1.09) is in a highly conserved intronic region within Ceramide Synthase 5 (*CERS5*), where it is a known eQTL for CERS5.[85] Ceramides are a key group of enzymes which are involved in cell death, differentiation and senescence.[86,87] *CERS5* overexpression facilitates apoptosis and autophagy, and it is overexpressed in several tumours, including colorectal and colon cancers.[88] Inhibition of CERS5 could prevent the normal apoptotic response of damaged endothelium following vascular injury.

rs72683923 ($P = 3.62 \times 10^{-14}$, OR = 1.28), is an exonic synonymous SNP in cyclin-dependent kinase-like 1 (CDKL1), which interacts with cyclin to regulate cell cycle, differentiation and apoptosis[89]. CDKL1 disruption inhibits cell proliferation, promoting apoptosis in breast cancer and melanoma. Further, *in vitro* knockout of CDKL1 suppresses cell proliferation and promotes apoptosis.[90] Therefore, one can postulate that decreased CDKL1 activity in endothelial cells may lead to increased apoptosis and predispose to varicose veins development.

2.4.8. Genetic risk score

In the USA, over two million participants have advanced chronic venous disease[91], and around half a million require invasive surgical procedures annually. The weighted genetic risk score derived from the replicated signals was found to correlate with disease severity, with varicose veins cases managed surgically possessing a higher genetic burden than those managed non-surgically. This finding suggests that those who were *phenotypically* severe were also *genotypically* more susceptible. This represents a proof-of-principle, demonstrating the feasibility for data-driven prognostication in enabling the identification of varicose veins cases that are more likely to require surgical intervention. This could foreseeably guide medical and surgical management, such as the use of early preventive approaches in high risk participants; these might include prophylactic compression stockings or early ablation procedures to mitigate risk of venous ulceration. Indeed, the efficacy of early endovenous ablation in improving outcomes of venous leg ulcers has been shown.[92] However, among cases with ulceration, the wGRS did not correlate with severity, suggesting i) other variants not identified in this study may be involved in ulceration risk, ii) ulceration may be less sensitive to genetic contributions and more a product of non-genetic risk factors (such as orthostatic professions), or iii) this part of the study was underpowered to detect ulceration-specific loci, given there were only 596 cases with ulceration (2.65% of overall cases).

2.4.9 Strengths and limitations

Several limitations of the present study must be recognised. Firstly, while the discovery cohort in UK Biobank used a combination of hospital diagnostic, operative and self-report codes, varicose veins cases in the 23andMe cohort were defined *solely* by self-report codes, meaning that the phenotyping for the replication study was necessarily less stringent. Moreover, instead of performing a formal meta-analysis between the discovery and replication GWAS, the association of only the 118 independent lead GWS SNPs from the UK Biobank cohort were independently tested. Thus, sub-threshold signals in the discovery analysis that may have reached significance in the replication GWAS, or under meta-analysis, were not identified. Finally, the unavailability of the full summary statistics for the replication GWAS restricted our *in silico* analyses to the summary statistics from the discovery GWAS alone.

However, several strengths go in some way to lend credence to the study findings. The present study was performed in a total of 135,514 cases and 675,111 controls, representing the largest association study of varicose veins by a substantial margin. Moreover, the false positive rate was rigidly controlled by reporting only the loci that were associated in the discovery cohort at genome-wide levels of significance *and* that subsequently replicated in 23andMe, hence the 49 variants reported here are likely to represent true signals. This notion is substantiated by the fact that the associated loci mapped to a plethora of biologically plausible genes, which show clustering across several connected pathways. This study represents a fundamental step in the use of genetic risk scoring in enabling better prognostication and decision-making in the management of varicose veins.

In this study, the inclusion of operative codes for phenotyping in the UK Biobank discovery cohort enabled a considerably larger number of cases to be identified than a previous GWAS that also used the UK Biobank resource but relied solely on ICD diagnostic codes (22,473 cases vs 9,577 cases).[5] As a fundamental principle of case ascertainment in surgical diseases, it is necessary to identify participants that have undergone surgery for a disease: given the inevitable risk of complications, surgery is generally reserved for participants at the phenotypically severe end of the disease that may have failed non-surgical treatment. The weighted genetic risk score demonstrates that varicose veins cases that had undergone surgery were also, on the whole, genotypically more severe. This finding lends further validity to the forty-nine identified signals.

2.5. Conclusion

This chapter presents the largest association study of varicose veins, a common disease associated with significant patient morbidity, reduced quality-of-life, and high socioeconomic burden. Forty-nine variants at 46 susceptibility loci were discovered to associate with varicose veins, with associated genomic regions mapped to new genes and pathways that are involved in angiogenesis, lymphangiogenesis, extracellular matrix regulation, inflammation, vascular smooth muscle cell activity, and apoptosis. Identified genes and pathways demonstrate striking representation along biologically viable pathways, suggesting they are eminently plausible contributors to the pathogenesis of varicose veins. A number of genes represent promising candidates for further investigation of venous biology. Notably, *VEGFA*, *COL27A1*, *EFEMP1*, *PPP3R1* and *NFATC2* represent probable 'key players' as potential therapeutic targets in the treatment of varicose veins. Lastly, the demonstration that genetic risk score correlates with disease severity represents a fundamental step towards improved prognostication of varicose veins patients.

2.6. Chapter References

1 Laing W. Chronic Venous Diseases of the Leg. Off. Heal. Econ. London. 1992. https://www.ohe.org/publications/chronic-venous-diseases-leg (accessed Jan 1, 2020).

2 O'Donnell TF, Balk EM, Dermody M, Tangney E, Iafrati MD. Recurrence of varicose veins after endovenous ablation of the great saphenous vein in randomized trials. J. Vasc. Surg. Venous Lymphat. Disord. 2016; **4**: 97–105.

3 Scott TE, LaMorte WW, Gorin DR MJ. Risk factors for chronic venous insufficiency: a dual case-control study. *J Vasc Surg* 1995; **22**: 622–8.

4 Lim CS, Davies AH. Pathogenesis of primary varicose veins. *Br J Surg* 2009; **96**: 1231–42.

5 Fukaya E et al. Clinical and Genetic Determinants of Varicose Veins. *Circulation* 2018; : 1–12.

6 Shadrina AS, Sharapov SZ, Shashkova TI, Tsepilov YA. Varicose veins of lower extremities: Insights from the first large-scale genetic study. *PLoS Genet* 2019; **15**: e1008110.

7 Ellinghaus E, Ellinghaus D, Krusche P, *et al.* Genome-wide association analysis for chronic venous disease identifies EFEMP1 and KCNH8 as susceptibility loci. *Sci Rep* 2017; **7**. DOI:10.1038/srep45652.

8 Allen NE, Sudlow C, Peakman T, Collins R. UK biobank data: Come and get it. Sci. Transl. Med. 2014; **6**. DOI:10.1126/scitranslmed.3008601.

9 Bycroft C, Freeman C, Petkova D, *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* 2018; **562**: 203–9.

10 Sudlow C, Gallacher J, Allen N, *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and

Old Age. *PLoS Med* 2015; **12**. DOI:10.1371/journal.pmed.1001779.

11 Wiberg A, Ng M, Schmid AB, *et al.* A genome-wide association analysis identifies 16 novel susceptibility loci for carpal tunnel syndrome. *Nat Commun* 2019; **10**: 1–12.

12 Loh PR, Tucker G, Bulik-Sullivan BK, *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet* 2015; **47**: 284–90.

13 Durand EY, Do CB, Mountain JL, Macpherson JM. Ancestry Composition: A Novel, Efficient Pipeline for Ancestry Deconvolution. *bioRxiv* 2014; : 010512.

14 Auton A, Abecasis GR, Altshuler DM, *et al.* A global reference for human genetic variation. Nature. 2015; **526**: 68–74.

15 O'Connell J, Sharp K, Shrine N, *et al.* Haplotype estimation for biobank-scale data sets. *Nat Genet* 2016; **48**: 817–20.

16 Walter K, Min JL, Huang J, *et al.* The UK10K project identifies rare variants in health and disease. *Nature* 2015; **526**: 82–9.

17 Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* 2007; **81**: 1084–97.

18 Loh PR, Palamara PF, Price AL. Fast and accurate long-range phasing in a UK Biobank cohort. *Nat Genet* 2016; **48**: 811–6.

19 Das S, Forer L, Schönherr S, *et al.* Next-generation genotype imputation service and methods. *Nat Genet* 2016; **48**: 1284–7.

20 Mägi R, Morris AP. GWAMA: Software for genome-wide association meta-analysis. *BMC Bioinformatics* 2010; **11**. DOI:10.1186/1471-2105-11-288.

21 Watanabe K, Taskesen E, Van Bochoven A, Posthuma D. Functional mapping

and annotation of genetic associations with FUMA. *Nat Commun* 2017; **8**: 1–10.

22 D. W, J. M, J. M, *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res*; **42**: D1001–6.

23 Wang KLMHH. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010; **38**: e164.

24 Boyle AP, Hong EL, Hariharan M, *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* 2012; **22**: 1790–7.

25 Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* 2018; **47**: D886–94.

26 Ernst J, Kellis M. Chromatin-state discovery and genome annotation with ChromHMM. *Nat Protoc* 2017; **12**: 2478–92.

27 Kircher M, Witten DM, Jain P, O'roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014; **46**: 310–5.

28 Cunningham F, Achuthan P, Akanni W, *et al.* Ensembl 2019. *Nucleic Acids Res* 2019; **47**: D745–51.

29 de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLoS Comput Biol* 2015; **11**: 1–20.

30 Zhu Z, Zhang F, Hu H, *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* 2016; **48**: 481–7.

31 Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst* 2015; **1**: 417–25.

32 Fang H, Knezevic B, Burnham KL, Knight JC. XGR software for enhanced

interpretation of genomic summary data, illustrated by application to immunological traits. *Genome Med* 2016; **8**: 1–20.

33 Bulik-Sullivan B, Loh PR, Finucane HK, *et al.* LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* 2015; **47**: 291–5.

34 Zheng J, Erzurumluoglu AM, Elsworth BL, *et al.* LD Hub: A centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* 2017; **33**: 272–9.

35 Carvalho-Silva D, Pierleoni A, Pignatelli M, *et al.* Open Targets Platform: new developments and updates two years on. *Nucleic Acids Res* 2019; **47**: D1056–65.

36 De Jager PL, Chibnik LB, Cui J, *et al.* Integration of genetic risk factors into a clinical algorithm for multiple sclerosis susceptibility: a weighted genetic risk score. *Lancet Neurol* 2009; **8**: 1111–9.

37 Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res* 2003; **31**: 3812–4.

38 Adzhubei IA, Schmidt S, Peshkin L, *et al.* A method and server for predicting damaging missense mutations. Nat. Methods. 2010; **7**: 248–9.

39 Lukacs V, Mathur J, Mao R, *et al.* Impaired PIEZO1 function in patients with a novel autosomal recessive congenital lymphatic dysplasia. *Nat Commun* 2015; **6**: 1–7.

40 Aguet F, Brown AA, Castel SE, *et al.* Genetic effects on gene expression across human tissues. *Nature* 2017; **550**: 204–13.

41 Morris DL, Sheng Y, Zhang Y, *et al.* Genome-wide association meta-analysis in

Chinese and European individuals identifies ten new loci associated with systemic lupus erythematosus. *Nat Genet* 2016; **48**: 940–6.

42 Chang MY, Chiang PT, Chung YC, *et al.* Apoptosis and Angiogenesis in Varicose Veins Using Gene Expression Profiling. *Fooyin J Heal Sci* 2009; **1**: 85–91.

43 Ferrara N. Molecular and biological properties of vascular endothelial growth factor. J. Mol. Med. 1999; **77**: 527–43.

44 Lim CS, Kiriakidis S, Sandison A, Paleolog EM, Davies AH. Hypoxia-inducible factor pathway and diseases of the vascular wall. J. Vasc. Surg. 2013; **58**: 219–30.

45 Kowalewski R, Małkowski A, Sobolewski K, Gacko M. Vascular endothelial growth factor and its receptors in the varicose vein wall Czynnik wzrostu śródbłonka naczyniowego i jego receptory w ścianie żylaków kończyn. 2011.

46 Esser S, Wolburg K, Wolburg H, Breier G, Kurzchalia T, Risau W. Vascular endothelial growth factor induces endothelial fenestrations in vitro. *J Cell Biol* 1998; **140**: 947–59.

47 Hollingsworth SJ, Powell GL, Barker SGE, Cooper DG. Primary varicose veins: Altered transcription of VEGF and its receptors (KDR, flt-1, soluble flt-1) with sapheno-femoral junction incompetence. *Eur J Vasc Endovasc Surg* 2004; **27**: 259–68.

48 Howlader MH, Coleridge Smith PD. Relationship of plasma vascular endothelial growth factor to CEAP clinical stage and symptoms in patients with chronic venous disease. *Eur J Vasc Endovasc Surg* 2004; **27**: 89–93.

49 Kim I, Moon SO, Kim SH, Kim HJ, Koh YS, Koh GY. Vascular Endothelial Growth Factor Expression of Intercellular Adhesion Molecule 1 (ICAM-1),

Vascular Cell Adhesion Molecule 1 (VCAM-1), and E-selectin through Nuclear Factor-κB Activation in Endothelial Cells. *J Biol Chem* 2001; **276**: 7614–20.

50 Spinner MA, Sanchez LA, Hsu AP, *et al.* GATA2 deficiency: A protean disorder of hematopoiesis, lymphatics, and immunity. *Blood* 2014; **123**: 809–21.

51 Harvey NL, Srinivasan RS, Dillard ME, *et al.* Lymphatic vascular defects promoted by Prox1 haploinsufficiency cause adult-onset obesity. *Nat Genet* 2005; **37**: 1072–81.

52 Wigle JT, Oliver G. Prox1 function is required for the development of the murine lymphatic system. *Cell* 1999; **98**: 769–78.

53 Bazigou E, Lyons OTA, Smith A, *et al.* Genes regulating lymphangiogenesis control venous valve formation and maintenance in mice. *J Clin Invest* 2011; **121**: 2984–92.

54 Sathish Srinivasan R, Oliver G. Prox1 dosage controls the number of lymphatic endothelial cell progenitors and the formation of the lymphovenous valves. *Genes Dev* 2011; **25**: 2187–97.

55 Sabine A, Agalarov Y, Maby-ElHajjami H, *et al.* Mechanotransduction, PROX1, and FOXC2 Cooperate to Control Connexin37 and Calcineurin during Lymphatic-Valve Formation. *Dev Cell* 2012; **22**: 430–45.

56 Fang J, Dagenais SL, Erickson RP, *et al.* Mutations in FOXC2 (MFH-1), a forkhead family transcription factor, are responsible for the hereditary lymphedema-distichiasis syndrome. *Am J Hum Genet* 2000; **67**: 1382–8.

57 Serra R, Buffone G, De Franciscis A, *et al.* A genetic study of chronic venous insufficiency. *Ann Vasc Surg* 2012; **26**: 636–42.

58 Mellor RH, Brice G, Stanton AWB, *et al.* Mutations in FOXC2 Are Strongly Associated With Primary Valve Failure in Veins of the Lower Limb. 2007; : 1912–

21.

59 Segiet OA, Brzozowa-Zasada M, Piecuch A, Dudek D, Reichman-Warmusz E, Wojnicz R. Biomolecular mechanisms in varicose veins development. *Ann Vasc Surg* 2015; **29**: 377–84.

60 Oklu R, Habito R, Mayr M, *et al.* Pathogenesis of varicose veins. *J Vasc Interv Radiol* 2012; **23**: 33–9.

61 Gandhi RH, Irizarry E, Nackman GB, Halpern VJ, Mulcare RJ, Tilson MD. Analysis of the connective tissue matrix and proteolytic activity of primary varicose veins. *J Vasc Surg* 1993; **18**: 814–20.

62 Pace JM, Corrado M, Missero C, Byers PH. Identification, characterization and expression analysis of a new fibrillar collagen gene, COL27A1. *Matrix Biol* 2003; **22**: 3–14.

63 Markovic JN, Shortell CK. Genomics of varicose veins and chronic venous insufficiency. Semin. Vasc. Surg. 2013; **26**: 2–13.

64 De Vega S, Iwamoto T, Yamada Y. Fibulins: Multiple roles in matrix structures and tissue functions. Cell. Mol. Life Sci. 2009; **66**: 1890–902.

65 De Vega S, Suzuki N, Nonaka R, *et al.* A C-terminal fragment of fibulin-7 interacts with endothelial cells and inhibits their tube formation in culture. *Arch Biochem Biophys* 2014; **545**: 148–53.

66 Ellinghaus E, Ellinghaus D, Krusche P, *et al.* Genome-wide association analysis for chronic venous disease identifies EFEMP1 and KCNH8 as susceptibility loci. *Sci Rep* 2017; **7**: 1–9.

67 Zhang Y, Marmorstein LY. Focus on molecules: Fibulin-3 (EFEMP1). Exp. Eye Res. 2010; **90**: 374–5.

68 Albig AR, Neil JR, Schiemann WP. Fibulins 3 and 5 antagonize tumor

angiogenesis in vivo. *Cancer Res* 2006; **66**: 2621–9.

69 Parra JR, Cambria RA, Hower CD, *et al.* Tissue inhibitor of metalloproteinase-1 is increased in the saphenofemoral junction of patients with varices in the leg. *J Vasc Surg* 1998; **28**: 669–75.

70 Lee AJ, Evans CJ, Allan PL, Ruckley C V, Fowkes FGR. Lifestyle factors and the risk of varicose veins: Edinburg Vein Study. *J Clin Epidemiol* 2003; **56**: 171–9.

71 Sisto T, Reunanen A, Laurikka J, *et al.* Prevalence and risk factors of varicose veins in lower extremities: Mini-Finland health survey. *Eur J Surgery, Acta Chir* 1995; **161**: 405–14.

72 Gao L-B, Tian S, Gao H-H, Xu Y-Y. Metformin inhibits glioma cell U251 invasion by downregulation of fibulin-3. *Neuroreport* 2013; **24**: 504–8.

73 Engelhardt KR, McGhee S, Winkler S, *et al.* Large deletions and point mutations involving the dedicator of cytokinesis 8 (DOCK8) in the autosomal-recessive form of hyper-IgE syndrome. *J Allergy Clin Immunol* 2009; **124**. DOI:10.1016/j.jaci.2009.10.038.

74 Okada Y, Shimane K, Kochi Y, *et al.* A Genome-Wide Association Study Identified AFF1 as a Susceptibility Locus for Systemic Lupus Eyrthematosus in Japanese. *PLoS Genet* 2012; **8**: e1002455.

75 Lewis RS. Calcium Signaling Mechanisms in T Lymphocytes. *Annu Rev Immunol* 2001; **19**: 497–521.

76 Sayer GL, Smith PDC. Immunocytochemical Characterisation of the Inflammatory Cell Infiltrate of Varicose Veins. *Eur J Vasc Endovasc Surg* 2004; **28**: 479–83.

77 Jacobs BN, Andraska EA, Obi AT, Wakefield TW. Pathophysiology of varicose

veins. *J Vasc Surg Venous Lymphat Disord* 2017; **5**: 460–7.

78 Wali MA, Eid RA. Smooth muscle changes in varicose veins: An ultrastructural study. *J Smooth Muscle Res* 2001; **37**: 123–35.

79 Wali MA, Eid RA. Intimal changes in varicose veins: An ultrastructural study. *J Smooth Muscle Res* 2002; **38**: 63–74.

80 Todd SC, Doctor VS, Levy S. Sequences and expression of six new members of the tetraspanin/TM4SF family. *Biochim Biophys Acta - Gene Struct Expr* 1998; **1399**: 101–4.

81 Zhao J, Wu W, Zhang W, *et al.* Selective expression of TSPAN2 in vascular smooth muscle is independently regulated by TGF-b1/smad and myocardin/serum response factor. *FASEB J* 2017; **31**: 2576–91.

82 Halayko AJ, Solway J. Invited review: Molecular mechanisms of phenotypic plasticity in smooth muscle cells. J. Appl. Physiol. 2001; **90**: 358–68.

83 Ascher E, Jacob T, Hingorani A, Gunduz Y, Mazzariol F, Kallakuri S. Programmed cell death (apoptosis) and its role in the pathogenesis of lower extremity varicose veins. In: Annals of Vascular Surgery. Springer New York, 2000: 24–30.

84 Ascher E, Jacob T, Hingorani A, Tsemekhin B, Gunduz Y. Expression of molecular mediators of apoptosis and their role in the pathogenesis of lower-extremity varicose veins. *J Vasc Surg* 2001; **33**: 1080–6.

85 Brachtendorf S, Wanger RA, Birod K, *et al.* Chemosensitivity of human colon cancer cells is influenced by a p53-dependent enhancement of ceramide synthase 5 and induction of autophagy. *Biochim Biophys Acta - Mol Cell Biol Lipids* 2018; **1863**: 1214–27.

86 Hannun YA. Functions of ceramide in coordinating cellular responses to stress.

Science (80-) 1996; **274**: 1855–9.

87 Hannun YA, Obeid LM. Sphingolipids and their metabolism in physiology and disease. Nat. Rev. Mol. Cell Biol. 2018; **19**: 175–91.

88 Chen L, Chen H, Li Y, Li L, Qiu Y, Ren J. Endocannabinoid and ceramide levels are altered in patients with colorectal cancer. *Oncol Rep* 2015; **34**: 447–54.

89 Montini E, Andolfi G, Caruso A, *et al.* Identification and characterization of a novel serine-threonine kinase gene from the Xp22 region. *Genomics* 1998; **51**: 427–33.

90 Wang Y, Huang Q. Downregulation of CDKL1 promotes gastric cell apoptosis through inhibiting cell growth and colony formation. 2017.

91 Gloviczki P, Comerota AJ, Dalsing MC, *et al.* The care of patients with varicose veins and associated chronic venous diseases: Clinical practice guidelines of the Society for Vascular Surgery and the American Venous Forum. *J Vasc Surg* 2011; **53**: 2S-48S.

92 Gohel MS, Heatley F, Liu X, *et al.* A randomized trial of early endovenous ablation in venous ulceration. *N Engl J Med* 2018; **378**: 2105–14.

.

2.7. Chapter Appendix

The appendix for this chapter is provided as an online supplement at the following URL: **bit.ly/WAhmed_C2Appendix**

Table of Contents

1. Appendix Tables

Appendix Table 2.1. Additional variants associated with varicose veins in discovery cohort

Appendix Table 2.2. Varicose veins associated exonic variants at the replicated loci

Appendix Table 2.3. Predicted functional intronic and intergenic variants at the replicated loci

Appendix Table 2.4. Genome-wide gene-based association analysis in MAGMA

Appendix Table 2.5. Summary-based Mendelian Randomisation (SMR) using eQTL data from GTEx v7 tibial artery

Appendix Table 2.6. Genes mapped to the varicose veins-associated loci using the four mapping strategies

Appendix Table 2.7. Enriched gene sets from genome-wide gene-based enrichment analysis in MAGMA v1.07

Appendix Table 2.8. Genetic correlation between varicose veins and other phenotypes

Appendix Table 2.9. Enriched drug pathways from the drug target enrichment analysis

Appendix Table 2.10. Tractability information for targets in the drug-target enrichment analysis

Appendix Table 2.11. Pharmacologically-active targets identified in the drug-target enrichment analysis

Appendix Table 2.12. Functional categories of the gene clusters

Chapter 3: Genome-wide association analysis of haemorrhoids

3.1. Introduction

3.1.1. Rationale and aims

In **Chapter 1**, I substantiate haemorrhoids disease as a complex disorder with a multi-factorial aetiology, including elastic tissue dysfunction (the sliding anal lining theory).[1,2] Several studies report a positive family history among patients with haemorrhoids[3–6], however heritability estimates are lacking for haemorrhoids, as are, candidate genes involved in its pathobiology. The aim of this chapter is to undertake the first ever genome-wide association study (GWAS) of haemorrhoids to advance our understanding of the genetic architecture of haemorrhoids and to discover clinically-relevant biologic pathways, prioritise targets for therapeutic development, and to enhance personalised medicine approaches to haemorrhoids through genetic risk scoring.

3.2. Methods

3.2.1. Ethics and consent

The research and consent procedures of the UK Biobank are provided in **Chapter 2 (Section 2.2.1.)**.[7]

3.2.2. Study participants

A complete description of the study participants of the UK Biobank cohort are provided in **Chapter 2 (Section 2.2.2.)**.[8]

Haemorrhoids cases were defined as such if they had at least one of the following diagnostic or operative codes consistent with haemorrhoids (**Table 3.1)**:

1. Primary and/or secondary ICD-10 codes for haemorrhoids *(I84.0, I84.1, I84.2, I84.3, I84.4, I84.5, I84.7, I84.8, I84.9, K64.0, K64.1, K64.2, K64.3, K64.8 and K64.9)*
2. Primary and/or secondary OPCS code for haemorrhoids surgery *(H51.1, H51.3, H51.8, H51.9, H52, H53.2, H53.3, H53.8, H53.9, L70.3)*
3. Self-reported operation code for haemorrhoids surgery *(1483)*
4. Self-reported non-cancer illness code for haemorrhoids *(1505)*

In total, 39,950 UK Biobank participants had at least one diagnostic or operative code indicative of haemorrhoids and were therefore defined as cases.

Table 3.1. Diagnostic codes used for haemorrhoids case definition. The total number of participants with each of the diagnostic codes are described below. A total of 39,950 participants possessed at least one of the diagnostic codes for haemorrhoids.

Source of Data	UK Biobank Data Field	Code	Description	N
Primary ICD-10	41202	I84.0	Internal thrombosed haemorrhoids	19019
		I84.1	Internal haemorrhoids with other complications	
		I84.2	Internal haemorrhoids without complication	
		I84.3	External thrombosed haemorrhoids	
		I84.4	External haemorrhoids with other complications	
		I84.5	External haemorrhoids without complication	
		I84.7	Unspecified thrombosed haemorrhoids	
		I84.8	Unspecified haemorrhoids with other complications	
		I84.9	Unspecified haemorrhoids without complication	
		K64.0	First degree haemorrhoids	
		K64.1	Second degree haemorrhoids	
		K64.2	Third degree haemorrhoids	
		K64.3	Fourth degree haemorrhoids	
		K64.8	Other specified haemorrhoids	
		K64.9	Haemorrhoids, unspecified	
Secondary ICD-10	41204	As above	As above	16425
Primary OPCS	41200	H51.1	Haemorrhoidectomy	10198
		H51.3	Stapled haemorrhoidectomy	
		H51.8	Other specified excision of haemorrhoid	
		H51.9	Unspecified excision of haemorrhoid	
		H52	Destruction of haemorrhoid	
		H53.2	Forced manual dilation of anus for haemorrhoid	
		H53.3	Manual reduction of prolapsed haemorrhoid	
		H53.8	Other specified other operations on haemorrhoid	
		H53.9	Unspecified other operations on haemorrhoid	
		L70.3	Ligation of artery NEC	
Secondary OPCS	41210	As above	As above	2491

Non-cancer illness (self-report)	20002	1505	Haemorrhoids / piles	2283
Operation (self-report)	20004	1483	Haemorrhoidectomy / piles surgery/ banding of piles	9662
Total unique cases (excluding overlap)				**39950**

Of the 39,950 haemorrhoids cases identified in UK Biobank, 31,652 cases passed quality control (outlined in **3.2.4 Quality Control**), with the remaining 369,931 post-QC participants that did not possess a diagnostic or operative code indicative of haemorrhoids being defined as controls.

3.2.3. Genotyping

A complete description of the genotyping procedure for the UK Biobank cohort is provided in **Chapter 2 (Section 2.2.3.).** [8]

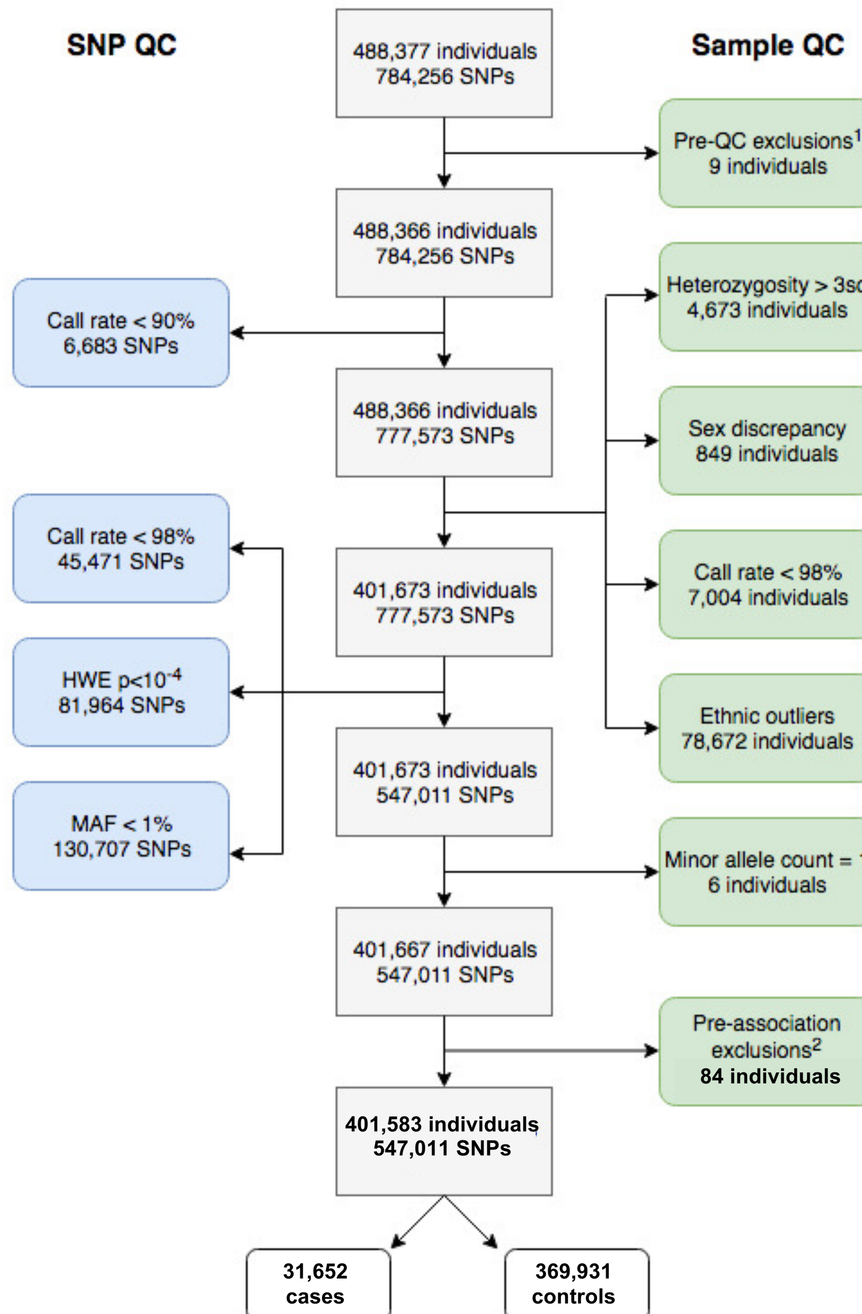
3.2.4 Quality control

A complete description of the quality control (QC) procedure implemented centrally by UK Biobank and locally by our group is provided in **Chapter 2 (Section 2.2.4.).**

After central (UK Biobank)[8], and local QC implementation, 86,794 participants from the GWAS analysis and 230,562 genotyped SNPs were excluded (**Figure 3.1**). In summary, 547,011 genotyped variants and 401,583 participants of white British

ancestry passed the QC and were included in the subsequent association analysis (See **3.2.6 Association Analysis**).

Figure 3.1. Overview of quality control pipeline. Excluded SNPs are presented in the blue panels and excluded samples are in the green panels. [1]Pre-QC exclusions: 3 participants with invalid IDs and sex, and 8 participants who had withdrawn from the study were excluded prior to QC. [2]Pre-association exclusions: 11 participants who were not present in UK Biobank's sample file accompanying the imputed genotype data were excluded prior to association, alongside 73 participants who subsequently withdrew consent for participation in UK Biobank.



3.2.5. Imputation

A complete description of the imputation methodology is provided in **Chapter 2 (Section 2.2.5)**.[8]

3.2.6. Association analysis

Genome-wide association testing was performed across a total of 8,944,561 SNPs (547,011 directly genotyped (MAF ≥ 0 .01) and 8,397,550 imputed SNPs (MAF ≥ 0.01, INFO Imputation score ≥ 0.90)) using a linear mixed non-infinitesimal model implemented in BOLT-LMM v2.323.[9,10]

Further methodological details pertaining to the association analysis performed for haemorrhoids in this chapter are provided in **Chapter 2 (Section 2.2.13)**.

3.2.7. Genomic risk loci definition

A complete description of the methods used for genomic risk loci definition is provided in **Chapter 2 (Section 2.2.7)**.

3.2.8. Functional annotation of SNPs

A complete description of the methods used for the functional annotation of variants is provided in **Chapter 2 (Section 2.2.8)**.

3.2.9. Candidate gene mapping

A complete description of the methods used for candidate gene mapping approach is provided in **Chapter 2 (Section 2.2.9)**.

3.2.10. Gene set, tissue and pathway analyses

This section follows the methodology described in **Chapter 2 (Section 2.2.10)**.

3.2.11. SNP-based heritability analysis

Methodological details on the SNP heritability analysis is described in **Chapter 2 (Section 2.2.11.)**.

3.2.12. Genetic correlation analysis

Full methodological details regarding the genetic correlation analysis are consistent with, and provided in, **Chapter 2 (Section 2.2.12.)**.

89 traits across thirteen trait categories from the LDHub[11] database were tested for correlation with haemorrhoids: anthropometric, autoimmune, cancer, education, haematological, hormone, lipids, personality, psychiatric, reproductive, sleeping, smoking behaviour. Traits categories were pre-defined based on associations in the literature. To account for multiple testing, a Bonferroni correction of $P < 3.8\times10^{-3}$ (0.05/13) was applied.

3.2.13. Drug-target enrichment analysis

The drug-target enrichment analysis performed for haemorrhoids in this chapter followed the methods described in **Chapter 2 (Section 2.2.13.).**[12]

3.2.14. Genetic risk score

The weighted genetic risk score (wGRS) methodology implemented in this chapter mirrors those described in **Chapter 2** (**Section 2.2.14.)**. The wGRS was compared between four groups of participants from the GWAS: i) all cases vs all controls; ii) surgical cases vs non-surgical cases.

3.2.15. URLs

ANNOVAR, www.annovar.openbioinformatics.org/en/latest/; BOLT-LMM, www.data.broadinstitute.org/alkesgroup/BOLT-LMM/; CADD, cadd.gs.washington.edu/; ENSEMBL, www.ensembl.org/index.html; flashpca, github.com/gabraham/flashpca/; FUMA, www.fuma.ctglab.nl/; GERP, http://mendel.stanford.edu/SidowLab/downloads/gerp/; GTEx Portal, www.gtexportal.org/home/; GWAMA, www.genomics.ut.ee/en/tools/gwama; Human Genome Variation Society (HGVS), www.varnomen.hgvs.org/; HRC, www.haplotype-reference-consortium.org/ ; LD Hub, www.ldsc.broadinstitute.org/ldhub/; LD Link, www.ldlink.nci.nih.gov/; MAGMA, www.ctg.cncr.nl/software/magma; Open Targets Platform, www.targetvalidation.org/; PLINK, www.pngu.mgh.harvard.edu/~purcell/plink/; Polyphen-2, www.genetics.bwh.harvard.edu/pph2; QCTOOL,www.well.ox.ac.uk/~gav/qctool_v2/#overview; R, www.r-project.org;

RegulomeDB, www.regulomedb.org/; SHAPEIT3, jmarchini.org/shapeit3/; SIFT, www.sift.bii.a-star.edu.sg/; UK Biobank, www.ukbiobank.ac.uk/; XGR, www.galahad.well.ox.ac.uk:3040; 1000 Genomes Project, www.1000genomes.org;

3.3. Results

3.3.1. Twelve novel haemorrhoids associated loci

A single-stage genome-wide association analysis was performed in UK Biobank consisting of 31,652 cases and 369,931 controls of white British ancestry. The analytic workflow for the GWAS is provided in **Figure 3.2**. Association testing for haemorrhoids was conducted across 547,011 directly genotyped SNPs (MAF ≥ 0 .01) and 8,397,550 imputed SNPs (MAF ≥ 0.01, INFO Imputation score ≥ 0.90).[10] The analysis yielded genome-wide level associations ($P < 5\times10^{-8}$) at 12 risk loci (882 variants). Conditional regression analysis demonstrated an additional independent residual signal at locus 7q11.23 (rs77689666, $P_{Cond} = 2.90\times10^{-9}$, OR = 1.10). Thus, in summary, 13 independent signals at 12 novel risk loci associated with haemorrhoids (**Table 3.2**). The λ_{GC} demonstrated moderate inflation (1.15), however the LDSC intercept of 1.01 and an attenuation ratio of 0.07 suggests that this is due to the large sample size of the cohort and polygenicity, rather than population stratification (**Figure 3.3**).[13]

The most significant association signal from the GWAS came from locus 9q34.2 (index SNP rs687621, $P = 3.3\times10^{-26}$, OR = 1.10) which was mapped in the genome-wide, gene-based MAGMA test to *CACFD1*. *CACFD1* encodes the flower membrane protein, human flower (hFWE), the negative expression of which has been found to reduce tumour growth and metastasis, and impart sensitisation to chemotherapy.[14] Of the 13 index variants, the variant described above, rs687621, was genotyped with the remaining twelve imputed variants having robust imputation scores of between 0.913 and 0.998. All index variants were common, with minor allele frequencies in Europeans ranging from 6% to 49%. Odds ratios for the effect alleles range from 1.05

to 1.13 which is in keeping with the effect sizes typically seen in other GWAS (median OR ~1.33) (A Manhattan plot is provided in **Figure 3. 4**). Regional plots for all 13 associated loci are provided in **Figure 3.5**.

Figure 3.2. Haemorrhoids GWA study design and analysis workflow. A single-stage GWAS of haemorrhoids was conducted in UK Biobank, identifying 13 independent variants, which were interrogated further in subsequent analyses.

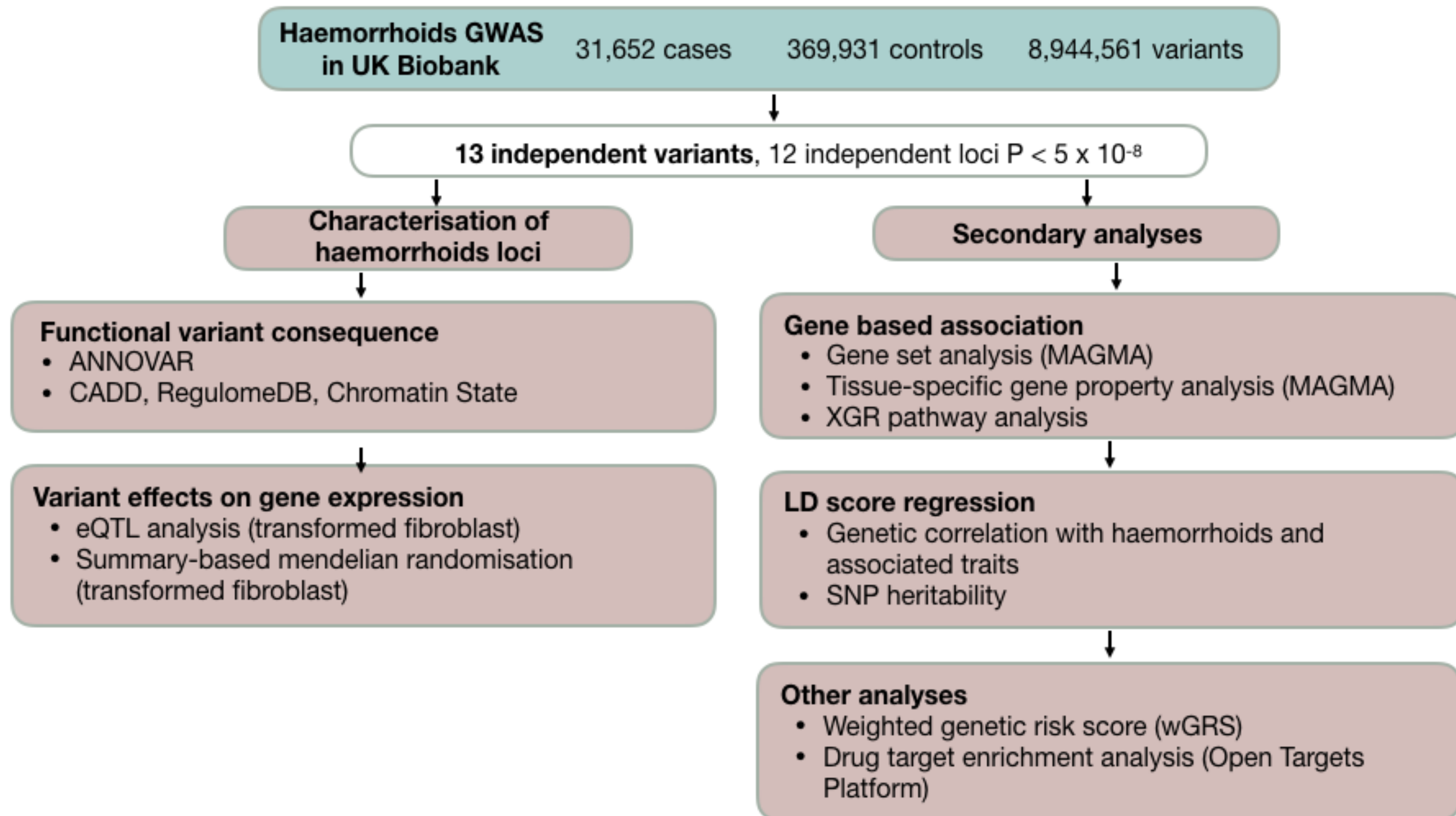


Table 3.2. Thirteen novel signals at 12 loci significantly associated with haemorrhoids in 31,652 cases and 369,931 controls in UK Biobank

Chromosome	Position[a]	rsID	EA[b]	NEA[c]	EAF[d]	Info[e]	OR (95% CI)	P-value	Mapped genes[f]
1	204559707	rs10900600	T	G	0.69	0.995	1.06 (1.04-1.07)	1.50×10^{-9}	*LRRN2, MDM4±, PIK3C2B*
2	67881931	rs114233333	C	G	0.06	0.995	1.11 (1.07-1.14)	7.90×10^{-9}	-
5	37422640	rs112047495	T	C	0.64	0.997	1.06 (1.04-1.08)	6.10×10^{-11}	*C5orf42±, NIPBL±, NUP155±, WDR70±*
5	50759375	rs17824230	G	C	0.87	0.989	1.07 (1.05-1.10)	2.10×10^{-8}	-
7	73306093	rs75606842	A	G	0.22	0.977	1.07 (1.05-1.09)	1.00×10^{-10}	-
7[#]	73434287	rs77689666	G	A	0.06	0.990	1.10 (1.07-1.14)	2.90×10^{-9}	*ELN±*
7	100632790	rs4556017	C	T	0.15	0.984	1.08 (1.06-1.11)	9.10×10^{-12}	*ACHE, MUC3A, MUC12*
8	71651344	rs4612371	G	C	0.46	0.913	1.05 (1.03-1.07)	4.00×10^{-8}	*LACTB2, XKR9±*
8	105879946	rs12375337	T	A	0.31	0.996	1.05 (1.04-1.07)	4.20×10^{-9}	*RP11-127H5.1±*
9	22124504	rs1333047	A	T	0.51	0.998	1.06 (1.04-1.08)	4.60×10^{-12}	-
9	136137065	rs687621	A	G	0.68	G	1.10 (1.08-1.12)	3.30×10^{-26}	*CACFD1*
12	66409367	rs11176001	C	A	0.87	0.990	1.13 (1.10-1.15)	4.70×10^{-22}	-
15	67441750	rs72743461	A	C	0.24	0.998	1.06 (1.04-1.08)	1.60×10^{-8}	*RP11-342M21.2, SMAD3±*

[a]Based on NCBI Genome Build 37 (hg19).
[b]The effect allele.
[c]The non-effect allele.
[d]The effect allele frequency.
[e]The SNP INFO score for imputed SNPs; G = genotyped SNP.
[f]The 17 genes (in alphabetical order) prioritised at these loci based on positional mapping, eQTL mapping and MAGMA gene mapping (see Methods).
[±]Where overlap in gene mapping strategies occurred, the gene(s) with the highest level of overlap are depicted.
[#]Denotes a residual significant signal following conditional regression analysis at the lead SNP at the locus.

Figure 3.3. Quantile-quantile (Q-Q) plot of associated variants.

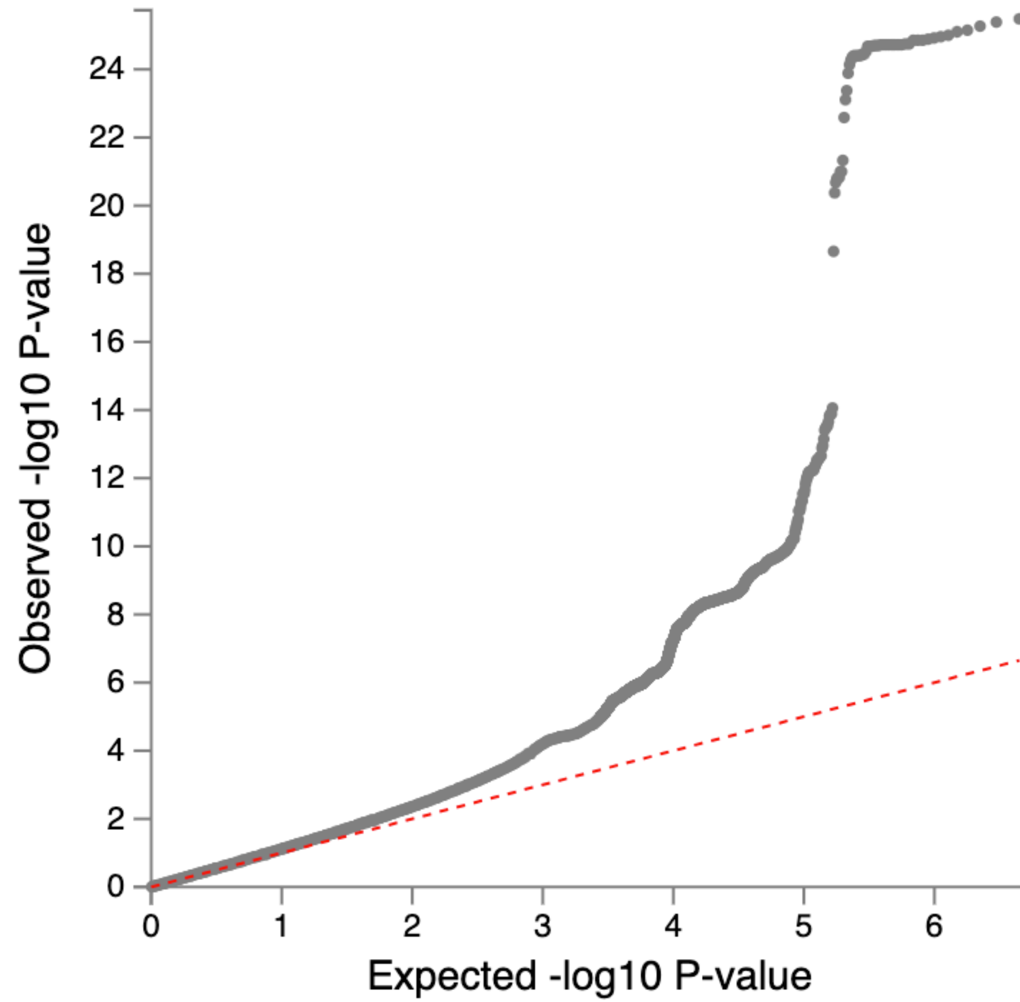


Figure 3.4. Results of the genome-wide association study in haemorrhoids. A Manhattan plot showing genomic position on each of the autosomes plotted against association signal strength ($-\log_{10}$ P-value). The red line refers to the genome-wide significance threshold of $P < 5\times10^{-8}$. Of the prioritised genes (described elsewhere in this chapter), the most promising gene candidate is highlighted in the Manhattan plot based on literature. Bold genes are those that were prioritised using the four mapping strategies; *CDKN2B-AS1* was prioritised based on a detailed literature search of proximal genes at this locus.

