An Integrative Genomic Study of Dupuytren's Disease

I n a u g u r a l - D i s s e r t a t i o n

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Juanjiangmeng Du

aus Baotou (China)

2016

Berichterstatter:

Prof. Dr. Peter Nürnberg

Prof. Dr. Angelika Anna Noegel

Vorsitz der Prüfung:

Prof. Dr. Wolfgang Werr

Beisitzerin:

Dr. Birgit Budde

Tag der letzten mündlichen Prüfung:

18.01.2017

Die vorliegende Arbeit wurde in der Zeit von November 2013 bis November 2016 unter Anleitung von Prof. Dr. Peter Nürnberg am Cologne Center for Genomics (CCG) der Universität zu Köln, angefertigt.

To patients with complex diseases, we are here for you.

Zusammenfassung

Morbus Dupuytren (DD) ist eine Fibromatose des palmaren Bindegewebes, die zu Flexionskontrakturen der Finger führt. DD hat eine starke genetische Komponente mit einer geschätzten Erblichkeit von etwa 80%. Eine frühere kollaborative genom-weite Assoziationsstudie (GWAS) identifizierte 9 Suszeptibilitäts-Loci, die zu einem erhöhten Risiko für DD beitragen. Allerdings können diese Loci zusammen nur einen Bruchteil der Erblichkeit erklären. Die spezifischen Gene und funktionellen Varianten für DD blieben hier unklar.

Ziel der vorliegenden Forschungsarbeit ist es daher, die zugrunde liegende genetische Architektur von DD auf verschiedenen Ebenen systematisch zu untersuchen: i) Identifizierung von funktionellen Varianten, die zu einem starken GWAS-Assoziationssignal an einem DD-Risiko-Locus führen, ii) Priorisierung von Genen mit seltenen Varianten, die mit dem DD-Phänotyp funktionell in Verbindung gebracht werden können, und iii) Charakterisierung der transkriptionellen Deregulation in DD.

Ein Intervall auf Chromosom 7, am 7p14.1, in dem sich der Einzelnukleotid-Polymorphismus (SNP) rs16879765 (G>A) mit der stärksten signifikanten Assoziation mit DD in der genannten GWAS befindet, wurde als DD-Suszeptibilitäts-Locus identifiziert. Zuerst haben wir daher eine gezielte Anreicherungsstrategie mit anschließender NGS sequenzierung mittels (targeted NGS) verfolgt, um eine Region von 500 kb an dem Locus auf 7p14.1 zu untersuchen. Eine seltene aminosäure-ändernde Variante, rs149095633 (p.P121L, auf Haplotypen durch das rs16879765*A Risiko-Allel markiert), und ein häufiger eQTL-Kandidat, rs2044831 (in mäßigem Kopplungsungleichgewicht (LD) mit rs16879765), wurden in *EPDR1* identifiziert, einem funktionellen Kandidatengen, das am kontraktilen Phänotyp der DD-Primärzellen beteiligt ist.

Zweitens führten wir eine Pilotstudie zur Exomsequenzierung (WES) bei 50 DD-Patienten mit mutmaßlich ausgeprägter genetischer Prädisposition durch und priorisierten Kandidatengene mit seltenen Varianten. 3919 seltene kodierende Varianten wurden als nachteilig vorhergesagt. 1774 Gene mit mehr als 2 Varianten wurden mit „Human Phenotype Ontology“ und „Gene Intolerance Score EvolTol“ nach geeigneten Phänotyp-Kategorien und palmarer Expression gefiltert. Als Ergebnis wurden 12 Gene als DD-Kandidatengene mit potenziell pathogenen seltenen Varianten priorisiert. Insbesondere wurden 6 davon als funktionell wichtige Gene für die DD-Entwicklung eingestuft.

