# The Molecular Genetics of Dupuytren Disease

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

Doctor of Philosophy




Michael Ng

St Cross College

Hilary Term 2018

# Abstract

The molecular genetics of Dupuytren Disease

*Doctor of Philosophy in Molecular and Cellular Medicine*

*Michael Tsz Hin Ng, St Cross College, Hilary 2018*

Dupuytren disease (DD) is a fibromatosis of the fascia of the hand. It affects around 4% of the population in the UK. Treatment for DD is dominated by surgery, but the recurrence rate after surgery is high. There are also no existing treatments to prevent progression of early disease, and reduce the need for surgery. At the cellular level, the fibrotic environment in DD is similar to that of other tissue specific fibrotic condition of other organs, such as lung and liver where myofibroblasts and some immune cells are found residing in a rich extracellular matrix. Although myofibroblast cells are thought to be the driving force in the development of DD, no effective biomarkers have been identified to date. Many different cell types have been suggested to be the origin of the myofibroblast cell population, but often without any solid evidence. As a result, the aetiology of DD remains poorly understood. The advent of high through put sequencing and genotyping techniques provided me with a new approach to the investigation of the pathophysiology of DD.

DD is a complex genetic disease, with a heritability of 80% - very similar to type I diabetes. Currently, the most effective method to interrogate a complex genetic disease is a GWAS. In the second chapter of my thesis, I report a total of 17 novel associated variants discovered with this approach. I prioritised variants for functional mechanistic studies through fine mapping and *in silico* analysis using epigenomic and transcriptomic public databases. Follow-up functional analysis on *SFRP4* and *DDR2*, in chapters II and III respectively, assess transcriptomic expression and various fibrotic phenotypes such as cell migration in diseased myofibroblasts resected from patients with DD. These analyses

have revealed the pro-fibrotic nature of these proteins, and suggest potential new therapeutic targets. In studying DDR2, I have also demonstrated that different collagen types can alter the myofibroblast cell function from pro-fibrotic to pro-inflammatory.

To further investigate the 17 novel variants that I discovered in the GWAS, I produced DD specific epigenomic data, and discovered new associations between variants and their targeted transcriptome. I also discovered that most variants control multiple genes concurrently.

As mentioned earlier, DD is pathophysiologically similar to many other organ-specific fibrotic diseases, but this has never been investigated from a genetic viewpoint. In chapter VI, I explored this avenue and identified a total of 18 cross-fibrosis associated variants, including some associations that would not have been discovered by studying a single disease in isolation. In the penultimate chapter, I have derived a weighted genetic risk score that is associated with recurrence after surgery, independent of high-risk clinical features. This may have clinical implications in the short term for patient counselling. However, the greatest impact of this work is likely to come after more functional mechanistic work that may lead to the discovery of better treatment methods.

In summary, I used genetic approaches to successfully provide a plethora of new information regarding DD development.

# Acknowledgements

I would like to give thanks to my supervisor Professor Dominic Furniss and Dr James Dunford, who gave me the opportunity to achieve this academic milestone. Especially Dominic, I am incredibly grateful for his managerial wisdom, advices, and most importantly his unwavering support throughout this eventful and fulfilling journey.

I would also like to extend my thanks to Dr Sarah Snelling, who has generously given me her valuable time to examine my work, and her expertise in molecular biology.

Lastly, I would like to thank my mum and dad, who put me through eight years of education in the UK and gave me the moral support when I needed it the most.

# Disclaimer

Every piece of work that is described in this thesis was conducted solely by me, Michael Ng apart from the following,

Chapter 1. Consenting Dupuytren disease patients for their genetic material and tissue sample for the functional work. Genotyping BSSH-GODD samples on the DNA array chip.

Chapter 4. Any work pertaining to the acquisition of the RNA sequencing data.

Chapter 5. The acquisition of medical records of all participants that were included in the study.

# Contents

**Chapter 1 Introduction ..... 15**

**1.1 Epidemiology of Dupuytren disease (DD) ..... 15**

**1.2 Evidence of genetic predisposition to Dupuytren disease ..... 17**

**1.3 Genetics of complex diseases ..... 19**

**1.4 The microscopic environment of fibrotic diseases ..... 21**

**1.5 The extracellular matrix of fibrotic tissue ..... 25**

**1.6 Currently known signalling pathways of fibrosis ..... 27**

1.6.1 Transforming growth factor beta pathway ..... 28

1.6.2 WNT signalling pathway ..... 30

**Chapter 2 The genetics of Dupuytren Disease ..... 34**

**2.1 Abstract ..... 34**

**2.2 Introduction ..... 36**

**2.3 Aims and hypotheses ..... 43**

**2.4 Materials and methods ..... 44**

2.4.1 Ethical approval ..... 44

2.4.2 Saliva DNA extraction ..... 44

2.4.3 Discovery phase genome wide association study ..... 45

2.4.4 Genome wide association study data quality control ..... 46

2.4.5 GWAS analysis ..... 47

2.4.6 Replication and meta-analysis ..... 48

2.4.7 Pre-phasing, imputation and post-imputation quality control ..... 49

2.4.8 Bayesian framework imputation analysis and fine mapping ..... 50

2.4.9 *In silico* analysis ..... 50

2.4.10 Myofibroblast tissue culture ..... 51
2.4.11 Tissue processing and immunohistochemistry ..... 52
2.4.12 Immunocytochemistry ..... 53
2.4.13 RNA extraction and gene expression assay ..... 54
2.4.14 Enzyme-linked immunosorbent assay (ELISA) ..... 55
2.4.15 Statistical analysis of functional work ..... 56
**2.5 Results ..... 57**
2.5.1 Quality control of the genotyped data ..... 57
2.5.2 Genome wide association study results ..... 60
2.5.3 Fine mapping of the DD associated loci ..... 74
2.5.4 The functional evidence of the most significantly associated locus with Dupuytren disease ..... 90
2.5.5 The molecular function of SFRP4 in myofibroblasts ..... 92
2.5.6 *In silico* analysis of the remaining associated loci ..... 99
**2.6 Discussion ..... 103**
**Chapter 3 Discoidin Domain Receptors and their functions in DD ..... 111**
**3.1 Abstract ..... 111**
**3.2 Introduction ..... 113**
**3.3 Aims and hypotheses ..... 119**
**3.4 Materials and methods ..... 120**
3.4.1 Protein knockdown ..... 120
3.4.2 RNA extraction ..... 121
3.4.3 Reverse transcription and qPCR ..... 122
3.4.4 Western blot ..... 123
3.4.5 Tissue collection, cell disaggregation and myofibroblast cell culture ..... 123
3.4.6 Proliferation assay ..... 124

3.4.7 Migration assay ..... 125
3.4.8 Traction force microscopy ..... 126
3.4.9 Sequencing library preparation and high-throughput sequencing ..... 127
3.4.10 RNA-seq data analysis ..... 129
**3.5 Results ..... 131**
3.5.1 The confirmation of DDR2 knockdown ..... 131
3.5.2 The function of DDR2 in DD myofibroblast proliferation ..... 132
3.5.3 The function of DDR2 in DD myofibroblast migration ..... 134
3.5.4 The function of DDR2 in cell contraction ..... 137
3.5.5 DDR2 and downstream gene regulation ..... 139
**3.6 Discussion ..... 145**
**Chapter 4 Genetic fine mapping of DD myofibroblasts and fibroblasts .. 148**
**4.1 Abstract ..... 148**
**4.2 Introduction ..... 151**
**4.3 Aims and hypotheses ..... 156**
**4.4 Materials and methods ..... 157**
4.4.1 Tissue culture and sample processing for ATAC-seq ..... 157
4.4.2 Sample processing for circularised chromatin conformation capture sequencing (4C-seq) ..... 159
4.4.3 ATAC sequencing data processing ..... 162
4.4.4 4C sequencing data processing ..... 163
4.4.5 Integration of RNA-seq data with ATAC-seq data ..... 164
**4.5 Results ..... 165**
4.5.1 Quality control of ATAC-seq data ..... 165
4.5.2 The distribution of open chromatin regions in myofibroblast cells ..... 169
4.5.3 Functional confirmation of Dupuytren associated variants ..... 174

4.5.4 Differential chromatin accessibility between Dupuytren myofibroblasts and fibroblasts from the palmar skin ..... 180

**4.6 Discussion ..... 186**

**Chapter 5 Dupuytren disease genetic risk score and recurrence ..... 190**

**5.1 Abstract ..... 190**

**5.2 Introduction ..... 192**

**5.3 Aims and hypotheses ..... 194**

**5.4 Materials and methods ..... 195**

5.4.1 Patient cohorts, clinical data and ethical approval ..... 195

5.4.2 Genetic risk score ..... 196

5.4.3 Imputation of the missing diatheses ..... 196

5.4.4 Statistical analysis ..... 197

**5.5 Results ..... 199**

5.5.1 The demographic of diatheses in DD patients ..... 199

5.5.2 Weighted GRS can predict DD diatheses ..... 203

5.5.3 Weighted GRS and DD diatheses can predict DD recurrence ..... 205

**5.6 Discussion ..... 208**

**Chapter 6 The shared genetic epidemiology of fibrotic diseases ..... 210**

**6.1 Abstract ..... 210**

**6.2 Introduction ..... 212**

6.2.1 Adhesive capsulitis ..... 213

6.2.2 Endometriosis ..... 214

6.2.3 Hepatic fibrosis ..... 214

6.2.4 Idiopathic pulmonary fibrosis ..... 216

6.2.5 Systemic sclerosis ..... 217

6.2.6 Uterine fibroids ..... 218

**6.3 Aims and hypotheses ..... 220**

**6.4 Materials and methods ..... 221**

6.4.1 Quality control and association analysis ..... 221

6.4.2 Phenotypic correlation ..... 222

6.4.3 Meta-analysis ..... 223

6.4.4 LD score regression and Cross-phenotype association analysis (CPASSOC) ..... 223

6.4.5 *In silico* analysis of the cross-trait associations ..... 224

**6.5 Results ..... 225**

6.5.1 Phenotypic correlation between fibrotic conditions ..... 225

6.5.2 Genome wide associations of multiple fibrotic conditions ..... 229

6.5.3 Cross-fibrotic genetic associations ..... 254

6.5.4 *In silico* analysis of the cross-phenotype associations ..... 269

**6.6 Discussion ..... 274**

**Chapter 7 Discussion and future directions ..... 281**

**Appendix Publications associated with this thesis ..... 288**

**Bibliography ..... 289**

# Abbreviations

1KG 1000 genome

4C Circularised chromatin conformation capture

AC Adhesive capsulitis

ACTA2 Alpha smooth muscle actin

ADAM A disintegrin and metalloproteinase

ADAMT ADAM with thrombospondin motifs

ASMA Alpha smooth muscle actin

ATAC Assay for transposase-accessible chromatin

BF Bayes factor

BSSH-GODD British Society for Surgery of the Hand, Genetics of Dupuytren's Disease

CEU 30 trios of Utah residents of northern and western European ancestry

CHIP Chromatin immunoprecipitation

cM Centimorgan

COL Collagen

Ct Cycling threshold

CTGF Connective tissue growth factor

DD Dupuytren disease

DDR Discoidin domain receptor

DM Diabetes mellitus

DNA Deoxyribonucleic acid

EAF Effective allele frequency

ECM Extracellular matrix

EGA European Genome Archive

ELISA Enzyme-linked immunosorbent assay

EMT Endothelial mesenchymal transition

ENDO Endometriosis

EPDR1 Ependymin related 1

eQTLs Expression quantitative trait loci

ERAD Endoplasmic-reticulum-associated protein degradation

FBS Fetal bovine serum

FFBE Formalin fixed paraffin embedded

FGF Fibroblast growth factor

FZD Frizzled

GCTA Genome wide complex trait analysis

GO Gene ontology

GODDAF Genetic Origin of Dupuytren Disease

GRAIL Gene Relationships Across Implicated Loci

GRS Genetic risk score

GWAS Genome wide association study

HA Hyaluronic acid

HES Hospital episode statistics

HF Hepatic fibrosis

HRC Haplotype Reference Consortium

HWE Hardy Weinberg equilibrium

INDELs Insertions and deletions

IPF Idiopathic pulmonary fibrosis

LCR Locus control region

LD Linkage disequilibrium

LDSC LD score regression

LMM Linear mixed model

MAC Minor allele count

MAF Minor allele frequency

Mb Megabase pair

MCAR Missing completely at random

miRNA Micro RNA

MMP Matrix metalloproteinase

mRNA Messenger RNA

NAFLD Non-alcoholic fatty liver disease

OR Odds ratio

PBS Phosphate buffered saline

PC Principle component

PCA Principle component analysis

PDGF Platelet-derived growth factor

PF Pulmonary fibrosis

PRS Polygenic risk score

PS Palmar skin

QC Quality control

Q-Q Quantile quantile

qPCR Quantitative polymerase chain reaction

RNA Ribonucleic acid

ROM Range of motion

RTK Receptor tyrosine kinase

SE Standard error

SEM Standard error of the mean

seq Sequencing

SFRP4 Secreted frizzled-related protein 4

SNP Single nucleotide polymorphism

SSc Systemic sclerosis

TF Transcription factor

TFD Total flexion deformity

TFM Traction force microscopy

TGF-β Transforming growth factor beta

TSS Transcription start site

UKHLS UK Household Longitude Study

| | |
|---|---|
| UF | Uterine fibroids |
| UV | Ultraviolet |
| VCF | Variant call format |
| VEGF | Vascular endothelial growth factor |
| wGRS | Weighted GRS |

## 1.1 Epidemiology of Dupuytren disease (DD)

Dupuytren disease is fibromatosis in the palmar fascia of the hand, resulting in contracture of the affected digit with a limited range of movement (Figure 1.1.1). The clinical presentation of DD consists of a fibrous cord and a cellular nodule.

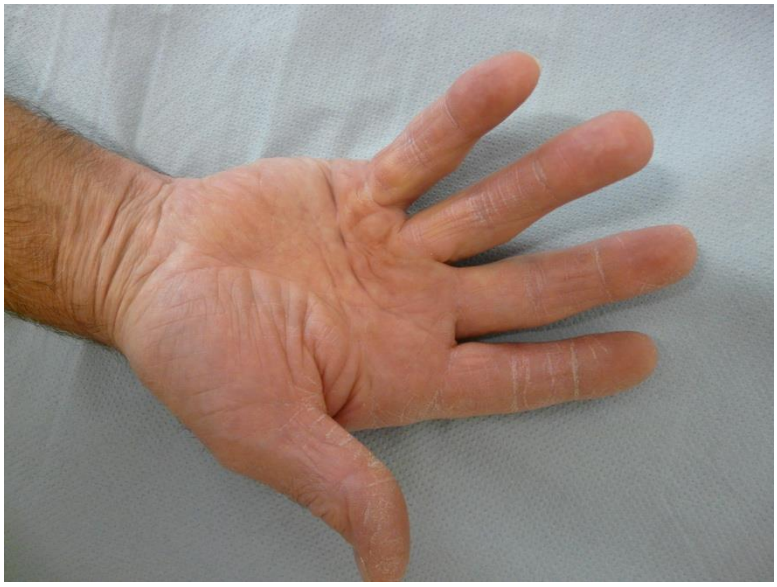


Figure 1.1.1 Picture demonstrating a DD contracture in the fifth digit of the hand (picture provided by Professor Dominic Furniss).

Many different clinical staging and grading systems have been introduced for DD since the naming of the condition, from the first classification proposed by Iselin et al to the more commonly used Tubiana [1]. This classification is divided into five categories, N and one to four, where N represents nodule only, and one to four describe the total flexion deformity (TFD) of the contracture, from the least severe, zero to 45 degrees, to the most severe, greater than 135 degrees. DD is commonly diagnosed in the UK and in Europe, mainly amongst white Caucasians with a prevalence around 4% [2]. Populations of other non-Caucasian ethnicities such as Blacks [3], [4] Asians [5 6] and Hispanics [7] have all reported cases of DD, but at a much lower rate than Caucasians. The age of onset of DD is commonly in the fourth to seventh decades of life, though it has been reported in children[8]. Sex also seems to be one of the deterministic factors in DD, where men are up to five times more likely to develop DD than women [9-11]. In addition, severe symptoms are more commonly found in men than in women [12]. DD has been reported to be associated with a plethora of non-genetic factors, as well as other complex diseases, including heavy manual work[13-15], hand injury, smoking, alcohol consumption, epilepsy, adhesive capsulitis, diabetes mellitus, hypertension, HIV, Ledderhose disease, knuckle pads and Peyrone disease[2]. Current treatment options for DD are limited, and the most common procedures are range of movement (ROM) exercises, injection of collagenase (*Clostridium histolyticum*), and operative procedures including needle aponeurotomy, fasciectomy and dermofasciectomy. The treatment decision is made based on the severity of symptoms. DD recurrence does occur but the reported prevalence varies from study to study, and this largely due to the inconsistency of the diagnostic criteria, and variation in treatment regimen. A new

criterion has been suggested in 2017, redefining the recurrence as a new contracture of >20 degrees in a previously treated joint with a palpable cord [16].

## 1.2 Evidence of genetic predisposition to Dupuytren disease

In addition to the non-genetic risk factors described above, multiple sources of evidence pointing to host genetic predisposition to DD development have also been reported in the literature. One of the early reported pieces of evidence suggesting DD has a genetic component, was a case report presenting a pair of monozygotic twins affected by DD [17]. The results of this case study was then further supported by evidence presented in two separate twin epidemiological studies performed in the UK [18] and Denmark [19]. The latter study has also addressed a genetic phenomenon shown in DD, a discordance in phenotype between monozygotic twins showing a non-Mendelian mode of inheritance, an issue which was first raised in a case report by Lyall [20]. This Danish study has also concluded that the heritability, a proportion of phenotypic variance explainable by host genetic predisposition, of DD is 80% [19], which is significantly higher than other well-known complex diseases such as type II diabetes. Capstick et al have taken this one step further by quantifying the risk of genetic predisposition with sibling recurrence rate, and concluded that a sibling of an affected proband is 4.5 times more likely to be affected by DD than a member of the general population[21]. Another epidemiological study using retrospective clinical data has also demonstrated that a proband with a positive family history of DD has a significantly lower age of onset than a proband without [22].

With strong evidence that genetic factors play a role in the development of DD, numerous attempts have been made to dissect its genetic cause using various approaches, such as linkage with microsatellite markers and candidate gene association studies. Results of these early genetic studies yielded exciting results, where a linkage study using DNA collected from a family of five generations has successfully narrowed down genetic cause to a 6cM region on chromosome 16q [23]. Studies using the candidate gene approach have also yielded some successes in associating polymorphisms of the human genome with DD including *TGFB1*, *TGFBRI*, *Zf9 and HLA-DRB-1,* though these associations suffered from the lack of statistical power and require replication for their validity [24]. Also, using in-situ hybridization techniques, multiple research groups have identified several copy number changes on multiple chromosomes in fibroblast cells that were disaggregated from affected DD tissue[25-28]. This warranted a comprehensive DD genetic association study, and a pan-European cohort GWAS was published in 2011 [29] revealing a total of nine novel associated loci. Using gene relationships across implicated loci (GRAIL) analysis, this study has also predicted six out of the nine regions may functionally interfere with genes such as R-spondin 2 (*RSPO2)*, *WNT2*, *WNT4* and *WNT7B* of the WNT signalling pathway, which has been associated with the regulation of many biological processes including cell migration and proliferation. The authors have subsequently concluded this pathway is likely to be responsible for the manifestation of DD and warranted further investigation of this pathway. Significant and suggestive associations of this study have been confirmed by a later meta-analysis that comprised of three European GWAS datasets but no new significant variants were identified, again likely due to lack of statistical power[30].

Numerous gene expression studies have been conducted in an attempt to validate these gene-to-disease associations and yielded some promising results, showing a difference in gene expression between normal and DD tissue including *WNT2*, *WNT7B*, *SFRP4* [31] and *EPDR1*[32]. Other differentially expressed genes such as *WT1* [33] and multiple microRNAs (miRNAs) that regulate collagen production [34] have been reported, though the latter study also showed conflicting results with WNT pathway genes. No epigenetic study in DD has been published to date. Taken together, these data suggest that DD is a complex disease.

## 1.3 Genetics of complex diseases

A complex disease trait is a condition with an aetiology that is multifactorial, with both genetic and non-genetic predisposing factors, and polygenic. Many highly studied conditions are classified as a complex disease, including Crohn's disease, Schizophrenia, Alzheimer's and diabetes. Complex diseases do not follow the traditional Mendelian mode of inheritance but instead a liability threshold model[35], whereby the manifestation and the severity of the disease is governed by the amount of risk factors an individual is carrying. This threshold model is not limited to gene-to-gene interaction but also gene-to-environment interaction, for example risk allele carriers of the *APOE* gene are more susceptible to develop Alzheimer's with a high cholesterol diet than with a lower one [36 37]. Genetic risk factors in complex diseases have been identified in various forms, such as deletion and insertion, copy number

variation, methylation and most commonly single nucleotide polymorphisms. The traditional candidate gene approach and molecular genetic techniques such as karyotyping, comparative genomic hybridisation, and restriction fragment length polymorphism-based methods would be inadequate for the modern complex diseases association studies due to the shear amount, around 10 million, of unique polymorphisms, or methylation sites existing in the human genome. This has led to the development of high throughput microarray platforms, which are capable of genotyping over half a million SNPs of a DNA sample. This level of coverage is widely considered sufficient for discovering novel associated variants partly due to the design of the microarray chips. Using TagSNPs, a single variant on a chip can effectively [38] determine the nucleotide of all alleles within the same LD block on average between 100kb to 1mb apart in distance. Nevertheless, it is still far from genomic level and researchers often have to decide which platform to select based on their research question.

Sanger sequencing has been the gold standard of genome sequencing for over two decades, but it is seldom used for disease oriented genotyping due to its costly process in money and time, as the procedure involves cloning fragmented sequences into contigs. Even though the invention of high throughput sequencing methods has drastically reduced the time required to sequence an entire human genome, it is presently not a viable option to sequence thousands of samples for their genotype, especially for one disease, given often the interesting functional rare variants only exists in less than 0.01% of any tested populations. Instead, multiple sets of genomic

sequencing data from different resources have been amalgamated into one dataset, which have been used in imputation for GWAS that was obtained using array-based techniques [39]. Sequencing, or the more advanced single-cell sequencing techniques are often reserved for studying quantitative traits, such as differential gene expression, to complement genotyping data gathered using the microarray method.

## 1.4 The microscopic environment of fibrotic diseases

To date, no known gene to environment interactions have been discovered in DD. Pathologically, DD has many similarities with other fibrotic conditions, and therefore existing knowledge of this research area can be used to complement and facilitate DD genetic research. Fibrosis, under normal circumstances, is a tightly regulated event that is activated during wound healing. Its main stages include the attraction of progenitor cells for transdifferentiation to myofibroblasts, and laying down and contraction of extracellular matrix (ECM) molecules to form strong fibril bundles at the wound site to achieve wound healing [40 41]. A fibrotic condition, on the other hand, is an abnormal chronic pathological event, whereby excessive laying down and contraction of ECM leads to a loss of function in the affected organ [40 41]. Occasionally, changes in resident cell plasticity can lead to cancerous transformation, for example in liver fibrosis [42].

The triggers of a fibrotic event can originate from various sources including mechanical injury, toxins, autoimmune, infections and radiation [40 43], however, all share a common theme: inflammation. Fibrosis without inflammation is possible and has been reported in an in vivo study using a mouse model with an inducible *wnt1* expression inside the kidney tubule cells [44]. The findings of this study were not well received within the fibrosis research community, claiming that it would be improbable for this particular profibrotic molecule to be overexpressed in diseased tissue[45], let alone initiating the fibrotic process. However, evidence from genetic association studies on fibroses has shown a mixture of profibrotic and proinflammatory genes, suggesting that inflammation is not the only important pathway in fibrosis. Histologically, tissue from all known fibrotic conditions contain immune cells, and it is difficult to refute the existence of such populations and their possible roles during the early stages of the disease development process [46]. As a result, immunologists are inclined to favour the low chronic inflammatory initiator model [46 47], and disregard the possible disease initiating or perpetuating role of other non-inflammatory cells. To accommodate the mounting evidence of fibrosis associated susceptibility genes, some research groups have suggested the multiple-hit hypothesis, describing the development process with more than one disease initiator [48 49] where, of course, some converge to inflammatory pathways.

Cells from both innate and adaptive immunity pathways have been identified in fibroses, where they work in synchronicity from the early migration, cytokine producing stage of the disease to the later fibrotic, ECM producing stage [46]. Macrophages can

be found in both early and late stages, though with different roles throughout [50 51]. In the early stage of fibrosis, macrophages act as antigen presenting cells, where they produce pro-inflammatory cytokines such as Tumour Necrosis factor alpha (TNF-α) and Interleukin 1 (IL-1), promoting cell growth, cell migration and activation of Th1 lymphocytes [46 47 50], In contrast, at later stages of fibrosis, macrophages produce paracrine, pro-fibrotic factors such as Insulin-like growth factor (IGF-1), vascular endothelial growth factor alpha (VEGF-α) and transforming growth factor beta (TGF-β)[51]. Mast cells are a hub of secreted cytokines, such as IL-4, TNF-α, IL-10 and similar to macrophages, have both proinflammatory and profibrotic functions [52]. In idiopathic pulmonary fibrosis (IPF), the rare chymase tryptase double positive mast cells ($MC_{TC}$) have been identified in the affected lung, expressing profibrotic cytokines such as IL-4 and TGF-β [52 53]. This profibrotic response is attenuated when mast cells are depleted as shown in the Bleomycin treated mast cell deficient mouse model [54 55] . The role of Natural Killer cells has been shown to be negative regulatory; to terminate fibrosis through inducing apoptosis in myofibroblasts[56]. Similar to the described innate immune cells, adaptive immune cells have both profibrotic and proinflammatory roles in fibrosis. T lymphocytes, in particular, have been shown to regulate fibrosis through changes in relative abundance of the three subtypes Th1, Th2, and Th17. Each subtype possesses a specific function; the predominantly proinflammatory Th1 and Th17, and the profibrotic Th2. An imbalance in the subtypes leads to dysregulation of the inflammatory response, and progression to fibrosis[46]. B lymphocytes are antibody producing cells and have been shown to be proinflammatory, and can facilitate autoimmune response through the formation of immune complexes between self-

antigen and antibody. Multiple IPF studies using Bleomycin treated mouse models have shown that the presence of B cells is essential for a fibrotic response [57 58].

Although a plethora of immune cells have been identified in fibrotic tissue, they constitute only a small subset of cells in comparison to myofibroblasts, a cell type that exists only in fibrotic tissue. During early stages of fibrosis or wound healing, progenitor cells in the vicinity of the fibrotic area or in circulation are activated by cytokines from platelets and immune cells, and transformed into proto-myofibroblasts, which can then later mature into myofibroblasts[59 60] Progenitor cells can be generic cells, such as fibrocytes, fibroblasts and pericytes as well as specialised cells that reside in the affected organs, such as alveolar endothelial cells in IPF, hepatic stellate cells in hepatic fibrosis. Regardless of the origins of the progenitor cells, all myofibroblasts possess similar functions, which include proliferation, migration, interaction with immune cells and the ECM, cytokine production, ECM production and degradation, and contraction. Recent data from single cell sequencing studies on various fibroses has revealed new insights on additional and more defined subtypes of cells in IPF [61], as well as casting doubts into the precursors of myofibroblasts in a hepatic fibrosis mouse model[62].

## 1.5 The extracellular matrix of fibrotic tissue

The ECM is a vital structure that provides an architectural support for tissue, a scaffold for resident cells to perform their appropriate functions, and a repository of signalling molecules and enzymes. The composition of ECM is mostly collagen, which is accompanied by a plethora of glycoproteins, glycosaminoglycans and protecoglycans such as elastin, fibrillin, fibronectin, hyaluronic acid and tenascin. Each ECM molecule has a specific role, and together they are modulators of cell behaviour, regulating biological processes such as apoptosis, contraction, differentiation and proliferation [41]. One clear difference between normal tissue homeostasis and fibrosis is that the deposition of a subset of ECM molecules such as collagens, fibronectins and hyaluronic acid (HA) are elevated. Collagens are a huge family of proteins that provide the essential structural support in tissue, and are divided into two classes, fibrillar and non-fibrillar. The former currently consists of members including type I, II, III, V, XI, XXIV and XXVII. Fibrillar collagens are often found in their triple helical form, and in terms of functionality, type I collagen is primarily involved in providing the rigid structure of the wound area, whereas type III is responsible for mediating cell contraction and differentiation[63 64]. Similar to collagen, fibronectin molecules are the building blocks of the ECM scaffold, and their interaction with myofibroblasts via cell surface receptors of the integrin family and the RGF-motif is key to altering a cell's fate and cell positioning [41]. Growth factors including Transforming growth factor beta (TGF-β), Fibroblast growth factor (FGF), Platelet-derived growth factor (PDGF) and Vascular endothelial growth factor (VEGF) have all been reported to be able to interact with

different splice variants of fibronectin, and ultimately promote the phenotype of myofibroblasts[65]. HA is a linear non-sulfated GAG protein, and an important ECM molecule that plays various roles throughout different stages of the fibrotic process. During the early inflammatory stage, the degradation products of HA have been shown to promote the fibrotic response by increasing cytokine production, such as IL-6, through the interaction with receptor Toll-Like receptors (TLR) 2 and 4 [66]. HA has also been reported to attract fibroblasts to the fibrotic or wound areas in a concentration dependent [67] and has a cell to ECM adhesion stabilising function[64]. Beside the ECM composition, stiffness of the ECM also has a functional role - to activate latent TGF-β to promote fibroblast gene expression and myofibroblasts differentiation [68 69]. Other known profibrotic signalling molecules such as YAP and TAZ, downstream effectors of the RhoGTPases of the Hippo pathway, have also been shown to be mechanosensitive [70].

The remodelling of ECM is also a crucial part of the regulatory process in tissue homeostasis and fibrosis by maintaining the composition and structure of the ECM, and also through facilitating the release of the stored profibrotic growth factors and proinflammatory cytokines. The groups of enzymes that are involved in such processes are matrix metalloproteinases (MMPs), adamalysins and meprins [41]. MMPs are a large family of proteinases that are specialised in ECM degradation. Some MMPs are membrane-bound, whilst others are secreted enzymes. The list of substrates that MMPs target is extensive, though mainly consists of ECM molecules such as collagens or fibronectins. However, non-ECM molecules including DDR, DKK,

and SFRP have also been reported as substrates, suggesting an ECM regulating role beyond direct ECM degradation. To further complicate the characterisation of MMPs, pro-MMPs have been shown to possess a non-proteolytic but fibrosis inducing function [71] Similar to MMPs, ADAMs and ADAMTS of the Adamalysins protein family also possess the functions of collagen degradation and ECM remodelling [41]. Some ADAMs are ectodomain sheddases, which can cleave transmembrane proteins and release cytokines, growth factors and receptors into the ECM, for example the release of soluble TNF-$\alpha$ by ADAM-17[72], thus they are capable of indirectly promoting a fibrotic response. ADAMTS, on the other hand, are mostly specialised as aggrecanases, which function as a cleaving enzyme of aggrecans, or as N-propeptidases which cleave procollagen to collagen to allow the formation of collagen fibrils.

## 1.6 Currently known signalling pathways of fibrosis

It is clear that fibrosis is a perturbation of an intricate, highly regulated tissue homeostatic mechanism. Fibrosis is a multi-stage disease that involves a multitude of inflammatory, fibrotic and ECM components. The activation of their corresponding downstream signalling pathways in turn activates a network of proinflammatory or profibrotic signalling cascades that ultimately manifest as fibrotic phenotypes.

### 1.6.1 Transforming growth factor beta pathway

One well studied proinflammatory and profibrotic mediator is transforming growth factor beta (TGF-β). This is due to its overexpression in different fibrotic tissue such as IPF, hepatic fibrosis, ocular, and renal fibrosis [73]. TGF-β is a superfamily of proteins that include three isoforms of TGF-β 1 2 and 3, activins, inhibins, bone morphgenic proteins, growth differentiation factors and glial-derived neurotrophic factors. Amongst the three isoforms, TGF-β1 is the most abundant and has been identified in most cell types, for example immune cells, and myofibroblasts that reside in fibrotic tissue. When TGF-β1 is synthesised, it is transformed into its latent form called large latent form complex (LLT) through the binding with pro-peptide LAP and LTBP [73], which is then later activated and released enzymatically in the ECM, for example through cleaving by MMP-2 and -9[74]. The binding of TGF-β to TGF-β receptor type II triggers the formation of heterotetramer, phosphorylation of the type I receptor and subsequent activation of the receptor regulated-Smad (RSmad)2/3 pathway[75]. The nuclear translocation of the RSmad 2/3 and 4 complex in turn promotes the expression of many known fibrotic genes including *CTGF*, *COL1A1* and *COL3A1* and downstream biological processes such as ECM production, myofibroblast differentiation, proliferation, and activation of CD4+ T-cells, leading to fibrosis in different organs[76 77]. These profibrotic phenotypes are the outcome of an active TGF-β Smad-dependent pathway in conjunction with various higher-order signalling pathways, including transcriptional factors YAP/TAZ of the Hippo pathway and β–catenin of the WNT signalling pathway[73]. Altogether this evidence emphasises the importance of TGF-β1

expression and activation and its canonical signalling pathway in early fibrotic development [78].

Non canonical TGF-β signalling pathways, a term for signalling pathways that involve non Smad molecules, have also shown profibrotic functions similar to the canonical pathway. To date, five non-canonical pathways have been identified: the ERK1/2 pathway, p38 MAPK pathway, JNK pathway, RhoA/ROCK pathway and PI3K/AKT pathway. In particular, the PI3K/AKT, p38 and JNK pathways are regulated by the autoubiquitination of TRAF6[79], whereas the ERK 1/2 pathway is regulated by the phosphorylation of shcA [80]. Collectively, the activity of TRAF6 TAK1 signalling pathway correlates with the gene expression of fibronectin, collagens, aVβ6 integrin and MMP2, which can promote LLT activation, endothelial mesenchymal transition (EMT) and autophagy, which is crucial for apoptosis inhibition and cell survival [79 81 82]. On the other hand, the activity of RhoA/ROCK pathway is regulated by the phosphorylation of Smad, which was demonstrated in a Smad/ROCK inhibition study [83], and is essential for *CTGF* and *α-SMA* expression, and stress fibre alignment, contraction, and myofibroblast differentiation [84]. The function of ERK1/2 signalling is generally unclear though recent studies have demonstrated it is an upstream positive regulatory molecule of ROCK1 in a cardiac model[85], and a regulator of lung fibroblast to myofibroblast EMT through the splicing of CD44v6 in an IPF model[86].

Despite the clear proinflammatory and profibrotic nature of TGF-β, numerous inhibition strategies of this molecule have been suggested and applied in mouse models, and in human clinical trials. These studies have used small molecule inhibitors and antibodies against the receptors and the ligand itself, such as Fresolimumab, Nintedanib, and Pirfenidone [87 88]. The results were mostly positive, with reports showing recession of fibrosis phenotypic response in a renal diabetic mouse model[81 89], in renal fibrosis [90] and pulmonary fibrosis patients [91] though showing mixed effects in systemic sclerosis patients[92 93]. Researchers speculated the negative effects demonstrated in systemic sclerosis patients were a result of a heightened inflammatory response that TGF-β was in place to suppress and concluded that global inhibition of TGF-β was not a suitable strategy. Furthermore, although Pirfenidone has been approved for treating IPF since 2014, its mechanism of action was not clear until recently [94 95]. To overcome the potential issues arising from direct inhibition of TGF-β, α-integrin aVβ6 targeting antibody STX-100 has also been developed specifically for preventing latent TGF-β activation and is currently in a phase II clinical trial.

### 1.6.2 WNT signalling pathway

The WNTs are a protein family of developmental process regulators, and their biological importance is demonstrated by the organ defects and loss of function found in various knockout studies[96]. In fibrosis, the list of WNT functions extends from cell development, differentiation, and proliferation to cell migration [97]. WNT signalling has

been divided into two major pathways: canonical and non-canonical. The canonical pathway is defined as β-catenin dependent and the non-canonical pathway β-catenin independent [96]. There are in total 19 WNT ligands, each of which has different binding affinity to WNT receptors and their selective binding determines whether the canonical or non-canonical pathway is activated. Some WNTs are capable of activating both pathways, and have been demonstrated to operate in a concentration dependent manner [98 99] though their precise mechanism of action remains poorly understood. WNT receptors, which also in part determine the signal transduction pathway, act mainly as dimers consisting of a core G protein coupled transmembrane molecule from the Frizzled (FZD) family, and a co-receptor LRP5/6 or ROR2. The canonical pathway is activated when the binding of receptor FZD and LRP5/6 to their substrates occurs. This subsequently phosphorylates the intracellular domain of LRP, resulting in the recruitment and the inactivation of signalling molecule AXIN. This promotes the dissociation of the degradation complex from β-catenin to allow its stabilization and subsequent translocation into nucleus to form a complex with transcription enhancer TCF/LEF (Figure 1.6.2.1). In fibrosis, this leads to the binding of POL II transcription complex in promoter regions of profibrotic gene such as *COL1A1* and *ACTA2*. As for non-canonical pathways, WNT activity can be transduced either through the binding of ROR2 alone or ROR2-FZD dimers. Several downstream effectors, such as JNK, RHOA of the planar cell polarity pathway are phosphorylated in response to this activation, and they are the upstream regulators of fibrotic functions such as ECM deposition [100 101]. The activity of WNT pathways is tightly regulated by families of secreted inhibitors DKKs, SFRPs, and WIF1. The inhibition of each substrate is selective, where DKK binds specifically to LRP, and SFRP and WIF1 to the WNT

ligands. Mutations in WNTs and the WNT pathway inhibitors have been shown to be implicated in the development of a wide range of fibrotic disease in mouse models and in humans, further emphasising the central role that WNT plays in fibrosis [96]. As mentioned above, numerous components of the canonical and non-canonical WNT pathways interact with molecules of the SMAD2/3 and HIPPO pathways to form a complex inflammatory, profibrotic, and cell motility signalling network[73].

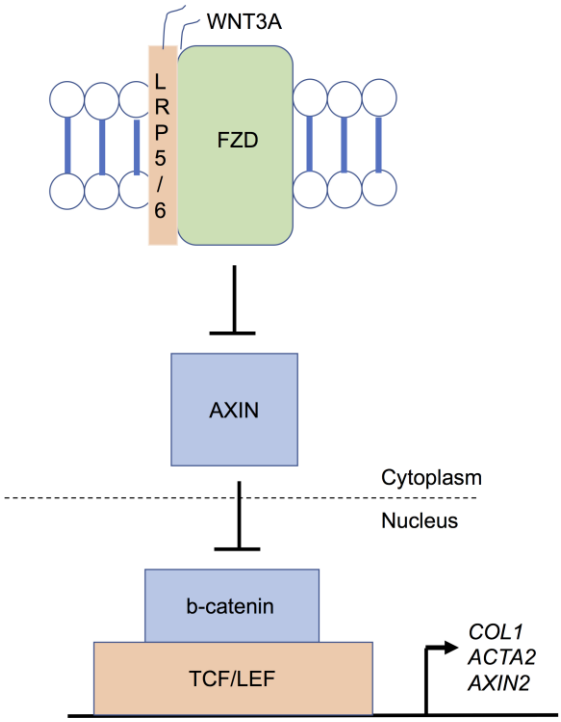



Figure 1.6.2.1 Picture depicting an abstract of the WNT canonical signalling pathway. (Adapted from Park et al [102])

Several protein knockdown studies on various components of the WNT canonical and non-canonical signalling pathways have demonstrated the profibrotic role of WNT pathways. For example, knockdown of FZD7/8, β-catenin, and β-catenin targeting

promoter regions have been shown to suppress the fibrotic phenotype in Bleomycin-induced IPF mouse models, [103 104]. Clinical trials of anti-WNT treatments have been primarily conducted in cancer and showed some reduction in disease progression [96 105]. Attempts to use WNT inhibitors to treat other non-cancer illnesses, such as neurodegenerative disorders, have also been conducted but showed inconsistent outcomes [106] . No clinical trials of WNT inhibition have been conducted or are currently proposed for fibrosis, due to the potential risks of developing unexpected side effects [107].

# Chapter 2 The genetics of Dupuytren Disease

## 2.1 Abstract


Introduction

Dupuytren disease (DD) is fibrosis of the palmar fascia of the hand. It is commonly diagnosed in Europe with a prevalence of around 10%. DD is highly multifactorial and polygenic. However, many pathophysiological aspects of DD are unclear. Using population genetics techniques, I aimed to identify new molecular targets to find new insights into the DD development process.

Materials and methods

Genome wide association was performed on a total of 8557 samples using the Illumina HumanExomeCore genotype chip. Replication was performed using PCR based methods on 10601 samples. Genetic associations were calculated under the logistic regression model using Plink. Meta-analysis was performed using fixed effect inverse-variance method, and was tested for heterogeneity using Breslow Day. Imputation was performed using Haplotype Reference Consortium (HRC) reference panel version 1.1.

For follow-up functional work, RNA and protein expression were measured using Taqman assays and ELISA respectively. Immunohistochemistry was performed on 5μm FFPE DD nodule tissue using SFRP4 monoclonal antibodies.

## Results

A total of 26 associated regions were identified in the GWAS, 17 of which were novel. Imputation and fine mapping analysis have confirmed multiple 99% credible sets. Molecular functional work has shown that SFRP4 RNA and protein expression are indirectly governed by the alleles at variant rs16879765 (pMeta = 3.38 $x10^{-81}$). An additional stimulation experiment has shown that SFRP4 can decrease the production of WNT mediated *CTGF* expression.

## Discussion

My GWAS has revealed new exciting molecular targets that extend our current understanding of the pathology of DD. Follow-up functional work has demonstrated the involvement and importance of the non-canonical WNT signalling pathway, possibly in the early development of DD. However, addition genomic and epigenomic information is required for further functional work.

## 2.2 Introduction

Dupuytren disease is a fibrotic condition in the palmar fascia of the hand commonly diagnosed in Caucasians. The clinical presentation of DD is a flexion contracture of the digits, with the fingers being slowly drawn into the palm, which ultimately causes functional problems [108]. Pathologically, DD primarily consists of two major components: a collagen rich cord and a cell rich nodule [109]. DD presents many similarities when compared to fibrosis of other organs, including infiltration of inflammatory cells as well as an increase in deposition of ECM components such as collagen I, III, V, fibronectin, laminin and tenascin[110-112]. The variety of inflammatory cells found in DD tissue is diverse, and consists of leukocytes, monocytes, macrophages, B cells, T cells and mast cells. In total, they make up around 10% of cells found in the nodule [113]. Working in concert with the ECM, which plays a crucial fibrotic mediatory role by storing various growth factors, they provide necessary conditions for myofibroblasts to differentiate and to contract [64]. Myofibroblasts are the key cells in DD and are responsible for many features of the fibrotic phenotype, including cellular migration, ECM production, and contraction of the ECM [60]. Removing, or increasing apoptosis, of this cell population can facilitate resolution of fibrosis, as observed in a lung fibrosis mouse model [114]. The majority of DD myofibroblasts are found congregating mostly in nodules rather than in cords as studies in the literature have suggested they undergo apoptosis due to the stiffness and high contractile stress in the ECM of the cords[115]. The function of myofibroblasts primarily involves interacting with and contracting the ECM via a variety of cell surface

receptors, including integrins and collagen receptors[64]. Similar to other fibrotic conditions, little is known with regards to the progenitors of myofibroblasts, where pericytes residing in the adjacent tissue, smooth cells within the nearby blood vessels, fibroblasts from the underlying dermis and fibrocytes from bone marrow have all been suggested in the literature [116]. The differentiation processes seem to vary between different progenitor cell types and happens in stages. The early “inactivated” myofibroblasts are called proto-myofibroblasts. Inflammatory cytokines such as TGF-β have been reported in the literature to initiate this differentiation process, via endothelial-mesenchymal transition (EMT). Even though the majority of the cell population in DD tissue are myofibroblasts, they are poorly defined at a cellular level, with no well-defined cell surface markers to allow easy identification. Alpha smooth muscle actin (ASMA) has been regularly used as a biomarker largely because it is often enriched in this cell type, but it is unreliable as not all ASMA positive cells are contractile. The current best strategy involves a combination of positive and negative marker selection [60]

One of the main features of complex disease traits is that they are all multifactorial, with both genetic and non-genetic factors predisposing to disease. In the fibrosis model, there is an initial toxic insult to the tissue, which may resolve, or progress to chronic inflammation [117]. The initiating factors include autoimmune, infection, injury, radiation and toxins [40]. As not everyone has a genetic risk to develop fibrosis, individuals possessing the protective alleles are likely to resolve normally despite being exposed to chronic inflammation. This is reflected in fibrosis clinical trials where

anti-inflammatory treatment regimens show little to no efficacy [118]. Therefore, it is imperative to investigate and understand the function of molecules within and outwith the inflammatory pathways. It is undeniable that there is no fibrosis without chronic inflammation but conversely, it is also true that there is no fibrosis without genetic predisposition. A typical approach to identify new molecular targets is through population genetics, or a technique known as genome wide association study or GWAS.

There is no shortage of evidence that genetic factors play an important role in the development of DD. A sibling recurrence study in the UK has reported that a sibling of a DD patient has a relative risk to develop DD of 4.48 times higher than a random individual in the population [21]. This was further supported by a Danish twin study, where the authors estimated DD heritability to be 80% [19]. Furthermore, a pilot genome wide association study of three European cohorts [29] has pinpointed nine associated loci.

In additional to being multifactorial, another feature of complex disease traits is that they are also highly polygenic, whereby many polymorphisms of multiple genes can contribute to their genetic predisposition. The proportion of variance explained at each locus is often very small, and therefore large sample size is the first element to consider in a GWAS design. A cohort between 10000 and 100000 people would be required to locate rare associated variants, which are often linked to a moderately high

OR ($\beta$ = 0.1) and an interesting biological function [119]. Conversely, even with a large sample size many GWA studies in the literature lack the power to detect rare variants and there are two possible explanations; the limitation of the current genotyping technology and calling algorithms. To circumvent this issue, multiple replication studies and follow-up meta-analyses have been employed to be the predominant solution to identify the missing heritability [120]. This has led to an increase in popularity of using large public genotype databases, such as the UKBiobank, but there are multiple drawbacks, where samples can be misclassified and the same sample may appear in multiple cohorts within the same study. These factors should be taken into account during the study design stage and in power calculations. In terms of genotyping depth, array-based methods are often used and are sufficient due to linkage disequilibrium, though whole genome sequencing should be considered for discovering novel and rare variants because of the limitation of the former platform [121]. By exploiting the uniqueness of linkage disequilibrium (LD) blocks in the human genome, fine mapping by imputation using reference panels that were generated from the genomic sequence of as few as a thousand samples of a similar ethnic background is an accurate and cheap procedure to increase the genotyping resolution of GWAS data that was collected using array-based methods.

With regards to downstream functional work, for intergenic and intronic variants (as in the vast majority of GWAS associated variants), this should be prioritised based on the existing knowledge of the variants and their interactions with transcriptional factors or transcriptional complex machinery. For exonic variants, their effect on known

protein functional domains should be assessed, as well as the likelihood of a deleterious effect on protein function. Further in-depth discussion on the methods, advantages and disadvantages of current *in-silico* analysis after GWAS is found in Chapter 5.

Translational work from GWAS results have historically proven challenging, but recent advances in statistical methods, better fine mapping resolution and high through-put and single cell sequencing techniques have led to improvements in treatment regiments of many complex disease traits. Here I used autoimmune diseases, psoriasis and ankylosing spondylitis as an example. GWAS discoveries have led to the use of inhibitors that target the components of the IL-23 pathway, specifically IL-17, and the p19 and p40 subunit of IL-23. These inhibitors have demonstrated better efficacy than the current biologics in phase-II clinical trials [122].

Other than molecular functional work, GWAS data can be used in various non-functional analyses to further advance our understanding of the disease trait. As addressed earlier, heritability, or more specifically narrow sense heritability ($h^2$), is used to dissect the genetic architecture and to emphasise the importance of genetic predisposition in a complex disease trait. Estimation of heritability can be done using variants of the entire genome, which addresses the total genetic variation, or in part by restricting the variant selection to include only the associated variants to identify the proportion of missing heritability. Estimators can be used to aid the design of

further studies. Various methods have been developed to calculate heritability, from the traditional ANOVA method to the more recent restricted estimation maximum likelihood (REML) under the linear mixed model (LMM), which boasts the precision by the estimators of standard error (SE) and the construction of confidence intervals. Examples of LMM methods included GCTA[123], Fast-LMM [124] and the non-asymptotic ALBI [125]. Another use of the GWAS data is to detect and identify pleiotropy, which ultimately can be used to aid making future diagnostic decisions, detecting associated diseases. This is based on the idea that an individual's genetic make-up, i.e. polygenic risk score (PRS) can be used to predict their likelihood to develop related diseases, or to predict the rate of disease progression. With extensive validation, PRS can potentially be used for disease prevention and to improve upon existing treatment methods. Lastly, GWAS data can also be used to find causal relationships amongst different traits with the same associated loci using a suite of methods termed Mendelian Randomisation [126].

There are a lot of unanswered questions in DD aetiology, the process of pathological development, and the cell types involved. Population genetics methods are powerful tools to find new biological targets that might form the basis of new treatments, and insights into biological processes. GWAS results were not easily translatable in the past due to the limitation of the genotyping technology available at the time, but this has been vastly improved in recent years due to the advancements in genotyping and sequencing techniques. Ultimately, the results of this chapter have huge potential to

improve future diagnostic and treatment decisions. It will also further our understanding in fibrosis as a whole.

## 2.3 Aims and hypotheses

DD is a complex genetic condition and the current most appropriate method to interrogate a complex genetic disease is through employing GWAS. Our aim was to first identify the variants that predispose to DD development, and subsequently prioritise these variants for functional studies using existing knowledge of fibrosis biology, and the results from *in silico* analysis and utilising our epigenomic and transcriptomic data. Our second aim was to attempt to identify the biological function of one of the discovered variants.

My hypothesis was that because this study was a well powered extension of a previous DD GWAS with a similar participating demographic, I would have the power to detect novel associated variants in addition to the existing variants. Some of these novel variants would be located in regions that are in the proximity of genes known to be important in the pathogenesis of fibrosis, but most would be intronic and intergenic. Thus, it would be challenging to dissect their biological function and to single out the causal variants. Despite the difficulty, I aimed to further investigate a variant potentially related to known biological pathways implicated in fibrosis to maximise my chances of success.

## 2.4 Materials and methods

### 2.4.1 Ethical approval

Patient samples that were used in this chapter were approved by the Research Ethics Committee or equivalent, which included the Oxfordshire Research Ethics Committee B/09/H0605/65 for the British Society for Surgery of the Hand Genetics of Dupuytren's Disease (BSSH-GODD), the Medical Ethics Committee (METc) 2007/067 for the Genetic Origin of Dupuytren Disease (GODDAF) and the University of Cologne 14/292 for the German Dupuytren study.

### 2.4.2 Saliva DNA extraction

Saliva samples were collected using Oragene OG250 DNA kit (DNAgenotek, Ontario, Canada), and DNA was extracted according to manufacturer's instructions. Briefly, saliva samples were super pasteurised at 50$^{o}$C overnight. 1:25$^{th}$ ratio of prepIT.L2P lysis buffer (DNAgenotek, Ontario, Canada) to saliva was then added and mixed by vortexing gently. The mixture was incubated on ice for 10 minutes and pelleted by centrifugation at 3250g for 30 minutes. The supernatant was then transferred into a new 15ml falcon tube. One volume of 100% ethanol was added to the supernatant

and mixed by inversion 10 times. Samples were incubated at room temperature for 10 minutes, allowing DNA to precipitate. The DNA was pelleted by centrifugation for 30 minutes at 3250g, following by two 70% ethanol washes. The DNA pellet was air dried at room temperature and re-suspended with 500 μl of Tris-EDTA (TE) buffer. Extracted DNA was frozen at -20$^{o}$C until further analysis.

### 2.4.3 Discovery phase genome wide association study

The concentration of each DNA sample was determined using Nanodrop 3000 spectrophotometer (ThermoFisher Scientific, Paisley, UK). DNA was diluted to 100ng/μl and aliquoted onto barcoded 96-well plates, which were couriered frozen on dry ice to the microarray facility at the Sanger Institute (Wellcome Trust, Hinxton, UK). The quality of each sample was assessed using the MassArray system (Agena Bioscience, Hamburg, Germany). Samples that had gender discrepancies between assayed and self-reported or were degraded, were discarded before further analysis. Samples that passed the MassArray system quality control were genotyped on the HumanCoreExome microarray chip (Illumina, Cambridge, UK).

### 2.4.4 Genome wide association study data quality control

Quality control of the genotyping data was performed using Bash, Perl, Plink and R unless stated otherwise. The outline of the quality control steps is described in figure 2.5.3. SNPs with a call rate less than 90 percent were removed. Remaining variants were aligned to the positive strand using the Illumina HumanCoreExome v37 reference panel and scripts provided by Dr William Rayner (Wellcome Trust Centre of Human Genetics, Oxford). In the initial sample quality control (QC), contaminated and poor quality samples were removed using various thresholds including; mean minor allele frequency (MAF) and minor allele count equals one (MAC = 1) three times greater than standard deviation, less than 98% call rate, sex and identity mismatch and ethnic outliers, which were selected by manual visualisation of the scatter plot that were created using the first two principle components (PC) of the 1000 genome reference panel. Samples that were not situated around or within the Utah Resident with Northern and Western European Ancestry (CEU) cluster in the principle component analysis were identified as not Caucasian, or an ethnic outlier. In the second round of QC, thresholds including SNPs with less than 98% call rate, cluster separation score < 0.4 (where a low score indicates high overlaps between genotyping clusters), that were duplicated, non-autosomal and significantly deviated from the Hardy Weinberg equilibrium (HWE; $p<0.0001$) were applied. The remaining good quality samples were merged with the UKHLS control dataset. This was acquired from the European Genome Archive (EGA) depository (Cambridge, UK) and was provided pre-processed using the same stringent thresholds as described above. After merging,

samples relatedness was determined using IBD estimation. All but one of the close to third-degree related samples, or PI_HAT > 0.185, were removed. PCA was performed again on the merged dataset to remove additional ethnic outliers. Rare alleles, MAF <0.05, were removed.

### 2.4.5 GWAS analysis

GWA analysis was performed using Plink, and diagrams were created using R and LocusZoom unless stated otherwise. Genetic maps and recombination rates genomic panels were referenced to the hg19 assembly, and were acquired from the UCSC Genome database and Hapmap consortium. Association analysis was performed adjusted for population stratification using principle components of the QC filtered dataset and sex as covariates, and the multivariate logistic regression model. The quantity of PCs applied in the reported association analysis was determined using two methods, first by a decrease in variance between PCs and second by the lambda differences between results of logistic regressions that were adjusted with an increasing number of PCs. Q-Q plot was created for better visualisation of lambda. Manhattan plot and regional plots were created for visualising the location of the associated SNPs in relation to recombination hotspots and gene locations. To account for the number of association tests that were performed, a p value < $5x10^{-8}$ was deemed statistically significant and p between $5x10^{-6}$ and $5x10^{-8}$ as suggestive. Conditional analysis was performed to identify potential hidden associated regions. This was performed by applying the allelic dosage of the most significant SNPs of each

associated region as a covariate in addition to the covariates mentioned earlier. The genotyping integrity of the associated SNPs was confirmed by manual visualisation of scatter plots that were created using their corresponding intensities.

### 2.4.6 Replication and meta-analysis

DNA samples from Germany and the Netherlands were a gift from our collaborators (Dr Hennies, University of Cologne; Professor Werker, University of Groningen). They were diluted to 100ng/µl prior to being plated on 96 well plates (ThermoFisher Scientific, Paisley, UK), which were couriered frozen on dry ice to the genotyping facility at the Sanger Institute (Wellcome Trust, Hinxton, UK). Genotyping was performed using the MassArray System (Agena Bioscience, Hamburg, Germany). The assay panel was selected based on the results of the discovery phase GWAS. This included suggestive and genome wide significant variants and their corresponding linkage disequilibrium variants. Due to the limitation of the genotyping system, not all initially selected SNPs were included in the final assay, and variant rs2598107 was genotyped separately from all other variants. Prior to association analysis, genotyped samples were filtered using a list of QC metrics. This removed samples and SNPs with call rates less than 90%, and that were not following the HWE ($p < 0.0001$). Using logistic regression as a model and sex as a covariate, association analysis performed twice; once on each cohort individually, and then together. Effect size heterogeneity between cohorts was tested using Breslow-Day. Meta-analysis was performed on both

the discovery and replication cohort using the inversed variance method. Again, Breslow-Day was used to test for heterogeneity.

### 2.4.7 Pre-phasing, imputation and post-imputation quality control

Imputation was performed on the post QC filtered discovery GWAS dataset. This process was split into pre-phasing and imputation. Prior to pre-phasing, a list of additional QC metrics was applied on to the dataset. This included removing variants that were monomorphic (MAF = 0), singletons (MAC = 1), INDELS or duplicated. SNP IDs, reference allele and strand orientation of each variant were then updated to match the HRC reference panel version 1.1 using a script provided by Dr William Rayner (Wellcome Trust Centre of Human Genetics, Oxford). The QC filtered dataset was divided by its chromosome number into 22 files, which were separately pre-phased using SHAPEIT v2.1.2 (r837) with the following parameters; effective size equals 11418 and states equals 200. The pre-phased dataset was converted into VCF, compressed, indexed with Tabix and concatenated using Bcftools prior to being uploaded to the imputation server at the Sanger Imputation Service (Wellcome Trust, Hinxton, UK). The Imputation was performed using the PBWT algorithm [39] and the HRC reference panel release 1.1. Post imputation QC metrics were applied to the imputed dataset to remove SNPs with call rate less than 95%, MAF < 0.01 or that were deviated from the HWE ($p < 0.000001$). SNPs with an info score less than 0.3 were also omitted from further analysis.

### 2.4.8 Bayesian framework imputation analysis and fine mapping

Bayes factor (BF) of each imputed SNP was calculated using Snptest v2.5.2 with the assumption that the prior distribution is additive. Posterior odds were calculated on 1Mb regions centred on SNPs with the highest BF using the formula below,

$$P = \frac{BF_i}{\sum_{J=1}^{n} BF_J}$$

where i is the SNP of interest and J is the SNPs of the fine mapping region. 99% credible set of each 1Mb region was calculated by adding up the posterior odds, starting from the highest BF of the region to the lowest until the total reaches 0.99. Similar to the method that was described in the GWAS analysis section, conditional analysis was performed on regions that were harbouring multiple associated regions using the lead SNP as a covariate with the exception of allelic dosage uncertainty. Posterior odds and 99% credible sets of these dataset were created using BFs reported in the post conditional analysis.

### 2.4.9 *In silico* analysis

The possible biological function of each variant that was reported in the 99% credible sets were first investigated using 3DSNP for the annotation of genomic features including enhancer/promoter binding, transcriptional factor binding site and changes in motif binding element. Any variant scored above 10 in 3DSNP was further analysed

for gene interactions using the virtual 4C database in the 3D genome browser. The interactions were further confirmed at expression level using GTEx, selecting tissues of potential relevance to the pathogenesis of DD.

### 2.4.10 Myofibroblast tissue culture

Dupuytren tissue of consented patients (REC 07/H0706/81), who had undergone fasciectomy was collected in DMEM (Lonza, Slough, UK). Tissue was diced into 1mm pieces with surgical scalpels and digested in 20μg/ml collagenase (Worthington Biochemcials, NJ, US) at 37ºC with agitation at 120rpm. Disaggregated cells were filtered from the digested extracellular matrix using a 40μm cell strainer (Fisher Scientific, Loughborough, UK), pelleted by centrifugation at 300g for 5 minutes and re-suspended in 10ml 10% FBS supplemented DMEM. The suspension was plated onto a 10cm$^2$ dish and incubated at 37ºC with 5% $CO_2$. When the cells reached 80% confluence, they were detached from the petri dish using 2ml of Tryple (Life Technologies, Paisley, UK), pelleted by centrifugation, re-suspended in 20ml 10% FBS DMEM, and reseeded into 150cm$^2$ culture flasks (Corning, NY, US). Cells were passaged again when the confluence reached 80%, one in four using the same method as described above.

### 2.4.11 Tissue processing and immunohistochemistry

Dupuytren tissue was collected following the same procedure as described in section 2.4.8. Tissue was first rinsed with PBS and fixed in 10% Formalin overnight. Fixed tissue was stored in 70% ethanol until further processing for a maximum of two days. Tissue processing was performed using tissue processor ASP3000 (Leica, Milton Keynes, UK). Briefly, tissue was dehydrated by submerging in 70%, 90% and 100% ethanol for 45 minutes each, immediately followed by three times 100% ethanol incubation for one hour. Ethanol was then removed by submerging the sample in xylene for 45 minutes, one hour and 1 hour 15 minutes. In the final step, the sample was submerged in paraffin embedded wax for one hour. Formalin Fixed Paraffin Embedded (FFPE) blocks were created using Histowax, using the bottom half of the Shandon cassettes (ThermoFisher Scientific, Paisley, UK) and small metal containers as a mould. For creating tissue slides, FFPE blocks were dissected into 5μm sections using a microtome (Leica, Milton Keynes, UK), air dried on poly-L-lysine slides (ThermoFisher Scientific, Paisley, UK), and at 37$^{o}$C overnight. The day before staining, tissue slides were baked at 65$^{o}$C for two hours. To increase reproducibility, an autostainer machine (Dako, CA, US) was used to conjugate tissue sections with the desired antibodies. Briefly, antigens were retrieved using the heat induced epitope retrieval method by submerging sample slides in Antigen Retrieval Solution at 90$^{o}$C (Dako, CA, US). Slides were then incubated with isotype control antibody and 1 in 200 EnVision FLEX Antibody Diluent buffer (Dako, CA, US) diluted primary antibodies against the desired protein. This was followed by a wash step using FLEx wash buffer

before the incubation with HRP conjugated secondary antibody, and another wash step prior to incubation with HRP substrate DAB+ chromogen solution (Dako, CA, US). After the brown colour has developed, samples were counter stained with haematoxylin. Immuno-stained slides were dehydrated with one minute of 70%, 90% and 100% ethanol wash, followed by two times 100% xylene wash. A coverslip (VWR, Leicester, UK) was then mounted onto the slide using DPX Mountant (Sigma Aldrich, Dorset, UK), which was then air dried overnight. Images of the tissue section were acquired using an Axio imager brightfield microscope (Carl Zeiss, Cambridge, UK) with either a 20x or 40x objective.

### 2.4.12 Immunocytochemistry

Myofibroblast cells were dissociated from tissue culture flask using Tryple, pelleted, counted using a haemocytometer and resuspended in 10% FBS at 25000 cells per ml. 200 μl of this re-suspension were seeded on chamber slides (ThermoFisher Scientific, Paisley, UK) or 35mm glass Fluorodish (World precision instruments, NL, US). The seeded cells were incubated at normoxic tissue culture condition for 12 hours or until attached. From this point, two times PBST and one PBS washes were applied between every step unless stated otherwise. Attached cells were fixed using 1% paraformaldehyde for 20 minutes, and blocked using 1% BSA in PBS for one hour. Fixed cells were incubated with primary antibodies that were diluted 1 in 200 in 0.1% triton PBS for one hour at room temperature and subsequently with HRP conjugated secondary antibodies in 1% BSA PBS. Cells were counterstained with Actin stain 488

phalloidin (Cytoskeleton, CO, US) at 1 in 1000 dilution in 1% bovine serum albumin (BSA) PBS for 30 minutes. 20x and 40x Images were acquired using a Axio imager brightfield microscope (Carl Zeiss, Cambridge, UK) and LSM 710 confocal microscope (Carl Zeiss, Cambridge, UK) respectively.

### 2.4.13 RNA extraction and gene expression assay

RNA extraction was performed using the two commercially available extraction kits, Direct-zol miniprep or 96-well kit (Zymo Research, CA, US), according to manufacturer's instructions. Methods described below are from the miniprep kit as both protocols are mostly identical. TRIzol (ThermoFisher Scientific, Paisley, UK) was added onto the cell sample and was mixed gently by pipetting up and down. The homogenate was stored at -80$^{o}$C until extraction. At the day of extraction, it was centrifuged at 12000g for 1 minute, and the supernatant was transferred to an RNase-free Eppendorf tube. One volume of ethanol was added to the solution and mixed by pipetting up and down several times. The mixture was transferred to a Zymo-spin IIC column / collection tube and centrifuged at 12000g for one minute. The column was washed with 400µl RNA Wash buffer, centrifuged and DNase treated at room temperature for 15 minutes. The column was washed three more times, once with 400µl of RNA Pre-Wash buffer and two times with 700µl of RNA wash buffer. RNA was eluted using 20µl of DNase/RNase free water, and were stored at -80$^{o}$C until further use. Extracted RNA was reversed transcribed into cDNA using high capacity

RNA-to-cDNA reverse transcription kit (ThermoFisher scientific, Paisley, UK), up to a maximum of 200ng.

Gene expression assay was performed using Taqman (ThermoFisher Scientific, Paisley, UK) based method on 384 well plates and Viia7. Each well contained 5 μl of Fast Advance Taqman master mix, 0.5 μl of Taqman primers and probes of the gene of interest and 4 μl of 4ng cDNA. Two housekeeping genes 18S and GAPDH were included in each assay. The "Fast" pre-set cycling condition was used. Relative expression, or $2^{-\text{delta Ct}}$, was calculated, where delta Ct was subtraction of the Ct value of the housekeeping gene with gene of interest.

### 2.4.14 Enzyme-linked immunosorbent assay (ELISA)

SFRP4 sandwich ELISA kit (Phadia Gmbh, Freiburg, Germany) was a gift from Dr Hoffmann (University of Freiberg, Freiburg, Germany). Intracellular protein was extracted from experimental cells by scraping in 100 μl of RIPA buffer on ice. The supernatant was also collected and was stored at -80$^{o}$C until further use. Protein concentration was determined using a BCA protein assay kit (Millipore, Watford, UK). Briefly, 2 μl of sample and protein standards were incubated with 100 μl of 50 to one mixture of Bicinchoninic acid and Copper II sulfate at 37$^{o}$C for 15 minutes. The colorimetric absorbance of this mixture was then measured at 562nm using plate reader spectrophotometer CLARIOstar (BMG Labtech, Aylesbury, UK). ELISA was

performed according to the manufacturer's instructions. All reagents mentioned below were provided as part of the ELISA kit. Sample was diluted 1 in 10 using sample diluent and added onto wells of the SFRP4 antibody coated 8 well strips. After 60 minutes of incubation at room temperature and three washes with the wash buffer, 100 μl of primary antibodies was added onto each well and further incubated at room temperature for 60 minutes. This was followed by three more washes with the wash buffer and 30 minutes of incubation with 100 μl of HRP conjugates. HRP substrate, 3,3',5,5'-Tetramethylbenzidine (TMB), was added to each well and the strips were incubated for 30 minutes in the dark. The absorbance of samples was measured at 450nm. Absolute concentration was calculated using the protein weight derived from the standard curve.

### 2.4.15 Statistical analysis of functional work

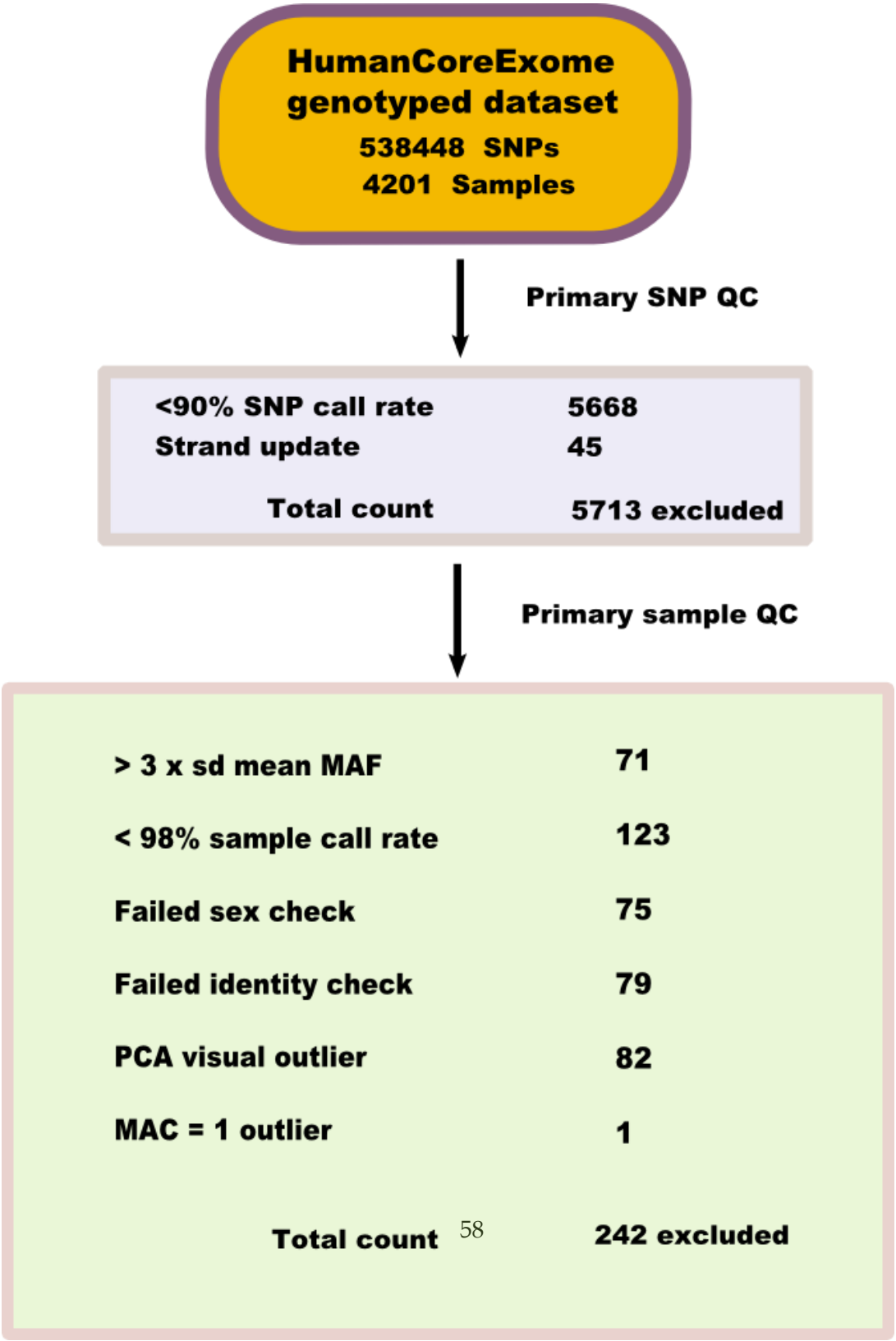
Shapiro-Wilk test was used to test the normality of the data. Levene's test was used to test the equality of variances between samples. The differences between two groups of means were tested using student t-test when variances between samples were equal. In all situations, a p-value less than 0.05 was considered statistically significant. Bonferroni correction was employed when the same sample has undergone multiple testing.

## 2.5 Results

### 2.5.1 Quality control of the genotyped data

Prior to performing the association analysis, array data containing genotypes of the tested samples was subjected to stringent QC to remove spurious data points. The outline of the QC workflow is shown in figure 2.5.1.1, describing in detail each threshold being applied and the corresponding number of samples or SNPs that were removed as a result of the application. Genotyping was performed on the Illumina HumanCoreExome chip, which contained 538,448 SNPs and was used to assay 4201 case samples of the BSSH-GODD cohort. QC was performed in four major steps and each step contained a group of either SNP or sample QC thresholds. In the first SNP QC step, 5,713 SNPs were removed including those which were poorly called (less than 90% call rate) or due to allele mismatch to the reference panel. This was then followed by the first sample QC step, where a total of 242 samples were removed because of sample contamination (greater than three times SD of mean MAF, or MAC equals one), inadequate rate of calling (less than 98% call rate), sex and identity mismatch or ethnically not Caucasian (PCA visual outlier). In the secondary SNP QC step, a further 35,705 SNPs were excluded from the dataset due to being poorly called (less than 98% call rate), not following the Hardy Weinberg equilibrium (HWE, $p< 0.0001$), poorly clustered during genotype calling (Cluster separation score < 0.4), heterozygous haploid on the X chromosome, not autosomal or duplicated. The dataset

was merged with the control dataset, UKHLS (which had been subjected to identical QC), before a final round of sample QC, where an additional 293 samples were removed because of relatedness to another sample to the third degree or closer, ethnic outlier, and poor SNP call rate.