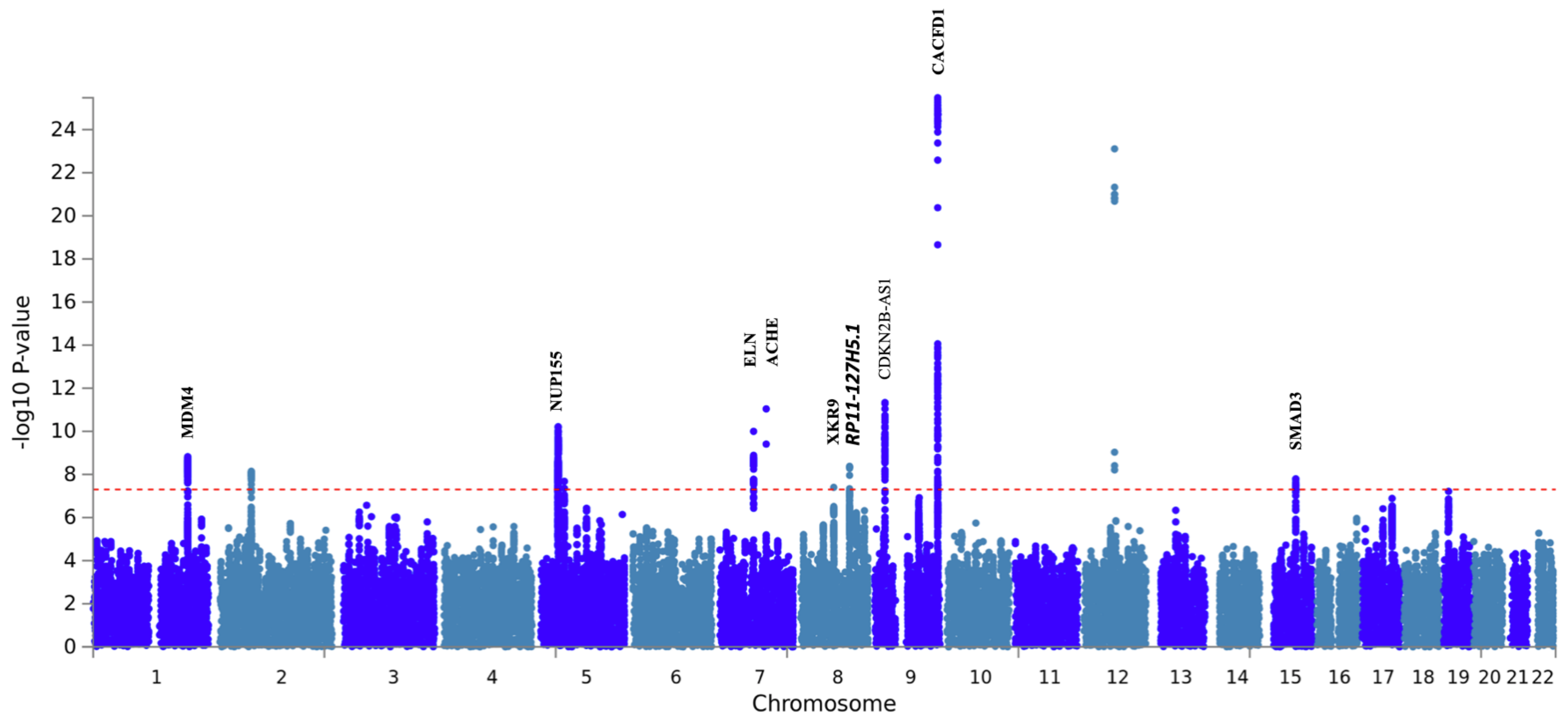
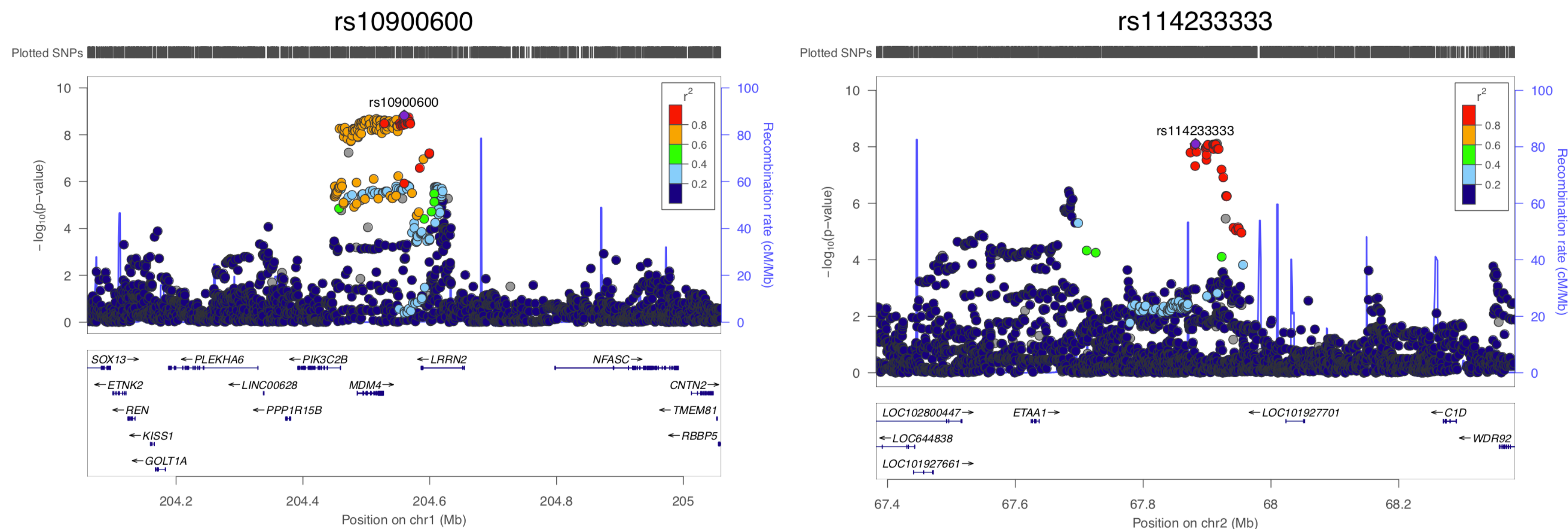
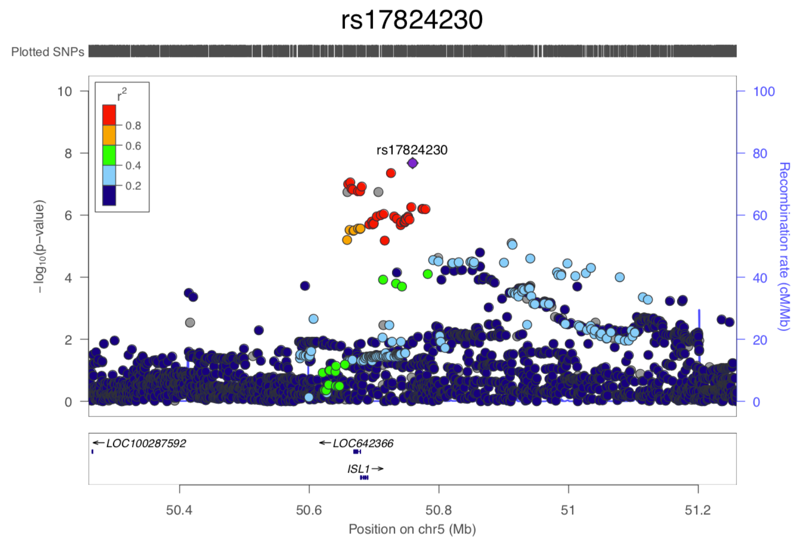
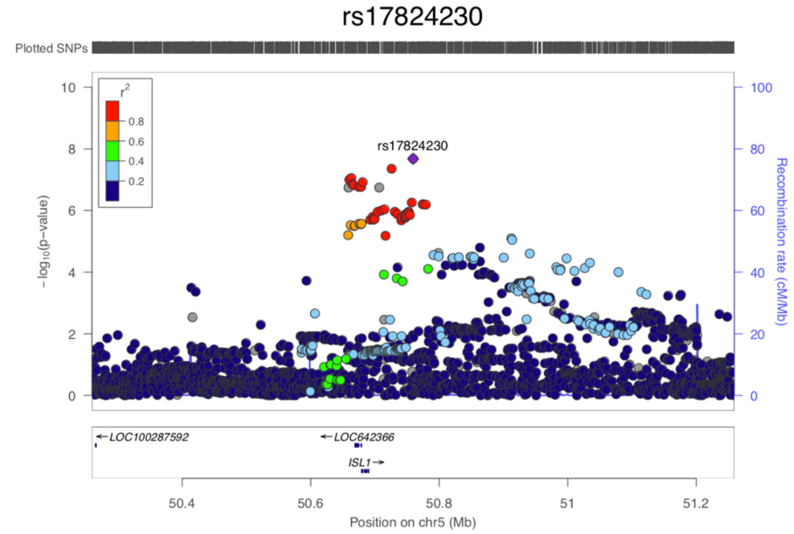


Figure 3.5. Regional Locus Zoom plots of all haemorrhoids associated signals. LocusZoom plots of the 13 independent genome-wide significant variants at the 12 haemorrhoids associated susceptibility loci. Plots are ordered by chromosome number and genomic position. SNP position is shown on the x-axis, and strength of association on the y-axis (-$\log_{10}$ P-value). The linkage disequilibrium (LD) relationship between the lead SNP and the surrounding SNPs is indicated by the r^2 legend. In the lower panel of each sub-figure, genes within 500kb on either side of the index SNP are shown. The position on each chromosome is depicted in relation to Human Genome build hg19 (GRCh37).

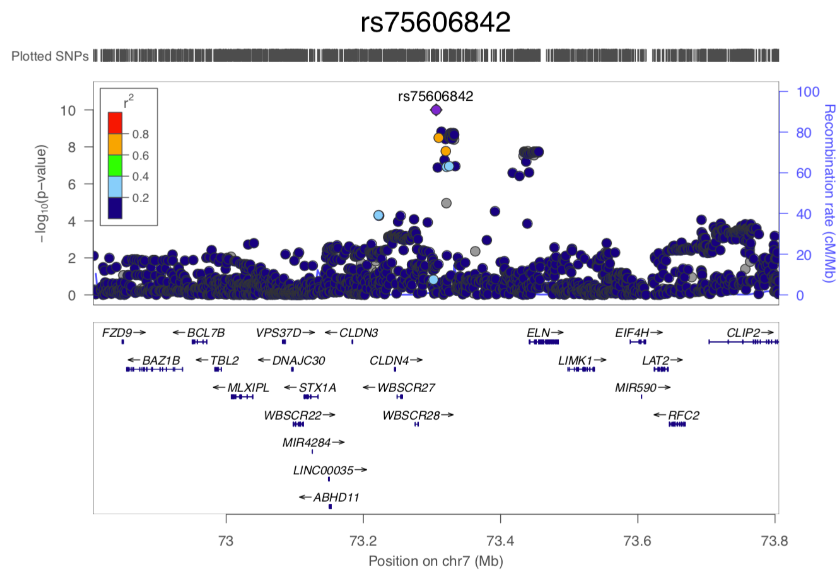




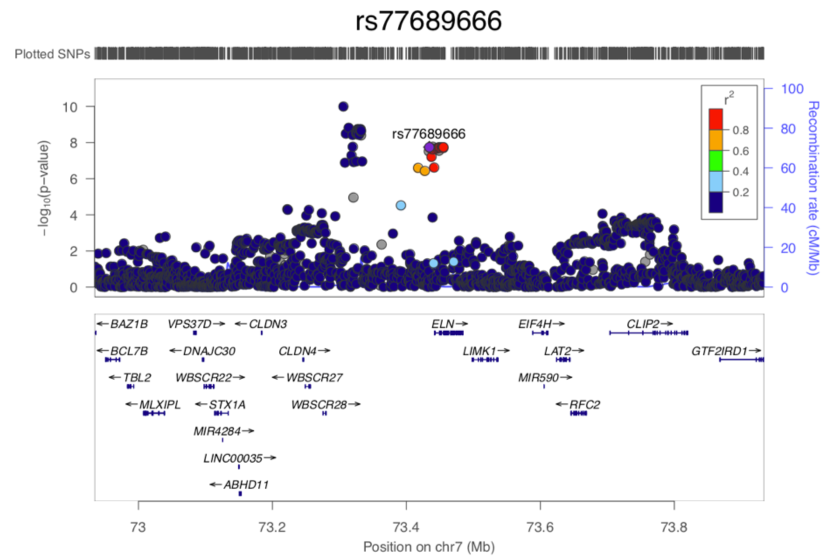
rs17824230
Plotted SNPs
rs17824230
r2
0.8
0.6
0.4
0.2
-log10(p-value)
Recombination rate (cM/Mb)
LOC100287592
LOC642366
ISL1
Position on chr5 (Mb)



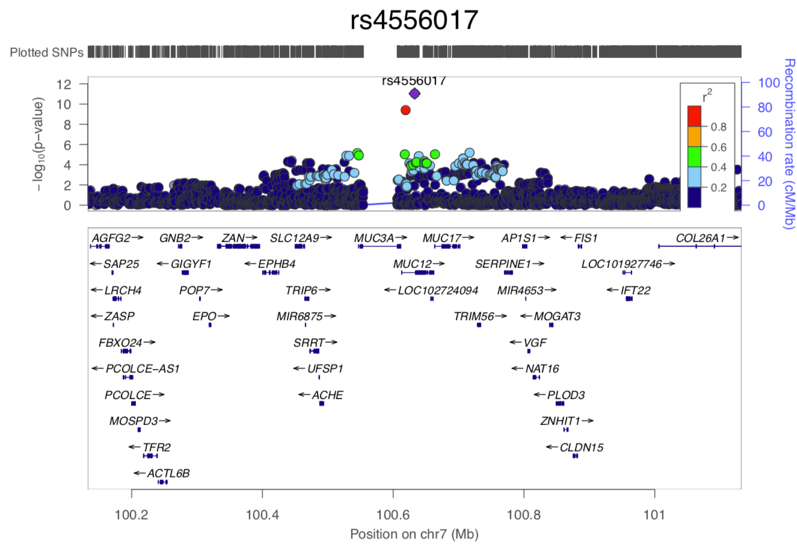
rs17824230
Plotted SNPs
rs17824230
r2
0.8
0.6
0.4
0.2
-log10(p-value)
Recombination rate (cM/Mb)
LOC100287592
LOC642366
ISL1
Position on chr5 (Mb)

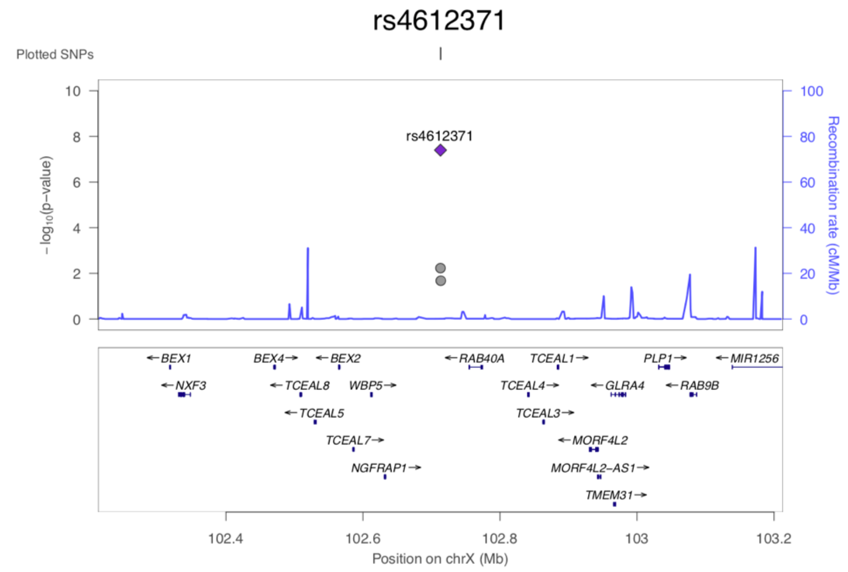


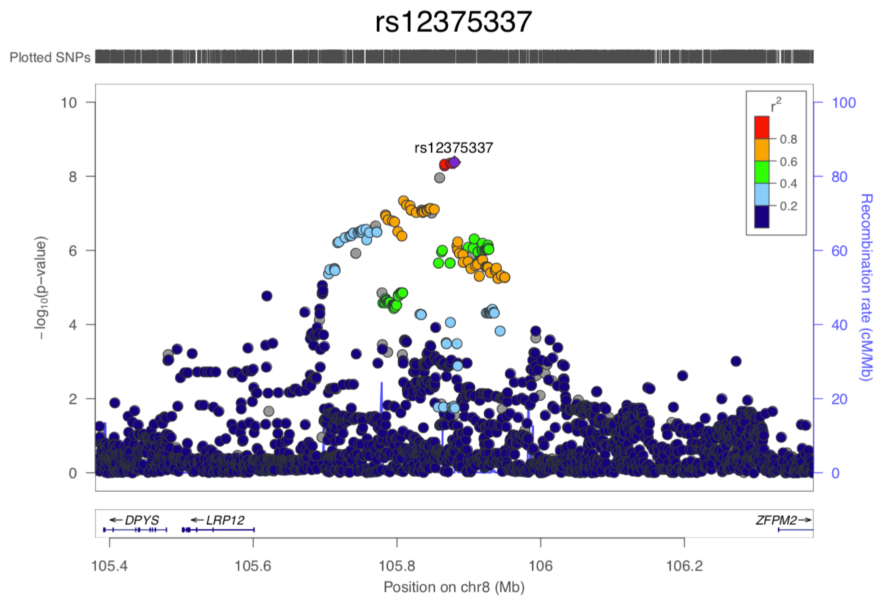
rs75606842
Plotted SNPs
rs75606842
r2
0.8
0.6
0.4
0.2
-log10(p-value)
Recombination rate (cM/Mb)
FZD9
BCL7B
VPS37D
CLDN3
ELN
EIF4H
CLIP2
BAZ1B
TBL2
DNAJC30
CLDN4
LIMK1
LAT2
MLXIPL
STX1A
WBSCR27
MIR590
WBSCR22
WBSCR28
RFC2
MIR4284
LINC00035
ABHD11
Position on chr7 (Mb)

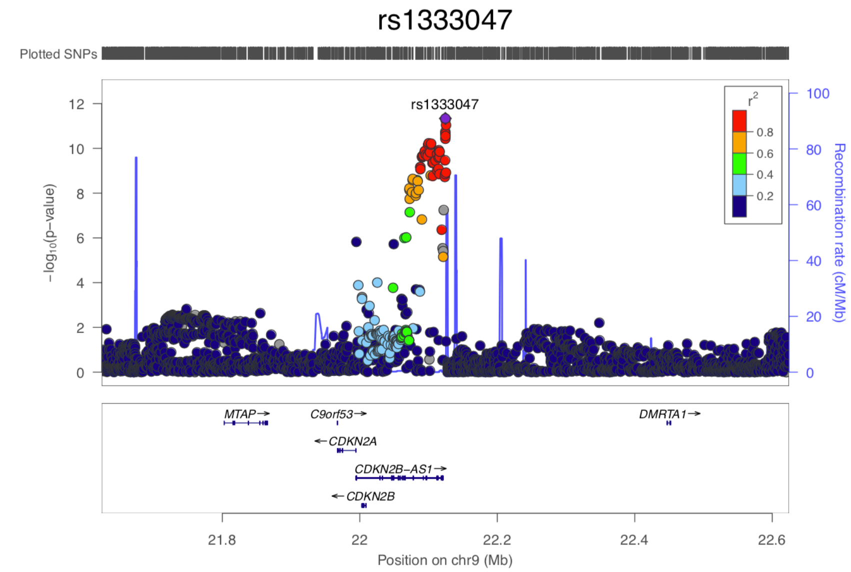


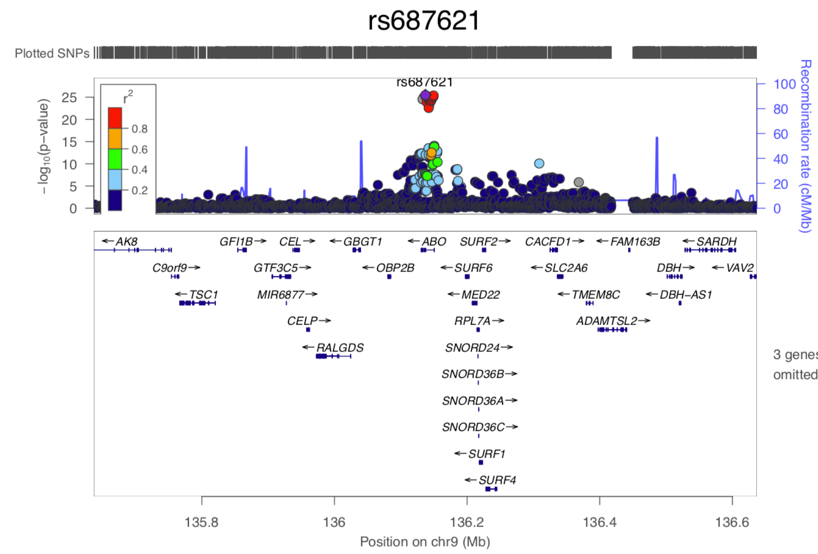
rs77689666
Plotted SNPs
rs77689666
r2
0.8
0.6
0.4
0.2
-log10(p-value)
Recombination rate (cM/Mb)
BAZ1B
VPS37D
CLDN3
ELN
EIF4H
CLIP2
BCL7B
DNAJC30
CLDN4
LIMK1
LAT2
GTF2IRD1
TBL2
WBSCR22
WBSCR27
MIR590
MLXIPL
STX1A
WBSCR28
RFC2
MIR4284
LINC00035
ABHD11
Position on chr7 (Mb)

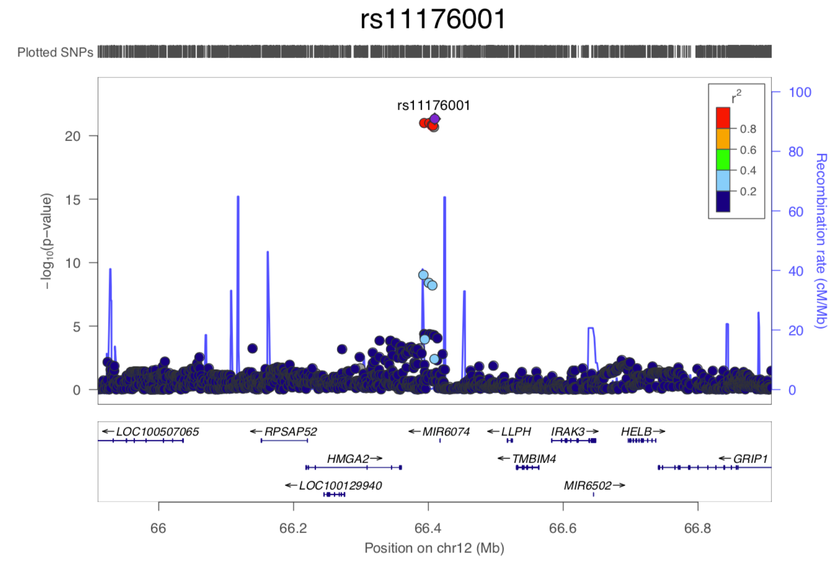

rs4556017
Plotted SNPs
rs4556017
−log10(p−value)
Recombination rate (cM/Mb)
r2
AGFG2 GNB2 ZAN SLC12A9 MUC3A MUC17 AP1S1 FIS1 COL26A1
SAP25 GIGYF1 EPHB4 MUC12 SERPINE1 LOC101927746
LRCH4 POP7 TRIP6 LOC102724094 MIR4653 IFT22
ZASP EPO MIR6875 TRIM56 MOGAT3
FBXO24 SRRT VGF
PCOLCE−AS1 UFSP1 NAT16
PCOLCE ACHE PLOD3
MOSPD3 ZNHIT1
TFR2 CLDN15
ACTL6B
Position on chr7 (Mb)

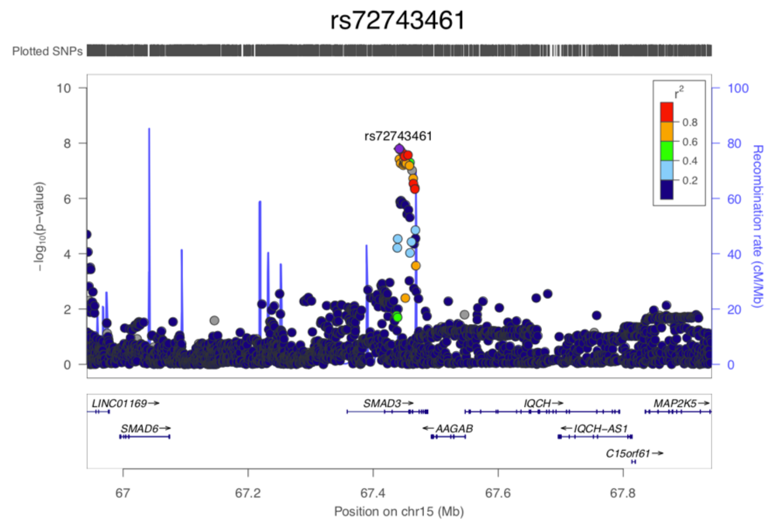

rs4612371
Plotted SNPs
rs4612371
−log10(p−value)
Recombination rate (cM/Mb)
BEX1 BEX4 BEX2 RAB40A TCEAL1 PLP1 MIR1256
NXF3 TCEAL8 WBP5 TCEAL4 GLRA4 RAB9B
TCEAL5 TCEAL3
TCEAL7 MORF4L2
NGFRAP1 MORF4L2−AS1
TMEM31
Position on chrX (Mb)


rs12375337
Plotted SNPs
rs12375337
−log10(p−value)
Recombination rate (cM/Mb)
r2
DPYS LRP12 ZFPM2
Position on chr8 (Mb)


rs1333047
Plotted SNPs
rs1333047
−log10(p−value)
Recombination rate (cM/Mb)
r2
MTAP C9orf53 DMRTA1
CDKN2A
CDKN2B−AS1
CDKN2B
Position on chr9 (Mb)


rs687621
Plotted SNPs
rs687621
−log10(p−value)
Recombination rate (cM/Mb)
r2
AK8
C9orf9
TSC1
GFI1B
GTF3C5
MIR6877
CELP
RALGDS
CEL
GBGT1
OBP2B
ABO
SURF6
MED22
RPL7A
SNORD24
SNORD36B
SNORD36A
SNORD36C
SURF1
SURF4
SURF2
CACFD1
SLC2A6
TMEM8C
ADAMTSL2
FAM163B
DBH
DBH−AS1
SARDH
VAV2
3 genes omitted
Position on chr9 (Mb)

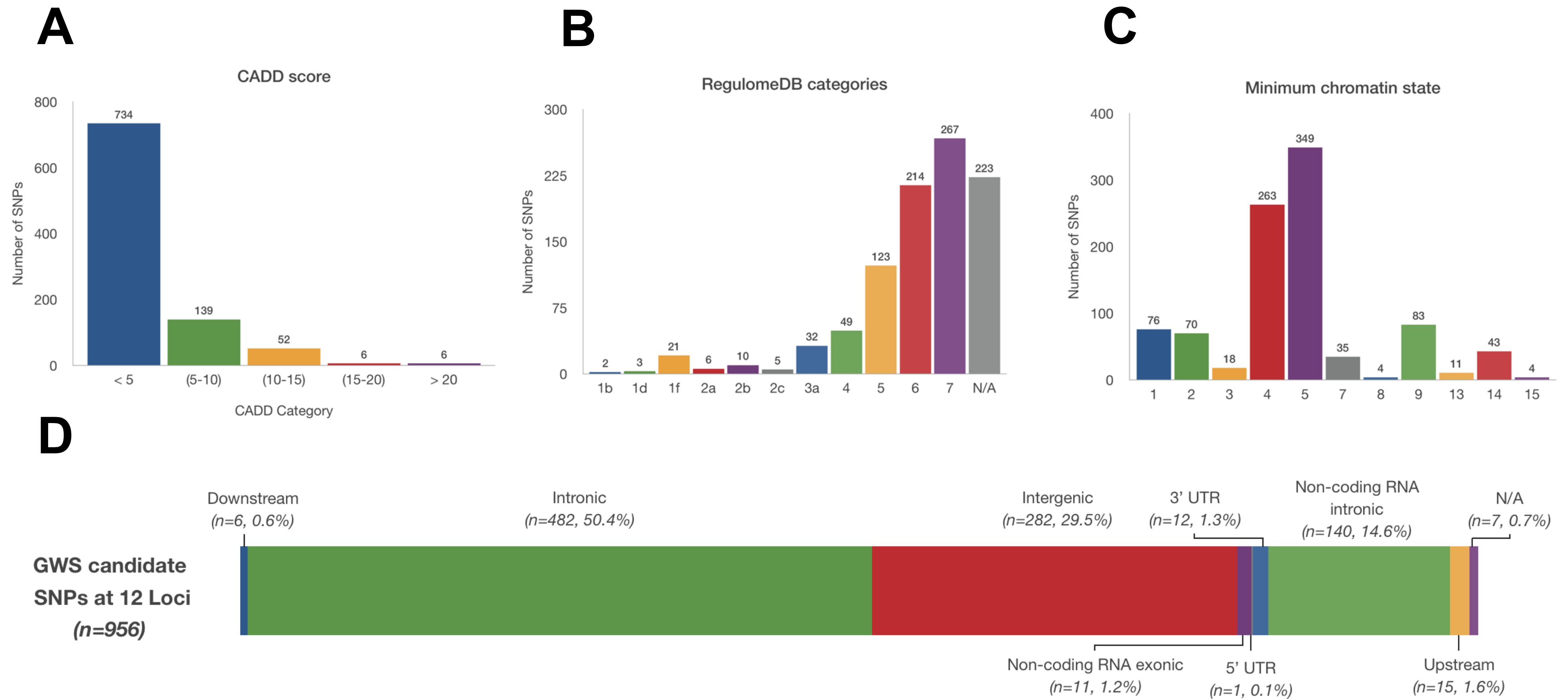

rs11176001
Plotted SNPs
rs11176001
−log10(p−value)
Recombination rate (cM/Mb)
r2
LOC100507065
RPSAP52
HMGA2
LOC100129940
MIR6074
LLPH
TMBIM4
IRAK3
MIR6502
HELB
GRIP1
Position on chr12 (Mb)


rs72743461
Plotted SNPs
rs72743461
−log10(p−value)
Recombination rate (cM/Mb)
r2
LINC01169
SMAD6
SMAD3
AAGAB
IQCH
IQCH−AS1
C15orf61
MAP2K5
Position on chr15 (Mb)

3.3.2. In silico annotation

Localising associated SNPs using ANNOVAR[15] implemented in FUMA GWAS (Functional Mapping and Annotation of GWAS) v1.3.3[16] yielded 609 genome-wide significant ($P < 5\times10^{-8}$) candidate SNPs (956 in total) across the 12 risk loci (**Figure 3.6**).[15,16] All variants were non-coding, with the majority lying in intronic and intergenic regions (76.5%, n = 466), and a proportion annotated as intronic or exonic non-coding RNAs (20.9%, n = 127) (**Figure 3.6**). Of the 466 intronic-intergenic variants, 98.9% (n = 461) resided in open chromatin regions depicted by a minimum 15-core chromatin state of 1 to 7 in 127 tissue/cell types from the Roadmap Epigenomics Consortium ChroHMM model.[17] 33 intronic-intergenic candidate variants had a CADD score[18] ≥ 12.37 suggesting they may be deleterious (**Appendix Table 3.1.**), of which 16 were genome-wide significant – four of these demonstrated potential regulatory activity depicted by a RegulomeDB[19] score of 2b or less (*likely to affect binding):* rs7532236 (*RP11-430C7.4*), rs11750212 (*WDR70*), rs1866316 (*SMAD3*), rs17293632 (*SMAD3*).

Figure 3.6. Functional annotation of the 956 candidate SNPs at the 12 haemorrhoids associated risk loci. Functional consequences of the SNPs on genes were obtained by performing ANNOVAR gene-based annotation using Ensembl genes (build 85) in FUMA. A) CADD scores, B) RegulomeDB scores and C) 15-core chromatin state were annotated to all 956 SNPs in 1000G phase 3 by FUMA through matching chromosome, position, reference, and alternative alleles. D) Positional classification of the 956 SNPs.



3.3.3. Gene mapping

Candidate SNPs in FUMA v1.3.3[16] were mapped positionally to 13 protein-coding genes based on genomic proximity. Three genes were prioritised based on containing variants that are known to affect expression of these genes (eQTLs) within fibroblast tissue from the GTEx consortium[20] and GENCORD collection[21] or skeletal muscle tissues from the GTEx consortium[20] ($P_{eqtl} < 5\times10^{-8}$) — two of these genes (*ACHE*, *LACTB2*) were *not* prioritised in the positional or subsequent MAGMA mapping approaches (i.e. lay outside the confines of the 10kb positional window from the lead SNP and were not picked up in the genome-wide gene association test). Performing a genome-wide, gene-based association test in MAGMA v1.07[22], 18 genes reached the genome-wide significance threshold in MAGMA ($P < 2.64\times10^{-6}$), 11 of which resided within the realms of our susceptibility loci (**Appendix Table 3.2.**), with two genes (*LRRN2* and *CACFD1*) being prioritised only at this genome-wide gene mapping level (**Figures 3.7 and 3.8**). The fourth strategy, summary-based mendelian randomisation (SMR)[23] identified no genes that met the SMR-significance threshold ($P < 0.05/4323 = 2.68\times10^{-6}$) and associated with haemorrhoids through pleiotropy.

In summary, 17 unique genes were mapped to eight of the 12 haemorrhoids susceptibility loci by at least one mapping approach. Nine genes were mapped by two or more gene-mapping strategies, with one gene (*XKR9*), being mapped by three gene-mapping approaches (**Table 3.3 and Figure 3.9**).

Figure 3.7. MAGMA gene-based association analysis Quantile-Quantile plot. Quantile-Quantile (Q-Q) plot for the genome-wide, gene-based association test computed by MAGMA v1.07.[22]

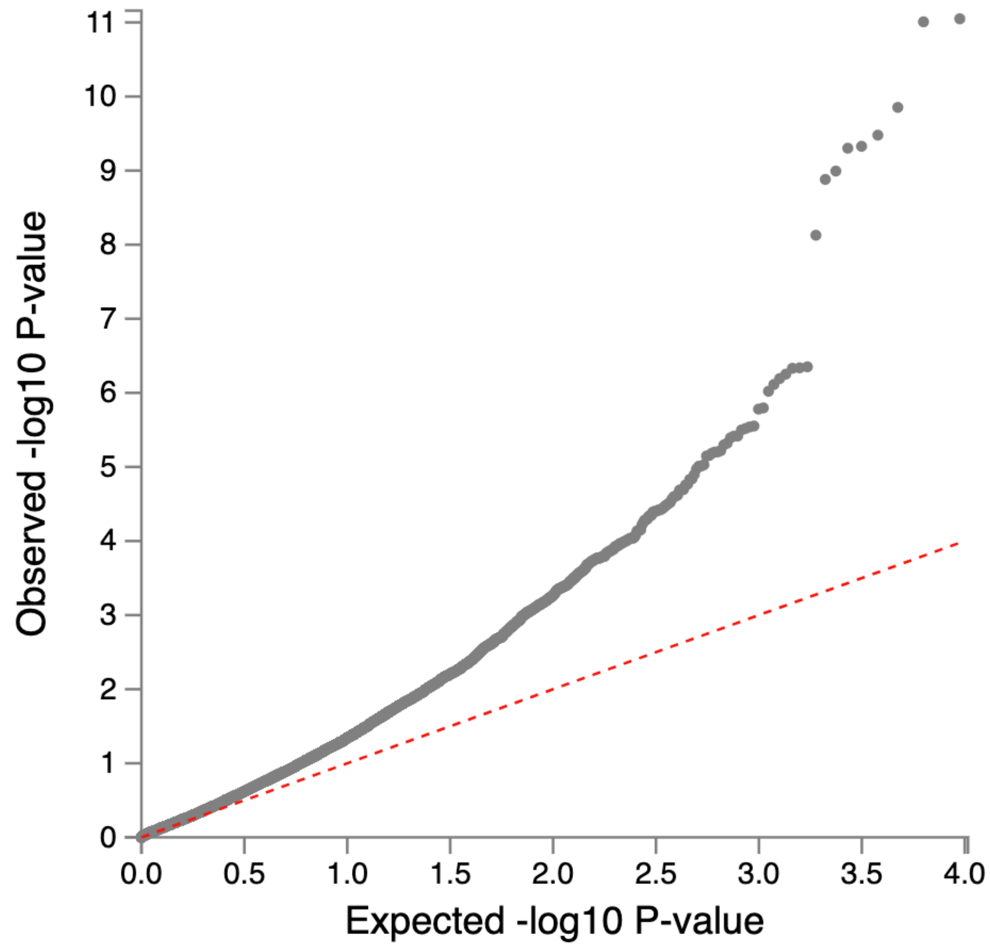


Figure 3.8. MAGMA gene-based association analysis Manhattan plot. Manhattan plot for the genome-wide, gene-based association test computed by MAGMA v1.07.[22] Input SNPs were mapped to 18918 protein coding genes. Genome wide significance (red dashed line in the plot) was defined at $P = 2.64\times10^{-6}$ (0.05/18918). The 18 significant MAGMA genome-wide associated genes are depicted above the red line; 11 of the 18 MAGMA mapped genes resided within the realms of the 12 genomic risk loci.

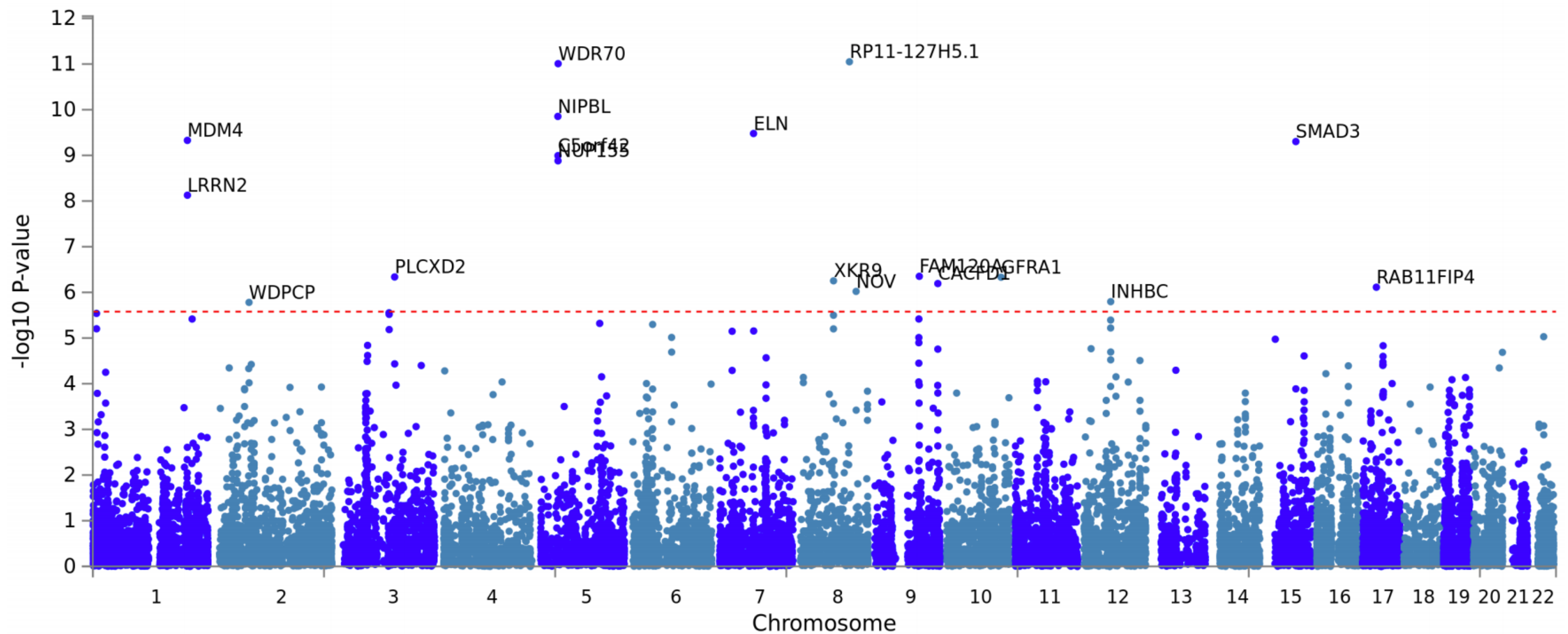
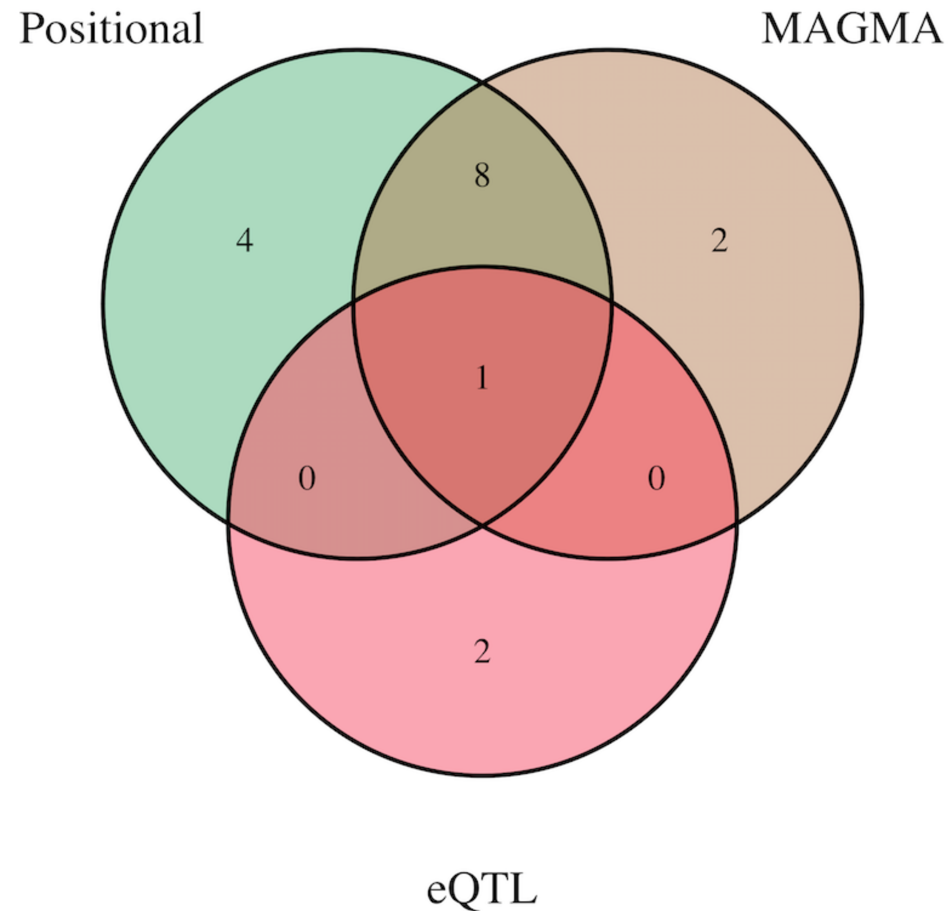


Table 3.3. Genes mapped to the haemorrhoids associated loci using the four mapping strategies. 17 unique genes were mapped to eight of the 12 associated loci by one or more gene mapping strategies (see Methods). 13 genes were mapped via positional mapping in FUMA, 3 genes were mapped via eQTL mapping in FUMA, 11 genes were mapped using MAGMA. No genes were mapped using SMR. Overlap between the four different mapping strategies is shown.

Chromosome	Lead SNP	Position	13 FUMA Positionally Mapped Genes	3 FUMA eQTL Mapped Genes	11 MAGMA Mapped Genes	Number of Gene Mapping Approaches
1	rs10900600	204559707			LRRN2	1
1	rs10900600	204559707	*MDM4*		*MDM4*	2
1	rs10900600	204559707	*PIK3C2B*			1
5	rs112047495	37422640	*C5ORF42*		*C5ORF42*	2
5	rs112047495	37422640	*NIPBL*		*NIPBL*	2
5	rs112047495	37422640	*NUP155*		*NUP155*	2
5	rs112047495	37422640	*WDR70*		*WDR70*	2
7	rs77689666	73434287	*ELN*		*ELN*	2
7	rs4556017	100632790		*ACHE*		1
7	rs4556017	100632790	*MUC3A*			1
7	rs4556017	100632790	*MUC12*			1
8	rs4612371	71651344		*LACTB2*		1
8	rs4612371	71651344	*XKR9*	*XKR9*	*XKR9*	3
8	rs12375337	105879946	*RP11-127H5.1*		*RP11-127H5.1*	2
9	rs687621	136137065			*CACFD1*	1
15	rs72743461	67441750	*RP11-342M21.2*			1
15	rs72743461	67441750	*SMAD3*		*SMAD3*	2

Figure 3.9. Venn diagram for the 17 mapped genes prioritised at the haemorrhoids associated loci. 17 unique genes were mapped to eight of the 12 associated loci by one or more gene mapping strategies (see Methods). 13 genes were mapped via positional mapping in FUMA, 3 genes were mapped via eQTL mapping in FUMA, 11 genes were mapped using MAGMA. No genes were mapped using SMR. Overlap between the three different mapping strategies is shown in the Venn diagram.



3.3.4. Gene set, pathway and tissue-specific enrichment

Gene-set analysis in MAGMA v1.07[22] determined the convergence of MAGMA prioritised genes within 15,496 gene sets (5500 curated gene sets and 9995 GO terms) from MSigDB v8.0. [24] One curated gene set was significantly enriched: '*Genes up-regulated and displaying increased copy number in glioblastoma samples (TCGA_Glioblastoma_Copy_Number_Up (M5536))'* (P = 7.15×10^{-8}, n = 72 genes) (**Appendix Table 3.3.**). Moreover, tissue expression analysis in MAGMA[22] using GTEx v8.0 30 general tissue types[20] demonstrated blood vessel to be the most enriched tissue (P = 3.07×10^{-6}), with all three vascular tissue types being significantly enriched in the separate GTEx v8.0 54 tissue types expression analysis[20]: Tibial Artery (P = 5.20×10^{-6}, most enriched tissue), Aorta (P = 7.83×10^{-5}, 3rd most enriched tissue), Coronary Artery (P = 2.64×10^{-4}, 6th most enriched tissue) (**Figure 3.10**). Moreover, of all remaining significantly enriched non-vascular tissues across both general and specific GTEx tissues; all tissues reside in hollow organs where smooth muscle plays an important role in function (namely uterus, oesophagus, oesophago-gastric junction, and cervix).

Performing gene set enrichment analysis in FUMA *GENE2FUNC*[16] demonstrated the 17 prioritised genes to cluster in several GWAS Catalog reported genes. Striking enrichment was noted of 11 of 18 prioritised genes in gene sets for six phenotypes reported in the GWAS Catalog[25], most significantly diverticular disease (P = 3.20×10^{-4}, 2nd most significant, *SMAD3*, *WDR70*, *ELN*, *LACTB2*) and fracture non-union (P = 3.20×10^{-4}; *ACHE*, *MUC3A*, *MUC12*). Pathway analysis in eXploring Genomic Relations (XGR)[26] furthermore yielded five enriched canonical pathways including

those related to regulation of cytoplasmic and nuclear SMAD2/3 signalling ($P = 4.8\times10^{-4}$, $Z = 5.21$, $FDR = 7.1\times10^{-3}$) extracellular matrix biology. ($P = 3.6\times10^{-3}$, $Z = 2.88$, $FDR = 1.3\times10^{-2}$) (**Table 3.4)**. Moreover, gene expression heatmap of the 17 prioritised genes in both GTEx v8.0[20] 30 general tissue types and 54 specific tissue types demonstrated *ELN* and *SMAD3* genes to have the highest expression in blood vessel tissue (**Figure 3.11**).

Figure 3.10. MAGMA tissue expression analysis. Tissue Expression Analysis of haemorrhoids GWAS data computed by MAGMA v1.07. A) 54 specific and B) 30 general tissue types. This analysis examines the relationship between genes containing significant genetic associations from the MAGMA association test and their expression levels across various tissues from the GTEx consortium. Gene-property analysis is performed using average expression of genes per tissue type as a gene covariate. Gene expression values are log2 transformed average RPKM (Read Per Kilobase Per Million) per tissue type after winsorization at 50, and are based on GTEx v8 RNA-Seq data across 54 specific tissue types and 30 general tissue types. The dotted line indicates the Bonferroni-corrected α level, and the tissues that meet this significance threshold are highlighted in red.