Drittens führten wir eine umfassende Transkriptomstudie durch RNA-Sequenzierung in 50 DD-/Kontroll-Biopsieproben durch. Unter Verwendung von Genexpressionsprofilen wurde durch eine Perturbationsanalyse der Hippo-Signalweg als ein Schlüsselmechanotransduktionsweg vorgeschlagen, der die profibrotische Mikroumgebung in DD vorbereitet. Der TGFβ-Signalweg und ECM-Rezeptor-Wechselwirkungen wurden als essenziell für die Gewebefibrose vorhergesagt. Darüber hinaus wurden durch die Analyse von alternativem Splicing (AS) charakteristische Isoformprofile und Isoformverwendungen in DD-Gewebe gefunden. Diese Isoformen könnten an

einem Mechanismus beteiligt sein, mit dem sich Zellen in betroffenem Gewebe an Fibrose anpassen und die Fibroseprogression weiter fördern.

Zusammengefasst charakterisiert diese Studie erstmals einen GWAS-Risikolocus für DD und bietet einen Ansatz für die Identifizierung funktioneller Varianten für DD im Post-GWAS-Zeitalter. Die explorative Priorisierung von DD-assoziierten Kandidatengenen unterstützt die Annahme, dass seltene Varianten einen Beitrag zur Entwicklung der Krankheit leisten, und benennt Kandidatengene für Folgestudien. Zudem liefert diese Studie neue Anhaltspunkte für einen Zusammenhang zwischen wichtigen physiologischen Pfaden und der transkriptionellen Regulation von DD und bietet einen ersten Einblick in spezifisches AS in betroffenem Gewebe und die damit verbundenen AS-Regulationsmechanismen. Unsere Studie stellt einen ersten Schritt zur Integration versoliedener genomischer Ansätze mit dem Ziel dar, um die mechanistischen Verbindungen zwischen der genetischer Prädisposition und der Ausprägung von DD aufzuklären.

Abstract

Dupuytren's Disease (DD) is a fibromatosis in the palmar connective tissue that leads to flexion contractures of fingers. DD has a strong genetic component with an estimated heritability of about 80%. A previous collaborative genome-wide association study (GWAS) has mapped 9 susceptibility loci that were shown to contribute to the increased risk of DD. However, these loci together can only explain a small fraction of heritability. Moreover, the respective genes and functional variants underlying DD remained unclear.

Therefore, the present study aims to systematically investigate the genetic architecture of DD at different levels by: i) identifying functional variants contributing to a strong GWAS association signal at a DD risk locus, ii) prioritizing DD phenotype-related genes with rare variant burdens, and iii) characterizing the transcriptional deregulation in DD.

An intervall on chromosome 7, 7p14.1, tagged by rs16879765 (G>A), which was most significantly associated with DD in the previous GWAS, has been identified as a DD susceptibility locus. Therefore, we first used a target-enrichment strategy coupled with next generation sequencing (targeted NGS) to assess a 500kb region at 7p14.1. A rare non-synonymous variant, rs149095633 (p.P121L, on haplotypes tagged by the rs16879765*A risk allele), and a common eQTL candidate, rs2044831 (in moderate linkage disequilibrium with rs16879765), were identified in *EPDR1*, a functional candidate gene contributing to the contractile phenotype of DD primary cells.

Second, we performed a pilot whole exome sequencing (WES) study in 50 DD patients with suspected high genetic predisposition and prioritized candidate genes with rare variant burden. 3919 rare coding variants were predicted to be deleterious. 1774 genes with gene burden greater than 2 were filtered for suitable phenotype classes and palmar expression according to Human Phenotype Ontology and the gene intolerance score EvolTol. As a result, 12 genes were prioritized as DD candidate genes with potentially pathogenic rare variants. In particular, 6 of these genes were suggested as functionally important for DD development.