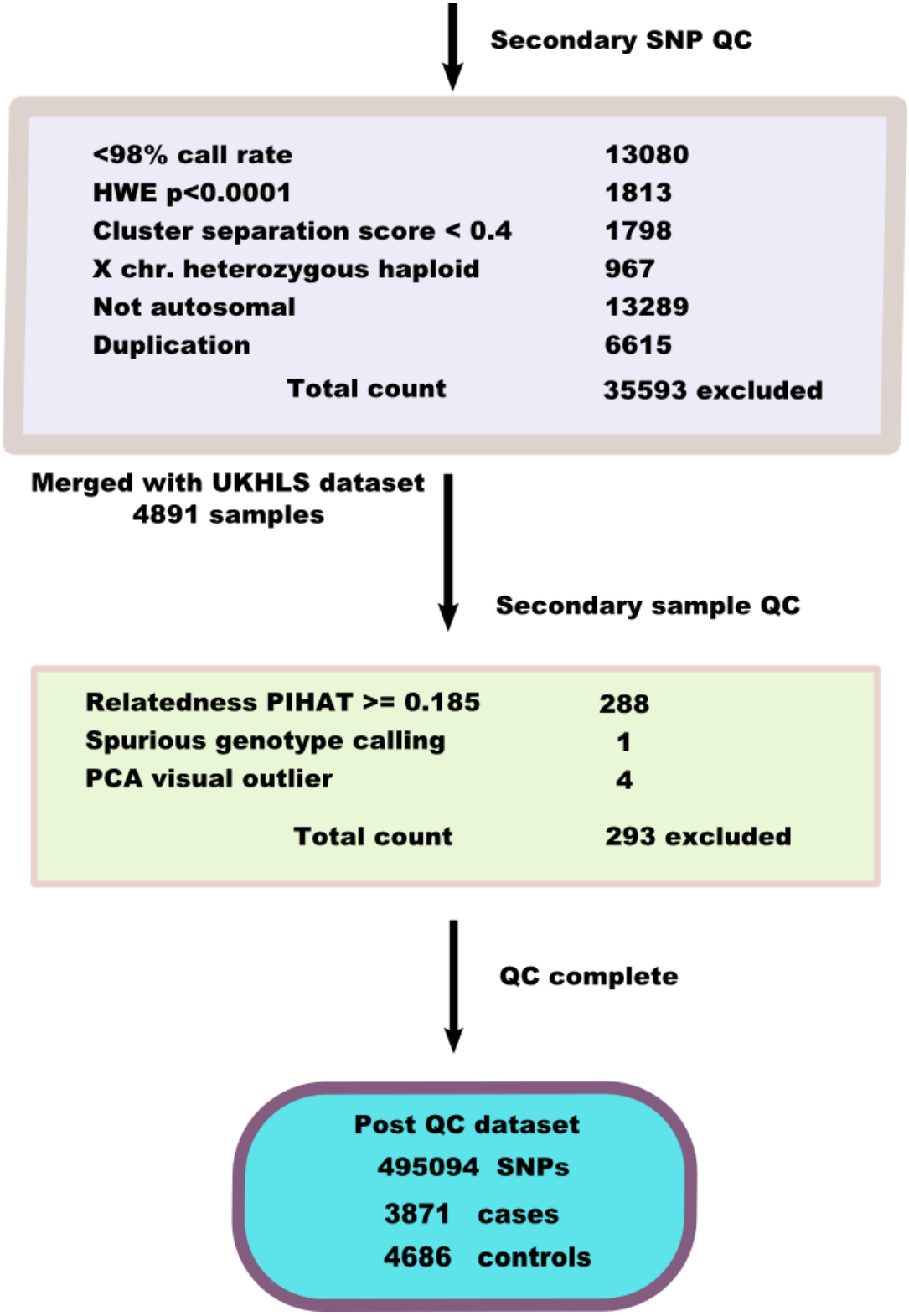



Figure 2.5.1.1 Workflow depicting the filtering thresholds used in the quality control steps of the discovery phase GWAS and the number of samples and SNPs that were removed as a result prior to association analysis being performed.

## 2.5.2 Genome wide association study results

In the discovery phase GWAS, association analysis was performed on a total of 8,557 samples, 3,871 cases and 4,686 controls, on 495,094 SNPs. A total of 15 Dupuytren associated variants were identified, seven of which were novel (Table 2.5.2.1; Figure 2.5.2.1). The associations were calculated using logistic regression with sex and the first two principle components as covariates, which were adequate as only a small trace of hidden population stratification was detected (genomic inflation factor $\lambda$ = 1.091) in the results (Figure 2.5.2.2). As part, if not all, of the genomic inflation could due to the polygenicity of DD, therefore no p value adjustment using genomic inflation factor has been made to the summary statistics. Two independent associated loci, with no linkage to the lead SNP (r2 = 0, D' = 0), were identified after performing conditional analysis on SNP rs16978765 (rs2598107, OR=1.48, Pcond=$6.85 \times 10^{-31}$) on chromosome 7 and rs629535 (rs2912522, OR=0.73, Pcond=$1.31 \times 10^{-14}$) on chromosome 8 (Figure 2.5.2.3). Conditional analyses at the other loci did not reveal any other additional signals. Replication was performed at genome-wide significant and suggestive SNPs in three European cohorts from Germany, the Netherlands and the United Kingdom, amounting to a total of 4,041 cases and 8,251 controls. I successfully replicated 25 out of the 26 associated variants that were identified in the discovery phase. Genome wide significant variant rs246105 and four suggestive SNPs were not assayed in the replication due to restriction of the primer design and limitations of the Mass-spectrometry, PCR-based DNA genotyping array platform. The inverse variance fixed effect meta-analysis using the results of both the discovery and

replication phase descriptive statistics further confirmed a total of nine previously reported associated loci and an additional 17 novel loci (Table 2.5.2.1). Genome-wide Complex Trait Analysis (GCTA) was performed on all common autosomal variants (MAF≥0.05), and the 26 positively associated signals, demonstrating heritability of 53.1% and 11.3% respectively. All 26 DD associated variants were further confirmed as genuine signals through the visualisation of the intensity plots, where three clear distinctive genotype clusters were observed, confirming accuracy of genotyping and genotype calling (Figure 2.5.2.4).

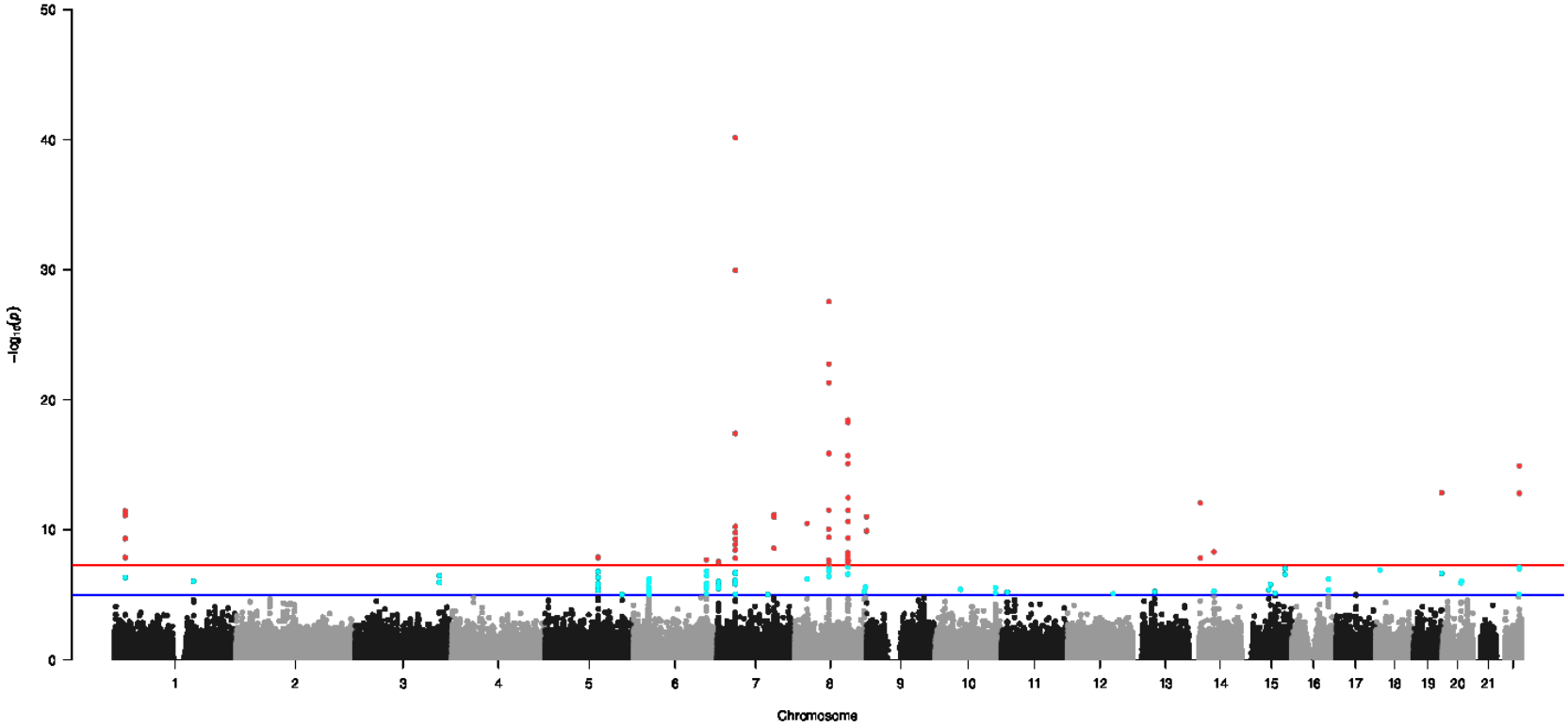



Figure 2.5.2.1 Manhattan plot showing the distribution of the genome wide significant and suggestive DD associated SNPs. Genome wide significant SNPs ($p < 5 \times 10^{-8}$) are coloured red. Suggestive SNPs ($5 \times 10^{-5} < p < 5 \times 10^{-8)}$ are coloured blue. The red and blue horizontal lines mark the significant and suggestive p value threshold at $5 \times 10^{-8}$ and $5 \times 10^{-5}$ respectively.

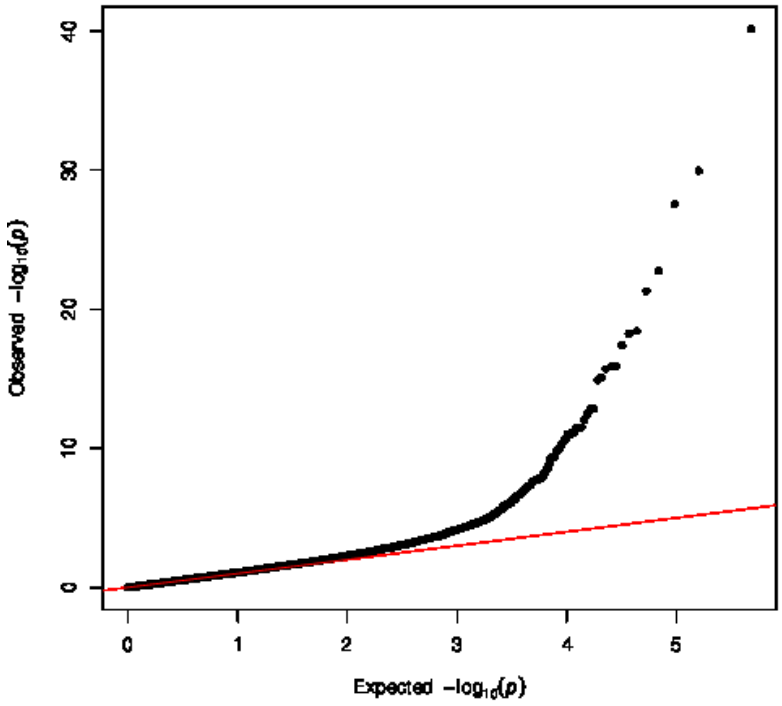
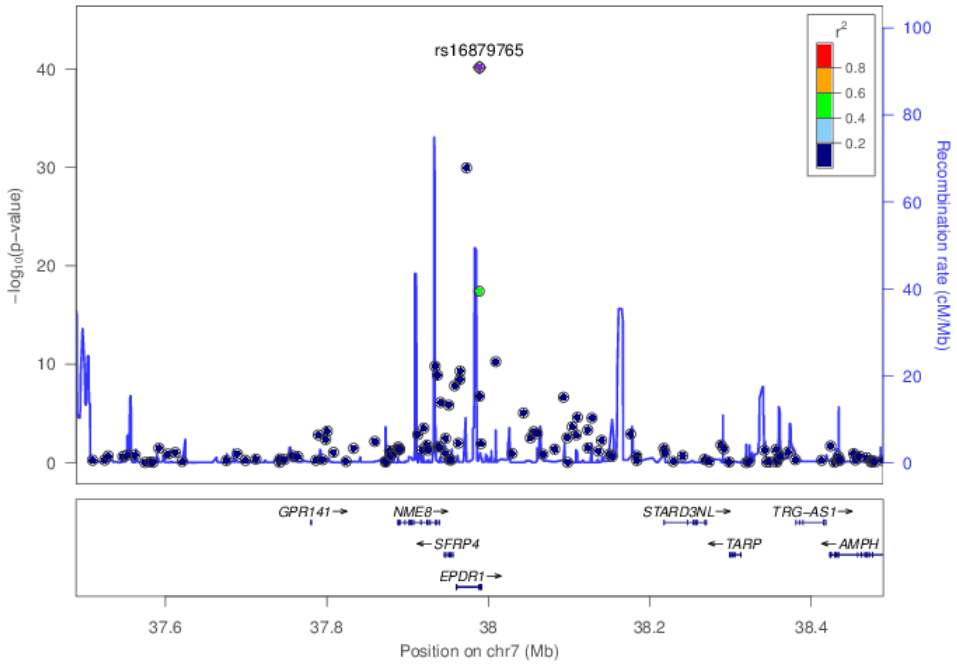
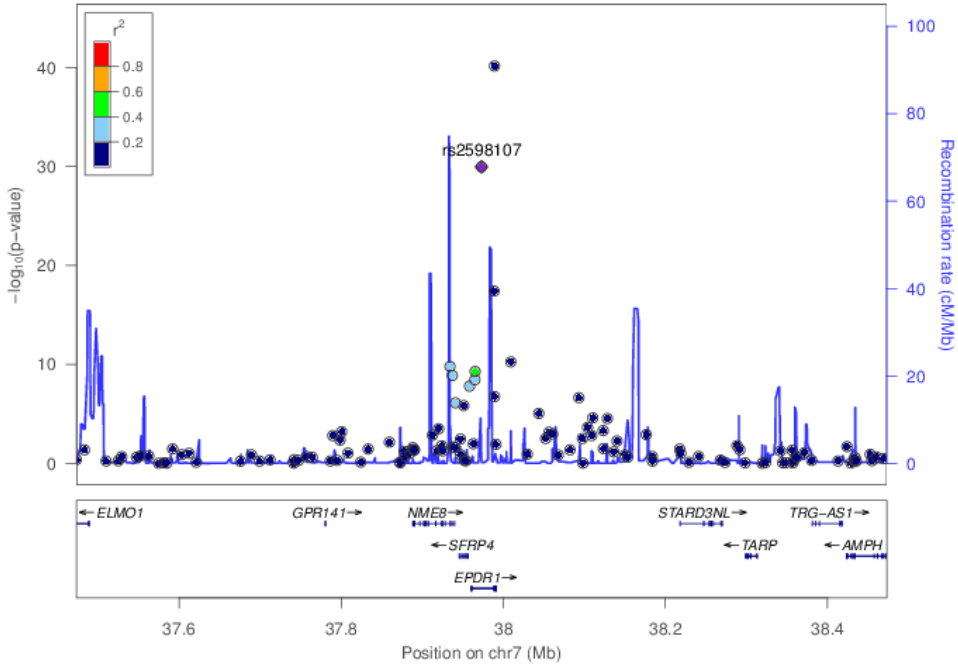



Figure 2.5.2.2 Quantile-Quantile (Q-Q) plot depicting the relationship between the observed and expected p values of the discovery GWAS results. The genomic inflation factor, or lambda, was 1.091, indicating either there is some minor hidden population stratification remaining in the results or due to the polygenicity of the disease.

a)


rs16879765
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
GPR141
NME8
STARD3NL
TRG−AS1
SFRP4
TARP
AMPH
EPDR1
37.6
37.8
38
38.2
38.4
Position on chr7 (Mb)


rs2598107
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
ELMO1
GPR141
NME8
STARD3NL
TRG−AS1
SFRP4
TARP
AMPH
EPDR1
37.6
37.8
38
38.2
38.4
Position on chr7 (Mb)

b)

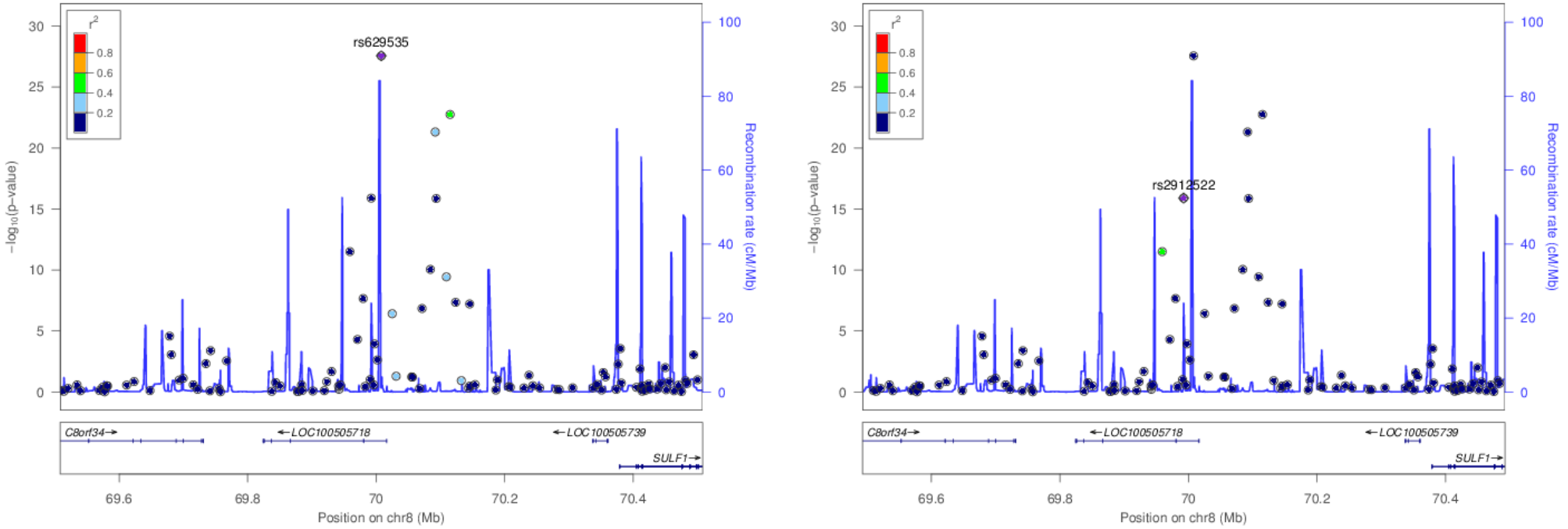



Figure 2.5.2.3 Regional plots depicting the location and the genes in the proximity of a) the most significant associated loci of the discovery GWAS, rs16879765 (left panel), and the associated region of lead SNP rs2598107 after the conditional analysis was performed (right panel). b) The significant associated regions before (left panel) and after (right panel) conditional analysis was performed on the chromosome 8 locus using lead SNP rs629535.

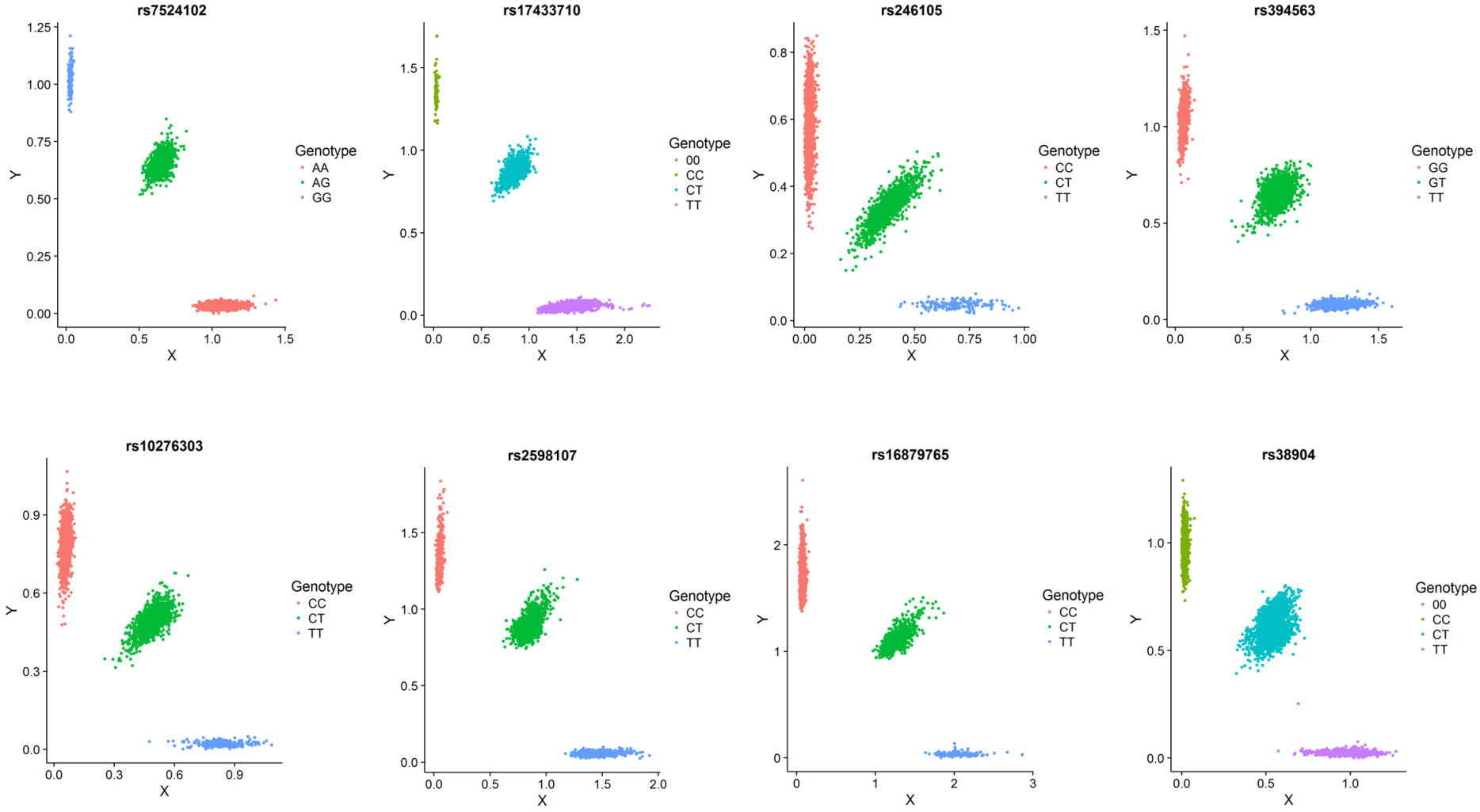
Table 2.5.2.1 The effect allele frequency (EAF), chromosome location, risk allele, p values, and odds ratio of all 26 associated variants that were discovered in the multi stage DD GWAS

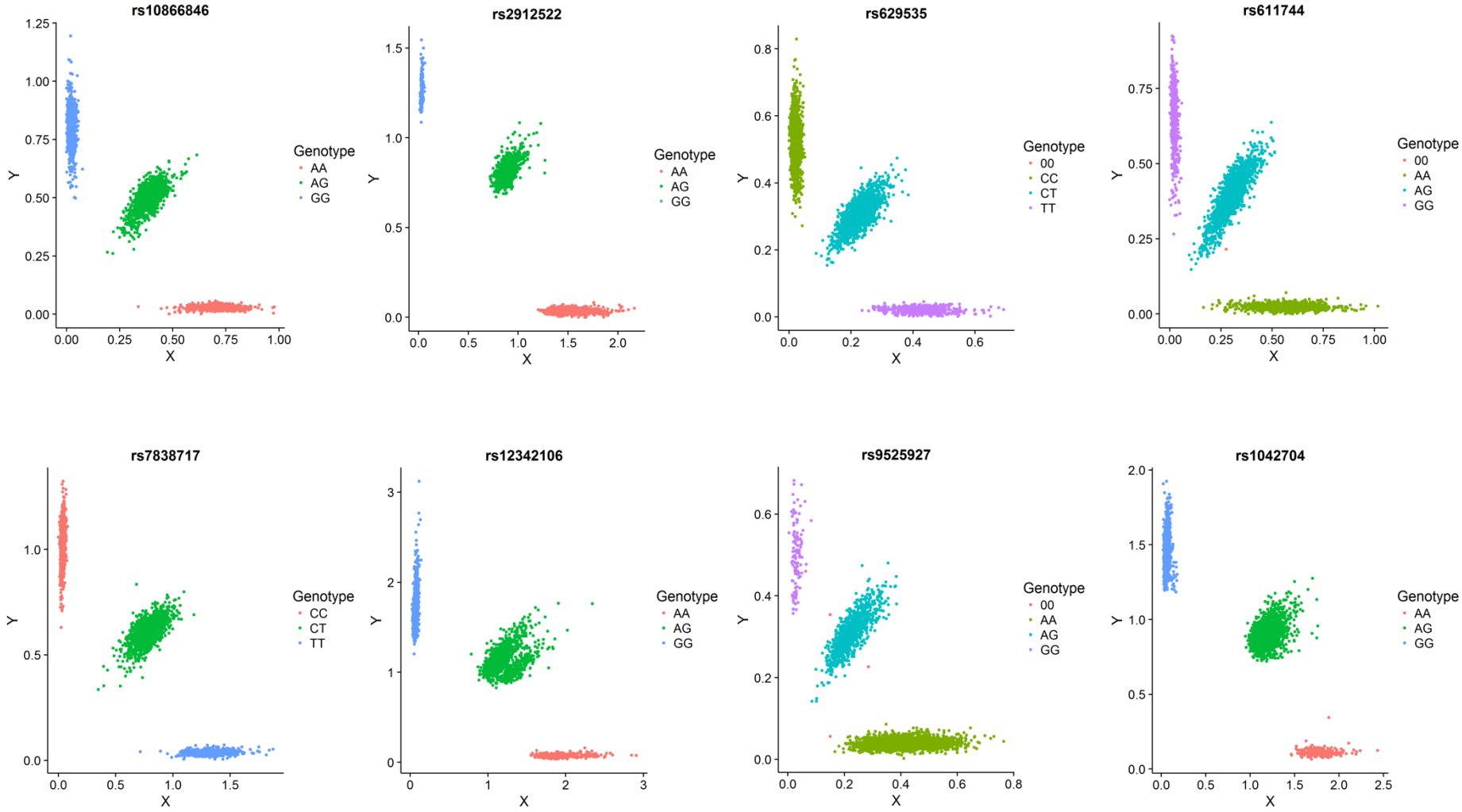
| | | | | | Discovery | | Replication | | Meta-analysis | | | Genes in proximity |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Chromosome | Position | rsID | Allele | EAF | P | OR | P | OR | P | OR | (95% CI) | |
| 1 | 22698447 | rs7524102 | G | 0.214 | $7.68X10^{-12}$ | 1.332 | $6.43X10^{-05}$ | 1.448 | $3.00X10^{-15}$ | 1.351 | 1.254-1.456 | *WNT4, ZBTB40* |
| 1 | 162672011 | rs17433710 | C | 0.120 | $9.13X10^{-07}$ | 0.791 | $3.73X10^{-05}$ | 0.833 | $1.99X10^{-10}$ | 0.813 | 0.763-0.867 | *DDR2* |
| 5 | 108672946 | rs246105 | T | 0.201 | $1.34X10^{-08}$ | 0.796 | | | | | | *PJA2* |
| 6 | 149797014 | rs394563 | T | 0.411 | $2.03X10^{-08}$ | 0.828 | $1.02X10^{-10}$ | 0.827 | $1.14X10^{-17}$ | 0.828 | 0.793-0.864 | *ZC3H12D, TAB2* |

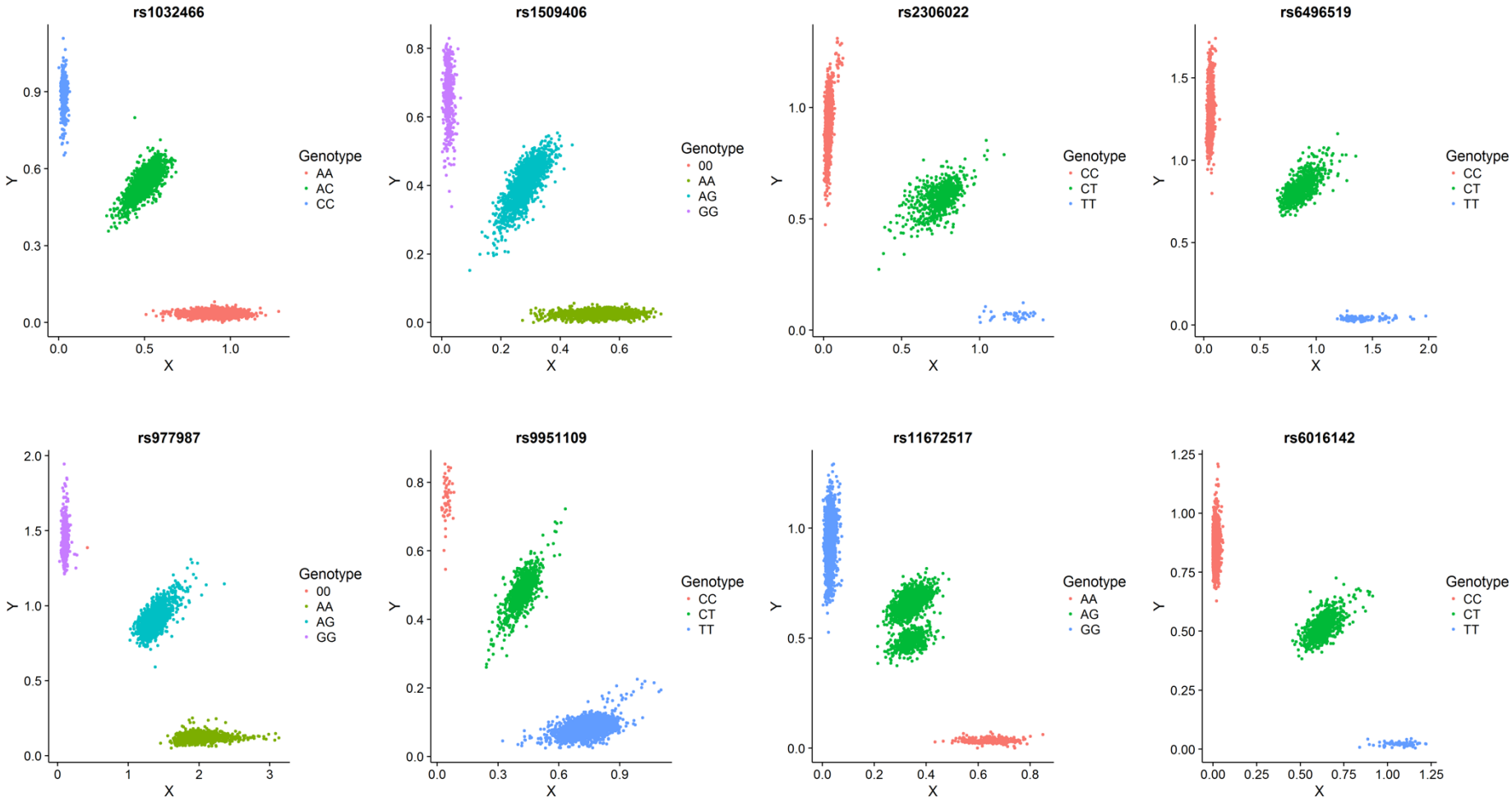
| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 7 | 3318658 | rs10276303 | T | 0.260 | $2.89X10^{-08}$ | 0.817 | $6.00X10^{-09}$ | 0.831 | $9.63X10^{-16}$ | 0.825 | 0.787-0.865 | *SDK1, CARD11* |
| 7 | 37973014 | rs2598107 | T | 0.447 | $1.11X10^{-30}$ | 1.478 | $1.82X10^{-15}$ | 1.475 | $1.55X10^{-44}$ | 1.477 | 1.399-1.560 | *SFRP4, EPDR1* |
| 7 | 37989095 | rs16879765 | T | 0.178 | $7.15X10^{-41}$ | 1.926 | $2.82X10^{-42}$ | 1.837 | $3.38X10^{-81}$ | 1.877 | 1.759-2.002 | *SFRP4, EPDR1* |
| 7 | 116892846 | rs38904 | C | 0.464 | $1.02X10^{-11}$ | 1.254 | $6.52X10^{-13}$ | 1.253 | $4.24X10^{-23}$ | 1.253 | 1.199-1.311 | *WNT2* |
| 8 | 25845675 | rs10866846 | A | 0.421 | $3.14X10^{-11}$ | 1.249 | $1.78X10^{-06}$ | 1.148 | $1.75X10^{-15}$ | 1.190 | 1.140-1.242 | *EBF2* |
| 8 | 69992380 | rs2912522 | G | 0.201 | $1.29X10^{-16}$ | 0.720 | $3.26X10^{-14}$ | 0.751 | $4.09X10^{-29}$ | 0.736 | 0.698-0.777 | *LOC100505718* |
| 8 | 70007938 | rs629535 | T | 0.351 | $2.84X10^{-28}$ | 1.477 | $1.17X10^{-15}$ | 1.275 | $4.31X10^{-40}$ | 1.357 | 1.297-1.420 | *LOC100505718* |
| 8 | 109228008 | rs611744 | G | 0.402 | $3.70X10^{-19}$ | 0.737 | $9.92X10^{-16}$ | 0.794 | $1.15X10^{-32}$ | 0.770 | 0.737-0.804 | *EIF3E, RSPO2* |
| 8 | 145504343 | rs7838717 | T | 0.405 | $2.55X10^{-06}$ | 1.173 | $3.91X10^{-09}$ | 1.188 | $4.81X10^{-14}$ | 1.182 | 1.131-1.234 | *BOP1, HSF1* |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 9 | 1201156 | rs12342106 | A | 0.308 | $9.76X10^{-12}$ | 1.289 | $6.40X10^{-16}$ | 1.290 | $3.78X10^{-26}$ | 1.290 | 1.230-1.352 | *LINC01230, DMRT1* |
| 13 | 44842503 | rs9525927 | G | 0.167 | $5.80X10^{-06}$ | 0.823 | $6.76X10^{-06}$ | 0.842 | $1.82X10^{-10}$ | 0.833 | 0.788-0.881 | *MIR8079, SMIM2* |
| 14 | 23312594 | rs1042704 | A | 0.248 | $8.72X10^{-13}$ | 1.326 | $1.12X10^{-08}$ | 1.213 | $2.49X10^{-19}$ | 1.259 | 1.198-1.324 | *MMP14* |
| 14 | 51074461 | rs1032466 | C | 0.306 | $4.90X10^{-09}$ | 0.812 | $6.82X10^{-10}$ | 0.824 | $1.96X10^{-17}$ | 0.818 | 0.781-0.857 | *ATL1*, *MAP4K5* |
| 15 | 56229760 | rs1509406 | G | 0.356 | $4.03X10^{-06}$ | 1.175 | $1.41X10^{-05}$ | 1.154 | $2.59X10^{-10}$ | 1.164 | 1.110-1.220 | *NEDD4* |
| 15 | 68628163 | rs2306022 | T | 0.110 | $7.59X10^{-06}$ | 1.286 | $2.57X10^{-06}$ | 1.266 | $8.70X10^{-11}$ | 1.275 | 1.185-1.372 | *ITGA11* |
| 15 | 89238184 | rs6496519 | T | 0.164 | $9.35X10^{-08}$ | 0.795 | $1.42X10^{-10}$ | 0.789 | $7.18X10^{-17}$ | 0.791 | 0.749-0.836 | *ISG20, ACAN* |
| 16 | 75506593 | rs977987 | G | 0.403 | $6.24X10^{-07}$ | 1.184 | $8.84X10^{-05}$ | 1.120 | $4.82X10^{-10}$ | 1.146 | 1.098-1.197 | *CHST6, TMEM170A* |
| 18 | 9762933 | rs9951109 | C | 0.133 | $1.24X10^{-07}$ | 0.776 | $8.89X10^{-05}$ | 0.852 | $1.43X10^{-10}$ | 0.820 | 0.771-0.871 | *RAB31* |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 19 | 57678194 | rs11672517 | A | 0.284 | $1.42X10^{-13}$ | 1.331 | $2.71X10^{-05}$ | 1.384 | $1.99X10^{-17}$ | 1.341 | 1.254-1.435 | *DUXA, ZIM3* |
| 20 | 38300807 | rs6016142 | T | 0.132 | $1.19X10^{-06}$ | 1.282 | $1.98X10^{-08}$ | 1.274 | $1.15X10^{-13}$ | 1.277 | 1.197-1.363 | *LINC01370* |
| 20 | 39320751 | rs6102095 | A | 0.125 | $8.54X10^{-07}$ | 0.792 | $1.10X10^{-13}$ | 0.710 | $1.96X10^{-18}$ | 0.748 | 0.701-0.799 | *MAFB* |
| 22 | 46459132 | rs7291412 | T | 0.413 | $1.24X10^{-15}$ | 1.316 | $1.14X10^{-16}$ | 1.269 | $1.53X10^{-30}$ | 1.288 | 1.234-1.345 | *WNT7B* |


rs7524102
Genotype
AA
AG
GG
rs17433710
Genotype
00
CC
CT
TT
rs246105
Genotype
CC
CT
TT
rs394563
Genotype
GG
GT
TT
rs10276303
Genotype
CC
CT
TT
rs2598107
Genotype
CC
CT
TT
rs16879765
Genotype
CC
CT
TT
rs38904
Genotype
00
CC
CT
TT
X
Y


rs10866846
Genotype
AA
AG
GG
rs2912522
Genotype
AA
AG
GG
rs629535
Genotype
00
CC
CT
TT
rs611744
Genotype
00
AA
AG
GG
rs7838717
Genotype
CC
CT
TT
rs12342106
Genotype
AA
AG
GG
rs9525927
Genotype
00
AA
AG
GG
rs1042704
Genotype
AA
AG
GG
X
Y


rs1032466
Genotype
AA
AC
CC
rs1509406
Genotype
00
AA
AG
GG
rs2306022
Genotype
CC
CT
TT
rs6496519
Genotype
CC
CT
TT
rs977987
Genotype
00
AA
AG
GG
rs9951109
Genotype
CC
CT
TT
rs11672517
Genotype
AA
AG
GG
rs6016142
Genotype
CC
CT
TT
X
Y

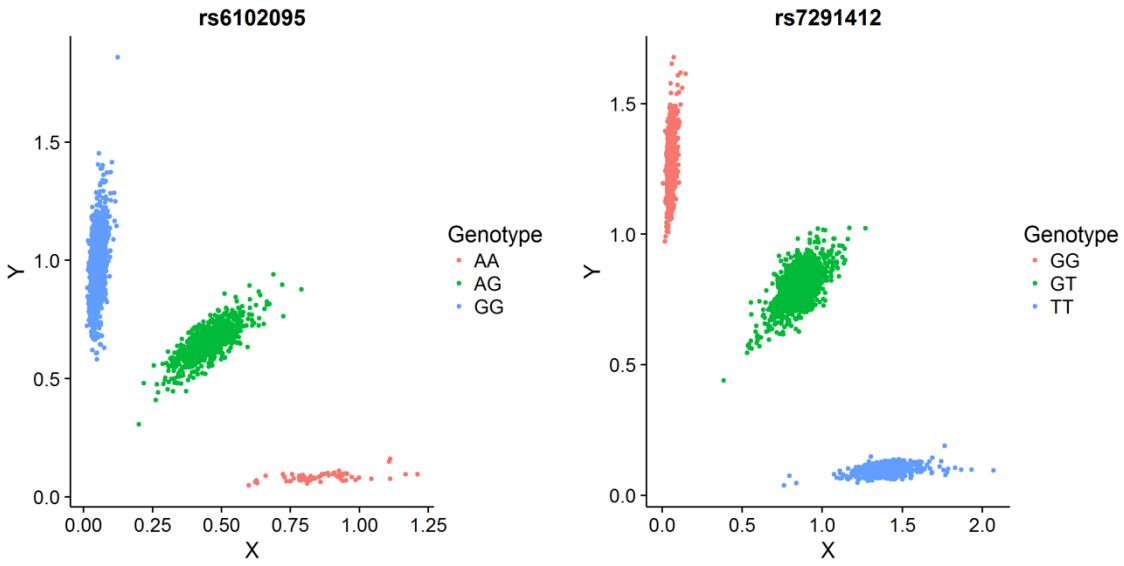
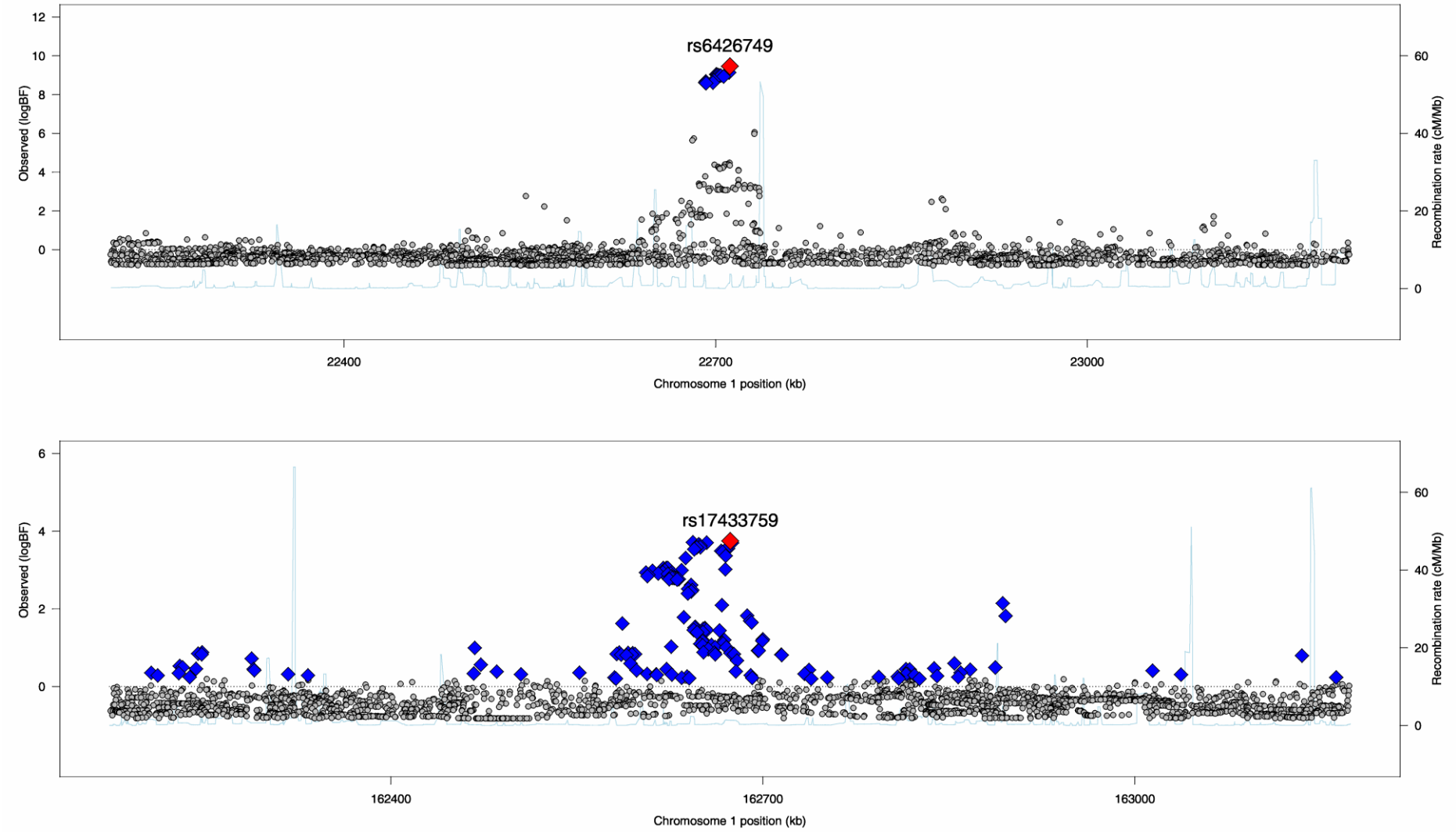
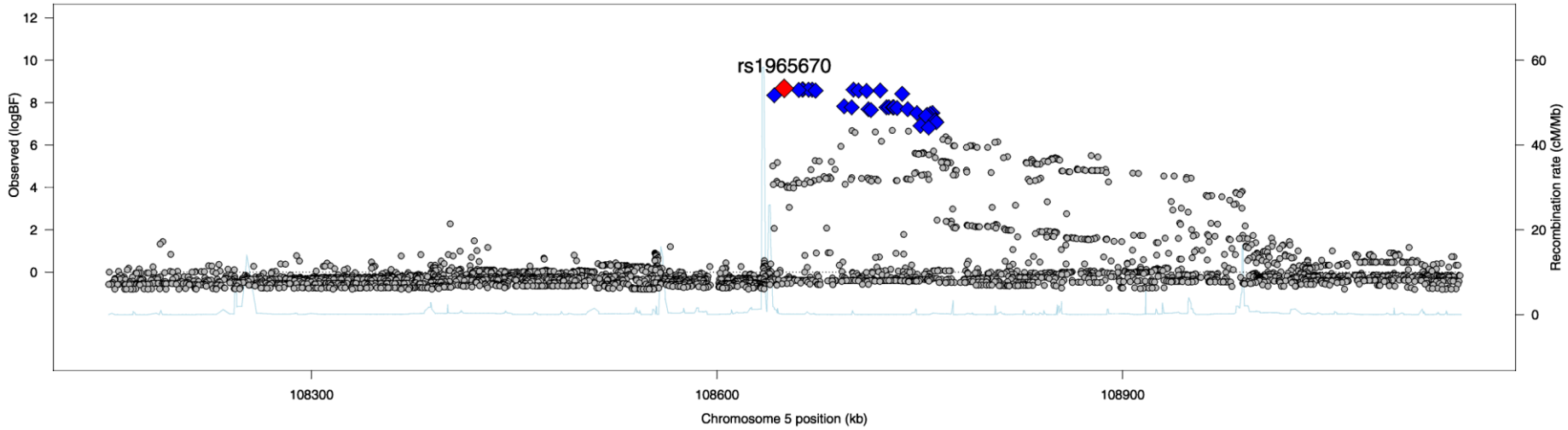


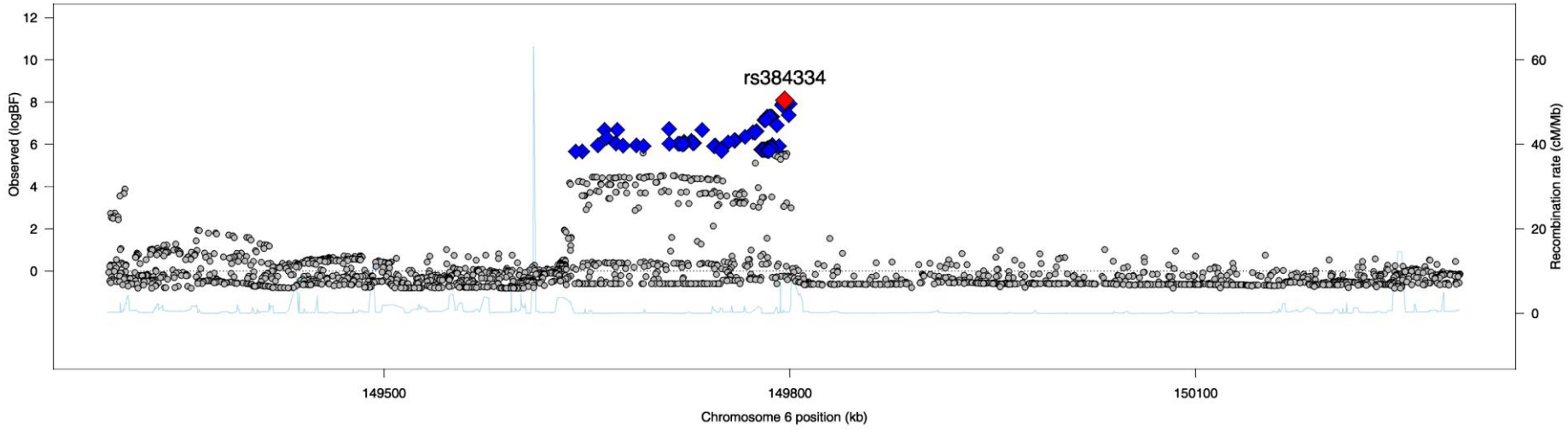

Figure 2.5.2.4 Scatter plots showing the fidelity of the genotype calls of the 26 reported associated variants. Each dot in the scatter plot represents the raw intensity of a genotype from a single sample in the discovery phase GWAS.

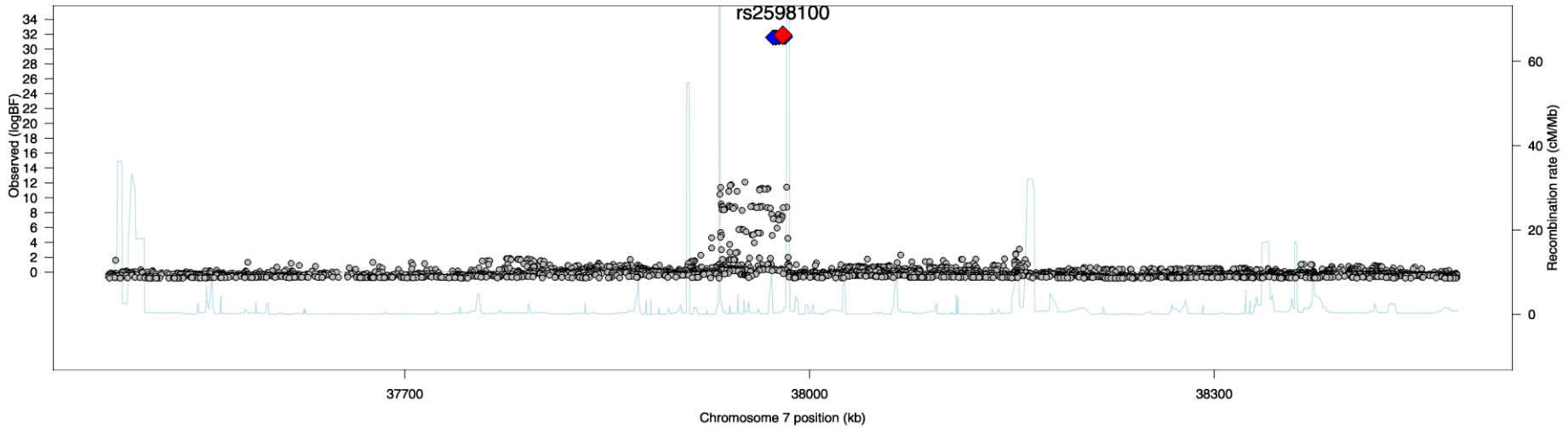
### 2.5.3 Fine mapping of the DD associated loci

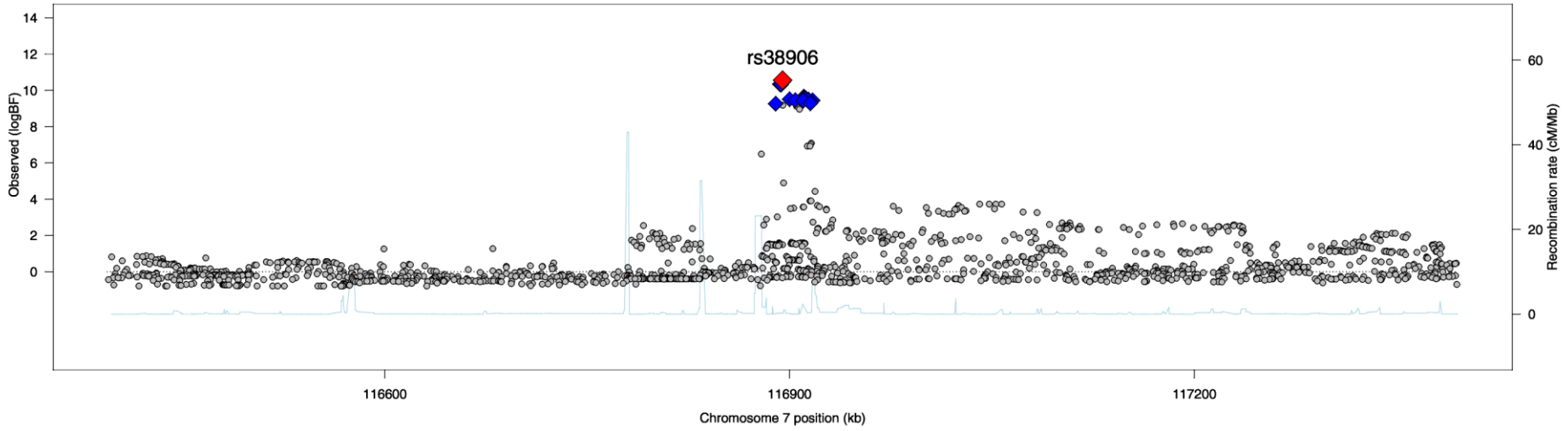
To rationalise the possible causal variants for functional work, fine mapping was performed by imputation and the generation of 99% credible sets on the post QC discovery phase GWAS dataset. Imputation was performed using version 1.1 of the Haplotype Reference Consortium (HRC) reference panel, and only on the common SNPs (n=237,613) of the pre-phased Dupuytren GWAS dataset, which contained no multi-allelic variants or singletons. After imputation, a list of QC thresholds was applied to the dataset to remove SNPs with a low imputation confidence score (INFO<0.4), MAF less than 1%, and that are deviated from HWE ($p < 0.000001$). Bayes factors were calculated on the QC dataset containing 7,218,238 SNPs, and 99% credible sets of 1Mb associated regions centred on the lead associated SNP were subsequently created (Figure 2.5.3.1). Similar to the discovery GWAS workflow, conditional analyses were performed on regions harbouring multiple associated regions on chromosome 7 (rs117402009) and 8 (rs2472141), and residual associations rs2598100 ($BF_{cond} = 7.29 \times 10^{31}$) and rs2981040 ($BF_{cond} = 1.67 \times 10^{10}$) were identified. Each 99% credible set contains a range of variants, from one up to 293. The lone variant in the 99% credible set on chromosome 14 (rs1042704), is a coding variant in MMP14, and these data strongly suggest it is causative.

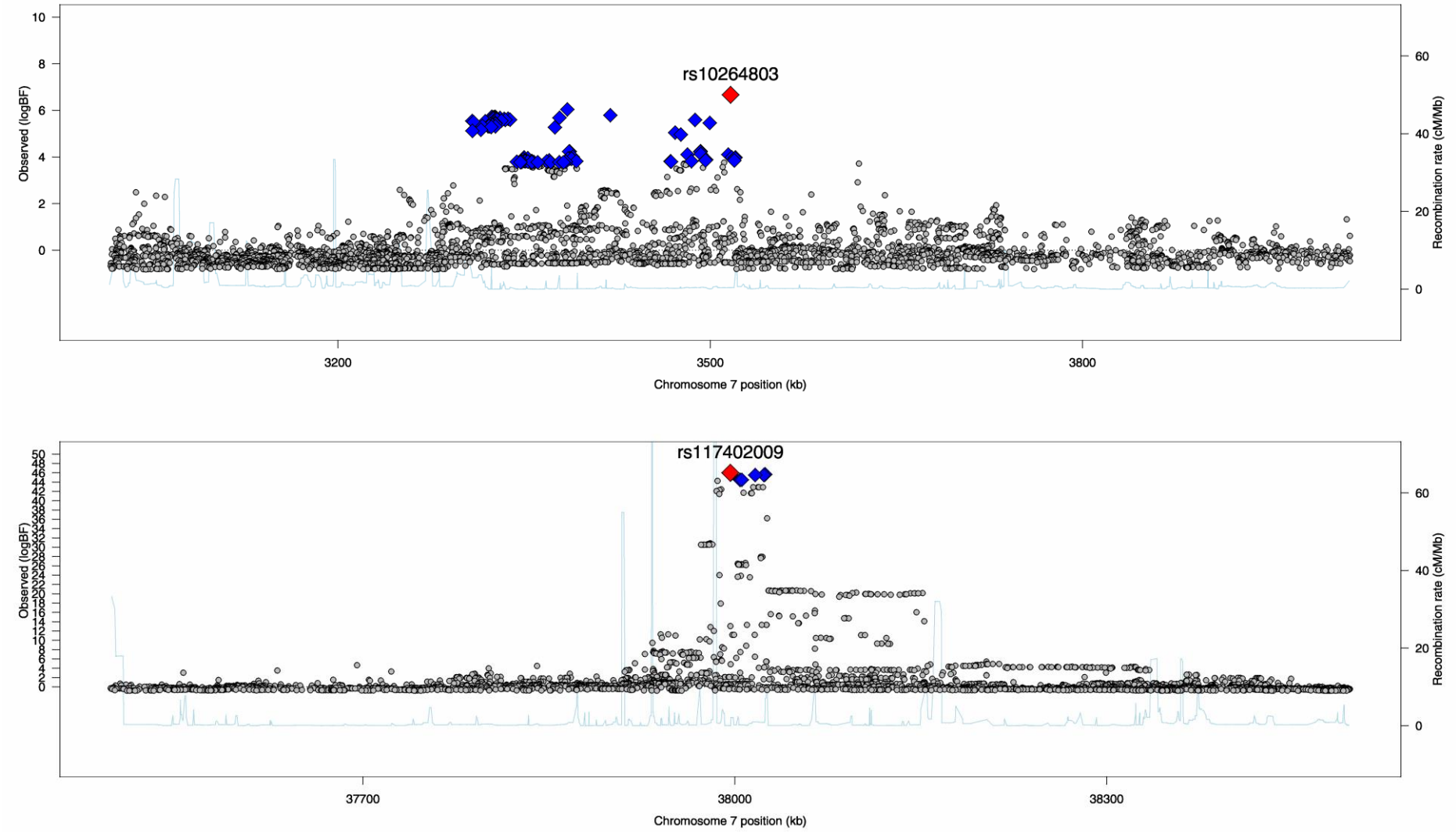

rs6426749
Observed (logBF)
Recombination rate (cM/Mb)
22400
22700
23000
Chromosome 1 position (kb)
rs17433759
Observed (logBF)
Recombination rate (cM/Mb)
162400
162700
163000
Chromosome 1 position (kb)

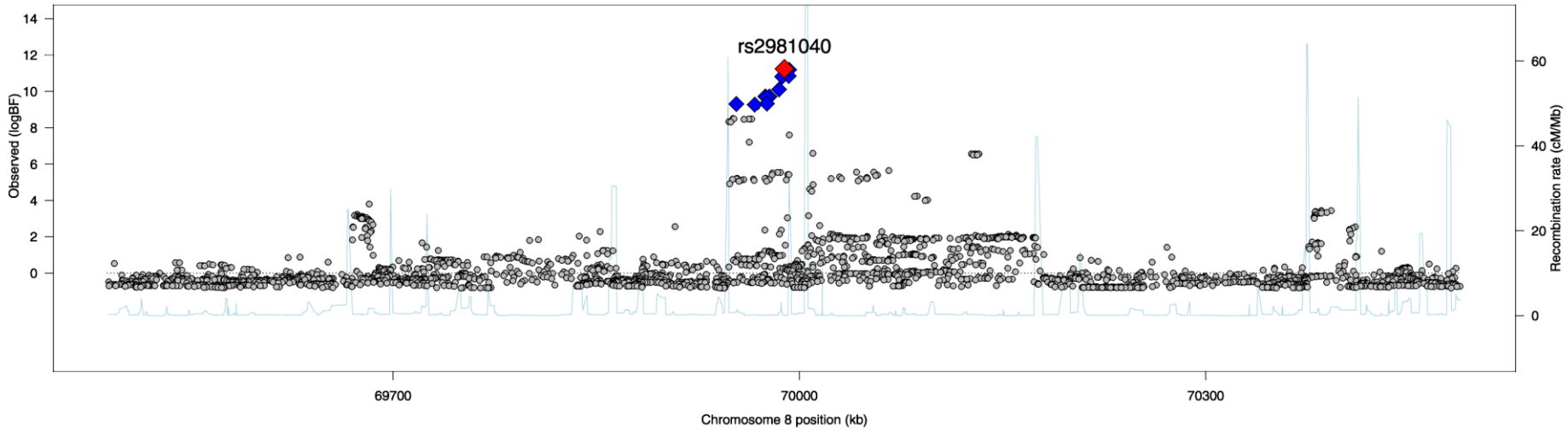

rs1965670
Observed (logBF)
Recombination rate (cM/Mb)
108300
108600
108900
Chromosome 5 position (kb)

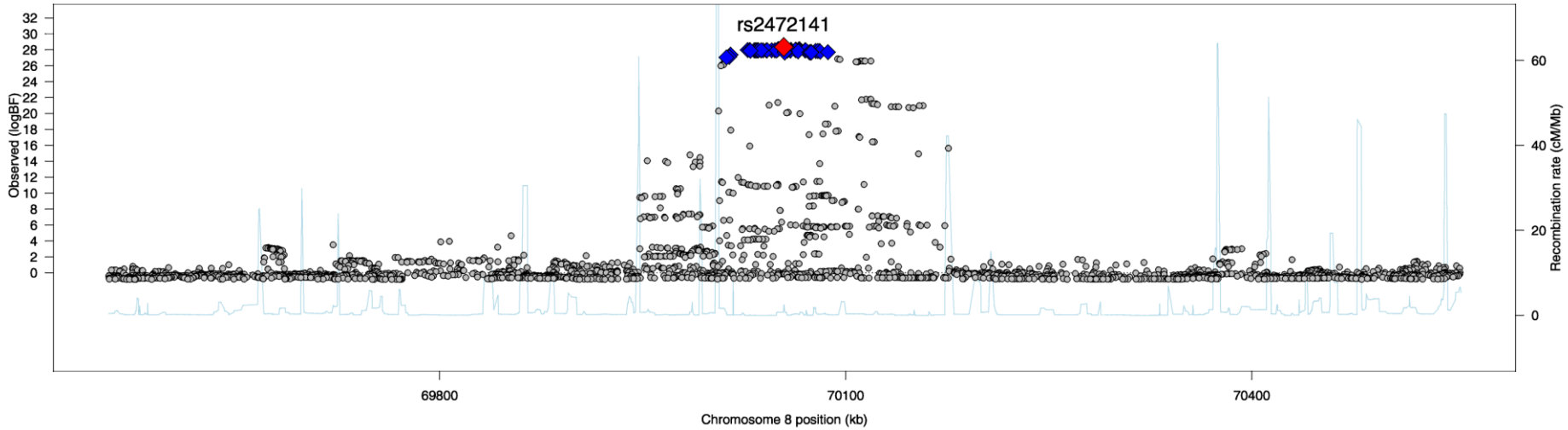

rs384334
Observed (logBF)
Recombination rate (cM/Mb)
149500
149800
150100
Chromosome 6 position (kb)

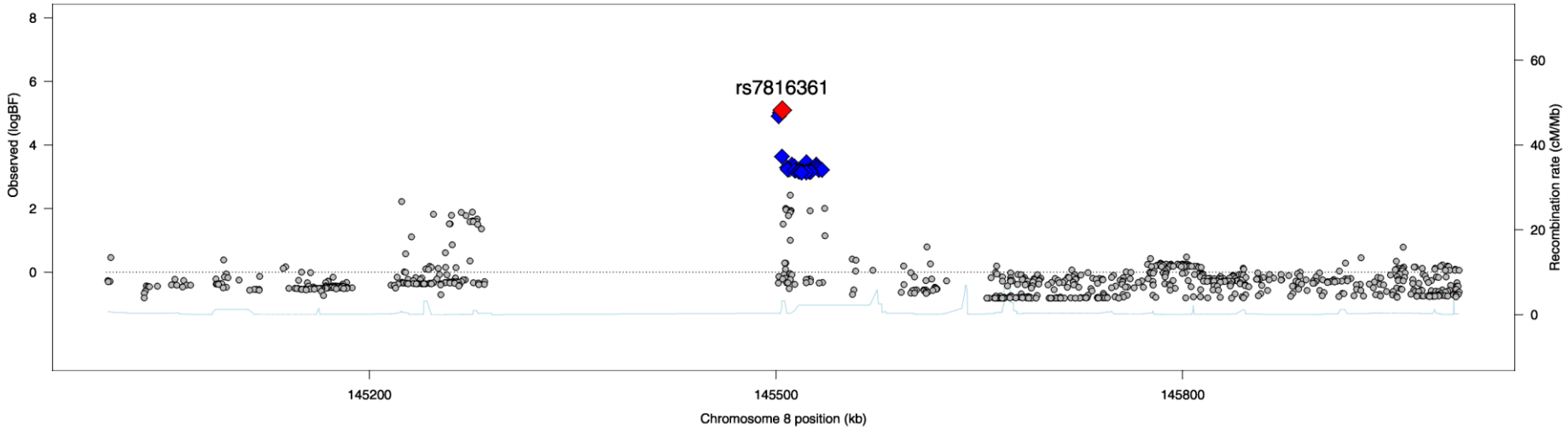

rs2598100
Observed (logBF)
Recombination rate (cM/Mb)
37700
38000
38300
Chromosome 7 position (kb)

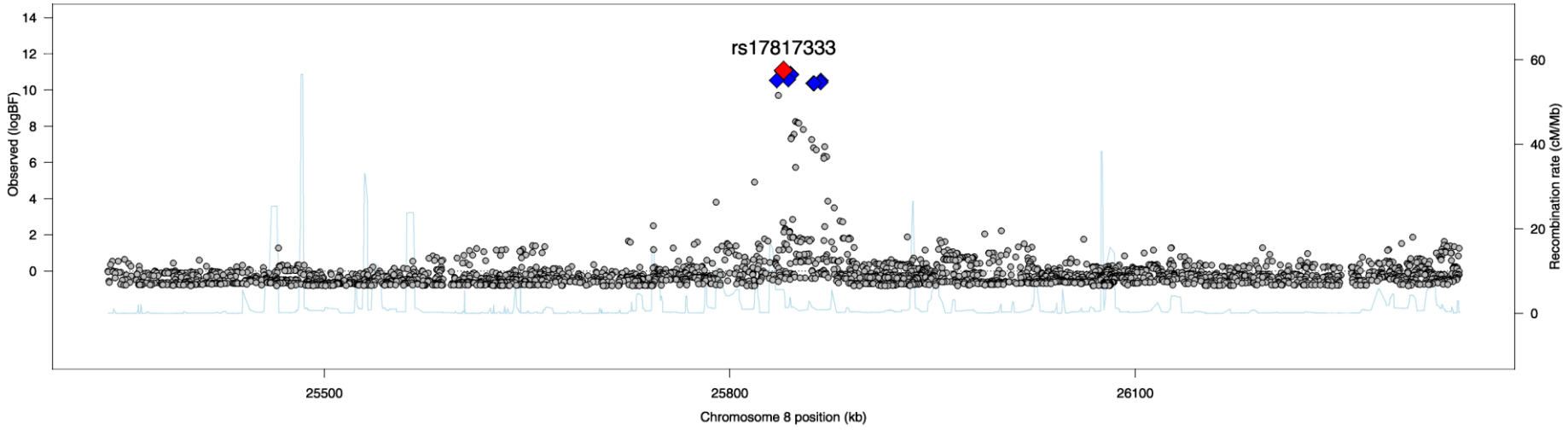

rs38906
Observed (logBF)
Recombination rate (cM/Mb)
116600
116900
117200
Chromosome 7 position (kb)

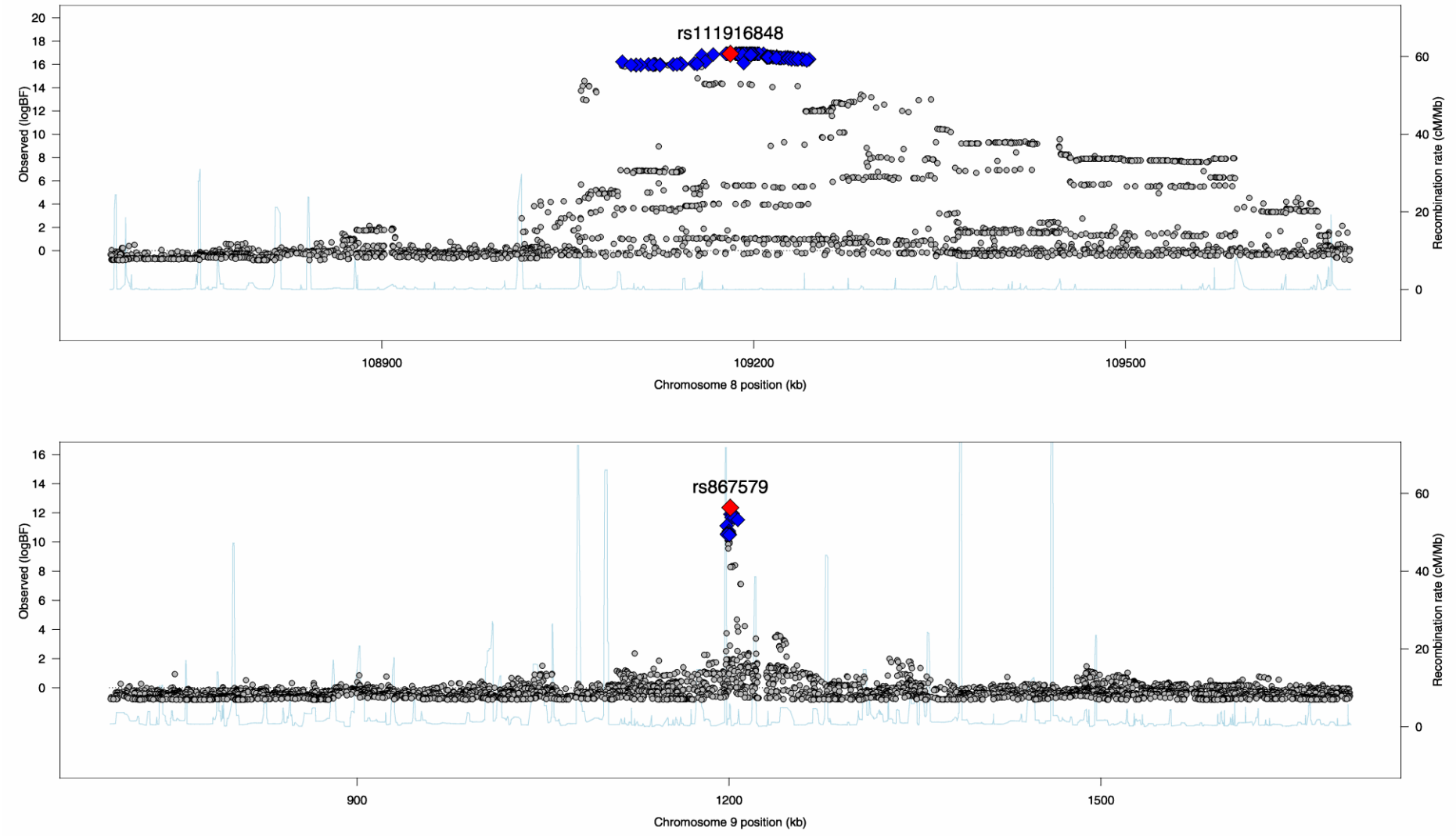

rs10264803
Observed (logBF)
Recombination rate (cM/Mb)
Chromosome 7 position (kb)
3200
3500
3800
rs117402009
37700
38000
38300

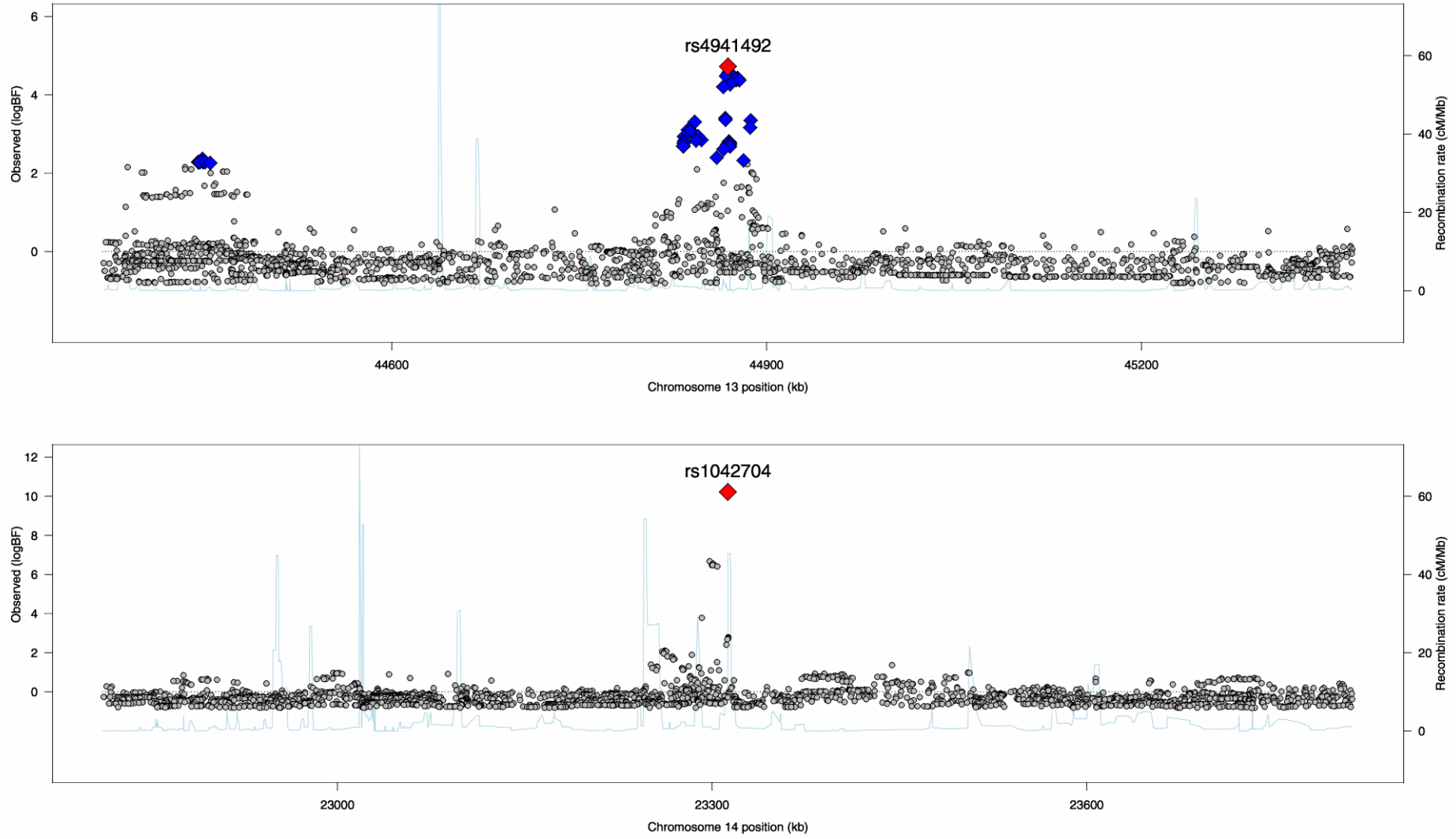

rs2981040
Observed (logBF)
Recombination rate (cM/Mb)
Chromosome 8 position (kb)
69700
70000
70300