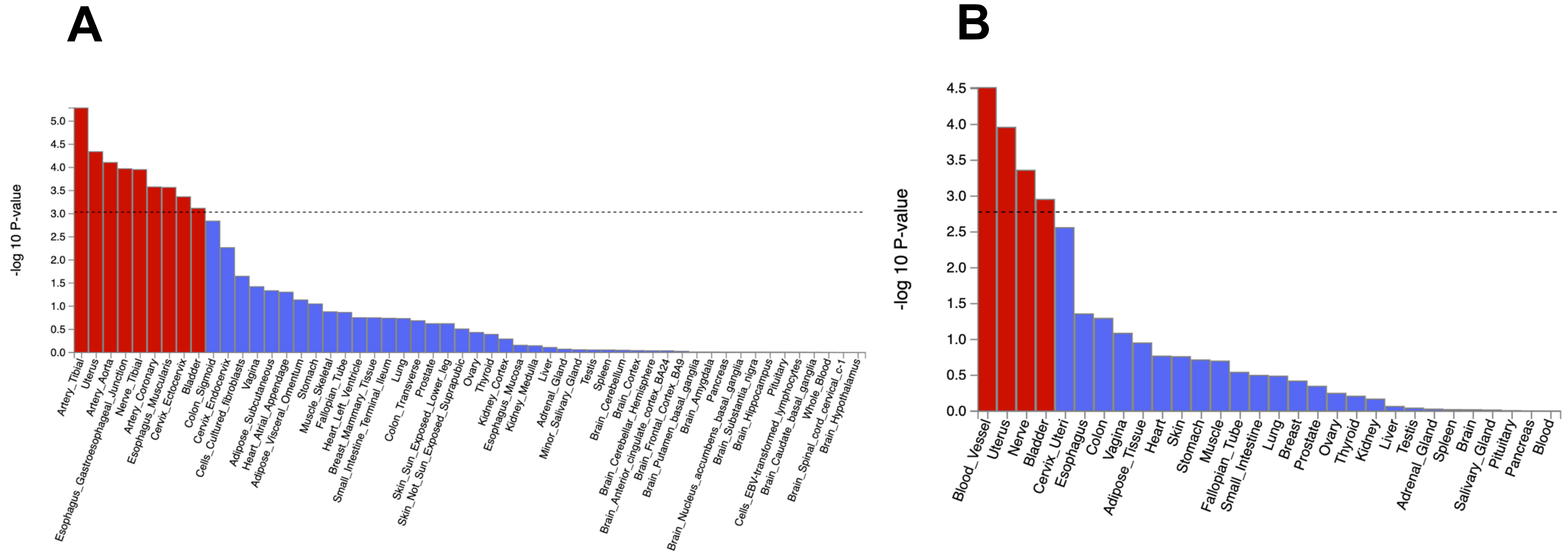
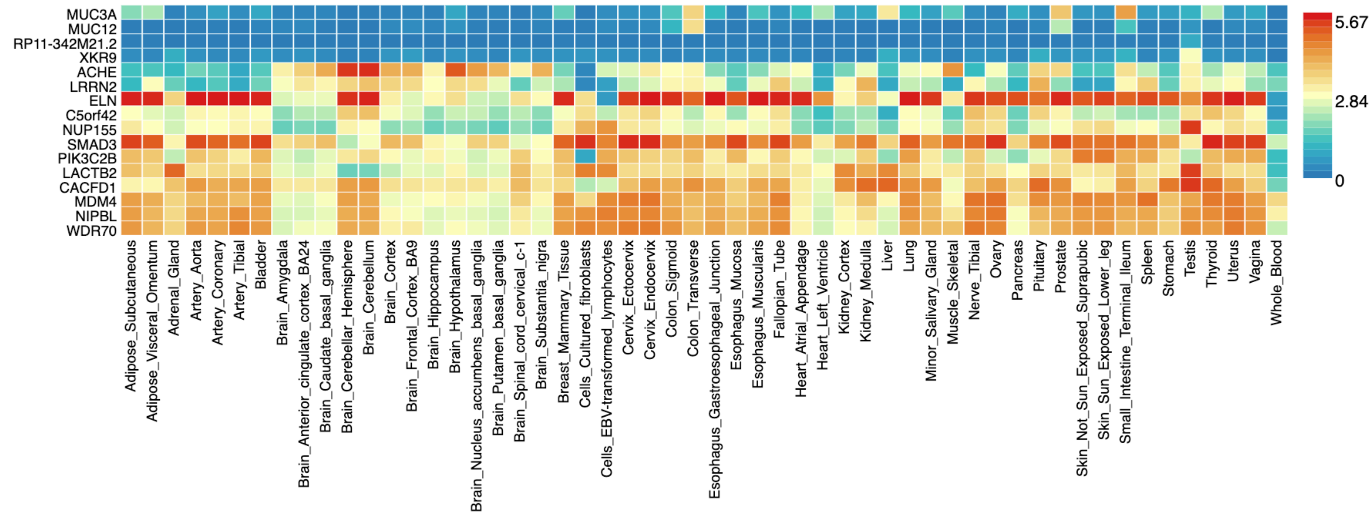


Table 3.4. Gene-based enrichment analysis in XGR. Pathway analysis of the 17 prioritised haemorrhoids associated genes was performed in eXploring Genomic Relations (XGR) for canonical pathways using a hypergeometric distribution test.

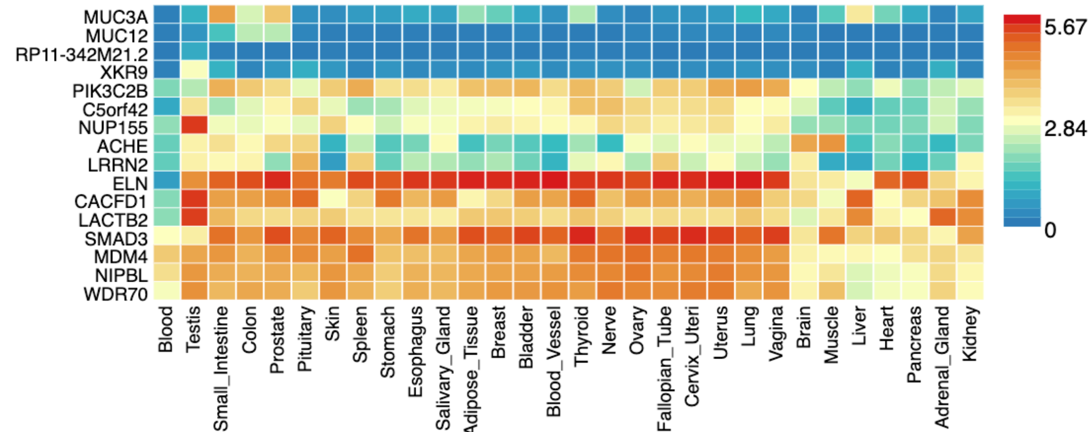
Biological Process	**Z-Score**	**P-Value**	**FDR**	**Number of overlapped genes**	**Genes**
Regulation of cytoplasmic and nuclear SMAD2/3 signalling	5.21	4.8×10^{-4}	7.1×10^{-2}	1	*SMAD3*
Genes encoding proteins affiliated structurally or functionally to extracellular matrix proteins	2.88	3.6×10^{-3}	1.3×10^{-2}	2	*MUC12, MUC3A*
Signalling events mediated by Stem cell factor receptor (c-Kit)	2.77	4.5×10^{-3}	1.3×10^{-2}	1	*PIK3C2B*
p53 pathway	2.56	5.7×10^{-3}	1.3×10^{-2}	1	*MDM4*
ATF-2 transcription factor network	2.56	5.7×10^{-3}	1.3×10^{-2}	1	*ACHE*

Figure 3.11. Heatmap of gene expression of the 17 prioritised genes across GTEx tissues. The average expression of the 17 prioritised genes at the haemorrhoids associated loci across A) 54 specific and B) 30 general GTEx v8.0 tissue types are shown. Average expression value per tissue type per gene are shown following winsorization at 50 and log 2 transformation with pseudocount 1. Expression value is depicted in Transcripts per Million, and both genes and tissues have been ordered by hierarchical clustering.

A



B



3.3.5 Genetic correlations with haemorrhoids associated phenotypes

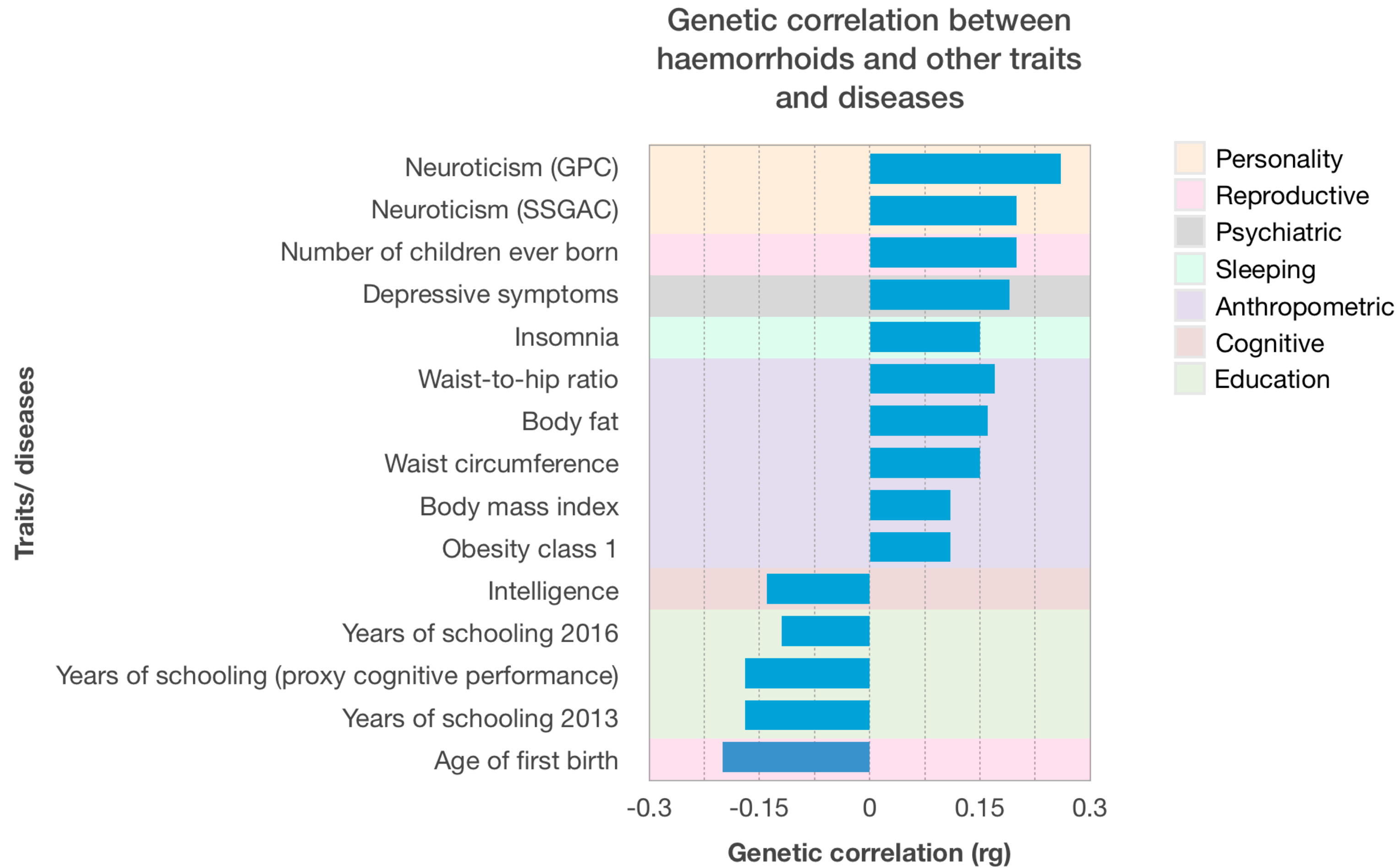
The contribution of common variants to haemorrhoids risk were examined using LD Score regression[27]. Using LD scores from ~1.2 million variants found in European populations, the SNP-based heritability (h^2_{SNP}) for haemorrhoids was estimated in the UK Biobank population to be 2.37% (S.E. = 0.18%). Correlation testing between haemorrhoids and other traits using GWAS summary statistics from LD Hub highlighted fifteen traits/diseases across seven trait categories significantly associated with haemorrhoids ($P_{bon} < 3.85\times10^{-3}$).[11]

Of the 89 traits tested across seven trait categories (**Appendix Table 3.4.**), 15 traits across seven trait categories were significantly correlated ($P_{bon} < 3.85\times10^{-3}$). Four categories (personality, psychiatric, sleeping, anthropometric) contained traits that were positively correlated with haemorrhoids (**Figure 3.12**). Traits closely related to mental health were amongst those positively correlated with haemorrhoids; with neuroticism sharing a genetic overlap of ~20 - 26%, depressive symptoms ~19%, and insomnia ~15%. The largest quantity of positively-associated traits was in the anthropometric category, with several measures relating to bodyweight sharing a genetic overlap of between 11 and 17%.

Of note, the reproductive category contained traits that were both positively and negatively correlated with haemorrhoids, with number of children ever born (parity status) having a positive correlation (~20% genetic overlap) and age at first birth having a negative genetic correlation (-20%). In the cognitive and education trait categories, several measures of literacy and cognitive performance negatively

correlated with haemorrhoids (years of schooling -12-17% and intelligence -14% genetic overlap).

Figure 3.12. Genetic correlation between haemorrhoids and other traits and diseases. Genetic correlation (r_g) between haemorrhoids and publicly available phenotypes in LD Hub, using LDSC regression. Fifteen traits met a Bonferroni-corrected significance $P_{bon} < 3.85\times10^{-3}$ and are depicted. GPC, Genetics of Personality Consortium; SSGAC, Social Science Genetics Association Consortium.



3.3.6. Drug target enrichment analysis

Interrogation within the Open Targets Platform underlined the potential for therapeutic targeting of the protein products of 16 of the 17 prioritised genes. Forty-eight drug pathways reached a nominal significance threshold (**Appendix Table 3.5.**), with the mucin drug targets MUC12 and MUC3A in the '*Defective GALNT3 causes familial hyperphosphataemic tumoral calcinosis (HFTC)*' pathway being most enriched (P = 3.3×10^{-4}). Pharmacological tractability data for 13 gene targets were available, with three (*ACHE*, *PIK3C2B* and *SMAD3*) predicted tractable to small molecule targeting and six predicted tractable by antibody targeting to a high confidence (*ACHE*, *ADRA2B*, *ELN*, *MUC12*, *MUC3A*, *PIK3C2B*) (**Figure 3.13**). Three gene targets were pharmacologically active with 43 known pharmaceutical interactions (*ACHE*, *ADRA2B*, *ELN*), with ADRA2B (encoded by *RP11-127H5.1*) having the most drug interactions (n= 30), followed by ACHE (n= 12) and ELN (n= 1) (**Appendix Table 3.6.**). Of note, the structural protein ELN is an established target by the elastin proteolytic enzyme, vonapanitase, which has been independently investigated in five completed, stage 1-3 clinical trials (clinicaltrials.gov) for the management of chronic kidney disease.

Figure 3.13. Tractability information for gene targets in the drug-target enrichment analysis. Of the 17 prioritised genes interrogated in the Open Targets Platform, tractability information on 13 genes was available; the tractability of each of these genes to small molecule and/or antibody targeting is shown in the below figure.

Small molecule
Antibody
Target symbol
Clinical precedence
Discovery precedence
Predicted tractable
Clinical precedence
Predicted tractable high confidence
Predicted tractable mid-low confidence
ACHE
ADRA2B
CACFD1
CPLANE1
ELN
LRRN2
MDM4
MUC12
MUC3A
NUP155
PIK3C2B
SMAD3
XKR9

3.3.7. Genetic risk score for haemorrhoids

The 13 independent significant signals associated with haemorrhoids were used to calculate a weighted genetic risk score (wGRS) for all 401,658 participants in UK Biobank. As expected, the wGRS for haemorrhoids cases (0.856) was higher than in controls (0.835; $P = 7.63 \times 10^{-135}$). Moreover, the haemorrhoids cases that had undergone surgery had a higher wGRS (0.865) compared to haemorrhoids cases that had not undergone a surgical procedure to manage their disease (0.848; $P = 4.58 \times 10^{-27}$) (**Table 3.5**).

Table 3.5. Weighted genetic risk score for haemorrhoids in the UK Biobank cohort

Group	Haemorrhoids cases	Controls	P-Value†	Haemorrhoids cases with operation code	Haemorrhoids cases without operation code	P-Value§
N	31,652	369,931		14,735	16,917	
Mean wGRS* (standard deviation)	0.835 (0.144)	0.856 (0.145)	7.63×10^{-135}	0.865 (0.145)	0.848 (0.145)	4.58×10^{-27}

*wGRS: weighted genetic risk score. †Unpaired two-tailed t-test between haemorrhoids cases and controls. §Unpaired two-tailed t-test between haemorrhoids cases with an operation code and haemorrhoids cases without an operation code.

3.4. Discussion

3.4.1. Summary

Haemorrhoids are a major health burden associated with high morbidity, reduced quality-of-life, and high healthcare costs.[28] Effective medical therapies are lacking in the armamentarium of the colorectal surgeon and are increasingly required. Performing the first genome-wide association analysis of haemorrhoids to date, in 31,652 cases and 369,931 control, 13 independent susceptibility signals reaching genome-wide significance at 12 risk loci were identified. Further, using *in silico* analyses, strong evidence of functionality in the haemorrhoids-associated regions was demonstrated; showing that genes in these associated regions are significantly enriched for expression in vascular tissue, tissues associated with hollow organs with a high preponderance of smooth muscle, and extracellular matrix regulation pathways. Genetic correlation analyses demonstrated a strong and positive genetic correlation between haemorrhoids and several mental-health related phenotypes, parity status, and weight. Genetic risk scoring of patients with haemorrhoids undergoing surgery correlated with a more severe phenotype - representing a first step in personalised medicine approaches to managing haemorrhoids. The majority of prioritised genes demonstrate pharmaceutical potential, with three genes (*ACHE*, *ADRA2B*, *ELN*) under investigation in active pharmaceutical research efforts.

3.4.2. Extracellular Matrix Remodelling

Haemorrhoids are fibrovascular cushions in the rectal mucosa which constitute a complex arrangement of smooth muscle, connective tissue and elastic fibres with arteriovenous communications.[29] Anatomical studies demonstrate haemorrhoidal veins and the surrounding musculature to be supported by an extracellular lattice of longitudinally organised collagen and elastic fibres.[30] Diseased haemorrhoid tissue has been shown to have an abnormal ratio of type I to III collagen[31], suggesting a role for collagen metabolism and resulting mechanical instability and loss of tensile strength as contributors to haemorrhoid disease pathobiology.[31,32] Indeed, haemorrhoid disease has also been reported in patients with Ehlers-Danlos Syndrome (which is characterised by mutations in collagen genes).[33] The ECM of the anal submucosa degrades with age, with adult anal specimens found to have a less structured appearance of collagen fibres[34] — this also aligns with the peak incidence of haemorrhoids in middle to old age.[35]

In the ECM of connective tissues, collagen (tensile strength) works in concert with elastin (elastic recoil) to determine the geometric arrangement and mechanical properties of tissues. Indeed matrix metalloproteinases (MMPs) negatively regulate ECM composition, and gene expression of MMPs -1, -2 and -3 have been shown to be increased in grade I and II haemorrhoids disease with MMPs -2, -8, -9 and NGAL levels shown to be elevated in grade 3 haemorrhoids.[32] Studied to a lesser degree in the biology of haemorrhoids is the contribution of elastin. The association analysis identified rs77689666-G ($P = 2.90\times10^{-9}$, OR = 1.10, EAF = 0.06), a variant residing in an intergenic region at 7q11.23, ~8kb upstream of *ELN*. *ELN* encodes elastin, which

is the core component of elastic fibres and provides elastic recoil to tissues such as haemorrhoidal veins. This is supported by the fact *ELN* was the most differentially-expressed gene in blood vessel tissue of all 17 prioritised genes, highlighting its roles in the elasticity of vascular tissues. Moreover, the drug-target enrichment analyses revealed elastin to be among the three prioritised genes targets with confirmed pharmaceutical interactions. XGR pathway analysis enriched canonical gene sets involved in structural components of the ECM. Disruption of the muscular and elastic constituents of haemorrhoids is thought to lead to distal shifting of the vascular padding (sliding anal lining theory[2]), a common feature of haemorrhoid disease.[36] Loss of integrity of the ECM, and in particular of elastin during ageing has been posited as an important pathway in the pathobiology of haemorrhoids.[30] Elastin may therefore represent an important player in the biology of haemorrhoids. Moreover, collagen and elastin expression in rectal submucosa has also been shown to be significantly reduced in obstructive defecation syndrome[37], another disorder of rectum.

Epidemiological observations have highlighted haemorrhoids to be associated with other disorders of connective tissue, namely: hernia[38], varicose veins[39,40], , genitourinary prolapse[41,42], and diverticular disease.[43,44] The G minor allele of SNP rs77689666 also significantly associates with and confers risk in Europeans for diverticular disease in a previous GWAS by Maguire[45] ($P = 1.65\times10^{-9}$, OR = 1.11) and in the hernia GWAS ($P = 3.2\times10^{-10}$, OR = 1.08) (*Please refer to* ***Chapter 4***). These results are consistent with the hypothesis of a degree of shared pathophysiology between haemorrhoids and other associated elastic tissue disorders.

3.4.3. TGFβ-Signalling Pathway

The transforming growth factor-β (TGF-β) signalling pathway plays a central role in normal physiological and disease processes via regulation of core cellular processes such as growth, differentiation, migration, apoptosis, and ECM remodelling.[46,47] The TGF-β signalling pathway constitutes a complex family of 33 polypeptide growth factors, including Smads which are key intracellular mediators of the response to TGF-β in ECM-producing mesenchymal cells (myofibroblasts and fibroblasts).[48,49] Target genes known to be Smad-responsive include key fibrillar ECM glycoproteins such as collagen and fibronectin, MMPs, and tissue inhibitors of MMPs (TIMP-1). [50–53] Indeed, TGF-β signalling, has been shown to play a central pathogenic role in the development of aortic aneurysm in Marfan syndrome.[54]

SNP rs72743461-A is significantly associated with haemorrhoids, a variant residing within intron 1 of *SMAD3* at 15q22.33 ($P = 1.6\times10^{-8}$, OR = 1.06, EAF = 0.24). The TGF-β/Smad3 signalling pathway is a key component of tissue fibrogenesis[52], acting as a potent stimulator of ECM protein accumulation.[55–58] Of note, two intronic variants were identified residing < 1kb from - and in high linkage with - the lead SNP at this locus: rs17293632 ($P = 1.7\times10^{-8}$, OR = 1.06, EAF = 0.24, $r^2 = 1.00$) and rs1866316 ($P = 3.8\times10^{-8}$, OR = 1.05, EAF = 0.30, $r^2 = 0.71$) which are among the top four variants from the GWAS predicted *both* deleterious (CADD score[18] = 22.6 and 14.4, respectively) and demonstrating strong regulatory potential (RegulomeDB[19] score = 2A and 2B, respectively). Disruption of TGF-β/Smad3 signalling pathway through loss of Smad3 is thought to reduce fibrogenic mesenchymal cell activation and confer resistance to the development of colorectal fibrosis[59], and induce resistance to tissue

fibrosis in other organs[52] (such as skin[60], kidney[61], lung[62], and liver[63]). One can therefore hypothesise that overexpression of *SMAD3,* and activation of the TGF-β/Smad3 signalling pathway disrupts the ECM architecture of haemorrhoidal veins and may therefore predispose to risk of haemorrhoids development. Aside from its role in ECM regulation, the TGF-β-signalling pathway is a fundamental pathway involved in colorectal cancer risk[64], with SMAD3 in particular having been demonstrated to associate with survival in colorectal cancer.[65] Several variants in *SMAD3* have been shown to be associated with differential miRNA expression levels in both normal colorectal mucosa as well as tumour tissue.[66]

Endoglin (CD105) is a membrane glycoprotein expressed in vascular endothelial cells where it forms a core component of the receptor that binds TGF-β1 and TGF-β3. Endoglin is a proliferation-associated antigen on vascular endothelium and is a specific marker for neovascularisation.[67,68] Haemorrhoids samples have been found to have overexpression of endoglin[69]; supporting the involvement of the TGF-β signalling pathway in haemorrhoids and also suggesting a role for neovascularisation as well as extracellular matrix disruption in haemorrhoids biology.

3.4.4. Internal anal sphincter tonicity

Freckner and Euler[70] suggest the internal anal sphincter (IAS) to contribute as much as 85% of resting anal pressure in normal individuals. Prevailing manometric studies have demonstrated the presence of ultra-slow waves and high resting pressures in the anal canal of patients with haemorrhoids; which has been suggested to be due to IAS hypertonicity[71–73] (with one study by Teramoto also describing external AS hypertrophy[74]). High resting anal pressure in the anal canal are thought to impair venous return from haemorrhoidal veins during defecation.[1] Hancock demonstrated resting anal pressures to decline after manual anal dilation, further supporting the role of IAS hypertonicity in haemorrhoids pathobiology.[75]

The association analysis discovered rs4556017-T ($P = 9.10\times10^{-12}$, OR = 1.08, EAF = 0.15), which resides in intron 1 of *MUC12* at 7q22.1. Via eQTL mapping, rs4556017 was identified to be a robust eQTL for *ACHE* (~140kb upstream from *MUC12*) in GTEx v8 aorta tissue ($P_{eQTL}= 1.8\times10^{-15}$, FDR = 5.55×10^{-39}). *ACHE* encodes acetylcholinesterase which catalyses the degradation of acetylcholine in the neuromuscular junction leading to termination of synaptic transmission. Acetylcholine is thought to relax smooth muscle in the IAS by stimulating NO synthesis.[76] Detailed 3D reconstruction of the human autonomic innervation of the IAS has been previously performed.[77] NO is the predominant inhibitory neurotransmitter of non-adrenergic non-cholinergic enteric neurones that mediate relaxation of the IAS during defecation.[78] Over-expression of the *ACHE* gene in patients with haemorrhoids may therefore have some role to play in the pathobiology of haemorrhoids by imparting a higher resting anal pressure via reduced inhibitory signalling of the IAS tone. The ACHE gene

product was found to have confirmed pharmaceutical interactions in the drug-target enrichment analysis. Indeed, injection of botulinum toxin into the anal sphincter — which prevents release of acetylcholine from presynaptic nerve endings and noradrenaline from sympathetic nerve endings[79]—has been shown to be highly effective in rapid relief of pain from thrombosed external haemorrhoids (thought to be due to a reduction in anal resting pressure).[80] ACHE may therefore represent an early target and warrant further research as a potential therapeutic avenue for haemorrhoids disease.

It is encouraging that through eQTL mapping, this study has implicated for the first time, a biological candidate in support of the IAS hypertonicity theory. However, an important caveat is that several studies have failed to identify differences in IAS thickness in patients with haemorrhoids compared to healthy controls.[81,82] This suggests alternative factors may be contributing to the high anal pressure in patients with haemorrhoids. The vascular hyperplasia theory proposed by Stelzner[29], represents an alternative view implicating a role for vascular distension and increased pressure of the anal cushions themselves in the increased anal pressure seen in patients with haemorrhoids. Sun et al[100], demonstrated in their study that the high anal pressure in patients with haemorrhoids were indeed due to higher pressure in the *vascular cushions*, rather than the IAS. However, it is very likely that *both* these factors hold part of the truth and contribute to overall increased anal pressures seen in haemorrhoids.

3.4.5. Haemorrhoids and arterial dilating diseases

rs1333047 is an intergenic variant that resides ~4kb upstream of *CDKN2B-AS1* (CDKN2B Antisense RNA 1), located near the *CDKN2B-CDKN2A* gene cluster. This region at 9p21.3 is known to be heavily associated with cardiovascular disease, endometriosis, periodontitis, glaucoma, colorectal cancer, and several aneurysm phenotypes. Striking overlap between the lead SNP at this *CDKN2B-AS1* region and variants previously known to be associated with several aneurysm phenotypes is noted (all variants significantly associated with haemorrhoids and in high LD (r^2 > ~0.8) with rs1333047). All associated variants reside in an ~4.2kb window adjacent to rs1333047: Abdominal aortic aneurysm: rs10757278[83] (P = 2.2×10^{-11}, r^2 = 0.97), rs10757274[84] (P = 1.9×10^{-10}, r^2 = 0.84) ; intracranial aneurysm: rs10733376[85] (P = 2.6 $\times10^{-10}$, r^2 = 0.87), rs10757272[86] (P = 8.5 $\times10^{-10}$, r^2 = 0.85) & rs6475606 (P = 1×10^{-8}, r^2 = 0.79); and combined intracranial, abdominal and thoracic aneurysms (pleiotropy): rs7866503[87] (P = 2.1×10^{-10}, r^2 = 0.86). The dilating venous disorders (namely haemorrhoids, varicose veins, varicoceles) are a phenomenon which occur in a different vascular territory - and with differing clinical manifestations - to the dilating arterial disorders, namely arterial aneurysms and coronary artery ectasia. However, it is understood that they share an overlapping pathobiology, with pathological weakness in vascular wall conferring disease risk.[88] Implicated and shared pathological processes underpinning both arterial and venous dilating disease include ECM remodelling[2,89], oxidative stress[90], increased inflammatory processes[91,92], and increased NO stimulation[93,94] in the vessel wall.

This finding lends support and brings together the previous associations implicated in ECM remodelling, TGF-β signalling and IAS tonicity. Lastly, in an analysis of the pheWAS catalogue, Salnikova, *et al*.[95] identified a pronounced signal (rs4977574, P = 7.0×10^{-4}) at *CDKN2B-AS1*, associated with 2796 haemorrhoids cases and 642 female stress urinary incontinence patients. It is encouraging that in this GWA study, variant rs4977574 at *CDKN2B-AS1* was significantly associated with haemorrhoids and in high LD with the lead SNP at this locus (P = 1.9×10^{-10}, r^2 = 0.84).

3.4.6. Haemorrhoids and colorectal cancer

Cancer of the haemorrhoidal veins is rare[96], however haemorrhoids can often occur alongside colorectal cancer. Rectal bleeding is a common symptom in haemorrhoids; it can be indicative of benign pathology but also of proximal pathology.[97] Discriminating haemorrhoids from anorectal malignancy can be difficult, with coincidental pathology occurring in a large proportion of patients, especially in the elderly.[98] These results support the need for great care in investigating these patients to ensure malignancy (or inflammatory bowel disease) is ruled out.[99] Several studies have previously demonstrated the preponderance of haemorrhoids in colorectal cancer patients[100], and while this may have an associated risk of confounding (due to similar symptomatology), it is an important observation nonetheless.

Of the 12 risk loci associated with haemorrhoids in this study, eight loci are mapped to genes which are associated with malignancy, of which six loci are associated with colorectal malignancy. Mucins are epithelial glycoproteins that are highly-expressed in colorectal cancer.[101] The mucins MUC3A (transmembrane mucin) and MUC12 (membrane-bound mucin) - to which rs4556017 is mapped - are evolutionarily related to each other and thought to be down-regulated in colorectal cancers.[102] ELN (lead SNP: rs77689666) is a component of ECM glycoproteins and is implicated in the tumour microenvironment. ELN gene expression is increased in colorectal tumours, with MMP9 gene expression found to be elevated and MMP12 gene expression down-regulated.[103] Slattery *et al.*[104] found variants in *SMAD3* to confer colon cancer-specific survival. Locus 8q13.3 (lead SNP rs4612371) was eQTL mapped to *LACTB2* via associated SNP rs6999140 in several vascular and cultured fibroblast tissues (GTEx

v8.0 Tibial Artery, P_{eQTL} = 7.34×10^{-17}, FDR = 2.16×10^{-15}; GTEx v8.0 Cells Cultured Fibroblasts, P_{eQTL} = 1.11×10^{-8}, FDR = 2.03×10^{-9}; GTEX v8.0 Aorta Artery, P_{eQTL} = 2.22×10^{-8}, FDR = 1.34×10^{-5})). A recurrent in-frame gene fusion of *LACTB2-NCOA2* has been found to promote colorectal carcinogenesis via inactivation of the negative growth regulatory gene *NCOA2* in colorectal cancer.[105] MDM4 and the nucleoporin Nup155 are intimately involved in in the p53 pathway — with MDM4 regulating p53 activity and Nup155 acting as a p53 repression target.[106] Finally, *CACFD1* mapped by rs687621 (P=3.3×10^{-26}, OR=1.10), encodes human flower (hFWE) - inhibition of flower protein expression has been found to reduce tumour growth and metastasis, including imparting sensitivity to chemotherapy.[14]

3.4.7. Heritability and genetic correlations of haemorrhoids disease

Despite being a highly common condition with a significant global health burden[99], to date, no family or twin studies have attempted to demonstrate the population variance for haemorrhoids disease attributable to genetic differences. Indeed several studies have suggested haemorrhoids has *no* familial component[2,107], or that any familial component is confounded by common lifestyle habits[108] or sufferers increased awareness of their parents' anal health.[3] Through LDSC regression[27], the SNP-based heritability (h^2_g) for haemorrhoids - based on ~10% of the common frequency variants in the GWAS - was found to be 2.37% (0.18%). Since common variation accounts for a small proportion of the overall narrow-sense heritability of complex diseases[109], it is likely that this analysis does not capture the *full* extent of the heritability for haemorrhoids. In this large cohort, haemorrhoids has been found to have a notable genetic architecture, with several significantly associated putative and biologically-plausible risk loci.

Genetic correlation analyses identified several traits which significantly correlated with haemorrhoids. A striking positive correlation was seen between haemorrhoids and weight-related phenotypes (Genetic overlap 11 - 17%). This is supported by the fact that obesity is an independent risk factor for haemorrhoids[110,111]; postulated to impede venous return through increased intraabdominal pressure, and imparting increased stress on rectal musculature.[111] To this end, pregnancy is also a risk factor for haemorrhoids for these reasons[3], and thought to also be compounded by hormonal changes.[3] The correlation analysis highlighted parity status (*number of children ever born*) to be positively correlated (~20% genetic overlap) with haemorrhoids and age

of first birth to be negatively correlated (-20%) with haemorrhoids. These findings are likely due to the fact that a younger age of birth relates to a higher parity status, increasing risk of haemorrhoids.

Several mental-health related phenotypes (neuroticism, depression, and insomnia) were among the most significant and positively correlated (~15 - 26% genetic overlap) with haemorrhoids. Recurrent haemorrhoid symptoms such as anal bleeding, itchiness and pain - made worse by high recurrence following surgery - may impact quality-of-life in patients with haemorrhoids. Lee *et al*. demonstrate in a national cross-sectional study (n = 17,228 (2480 haemorrhoids cases)), that both self-reported depression and physician diagnosed depression significantly correlate with haemorrhoids.[111] A genetic susceptibility to haemorrhoids may therefore include a heightened risk of mental health sequelae in these patients.

3.4.8. Genetic risk score for haemorrhoids correlates with disease severity

In the USA (1999), ~23 million adults (~13% of the USA population) were projected to have haemorrhoids[112], of which 21% (7.7 million) were estimated to have had surgical intervention for haemorrhoids.[113] The genetic risk score derived from the association signals was higher among patients with haemorrhoids requiring surgery (therefore likely to represent the phenotypically more severe end of the disease spectrum), than those managed outwith surgery. This provides the proof-of-concept that the use of genetic risk scoring may enable pinpointing of patients that have a genetic susceptibility towards more severe haemorrhoidal disease that is likely to require surgical intervention. This could foreseeably enable early lifestyle modification, such as weight loss and increased fibre intake, which may show greater impact early in the disease course[32], and moreover enable early intervention when symptoms appear and guide surgical plans.

3.4.9. Strengths and limitations

There are several limitations that warrant further discussion. In the GWAS of varicose veins (**Chapter 2**), ~1/3 of all discovery loci from UK Biobank replicated in an independent cohort. Independent replication of this haemorrhoids GWAS is therefore necessary to control the false positive rate. Further, an independent cohort would enable the genetic risk score to be validated externally, limiting bias. Another weakness is that the genetic factors may be a small contributor to overall disease risk – this suggests that environmental factors, in concert with genetic factors, contribute to overall haemorrhoids disease risk. In the absence of current family or twin studies, and the use of common variation in this study, it is likely that this analysis has underestimated the full extent of the heritability of haemorrhoids.

Several strengths go in some way to lend credence to the study findings. Firstly, the use of the UK Biobank has enabled among the largest study to date of this disease model in the literature, allowing the identification of several robust signals. The rigorous bioinformatic analyses employed has enabled characterisation of biologically viable candidates which demonstrate natural clustering and are supported by previous studies. Lastly, the genetic risk score provides a partial validation of the association, paving the way for genetic risk stratification in personalised medicine approaches to haemorrhoids.

3.5. Conclusion

This study represents the first association analysis of haemorrhoids, a disease with a substantial global prevalence, patient morbidity, and high attributed healthcare costs. Moreover, several of the loci appear to cluster in shared pathways which demonstrate biological plausibility, including extracellular matrix remodelling, TGF-β signalling, and internal anal sphincter hypertonicity. These results lend credence to the study findings and are an important step in highlighting core players in haemorrhoids pathobiology. The weighted genetic risk score correlated with disease severity, indicating that individuals with a higher genetic burden were more likely to be on the phenotypic extreme of haemorrhoids – an important step in personalised medicine approaches to haemorrhoids. Further independent replication analyses and characterisation of loci through functional work are required to characterise these pathways further.

3.6. Chapter References

1. Margetis, N. Pathophysiology of internal hemorrhoids. *Ann. Gastroenterol.* **32**, 264–272 (2019).
2. Pata, F. *et al.* Anatomy, Physiology and Pathophysiology of Haemorrhoids. *Rev. Recent Clin. Trials* **15**, (2020).
3. Loder, P. B., Nicholls, R. J. & Phillips, A. K. S. *Haemorrhoids: pathology, pathophysiology and aetiology*. *British Journal of Surgery* **81**, (1994).
4. Bruch, H. P. & Roblick, U. J. Pathophysiologie des Hämorrhoidalleidens. *Chirurg* **72**, 656–659 (2001).
5. Khan, N. M. M. Injection sclerotherapy versus electrocoagulation in the management outcome of early haemorrhoids - PubMed. *J Pak Med Assoc* **56**, 579–582 (2006).
6. Yildiz, T., Aydin, D. B., Ilce, Z., Yucak, A. & Karaaslan, E. External hemorrhoidal disease in child and teenage: Clinical presentations and risk factors. *Pakistan J. Med. Sci.* **35**, 696–700 (2019).
7. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med.* **12**, (2015).
8. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
9. Loh, P. Mixed-model association for biobank-scale datasets. **50**, 906–908 (2018).
10. Loh, P. R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
11. Zheng, J. *et al.* LD Hub: A centralized database and web interface to perform

LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272–279 (2017).

12. Carvalho-Silva, D. *et al.* Open Targets Platform: new developments and updates two years on. *Nucleic Acids Res.* **47**, D1056–D1065 (2019).
13. Bulik-Sullivan, B. *et al.* LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
14. Madan, E. *et al.* Flower isoforms promote competitive growth in cancer. *Nature* **572**, 260–264 (2019).
15. Wang, K. L. M. H. H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
16. Watanabe, K., Taskesen, E., Van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1–10 (2017).
17. Ernst, J. & Kellis, M. Chromatin-state discovery and genome annotation with ChromHMM. *Nat. Protoc.* **12**, 2478–2492 (2017).
18. 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 (2018).
19. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22**, 1790–1797 (2012).
20. Aguet, F. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
21. Popadin, K. Y. *et al.* Gene Age Predicts the Strength of Purifying Selection

Acting on Gene Expression Variation in Humans. *AJHG* **95**, 660–674 (2014).

22. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLoS Comput. Biol.* **11**, 1–20 (2015).

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

24. Liberzon, A. *et al.* The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst.* **1**, 417–425 (2015).

25. D., W. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006

26. Fang, H., Knezevic, B., Burnham, K. L. & Knight, J. C. XGR software for enhanced interpretation of genomic summary data, illustrated by application to immunological traits. *Genome Med.* **8**, 1–20 (2016).

27. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).

28. Burkitt, D. P. & Graham-Stewart, C. W. Haemorrhoids-postulated pathogenesis and proposed prevention. *Postgrad. Med. J.* **51**, 631–636 (1975).

29. STELZNER, F. THE CORPUS CAVERNOSUM RECTI. *Dis. Colon Rectum* **7**, 398–9 (1964).

30. Lierse, W. Anatomie und Pathophysiologie des Haemorrhoidalleidens. *Langenbecks Arch Chir 1989;* **1989**, 769–72 (1989).

31. Willis, S., Junge, K., Ebrahimi, R., Prescher, A. & Schumpelick, V. Haemorrhoids - a collagen disease? *Color. Dis.* **12**, 1249–1253 (2010).

32. Serra, R. *et al.* Hemorrhoids and matrix metalloproteinases: A multicenter study on the predictive role of biomarkers. *Surg. (United States)* **159**, 487–494 (2016).

33. Plackett, T. P., Kwon, E., Gagliano, R. A. & Oh, R. C. Case Report Ehlers-

Danlos Syndrome-Hypermobility Type and Hemorrhoids. (2014). doi:10.1155/2014/171803

34. Morgado, P. J., Suárez, J. A., Gómez, L. G. & Morgado, P. J. Histoclinical basis for a new classification of hemorrhoidal disease. *Dis. Colon Rectum* **31**, 474–480 (1988).

35. Johanson, J. F. & Sonnenberg, A. The prevalence of hemorrhoids and chronic constipation. An epidemiologic study. *Gastroenterology* **98**, 380–386 (1990).

36. Thomson, W. H. F. The nature of haemorrhoids. *Br. J. Surg.* **62**, 542–552 (1975).

37. Li, J., Lin, H. & Ren, D. Expression of collagen and elastin fibers in the rectum of patients with obstructed defecation syndrome and its significance. *Zhonghua Wei Chang Wai Ke Za Zhi* **18**, 1215–9 (2015).

38. Abramson, J. H., Gofin, J., Hopp, C., Makler, A. & Epstein, L. M. The epidemiology of inguinal hernia. A survey in western Jerusalem. *J. Epidemiol. Community Health* **32**, 59–67 (1978).

39. Ekici, U., Kartal, A. & Ferhatoglu, M. F. Association Between Hemorrhoids and Lower Extremity Chronic Venous Insufficiency. *Cureus* **11**, (2019).

40. Burkitt, D. P. Varicose Veins Deep Vein Thrombosis, and Haemorrhoids: Epidemiology and Suggested Aetiology. *Br. Med. J.* **2**, 556 (1972).

41. Heslop, J. PILES AND RECTOCELES. *ANZ J. Surg.* **57**, 935–938 (1987).

42. Miedel, A., Tegerstedt, G., Mæhle-Schmidt, M., Nyrén, O. & Hammarström, M. Nonobstetric Risk Factors for Symptomatic Pelvic Organ Prolapse. *Obstet. Gynecol.* **113**, 1089–1097 (2009).

43. Humphreys, D. M. Diverticular disease: Three studies: Part I—Relation to other disorders and fibre intake. *Br. Med. J.* **1**, 424–425 (1976).

44. S, R. J. vd L. G. D. Diverticular disease. Pathology and clinical aspects based on 368 autopsy cases. *Zentralbl Chir* **116**, 991–8 (1991).

45. Maguire, L. H. *et al.* Genome-wide association analyses identify 39 new susceptibility loci for diverticular disease. *Nat. Genet.* **50**, 1359–1365 (2018).

46. Massagué, J., Blain, S. W. & Lo, R. S. TGFβ signaling in growth control, cancer, and heritable disorders. *Cell* **103**, 295–309 (2000).

47. Derynck, R. & Akhurst, R. J. Differentiation plasticity regulated by TGF-β family proteins in development and disease. *Nat. Cell Biol.* **9**, 1000–1004 (2007).

48. Ono, M. Molecular links between tumor angiogenesis and inflammation: Inflammatory stimuli of macrophages and cancer cells as targets for therapeutic strategy. *Cancer Science* **99**, 1501–1506 (2008).

49. Derynck, R. & Zhang, Y. Intracellular signalling: The mad way to do it. *Curr. Biol.* **6**, 1226–1229 (1996).

50. Wells, R. G. Fibrogenesis. V. TGF-β signaling pathways. *American Journal of Physiology - Gastrointestinal and Liver Physiology* **279**, (2000).

51. Roberts, A. B., Russo, A., Felici, A. & Flanders, K. C. Smad3: A key player in pathogenetic mechanisms dependent on TGF-β. in *Annals of the New York Academy of Sciences* **995**, 1–10 (New York Academy of Sciences, 2003).

52. Flanders, K. C. Smad3 as a mediator of the fibrotic response. *International Journal of Experimental Pathology* **85**, 47–64 (2004).

53. Inagaki, Y. & Okazaki, I. Emerging insights into transforming growth factor β Smad signal in hepatic fibrogenesis. *Gut* **56**, 284–292 (2007).

54. Benke, K. *et al.* The role of transforming growth factor-beta in Marfan syndrome. *Cardiol. J.* **20**, 227–234 (2013).

55. Powell, D. W. *et al.* Myofibroblasts. II. Intestinal subepithelial myofibroblasts.

American Journal of Physiology - Cell Physiology **277**, 183–201 (1999).

56. Pucilowska, J. B., Williams, K. L. & Lund, P. K. Fibrogenesis IV. Fibrosis and inflammatory bowel disease: Cellular mediators and animal models. *Am. J. Physiol. - Gastrointest. Liver Physiol.* **279**, (2000).
57. Vetuschi, A. *et al.* Smad3-null mice lack interstitial cells of Cajal in the colonic wall. *Eur. J. Clin. Invest.* **36**, 41–48 (2006).
58. Mckaig, B. C., Hughes, K., Tighe, P. J. & Mahida, A. Y. R. Differential expression of TGF-β isoforms by normal and inflammatory bowel disease intestinal myofibroblasts. *Am. J. Physiol. - Cell Physiol.* **282**, (2002).
59. Latella, G. *et al.* Smad3 loss confers resistance to the development of trinitrobenzene sulfonic acid-induced colorectal fibrosis. *Eur. J. Clin. Invest.* **39**, 145–156 (2009).
60. Lakos, G. *et al.* Targeted disruption of TGF-β/Smad3 signaling modulates skin fibrosis in a mouse model of scleroderma. *Am. J. Pathol.* **165**, 203–217 (2004).
61. Inazaki, K. *et al.* Smad3 deficiency attenuates renal fibrosis, inflammation, and apoptosis after unilateral ureteral obstruction. *Kidney Int.* **66**, 597–604 (2004).
62. Zhao, J. *et al.* Smad3 deficiency attenuates bleomycin-induced pulmonary fibrosis in mice. *Am. J. Physiol. - Lung Cell. Mol. Physiol.* **282**, 585–593 (2002).
63. Lee, J. Il *et al.* Role of Smad3 in platelet-derived growth factor-C-induced liver fibrosis. *Am. J. Physiol. Physiol.* **310**, C436–C445 (2016).
64. Tsushima, H. *et al.* High levels of transforming growth factor in patients with colorectal cancer: Association with disease progression. *Gastroenterology* **110**, 375–382 (1996).
65. Slattery, M. L., Lundgreen, A., Herrick, J. S., Wolff, R. K. & Caan, B. J. Genetic variation in the transforming growth factor-β signaling pathway and survival after

diagnosis with colon and rectal cancer. *Cancer* **117**, 4175–4183 (2011).

66. Slattery, M. L. *et al. Genetic variants in the TGFβ-signaling pathway influence expression of miRNAs in colon and rectal normal mucosa and tumor tissue*.

67. Brewer, C. Endoglin expression as a measure of microvessel density in cervical cancer. *Obstet. Gynecol.* **96**, 224–228 (2000).

68. Akagi, K. *et al.* Estimation of angiogenesis with anti-CD 105 immunostaining in the process of colorectal cancer development. *Surgery* **131**, S109–S113 (2002).

69. Chung, Y. C., Hou, Y. C. & Pan, A. C. H. Endoglin (CD105) expression in the development of haemorrhoids. *Eur. J. Clin. Invest.* **34**, 107–112 (2004).

70. Frenckner, B. & Euler, C. V. Influence of pudendal block on the function of the anal sphincters. *Gut* **16**, 482–489 (1975).

71. Hancock, B. D. Internal sphincter and the nature of haemorrhoids. *Gut* **18**, 651–655 (1977).

72. Gibbons, C. P., Bannister, J. J. & Read, N. W. Role of constipation and anal hypertonia in the pathogenesis of haemorrhoids. *Br. J. Surg.* **75**, 656–660 (1988).

73. Lane, R. H. Measurement of anal pressure in patients with haemorrhoids. *Schweiz. Rundsch. Med. Prax.* **71**, 112–5 (1982).