Third, we carried out an elaborate transcriptome study in 50 DD/control biopsy samples by RNAseq. By pathway perturbation analysis using gene expression profiles, the Hippo signaling pathway was suggested as a key mechanotransdution pathway to prepare the profibrotic microenvironment in DD. The TGFβ pathway and ECM-receptor interactions were predicted as pathways essential for tissue fibrosis. Moreover, by alternative splicing (AS) analysis, DD tissue was suggested to harbor distinctive isoform profiles and isoform usage, which might provide a mechanism for cells in disease tissue to adapt to fibrosis and further promote fibrosis progression.

In summary, this study characterized for the first time a GWAS risk locus for DD and provided an approach for identifying functional variants for DD in the post-GWAS era. The exploratory prioritization

of DD-related candidate genes supported the assumption that rare variants can contribute to the development of the disease and nominated candidate genes for follow-up studies. Furthermore, this study proposed key physiological pathways involved in transcriptional regulation of DD and gave a first insight into disease tissue-specific AS and possible AS regulation mechanisms. Overall, our study represents a first step in integrating various genomic approaches to elucidate the mechanistic links between the genetic predisposition and the development of DD.

Table of Contents

List of Abbreviations

AS	Alternative splicing
ASE	Allele specific expression
α-SMA	α-smooth muscle actin
BMP	Bone morphogenetic protein
cDNA	Complementary DNA
CT	Carpal tunnel syndrome (control)
CTcell	CT control — primary cells from CTtis
CTtis	CT control — palmar fat tissue
DD	Dupuytren's disease (case)
DDcell	DD patients — primary cells from DDtis
DDfat	DD patients — matched perinodular fat
DDtis	DD patients — palmar nodule disease tissue
DEG	Differentially expressed genes
ECM	Extracellular matrix
FPKM	Fragments Per Kilobase per Million
GO	Gene ontology
GSEA	Gene set enrichment analysis
GWAS	Genome-wide association study
HPO	Human phenotypic ontology
IF	Isoform fraction
LD	Linkage disequilibrium
MAF	Minor allele frequency
NGS	Next generation sequencing
NMD	Nonsense mediated decay
RNA-seq	RNA sequencing using NGS
RNP	Ribonucleoprotein
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
Targeted NGS	Target enrichment and NGS
TGFβ	Transforming growth factor beta
TF	Transcription factor
WES	Whole exome sequencing

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Dupuytren's disease (DD) — a complex fibrosis disorder

Most common diseases, such as cancer, diabetes, Alzheimer disease and epilepsy, have a complex etiology. Unlike rare Mendelian disorders characterized by single-gene mutations, complex disorders are caused by a combination of genetic and environmental factors (e.g. aging, nutrition and life style). Aging is a major risk factor for many chronic diseases, yet, the basic mechanisms that drive aging remained largely unknown due to several reasons such as the complexity of the aging process at molecular, cellular and organ level and the considerable heterogeneity among individuals[1].

Dupuytren's Disease (DD) is a common connective tissue fibrosis disorder with a prevalence of 2.5% in Germany[2]. The worldwide prevalence of DD is related to geographic location with highest prevalence in Scandinavia, UK, Ireland, Australia and North America [3-5]. Other factors including trauma and exposure to vibrating machinery are also suggested to increase the danger of developing DD[6]. The manifestation rate of DD is positively correlated with increased age peaking in the sixth and seventh decade of life in men and women, respectively[7]. However, unlike many multifocal aging-associated diseases, DD is specially localized in the hands.

DD patients commonly first display benign nodules in the palms (Figure 1-1). These nodules are characterized by a high amount of proliferating fibroblasts and their differentiation towards myofibroblasts followed by progressive formation of cords consisting mostly of extracellular matrix (ECM) proteins[8]. The maturation of cords is accompanied by fibrosis and ultimately leads to digital contractures resulting in hand deformity and impaired hand function in daily life[6]. The treatment of DD consists largely of surgical removal of the contracted tissue, which is unfortunately accompanied with a high risk of neurovascular injury and DD recurrence (8% — 66%)[9,10]. However, the availability of affected tissues as well as the unifocal feature of DD enables us to use this disease as a unique model for studying the molecular mechanisms of aging-related disorders in connective tissue.