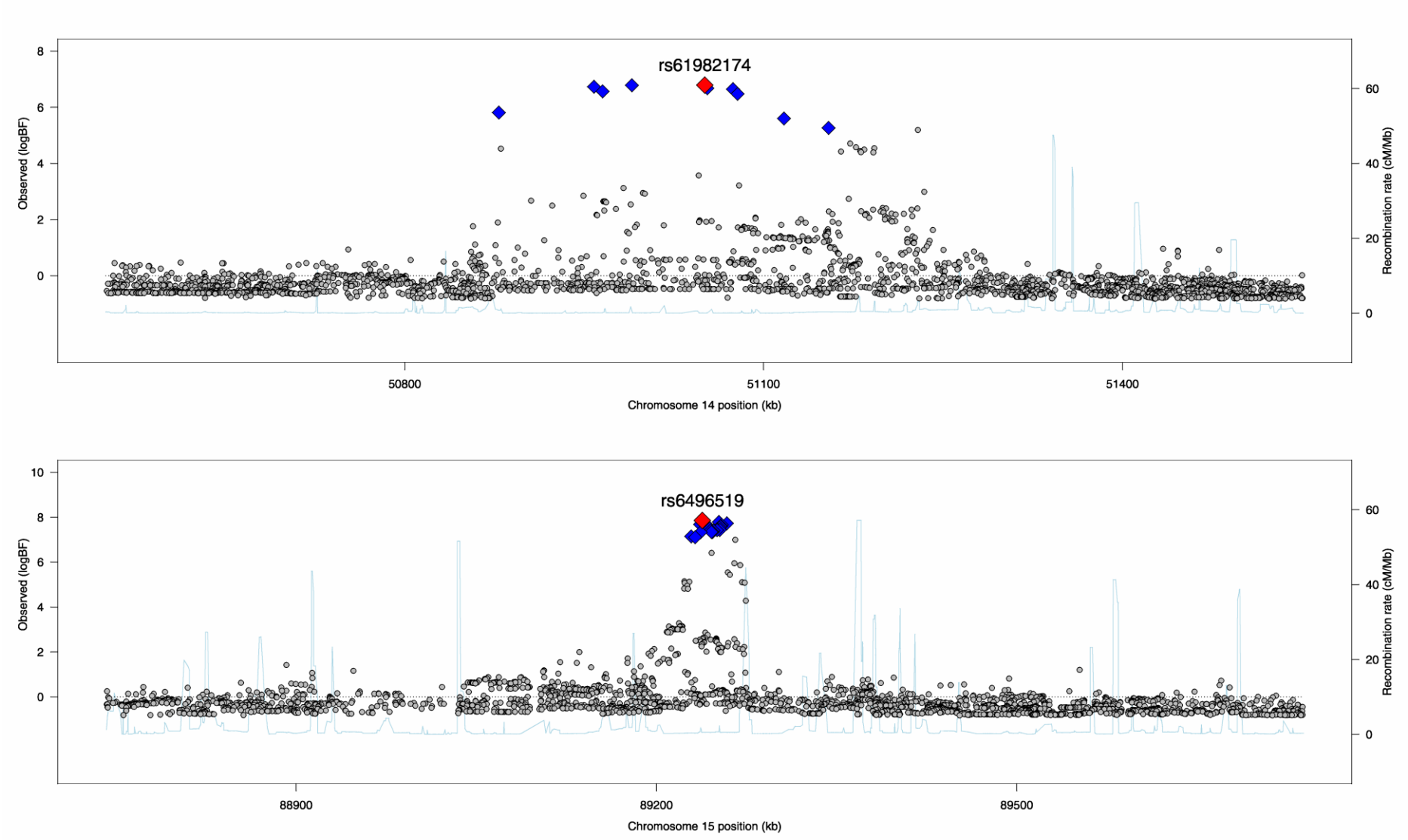

rs2472141
Observed (logBF)
Recombination rate (cM/Mb)
Chromosome 8 position (kb)
69800
70100
70400

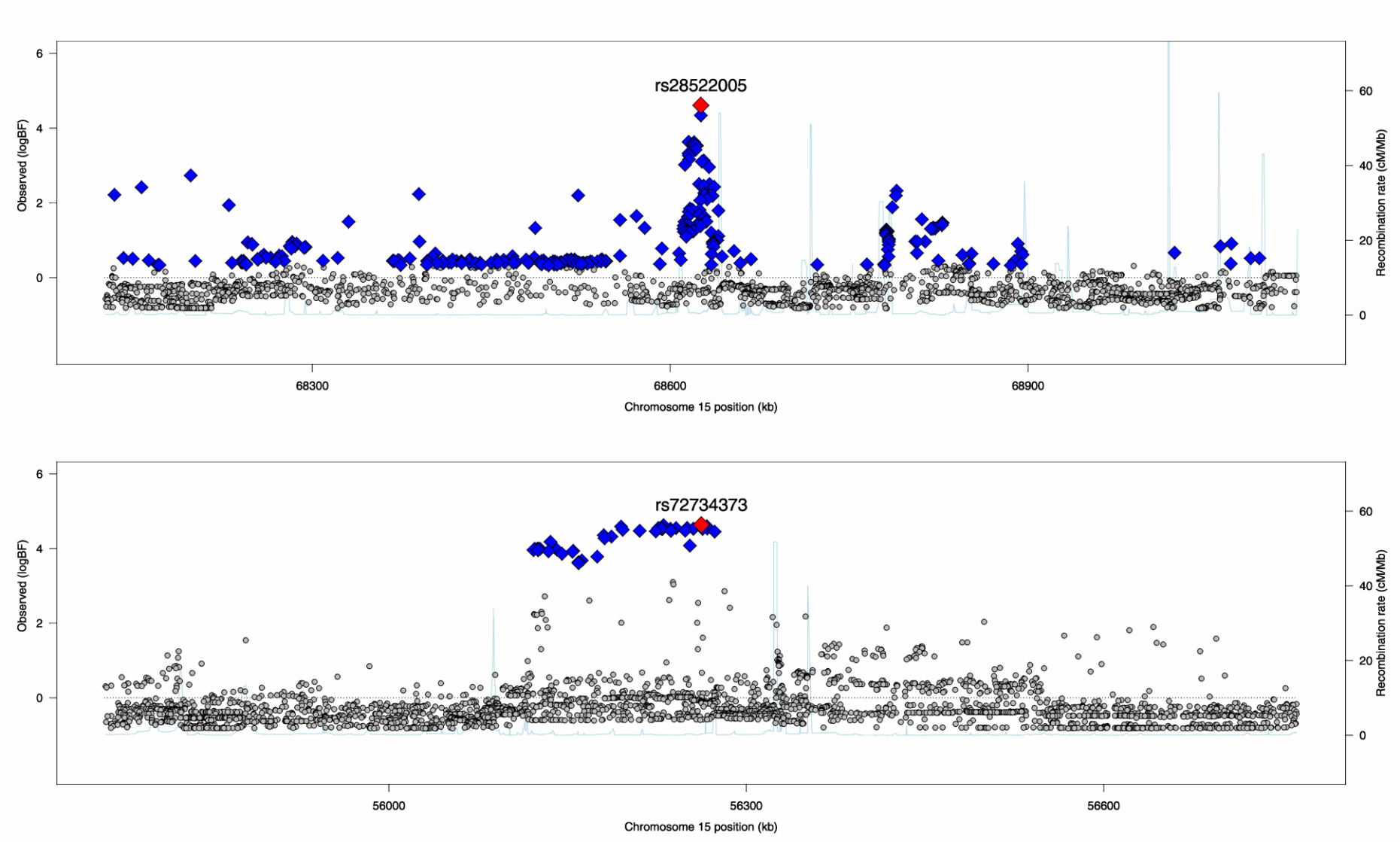

rs7816361
Observed (logBF)
Recombination rate (cM/Mb)
145200
145500
145800
Chromosome 8 position (kb)

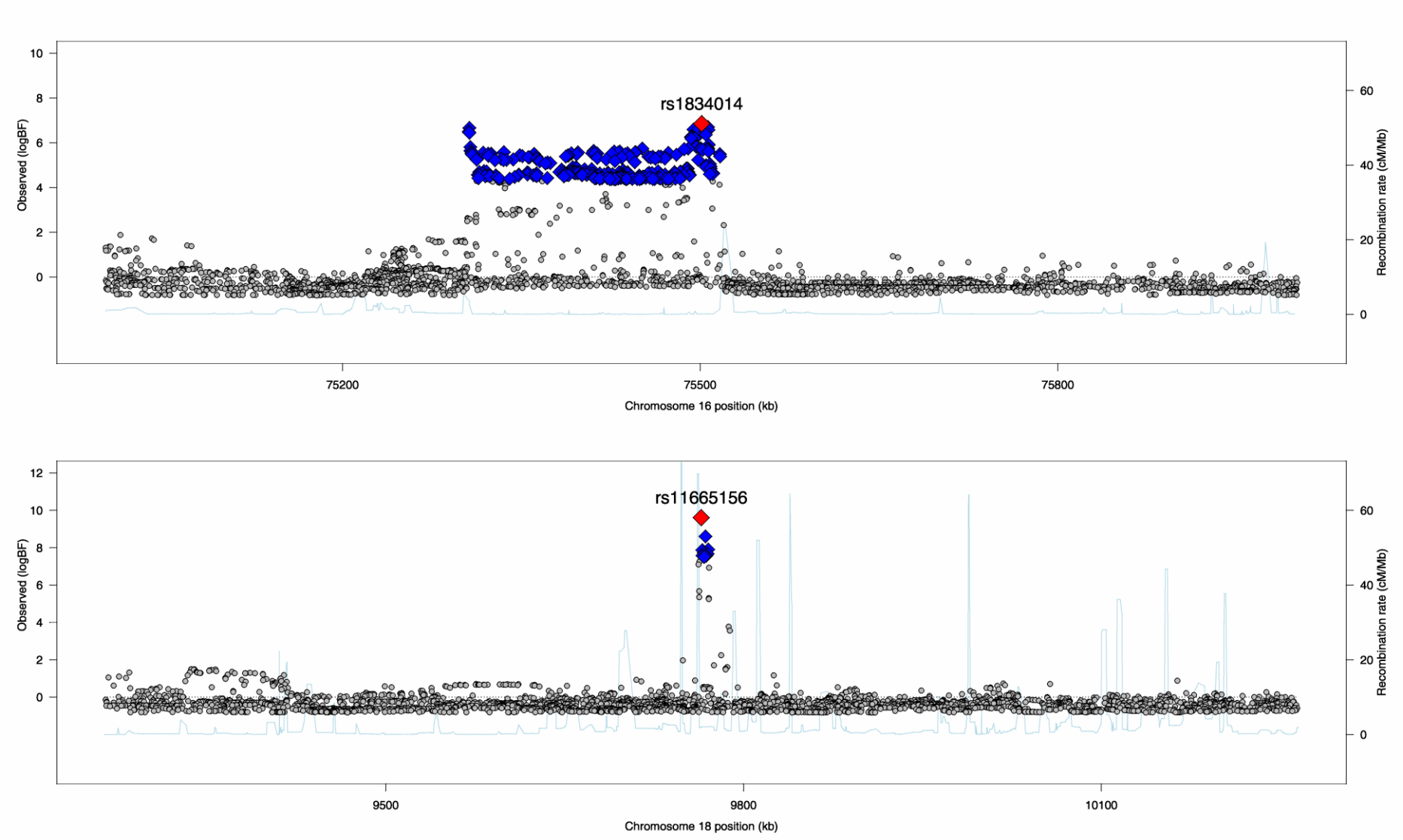

rs17817333
Observed (logBF)
Recombination rate (cM/Mb)
25500
25800
26100
Chromosome 8 position (kb)

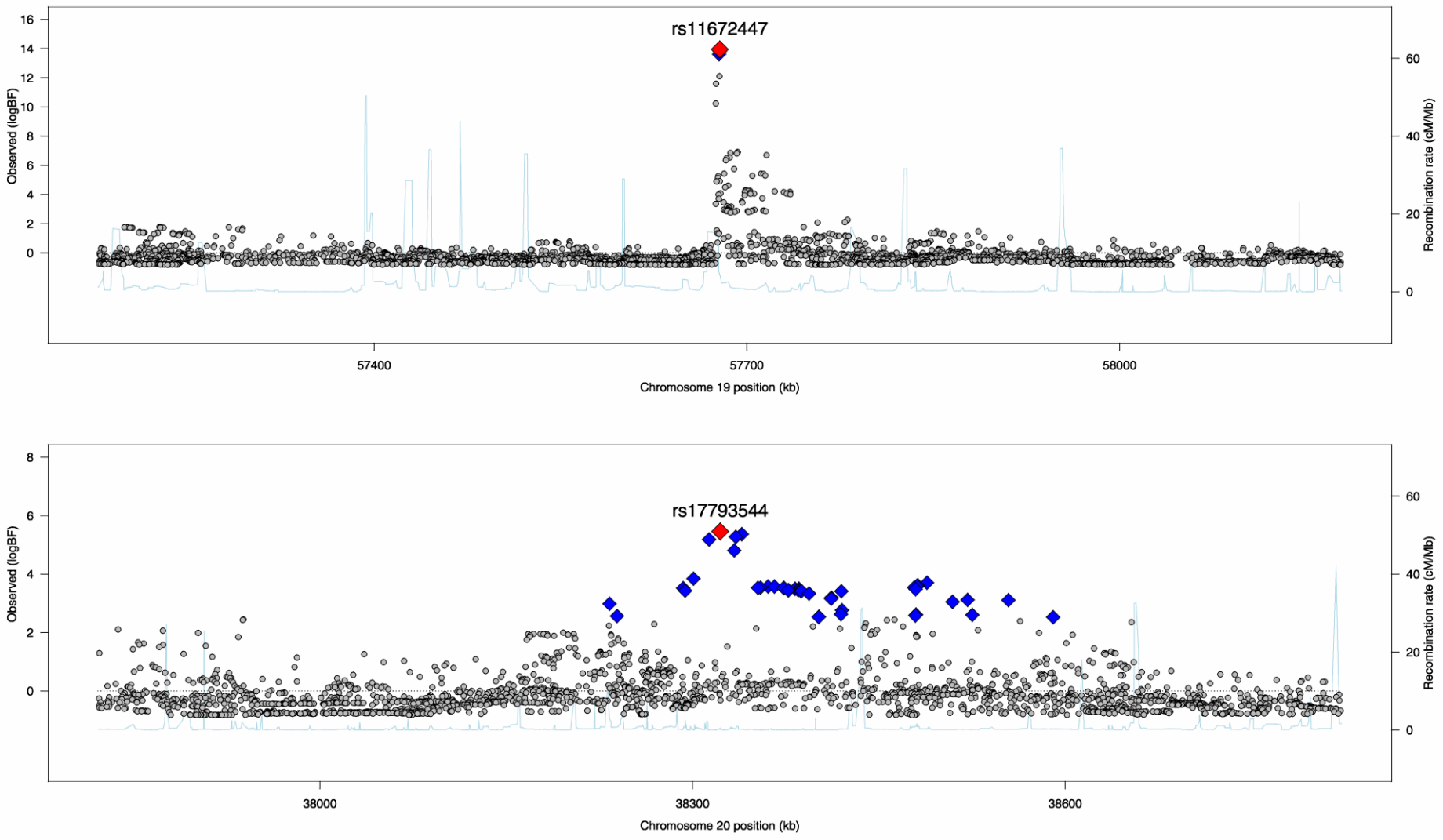

rs111916848
Observed (logBF)
Recombination rate (cM/Mb)
108900
109200
109500
Chromosome 8 position (kb)
rs867579
Observed (logBF)
Recombination rate (cM/Mb)
900
1200
1500
Chromosome 9 position (kb)

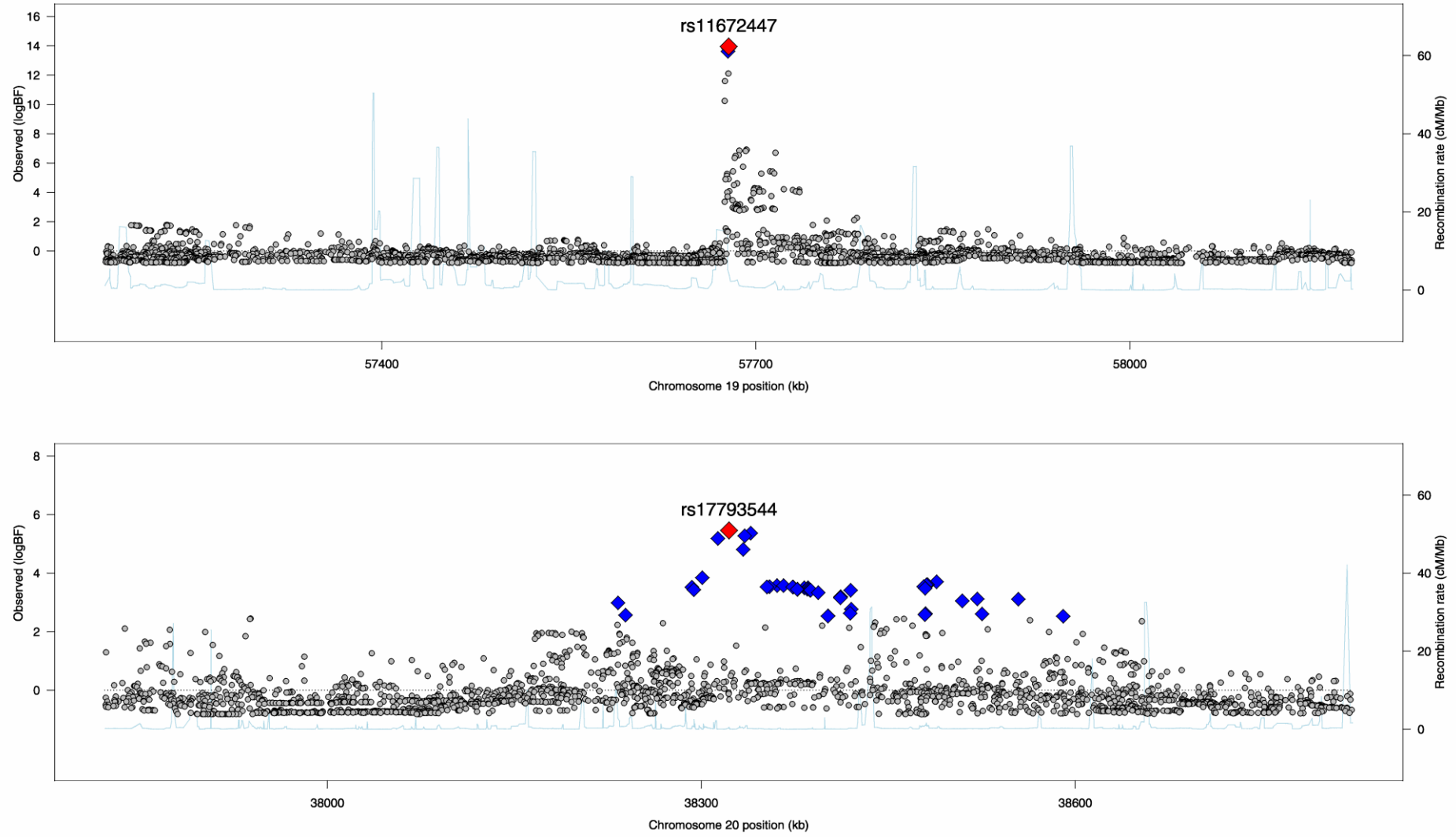

rs4941492
Observed (logBF)
Recombination rate (cM/Mb)
Chromosome 13 position (kb)
44600
44900
45200
rs1042704
Observed (logBF)
Recombination rate (cM/Mb)
Chromosome 14 position (kb)
23000
23300
23600

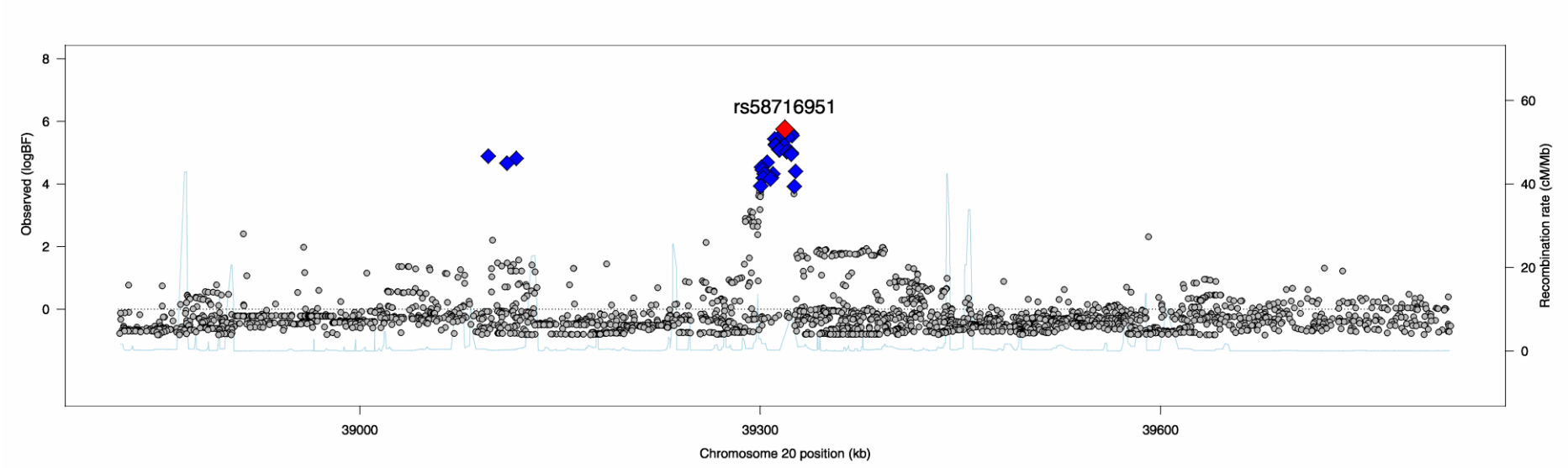

rs61982174
Observed (logBF)
Recombination rate (cM/Mb)
50800
51100
51400
Chromosome 14 position (kb)
rs6496519
Observed (logBF)
Recombination rate (cM/Mb)
88900
89200
89500
Chromosome 15 position (kb)


rs28522005
Observed (logBF)
Recombination rate (cM/Mb)
Chromosome 15 position (kb)
68300
68600
68900
rs72734373
56000
56300
56600


rs1834014
Observed (logBF)
Recombination rate (cM/Mb)
75200
75500
75800
Chromosome 16 position (kb)
rs11665156
Observed (logBF)
Recombination rate (cM/Mb)
9500
9800
10100
Chromosome 18 position (kb)


rs11672447
Observed (logBF)
Recombination rate (cM/Mb)
Chromosome 19 position (kb)
57400
57700
58000
rs17793544
Observed (logBF)
Recombination rate (cM/Mb)
Chromosome 20 position (kb)
38000
38300
38600


rs11672447
Observed (logBF)
Recombination rate (cM/Mb)
57400
57700
58000
Chromosome 19 position (kb)
rs17793544
Observed (logBF)
Recombination rate (cM/Mb)
38000
38300
38600
Chromosome 20 position (kb)


rs58716951
Observed (logBF)
Recombination rate (cM/Mb)
Chromosome 20 position (kb)
39000
39300
39600
8
6
4
2
0
60
40
20
0

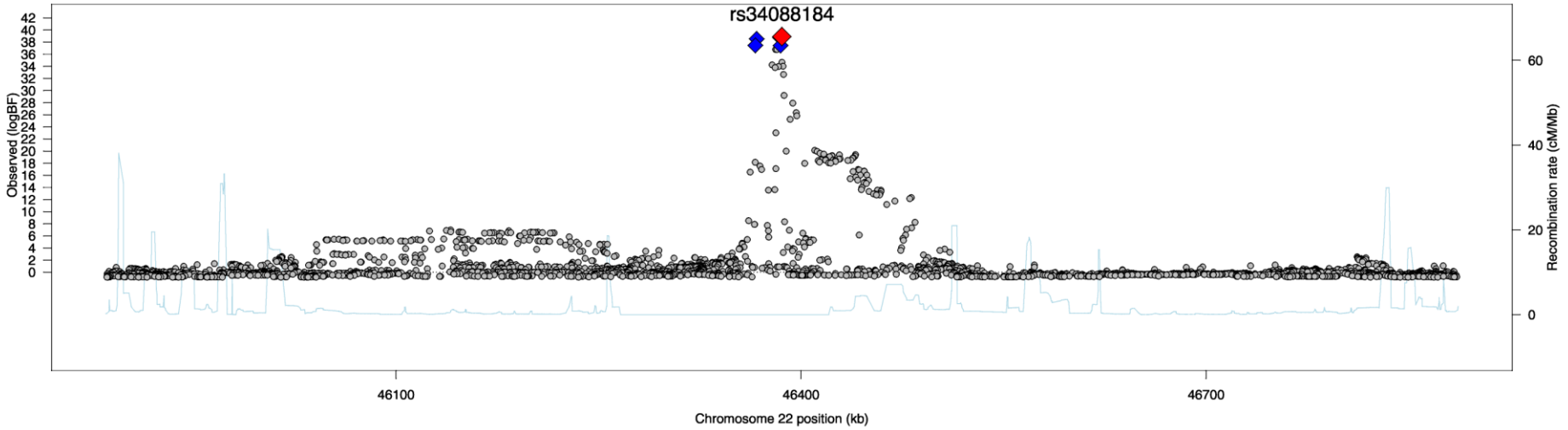



Figure 2.5.3.1 Regional plots showing 99% credible sets of all 26 associated regions that were generated after imputation for fine mapping analysis. The lead SNP is shown in red, other SNPs in the 99% credible set are shown in blue, and other SNPs not part of the 99% credible set are shown in grey.

### 2.5.4 The functional evidence of the most significantly associated locus with Dupuytren disease

On chromosome 7, the most significant DD associated variant rs16879765 (pMeta = $3.38 \times 10^{-81}$) is located in the second intronic region of *EPDR1* and within 50Mb of the promoter of *SFRP4*. Using the FibroP DNase I hypersensitive site track on the UCSC genome browser revealed no clear biological function of this variant as it is not located in an exon or an open chromatin region, indicating a lack of transcriptional activity of this region in this cell type. This evidence strongly suggested that the causative variant is not rs16879765 but most likely any of the 10 variants within the same linkage disequilibrium block reported in the credible set analysis (Figure 2.5.4.1). Analysis of the 10 fine mapped variants using 3DSNP and regulomeDb provided some evidence that eight are located within an enhancer region, and in particular rs117402009 and rs17171240 could affect the binding motif of transcription factors BCL6 and MYBL1. Further investigation using 3D genomic browser and eQTL from GTEx on rs117402009 and rs17171240 revealed that these two variants are located in the same 3'UTR region of *EPDR1*, and the expression level of *EPDR1* is affected by their polymorphisms but only in stomach tissue. Further *in silico* analysis using Virtual4C track, which was created using Hi-C data from lung cells, revealed a chromatin loop between this region and the 3'UTR of *SFRP4*, which has a cis-interaction with the promoter region of both *EPDR1* and *SFPR4*.

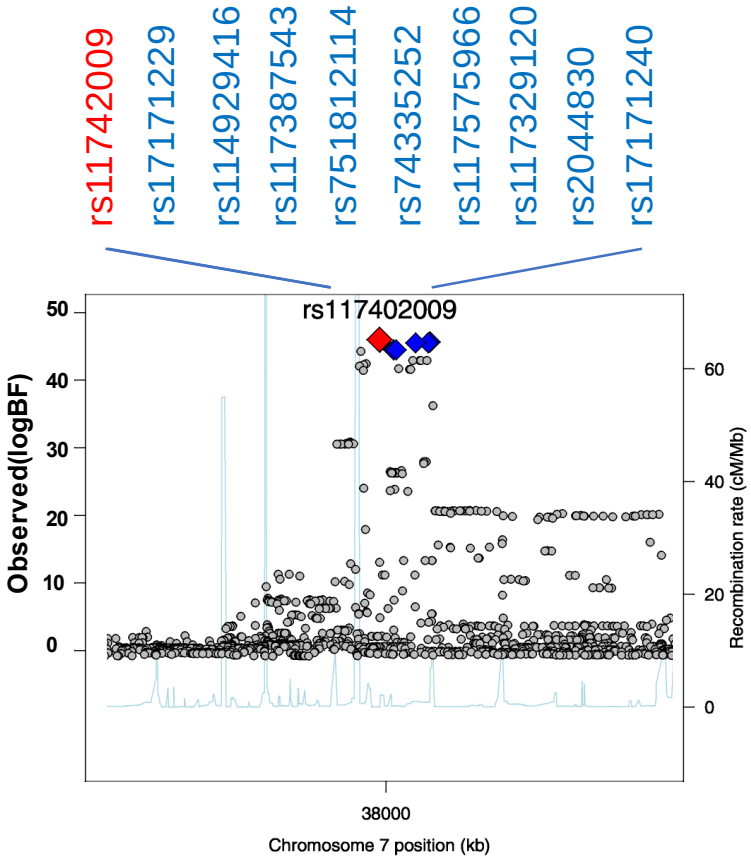



Figure 2.5.4.1 Regional plot depicting the order of the SNPs in the 99% credible set and same LD block as rs16879765.

### 2.5.5 The molecular function of SFRP4 in myofibroblasts

As there was lack of concrete *in silico* evidence to indicate which variant was causative, initial functional work around the *EPDR1* region on chromosome 7 was performed based on the available genotypes of rs16879765. Firstly, the basal expression of nearby genes *EPDR1, NME8* and *SFRP4* was measured using cultured myofibroblast cells that were disaggregated from surgically resected DD tissue. Out of the three genes tested, *NME8* expression was undetectable (Ct>38) and therefore excluded from further analysis. Differential analysis was performed on 26 samples, consisting samples of each genotype at rs16879765 (CC n=10; CT n=9; TT n=7), on *EPDR1* and *SFRP4*. The mean expression of *SFRP4* of cells carrying homozygous TT was significantly higher than those carrying homozygous CC (Figure 2.5.5.1) but neither of the homozygous genotypes was associated with significantly different expression compared to the heterozygous CT genotype. *EPDR1*, which was expressed at a much lower basal level in comparison to *SFRP4* in our experimental system, also demonstrated a similar differential pattern between the two homozygous genotypes (Figure 2.5.5.1) but did not reach statistical significance.

Prior to performing any SFRP4 protein functional work, *SFRP4* expression was confirmed histologically on Formalin fixed paraffin embedded (FFPE) DD tissue (Figure 2.5.5.2) and in DD derived primary myofibroblast cells (Figure 2.5.5.3) using SFRP4 specific antibodies and either DAB stain or fluorescence conjugated secondary antibody. Differential protein expression between the homozygous genotypes was

confirmed using the supernatant of the cultured myofibroblast cells (CC=9; TT= 6) and their absolute concentration, which was measured using the SFRP4 ELISA kit after confirming its working range. In contrast to the differential expression pattern detected at the mRNA level, the mean SFRP4 protein expression in the supernatant with cells carrying genotype CC was significantly higher than genotype TT (Figure 2.5.5.4) after 7 days of culture. Part of the SFRP4 biological function was also tested in my culture system in the hope of unravelling some of its relevance to the WNT canonical signalling pathway and DD development. Due to the antagonistic nature to WNT3A of SFRP4, myofibroblast cells were first treated with 200ng/ml WNT3A for six hours, which resulted in a significant elevation in mRNA expression of *ACTA2*, *AXIN2* and *CTGF* but not other fibrotic genes such as *COL1A1*, *COL3A1* and *ACTB*. Elevation in *CTGF* was attenuated by 8 μg/ml SFRP4 or 100ng/ml DKK1 (Figure 2.5.5.5).

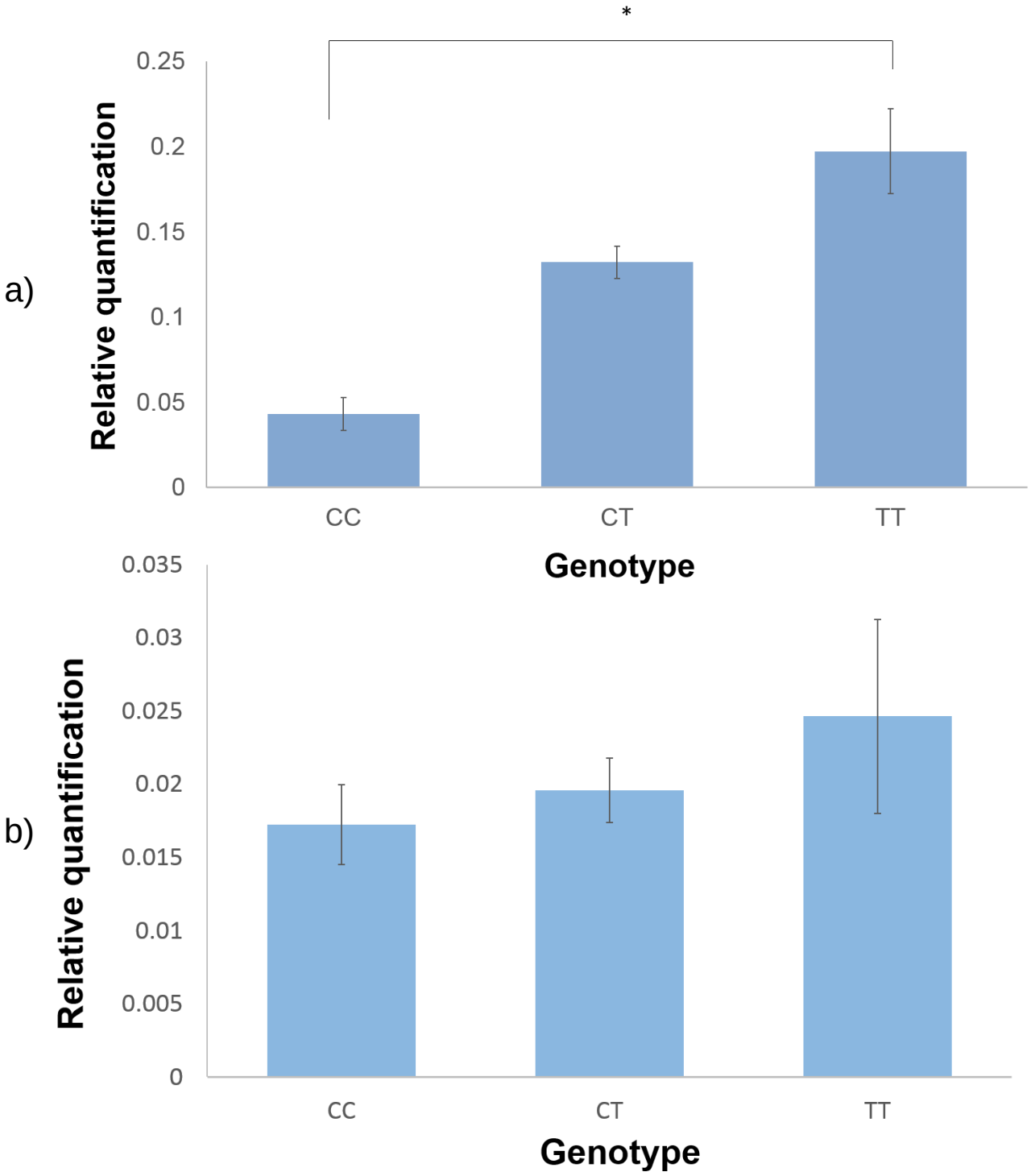



Figure 2.5.5.1 Bar chart depicting the relative expression of a) *SFRP4* and b) *EPDR1* to housekeeping gene *18S* stratified by genotype at rs16879765 (CC = 10, CT = 9, TT = 7). *p < 0.05

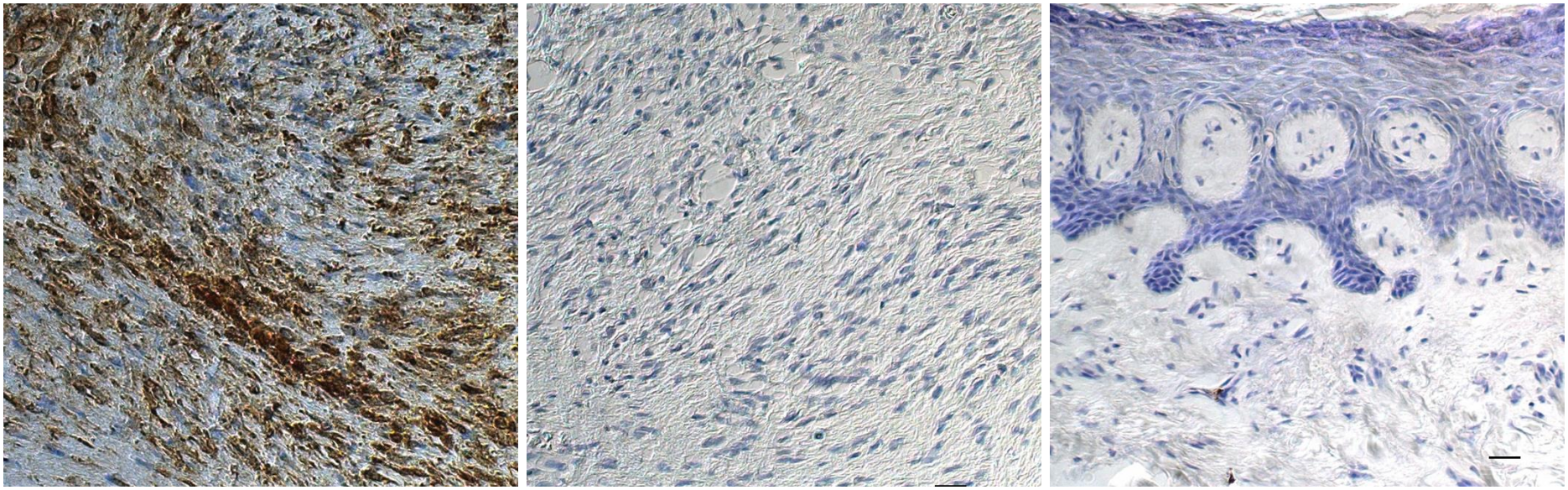


Figure 2.5.5.2 20x microscopic images of SFRP4 DAB stained FFPE DD tissue (left panel), isotype control (centre panel), and SFRP4 DAB stained palmar skin tissue (negative control; right panel).

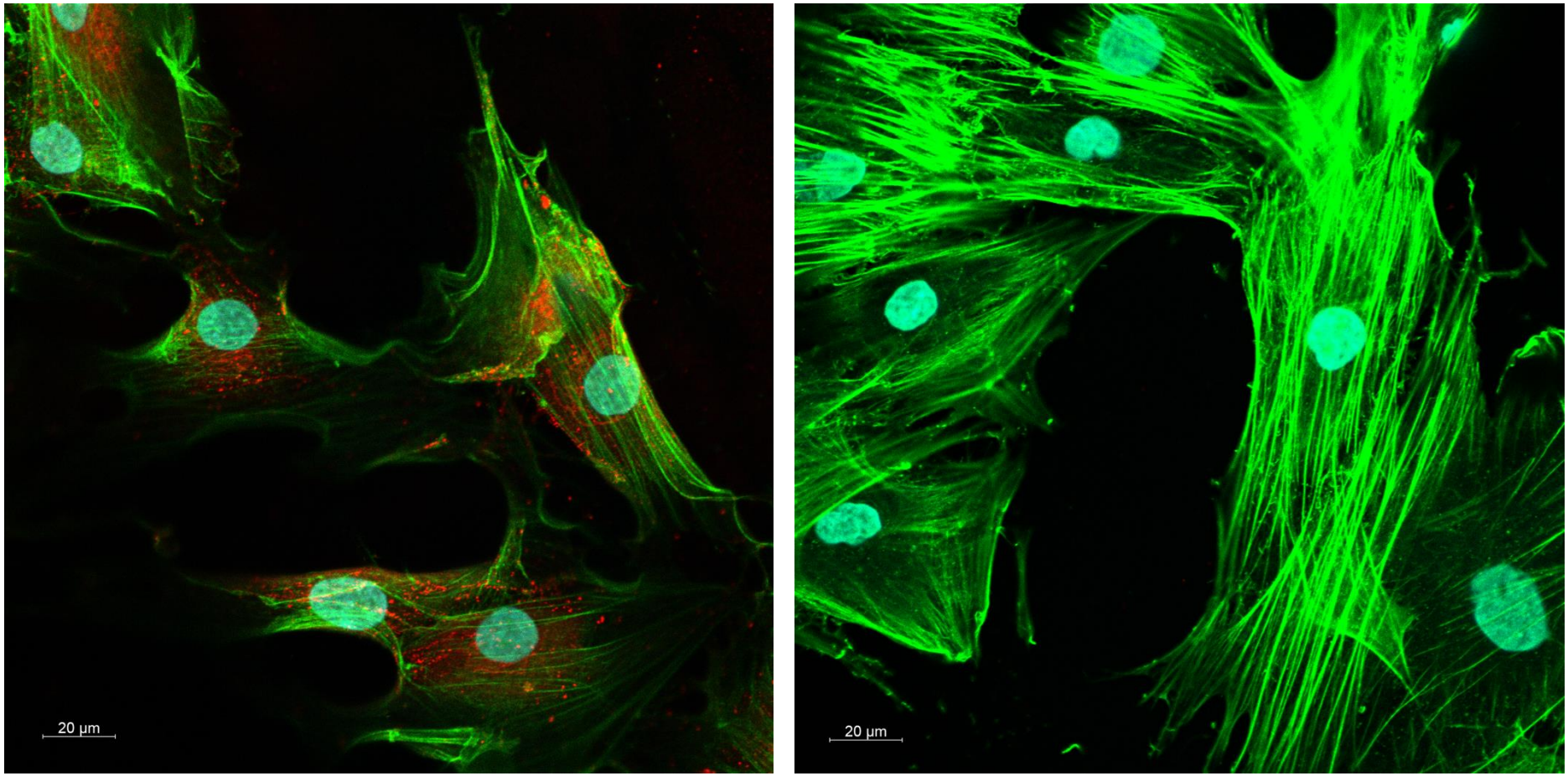



Figure 2.5.5.3 40x confocal images of cultured myofibroblast cells fluorescently labelled with SFRP4 primary antibody (left panel), and without primary antibody (negative control; right panel). SFRP4 is stained red, nuclei are stained blue (DAPI), and smooth muscle actin is stained green (phalloidin 488).

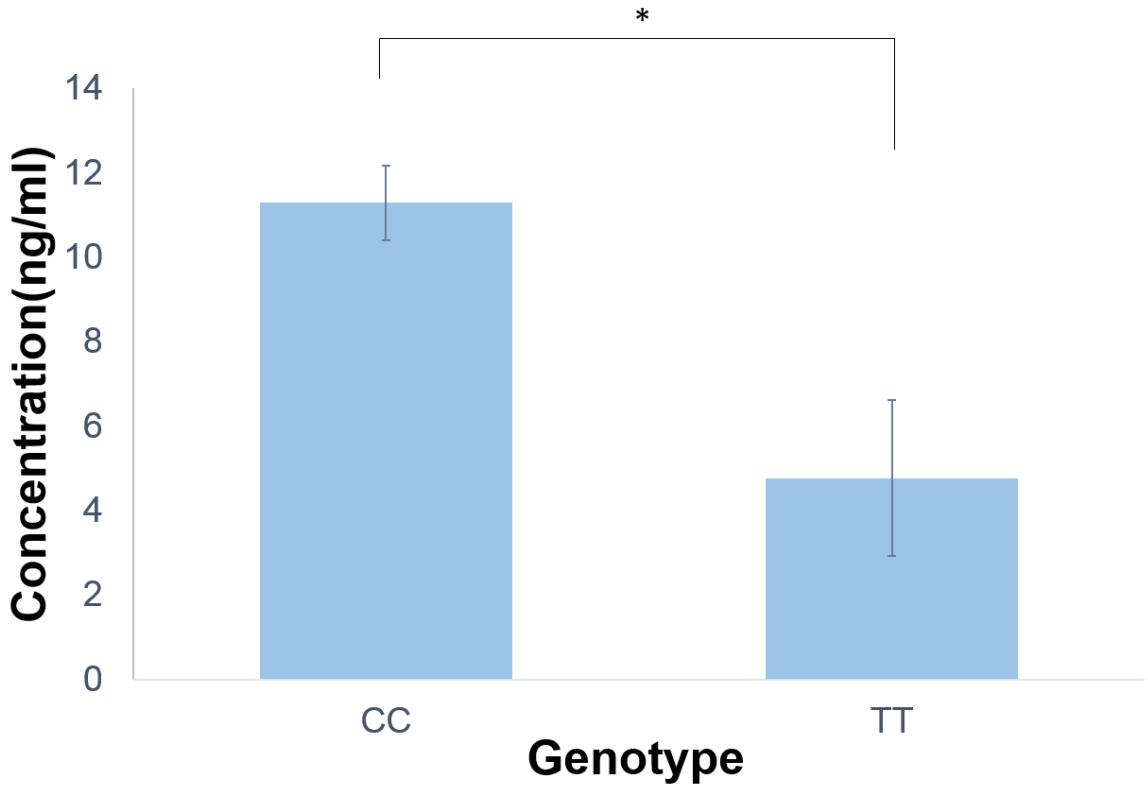
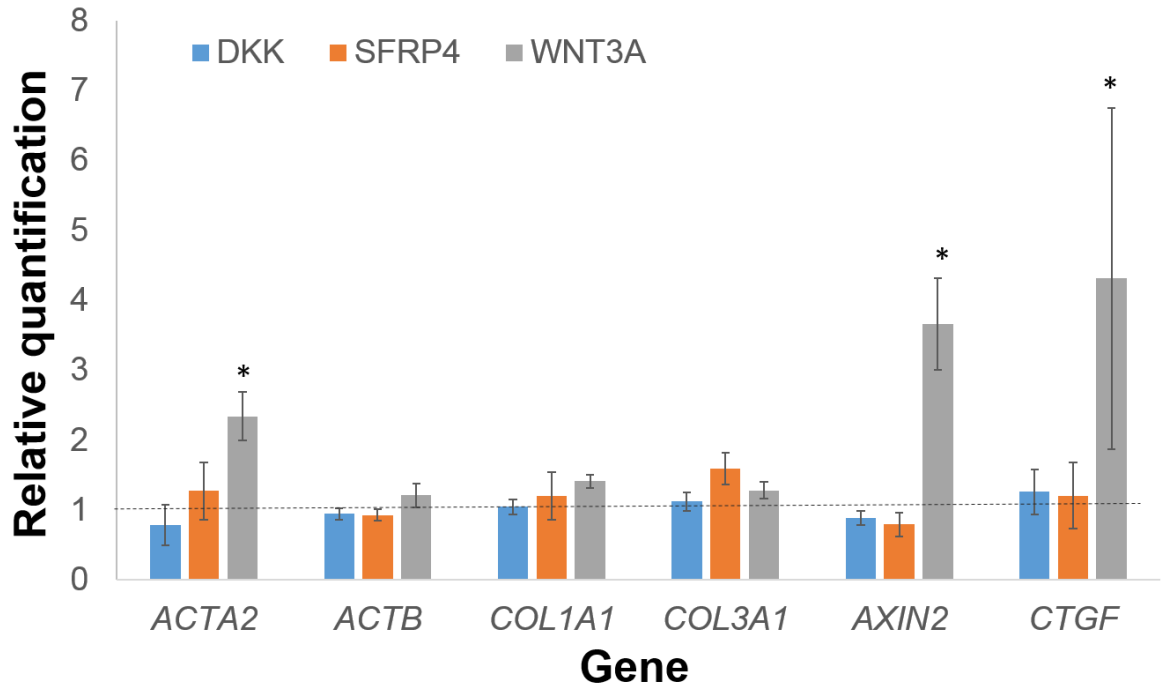



Figure 2.5.5.4 Bar chart showing the significant difference between the level of SFRP4 in the supernatant collected on day 7 that was used to culture myofibroblast cells carrying CC genotype compared to TT genotype (*$p < 0.05$).

a)




b)

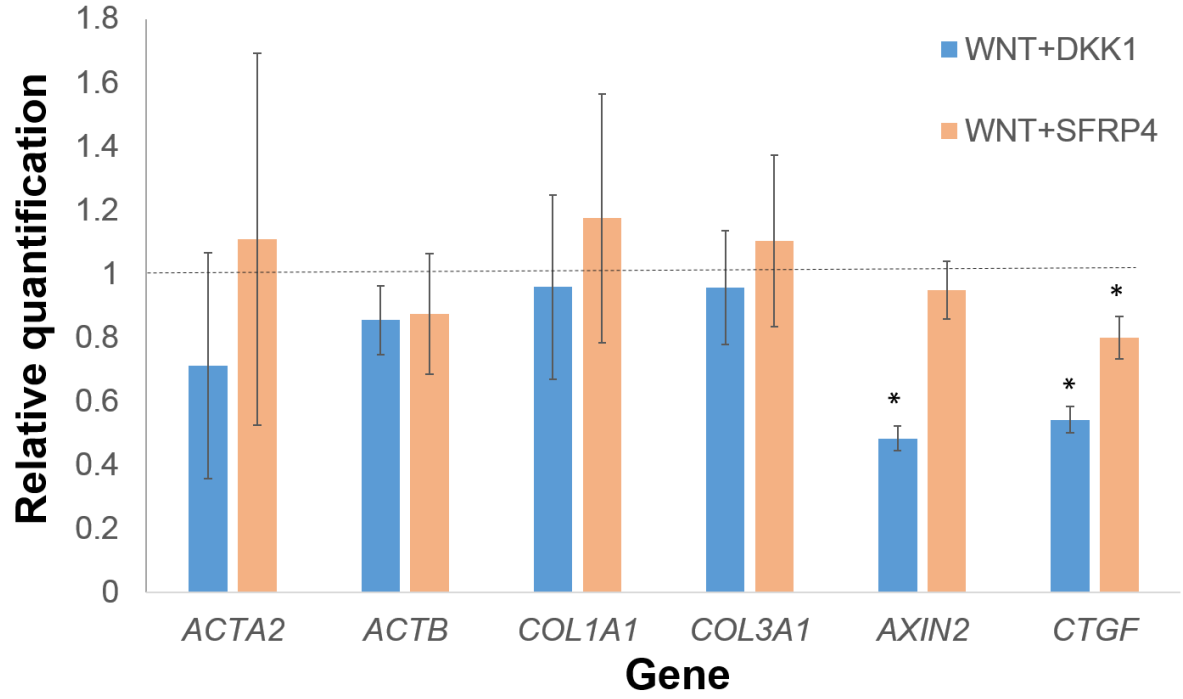



Figure 2.5.5.5 Bar charts showing a) a significant elevation in *AXIN2* and *CTGF* expression 6 hours after stimulation with 200ng/ml WNT3A and b) attenuation in *CTGF* in the presence of 8 μg/ml SFRP4. In both instances the horizontal line indicates no changes in expression.

### 2.5.6 *In silico* analysis of the remaining associated loci

Out of the 26 associated loci that were reported in the fine mapping analysis, 17 contain variants that are located within an open chromatin region or in an element that can alter a sequence motif, revealing a possible transcription regulatory function. Further analysis using 3D genome browser has demonstrated that all the tested variants can interact with three or more genes (Table 2.5.6.1), of which some are located beyond 500kb upstream or downstream of the lead SNP. For example, the region containing rs2306023, which is in the 99% credible set of rs2306022 on chromosome 15, is situated in an intron of *ITGA11* but interacts with the promoters *of CLN6*, *MIR4382* and *MIR548H4*, which was approximately half an Mb away.

Some variants not only demonstrated co-localisation with an open chromatin region but may also potentially affect a binding motif sequence. For example, on chromosome 1, rs6684375 (from the rs7524102 99% credible set) was predicted to change the motif sequence of AP2-alpha, and is located in an intergenic enhancer element interacting with multiple genes including the *WNT4*, *MIR4253* and more. Further eQTL analysis using GTEx has confirmed that the alleles of this variant affect the expression of *WNT4* in thyroid tissue, further validating the likely functional effect of this variant. Another potentially interesting variant on chromosome 1 is rs61811977, located in the 99% credible set of rs17433710. Although there is no predicted TF binding sequence motif change produced by this variant, the proximal region is an enhancer, which interacts

with not only *DDR2* but also *HSD17B7* further downstream, both of which have been associated with a functional role in the development of fibrotic disease.

Other than the intronic and intergenic variants, an exonic variant was also reported in my GWAS: rs104270. This variant is on chromosome 14, situated in the fifth exon of *MMP14* – a membrane bound matrix metalloproteinase with important roles in matrix degradation and homeostasis in a wide variety of diseases from Rheumatoid arthritis, to cancer invasion [127 128]. Interestingly, this SNP was the only SNP present in the 99% credible set analysis, and is therefore statistically highly likely to be causative. An Adenosine to Guanosine change at this position causes a non-synonymous change in the MMP14 amino acid sequence at position 273 from Glutamine to Glutamic acid. This alters the Hemopexin domain of MMP14, which cell surface dimerisation could therefore be affected. Further investigation into the functional consequences of this variant may lead to important therapeutic advances in DD, and also in other diseases where MMP14 is a key molecule.

Table 2.5.6.1 Summary of the *in silico* assay on every SNP in the 99% credible sets. A Y (Yes) and N (No) in the Enhancer/promoter, TFBS and Motif column indicate one or more variants of that particular credible set are located in the enhancer/promoter, transcription factor binding site, or cause a change in the binding site motif respectively. Gene interactions were curated by visual inspection using the virtual 4C track in the 3D genome browser. For rs6496519, gene interaction was reported in 3DSNP, but wasn't observed in the virtual 4C track.

| Chromosome | Position | rsID | Enhancer/promoter | TFBS | Motif | Mulitple genes interactions |
|---|---|---|---|---|---|---|
| 1 | 22698447 | rs7524102 | Y | N | N | *WNT4, CDC42, CELA3B, ZBTB40…* |
| 1 | 162672011 | rs17433710 | Y | Y | N | *DDR2, HSD17B7* |
| 5 | 108672946 | rs246105 | Y | N | N | *PJA2, FER, MAN2A1* |
| 6 | 149797014 | rs394563 | Y | Y | N | *TAB2, PPIl4, GINM1, SUMO4, ZC3H12…* |
| 7 | 3318658 | rs10276303 | N | N | Y | *CARD11, GNA12, SDK1* |
| 7 | 37973014 | rs2598107 | Y | N | N | - |
| 7 | 37989095 | rs16879765 | N | N | N | - |
| 7 | 116892846 | rs38904 | Y | Y | N | *WNT2, ST7,CFTR,CTTNBP2…* |
| 8 | 25845675 | rs10866846 | Y | N | N | *EBF2, PPP2R2A, BNIP3L* |
| 8 | 69992380 | rs2912522 | Y | Y | N | *LOC100505718, SULF1, SLCO5A1, C8orf34* |
| 8 | 70007938 | rs629535 | N | N | N | - |
| 8 | 109228008 | rs611744 | N | N | N | - |
| 8 | 145504343 | rs7838717 | Y | Y | Y | *BOP1, FAM2031, MAF1, HSF1…* |
| 9 | 1201156 | rs12342106 | N | N | N | - |
| 13 | 44842503 | rs9525927 | Y | Y | Y | *SERP2, LACC1, TSC22D1, LINC00284…* |

| | | | | | | |
|---|---|---|---|---|---|---|
| 14 | 23312594 | rs1042704 | N | N | N | *(MMP14 exonic)* |
| 14 | 51074461 | rs1032466 | Y | N | N | *MAP4K5, ATL1, CDKL1, ATP5S…* |
| 15 | 56229760 | rs1509406 | Y | Y | N | *NEDD4, RFX7, PYGO1, PRTG* |
| 15 | 68628163 | rs2306022 | Y | N | Y | *ITGA11, CALML4, ANP32A* |
| 15 | 89238184 | rs6496519 | Y | Y | N | ** (ACAN from 3D SNP)* |
| 16 | 75506593 | rs977987 | Y | Y | N | *CHST6, CFDP1, BCAR1, LDHD…* |
| 18 | 9762933 | rs9951109 | Y | Y | Y | - |
| 19 | 57678194 | rs11672517 | N | N | N | - |
| 20 | 38300807 | rs6016142 | N | N | N | - |
| 20 | 39320751 | rs6102095 | Y | Y | Y | *MAFB, TOP1, PLCG1, ZHX3* |
| 22 | 46459132 | rs7291412 | Y | Y | Y | *WNT7B, ATXN10, PPPARA, GTSE1* |

## 2.6 Discussion

My GWAS, which was the largest Dupuytren disease genetic study to date, reported a total of 26 associated loci. In addition to the nine loci that were reported by Dolmans et al in 2011, a further 17 were discovered. The discovery of the novel variants was a result of the increase in mapping resolution and sample size. There was little to no detectable difference between DD patients in Germany, the Netherlands and the UK, suggesting the homogeneity of DD genetics, at least with the common causative alleles amongst Western European Caucasian populations. The result of the heritability estimation analysis has provided further evidence that host genetic predisposition is required for the development of DD. However, in spite of the moderately large sample size, more than 40% of the population variance remains missing, and it would require GWA studies of a larger scale to overcome this problem. If performing additional GWAS is possible, utilising multi-ethnic samples [5] should be considered in the study design as it would increase the statistical power of the current findings, aid fine mapping and functional work prioritisation. A further approach could include meta-analysis of GWAS summary statistics in other fibrotic conditions, as discussed in later chapters.

From the results of my *in silico* analysis, functional work and recent reports in the literature, multiple components of the WNT signalling pathway have been implicated in the pathophysiology of DD. Three pro-fibrotic molecules in particular; WNT4, a

ligand of frizzled receptors; SFRP4, a WNT pathway antagonist, and RSPO2, a WNT pathway agonist, were suggested in my GWAS results. WNT4 has been shown highly expressed in Dupuytren disease nodule tissue compared to the normal tissue of the same individual [129]. Like other WNT ligands, it's debatable whether WNT4 activates the canonical or non-canonical pathway. However, it would not be surprising that this ligand activates both pathways as there is increasingly more evidence supporting the theory that both pathways are involved in fibrosis in tandem rather than independently. With regards to biological function, WNT4 has been thoroughly investigated in renal fibrosis. One study [130] has provided some evidence that WNT4 is involved in the process of myofibroblast differentiation via the β-catenin canonical pathway, though it is not clear whether it might be working in concert with other WNTs. WNT4 has also been shown to be up-regulated by TGF-β [131] through the canonical pathway via SMAD2/3.

SFRP4 is an antagonist of the WNT signalling pathway. Unlike members of the DKK protein family, which function to completely abrogate the FZL activated WNT signalling pathways, SFRP4 selectively inhibits WNT3A and WNT7A driven activation. WNT3A is expressed in DD tissue and has been demonstrated to have the ability to activate both the canonical signalling pathway through the binding of co-receptor LRP5/6, as well the non-canonical pathway with co-receptor ROR2. In our GWA analysis, the T allele at rs16879765 is associated with an increased risk of DD development. Further protein functional analysis has shown that DD nodule-derived myofibroblasts carrying the TT alleles showed significantly lower SFRP4 expression in comparison to those

carrying CC. Using the same type of myofibroblasts, the presence of SFRP4 was shown to attenuate WNT3A mediated *CTGF* expression, which suggests an anti-fibrotic effect nature of this molecule. This together with the results from the protein work provided evidence why individuals carrying the allele T at rs16879765 are more susceptible to develop DD than individuals carrying C. Other than *CTGF,* expression of many pro-fibrotic genes are governed by the nuclear translocation of YAP/TAZ, and the binding of transcriptional factor TEAD in the Hippo pathway [102 132]. Although YAP/TAZ pathway activation can be induced by the activation of TGF-β receptor via SMAD2/3 phosphorylation, it can be attenuated through the inhibition of LRP5/6 by a WNT pathway antagonist, DKK1 [133]. This further suggests that the potential of recombinant SFRP4 being used for targeted therapies in the treatment of DD, as it may be effective up to an approximately fifth of Caucasian Dupuytren patients. With regards to specificity, even though SFRP4 has the binding affinity only for a small subset of WNTs, extensive *in vitro* and *in vivo* work should be performed before applying it to patients. This is largely due to the importance of WNT pathway in the regulation of homeostasis in adult tissue. Severe dysregulation in WNT pathway can lead to detrimental developmental defects [134]. Further functional work focusing on converging signalling pathways downstream or upstream of SFRP4 could potentially lead to the discovery of a treatment that is effective on a wider population of DD patients.

Other than SFRP4, there is some *in-silico* evidence also suggesting the alleles at rs16879765 can affect the expression of *EPDR1*. The predicted causal SNP in LD with

rs16879765 is in a 5' – 3' loop that interacts with the possible bidirectional promoter [135] of both *EPDR1* and *SFRP4*. However, significant differential expression of *EPDR1* was not detected between the two homozygous genotypes of this variant in my analysis. This was likely due to the lack of samples, and statistical power to test the small effect size between the two. This issue was exacerbated by three factors; the basal expression of *EPDR1*, which was more than 10 fold lower than *SFRP4*, the biological variation between the tested cells that were derived from different DD patients, and the nature of a bidirectional promoter, where often the expression of the involved genes are not evenly expressed with one being several fold higher than another [135], This could be due to the competition in the recruitment of transcriptional machinery and chromatin remodelling molecules. Despite being statistically insignificant, I observed a similar differential expression pattern to that seen for *SFRP4,* which suggests that the T allele at rs16879765 could promote not only the expression of *SFRP4* but also *EPDR1*, thus confirming the prediction of the *in-silico* analysis. *EPDR1* is evolutionarily highly conserved. The function of EPDR1 is not clear, and most research has been performed in relation to developmental defects in zebrafish, though a recent publication using DD tissue and gene silencing technique has demonstrated its involvement in the contractility of myofibroblasts [32]. It is too early to speculate whether EPDR1 could be used as a target, or a biomarker, in translational medicine of DD without further functional evidence from extensive experimental work.

Poor correlation between mRNA and protein expression level [136] is a recurring problem in molecular research. At times, the concordance rate can be as low as 40%

[137] even when cells are at their steady state. This issue was seen in my functional work on SFRP4, where an inverse correlation was observed. The amount of RNA and protein produced and secreted is governed by post transcriptional, post translational and degradation regulation. In general, slow transcription rate of mRNA, fast translation rate of protein, and variable stability of RNA and protein could all contribute to this problem [138] but are dismissed in the current context as SFRP4 appeared to be a stable secreted protein in our ELISA results, and has been shown to be capable extracellular accumulation over long periods of time. One possible explanation for this inverse correlation is that SFRP4 could possess an unknown binding affinity to other ligands that can form a negative feedback loop. This is not entirely impossible as other member of the same family, SFRP1 has been shown to have other substrates including RANKL [139] and molecules with a DGR binding motif [140]. The other plausible explanation is the result of an elevated activity in the Endoplasmic-reticulum-associated protein degradation (ERAD) system due to the excess intracellular SFRP4 production in CC carrier cells [141].

ECM remodelling and sensing, cell migration and contractions are well known biological processes that play an integral part in fibrosis development and that is no exception in DD. Expression or protein conformation changes in major components of these processes have been reported to associate with DD in my GWAS. For example, Discoidin domain receptors (DDR) are a family of collagen tyrosine kinase receptors that have a binding affinity mostly for fibrillar collagen. This is interesting as collagens are a vast protein family that consists of more than 16 members and only type I, II, III

and V are classified fibrillar. In DD, the majority of collagens in cord and nodule are type I and type III, which are substrates of the second member of the DDR family, DDR2 alongside type II and V, and non-fibrillar collagen type X [142]. Dysregulation of the DDR2 signalling pathway has been reported to play a role in hepatic fibrosis after acute liver injury [143] in a mouse model. Furthermore, DDR2 mediated pro-fibrotic activity has been demonstrated to work in concert with various molecules of the inflammatory signalling pathway in hepatic stellate cells and lung fibroblasts [144]. DDR2 is an excellent therapeutic target as high specificity and efficacy small molecular inhibitors to this molecule are readily available for use in clinical trials [145]. The genetics of DDR2 and its molecular function in DD development will be further discussed in the next chapter.

Matrix metalloproteinases are a large family of proteases that are mainly characterised by their proteolytic function specifically for ECM and cell surface molecules. MMPs can be secreted or membrane bound, where each member of the family has an affinity for different substrates and thus different functions. The DD associated MMP14 in particular, also known as MT1-MMP, is a transmembrane protein with more than 40 known different substrates[146], which includes the aforementioned CTGF, fibrillar collagens and DKK1. The proteolytic function of MMP14 has been demonstrated to increase or promote the activity of its substrate, for example the activation of MMP2 from pro-MMP2, the cellular migration and invasion activity of CD44 [146 147]. The breakdown of DKK1 by MMP14 has also been shown to have conserved its LRP5/6 binding domain and possibly its function as a WNT signalling pathway inhibitor [148]. In

clinical trials, broad-range MMP inhibitors have been used to treat cancer patients, where some have progressed into developing DD, which suggests the importance of MMPs and their anti-fibrotic effect [149]. MMP14 is highly expressed in DD. Studies have shown that knockdown of MMP14 doesn't affect the rate of collagen degradation, which suggests its non-proteolytic function may be involved in the regulation of the pro-fibrotic aspect of DD [149]. MMP14 is an interesting molecule that has functions which perfectly fit the aetiology of DD, and the pathway may represent a useful therapeutic target.

As mentioned in the introduction, genetic predisposition and inflammation are the main elements of DD development. Although there wasn't a clear causal variant to gene interaction presented in my *in-silico* analysis, *GCKR* should be followed up for further functional work due to the interesting properties of the protein counterpart. GCKR is a regulatory molecule that has been demonstrated to be activated by TNF-$\alpha$[150], a potent cytokine. It has also been demonstrated to be involved in the regulation of WNT signalling pathway in B lymphocytes [151] and its expression correlates with the expression of E-cadherin, which is associated with EMT [152]. Further evidence has been observed in Non-alcoholic fatty liver disease (NAFLD) patients, where high expression in GCKR facilitates the development of fibrosis as well as correlates with severity [153]. As inflammation is not a tissue specific event, it is not surprising that not many inflammatory genes are genetically associated with DD.

Transcription regulation is a complex machinery, formed by a network of DNA-DNA interactions between promoters, enhancers, and insulators within one gene or multiple genes. Multiple types of interaction loops have been observed in human cells, which includes looping between promoter-enhancer, promoter-promoter, enhancer-enhancer, insulator-insulator [154]. Most intronic or intergenic variants in my GWAS credible sets are not situated within regions of these loops, and therefore only a handful could be functional. This helped narrow down the number of targets for follow-up functional work. However, certain regions do pose a technical challenge such as complex regions on chromosome six, or high gene density areas, as interaction loops identified at such locations often span across multiple binding elements of multiple genes. Nevertheless, as chromatin region remodelling and gene expression are often cell type and tissue specific, it is imperative to acquire DD specific eQTLs and chromatin enrichment profiles before embarking on lengthy functional assays. This could be achieved by investigating the RNA and protein expression of an individual cell within DD tissue using single cell sequencing and multiple marker cell sorting.

Despite the high genetic heritability in DD, the genetic predisposition to DD had been under-researched. My GWAS results have revealed exciting new targets and initial molecular work has provided further information on the mechanism of early DD development.

# Chapter 3 Discoidin Domain Receptors and their functions in DD

## 3.1 Abstract


Introduction

Discoidin domain receptor 2 (DDR2) is a tyrosine kinase receptor that binds specially to fibrillar collagens. In DD, the ratio between type I and III collagen is often altered but their biological functions remain unknown. *DDR2* has been implicated in DD development through my GWAS. These three molecules are the ideal candidates for further understanding the biology of DD.

Materials and methods

*DDR2* was knocked down using siRNA in myofibroblast cells that were freshly disaggregated from DD tissue. Cell contraction, migration, and proliferation were assessed in both *DDR2* knockdown (KD) and wildtype cells in the presence of either collagen I, or collagen III. Cell contraction was measured using traction force microscopy. Cell migration was assessed using a wound healing assay. Cell

proliferation was analysed using Alamar blue and Cyquant. Transcriptomic expression was measured using RNA-seq and differential analysis was performed using DEseq2.

## Results

DDR2 KD cells demonstrated a change in migration rate in the presence of collagen I, but not III. A morphology change was observed in some KD cells but no difference in overall contractile force was detected. No proliferation changes were observed in the presence of either collagen. RNA-seq analyses revealed that collagen I promoted angiogenesis and cell proliferation via DDR2, whereas collagen III promotes a pro-inflammatory and autoimmune response.

## Discussion

The role of collagen I, III and DDR2 in DD have been investigated in this chapter. Collagen I signalling via DDR2 promotes cell proliferation, and migration in DD as it would in normal tissue homeostasis. Collagen III, by contrast, appears to promote inflammation, and therefore may be crucial to perpetuating fibrosis. Follow-up immune cell co-culture work should be performed to further support this finding.

## 3.2 Introduction

Extracellular matrix deposition is an integral part of fibrosis in general, and specifically in Dupuytren disease. The composition of cells and structure of ECM are vastly different between fibrotic and normal tissue due to a shift in purpose, from maintaining homeostasis to repair [41]. Fibrosis is often described as a failure of resolution of inflammation, which could be due to various reasons including: abnormal excess production of ECM molecules; defects in enzymes that are responsible for the remodelling ECM such as MMPs and TIMPs; persistent inflammatory response; reduction in hormone protein levels; or elevated positive feedback from ECM due to an increase in ECM matrix sensing receptors. During the early fibrotic event, recruited immune cells increase the production of pro-fibrotic cytokines, such as TGF-β, and facilitate the transdifferentiation of epithelial, endothelial or other progenitor cells into myofibroblasts, which in turn increases the production of ECM molecules such as fibronectin, tenascin and collagen [64]. As a result, the stiffness of the tissue increases, driven by the highly contractile myofibroblast cells that promote the formation of rigid collagen bundles [64]. The stiffened fibrous matrix affects the cell apical polarity as well as cell to cell adhesion stability [64]. It also provides an environment for cells to migrate to the area.

Collagen, a triple helical molecule, is crucial for maintaining the integrity of tissue due to its biophysical properties [155]. Each chain of the helix is a multidomain protein, consisting of either one or two COL1 domains and two flanking propeptide domains. The composition of individual chains determines the type of the collagen, for example collagen I consists of two $\alpha 1$(I) chains and one $\alpha 2$(I) chain, where collagen II is a homotrimer of $\alpha 1$(II) chains. Collagens are organised into two major groups of secondary structure by telopeptides: fibrillar collagens are commonly found in skin and fascia, whereas non-fibrillar collagens are commonly found in basement membranes [155]. Although there are many types of collagen, over 80% in humans is type I, which belongs to the fibrillar collagen family along with types II, III, V, and others. In DD, collagen type I is found in abundance in the cord part of DD tissue, arranged in organized bundles of fibrils. In contrast, the nodule part of DD tissue harbours less collagen type I but more of its precursor protein proto-collagen type I [109]. Interestingly, both cord and nodule express a similar level of collagen type III. The ratio between type I and III collagen changes as the affected tissue changes from normal to fibrotic. This change has been observed in tissues during wound healing, other fibrotic conditions, and in *in vitro* studies in DD tissue [156]. The magnitude of this shift has been reported from 90% collagen I to 10% collagen III in normal tissue to 65:35% in DD [156] and is thought to be due to the reduction in collagen I production as opposed to an increase in the expression of collagen III [109]. Collagen is the foundation of fibrosis and the biological shift in expression of different collagen types indicates the course of DD development may be modulated by how myofibroblast cells interact with this molecule. One group of cell surface receptors that are known to interact with all types of fibrillar collagen are the discoidin domain receptors, or DDRs.