74. Teramoto, T., Parks, A. G. & Swash, M. Hypertrophy of the external anal sphincter in haemorrhoids: a histometric study. *Gut* **22**, 45–48 (1981).

75. Hancock, B. D. & Smith, K. The internal sphincter and Lord's procedure for haemorrhoids. *Br. J. Surg.* **62**, 833–836 (1975).

76. Bhardwaj, R., Vaizey, C. J., Boulos, P. B. & Hoyle, C. H. V. Neuromyogenic properties of the internal anal sphincter: Therapeutic rationale for anal fissures. *Gut* **46**, 861–868 (2000).

77. Moszkowicz, David; Peschaud, Fredrique; Bessede, Thomas; Benoit, Gerard; Alsaid, B. Internal Anal Sphincter Parasympathetic-Nitrergic and Sympathetic-Adrenergic Innervation: A 3-Dimensional Morphological and Functional Analysis. *Dis Colon Rectum* **55**, 473–481 (2012).
78. O'Kelly, T., Brading, A. & Mortensen, N. Nerve mediated relaxation of the human internal anal sphincter: The role of nitric oxide. *Gut* **34**, 689–693 (1993).
79. Hallett, M. One Man's Poison — Clinical Applications of Botulinum Toxin. *N. Engl. J. Med.* **341**, 118–120 (1999).
80. Patti, R. *et al.* Randomized clinical trial of botulinum toxin injection for pain relief in patients with thrombosed external haemorrhoids. *Br. J. Surg.* **95**, 1339–1343 (2008).
81. Sun, W. M., Peck, R. J., Shorthouse, A. J. & Read, N. W. Haemorrhoids are associated not with hypertrophy of the internal anal sphincter, but with hypertension of the anal cushions. *Br. J. Surg.* **79**, 592–594 (1992).
82. Sun, WM; Read, NW; Shorthouse, A. Hypertensive anal cushions as a cause of the high anal canal pressures in patients with haemorrhoids. *Br. J. Surg.* **7**, 458–462 (1990).
83. Gretarsdottir, S. *et al.* Genome-wide association study identifies a sequence variant within the DAB2IP gene conferring susceptibility to abdominal aortic aneurysm. *Nat. Genet.* **42**, 692–697 (2010).
84. Jones, G. T. *et al.* Meta-Analysis of Genome-Wide Association Studies for Abdominal Aortic Aneurysm Identifies Four New Disease-Specific Risk Loci. *Circ. Res.* **120**, 341–353 (2017).
85. Foroud, T. *et al.* Genome-wide association study of intracranial aneurysm identifies a new association on chromosome 7. *Stroke* **45**, 3194–3199 (2014).

86. Low, SK; Takahashi, A; Cha, PC; Zembutsu, H; Kamatani, N; Kubo, M; Nakamura, Y. Genome-wide association study for intracranial aneurysm in the Japanese population identifies three candidate susceptible loci and a functional genetic variant at EDNRA. *Hum Mol Genet* **21**, 2102–2110 (2012).

87. van 't Hof, F. N. G. *et al.* Shared Genetic Risk Factors of Intracranial, Abdominal, and Thoracic Aneurysms. *J. Am. Heart Assoc.* **5**, (2016).

88. Yetkin, E. & Ileri, M. Dilating venous disease: Pathophysiology and a systematic aspect to different vascular territories. *Med. Hypotheses* **91**, 73–76 (2016).

89. Yetkin, E. & Waltenberger, J. Novel insights into an old controversy: Is coronary artery ectasia a variant of coronary atherosclerosis? *Clinical Research in Cardiology* **96**, 331–339 (2007).

90. Emeto, T. I., Moxon, J. V., Au, M. & Golledge, J. Oxidative stress and abdominal aortic aneurysm: Potential treatment targets. *Clin. Sci.* **130**, 301–315 (2016).

91. Cebral, J. *et al.* Flow conditions in the intracranial aneurysm lumen are associated with inflammation and degenerative changes of the aneurysm wall. *Am. J. Neuroradiol.* **38**, 119–126 (2017).

92. Lim, C. S. & Davies, A. H. Pathogenesis of primary varicose veins. *Br. J. Surg.* **96**, 1231–1242 (2009).

93. Kuhlencordt, P. J. *et al.* Accelerated Atherosclerosis, Aortic Aneurysm Formation, and Ischemic Heart Disease in Apolipoprotein E/Endothelial Nitric Oxide Synthase Double-Knockout Mice. *Circulation* **104**, 448–454 (2001).

94. Jacob, T., Hingorani, A. & Ascher, E. Overexpression of transforming growth factor-β1 correlates with increased synthesis of nitric oxide synthase in varicose veins. *J. Vasc. Surg.* **41**, 523–530 (2005).

95. Salnikova, L. E., Khadzhieva, M. B. & Kolobkov, D. S. Biological findings from

the PheWAS catalog: focus on connective tissue-related disorders (pelvic floor dysfunction, abdominal hernia, varicose veins and hemorrhoids). *Hum. Genet.* **135**, 779–795 (2016).

96. Gujral, D. M., Bhattacharyya, S., Hargreaves, P. & Middleton, G. W. Metastatic rectal adenocarcinoma within haemorrhoids: A case report. *J. Med. Case Rep.* **2**, 128 (2008).

97. Pfenninger, J. L. & Zainea, G. G. *Common Anorectal Conditions: Part I. Symptoms and Complaints*. *American Family Physician* **63**, (2001).

98. Koning, M. & Loffeld, R. Rectal bleeding in patients with haemorrhoids. Coincidental findings in colon and rectum. *Fam Pr.* **27**, 260–262 (2010).

99. Tol, R. R. *et al.* European Society of ColoProctology: guideline for haemorrhoidal disease. *Color. Dis.* **22**, 650–662 (2020).

100. Kune, G. A., Kune, S. & Watson, L. F. Colorectal Cancer Risk, Chronic Illnesses, Operations, and Medications: Case Control Results from the Melbourne Colorectal Cancer Study. *Cancer Res.* **48**, (1988).

101. Matsuyama, T. *et al.* MUC12 mRNA expression is an independent marker of prognosis in stage II and stage III colorectal cancer. *Int. J. Cancer* **127**, 2292–2299 (2010).

102. Byrd, J. C. & Bresalier, R. S. Mucins and mucin binding proteins in colorectal cancer. *Cancer and Metastasis Reviews* **23**, 77–99 (2004).

103. Li, J. *et al.* Elastin is a key factor of tumor development in colorectal cancer. doi:10.1186/s12885-020-6686-x

104. Slattery, M. L. & Lundgreen, A. The influence of the CHIEF pathway on colorectal cancer-specific mortality. *PLoS One* **9**, (2014).

105. Yu, J. *et al.* Disruption of NCOA2 by recurrent fusion with LACTB2 in colorectal

cancer. *Oncogene* **35**, 187–195 (2016).

106. Holzer, K. *et al.* Nucleoporin Nup155 is part of the p53 network in liver cancer. *Nat. Commun.* **10**, 1–13 (2019).

107. Kaidar-Person, O., Person, B. & Wexner, S. D. Hemorrhoidal Disease: A Comprehensive Review. *Journal of the American College of Surgeons* **204**, 102–117 (2007).

108. Johanson, J. F. & Sonnenberg, A. Constipation is not a risk factor for hemorrhoids: A case-control study of potential etiological agents. *Am. J. Gastroenterol.* **89**, 1981–1986 (1994).

109. Yang, J., Zeng, J., Goddard, M. E., Wray, N. R. & Visscher, P. M. Concepts, estimation and interpretation of SNP-based heritability. *Nature Genetics* **49**, 1304–1310 (2017).

110. Riss, S. *et al.* The prevalence of hemorrhoids in adults. doi:10.1007/s00384-011-1316-3

111. Lee, J.-H., Kim, H.-E., Kang, J.-H., Shin, J.-Y. & Song, Y.-M. Factors Associated with Hemorrhoids in Korean Adults: Korean National Health and Nutrition Examination Survey. *Korean J Fam Med* **35**, 227–236 (2014).

112. LeClere, F. B., Moss, A. J., Everhart, J. E. & Roth, H. P. Prevalence of major digestive disorders and bowel symptoms, 1989. *Adv. Data* 1–15 (1992).

113. Sandler, R. S. & Peery, A. F. Rethinking What We Know About Hemorrhoids. *Clinical Gastroenterology and Hepatology* **17**, 8–15 (2019).

.

3.7. Chapter Appendix

The appendix for this chapter is provided as an online supplement at the following URL: **bit.ly/WAhmed_C3Appendix**

Table of Contents

1. Appendix Tables

Appendix Table 3.1. Predicted functional Intronic-intergenic candidate variants

Appendix Table 3.2. Genome-wide gene-based association analysis in MAGMA

Appendix Table 3.3. Enriched gene sets from genome-wide gene-based enrichment analysis in MAGMA v1.07

Appendix Table 3.4. Genetic correlations between haemorrhoids and other phenotypes

Appendix Table 3.5. Enriched drug pathways from the drug target enrichment analysis

Appendix Table 3.6. Pharmacologically-active targets identified in the drug-target enrichment analysis

Chapter 4: The shared genetic architecture of hernia phenotypes

---

4.1. Introduction

4.1.1. Rationale and aims

In **Chapter 1** I provided evidence for hernia being a complex disease model, with known familial clustering. Patients with a positive family history are at an eight-fold increased risk of groin hernia[1,2], and have an increased susceptibility to both contralateral and recurrent inguinal hernia.[3–5] Moreover, both concordant and discordant standardised incidence ratios (SIR) for multiple abdominal wall hernia (AWH) subtypes have been found to be higher than 2 in a study by Zöller *et al*[6], suggesting multiple hernia susceptibility, which is supported by clinical observational data.[7,8]

The broad aim of this chapter is to study the shared genetic architecture between four hernia subtypes. My hypothesis is that there will be certain genetic regions that predispose to more than one hernia subtype.

I will investigate this by:

1) Performing individual GWAS of four hernia subtypes (inguinal, femoral, umbilical, and hiatus hernia) using the UK Biobank resource
2) Look for evidence of shared genetics between these hernia subtypes by performing:

   i) Combined GWAS analyses of all four individual hernia subtypes in UK biobank and participants with multiple overlapping hernia phenotypes, as well as a GWAS of multiple overlapping hernia phenotypes alone

ii) A multi-trait analysis of all four individual hernia phenotypes in UK Biobank to uncover genetic regions of shared susceptibility

iii) A multivariate meta-analysis of all four individual hernia phenotypes in UK Biobank to uncover genetic regions of shared susceptibility

4.2. Methods

4.2.1. Ethics and consent

The research ethics and consent procedures of the UK Biobank are provided in **Chapter 2 (Section 2.2.1.)**.[9]

4.2.2. Study participants

A complete description of the study participants of the UK Biobank cohort are provided in **Chapter 2 (Section 2.2.2.)**.[10]

The UK Biobank cohort was used to define three different sets of hernia case-control cohorts, as described below:

i) **Individual hernia** – these were four sets of hernia cases that had diagnostic and/or operative codes for only *one* of the four hernia types. i.e. ether inguinal, femoral, umbilical or hiatus hernia. In other words, for all four hernia cases, cases with phenotype coding for more than one hernia were removed from the cohort. All cases in this cohort were matched 1:5 to non-hernia controls.

ii) **Overlap hernia –** this cohort consisted of hernia cases with diagnostic and/or operative coding for *more than* one of the four hernia types. In other words, cases with phenotype coding for only one of the four hernia types were removed from the cohort. All cases in this cohort were matched 1:5 to non-hernia controls.

iii) **Umbrella hernia** – this cohort consisted of the full set of hernia cases in UK Biobank (i.e. any individual with at least one diagnostic code for any hernia subtype was included in this group). The umbrella cohort was therefore constructed by combining cohorts (i) and (ii). All cases in this cohort were matched 1:5 to non-hernia controls.

i) Individual hernia cohort

The four individual hernia cases were defined if they had at least one of the following diagnostic or operative codes consistent with either inguinal, femoral, umbilical or hiatus hernia (**Appendix Table 4.1)**:

1. Primary and/or secondary ICD-10 codes for either inguinal, femoral, umbilical or hiatus hernia.
2. Primary and/or secondary OPCS code for either inguinal, femoral, umbilical or hiatus hernia.
3. Self-reported operation code for either inguinal, femoral, umbilical or hiatus hernia surgery.
4. Self-reported non-cancer illness code for either inguinal, femoral, umbilical or hiatus hernia.

For the four hernia case cohorts, after quality control (described in **Section 4.2.4**), 23,007 participants had diagnostic or operative codes for inguinal hernia, 1,578 for femoral hernia, 7,432 for umbilical hernia, and 31,543 for hiatus hernia. The final individual hernia cohorts were then defined by removing all overlapping hernia cases (a total of 4,216 overlapping cases were removed from the inguinal hernia cohort, 605

from the femoral hernia cohort, 2,076 from the umbilical hernia cohort and 4,596 from the hiatus hernia cohort). The final four individual hernia cohorts consisted of 18,791 inguinal hernia, 973 femoral hernia, 5,356 umbilical hernia, and 31,543 hiatus hernia cases **(Figure 4.1)**.

All individual hernia cases were matched 1:5 to controls in the UK Biobank cohort based on i) age (+/- 5 years), ii) sex, iii) genotyping platform. All control cohorts contained completely distinct participants.

The final individual hernia cohorts therefore consisted of the following participants:

- **Inguinal hernia:** 112,746 participants (18,791 cases and 93,955 controls)
- **Femoral hernia:** 5,838 participants (973 cases and 4,865 controls)
- **Umbilical hernia:** 32,136 participants (5,356 cases and 26,780 controls)
- **Hiatus hernia:** 193,788 participants (32,298 cases and 161,490 controls)

ii) Overlap hernia cohort

The overlap hernia cases were those cases that had diagnostic or operative codes for two or more hernia types and were subsequently removed from the individual hernia cohort. The final overlap cohort consisted of 5,219 cases (**Figure 4.2**), which were matched 1:5 (using the above described approach) to 26,095 non-hernia controls (total cohort 31,314 participants).

iii) Umbrella hernia cohort

The umbrella hernia cases were defined if they had diagnostic or operative codes for any hernia types, including those with an individual or overlapping hernia. The final umbrella cohort consisted of 62,637 cases after quality control (described in **Section 4.2.4**) (**Figure 4.3**), which were subsequently matched 1:5 (using the approach described previously) to 313,185 non-hernia controls (total cohort 375,822 participants)

Figure 4.1: Venn diagram of the four individual hernia case cohorts in UK Biobank. Following quality control (described in **Section 4.2.4**), a total of 62,637 participants in UK Biobank possessed a diagnostic and/or operative code for at least one of the four hernia subtypes. 5,219 cases possessed coding for two or more hernia subtypes (i.e. overlapping hernia) and were therefore removed (*grey*) to define the four individual hernia case cohorts (total 57,418).

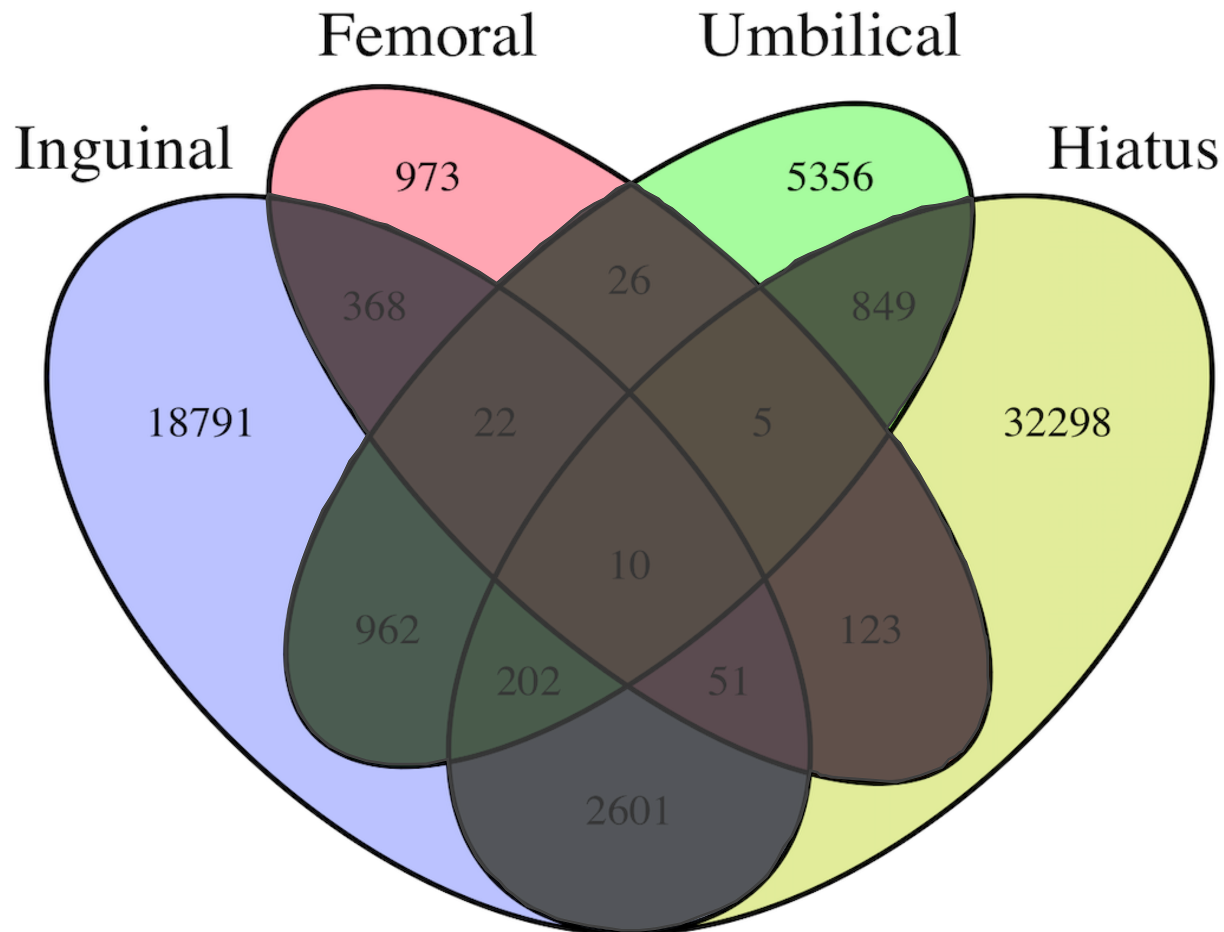


Figure 4.2: Venn diagram of the overlap hernia case cohort in UK Biobank. Following quality control (described in **Section 4.2.4**), a total of 5,219 cases possessed codes for two or more hernia subtypes (i.e. overlapping hernia) and were therefore included in the overlap hernia cohort. All cases with phenotype codes for only one hernia type were removed from this cohort (grey).

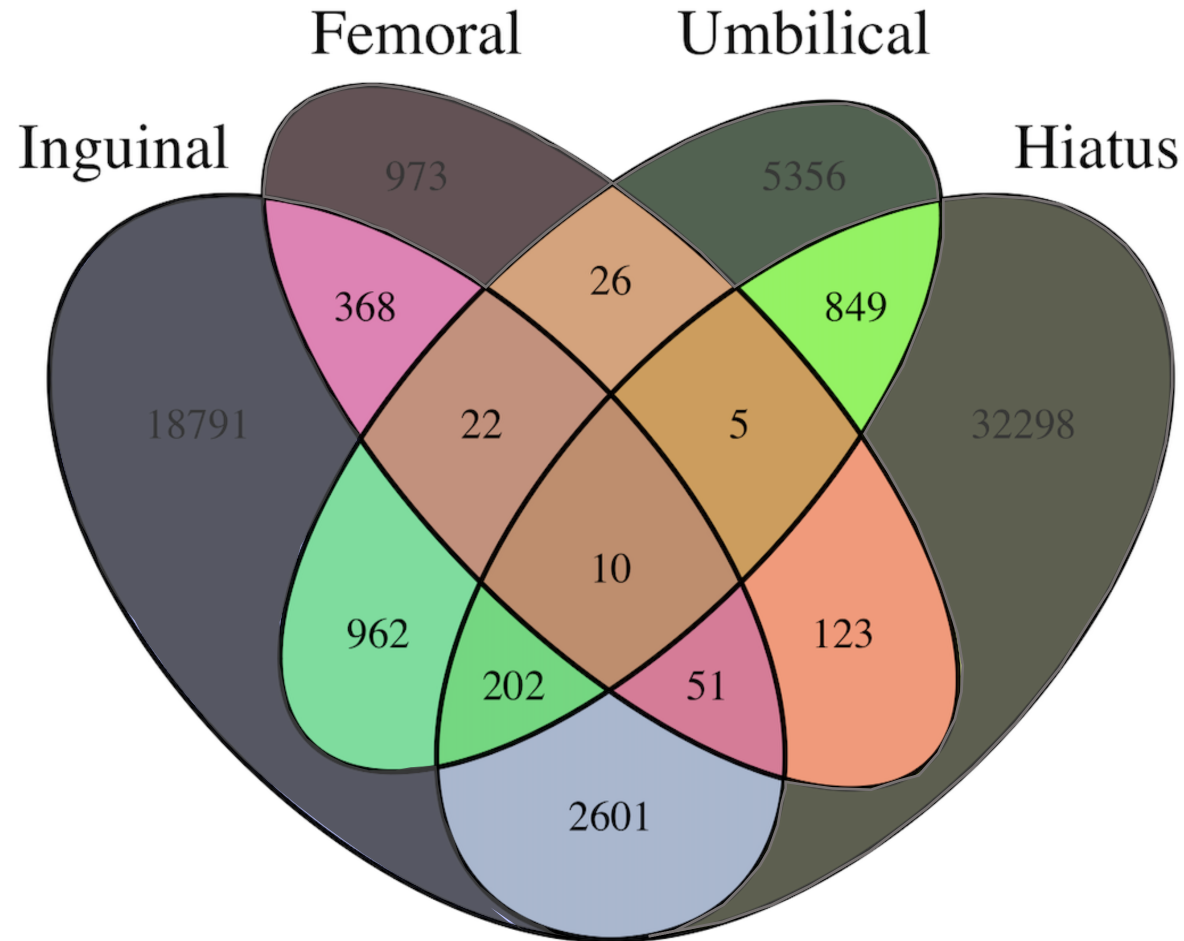
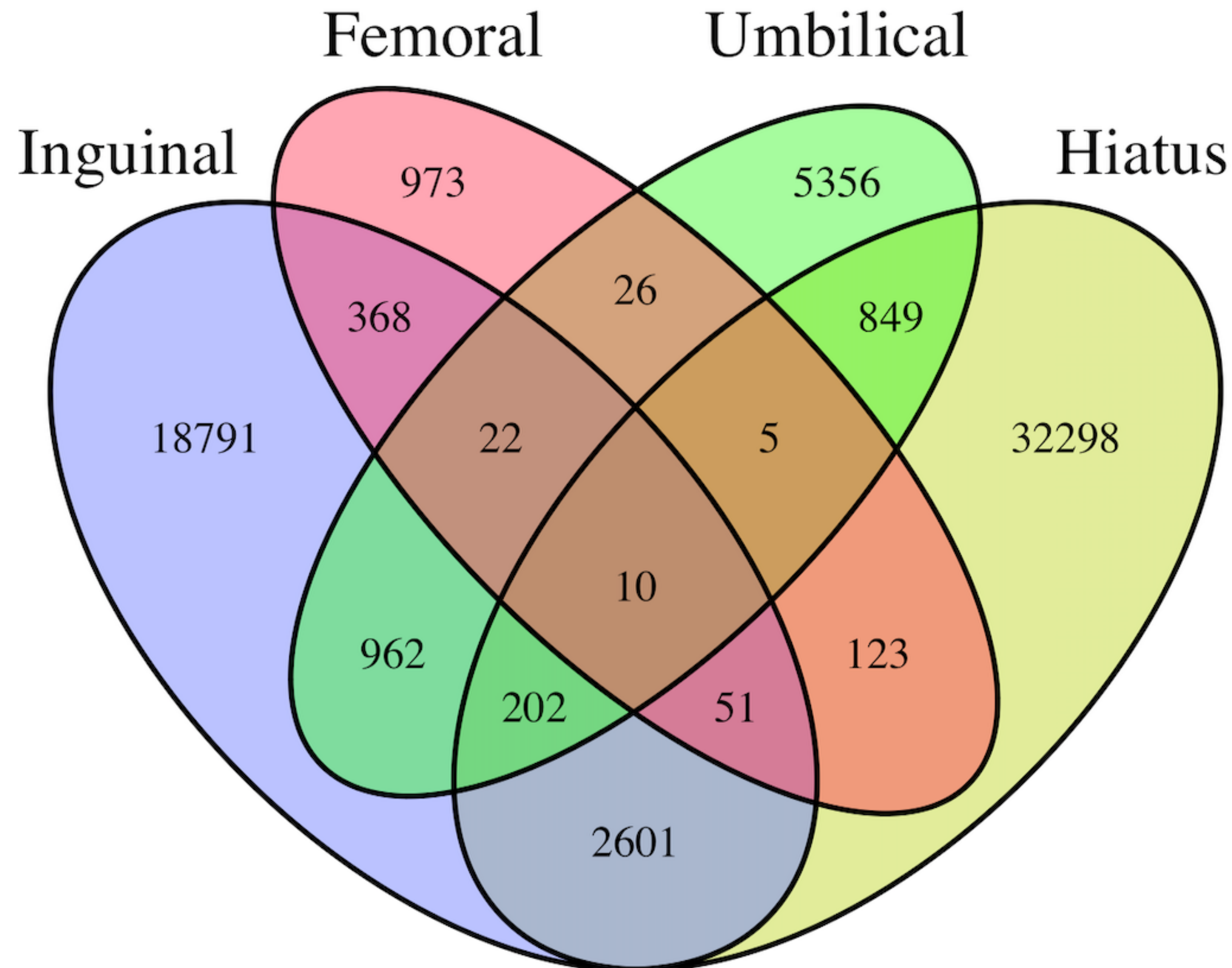


Figure 4.3: Venn diagram of the umbrella hernia case cohort in UK Biobank. Following quality control (described in **Section 4.2.4**), a total of 62,637 cases possessed codes for at least one hernia subtype and were therefore included in the umbrella hernia cohort.



4.2.3. Genotyping

A complete description of the genotyping procedure for the UK Biobank cohort is provided in **Chapter 2 (Section 2.2.3.).**[10]

4.2.4. Quality control

A complete description of the quality control (QC) procedure implemented centrally by UK Biobank and locally by our group is provided in **Chapter 2 (Section 2.2.4.).**[10]

After central (UK Biobank)[10], and local QC implementation, 86,794 participants and 230,562 genotyped SNPs were excluded from subsequent association analyses (See **Chapter 3**, **Figure 3.3**). In summary, a maximal set of 547,011 genotyped variants and 401,583 participants of white British ancestry passed the QC and were available for inclusion in the three hernia cohorts – matching was subsequently performed to define each cohort as described in **Section 4.2.2**.

4.2.5. Imputation

A complete description of the imputation methodology is provided in **Chapter 2 (Section 2.2.5)**.[10]

4.2.6. Association analyses in BOLT-LMM

Genome-wide association testing was performed across a total of ~9M SNPs (~500K genotyped (MAF ≥ 0 .01) and ~8.4M imputed SNPs (MAF ≥ 0.01, INFO Imputation

score ≥ 0.90)) for each of the four individual, overlap and umbrella cohorts using a linear mixed non-infinitesimal model implemented in BOLT-LMM v2.323.[11,12]

Further methodological details pertaining to the association analysis performed using BOLT-LMM[11,12] in this chapter are provided in **Chapter 2 (Section 2.2.13)**.

4.2.7. Genomic risk loci definition for BOLT-LMM studies

A complete description of the methods used for genomic risk loci definition for the individual, overlap and umbrella BOLT-LMM[11,12] analysis results is provided in **Chapter 2 (Section 2.2.7)**.

4.2.8. Functional annotation of BOLT-LMM studies

A complete description of the methods used for the functional annotation of variants for the individual, overlap and umbrella BOLT-LMM[11,12] analysis results is provided in **Chapter 2 (Section 2.2.8)**.

4.2.9. Candidate gene mapping of BOLT-LMM studies

A complete description of the methods used for the candidate gene mapping approach for the individual, overlap and umbrella BOLT-LMM[11,12] analysis results is provided in **Chapter 2 (Section 2.2.9)**.

4.2.10. Gene set, tissue and pathway analyses of BOLT-LMM studies

These analyses were performed for the individual, overlap, and umbrella analyses. This section follows the methodology described in **Chapter 2 (Section 2.2.10)**.

4.2.11. SNP-based heritability of BOLT-LMM studies

Methodological details on the SNP heritability analysis for each of the individual, overlap and umbrella summary statistics is described in **Chapter 2 (Section 2.2.11.)**.

4.2.12. Genetic risk score for hernia

The weighted genetic risk score (wGRS) methodology implemented in this chapter mirrors those described in **Chapter 2** (**Section 2.2.14.)**.[13] For each of the individual, overlap and umbrella summary statistics, the wGRS was compared between six groups of participants from the GWAS: i) all cases vs all controls; ii) surgical cases vs non-surgical cases, ii) cases with a single hernia (i.e. individual hernia cases) vs cases with that particular hernia type plus at least one more hernia type (i.e. overlapping hernia cases).

4.2.13. Multi-trait analysis in MTAG

To unpick the shared genetics behind the hernia phenotypes, MTAG (multi-trait analysis of GWAS) was implemented across the four individual hernia cohorts (total 57,418 cases and 287,090 controls).[14] The MTAG method enables joint analysis of summary statistics by combining several genetically correlated traits to augment the power to discover new susceptibility loci. MTAG produces trait-specific effect estimates across each SNP and through inverse-variance-weighted meta-analysis for each single-trait GWAS, MTAG outputs trait-specific association statistics.[14] The

MTAG analysis makes a homogeneity assumption across all SNPs, assuming an equal variance–covariance matrix of effect sizes across the traits. MTAG therefore works optimally when there is high genetic correlation between the input traits. The original authors have demonstrated analytically that even in scenarios where this assumption is not true, MTAG is a consistent and reliable estimator.[14] The final MTAG analysis was therefore performed across a shared 6,760,521 SNPs from each of the four individual hernia phenotypes. The genome-wide significant threshold for the MTAG association was set a $P < 5\times10^{-8}$. Pairwise testing was performed in GWAMA for all MTAG signals where the MTAG derived P-value was stronger than all four individual hernia trait P-values.[15]

4.2.14. Multivariate meta-analysis in metaUSAT

The metaUSAT multivariate method was used as an auxiliary meta-analysis method to further characterise potential regions of shared hernia susceptibility between the four individual hernia traits.[16] metaUSAT (meta-analysis unified score-based association test) performs a unified association test for each SNP (using the estimated correlation matrix to test association) across several trait summary statistics.[16] metaUSAT is data-adaptive, and was established to be robust to the association structure of correlated traits (less affected by the true (unknown) association structure) and is not dependent on individual-level data.[16] Unlike MTAG[14], metaUSAT does not assume homogeneity of effects across traits. metaUSAT outputs an approximate asymptotic P-value for the meta-analysis association and has been shown to maintain a low type I error in simulation experiments.[16] The metaUSAT meta-analysis was performed across the four individual hernia cohorts (total 57,418 cases and 287,090

controls) and 8,896,286 SNPs. The genome-wide significant threshold for the metaUSAT association was set a $P < 5\times10^{-8}$.

4.2.15. URLs

metaUSAT, https://github.com/RayDebashree/metaUSAT; MTAG, https://github.com/JonJala/mtag

4.3. Results

The analytic workflow implemented to unpick shared genetic biology between hernia phenotypes is summarised in **Figure 4.4**.

4.3.1. Association analysis of individual hernia phenotypes in BOLT-LMM

Association analysis of inguinal hernia yielded 24 susceptibility loci (3076 variants), 20 of which were previously unreported (all four previously associated inguinal hernia loci replicated in UK Biobank[17]) (**Table 4.1**). Conditional regression analysis[11,12] provided evidence for a further four independent signals at three inguinal hernia susceptibility loci: **2p16.1** (*EFEMP1*), **8p21.2** (*EBF2*) and **11p13** (*WT1*). In the femoral hernia analysis, 43 genome-wide significant variants clustering in a single ~10kb region at one independent locus was identified, **1q41** (lead variant rs7538503; P = 1.3×10^{-10}, OR = 1.42) (**Table 4.2**). Association analysis for umbilical hernia uncovered five novel independent susceptibility loci (277 variants) (**Table 4.3**). The statistically strongest signal was rs4846567, a predicted pathogenic, genotyped, regulatory region variant at **1q41** (*ZC3H11B*) (P = 1.7×10^{-18}, OR = 1.22, CADD = 14.9). Lastly, for the hiatus hernia association analysis, eight independent signals were discovered at eight susceptibility loci (all novel) (**Table 4.4**). Locus Zoom plots for all associations can be found in **Appendix Figure 4.1**.

In summary, 52 independent signals across 38 loci (34 novel) were found to be associated with the four individual hernia phenotypes (**Figure 4.5**). The strongest associated signal across all hernia phenotypes was rs6983815, which associated with inguinal hernia (P = 1.1×10^{-54}, OR = 1.19) at **8p21.2** (*EBF2*)— this region did not

associate with any of the other hernia phenotypes. The signal with the largest effect size was rs7538503 at **1q41** (*ZC3H11B*) ($P = 1.3×10^{-10}$, OR = 1.42), which associated with femoral hernia. The $λ_{GC}$ demonstrated nominal inflation levels across the four association analyses, ranging from 1.00-1.20 ($λ_{GC\text{-}femoral}$: 1.00; $λ_{GC\text{-}umbilical}$: 1.05; $λ_{GC\text{-}inguinal}$: 1.15; $λ_{GC\text{-}hiatus}$: 1.20), however the LDSC intercept range of 1.00-1.03 (Femoral: 1.00; Umbilical: 1.01; Inguinal: 1.02; Hiatus: 1.03) and an attenuation ratio of 0.08-0.19 (Umbilical: 0.08; Inguinal:0.11; Hiatus: 0.13; Femoral: 0.19) is fully in keeping with the effects of polygenicity and large sample size (**Appendix Figure 4.2**).[18]

Shared susceptibility between multiple hernia subtypes was seen at five loci (four loci demonstrating concordance in allelic effect directions). The most striking evidence of shared susceptibility was seen at **1q41** (near *ZC3H11B*), which associated with three of four hernia subtypes: inguinal hernia (lead variant rs2820441, $P = 6.6×10^{-13}$, OR = 1.09, EAF = 0.32(G)), femoral hernia (rs7538503 ($P = 1.3×10^{-10}$, OR = 1.42, EAF = 0.29(G)), and umbilical hernia (rs4846567, $P = 1.7×10^{-18}$, OR = 1.22, EAF = 0.31(T)). A further four loci showed overlap between two hernia subtypes. Shared susceptibility was demonstrated between inguinal and hiatus hernia at three loci: **2p16.1** (*EFEMP1*) (rs11899888, $P_{Inguinal} = 2.2×10^{-12}$, OR = 1.16, EAF = 0.16(G); rs10207635, $P_{Hiatus} = 1.3×10^{-8}$, OR = 1.07, EAF =0.13)), **6p22.2** (*MHC region*) (rs13212652, $P_{Inguinal} = 3.1×10^{-11}$, OR = 1.12, EAF=0.87(T); rs9393735, $P_{Hiatus} = 2.7×10^{-8}$, OR = 1.07, EAF = 0.86(G)), and **11p13** (*WT1*) (rs4140413, $P_{Inguinal} = 2.4×10^{-20}$, OR =1.11, EAF = 0.63(G); rs11031796, $P_{Hiatus} = 3.6×10^{-16}$, OR = 1.07, EAF = 0.62(G)). A shared susceptibility locus was also demonstrated between umbilical and hiatus hernia at locus **7q33** (*CALD1*) (rs12707188, $P_{Umbilical} = 5.0×10^{-15}$, OR = 1.19, EAF = 0.37(T); rs4728341,

$P_{Hiatus} = 3.9\times10^{-10}$, OR = 1.06, EAF = 0.55(T)), although these associations were intriguingly in opposite effect directions.

Figure 4.4: Hernia GWA study design and shared genetics analysis workflow. The three analysis approaches to characterise the shared genetic biology of hernia in UK Biobank are depicted.

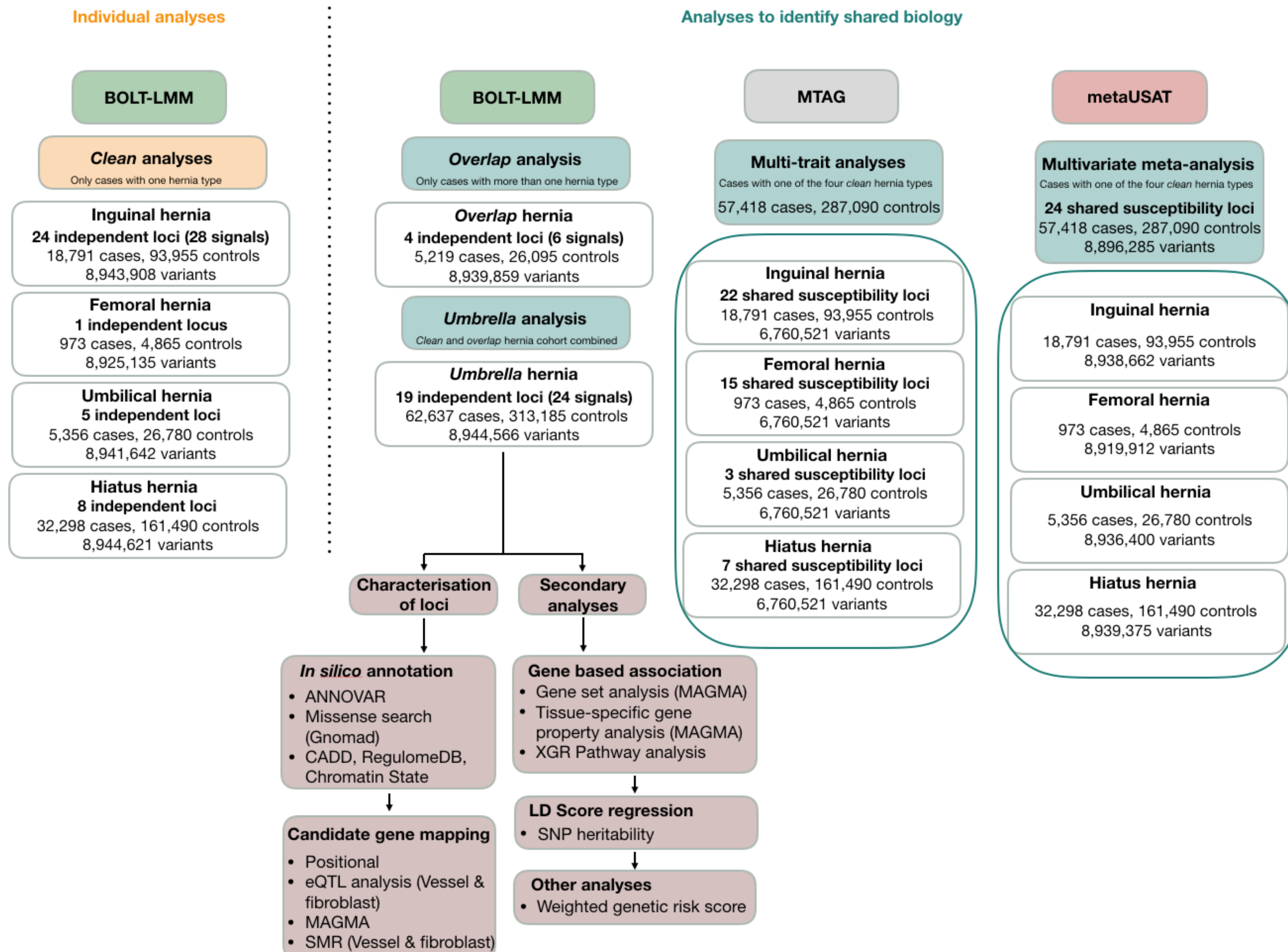


Table 4.1. Twenty-eight signals at 24 loci associated with inguinal hernia in 18791 cases and 93955 controls in UK Biobank

Chromosome	Position[a]	rsID	EA[b]	NEA[c]	EAF[d]	Info[e]	OR (95% CI)	P-value	Mapped genes[f]
1	**9443340**	**rs1106370**	**A**	**G**	**0.42**	**0.990**	**1.07 (1.04-1.09)**	1.0×10^{-8}	***SPSB1***
1	**219734960**	**rs2820441**	**C**	**A**	**0.32**	**G**	**1.09 (1.06-1.11)**	6.6×10^{-13}	**-**
2	**43665943**	**rs76684055**	**G**	**A**	**0.90**	**0.998**	**1.12 (1.08-1.16)**	2.8×10^{-10}	***THADA, ZFP36L2***
2[#]	56102744	rs11899888	G	A	0.16	0.987	1.16 (1.13-1.20)	2.2×10^{-12}	*EFEMP1, PNPT1*
2	56106928	rs59985551	C	T	0.78	0.998	1.19 (1.16-1.22)	4.7×10^{-40}	*EFEMP1, PNPT1*
2[#]	56197200	rs7564964	A	G	0.57	0.995	1.13 (1.10-1.15)	2.2×10^{-10}	-
3	**55602137**	**rs61613824**	**A**	**T**	**0.37**	**0.987**	**1.08 (1.05-1.10)**	1.1×10^{-10}	***ERC2***
3	**56141843**	**rs7647972**	**C**	**G**	**0.70**	**0.991**	**1.09 (1.06-1.11)**	8.9×10^{-12}	***CCDC66, ERC2***
3	**100297679**	**rs13083051**	**T**	**C**	**0.92**	**0.986**	**1.12 (1.08-1.17)**	2.9×10^{-8}	***TMEM45A***
4	**4949339**	**rs4330303**	**G**	**A**	**0.68**	**0.975**	**1.07 (1.04-1.09)**	2.4×10^{-8}	**-**
4	**174616174**	**rs56063997**	**C**	**T**	**0.36**	**0.988**	**1.07 (1.05-1.10)**	3.6×10^{-10}	**-**
5	64355060	rs370763	A	T	0.67	0.998	1.10 (1.08-1.13)	3.3×10^{-17}	*ADAMTS6, CWC27*
6	**6743149**	**rs1294421**	**T**	**G**	**0.40**	**G**	**1.07 (1.05-1.10)**	5.6×10^{-10}	**-**
6	**26099279**	**rs13212652**	**T**	**G**	**0.87**	**1.000**	**1.12 (1.08-1.15)**	3.1×10^{-11}	***BTN2A1, BTN3A2, C6orf15, HFE, HIST1H1A, HIST1H1B, HIST1H1C, HIST1H1T, HIST1H2AB, HIST1H2AC, HIST1H2AJ, HIST1H2AL, HIST1H2BB, HIST1H2BC, HIST1H2BL, HIST1H2BN, HIST1H3A, HIST1H3B, HIST1H3C, HIST1H3I, HIST1H3J, HIST1H4A, HIST1H4B, HIST1H4C, HIST1H4L, LRRC16A, OR2B2, PGBD1, SCGN, SLC17A1, SLC17A2, SLC17A3, SLC17A4, TRIM26, TRIM31, TRIM38, ZKSCAN3, ZKSCAN4, ZKSCAN8, ZNF165, ZNF322, ZSCAN12, ZSCAN16, ZSCAN31, ZSCAN9***

6	**32808299**	**rs45506201**	**G**	**A**	**0.90**	**0.996**	**1.11 (1.07-1.15)**	**5.6×10^{-9}**	***APOM, BRD2, C4A, C4B, C6orf10, GPANK1, HLA-DMA, HLA-DMB, HLA-DOB, HLA-DQB1, HLA-DQB2, HLA-DRA, HLA-DRB5, LSM2, LY6G5B, MICB, MSH5, MSH5-SAPCD1, NOTCH4, PSMB8, PSMB9, RNF5, SKIV2L, TAP1, TAP2, VARS, VWA7, XXbac-BPG181M17.5***
6	**45481873**	**rs62400367**	**A**	**G**	**0.85**	**0.983**	**1.09 (1.06-1.12)**	**2.9×10^{-8}**	***RUNX2***
6	**143676186**	**rs6570555**	**A**	**T**	**0.43**	**0.995**	**1.08 (1.06-1.11)**	**7.8×10^{-13}**	***AIG1***
7	**25681464**	**rs10951081**	**C**	**A**	**0.33**	**0.968**	**1.07 (1.05-1.10)**	**5.3×10^{-9}**	**-**
7	**73540726**	**rs3895707**	**C**	**T**	**0.91**	**0.987**	**1.11 (1.07-1.15)**	**2.3×10^{-8}**	***ELN, LIMK1***
8[#]	25435170	rs10481336	C	T	0.21	0.986	1.10 (1.08-1.13)	1.6×10^{-15}	*CDCA2, DOCK5, GNRH1, KCTD9*
8	25717620	rs6983815	A	T	0.41	0.996	1.19 (1.17-1.22)	1.1×10^{-54}	*EBF2*
9	**16766118**	**rs7850168**	**C**	**A**	**0.08**	**0.980**	**1.12 (1.08-1.17)**	**1.7×10^{-8}**	***BNC2***
11[#]	32350027	rs7924571	C	A	0.78	0.995	1.08 (1.06-1.11)	8.8×10^{-9}	*CCDC73, EIF3M*
11	32459228	rs4140413	G	T	0.63	0.988	1.11 (1.09-1.14)	2.4×10^{-20}	*CCDC73, EIF3M, WT1*
12	**66328027**	**rs12810758**	**T**	**C**	**0.23**	**0.991**	**1.08 (1.05-1.11)**	**3.2×10^{-9}**	***AC090673.2, HMGA2***
13	**32398964**	**rs796861335**	**C**	**CT**	**0.41**	**0.960**	**1.06 (1.04-1.09)**	**4.6×10^{-8}**	***FRY, RXFP2***
16	**84856552**	**rs4238714**	**C**	**T**	**0.42**	**0.992**	**1.09 (1.06-1.11)**	**2.8×10^{-13}**	***CRISPLD2***
17	**12191339**	**rs12453693**	**T**	**C**	**0.31**	**0.995**	**1.08 (1.06-1.11)**	**3.0×10^{-11}**	**-**

[a]Based on NCBI Genome Build 37 (hg19).
[b]The effect allele.
[c]The non-effect allele.
[d]The effect allele frequency.
[e]The SNP INFO score for imputed SNPs; G = genotyped SNP.
[f] The 101 genes prioritised at these loci based on positional mapping, eQTL mapping, MAGMA gene mapping and summary-based mendelian randomisation (see Methods).
[#]Denotes the four residual significant signals following conditional regression analysis at the lead SNP at the locus.
Bold loci are those that have not been previously reported.

Table 4.2. One previously unreported locus significantly associated with femoral hernia in 973 cases and 4865 controls in UK Biobank.

Chromosome	Position[a]	rsID	EA[b]	NEA[c]	EAF[d]	Info[e]	OR (95% CI)	P-value	Mapped genes[f]
1	219788530	rs7538503	G	A	0.29	0.995	1.42 (1.27-1.58)	1.3×10^{-10}	-

[a]Based on NCBI Genome Build 37 (hg19).
[b]The effect allele.
[c]The non-effect allele.
[d]The effect allele frequency.
[e]The SNP INFO score for imputed SNPs; G = genotyped SNP.
[f] No genes were prioritised at this loci based on positional mapping, eQTL mapping, MAGMA gene mapping and summary-based mendelian randomisation (see Methods).

Table 4.3. Five previously unreported loci significantly associated with umbilical hernia in 5,356 cases and 26,780 controls in UK Biobank.