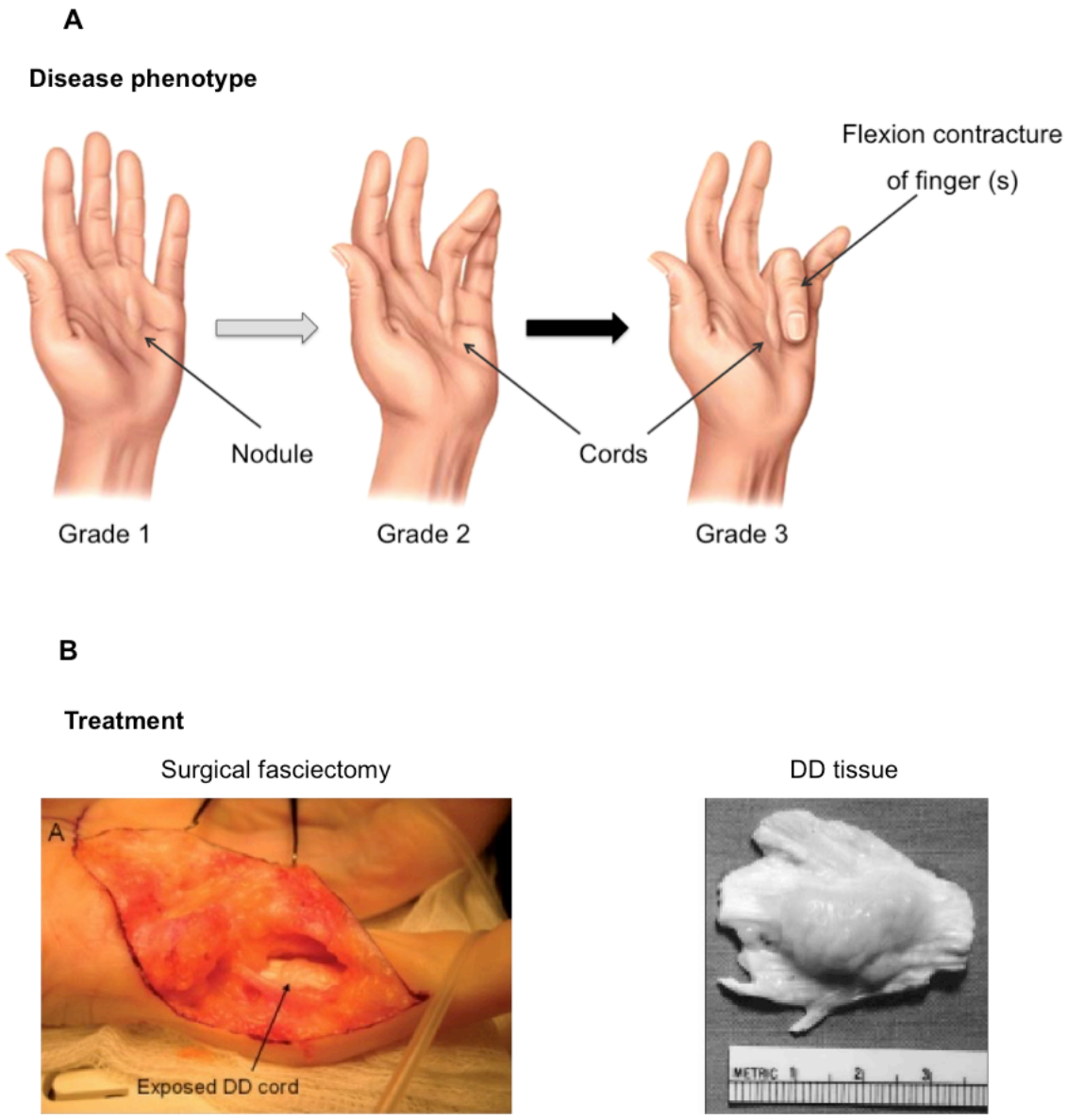


Figure 1-1 The clinical presentation of DD

(A) Benign nodules commonly first displayed in the palms of DD patients (Grade 1). The nodules are characterized by a high amount of proliferating fibroblasts and their differentiation towards myofibroblasts, followed by the progressive formation of cords consisting mostly of extracellular matrix proteins (Grade 2). The maturation of cords ultimately leads to digital contractures resulting in the hand deformity (Grade 3). The figure was adapted from http://www.chennaiplasticsurgery.in/Dupuytren's%20contracture.html.

(B) Treatment of DD consists largely of surgical removal of the contracted tissue, named as DD tissue. The figure was adapted from Shil et.al. [6]

1.1.1 Major cell types in DD — activated fibroblasts and myofibroblasts

In normal connective tissues, the resident fibroblasts, which are embedded in loosely arranged ECM, maintain tissue homeostasis not only by synthesizing ECM proteins including interstitial collagens, proteoglycans and adhesive non-collagenous proteins, but also by exerting mild contractile force on this ECM[11]. Very low levels of growth factors, cytokines and blood plasma proteins percolate through this microenvironment, which maintains fibroblasts in a 'quiescent' state[11].

During normal tissue repair such as wound healing, fibroblasts become activated and a set of cellular and extracellular events take place to repair the damage including four sequential yet overlapping phases: inflammation, proliferation, contraction and remodeling[11]. The cellular contraction accounts for a key phase in normal wound healing as it enables wound closure, however, if contraction is abnormally persistent, it can lead to tissue fibrosis[6], for example, DD.