DDR is a family of receptor tyrosine kinases (RTKs), which consist of two members DDR1 and DDR2. Both receptors have the binding affinity to all fibrillar collagen, I, II, III and V via the GVMGFO motif while in its single triple helical form [157]. DDRs can also bind to non-fibrillar collagens but with unique specificity, where DDR1 has affinity for collagen IV, VI and VIII, and DDR2 for collagen X [158]. In addition, DDR2 can bind to collagen via other peptides [159]. The activation of DDRs is a lengthy process, but its signalling transduction activity can persist for a long duration of up to 18 hours once activated. Unlike most RTKs, DDRs are constitutively dimerised and are activated through ligand induced clustering rather than ligand induced conformation change [160]. This is in contrast to another common family of collagen binding surface receptors; integrins. Although both DDRs and integrins have affinity for collagen binding, they possess different binding sites and can bind collagen simultaneously. Studies have reported that the activation of DDR1 can enhance $\alpha 2\beta 1$ integrin activation and subsequently increase cell adhesion, rather due to an elevation in expression of $\alpha 2\beta 1$[161]. Both DDRs have an extracellular globular collagen binding discodin domain, an extracellular and intracellular juxtamembrane domain and an intracellular catalytic kinase domain. The binding of collagen to the recognition site triggers autophosphorylation of the kinase domain as well as the intracellular juxtamembrane domain. In osteoblast differentiation, the activation of DDR2 is required for activating the MAPK pathway and the transcriptional factor RUNX2 that is essential for this process [162]. The mRNA expression of *DDR*s is regulated chemically as well as mechanically, such as by the stiffness of ECM [163], hypoxia,[164] binding of collagen I, TGF-$\beta$ stimulation[165]. Upon DDR activation, expression of members of the metalloproteinase family, in particular, *MMP1*, *2*, *9*, *10*, *13* and *14* have all been

reported to be elevated [166-170]. This family of proteins can influence cell invasion, migration and proliferation, ECM degradation and, in vivo processes such as tissue degradation in osteoarthritis [171] and reduction in plaque stability in atherosclerosis[172].

Apart from cell adhesion, DDRs like all RTKs have a variety of cellular functions including cell proliferation, cell migration, cell survival and cell contraction, and play an important role in processes such as the immune response, skeletal development and wound healing [173 174]. Because of these myriad of functions, and the wide expression of the ligand collagen, it is not surprising that DDRs are expressed in many human tissues. However, the expression of subtypes of DDRs is cell type specific: DDR1 can be found only in epithelial cells, and DDR2 in mesenchymal cells[175]. In addition, DDR2 has also been identified in myofibroblasts though this could be due to the plasticity of this cell type [144]. Both members of the DDR family have been reported to possess the ability to promote EMT following ligand binding [144], which is a key process in wound healing. Abnormal expression of DDRs has been observed in various conditions such as atherosclerosis, cancer, and fibrosis. The role of DDRs in different conditions greatly varies. In breast cancer, the roles of DDRs include cell migration, metastasis, cell invasion, and EMT through the phosphorylation of SNAIL1 [158 176]. In atherosclerosis, excess DDR1 can lead to the destabilisation of collagen I, which in turn increases instability of the atherosclerotic plaque, through elevation in the inflammatory response by recruiting macrophages to the plaque [177 178]. In idiopathic pulmonary fibrosis, DDR2 has been shown to increase angiogenesis, cell migration, and the recruitment of fibroblasts to the fibrotic area. The same study also reported

that collagen type 1 induced myofibroblast differentiation is TGF-β independent, thus iterating its potential importance in therapeutic use of anti-inflammatory drugs[144].

Other than IPF, DDR2 is also likely to have a role in the development of DD, as suggested by the results of our recent multi-European cohort GWAS (Chapter 2). The *DDR2* gene is located on chromosome 1 at q23.3 and has 19 exons, of which exons 4-11 encode the extracellular domain, exon 12 encodes the transmembrane domain, exons 13 and 14 encode the intracellular juxtamembrane domain, and exons 15-19 encode the kinase domain. Although, the span of this gene is over 100kb, the coding exons 4-19 are all congregated in the last 30kb, which together with existing DNase I data [179] suggest that the majority of the transcription regulatory elements of this gene are either upstream of this region or in the 3'UTR. The most statistically significant DD associated variant is rs17433710, which is at position 162672011 (hg19), intronic between exons 2 and 3. The risk allele C is present in 15% of European populations [180] and 12% in our Dupuytren discovery phase GWAS data. Further fine mapping and in silico analysis using the genetic data of this GWAS and HRC imputation panel has revealed a small 99% credible set of LD variants congregating in an open chromatin region between position 162666000 and 162667000 (hg19), suggesting the regulatory function of this association. *DDR2* polymorphisms have also been reported to be associated with a handful of other diseases and conditions including osteoporosis [181], age at smoking initiation [182] and lung cancer, though more supporting evidence is required here [183]. Somatic mutations in *DDR2* have also associated with lung cancer progression [184].

DDR2 is an exciting protein to study in DD due to its function as a collagen receptor, as well as the existing knowledge of its role in tissue homeostasis and other diseases such as IPF.

## 3.3 Aims and hypotheses

DDR2 is an excellent candidate molecule for functional work primarily due to the location of the associated variant, and DDR2's known biological function as a transmembrane receptor for fibrillar collagen. My aim was to identify the phenotypic changes, namely proliferation, contraction and migration in DD myofibroblast cells according to DDR2 status (wild type or knockdown) and the presence of different type I or III collagen. My aim was also to look at the effect of *DDR2* knockdown on RNA expression by screening the transcriptomic profile at whole genomic level.

The functions of DDR2 in fibrosis are partially documented in the literature, and as a result it is clear that it plays a pathophysiological role. My hypotheses were two fold. Firstly, DDR2 status would affect only one or two of the tested phenotypes in myofibroblasts but not all, as these phenotypes don't physiologically happen in parallel. My second hypothesis was that collagen I and III each plays a different regulatory role in fibrosis development, but both are mediated through the binding of DDR2.

## 3.4 Materials and methods

### 3.4.1 Protein knockdown

*DDR2* knockdown was performed using Accell siRNA (Dharmacon, Lafayette, US) and transfection reagent Interferin (Polyplus, France). 5nM of lyophilised Accell siRNA was re-suspended in 50µl of 1x siRNA buffer (Dharmacon, Lafayette, US) to create a 100µM stock according to manufacturer's instructions. Re-suspended siRNA was stored at -20$^{o}$C until further use. Prior to transfection, the 100µM solution was diluted to 1µM using 1x siRNA buffer (Dharmacon, Lafayette, US). The final working concentration of the siRNA in each transfection reaction was 10nM. The transfection mix was created by combining an appropriate amount of 1µM siRNA and Interferin in Opti-MEM (ThermoFisher Scientific, Paisley, UK). The amount of Interferin required for each reaction was determined by the area of the tissue culture apparatus according to the manufacturer's protocol. The transfection mix was vortexed for 10 seconds and incubated at room temperature for 10 minutes. Meanwhile, culture media of previously seeded cells was replaced with phenol red free 10% FBS DMEM (ThermoFisher Scientific, Paisley, UK). The transfection mix was then added on the seeded cells. The knockdown efficiency was checked at RNA and protein level by qPCR and Western Blotting respectively. The RNA was collected 48 hours post-transfection and protein at 72 hours.

### 3.4.2 RNA extraction

RNA extraction was performed using two commercially available extraction kits, Direct-zol miniprep and 96-well kit (Zymo research, CA, US), according to the manufacturer's instructions. Below is described the instructions of the miniprep kit as both protocols were almost identical. TRIzol (ThermoFisher Scientific, Paisley, UK) was added onto the cell sample and mixed gently by pipetting up and down. The cell homogenate could be stored at -80$^{o}$C if extraction was performed on a later date. The protocol was performed at room temperature according to manufacturer's instructions. Briefly, cell homogenate was centrifuged at 12000g for 1 minute, and supernatant was carefully transferred to a DNAse/RNase-free Eppondorf tube. One volume of ethanol was added to the solution, and mixing by pipetting up and down several times. The mixed solution was transferred to a Zymo-spin IIC column in a collection tube, and centrifuged at 12000g for 1 minute. The column was washed with 400µl RNA Wash buffer, centrifuged, and DNase treated at room temperature for 15 minutes. The column was washed three more times, once with 400µl of RNA PreWash buffer and two times with 700µl of RNA wash buffer. 20µl of DNase/RNase free water was added to elute RNA off the column. RNA samples were stored at -80$^{o}$C until further use.

### 3.4.3 Reverse transcription and qPCR

Reverse transcription was performed using the High-Capacity cDNA reverse transcription kit (ThermoFisher Scientific, Paisley, UK). A typical reverse transcription reaction was performed in a 0.2ml tube, consisting 2µl 10X RT buffer, 0.8µl 25X dNTP mix, 1.0µl 10X RT random primers, 1.0µl MultiScribe reverse transcriptase, 1.0µl RNase inhibitor, 3.2µl Nuclease-free water, and 10µl RNA. This was prepared in bulk to minimise inaccuracy due to pipetting small volumes. The reverse transcription was performed in a thermal cycler using the following cycling conditions, 10' at 25$^{o}$C, 120' at 37$^{o}$C and 5' at 85$^{o}$C. Real time qPCR was performed using Taqman assays (ThermoFisher Scientific, Paisley, UK). A typical Taqman reaction contained 5µl Taqman Fast Advanced Master Mix, 0.5µl gene specific Taqman probes, 0.5µl water and 4µl RNA. The reaction mix was aliquoted into a MicroAmp Optical 384-well plate (ThermoFisher Scientific, Paisley, UK) and was assayed using ViiA 7 (ThermoFisher Scientific, Paisley, UK) and Taqman cycling conditions. *18S* and *GAPDH* were used as housekeeping genes. Relative quantitation ($\Delta$Ct) and fold expression change ($2^{-\Delta\Delta Ct}$) were calculated using the formulas below, where G is gene of interest, C is housekeeping gene, S is stimulated and US is unstimulated.

$$\Delta Ct = Ct_G - Ct_C$$

$$\Delta\Delta Ct = \Delta Ct_S - \Delta Ct_{US}$$

### 3.4.4 Western blot

Protein samples were collected using SDS-PAGE sample buffer supplemented with 100mM DTT (Cambridge Biosciences, Cambridge, UK) and denatured at 95$^{o}$C for 20 minutes. Samples were loaded onto wells of a 4-20% Criterion TGX gel (Bio-Rad, Watford, UK) and the gel migration was performed using electrophoresis at 100V for 90 minutes. Protein was transferred to a PVDF membrane on the Trans-Blot turbo transfer system (Bio-Rad, Watford, UK) using the preset mixed weight protocol. The membrane was blocked with 5% BSA in PBST for one hour at room temperature. Primary antibody incubation was performed in 5% BSA PBST overnight at 4$^{o}$C, followed by two washes of PBST for 5 minutes. Secondary antibody incubation was performed in 5% BSA PBST at room temperature for one hour. After two washes with PBST, the membrane was then dab dried using hand towels and covered in 2ml of Clarity Western ECL blotting substrates (Bio-Rad, Watford, UK) for two minutes. The chemiluminescent substrates were imaged using the UVITEC gel documentation system (Cleaver Scientific Ltd, Rugby, UK). The bands on the image were quantified using ImageJ.

### 3.4.5 Tissue collection, cell disaggregation and myofibroblast cell culture

Affected fascia from Dupuytren disease patients was surgically resected, and collected, under ethics approval by the Research Ethics Committee (REC

07/H0706/81), directly from the operating theatre and stored in DMEM (Lonza, Slough, UK) at 4$^{o}$C until disaggregation. Tissue was finely minced and incubated in sterile filtered 20mg/ml collagenase (Worthington, NJ, US) in 0.5% FBS (Labtech, East Sussex, UK) DMEM at 37$^{o}$C with agitation at 180rpm for three hours. Disaggregated cells were filtered from debris, which contained extracellular matrix, using 70µm Falcon cell strainers (ThermoFisher Scientific, Paisley, UK), pelleted by centrifugation at 300g and re-suspended in 10ml 10% FBS DMEM (Lonza, Slough, UK). The suspension was transferred to a 10cm$^{2}$ petri-dish and incubated under normal tissue culture atmosphere; 37$^{o}$C and 5% $CO_2$. When the petri-dish reached 90% confluence, myofibroblast cells were transferred to a 150cm$^{2}$ cell culture flask (Corning, NY, US). This was performed by detaching the myofibroblast cells using Tryple (ThermoFisher Scientific, Paisley, UK), pelleting with centrifugation at 300g and re-suspension in 10% FBS DMEM. Cells were passaged up to a maximum number of three times.

### 3.4.6 Proliferation assay

Myofibroblast cells were first detached using Tryple, counted and re-suspended in 10% DMEM (Lonza, Slough, UK). Counted cells were then seeded onto 96 well plates at 5000 cells per well in 100µl and incubated for 6 hours or until attached. Media was then replaced with 0.5% DMEM and incubated for 18 hours. Conditioned media containing either collagen I or III (Advanced Biomatrix, CA, US) was added onto each well according to experimental plan. Proliferation assay was performed using two methods, PrestoBlue (ThemoFisher Scientific, Paisley, UK) and CyQUANT

(ThermoFisher Scientific, Paisley, UK). Working PrestoBlue solution was prepared by diluting one part of the stock solution with nine parts of DMEM (Lonza, Slough, UK). Working CyQUANT solution was prepared by diluting one part of 20X cell-lysis buffer with 19 parts of sterile water and 1/400th of CyQUANT GR dye. At each time point, PrestoBlue solution was added directly onto the cells and incubated in tissue culture atmosphere for 45 minutes. The fluorophores of the solution were excited at 535nm and measured at 615nm using FLUOstar Omega microplate reader (BMG Labtech, Bucks, UK). Cells were then washed with 100µl of PBS, aspirated, and stored at -70$^{o}$C until frozen. 200µl of the working CyQUANT solution was added to the thawed samples, and was incubated at room temperature away from light for five minutes. This solution was then excited at 480nm and measured at 520nm using FLUOstar Omega microplate reader. Percentage proliferation was calculated by dividing the fluorescent unit of the stimulated cells with the unstimulated. ANOVA was used to compare the means of each group, where a p value less than 0.05 was considered significant.

### 3.4.7 Migration assay

Myofibroblast cells were detached from the tissue culture apparatus using Tryple (ThermoFisher Scientiific, Paisley, UK), pelleted by centrifugation at 300g, re-suspended in 10% FBS DMEM (Lonza, Slough, UK) and then counted using a haemocytometer. 70µl of 200 thousand cells per millilitre were plated onto both chambers of the Culture-Insert 2 well in µ-Dish 35mm (Ibidi, Martinsried, Germany).

Cells were left to settle at room temperature for 30 minutes before being transferred to normal cell culture atmosphere and incubated for 6 hours. After the cells attached, the 10% DMEM media was replaced with 0.5% DMEM and then incubated for 18 hours. 2ml of 0.5% FBS DMEM (Lonza, Slough, UK) supplemented with collagens was added to the dish and the insert was gently removed using a pair of sterile tweezers. The empty compartment between the two populations of cells on the dish was imaged using a brightfield microscope (Nikon, Kingston-upon-Thames, UK) with a 4X objective. Three to four monochrome images were captured to cover the entire length of the compartment. The culture dish was returned to normal culture atmosphere for 12 hours. The centre compartment was imaged again under the same conditions. The area of the centre compartment was determined using Cell Profiler, and the percentage of area closure was calculated by dividing the size of the migrated area with the size of the empty area. The difference in mean percentage of area closure between knockdown and wild type cells was examined using T test. ANOVA was used to test the statistical significance in means between wild type cells that were unstimulated and stimulated with collagens. A p value less than 0.05 was considered significant.

### 3.4.8 Traction force microscopy

Myofibroblast cells were detached using Tryple, pelleted, counted and re-suspended in 10% FBS DMEM (Lonza, Slough, UK). 100µl of 4000 cells were seeded onto a collagen coated, 1µm Fluorospheres filled, 2KPa SoftTrac softview 35 dish (Matrigen,

Cambridge, UK). Cells were left to settle onto the dish at room temperature for 30 minutes before being transferred to normal cell culture atmosphere for 12 hours. Prior to imaging, cells were labelled using 10 µM Calcein AM (Cambridge BioVision, CA, US) in 10% FBS DMEM (Lonza, Slough, UK) for 45 minutes. Imaging was performed using LSM710 confocal microscope (ZEISS, Cambridge, UK) with a 40X water immersion objective. 20 images were captured for each condition before and after the cells were detached from the dish, which was achieved by gently adding 100µl Tryple without disrupting its position on the stage. Each pair, before and after cell detachment, of images were employed to construct traction force field using a series of ImageJ plugins including Template Matching, PIV and FTTC [185] following the author's instructions. Total traction force of each cell of interest was then calculated by adding all traction forces in the visible field. Mann Whitney rank sum test was used to analyse the statistical significance in median total traction force between tested groups, where a $p < 0.05$ was considered significant.

### 3.4.9 Sequencing library preparation and high-throughput sequencing

The quality of RNA samples was assessed using Tapestation and High Sensitivity RNA ScreenTapes (Agilent Technologies, Stockport, UK). Briefly, 2µl RNA sample was combined with 2µl of High Sensitivity RNA ScreenTape sample buffer in optical tube strips (Agilent Technologies, Stockport, UK) and loaded into a ScreenTape via a Tapestation. Samples with an RNA integrity number (RIN) value greater than 8.0 were considered good quality. Between 100ng and 1µg RNA sample was prepared for

sequencing using KAPA stranded RNA-Seq library preparation kit with RiboErase (KAPA Biosystems, MA, US). Briefly, rRNA of each sample was depleted by oligo hybridisation, immediately followed by a clean-up process using 2.2X Sera-mag beads (GE Healthcare, Buckinghamshire, UK) in 20% PEG 8000 (Sigma-Aldrich, Dorset, UK), 2.5M NaCl, 10mM Tris-HCl, and 1mM EDTA. The hybridised oligonucleotides were then removed by DNase digestion at 37$^{o}$C, and purified using 2.2X Sera-mag beads. Purified RNA was eluted from the beads with RNase free water, fragmented at 94$^{o}$C for 6 minutes and cDNA synthesis was initiated immediately with first and second strand created in separate reactions. This was followed by a 1.8X Sera-mag beads clean-up and A-tailing. A-tailed samples were ligated with the adapters provided in the NEBNext Multiplex Oligos for lllumina kit (New England Biolabs, Hitchin, UK), and was purified twice with 1X Sera-mag beads. Library amplification was performed using the protocol described in the NEBNext kit for 13 cycles. This number of cycles was used specifically for input RNA up to 200ng. Amplified samples were cleaned using 1X Sera-mag beads. The sequencing-ready samples were assessed for DNA fragment size using D1000 High Sensitivity ScreenTape (Agilent Technologies, Stockport, UK) and their concentrations were determined using KAPA library quantification kit (KAPA Biosystems, MA. US) and ViiA7. Multiple samples were pooled to create a single 4nM sample, whose concentration was confirmed using the above method. The sample was denatured and diluted to 1.8pM according to the Denature and Dilute Library guide for NextSeq system (Illumina, Essex, UK). This was loaded on to a 75 cycle NextSeq 500/550 v2 kit (Illumina, Essex, UK) and sequenced using a high-output flow cell in a NextSeq 500 desktop sequencer (Illumina, Essex, UK). The sequencing run

was set up using Illumina's proprietary software BaseSpace (Illumina, Essex, UK) and the paired-end setting was used.

### 3.4.10 RNA-seq data analysis

Myofibroblast cells were detached from the tissue culture apparatus using the protocol as described in section 3.4.7. Resuspended cells were plated onto a 6-well plate and were incubated at normoxic condition for 6 hours. For the basal RNA-seq experiment, fresh media was added onto the plate and cells were further incubated for 6 hours before being harvested for RNA. As for the collagen stimulation experiments, collagen supplemented media was added and cells were incubated for 12 hours before being harvested. The RNA was then processed for sequencing using the method as described in section 3.4.2 and 3.4.9.

Raw DNA sequence Fastq files were acquired from Basespace (Illumina, Cambridge, UK) after having been deposited from the NextSeq sequencer. Fastq files of the same sample and from the same read were catenated into one fastq file using Bash. The amalgamated Fastq file was then checked for adapter contaminations and any other problems using FastQC. If required, adapters were remove using Cutadapt. Fragments in the fastq files were then aligned to hg19 assembly reference panel using Hisat2 [186] in a pairwise fashion. Aligned fragments were then sorted, indexed and converted to bam files using Samtools. Gene counts were performed using

FeatureCounts [187]. Differential analysis was performed in R using DESeq2 [188]. For visualisation, heatmaps were created using gplots. KEGG pathway enrichment analysis was performed using GAGE [189 190].

## 3.5 Results

### 3.5.1 The confirmation of DDR2 knockdown

Prior to being utilised in various phenotypic experiments, DDR2 knockdown and wild type (untransfected) DD myofibroblast cells were assessed for their DDR2 protein and RNA expression level using quantitative PCR (qPCR) and western blot. RNA and protein were harvested 48 and 72 hours post siRNA transfection, respectively. The knockdown efficiency at RNA level was between 70% and 90% (Figure 3.5.1.1), and between 35% and 60% at protein level (Figure 3.5.1.2), which was as expected.

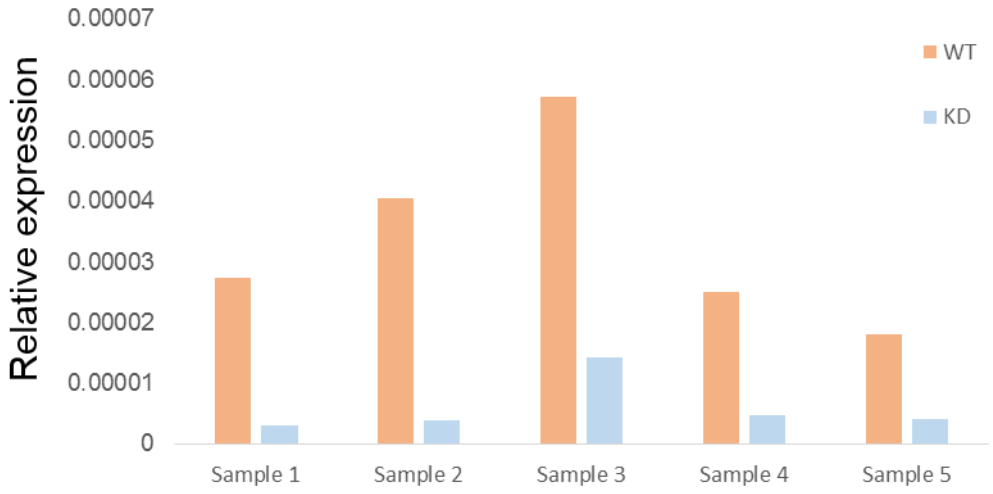



Figure 3.5.1.1 A bar chart demonstrating the relative expression of *DDR2* to housekeeping gene *18S* of *DDR2* knockdown (KD) and siRNA untransfected cells (WT). The overall knockdown efficiency at RNA level was ranging between 70% and 90%.

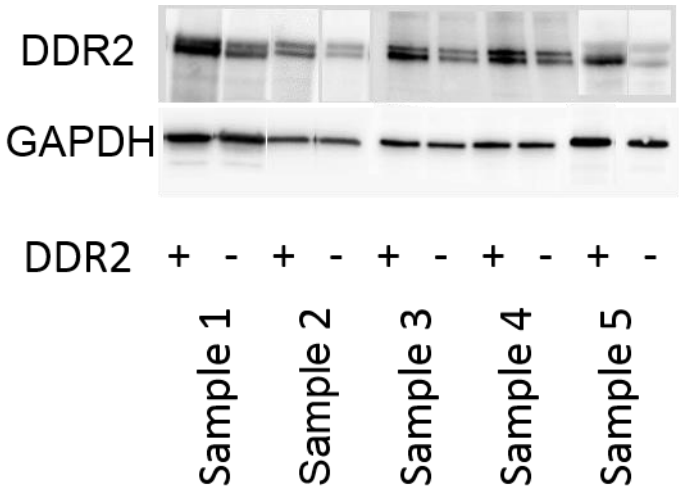



Figure 3.5.1.2 A picture showing the protein expression before and after DDR2 knockdown in each pair of samples, where a weak band indicates low expression. The protein knockdown efficiency was approximately between 35% and 60 %. A minus sign indicates DDR2 knockdown.

## 3.5.2 The function of DDR2 in DD myofibroblast proliferation

The potential cell proliferation role of DDR2 and its substrates collagen I and III in myofibroblast cells was examined using two different measurements, DNA concentration (Cyquant) and NADH dehydrogenase activity (Alamar blue). Knockdown cells were cultured up to 72 hours post stimulation, until the effect of the DDR2 siRNA began to dissipate. Both assays demonstrated similar pattern in the proliferative activity out of all the scenarios tested. No statistical differences were observed between 20μg/ml collagen I stimulated knockdown and wild type DD

myofibroblast cells (Figure 3.5.1.1), or between 20μg/ml collagen III stimulated knockdown and wild type myofibroblast cells (Figure 3.5.1.2).

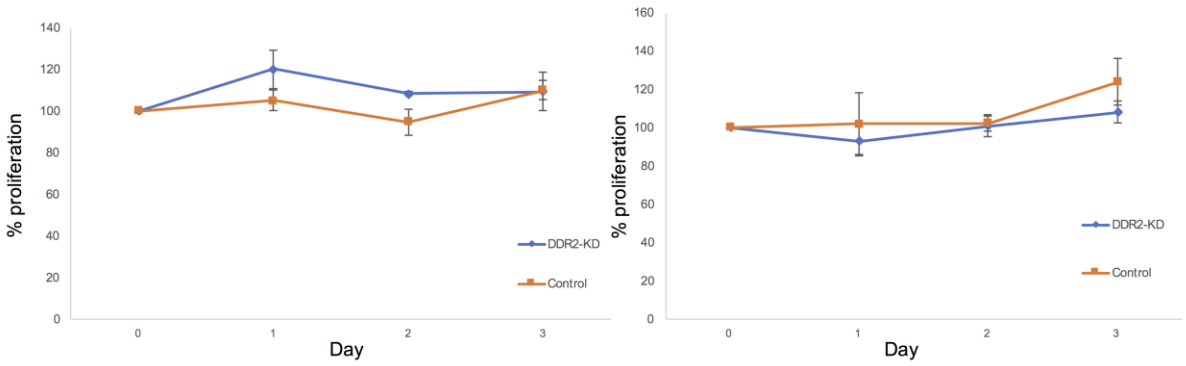



Figure 3.5.1.1 Line graphs demonstrating the proliferation activity of myofibroblast cells with and without DDR2 knockdown up to three days post collagen I stimulation using two different assay methods (Left: Alamar Blue; Right: Cyquant). Each group contains three biological replicates (n=3) with error bars representing the standard error of the mean (SEM).

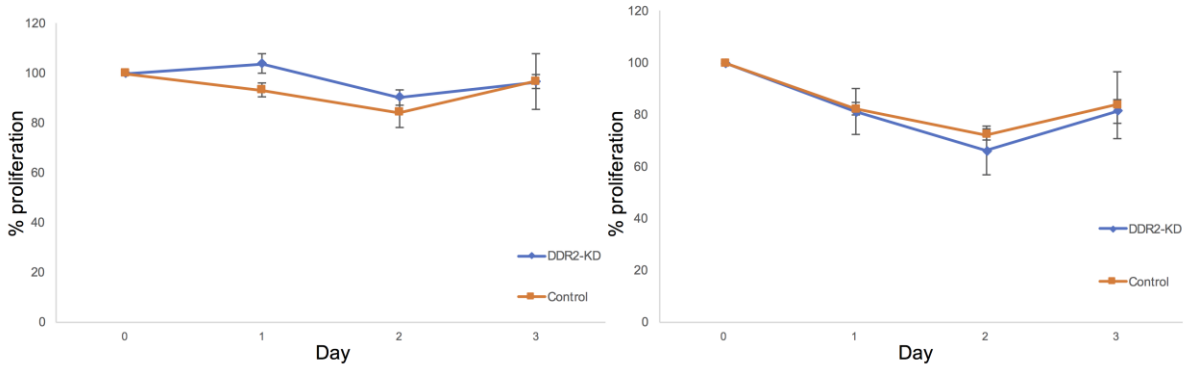



Figure 3.5.1.2 Line graphs demonstrating the proliferation activity of myofibroblast cells with and without DDR2 knockdown up to three days post collagen III stimulation

using two different assay methods (Left: Alamar Blue; Right: Cyquant). Each group contains three biological replicates (n=3) with error bars representing SEM.

### 3.5.3 The function of DDR2 in DD myofibroblast migration

Cell migration was quantified using percentage of area closure between the two populations of DD myofibroblast that were seeded into the 2-cell Ibidi chamber, where higher percentage equates to higher cell motility. Migration was measured 12 hours after the separation compartment was removed as the percent of area closure reaches 100% between time points 12 and 18 hours (Figure 3.5.2.1). In the 20μg/ml collagen I stimulation experiment (Figure 3.5.2.2), DDR2 knockdown myofibroblast cells (n=5) demonstrated a mean 39% area closure 12 hours post stimulation comparing to the mean 63% area closure demonstrated in the wild type cells, showing a 1.6 fold decrease (FDR p=0.022). In the 20μg/ml collagen III experiment (Figure 3.5.2.2), the DDR2 knockdown myofibroblast cells demonstrated a 47% area closure compared to the 60% area closure that was identified in the wild type cells, demonstrating a 13% reduction in migration. However, this comparison was not statistically significant (p=0.09). Although there was a visual difference in migration rate between the collagen I or collagen III stimulated wild type cells and the non-stimulated wild type cells, no statistical significant differences were detected. This was partly due to the limitation of the initial size of the “wound area” (Figure 3.5.2.3).

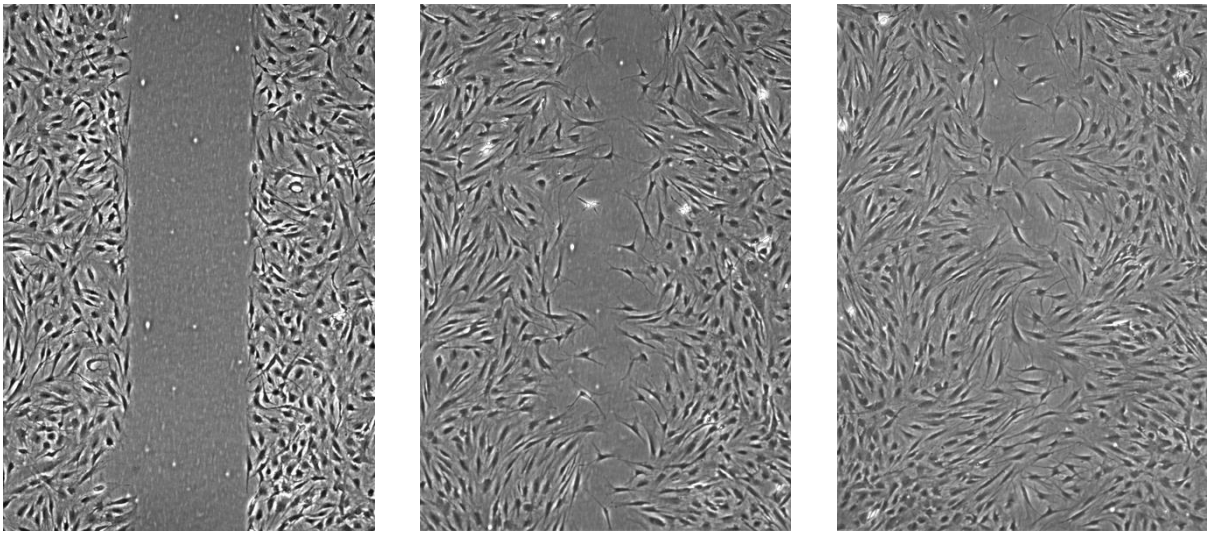


Figure 3.5.2.1 Pictures showing the migration of wild type myofibroblasts over the course of 18 hours (Left t=0h; Middle t=12h; Right t=18h). The percentage of area closure reaches typically reaches 100% between 12 and 18 hours for healthy DD myofibroblasts.

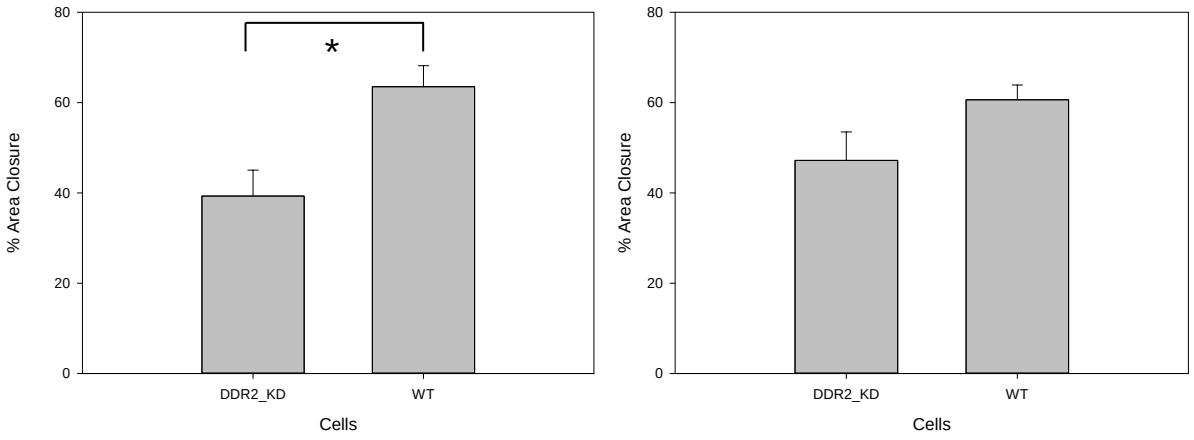



Figure 3.5.2.2 Bar chart showing the percentage of area closure of DDR2 KD and wild type cells 12 hours post left) collagen I stimulation (n=5) and right collagen III stimulation (n=5). WT cells migrated 1.6 fold further than the DDR2 knockdown cell while under collagen I stimulation (*p=0.022). Error bars represent SEM.

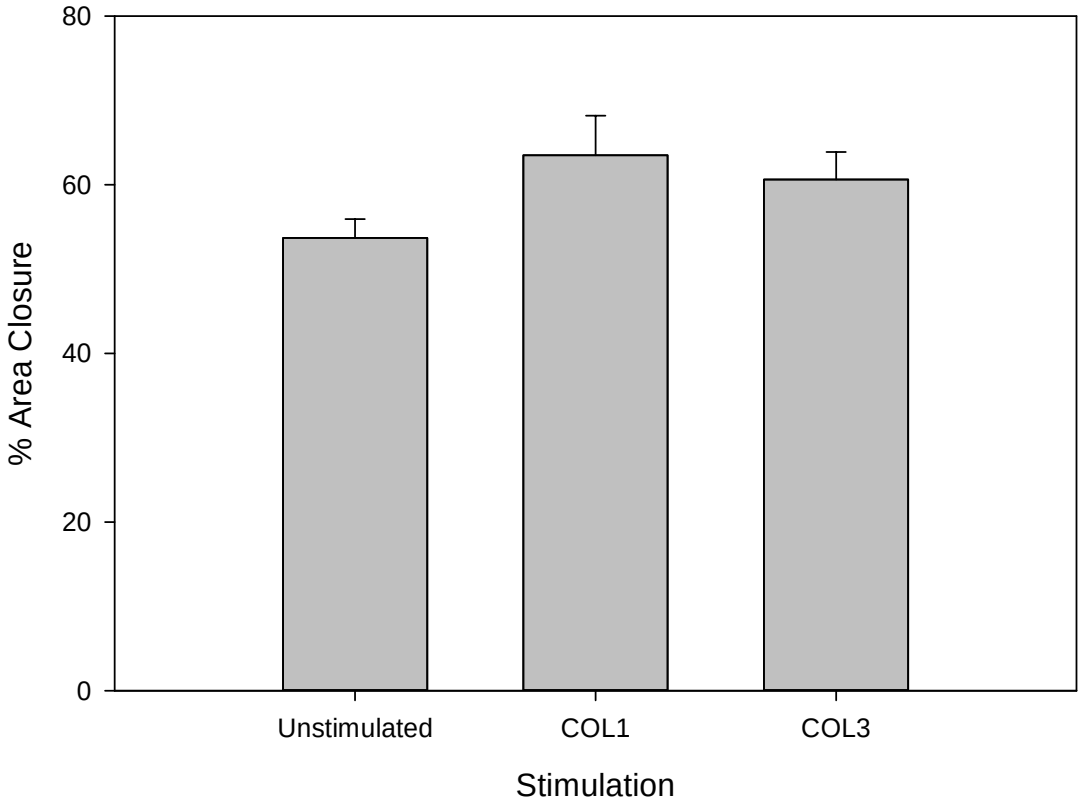



Figure 3.5.2.3 Bar chart showing the percentage of area closure (t=12) of wild type cells in three different conditions; unstimulated, collagen I stimulated, and collagen III stimulated (unstimulated, n=4; collagen 1 & III, n=5). Error bars represent SEM.

### 3.5.4 The function of DDR2 in cell contraction

Despite the binding affinity to collagen I and functional interaction with Integrins, the expression of DDR2 has been speculated to affect the contractility of DD myofibroblasts, which was determined using traction force microscopy (TFM). Six hours prior imaging, DDR2 knockdown and wildtype cells were seeded onto collagen I coated hydrogels that were filled with fluorospheres. High resolution (40X) confocal images were obtained before and after the cells had been detached to capture the magnitude of the displacement of the fluorospheres due to the relaxing of the hydrogel. The displacement fields were then reconstructed into traction force fields, and the total traction force was calculated by summing all the visible traction force surrounding a target cell. There were no statistical differences in the median total traction force between DDR2 knockdown (DDR2-KD) and wildtype cells (WT) (Figure 3.5.3.1). Although DDR2 expression level didn't seem to affect myofibroblast ability to contract collagen I matrices, I observed some morphological and force placement differences between some of the DDR2 KD and wildtype cells, where the latter appeared to be more dendritic-like with the focal adhesion points situated towards the mid-end of the dendritic network, whereas the former were more granulated with the focal adhesion points situated on the perimeter of the cell (Figure 3.5.3.2).

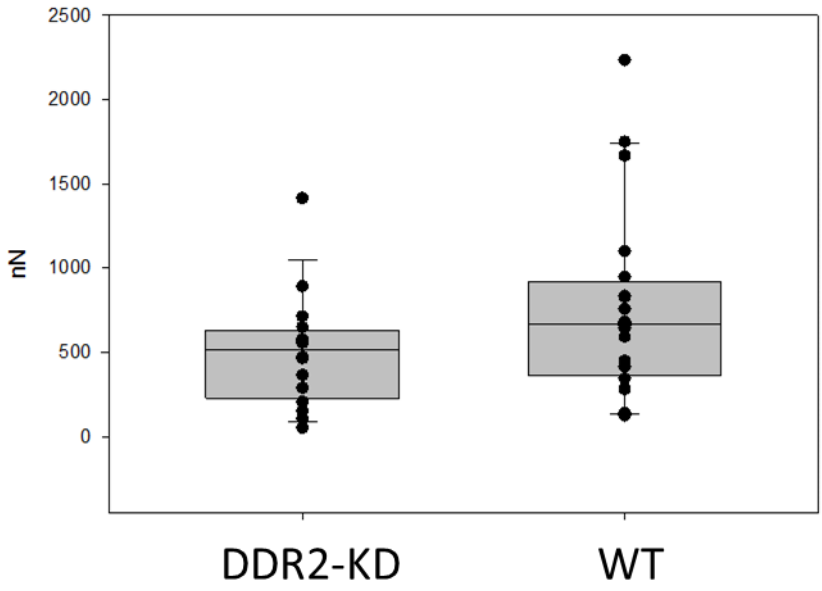



Figure 3.5.3.1 Dot plot showing the distribution of total traction force between DDR2-KD (n=15) and wild type myofibroblast cells (n=20). No statistical differences were detected between the two groups.

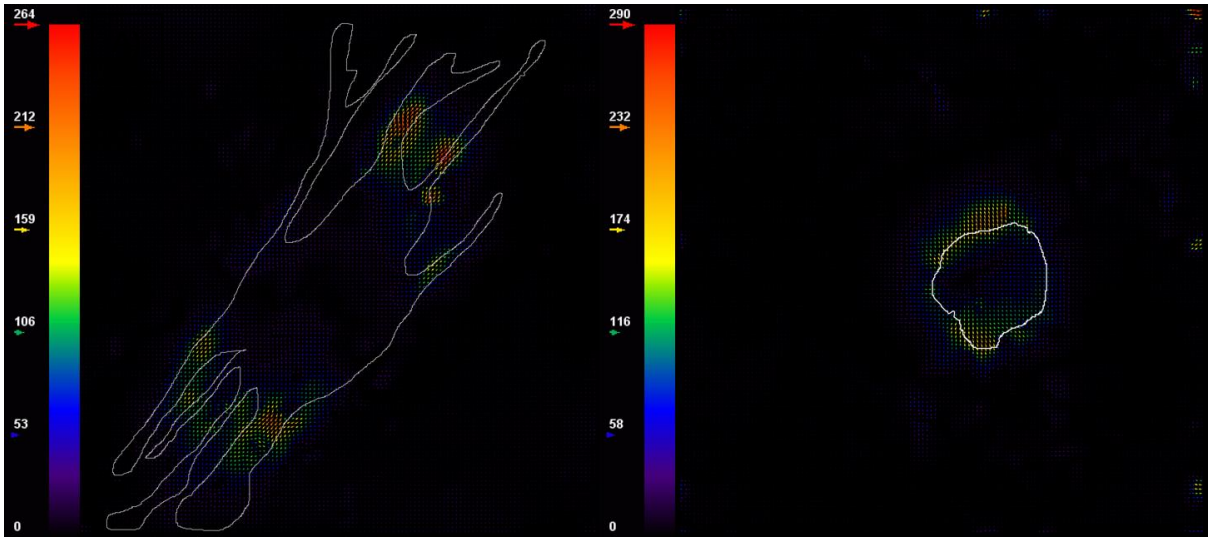



Figure 3.5.3.2 Pictures showing the focal adhesion points of a A) wildtype myofibroblast cell and B) knockdown DDR2 myofibroblast cell. The outline of the cell was sketched using a white line.

### 3.5.5 DDR2 and downstream gene regulation

Gene expression that was regulated by DDR2 expression was examined using RNA-seq on both the DDR2 knockdown and wild type myofibroblast cells. At basal level, a total of 43 differentially expressed genes were detected, where 39 were upregulated and 4 were downregulated (Figure 3.5.4.1). Some of the upregulated genes belong to the known fibrotic pathways, including *CDK4*, *TGFB* and *WISP1* of the SMAD and WNT pathway. No significantly enriched gene ontology (GO) pathways were identified from the gene list, possibly due to the absence of a DDR2 substrate.

The list of differentially expressed genes was expanded when a similar experiment was performed with 12 hours 20μg/ml collagen stimulation. In the collagen I stimulation experiment, a total of 64 differentially expressed genes (including *DDR2*) were detected, where 56 were downregulated and 8 were upregulated in the DDR2 knockdown cells when being compared to wild type (Figure 3.5.4.2. left panel). In a similar experiment using collagen III as a stimulant, a total of 197 differentially expressed genes were detected, where 123 were downregulated and 74 were upregulated in the DDR2 knockdown cells. Twenty-five of these differentially expressed genes were common in both collagen stimulation scenarios, demonstrating a similar fold change as well as direction (Figure 3.5.4.3, right panel). GO enrichment pathway analysis have revealed that a proportion of the genes are involved in enhancing tissue and organ integration processes. To examine whether DDR2 downstream targets are collagen type specific, I compared the differentially expressed

gene list between the aforementioned collagen I and collagen III experiment. Out of the 197 genes that were identified in the collagen III stimulation experiment, 172 have been identified to be collagen III specific. Pathways that are downregulated in the absence of DDR2 during collagen III stimulation include exocytosis regulation, extracellular matrix organisation, response to xenogens, leukocyte activation, antigen presenting of MHC class I.

On the other hand, in the collagen I experiment, 39 out of the 56 differentially expressed genes were collagen I specific. The list of downregulated pathways in the absence of DDR2 include extracellular matrix organisation, extracellular structure organisation, regulation of angiogenesis and upregulation of WNT signalling pathway (Figure 3.5.4.3).

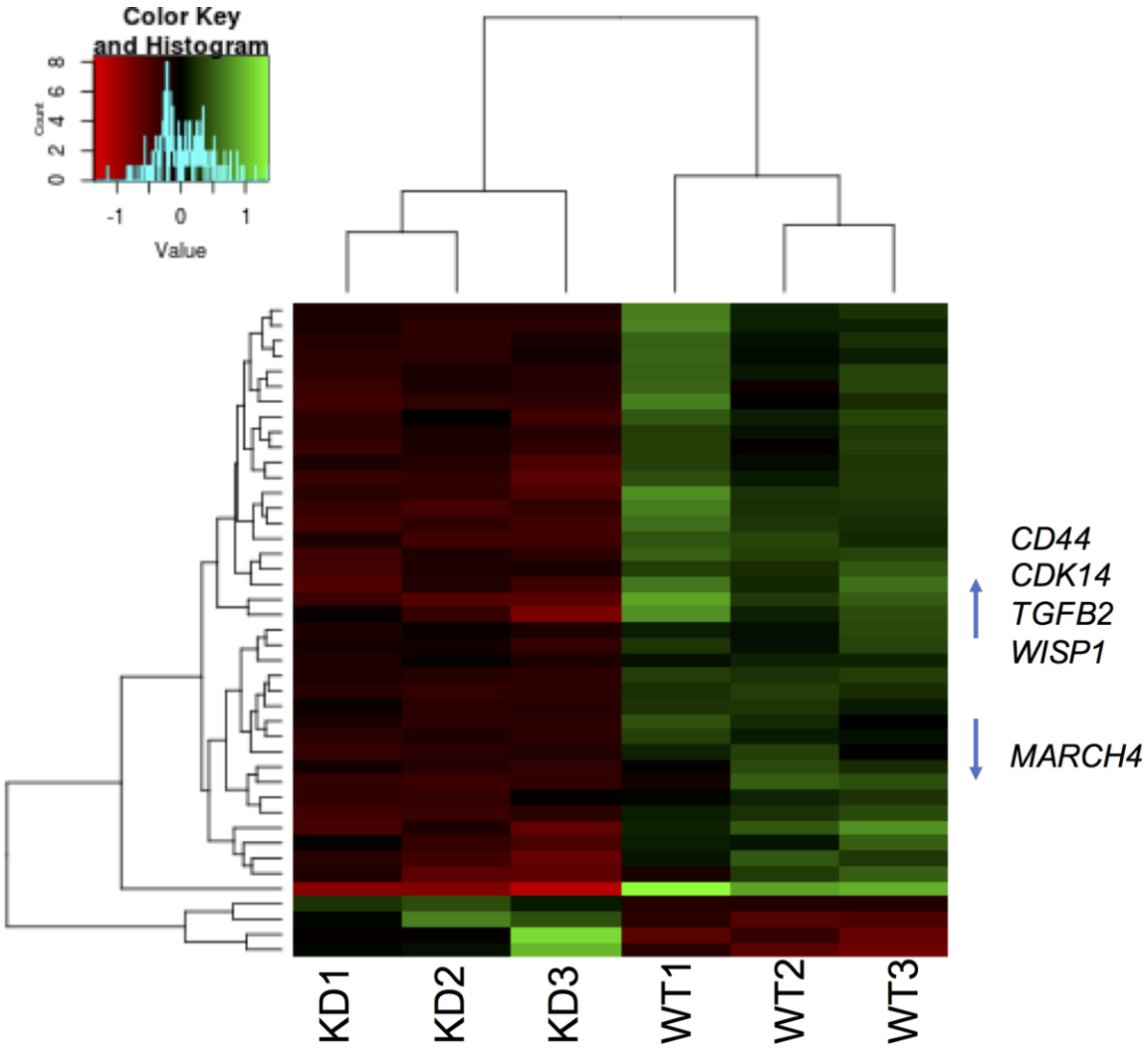



Figure 3.5.4.1 Heatmap showing the 43 differentially expressed genes between the DDR2 knockdown and wild type myofibroblast cells. Four out of the 43 differentially expressed genes were downregulated and 39 were upregulated in the knockdown cells when compared to wild type.

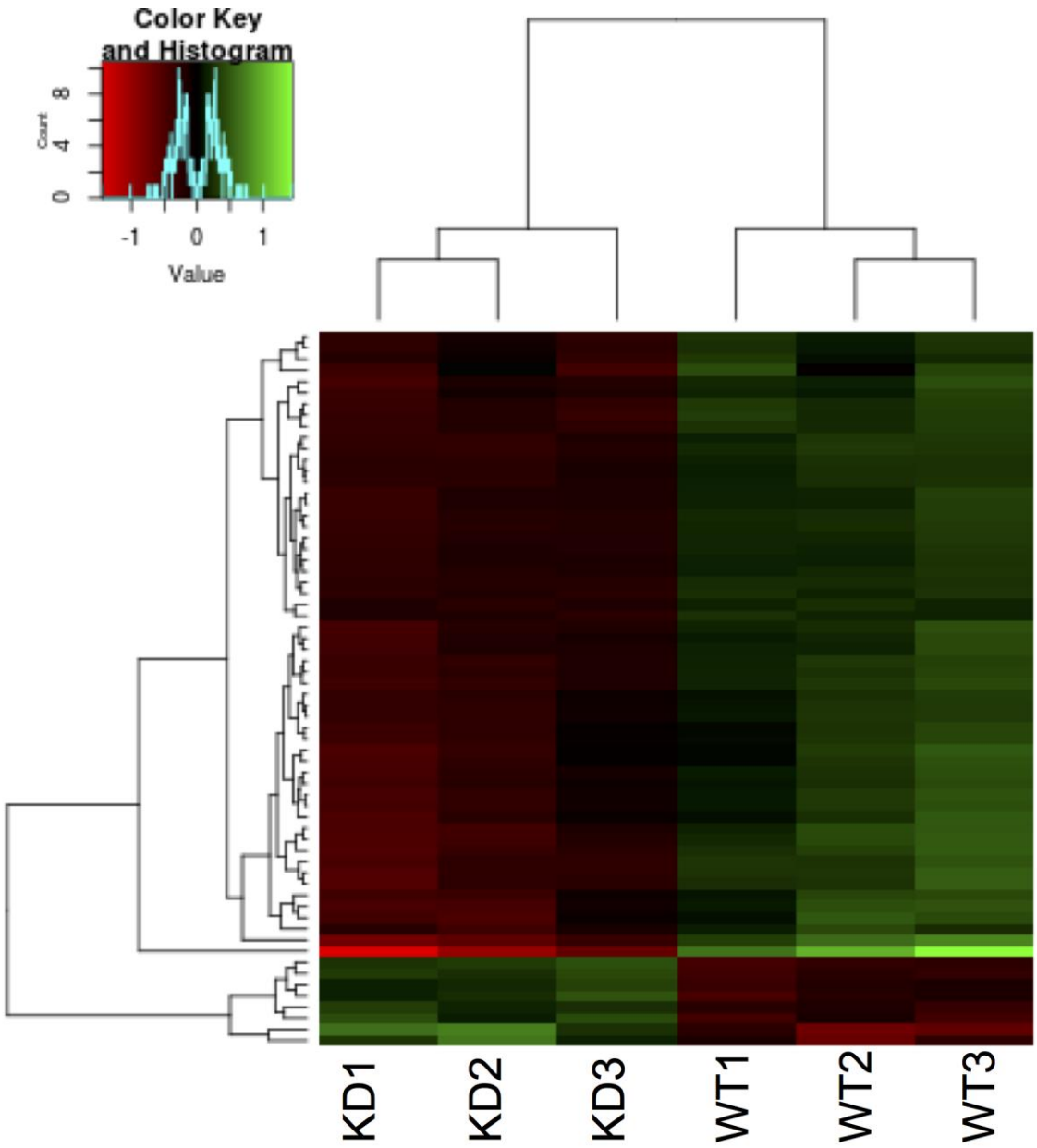



Figure 3.5.4.2 Heatmap depicting the clustering of the differentially expressed genes between DDR2 knockdown and wild type myofibroblast cells post collagen I stimulation. A total of 64 differentially expressed genes (including DDR2) were detected, 56 of which were downregulated in the knockdown cells when compared to the wild type cells.

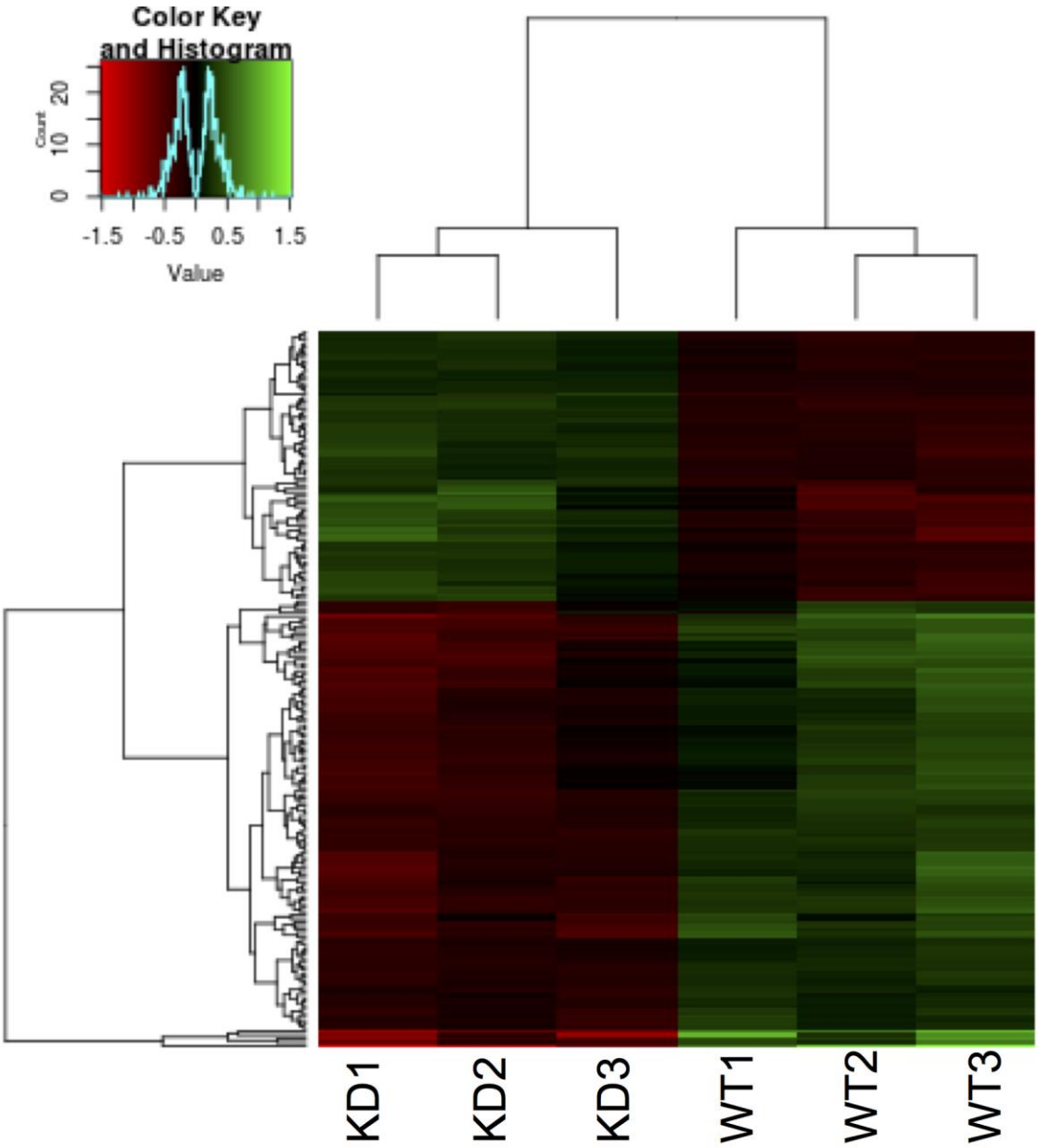
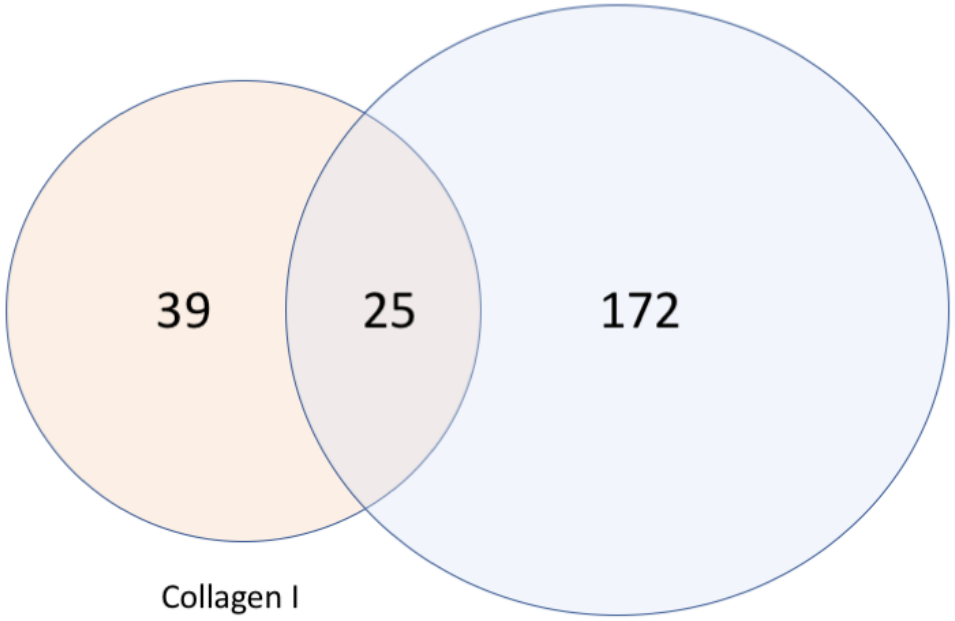



Figure 3.5.4.3 Heatmap depicting the clustering of the differentially expressed genes between DDR2 knockdown and wild type myofibroblast cells post collagen III stimulation. 197 genes were differentially expressed, where 123 were downregulated in the knockdown cells when compared to the wild type cells.

.




Collagen I GO enriched biological processes

| | |
|---|---|
| Negative regulation of canonical Wnt signaling pathway | $6.56x10^{-04}$ |
| Regulation of angiogenesis | $8.83x10^{-04}$ |
| Regulation of vasculature development | $9.90x10^{-04}$ |
| Positive regulation of vasculature development | $1.10x10^{-02}$ |
| Appendage development | $1.13x10^{-02}$ |
| Regulation of fibroblast proliferation | $1.23x10^{-02}$ |
| Limb development | $1.27x10^{-02}$ |
| Extracellular matrix organization | $1.43x10^{-02}$ |
| Extracellular structure organization | $1.74x10^{-02}$ |

Collagen III GO enriched biological processes

| | |
|---|---|
| Response to organic substance | $2.63x10^{-03}$ |
| Cellular response to type I interferon | $3.68x10^{-03}$ |
| Regulated exocytosis | $5.89x10^{-03}$ |
| Extracellular matrix organization | $6.43x10^{-03}$ |
| Regulation of viral genome replication | $6.48x10^{-03}$ |
| Cellular response to chemical stimulus | $8.32x10^{-03}$ |
| Positive regulation of biological process | $9.57x10^{-03}$ |
| Extracellular structure organization | $9.81x10^{-03}$ |
| Leukocyte activation | $1.57x10^{-02}$ |
| Regulation of viral entry into host cell | $1.58x10^{-02}$ |
| Antigen processing and presentation of endogenous peptide antigen via MHC class I | $1.79x10^{-02}$ |

| | |
|---|---|
| Anatomical structure formation involved in morphogenesis | $1.98X10^{-03}$ |
| System development | $2.95X10^{-02}$ |
| Response to growth factor | $4.39X10^{-02}$ |

Common GO enriched biological processes

| | |
|---|---|
| Positive regulation of multicellular organismal process | $1.95x10^{-02}$ |

Figure 3.5.4.4 Venn diagram demonstrating a list of collagen I and III upregulated and downregulated GO biological processess that were presented in the DDR2 knock down myofibroblast cells.

## 3.6 Discussion

In this chapter, I examined and successfully revealed some of the fibrotic driving biological functions that DDR2 exhibits in DD myofibroblast cells. Interestingly, these biological functions appeared to be collagen type specific, possibly due to the capability of DDR2 being able to bind other collagen peptides beside the GVMGFO motif domain, as observed in DDR1 [191]. To date, multiple downstream signalling pathways of DDR2 have been identified though they were mostly through the activation of collagen I. As a result, the biological function of collagen III in fibrosis development is not clear, and its understanding is crucial for DD research as altered collagen I : III ratio is one of the primary features found in affected tissue. One of the main concerns of using commercial collagen III in experiments for comparisons with collagen I is that it often contains small traces of the collagen I, which in our case was around 15%. Although this could be problematic, because the amount of collagen I was relatively small in comparison to collagen III, I was able to segregate their functional differences when testing for differentially expressed genes with only a small percentage of overlaps, which was expected.

From the results of the *DDR2* knockdown RNA-seq data, it is clear that both collagen I and III can induce extracellular matrix remodelling via binding with DDR2, which might facilitate the motility of myofibroblasts through ECM to its target destination, release the stored cytokines and growth factors, and strengthen the structure of the

ECM. Dissecting further the differences in enriched pathways between collagen I and III stimulation, I have revealed that collagen I plays a role in promoting cell proliferation, and angiogenesis, which both were previously demonstrated in a IPF study[144], as well as negative regulation of the WNT signalling pathway. Together this suggested that collagen I pauses the production of fibrotic genes such as *COL1A1*, *Alpha smooth muscle actin* (*ACTA2)* and switches the myofibroblast cell function into promoting proliferation and positioning in the ECM. This behaviour was replicated in the acute migration experiment where I compared DDR2 knockdown with wild type myofibroblast cells under the influence of collagen I. However, I couldn't show any differences in proliferation rate, at least in part due to the limitation of the siRNA transfection system I used. Unfortunately, there isn't a perfect solution for this issue as using a stable transfection system would also require the use of a cell line instead of primary cells, which are susceptible to phenotypic drift [192].

The functional role of collagen III, in contrast, is to promote a pro-inflammatory and autoimmune response in myofibroblast cells and the surrounding immune cells. This includes increasing the expression of genes that play a role in leukocyte activation, endogenesis, and antigen presenting via MHC class I. Some of the genes whose expression is increased are involved in pro-inflammatory cytokine signalling, including *IL1R* and *IL7R*. Besides facilitating a pro-inflammatory response, differentially expressed collagen genes such as *COL5A3* and *COL8A1* have also been identified, suggesting collagen III is also pro-fibrotic: *COL8A1* has previously been reported to be upregulated in DD cells [193].

However, the full biological function of both collagens in DD is unclear. Neither of the collagen stimulation data demonstrated any differentially expressed genes that could be implicated in cell contraction. I also didn't observe any changes in traction forces on a collagen I surface between the knockdown and the wild type fibroblast cells. However, the placements of the traction force were different between the elongated fibroblast like cells that are mostly found in the wild type myofibroblast and the more granulated cells that were presented in DDR2 knockdown, where the focal adhesion points were mainly in the middle, or between, of the dendritic-like network rather than on the perimeter of the cells.

It is evident that collagen I and III play a crucial role via DDR2 in DD development, through promotion of cell proliferation and inflammation. The evidence from this chapter has suggested that the chronic inflammation that is observed in DD patients is perpetuated by collagen III via DDR2, which in turn leads to a longer pro-fibrotic response. This suggests that a reduction of the collagen III expression or inhibition of signalling via DDR2 in DD patients could be a potential therapeutic target. Further in vivo work by co-culturing collagen III stimulated myofibroblast with immune cells, such as T cells, would support this finding, and be helpful in elucidating the pathophysiological mechanism.

# Chapter 4 Genetic fine mapping of DD myofibroblasts and fibroblasts

## 4.1 Abstract


Introduction

The identification of a functional mechanistic association between a risk variant and a disease of interest requires cell type and organ specific genomic landscape data. Comprehensive genomic landscape evaluation consists of chromatin accessibility, chromatin contact mapping, chromatin modifications, and transcriptomic profiles. By employing high throughput sequencing techniques, I aimed to acquire the aforementioned genomic landscape data in Duputyren disease derived myofibroblasts and nearby control cells, and perform a comprehensive biological function evaluation of the previously discovered Dupuytren disease associated genetic variants.

Materials and methods

Open chromatin regions of Dupuytren myofibroblasts, and fibroblasts from non-affected palmar skin were analysed using ATAC-seq. Spearman's correlation was used to identify correlation between primary cells and cultured cells. Genomic features

were identified using Chipseeker and motif finding using Homer. Differential open chromatin accessibility was identified using Diffbind. Visualisation was using R and IGV. Cis and trans interactions with the region harbouring associated variant rs61811977 were investigated using circularised chromatin conformation capture (4C) sequencing, and the data collected in 4C-ker and visualised using WashU epigenomic browser and R.

## Results

Most open chromatin ATAC-seq fragments were found located in the intergenic, enhancer and promoter regions. Motif finding analysis revealed the most significant transcription factor binding these open chromatin regions to be Basic leucine zipper transcription factor (BATF). Other biologically interesting enriched binding motifs included SMAD2/3 and TEA Domain Family Member 1 (TEAD1). 4C analysis has revealed interaction between the bait enhancer region around variant rs61811977 and multiple promoters. Integration between ATAC-seq and RNA-seq data revealed 136 common differentially expressed genes whose promoters were also within open chromatin: 81 were myofibroblast specific, and 55 were fibroblast specific.

Discussion

My genomic landscape data have provided information as to how DD myofibroblasts are regulated. Further differential and integrative analysis has formed a connection between DD associated variants and biological function of *EPDR1*, *SFRP4* and *MAFB*. Both *DDR2* and *HSD17B7* have known pro-fibrotic functions but remain poorly understood in DD.

## 4.2 Introduction

The translation of disease associated variants into a meaningful biological function has historically been a challenging task in genetics research. This is primarily due to the lack of complementary expression and regulatory information such as areas of open chromatin, three-dimensional chromatin structure, and transcriptomic data for the surrounding genomic region. Reasons for this shortage include technical problems in isolating appropriate regions and sequencing the subsequent DNA library, and the fact that there are multiple cell types present in the disease tissue of interest, meaning that the cell in which a particular variant exerts its effect is often not clear. However, in recent years improved genome-wide methods have been introduced, improving the assessment of function of SNPs [194]. The selection of an appropriate cell type to assess genomic landscape data is crucial when designing a landscape profiling study, as regulatory and expression information is often cell type and organ specific, and sensitive to phenotypic drift *in vitro*; a phenomenon that can occur as the passage number of cells increases during cell culture or simply due to the limitations of laboratory *in vitro* culture conditions [192]. The analysis of the landscape data should be conservative to avoid over interpretation as both RNA and protein expression can naturally be affected by a myriad of post transcriptional, and post translational regulations and modifications, including miRNA mediated degradation, the stability of mRNA, transcription and translation rate, amino acid glycosylation, etc. A comprehensive profile containing transcriptomic, chromatin accessibility and modification profiles can provide the key necessary bridging information between

biological functions and associated variants, and this amalgamation of profiles are the basis of many SNP prioritisation methods such as Enlight [195], FUMA[196] and GREGOR[197]. Broadly speaking, these methods interrogate the similarities between genomic location of the associated variants, allele specific expression quantification trait loci (eQTL), chromatin open regions, chromatin modification, and predicted transcriptional factor binding motifs. These data are integrated to generate a list of candidate genes, which are then collated to form a network of enriched pathways relevant to the disease of interest. By using publicly available databases such as Haploreg[198], RegulomeDB[199], GTEx[200], GWAS3D[201], these methods provide a platform to link causal variants, intermediate transcriptional regulation, the involved cell specific chromatin modifiers and transcriptional factors, and differentially expressed genes of the downstream pathways, which are far superior to text-based prioritisation methods such as GRAIL [202]. However, extra caution should be exercised when using these prioritisation methods to predict biological functions as often their integrated landscape data was not generated from cells relevant to the disease of interest of the user. The datasets used are commonly derived from blood, as this is the most easily accessible human tissue, and are rarely relevant to musculoskeletal conditions.

The transcription of DNA is a tightly regulated process that is governed by layers of epigenetic code, such as DNA methylation, nucleosome positioning and histone modifications. Underneath the sequestered inactive genomic regions are hidden regulatory elements that are responsible for controlling gene transcriptional activity.

These regulatory elements are categorised into promoters, enhancers and insulators depending on their regulatory function[203]. They interact with each other via transcription factors such as activators, repressors and silencers to form various DNA loops e.g. promoter-promoter, enhancer-insulator and insulator-promoter, to permit or inhibit the transcriptional machinery complex from being recruited to the target site and initiate RNA transcription[154 204]. Many enzymatic and non-enzymatic methods have been developed for identifying of open chromatin regions and to predict interactions between regulatory elements and transcription factors, including the sequencing of DNase I hypersensitive sites, MNase assisted isolation of nucleosome (MAINE), Formaldehyde assisted isolation of regulatory element (FAIRE), assay for Transposase-Accessible Chromatin (ATAC), by utilising their DNA fragmenting properties. Newer methods such as ATAC-seq are generally used more frequently because they are simpler to perform, the data created has less background noise, and most importantly they require fewer input cells and are thus more likely to preserve the original phenotype of the investigating cells and lower the chances of phenotypic drift.

Tn5 is an engineered hyperactive Transposase, which when utilised on an intact nucleus excises itself and confers the attached oligo sequences to the 5’ and 3’ ends of open chromatin regions, thereby allowing the exploration of the unknown open chromatin regions through sequencing [205]. One of the applications of open chromatin data is for footprinting CTCF, a zinc-finger binding protein which can possess the properties of transcription activators, insulators and repressors and is crucial for

regulating a myriad of genomic processes including transcription regulation, imprinting, enhancer-promoter interaction mediated DNA-looping and general organisation of the chromatin architecture [206-208]. There are more than 30,000 CTCF binding sites across the human genome and abnormal expression and mutation of *CTCF* have been associated with debilitating conditions and diseases, such as colorectal, endometrial cancer, etc.[209 210]. Other techniques, such as variants of chromatin immunoprecipitations (ChIP), have been developed for investigating chromatin accessibility that is indicated by specific histone marks, as well as various chromatin conformation captures (3C and variants) for identifying global chromatin modification and cis or trans interaction between DNA regulatory elements, distal or proximal, and the promoter of genes of interest.

Other than the merit of high fidelity, newer sequencing methods also make integration between different genomic landscape data possible, thereby maximising the usefulness of individual profile and concomitantly filling the gaps presented in each. This integrative process was successfully demonstrated in a pancreatic research study, where the authors revealed and validated novel signatures of different islet cells types and subsequently their contribution to the pathophysiology of diabetes. Utilising the same genomic profiles, they have also successfully established a clear connection between multiple diabetes associated variants with the dysfunction of β-cells by assessing the overlaps with previously published GWAS results [211]. Despite its practicality, no genomic landscape data of Dupuytren disease cells has been collected

to date and its acquisition will help answer many questions regarding fibrosis development.

## 4.3 Aims and hypotheses

Part of the challenge of associating a causal variant to a disease is the lack of disease specific cell regulatory information at epigenomic and transcriptomic level. My first aim of this chapter was to curate this level of data using DD myofibroblast cells. My second aim of this chapter was to segregate the differences in open chromatin regions and expressed genes between DD myofibroblasts and palmar fibroblasts, and between passaged cells and primary cells. My final aim was to investigate a particular open chromatin region that contains a credible set of DD-associated variants for its interaction with nearby genes hoping to further understand its biological function.

The curation of epigenomic data for a disease specific study has always been a laborious task that has often been neglected. My hypothesis was that this type of data, together with transcriptomic data from RNA-seq, would reveal a set of novel genes in the myofibroblast cell that is important in the pathogenesis of DD, refining our understanding of the individuality of their biological function.

## 4.4 Materials and methods

### 4.4.1 Tissue culture and sample processing for ATAC-seq

For ATAC-seq performed using cells from fresh Dupuytren nodules, the sample was finely minced with surgical scalpels and homogenised using gentleMACS c-tube (Miltenyi Biotec, Surrey, UK) utilising the pre-set ovarian cancer tissue protocol. The sample was then placed on ice for 30 seconds and was homogenised again with the same setting. This step was then repeated for a third time and the fully homogenized sample was then kept on ice until further processing.

For tissue culture samples, cells were disassociated from the culture flask using Tryple (Fisher Scientific, Paisley, UK) and centrifuged at 300g for five minutes. The sample was immediately placed on ice. Disaggregated cells were washed twice using 5ml of ice cold PBS and pelleted at 300g for five minutes. For every 2 million cells, 1ml of HS lysis buffer, which consisted of 50mM KCl, 10mM MgSO4.7$H_2O$, 5mM HEPES, 0.1% IGEPAL CA630, 40μl of 25x Protease inhibitor cocktail (Sigma Aldrich, Dorset, UK) and 3μl of DTT (Sigma Aldrich, Dorset, UK) was prepared. The pellet was gently resuspended in the lysis buffer by pipetting up and down, followed by an incubation at room temperature for two minutes and 30 seconds. Samples were then pelleted at 1000g for five minutes at 4$^oC$ and washed with RSB, which consisted of 10mM NaCl,

10mM Tris pH 7.4 and 3mM $MgCl_2$, twice. Prior to the penultimate wash, 10μl was withdrew from the resuspension, mixed with an equal volume of Tryphan blue (ThermoFisher scientific, Paisley, UK) and examined for successful lysis on a haemocytometer, where cells would appear blue, and counted for later use. If cells were not completely lysed, the lysis step would be repeated once more but reduced down to only 30 seconds each time to avoid over lysis. Pelleted nuclei were re-suspended in nuclease free water (GE Healthcare Life sciences, Little Chalfont, UK) at 50 to 100 thousand cells per 5μl depending on the activity and concentration of the Tn5 Transposase (Illumina, Cambridge, UK).