Chromosome	Position[a]	rsID	EA[b]	NEA[c]	EAF[d]	Info[e]	OR (95% CI)	P-value	Mapped genes[f]
1	219750717	rs4846567	T	G	0.31	G	1.22 (1.17-1.28)	1.7×10^{-18}	-
2	146365492	Not available	C	CAA	0.56	0.990	1.13 (1.08-1.17)	2.5×10^{-8}	-
2	199676405	rs778276885	AT	A	0.51	0.996	1.12 (1.08-1.17)	3.9×10^{-8}	-
7	134591097	rs12707188	T	C	0.37	0.998	1.19 (1.14-1.24)	5.0×10^{-15}	*CALD1*
12	78154757	rs2887596	T	C	0.47	0.998	1.12 (1.08-1.17)	2.7×10^{-8}	-

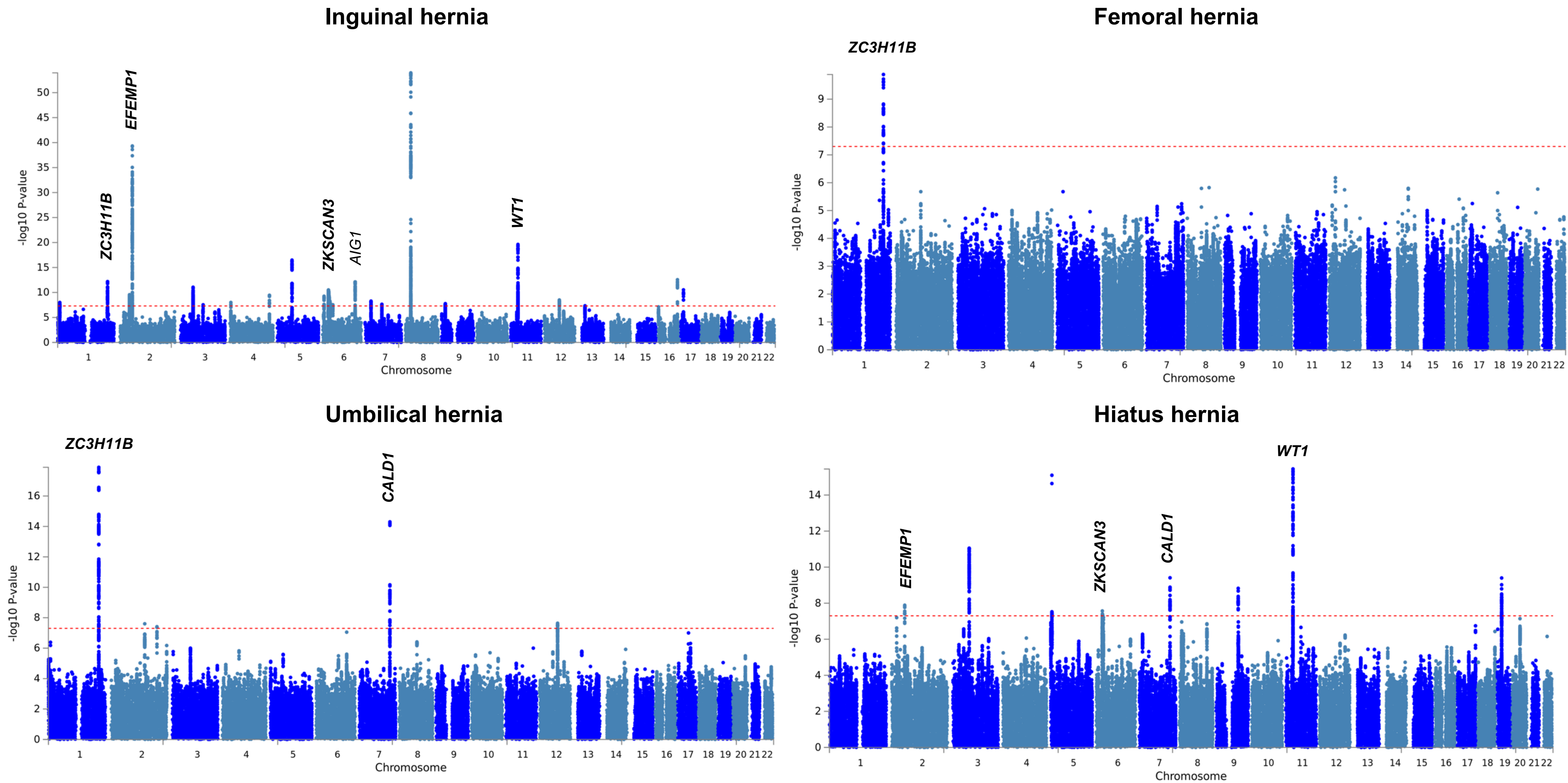
[a]Based on NCBI Genome Build 37 (hg19).
[b]The effect allele.
[c]The non-effect allele.
[d]The effect allele frequency.
[e]The SNP INFO score for imputed SNPs; G = genotyped SNP.
[f] One gene was prioritised at these loci based on positional mapping and MAGMA gene mapping (see Methods).

Table 4.4. Eight previously unreported loci significantly associated with hiatus hernia in 32,298 cases and 161,490 controls in UK Biobank.

Chromosome	Position[a]	rsID	EA[b]	NEA[c]	EAF[d]	Info[e]	OR (95% CI)	P-value	Mapped genes[f]
2	56040035	rs10207635	T	A	0.13	1.000	1.07 (1.05-1.10)	1.3×10^{-8}	*EFEMP1*
3	70920485	rs4499560	A	T	0.31	0.984	1.07 (1.05-1.08)	8.9×10^{-12}	-
5	4977446	rs42202	A	G	0.08	0.986	1.14 (1.10-1.18)	8.0×10^{-16}	-
6	26582327	rs9393735	A	G	0.86	G	1.07 (1.05-1.10)	2.7×10^{-8}	*BTN2A1, BTN3A2, HIST1H2BN, HIST1H4L, HMGN4, OR2B2, ZNF311, ZNF391*
7	134605106	rs4728341	T	C	0.55	0.965	1.06 (1.04-1.07)	3.9×10^{-10}	*CALD1*
9	96624645	rs4075733	C	T	0.46	0.996	1.05 (1.04-1.07)	1.5×10^{-9}	-
11	32479807	rs11031796	G	A	0.62	0.998	1.07 (1.06-1.09)	3.6×10^{-16}	*WT1*
19	18787981	rs2891698	G	A	0.47	0.998	1.06 (1.04-1.07)	4.0×10^{-10}	*CRTC1, KLHL26, TMEM59L, UBA52*

[a]Based on NCBI Genome Build 37 (hg19).
[b]The effect allele.
[c]The non-effect allele.
[d]The effect allele frequency.
[e]The SNP INFO score for imputed SNPs; G = genotyped SNP.
[f]The 15 genes prioritised at these loci based on positional mapping, eQTL mapping and MAGMA gene mapping (see Methods).

Figure 4.5. Manhattan plots for the four individual hernia analyses in UK Biobank. Manhattan plots are annotated with the gene names of loci that demonstrate shared susceptibility across two or more individual analyses (*Bold*). The 6q24.2 (*AIG1)* locus is plotted for inguinal hernia because it shows shared susceptibility with the overlap hernia analysis (discussed later).



4.3.2. In silico annotation of individual hernia loci

Localisation of associated variants across the four individual hernia phenotypes was performed in FUMA *SNP2GENE* v1.3.6[19]. ANNOVAR[20] identified genome-wide significant exonic variants for three hernia subtypes: inguinal, umbilical and hiatus hernia. For inguinal hernia, 11 variants were exonic— five of which were non-synonymous variants (**Appendix Table 4.2**). Most notably, rs7578597 (P = 4.20×10-8, OR = 1.12), which is in high LD with the lead variant (rs76684055) at locus 2p21 (r^2 = 0.88), resides in an evolutionarily constrained region (Genomic Evolutionary Rate Profiling (GERP) Score = 4.0) within *THADA*. rs7578597 causes a p.Thr1187Ala substitution that is predicted to have a *'deleterious' (*SIFT[21]) and *'possibly damaging'* (PolyPhen[22]) consequence on THADA protein structure and function (**Appendix Table 4.3**). Moreover, three additional intronic/intergenic variants demonstrated strong pathogenic and regulatory potential (Combined Annotation Dependent Deletion (CADD) Score[23] > 12.37, RegulomeDB (RDB)[24] = 2B or less) near the *THADA* and *SPSB1* loci (all in high linkage with lead variant (r^2 > 0.89)). For both umbilical and hiatus hernia, one missense variant was identified which significantly associated with both hernia types: rs6973420 ($P_{Umbilical} = 2.70\times10^{-10}$, OR = 1.14 (G); $P_{Hiatus} = 2.40\times10^{-8}$, OR = 1.05 (A)), which resides within exon 5 of *CALD1*, resulting in a p.His397Arg substitution that is predicted to have a '*tolerated (low confidence)'* effect on protein function (SIFT[21]) (**Appendix Table 4.2**). Aside from exonic variants, one intronic variant significantly associated with umbilical hernia was identified near *CALD1,* rs12532492 ($P_{Umbilical} = 1.4\times10^{-8}$, OR = 1.13, CADD = 16.02, RDB = 1F, r^2_{index} = 0.64) which suggests strong functionality (**Appendix Table 4.3**). Moreover, for hiatus hernia, one non-coding variant, rs1522552 ($P = 3.40\times10^{-11}$, OR = 1.06, CADD = 19.71, RDB = 2B) was suggested to have strong functionality at locus 3p13, and is in high

LD with the lead variant at this locus (rs4499560, $r^2 = 0.96$) (**Appendix Table 4.3**). For femoral hernia, no exonic variants were identified; however, three potentially robust functional intergenic variants were recognised within a 47kb cluster (all in high LD with the index variant) at the 1q41 locus near *ZC3H11B*: rs2785986 ($P = 1.90\times10^{-8}$, OR= 1.34, $r^2_{index} = 0.66$, CADD = 14.7), rs4846567 ($P = 3.30\times10^{-9}$, OR = 1.38, $r^2_{index} = 0.83$, CADD = 14.9), and rs2820443 ($P = 2.20\times10^{-9}$, OR = 1.38, $r^2_{index} = 0.85$, CADD = 13.0) (**Appendix Table 4.3**).

4.3.3. Candidate gene mapping of individual hernia loci

Associated signals for the four hernia types were mapped to putative protein-coding genes using four mapping strategies (as previously described in **Section 4.2.9**). In total, 101 unique genes were mapped to 18 of the 24 inguinal hernia associated loci by positional mapping[19] (n = 53), eQTL mapping in GTEx[25] v8 Cells Transformed Fibroblast and Skeletal Muscle Tissue (n = 42), MAGMA[26] (n = 64), and Summary-based Mendelian Randomisation (SMR)[27] using GTEx v7 Cells Transformed Fibroblast and Skeletal Muscle Tissue (n = 3) (**Appendix Tables 4.4, 4.5 and 4.6**). Overlap was demonstrated across the mapping strategies with 29 genes mapped by two or more strategies, and four genes mapped by three mapping strategies (*THADA*, *EFEMP1*, *HLA-DMA*, *CRISPLD2*) (**Appendix Figure 4.3)**. For femoral hernia, no genes were prioritised at the **1q41** susceptibility locus, although the index signal rs7538503 ($P = 1.7\times10^{-18}$, OR = 1.22, CADD = 14.9) is a regulatory region variant narrowly residing outside the positional mapping window from *ZC3H11B (*~11.25kb downstream). For umbilical hernia, one gene (*CALD1*) was prioritised at the **7q33** locus by both positional and MAGMA mapping (**Appendix Table 4.7)**. For hiatus hernia, 15 unique genes were prioritised at five of eight genome-wide significant susceptibility loci by position (n = 5), eQTL (n = 4), and MAGMA mapping (n = 11) (**Appendix Tables 4.8 and 4.9**)., with five putative genes mapped by more than one approach (*BTN3A2*, *CALD1*, *WT1*, *KLHL26*, *CRTC1*) (**Appendix Figure 4.4)**. In summary, 117 unique genes were prioritised using the four mapping strategies at 12 of 38 susceptibility loci associated with the four individual hernia phenotypes.

4.3.4. SNP-based heritability of individual hernia phenotypes

The contribution of common variants to hernia risk for each of the individual GWAS were estimated using LD Score regression.[18] Using LD scores from ~1.2 million common low-LD variants from each of the studies, the SNP-based heritability (h^2g) for each individual hernia phenotype was calculated in UK Biobank for the first time. The largest h^2g was seen for femoral hernia at 12.79% (S.E. = 7.85%), followed by umbilical hernia which had a h^2g of 9.80% (1.67%). Interestingly, the association analyses which yielded the largest number of susceptibility signals— inguinal hernia and hiatus hernia, had the lowest SNP-heritability at 7.43% (0.83%) and 5.06% (0.34%).

4.3.5. Genetic risk score for the individual hernia phenotypes

The lead signals from all four individual hernia analyses in UK Biobank were used to calculate a weighted genetic risk score (wGRS) for each individual hernia phenotype (**Table 4.5**). As expected, the wGRS for all four hernia cases was higher than in controls (inguinal hernia: 3.070 vs 2.967 ($P = 5.53\times10^{-322}$); hiatus hernia: 0.455 vs 0.442 ($P = 9.54\times10^{-79}$); umbilical hernia: 0.650 vs 0.599 ($P = 2.02\times10^{-50}$); femoral hernia: 0.710 vs 0.565 (unweighted) ($P = 1.24\times10^{-9}$). Furthermore, it was also found that all hernia cases (across all four hernia phenotypes) that had undergone surgery had a higher wGRS compared to hernia cases that had not undergone surgery (inguinal hernia: 3.072 vs 3.013 ($P = 5.46\times10^{-6}$); femoral hernia: 0.739 vs 0.597 (*unweighted*) ($P = 6.69\times10^{-3}$); hiatus hernia: 0.464 vs 0.454 ($P = 4.77\times10^{-3}$); umbilical hernia: 0.653 vs 0.630 ($P = 2.13\times10^{-2}$).

Table 4.5. Genetic risk score for the four individual hernia phenotypes in the UK Biobank cohort.

Inguinal hernia

Group	Inguinal hernia cases	Controls	P-value†	Inguinal hernia cases with operation code	Inguinal hernia cases without operation code	P-value§
N	18,791	93,955		18,082	709	
Mean wGRS (standard deviation)	3.070 (0.332)	2.967 (0.334)	5.53×10^{-322}	3.072 (0.332)	3.013 (0.337)	5.46×10^{-6}

Femoral hernia

Group	Femoral hernia cases	Controls	P-value†	Femoral hernia cases with operation code	Femoral hernia cases without operation code	P-value§
N	973	4,865		774	199	
Mean GRS (standard deviation)	0.710 (0.681)	0.565 (0.636)	1.24×10^{-9}	0.739 (0.688)	0.597 (0.642)	6.69×10^{-3}

Umbilical hernia

Group	Umbilical hernia cases	Controls	P-value†	Umbilical hernia cases with operation code	Umbilical hernia cases without operation code	P-value§
N	5,356	26,780		4,749	607	
Mean wGRS (standard deviation)	0.650 (0.228)	0.599 (0.223)	2.03×10^{-50}	0.653 (0.228)	0.630 (0.229)	2.13×10^{-2}

Hiatus hernia

Group	Hiatus hernia cases	Controls	P-value†	Hiatus hernia cases with operation code	Hiatus hernia cases without operation code	P-value§
N	32,398	161,490		1,311	30,987	
Mean wGRS (standard deviation)	0.455 (0.115)	0.442 (0.114)	9.54×10^{-79}	0.464 (0.115)	0.454 (0.115)	4.77×10^{-3}

*wGRS: weighted genetic risk score. GRS: unweighted genetic risk score (for femoral hernia analysis only). †Unpaired two-tailed t-test between hernia cases and controls. §Unpaired two-tailed t-test between hernia cases with an operation code and hernia cases without an operation code.

4.3.6. Combined association analysis of hernia phenotypes in BOLT-LMM

To understand the shared genetic architecture of hernia risk, a second set of association analyses were performed in BOLT-LMM[11,12]: **i) Overlap hernia**: all participants in UK Biobank with diagnostic or operative codes for two or more hernia subtypes; **ii) Umbrella hernia**: a complete set (umbrella) of hernia cases from all four individual hernia cohorts (described in **Section 4.3.1**) and the overlapping hernia (overlap) cohort, i.e. all participants in UK Biobank with diagnostic or operative codes for at least one hernia subtype.

Overlap hernia analysis

5,219 cases with more than one hernia subtype and 26,095 matched controls in UK Biobank that were not included in the individual hernia analyses were tested for association. Association analysis across 8,939,859 imputed/genotyped common variants yielded four genome-wide significant loci (516 variants), with conditional regression analysis identifying a further two signals at locus **2p16.1** (*EFEMP1*) (**Table 4.6**). Regional Locus Zoom Plots for all six overlap hernia signals are provided in **Appendix Figure 4.5**. The three strongest association signals were seen at locus **2p16.1** (*EFEMP1*) (rs1346786 ($P = 7.6\times10^{-20}$, OR = 1.24)), **1q41** (*ZC3H11B*) (rs1415287 ($P = 1.2\times10^{-16}$, OR = 1.21)), and locus **11p13** (*WT1*) (rs3858458, $P = 8.4\times10^{-13}$, OR = 1.17), all of which demonstrated shared susceptibility with individual hernia loci (**Figure 4.6**). The **1q41** (*ZC3H11B*) locus was associated with inguinal, femoral and umbilical hernia individually, with the **2p16.1** (*EFEMP1*) and **11p13** (*WT1*) loci being significantly associated in the individual analyses with inguinal and hiatus hernia. The final significant overlap hernia locus, **6q24.2** (*AIG1*) (rs4896643, P =

3.6×10^{-8}, OR = 1.12) only demonstrated a significant association with inguinal hernia in the individual association analysis (rs6570555, $P_{Inguinal} = 7.8\times10^{-13}$, OR = 1.08).

Umbrella hernia analysis

A combined cohort (umbrella) of 62,637 hernia cases with individual and overlap hernia phenotypes and 313,185 matched controls was tested for association in UK Biobank. The final umbrella analysis was performed across 375,822 participants and 8,420,566 imputed (MAF ≥ 0.01, INFO Imputation score ≥ 0.90) and 524,000 genotyped SNPs (MAF ≥ 0.01), and discovered 25 independent signals at 19 genome-wide significant loci (4785 variants) (**Table 4.7**) (Regional Locus Zoom Plots provided in **Appendix Figure 4.6**). Nine loci were not previously discovered in the individual or overlap hernia analyses (**Figure 4.6**), and the new top locus was **1q41** (*TGFB2*) (rs2799098, $P = 9.3\times10^{-15}$, OR = 1.06). Five of the 10 previously associated loci (in the individual and/or overlap cohort) demonstrated significant overlap in the umbrella analysis. The most striking overlap was seen at the **1q41** (*ZC3H11B*) locus (rs2820441, $P_{Umbrella} = 2.7\times10^{-23}$, OR = 1.07) which coincided with four previous hernia phenotypes (inguinal, femoral, umbilical, and overlap hernia). The **11p13** (*WT1*) locus was the strongest associated signal (rs66798575, $P = 1.6\times10^{-40}$, OR= 1.09) and showed overlap with three of the previous phenotypes (inguinal, hiatus and overlap hernia). This was followed by the **2p16.1** (*EFEMP1*) locus, which also showed commonality with three phenotypes (inguinal, hiatus and overlap hernia), and was the second most significant umbrella signal (rs75439645, $P = 2.4\times10^{-38}$, OR = 1.12). Two umbrella loci overlapped with two previous hernia phenotypes: the **6p22.2** (*MHC*) locus (rs28360634, $P_{Umbrella} = 1.7\times10^{-17}$, OR = 1.09) which overlapped with inguinal

and hiatus individual analyses and the **6q24.2** locus (*AIG1*) (rs6917403, $P_{Umbrella}$ = 2.9×10^{-12}, OR = 1.04) which overlapped with inguinal and overlap hernia analyses.

Across both overlap and umbrella hernia analyses, the λ_{GC} was 1.05 and 1.20, respectively, with an LDSC intercept of 1.01 and 1.03 and an attenuation ratio of 0.15 and 0.10 (**Appendix Figures 4.7 and 4.8)**.[18] These findings are consistent with the individual analyses performed in BOLT-LMM.

Table 4.6. Six signals (four loci) significantly associated with overlap hernia in 5,219 cases and 26,095 controls in UK Biobank

Chromosome	Position[a]	rsID	EA[b]	NEA[c]	EAF[d]	Info[e]	OR (95% CI)	P-value	Mapped genes[f]
1	219742537	rs1415287	T	C	0.31	0.998	1.21 (1.16-1.26)	1.2×10^{-16}	-
2[#]	56040099	rs10199082	C	T	0.14	G	1.29 (1.22-1.37)	1.9×10^{-10}	*EFEMP1*
2	56108333	rs1346786	C	T	0.71	0.994	1.24 (1.18-1.29)	7.6×10^{-20}	*EFEMP1*
2[#]	56194773	rs981037	T	C	0.58	0.993	1.21 (1.16-1.27)	1.0×10^{-9}	-
6	143670001	rs4896643	C	G	0.45	0.993	1.12 (1.08-1.17)	3.6×10^{-8}	*AIG1*
11	32484594	rs3858458	C	T	0.63	0.981	1.17 (1.12-1.22)	8.4×10^{-13}	*WT1*

[a]Based on NCBI Genome Build 37 (hg19).
[b]The effect allele.
[c]The non-effect allele.
[d]The effect allele frequency.
[e]The SNP INFO score for imputed SNPs; G = genotyped SNP.
[f]The four genes prioritised at these loci based on positional mapping, eQTL mapping and MAGMA gene mapping (see Methods). [#]Denotes the two residual significant signals following conditional regression analysis at the lead SNP at the locus.

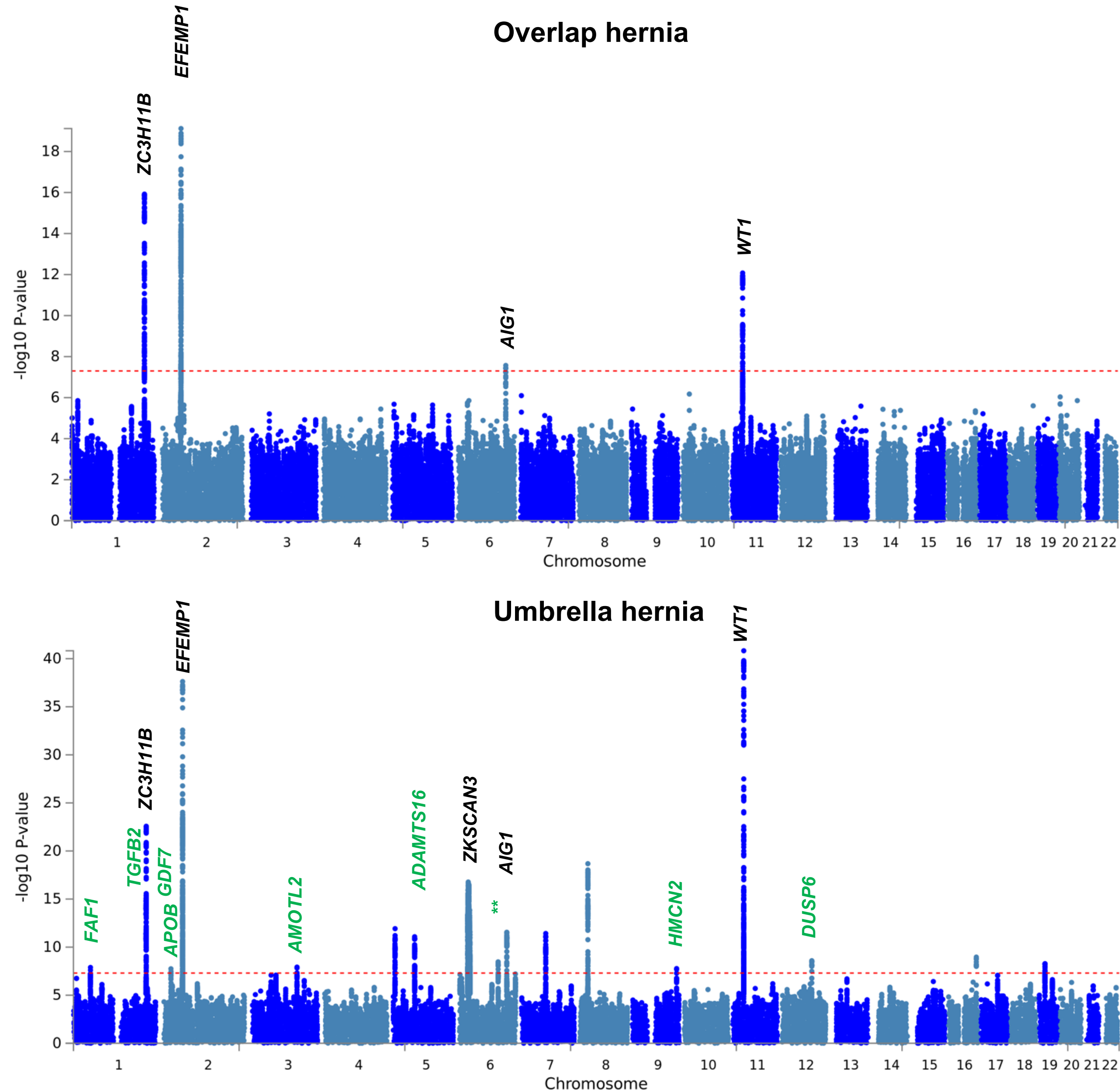
Table 4.7. Twenty-five signals at 19 loci significantly associated in the umbrella hernia analysis

Chr	Position[a]	rsID	EA[b]	NEA[c]	EAF[d]	Info[e]	OR (95% CI)	P-value	Mapped genes[f]	Significant in individual or overlap GWAS[g]
1	51477643	rs13376700	A	T	0.43	0.992	1.04 (1.02-1.05)	1.3×10^{-8}	*CDKN2C, EPS15, FAF1, NRD1*	-
1	218521609	rs2799098	A	G	0.82	G	1.06 (1.05-1.08)	9.3×10^{-15}	*RRP15, TGFB2*	-
1	219734960	rs2820441	C	A	0.32	G	1.07 (1.05-1.08)	2.7×10^{-23}	-	*IH,FH,UH,OH*
2	20878406	rs3072	C	T	0.36	0.994	1.04 (1.02-1.05)	1.8×10^{-8}	*C2orf43, GDF7*	-
2	21239884	rs76622701	A	T	0.56	0.979	1.04 (1.02-1.05)	3.5×10^{-8}	*APOB*	-
2	56048944	rs75439645	A	G	0.13	0.999	1.12 (1.10-1.14)	2.4×10^{-38}	*CCDC104, EFEMP1, PNPT1, SMEK2*	*IH,HH,OH*
2[#]	56106928	rs59985551	C	T	0.78	0.998	1.09 (1.08-1.11)	5.1×10^{-9}	*CCDC104, CLHC1, EFEMP1, PNPT1, RTN4, SMEK2*	*IH,HH,OH*
2[#]	56193665	rs13431149	C	A	0.60	0.991	1.07 (1.05-1.08)	2.5×10^{-23}	*CCDC104, PNPT1, SMEK2*	*IH,HH,OH*
3	134372486	rs9883955	G	T	0.63	G	1.04 (1.02-1.05)	1.2×10^{-8}	*AMOTL2, ANAPC13, CEP63, EPHB1, KY*	-
5[#]	4881885	rs570260	G	A	0.34	G	1.03 (1.02-1.05)	6.2×10^{-10}	-	*HH*
5[#]	4977446	rs42202	A	G	0.08	0.986	1.08 (1.06-1.11)	1.7×10^{-11}	-	*HH*
5[#]	5145100	rs1834922	G	A	0.35	0.999	1.04 (1.02-1.05)	9.0×10^{-10}	*ADAMTS16*	
5	5350637	rs7715383	C	G	0.10	0.970	1.08 (1.06-1.10)	1.2×10^{-12}	-	
5	64355060	rs370763	A	T	0.67	0.998	1.05 (1.03-1.06)	8.3×10^{-12}	*ADAMTS6*	*IH*
6	27332891	rs28360634	T	C	0.89	1.000	1.09 (1.07-1.11)	1.7×10^{-17}	*ABT1, APOM, APOM, BTN1A1, BTN2A1, BTN2A2, BTN3A1, BTN3A2, BTN3A3, C4A, C4B, C6orf15, C6orf48, CCHCR1, CLIC1, DDR1, DPCR1, HCG27, HFE, HIST1H1A, HIST1H1B, HIST1H2AG, HIST1H2AI, HIST1H2AJ, HIST1H2AK, HIST1H2AL, HIST1H2AM, HIST1H2BC, HIST1H2BF,*	*IH,HH*

									HIST1H2BJ, HIST1H2BL, HIST1H2BM, HIST1H2BN, HIST1H2BO, HIST1H3C, HIST1H3H, HIST1H3I, HIST1H3J, HIST1H4A, HIST1H4J, HIST1H4K, HIST1H4L, HLA-A, HLA-B, HLA-C, HLA-DMA, HLA-DMB, HLA-DRA, HMGN4, LRRC16A, LSM2, MSH5, MSH5-SAPCD1, NKAPL, OR12D3, OR2B2, OR2B6, PBX2, PGBD1, POM121L2, POU5F1, PRRC2A, PRSS16, PSORS1C1, RNF5, SCAND3, SFTA2, SLC17A1, SLC17A2, SLC17A3, SLC17A4, TRIM26, TRIM27, TRIM31, TRIM38, TRIM39, TRIM39-RPP21, TUBB, VARS, VWA7, ZFP57, ZKSCAN3, ZKSCAN4, ZKSCAN8, ZNF165, ZNF184, ZNF192P1, ZNF322, ZNF391, ZSCAN12, ZSCAN16, ZSCAN23, ZSCAN31, ZSCAN9	
6	117507982	rs200889152	C	A	0.38	0.991	1.04 (1.03-1.05)	3.4×10^{-9}	-	-
6	143653287	rs6917403	A	G	0.42	0.987	1.04 (1.03-1.06)	2.9×10^{-12}	*AIG1*	*IH,OH*
7[#]	73445942	rs2356532	G	A	0.06	0.996	1.09 (1.06-1.11)	1.1×10^{-8}	*ELN*	*IH*
7	73474825	rs17855988	G	C	0.90	0.963	1.08 (1.05-1.10)	3.8×10^{-12}	*ELN, LIMK1*	*IH*
8	25693744	rs4368985	T	A	0.40	0.997	1.06 (1.05-1.07)	2.1×10^{-19}	*EBF2*	*IH*
9	133038387	rs9299329	G	A	0.50	0.979	1.04 (1.02-1.05)	1.7×10^{-8}	*HMCN2*	-
11	32451920	rs66798575	T	G	0.64	0.973	1.09 (1.08-1.10)	1.6×10^{-40}	*CCDC73, EIF3M, WT1*	*IH, HH,OH*
12	89767237	rs797267	G	A	0.19	0.996	1.05 (1.03-1.06)	2.6×10^{-9}	*DUSP6*	-
16	84855477	rs1874013	G	T	0.38	0.994	1.04 (1.03-1.05)	1.1×10^{-9}	*CRISPLD2*	*IH*
19	18824038	rs34482977	C	G	0.81	0.992	1.05 (1.03-1.06)	5.3×10^{-9}	*CRLF1, CRTC1, KLHL26, SSBP4*	*HH*

[a]Based on NCBI Genome Build 37 (hg19).
[b]The effect allele.
[c]The non-effect allele.
[d]The effect allele frequency.
[e]The SNP INFO score for imputed SNPs; G = genotyped SNP.
[f]The 138 genes prioritised at these loci based on positional mapping, eQTL mapping, MAGMA gene mapping and summary-based Mendelian randomisation (see Methods).
[g]where IH = Inguinal Hernia Individual, FH = Femoral Hernia Individual, UH = Umbilical Hernia Individual, HH = Hiatus Hernia Individual, OH = overlapping hernia analysis.
[#]Denotes the six residual significant signals following conditional regression analysis at the lead SNP at the locus

Figure 4.6. Manhattan plots of the combined hernia analyses performed in BOLT-LMM. Manhattan plots are annotated with the gene names of loci that demonstrate shared susceptibility across two or more individual analyses (*black*). For the umbrella plot, the nine loci that were not discovered in the individual or overlap analyses are highlighted with the gene name in green *(or ** where no gene was prioritised at this locus).*



4.3.7. In silico annotation of combined hernia loci

Associated variants in the overlap and umbrella analyses were annotated within FUMA *SNP2GENE* v1.3.6.[19] 187 genome-wide significant ($P < 5\times10^{-8}$) candidate SNPs were identified by FUMA to be in LD ($r^2 > 0.6$) with the lead variant at each of the four overlap hernia loci. No exonic variants were localised, however, six intronic-intergenic variants had predicted deleterious effects and were in high LD with the index variant at each overlap hernia loci, including three at locus **1q41** (*ZC3H11B*) (rs4846567 ($P = 1.40\times10^{-16}$, $r^2_{index} = 0.99$, CADD[23] = 14.9); rs2820443 ($P = 3.40\times10^{-16}$, $r^2_{index} =0.99$, CADD = 13.0); rs2785986 ($P = 1.60\times10^{-15}$, $r^2_{index} = 0.76$, CADD = 14.7), and two at locus **2p16.1** (*EFEMP1*) (rs3791679 ($P = 1.40\times10^{-17}$, $r^2_{index} = 0.66$, CADD =17.8) and rs7422809 ($P = 1.30\times10^{-14}$, $r^2_{index} = 0.68$, CADD=15.2) (**Appendix Table 4.10**).

For the umbrella association analysis, FUMA identified 877 genome-wide significant candidate SNPs in LD with the lead SNP at the 19 umbrella susceptibility loci. 18 high-LD exonic variants were discovered (**Appendix Table 4.11**), 15 of which resided at the MHC locus (**6p22.2**) — of these, four variants resulted in amino acid substitutions that were predicted to have damaging (PolyPhen[22]) and deleterious (SIFT[21]) consequences on *BTN2A1* (rs13195401, rs13195402) and *OR2B2* (rs34788973, rs61742093). Of the non-MHC exonic variants, rs17855988 (P=3.80E-12, OR=1.07, $r^2_{index} = 1.0$, CADD[23] = 25.9, RDB[24] = 2B), results in a p.Gly581Arg substitution in the 25th exon of *ELN* that is predicted by SIFT to have a deleterious (low confidence) consequence on elastin function. Predicted functional intronic and intergenic variants associated with umbrella hernia are provided in **Appendix Table 4.12**.

4.3.8. Candidate gene mapping of combined hernia loci

Genes were prioritised at the overlap and umbilical hernia associated loci using four mapping strategies (identical to individual hernia). In summary, three unique genes were mapped at the three of four of the overlap hernia associated loci, with two genes (*WT1* and *EFEMP1*) being prioritised by more than one mapping approach (**Appendix Tables 4.13 and 4.14; Appendix Figure 4.9**). A total of 129 unique genes were mapped to 20 of the 25 signals from the umbrella hernia analysis (**Appendix Tables 4.15 and 4.16**), with 38 genes being mapped by two strategies, and seven genes showing overlap across three mapping strategies (*EFEMP1*, *ADAMTS16*, *BTN2A1*, *HIST1H2BN*, *HMGN4*, *ZSCAN31*, *CRISPLD2*) (**Appendix Table 4.17; Appendix Figure 4.10)**.

4.3.9. Gene set, pathway and tissue enrichment analysis of combined hernia loci

Overlap hernia: Gene-set analysis performed in MAGMA v1.07[26] delineated the convergence of MAGMA mapped genes within 15,496 gene sets (5500 curated gene sets and 9995 GO terms) from MSigDB v8.0.[28] Two biological processes gene sets were significantly enriched: *'Negative regulation of cell proliferation in kidney development'* (P = 5.76×10^{-7}, n = 5 genes) and *'Diaphragm development'* (P = 3.00×10^{-6}, n = 9 genes) (**Appendix Table 4.18)**. Gene set enrichment analysis in FUMA *GENE2FUNC*, identified all three prioritised genes to cluster with GWAS Catalog reported genes related to pulse pressure. Moreover, XGR analysis enriched two canonical pathways[29]: *'Regulation of Telomerase'* (P = 3.80×10^{-4}, Z = 4.88, FDR = 1.0×10^{-3}, *WT1*) and *'Genes encoding structural ECM glycoproteins'* (P = 3.2×10^{-3}, Z = 2.7, FDR = 5.4×10^{-3}, *EFEMP1*) (**Appendix Table 4.19)**.

Umbrella hernia: MAGMA[26] gene-set analysis enriched 29 gene sets from MSigDB that met Bonferroni-correction (0.05/15496) (**Appendix Table 4.20**). The top biological processes gene ontology was *'Connective tissue development'* (P = 8.05×10^{-9}, n = 262 genes); the top curated gene set was *'Elastic fibre formation'* (P = 3.29×10^{-8}, n = 46 genes), and the top molecular functions gene ontology was *'BMP receptor binding'* (3.36×10^{-8}, n = 8 genes). Of note, enriched biological processes gene sets included *'Skeletal system development*' (P = 3.28×10^{-7}, n = 498 genes) and *'Thorax and anterior abdomen determination'* (P = 1.42×10^{-6}, n = 5 genes). Moreover, tissue expression analysis in MAGMA[26] using GTEx[25] v8.0 54 specific tissue types demonstrated adipose visceral omentum (P = 6.77×10^{-4}, most enriched) and adipose subcutaneous tissues to be enriched (P = 1.11×10^{-3}, 4th most enriched) (**Appendix**

Figure 4.11). GTEx v8.0 30 general tissue types analysis demonstrated Adipose tissue to be the most enriched tissue ($P = 6.31 \times 10^{-4}$) (**Appendix Figure 4.11**).

4.3.10. SNP-based heritability of combined hernia phenotypes

Using common, low-LD variants from the GWAS summary statistics, the narrow-sense SNP-based heritability (h^2g) for the overlap hernia phenotype in UK Biobank was 10.51% (1.85%), and for the umbrella hernia phenotype it was 3.35% (0.24%).

4.3.11. Genetic risk score to look for evidence of shared biology between hernia subtypes

As with the individual phenotypes, for both overlap and umbrella hernia analyses, a wGRS was constructed from the lead independent variants from the association analysis (**Table 4.8**). Again, as expected, both overlap (1.093 vs 1.005) and umbrella hernia cases (1.354 vs 1.325) had a higher wGRS than controls ($P = 1.17 \times 10^{-65}$ and $P = 4.20 \times 10^{-298}$, respectively). Moreover, as with the individual analyses, overlap cases that had undergone surgery (1.096 vs 1.036) and umbrella cases that had undergone surgery (1.366 vs 1.343) had a higher wGRS than cases that had not ($P = 3.98 \times 10^{-3}$ and $P = 4.87 \times 10^{-55}$, respectively).

For each of the four hernia types, the wGRS was compared between participants who only had that hernia type (individual) with those who had more than one hernia type (overlap) (**Table 4.9**). Across three of the four hernia phenotypes, the individual cases had a higher wGRS than for overlap cases: inguinal (3.070 vs 3.067, $P = 5.9 \times 10^{-1}$), femoral (0.710 vs 0.708, $P = 9.7 \times 10^{-1}$) and umbilical (0.650 vs 0.642, $P = 1.8 \times 10^{-11}$). However, for hiatus hernia (which was the best powered analysis), overlap cases had a higher wGRS than for individual cases (0.459 vs 0.455, $P = 1.70 \times 10^{-2}$).

Table 4.8. Weighted genetic risk score for combined hernia

Overlap hernia

Group	Overlap hernia cases	Controls	P-value†	Overlap hernia cases with operation code	Overlap hernia cases without operation code	P-value§
N	5,219	26,095		4,941	278	
Mean wGRS* (standard deviation)	1.093 (0.336)	1.005 (0.332)	1.17×10^{-65}	1.096 (0336)	1.036 (0.338)	3.98×10^{-3}

Umbrella hernia

Group	Umbrella hernia cases	Controls	P-value†	Umbrella hernia cases with operation code	Umbrella hernia cases without operation code	P-value§
N	62,637	313,185		29,857	32,780	
Mean wGRS* (standard deviation)	1.354 (0.180)	1.325 (0.179)	4.20×10^{-298}	1.366 (0.180)	1.343 (0.180)	4.87×10^{-55}

*wGRS: weighted genetic risk score. †Unpaired two-tailed t-test between hernia cases and controls. §Unpaired two-tailed t-test between hernia cases with an operation code and hernia cases without an operation code.

Table 4.9. Weighted genetic risk score comparing cases with a single hernia with those with multiple hernias

Inguinal hernia

Group	Inguinal hernia individual cases	Inguinal hernia overlap cases	P-value†
N	18,791	4,216	
Mean wGRS* (standard deviation)	3.070 (0.332)	3.067 (0.332)	5.85×10^{-1}

Femoral hernia

Group	Femoral hernia individual cases	Femoral hernia overlap cases	P-value†
N	973	605	
Mean wGRS* (standard deviation)	0.710 (0.681)	0.708 (0.705)	9.67×10^{-1}

Umbilical hernia

Group	Umbilical hernia individual cases	Umbilical hernia overlap cases	P-value†
N	5,356	2,076	
Mean wGRS* (standard deviation)	0.650 (0.228)	0.642 (0.227)	1.81×10^{-1}

Hiatus hernia

Group	Hiatus hernia individual cases	Hiatus hernia overlap cases	P-value†
N	32,398	3,841	
Mean wGRS* (standard deviation)	0.455 (0.115)	0.459 (0.115)	1.70×10^{-2}

*wGRS: weighted genetic risk score. †Unpaired two-tailed t-test between hernia cases with only one hernia type and hernia cases with a particular hernia type and at least one more.

4.3.12. Multi-trait analysis of the individual hernia phenotypes in MTAG to uncover shared genetic biology

To delineate the shared genetic biology between the four individual hernia phenotypes in greater detail, multi-trait analysis was performed in MTAG across 6,760,521 of the ~9M SNPs from each of the individual hernia phenotypes.[14] The final MTAG analysis consisted of 32,298 hiatus hernia cases, 18,791 inguinal hernia cases, 5,356 umbilical hernia cases, and 973 femoral hernia cases (total 57,418 individual hernia cases) and 287,090 matched controls from UK Biobank. Forty-seven loci were discovered across the four MTAG multi-trait analyses (22 inguinal (**Table 4.10**); 15 femoral (**Table 4.11**); 3 umbilical (**Table 4.12**); 7 hiatus (**Table 4.13**); Manhattan plots provided in **Figure 4.7**; Regional Locus Zooms Plots provided in **Appendix Figure 4.12**). Locus **2p16.1** (*EFEMP1*) and **11p13** (*WT1*) were significant across all four MTAG analyses (**Figure 4.8)**, and locus **1q41** (*ZC3H11B*) and **6p22.2** (*MHC*) were found to be significant in three of four analyses.

Of the twenty-two loci significantly associated with inguinal hernia in MTAG, 14 were more significant under multi-trait analysis (than any of the four individual phenotypes alone), providing evidence for shared genetics (**Table 4.10**). This includes locus **1q41** (*TGFB2*) which was sub-threshold across the four individual hernia phenotypes ($P_{Inguinal} = 3\times10^{-7}$, $P_{Femoral} = 8.8\times10^{-1}$, $P_{Umbilical} = 7.2\times10^{-2}$, $P_{Hiatus} = 2.7\times10^{-5}$), but became significant under multi-trait analysis (rs3121580, $P_{MTAG} = 3.41\times10^{-8}$), with pairwise testing confirming the signal at this locus to be contributed most significantly by inguinal and hiatus hernia (**Figure 4.9)**. Of note, **1q41** (*TGFB2*) was also discovered as a new locus in the femoral hernia MTAG analysis (rs2799098, P_{MTAG} =

4.66×10^{-8}; **Appendix Figure 4.13**) (with pairwise analysis demonstrating that both inguinal and femoral hernia contribute to the signal), as well as in the umbrella analysis, where it was the top associated signal (rs2799098, $P_{Umbrella} = 9.3 \times 10^{-15}$). Of the 15 loci discovered to associate with femoral hernia in MTAG (**Table 4.11**), 14 (bar **1q41** (*TGFB2*)) significantly associated with inguinal hernia in the individual analysis .

For the remaining two hernia phenotypes, no new loci became significant under multi-trait MTAG analysis that were not discovered in either of the four individual analyses.[14] For umbilical hernia, three loci were discovered under multi-trait analysis, with locus **1q41** (*ZC3H11B*) (rs4846567, $P_{MTAG} = 8.85 \times 10^{-24}$) becoming more significant under multi-trait analysis than for any individual traits (**Table 4.12**). Two loci were newly discovered to associate with umbilical hernia under multi-trait analysis (**2p16.1** (*EFEMP1*) and **11p13** (*WT1*)), both of which previously associated with inguinal and hiatus hernia in the individual analyses. Lastly, of the seven loci associated with hiatus hernia under multi-trait analysis (**Table 4.13**), three loci became more significant under MTAG than in the individual hiatus analysis alone (**2p16.1** (*EFEMP1*), **6p22.1** (MHC region), and **11p13** (*WT1*)) -- with **11p13** (*WT1*) becoming more significant under multi-trait analysis than across any of the individual individual analyses.

Table 4.10. Twenty-two loci significantly associated with inguinal hernia in the MTAG multi-trait analysis of 57,418 cases and 287,090 controls in UK Biobank. Statistically significant signals from the inguinal hernia MTAG analysis are shown in the left-hand column. The central column shows the association p-values for those SNPs in the individual hernia analyses, with the direction of effect indicated by a + or – sign. Pairwise testing was performed for MTAG signals that were more statistically significant than in any of the individual analyses (right-hand column). Colour code: Green = MTAG more significant than individual analysis, Red = MTAG less significant, Blue = new locus not previously associated with an individual trait.

Inguinal hernia MTAG								Four individual traits					Pairwise analysis between individual traits (P-value)		
Chr	**Pos**	**rsID**	**EA**	**BETA**	**SE**	**P**	**Candidate genes**	**IH**	**FH**	**UH**	**HH**	**BETA Direction**	**FH and IH**	**UH and IH**	**HH and IH**
1	9446761	rs12134602	G	0.00865193	0.00153976	1.92E-08	*SPSB1*	1.2E-08	8.6E-01	2.7E-01	8.7E-01	+-++	-	-	-
1	218492121	rs3121580	T	-0.0110874	0.00200888	3.41E-08	*(TGFB2, RRP15)*	3.0E-07	8.8E-01	7.2E-02	2.7E-05	----	4.92E-07	8.08E-08	**1.12E-10**
1	219734960	rs2820441	A	-0.0143895	0.00160964	3.91E-19	- *(ZC3H11B)*	6.6E-13	2.7E-09	2.0E-15	4.5E-01	----	1.03E-16	**7.95E-24**	5.44E-07
2	43665943	rs76684055	G	0.01513517	0.00248842	1.19E-09	*THADA, ZFP36L2*	2.8E-10	9.4E-01	2.3E-01	1.2E-01	+--+	-	-	-
2	56106928	rs59985551	C	0.02461083	0.00180193	1.81E-42	*EFEMP1*	4.7E-40	1.0E-02	2.2E-02	7.2E-02	++++	**2.01E-41**	1.97E-37	1.53E-21
3	55585396	rs4271886	C	-0.0103582	0.00157803	5.24E-11	*ERC2*	1.2E-10	1.6E-01	7.0E-01	5.4E-01	----	**4.55E-11**	4.26E-09	9.15E-06
3	56139250	rs4974167	A	0.00988327	0.00163651	1.55E-09	*ERC2*	1.3E-11	2.0E-01	3.6E-01	2.2E-02	+---	-	-	-
3	100297679	rs13083051	T	0.01570352	0.00281045	2.30E-08	*GPR128, NIT2, TMEM45A*	2.9E-08	3.7E-01	1.5E-02	1.8E-01	+-++	1.99E-07	**1.58E-09**	8.87E-06

4	4965364	rs6446301	C	-0.0090194	0.00163458	3.43E-08	-	2.6E-08	7.6E-01	8.4E-01	4.8E-01	----	-	-	-
4	174616174	rs56063997	T	-0.0093771	0.00156848	2.25E-09	-	3.6E-10	3.2E-01	1.4E-01	3.4E-01	-+-+	-	-	-
5	64351146	rs422529	G	-0.0133784	0.00159385	4.71E-17	- *(ADAMTS6)*	5.7E-17	3.7E-01	5.6E-01	3.2E-01	----	**6.34E-17**	1.67E-14	3.28E-09
6	6743149	rs1294421	T	0.0093173	0.00153569	1.30E-09	-	5.6E-10	8.0E-01	6.4E-01	4.4E-01	+-++	-	-	-
6	26099279	rs13212652	T	0.01550491	0.00223625	4.11E-12	*HFE, HIST1H1A, HIST1H2AB, HIST1H3B, HIST1H4B, SCGN, SLC17A1, SLC17A2, SLC17A3, TRIM38*	3.1E-11	2.6E-01	9.7E-01	7.9E-05	++-+	1.79E-11	4.74E-09	**6.56E-13**
6	32343236	rs2894251	A	0.01309947	0.00215632	1.24E-09	*C2, BTNL2, C6orf10, HLA-DQA1, HLA-DQB1, HLA-DRA, NOTCH4, PSMB8, TAP2, TNXB*	1.3E-08	8.3E-02	9.4E-01	3.4E-04	++-+	3.25E-09	6.02E-07	**2.84E-10**
6	143653287	rs6917403	A	0.01063646	0.00152724	3.30E-12	*AIG1*	9.9E-13	6.3E-02	1.5E-02	1.3E-01	+-++	-	-	-
7	25681464	rs10951081	C	0.0090193	0.00016197	2.57E-08	-	5.3E-09	9.7E-02	1.3E-03	9.5E-01	++--	-	-	-

7	73422593	rs76027228	C	0.01692608	0.0026912 7	3.19E-10	*ELN*	2.4E-08	5.5E-02	8.4E-04	2.0E-02	++++	4.75E-09	**8.81E-11**	1.46E-07
8	25707778	rs10746560	A	-0.0225574	0.00152869	2.82E-49	*EBF2*	1.5E-54	4.7E-01	3.8E-01	4.0E-01	-+++	-	-	-
11	32459228	rs4140413	G	0.01611576	0.00156696	8.26E-25	*WT1*	2.4E-20	2.1E-03	5.2E-02	1.7E-13	++++	3.46E-22	1.16E-19	**1.69E-30**
12	66328027	rs12810758	C	-0.0104211	0.00179918	6.95E-09	*AC090673.2, HMGA2*	3.2E-09	5.4E-01	5.2E-01	2.1E-01	-+--	-	-	-
16	84856552	rs4238714	T	-0.0113588	0.00153157	1.20E-13	*CRISPLD2*	2.8E-13	4.5E-01	1.0E-01	6.4E-01	----	**3.08E-13**	5.35E-13	1.31E-06
17	12191339	rs12453693	C	-0.0102141	0.00162295	3.10E-10	-	3.0E-11	6.0E-01	8.5E-01	8.1E-01	-+++	-	-	-

[a]Based on NCBI Genome Build 37 (hg19).
[b]The effect allele.
[c]The non-effect allele.
[d]The effect allele frequency.
[e]The SNP INFO score for imputed SNPs; G = genotyped SNP.
[f]Genes mapped to these loci based on positional mapping in FUMA (see Methods).