The digital contracture phenotype of DD is mainly caused by an increased amount of activated fibroblasts and their differentiated myofibroblasts in the palmar fascia. The myofibroblasts, which display molecular and cellular phenotypes between fibroblasts and smooth muscle cells, are characterized by a flattened spindle shape, excessive ECM deposition and expression of α-smooth muscle actin (α-SMA, an early differentiation marker of vascular smooth muscle cells)[12,13]. The expression of α-SMA is a molecular hallmark to distinguish myofibroblasts from 'quiescent' fibroblasts. Upon tissue injury, the resident 'quiescent' fibroblasts are activated by both cytokine signaling and mechanical change of microenvironment and differentiate into a myofibroblast phenotype[11].

Then, the differentiated myofibroblasts incorporate α-SMA in stress fibers and develop adhesion complexes with ECM termed 'fibronexus' *in vivo* or 'super mature focal adhesions' *in vitro*. The adhesion complexes bridge the myofibroblasts' internal microfilaments with ECM and thus enable these cells to generate contractile forces surrounding the ECM. In pathological situations, the contractile activity in myofibroblasts is abnormally maintained over time and enhanced by the deposition of excessive ECM in their surrounding niche[14,15].

The origins of activated fibroblasts and myofibroblasts in fibrotic tissues still remain unclear. Recent studies on some fibrosis disorders and cancer have suggested a heterogeneous origin of activated fibroblasts and myofibroblasts in tissue fibrosis. Proposed major contributors include bone marrow-derived fibroblasts, adipocytes, epithelial cells and endothelial cells (Figure 1-2)[11,16]. However, the molecular mechanisms by which this differentiation occurs in DD remain not well understood.