The Tn5 enzyme was a part of the Nextera XT library prep kit (Illumina, Cambridge, UK), and a typical reaction was set up by adding the reagents in the following order, 10μl reagent TD, 5μl cell suspension and 5μl TDM. After mixing gently by pipetting up and down, the reaction mix was tagmented at 37°C for 30 minutes and held at 10°C. To neutralize the Tn5 enzyme and stop the reaction, 5μl of NT buffer was added and incubated at room temperature for five minutes. The tagmented sample was cleaned using Clean & Concentrator 5 kit (Zymo research, CA, USA) according to manufacturer's instructions. Briefly, five volumes of DNA binding buffer were added to the tagmented DNA and the mixture was transferred to a Zymo-spin column. With a centrifugation step at 10000g between each step, the column was washed with 200μl DNA wash buffer and eluted with 10μl nuclease free water. The sample was barcoded using a pair of primers which included a universal and an index PCR primer (IDT technologies, Leuven, Belgium) as described previously [205].

The PCR reaction mix consisted of 10μl tagmented DNA, 2.5μl 0.5 mM universal and index primer and 25μl NEBnext High Fidelity PCR master mix (New England Biolabs, Hitchin, UK). The PCR cycling conditions were 72ºC for 5 minutes, 98ºC for 30 seconds, followed by 10 cycles of 98ºC for 10 seconds, 63ºC for 30 seconds and 72ºC for 30 seconds, and the final extension step was at 72ºC for 5 minutes. The sample was purified again with Clean & Concentrator 5 kit (Zymo research, CA, USA) and its fragment size distribution was determined using high sensitivity Tapestation tapes (Agilent Technologies, Winnerish, UK). Concentration was determined using qPCR method with the KAPA library quantification kit (Roche, West Sussex, UK). Samples were pooled at a final concentration of 4nM, or minimum 500pM, prior to being denatured according to the Denature and Dilute library guide (Illumina, Cambridge, UK). Briefly, the sample was mixed with an equal volume of 0.2N NaOH, followed by an equal volume of pH7 200mM Tris-HCl and then finally diluted to 20pM in buffer HT1. Diluted and denatured sample were loaded onto a 75 cycle Nextseq 500/550 high output kit v2 (Illumina, Cambridge, UK).

### 4.4.2 Sample processing for circularised chromatin conformation capture sequencing (4C-seq)

Cultured cells were disaggregated and pelleted as described in section 4.3.1. After resuspending the pellet with DMEM, the number of cells was counted, and the suspension was pelleted again by centrifugation and resuspended in 5ml PBS

supplemented with 10% fetal bovine serum. Cells were fixed by adding 5ml of 4% formaldehyde (EMS, PA, US) and gently mixed on a shaker for 10 minutes at room temperature. The cross-linking reaction was then quenched by adding 1ml of 1.425M glycine (Sigma Aldrich, Paisley, UK) and the sample was then immediately placed on ice. Cells were pelleted at 4$^{o}$C at 400g for 8 minutes and resuspended in 5ml of ice cold lysis buffer, which consisted of 50mM Tris, 150mM NaCl, 5mM EDTA, 1% IGEPAL, 1.15 Triton X-100, 100μl 100X protease inhibitor. After 12 minutes of incubation, the lysed cells were pelleted and resuspended in 360μl of nuclease free water and 60μl of 10x endonuclease restriction enzyme buffer 1 (New England Biolabs, Hitchin, UK). Fifteen microliters of 10% SDS (Sigma Aldrich, Paisley, UK) were added into the sample and the reaction was mixed by pipetting up and down vigorously until no visible cell aggregates were observed.

The reaction mix was incubated at 37$^{o}$C for one hour with agitation at 1000 rpm. The SDS was quenched by adding 150 μl of 10% Triton X-100 and incubated for one further hour at 1000rpm. After the addition of 600μl of 1x restriction enzyme buffer 1, 400U of the first restriction enzyme DpnII was added and the sample was incubated at 37$^{o}$C for four hours with agitation at 400rpm. 200U of DpnII (New England Biolabs, Hitchin, UK) were then added prior to overnight incubation and a second time, four hours prior enzyme inactivation at 65$^{o}$C for 20 minutes.

Ligation of the digested product was performed in a reaction mix, which consisted of 700μl 10x T4 ligation buffer, 7ml nuclease free water and 50U T4 DNA ligase (New England Biolabs, Hitchin, UK), and was incubated overnight at 16°C. The successfully ligated product was then de-crosslinked with 30μl of 10mg/ml proteinase K at 65°C, followed by an RNA digestion step with 30μl of RNAseA at 37°C for 45 minutes. 7ml of phenol chloroform was added to sample, which was centrifuged and the water phase was transferred to a new falcon tube. Precipitation of DNA was achieved by adding 0.2M NaAC, 28μl of 5mg/ml glycogen and 35ml of 100% ethanol to the sample and incubation at -80°C until the mixture was completely frozen. The precipitated DNA was pelleted at 8350g in a 4°C centrifuge for 20 minutes. After removing the supernatant, the pellet was purified using 70% ethanol, and was centrifuged at 3300g at 4°C. The purified DNA pellet was dissolved in 150μl Tris pH 7.5 at 37°C.

A second round of DNA digestion was performed by adding 50μl of restriction enzyme buffer 2 (Thermo Fisher Scientific, Hemel Hempstead, UK), 60U DpnII (Thermo Fisher Scientific, Hemel Hempstead, UK) and 295μl of nuclease free water to the previously purified DNA, followed by incubation overnight at 37°C. After restriction enzyme inactivation at 65°C for 25 minutes, 12.1ml of nuclease free water, 1.4ml 10X ligation buffer and 100U of T4 DNA ligase were added to the sample, and together were incubated at 16°C overnight. The sample was then precipitated and purified using the same method as described above and dissolved in 75μl 10mM pH 7.5 Tris. Samples were purified using Clean & Concentrator-5 kit (Zymo Research, CA, US). Bait primer sequences were designed using a web-based primer design engine that was created

by the Nobrega lab (Department of Human Genetics, University of Chicago, US; URL: https://mnlab.uchicago.edu/4Cpd/) and are shown in table 4.3.2.1. Samples were then barcoded using indexes from the NEBNext Multiplex Oligos for Illumina Index Primers Set 1 (New England Biolabs, Hitchin, UK) according to manufacturer's instructions. Indexed samples were purified twice using Sera-mag beads at 1:1 ratio to the sample as described in the RNA-seq section in chapter 3.4.9. Purified samples were pooled together before sequencing if necessary.

Table 4.3.2.1) Bait primers used for 4C-seq

| | |
|---|---|
| 4C_chr1_Bait_FWD | CTCACAATTTAGAGACGATC |
| 4C_chr1_Bait_REV | GTGCTAATCCTCTGTTTGAG |

## 4.4.3 ATAC sequencing data processing

Raw DNA sequence Fastq files were acquired from Basespace (Illumina, Cambridge, UK) after having been deposited from the NextSeq sequencer. Sequences that were acquired from the same read were catenated, and subsequently trimmed from any adaptor sequences using Cutadapt. The trimmed sequences were then aligned to the human genome hg19 reference panel using Bowtie2[212], sorted numerically from smallest to largest by chromosome number and position, and converted into bam

format using Samtools. As we were only interested in chromatin accessibility of the autosomal chromosomes, duplicated alignments and fragments that had been aligned to the mitochondrial genomic DNA were removed. The reads were then shifted +4 bp with the positive strand and -5bp with the negative strand using R package ATACseqQC [213], and indexed using Samtools. For quality control, Fraction of reads in peaks (FRip) was calculated, where a value > 0.3 deemed as good quality. Aligned fragments were split by size using Samtools for generating various QC plots using ATACseqQC. Enrichment peaks were offset by 75bp and called using MACS2 [214] and were annotated to gene level using annotatePeaks.pl provided in the package Homer [215]. Spearman's correlation and PCA were calculated using Deeptools [216] to highlight similarities between samples of various passages. Differential peaks between samples were called using Diffbind [217] and visualised using a Venn diagram. Binding motifs of the differential peaks were predicted using the script findMotifsGenome.pl from Homer and $p<0.05$ was regarded as significant. Sample peaks were pooled before being annotated for genomic features using R package CHIPseeker [218]. Open chromatin region peaks were smoothed out using BamCoverage from Deeptools and imported into Integrative genome browser [219] for visualisation.

### 4.4.4 4C sequencing data processing

4C sequences were acquired as described in section 4.3.3 and were trimmed off from the bait adaptor sequences using Cutadapt. Trimmed sequence were aligned to

fragments of the DpnII reduced hg19 human genome reference panel, which was created using reduced_genome.sh [220] and Bowtie2. Aligned sequences were counted using sam_bedGraph.sh. Self-ligated and undigested fragments were identified using the coordinates of the bait primer sequences and were removed from the count files. Interactions between bait and target regions were analysed using R package 4C-ker [220] and visualised using R ggplot2 and WashU Epigenome browser (URL: http://epigenomegateway.wustl.edu/browser/).

### 4.4.5 Integration of RNA-seq data with ATAC-seq data

Raw RNA sequences from myofibroblasts, and palmar skin fibroblasts were a gift from the Nanchahal / McCann group. Quality control and data processing of raw data was identical to the method described in chapter 3.4.3. Differential analysis was performed using DESeq2 [188] . Differentially expressed RNA peaks were visualised using IGV.

## 4.5 Results

### 4.5.1 Quality control of ATAC-seq data

To examine the quality of the ATAC-seq samples, fragments that were successfully aligned to the hg19 assembly of the human genome were divided into two populations; below 147bp and above 147bp. They represent the theoretical length of nucleosome-free and mononucleosome regions, which are commonly identified in intact nuclei during the tagmentation process of ATAC-seq library preparation. With regards to genomic location, fragments of the nucleosome regions are primarily congregated within 1kb of transcriptional start sites (TSS), which is in contrast to the mononucleosome fragments, indicating the integrity of the mononucleosomes within our samples (Figure 4.5.1.1). The open chromatin regions in the proximal of TSS are evenly distributed across the entire genome of all the tested myofibroblasts samples as demonstrated in the heatmap (figure 4.5.1.2). There is a strong correlation in the distribution of open chromatin regions across the entire genome between primary, passage one and two DD myofibroblast cells but a drop was observed in passage three (Spearman correlation, figure 4.5.1.3), suggesting a phenotypic drift in cell behaviour.

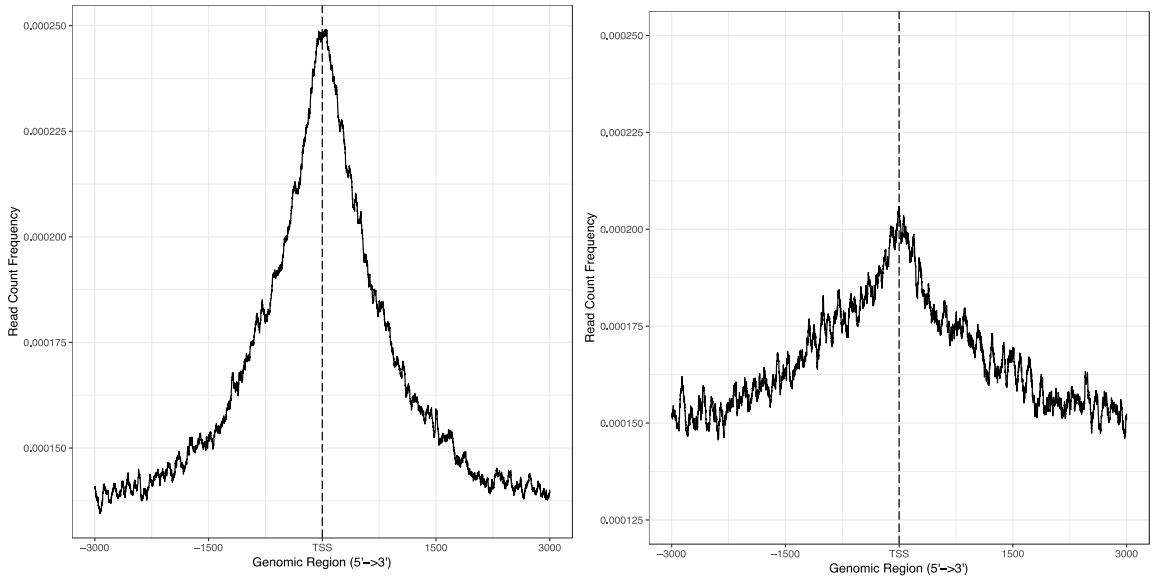



Figure 4.5.1.1 Left) Typical distribution of fragment shorter than 147bp, representing nucleosome free regions. Right) longer than 147bp, representing the mononucleosome regions.

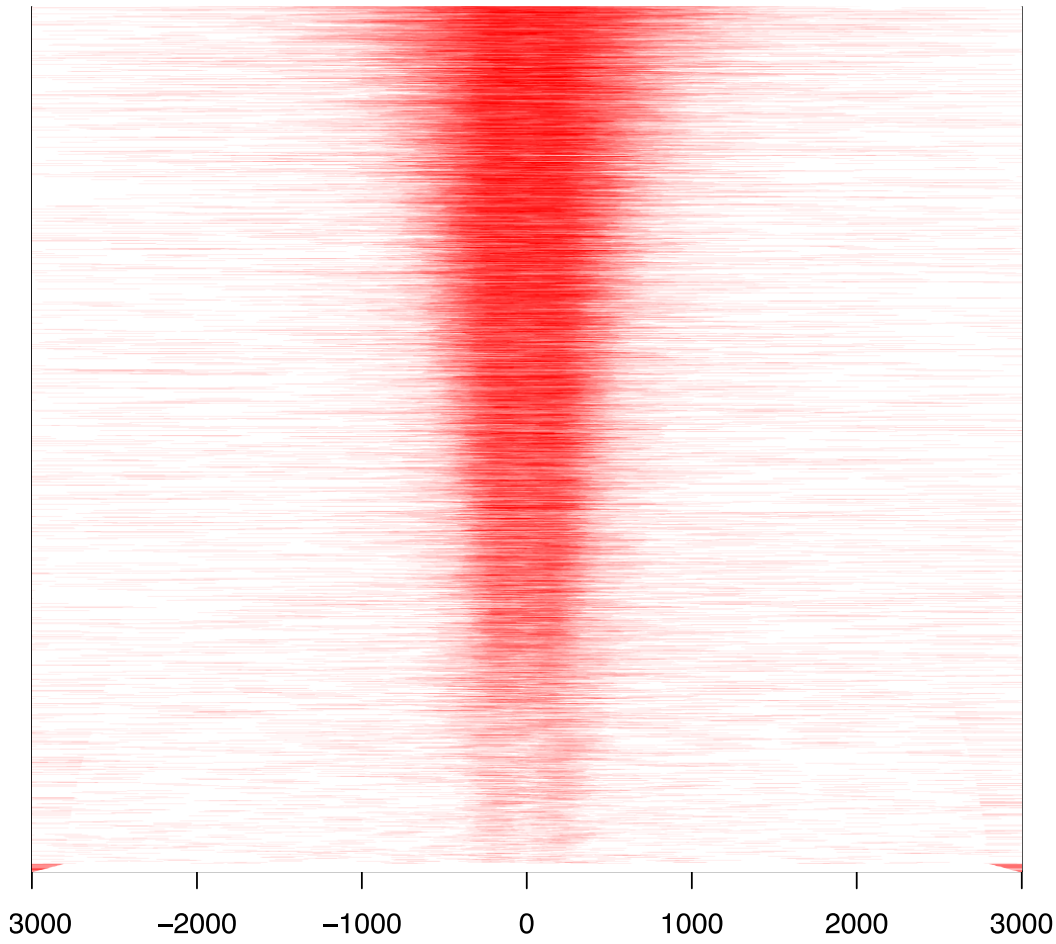



Figure 4.5.1.2 Heatmap demonstrating the distribution of ATAC-seq fragments 3000bp upstream and downstream of TSS across the entire genome.

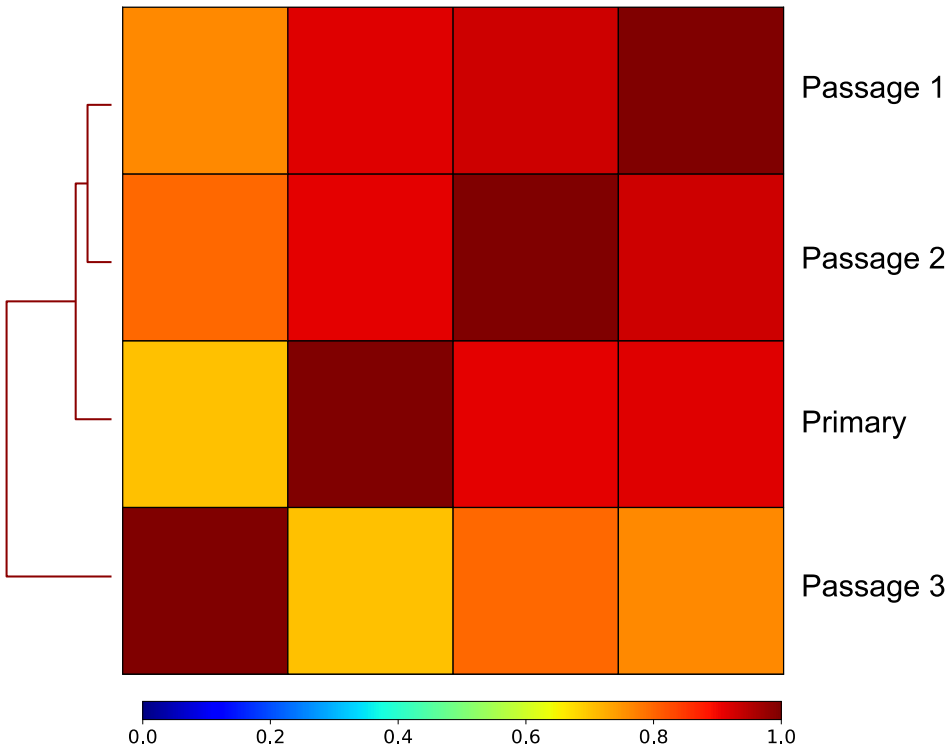



Figure 4.5.1.3 Heatmap depicting a strong correlation of open chromatin accessibility between primary cells from Dupuytren tissue, passage 1 and passage 2 myofibroblast cells. A decrease in correlation can be observed in passage 3 cells.

### 4.5.2 The distribution of open chromatin regions in myofibroblast cells

In order to strengthen the signal of the open chromatin peaks that were generated from the ATAC-seq experiment, functional annotation was performed on pooled data of all the tested passage one and two myofibroblast cells, which comprises a total of 10 samples. The majority of the open chromatin regions are situated in the proximal region of the TSS, introns and distal intergenic regions, indicating the active binding of polymerase transcription machinery to the genome. Approximately 2 ½% of the open chromatin regions annotated are situated within exons, which is a strong indication of corresponding genes were actively being transcribed at the time when the cells were being harvested (Figure 4.5.2.1). Transcription factor binding occurred mostly proximal to the TSS, or in enhancer regions, suggesting long range enhancer-promoter interactions (Figure 4.5.2.2). With regards to biological function, the locations of the open chromatin regions are primarily situated near genes that are responsible for cell growth, cell survival and proliferation as shown in the results of the Reactome pathway analysis (Figure 4.5.2.3). Further analysis to identify the enriched transcription factor binding motifs have revealed that many open chromatin regions are situated within the enhancer and promoter of pro-fibrotic or pro-inflammatory genes including *BATF*, *TEAD1* of the Hippo signalling pathway and *SMAD2/3* of the TGF-β signalling pathway(Table 4.5.2.1). Using GO analysis on the list of binding targets of the top three transcription factor binding motifs have revealed a list of GO biological processes that are mainly related to cell growth and motility (Table 4.5.2.2).

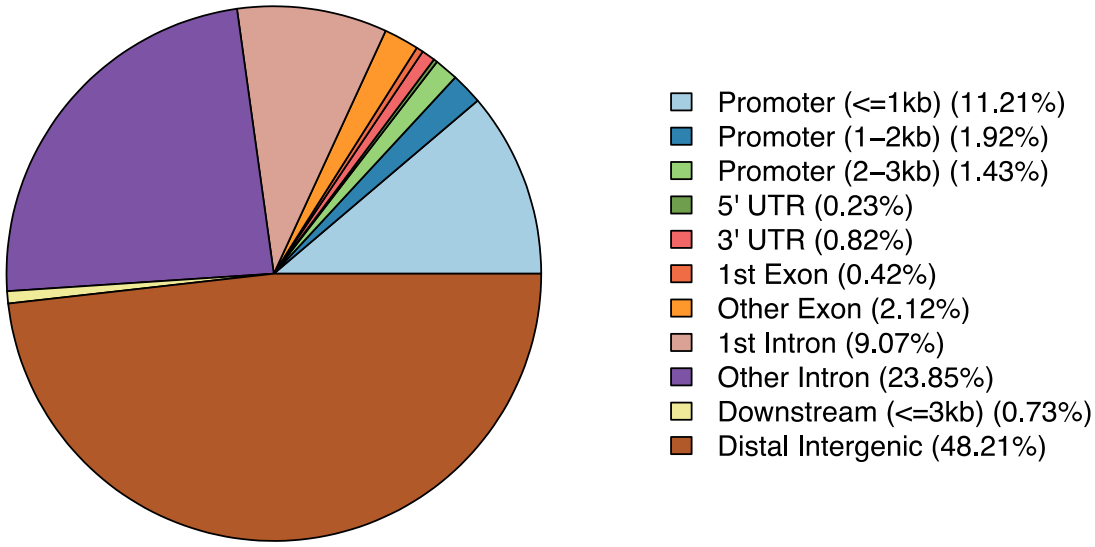



Figure 4.5.2.1 Pie chart showing the distribution of open chromatin peaks in Dupuytren myofibroblast cells.

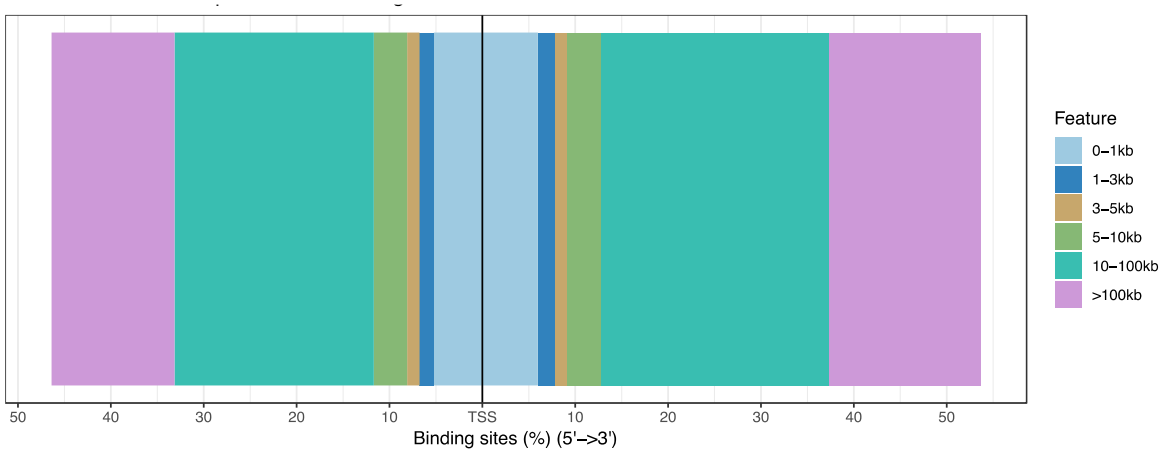



Figure 4.5.2.2 Diagram showing the distribution of transcriptional factor binding sites (%) in respect of the TSS. Each colour represents a range of physical distances, e.g. turquoise is between zero and 1kb upstream or downstream of the TSS.

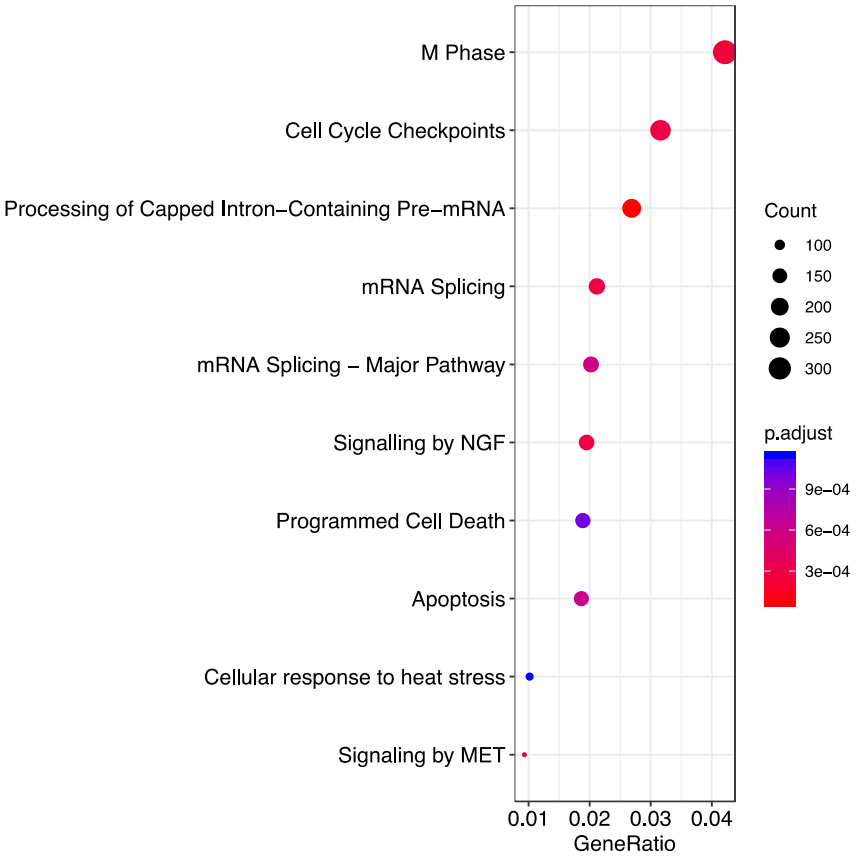



Figure 4.5.2.3 Top 10 enriched pathways (FDR < 0.05) in Dupuytren myofibroblast cells based on the enriched open chromatin peak location across the entire genome using R package Chipseeker and Reactome pathway analysis. A p value less than 0.05 indicate a statistically significant pathway enrichment over a set of random genes. The definition of each pathway and its constituent gene sets can be found in the Molecular Signature Database (MSigDB).

(http://software.broadinstitute.org/gsea/msigdb/genesets.jsp?collection=CP:REACTOME)

Table 4.5.2.1 Top 15 transcription factor binding motifs that were enriched in the assayed Dupuytren myofibroblasts. A p value less than 0.05 indicate a statistically significant motif enrichment over a random background region.

| Rank | Binding motif | p-value | Percent of targets(%) | Transcription factor |
|---|---|---|---|---|
| 1 |  | $1x10^{-6632}$ | 30.42% | BATF |
| 2 |  | $1x10^{-668}$ | 13.20% | SP2 |
| 3 |  | $1x10^{-664}$ | 14.80% | TEAD1 |
| 4 |  | $1x10^{-578}$ | 9.27% | NFY |
| 5 |  | $1x10^{-535}$ | 12.46% | RUNX2 |
| 6 |  | $1x10^{-504}$ | 9.05% | CEBPE |
| 7 |  | $1x10^{-383}$ | 21.30% | Elk1 |
| 8 |  | $1x10^{-271}$ | 41.32% | PB0114.1_Egr1_2 |
| 9 |  | $1x10^{-251}$ | 1.71% | GFY |
| 10 |  | $1x10^{-246}$ | 27.87% | NF1-halfsite(CTF) |
| 11 |  | $1x10^{-213}$ | 7.88% | Atf7 |
| 12 |  | $1x10^{-129}$ | 59.25% | Smad3 |
| 13 |  | $1x10^{-120}$ | 0.58% | ZBTB33 |
| 14 |  | $1x10^{-117}$ | 2.49% | NRF1 |
| 15 |  | $1x10^{-67}$ | 0.10% | Smad2 |

Table 4.5.2.2 Top 10 common biological processes that are regulated by the top three transcription factors BATF, SP2 and TEAD1. The definition of each GO process can be found on the Gene Ontology Consortium website (http://geneontology.org/). For example, regulation of localisation is a collection of processes that include the regulation of sequestering of calcium or zinc ion within a cell.

| GO biological process | p-value |
|---|---|
| Regulation of localisation | $1.35 \times 10^{-06}$ |
| Cellular process | $1.39 \times 10^{-06}$ |
| Regulation of cell motility | $1.66 \times 10^{-06}$ |
| Regulation of cellular component movement | $1.99 \times 10^{-06}$ |
| Regulation of locomotion | $2.72 \times 10^{-06}$ |
| Regulation of cellular component organization | $3.77 \times 10^{-06}$ |
| Cell development | $8.36 \times 10^{-06}$ |
| Cell projection morphogenesis | $8.54 \times 10^{-06}$ |
| Plasma membrane bounded cell projection morphogenesis | $9.18 \times 10^{-06}$ |
| Cell differentiation | $9.63 \times 10^{-06}$ |

### 4.5.3 Functional confirmation of Dupuytren associated variants

As detailed in the DD GWAS results section (chapter 2.5.5), nine out of the twenty-six 99% credible sets were predicted in possession of a biological function due to their positioning on the chromosome, within an open chromatin region. Here, I combine the physical location of these variants and ATAC-seq and RNA-seq genomic tracks of the surrounding regions that were created using DD myofibroblast cells to provide more evidence in support of the potential biological functions of the variants. Out of the 9 regions tested, 4 associated variants rs61811977 on chromosome 1(Figure 4.5.3.1), rs9691008 on chromosome 7 (Figure 4.5.3.2), rs2912522 on chromosome 8 (Figure 4.5.3.3), and rs79291578 (Figure 4.5.3.4) on chromosome 20 are situated in an enhancer region that is capable of affecting transcription of nearby genes, which are also being actively transcribed.

Further descriptive analysis was performed to establish an association between the variant rs61811977 and its nearby genes using DD myofibroblast cells (n=1) and circularised chromatin conformation capture (4C). Results showed that there are clear interactions between the region where rs61811977 is situated and the promoter region of DDR2, with much weaker interactions with the promoter of C1orf110. This indicates that the possible functional consequences of rs61811977 are likely to be a consequence of altered *DDR2* transcription (Figure 4.5.3.5).

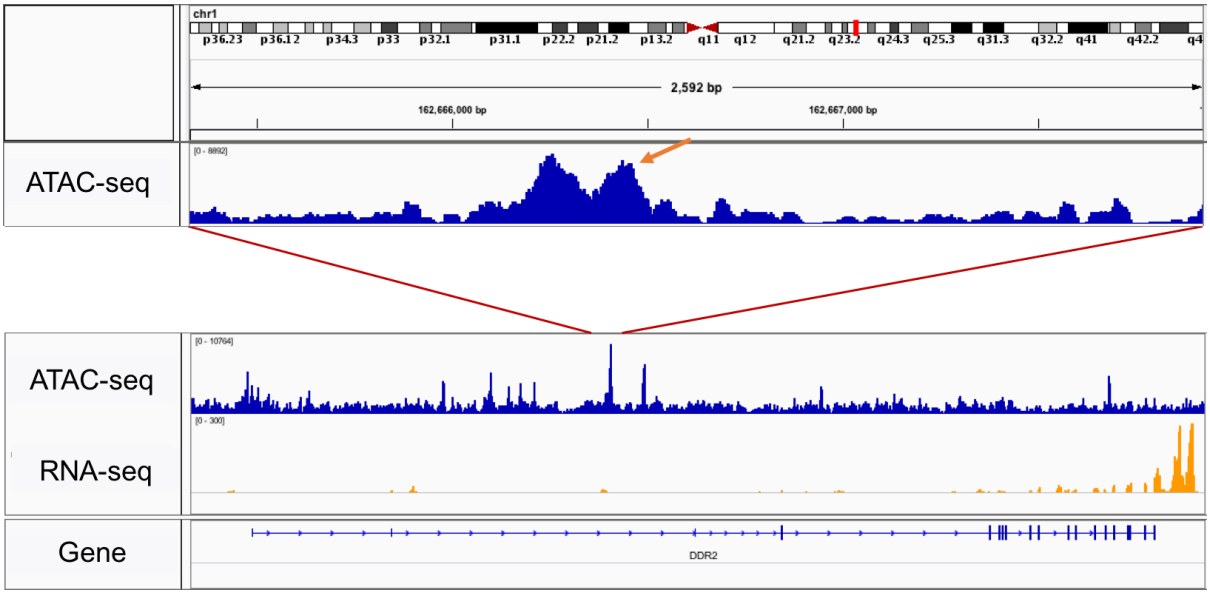



Figure 4.5.3.1 A section of chromosome one showing the genomic region surrounding gene *DDR2*, which is actively transcribed in the DD myofibroblast cells. The orange arrow indicates the location of the DD associated variant rs61811977, which is within an active enhancer region, indicating its high potential of affecting *DDR2* expression. The following description also applies to figure 4.5.3.2 – 4.5.3.4; Orange peaks in the RNA-seq track indicate mRNA transcripts, and blue peaks in the ATAC-seq tracks indicate open chromatin regions. In either instances, the higher the peaks, the higher the accessibility or the higher transcript level was detected. The top part of the figure is a close up look of the region below, which is marked by two red lines to emphasise the region of interest.

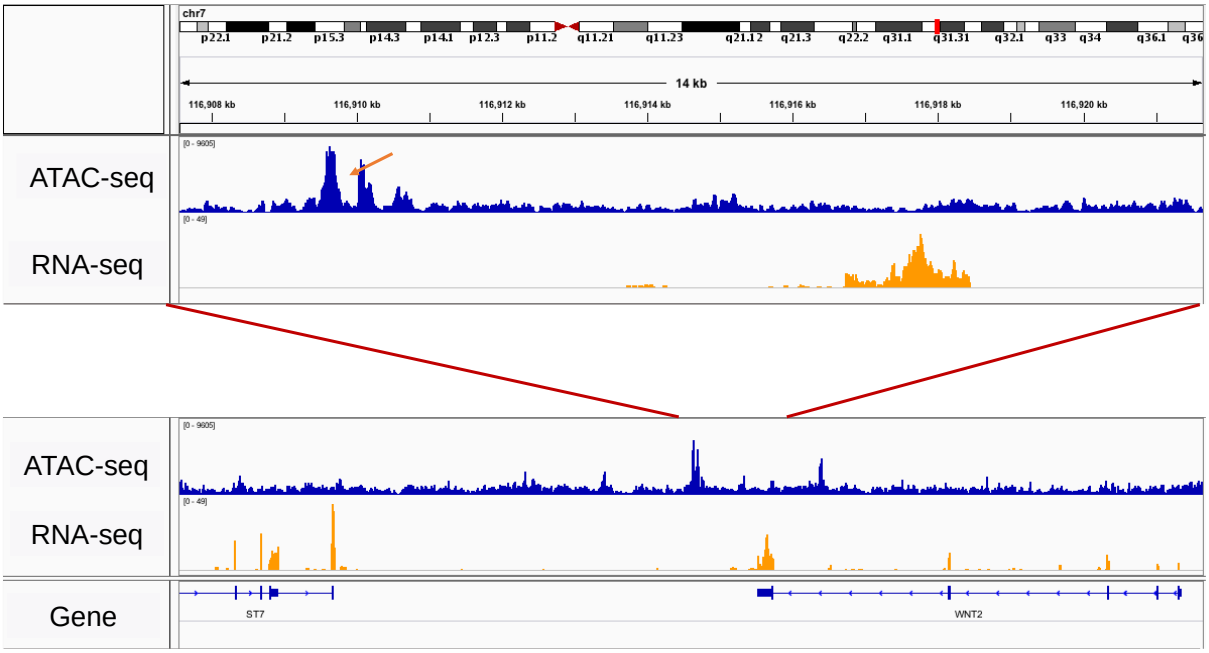



Figure 4.5.3.2 Section on chromosome 7 showing the genomic region in an intergenic region between gene *WNT2* and *ST7,* which are actively transcribed in DD myofibroblast cells. The orange arrow indicates the location of DD associated variant rs9691008, which is situated between two open chromatin enrichment peaks, indicating this region is actively bound by transcription factors.

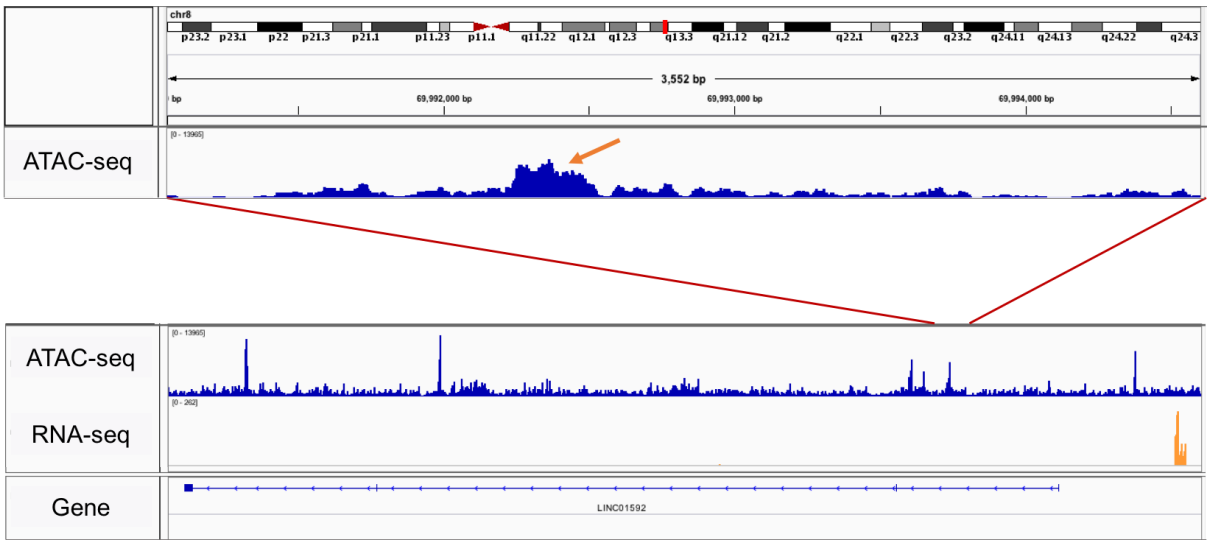



Figure 4.5.3.3 Section on chromosome 8 showing the genomic region surrounding a long intergenic non-coding RNA *LINC01592*. The orange arrow indicates the location of DD associated variant rs2912522. Although there are multiple open chromatin regions along the *LINC01592*, no *LINC01592* transcripts has been detected in the myofibroblasts. Interestingly, some RNA transcripts have been detected between 40kb and 45kb downstream but they have not yet been annotated and their function is unknown.

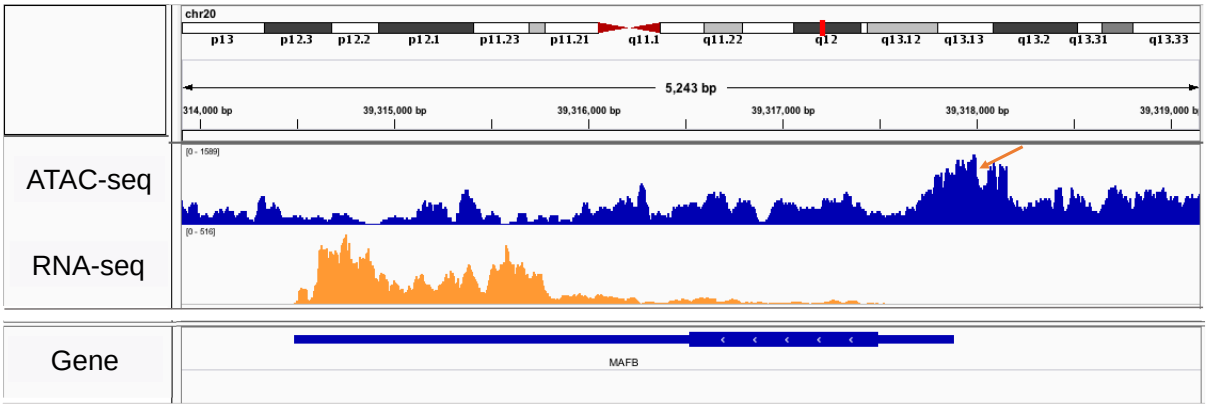



Figure 4.5.3.4 Section on chromosome 20 showing the genomic region surrounding gene *MAFB*, a transcriptional factor which is actively transcribed in DD myofibroblast cells. The orange arrow indicates the location of DD associated variant rs79291578, which is situated within the promoter region of *MAFB*.

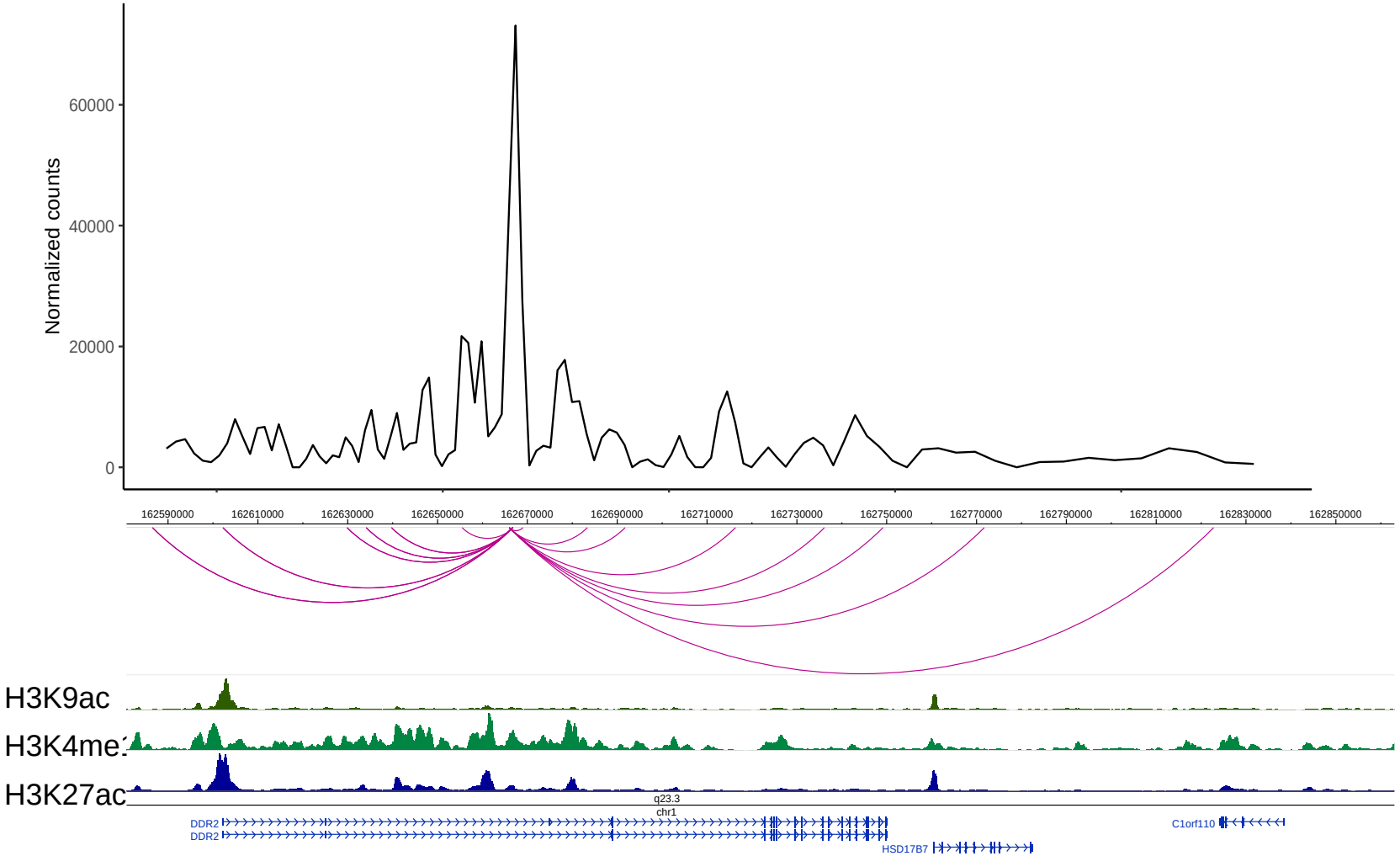



Figure 4.5.3.5 Diagram showing the interaction between the chromatin region that rs61811977 is located and cis-interaction with the nearby genes. The three CHIP-seq tracks, H3K9ac, H3K4me and H3K27ac of NHLF cells were submitted by Dr Bradley Berstein [221] for the WashU epigenome browser and were included in the diagram to indicate the locations of enhancer and promoter region.

### 4.5.4 Differential chromatin accessibility between Dupuytren myofibroblasts and fibroblasts from the palmar skin

Differential open chromatin region peak analysis was performed between Dupuytren myofibroblasts (n=10) and palmar skin fibroblasts (n=3) to identify the unique transcriptional signature of either cell type. Subsequent gene annotation on the differential peaks have revealed a total of 13,580 genes that are significantly different between Dupuytren myofibroblasts and palmar skin fibroblasts. Further analysis was performed to investigate whether the same genes were differentially expressed in the RNA-seq data, which has a total of 15 samples (DD myofibroblasts n=8; palmar fibroblasts n=7). Differential analysis using DESeq2 revealed a total of 223 differentially expressed genes, amongst which 136 also showed differences in chromatin accessibility in the ATAC-seq data (Figure 4.5.4.1). Out of these 136 genes, 81 were myofibroblast cell specific and 55 were palmar fibroblast cell specific (Figure 4.5.4.2). Some of the common genes are known to play a role in matrix remodelling and cell proliferation including *ADAMTS5*, *MMP1*, *MMP3, MMP13*, *MAFB*, *EPDR1* and *SFRP4*, where the latter three were nearby genes to the reported DD associated variants. Using *EPDR1*(Figure 4.5.4.3), *SFRP4* (Figure 4.5.4.4) *and MAFB* (Figure 4.5.4.5) as examples, a reduction in open chromatin region accessibly in the gene promoter area was observed in palmar skin ATAC-seq data compared to DD myofibroblasts ATAC-seq. In all instances, the reductions correlate well with the reduction in the corresponding transcript level as observed in the RNA-seq data.




Figure 4.5.4.1 Venn diagram showing the number of genes that showed differential chromatin accessibility in ATAC-seq, and / or are differentially expressed in RNA-seq between DD myofibroblasts and palmar skin fibroblasts.

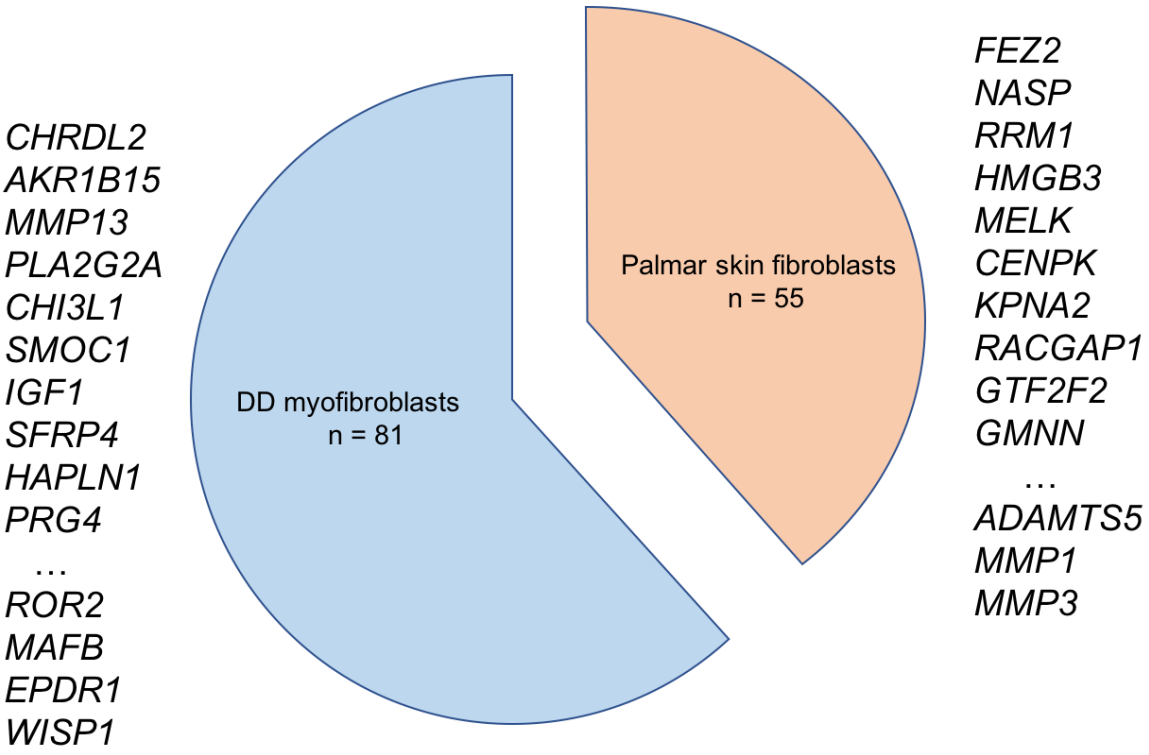



Figure 4.5.4.2 Pie chart showing the distribution of the differential genes by both open chromatin accessibility and RNA transcript level. The gene lists on either side of the diagram are the top 10 most upregulated, and biologically relevant to current knowledge in DD, genes corresponding to the closest cell group.

.

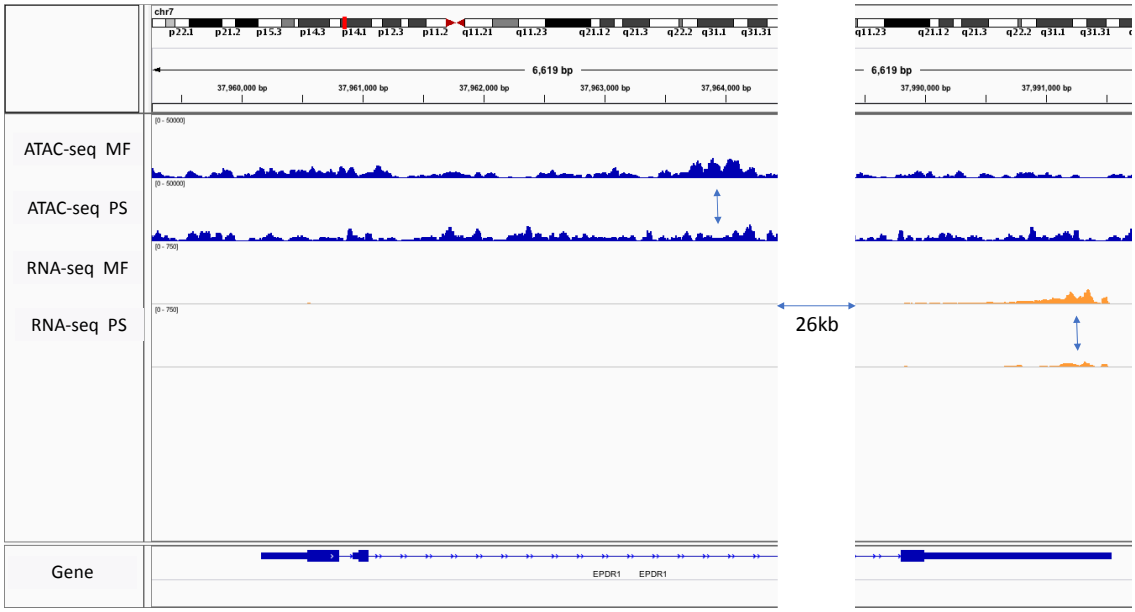



Figure 4.5.4.3 Diagram showing the differential open chromatin and RNA expression region in gene *EPDR1*. The differential regions are marked with blue two-way arrows. RNA-seq peaks were orange and ATAC-seq peaks were blue. MF is myofibroblasts and PS is palmar skin fibroblasts.

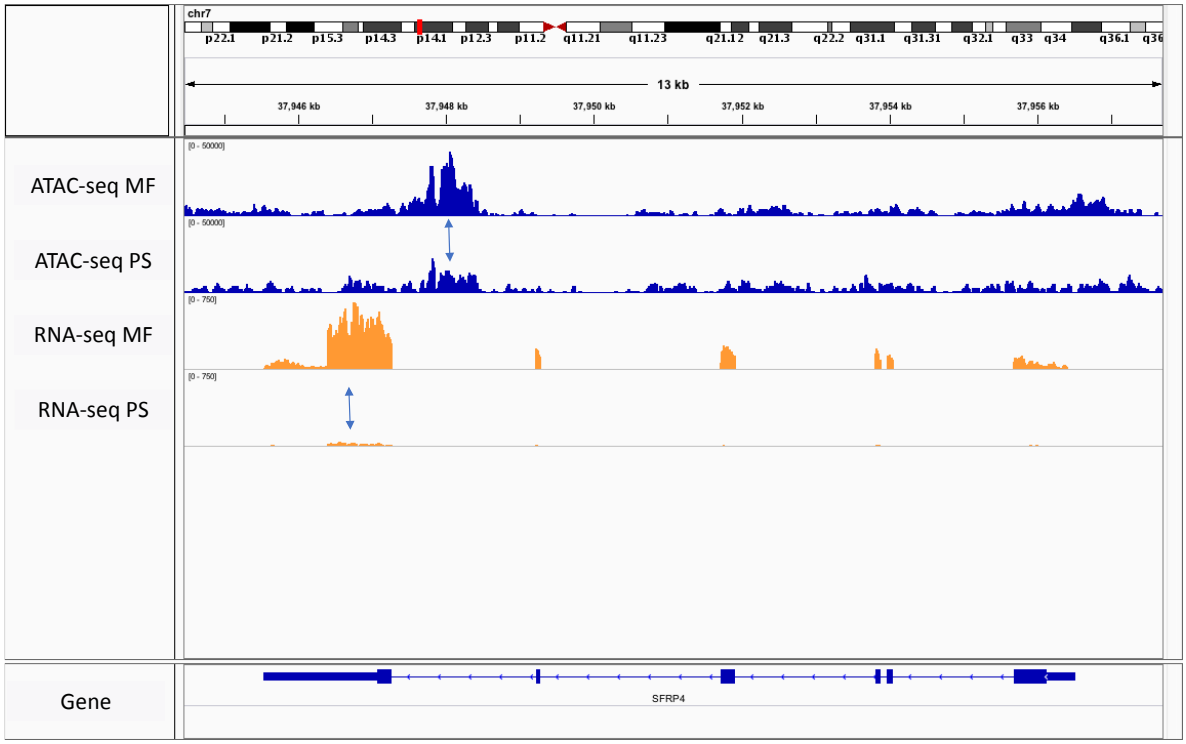



Figure 4.5.4.4 Diagram showing the differential open chromatin and RNA expression region in gene *SFRP4*. The differential regions are marked with blue two-way arrows. Peaks are annotated the same way as figure 4.5.4.3.

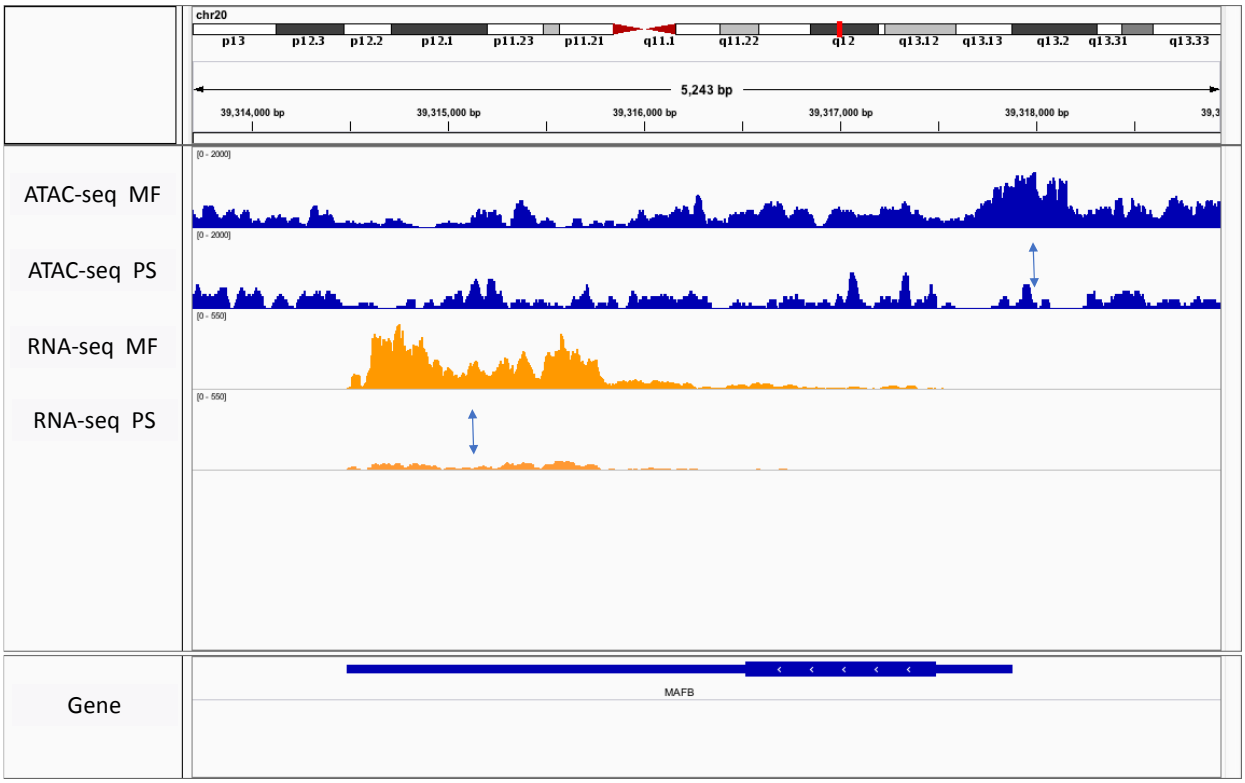



Figure 4.5.4.5 Diagram showing the differential open chromatin and RNA expression region in gene *MAFB*. The differential regions are marked with blue two-way arrows. Peaks are annotated the same way as figure 4.5.4.3.

## 4.6 Discussion

In this chapter, I generated and analysed DD cell-type specific genomic landscape data, which revealed previously unknown epigenomic regulatory mechanisms, providing new insights for the earlier DD functional work and a layer of information for designing future studies examining DD or fibrotic disease in general. Furthermore, these data also provided the unadulterated first-hand genomic information of DD cells during cell culture and in physiological status i.e. in tissue, without the noise generated from trans-effectors and post transcriptional mechanisms that transcriptomic data inherently carries [222 223]. Results of the pathway enrichment and transcription motif analysis suggested that DD cells in vitro, in our culture system, maintained to some extent their fibrotic functions such as cell contraction, motility and cell survival. This is without the presence of immune cells, suggesting either the non-plasticity of DD cells, within a limited time span, or the rigidity[224] of the tissue culture plastics [225].

The most significant enriched transcription factor binding motif found in DD myofibroblasts is BATF, which is a basic zinc leucine zipper TF belonging to the activator protein (AP-1) family. This TF lacks a transactivation domain and requires dimerization with JUN for the binding of AP1 sites[226] or with the aid of IRF4 to AP1-IRF composite elements (AICE)[227]. Although the genomic function, whether activation or repression, of BATF is variable [226], the biological function of BATF has been mainly assessed in B cells [228] and T-helper cells [229] due to its class switching and Th cells

proliferation activity. Some possible fibrosis related functions of BATF have been shown in an OA study, where it upregulated expression of *ADAMTS5*, *MMP3* and *MMP13* [230]. These are some of the genes that were also found differentially expressed between myofibroblast cells and palmar fibroblast cells in both my ATAC-seq and RNA-seq data. This suggests that the inhibition of BATF dimerization, or signalling molecules within pathways upstream of BATF translocation, could affect the profibrotic phenotype of DD cells. Future research, in particular the addition of BATF CHIP-seq data from DD myofibroblast cells, is required to support this speculation.

The integration of ATAC-seq and RNA-seq data has also provided some new insights into three genes*,* *MAFB*, *EPDR1* and *SFRP4* that are in close proximity to DD associated GWAS variants. In all three cases, I observed a positive correlation between ATAC-seq and RNA-seq data, with higher expression and open chromatin accessibility detected in the myofibroblast cells than in palmar fibroblasts. Interestingly, the variable open chromatin region in *SFRP4*, which is in the last intron of *SFRP4* is several folds higher than the variable region in *EPDR1* with very little differences in the promoter region. This difference was reflected in the RNA-seq data at their transcriptomic level and in the qPCR results reported in chapter 1, suggesting that both genes are affected by a common causal variant. However, further extensive *in silico* and functional work is required to support this speculation. Nevertheless in all three instances, the ATAC-seq data suggested that the TFs binding to either variable region were acting as activators rather than repressors [231].

Although it is well understood that myofibroblast cells are different from fibroblasts, whether it is at genomic transcript level or in biological functions, not all aspects are clear [60] and this includes the uncertainty of biomarkers. Part of the reason is that myofibroblast is a generic term for activated cells that are found in fibrotic tissue of any organ expressing currently known fibrotic genes, such as *ED-A* & *ACTA2*, with elevated collagen production and cell contraction. To further complicate this uncertainty, myofibroblasts may originate from many cell types in the nearby normal tissue as well as those in circulation [46 232], which are themselves genetically and phenotypically diverse. My data has successfully isolated a subset of genes and their corresponding regulatory regions that separates the DD myofibroblasts and palmar fibroblasts. Some ATAC-seq data has also been acquired from non-palmar skin fibroblast cells, as there is some evidence that dermal fibroblasts could harbour Dupuytren disease cells [233]. I will analyse these data in the future, when there are sufficient data points.

The pilot genomic interaction experiment has also been a success, associating the causal region pertaining to rs61811977 with promoters of *DDR2*, further justifying the necessity of investigating the molecular function of this gene. However, the 4C results also suggested the possibility of a locus control region (LCR) property, demonstrating multiple interactions with promoters of other genes downstream, including *HSD17B7*, and *CCDC190* (*C1orf110*). The function of *CCDC190* is unclear, but it is highly conserved amongst mammals. HSD17B7 is a hydroxysteroid dehydrogenase enzyme that facilitates the conversion process of estrogen to potent estradiol. The elevation of

estradiols is known to cause endometriosis, but not known in the development of DD [234], and DD is well known to be more common in males than females. Nevertheless, it is interesting that both genes are potentially regulated by the same region.

With the recent breakthrough of single-cell sequencing technologies and big data analysis methods, the focus of fibrosis research has increasing been into dissecting cell subtypes [61]. However, like many fibrosis genetics studies in the past, the assay material was transcriptomic. As mentioned earlier, RNA transcript levels are subjected to many biological processes such as the availability and binding of TFs, various post-transcription processes, etc. and therefore the link between a GWAS disease associated variant and its biological function may be complex. Whilst such single-cell sequencing methods will undoubtedly improve our understanding of the pathogenesis of fibrosis, it is imperative that other genomic landscape data is also collected at the single-cell level as techniques are developed.

# Chapter 5 Dupuytren disease genetic risk score and recurrence

## 5.1 Abstract


Introduction

DD surgery has a high recurrence rate, and repeat surgery increases the chances of complications. In this chapter, I generated weighted genetic risk score (wGRS) using the 26 associated SNPs that were discovered in my DD GWAS. High wGRS score and other high risk clinical features were tested for association with recurrence.

Materials and methods

Weighted genetic risk scores were calculated using the logOR and the number of risk alleles present at each of the genotyped associated variants. Clinical data was obtained from patient questionnaires and medical records. Association between wGRS, clinical features, and recurrence was analysed using univariate and multivariate logistic regression models, in a UK and Dutch cohort with DD.

Results

Increasing wGRS was associated with high risk clinical features including bilateral disease, age of onset under 50, ectopic disease and a positive family history ($p<0.005$). The association between recurrence and high wGRS remained significant in the regression model after adjusting for high risk clinical features ($p=0.004$).

Discussion

DD wGRS is a good independent predictor of the risk of recurrence after surgery. The applications of wGRS might include pre-surgical counselling, and could be used to make diagnostic and treatment decisions in the future.

## 5.2 Introduction

Although it would be fascinating to understand the biology behind every DD associated variant and their mechanism of action that contributes to the development of a fibrotic phenotype, prioritising functional work is not the sole application of GWAS summary statistics. One of the most frequently employed analyses by clinicians and epidemiologists is the calculation of genetic risk score (GRS), which is a method of predicting the likelihood of disease using genotype data [235]. GRS can be applied to almost any results from GWAS as it doesn't require a high-resolution dataset based on the premise that the associated variants are in linkage equilibrium with the causal variants. GRS can be weighted, by the total count of variants, and non-weighted, where both calculated by aggregating the number of risk alleles that a subject carries at each disease associated variant multiplying by the magnitude and direction of their effect. Although both methods convey similar outcome, the weighted GRS (wGRS) is more commonly used as it is more likely to confer better correlation with the susceptibility of an individual to developing a condition. GRS calculations have been employed in many complex disease studies, with some showing promising results, successfully segregating the genetic predictability between type I and type II diabetes, and between monogenic and type I diabetes [236 237], although others have only shown modest predictability or clinical usefulness. Although possessing a small heritability and a modest amount of associations contributing to polygenicity, the current set of DD associated variants were hypothesised to be predictors for disease recurrence - the recurrence rate of DD after surgical treatment is high, ranging from 21% to 85% at

5 years [238]. Other clinical features, or DD diatheses, have also been identified to be the predictors of DD recurrence, including age of onset under 50 years, bilateral disease, ectopic disease beyond the hand (including Ledderhose and Peyronie), and positive family history of DD [239]. These clinical features could be themselves be predicated by the wGRS or could confound the ability of the wGRS to predict DD recurrence. Because of high revision surgical complication rates [240], there is a need for a reliable DD prognostic indicator, which currently can be utilised for pre-surgical counselling, and in the future, may be used to select specific treatments for individual patients.

## 5.3 Aims and hypotheses

The aims of this chapter were two-fold; firstly, to calculate wGRS using the available genotype resources that were created in our GWAS, comprising two cohorts BSSH-GODD and GODDAF. Secondly to find a relationship between wGRS and the aforementioned diatheses, wGRS, and DD disease recurrence, and between diatheses and DD disease recurrence.

I hypothesised that wGRS is a good predictor of DD recurrence and would be useful for pre-surgical counselling, and treatment decision making when more effective treatments and prevention methods have been developed.

## 5.4 Materials and methods

### 5.4.1 Patient cohorts, clinical data and ethical approval

Two patient populations were used in the analyses; the UK cohort, British Society for Surgery of the Hand Genetics of Dupuytren's Disease (BSSH-GODD) and the Dutch cohort, Genetic Origin of Dupuytren Disease (GODDAF)[29]. Both studies were approved by the Research Ethic Committee or the equivalent, and consent was obtained from all participants. All selected cases must have had at least one surgical operation for DD. Patient clinical data of the BSSH-GODD study was extracted from the questionnaires that were completed by the participants containing information such as presence of ectopic fibrosis, bilateral disease, age of onset under 50 years, positive family history and the number of operations undertaken. Similar information was extracted from medical notes in the Dutch cohort. In both cohorts, participants were defined as recurrent cases if the number of operations undertaken on the same hand was greater than one. Patients that had been diagnosed a recurrence by the treating surgeons in the Dutch cohort were also included.

### 5.4.2 Genetic risk score

Genetic risk score of each patient was calculated using the logOR and the number of risk alleles that were reported in the previously reported summary statistics of the DD GWAS [241]. wGRS of each participant was calculated using the following formula,

$$\frac{26}{n} \times \sum_{i=1}^{26} W_i X_i$$

where i is the SNP, n is the number of SNPs that are included in the calculation, W is weight or log(OR) and X is the number of risk alleles.

### 5.4.3 Imputation of the missing diatheses

Due to the presence of missing data points in clinical features and wGRS, imputation was performed using multiple imputation by chained equation (MICE) [242]. 48 samples were removed prior to imputation as they violated the missing completely at random (MCAR) assumption because only one SNP was genotyped. MICE was performed on all aforementioned clinical features, which included family history, early onset, bilateral disease, recurrence, sex, and wGRS, with the addition of ectopic disease only in GODDAF as none were missing in GODD. To ensure the quality of the imputation, 75

imputed datasets with each dataset having undergone 100 iterations were created. Visual inspection of imputed datasets revealed good convergence with narrow variance, indicating reliable imputation.

### 5.4.4 Statistical analysis

Statistical analysis was performed using Stata MP14 and R unless stated otherwise. Manual visualisation of the quantile-quantile plot and Levene's test were employed to ensure the normality and the equality of variances in the wGRS data before statistical analysis was performed. Two-tailed student t-test was employed to compare the mean wGRS between males and females, and cases and controls. A p value less than 0.05 was deemed statistically significant.

wGRS data was tested as a continuous variable and as categories, which were created using three thresholds; the mean of the wGRS, and mean plus and minus one standard deviation.

Associations between continuous wGRS or each DD diathesis and recurrence were tested using univariate logistic regression. Multiple logistic regression was used to test associations between categorised wGRS, clinical diatheses and recurrence. Each analysis was performed separately in each cohort, and the coefficients were

subsequently meta-analysed using fixed effect inverse variance method. A p value less than 0.05 was considered significant.

## 5.5 Results

### 5.5.1 The demographic of diatheses in DD patients

The occurrence of each diathesis was similar between BSSH-GODD and GODDAF (Table 5.5.1.1). The ratio between male to female was approximately three to one. The most commonly occurring diathesis feature was bilateral disease, which was reported in 63% of BSSH-GODD and 40% of GODDAF. The least commonly occurring diathesis feature was ectopic disease, which was reported in 11.7% of BSSH-GODD and 12.7% of GODDAF. The number of SNPs contributing to the wGRS calculation was primarily 26 in the BSSH-GODD and 22 in GODDAF due to a logistical limitation at the time of genotyping the GODDAF samples. Both cohorts demonstrated approximately 20% of missingness (Table 5.5.1.2).

Plotting observed against expected wGRS demonstrated a Gaussian distribution in both the BSSH-GODD and GODDAF datasets (Figure 5.5.1.1). The spread of wGRS between BSSH-GODD and GODDAF were almost identical, ranging from 3.48 to 9.76 and 3.00 to 9.77 in BSSH-GODD and GODDAF respectively. In BSSH-GODD, a significant difference in mean wGRS was observed between DD patients with and without recurrence (6.32 vs 5.44; $p < 0.001$), as well as between females and males (6.42 vs 6.29; $p<0.001$).

Table 5.5.1.1 Table showing the distribution of wGRS and DD diatheses in GODD and GODDAF study.

| Characteristic | BSSH-GODD (n=6126) | GODDAF (n=1730) |
|---|---|---|
| Male (%) | 4797 (78.3%) | 1311 (75.8%) |
| Missing (%) | 25 (0.4%) | 3 (0.2%) |
| Bilateral disease (%) | 3860 (63%) | 692 (40%) |
| Missing (%) | 36 (0.6%) | 606 (35%) |
| Early age of onset (%) | 1553 (25.3%) | 627 (36.2%) |
| Missing (%) | 1140 (18.6%) | 126 (7.3%) |
| Ectopic disease (%) | 719 (11.7%) | 219 (12.7%) |
| Missing (%) | 0 (0%) | 618 (35.7%) |
| Positive Family history (%) | 2391 (39%) | 846 (48.9%) |
| Missing (%) | 36 (0.6%) | 447 (25.8%) |
| Recurrence (%) | 1055 (17.2%) | 239 (13.8%) |
| Missing (%) | 2682 (43.8%) | 1007 (58.2%) |

Table 5.5.1.2 Table showing the distribution of SNP frequency that was used as part of the wGRS calculation process in BSSH-GODD and GODDAF.

| | **Frequency** | |
|---|---|---|
| **SNP Count** | **BSSH-GODD** | **GODDAF** |
| 19 | 5 | 1 |
| 20 | 3 | 9 |
| 21 | 8 | 114 |
| 22 | 61 | 1336 |
| 23 | 1060 | - |
| 24 | 2 | - |
| 25 | 58 | - |
| 26 | 3808 | - |
| Missing | 1121 | 270 |
| **Total** | **6126** | **1730** |

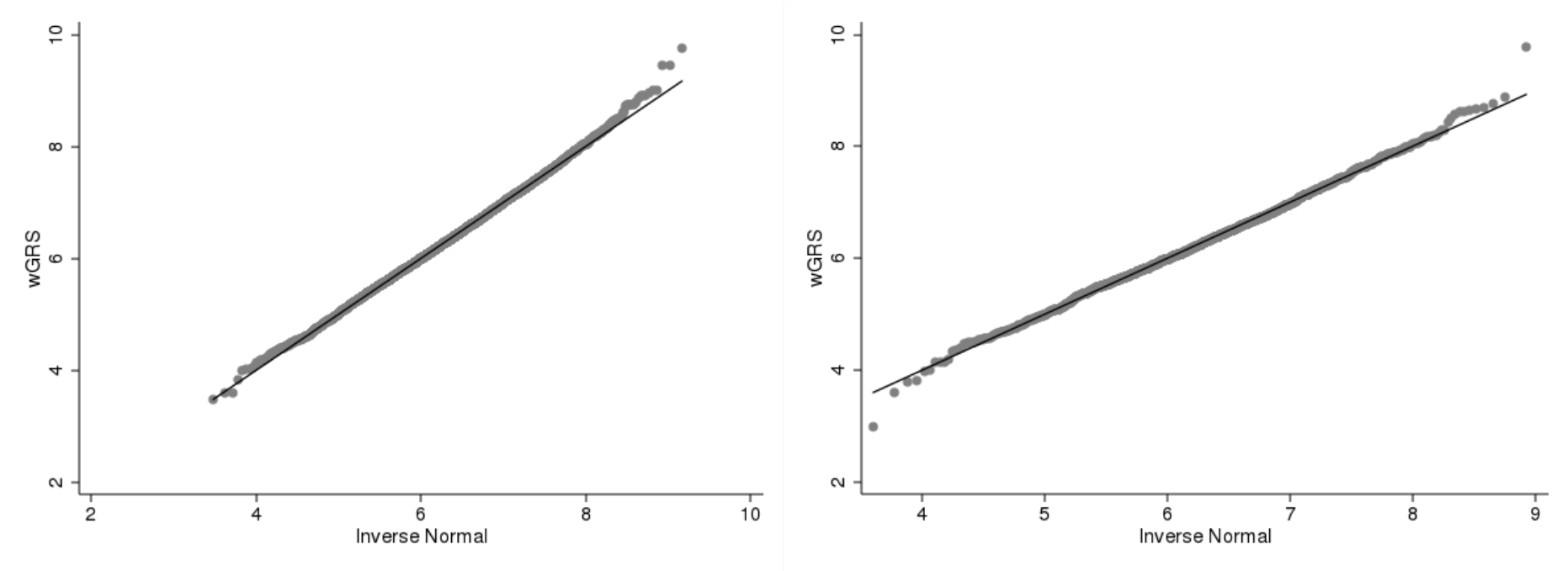



Figure 5.5.1.1 Quantile-quantile plot showing the Gaussian distribution of the left) BSSH-GODD and right) GODDAF wGRS data.