Table 4.11. Fifteen loci significantly associated with femoral hernia in the MTAG multi-trait analysis of 57,418 cases and 287,090 controls in UK Biobank. Statistically significant signals from the femoral hernia MTAG analysis are shown in the left-hand column. The central column shows the association p-values for those SNPs in the individual hernia analyses, with the direction of effect indicated by a + or – sign. Pairwise testing was performed for MTAG signals that were more statistically significant than in any of the individual analyses (right-hand column). Colour code: Green = MTAG more significant than individual analysis, Red = MTAG less significant, Blue = new locus not previously associated with an individual trait.

Femoral hernia MTAG								Four Individual Traits					Pairwise analysis between individual traits (P-value)		
Chr	**Pos**	**rsID**	**EA**	**BETA**	**SE**	**P**	**Mapped genes**	**IH**	**FH**	**UH**	**HH**	**BETA Direction**	**IH and FH**	**UH and FH**	**HH and FH**
1	218521609	rs2799098	G	-0.0166534	0.00304786	4.66E-08	*TFGB2, RRP15*	1.1E-07	9.9E-01	9.6E-02	8.7E-06	-+--	**2.20E-07**	0.126445	1.21E-05
1	219734960	**rs2820441**	A	-0.0224528	0.00248037	1.40E-19	- *(ZC3H11B)*	6.6E-13	2.7E-09	2.0E-15	4.5E-01	----	1.03E-16	**6.79E-22**	0.078315
2	43451957	rs149290349	G	0.02435397	0.00426067	1.09E-08	*PLEKHH2, THADA, ZFP36L2*	3.5E-09	4.5E-01	3.2E-01	3.5E-02	++-+	-	-	-
2	56106928	rs59985551	C	0.03669374	0.00282008	1.05E-38	*EFEMP1*	4.7E-40	1.0E-02	2.2E-02	7.2E-02	++++	-	-	-
3	55585396	rs4271886	C	-0.0152931	0.00243215	3.22E-10	*ERC2*	1.2E-10	1.6E-01	7.0E-01	5.4E-01	----	-	-	-
5	64351146	rs422529	G	-0.0193603	0.00245173	2.87E-15	- *(ADAMTS6)*	5.7E-17	3.7E-01	5.6E-01	3.2E-01	----	-	-	-
6	6740366	rs1294414	A	0.01331708	0.00238437	2.33E-08	-	1.2E-09	9.0E-01	6.4E-01	3.7E-01	++++	-	-	-

6	26099472	rs35402046	G	0.02334313	0.00341065	7.69E-12	*HFE, HIST1H1A, HIST1H2AB, HIST1H3B, HIST1H4B, SCGN, SLC17A1, SLC17A2, SLC17A3, TRIM38*	4.6E-11	2.3E-01	1.0E+00	2.4E-05	++++	**2.35E-11**	0.632349	1.29E-05
6	143625532	rs7383094	G	0.01477797	0.00237303	4.74E-10	*AIG1*	5.2E-12	3.3E-01	4.9E-02	4.1E-02	+-++	-	-	-
7	25681464	rs10951081	C	0.0137399	0.00250771	4.28E-08	-	5.3E-09	9.7E-02	1.3E-03	9.5E-01	++--	-	-	-
7	73422593	rs76027228	C	0.02588356	0.00419955	7.12E-10	*ELN*	2.4E-08	5.5E-02	8.4E-04	2.0E-02	++++	**4.75E-09**	1.33E-04	0.008968
8	25706115	rs11135895	A	-0.0315961	0.00236289	8.84E-41	*EBF2*	2.1E-54	5.1E-01	3.8E-01	4.4E-01	-+++	-	-	-
11	32459228	rs4140413	G	0.02491545	0.00238989	1.90E-25	*WT1*	2.4E-20	2.1E-03	5.2E-02	1.7E-13	++++	**3.46E-22**	0.002738	7.03E-15
16	84856552	rs4238714	T	-0.0162622	0.00235335	4.84E-12	*CRISPLD2*	2.8E-13	4.5E-01	1.0E-01	6.4E-01	----	-	-	-
17	12191339	rs12453693	C	-0.0142353	0.00251839	1.58E-08	-	3.0E-11	6.0E-01	8.5E-01	8.1E-01	-+++	-	-	-

[a]Based on NCBI Genome Build 37 (hg19).
[b]The effect allele.
[c]The non-effect allele.
[d]The effect allele frequency.
[e]The SNP INFO score for imputed SNPs; G = genotyped SNP.
[f]Genes mapped to these loci based on positional mapping in FUMA (see Methods).

Table 4.12. Three loci significantly associated with umbilical hernia in the MTAG multi-trait analysis of 57,418 cases and 287,090 controls in UK Biobank. Statistically significant signals from the umbilical hernia MTAG analysis are shown in the left-hand column. The central column shows the association p-values for those SNPs in the individual hernia analyses, with the direction of effect indicated by a + or – sign. Pairwise testing was performed for MTAG signals that were more statistically significant than in any of the individual analyses (right-hand column). Colour code: Green = MTAG more significant than individual analysis, Red = MTAG less significant, Blue = new locus not previously associated with an individual trait.

Umbilical hernia MTAG								Four Individual Traits					Pairwise analysis between individual traits (P-value)		
Chr	Pos	rsID	EA	BETA	SE	P	Mapped genes	IH	FH	UH	HH	BETA Direction	IH and UH	FH and UH	HH and UH
1	219750717	rs4846567	G	-0.0279967	0.00278473	8.85E-24	- *(ZC3H11B)*	9.3E-12	3.3E-09	1.7E-18	6.5E-01	----	3.99E-24	**3.22E-25**	1.65E-04
2	56048944	rs75439645	G	-0.028864	0.00381218	3.69E-14	*EFEMP1*	3.6E-21	2.3E-03	8.4E-04	1.7E-08	----	-	-	-
11	32459228	rs4140413	G	0.01830575	0.00268096	8.61E-12	*WT1*	2.4E-20	2.1E-03	5.2E-02	1.7E-13	++++	-	-	-

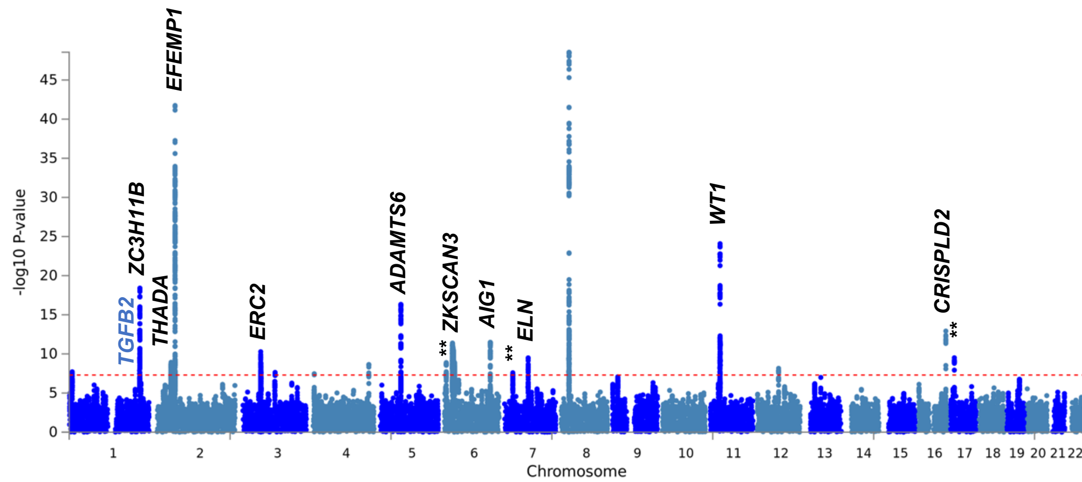
[a]Based on NCBI Genome Build 37 (hg19).
[b]The effect allele.
[c]The non-effect allele.
[d]The effect allele frequency.
[e]The SNP INFO score for imputed SNPs; G = genotyped SNP.
[f]Genes mapped to these loci based on positional mapping in FUMA (see Methods).

Table 4.13. Seven loci significantly associated with hiatus hernia in the MTAG multi-trait analysis of 57,418 cases and 287,090 controls in UK Biobank. Statistically significant signals from the hiatus hernia MTAG analysis are shown in the left-hand column. The central column shows the association p-values for those SNPs in the individual hernia analyses, with the direction of effect indicated by a + or – sign. Pairwise testing was performed for MTAG signals that were more statistically significant than in any of the individual analyses (right-hand column). Colour code: Green = MTAG more significant than individual analysis, Red = MTAG less significant, Blue = new locus not previously associated with an individual trait.

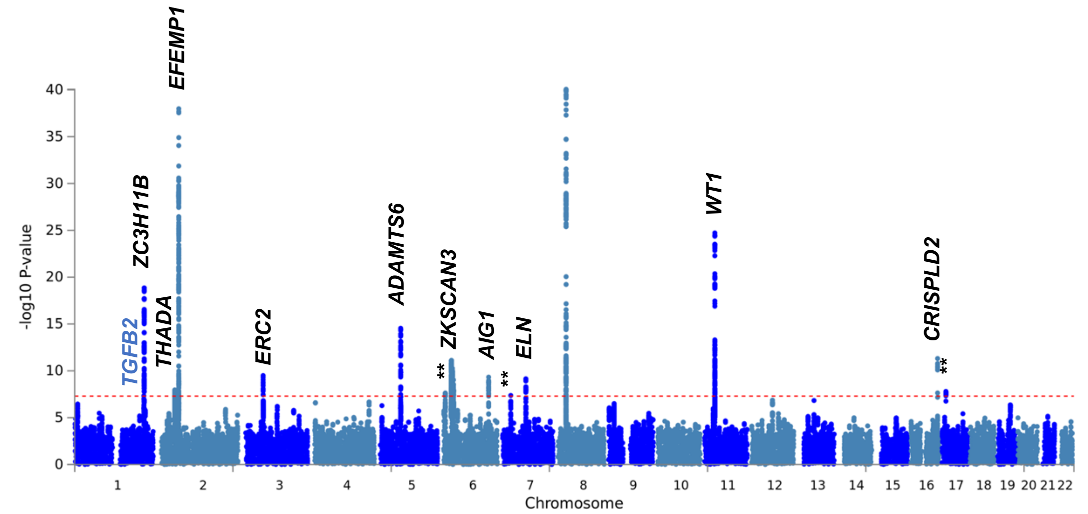
Hiatus hernia MTAG								Four individual traits					Pairwise analysis between individual traits (P-value)		
Chr	Pos	rsID	EA	BETA	SE	P	Candidate Genes	IH	FH	UH	HH	BETA Direction	IH and HH	FH and HH	UH and HH
2	56048944	rs75439645	G	-0.0118023	0.00175903	1.95E-11	*EFEMP1*	3.6E-21	2.3E-03	8.4E-04	1.7E-08	----	-	-	-
3	70951405	rs6805430	C	-0.0087604	0.00128815	1.04E-11	*FOXP1*	1.9E-01	1.2E-02	9.0E-02	1.0E-11	+---	-	-	-
5	4977446	rs42202	A	0.01758426	0.0022514	5.70E-15	-	8.4E-01	9.4E-02	5.7E-01	8.0E-16	-+-+	-	-	-
6	28825573	rs3132387	G	0.01135776	0.00192355	3.54E-09	*HIST1H1B, HIST1H3I, OR12D2, OR12D3, OR2B2, OR2B6, PGBD1, ZKSCAN4, ZSCAN12, ZSCAN9*	6.8E-10	4.9E-01	3.4E-01	5.6E-08	++++	-	-	-
9	96624645	rs4075733	C	0.00706618	0.00119858	3.74E-09	-	5.9E-01	1.4E-01	4.6E-01	1.5E-09	--++	-	-	-
11	32479807	rs11031796	G	0.01073813	0.00122954	2.47E-18	*WT1*	2.1E-15	9.7E-04	4.6E-01	3.6E-16	++++	**1.36E-29**	8.75E-18	5.22E-15
19	18787981	rs2891698	G	0.00744858	0.00119487	4.55E-10	*CRLF1, CRTC1, KLHL26, TMEM59L*	6.2E-01	5.5E-01	4.8E-01	4.0E-10	++++	-	-	-

Figure 4.7. Manhattan plots for the inguinal, femoral, umbilical and hiatus hernia MTAG multi-trait analysis summary statistics. Manhattan plots are annotated with the gene names of loci that demonstrate shared susceptibility across two or more MTAG analyses (*black gene names* or ** where no genes are proximate to the lead signal). Blue gene names are those loci that were previously not discovered in any of the individual hernia analyses (i.e. new loci).

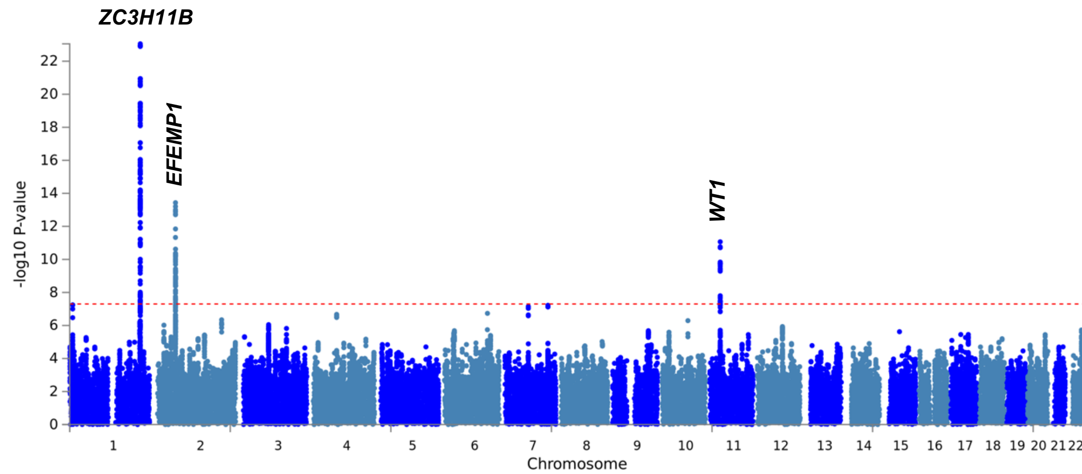
Inguinal hernia (*MTAG*)



Femoral hernia (*MTAG*)



Umbilical hernia (*MTAG*)



Hiatus hernia (*MTAG*)

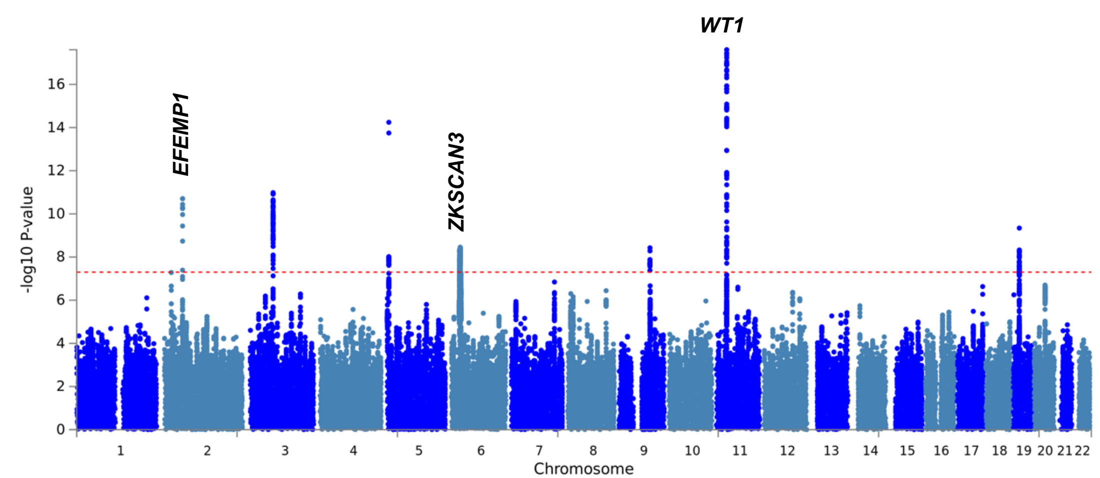
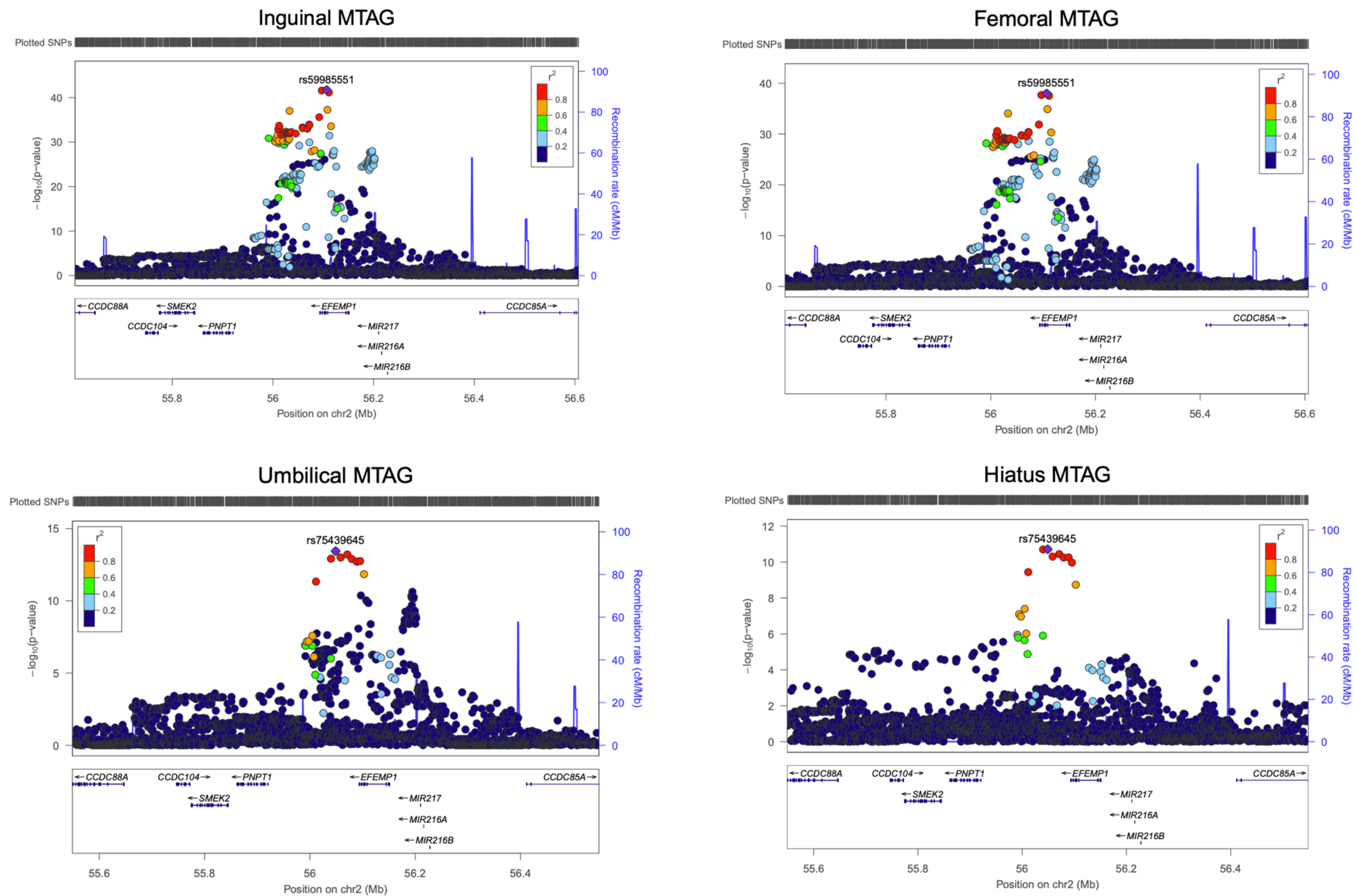
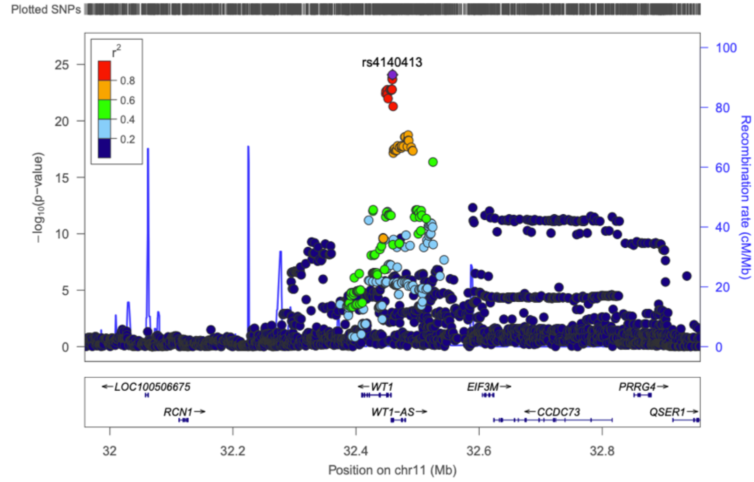


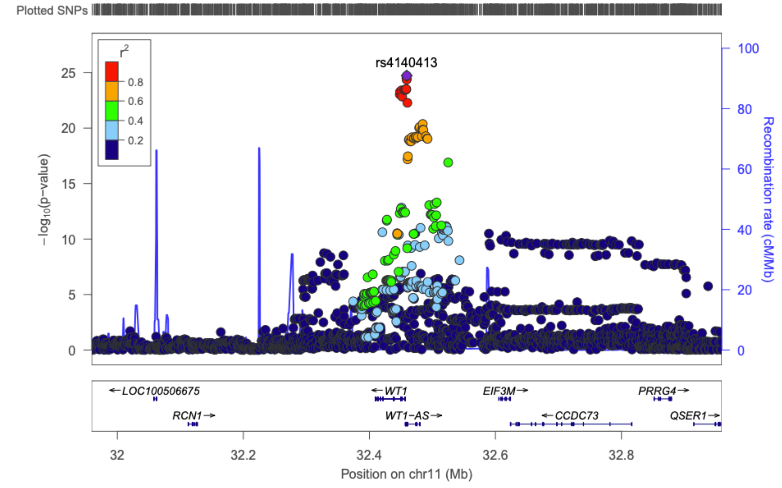
Figure 4.8. Two loci consistently significant across the four MTAG analyses.

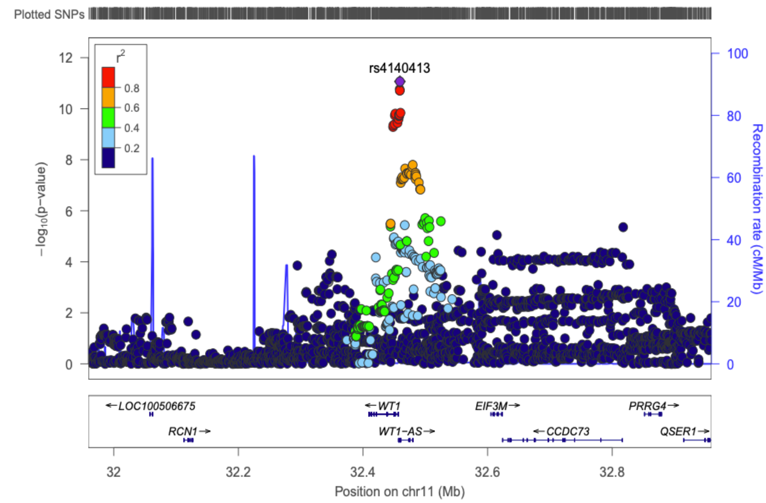
***2p16.1** (EFEMP1)*



11p13 *(WT1))*







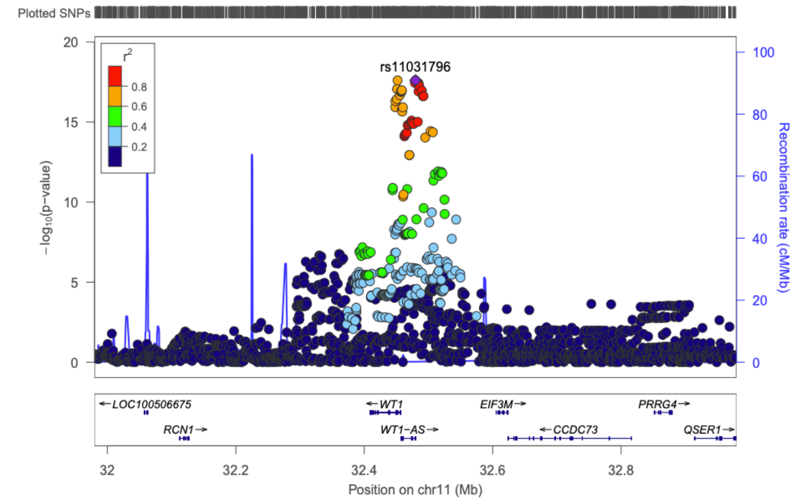
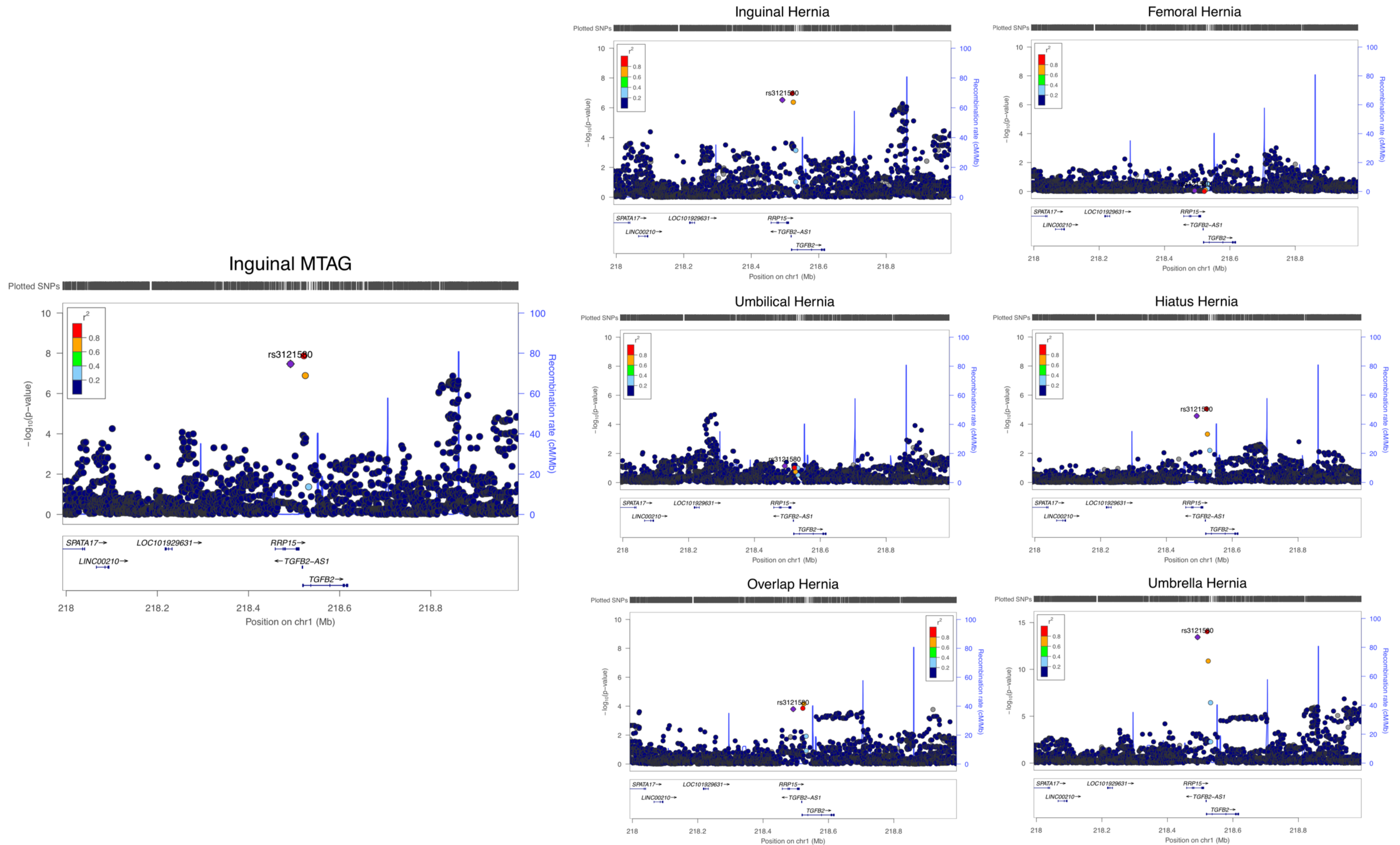


Figure 4.9. New 1q41 (*TGFB2*) locus discovered to associate with inguinal hernia through MTAG multi-trait analysis.



4.3.13. Multivariate meta-analysis of the individual hernia phenotypes in metaUSAT to uncover shared genetic biology

Multivariate meta-analysis of the four individual hernia traits was performed in metaUSAT[16], across a total of 57,418 individual hernia cases and 287,090 matched controls in UK Biobank. Twenty-four susceptibility loci (3645 variants) were discovered genome-wide significant ($P_{metaUSAT}$ < 5×10^{-8}), with one-third of metaUSAT loci (**1q41** (*ZC3H11B*); **2p16.1** (*EFEMP1*), **3p14.3** (*ERC2*); **3p13**; **5p15.33**; **6q24.2** (*AIG1*); **7p15.2** (*LOC646588*); **7q33** (*CALD1*)) becoming more significant than in any of the previous six BOLT-LMM analyses (4 x individual, overlap, and umbrella hernia), providing robust evidence of shared genetics (**Table 4.14; Figure 4.9**). Furthermore, six loci ((***1q41*** *(TGFB2),* **2p24.1** (*GDF7*), **5p15.33** (near *CEP72*), **5p15.32** (near *ADAMTS16*), **7q33** (*CALD1*), **12q21.33** (*DUSP6*)) were sub-threshold across the four individual analyses, and became significant under meta-analysis.

Intriguingly, the meta-analysis approach yielded an entirely new putative locus, **5p15.33** (rs72703080, $P_{metaUSAT}$ = 3.68×10^{-8}, ~20kb upstream from *CEP72*), which was sub-threshold across all six BOLT-LMM analyses and the four MTAG multi-trait analyses (**Figure 4.10**). Furthermore, for the majority of loci (n = 15), the metaUSAT association was more significant than all four MTAG analyses; importantly, metaUSAT discovered five loci that MTAG did not: **2p24.1** (*GDF7*), **5p15.33** (near *CEP72*), **5p15.32** (near *ADAMTS16*), **7q33** (*CALD1*), **12q21.33** (*DUSP6*) (the umbrella analysis was able to discover all loci except the new metaUSAT locus **5p15.33**) (**Table 4.14**).

As well as discovering new putative loci, metaUSAT demonstrated strong enrichment for loci that demonstrated significant overlap in the previous analyses.[16] This includes six loci that showed the greatest overlap with the individual and overlap susceptibility loci (**Figure 4.11**). This includes all five loci that imparted shared susceptibility to two or more hernia phenotypes in the individual analyses (discussed in **Section 4.3.1**): **1q41** (*ZC3H11B*); **2p16.1** (*EFEMP1*); **6p22.2** (*ZKSCAN3* (MHC region)); **11p13** (*WT1*); and **7q33** (*CALD1*) (note: three of these (**1q41** (*ZC3H11B*), **2p16.1** (*EFEMP1*), and **11p13** (*WT1*)) were also significant in the overlap analysis)**.** The sixth locus that metaUSAT enriched which showed the highest overlap in previous analyses was locus (**6q24.2** (*AIG1*) (**Table 4.14**), which happens to be the fourth and final overlap hernia locus and was also associated with inguinal hernia (discussed in **Section 4.3.6)** .

Table 4.14. 24 genome-wide significant loci discovered in the metaUSAT multi-trait meta-analysis of inguinal, femoral, umbilical, hiatus hernia in 57,418 cases and 287,090 controls in UK Biobank. Statistically significant signals from the metaUSAT analysis are shown in the left-hand column. The central column shows the association p-values for those SNPs in the BOLT-LMM analyses, with the direction of effect indicated by a + or – sign across the six BOLT analyses. The metaUSAT signals were compared against for their relevant association strength in the MTAG analysis. (right-hand column). Candidate genes are those selected from the prioritised genes (using the four mapping strategies described previously for all BOLT-discovered loci) or genes in proximity as identified within the UCSC genome browser.

metaUSAT analysis							BOLT-LMM analyses							MTAG analyses (P-value)				Candidate Gene
							Four individual hernias (P-value)				Combined Hernia (P-value)		BETA Direction[f]					
Chr[a]	BP[a]	SNP	A1[b]	A0[c]	T-statistic[d]	P-value[e]	IH	FH	UH	HH	OH	Umbrella		IH	FH	UH	HH	
1	218521609	rs2799098	G	A	3.53E-10	**5.18E-10**	1.10E-07	9.90E-01	9.60E-02	8.70E-06	1.40E-04	**9.30E-15**	-+----	**1.35E-08**	**4.66E-08**	1.14E-05	7.76E-07	*TGFB2*
1	219754012	rs559230165	C	CT	8.23E-33	**3.28E-34**	**1.50E-11**	**2.20E-09**	**1.30E-18**	7.90E-01	**1.70E-15**	**1.90E-21**	------	—#	—#	—#	—	*ZC3H11B*
2	20878406	rs3072	T	C	2.23E-08	**3.48E-08**	1.70E-02	9.40E-01	9.00E-01	6.30E-08	3.00E-02	**1.80E-08**	--+---	6.21E-03	3.61E-03	2.15E-02	5.24E-08	*GDF7*
2	43665943	rs76684055	G	A	6.96E-09	**1.07E-08**	**2.80E-10**	9.40E-01	2.30E-01	1.20E-01	8.30E-01	3.80E-05	+--+++	**1.19E-09**	**1.19E-08**	2.14E-01	6.36E-02	*THADA*
2	56106928	rs59985551	C	T	1.10E-39	**2.04E-41**	**4.70E-40**	1.00E-02	2.20E-02	**7.20E-02±**	**8.30E-18**	**2.70E-33**	++++++	**1.81E-42**	**1.05E-38**	**9.23E-11**	1.80E-03	*EFEMP1*
3	56141843	rs7647972	C	G	1.71E-11	**6.97E-12**	**8.90E-12**	1.20E-01	4.90E-01	1.10E-02	2.70E-01	5.70E-02	+---++	—#	—#	—	—	*ERC2*
3	70951945	rs5007038	A	T	7.78E-12	**3.07E-12**	1.80E-01	9.90E-03	9.30E-02	**9.60E-12**	7.30E-01	1.80E-07	+-----	—	—	—	—#	- (~70kb from *FOXP1*)
4	174606591	rs12649191	T	C	1.17E-08	**1.81E-08**	**6.20E-10**	2.60E-01	1.60E-01	2.90E-01	2.00E-03	2.90E-04	-+-+--	**4.82E-09**	4.60E-07	8.48E-03	6.72E-01	- (~300kb *HAND-AS1*)
5	595238	rs72703080	A	G	2.35E-08	**3.68E-08**	4.40E-01	4.30E-01	5.00E-04	1.70E-07	7.40E-01	7.30E-03	+-+---	5.61E-01	9.41E-01	1.15E-01	3.55E-06	- (~20kb from *CEP72*)
5	4977446	rs42202	A	G	1.44E-15	**2.71E-14**	8.40E-01	9.40E-02	5.70E-01	**8.00E-16**	8.40E-04	**1.90E-11**	-+-+++	4.93E-01	1.31E-01	3.80E-02	**5.70E-15**	- (~100kb from *ADAMTS16*)
5	5350637	rs7715383	G	C	1.97E-09	**2.97E-09**	5.00E-05	7.00E-01	4.80E-02	2.30E-07	1.40E-01	**1.20E-12**	------	—	—	—	—	- (~25kb from *ADAMTS16*)
5	64355060	rs370763	T	A	5.98E-15	**2.84E-14**	**3.30E-17**	4.60E-01	6.00E-01	2.80E-01	7.70E-06	**8.30E-12**	------	—#	—#	—	—	*ADAMTS6*

6	27352750	rs71559024	G	A	3.17E-14	**3.64E-14**	**2.20E-10**	8.80E-01	3.30E-01	**8.10E-08±**	2.80E-03	**2.10E-17**	++++++	**2.04E-11**	**7.33E-11**	1.10E-05	**5.27E-09**	*ZKSCAN3*
6	143676186	rs6570555	A	T	8.89E-13	**3.42E-13**	**7.80E-13**	4.70E-02	1.80E-02	1.80E-01	**2.20E-07±**	**1.00E-11**	+-++++	—#	—#	—	—	*AIG1*
7	25681464	rs10951081	C	A	1.32E-09	**1.98E-09**	**5.30E-09**	9.70E-02	1.30E-03	9.50E-01	6.00E-01	1.40E-02	++--++	**2.57E-08**	**4.28E-08**	2.64E-01	9.60E-01	- (*LOC646588*)
7	73422593	rs76027228	C	T	3.32E-10	**4.83E-10**	**2.40E-08**	5.50E-02	8.40E-04	2.00E-02	7.20E-03	**8.60E-12**	++++++	**3.19E-10**	**7.12E-10**	8.74E-08	1.67E-03	*ELN*
7	134593511	rs4472440	C	G	9.20E-20	**1.54E-20**	3.10E-01	9.90E-01	**7.10E-15**	**1.30E-08**	1.10E-01	6.10E-02	---+++	—	—	—	—	*CALD1*
8	25717620	rs6983815	T	A	7.26E-52	**3.47E-54**	**1.10E-54**	4.10E-01	4.30E-01	4.40E-01	3.10E-04	**1.00E-18**	-+++--	—#	—#	—	—	*EBF2*
9	96624645	rs4075733	C	T	2.45E-09	**3.71E-09**	5.90E-01	1.40E-01	4.60E-01	**1.50E-09**	2.50E-01	3.20E-04	--++-+	8.31E-01	9.65E-01	1.81E-02	**3.74E-09**	- (~50kb from *BARX1*)
11	32458278	rs5030123	G	GT GT	3.73E-32	**1.60E-33**	**2.00E-19**	6.30E-03	1.40E-01	**7.70E-16**	**1.20E-12**	**1.50E-41**	++++++	—#	—#	—#	—#	*WT1*
12	89767237	rs797267	A	G	2.25E-08	**3.51E-08**	1.00E-02	5.60E-04	4.20E-01	1.70E-06	3.00E-01	**2.60E-09**	------	3.74E-04	2.89E-05	1.92E-03	4.31E-07	*DUSP6*
16	84856552	rs4238714	T	C	1.73E-11	**7.05E-12**	**2.80E-13**	4.50E-01	1.00E-01	6.40E-01	4.70E-04	**1.60E-09**	------	**1.20E-13**	**4.84E-12**	2.19E-04	2.19E-01	*CRISPLD2*
17	12191339	rs12453693	C	T	4.73E-09	**7.22E-09**	**3.00E-11**	6.00E-01	8.50E-01	8.10E-01	6.10E-02	1.20E-04	-+++--	**3.10E-10**	**1.58E-08**	1.00E-01	8.06E-01	-
19	18787981	rs2891698	G	A	6.85E-10	**1.02E-09**	6.20E-01	5.50E-01	4.80E-01	**4.00E-10**	1.00E-01	**1.60E-08**	++++++	2.44E-01	1.10E-01	5.31E-03	**4.55E-10**	*KLHL26*

[a]*Based on NCBI Genome Build 37 (hg19).*
[b]*The reference allele.*
[c]*The alternate allele.*
[d]*The metaUSAT test statistic (scalar)*
[e]*The p-value of association based on the metaUSAT statistic*
fThe effect size direction in the six BOLT-LMM association analyses (IH, UH, FH, HH, OH, Umbrella) with respect to the reference allele
[g]*Genes were selected based on those genes mapped in the six BOLT-LMM analyses, and subsequently prioritised based on the existing literature.*

Bold P-values are those variants identified by metaUSAT that are genome-wide significant ($P < 5\times10^{-8}$) in a particular analysis
±Denotes three loci where the reference metaUSAT SNP is not significant in the individual *BOLT-LMM hernia analysis, however the locus contains genome-wide significant SNP associations.*
__Denotes loci where the metaUSAT identified variant was not available in the MTAG Analysis
 __# Denotes loci where the metaUSAT identified variant was not available in the MTAG Analysis, however the locus was significant in the analysis.

Green P-values depict those that are the most significant across the three analysis approaches (metaUSAT, BOLT-LMM. MTAG).
The following shorthand notations are used: Inguinal Hernia, IH; Femoral Hernia, FH; Umbilical Hernia, UH; Hiatus Hernia, HH; Overlap Hernia, OH; Umbrella Hernia, Umbrella.

Figure 4.9. 24 loci discovered to confer shared hernia susceptibility after multivariate meta-analysis in 57,418 cases and 287,090 controls in metaUSAT. Each metaUSAT locus is annotated according to whether it was genome-wide significant in the individual, overlap or umbrella analyses (legend).