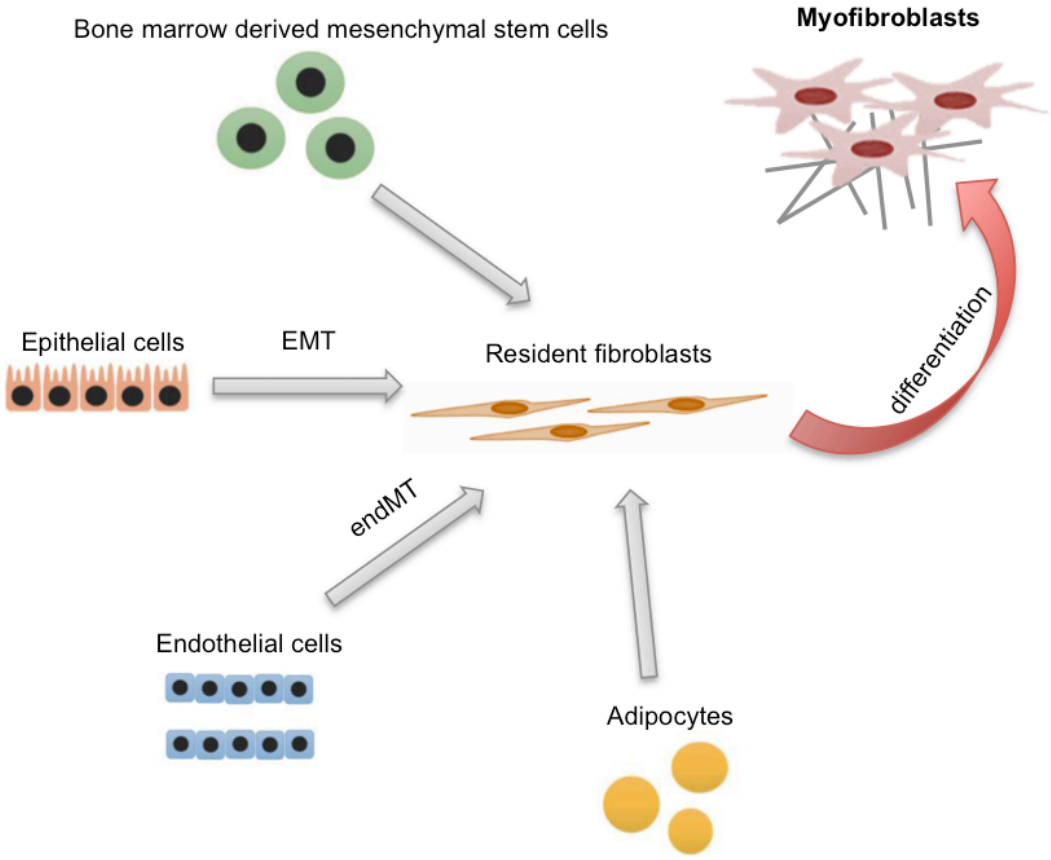


Figure 1-2 Possible origins of myofibroblasts

In some fibrosis disorders and cancer, myofibroblasts are suggested to originate directly or indirectly from various cells such as resident fibroblasts, adipocytes, bone marrow derived mesenchymal stem cells, epithelial cells (through epithelial mesenchymal transition: EMT) and endothelial cells (through endothelial mesenchymal transition: endMT) etc. The figure was adapted from Shiga et al.[16]

1.1.2 DD — associated with low BMI and type 2 diabetes

DD appears to be associated with other complex traits[17]. In the Reykjavik Study including 1297 men, elevated fasting blood glucose low body mass index (BMI, a marker for adiposity), were significantly correlated with the presence of DD[18]. The association between DD and low adiposity suggests common genes or pathways involved in diseases susceptibility.

1.2 The genetic background of DD

Compelling evidence, including frequent familial clustering and high Caucasian prevalence compared to other ethnic groups[5,18], consistently suggests genetic factors to contribute to the onset of DD.

In order to identify genetic risk loci associated with Dupuytren´s disease, a collaborative genome-wide association study (GWAS) including 2,325 DD cases and 11,562 controls was conducted by groups in Groningen (the Netherlands), Oxford (UK) and our group (Germany) to identify associations between the disease and common genetic markers. As a result, 11 single nucleotide polymorphisms (SNPs) in 9 different loci were found to be significantly associated ($p<5.0x10^{-8}$) with DD, implicating a genetic susceptibility to DD.