### 5.5.2 Weighted GRS can predict DD diatheses

The association between wGRS and DD diatheses was examined using wGRS as a continuous variable and also as categories one to four. In the multivariate regression model, all diatheses demonstrated significant association with the top two wGRS groups, with the exception of age of onset under 50 where only the highest wGRS showed significance (Table 5.5.2.1). Testing the association using wGRS as a continuous variable demonstrated a similar result, with bilateral disease, ectopic disease and positive family history demonstrating an association (Table 5.5.2.2).

Table 5.5.2.1 Table showing the associations between DD diatheses and categories of wGRS.

| Diathesis | wGRS Group | OR (95% CI) | p value |
|---|---|---|---|
| Bilateral disease | Base | | |
| | 2 | 1.26 (1.07-1.50) | 0.007 |
| | 3 | 1.45 (1.22-1.73) | <0.001 |
| | 4 | 1.78 (1.44-2.19) | <0.001 |
| Early age of onset | Base | | |
| | 2 | 1.11 (0.92-1.35) | 0.274 |
| | 3 | 1.21 (1.00-1.47) | 0.059 |
| | 4 | 1.38 (1.10-1.73) | 0.006 |
| Ectopic disease | Base | | |
| | 2 | 1.20 (0.90-1.59) | 0.220 |
| | 3 | 1.52 (1.14-2.01) | 0.004 |
| | 4 | 1.99 (1.46-2.70) | <0.001 |
| Positive family history | Base | | |
| | 2 | 1.14 (0.96-1.35) | 0.137 |
| | 3 | 1.29 (1.08-1.53) | 0.004 |
| | 4 | 1.38 (1.13-1.68) | 0.002 |

Table 5.5.2.2 Table showing the associations between wGRS, as a continuous variable, and DD diatheses.

| Diathesis | OR (95% CI) | p value |
|---|---|---|
| Bilateral Disease | 1.26 (1.17-1.35) | <0.001 |
| Early Age of Onset | 1.13 (1.04-1.23) | 0.004 |
| Ectopic Disease | 1.31 (1.18-1.46) | <0.001 |
| Positive Family History | 1.14 (1.07-1.22) | <0.001 |

## 5.5.3 Weighted GRS and DD diatheses can predict DD recurrence

Each DD diatheses feature demonstrated an increased odds for recurrence requiring further surgical intervention (Table 5.5.3.1). In the association analysis between wGRS and recurrence, the top two groups of wGRS demonstrated an increase risk in both BSSH-GODD and GODDAF, demonstrating a maximum of odds ratio at 1.69 ($p<0.001$) and 1.88 ($p=0.019$) respectively. When the same analysis was repeated with the previously identified significant diatheses as covariates, the risk of wGRS group 3 and 4 experiencing recurrence remained significant in BSSH-GODD, demonstrating an odds ratio of 1.35 ($p=0.009$) and 1.46 ($p=0.008$) but failed to be replicated in GODDAF, most likely due to small sample size (Table 5.5.3.1). This association pattern was replicated when modelling using univariate logistic regression

and wGRS as a continuous variable, showing a moderate risk of recurrence for each unit increase in wGRS (Table 5.5.3.2). Meta-analyses (Table 5.5.3.2) yielded statistical significance in both DD diatheses unadjusted (OR 1.24; 95% CI 1.12-1.39; p<0.001) and adjusted analysis (OR 1.17; 95% CI 1.05-1.31; p=0.004).

Table 5.5.3.1 The risk of DD recurrence by categorical wGRS in BSSH-GODD and GODDAF. DD diatheses including age of onset under 50, bilateral disease, ectopic disease and positive family history were used as covariates in the adjusted model.

| Cohort | wGRS category | Unadjusted | | Adjusted | |
|---|---|---|---|---|---|
| | | OR (95% CI) | p value | OR (95% CI) | p value |
| BSSH-GODD | Base | | | | |
| | 2 | 1.23 (1.00-1.53) | 0.06 | 1.16 (0.93-1.46) | 0.177 |
| | 3 | 1.48 (1.17-1.86) | 0.001 | 1.35 (1.06-1.71) | 0.009 |
| | 4 | 1.69 (1.28-2.23) | <0.001 | 1.46 (1.09-1.96) | 0.008 |
| | | | | | |
| GODDAF | Base | | | | |
| | 2 | 1.43 (0.92-2.22) | 0.107 | 1.31 (0.82-2.09) | 0.257 |
| | 3 | 1.50 (0.94-2.42) | 0.09 | 1.09 (0.80-2.18) | 0.276 |
| | 4 | 1.88 (1.11-3.17) | 0.019 | 1.58 (0.90-2.77) | 0.111 |

Table 5.5.3.2 The association between wGRS, as a continuous variable, and recurrence in BSSH-GODD, GODDAF, and after meta-analysis. Adjustment was performed for diathesis factors as described in table 5.5.3.1.

| | Unadjusted | | Adjusted | |
|---|---|---|---|---|
| Cohort | OR (95% CI) | p value | OR (95% CI) | p value |
| BSSH-GODD | 1.25 (1.13-1.38) | <0.001 | 1.18 (1.06-1.31) | 0.002 |
| GODDAF | 1.22 (1.01-1.46) | 0.035 | 1.14 (0.94-1.39) | 0.194 |
| Meta-analysis | 1.24 (1.12-1.39) | <0.001 | 1.17 (1.05-1.31) | 0.004 |

## 5.6 Discussion

In this chapter, I have successfully demonstrated that DD wGRS can predict recurrence of DD after surgery. This prediction was independent of any of the clinical DD diathesis features that were included in the study, namely age of onset under 50, bilateral disease, ecoptic disease and positive family history. These results also extended the list of diatheses features predicted by wGRS to also include bilateral disease [11], which was primarily due to an increase in power of the contributing DD GWAS dataset [241].

There were some differences in definition of recurrence between the participating cohorts, BSSH-GODD and GODDAF. In BSSH-GODD, recurrence was defined as patients with more than one operation on the same hand, whereas in GODDAF the inclusion criteria were more lenient by also including patents that were given a clinical diagnosis of recurrence by the treating surgeon. As a result, the GODDAF study may have included patients that have declined surgical treatment, which in turn could increase the number of cases classified as recurrent, but may have also increased the chances of misclassification bias, if the criteria were not strictly adhered to by the diagnosing surgeon. Furthermore, in the BSSH-GODD study, some patients may have had separate operations on different fingers of the same hand, and therefore have been clasified as having recurrence, when in reality their disease had extended. Nevertheless, despite different definitions of recurrence, the distribution of all the

diatheses, gender of DD cases, and the wGRS itself were similar between the two cohorts.

There were some methodological limitations of this study. Firstly, I didn't observe a significant association between recurrence and wGRS in the diathesis adjusted model in the GODDAF cohort. This was primarily due to its small sample size and the high counts of missing data, though overall it didn't affect the final outcome as the magnitude and direction of log odds were similar to the BSSH-GODD. Secondly, the gene set that was used in this study was tailored to predict DD susceptibly as a whole, not recurrence *per se*. Despite the high recurrence rate of DD, some included SNPs might well not be associated with the recurrence of DD at all.

In conclusion, I have succesfully demonstrated that the wGRS predicts the presence of all known clinical diathesis features, extending previous work to include bilateral disease. Importantly, I have also demonstrated that increasing wGRS is associated with recurrence after surgery for DD, even when adjusting for these well known clinical predictors of recurrence. The usefullness of this information is currently limited to pre-surgical counselling, however in the future it may be useful to stratify patients for trials of novel therapeutics or surgical interventions aimed at reducing recurrence.

# Chapter 6 The shared genetic epidemiology of fibrotic diseases

## 6.1 Abstract


Introduction

Fibroses are complex diseases that affect millions in the UK alone. Untreated chronic fibrosis can lead to poor prognosis, such as reduction in organ function, cancer and death. Despite the similarities between most fibrotic conditions, little effort has been made to identify the common aetiology. Using genetic associations, we aimed to identify causal SNPs across multiple fibrotic phenotypes and reveal common targetable pathways. Here we examined the correlations between adhesive capsulitis (AC), Dupuytren disease (DD), endometriosis (Endo), hepatic fibrosis (HF), pulmonary fibrosis (PF) systemic sclerosis (SSc), and uterine fibroids (UF).

Materials and methods

We first conducted a phenotypic correlation analysis across seven fibrotic conditions using the HES data from participants of the UKBiobank cohort. Genetic association analysis of each condition was then performed on the UKBiobank imputed genetic

data using BOLT-LMM. Meta-analysis was conducted using additional published GWAS summary statistics to increase statistical power. Two-stage cross phenotypic genetic association analysis was performed to identify cross phenotypic associated SNPs using LD score regression and CPASSOC.

## Results

Novel associations have been identified in adhesive capsulitis, Dupuytren disease and endometriosis GWAS. LD score regression analysis showed significant correlations between endometriosis and uterine fibroids. Further genetic correlation analysis has revealed 12 cross-phenotypic associations. These cross phenotypic associations were found between AC and DD, DD and PF, Endo and UF, Endo and PF.

## Discussion

Multiple associations have been identified that predispose to more than one fibrotic condition, strongly indicating a common pathway amongst them. Function of the associated genes included ECM and hormonal sensing receptors, MHCII vesicle trafficking molecules, and cytoskeleton nuclear lamina linking proteins. These proteins are likely to have a molecular role in fibrosis development.

## 6.2 Introduction

Besides Dupuytren disease, fibroses have also been diagnosed in a comprehensive list of other organs, which include eyes, liver, lung, kidney, shoulder and uterus. The aetiology of most fibrotic conditions is poorly understood but all have a similar developmental process and that can be divided into three stages, initiator, persistent inflammation, and abnormal fibrotic resolution and repair process. An initiator can be any of the following factors including autoimmune, injury, infection, idiopathic, toxin and radiation. In addition, genetic and environmental factors are both required to initiate and perpetuate a fibrotic disease. Similar molecular components, and cell differentiation processes have also been identified in different fibrotic conditions, including epithelial or endothelial to mesenchymal transition (EMT), elevated myofibroblast counts, excessive production of collagen I and III, pro-inflammatory and pro-fibrotic cytokines [40] in the involved tissue.

Despite the similarities, treatment methods for different fibrotic conditions vary enormously, ranging from surgical intervention [243] , anti-fibrotic, anti-inflammatory and anti-hormonal medication [244 245] to UV therapy[246], and many conditions lack an effective treatment. And thus a genetics study of multiple fibrotic diseases to attempt to identify common pathways that might lead to new treatments is therefore warranted. The identification of common associations between fibrotic conditions may reveal some novel fundamental fibrosing developmental mechanism, which could lead to an

improvement in treatment and possibly the development of a preventative strategy. Here, I briefly describe the aetiology and epidemiology of six fibrotic conditions: adhesive capsulitis, endometriosis, hepatic fibrosis, pulmonary fibrosis, systemic sclerosis and uterine fibroids.

### 6.2.1 Adhesive capsulitis

Adhesive capsulitis (AC), or frozen shoulder, is a fibrotic condition in the shoulder, and its clinical presentation is thickening in the capsule of the glenohumeral joint. This often results in contraction of the capsule and restriction in motion of the affected joint and arm. It is also acutely painful [247]. The prevalence rate is around 10% in the UK, and it mainly affects patients between age 40 and 65 [248]. Interestingly, the fibrosis tends to be self-limiting, with resolution within 36 months, though symptoms can be persistent [247]. Recurrence is uncommon[249]. It has been associated with both type I and II Diabetes mellitus (DM), manifesting in up to 30% DM patients [250]. Synovial inflammation has been suggested to be the initiator of this condition. Similar to DD, an increase in type-III collagen disposition has been found in the affected area, which in the case of AC is the anterosuperior capsule[251]. Treatments include shoulder distention and manipulation, or open surgical release. No known genotyping analysis has been performed on this disease to date.

### 6.2.2 Endometriosis

Endometriosis is a fibrotic condition that is characterised by the deposition of endometrial lining cells in any organs outside the uterus [252]. Patients can suffer from severe dysmenorrhea, dysuria, dyschezia, and intermenstrual bleeding. The lifetime risk is around 10% -15% of woman, and it mainly presents between 18 – 45 years of age. Endometriosis is staged clinically between stage 1 (mild) and stage 4 (severe). Fibrosis is a prominent feature of more advanced stage 3 or 4 disease [252]. Genetic and environmental factors are again important, and tall height, low BMI, and heavy alcohol and caffeine intake have been shown to associate with this disease. Retrograde menstruation and autoimmune disorder have been suggested to be its aetiology. GWAS have previously discovered 19 loci predisposing to endometriosis [253]. Analysis of the loci suggests genes such as estrogen receptor one (*ESR1*) [253] of the steroid hormone pathway, and *WNT4* of the WNT signalling pathway [254] may have a role in its development. Treatments of for this disease endometriosis included hormonal therapy and, in more severe cases, laparoscopic surgery.

### 6.2.3 Hepatic fibrosis

Hepatic fibrosis is fibrosis in the liver, which manifestation is due to chronic injury and excessive deposition of disorganised extracellular matrix. When untreated, this condition can progress into cirrhosis and hepatocellular carcinoma, resulting in loss of

organ function, and ultimately death. A wide variety of factors can initiate hepatic fibrosis, including alcohol, viral infection, (e.g. hepatitis C), autoimmune, cholestatic and metabolic disorders (e.g. obesity), meaning that it is a heterogeneous disorder. Genetic polymorphisms, alcohol consumption and obesity are the factors contributing to the development of this disease[255]. The prevalence rate ranges from 5%-10 % [256], and the affected individuals are commonly 40 years of age and older. The best treatment is to eliminate the initiating factor, for example by stopping alcohol consumption, eradicating hepatitis virus infection, or massive weight loss. Medical intervention for end stage disease lacks efficacy, meaning that patients often require liver transplantation. However, though the use of anti-inflammatory and anti-fibrotic drugs such as multi-tyrosine kinase and TGF-β inhibitors has shown promising results [257]. The mechanism of actions of these anti-fibrotic drugs primarily involves inhibiting the differentiation of myofibroblast and proliferation of hepatic stellate cells. Host genetics plays a major role in the development of hepatic fibrosis. GWAS have collectively identified 5 loci predisposing to hepatic fibrosis [258-260] , and follow-up functional studies have suggested that an increase in *PNPLA3* and *RNF7* are protective against liver fibrosis due to its retinol release, and the antioxidant utility action of the protein [261].

### 6.2.4 Idiopathic pulmonary fibrosis

Idiopathic pulmonary fibrosis (IPF) is a type of interstitial lung disease. The prevalence of IPF in the UK is moderate, reported at approximately 38.2 per 100000 people, with an incidence rate increasing with age. The mortality rate is high with five-year and 10-year survival rates at 34% and 19% respectively[262]. It is more commonly diagnosed in men than in women. Clinical presentation of IPF includes dyspnea, fine crackles in the lower posterior of the lung and clubbed fingers[263]. The aetiology of IPF is poorly understood and treatment options are limited. There is much debate whether inflammation is involved in the initial developmental process, where some researchers have suggested that the activation of fibroblasts and myofibroblasts is due to an aberrant activation event in the lung epithelium[264]. Nevertheless, this infiltration of myofibroblast cells elevates the expression and deposition of ECM in the alveoli, leading to lung scarring and a reduction in organ function. Genetic and environmental factors have been associated with IPF, and to date a total of 17 independent associated loci have been reported from GWAS. The most recently identified associated variant is located on chromosome 15 in a region near genes *AKAP13* and *KLHL25*. Follow-up functional work in this study has demonstrated the risk allele A of rs62025270 is associated with an increase in *AKAP13* expression in lung tissue. AKAP13 has been previously demonstrated to play a role in a pro-fibrotic signalling pathway [265].

### 6.2.5 Systemic sclerosis

Systemic sclerosis (SSc), or scleroderma, is a rare, high mortality rate, fibrotic condition with a prevalence rate around 13 per million in men and 48 per million in women in the UK[266]. The onset of disease is mainly between 40 – 50 years of age. Clinical symptoms can be manifested in subcutaneous skin, and any internal organs, including lung, heart, joints, where patients often develop decreased tissue perfusion[267]. Chemical, toxin, and viral infection mediated autoimmune response against endothelial cells have all been demonstrated to be the initiators of SSc[268]. Pathologically, chronic inflammation of endothelial cells is one of the key events in the development of SSc, driving endothelial to mesenchymal transition (EMT). There is no effective treatment for SSc to date, and current regiments using corticosteroids and anti-fibrotic drugs lack efficacy[269]. In terms of mode of inheritance, a familial genetic study has demonstrated that first degree relatives and siblings of individual with SSc carry a risk around 15-fold higher than to those without [270]. An array of genes, including *IL6, STAT4 , TLR2* and *TNFAIP3* have been identified from GWAS studies and some follow up experiments pertaining to *TNFAIP3* identified its functions in promoting *MMP1* and *MMP3* expression as being critical in pathogenesis [271].

### 6.2.6 Uterine fibroids

Uterine fibroid (UF), or Leiomyoma, is a non-malignant smooth muscle tumour in the uterus, with a dense collagen rich matrix. The lifetime risk of developing UF is between 5.4% - 77% depending on the disease definition and population studied, and is higher in black women than in Caucasian women [272]. The most common age of onset is between 18 and up to 55 [272]. Similar to endometriosis patients, UF patients also develop symptoms such as dysmenorrhea, dyspareunia, pelvic pain, and have high rate of miscarriage and infertility. Lifestyle choices, environmental and genetic components have been associated with the development of UF, where low vitamin D level, obesity and those with a positive family history than those without a proband[272] all have been shown with to have an elevated risk [272]. Treatments include over the counter pain killers, anti-hormonal medication and for more severe cases, surgery such as hysterectomy and myomectomy. Pathologically, UF tissue has elevated expression estrogen receptors and progesterone receptors [273], suggesting the role of hormones in fibrosis development. Linkage and cytogenetic studies in uterine fibroid tumours have identified multiple causal regions, including a trisomy on chromosome 12, a translocation between chromosome 12 and 14 and a deletion on chromosome 3 and 7 [274]. To date, GWAS have revealed five UF associated loci from three studies with participants of different ethnicity [275-277]. Interesting potential genes include *SLK*, which is involved in myogenic differentiation and regulation of extracellular matrix deposition.

Despite the similarity between fibrotic conditions at the phenotypic level, little research effort has been made to identify their common genetic predisposition. With the aid of the recently published UKBiobank genetic dataset of more than 500,000 participants and multiple specific GWAS datasets, we aimed to identify their cross disease genetic and phenotypic correlation.

## 6.3 Aims and hypotheses

Understanding the differences and similarities between various fibrotic conditions has long been warranted in fibrotic research. However, we didn't have the resources until recent times. My aim was to identify common causal variants, and their predicted functions across multiple fibrotic conditions: adhesive capsulitis, Dupuytren disease, endometriosis, hepatic fibrosis, uterine fibroids and systemic sclerosis. These variants will be the foundation of further functional work beyond the scope of this thesis.

My hypotheses were two fold. Firstly, I hypothesised that there would be multiple pairwise phenotypic associations across all the fibrotic conditions that have been analysed – that is patients with one fibrotic condition were more likely than chance to suffer from another fibrotic condition. Secondly, I also hypothesised that this phenotypic association would be the result of shared genotypic predisposition. These genetic associations would most likely be intronic or intergenic, and reside near or in genes that have a role in developmental, inflammatory and ECM remodelling signalling pathways. Genetic correlation between endometriosis and uterine fibroids was predicted to be more significant than the other fibrotic conditions due to the similar organ involvement between the two conditions.

## 6.4 Materials and methods

### 6.4.1 Quality control and association analysis

Genetic association analysis was performed using the UKBiobank dataset, comprising imputed data and individual patient records, which was obtained directly from the UKBiobank data showcase. Patients that were diagnosed with adhesive capsulitis, hepatic fibrosis, pulmonary fibrosis and systemic sclerosis were identified using a list of identifier codes including ICD-10, OPCS, and self-reported illness and surgery (Table 6.5.1.1), which was assembled by expert consensus from clinically qualified specialists within our group. The summary statistics of endometriosis and uterine fibroids GWAS were provided by our collaborators Dr Nilufer Ramioglu and Dr James Cook. Association analysis was performed on each disease separately after the genotyping data had undergone quality control. This was performed using the following filtering thresholds which were provided by Dr Akira Wiberg; sample and SNPs with less than 98% call rate, heterozygosity, HWE, MAC equals one outliers, ethnically not Caucasian and sex mismatch. Association analysis was performed using linear mixed model with non-infinitesimal modelling (BOLT-LMM v2.3) [278], with sex and genotyping batch as covariates. Conditional analysis was performed using the same parameters with the addition of lead SNP allelic dosage as the third covariate. SNPs with MAF less than $1 \times 10^{-3}$ and an info score less than 0.9 (for imputed SNPs) were excluded from the test. Attenuation ratio was calculated using

LDSC intercept and was used to calculate the amount of genomic inflation factor contributing to polygenicity. Manhattan plots were created for the results of each fibrotic association analysis.

SNPs that are part of the 1KG panel in the dataset were not included in the final analysis due to a position numbering error by the UKBiobank. Quality control was performed on the summary statistics, including those from our collaborators, to exclude any SNPs with a standard error greater than five, MAC less than three and an info score less than 0.4 from subsequent analyses. A linear model was used for performing the association analysis, betas and SEs were converted into log odds scale before being utilised in the meta-analysis. Heritability of each condition was estimated using LDSC.

### 6.4.2 Phenotypic correlation

Phenotypic correlation was performed using R. Utilising the Hospital Episode Statistics (HES) data provided in the UKBiobank showcase, patients reported or diagnosed with more than one fibrotic condition were identified. Analysis was performed using chi-square test with one degree of freedom. Multiple testing was accounted for using Bonferroni correction, where a p less than $5 \times 10^{-3}$ was deemed statistically significant. The analysis was performed twice, including females only and then the entire cohort to test for the possible gender bias of the endometriosis and uterine fibroid cohorts.

### 6.4.3 Meta-analysis

Meta-analyses were performed using summary statistics of the UKBiobank and previously published GWAS datasets that were available to our research group. This included one systemic sclerosis [279], one Dupuytren disease [241], three pulmonary fibrosis [265 280 281] and four endometriosis datasets [12]. Packages METAL[282] and GWAMA[283] were used for p-value based and beta-based inverse variance weighted meta-analysis methods respectively. Discrepancies in p-values between results from the two meta-analysis methods were assessed visually by plotting the p values generated from both meta-analysis methods against each other.

### 6.4.4 LD score regression and Cross-phenotype association analysis (CPASSOC)

Genome wide genetic correlation was calculated using logistic score regression LDSC v1.0.0 [284 285] and the summary statistics generated in chapter 6.4.3. The analysis was performed between fibrotic conditions in a pairwise fashion, where a p value less than 0.05 after being corrected for multiple testing indicates a correlation. Individual SNP overlaps between conditions were identified using beta and standard error from the summary statistics and CPASSOC [286] with the invocation of the SHet function. A p-value of less than $5 \times 10^{-8}$ indicated a significant association that may be contributed by one or more of the tested conditions, which were subsequently identified using bash

scripts. We considered a locus “shared” between two or more fibrotic conditions if it met the following criteria: 1. Overall CPASSOC p-value $<5x10^{-8}$; 2. Individual disease GWAS p-value $<1x10^{-4}$, i.e. at least suggestive of association. Forest plots and Manhattan plots were created for visualising the significant associations using R.

### 6.4.5 *In silico* analysis of the cross-trait associations

*In silico* analysis was performed using online databases to identify SNP functions to aid planning of future molecular functional work. All significantly associated SNPs that were identified in the CPASSOC analysis were included. The selected databases comprised 3DSNP, 3D Genome Browser, Haploreg, FuncPred, GTEx, regulomeDB, scRNASeqDB and SNP2TFBS, in order to detect predicted changes in nearby gene expression, promoter enhancer interaction and transcriptional factor binding sites. The cell type that these variants might affect have also been predicted.

## 6.5 Results

### 6.5.1 Phenotypic correlation between fibrotic conditions

In the June 2017 release of the UKBiobank HES data, there were a total of 502,619 samples, of which 273,455 were female. A list of disease diagnostic codes (ICD-10), surgical operation codes (OPCS) and self-diagnosed illness codes were generated to identify individuals who have had been diagnosed with the following fibrotic conditions; adhesive capsulitis, Dupuytren disease, endometriosis, hepatic fibrosis, pulmonary fibrosis, uterine fibroids and systemic sclerosis (Table 6.5.1.1). The proportion of cases between males and females for each fibrotic condition followed the published statistics. For example, there were three times more Dupuytren disease cases reported in males than females (Table 6.5.1.2)[2] and more systemic sclerosis cases in females than males [266]. The phenotypic correlation analysis was performed twice because of the gender specific nature of endometriosis and uterine fibroids. Multiple significant correlations were identified in both analyses, which included between adhesive capsulitis, Dupuytren disease and endometriosis; endometriosis and uterine fibroids; pulmonary fibrosis, hepatic fibrosis and systemic sclerosis (Table 6.5.1.3).

Table 6.5.1.1. A list of codes definition that were used to identify individuals, which have been diagnosed or self-reported with the corresponding disease.

| | ICD-10 | OPCS | Self-reported cases | Self-reported operations |
|---|---|---|---|---|
| Adhesive capsulitis | M750 | W781 | | |
| | M779 | W789 | | |
| Dupuytren disease | M720 | T521 | 1544 | 1535 |
| | M7201 | T522 | | |
| | M7204 | T525 | | |
| | | T526 | | |
| | | T561 | | |
| | | T562 | | |
| Endometriosis | N800 | | 1402 | |
| | N801 | | | |
| | N802 | | | |
| | N803 | | | |
| | N804 | | | |
| | N805 | | | |
| | N806 | | | |
| | N807 | | | |
| | N808 | | | |
| | N809 | | | |
| Hepatic fibrosis | K740 | | 1158 | |
| | K741 | | 1506 | |
| | K742 | | | |
| | K743 | | | |
| | K744 | | | |
| | K745 | | | |
| | K746 | | | |
| Pulmonary fibrosis | J841 | | | |
| System sclerosis | L940 | | 1384 | |
| | L941 | | | |
| Uterine fibroids | D25 | | 1351 | |
| | D250 | | | |
| | D251 | | | |
| | D252 | | | |
| | D259 | | | |

Table 6.5.1.2. Summary statistics of each fibrotic condition

| | Cases (% total UKBiobank participants) | Females (% total female UKBiobank participants) |
|---|---|---|
| Adhesive capsulitis | 2493 (0.5%) | 1565 (0.57%) |
| Dupuytren disease | 4389 (0.87%) | 1051 (0.38%) |
| Endometriosis | 7982 (1.59%) | 7982 (2.91%) |
| Hepatic fibrosis | 1259 (0.25%) | 597 (0.22%) |
| Pulmonary fibrosis | 1059 (0.21%) | 445 (0.16%) |
| Systemic sclerosis | 217 (0.04%) | 183 (0.07%) |
| Uterine fibroids | 19533 (3.89%) | 19533 (7.14%) |

Table 6.5.1.3. The p values of the pairwise phenotypic correlation between various fibrotic conditions within the UKBiobank dataset (chi-squared test). This analysis was performed twice, stratifying by cases a) of both sexes, and b) females only. Accented colour cells represent significant correlations

a)

| Females & Males | Adhesive capsulitis | Dupuytren's disease | Endometriosis | Hepatic fibrosis | Uterine fibroids | Pulmonary fibrosis | Systemic sclerosis |
|---|---|---|---|---|---|---|---|
| Adhesive capsulitis | - | | | | | | |
| Dupuytren's disease | $2.54x10^{-17}$ | - | | | | | |
| Endometriosis | $1.94x10^{-07}$ | $5.93x10^{-8}$ | - | | | | |
| Hepatic fibrosis | 0.007 | 0.069 | 0.114 | - | | | |
| Uterine fibroids | $6.67x10^{-03}$ | $1.19x10^{-13}$ | 0 | 0.892 | - | | |
| Pulmonary fibrosis | 0.583 | 0.363 | 0.655 | $4.91x10^{-43}$ | $4.21x10^{-4}$ | - | |
| Systemic sclerosis | 0.372 | 0.514 | $2.56x10^{-3}$ | 0.048 | $2.05x10^{-4}$ | $1.86x10^{-89}$ | - |

b)

| Females only | Adhesive capsulitis | Dupuytren's disease | Endometriosis | Hepatic fibrosis | Uterine fibroids | Pulmonary fibrosis | Systemic sclerosis |
|---|---|---|---|---|---|---|---|
| Adhesive capsulitis | - | | | | | | |
| Dupuytren's disease | $1.01x10^{-8}$ | - | | | | | |
| Endometriosis | $7.39x10^{-05}$ | 0.297 | - | | | | |
| Hepatic fibrosis | 0.161 | 0.845 | 0.281 | - | | | |
| Uterine fibroids | 0.270 | 0.911 | 0 | 0.394 | - | | |
| Pulmonary fibrosis | 0.731 | 0.323 | 0.571 | $3.33x10^{-16}$ | 0.018 | - | |
| Systemic sclerosis | 0.305 | 0.723 | 0.108 | 0.011 | 0.089 | $7.14x10^{-71}$ | - |

### 6.5.2 Genome wide associations of multiple fibrotic conditions

After quality control, of the seven association analyses, hepatic fibrosis had the least amount of cases (n=968) and uterine fibroids had the most (n =15184) (Table 6.5.2.1). A small number of imputed variants (n=25) that passed QC reached statistical significance of association without any supporting genome wide significant genotyped variants in the same associated region. These variants are of uncertain status, and require further work to confirm if they represent true associations. They are tabulated in the tables below (marked “imputed”) but not reported here as confident associated variants.

After removing the rare variants and poorly imputed SNPs, a total of 11,702,081 biallelic SNPs were analysed in each condition. In the adhesive capsulitis association analysis, one novel associated region was identified, with the lead SNP rs62228062 (Table 6.5.2.2; Figure 6.5.2.1). The attenuation ratio is 0.3 (LDSC intercept = 1.01) indicating 70% of the genomic inflation factor was attributable to polygenicity ($\lambda$gc = 1.05). The estimated heritability of this SNP is 0.0031 (se=0.0011).

In the Dupuytren disease association analysis, 23 novel associations and all 26 previously published loci were identified. The new novel associations include one region on chromosome 5,10,11,19,20 and 21, two regions on chromosome 1, 2, 4, 8, 9, 12, 15 and three regions on chromosome 6 (Table 6.5.2.3; Figure 6.5.2.2). The

LDSC intercept was less than 1 and the genomic inflation factor was 1.09, indicating the entire λgc was attributable to polygenicity. The estimated heritability of these SNPs is 0.0302 (se=0.0032).

In the endometriosis association analysis, 9 associated loci were identified (Table 6.5.2.4; Figure 6.5.2.3), where both on chromosome 10, rs1802669 and rs792234 are novel. The LDSC intercept was less than 1 and the genomic factor was 1.02, indicating the entire λgc was attributable to polygenicity. The estimated heritability of these SNPs is 0.0234 (se=0.0024).

In the hepatic fibrosis association analysis, two associated loci have been identified (Table 6.5.2.5; Figure 6.5.2.4); rs80215559 (OR=1.63, $p=8.7 \times 10^{-9}$) and rs738408 (OR=1.43, $p=7.2 \times 10^{-11}$). The estimated heritability ($h^2$) was 0.0008 (se=0.0011). There was no hidden population stratification (λgc ~ 1) in the analysed dataset.

In the idiopathic pulmonary fibrosis association analysis, 15 associated loci were identified, where seven of which were novel (Table 6.5.2.6; Figure 6.5.2.5). There was no hidden population stratification (λgc ~ 1). The estimated heritability of these SNPs is 0.0011(se= 0.001).

In the system sclerosis association analysis, one association, rs10174238 (OR=1.48, $p=8.77x10^{-14}$) was identified (Table 6.5.2.7; Figure 6.5.2.6), which has previously been reported. There was no hidden population stratification ($\lambda gc \sim 1$). The estimated heritability of this SNP is 0.0044 (se = 0.0013).

In the uterine fibrosis association analysis, 15 associations were identified (Table 6.5.2.8; Figure 6.5.2.7), where 13 of which have not been previously reported. This included one region on chromosome 1, 3, 5, 6, 7, 8, 12, 13, 16 and 22, 2 regions on chromosome 4, and 3 regions on chromosome 11. The attenuation ratio was 0.07 (LDSC intercept = 1.007) indicating 92% of the genomic inflation factor was attributable to polygenicity ($\lambda gc = 1.05$). The estimated heritability of these SNPs was 0.0204 (se = 0.0028).

Table 6.5.2.1 Table showing the total number of cases and controls used in each association study.

| Condition | Cases | Controls |
|---|---|---|
| Adhesive capsulitis | 1859 | 399799 |
| Duputren disease | 7607 | 402608 |
| Endometriosis | 14609 | 241793 |
| Hepatic fibrosis | 968 | 400690 |
| Idiopathic pulmonary fibrosis | 3600 | 409379 |
| Systemic sclerosis | 1449 | 405064 |
| Uterine fibroids | 15184 | 205752 |

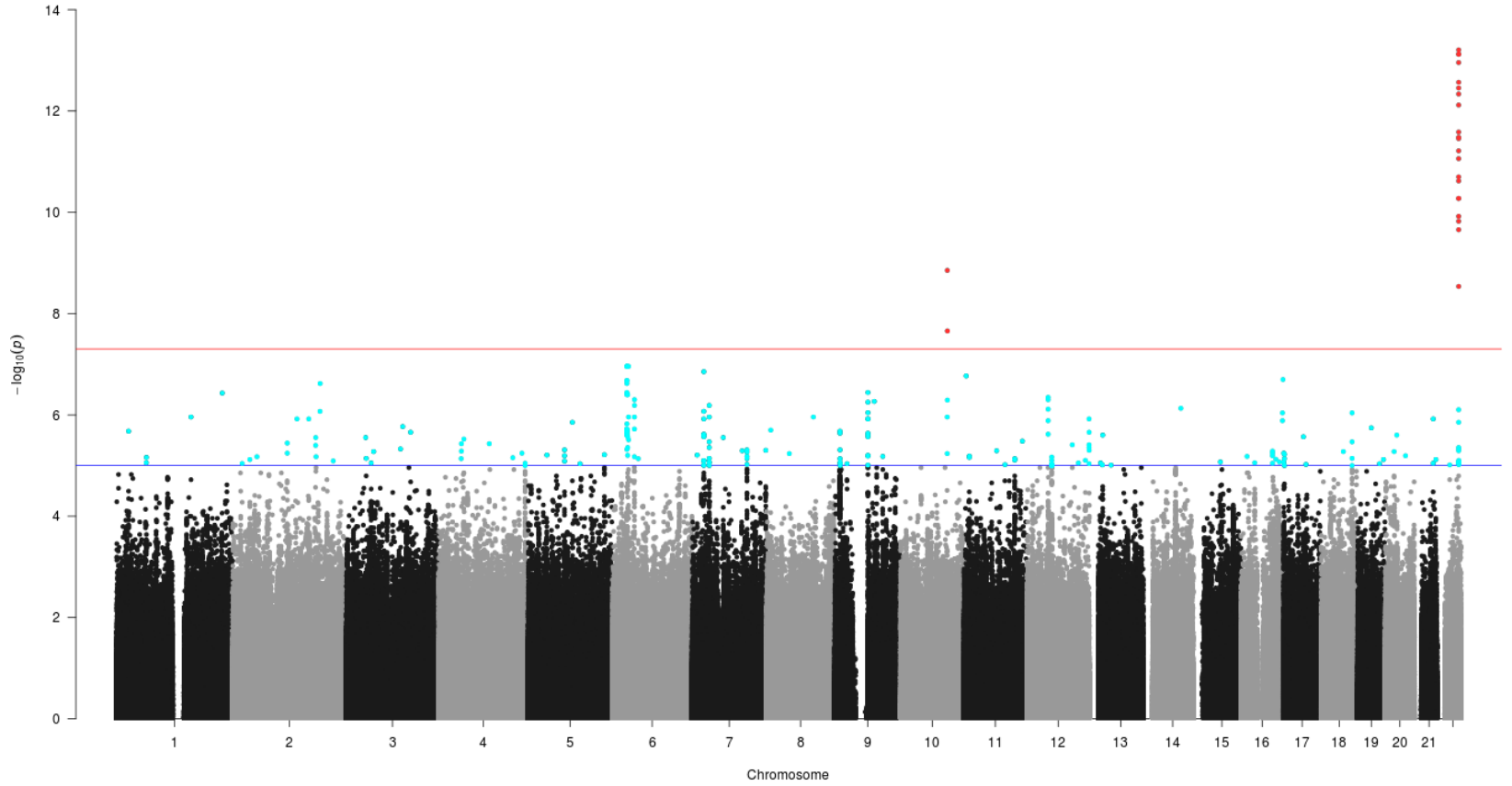



Figure 6.5.2.1 Manhattan plot showing the location and -$\log_{10}$ p-values of each SNP reported in the Adhesive Capsulitis association analysis. Genomic wide significant associations ($p < 5 \times 10^{-8}$) are highlighted red and suggestive associations ($p < 5 \times 10^{-5}$) are highlighted blue. The blue line and the red line mark the thresholds $5 \times 10^{-5}$ and $5 \times 10^{-8}$, respectively.

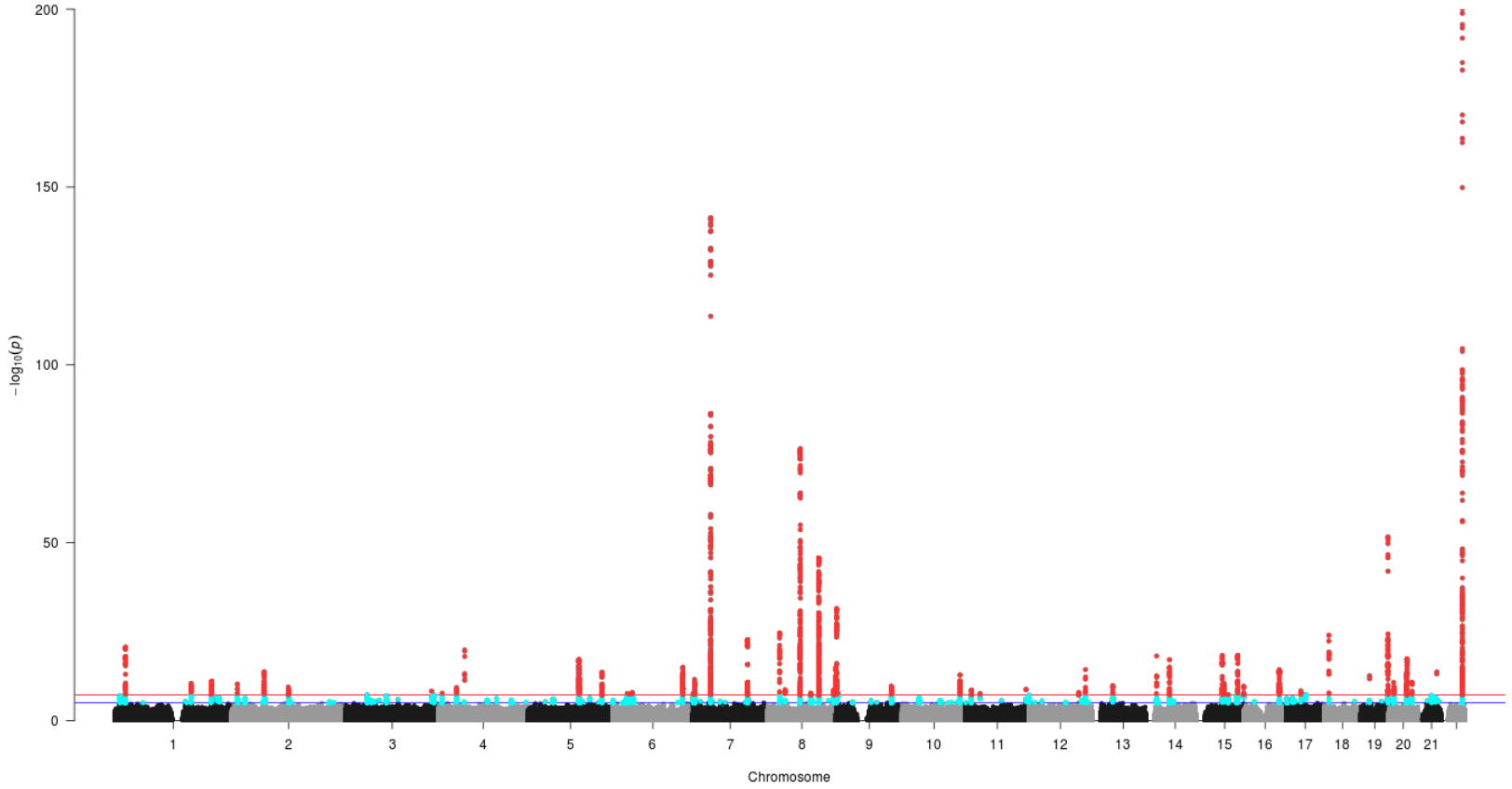



Figure 6.5.2.2 Manhattan plot showing the location and -$\log_{10}$ p-values of each SNP reported in the Dupuytren Disease association analysis. Genomic wide significant associations ($p < 5 \times 10^{-8}$) are highlighted red and suggestive associations ($p < 5 \times 10^{-5}$) are highlighted blue. The blue line and the red line mark the thresholds $5 \times 10^{-5}$ and $5 \times 10^{-8}$, respectively.

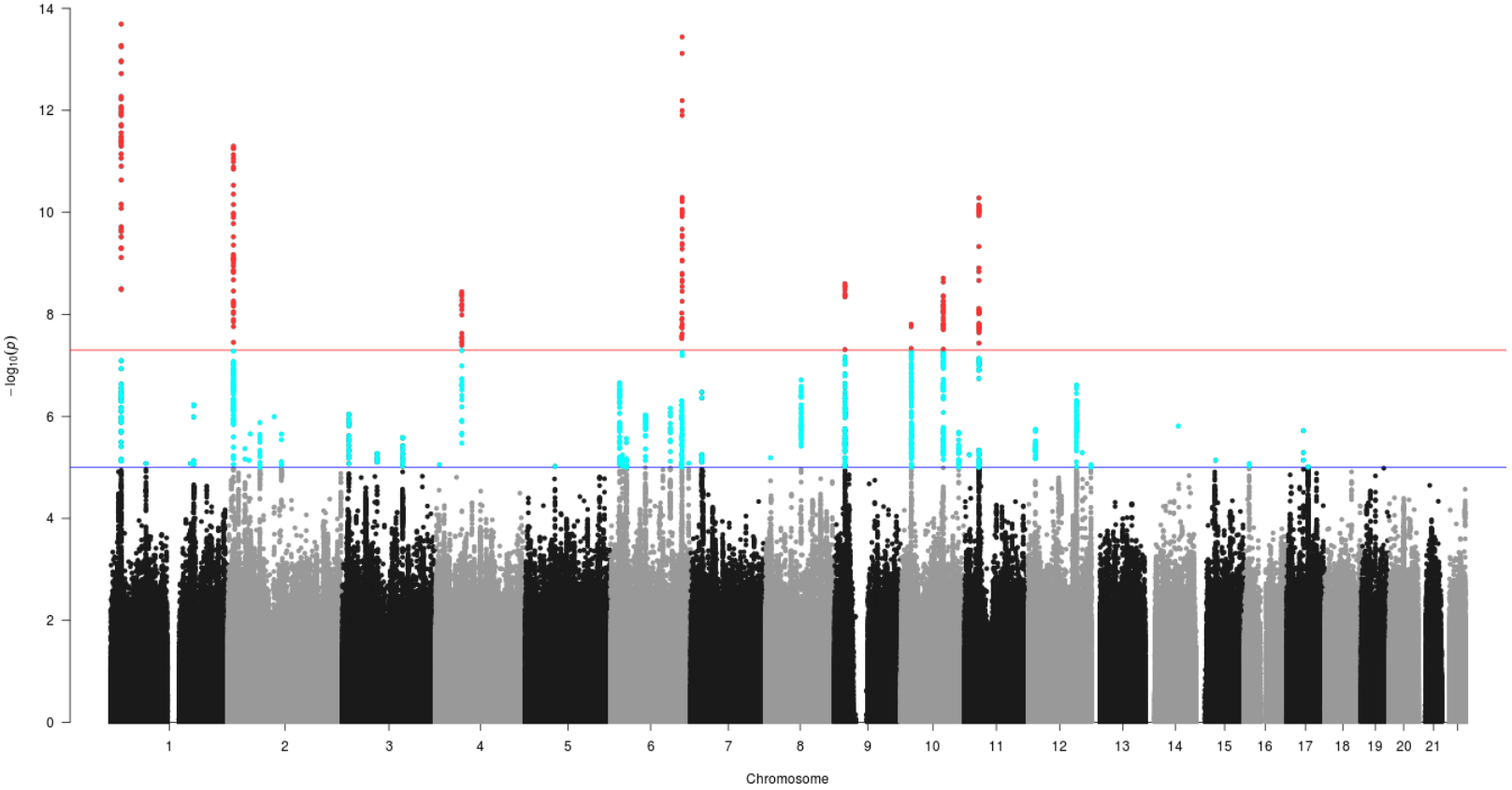



Figure 6.5.2.3 Manhattan plot showing the location and -$\log_{10}$ p-values of each SNP reported in the Endometriosis association analysis. Genomic wide significant associations ($p < 5 \times 10^{-8}$) are highlighted red and suggestive associations ($p < 5 \times 10^{-5}$) are highlighted blue. The blue line and the red line mark the thresholds $5 \times 10^{-5}$ and $5 \times 10^{-8}$, respectively.

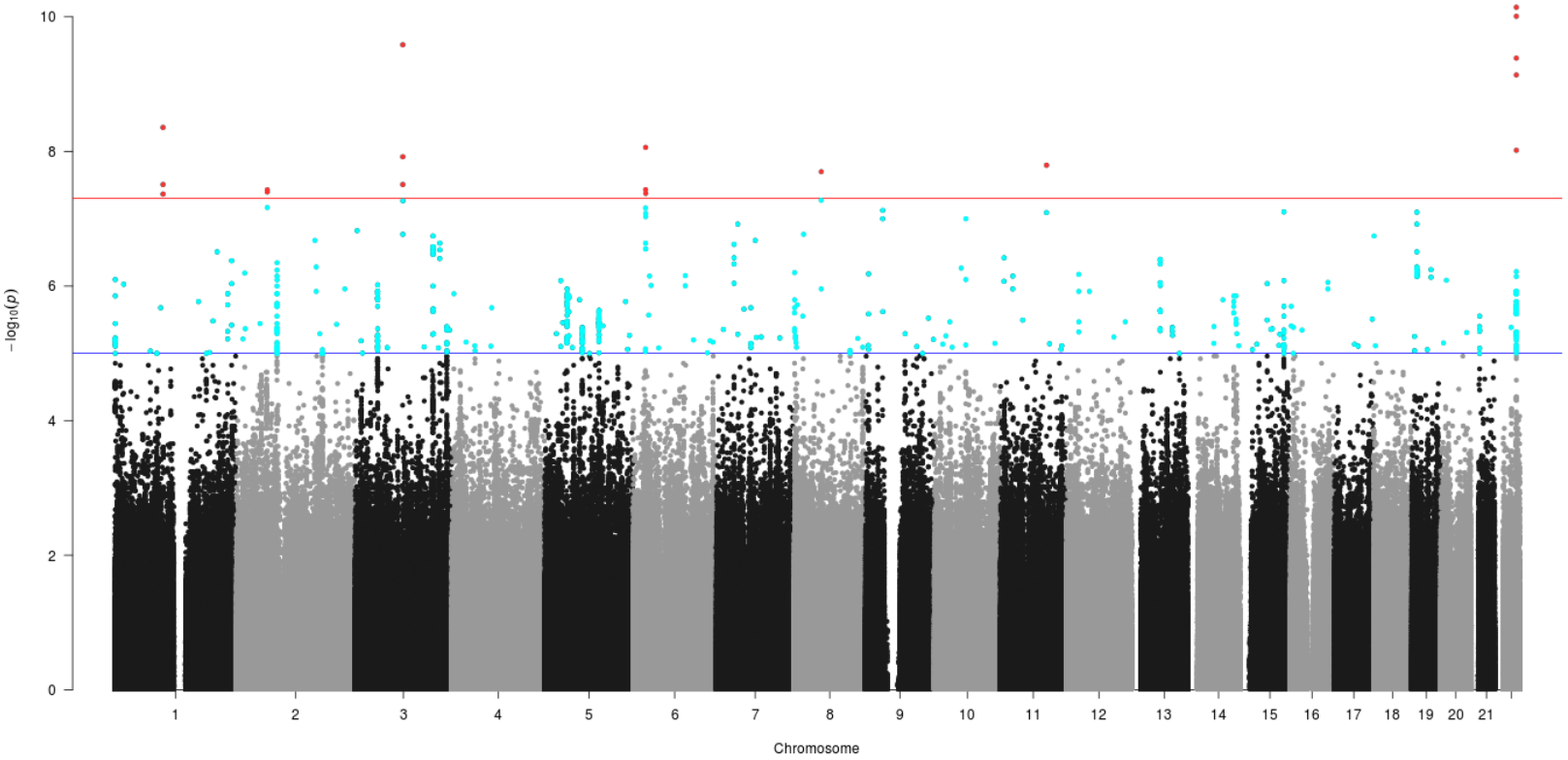



Figure 6.5.2.4 Manhattan plot showing the location and -$\log_{10}$ p-values of each SNP reported in the hepatic fibrosis association analysis. Genomic wide significant associations ($p < 5 \times 10^{-8}$) are highlighted red and suggestive associations ($p < 5 \times 10^{-5}$) are highlighted blue. The blue line and the red line mark the thresholds $5 \times 10^{-5}$ and $5 \times 10^{-8}$, respectively.

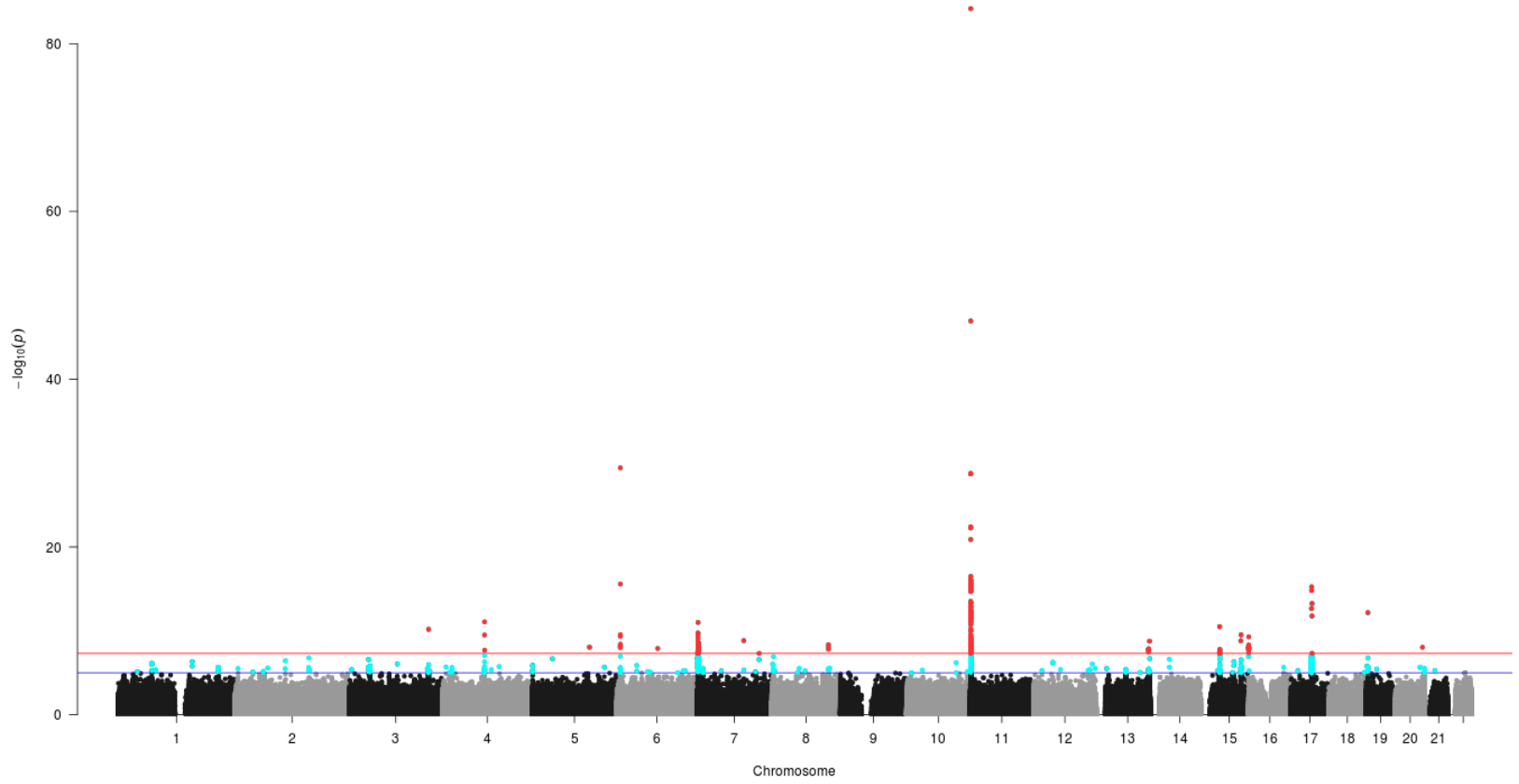



Figure 6.5.2.5 Manhattan plot showing the location and -$\log_{10}$ p-values of each SNP reported in the idiopathic pulmonary fibrosis association analysis. Genomic wide significant associations ($p < 5 \times 10^{-8}$) are highlighted red and suggestive associations ($p < 5 \times 10^{-5}$) are highlighted blue. The blue line and the red line mark the thresholds $5 \times 10^{-5}$ and $5 \times 10^{-8}$, respectively.

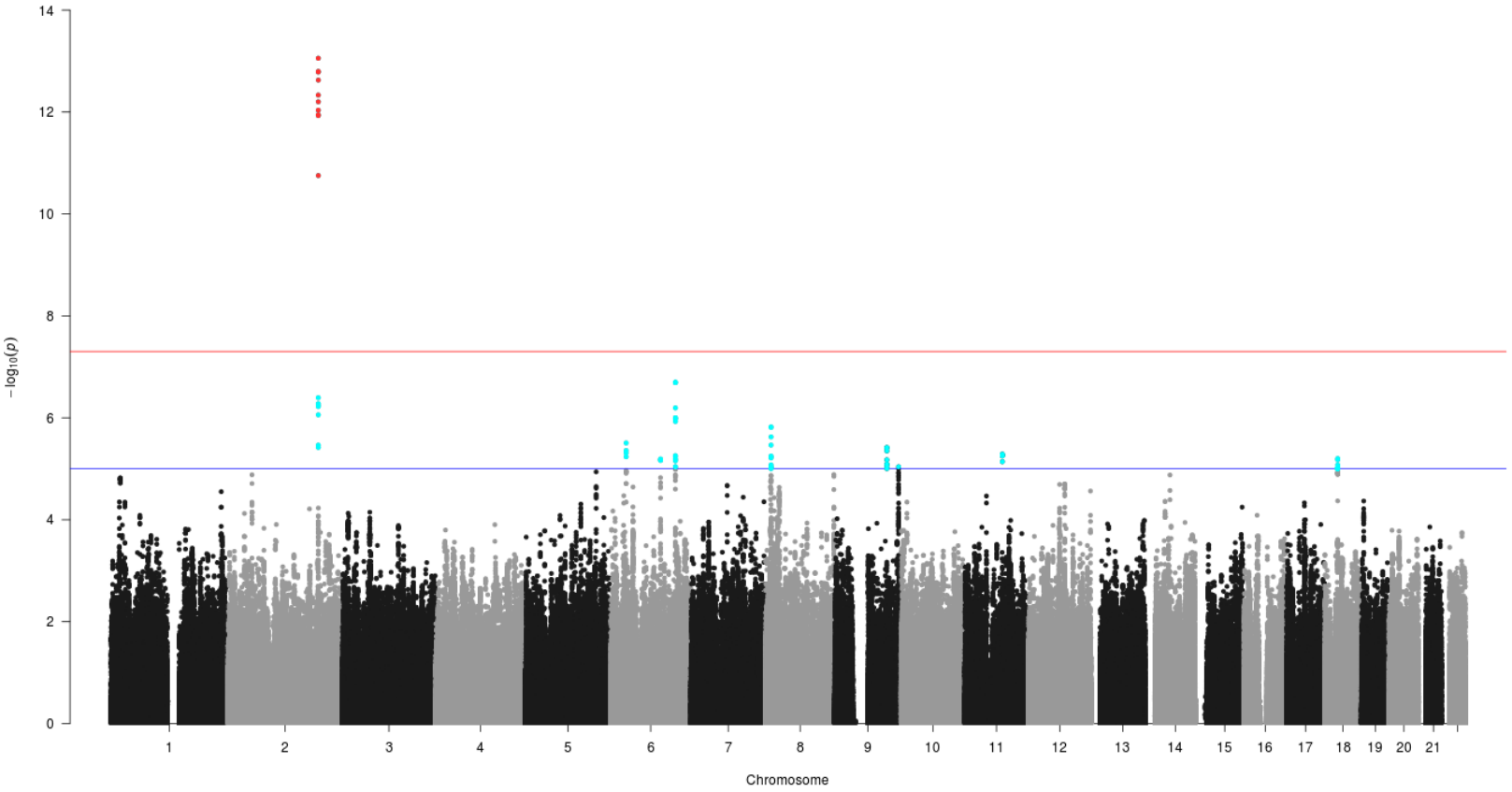



Figure 6.5.2.6 Manhattan plot showing the location and -$\log_{10}$ p-values of each SNP reported in the system sclerosis association analysis. Genomic wide significant associations ($p < 5 \times 10^{-8}$) are highlighted red and suggestive associations ($p < 5 \times 10^{-5}$) are highlighted blue. The blue line and the red line mark the thresholds $5 \times 10^{-5}$ and $5 \times 10^{-8}$, respectively.

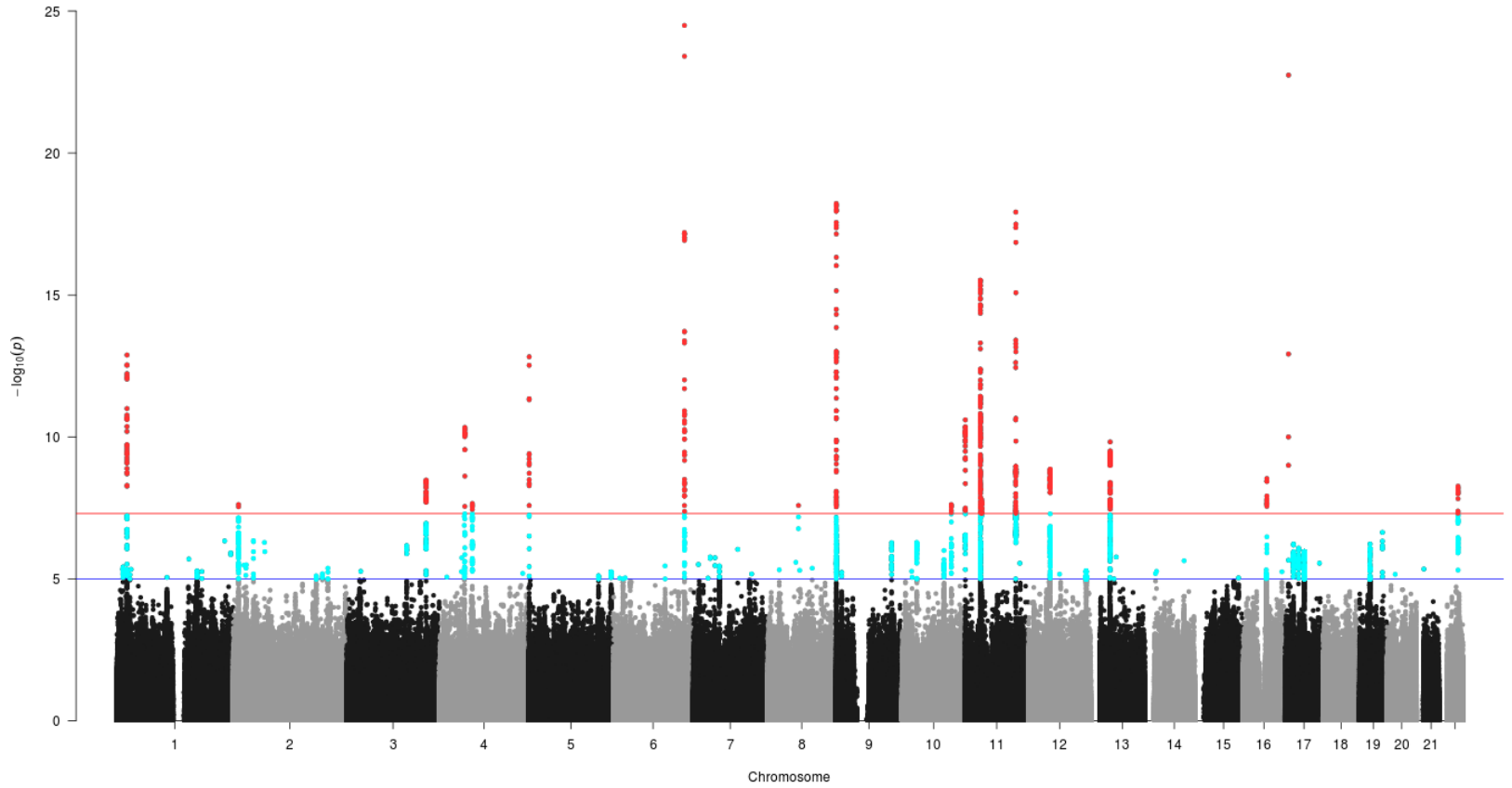



Figure 6.5.2.7 Manhattan plot showing the location and -$\log_{10}$ p-values of each SNP reported in the uterine fibroids association analysis. Genomic wide significant associations ($p < 5 \times 10^{-8}$) are highlighted red and suggestive associations ($p < 5 \times 10^{-5}$) are highlighted blue. The blue line and the red line mark the thresholds $5 \times 10^{-5}$ and $5 \times 10^{-8}$, respectively.