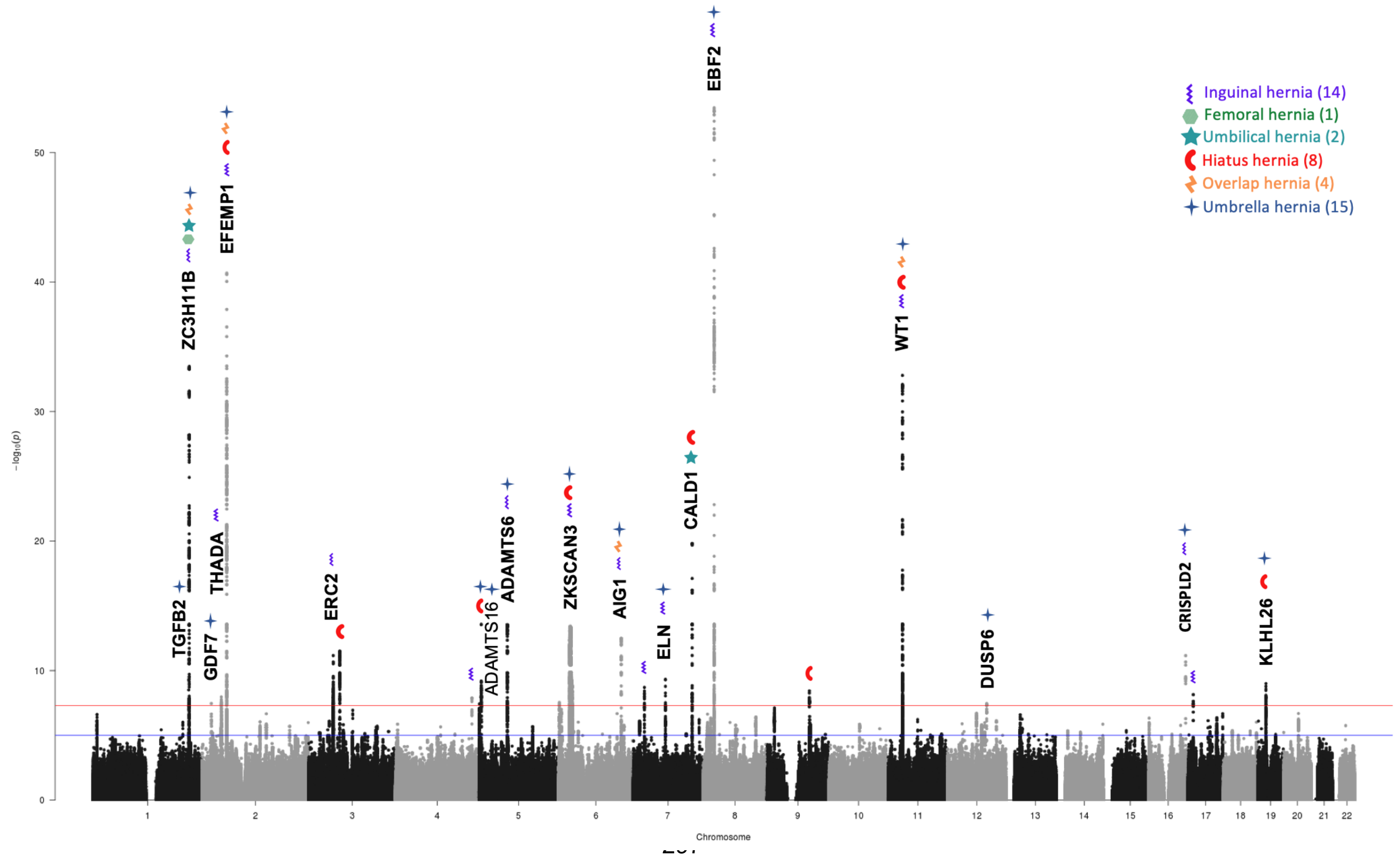


Figure 4.10. New putative shared hernia susceptibility locus 5p15.33 discovered through multivariate meta-analysis in metaUSAT

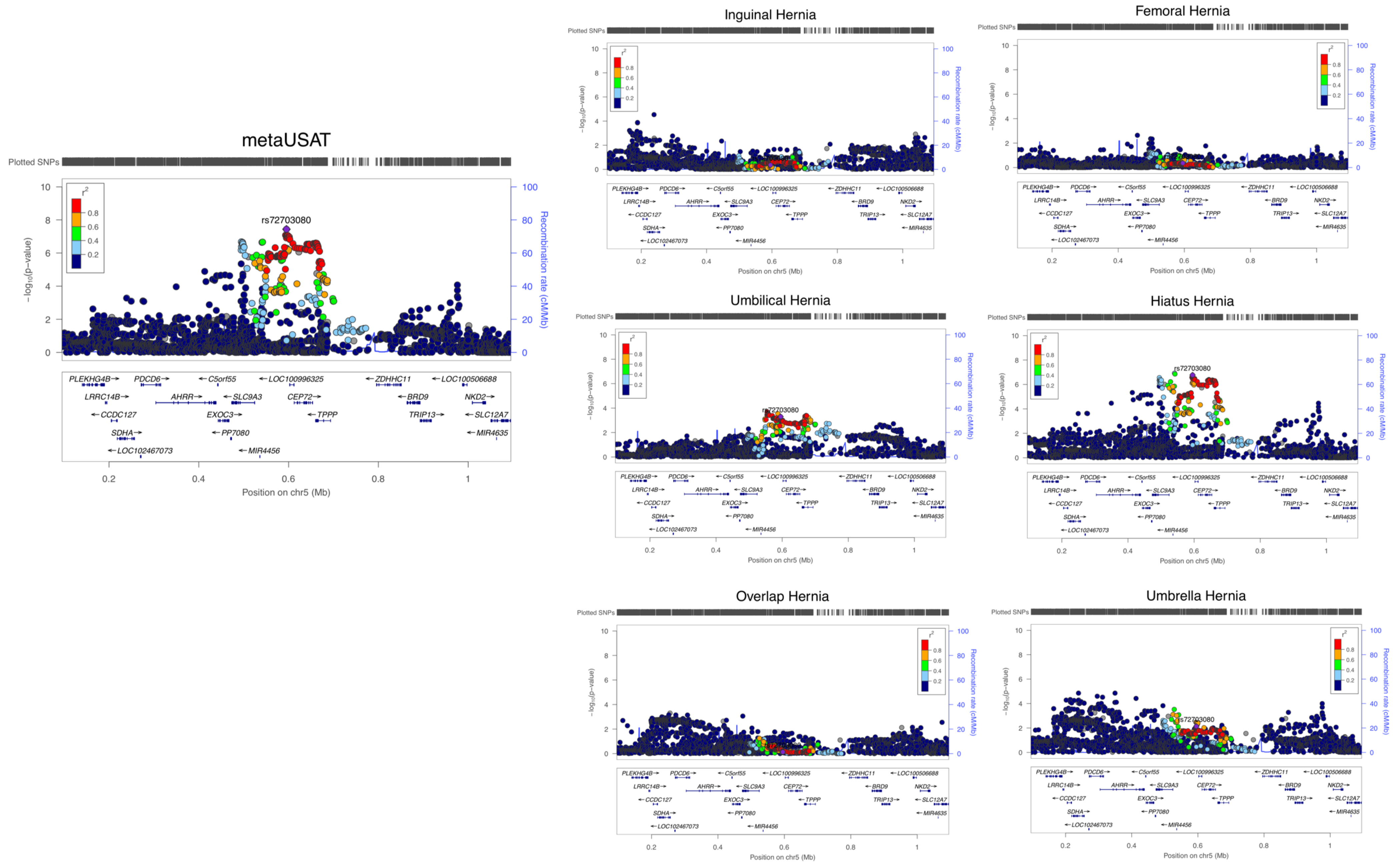
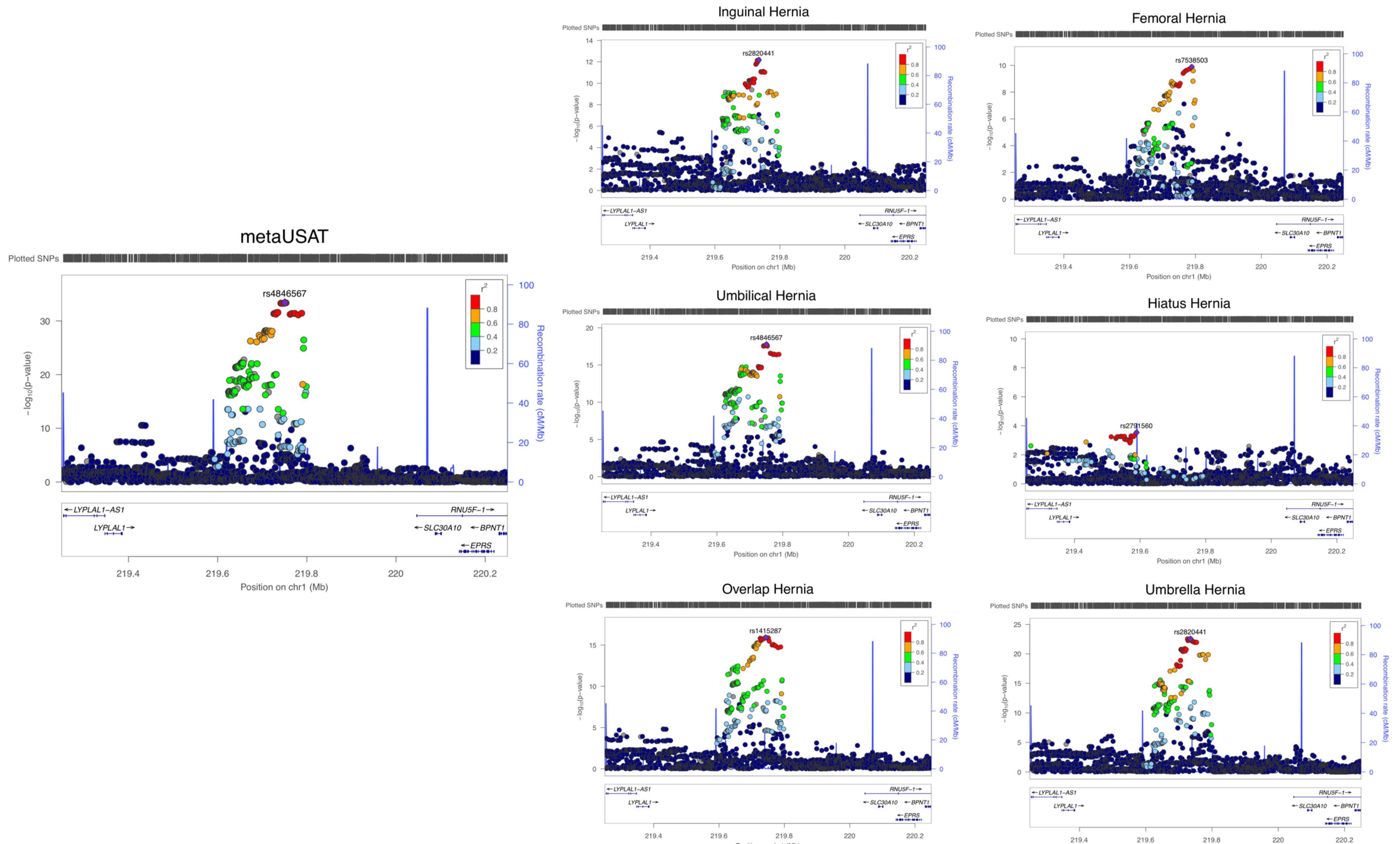
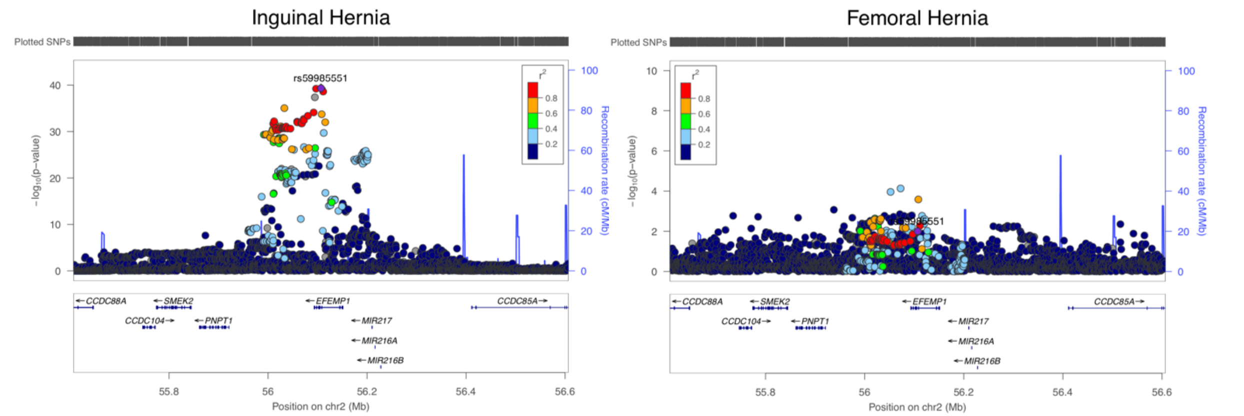
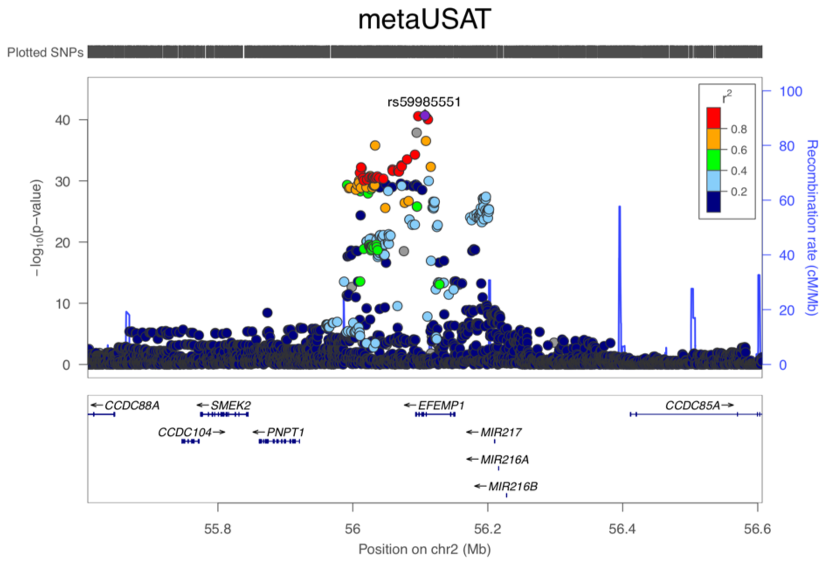


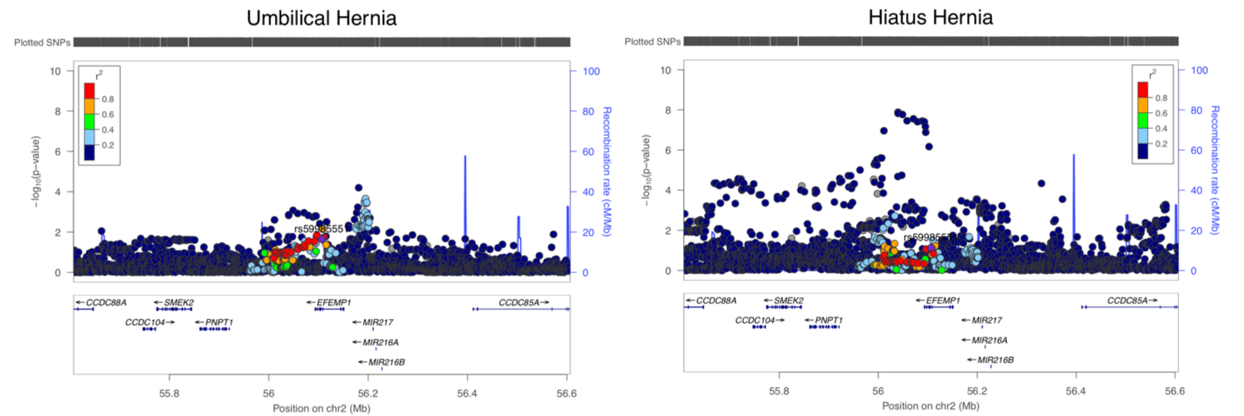
Figure 4.11. Six loci showing the greatest degree of overlap across all analyses.
1q41 (*ZC3H11B*) – metaUSAT SNP rs559230165 was not available in the reference panel hence the lowest P-value SNP was plotted for each Locus Zoom. *ZC3H11B* (the most proximal gene) was also not available in the gene panel.

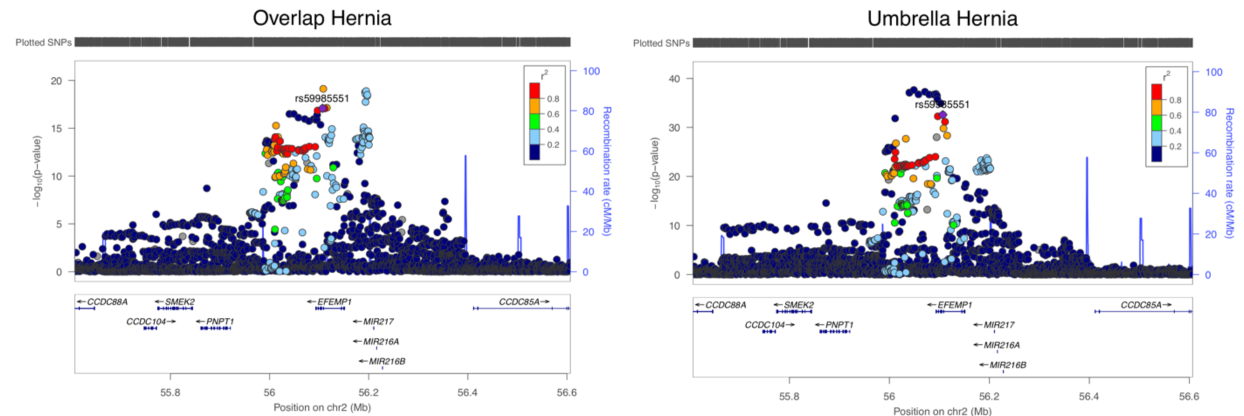


2p16.1 (*EFEMP1*)

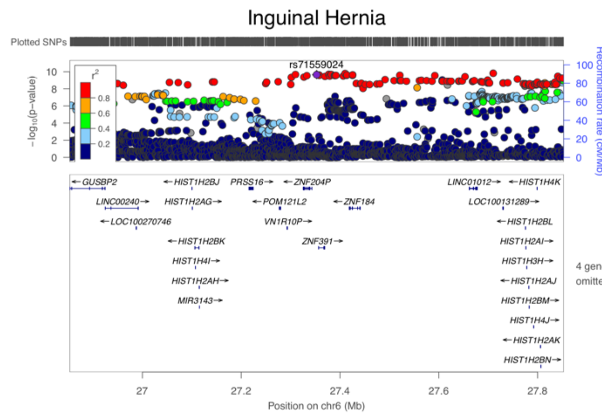


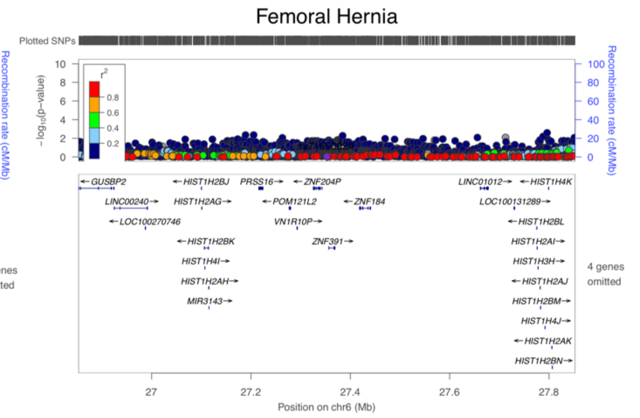


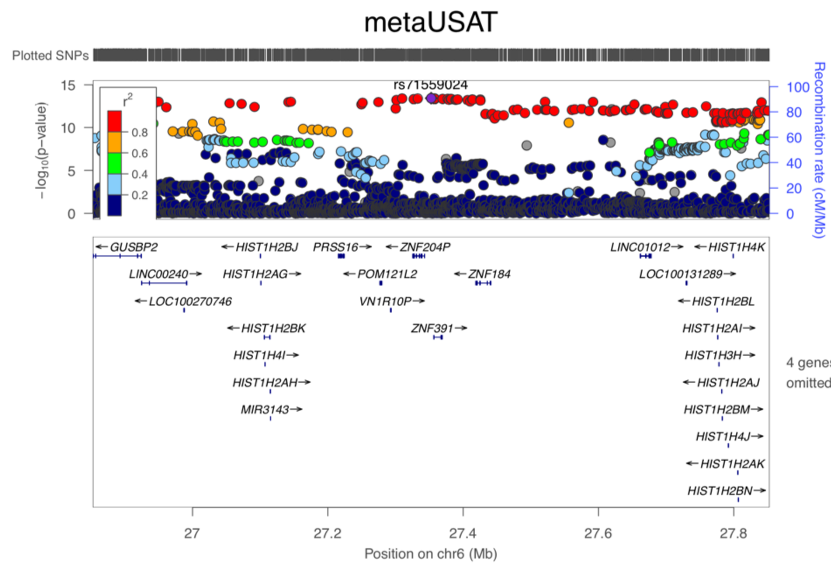


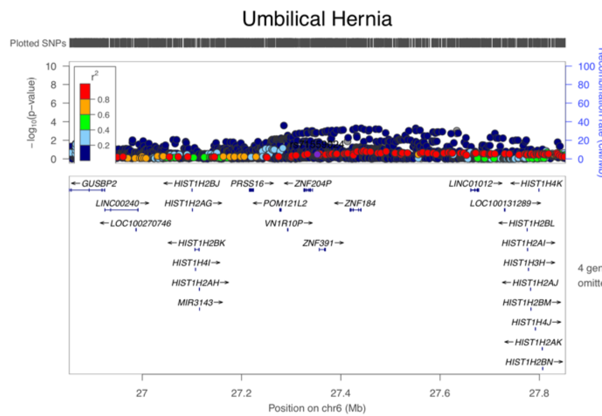


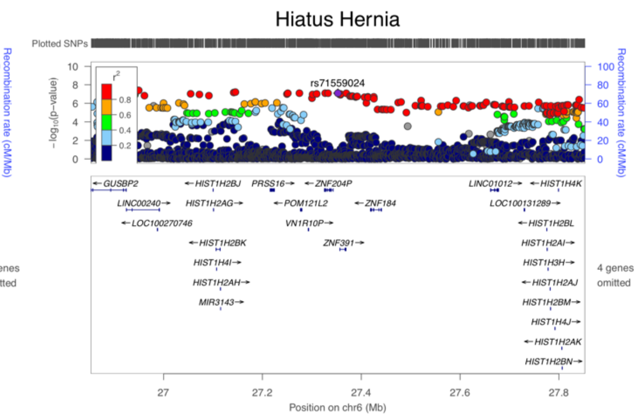
6p22.2 (*ZKSCAN3* (MHC region))

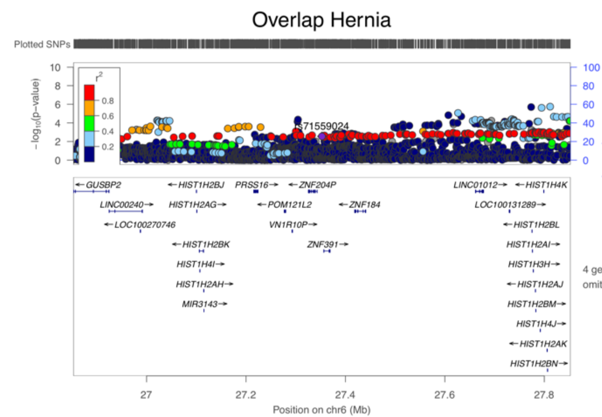


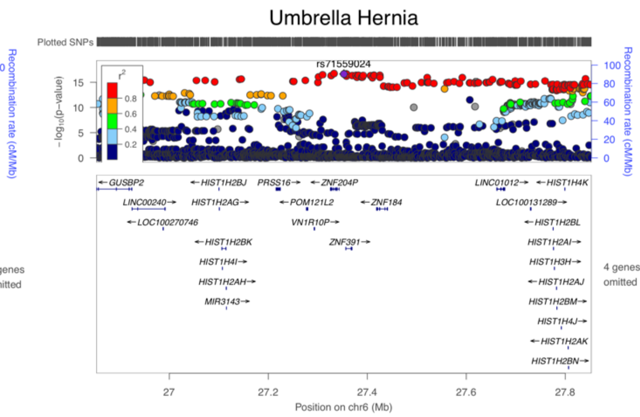




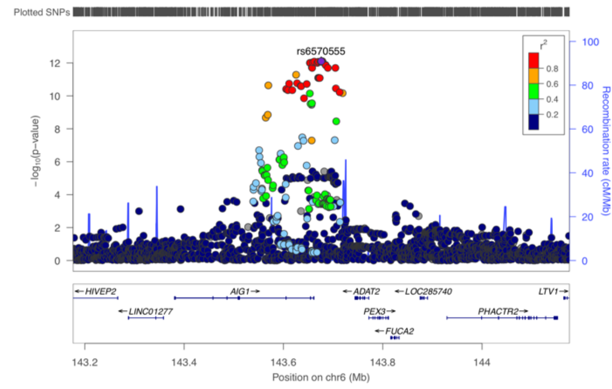


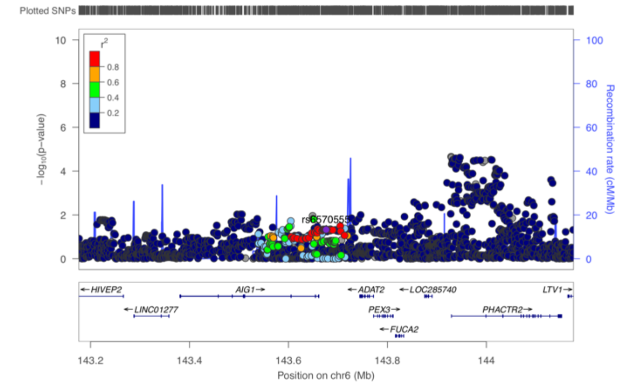


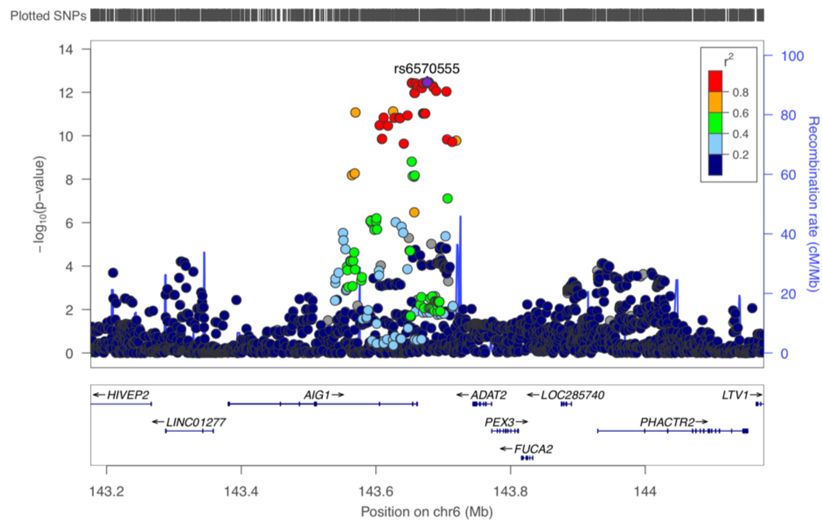


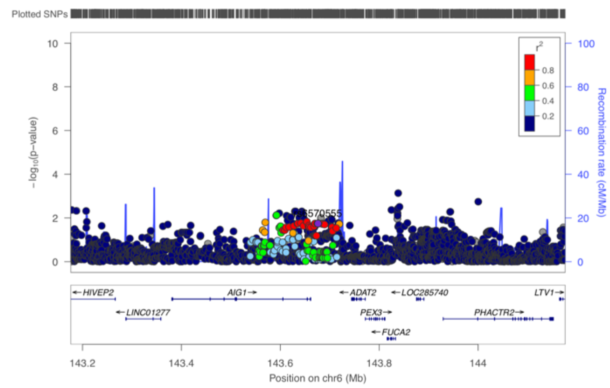


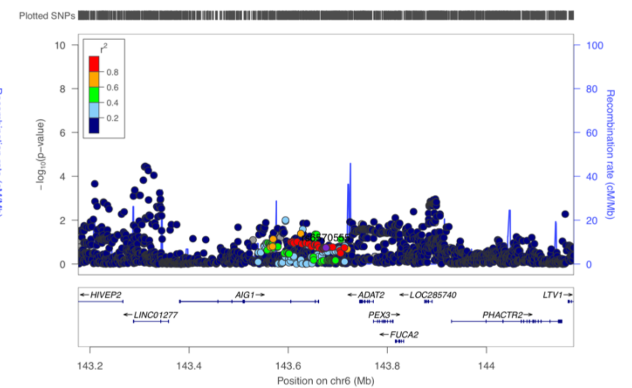
6q24.2 (*AIG1*)

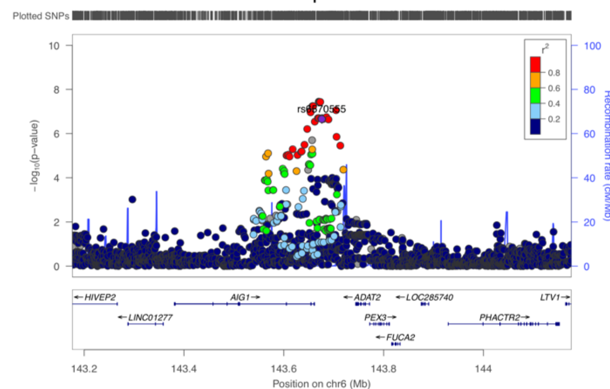


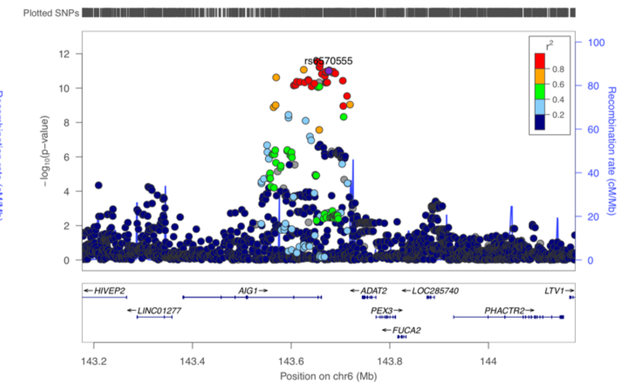




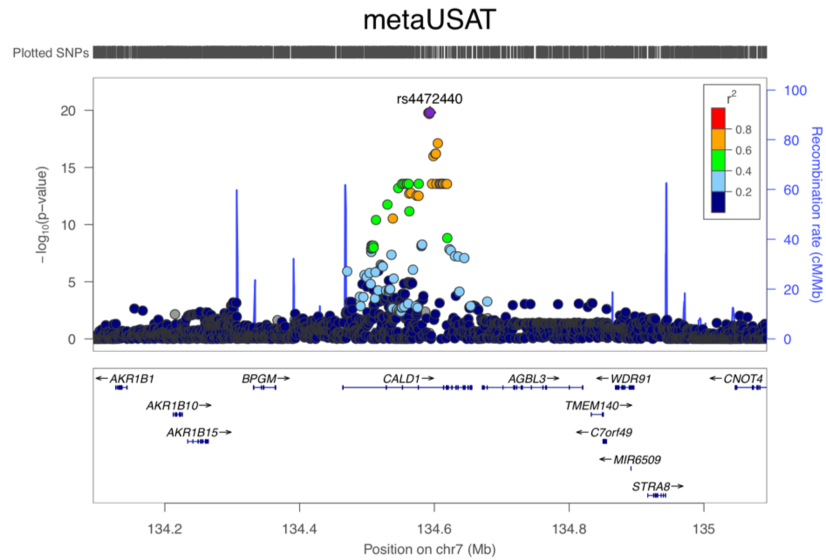


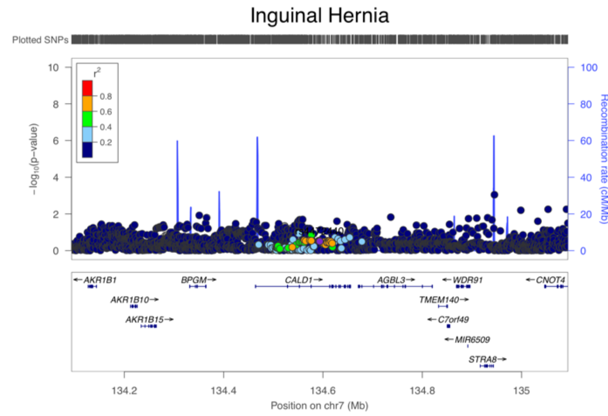


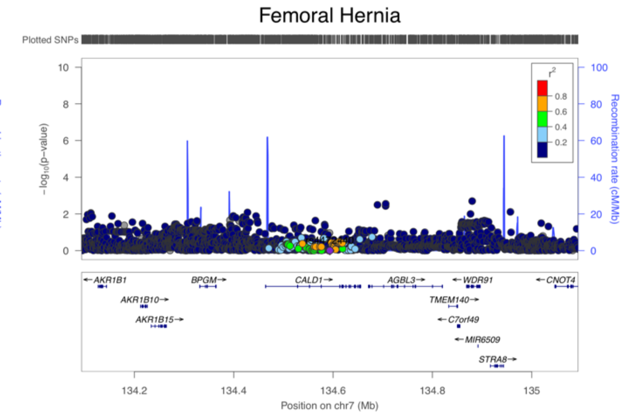


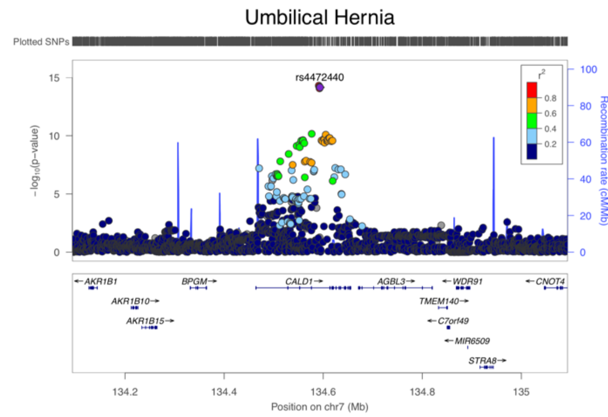


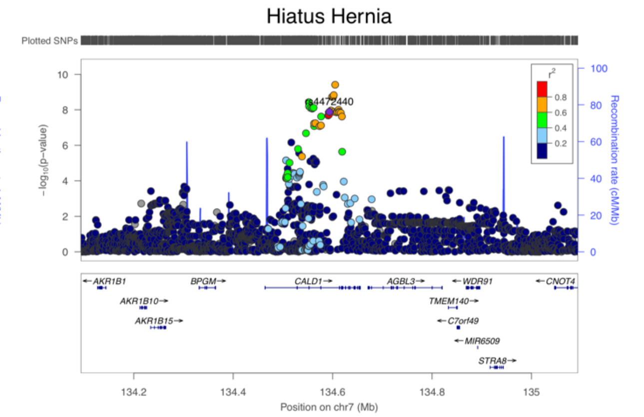
7q33 (*CALD1*)

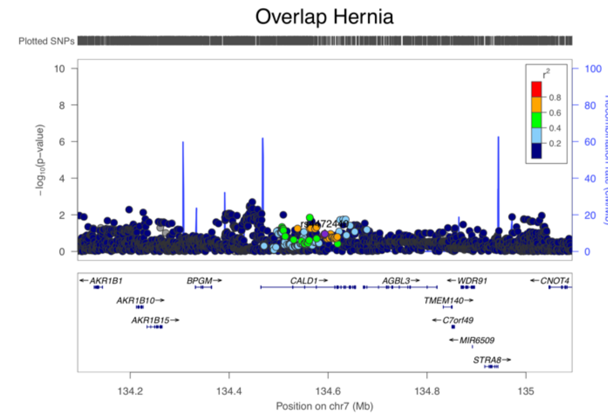


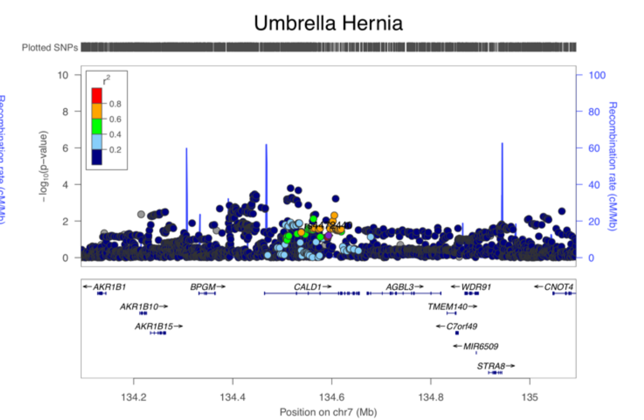




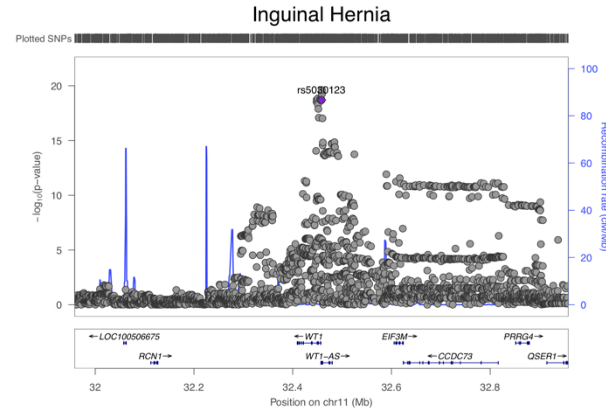


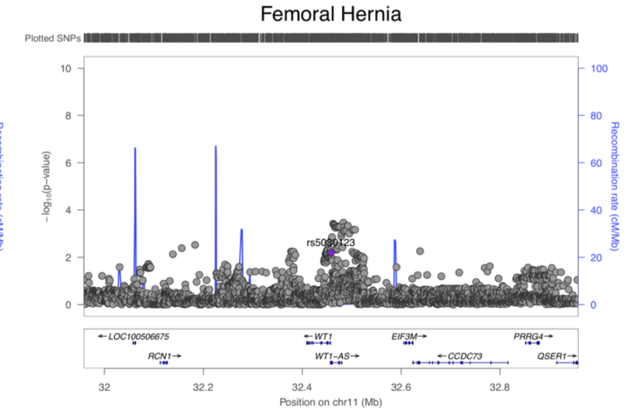


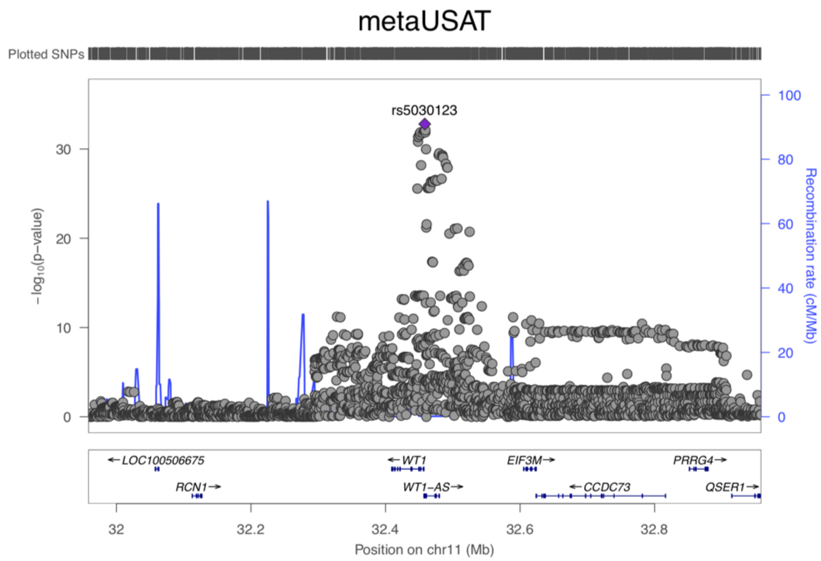


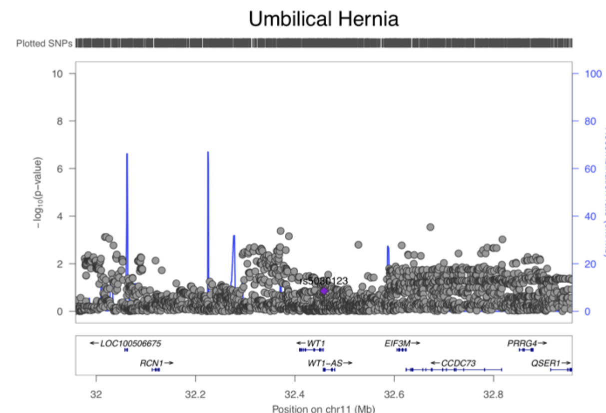


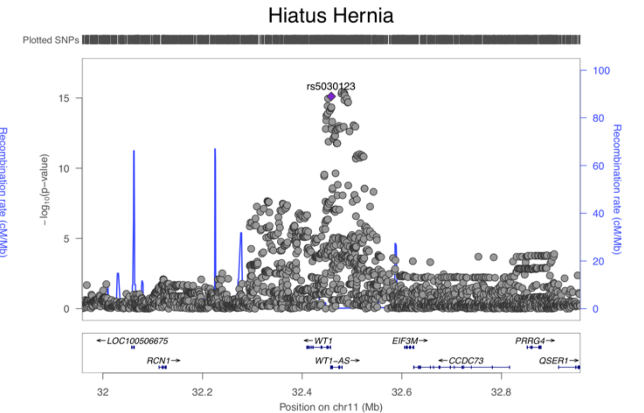
11p13 (*WT1*). *LD information was not available for SNP rs5030123 in the reference panel.*

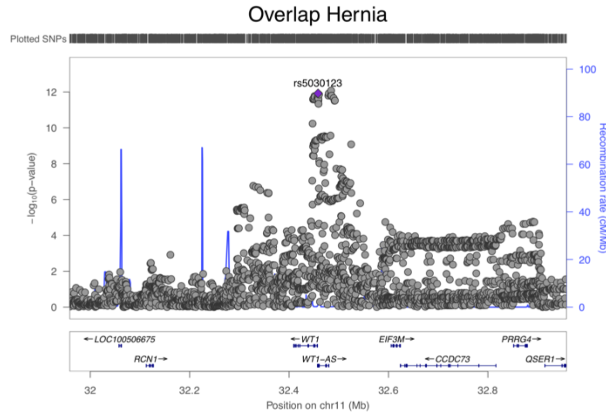


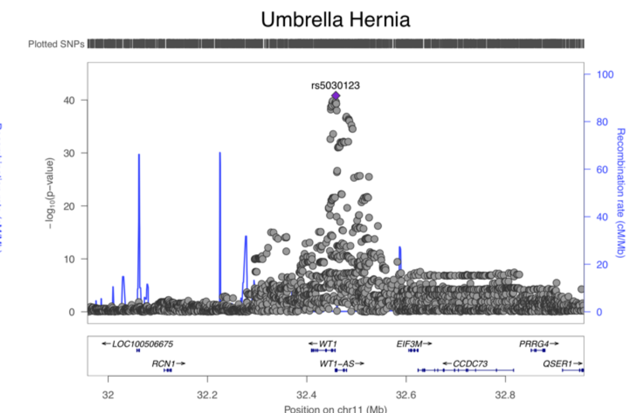












4.4. Discussion

4.4.1. Summary

Hernias are complex diseases with substantial genetic and non-genetic components. This chapter led to the discovery of fundamental molecular genetic candidates which collectively advance our existing understanding of the pathobiology of four common hernia subtypes, and more widely, hernia in general. In the first individual analyses, 38 susceptibility loci (34 of them novel) were found to associate individually with inguinal, femoral, umbilical, and hiatus hernia, representing substantial progress in characterising the genetic basis of these common diseases. Further, using both linear, multi-trait and multivariate analysis approaches to unravel the shared genetic biology of the four hernia phenotypes, six biologically-relevant loci (**1q41** (*ZC3H11B*); **2p16.1** (*EFEMP1*); **6p22.2** (*ZKSCAN3* (MHC region)); **11p13** (*WT1*); and **7q33** (*CALD1*), and **6q24.2** (*AIG1*)) were prioritised as putative loci demonstrating the highest evidence of shared susceptibility to multiple hernia phenotypes. Moreover, a further six loci (***5p15.33*** *(CEP72),* ***2p24.1*** *(GDF7),* ***1q41*** *(TGFB2),* ***3q22.2*** *(AMOTL2),* ***5p15.32*** *(ADAMTS16) and* ***12q21.33*** *(DUSP6))* were discovered under shared genetic analyses that were not discovered in the individual analyses and represent putative genetic elements in shared hernia biology. The *in silico* analyses provide data suggesting functional variants at the susceptibility loci. Pathway analyses enriched genes implicated in connective and elastic tissue biology, and thorax and anterior abdomen development, with tissue expression analyses markedly enriching adipose tissue. Finally, using genetic risk scoring, hernia severity (defined by surgical intervention) was discovered to correlate with genetic burden, highlighting the role of genetics in this unique complex disease model.

4.4.2. Six loci with evidence of shared susceptibility towards multiple hernia subtypes

1q41 (*ZC3H11B*). This locus imparted the highest susceptibility across three of four individual hernia phenotypes (umbilical, femoral and umbilical). Moreover, investigation for shared biology demonstrated this locus to be significant in the overlap and umbrella analyses, three of four MTAG analyses and the metaUSAT analysis. This locus also demonstrated strong evidence of functionality, with the highest effect size lead variant of all individual analyses, rs7538503 ($P_{Femoral} = 1.3\times10^{-10}$, OR = 1.42), associated with femoral hernia at this locus. Moreover, in the umbilical hernia individual analysis, the lead signal at this locus, rs4846567, is a predicted pathogenic regulatory region variant ($P = 1.7\times10^{-18}$, OR = 1.22, CADD = 14.9), and is also a genome-wide significant eQTL for *ZC3H11B* in GTEx v8 testis tissue ($P_{eQTL} = 1.5\times10^{-8}$). *ZC3H11B* demonstrates significant association with several myopia endophenotypes, including axial length, refractive error, corneal astigmatism, and spherical equivalent.[30] Accelerated connective tissue remodelling of the posterior sclera is thought to lead to axial elongation, which is a key feature of myopia. This pathological remodelling leads to stiffening of ocular connective tissue and is thought to be the underlying process behind myopia leading to glaucoma[31]. Intriguingly, at least two marfanoid-like syndromes, have been described with co-existing myopia and hernia.[32,33]

2p16.1 (*EFEMP1*). *EFEMP1* imparted susceptibility to inguinal and hiatus individual hernia phenotypes, as well as the overlap, umbrella, metaUSAT, and importantly, all four MTAG analyses. EFEMP1, which is discussed in detail in the previous two

chapters, encodes fibulin-3, a core extracellular matrix glycoprotein. Notably, an almost identical association profile is seen for *EFEMP1* as with *ZC3H11B*, with aberrant expression of *EFEMP1* associated with several ophthalmic phenotypes, including age-related macular degeneration, and several endophenotypes of myopia and glaucoma, as well as anthropometric measures of height and weight which are also associated with *ZC3H11B*. *EFEMP1* was discovered as a susceptibility locus by Jorgensen and colleagues in their GWA study of inguinal hernia (rs2009262, P = 1.45×10^{-17}, OR = 1.15), and this signal was replicated in this larger study (P = 6.1×10^{-33}, OR = 1.18).[34] Jorgensen demonstrated aberrant accumulation of EFEMP1 in mouse connective tissue[34], and suggested a role for EFEMP1 in hernia pathobiology through its role in collagen and elastin fibre homeostasis. Collagen is an important component of abdominal fascia[35], and indeed the transversalis fascia has been demonstrated to have a decreased type I:III collagen ratio in inguinal hernia[36], with heightened expression of type III collagen mRNA also described.[37] Disruption in the complex balance of collagen in fascia is therefore strongly implicated in inguinal hernia biology.[38] As well as collagen, fibulin-3 binds tropoelastin[39], the monomeric unit of elastin fibres. Diminished elastic fibre assembly in fascia of EFEMP1 knockout mice has been described, and these mice invariably develop inguinal hernia.[40] Therefore, the overexpression of collagen and decreased expression of elastin in fascial tissue may play a fundamental role in hernia biology.

11p13 (*WT1*). Alongside *EFEMP1*, locus 11p13 (*WT1*) was significantly associated in all four multi-trait MTAG analyses. WT1 was discovered by Jorgensen as one of four susceptibility loci associated with inguinal hernia in their GWA study (P = 3.69×10^{-14}, OR = 1.11)[34], a signal which was replicated in the present study (P = 4.0×10^{-20}, OR =

1.11). This locus was also associated with hiatus hernia ($P = 1.1\times10^{-13}$, OR = 1.07). Of note, several *WT1* deletion syndromes, including Denys-Drash[41], WAGR[42], and Meachem syndromes[43], are characterised by congenital hiatus hernia. Moreover, a variant near *WT1-AS* (~15kb from the lead variant at 11p13) has been discovered to significantly associate (rs3858461, $P = 1.0\times10^{-8}$) with medication use for peptic ulcer and gastro-oesophageal reflux disease (GORD)[44], which are the principal clinical sequelae of hiatus hernia. Of note, this variant was more significantly associated with hiatus and inguinal hernia in the individual analyses ($P_{Hiatus} = 2.4\times10^{-13}$, OR = 1.06; $P_{Inguinal} = 2.9\times10^{-9}$, OR = 1.07). WT1 is implicated in collagen homeostasis via its inhibition of MMP2 and activation of TIMP3, which lead to collagen accumulation. Deranged expression of MMPs which break down collagen, and their inhibitors, is noted in hernia fibroblasts.[37] EFEMP1 is thought to buttress the inhibitory effect of WT1 on MMPs through stimulating TIMP3 activity[45] - thus connecting these two shared loci. Taken together, these loci therefore impede MMPs and are suggested to disrupt the collagen make-up of fascia leading to hernia pathogenesis.

7q33 (*CALD1*). Susceptibility signals near *CALD1* were discovered to associate with umbilical and hiatus hernia (rs12707188, $P_{Umbilical} = 5.0\times10^{-15}$, OR = 1.19, EAF = 0.37(T); rs4728341, $P_{Hiatus} = 3.9\times10^{-10}$, OR = 1.06, EAF = 0.55 (T)), though intriguingly their effects are in opposite directions. The metaUSAT analysis further suggested shared genetic architecture for hernia at this locus. *CALD1* encodes caldesmon-1, which is a calmodulin- and actin-binding protein and is highly expressed in smooth muscle and non-muscle tissue[46], with important roles in cell motility, migration and reorganisation of actin cytoskeleton and smooth muscle contraction.[47] The patency of the processus vaginalis in inguinal hernia has been linked to smooth muscle cell

(SMC) persistence[48,49], and immunohistochemical studies have demonstrated smooth muscle phenotypic modulation to a synthetic phenotype in hernia; several markers for mature smooth muscle cells, including alpha-smooth muscle actin, desmin, and caldesmon, were found to be elevated in inguinal hernia sacs.[50,51] Moreover, new-born mice that are caldesmon-1 null have occasionally been demonstrated to be viable, however with fatalities occurring within the first 5-7 hours in the majority of homozygotes due to severe umbilical hernia.[52] *CALD1* has also been associated with diverticular disease, which is another elastic tissue disease characterised by SMC phenotypic switch to a secretory phenotype, specifically a reduction in smooth muscle myosin heavy chain (a marker of mature SMCs).[53,54] Therefore impaired contractility of smooth muscle and pathological SMC phenotypic transition may have an important role to play in hernia formation.

6q24.2 (*AIG1*). *AIG1* represents one of four overlap loci, demonstrating robust evidence of shared genetics via an association with inguinal and umbrella hernia and enrichment in the multivariate metaUSAT analysis. AIG1 (Androgen-inducible gene 1) has been cloned from human dermal papilla cells[55], where it has been shown to be inducible by dihydrotestosterone.[56] AIG1 mRNA shows high expression in testis, ovary, heart and colon.[55] Complete androgen insensitivity syndrome (CAIS) is characterised by masculinisation of the external genitalia caused by an inability to respond to androgens and a patent processus vaginalis.[57] Most characteristically, CAIS presents in young females with bilateral inguinal hernia; indeed, ~1-2% of all females with bilateral inguinal hernia are thought to have CAIS.[58] Disruption of AIG1 expression may therefore have some role to play in androgen sensitivity and therefore hernia pathobiology.

6p22.2 (*ZKSCAN3* (MHC region)). The MHC region imparted susceptibility to both inguinal and hiatus individual phenotypes (rs13212652, $P_{Inguinal} = 3.1\times10^{-11}$, OR = 1.12, EAF = 0.87(T); rs9393735, $P_{Hiatus} = 2.7\times10^{-8}$, OR = 1.07, EAF = 0.86(G)). The exact role of the MHC region in hernia is difficult to discern, however, common variants in the MHC locus have been found to heavily associate with Barrett's oesophagus[59], which is strongly linked with hiatus hernia.

4.4.3. Six loci discovered through multi-trait analysis that weren't discovered in the individual analyses

5p15.33 *(near CEP72).* A putative new locus **5p15.33** was discovered to associate with hernia in the multivariate analysis (rs72703080, $P_{metaUSAT}$ = 3.68×10^{-8}, ~20kb upstream from *CEP72*). *CEP72* encodes a centriolar satellite protein that is necessary for regulating microtubule-organising activity and centrosome integrity.[60] Using comparative genomic hybridisation, Choi *et al* discovered copy number aberrations (gains) at **5p15.33** in patients with ruptured intracranial aneurysms.[61] Moreover, the *CEP72* region was previously implicated in a GWA meta-analysis of Barrett's oesophagus and oesophageal adenocarcinoma.[62] Hiatus hernia plays a key role in oesophageal mucosal injury in GORD, which predisposes to Barrett's oesophagus.[63] Indeed, the size of hiatus hernia is significantly associated with progression of Barrett's oesophagus to high-grade dysplasia or malignancy.[64] To this end, a tangible contributor to shared hernia risk has been identified through multivariate meta-analysis.