The top GWAS SNP, ranked by the lowest p-value was found in the intronic region of *EPDR1*/*SFRP4* (rs16879765, $p=5.6\times10^{-39}$; odds ratio, 1.98). Five further loci were found to harbor genes encoding proteins of the WNT/β-catenin signaling pathway[19], which are *WNT4* (rs7524102, $p=2.8\times10^{-9}$; odds ratio, 1.28), *WNT2* (rs4730775, $p=3.0\times10^{-8}$; odds ratio, 0.83), *RSPO2* (rs611744, $p=7.9\times10^{-15}$; odds ratio, 0.75), *WNT7B* (rs6519955, $p=3.2\times10^{-33}$; odds ratio, 1.54) and *SULF1* (rs2912522, $p=2.0\times10^{-13}$; odds ratio, 0.72), suggesting that variants in genes involved in the WNT/β-catenin signaling pathway may cause a predisposition to DD.

In total, the 9 risk loci identified in this GWAS could explain about 1% of the heritability of DD. Moreover, all the 9 loci were further validated in a study[20] combining three imputed GWAS data sets including 1580 cases and 4480 controls from the Netherlands (the previous Groningen study)[19], Switzerland and Germany. Six of these loci reached genome-wide significance ($p<5x10^{-8}$). In sum, these GWAS results provided the first solid piece of evidence, which greatly improved our understanding of the genetic basis of DD.

1.3 The unclear genetic architecture of DD

1.3.1 The missing genetic causes of GWAS loci

A more recent association study, involving 30,330 Danish monozygotic and heterozygotic twins, was conducted to assess the relative contribution of genes and environment in the etiology of DD[21]. A high heritability (80%) was estimated for DD occurrence suggesting a major role of genetic factors in the development of DD[21]. Thus, a large portion of the genetic heritability was not explained by previous GWAS leading to the arising question of how the 'missing heritability' of DD can be explained.

One explanation refers back to the initial basis of conducting GWAS, the 'common disease–common variant' hypothesis, which states that most of the genetic risks for common diseases are due to loci where common functional variants (minor allele frequency [MAF] > 5%) exist at each locus[22]. Therefore, GWAS applies an indirect association approach to identify tag SNPs associated with the disease phenotype. However, the GWAS tag SNPs are not necessarily biologically meaningful, but strongly correlated with the underlying functional variants[22]. This correlation is known as linkage disequilibrium (LD), the non-random combination between alleles at linked loci along a chromosome, which in part, reflects their proximity and the correspondingly low probability of recombination breaking the haplotype on which they are found.[23]

Recently, the fast development of next generation sequencing (NGS) has created the opportunity to directly map the variants. In particular, by target enrichment and NGS (Targeted NGS) of a GWAS-associated locus, the variability of the entire locus can be exhaustively identified, including coding and noncoding regions comprising all common and rare variants. Rare genetic variants affecting protein coding were recently identified to have a strong effect on the susceptibility of complex diseases, for example, Alzheimer's disease[24]. In addition, common variants associated with gene expression are increasingly shown to harbor regulatory roles by manifesting themselves as cause of gene expression differences[25]. Mapping of expression quantitative trait loci (eQTL) is one way to identify regulatory variants affecting gene expression. Several comprehensive projects including Genotype-Tissue Expression (GTEx), have established as a resource databases to search for eQTL in human normal tissues[25,26]. Combining eQTL data with disease associated genetic variants provides an opportunity to identify common regulatory variants influencing gene expression at GWAS loci.

Given that consistent evidence of several susceptibility loci linked to DD, fine mapping of GWAS-associated loci should be performed to identify functional rare and common variations represented by GWAS association signals.