Table 6.5.2.2 Significant associated SNPs and the relevant information reported in the adhesive capsulitis GWAS. Associated loci with at least one significant genotyped variant in the region was marked genotyped in the imputed/genotyped column. An asterisk next to the rsID number indicates novel association.

| rsID | Chromosome | Position | EAF | Ref. allele | OR | 95% CI | p | Imputed/Genotyped |
|---|---|---|---|---|---|---|---|---|
| rs1543563 | 10 | 100016274 | 0.001 | A | 21.02 | 7.85-56.31 | $1.40x10^{-09}$ | Imputed |
| rs62228062* | 22 | 46381234 | 0.26 | G | 1.33 | 1.23-1.43 | $6.20x10^{-14}$ | Genotyped |

Table 6.5.2.3 Significant associated SNPs and the relevant information reported in the Dupuytren disease GWAS. Associated loci with at least one significant genotyped variant in the region was marked genotyped in the imputed/genotyped column. An asterisk next to the rsID number indicates novel association.

| rsID | Chromosome | Position | EAF | Ref. allele | OR | 95% CI | p | Imputed/Genotyped |
|---|---|---|---|---|---|---|---|---|
| rs11802513 | 1 | 22701146 | 0.17 | A | 1.26 | 1.2-1.32 | $2.08X10^{-21}$ | Genotyped |
| rs4657218 | 1 | 162637785 | 0.11 | T | 0.81 | 0.76-0.86 | $3.05X10^{-11}$ | Genotyped |
| rs1172134* | 1 | 205248032 | 0.09 | A | 1.24 | 1.17-1.33 | $1.26X10^{-11}$ | Genotyped |
| rs3849275* | 1 | 206605665 | 0.46 | G | 1.13 | 1.09-1.18 | $8.09X10^{-12}$ | Genotyped |
| rs17605696* | 2 | 12039678 | 0.34 | C | 1.14 | 1.09-1.18 | $4.85X10^{-11}$ | Imputed |
| rs72895448* | 2 | 69268277 | 0.04 | A | 1.4 | 1.28-1.53 | $1.73X10^{-14}$ | Genotyped |
| rs10174596* | 2 | 120507425 | 0.27 | C | 0.87 | 0.83-0.91 | $3.30X10^{-10}$ | Genotyped |
| rs7630096* | 3 | 183199517 | 0.36 | T | 1.12 | 1.08-1.17 | $4.62X10^{-09}$ | Imputed |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| rs73214051* | 4 | 7935754 | 0.24 | C | 0.88 | 0.84-0.92 | $1.74X10^{-08}$ | Imputed |
| rs1386625* | 4 | 38361120 | 0.09 | A | 1.22 | 1.14-1.3 | $4.60X10^{-10}$ | Genotyped |
| rs73818546* | 4 | 55941418 | 0.14 | A | 0.76 | 0.72-0.81 | $1.25X10^{-20}$ | Genotyped |
| rs1965670 | 5 | 108649798 | 0.23 | T | 0.82 | 0.78-0.85 | $5.83X10^{-18}$ | Genotyped |
| rs6881996* | 5 | 157888115 | 0.25 | T | 1.18 | 1.13-1.23 | $2.37X10^{-14}$ | Genotyped |
| rs886423* | 6 | 30782205 | 0.19 | C | 1.14 | 1.09-1.2 | $2.18X10^{-08}$ | Genotyped |
| rs733590* | 6 | 36645203 | 0.33 | C | 1.11 | 1.07-1.16 | $3.55X10^{-08}$ | Imputed |
| rs6916515* | 6 | 41666065 | 0.13 | A | 1.16 | 1.1-1.23 | $1.32X10^{-08}$ | Imputed |
| rs1626823 | 6 | 149710663 | 0.47 | C | 1.16 | 1.12-1.21 | $9.28X10^{-16}$ | Genotyped |
| rs9770861* | 7 | 582259 | 0.36 | G | 0.86 | 0.82-0.91 | $2.23X10^{-08}$ | Imputed |
| rs10264803 | 7 | 3516425 | 0.29 | T | 0.86 | 0.82-0.89 | $2.61X10^{-12}$ | Genotyped |
| rs2598100 | 7 | 37980161 | 0.35 | G | 1.43 | 1.36-1.50 | $1.5X10^{-50}$ | Genotyped |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| rs2044830 | 7 | 38024404 | 0.1 | A | 2.1 | 1.98-2.23 | $4.32X10^{-142}$ | Genotyped |
| rs12706146 | 7 | 116910648 | 0.46 | G | 0.82 | 0.79-0.85 | $1.57X10^{-23}$ | Genotyped |
| rs901177 | 8 | 25867383 | 0.3 | T | 1.23 | 1.19-1.29 | $2.03X10^{-25}$ | Genotyped |
| rs7386644 | 8 | 37482036 | 0.16 | C | 1.16 | 1.1-1.22 | $1.85X10^{-09}$ | Genotyped |
| rs2472141 | 8 | 70054216 | 0.28 | T | 1.47 | 1.41-1.53 | $3.34X10^{-77}$ | Genotyped |
| rs10096149* | 8 | 92129436 | 0.28 | C | 0.88 | 0.85-0.92 | $1.82X10^{-08}$ | Imputed |
| rs674780 | 8 | 109185463 | 0.45 | T | 0.75 | 0.73-0.78 | $1.88X10^{-46}$ | Genotyped |
| rs62524109* | 8 | 141807467 | 0.43 | G | 0.89 | 0.85-0.92 | $2.41X10^{-09}$ | Genotyped |
| rs7816361 | 8 | 145504690 | 0.48 | C | 0.85 | 0.82-0.88 | $1.26X10^{-15}$ | Genotyped |
| rs12340528 | 9 | 1202855 | 0.25 | A | 1.29 | 1.23-1.34 | $3.08X10^{-32}$ | Genotyped |
| rs74573085* | 9 | 4366497 | 0.12 | C | 1.17 | 1.11-1.24 | $2.06X10^{-08}$ | Genotyped |
| rs1335332* | 9 | 117884015 | 0.33 | T | 1.13 | 1.09-1.18 | $2.20X10^{-10}$ | Genotyped |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| rs7910243* | 10 | 123419652 | 0.18 | A | 1.19 | 1.14-1.25 | $1.37X10^{-13}$ | Genotyped |
| rs7131034* | 11 | 12183036 | 0.22 | A | 0.87 | 0.83-0.91 | $2.20X10^{-09}$ | Genotyped |
| rs139904747* | 11 | 30698562 | 0.001 | G | 7.53 | 3.7-15.3 | $2.40X10^{-08}$ | Imputed |
| rs583683* | 11 | 128645583 | 0.44 | C | 1.12 | 1.08-1.16 | $1.69X10^{-09}$ | Imputed |
| rs2374545* | 12 | 106140551 | 0.38 | A | 0.89 | 0.86-0.92 | $1.10X10^{-08}$ | Genotyped |
| rs77660995* | 12 | 120672843 | 0.02 | T | 1.62 | 1.43-1.83 | $4.16X10^{-15}$ | Genotyped |
| rs4942308 | 13 | 44866957 | 0.32 | C | 0.87 | 0.84-0.91 | $1.33X10^{-10}$ | Imputed |
| rs1042704 | 14 | 23312594 | 0.21 | A | 1.22 | 1.17-1.28 | $6.38X10^{-19}$ | Genotyped |
| rs61984263 | 14 | 50878681 | 0.38 | C | 0.84 | 0.81-0.87 | $6.35X10^{-18}$ | Genotyped |
| rs8032158 | 15 | 56194877 | 0.32 | C | 1.19 | 1.15-1.24 | $4.85X10^{-19}$ | Genotyped |
| rs4775130* | 15 | 59503103 | 0.36 | T | 0.87 | 0.84-0.91 | $2.69X10^{-11}$ | Genotyped |
| rs28522005 | 15 | 68625715 | 0.11 | G | 1.17 | 1.1-1.24 | $4.66X10^{-08}$ | Imputed |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| rs10852113 | 15 | 89254370 | 0.19 | A | 0.8 | 0.76-0.84 | $4.82X10^{-19}$ | Genotyped |
| rs4965806* | 15 | 101752435 | 0.38 | A | 0.88 | 0.84-0.91 | $2.03X10^{-10}$ | Genotyped |
| rs8050662 | 16 | 75360734 | 0.14 | C | 1.23 | 1.17-1.3 | $4.23X10^{-15}$ | Genotyped |
| rs8073901* | 17 | 30963534 | 0.17 | A | 0.86 | 0.82-0.9 | $4.04X10^{-09}$ | Genotyped |
| rs11665156 | 18 | 9764516 | 0.07 | T | 0.68 | 0.63-0.73 | $8.35X10^{-25}$ | Genotyped |
| rs78030362* | 19 | 18575193 | 0.07 | G | 1.3 | 1.21-1.4 | $2.35X10^{-13}$ | Genotyped |
| rs11672447 | 19 | 57678212 | 0.23 | G | 1.4 | 1.34-1.46 | $2.17X10^{-52}$ | Genotyped |
| rs35308216* | 20 | 10629525 | 0.08 | C | 0.78 | 0.73-0.84 | $1.71X10^{-11}$ | Genotyped |
| rs16988531 | 20 | 38478405 | 0.04 | A | 1.39 | 1.27-1.52 | $1.12X10^{-13}$ | Genotyped |
| rs6029275 | 20 | 39319185 | 0.15 | T | 0.79 | 0.75-0.83 | $3.92X10^{-18}$ | Genotyped |
| rs73130881* | 20 | 50184300 | 0.27 | G | 1.15 | 1.1-1.2 | $1.45X10^{-11}$ | Imputed |
| rs117870289* | 21 | 39983448 | 0.01 | T | 1.8 | 1.55-2.1 | $1.92X10^{-14}$ | Genotyped |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| rs10453460 | 22 | 46366165 | 0.25 | T | 2.43 | 2.33-2.54 | 0 | Genotyped |

Table 6.5.2.4 Significant associated SNPs and the relevant information reported in the Endometriosis GWAS. Associated loci with at least one significant genotyped variant in the region was marked genotyped in the imputed/genotyped column. An asterisk next to the rsID number indicates novel association.

| rsID | Chromosome | Position | EAF | Ref. allele | OR | 95% CI | p | Imputed/Genotyped |
|---|---|---|---|---|---|---|---|---|
| rs61768001 | 1 | 22465820 | 0.28 | C | 1.14 | 1.1-1.19 | $2.03x10^{-14}$ | Genotyped |
| rs9789525 | 2 | 11724580 | 0.34 | T | 0.9 | 0.88-0.93 | $5.04x10^{-12}$ | Genotyped |
| rs9312655 | 4 | 56002665 | 0.41 | C | 0.91 | 0.89-0.94 | $3.63x10^{-09}$ | Genotyped |
| rs58415480 | 6 | 152562271 | 0.35 | G | 1.15 | 1.11-1.19 | $3.63x10^{-14}$ | Genotyped |
| rs7742927 | 6 | 152630180 | 0.15 | A | 1.12 | 1.08-1.15 | $6.44x10^{-13}$ | Genotyped |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| rs10122243 | 9 | 22158924 | 0.34 | T | 1.08 | 1.05-1.11 | $2.51 \times 10^{-09}$ | Genotyped |
| rs1802669* | 10 | 21827796 | 0.21 | A | 1.08 | 1.05-1.11 | $1.56 \times 10^{-08}$ | Genotyped |
| rs792234* | 10 | 90151021 | 0.15 | A | 1.08 | 1.05-1.11 | $1.95 \times 10^{-09}$ | Genotyped |
| rs10835649 | 11 | 30329175 | 0.15 | G | 0.88 | 0.85-0.91 | $5.22 \times 10^{-11}$ | Genotyped |

Table 6.5.2.5 Significant associated SNPs and the relevant information reported in the hepatic fibrosis GWAS. Associated loci with at least one significant genotyped variant in the region was marked genotyped in the imputed/genotyped column. An asterisk next to the rsID number indicates novel association

| rsID | Chromosome | Position | EAF | Ref. allele | OR | 95% CI | p | Imputed/Genotyped |
|---|---|---|---|---|---|---|---|---|
| rs575372939 | 1 | 98859277 | 0.002 | A | 31.23 | 9.89-98.7 | $4.40 \times 10^{-09}$ | Imputed |
| rs188607976 | 2 | 63325258 | 0.002 | C | 18.36 | 6.52-51.73 | $3.70 \times 10^{-08}$ | Imputed |
| rs146184240 | 3 | 97794660 | 0.004 | C | 11.13 | 5.28-23.52 | $2.60 \times 10^{-10}$ | Imputed |
| rs80215559* | 6 | 25918225 | 0.08 | C | 1.63 | 1.38-1.93 | $8.70 \times 10^{-09}$ | Genotyped |
| rs186619282 | 8 | 55834462 | 0.002 | G | 46.79 | 12.21-179 | $2.00 \times 10^{-08}$ | Imputed |
| rs35354571 | 11 | 94643561 | 0.006 | A | 6.49 | 3.39-12.42 | $1.60 \times 10^{-08}$ | Imputed |
| rs738408 | 22 | 44324730 | 0.23 | T | 1.43 | 1.28-1.60 | $7.20 \times 10^{-11}$ | Genotyped |

Table 6.5.2.6 Significant associated SNPs and the relevant information reported in the idiopathic pulmonary fibrosis GWAS. Associated loci with at least one significant genotyped variant in the region was marked genotyped in the imputed/genotyped column. An asterisk next to the rsID number indicates novel association.

| rsID | Chromosome | Position | EAF | Ref. allele | OR | 95% CI | p | Imputed/Genotyped |
|---|---|---|---|---|---|---|---|---|
| rs6793295* | 3 | 169518455 | 0.28 | C | 1.31 | 1.21-1.43 | $6.63x10^{-11}$ | Genotyped |
| rs2609255* | 4 | 89811195 | 0.22 | G | 1.29 | 1.2-1.4 | $8.60x10^{-12}$ | Genotyped |
| rs547189713 | 5 | 122438277 | 0.001 | C | 184 | 31.17-1090 | $9.01x10^{-09}$ | Imputed |
| rs2076295 | 6 | 7563232 | 0.46 | G | 1.45 | 1.36-1.55 | $3.63x10^{-30}$ | Genotyped |
| rs117761768 | 6 | 87174399 | 0.001 | G | 168 | 28.81-983 | $1.29x10^{-08}$ | Imputed |
| rs10950503* | 7 | 2039594 | 0.42 | A | 1.25 | 1.17-1.34 | $1.01x10^{-11}$ | Genotyped |
| rs2283017 | 7 | 99574758 | 0.4 | G | 1.22 | 1.14-1.3 | $1.47x10^{-09}$ | Genotyped |
| rs190071296 | 7 | 132239089 | 0.004 | A | 15.62 | 5.82-41.94 | $4.96x10^{-08}$ | Imputed |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| rs9987332* | 8 | 120933963 | 0.4 | A | 0.82 | 0.77-0.87 | $4.46 \times 10^{-09}$ | Genotyped |
| rs35705950 | 11 | 1241221 | 0.14 | T | 4.72 | 4.04-5.51 | $6.55 \times 10^{-85}$ | Genotyped |
| rs9515368* | 13 | 111732771 | 0.27 | A | 1.38 | 1.23-1.55 | $1.38 \times 10^{-08}$ | Genotyped |
| rs1278769* | 13 | 113536627 | 0.21 | A | 0.78 | 0.72-0.84 | $1.72 \times 10^{-09}$ | Genotyped |
| rs1007177 | 15 | 40651838 | 0.33 | A | 0.78 | 0.73-0.84 | $3.21 \times 10^{-11}$ | Genotyped |
| rs338527 | 15 | 86261659 | 0.39 | G | 1.23 | 1.15-1.32 | $3.00 \times 10^{-10}$ | Genotyped |
| rs62031426* | 16 | 92195 | 0.06 | T | 2.14 | 1.68-2.72 | $5.25 \times 10^{-10}$ | Genotyped |
| rs1981997 | 17 | 44056767 | 0.22 | A | 0.71 | 0.65-0.77 | $5.88 \times 10^{-16}$ | Genotyped |
| rs415430 | 17 | 44859144 | 0.21 | C | 0.72 | 0.66-0.78 | $5.58 \times 10^{-14}$ | Genotyped |
| rs2109069 | 19 | 4719443 | 0.32 | A | 1.28 | 1.2-1.37 | $6.83 \times 10^{-13}$ | Genotyped |
| rs201581384 | 20 | 57766673 | 0.003 | T | 23.13 | 7.92-67.53 | $9.26 \times 10^{-09}$ | Imputed |

Table 6.5.2.7 Significant associated SNP and the relevant information reported in the systemic sclerosis GWAS. This locus has been previously reported.

| rsID | Chromosome | Position | EAF | Ref. allele | OR | 95% CI | P | Imputed/Genotyped |
|---|---|---|---|---|---|---|---|---|
| rs10174238 | 2 | 191973034 | 0.23 | G | 1.48 | 1.33-1.63 | $8.77 \times 10^{-14}$ | Genotyped |

Table 6.5.2.8 Significant associated SNPs and the relevant information reported in the uterine fibroids GWAS. Associated loci with at least one significant genotyped variant in the region was marked genotyped in the imputed/genotyped column. An asterisk next to the rsID number indicates novel association.

| rsID | Chromosome | Position | EAF | Ref. allele | OR | 95% CI | p | Imputed/Genotyped |
|---|---|---|---|---|---|---|---|---|
| rs2235529* | 1 | 22450487 | 0.16 | T | 1.12 | 1.09-1.16 | $1.30x10^{-13}$ | Genotyped |
| rs10929757 | 2 | 11702661 | 0.42 | A | 0.93 | 0.91-0.95 | $2.40x10^{-08}$ | Imputed |
| rs10936600* | 3 | 169514585 | 0.24 | T | 0.92 | 0.89-0.94 | $3.30x10^{-09}$ | Genotyped |
| rs62323680* | 4 | 54546192 | 0.067 | A | 1.16 | 1.11-1.22 | $4.60x10^{-11}$ | Genotyped |
| rs2202282* | 4 | 70634441 | 0.50 | T | 1.06 | 1.04-1.09 | $2.20x10^{-08}$ | Genotyped |
| rs2736100* | 5 | 1286516 | 0.50 | A | 0.91 | 0.89-0.93 | $1.50x10^{-13}$ | Genotyped |
| rs58415480* | 6 | 152562271 | 0.15 | G | 1.18 | 1.14-1.22 | $3.20x10^{-25}$ | Genotyped |
| rs75054103 | 8 | 66333154 | 0.001 | A | 2.55 | 1.84-3.52 | $2.60x10^{-08}$ | Imputed |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| rs7030354* | 9 | 804231 | 0.40 | T | 1.11 | 1.08-1.13 | $6.00 \times 10^{-19}$ | Genotyped |
| rs9419958 | 10 | 105675946 | 0.13 | T | 1.1 | 1.06-1.14 | $2.40 \times 10^{-08}$ | Imputed |
| rs507139 | 11 | 225196 | 0.03 | A | 0.85 | 0.82-0.89 | $2.50 \times 10^{-11}$ | Genotyped |
| rs10835889* | 11 | 32370380 | 0.07 | A | 1.14 | 1.1-1.17 | $3.00 \times 10^{-16}$ | Genotyped |
| rs2553772* | 11 | 35085453 | 0.16 | T | 0.93 | 0.91-0.95 | $1.60 \times 10^{-08}$ | Genotyped |
| rs141379009* | 11 | 108149207 | 0.46 | G | 1.37 | 1.28-1.47 | $1.20 \times 10^{-18}$ | Genotyped |
| rs7485599* | 12 | 46830479 | 0.30 | A | 0.92 | 0.9-0.94 | $1.40 \times 10^{-09}$ | Genotyped |
| rs7986407* | 13 | 41179798 | 0.31 | G | 1.08 | 1.05-1.11 | $1.50 \times 10^{-10}$ | Genotyped |
| rs66998222* | 16 | 51481596 | 0.20 | A | 0.91 | 0.89-0.94 | $2.90 \times 10^{-09}$ | Genotyped |
| rs78378222 | 17 | 7571752 | 0.012 | G | 1.71 | 1.54-1.9 | $1.80 \times 10^{-23}$ | Imputed |
| rs12484951 | 22 | 40703075 | 0.21 | G | 1.08 | 1.05-1.11 | $5.40 \times 10^{-09}$ | Genotyped |

### 6.5.3 Cross-fibrotic genetic associations

In the pairwise LD score regression analysis, endometriosis and uterine fibroid demonstrated a significant positive correlation ($r^2 = 0.3534$). Further genetic correlation analyses using CPASSOC revealed 18 cross phenotypic associated regions (Figure 6.5.3.1). The breakdown of these correlations were four loci shared between adhesive capsulitis and Dupuytren disease, one between Dupuytren disease and endometriosis, one between Dupuytren disease and uterine fibroid, eight between endometriosis and uterine fibroids, two between uterine fibroid and pulmonary fibrosis, and two between Dupuytren disease and pulmonary fibrosis (Table 6.5.3.2). Of note, four loci were identified in the cross-fibrosis analysis that where sub-threshold of statistical significance in the individual GWAS analysis. This means that the contribution of these variants to fibrosis may have been overlooked without this analysis (Table 6.5.3.1 and 6.5.3.2). No correlations between hepatic fibrosis, system sclerosis and other fibrotic conditions were identified. As the threshold to identify the secondary contributing fibrotic condition was somewhat lenient ($p < 1 \times 10^{-4}$, see Chapter 6.4.4), visual inspection on the 1Mb region surrounding the lead SNP was employed to validate the genuineness of the signal.

Seven cross fibrotic signals included Dupuytren disease as a contributing fibrotic condition (Figure 6.5.3.2-6.5.3.8) include rs61811971 (CPASSOC $p=1.33\times10^{-11}$), rs6881996 (CPASSOC $p=1.68\times10^{-17}$), rs507139 (CPASSOC $p=4.33\times10^{-10}$), rs4888370 (CPASSOC $p=3.59\times10^{-15}$), rs117870289 (CPASSOC $p=1.64\times10^{-7}$) and

rs10453460 (CPASSOC p=0). Genes situated near these signals include *DDR2* (Figure 6.5.3.3), *SIRT3* (Figure 6.5.3.4), *ITGA11* (Figure 6.5.3.5), *CFDP1* (Figure 6.5.3.6), *ERG* (Figure 6.5.3.7) and *WNT7B* (Figure 6.5.3.8).

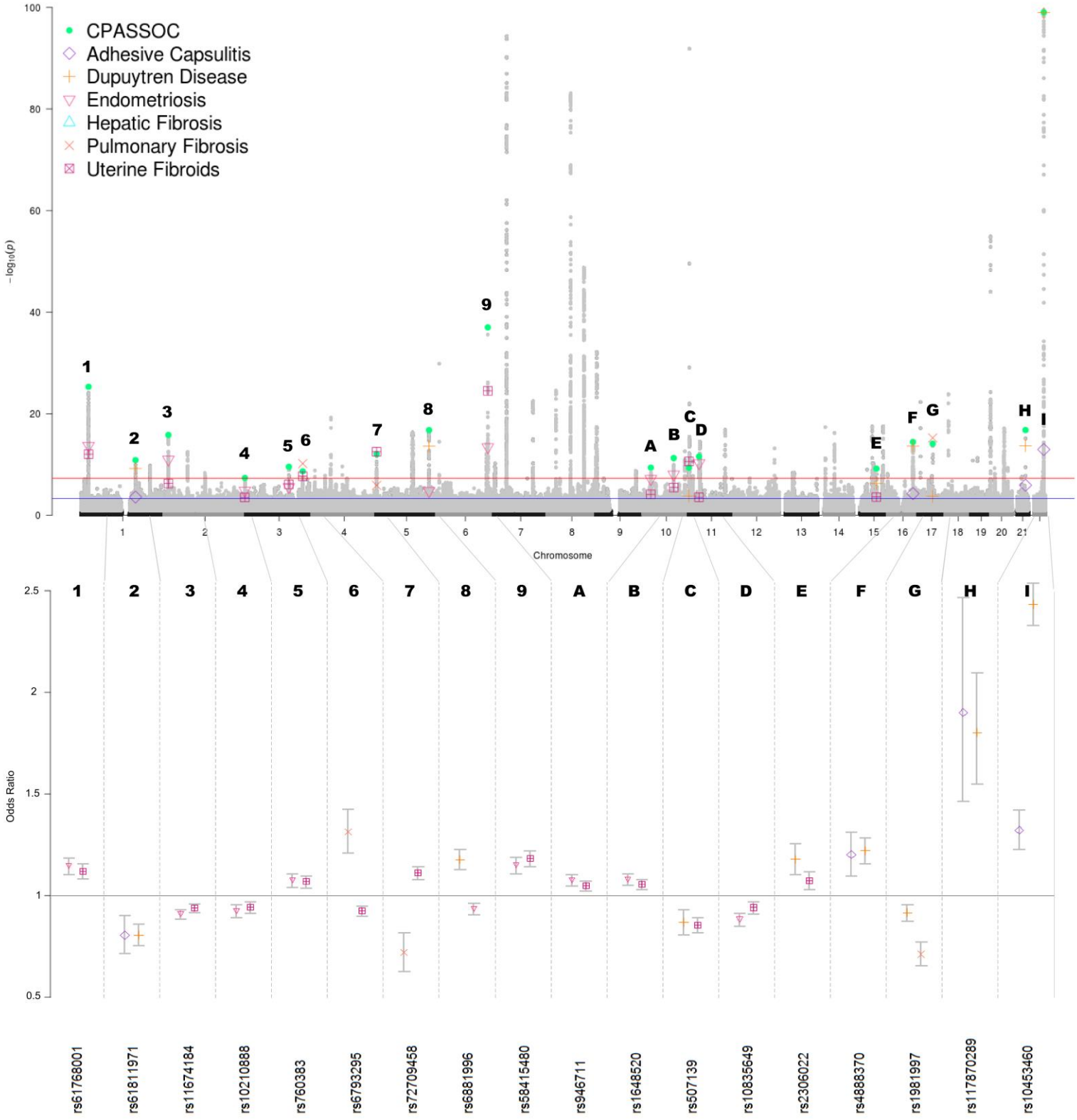



Figure 6.5.3.1 Top) Manhattan plot demonstrating the location of cross-phenotypic associations between different fibrotic conditions. Each filled grey circle was a variant. Lead significant

Figure 6.5.3.1 cont'd) associations that were reported in the genetic correlation analysis were coloured green and were accompanied by the significance of the contributing fibrotic condition variant. Bottom) Forest plot of each cross fibrotic associated SNP from individual GWAS, using the same legend as shown in the Manhattan plot. The points represent the odds ratio of effect of the SNP, and the error bars 95% confidence intervals of the odds ratio.

Table 6.5.3.1 Significant genetic correlations between more than one fibrotic conditions. Highlighted in bold are four loci whose association is below the threshold for genome-wide significance in the individual GWAS, but is significant in the cross-fibrotic analysis. AC - Adhesive capsulitis; DUP - Dupuytren disease; ENDO - Endometriosis; PF - Pulmonary fibrosis; UF - uterine fibroids.

| rsID | Chromosome | Position | CPASSOC p | Condition 1 | p | Condition 2 | p |
|---|---|---|---|---|---|---|---|
| rs61768001 | 1 | 22465820 | $4.94X10^{-26}$ | ENDO | $2.03X10^{-14}$ | UF | $9.20X10^{-13}$ |
| rs61811971 | 1 | 162649626 | $1.33X10^{-11}$ | AC | $2.50X10^{-04}$ | DUP | $6.15X10^{-10}$ |
| rs11674184 | 2 | 11721535 | $1.52X10^{-16}$ | ENDO | $1.02X10^{-11}$ | UF | $5.40X10^{-07}$ |
| **rs10210888** | **2** | **239901835** | **$4.40X10^{-08}$** | **ENDO** | **$1.31X10^{-05}$** | **UF** | **$3.10X10^{-04}$** |
| **rs760383** | **3** | **128119565** | **$2.79X10^{-10}$** | **ENDO** | **$2.59X10^{-06}$** | **UF** | **$8.10X10^{-07}$** |
| rs6793295 | 3 | 169518455 | $2.17X10^{-09}$ | PF | $6.63X10^{-11}$ | UF | $2.00X10^{-08}$ |
| rs73214051 | 4 | 7935754 | $1.48X10^{-09}$ | AC | $4.90X10^{-04}$ | DUP | $1.74X10^{-08}$ |
| rs72709458 | 5 | 1283755 | $8.37X10^{-13}$ | PF | $1.27X10^{-06}$ | UF | $3.00X10^{-13}$ |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| rs6881996 | 5 | 157888115 | $1.68X10^{-17}$ | DUP | $2.37X10^{-14}$ | ENDO | $1.60X10^{-05}$ |
| rs58415480 | 6 | 152562271 | $9.86X10^{-38}$ | ENDO | $3.63X10^{-14}$ | UF | $3.20X10^{-25}$ |
| **rs946711** | **10** | **21806832** | **$4.13X10^{-10}$** | **ENDO** | **$6.84X10^{-08}$** | **UF** | **$7.50X10^{-05}$** |
| rs1648520 | 10 | 90143657 | $5.19X10^{-12}$ | ENDO | $8.70X10^{-09}$ | UF | $3.30X10^{-06}$ |
| rs507139 | 11 | 225196 | $4.33X10^{-10}$ | DUP | $1.55X10^{-04}$ | UF | $2.50X10^{-11}$ |
| rs10835649 | 11 | 30329175 | $2.55X10^{-12}$ | ENDO | $5.22X10^{-11}$ | UF | $2.90X10^{-04}$ |
| **rs2306022** | **15** | **68628163** | **$6.42X10^{-10}$** | **DUP** | **$5.59X10^{-07}$** | **UF** | **$2.50X10^{-04}$** |
| rs4888370 | 16 | 75308230 | $3.59X10^{-15}$ | AC | $5.20X10^{-05}$ | DUP | $2.34X10^{-14}$ |
| rs1981997 | 17 | 44056767 | $7.76X10^{-15}$ | DUP | $1.78X10^{-04}$ | PF | $5.88X10^{-16}$ |
| rs117870289 | 21 | 39983448 | $1.64X10^{-17}$ | AC | $1.20X10^{-06}$ | DUP | $1.92X10^{-14}$ |
| rs10453460 | 22 | 46366165 | 0 | AC | $1.10X10^{-13}$ | DUP | 0 |

Table 6.5.3.2 Table showing the reference alleles, OR and 95% confidence intervals of all 18 cross fibrotic significant associations. The contributing conditions 1 and 2 of each associated SNPs are the same as in table 6.3.3.1. Highlighted in bold are four loci whose association is below the threshold for genome-wide significance in the individual GWAS, but is significant in the cross-fibrotic analysis.

| | | | | **Condition 1** | | **Condition 2** | |
|---|---|---|---|---|---|---|---|
| **rsID** | **Chromosome** | **Position** | **ref. allele** | **OR** | **95% CI** | **OR** | **95% CI** |
| rs61768001 | 1 | 22465820 | C | 1.14 | 1.1-1.19 | 1.12 | 1.08-1.15 |
| rs61811971 | 1 | 162649626 | T | 0.8 | 0.71-0.9 | 0.8 | 0.75-0.86 |
| rs11674184 | 2 | 11721535 | G | 0.91 | 0.88-0.93 | 0.94 | 0.91-0.96 |
| **rs10210888** | **2** | **239901835** | **T** | **0.92** | **0.89-0.95** | **0.94** | **0.91-0.97** |
| **rs760383** | **3** | **128119565** | **G** | **1.07** | **1.04-1.11** | **1.07** | **1.04-1.1** |
| rs6793295 | 3 | 169518455 | C | 1.31 | 1.21-1.43 | 0.92 | 0.9-0.95 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| rs72709458 | 5 | 1283755 | T | 0.72 | 0.63-0.82 | 1.11 | 1.08-1.14 |
| rs6881996 | 5 | 157888115 | T | 1.18 | 1.13-1.23 | 0.93 | 0.9-0.96 |
| rs58415480 | 6 | 152562271 | G | 1.15 | 1.11-1.19 | 1.18 | 1.14-1.22 |
| **rs946711** | **10** | **21806832** | **C** | **1.07** | **1.04-1.1** | **1.05** | **1.02-1.07** |
| rs1648520 | 10 | 90143657 | T | 1.08 | 1.05-1.11 | 1.05 | 1.03-1.08 |
| rs507139 | 11 | 225196 | A | 0.87 | 0.81-0.93 | 0.85 | 0.82-0.89 |
| rs10835649 | 11 | 30329175 | G | 0.88 | 0.85-0.91 | 0.94 | 0.91-0.97 |
| **rs2306022** | **15** | **68628163** | **T** | **1.18** | **1.1-1.25** | **1.07** | **1.03-1.12** |
| rs4888370 | 16 | 75308230 | T | 1.2 | 1.1-1.31 | 1.22 | 1.16-1.28 |
| rs1981997 | 17 | 44056767 | A | 0.91 | 0.87-0.95 | 0.71 | 0.65-0.77 |
| rs117870289 | 21 | 39983448 | T | 1.9 | 1.46-2.47 | 1.8 | 1.55-2.1 |
| rs10453460 | 22 | 46366165 | T | 1.32 | 1.23-1.42 | 2.43 | 2.33-2.54 |

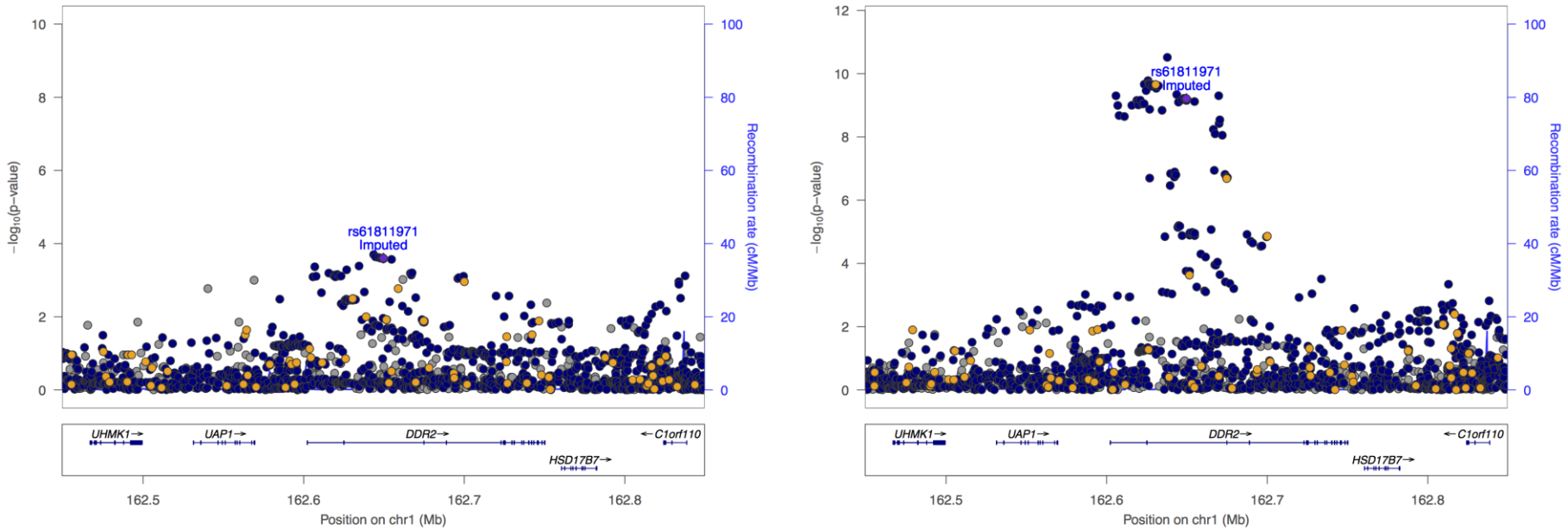



Figure 6.5.3.2 Regional plot showing the location of cross fibrotic association between left) adhesive capsulitis and right) Dupuytren disease on chromosome 1 at rs61811971. Yellow dot indicates the variant was genotyped and blue dot indicates the variant was imputed.

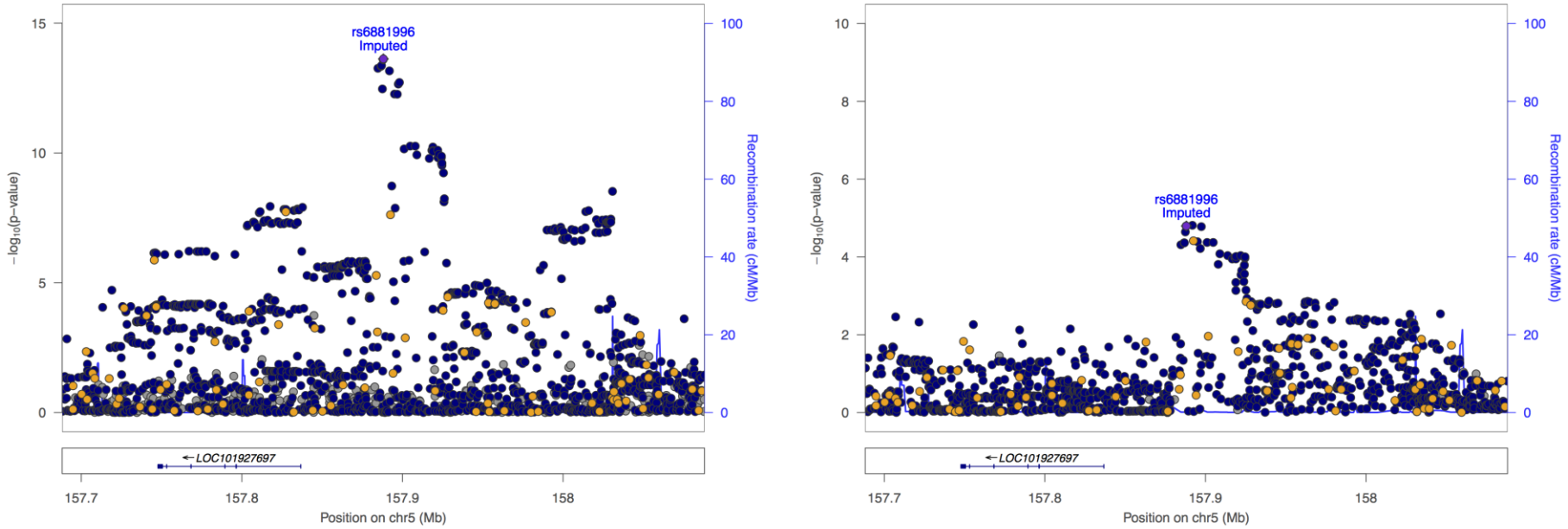



Figure 6.5.3.3 Regional plot showing the location of cross fibrotic association between left) Dupuytren disease and right) endometriosis on chromosome 5 at rs6881996. Yellow dot indicates the variant was genotyped and blue dot indicates the variant was imputed.

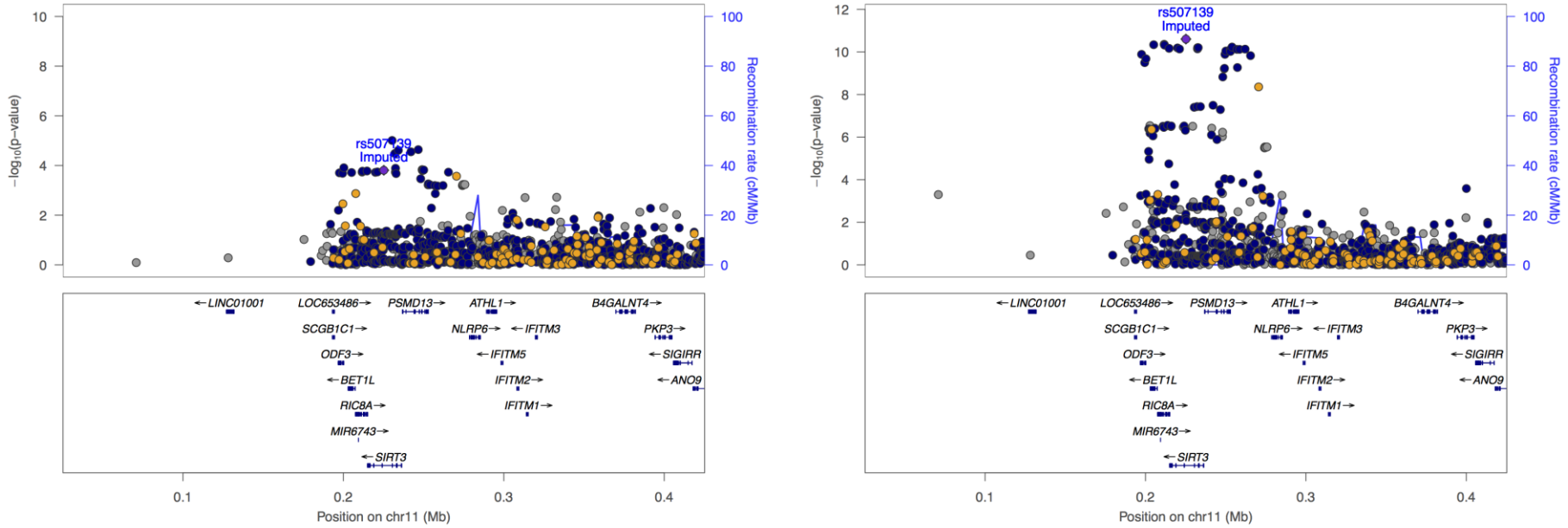



Figure 6.5.3.4 Regional plot showing the location of cross fibrotic association between left) Dupuytren disease and right) uterine fibroids on chromosome 11 at rs507139. Yellow dot indicates the variant was genotyped and blue dot indicates the variant was imputed.

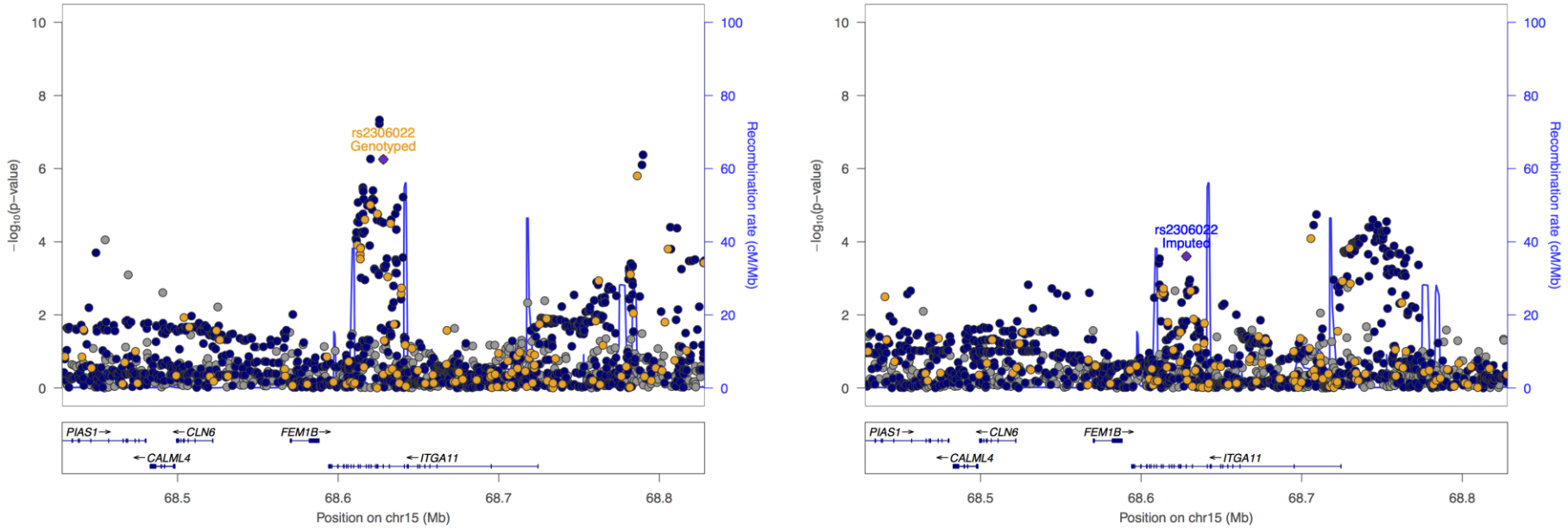



Figure 6.5.3.5 Regional plot showing the location of cross fibrotic association between left) Dupuytren disease and right) uterine fibroids on chromosome 15 at rs2306022. Yellow dot indicates the variant was genotyped and blue dot indicates the variant was imputed.

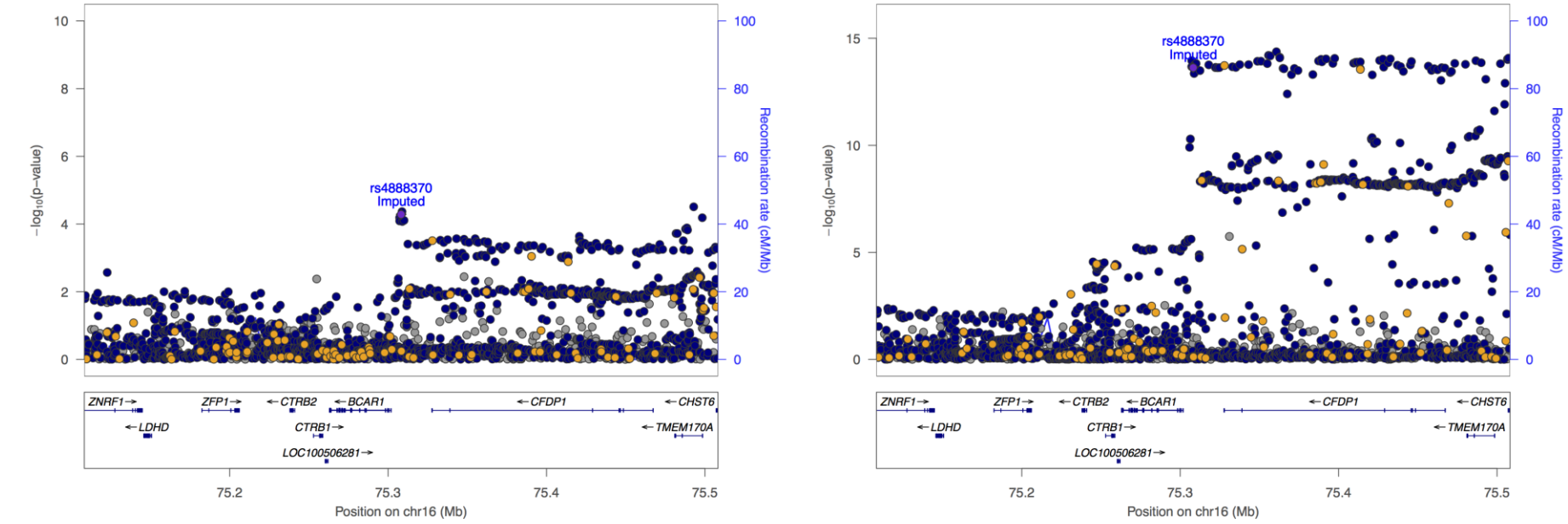



Figure 6.5.3.6 Regional plot showing the location of cross fibrotic association between left) adhesive capsulitis and right) Dupuytren disease on chromosome 16 at rs4888370. Yellow dot indicates the variant was genotyped and blue dot indicates the variant was imputed.

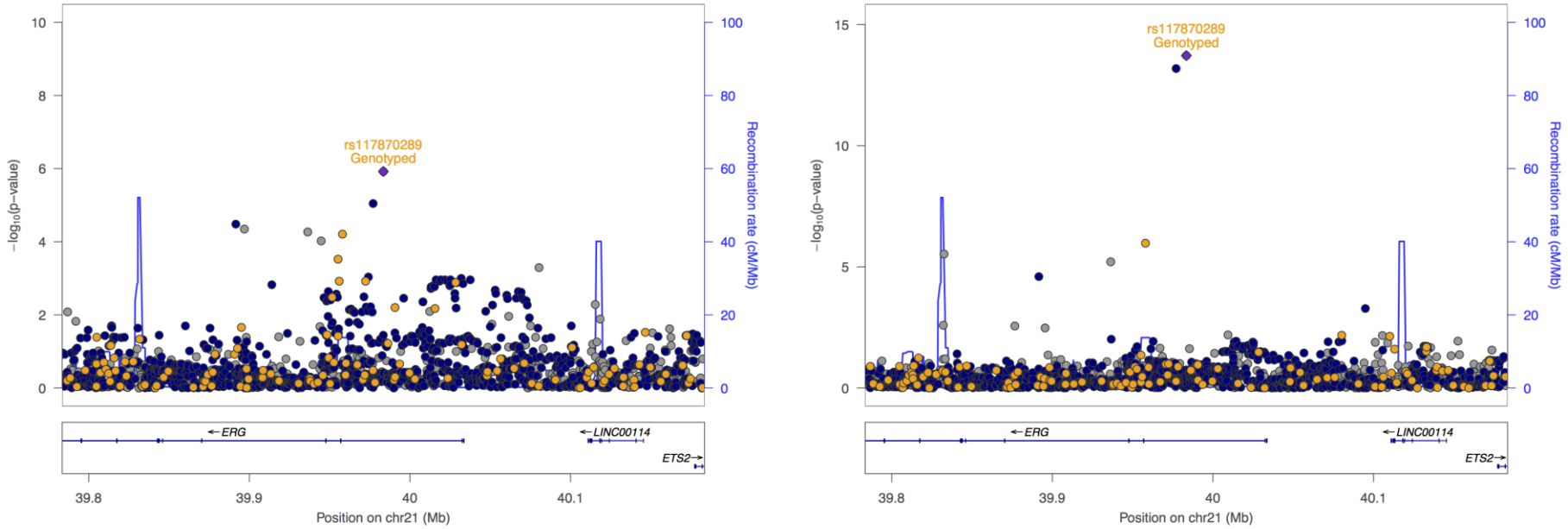



Figure 6.5.3.7 Regional plot showing the location of cross fibrotic association between left) adhesive capsulitis and right) Dupuytren disease on chromosome 21 at rs117870289. Yellow dot indicates the variant was genotyped and blue dot indicates the variant was imputed.

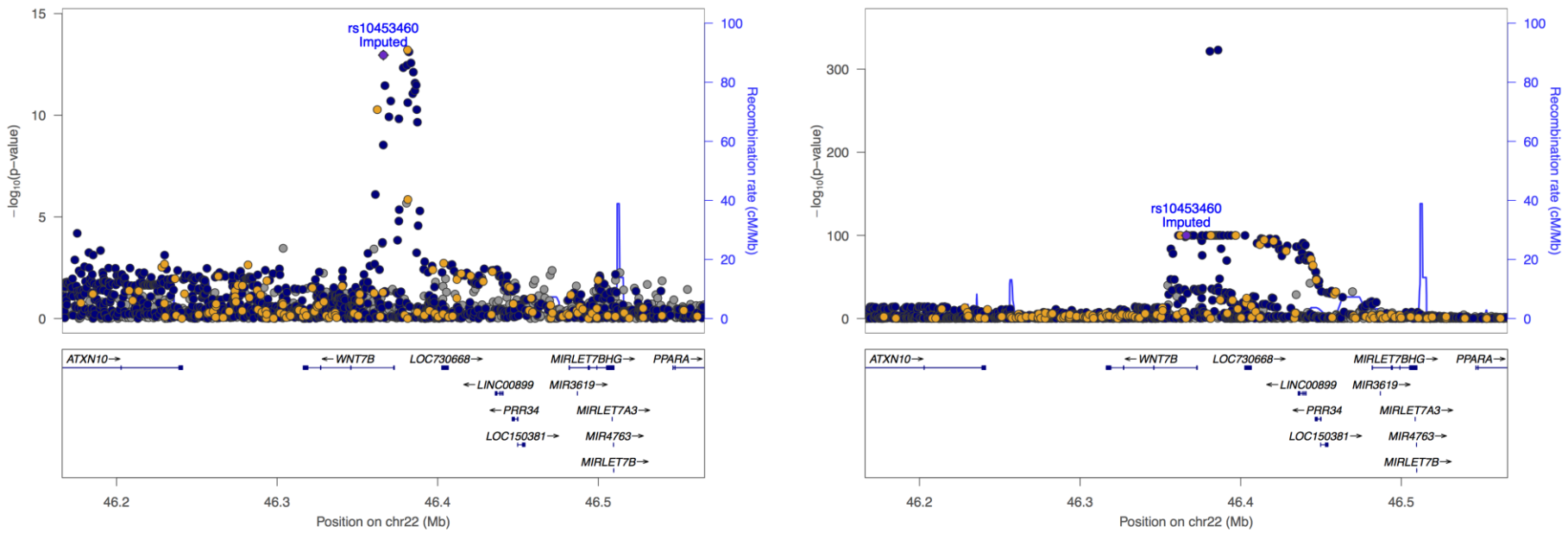



Figure 6.5.3.8 Regional plot showing the location of cross fibrotic association between left) adhesive capsulitis and right) Dupuytren disease on chromosome 22 at rs10453460. Yellow dot indicates the variant was genotyped and blue dot indicates the variant was imputed.

### 6.5.4 *In silico* analysis of the cross-phenotype associations

*In silico* analysis was performed in three stages, first using 3DSNP to filter out SNPs that have no functions, then virtual 4C to identify the predicted interacting gene(s). The cell types that express these genes were determined using scRNAseqdb and GTEx. Additional eQTL expression was predicted using GTEx. I have identified nine prioritised candidate SNPs, including the *DDR2* region, for further functional work. These SNPs (Table 6.5.4.1) are in the same associated loci as the lead SNP with genome-wide significance.

Most of the cross fibrotic associated SNPs are predicted to interact with enhancer or promoter regions of multiple genes (Table 6.5.4.1, columns 3-5). Analysis using 3DSNP program showed a total of 124 associated variants spreading across 14 loci are located within either an enhancer or a promoter. In addition, 11 loci were also predicted to affect a transcription factor binding site, and 9 of which with specific sequencing motifs being disrupted. For example, SNP rs61811977, 16844 bases downstream of rs61811971, and in very tight LD with the lead SNP ($r^2$=0.9), is an intronic variant. An A to T change of this SNP is predicted to alter the binding site of SP1 transcription factor. This site is in an enhancer region of *DDR2*, which is highly expressed in fibroblasts and myofibroblasts, and interacts with the *DDR2* promoter region (see Chapter 3).

Chromosome 3 variant rs12637184, which is in LD with rs6793295 ($r^2$=0.92), is located in the promoter region of *ACRT3*. Virtual 4C analysis has revealed that this SNP interacts with the promoter region of *LRRC34, MYNN, TERC* and other genes within the 1Mb region (Figure 6.5.4.1). Using eQTL data, I was able to confirm this variant is associated with the expression of *LRRC34* in myofibroblasts, and *MYNN* in cells from testicular tissue.

SNP rs71575922, is located within the intronic, enhancer region of one of the *SYNE1* splicing variants. Virtual 4C and chromatin anchor to anchor analysis revealed an interaction with the upstream enhancer regions of *SYNE1*, and a long-range interaction with the promoter of *MYCT1*.

SNP rs4071558, in tight LD with rs10835649 ($r^2$=1), is in the promoter region of *ARL14EP*. Analysis using the eQTL database has revealed an interaction with *ARL14EP* expression in myofibroblasts and lung fibroblasts.

SNP rs4888379, 32kb downstream of rs4888370, is located in the first intron of *CFDP1*. Virtual 4C analysis has demonstrated interactions with an enhancer region of *CFDP1* and the promoter region of *CHST6, BCAR1 and* many other genes up to 400kb downstream. *CFDP1* is reported highly expressed in myofibroblasts in the eQTL database.

SNP rs28971325, in tight LD with rs10453460 ($r^2$=0.89), is in the promoter region of *WNT7B* chromosome 22. On virtual 4C, this region interacts with the enhancer and promoter regions of genes in the surrounding 1Mb area, highly suggesting this might be a Locus Control Region (LCR).

Table 6.5.4.1 Table showing a summary of the cross-fibrotic associated SNPs and their interaction with genes nearby. The Enhancer/Promoter column shows the number of SNPs that are situated in either an enhancer or a promoter region. TFBS column indicates the number of SNPs that are within a transcription factor binding site, and the motif column indicates the number of SNPs that can disrupt a TF binding sequence motif. The next four column shows, the genes (enhancer or promoter) the queried SNP is interacting with, which cell type are the genes discovered in eQTL expressing, associated genes discovered in eQTL, and the SNP used for doing the query.

| Chromosome | CPASSOC snp | Enhancer/Promoter | TFBS | Motif | Virtual 4C Interacting genes | Fibrosing relevant cells | eQTL | Queried SNP |
|---|---|---|---|---|---|---|---|---|
| 1 | rs61768001 | 21 | 6 | 9 | *WNT4,CDC32,CELA3A,CELA3B,LINC00039* | Macrophage, B cell | *WNT4/NBPF2P* | rs2268177 |
| 1 | rs61811971 | 6 | 3 | 0 | *DDR2,HSD17B7* | Lung fibroblast, myofibroblast | - | rs61811977 |
| 2 | rs11674184 | 16 | 6 | 3 | *GREB1,GREB2,E2F6,NTSR2* | - | - | rs9789525 |
| 2 | rs10210888 | 0 | 0 | 0 | - | - | - | - |
| 3 | rs760383 | 0 | 0 | 1 | *EEFSEC,SEC61A1* | Myofibroblast | *EEFSEC,SEC61A1* | rs6785384 |
| 3 | rs6793295 | 3 | 2 | 1 | *ACTRT3,TERC,MYNN,LRRC34* | Myofibroblast | *LRRC34,MYNN* | rs12637184 |
| 4 | rs73214051 | 1 | 0 | 0 | *AFAP1* | Lung fibroblast | *AFAP1* | rs73214051 |
| 5 | rs72709458 | 0 | 0 | 0 | - | - | - | - |
| 5 | rs6881996 | 1 | 0 | 0 | *EBF1* | - | - | rs2946160 |
| 6 | rs58415480 | 8 | 3 | 2 | *ESR1,MYCT1,SYNE1* | - | - | rs71575922 |
| 10 | rs946711 | 4 | 2 | 1 | *MTTL10,DNAJC1,CASC10* | Foreskin fibroblast | *CASC10* | rs946711 |
| 10 | rs1648520 | 3 | 3 | 0 | *PTEN,LIPF,LIPK* | - | - | - |
| 11 | rs507139 | 5 | 1 | 8 | *NLRP6,BET1L,ATHL1,IFITM5,SIRT3* | T cell, B cell | *BET1L* | rs11551933 |
| 11 | rs10835649 | 2 | 3 | 0 | *ARL14EP,KCNA4,MPPED2* | T cell, Foreskin fibroblast | *ARL14EP* | rs4071558 |
| 15 | rs2306022 | 1 | 0 | 0 | *CLN6, ANP32A* | - | - | - |
| 16 | rs4888370 | 33 | 4 | 7 | *BCAR1,CTRB2,CHST6,CFDP1* | Myofibroblast | *CFDP1* | rs4888379 |
| 17 | rs1981997 | 0 | 0 | 0 | - | - | - | - |
| 21 | rs117870289 | 0 | 0 | 0 | - | - | - | - |
| 22 | rs10453460 | 20 | 10 | 12 | *WNT7B,PPARA,CELSR1,LINC00899* | - | - | rs28971325 |

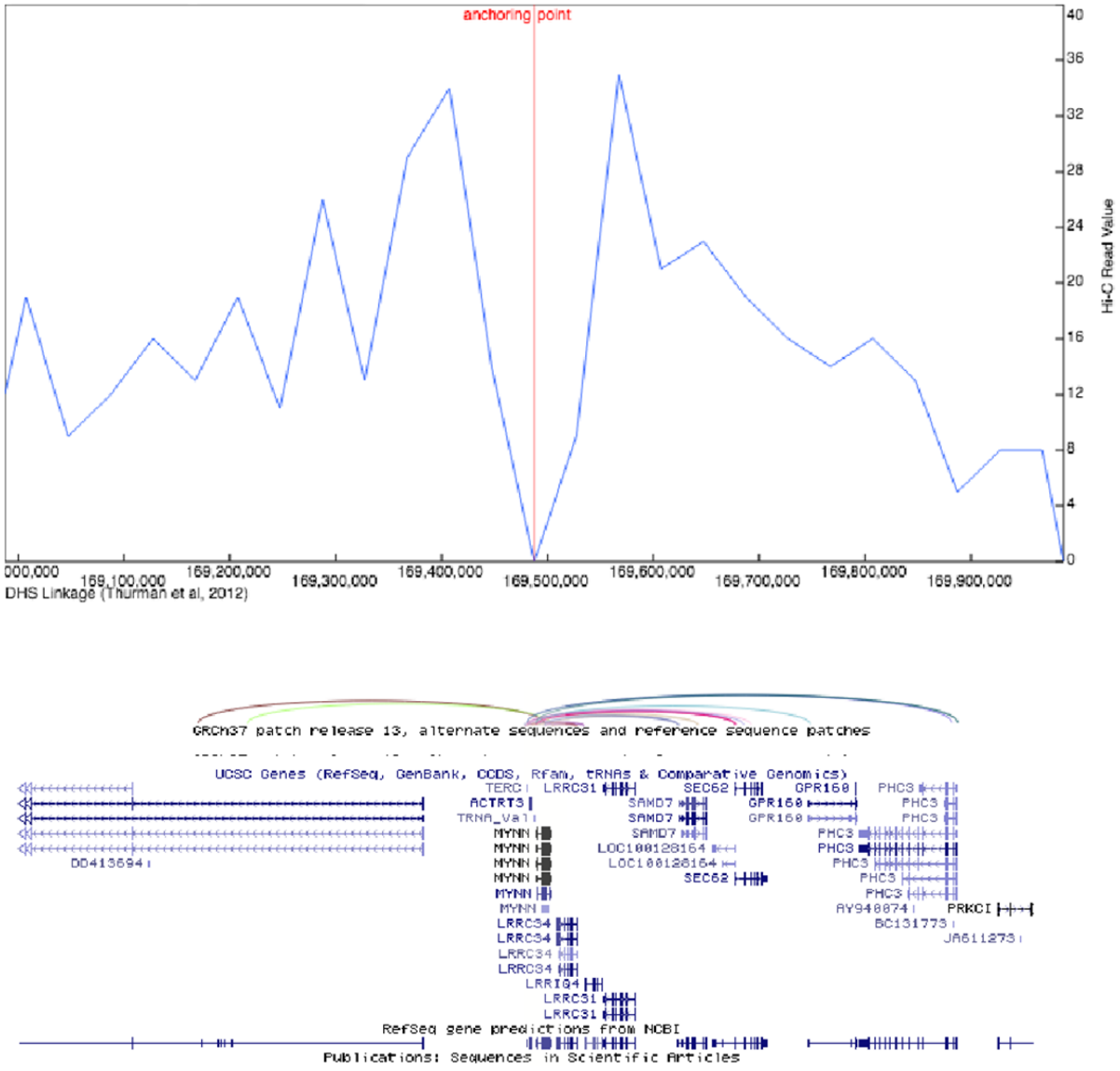



Figure 6.5.4.1 Top panel) Virtual 4C plot showing the distribution of Hi-C read counts 500kb upstream and downstream of the anchor point rs12637184. Bottom panel) DHS- linkage interaction plot of the same region. Both plots suggested that rs12637184 interacts with multiple genes including *LRRC34, MYNN, TERC* and many more within the 1Mb region.

## 6.6 Discussion

Due to an increase in sample size and the integration of multiple cohorts, novel associations have been identified for idiopathic pulmonary fibrosis, Dupuytren disease, endometriosis and uterine fibroids. I have also identified the first ever genetic variant associated with adhesive capsulitis. With regards to the association analysis of systemic sclerosis and hepatic fibrosis, I did not discover any new signals because the number of cases of both conditions in the analysis were relatively small, and as a result there wasn't any significant increase in power over previously published studies. As for the imputed-only genome wide significant variants, more statistically powered replications with a design partially targeting those regions are necessary to confirm their validity.

In terms of the cross fibrotic genetic analysis, interactions were observed between six out of the eight tested conditions but mainly between adhesive capsulitis and Dupuytren disease; Dupuytren disease and pulmonary fibrosis; endometriosis and uterine fibroids; and between uterine fibroids and pulmonary fibrosis. A total of 18 cross-fibrotic associations were identified.

As most associated variants are either intronic or intergenic, various multiple putative interactions including promoter-promoter, promoter-enhancer and enhancer-enhancer

were revealed in almost every associated region in my *in-silico* analysis, which was entirely expected for fibrotic conditions, and is typical of the results of most GWAS.

Unlike text-based gene prioritisation analysis methods such as GRAIL [202] rather than predicting the variant's functionality using primarily the number of appearances of genes surrounding the associated regions in the literature, a more hands-on, in-depth method, which has a similar workflow to FUMA [196] and GREGOR [197], was used in my analysis. The only disadvantage of this approach is that it was difficult to automate. This workflow gave me the freedom to use the most appropriate epigenomic and transcriptomic databases to identify DNA-DNA interactions and variant associated RNA expression enrichment in the most relevant tissue types to fibrotic disease. This should in future increase the chances of success in follow-up functional work built upon on the results of my *in silico* analysis.

Performing pathway enrichment analysis without the necessary information in the epigenomic regulation of the underlying genes can lead to time wasted on researching genes or proteins that are irrelevant to the disease. It is clear from the virtual 4C *in silico* analysis that most causal variants are in regions that interact and regulate the transcription of more than one gene.

Future functional work should be planned accordingly to maximise the chances of a positive outcome. For example, an initial experiment using CRISPR editing followed by expression array of appropriate genes may be more suitable for variants that

interact with multiple promoter regions. Finally, an appropriate cell type that has been disaggregated from tissue of the disease of interest is required to support the validity of the functional work, as TF binding and chromatin modification act in a tissue specific manner [287].

Similar to the results of most GWAS in complex disease, all the cross-phenotypic fibrotic associations in our analysis were either intronic or intergenic. The function of such variants has been predicted to be altering the level of RNA or protein expression, rather than a structural conformational change of a protein. This is because the latter often leads to a severe reduction in function of the protein and the respective physiological system, seen for example in monogenic diseases such as cystic fibrosis [288].

One clear example where the likely causal variant was intronic or intergenic in my analysis was the association between adhesive capsulitis and Dupuytren disease on chromosome 16 at rs4888379. The chromatin open region that was associated with this variant is widely enriched in most cell types of the human body, ranging from hepatic cells to skin cells. This indicates the crucial role of the genes that this region regulates. Not surprisingly, the protein functions of one of the nearby genes is involved in the basic structural genomic regulation of a cell. *CFDP1* is highly conserved in various species including mouse, chicken and zebrafish [289]. Other causal SNPs that have been previously identified within this region have been associated with

atherosclerosis. CFDP1 is a protein that belongs to the evolutionarily highly conserved Bucentaur (BCNT) family. Function of CFDP1 involves governing the organization of chromosomes [289]. CFDP is also crucial for craniofacial development and osteogenesis. Further work is required to understand the role this gene may play in fibrosis.

Chromosome 1 at rs61811977 was another association shared between adhesive capsulitis and Dupuytren disease. Although there isn't a direct association between this variant, the conditions and *DDR2* expression, there is a good amount of epigenomic and phenotypic evidence, reported in chapter 3 and 4 supporting the possible expression altering effect of this variant to *DDR2*, as well as demonstrating the molecular function of DDR2 in the development of fibrosis. This cross phenotypic association further cemented the idea that interaction between myofibroblasts and the extracellular matrix is crucial for fibrotic development, specifically contraction and cell migration, in both shoulder capsule and palmar fascia.

Disappointingly, no genetic association was found between three or more different fibrotic conditions, which agreed with the findings of my phenotypic correlation analysis. One clear explanation is that the gene expression controlled by associated variants is likely to be highly tissue and cell type specific. Using the endometriosis and uterine fibroids associated region on chromosome 6 as an example, the alternate allele of rs71575922 has been predicted to affect *ESR1* expression. This finding is in

line with other associated intronic variants within this region that have been reported in endometriosis and uterine fibroids [290 291] and breast cancer genetic studies [292]. *ESR1* possesses a tissue-dependent and differentially methylated region (T-DMRs) [293]. Its protein form, estrogen receptor alpha (ERα), is a nuclear receptor for 17-β estradiol and is expressed in smooth muscle cells and fibroblasts within the endometrial lining [294]. It governs the proliferation of endometrium. There are two types of estrogen receptors, ERα and ERβ, which can form homodimer ERαα, ERββ and heterodimer ERαβ. A change in ratio of ERα and ERβ has been demonstrated in endometriosis in the literature [295]. Similar expression patterns have also been reported in uterine fibroids, where ERα has been demonstrated to be highly expressed in these tumours. The resident cell types of uterine fibroids again comprise of smooth muscle cells and fibroblasts, suggesting a similar aetiology to endometriosis. One theory is that uterine fibroids are caused by hypoxia of the blood vessels in endometrium lining, leading to chronic inflammation. However, in both fibroses, aberrant ERα activation could also play a role in their aetiology. Estrogen receptors have different functions in different cell types including B-cell lymphopoiesis in the bone marrow, which can alter bone density, as well as thymus differentiation. This highlights the tissue specific nature of the SNP leading to expression change of *ESR1*. Myofibroblasts in different organs have different origins and thus likely to require a different trigger to be fibrotic. This has been demonstrated in a study, where the differentiation from fibroblasts to myofibroblasts was triggered by the activation of ERα[296].

However, *ESR1* was not the only gene that was associated with rs71575922 in my *in silico* analysis. *SYNE1* has been demonstrated to indirectly interact with an intronic region of a splice variant of *ESR1*. *SYNE1* has been reported to be highly expressed in placenta and SMC. Mutated *SYNE1* has been reported in various conditions including colon cancer, esophageal cancer, cerebellar ataxia, muscular dystrophy. *SYNE1* translates into Nesprin-1, a protein that plays a role in linking the nuclear envelope to the cytoskeleton and nuclear lamina. *SYNE1* knockout EDMD derived fibroblast cells have been shown to be less adherent and migratory than corresponding wild type cells [68]. In contrast, the *SYNE1* expression in endometriosis and uterine fibroids patients, who are carrying the risk alleles of this variant were predicted to be elevated, potentially leading to a more contractile and migratory fibrotic phenotype.

Uterine fibroids have also been identified to be genetically correlated with endometriosis on chromosome 11 at rs40711558, and pulmonary fibrosis on chromosome 3 at rs12637184. Virtual 4C and eQTL *in-silico* analysis have revealed that rs12637184 is in the promoter of *ARCT3*, which interacts with *LRCC34*. *LRCC34* is selectively expressed in fibroblasts and lymphoid cells, and highly conserved in mammals. No known function of this gene has been reported in the literature to date. Both genes seem to play a role in fibrosis in tandem. *ARL14EP* expression is associated with the risk allele of rs40711558. It is translated into an effector protein that forms a complex with a GTPase ARL14/ARF7. The known function of this complex is to control the export of MHC class II vesicles along the actin cytoskeleton in dendritic

cells. Not surprisingly, this gene is expressed in most tissues but only in a handful of cell types, which included endothelial cells. In a pulmonary fibrosis mouse model, different subsets of dendritic cells have been demonstrated to have a differential effect on fibrosis. CD11b+ dendritic cells had a higher turnover of metalloproteinases MMP-2 and -7, indicating an ECM regulatory role compared to CD103+ dendritic cells. Interestingly, both subsets were upregulated by TGF-β[297]. MHC class II presenting is crucial in fibrosis development. One study has shown that MHCII- αSMA hepatic stellate cells have a more myofibroblast like phenotype than MHCII+ αSMA hepatic stellate cells. In endometriosis and uterine fibroids, a decrease in *ARL14* expression in endothelial cells could potentially lead to a more myofibroblast-like phenotype.

As hypothesised, different fibrotic diseases have some shared genetic predisposition. My phenotypic and genetic correlation analysis has successfully revealed multiple correlations between them. Some of the likely associated genes have clear functions and may be targetable by inhibitors, such as DDR2 (Chapter 3). Others require further molecular genetic work to understand their functions. Patients carrying variants associating with multiple fibrotic conditions do not necessary develop symptoms, as multiple fibrosis hits are required to express the phenotype, e.g. inflammation, toxin, carrying the risk allele of a second causal SNP, etc. However, this analysis has revealed potentially important biological pathways in fibrosis. Further functional work is required to find new potential therapeutic targets in fibrotic disease.

# Chapter 7 Discussion and future directions

Although DD is a genetic disease, most current research is driven by existing knowledge in fibrosis from a cell biology viewpoint. Research interests in this field have long been divided into two major groups; matrix remodelling and immunology. As a result, the current view of DD is heavily biased towards these areas. The main goal of this thesis was to reduce this preconception by providing an unbiased genetic data to the DD community.

## Conclusion 1: Associated variants contribute to the disease by potentially influencing several genes simultaneously

One of the interesting observations from the GWAS, cross-fibrosis and the epigenomic chapters (Chapters 2, 4, 6) is that gene regulation under almost all circumstances does not happen independently but as a group of two, three or more. Genes that have no known biological relevance to fibrosis or relationship between one another are transcribed concurrently under the regulation of a single region. There are several examples of these gene sets in our list of DD associated variants, e.g. *SFRP4* and *EPDR1*, *DDR2* and *HSD17B7*, etc. In all cases, it questioned our reasoning of investigating the biological function of just one gene from the pair. It is evident that from the summary of my *in silico* analyses, and later the open chromatin and 4C work

on the *DDR2* region in chapter 4, this single regulatory region can be controlled by a single DD associated variant. The functional work I have performed thus far, albeit already showing interesting results regarding SFRP4 and DDR2, only answered some of the biology around these regions. Further work should be first focused extensively on the basic molecular function of EPDR1 and HSD17B7 in a DD setting and then on dissecting their cellular signalling pathways to reveal upstream effectors and possible therapeutically targetable receptors.

## Conclusion 2: DD undoubtedly is a complex disease

Further focusing on the aspect of gene regulation in DD, again using *SFRP4* and *EPDR1* as an example, it is evident that these two genes are controlled by a single region and their transcription level is vastly different, most likely due to a competition in the recruitment of TF and transcription machinery. The difference at the inter-gene level between cells carrying the risk and protective alleles at the associated variant, however, is comparatively subtle. This observation together with the location of the associated regions, where the majority are situated within intronic and intergenic regions, emphasises the aetiology of DD as a progressive delicate imbalance of many genes, befitting the term complex disease. Again, it would be interesting to find out how the protein products of some or all of these genes converge into a common pathway, which is response for initiating or perpetuating the development of DD.

## Conclusion 3: Anti-inflammatory agents are unlikely to be fully effective in treating DD

It is evident that DD tissue and cells are populated with a several immune cell types and cytokines [113]. Numerous immunological studies have shown fibrosis associated genes being transcribed because of an increase in inflammation or just the existence of a lymphocyte subtype, leading to the development of typical fibrosis phenotypes, such as an increase in myofibroblast differentiation [46]. Although it is probable that inflammation is a contributor to the development of fibrosis, my GWAS results did not reveal a plethora of inflammatory related genes. Indeed, the gene list derived from the GWAS and the ATAC-seq results (with the caveat that I examined for differences between fibroblast cell subtypes) only suggest a minor inflammatory component. The majority are involved in developmental pathways, gene transcription, TF binding, and hormonal pathways. The WNT developmental pathway has been suggested to play a role in fibrosis development without the support of inflammation [44], but since then no further research in this area has been performed. WNT signalling genes have appeared on numerous occasions during my research in this thesis; in the GWAS and cross-fibrosis analysis; in the results of the *in silico* analysis; and in the differential gene expression after *DDR2* knockdown. This repeated appearance of WNT pathway components highlights its fundamental importance in the pathogenesis of DD.

Although inflammatory pathways do converge with WNT pathways, transduced inflammatory signals don't always override WNT signalling. Further understanding the biology of each WNT component that is associated with DD would yield interesting biological results, but is unfortunately unlikely to lead to drug discovery in the near future due to the complexity of the pathway, and its critical roles in normal tissue development. Other future work should therefore focus on dissecting the hormonal pathways as there are already effective existing hormone receptor inhibitors, or on the pathways that regulate the translocation of the DD associated TFs. Inflammation is one important aspect of DD pathophysiology, but doesn't fully address its biology, and therefore inhibition is unlikely to provide a fully effective therapy.

I have successfully demonstrated that DDR2 is an excellent therapeutic target to inhibit, showing a reduction in the manifestations of a DD phenotype in patient-derived myofibroblasts on knockdown of *DDR2*. However, the effect of this knockdown seemed to be different between collagen I and III, and the genes that are upregulated by collagen I could be beneficial for the completion of the resolution phase in fibrosis tissue. Therefore, an adjuvant that masks collagen binding sites might be required when using DDR2 inhibitors clinically. Further work should be concentrated on exploring this hypothesis.

## Conclusion 4: More high-resolution DD epigenomic regulatory data is required to help dissect out the function of disease associated variants

Dissecting the function of an associated variant is a challenging task, as presented in the literature, which is littered with potentially interesting variants across a wide spectrum of complex disease, without any molecular follow-up work. The lack of diseased tissue specific epigenomic data is partly responsible for this issue. My epigenomic work in chapter 4 has illustrated part of this problem by demonstrating that only a small proportion of differentially expressed genes correlate with their corresponding enrichment in an open chromatin region in the TSS and enhancer. Transcriptomic data alone can, therefore, only be useful if the study design takes into account the allelic frequency of the associated variants, and this can be costly especially if the effect sizes are small. It would be useful to possess this level of eQTL data as a validation, but lack of such granular data can be offset by the collection of epigenomic data.

My bulk 4C and ATAC-seq experiments have revealed many details of the transcription regulatory system, providing the much-needed information to bridge the gap between the DD associated variants, the DD transcriptome, and TF binding sites. Despite its usefulness, this data inherently possesses a lot of “noise” from my cell culture system, as well as from the variety of cellular subtypes present in the sample.

The advent of single cell sequencing has helped overcome this issue, and can provide a much more refined picture, whether at epigenomic level or transcriptomic level, of a targeted cell population [298]. The refined cell population can be selected for further functional work for “cleaner” results through cell sorting. Nevertheless, imminent future work should include the curation of single cell transcriptomic and epigenomic data.

## Conclusion 5: Genetic screening may soon be routine in clinic prior to surgery

Genotyping at genomic level has existed over two decades, which has led to the discovery of many variants associated with diseases. Despite this success, routine genetic testing only applies mostly to major conditions such as cancer, neurological disease, and birth defects. The diagnosis, pre-treatment counselling and treatment decision-making of many conditions such as DD and type I diabetes that affect tens of thousands in the population, can be improved by the incorporation of wGRS. I demonstrated that wGRS was able to predict recurrence of disease in two independent populations with different definitions of recurrence, independent of Dupuytren Diathesis features. As larger cohorts of DD patients are investigated, more associated variants will be discovered, and it is likely that the accuracy of prediction, not only of who will get DD, but also who will recur after surgical intervention, will improve. Furthermore, it is worth highlighting that our GRS is derived from SNPs detected by their association with the development of DD, not recurrence after surgery per se. As

cohorts increase in size, a between-cases GWAS analysis of those with recurrent disease versus those who do not recur might reveal further interesting insights into the biology of fibrosis.

In summary, in this thesis I have successfully completed a large GWAS of DD, revealing 17 new genetic variants that pre-dispose to disease. I have used a variety of epigenetic methods to define the genomic landscape of DD derived myofibroblasts and aid interpretation of the genetic data. I have studied in detail two genomic loci revealed by my GWAS analysis, and suggest that DDR2 might be a useful therapeutic target. In addition, I have derived a weighted genetic risk score for DD that is able to predict recurrence of disease independently of well-known clinical predictors of recurrence. Finally, I have explored some of the genetic overlap between DD and other tissue specific fibrotic diseases. I hope that this work will provide a platform for future studies that will further define the biology of fibrosis, and lead to the development of new therapies in DD and other disabling fibrotic diseases.

# Appendix Publications associated with this thesis

Ng M, Thakkar D, Southam L, Werker P, Ophoff R, Becker K, et al. A Genome-wide Association Study of Dupuytren Disease Reveals 17 Additional Variants Implicated in Fibrosis. Am J Hum Genet. 2017;101(3):417-27

Riesmeijer SA1, Manley OWG, Ng M, Nolte IM, Broekstra DC, Werker PMN, Furniss D. A Weighted Genetic Risk Score Predicts Surgical Recurrence Independent of High-Risk Clinical Features in Dupuytren's Disease. Plast Reconstr Surg. 2019 Feb;143(2):512-518

# Bibliography


1. Ball C, Izadi D, Verjee LS, et al. Systematic review of non-surgical treatments for early dupuytren's disease. BMC Musculoskelet Disord 2016;**17**(1):345 doi: 10.1186/s12891-016-1200-y[published Online First: Epub Date]|.
2. Hindocha S, McGrouther DA, Bayat A. Epidemiological evaluation of Dupuytren's disease incidence and prevalence rates in relation to etiology. Hand (N Y) 2009;**4**(3):256-69 doi: 10.1007/s11552-008-9160-9[published Online First: Epub Date]|.
3. Mansur HG, Oliveira ER, Goncalves CB. Epidemiological analysis of patients with Dupuytren's disease. Rev Bras Ortop 2018;**53**(1):10-14 doi: 10.1016/j.rboe.2017.12.003[published Online First: Epub Date]|.
4. Mitra A, Goldstein RY. Dupuytren's contracture in the black population: a review. Ann Plast Surg 1994;**32**(6):619-22
5. Yeh CC, Huang KF, Ho CH, et al. Epidemiological profile of Dupuytren's disease in Taiwan (Ethnic Chinese): a nationwide population-based study. BMC Musculoskelet Disord 2015;**16**:20 doi: 10.1186/s12891-015-0476-7[published Online First: Epub Date]|.
6. Buntine JA. Dupuytren's disease in Asia (Re: ANZ J. Surg. 2010; 80: 495-9). ANZ J Surg 2011;**81**(5):393 doi: 10.1111/j.1445-2197.2011.05725.x[published Online First: Epub Date]|.
7. Weinstein AL, Haddock NT, Sharma S. Dupuytren's disease in the Hispanic population: a 10-year retrospective review. Plast Reconstr Surg 2011;**128**(6):1251-6 doi: 10.1097/PRS.0b013e318230bf0f[published Online First: Epub Date]|.
8. Mandalia VI, Lowdon IM. Dupuytren's disease in a child: a case report. J Pediatr Orthop B 2003;**12**(3):198-9 doi: 10.1097/01.bpb.0000049560.52224.34[published Online First: Epub Date]|.
9. Bayat A, McGrouther DA. Management of Dupuytren's disease--clear advice for an elusive condition. Ann R Coll Surg Engl 2006;**88**(1):3-8 doi: 10.1308/003588406X83104[published Online First: Epub Date]|.
10. Ross DC. Epidemiology of Dupuytren's disease. Hand Clin 1999;**15**(1):53-62, vi
11. Dolmans GH, de Bock GH, Werker PM. Dupuytren diathesis and genetic risk. J Hand Surg Am 2012;**37**(10):2106-11 doi: 10.1016/j.jhsa.2012.07.017[published Online First: Epub Date]|.
12. Anthony SG, Lozano-Calderon SA, Simmons BP, et al. Gender ratio of Dupuytren's disease in the modern U.S. population. Hand (N Y) 2008;**3**(2):87-90 doi: 10.1007/s11552-007-9076-9[published Online First: Epub Date]|.
13. Descatha A, Bodin J, Ha C, et al. Heavy manual work, exposure to vibration and Dupuytren's disease? Results of a surveillance program for musculoskeletal disorders. Occup Environ Med 2012;**69**(4):296-9 doi: 10.1136/oemed-2011-100319[published Online First: Epub Date]|.
14. Descatha A, Carton M, Mediouni Z, et al. Association among work exposure, alcohol intake, smoking and Dupuytren's disease in a large cohort study (GAZEL). BMJ Open 2014;**4**(1):e004214 doi: 10.1136/bmjopen-2013-004214[published Online First: Epub Date]|.