2p24.1 *(GDF7).* Lead variant rs3072 resides in a 3'UTR of *GDF7* and was discovered to associate with hernia in the umbrella and metaUSAT analyses (P = 1.80×10^{-8}, P= 3.48×10^{-8}, respectively). Of note, this variant demonstrates strong functionality as a robust eQTL for *GDF7* in GTEx aorta tissue (P_{eQTL} =5.4×10^{-9}). GDF7 encodes BMP12, a ligand in the bone morphogenetic protein pathway, which is involved in neural system development[65] and in tendon and ligament development and repair.[66] BMP12 is tenogenic, and plays an important role in differentiating mesenchymal stem cells into tenocytes[67], and is therefore used in tissue engineering approaches to tendon

repair.[68] *GDF7* has been identified through GWAS to associate with eight traits, with three of these characterised by connective and elastic tissue dysfunction: pelvic organ prolapse (rs9306894-G, P = 3×10^{-17}, OR = 1.11)[69], abdominal aortic aneurysm (rs13382862-A, P = 1.0×10^{-6}, OR = 1.10)[70], and diverticular disease (rs7255-T, P = 4×10^{-6}, OR = 1.06).[71] The T allele of rs7255 was also found to confer risk of Barrett's Oesophagus in the GWAS by Gharahkhani *et al* (p=9.0e-11, OR=1.14).[62] The BMP pathway has been heavily implicated in Barrett's Oesophagus[72], and several studies have identified polymorphisms in the *TBX5-GDF7* genomic region demonstrating a considerable connection with Barrett's.[73–75] Further, the lead variant rs3072 at this locus was sub-threshold, but strongly suggestive of association in the individual hiatus hernia analysis (P = 6.30×10^{-8}), lending evidence for a potential role for this locus in hiatus hernia biology and a wider role in shared hernia susceptibility.

1q41 (*TGFB2*). Using the multi-trait MTAG analysis approach, this new locus was discovered to associate with inguinal ((rs3121580, $P_{MTAG} = 3.41\times10^{-8}$) and femoral hernia (rs2799098, P = 4.66×10^{-8}), and was identified in the metaUSAT and umbrella analyses, where it was the most significant locus (rs2799098, $P_{metaUSAT} = 5.8\times10^{-10}$, $P_{Umbrella} = 9.3\times10^{-15}$). *TGFB2* encodes the TGF-β2 ligand which is a core component of the TGF-β signalling pathway. A signature of increased TGF-β signalling is seen in Marfan syndrome, Loeys-Dietz, and cutis laxa which are associated with aneurysmal changes caused by mutations that hamper smooth muscle cell contractile proteins.[76] *TGFB2* haploinsufficiency pathologically activates the TGF-β signalling pathway[77] leading to Loeys-Dietz syndrome type 4 (LDS4)[78], which is characterised by arterial vasculopathy (arterial aneurysms, dissection and tortuosity), and other widespread connective tissue pathology including hernias.[79] Of note, several association studies

have found *TGFB2* to associate with glaucoma endophenotypes, including intraocular pressure[80], central corneal thickness[81]; as well as FEV1/FVC ratio[82] and severe COPD[83], which has also been suggested as an independent risk factor for hernia pathology and severity.[1,84]

***3q22.2** (AMOTL2).* The umbrella analysis revealed the **3q22.2** locus (rs9883955, P = 1.2×10^{-8}), which was not identified in any of the other analyses. The scaffold protein, AMOTL2, forms a complex with VE-cadherin which regulates endothelial cell shape through mechanical coupling of adherens junctions to contractile actin fibres, and has been shown to be required for aortic lumen expansion.[85] Additionally, AMOTL2 plays a fundamental role in endothelial cell polarity, migration and proliferation during angiogenesis.[86] Several associations of *AMOTL2* have also been described with myopia-related phenotypes, including refractive error, spherical equivalent, and vertical cup-disc ratio.[30] Moreover, proximal deletions in 3q have been described in the literature, which most commonly occur in 3q22-23. *De novo* mutations in 3q22.1 result in a syndromic presentation of bilateral inguinal hernia[87] and an interstitial deletion of 3q23 has been described to result in BPES syndrome, which is characterised by diaphragmatic hernia.[88] This region on the long arm of chromosome 3 may therefore be of considerable interest in multiple hernia pathobiology.

***5p15.32** (ADAMTS16).* Two signals were discovered in the metaUSAT analysis in proximity to the *ADAMTS16* gene at **5p15.32**: rs42202 ($P_{metaUSAT} = 2.71\times10^{-14}$, ~100kb upstream from *ADAMTS16*) and rs7715383 ($P_{metaUSAT} = 2.97\times10^{-9}$, ~25kb downstream from *ADAMTS16*), both of which were discovered in the umbrella analysis (P = 1.7×10^{-11} and P = 1.2×10^{-12}, respectively). Intriguingly, only the upstream

signal (and not the downstream signal) was associated with hiatus hernia in the individual (rs42202, $P_{Hiatus} = 8.0 \times 10^{-16}$) and MTAG (rs42202, $P_{MTAG} = 5.70 \times 10^{-15}$) analyses. ADAMTS16 is a member of the ADAMTS family of multi-domain secreted metalloendopeptidases which are central remodelling enzymes of the extracellular matrix.[89] ADAMTS16 is co-expressed alongside WT1 in murine models and is thought to play an important role in murine genitourinary development.[90] Variants in *ADAMTS16* have been associated with urinary incontinence[91], a manifestation of pelvic floor dysfunction, which have been shown independently to lead to a higher prevalence of hiatus and inguinal hernia.[92] Several mutations have been described in the other 18 ADAMTS superfamily genes which result in distinct human genetic disorders.[89] Three additional ADAMTS proteins show striking implications for connective tissue biology. Mutations in *ADAMTS2* are responsible for dermatosparactic type Ehlers-Danlos Syndrome (type VIIC) by disrupting processing of procollagen molecules[93], and is typified by extreme skin fragility, joint laxity, and umbilical hernia. ADAMTS4 shows significant aggrecanase activity and is implicated in articular cartilage degradation and arthritis[94], and ADAMTS4 mRNA and protein have been found to be highly expressed in lumbar herniated intervertebral discs, suggesting their potential role in lumbar disc herniation.[95] ADAMTS6 protease converts procollagen to collagen[96]; using qRT-PCR and RNA-seq, Jorgensen and colleagues discovered ADAMTS6 to be a susceptibility locus for inguinal hernia and demonstrated reduced expression levels of *ADAMTS6 in* mouse connective tissue related to human transversalis fascia, suggesting a role for ADAMTS6 in collagen homeostasis and hernia biology.[34] These findings provide evidence for the role of ADAMTS proteins in susceptibility to hernia pathology and the potential role of *ADAMTS16* in shared hernia risk.

12q21.33 *(DUSP6).* rs797267 was discovered in the umbrella and metaUSAT analyses ($P = 2.6\times10^{-9}$ and $P = 3.51\times10^{-8}$, respectively), a variant which is a suggestive eQTL for *DUSP6* in tibial nerve tissue ($P_{eQTL}= 3.2\times10^{-5}$). *DUSP6* encodes MKP3, an extracellular signal regulated kinase (ERK)-specific MAPK phosphatase which is a key regulator of extracellular signals transduced to the nucleus. DUSP6 acts as a negative regulator of fibroblast growth factor receptor signalling and loss-of-function mutations in DUSP6 result in FGFR-like syndromes characterised by postnatal mortality, dwarfism, craniosynostosis and hearing loss.[97] Deletions of the 12q21 region have been attributed to cause cranio-facio-cutaneous syndrome[98], which, among other craniofacial manifestations, presents variably with umbilical hernia.[99]

4.4.4. Genetic risk scoring of hernia severity and multiple hernia risk

The lifetime risk of groin hernia is ~27% in males and 3% for females[100], and the incidence of surgical correction sits at 10 to 28 per 100,000 per year.[101] Due to the risk of complications such as postoperative pain, infection and recurrence, elective hernia repair is generally reserved for patients with more symptomatic disease. Moreover, emergency hernia surgery (associated with a substantial mortality rate[102]), is reserved for incarcerated and strangulated hernia, which are surgical emergencies. As expected, all individual, overlap and umbrella hernia cases had a higher genetic burden than matched controls. The genetic risk score analyses also correlated with disease severity, with surgical cases having a higher genetic burden than non-surgical cases in all analyses. These data provide insights into the role of genetic susceptibility in hernia development and severity, and provide an important proof-of-principle of genetic risk scoring in personalising therapeutic approaches to manage this highly prevalent disease. An important caveat is that the secondary genetic risk score analyses, by and large did not demonstrate a higher genetic burden among hernia patients with multiple hernia presentations compared to patients with only an individual hernia occurrence. The only case where multiple hernia was nominally associated with a higher genetic burden was for hiatus hernia, the best powered individual analysis. This may suggest that genetic risk scoring in a larger cohort and using the full extent of associated signals (i.e. a polygenic risk score) may enable more conclusive data. However, this remains to be determined, and in this study we were unable to determine genetic risk score correlation with multiple hernia.

4.4.5. Strengths and limitations

The approach described in this chapter has several limitations that must be addressed in future research. Firstly, despite multiple attempts, we were unable to source a replication cohort to test the many interesting associations discovered. In **Chapter 2**, ~1/3 of discovery loci replicated in an independent cohort, meaning some of the findings identified in this chapter may indeed be false positives. Secondly, the use of biobank-scale data for the study of shared genetic susceptibility to hernia means that invariably heterogeneity in cohort structure are apparent, meaning that hiatus and inguinal hernia (which are substantially more common in the UK population) were over-represented in our dataset and accounted for ~90% of the total cohort across the four hernia phenotypes. This means that in the several multi-trait approaches implemented to dissect shared genetics, many of the identified loci are only minimally associated with umbilical and femoral hernia, despite being highlighted as contributors of shared genetic susceptibility. A larger cohort size, with an equal balance across all four hernia subtypes, would prove useful in defining further the prioritised loci demonstrating high degrees of shared susceptibility as well as uncovering potentially new shared loci.

By using distinct analytic approaches to examine the genetic architecture of the four hernia subtypes, novel and informative insights were gained into the biology of hernia. This enabled sub-threshold (i.e. non-genome-wide significant) loci to be prioritised that would otherwise not have been discovered with traditional single-trait association analyses. Despite the evident lack of a replication cohort, by segregating the four hernia cohorts in UK Biobank to avoid overlap, one can have confidence in the validity

of several loci that were discovered across multiple hernia phenotypes. This is the case for the twelve loci that demonstrated the greatest degree of overlap across the different analyses, and the resulting clustering of many of these loci across functionally related ontologies. Furthermore, the enrichment of biological pathways previously implicated in hernia pathobiology provide further compelling evidence to support the veracity of these loci, and for a shared genetic susceptibility to hernia. Lastly, despite not correlating with shared hernia susceptibility, the genetic risk score correlated with disease severity, which provides important evidence for the role of genetics in hernia biology.

4.5. Conclusion

Herniae are a complex surgical disease with a substantial global health burden affecting patients across all age groups. This study represents an important advancement in the understanding of the genetic architecture of abdominal wall and hiatus hernia. Several loci were discovered as key genetic players in individual hernia susceptibility, and when analysed together, twelve loci demonstrated striking overlap across multiple hernia subtypes, representing fundamental candidates which confer a shared susceptibility to multiple hernia. These shared susceptibility loci cluster in functional categories, most prominently connective and elastic tissue homeostasis and dysfunction which were also enriched in the pathway analyses. Moreover, there is significant evidence of functionality in these shared regions, and the weighted genetic risk score constructed from the association signals also correlated with disease severity. This suggests that a *phenotypic* severity of hernia correlates with *genotypic* severity, which is an important finding to inform future personalised therapeutic approaches to hernia.

4.6. Chapter References

1. Lau, H., Fang, C., Yuen, W. K. & Patil, N. G. Risk factors for inguinal hernia in adult males: A case-control study. *Surgery* **141**, 262–266 (2007).
2. Burcharth, J., Pommergaard, H. C. & Rosenberg, J. The inheritance of groin hernia: A systematic review. *Hernia* **17**, 183–189 (2013).
3. Ikeda, H. *et al.* Risk of contralateral manifestation in children with unilateral inguinal hernia: Should hernia in children be treated contralaterally? *J. Pediatr. Surg.* **35**, 1746–1748 (2000).
4. Jansen, P. L., Klinge, U., Jansen, M. & Junge, K. Risk factors for early recurrence after inguinal hernia repair. *BMC Surg.* **9**, 18 (2009).
5. Maddox, M. M. & Smith, D. P. A long-term prospective analysis of pediatric unilateral inguinal hernias: Should laparoscopy or anything else influence the management of the contralateral side? *J. Pediatr. Urol.* **4**, 141–145 (2008).
6. Zöller, B., Ji, J., Sundquist, J. & Sundquist, K. Shared and nonshared familial susceptibility to surgically treated inguinal hernia, femoral hernia, incisional hernia, epigastric hernia, and umbilical hernia. *J. Am. Coll. Surg.* **217**, (2013).
7. Glassow, F. Femoral hernia. Review of 2,105 repairs in a 17 year period. *Am. J. Surg.* **150**, 353–356 (1985).
8. De Luca, L. *et al. Relationship Between Hiatal Hernia and Inguinal Hernia*. *Digestive Diseases and Sciences* **49**, (2004).
9. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med.* **12**, (2015).
10. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

11. Loh, P. Mixed-model association for biobank-scale datasets. **50**, 906–908 (2018).

12. Loh, P. R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).

13. De Jager, P. L. *et al.* Integration of genetic risk factors into a clinical algorithm for multiple sclerosis susceptibility: a weighted genetic risk score. *Lancet Neurol.* **8**, 1111–1119 (2009).

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

15. Mägi, R. & Morris, A. P. GWAMA: Software for genome-wide association meta-analysis. *BMC Bioinformatics* **11**, (2010).

16. Ray, D. & Boehnke, M. Methods for meta-analysis of multiple traits using GWAS summary statistics. *Genet. Epidemiol.* **42**, 134–145 (2018).

17. Jorgenson, E. *et al.* A genome-wide association study identifies four novel susceptibility loci underlying inguinal hernia. *Nat. Commun.* **6**, 1–9 (2015).

18. Bulik-Sullivan, B. *et al.* LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).

19. Watanabe, K., Taskesen, E., Van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1–10 (2017).

20. Wang, K. L. M. H. H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).

21. Ng, P. C. & Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).

22. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature Methods* **7**, 248–249 (2010).

23. 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 (2018).

24. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22**, 1790–1797 (2012).

25. Aguet, F. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

26. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLoS Comput. Biol.* **11**, 1–20 (2015).

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

28. Liberzon, A. *et al.* The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst.* **1**, 417–425 (2015).

29. Fang, H., Knezevic, B., Burnham, K. L. & Knight, J. C. XGR software for enhanced interpretation of genomic summary data, illustrated by application to immunological traits. *Genome Med.* **8**, 1–20 (2016).

30. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006

31. Grytz, R., Yang, H., Hua, Y., Samuels, B. C. & Sigal, I. A. Connective tissue remodeling in myopia and its potential role in increasing risk of glaucoma. *Current Opinion in Biomedical Engineering* **15**, 40–50 (2020).

32. Mégarbané, A. *et al.* Marfanoid habitus, inguinal hernia, advanced bone age, and distinctive facial features: A new collagenopathy? *Am. J. Med. Genet. Part*

A **158A**, 1185–1189 (2012).

33. Halal, F., Gervais, M.-H., Baillargeon, J., Lesage, R. & Opitz, J. M. Gastro-cutaneous syndrome: Peptic ulcer/hiatal hernia, multiple lentigines/café-aulait spots, hypertelorism, and myopia. *Am. J. Med. Genet.* **11**, 161–176 (1982).
34. Jorgenson, E. *et al.* A genome-wide association study identifies four novel susceptibility loci underlying inguinal hernia. *Nat. Commun.* **6**, (2015).
35. Antoniou, G. A. *et al.* Abdominal aortic aneurysm and abdominal wall hernia as manifestations of a connective tissue disorder. *J. Vasc. Surg.* **54**, 1175–1181 (2011).
36. Casanova, A. B., Trindade, E. N. & Trindade, M. R. M. Collagen in the transversalis fascia of patients with indirect inguinal hernia: a case-control study. *Am. J. Surg.* **198**, 1–5 (2009).
37. Rosch, R. *et al. A role for the collagen I/III and MMP-1/-13 genes in primary inguinal hernia? BMC Medical Genetics* (2002).
38. Bendavid, R. The Unified Theory of hernia formation. *Hernia.* **8**, 171–176 (2004).
39. Kobayashi, N. *et al.* A comparative analysis of the fibulin protein family: Biochemical characterization, binding interactions, and tissue localization. *J. Biol. Chem.* **282**, 11805–11816 (2007).
40. McLaughlin, P. J. *et al.* Lack of fibulin-3 causes early aging and herniation, but not macular degeneration in mice. *Hum. Mol. Genet.* **16**, 3059–3070 (2007).
41. Devriendt, K. *et al.* Diaphragmatic hernia in Denys-Drash syndrome. *Am. J. Med. Genet.* **57**, 97–101 (1995).
42. Scott, D. A. *et al.* Congenital diaphragmatic hernia in WAGR syndrome. *Am. J. Med. Genet. Part A* **134A**, 430–433 (2005).
43. Suri, M. *et al.* WT1 mutations in Meacham syndrome suggest a coelomic

mesothelial origin of the cardiac and diaphragmatic malformations. *Am. J. Med. Genet. Part A* **143A**, 2312–2320 (2007).

44. Wu, Y. *et al.* Genome-wide association study of medication-use and associated disease in the UK Biobank. *Nat. Commun.* **10**, (2019).

45. Klenotic, P. A., Munier, F. L., Marmorstein, L. Y. & Anand-Apte, B. Tissue inhibitor of metalloproteinases-3 (TIMP-3) is a binding partner of epithelial growth factor-containing fibulin-like extracellular matrix protein 1 (EFEMP1): Implications for macular degenerations. *J. Biol. Chem.* **279**, 30469–30473 (2004).

46. Hayashi, K. *et al.* Genomic structure of the human caldesmon gene. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 12122–12126 (1992).

47. Mayanagi, T. & Sobue, K. Diversification of caldesmon-linked actin cytoskeleton in cell motility. *Cell Adh. Migr.* **5**, 150–159 (2011).

48. Tanyel, F. C., Talim, B., Kale, G. & Büyükpamukçu, N. A Reevaluation of the Structures Accepted to Represent the Postnatal Gubernaculum. *Urol. Int.* **69**, 116–119 (2002).

49. Tanyel, F. C. *et al.* Inguinal hernia revisited through comparative evaluation of peritoneum, processus vaginalis, and sacs obtained from children with hernia, hydrocele, and undescended testis. *J. Pediatr. Surg.* **34**, 552–555 (1999).

50. Sfoungaris, D. *et al.* Differences in the development of the processus vaginalis between children with undescended testis and inguinal hernia. *J. Clin. Neonatol.* **5**, 31 (2016).

51. Mouravas, V. K. *et al.* Smooth muscle cell differentiation in the processus vaginalis of children with hernia or hydrocele. *Hernia* **14**, 187–191 (2010).

52. Guo, H. & Wang, C. L. A. Specific disruption of smooth muscle caldesmon

expression in mice. *Biochem. Biophys. Res. Commun.* **330**, 1132–1137 (2005).

53. Mimura, T. *et al.* Up-Regulation of Collagen and Tissue Inhibitors of Matrix Metalloproteinase in Colonic Diverticular Disease. *Dis. Colon Rectum* **47**, 371–379 (2004).

54. Hellwig, I. *et al.* Alterations of the enteric smooth musculature in diverticular disease. doi:10.1007/s00535-013-0886-y

55. Seo, J., Kim, J. & Kim, M. Cloning of androgen-inducible gene 1 (AIG1) from human dermal papilla cells. *Mol Cells* 35–40 (2001). Available at: https://pubmed.ncbi.nlm.nih.gov/11266118/. (Accessed: 10th September 2020)

56. A Arai, von Hintzenstern, F Kiesewetter, H Schell & O P Hornstein. In vitro effects of testosterone, dihydrotestosterone and estradiol on cell growth of human hair bulb papilla cells and hair root sheath fibroblasts. *Acta Derm Venereol* 338–341 (1990).

57. Viner, R. M., Teoh, Y., Williams, D. M., Patterson, M. N. & Hughes, I. A. Androgen insensitivity syndrome: a survey of diagnostic procedures and management in the UK. *Arch Dis Child* **77**, 305–309 (1997).

58. Jagiello, G. & Atwell, J. D. PREVALENCE OF TESTICULAR FEMINISATION. *The Lancet* **279**, 329 (1962).

59. Su, Z. *et al.* Common variants at the MHC locus and at chromosome 16q24.1 predispose to Barrett's esophagus. *Nat. Genet.* **44**, 1131–1136 (2012).

60. Oshimori, N., Li, X., Ohsugi, M. & Yamamoto, T. Cep72 regulates the localization of key centrosomal proteins and proper bipolar spindle formation. *EMBO J.* **28**, 2066–2076 (2009).

61. Choi, J. S., Kim, S. R., Jeon, Y. W., Lee, K. H. & Rha, H. K. Identification of DNA copy number aberrations by array comparative genomic hybridization in patients

with ruptured intracranial aneurysms. *J. Clin. Neurosci.* **16**, 295–301 (2009).

62. Gharahkhani, P. *et al.* Genome-wide association studies in oesophageal adenocarcinoma and Barrett's oesophagus: a large-scale meta-analysis. *Lancet Oncol.* **17**, 1363–1373 (2016).

63. Gordon, C., Kang, J. Y., Neild, P. J. & Maxwell, J. D. Review article: The role of the hiatus hernia in gastro-oesophageal reflux disease. *Alimentary Pharmacology and Therapeutics* **20**, 719–732 (2004).

64. Weston, A. P., Badr, A. S. & Hassanein, R. S. Prospective Multivariate Analysis of Clinical, Endoscopic, and Histological Factors Predictive of The Development of Barrett's Multifocal High-Grade Dysplasia or Adenocarcinoma. *Am. J. Gastroenterol.* **94**, 3413–3419 (1999).

65. Lo, L., Dormand, E. L. & Anderson, D. J. Late-emigrating neural crest cells in the roof plate are restricted to a sensory fate by GDF7. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 7192–7197 (2005).

66. Berasi, S. P. *et al.* Divergent activities of osteogenic BMP2, and tenogenic BMP12 and BMP13 independent of receptor binding affinities. *Growth Factors* **29**, 128–139 (2011).

67. Wang, Q. W., Chen, Z. L. & Piao, Y. J. Mesenchymal stem cells differentiate into tenocytes by bone morphogenetic protein (BMP) 12 gene transfer. *J. Biosci. Bioeng.* **100**, 418–422 (2005).

68. Hou, Y. *et al.* Tenomodulin highly expressing MSCs as a better cell source for tendon injury healing. *Oncotarget* **8**, 77424–77435 (2017).

69. Olafsdottir, T. *et al.* Genome-wide association identifies seven loci for pelvic organ prolapse in Iceland and the UK Biobank. *Commun. Biol.* **3**, (2020).

70. Jones, G. T. *et al.* Meta-Analysis of Genome-Wide Association Studies for

Abdominal Aortic Aneurysm Identifies Four New Disease-Specific Risk Loci. *Circ. Res.* **120**, 341–353 (2017).

71. Maguire, L. H. *et al.* Genome-wide association analyses identify 39 new susceptibility loci for diverticular disease. *Nat. Genet.* **50**, 1359–1365 (2018).

72. Castillo, D. *et al.* Activation of the BMP4 Pathway and Early Expression of CDX2 Characterize Non-specialized Columnar Metaplasia in a Human Model of Barrett's Esophagus. *J. Gastrointest. Surg.* **16**, 227–237 (2012).

73. Palles, C. *et al.* Polymorphisms near TBX5 and GDF7 are associated with increased risk for Barrett's esophagus. *Gastroenterology* **148**, 367–378 (2015).

74. Becker, J. *et al.* The Barrett-associated variants at GDF7< and TBX5 also increase esophageal adenocarcinoma risk. *Cancer Med.* **5**, 888–891 (2016).

75. Palles, C., Findlay, J. M. & Tomlinson, I. Common variants confer susceptibility to Barrett's esophagus: Insights from the first genome-wide association studies. *Adv. Exp. Med. Biol.* **908**, 265–290 (2016).

76. Loeys, B. L. *et al.* A syndrome of altered cardiovascular, craniofacial, neurocognitive and skeletal development caused by mutations in TGFBR1 or TGFBR2. *Nat. Genet.* **37**, 275–281 (2005).

77. Moustakas, A. & Heldin, C. H. The regulation of TGFβ signal transduction. *Development* **136**, 3699–3714 (2009).

78. Lindsay, M. E. *et al.* Loss-of-function mutations in TGFB2 cause a syndromic presentation of thoracic aortic aneurysm. *Nat. Genet.* **44**, 922–927 (2012).

79. Ritelli, M. *et al. Further delineation of Loeys-Dietz syndrome type 4 in a family with mild vascular involvement and a TGFB2 splicing mutation*. (2014). doi:10.1186/s12881-014-0091-8

80. Gao, X. R., Huang, H., Nannini, D. R., Fan, F. & Kim, H. Genome-wide

association analyses identify new loci influencing intraocular pressure. *Hum. Mol. Genet.* **27**, 2205–2213 (2018).

81. Gao, X. *et al.* Genome-wide association study identifies WNT7B as a novel locus for central corneal thickness in Latinos. *Hum. Mol. Genet.* **25**, 5035–5045 (2016).
82. Artigas, M. S. *et al.* Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function. *Nat. Genet.* **43**, 1082–1090 (2011).
83. Cho, M. H. *et al.* Risk loci for chronic obstructive pulmonary disease: A genome-wide association study and meta-analysis. *Lancet Respir. Med.* **2**, 214–225 (2014).
84. Blatnik, J. A., Krpata, D. M., Novitsky, Y. W. & Rosen, M. J. Does a history of wound infection predict postoperative surgical site infection after ventral hernia repair? *Am. J. Surg.* **203**, 370–374 (2012).
85. Hultin, S. *et al.* AmotL2 links VE-cadherin to contractile actin fibres necessary for aortic lumen expansion. doi:10.1038/ncomms4743
86. Wang, Y. *et al.* Angiomotin-like2 gene (amotl2) is required for migration and proliferation of endothelial cells during angiogenesis. *J. Biol. Chem.* **286**, 41095–41104 (2011).
87. Brett, M. S. *et al.* De novo 3q22.1 q24 deletion associated with multiple congenital anomalies, growth retardation and intellectual disability. *Gene* **517**, 82–88 (2013).
88. Wolstenholme, J. *et al.* Blepharophimosis sequence and diaphragmatic hernia associated with interstitial deletion of chromosome 3 (46,XY,del(3)(q21q23)). *J Med Genet* **31**, 647–648 (1994).

89. Mead, T. J. & Apte, S. S. ADAMTS proteins in human disorders. *Matrix Biology* **71–72**, 225–239 (2018).

90. Jacobi, C. L. J., Rudigier, L. J., Scholz, H. & Kirschner, K. M. Transcriptional regulation by the wilms tumor protein, Wt1, suggests a role of the metalloproteinase Adamts16 in murine genitourinary development. *J. Biol. Chem.* **288**, 18811–18824 (2013).

91. Richter, H. E. *et al.* Genetic contributions to urgency urinary incontinence in women. *J. Urol.* **193**, 2020–2027 (2015).

92. Segev, Y. *et al.* Are women with pelvic organ prolapse at a higher risk of developing hernias? *Int. Urogynecol. J. Pelvic Floor Dysfunct.* **20**, 1451–1453 (2009).

93. Colige, A. *et al.* Novel types of mutation responsible for the dermatosparactic type of Ehlers-Danlos syndrome (type VIIC) and common polymorphisms in the ADAMTS2 gene. *J. Invest. Dermatol.* **123**, 656–663 (2004).

94. Tortorella, M. D. *et al.* Purification and cloning of aggrecanase-1: A member of the ADAMTS family of proteins. *Science (80-.).* **284**, 1664–1666 (1999).

95. Hatano, E. *et al.* Expression of ADAMTS-4 (aggrecanase-1) and possible involvement in regression of lumbar disc herniation. *Spine (Phila. Pa. 1976).* **31**, 1426–1432 (2006).

96. Brocker, C. N., Vasiliou, V. & Nebert, D. W. Evolutionary divergence and functions of the ADAM and ADAMTS gene families. *Hum. Genomics* **4**, 43–55 (2009).

97. Li, C., Scott, D. A., Hatch, E., Tian, X. & Mansour, S. L. Dusp6 (Mkp3) is a negative feedback regulator of FgF-stimulated ERK signaling during mouse development. *Development* **134**, 167–176 (2007).

98. Rauen, K. A., Cotter, P. D., Bitts, S. M., Cox, V. A. & Golabi, M. Cardio-facio-cutaneous syndrome phenotype in an individual with an interstitial deletion of 12q: Identification of a candidate region for CFC syndrome. *Am. J. Med. Genet.* **93**, 219–222 (2000).

99. Chrzanowska, K., Fryns, J.-P. & Van den Berghe, H. Cardio-facio-cutaneous (CFC) syndrome: Report of a new patient. *Am. J. Med. Genet.* **33**, 471–473 (1989).

100. Primatesta, P. & Goldacre, M. J. Inguinal hernia repair: Incidence of elective and emergency surgery, readmission and mortality. *Int. J. Epidemiol.* **25**, 835–839 (1996).

101. Kingsnorth, A. & LeBlanc, K. Hernias: Inguinal and incisional. in *Lancet* **362**, 1561–1571 (Elsevier Limited, 2003).

102. Dahlstrand, U., Wollert, S., Nordin, P., Sandblom, G. & Gunnarsson, U. Emergency femoral hernia repair: A study based on a national register. *Ann. Surg.* **249**, 672–676 (2009).

4.7. Chapter Appendix

The appendix for this chapter is provided as an online supplement at the following URL: [***bit.ly/WAhmed_C4Appendix***](bit.ly/WAhmed_C4Appendix)

Table of Contents

1. Appendix Tables

Appendix Table 4.1. Phenotype codes used for four individual hernia case definitions

Appendix Table 4.2. Individual hernia associated exonic variants

Appendix Table 4.3. Predicted functional intronic and intergenic variants associated with the four individual hernia phenotypes

Appendix Table 4.4. Genome-wide gene-based association analysis for inguinal hernia in MAGMA

Appendix Table 4.5. Summary-based Mendelian Randomisation (SMR) for inguinal hernia using eQTL data from GTEx

Appendix Table 4.6. Genes mapped to the inguinal hernia-associated loci using the four mapping strategies

Appendix Table 4.7. Genome-wide gene-based association analysis for umbilical hernia in MAGMA

Appendix Table 4.8. Genome-wide gene-based association analysis for hiatus hernia in MAGMA

Appendix Table 4.9. Genes mapped to the hiatus hernia-associated loci using the four mapping strategies

Appendix Table 4.10. Predicted functional intronic and intergenic variants associated with overlap hernia

Appendix Table 4.11. Umbrella hernia associated exonic variants

Appendix Table 4.12. Predicted functional intronic and intergenic variants associated with umbrella hernia

Appendix Table 4.13. Genome-wide gene-based association analysis for overlap hernia in MAGMA

Appendix Table 4.14. Genes mapped to the overlap hernia-associated loci using the four mapping strategies

Appendix Table 4.15. Genome-wide gene-based association analysis for umbrella hernia in MAGMA

Appendix Table 4.16. Summary-based Mendelian Randomisation (SMR) for umbrella hernia using eQTL data from GTEx

Appendix Table 4.17. Genes mapped to the umbrella hernia-associated loci using the four mapping strategies

Appendix Table 4.18. Enriched gene sets from the genome-wide gene-based enrichment analysis of overlap hernia in MAGMA

Appendix Table 4.19. Gene-based enrichment analysis for overlap hernia associated genes in eXploring Genomic Relations.

Appendix Table 4.20. Enriched gene sets from the genome-wide gene-based enrichment analysis of umbrella hernia in MAGMA

2. Appendix Figures

Appendix Figure 4.1. Regional Locus Zoom plots for all four individual hernia associated signals

Appendix Figure 4.2. Quantile-quantile (Q-Q) plots for all four individual hernia analyses

Appendix Figure 4.3. Venn diagram for the 101 genes prioritised at the inguinal hernia associated loci

Appendix Figure 4.4. Venn diagram for the 15 genes prioritised at the hiatus hernia associated loci

Appendix Figure 4.5. Regional Locus Zoom plots of all overlap hernia associated signals

Appendix Figure 4.6. Regional Locus Zoom plots of all umbrella hernia associated signals

Appendix Figure 4.7. Quantile-quantile (Q-Q) plot of all overlap hernia associated signals

Appendix Figure 4.8. Quantile-quantile (Q-Q) plot of all overlap hernia associated signals

Appendix Figure 4.9. Venn diagram for the 3 genes prioritised at the overlap hernia associated loci

Appendix Figure 4.10. Venn diagram for the 129 genes prioritised at the umbrella hernia associated loci

Appendix Figure 4.11. MAGMA tissue expression analysis of umbrella hernia

Appendix Figure 4.12. Regional Locus Zoom plots for all four MTAG hernia associated signals

Appendix Figure 4.13. New 1q41(*TGFB2*) locus discovered to associate with femoral hernia through MTAG multi-trait analysis

Chapter 5: Conclusion

5.1. Conclusion

5.1.1. Summary

Chapter 1 I discussed the complexity of the extracellular matrix (ECM) and the mammalian matrisome, an organ with a fundamental role in structural integrity of tissues and roles exceeding far beyond this, including in important cellular processes. The ECM composition of tissues is delicately balanced by the complex interactions of proteases and their natural inhibitors. Dysfunction in the tightly controlled balance of elastic and collagen fibre components disrupts tissue homeostasis, which can alter the elastic properties of tissues and affect their ability to recover after deformation. This can prevent tissues from performing their normal physiological roles, which, in turn, can lead to disorders of elastic tissue. I then provided evidence for my three diseases of interest – varicose veins, haemorrhoids and hernia – being elastic tissue disorders with a complex aetiology and demonstrated a paucity in the understanding around their genetic susceptibility. Finally, I discussed the merits of GWAS as the current best way to study the genetic architecture of common disorders and introduce the UK Biobank as a truly ground-breaking resource that makes large-scale genetic association studies feasible.[1]

Chapter 2 To unravel the genetic architecture of varicose veins, I performed the largest two-stage GWAS of varicose veins in 810,625 individuals from the UK Biobank and replication in 23andMe. I identified 49 signals at 46 risk loci (29 novel) associated with varicose veins, and using a plethora of bioinformatic tools, I prioritised functional variants and annotated these loci to over 200 genes, and demonstrated therapeutic tractability of several important genes including several pertaining to ECM regulation.

Importantly, using the GWAS-derived variants as genetic predictors for varicose veins, I constructed a weighted genetic risk score (wGRS) that correlated with disease severity, which may foreseeably pave the way towards future personalised medicine approaches.

Chapter 3 I performed the first ever GWAS of haemorrhoids disease in over 400,000 individuals from the UK Biobank, I identified 12 novel loci associated with haemorrhoids, with prioritised genes showing remarkable clustering in pathways relating to ECM remodelling, TGF-β signalling, and internal anal sphincter tonicity. Moreover, almost one-fifth of prioritised genes had known pharmaceutical interactions and are currently being investigated in other disorders. Importantly, the wGRS again correlated with severity of haemorrhoids, and suggests that genetic susceptibility is an important contributor to haemorrhoids biology.

Chapter 4 Using my knowledge of GWAS from **Chapter 2 and 3**, I performed a novel GWAS to investigate the shared genetic susceptibility between multiple hernia phenotypes in UK Biobank (namely inguinal, femoral, umbilical, and hiatus hernia). I performed this principally through two approaches i) association studies of each hernia individually, ii) association analyses of the hernia phenotypes combined – using both linear mixed-model[2] approaches and multi-trait meta-analyses approaches (MTAG and metaUSAT).[3,4] I uncovered 38 loci (52 signals) associated with the four hernia traits individually, and through multi-trait meta-analysis identified six biologically relevant putative loci that demonstrate the highest degree of shared susceptibility across the hernia phenotypes. The multi-trait meta-analysis approaches also uncovered six predisposing genetic variants that were not revealed by the individual

analyses that demonstrate shared susceptibility. Through wGRS analysis, I demonstrate that genetic burden also correlates with disease severity across all examined hernia phenotypes.. All in all, these results provide convincing new evidence of a shared genetic susceptibility to hernia risk.

5.1.2. Implications

Despite the prominence of GWA studies over the past fifteen years, existing studies have largely overlooked the genetic susceptibility to these three surgical elastic tissue disorders.[5] In this MSc thesis I presented data which advances the field of study around surgical disorders and complex trait genetics. Collectively, I have identified a total of 114 independent associated signals at 86 susceptibility loci associated with varicose veins, haemorrhoids and individually to all four hernia phenotypes (inguinal, femoral, umbilical and hiatus hernia). This plethora of rich associations would not have been possible without augmenting the case numbers by the use of operative codes to identify additional individuals who had undergone surgery for the respective diseases. Indeed, the study of surgical disorders lends itself well to biobank-scale cohorts like the UK Biobank, due to the fact that individuals undergoing surgery are more likely to be on the phenotypically severe end of the spectrum and are therefore more likely to be true cases. However, this is often neglected in existing GWA studies, resulting in misclassification bias and weak cohort definitions.[6]

For the hernia analysis, several notable loci were identified only after multi-trait meta-analysis, and this finding is important in demonstrating the use of multi-trait approaches which are able to leverage larger sample sizes to delineate interesting

predisposing genetic variants that would not have been identified through single-trait approaches alone. In total, at the associated loci, I mapped over 400 putative genes, many of which demonstrated evidence of therapeutic tractability and warrant further investigation as potential therapeutic targets. The most striking clustering of genes was seen for those pertaining to ECM components, meaning these complex disorders are ones in which elastic tissue dysfunction plays a central role in their pathobiology – this too is a novel and insightful finding which helps to advance our understanding of their pathobiology. Aside from candidate discoveries, perhaps the most compelling result has been that for all three studied disorders, the wGRS correlated with disease severity, with patients undergoing surgery (across all three disorders) having a higher wGRS than cases managed non-surgically. These results represent an important proof-of-principle of the future utility of genetic risk scoring in personalised medicine approaches to these disorders.

5.1.2. Future plans

After my final year of medical school I hope to return to Oxford as an AFP doctor where I will extend my research enquires into the genetics of other elastic tissue disorders. I hope to perform a 'pan-elastinopathy' GWAS of all common elastic tissue disorders in UK Biobank (**Figure 1.2**), with replication from 23andMe and the NIH AllofUs Biobank (~1 million individuals) which is planned for initial release in two years' time.[7] These research efforts would enable a greater understanding of the genetic basis of common elastic tissue disorders, and guide both therapeutic approaches and risk prognostication efforts to improve patient outcomes in the long term.

5.2. Chapter references

1. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med.* **12**, (2015).
2. Loh, P. R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
3. Turley, P. *et al.* Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Genet.* **50**, 229–237 (2018).
4. Ray, D. & Boehnke, M. Methods for meta-analysis of multiple traits using GWAS summary statistics. *Genet. Epidemiol.* **42**, 134–145 (2018).
5. Grant, Y., Onida, S. & Davies, A. Genetics in chronic venous disease. *Phlebol. J. Venous Dis.* **32**, 3–5 (2017).
6. Fukaya, E. et al. Clinical and Genetic Determinants of Varicose Veins. *Circulation* 1–12 (2018). doi:10.1161/CIRCULATIONAHA.118.035584
7. Denny, J. C. *et al.* The “All of Us” Research Program. *N. Engl. J. Med.* **381**, 668–676 (2019).

5.3. Chapter Appendix

5.3.1 Academic development and contributions at Oxford

Academic achievements since the beginning of my MSc Programme at the University of Oxford.

i) Awards

- **Best Oral Presentation Prize**, Botnar Research Centre Student Symposium 2020, University of Oxford, NDORMS. Awarded best presentation prize alongside Mr Conrad Harrison. Presentation titled: 'Genome-wide association analysis and replication in 810,625 individuals identifies novel therapeutic targets for varicose veins'.
- **BJS Certificate for High Altmetric Score,** British Journal of Surgery. In recognition of Br J Surg. 2020 Jul;107(8):1023-1032. doi: 10.1002/bjs.11453.
- **Mitchell Powell Fund Award for Academic Development**, Lady Margaret Hall College, Oxford University -- £300 travelling grant to present at Royal College Surgeons, Surgical Research Society (SRS) 2020 Conference, Dublin, Ireland.
- **RCS/NRCM Conference Traveling Award**, National Research Collaborative Meeting (NRCM) 2019, Newcastle, UK; Royal College Surgeons/National Research Collaborative, £100

ii) Scholarships

- **Aziz Scholarship**, Aziz Foundation. Recipient of tuition fee scholarship from the Aziz Foundation to cover tuition fees for my MSc Programme. £7,730.
- **Wolfson Scholarship,** Wolfson Foundation, £5000 to cover maintenance.
- **Intercalated Degree Scholarship,** Royal College Surgeons of England, £5000 to cover maintenance.

iii) Grants

- **INSPIRE Small Grant Award,** The Academy of Medical Sciences, £20,000 awarded to develop INCEPT (clINical aCademia Elearning PlaTform), University of Edinburgh and STARSurg Collaborative (Co-applicant). Feb 2020 – Feb 2022.
- **Research Stipend,** NIHR Oxford Biomedical Research Centre (BRC), Musculoskeletal theme (2019-20). £5,500.

- **Bursary for Undergraduate Research (2018),** Royal College Surgeons of Edinburgh. £1500 awarded during the Summer of 2018 to visit the Furniss Group.
- **Student Research Grant (2018)**, British Association of Plastic, Reconstructive, Aesthetic Surgeons (BAPRAS). £500 awarded during the Summer of 2018 to visit the Furniss Group.

iv) Publications

Full-length papers:

1. Characteristics and outcomes of 627 044 COVID-19 patients with and without obesity in the United States, Spain, and the United Kingdom
MedRxiv Preprint. [Posted online 2020 Sep 3]. doi: 10.1101/2020.09.02.20185173
Recalde M*, Roel E*, Pistillo A,, Sena AG, Prats-Uribe A, Ahmed WUR, Alghoul H, Alser O, Areia C, Burn E, Casajust P, Dawoud D, DuVall SL, Falconer T, Fernández-Bertolín S, Golozar A, Gong M, Lai LYH, Lane JCE, Lynch KE, Matheny ME, Mehta PP, Morales DR, Natarjan K, Nyberg F, Posada JD, Reich CG, Schilling LM, Shah K, Shah NH, Subbian V, Zhang L, Zhu H, Ryan P, Prieto-Alhambra D, Kostka K*, Duarte-Salles T*.

2. Management of COMPlicAted intra-abdominal collectionS after colorectal surgery (COMPASS): protocol for a multicentre, observational, prospective, international study of drain placement practices in colorectal surgery
Colorectal Dis. 2020 Jul 27. doi: 10.1111/codi.15275. PMID: 32716111.
Ahmed WUR (writing and study management group, 2nd author); EuroSurg Collaborative

3. Perioperative non-steroidal anti-inflammatory drugs (NSAID) administration and acute kidney injury (AKI) in major gastrointestinal surgery: A prospective, multicenter, propensity matched cohort study
Annals of Surgery [Accepted: Jun 27, 2020]
Ahmed WUR (writing group, 2nd author); STARSurg Collaborative

4. Genome-wide association analysis and replication in 810,625 individuals identifies novel therapeutic targets for varicose veins
BioRxiv Preprint. [Posted online 2020 May 16]. doi: 10.1101/2020.05.14.095653
Ahmed WUR, Wiberg A, Ng M, Wang Wei, Auton A, 23andMe Research Team, Lee R, Handa A, Zondervan KT, Furniss D.

5. Perioperative intravenous contrast administration and the incidence of acute kidney injury in major gastrointestinal surgery: prospective, multicentre cohort study

Br J Surg. 2020. [published online ahead of print, 2020 Feb 5]. PMID: 32026470
Ahmed WUR (writing group, 2nd author); STARSurg Collaborative

6. Safety of Hospital Discharge Prior to the Return of Bowel Function After Colorectal Surgery: International, Prospective, Cohort Study
Br J Surg, 107 (2), e161-e169 Jan 2020. doi: 10.1002/bjs.11326. PMID: 31595986
Ahmed WUR (data collector); EuroSurg Collaborative

7. Core Content of the Medical School Surgical Curriculum: Consensus Report from the Association of Surgeons in Training (ASIT)
Int J Surg 09 Jan 2020. doi:10.1016/j.ijsu.2019.12.036. PMID: 31926325
Ahmed WUR (authorship consensus group); ASiT Medical Student and Foundation Doctor Consensus Group

8. REspiratory COmplications after abdomiNal Surgery (RECON): Study Protocol for a Multi-centre, Observational, Prospective, International Audit of Postoperative Pulmonary Complications Following Major Abdominal Surgery
Br J Anaesth. 2020 Jan; 124(1):e13-e16. doi: 10.1016/j.bja.2019.10.005. PMID: 31784036
Ahmed WUR (writing, steering & data management group, joint 2nd author); STARSurg Collaborative

9. Maximizing opportunities at medical school to support a career in surgery
Br J Hosp Med (Lond). 2019 Nov 2;80(11):670-673. doi: 10.12968/hmed.2019.80.11.670. PMID: 31707879
Ahmed WUR (writing group, 2nd author); STARSurg Collaborative

10. Safety and Efficacy of Non-Steroidal Anti-Inflammatory Drugs to Reduce Ileus After Colorectal Surgery
Br J Surg. 2019 Oct 9. [Epub ahead of print] doi: 10.1002/bjs.11326. PMID: 31595986
Ahmed WUR (data collector); EuroSurg Collaborative

Letters:

1. **Author response to: Does preoperative contrast administration really not affect the occurrence of acute kidney injury after major gastrointestinal surgery?**

Br J Surg. 2020;10.1002/bjs.11630. doi:10.1002/bjs.11630. PMID: 32379340
McLean KA, Ahmed WU, Akhbari M, Nepogodiev D.

v) Presentations

Oral Presentations:

1. Ahmed WUR, Wiberg A, Ng M, Wang Wei, Auton A, 23andMe Research Team, Lee R, Handa A, Zondervan KT, Furniss D. Genome-wide association analysis and replication in 810,625 individuals identifies novel therapeutic targets for varicose veins. Surgical Research Society (SRS) Meeting 2020, Dublin, Ireland, 13 Oct 20.

2. Ahmed WUR, Wiberg A, Ng M, Smart NJ, Zondervan KT, Furniss D. Genome-wide association analysis in 401,583 individuals identifies novel therapeutic targets for haemorrhoids. European Society of Coloproctology 15th Scientific and Annual Meeting, Vilnius, Lithuania, 21-23 Sep 20.

3. Ahmed WUR (Writing Group) on behalf of the STARSurg Collaborative. The Effects of Perioperative Intravenous Contrast on Acute Kidney Injury in Major Gastrointestinal Surgery. European Society of Coloproctology 14th Scientific and Annual Meeting, Vienna, Austria, 25-27 Sep 19.

4. STARSurg Collaborative (Ahmed WUR, Writing Group). Perioperative Non-Steroidal Anti-Inflammatory Drugs Administration and Acute Kidney Injury in Major Gastrointestinal Surgery: A National, Multicentre, Propensity Matched Cohort Study. European Society of Coloproctology 14th Scientific and Annual Meeting, Vienna, Austria, 25-27 Sep 19.

Poster Presentations:

1. Ahmed WUR, Wiberg A, Ng M, Wang Wei, Auton A, 23andMe Research Team, Lee R, Handa A, Zondervan KT, Furniss D. Genome-wide association study of varicose veins. Oxford Statistical Genetics Initiative 2020, Big Data Institute, University of Oxford. Feb 2020

Invited Speaker:

1. Ahmed WUR. Overcoming the Odds: From Luton Sixth Form College to Oxford University (Keynote). Luton Sixth Form College, Annual Heroes Awards Evening 2019. 16 Dec 2019.
2. Baker P, Scrimshire A, Nepogodiev D, Ahmed WUR (Co-Chair). ‘Big Data: Big Networks’ Conference Session. National Research Collaborative Meeting (NRCM) 2019 Conference ‘Frontiers in collaboration’, Newcastle, England. 6 Dec 2019
3. Ahmed WUR, Akhbari M. How to get involved with research and audit as a medical student (Workshop). Young Researchers Conference, Imperial College London, Society of Research and Academia. 30 Nov 2019.

VI) Appointments

- **Honorary Research Fellow**, University of Birmingham, Institute of Applied Health Research. Jan 2020 - Present.
- **Honorary Departmental Research Assistant**, University of Oxford, NDORMS, Sep 2018 - Sep 2019.

VII) Leadership and Management

- **CHOLECOVID Collaborative**, Steering Committee, Apr 2020– Present
- **EuroSurg Collaborative,** Steering Committee, Jan 2020 – Present
- **STARSurg Collaborative**, Steering Committee, Aug 2018 – Present.