15. Haines A, Levis C, Goldsmith CH, et al. Dupuytren's contracture and handwork: A case-control study. Am J Ind Med 2017;**60**(8):724-33 doi: 10.1002/ajim.22736[published Online First: Epub Date]|.
16. Kan HJ, Verrijp FW, Hovius SER, et al. Recurrence of Dupuytren's contracture: A consensus-based definition. PLoS One 2017;**12**(5):e0164849 doi: 10.1371/journal.pone.0164849[published Online First: Epub Date]|.
17. Couch H. Identical Dupuytren's Contracture in Identical Twins. Can Med Assoc J 1938;**39**(3):225-6
18. Hindocha S, Stanley JK, Watson S, et al. Dupuytren's diathesis revisited: Evaluation of prognostic indicators for risk of disease recurrence. J Hand Surg Am 2006;**31**(10):1626-34 doi: 10.1016/j.jhsa.2006.09.006[published Online First: Epub Date]|.
19. Larsen S, Krogsgaard DG, Aagaard Larsen L, et al. Genetic and environmental influences in Dupuytren's disease: a study of 30,330 Danish twin pairs. J Hand Surg Eur Vol 2015;**40**(2):171-6 doi: 10.1177/1753193414535720[published Online First: Epub Date]|.
20. Lyall HA. Dupuytren's disease in identical twins. J Hand Surg Br 1993;**18**(3):368-70
21. Capstick R, Bragg T, Giele H, et al. Sibling recurrence risk in Dupuytren's disease. J Hand Surg Eur Vol 2013;**38**(4):424-9 doi: 10.1177/1753193412453359[published Online First: Epub Date]|.
22. Becker K, Tinschert S, Lienert A, et al. The importance of genetic susceptibility in Dupuytren's disease. Clin Genet 2015;**87**(5):483-7 doi: 10.1111/cge.12410[published Online First: Epub Date]|.
23. Hu FZ, Nystrom A, Ahmed A, et al. Mapping of an autosomal dominant gene for Dupuytren's contracture to chromosome 16q in a Swedish family. Clin Genet 2005;**68**(5):424-9 doi: 10.1111/j.1399-0004.2005.00504.x[published Online First: Epub Date]|.
24. Bayat A, Watson JS, Stanley JK, et al. Genetic susceptibility to dupuytren disease: association of Zf9 transcription factor gene. Plast Reconstr Surg 2003;**111**(7):2133-9 doi: 10.1097/01.PRS.0000060531.98180.32[published Online First: Epub Date]|.
25. Dal Cin P, De Smet L, Sciot R, et al. Trisomy 7 and trisomy 8 in dividing and non-dividing tumor cells in Dupuytren's disease. Cancer Genet Cytogenet 1999;**108**(2):137-40
26. Casalone R, Mazzola D, Meroni E, et al. Cytogenetic and interphase cytogenetic analyses reveal chromosome instability but no clonal trisomy 8 in Dupuytren contracture. Cancer Genet Cytogenet 1997;**99**(1):73-6
27. Shih BB, Tassabehji M, Watson JS, et al. Genome-wide high-resolution screening in Dupuytren's disease reveals common regions of DNA copy number alterations. J Hand Surg Am 2010;**35**(7):1172-83 e7 doi: 10.1016/j.jhsa.2010.03.006[published Online First: Epub Date]|.
28. Shih B, Tassabehji M, Watson JS, et al. DNA copy number variations at chromosome 7p14.1 and chromosome 14q11.2 are associated with dupuytren's disease: potential role for MMP and Wnt signaling pathway. Plast Reconstr Surg 2012;**129**(4):921-32 doi: 10.1097/PRS.0b013e3182442343[published Online First: Epub Date]|.
29. Dolmans GH, Werker PM, Hennies HC, et al. Wnt signaling and Dupuytren's disease. N Engl J Med 2011;**365**(4):307-17 doi: 10.1056/NEJMoa1101029[published Online First: Epub Date]|.
30. Becker K, Siegert S, Toliat MR, et al. Meta-Analysis of Genome-Wide Association Studies and Network Analysis-Based Integration with Gene Expression Data Identify New Suggestive Loci and Unravel a Wnt-Centric Network Associated with Dupuytren's

Disease. PLoS One 2016;**11**(7):e0158101 doi: 10.1371/journal.pone.0158101[published Online First: Epub Date]|.
31. Ten Dam EJ, van Beuge MM, Bank RA, et al. Further evidence of the involvement of the Wnt signaling pathway in Dupuytren's disease. J Cell Commun Signal 2016;**10**(1):33-40 doi: 10.1007/s12079-015-0312-8[published Online First: Epub Date]|.
32. Staats KA, Wu T, Gan BS, et al. Dupuytren's disease susceptibility gene, EPDR1, is involved in myofibroblast contractility. J Dermatol Sci 2016;**83**(2):131-7 doi: 10.1016/j.jdermsci.2016.04.015[published Online First: Epub Date]|.
33. Crawford J, Raykha C, Charles D, et al. WT1 expression is increased in primary fibroblasts derived from Dupuytren's disease tissues. J Cell Commun Signal 2015;**9**(4):347-52 doi: 10.1007/s12079-015-0293-7[published Online First: Epub Date]|.
34. Riester SM, Arsoy D, Camilleri ET, et al. RNA sequencing reveals a depletion of collagen targeting microRNAs in Dupuytren's disease. BMC Med Genomics 2015;**8**:59 doi: 10.1186/s12920-015-0135-8[published Online First: Epub Date]|.
35. McCarthy MI, Abecasis GR, Cardon LR, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 2008;**9**(5):356-69 doi: 10.1038/nrg2344[published Online First: Epub Date]|.
36. Tindale LC, Leach S, Spinelli JJ, et al. Lipid and Alzheimer's disease genes associated with healthy aging and longevity in healthy oldest-old. Oncotarget 2017;**8**(13):20612-21 doi: 10.18632/oncotarget.15296[published Online First: Epub Date]|.
37. Van Cauwenberghe C, Van Broeckhoven C, Sleegers K. The genetic landscape of Alzheimer disease: clinical implications and perspectives. Genet Med 2016;**18**(5):421-30 doi: 10.1038/gim.2015.117[published Online First: Epub Date]|.
38. Slatkin M. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. Nat Rev Genet 2008;**9**(6):477-85 doi: 10.1038/nrg2361[published Online First: Epub Date]|.
39. McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet 2016;**48**(10):1279-83 doi: 10.1038/ng.3643[published Online First: Epub Date]|.
40. Wynn TA. Cellular and molecular mechanisms of fibrosis. J Pathol 2008;**214**(2):199-210 doi: 10.1002/path.2277[published Online First: Epub Date]|.
41. Bonnans C, Chou J, Werb Z. Remodelling the extracellular matrix in development and disease. Nat Rev Mol Cell Biol 2014;**15**(12):786-801 doi: 10.1038/nrm3904[published Online First: Epub Date]|.
42. Affo S, Yu LX, Schwabe RF. The Role of Cancer-Associated Fibroblasts and Fibrosis in Liver Cancer. Annu Rev Pathol 2017;**12**:153-86 doi: 10.1146/annurev-pathol-052016-100322[published Online First: Epub Date]|.
43. Wilson MS, Wynn TA. Pulmonary fibrosis: pathogenesis, etiology and regulation. Mucosal Immunol 2009;**2**(2):103-21 doi: 10.1038/mi.2008.85[published Online First: Epub Date]|.
44. He W, Dai C, Li Y, et al. Wnt/beta-catenin signaling promotes renal interstitial fibrosis. J Am Soc Nephrol 2009;**20**(4):765-76 doi: 10.1681/ASN.2008060566[published Online First: Epub Date]|.
45. Mack M. Inflammation and fibrosis. Matrix Biol 2018;**68-69**:106-21 doi: 10.1016/j.matbio.2017.11.010[published Online First: Epub Date]|.
46. Wick G, Grundtman C, Mayerl C, et al. The immunology of fibrosis. Annu Rev Immunol 2013;**31**:107-35 doi: 10.1146/annurev-immunol-032712-095937[published Online First: Epub Date]|.

47. Nanchahal J, Hinz B. Strategies to overcome the hurdles to treat fibrosis, a major unmet clinical need. Proc Natl Acad Sci U S A 2016;**113**(27):7291-3 doi: 10.1073/pnas.1607896113[published Online First: Epub Date]|.
48. Buzzetti E, Pinzani M, Tsochatzis EA. The multiple-hit pathogenesis of non-alcoholic fatty liver disease (NAFLD). Metabolism 2016;**65**(8):1038-48 doi: 10.1016/j.metabol.2015.12.012[published Online First: Epub Date]|.
49. Kropski JA, Lawson WE, Young LR, et al. Genetic studies provide clues on the pathogenesis of idiopathic pulmonary fibrosis. Dis Model Mech 2013;**6**(1):9-17 doi: 10.1242/dmm.010736[published Online First: Epub Date]|.
50. Wynn TA, Vannella KM. Macrophages in Tissue Repair, Regeneration, and Fibrosis. Immunity 2016;**44**(3):450-62 doi: 10.1016/j.immuni.2016.02.015[published Online First: Epub Date]|.
51. Mescher AL. Macrophages and fibroblasts during inflammation and tissue repair in models of organ regeneration. Regeneration (Oxf) 2017;**4**(2):39-53 doi: 10.1002/reg2.77[published Online First: Epub Date]|.
52. Bradding P, Pejler G. The controversial role of mast cells in fibrosis. Immunol Rev 2018;**282**(1):198-231 doi: 10.1111/imr.12626[published Online First: Epub Date]|.
53. Andersson CK, Andersson-Sjoland A, Mori M, et al. Activated MCTC mast cells infiltrate diseased lung areas in cystic fibrosis and idiopathic pulmonary fibrosis. Respir Res 2011;**12**:139 doi: 10.1186/1465-9921-12-139[published Online First: Epub Date]|.
54. O'Brien-Ladner AR, Wesselius LJ, Stechschulte DJ. Bleomycin injury of the lung in a mast-cell-deficient model. Agents Actions 1993;**39**(1-2):20-4
55. Reber LL, Daubeuf F, Pejler G, et al. Mast cells contribute to bleomycin-induced lung inflammation and injury in mice through a chymase/mast cell protease 4-dependent mechanism. J Immunol 2014;**192**(4):1847-54 doi: 10.4049/jimmunol.1300875[published Online First: Epub Date]|.
56. Fasbender F, Widera A, Hengstler JG, et al. Natural Killer Cells and Liver Fibrosis. Front Immunol 2016;**7**:19 doi: 10.3389/fimmu.2016.00019[published Online First: Epub Date]|.
57. O'Donoghue RJ, Knight DA, Richards CD, et al. Genetic partitioning of interleukin-6 signalling in mice dissociates Stat3 from Smad3-mediated lung fibrosis. EMBO Mol Med 2012;**4**(9):939-51 doi: 10.1002/emmm.201100604[published Online First: Epub Date]|.
58. Komura K, Yanaba K, Horikawa M, et al. CD19 regulates the development of bleomycin-induced pulmonary fibrosis in a mouse model. Arthritis Rheum 2008;**58**(11):3574-84 doi: 10.1002/art.23995[published Online First: Epub Date]|.
59. Hinz B. The role of myofibroblasts in wound healing. Curr Res Transl Med 2016;**64**(4):171-77 doi: 10.1016/j.retram.2016.09.003[published Online First: Epub Date]|.
60. Hinz B. Myofibroblasts. Exp Eye Res 2016;**142**:56-70 doi: 10.1016/j.exer.2015.07.009[published Online First: Epub Date]|.
61. Xu Y, Mizuno T, Sridharan A, et al. Single-cell RNA sequencing identifies diverse roles of epithelial cells in idiopathic pulmonary fibrosis. JCI Insight 2016;**1**(20):e90558 doi: 10.1172/jci.insight.90558[published Online First: Epub Date]|.
62. Kramann R, Machado F, Wu H, et al. Parabiosis and single-cell RNA sequencing reveal a limited contribution of monocytes to myofibroblasts in kidney fibrosis. JCI Insight 2018;**3**(9) doi: 10.1172/jci.insight.99561[published Online First: Epub Date]|.

63. Al-Qattan MM, Abd-Elwahed MM, Hawary K, et al. Myofibroblast expression in skin wounds is enhanced by collagen III suppression. Biomed Res Int 2015;**2015**:958695 doi: 10.1155/2015/958695[published Online First: Epub Date]|.
64. Klingberg F, Hinz B, White ES. The myofibroblast matrix: implications for tissue repair and fibrosis. J Pathol 2013;**229**(2):298-309 doi: 10.1002/path.4104[published Online First: Epub Date]|.
65. Hinz B. The extracellular matrix and transforming growth factor-beta1: Tale of a strained relationship. Matrix Biol 2015;**47**:54-65 doi: 10.1016/j.matbio.2015.05.006[published Online First: Epub Date]|.
66. Petrey AC, de la Motte CA. Hyaluronan, a crucial regulator of inflammation. Front Immunol 2014;**5**:101 doi: 10.3389/fimmu.2014.00101[published Online First: Epub Date]|.
67. Ghatak S, Maytin EV, Mack JA, et al. Roles of Proteoglycans and Glycosaminoglycans in Wound Healing and Fibrosis. Int J Cell Biol 2015;**2015**:834893 doi: 10.1155/2015/834893[published Online First: Epub Date]|.
68. Zhou Y, Hagood JS, Lu B, et al. Thy-1-integrin alphav beta5 interactions inhibit lung fibroblast contraction-induced latent transforming growth factor-beta1 activation and myofibroblast differentiation. J Biol Chem 2010;**285**(29):22382-93 doi: 10.1074/jbc.M110.126227[published Online First: Epub Date]|.
69. Zhou Y, Huang X, Hecker L, et al. Inhibition of mechanosensitive signaling in myofibroblasts ameliorates experimental pulmonary fibrosis. J Clin Invest 2013;**123**(3):1096-108 doi: 10.1172/JCI66700[published Online First: Epub Date]|.
70. Duscher D, Maan ZN, Wong VW, et al. Mechanotransduction and fibrosis. J Biomech 2014;**47**(9):1997-2005 doi: 10.1016/j.jbiomech.2014.03.031[published Online First: Epub Date]|.
71. Cheng Z, Limbu MH, Wang Z, et al. MMP-2 and 9 in Chronic Kidney Disease. Int J Mol Sci 2017;**18**(4) doi: 10.3390/ijms18040776[published Online First: Epub Date]|.
72. Moss ML, Jin SL, Milla ME, et al. Cloning of a disintegrin metalloproteinase that processes precursor tumour-necrosis factor-alpha. Nature 1997;**385**(6618):733-6 doi: 10.1038/385733a0[published Online First: Epub Date]|.
73. Piersma B, Bank RA, Boersema M. Signaling in Fibrosis: TGF-beta, WNT, and YAP/TAZ Converge. Front Med (Lausanne) 2015;**2**:59 doi: 10.3389/fmed.2015.00059[published Online First: Epub Date]|.
74. Yu Q, Stamenkovic I. Cell surface-localized matrix metalloproteinase-9 proteolytically activates TGF-beta and promotes tumor invasion and angiogenesis. Genes Dev 2000;**14**(2):163-76
75. Wrana JL. Signaling by the TGFbeta superfamily. Cold Spring Harb Perspect Biol 2013;**5**(10):a011197 doi: 10.1101/cshperspect.a011197[published Online First: Epub Date]|.
76. Morishima N, Mizoguchi I, Takeda K, et al. TGF-beta is necessary for induction of IL-23R and Th17 differentiation by IL-6 and IL-23. Biochem Biophys Res Commun 2009;**386**(1):105-10 doi: 10.1016/j.bbrc.2009.05.140[published Online First: Epub Date]|.
77. Cui H, Ge J, Xie N, et al. miR-34a promotes fibrosis in aged lungs by inducing alveolarepithelial dysfunctions. Am J Physiol Lung Cell Mol Physiol 2017;**312**(3):L415-L24 doi: 10.1152/ajplung.00335.2016[published Online First: Epub Date]|.
78. Zhou B, Liu Y, Kahn M, et al. Interactions between beta-catenin and transforming growth factor-beta signaling pathways mediate epithelial-mesenchymal transition and are

dependent on the transcriptional co-activator cAMP-response element-binding protein (CREB)-binding protein (CBP). J Biol Chem 2012;**287**(10):7026-38 doi: 10.1074/jbc.M111.276311[published Online First: Epub Date]|.
79. Choi ME, Ding Y, Kim SI. TGF-beta signaling via TAK1 pathway: role in kidney fibrosis. Semin Nephrol 2012;**32**(3):244-52 doi: 10.1016/j.semnephrol.2012.04.003[published Online First: Epub Date]|.
80. Lee MK, Pardoux C, Hall MC, et al. TGF-beta activates Erk MAP kinase signalling through direct phosphorylation of ShcA. EMBO J 2007;**26**(17):3957-67 doi: 10.1038/sj.emboj.7601818[published Online First: Epub Date]|.
81. Vega G, Alarcon S, San Martin R. The cellular and signalling alterations conducted by TGF-beta contributing to renal fibrosis. Cytokine 2016;**88**:115-25 doi: 10.1016/j.cyto.2016.08.019[published Online First: Epub Date]|.
82. Sorrentino A, Thakur N, Grimsby S, et al. The type I TGF-beta receptor engages TRAF6 to activate TAK1 in a receptor kinase-independent manner. Nat Cell Biol 2008;**10**(10):1199-207 doi: 10.1038/ncb1780[published Online First: Epub Date]|.
83. Ji H, Tang H, Lin H, et al. Rho/Rock cross-talks with transforming growth factor-beta/Smad pathway participates in lung fibroblast-myofibroblast differentiation. Biomed Rep 2014;**2**(6):787-92 doi: 10.3892/br.2014.323[published Online First: Epub Date]|.
84. Knipe RS, Tager AM, Liao JK. The Rho kinases: critical mediators of multiple profibrotic processes and rational targets for new therapies for pulmonary fibrosis. Pharmacol Rev 2015;**67**(1):103-17 doi: 10.1124/pr.114.009381[published Online First: Epub Date]|.
85. Luo S, Hieu TB, Ma F, et al. ZYZ-168 alleviates cardiac fibrosis after myocardial infarction through inhibition of ERK1/2-dependent ROCK1 activation. Sci Rep 2017;**7**:43242 doi: 10.1038/srep43242[published Online First: Epub Date]|.
86. Ghatak S, Markwald RR, Hascall VC, et al. Transforming growth factor beta1 (TGFbeta1) regulates CD44V6 expression and activity through extracellular signal-regulated kinase (ERK)-induced EGR1 in pulmonary fibrogenic fibroblasts. J Biol Chem 2017;**292**(25):10465-89 doi: 10.1074/jbc.M116.752451[published Online First: Epub Date]|.
87. Trachtman H, Fervenza FC, Gipson DS, et al. A phase 1, single-dose study of fresolimumab, an anti-TGF-beta antibody, in treatment-resistant primary focal segmental glomerulosclerosis. Kidney Int 2011;**79**(11):1236-43 doi: 10.1038/ki.2011.33[published Online First: Epub Date]|.
88. Mora AL, Rojas M, Pardo A, et al. Emerging therapies for idiopathic pulmonary fibrosis, a progressive age-related disease. Nat Rev Drug Discov 2017;**16**(11):755-72 doi: 10.1038/nrd.2017.170[published Online First: Epub Date]|.
89. Meng XM, Tang PM, Li J, et al. TGF-beta/Smad signaling in renal fibrosis. Front Physiol 2015;**6**:82 doi: 10.3389/fphys.2015.00082[published Online First: Epub Date]|.
90. Denton CP, Merkel PA, Furst DE, et al. Recombinant human anti-transforming growth factor beta1 antibody therapy in systemic sclerosis: a multicenter, randomized, placebo-controlled phase I/II trial of CAT-192. Arthritis Rheum 2007;**56**(1):323-33 doi: 10.1002/art.22289[published Online First: Epub Date]|.
91. King TE, Jr., Bradford WZ, Castro-Bernardini S, et al. A phase 3 trial of pirfenidone in patients with idiopathic pulmonary fibrosis. N Engl J Med 2014;**370**(22):2083-92 doi: 10.1056/NEJMoa1402582[published Online First: Epub Date]|.

92. Varga J, Pasche B. Antitransforming growth factor-beta therapy in fibrosis: recent progress and implications for systemic sclerosis. Curr Opin Rheumatol 2008;**20**(6):720-8 doi: 10.1097/BOR.0b013e32830e48e8[published Online First: Epub Date]|.
93. Yasuoka H. Recent Treatments of Interstitial Lung Disease with Systemic Sclerosis. Clin Med Insights Circ Respir Pulm Med 2015;**9**(Suppl 1):97-110 doi: 10.4137/CCRPM.S23315[published Online First: Epub Date]|.
94. Didiasova M, Singh R, Wilhelm J, et al. Pirfenidone exerts antifibrotic effects through inhibition of GLI transcription factors. FASEB J 2017;**31**(5):1916-28 doi: 10.1096/fj.201600892RR[published Online First: Epub Date]|.
95. Li C, Rezov V, Joensuu E, et al. Pirfenidone decreases mesothelioma cell proliferation and migration via inhibition of ERK and AKT and regulates mesothelioma tumor microenvironment in vivo. Sci Rep 2018;**8**(1):10070 doi: 10.1038/s41598-018-28297-x[published Online First: Epub Date]|.
96. Nusse R, Clevers H. Wnt/beta-Catenin Signaling, Disease, and Emerging Therapeutic Modalities. Cell 2017;**169**(6):985-99 doi: 10.1016/j.cell.2017.05.016[published Online First: Epub Date]|.
97. Li C, Bellusci S, Borok Z, et al. Non-canonical WNT signalling in the lung. J Biochem 2015;**158**(5):355-65 doi: 10.1093/jb/mvv081[published Online First: Epub Date]|.
98. Kestler HA, Kuhl M. Generating a Wnt switch: it's all about the right dosage. J Cell Biol 2011;**193**(3):431-3 doi: 10.1083/jcb.201103167[published Online First: Epub Date]|.
99. Eubelen M, Bostaille N, Cabochette P, et al. A molecular mechanism for Wnt ligand-specific signaling. Science 2018;**361**(6403) doi: 10.1126/science.aat1178[published Online First: Epub Date]|.
100. Vuga LJ, Ben-Yehudah A, Kovkarova-Naumovski E, et al. WNT5A is a regulator of fibroblast proliferation and resistance to apoptosis. Am J Respir Cell Mol Biol 2009;**41**(5):583-9 doi: 10.1165/rcmb.2008-0201OC[published Online First: Epub Date]|.
101. Spanjer AI, Baarsma HA, Oostenbrink LM, et al. TGF-beta-induced profibrotic signaling is regulated in part by the WNT receptor Frizzled-8. FASEB J 2016;**30**(5):1823-35 doi: 10.1096/fj.201500129[published Online First: Epub Date]|.
102. Park HW, Kim YC, Yu B, et al. Alternative Wnt Signaling Activates YAP/TAZ. Cell 2015;**162**(4):780-94 doi: 10.1016/j.cell.2015.07.013[published Online First: Epub Date]|.
103. Burgy O, Konigshoff M. The WNT signaling pathways in wound healing and fibrosis. Matrix Biol 2018;**68-69**:67-80 doi: 10.1016/j.matbio.2018.03.017[published Online First: Epub Date]|.
104. Guan S, Zhou J. Frizzled-7 mediates TGF-beta-induced pulmonary fibrosis by transmitting non-canonical Wnt signaling. Exp Cell Res 2017;**359**(1):226-34 doi: 10.1016/j.yexcr.2017.07.025[published Online First: Epub Date]|.
105. Krishnamurthy N, Kurzrock R. Targeting the Wnt/beta-catenin pathway in cancer: Update on effectors and inhibitors. Cancer Treat Rev 2018;**62**:50-60 doi: 10.1016/j.ctrv.2017.11.002[published Online First: Epub Date]|.
106. Lovestone S, Boada M, Dubois B, et al. A phase II trial of tideglusib in Alzheimer's disease. J Alzheimers Dis 2015;**45**(1):75-88 doi: 10.3233/JAD-141959[published Online First: Epub Date]|.
107. Kahn M. Can we safely target the WNT pathway? Nat Rev Drug Discov 2014;**13**(7):513-32 doi: 10.1038/nrd4233[published Online First: Epub Date]|.
108. Townley WA, Baker R, Sheppard N, et al. Dupuytren's contracture unfolded. BMJ 2006;**332**(7538):397-400 doi: 10.1136/bmj.332.7538.397[published Online First: Epub Date]|.

109. van Beuge MM, Ten Dam EJ, Werker PM, et al. Matrix and cell phenotype differences in Dupuytren's disease. Fibrogenesis Tissue Repair 2016;**9**:9 doi: 10.1186/s13069-016-0046-0[published Online First: Epub Date]|.
110. Berndt A, Kosmehl H, Katenkamp D, et al. Appearance of the myofibroblastic phenotype in Dupuytren's disease is associated with a fibronectin, laminin, collagen type IV and tenascin extracellular matrix. Pathobiology 1994;**62**(2):55-8 doi: 10.1159/000163879[published Online First: Epub Date]|.
111. Rehman S, Goodacre R, Day PJ, et al. Dupuytren's: a systems biology disease. Arthritis Res Ther 2011;**13**(5):238 doi: 10.1186/ar3438[published Online First: Epub Date]|.
112. Andrew JG, Andrew SM, Ash A, et al. An investigation into the role of inflammatory cells in Dupuytren's disease. J Hand Surg Br 1991;**16**(3):267-71
113. Verjee LS, Verhoekx JS, Chan JK, et al. Unraveling the signaling pathways promoting fibrosis in Dupuytren's disease reveals TNF as a therapeutic target. Proc Natl Acad Sci U S A 2013;**110**(10):E928-37 doi: 10.1073/pnas.1301100110[published Online First: Epub Date]|.
114. Glasser SW, Hagood JS, Wong S, et al. Mechanisms of Lung Fibrosis Resolution. Am J Pathol 2016;**186**(5):1066-77 doi: 10.1016/j.ajpath.2016.01.018[published Online First: Epub Date]|.
115. Verjee LS, Midwood K, Davidson D, et al. Myofibroblast distribution in Dupuytren's cords: correlation with digital contracture. J Hand Surg Am 2009;**34**(10):1785-94 doi: 10.1016/j.jhsa.2009.08.005[published Online First: Epub Date]|.
116. Prockop DJ. Inflammation, fibrosis, and modulation of the process by mesenchymal stem/stromal cells. Matrix Biol 2016;**51**:7-13 doi: 10.1016/j.matbio.2016.01.010[published Online First: Epub Date]|.
117. Wynn TA, Chawla A, Pollard JW. Macrophage biology in development, homeostasis and disease. Nature 2013;**496**(7446):445-55 doi: 10.1038/nature12034[published Online First: Epub Date]|.
118. Raghu G. Idiopathic pulmonary fibrosis: lessons from clinical trials over the past 25 years. Eur Respir J 2017;**50**(4) doi: 10.1183/13993003.01209-2017[published Online First: Epub Date]|.
119. Spencer CC, Su Z, Donnelly P, et al. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLoS Genet 2009;**5**(5):e1000477 doi: 10.1371/journal.pgen.1000477[published Online First: Epub Date]|.
120. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. Nature 2009;**461**(7265):747-53 doi: 10.1038/nature08494[published Online First: Epub Date]|.
121. Petersen BS, Fredrich B, Hoeppner MP, et al. Opportunities and challenges of whole-genome and -exome sequencing. BMC Genet 2017;**18**(1):14 doi: 10.1186/s12863-017-0479-5[published Online First: Epub Date]|.
122. McInnes IB, Sieper J, Braun J, et al. Efficacy and safety of secukinumab, a fully human anti-interleukin-17A monoclonal antibody, in patients with moderate-to-severe psoriatic arthritis: a 24-week, randomised, double-blind, placebo-controlled, phase II proof-of-concept trial. Ann Rheum Dis 2014;**73**(2):349-56 doi: 10.1136/annrheumdis-2012-202646[published Online First: Epub Date]|.
123. Yang J, Lee SH, Goddard ME, et al. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 2011;**88**(1):76-82 doi: 10.1016/j.ajhg.2010.11.011[published Online First: Epub Date]|.

124. Lippert C, Listgarten J, Liu Y, et al. FaST linear mixed models for genome-wide association studies. Nat Methods 2011;**8**(10):833-5 doi: 10.1038/nmeth.1681[published Online First: Epub Date]|.
125. Schweiger R, Kaufman S, Laaksonen R, et al. Fast and Accurate Construction of Confidence Intervals for Heritability. Am J Hum Genet 2016;**98**(6):1181-92 doi: 10.1016/j.ajhg.2016.04.016[published Online First: Epub Date]|.
126. Smith GD, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol 2003;**32**(1):1-22
127. Miller MC, Manning HB, Jain A, et al. Membrane type 1 matrix metalloproteinase is a crucial promoter of synovial invasion in human rheumatoid arthritis. Arthritis Rheum 2009;**60**(3):686-97 doi: 10.1002/art.24331[published Online First: Epub Date]|.
128. Suetsugu T, Koshizuka K, Seki N, et al. Downregulation of matrix metalloproteinase 14 by the antitumor miRNA, miR-150-5p, inhibits the aggressiveness of lung squamous cell carcinoma cells. International journal of oncology 2018;**52**(3):913-24 doi: 10.3892/ijo.2017.4232[published Online First: Epub Date]|.
129. van Beuge MM, Ten Dam EJ, Werker PM, et al. Wnt pathway in Dupuytren disease: connecting profibrotic signals. Transl Res 2015;**166**(6):762-71 e3 doi: 10.1016/j.trsl.2015.09.006[published Online First: Epub Date]|.
130. DiRocco DP, Kobayashi A, Taketo MM, et al. Wnt4/beta-catenin signaling in medullary kidney myofibroblasts. J Am Soc Nephrol 2013;**24**(9):1399-412 doi: 10.1681/ASN.2012050512[published Online First: Epub Date]|.
131. Padwal M, Cheng G, Liu L, et al. WNT signaling is required for peritoneal membrane angiogenesis. Am J Physiol Renal Physiol 2018;**314**(6):F1036-F45 doi: 10.1152/ajprenal.00497.2017[published Online First: Epub Date]|.
132. Zhao B, Ye X, Yu J, et al. TEAD mediates YAP-dependent gene induction and growth control. Genes Dev 2008;**22**(14):1962-71 doi: 10.1101/gad.1664408[published Online First: Epub Date]|.
133. Akhmetshina A, Palumbo K, Dees C, et al. Activation of canonical Wnt signalling is required for TGF-beta-mediated fibrosis. Nat Commun 2012;**3**:735 doi: 10.1038/ncomms1734[published Online First: Epub Date]|.
134. Clevers H, Nusse R. Wnt/beta-catenin signaling and disease. Cell 2012;**149**(6):1192-205 doi: 10.1016/j.cell.2012.05.012[published Online First: Epub Date]|.
135. Wei W, Pelechano V, Jarvelin AI, et al. Functional consequences of bidirectional promoters. Trends Genet 2011;**27**(7):267-76 doi: 10.1016/j.tig.2011.04.002[published Online First: Epub Date]|.
136. Maier T, Guell M, Serrano L. Correlation of mRNA and protein in complex biological samples. FEBS Lett 2009;**583**(24):3966-73 doi: 10.1016/j.febslet.2009.10.036[published Online First: Epub Date]|.
137. Liu Y, Beyer A, Aebersold R. On the Dependency of Cellular Protein Levels on mRNA Abundance. Cell 2016;**165**(3):535-50 doi: 10.1016/j.cell.2016.03.014[published Online First: Epub Date]|.
138. Koussounadis A, Langdon SP, Um IH, et al. Relationship between differentially expressed mRNA and mRNA-protein correlations in a xenograft model system. Sci Rep 2015;**5**:10775 doi: 10.1038/srep10775[published Online First: Epub Date]|.
139. Hausler KD, Horwood NJ, Chuman Y, et al. Secreted frizzled-related protein-1 inhibits RANKL-dependent osteoclast formation. J Bone Miner Res 2004;**19**(11):1873-81 doi: 10.1359/JBMR.040807[published Online First: Epub Date]|.

140. Chuman Y, Uren A, Cahill J, et al. Identification of a peptide binding motif for secreted frizzled-related protein-1. Peptides 2004;**25**(11):1831-8 doi: 10.1016/j.peptides.2004.07.010[published Online First: Epub Date]|.
141. Mueller S, Wahlander A, Selevsek N, et al. Protein degradation corrects for imbalanced subunit stoichiometry in OST complex assembly. Mol Biol Cell 2015;**26**(14):2596-608 doi: 10.1091/mbc.E15-03-0168[published Online First: Epub Date]|.
142. Leitinger B, Kwan AP. The discoidin domain receptor DDR2 is a receptor for type X collagen. Matrix Biol 2006;**25**(6):355-64 doi: 10.1016/j.matbio.2006.05.006[published Online First: Epub Date]|.
143. Olaso E, Ikeda K, Eng FJ, et al. DDR2 receptor promotes MMP-2-mediated proliferation and invasion by hepatic stellate cells. J Clin Invest 2001;**108**(9):1369-78 doi: 10.1172/JCI12373[published Online First: Epub Date]|.
144. Zhao H, Bian H, Bu X, et al. Targeting of Discoidin Domain Receptor 2 (DDR2) Prevents Myofibroblast Activation and Neovessel Formation During Pulmonary Fibrosis. Mol Ther 2016;**24**(10):1734-44 doi: 10.1038/mt.2016.109[published Online First: Epub Date]|.
145. Canning P, Tan L, Chu K, et al. Structural mechanisms determining inhibition of the collagen receptor DDR1 by selective and multi-targeted type II kinase inhibitors. J Mol Biol 2014;**426**(13):2457-70 doi: 10.1016/j.jmb.2014.04.014[published Online First: Epub Date]|.
146. Itoh Y. Membrane-type matrix metalloproteinases: Their functions and regulations. Matrix Biol 2015;**44-46**:207-23 doi: 10.1016/j.matbio.2015.03.004[published Online First: Epub Date]|.
147. Itoh Y. Discoidin domain receptors: Microenvironment sensors that promote cellular migration and invasion. Cell Adh Migr 2018:1-8 doi: 10.1080/19336918.2018.1460011[published Online First: Epub Date]|.
148. Butler GS, Dean RA, Tam EM, et al. Pharmacoproteomics of a metalloproteinase hydroxamate inhibitor in breast cancer cells: dynamics of membrane type 1 matrix metalloproteinase-mediated membrane protein shedding. Mol Cell Biol 2008;**28**(15):4896-914 doi: 10.1128/MCB.01775-07[published Online First: Epub Date]|.
149. Wilkinson JM, Davidson RK, Swingler TE, et al. MMP-14 and MMP-2 are key metalloproteases in Dupuytren's disease fibroblast-mediated contraction. Biochim Biophys Acta 2012;**1822**(6):897-905 doi: 10.1016/j.bbadis.2012.02.001[published Online First: Epub Date]|.
150. Shi CS, Kehrl JH. Tumor necrosis factor (TNF)-induced germinal center kinase-related (GCKR) and stress-activated protein kinase (SAPK) activation depends upon the E2/E3 complex Ubc13-Uev1A/TNF receptor-associated factor 2 (TRAF2). J Biol Chem 2003;**278**(17):15429-34 doi: 10.1074/jbc.M211796200[published Online First: Epub Date]|.
151. Shi CS, Huang NN, Harrison K, et al. The mitogen-activated protein kinase kinase kinase kinase GCKR positively regulates canonical and noncanonical Wnt signaling in B lymphocytes. Mol Cell Biol 2006;**26**(17):6511-21 doi: 10.1128/MCB.00209-06[published Online First: Epub Date]|.
152. Wang OH, Azizian N, Guo M, et al. Prognostic and Functional Significance of MAP4K5 in Pancreatic Cancer. PLoS One 2016;**11**(3):e0152300 doi: 10.1371/journal.pone.0152300[published Online First: Epub Date]|.
153. Kaliora AC, Kalafati IP, Gioxari A, et al. A modified response of NAFLD patients with non-significant fibrosis in nutritional counseling according to GCKR rs1260326. Eur J

Nutr 2018;**57**(6):2227-35 doi: 10.1007/s00394-017-1499-7[published Online First: Epub Date]|.
154. Mora A, Sandve GK, Gabrielsen OS, et al. In the loop: promoter-enhancer interactions and bioinformatics. Brief Bioinform 2016;**17**(6):980-95 doi: 10.1093/bib/bbv097[published Online First: Epub Date]|.
155. Shoulders MD, Raines RT. Collagen structure and stability. Annu Rev Biochem 2009;**78**:929-58 doi: 10.1146/annurev.biochem.77.032207.120833[published Online First: Epub Date]|.
156. Murrell GA, Francis MJ, Bromley L. The collagen changes of Dupuytren's contracture. J Hand Surg Br 1991;**16**(3):263-6
157. Konitsiotis AD, Raynal N, Bihan D, et al. Characterization of high affinity binding motifs for the discoidin domain receptor DDR2 in collagen. J Biol Chem 2008;**283**(11):6861-8 doi: 10.1074/jbc.M709290200[published Online First: Epub Date]|.
158. Jing H, Song J, Zheng J. Discoidin domain receptor 1: New star in cancer-targeted therapy and its complex role in breast carcinoma. Oncol Lett 2018;**15**(3):3403-08 doi: 10.3892/ol.2018.7795[published Online First: Epub Date]|.
159. Farndale RW, Lisman T, Bihan D, et al. Cell-collagen interactions: the use of peptide Toolkits to investigate collagen-receptor interactions. Biochem Soc Trans 2008;**36**(Pt 2):241-50 doi: 10.1042/BST0360241[published Online First: Epub Date]|.
160. Xu H, Abe T, Liu JK, et al. Normal activation of discoidin domain receptor 1 mutants with disulfide cross-links, insertions, or deletions in the extracellular juxtamembrane region: mechanistic implications. J Biol Chem 2014;**289**(19):13565-74 doi: 10.1074/jbc.M113.536144[published Online First: Epub Date]|.
161. Xu H, Bihan D, Chang F, et al. Discoidin domain receptors promote alpha1beta1- and alpha2beta1-integrin mediated cell adhesion to collagen by enhancing integrin activation. PLoS One 2012;**7**(12):e52209 doi: 10.1371/journal.pone.0052209[published Online First: Epub Date]|.
162. Zhang Y, Su J, Yu J, et al. An essential role of discoidin domain receptor 2 (DDR2) in osteoblast differentiation and chondrocyte maturation via modulation of Runx2 activation. J Bone Miner Res 2011;**26**(3):604-17 doi: 10.1002/jbmr.225[published Online First: Epub Date]|.
163. Kim D, You E, Jeong J, et al. DDR2 controls the epithelial-mesenchymal-transition-related gene expression via c-Myb acetylation upon matrix stiffening. Sci Rep 2017;**7**(1):6847 doi: 10.1038/s41598-017-07126-7[published Online First: Epub Date]|.
164. Chen SC, Wang BW, Wang DL, et al. Hypoxia induces discoidin domain receptor-2 expression via the p38 pathway in vascular smooth muscle cells to increase their migration. Biochem Biophys Res Commun 2008;**374**(4):662-7 doi: 10.1016/j.bbrc.2008.07.092[published Online First: Epub Date]|.
165. Makino K, Jinnin M, Aoi J, et al. Discoidin domain receptor 2-microRNA 196a-mediated negative feedback against excess type I collagen expression is impaired in scleroderma dermal fibroblasts. J Invest Dermatol 2013;**133**(1):110-9 doi: 10.1038/jid.2012.252[published Online First: Epub Date]|.
166. Xu L, Peng H, Glasson S, et al. Increased expression of the collagen receptor discoidin domain receptor 2 in articular cartilage as a key event in the pathogenesis of osteoarthritis. Arthritis Rheum 2007;**56**(8):2663-73 doi: 10.1002/art.22761[published Online First: Epub Date]|.
167. Majkowska I, Shitomi Y, Ito N, et al. Discoidin domain receptor 2 mediates collagen-induced activation of membrane-type 1 matrix metalloproteinase in human

fibroblasts. J Biol Chem 2017;**292**(16):6633-43 doi: 10.1074/jbc.M116.770057[published Online First: Epub Date]|.
168. Vogel W, Gish GD, Alves F, et al. The discoidin domain receptor tyrosine kinases are activated by collagen. Mol Cell 1997;**1**(1):13-23
169. Hou G, Vogel WF, Bendeck MP. Tyrosine kinase activity of discoidin domain receptor 1 is necessary for smooth muscle cell migration and matrix metalloproteinase expression. Circ Res 2002;**90**(11):1147-9
170. Ruiz PA, Jarai G. Collagen I induces discoidin domain receptor (DDR) 1 expression through DDR2 and a JAK2-ERK1/2-mediated mechanism in primary human lung fibroblasts. J Biol Chem 2011;**286**(15):12912-23 doi: 10.1074/jbc.M110.143693[published Online First: Epub Date]|.
171. Sato T, Konomi K, Yamasaki S, et al. Comparative analysis of gene expression profiles in intact and damaged regions of human osteoarthritic cartilage. Arthritis Rheum 2006;**54**(3):808-17 doi: 10.1002/art.21638[published Online First: Epub Date]|.
172. Ferri N, Carragher NO, Raines EW. Role of discoidin domain receptors 1 and 2 in human smooth muscle cell-mediated collagen remodeling: potential implications in atherosclerosis and lymphangioleiomyomatosis. Am J Pathol 2004;**164**(5):1575-85 doi: 10.1016/S0002-9440(10)63716-9[published Online First: Epub Date]|.
173. Olaso E, Lin HC, Wang LH, et al. Impaired dermal wound healing in discoidin domain receptor 2-deficient mice associated with defective extracellular matrix remodeling. Fibrogenesis Tissue Repair 2011;**4**(1):5 doi: 10.1186/1755-1536-4-5[published Online First: Epub Date]|.
174. Schmidt S, Friedl P. Interstitial cell migration: integrin-dependent and alternative adhesion mechanisms. Cell Tissue Res 2010;**339**(1):83-92 doi: 10.1007/s00441-009-0892-9[published Online First: Epub Date]|.
175. Schlessinger J. Direct binding and activation of receptor tyrosine kinases by collagen. Cell 1997;**91**(7):869-72
176. Zhang K, Corsa CA, Ponik SM, et al. The collagen receptor discoidin domain receptor 2 stabilizes SNAIL1 to facilitate breast cancer metastasis. Nat Cell Biol 2013;**15**(6):677-87 doi: 10.1038/ncb2743[published Online First: Epub Date]|.
177. Franco C, Ahmad PJ, Hou G, et al. Increased cell and matrix accumulation during atherogenesis in mice with vessel wall-specific deletion of discoidin domain receptor 1. Circ Res 2010;**106**(11):1775-83 doi: 10.1161/CIRCRESAHA.109.213637[published Online First: Epub Date]|.
178. Franco C, Britto K, Wong E, et al. Discoidin domain receptor 1 on bone marrow-derived cells promotes macrophage accumulation during atherogenesis. Circ Res 2009;**105**(11):1141-8 doi: 10.1161/CIRCRESAHA.109.207357[published Online First: Epub Date]|.
179. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;**489**(7414):57-74 doi: 10.1038/nature11247[published Online First: Epub Date]|.
180. International HapMap C. The International HapMap Project. Nature 2003;**426**(6968):789-96 doi: 10.1038/nature02168[published Online First: Epub Date]|.
181. Guo Y, Yang TL, Dong SS, et al. Genetic analysis identifies DDR2 as a novel gene affecting bone mineral density and osteoporotic fractures in Chinese population. PLoS One 2015;**10**(2):e0117102 doi: 10.1371/journal.pone.0117102[published Online First: Epub Date]|.

182. Jiang K, Yang Z, Cui W, et al. An exome-wide association study identifies new susceptibility loci for age of smoking initiation in African- and European-American populations. Nicotine Tob Res 2017 doi: 10.1093/ntr/ntx262[published Online First: Epub Date]|.
183. Sasaki H, Shitara M, Yokota K, et al. DDR2 polymorphisms and mRNA expression in lung cancers of Japanese patients. Oncol Lett 2012;**4**(1):33-37 doi: 10.3892/ol.2012.684[published Online First: Epub Date]|.
184. Hammerman PS, Sos ML, Ramos AH, et al. Mutations in the DDR2 kinase gene identify a novel therapeutic target in squamous cell lung cancer. Cancer Discov 2011;**1**(1):78-89 doi: 10.1158/2159-8274.CD-11-0005[published Online First: Epub Date]|.
185. Tseng Q, Duchemin-Pelletier E, Deshiere A, et al. Spatial organization of the extracellular matrix regulates cell-cell junction positioning. Proc Natl Acad Sci U S A 2012;**109**(5):1506-11 doi: 10.1073/pnas.1106377109[published Online First: Epub Date]|.
186. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods 2015;**12**(4):357-60 doi: 10.1038/nmeth.3317[published Online First: Epub Date]|.
187. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 2014;**30**(7):923-30 doi: 10.1093/bioinformatics/btt656[published Online First: Epub Date]|.
188. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol 2010;**11**(10):R106 doi: 10.1186/gb-2010-11-10-r106[published Online First: Epub Date]|.
189. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;**28**(1):27-30
190. Luo W, Friedman MS, Shedden K, et al. GAGE: generally applicable gene set enrichment for pathway analysis. BMC Bioinformatics 2009;**10**:161 doi: 10.1186/1471-2105-10-161[published Online First: Epub Date]|.
191. Xu H, Raynal N, Stathopoulos S, et al. Collagen binding specificity of the discoidin domain receptors: binding sites on collagens II and III and molecular determinants for collagen IV recognition by DDR1. Matrix Biol 2011;**30**(1):16-26 doi: 10.1016/j.matbio.2010.10.004[published Online First: Epub Date]|.
192. Yao L, Bestwick CS, Bestwick LA, et al. Phenotypic drift in human tenocyte culture. Tissue Eng 2006;**12**(7):1843-9 doi: 10.1089/ten.2006.12.1843[published Online First: Epub Date]|.
193. Lee LC, Zhang AY, Chong AK, et al. Expression of a novel gene, MafB, in Dupuytren's disease. J Hand Surg Am 2006;**31**(2):211-8 doi: 10.1016/j.jhsa.2005.09.007[published Online First: Epub Date]|.
194. Lowe WL, Jr., Reddy TE. Genomic approaches for understanding the genetics of complex disease. Genome Res 2015;**25**(10):1432-41 doi: 10.1101/gr.190603.115[published Online First: Epub Date]|.
195. Guo Y, Conti DV, Wang K. Enlight: web-based integration of GWAS results with biological annotations. Bioinformatics 2015;**31**(2):275-6 doi: 10.1093/bioinformatics/btu639[published Online First: Epub Date]|.
196. Watanabe K, Taskesen E, van Bochoven A, et al. Functional mapping and annotation of genetic associations with FUMA. Nat Commun 2017;**8**(1):1826 doi: 10.1038/s41467-017-01261-5[published Online First: Epub Date]|.

197. Schmidt EM, Zhang J, Zhou W, et al. GREGOR: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. Bioinformatics 2015;**31**(16):2601-6 doi: 10.1093/bioinformatics/btv201[published Online First: Epub Date]|.
198. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res 2012;**40**(Database issue):D930-4 doi: 10.1093/nar/gkr917[published Online First: Epub Date]|.
199. Boyle AP, Hong EL, Hariharan M, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res 2012;**22**(9):1790-7 doi: 10.1101/gr.137323.112[published Online First: Epub Date]|.
200. Consortium GT. The Genotype-Tissue Expression (GTEx) project. Nat Genet 2013;**45**(6):580-5 doi: 10.1038/ng.2653[published Online First: Epub Date]|.
201. Li MJ, Wang LY, Xia Z, et al. GWAS3D: Detecting human regulatory variants by integrative analysis of genome-wide associations, chromosome interactions and histone modifications. Nucleic Acids Res 2013;**41**(Web Server issue):W150-8 doi: 10.1093/nar/gkt456[published Online First: Epub Date]|.
202. Raychaudhuri S, Plenge RM, Rossin EJ, et al. Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. PLoS Genet 2009;**5**(6):e1000534 doi: 10.1371/journal.pgen.1000534[published Online First: Epub Date]|.
203. Riethoven JJ. Regulatory regions in DNA: promoters, enhancers, silencers, and insulators. Methods Mol Biol 2010;**674**:33-42 doi: 10.1007/978-1-60761-854-6_3[published Online First: Epub Date]|.
204. Ramani V, Shendure J, Duan Z. Understanding Spatial Genome Organization: Methods and Insights. Genomics Proteomics Bioinformatics 2016;**14**(1):7-20 doi: 10.1016/j.gpb.2016.01.002[published Online First: Epub Date]|.
205. Buenrostro JD, Wu B, Chang HY, et al. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr Protoc Mol Biol 2015;**109**:21 29 1-9 doi: 10.1002/0471142727.mb2129s109[published Online First: Epub Date]|.
206. Bell AC, Felsenfeld G. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. Nature 2000;**405**(6785):482-5 doi: 10.1038/35013100[published Online First: Epub Date]|.
207. Phillips JE, Corces VG. CTCF: master weaver of the genome. Cell 2009;**137**(7):1194-211 doi: 10.1016/j.cell.2009.06.001[published Online First: Epub Date]|.
208. Ren G, Jin W, Cui K, et al. CTCF-Mediated Enhancer-Promoter Interaction Is a Critical Regulator of Cell-to-Cell Variation of Gene Expression. Mol Cell 2017;**67**(6):1049-58 e6 doi: 10.1016/j.molcel.2017.08.026[published Online First: Epub Date]|.
209. Marshall AD, Bailey CG, Champ K, et al. CTCF genetic alterations in endometrial carcinoma are pro-tumorigenic. Oncogene 2017;**36**(29):4100-10 doi: 10.1038/onc.2017.25[published Online First: Epub Date]|.
210. Guo YA, Chang MM, Huang W, et al. Mutation hotspots at CTCF binding sites coupled to chromosomal instability in gastrointestinal cancers. Nat Commun 2018;**9**(1):1520 doi: 10.1038/s41467-018-03828-2[published Online First: Epub Date]|.
211. Ackermann AM, Wang Z, Schug J, et al. Integration of ATAC-seq and RNA-seq identifies human alpha cell and beta cell signature genes. Mol Metab 2016;**5**(3):233-44 doi: 10.1016/j.molmet.2016.01.002[published Online First: Epub Date]|.

212. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;**9**(4):357-9 doi: 10.1038/nmeth.1923[published Online First: Epub Date]|.
213. Ou J, Liu H, Yu J, et al. ATACseqQC: a Bioconductor package for post-alignment quality assessment of ATAC-seq data. BMC Genomics 2018;**19**(1):169 doi: 10.1186/s12864-018-4559-3[published Online First: Epub Date]|.
214. Zhang Y, Liu T, Meyer CA, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol 2008;**9**(9):R137 doi: 10.1186/gb-2008-9-9-r137[published Online First: Epub Date]|.
215. Heinz S, Benner C, Spann N, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell 2010;**38**(4):576-89 doi: 10.1016/j.molcel.2010.05.004[published Online First: Epub Date]|.
216. Ramirez F, Ryan DP, Gruning B, et al. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res 2016;**44**(W1):W160-5 doi: 10.1093/nar/gkw257[published Online First: Epub Date]|.
217. Ross-Innes CS, Stark R, Teschendorff AE, et al. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. Nature 2012;**481**(7381):389-93 doi: 10.1038/nature10730[published Online First: Epub Date]|.
218. Yu G, Wang LG, He QY. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics 2015;**31**(14):2382-3 doi: 10.1093/bioinformatics/btv145[published Online First: Epub Date]|.
219. Robinson JT, Thorvaldsdottir H, Winckler W, et al. Integrative genomics viewer. Nat Biotechnol 2011;**29**(1):24-6 doi: 10.1038/nbt.1754[published Online First: Epub Date]|.
220. Raviram R, Rocha PP, Muller CL, et al. 4C-ker: A Method to Reproducibly Identify Genome-Wide Interactions Captured by 4C-Seq Experiments. PLoS Comput Biol 2016;**12**(3):e1004780 doi: 10.1371/journal.pcbi.1004780[published Online First: Epub Date]|.
221. Roadmap Epigenomics C, Kundaje A, Meuleman W, et al. Integrative analysis of 111 reference human epigenomes. Nature 2015;**518**(7539):317-30 doi: 10.1038/nature14248[published Online First: Epub Date]|.
222. Buenrostro JD, Wu B, Litzenburger UM, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature 2015;**523**(7561):486-90 doi: 10.1038/nature14590[published Online First: Epub Date]|.
223. Zhang Z, Park E, Lin L, et al. A panoramic view of RNA modifications: exploring new frontiers. Genome Biol 2018;**19**(1):11 doi: 10.1186/s13059-018-1394-4[published Online First: Epub Date]|.
224. El-Mohri H, Wu Y, Mohanty S, et al. Impact of matrix stiffness on fibroblast function. Mater Sci Eng C Mater Biol Appl 2017;**74**:146-51 doi: 10.1016/j.msec.2017.02.001[published Online First: Epub Date]|.
225. Skardal A, Mack D, Atala A, et al. Substrate elasticity controls cell proliferation, surface marker expression and motile phenotype in amniotic fluid-derived stem cells. J Mech Behav Biomed Mater 2013;**17**:307-16 doi: 10.1016/j.jmbbm.2012.10.001[published Online First: Epub Date]|.
226. Murphy TL, Tussiwand R, Murphy KM. Specificity through cooperation: BATF-IRF interactions control immune-regulatory networks. Nat Rev Immunol 2013;**13**(7):499-509 doi: 10.1038/nri3470[published Online First: Epub Date]|.

227. Tussiwand R, Lee WL, Murphy TL, et al. Compensatory dendritic cell development mediated by BATF-IRF interactions. Nature 2012;**490**(7421):502-7 doi: 10.1038/nature11531[published Online First: Epub Date]|.
228. Ise W, Kohyama M, Schraml BU, et al. The transcription factor BATF controls the global regulators of class-switch recombination in both B cells and T cells. Nat Immunol 2011;**12**(6):536-43 doi: 10.1038/ni.2037[published Online First: Epub Date]|.
229. Ciofani M, Madar A, Galan C, et al. A validated regulatory network for Th17 cell specification. Cell 2012;**151**(2):289-303 doi: 10.1016/j.cell.2012.09.016[published Online First: Epub Date]|.
230. Rhee J, Park SH, Kim SK, et al. Inhibition of BATF/JUN transcriptional activity protects against osteoarthritic cartilage destruction. Ann Rheum Dis 2017;**76**(2):427-34 doi: 10.1136/annrheumdis-2015-208953[published Online First: Epub Date]|.
231. Lambert SA, Jolma A, Campitelli LF, et al. The Human Transcription Factors. Cell 2018;**172**(4):650-65 doi: 10.1016/j.cell.2018.01.029[published Online First: Epub Date]|.
232. Hinz B, Phan SH, Thannickal VJ, et al. The myofibroblast: one function, multiple origins. Am J Pathol 2007;**170**(6):1807-16 doi: 10.2353/ajpath.2007.070112[published Online First: Epub Date]|.
233. Wade R, Igali L, Figus A. Skin involvement in Dupuytren's disease. J Hand Surg Eur Vol 2016;**41**(6):600-8 doi: 10.1177/1753193415601353[published Online First: Epub Date]|.
234. Delvoux B, D'Hooghe T, Kyama C, et al. Inhibition of type 1 17beta-hydroxysteroid dehydrogenase impairs the synthesis of 17beta-estradiol in endometriosis lesions. J Clin Endocrinol Metab 2014;**99**(1):276-84 doi: 10.1210/jc.2013-2851

10.1210/jc.2013-2503[published Online First: Epub Date]|.
235. Horne BD, Anderson JL, Carlquist JF, et al. Generating genetic risk scores from intermediate phenotypes for use in association studies of clinically significant endpoints. Ann Hum Genet 2005;**69**(Pt 2):176-86 doi: 10.1046/j.1529-8817.2005.00155.x[published Online First: Epub Date]|.
236. Oram RA, Patel K, Hill A, et al. A Type 1 Diabetes Genetic Risk Score Can Aid Discrimination Between Type 1 and Type 2 Diabetes in Young Adults. Diabetes Care 2016;**39**(3):337-44 doi: 10.2337/dc15-1111[published Online First: Epub Date]|.
237. Patel KA, Oram RA, Flanagan SE, et al. Type 1 Diabetes Genetic Risk Score: A Novel Tool to Discriminate Monogenic and Type 1 Diabetes. Diabetes 2016;**65**(7):2094-99 doi: 10.2337/db15-1690[published Online First: Epub Date]|.
238. van Rijssen AL, ter Linden H, Werker PM. Five-year results of a randomized clinical trial on treatment in Dupuytren's disease: percutaneous needle fasciotomy versus limited fasciectomy. Plast Reconstr Surg 2012;**129**(2):469-77 doi: 10.1097/PRS.0b013e31823aea95[published Online First: Epub Date]|.
239. Degreef I, De Smet L. Risk factors in Dupuytren's diathesis: is recurrence after surgery predictable? Acta Orthop Belg 2011;**77**(1):27-32
240. Denkler K. Surgical complications associated with fasciectomy for dupuytren's disease: a 20-year review of the English literature. Eplasty 2010;**10**:e15
241. Ng M, Thakkar D, Southam L, et al. A Genome-wide Association Study of Dupuytren Disease Reveals 17 Additional Variants Implicated in Fibrosis. Am J Hum Genet 2017;**101**(3):417-27 doi: 10.1016/j.ajhg.2017.08.006[published Online First: Epub Date]|.

242. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med 2011;**30**(4):377-99 doi: 10.1002/sim.4067[published Online First: Epub Date]|.
243. Rodrigues JN, Becker GW, Ball C, et al. Surgery for Dupuytren's contracture of the fingers. Cochrane Database Syst Rev 2015(12):CD010143 doi: 10.1002/14651858.CD010143.pub2[published Online First: Epub Date]|.
244. Barratt SL, Creamer A, Hayton C, et al. Idiopathic Pulmonary Fibrosis (IPF): An Overview. J Clin Med 2018;**7**(8) doi: 10.3390/jcm7080201[published Online First: Epub Date]|.
245. Samuel CS, Royce SG, Hewitson TD, et al. Anti-fibrotic actions of relaxin. Br J Pharmacol 2017;**174**(10):962-76 doi: 10.1111/bph.13529[published Online First: Epub Date]|.
246. Gambichler T, Schmitz L. Ultraviolet A1 Phototherapy for Fibrosing Conditions. Front Med (Lausanne) 2018;**5**:237 doi: 10.3389/fmed.2018.00237[published Online First: Epub Date]|.
247. Manske RC, Prohaska D. Diagnosis and management of adhesive capsulitis. Curr Rev Musculoskelet Med 2008;**1**(3-4):180-9 doi: 10.1007/s12178-008-9031-6[published Online First: Epub Date]|.
248. Walker-Bone K, Palmer KT, Reading I, et al. Prevalence and impact of musculoskeletal disorders of the upper limb in the general population. Arthritis Rheum 2004;**51**(4):642-51 doi: 10.1002/art.20535[published Online First: Epub Date]|.
249. Lubiecki M, Carr A. Frozen shoulder: past, present, and future. Journal of orthopaedic surgery 2007;**15**(1):1-3 doi: 10.1177/230949900701500101[published Online First: Epub Date]|.
250. Zreik NH, Malik RA, Charalambous CP. Adhesive capsulitis of the shoulder and diabetes: a meta-analysis of prevalence. Muscles Ligaments Tendons J 2016;**6**(1):26-34 doi: 10.11138/mltj/2016.6.1.026[published Online First: Epub Date]|.
251. Rodeo SA, Hannafin JA, Tom J, et al. Immunolocalization of cytokines and their receptors in adhesive capsulitis of the shoulder. J Orthop Res 1997;**15**(3):427-36 doi: 10.1002/jor.1100150316[published Online First: Epub Date]|.
252. Vercellini P, Vigano P, Somigliana E, et al. Endometriosis: pathogenesis and treatment. Nat Rev Endocrinol 2014;**10**(5):261-75 doi: 10.1038/nrendo.2013.255[published Online First: Epub Date]|.
253. Sapkota Y, Steinthorsdottir V, Morris AP, et al. Meta-analysis identifies five novel loci associated with endometriosis highlighting key genes involved in hormone metabolism. Nat Commun 2017;**8**:15539 doi: 10.1038/ncomms15539[published Online First: Epub Date]|.
254. Painter JN, Anderson CA, Nyholt DR, et al. Genome-wide association study identifies a locus at 7p15.2 associated with endometriosis. Nat Genet 2011;**43**(1):51-4 doi: 10.1038/ng.731[published Online First: Epub Date]|.
255. Parola M, Pinzani M. Liver fibrosis: Pathophysiology, pathogenetic targets and clinical issues. Mol Aspects Med 2018 doi: 10.1016/j.mam.2018.09.002[published Online First: Epub Date]|.
256. Caballeria L, Pera G, Arteaga I, et al. High Prevalence of Liver Fibrosis Among European Adults With Unknown Liver Disease: A Population-Based Study. Clin Gastroenterol Hepatol 2018;**16**(7):1138-45 e5 doi: 10.1016/j.cgh.2017.12.048[published Online First: Epub Date]|.
257. Qu K, Huang Z, Lin T, et al. New Insight into the Anti-liver Fibrosis Effect of Multitargeted Tyrosine Kinase Inhibitors: From Molecular Target to Clinical Trials.

Front Pharmacol 2015;**6**:300 doi: 10.3389/fphar.2015.00300[published Online First: Epub Date]|.
258. Buch S, Stickel F, Trepo E, et al. A genome-wide association study confirms PNPLA3 and identifies TM6SF2 and MBOAT7 as risk loci for alcohol-related cirrhosis. Nat Genet 2015;**47**(12):1443-8 doi: 10.1038/ng.3417[published Online First: Epub Date]|.
259. Patin E, Kutalik Z, Guergnon J, et al. Genome-wide association study identifies variants associated with progression of liver fibrosis from HCV infection. Gastroenterology 2012;**143**(5):1244-52 e12 doi: 10.1053/j.gastro.2012.07.097[published Online First: Epub Date]|.
260. Ulveling D, Le Clerc S, Cobat A, et al. A new 3p25 locus is associated with liver fibrosis progression in human immunodeficiency virus/hepatitis C virus-coinfected patients. Hepatology 2016;**64**(5):1462-72 doi: 10.1002/hep.28695[published Online First: Epub Date]|.
261. Kupcinskas J, Valantiene I, Varkalaite G, et al. PNPLA3 and RNF7 Gene Variants are Associated with the Risk of Developing Liver Fibrosis and Cirrhosis in an Eastern European Population. J Gastrointestin Liver Dis 2017;**26**(1):37-43 doi: 10.15403/jgld.2014.1121.261.pnp[published Online First: Epub Date]|.
262. Strongman H, Kausar I, Maher TM. Incidence, Prevalence, and Survival of Patients with Idiopathic Pulmonary Fibrosis in the UK. Advances in therapy 2018;**35**(5):724-36 doi: 10.1007/s12325-018-0693-1[published Online First: Epub Date]|.
263. Nakamura Y, Suda T. Idiopathic Pulmonary Fibrosis: Diagnosis and Clinical Manifestations. Clin Med Insights Circ Respir Pulm Med 2015;**9**(Suppl 1):163-71 doi: 10.4137/CCRPM.S39897[published Online First: Epub Date]|.
264. Martinez FJ, Collard HR, Pardo A, et al. Idiopathic pulmonary fibrosis. Nature reviews Disease primers 2017;**3**:17074 doi: 10.1038/nrdp.2017.74[published Online First: Epub Date]|.
265. Allen RJ, Porte J, Braybrooke R, et al. Genetic variants associated with susceptibility to idiopathic pulmonary fibrosis in people of European ancestry: a genome-wide association study. Lancet Respir Med 2017;**5**(11):869-80 doi: 10.1016/S2213-2600(17)30387-9[published Online First: Epub Date]|.
266. Ranque B, Mouthon L. Geoepidemiology of systemic sclerosis. Autoimmun Rev 2010;**9**(5):A311-8 doi: 10.1016/j.autrev.2009.11.003[published Online First: Epub Date]|.
267. Arad U, Balbir-Gurman A, Doenyas-Barak K, et al. Anti-neutrophil antibody associated vasculitis in systemic sclerosis. Semin Arthritis Rheum 2011;**41**(2):223-9 doi: 10.1016/j.semarthrit.2010.11.001[published Online First: Epub Date]|.
268. Pattanaik D, Brown M, Postlethwaite BC, et al. Pathogenesis of Systemic Sclerosis. Front Immunol 2015;**6**:272 doi: 10.3389/fimmu.2015.00272[published Online First: Epub Date]|.
269. Yanaba K. Strategy for treatment of fibrosis in systemic sclerosis: Present and future. J Dermatol 2016;**43**(1):46-55 doi: 10.1111/1346-8138.13026[published Online First: Epub Date]|.
270. Broen JC, Coenen MJ, Radstake TR. Genetics of systemic sclerosis: an update. Curr Rheumatol Rep 2012;**14**(1):11-21 doi: 10.1007/s11926-011-0221-7[published Online First: Epub Date]|.
271. Wei P, Yang Y, Guo X, et al. Identification of an Association of TNFAIP3 Polymorphisms With Matrix Metalloproteinase Expression in Fibroblasts in an Integrative Study of Systemic Sclerosis-Associated Genetic and Environmental Factors. Arthritis

Rheumatol 2016;**68**(3):749-60 doi: 10.1002/art.39476[published Online First: Epub Date]|.
272. Stewart EA, Cookson CL, Gandolfo RA, et al. Epidemiology of uterine fibroids: a systematic review. BJOG 2017;**124**(10):1501-12 doi: 10.1111/1471-0528.14640[published Online First: Epub Date]|.
273. Borahay MA, Asoglu MR, Mas A, et al. Estrogen Receptors and Signaling in Fibroids: Role in Pathobiology and Therapeutic Implications. Reprod Sci 2017;**24**(9):1235-44 doi: 10.1177/1933719116678686[published Online First: Epub Date]|.
274. Medikare V, Kandukuri LR, Ananthapur V, et al. The genetic bases of uterine fibroids; a review. J Reprod Infertil 2011;**12**(3):181-91
275. Eggert SL, Huyck KL, Somasundaram P, et al. Genome-wide linkage and association analyses implicate FASN in predisposition to Uterine Leiomyomata. Am J Hum Genet 2012;**91**(4):621-8 doi: 10.1016/j.ajhg.2012.08.009[published Online First: Epub Date]|.
276. Cha PC, Takahashi A, Hosono N, et al. A genome-wide association study identifies three loci associated with susceptibility to uterine fibroids. Nat Genet 2011;**43**(5):447-50 doi: 10.1038/ng.805[published Online First: Epub Date]|.
277. Hellwege JN, Jeff JM, Wise LA, et al. A multi-stage genome-wide association study of uterine fibroids in African Americans. Hum Genet 2017;**136**(10):1363-73 doi: 10.1007/s00439-017-1836-1[published Online First: Epub Date]|.
278. Loh PR, Tucker G, Bulik-Sullivan BK, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat Genet 2015;**47**(3):284-90 doi: 10.1038/ng.3190[published Online First: Epub Date]|.
279. Terao C, Kawaguchi T, Dieude P, et al. Transethnic meta-analysis identifies GSDMA and PRDM1 as susceptibility genes to systemic sclerosis. Ann Rheum Dis 2017;**76**(6):1150-58 doi: 10.1136/annrheumdis-2016-210645[published Online First: Epub Date]|.
280. Fingerlin TE, Murphy E, Zhang W, et al. Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis. Nat Genet 2013;**45**(6):613-20 doi: 10.1038/ng.2609[published Online First: Epub Date]|.
281. Noth I, Zhang Y, Ma SF, et al. Genetic variants associated with idiopathic pulmonary fibrosis susceptibility and mortality: a genome-wide association study. Lancet Respir Med 2013;**1**(4):309-17 doi: 10.1016/S2213-2600(13)70045-6[published Online First: Epub Date]|.
282. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 2010;**26**(17):2190-1 doi: 10.1093/bioinformatics/btq340[published Online First: Epub Date]|.
283. Magi R, Morris AP. GWAMA: software for genome-wide association meta-analysis. BMC Bioinformatics 2010;**11**:288 doi: 10.1186/1471-2105-11-288[published Online First: Epub Date]|.
284. Bulik-Sullivan B, Finucane HK, Anttila V, et al. An atlas of genetic correlations across human diseases and traits. Nat Genet 2015;**47**(11):1236-41 doi: 10.1038/ng.3406[published Online First: Epub Date]|.
285. Bulik-Sullivan BK, Loh PR, Finucane HK, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet 2015;**47**(3):291-5 doi: 10.1038/ng.3211[published Online First: Epub Date]|.
286. Zhu X, Feng T, Tayo BO, et al. Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. Am J Hum Genet 2015;**96**(1):21-36 doi: 10.1016/j.ajhg.2014.11.011[published Online First: Epub Date]|.

287. Wang X, Lin P, Ho JWK. Discovery of cell-type specific DNA motif grammar in cis-regulatory elements using random Forest. BMC Genomics 2018;**19**(Suppl 1):929 doi: 10.1186/s12864-017-4340-z[published Online First: Epub Date]|.
288. Slieker MG, Sanders EA, Rijkers GT, et al. Disease modifying genes in cystic fibrosis. J Cyst Fibros 2005;**4 Suppl 2**:7-13 doi: 10.1016/j.jcf.2005.05.006[published Online First: Epub Date]|.
289. Messina G, Celauro E, Atterrato MT, et al. The Bucentaur (BCNT) protein family: a long-neglected class of essential proteins required for chromatin/chromosome organization and function. Chromosoma 2015;**124**(2):153-62 doi: 10.1007/s00412-014-0503-8[published Online First: Epub Date]|.
290. Zhao L, Gu C, Huang K, et al. Association between oestrogen receptor alpha (ESR1) gene polymorphisms and endometriosis: a meta-analysis of 24 case-control studies. Reprod Biomed Online 2016;**33**(3):335-49 doi: 10.1016/j.rbmo.2016.06.003[published Online First: Epub Date]|.
291. Rafnar T, Gunnarsson B, Stefansson OA, et al. Variants associating with uterine leiomyoma highlight genetic background shared by various cancers and hormone-related traits. Nat Commun 2018;**9**(1):3636 doi: 10.1038/s41467-018-05428-6[published Online First: Epub Date]|.
292. Li T, Zhao J, Yang J, et al. A Meta-Analysis of the Association between ESR1 Genetic Variants and the Risk of Breast Cancer. PLoS One 2016;**11**(4):e0153314 doi: 10.1371/journal.pone.0153314[published Online First: Epub Date]|.
293. Maekawa R, Sato S, Okada M, et al. Tissue-Specific Expression of Estrogen Receptor 1 Is Regulated by DNA Methylation in a T-DMR. Mol Endocrinol 2016;**30**(3):335-47 doi: 10.1210/me.2015-1058[published Online First: Epub Date]|.
294. Vitkus S, Yeh CR, Lin HH, et al. Distinct function of estrogen receptor alpha in smooth muscle and fibroblast cells in prostate development. Mol Endocrinol 2013;**27**(1):38-49 doi: 10.1210/me.2012-1212[published Online First: Epub Date]|.
295. Bulun SE, Monsavais D, Pavone ME, et al. Role of estrogen receptor-beta in endometriosis. Semin Reprod Med 2012;**30**(1):39-45 doi: 10.1055/s-0031-1299596[published Online First: Epub Date]|.
296. Novotny M, Vasilenko T, Varinska L, et al. ER-alpha agonist induces conversion of fibroblasts into myofibroblasts, while ER-beta agonist increases ECM production and wound tensile strength of healing skin wounds in ovariectomised rats. Exp Dermatol 2011;**20**(9):703-8 doi: 10.1111/j.1600-0625.2011.01284.x[published Online First: Epub Date]|.
297. Bain CC, Montgomery J, Scott CL, et al. TGFbetaR signalling controls CD103(+)CD11b(+) dendritic cell development in the intestine. Nat Commun 2017;**8**(1):620 doi: 10.1038/s41467-017-00658-6[published Online First: Epub Date]|.
298. Trapnell C. Defining cell types and states with single-cell genomics. Genome Res 2015;**25**(10):1491-8 doi: 10.1101/gr.190595.115[published Online First: Epub Date]|.