1.3.2 The contribution of rare variants and candidate genes to DD

Over the past several years, the 'common disease–common variant' hypothesis was increasingly

challenged by the 'common disease–rare variant' hypothesis, as a growing body of evidence suggests that rare variants with a MAF of less than 5% also play an important role in complex trait etiology[27]. Currently, the widely accepted model for the genetic architecture of common diseases is the 'broad sense heritability' model, which refers to not only genetic components (which includes a large number of rare variants with a large-effect size and common variants with a small-effect size), but also the interactions between genetic components and environmental factors[22].

For DD, common variants tagging genetic risk loci have been captured by GWAS, however, all of them are in the non-coding regions of the genome and many of them are in intergenic regions. This makes it hard to address their functional consequences and identify candidate genes contributing to the DD phenotype. On the hand, the impact of rare variants on DD phenotype remains unanswered since the low-frequency variants cannot be completely represented by GWAS using genotyping arrays.

In the past few years, whole exome sequencing (WES) of all protein-coding regions has been considered a powerful tool to discover genes underlying unsolved Mendelian disorders[28]. More recently, WES was also used to prioritize candidate genes[29,30] and evaluate the contribution of rare variants to genetic heritability in complex diseases[31,32]. To date, no published WES studies on rare genetic variants have been reported in DD.

1.3.3 Transcriptome characterization of DD

Identifying genes carrying genetic variants is only one part of genetic studies. The ultimate goal is to understand the functional relevance of these candidates and the underlying disease biology. Whole transcriptome analysis of disease tissues/cells can provide the foundation for the understanding of the contribution of genetic variants, gene function and pathogenic mechanisms involved in diseases.

1.3.3.1 Findings and problems in previous gene expression profiling studies in DD

Global gene expression in DD has been studied in DD tissue biopsies or patient-derived cells using microarrays in several groups[8,33]. These studies have provided important insights for deregulated genes and pathways involved in DD.

An elevated level of β-catenin was detected in the DD tissues compared to patient-matched fat controls as well as in primary DD cells derived from DD tissues on collagen lattice[34]. No genetic variants were identified in *CTNNB1*, which encodes β-catenin, in DD patients by Sanger sequencing. Therefore, the altered expression of β-catenin may be explained by alterations of WNT/β-catenin pathway components or other pathway components, which modulate β-catenin stability[34], for example, by transforming growth factor-beta (TGFβ) pathway components[35-37]. In accordance with virtually all other fibrosis disorders, the TGFβ pathway is considered as a master regulator in DD by regulating fibroblasts proliferation and their differentiation into myofibroblasts[12,38]. Additionally, the increased β-

catenin expression observed in DD cells collagen lattice cultures suggests a potential role of ECM in modulating β-catenin levels in DD[34].

A number of microarray studies have been attempted to explore the abnormalities of ECM in DD. The high expression of various ECM components, such as *COL1A1*[33,39], *COL3A1*[33,39], *COL4A2*[39], *COL5A1*[39] and *TNC*[33,39], was observed in DD nodules. Increased expression of *MMP14* was also detected in DD nodules and associated with recurrence of nodules following surgical intervention in DD cases[40]. These studies provided important insights in the abnormal characteristics of the ECM in DD relevant tissues. However, the exact cause of the differential expression and the function of these components are still not clear, which may be due to several reasons.

First, it is difficult to define the function of a candidate gene using pure expression profiling information, which, by its nature, only captures the mRNA activity in tissues at the disease state without knowing genetic or mechanistic dysfunction. Including genetic information, for example WES data of patients, should facilitate the prioritization of the functional genes involved in transcriptional regulation.

Second, the controls used in previous expression profiling studies were insufficient. In some studies, the gene expression profiling for DD tissues (DDtis) was compared with internal controls, the unaffected palmar fat (DDfat) tissues from the same patients. While in other studies, the external controls — palmar fat tissues from patients undergoing open surgical carpal tunnel release (CTtis) were used as healthy tissue controls, since the truly healthy palmar tissues from healthy people are not available due to practical issues. Very rarely, both internal and external tissue controls were included in expression analysis[41].

Third, the absence of appropriate animal models, owing to the unique anatomical and functional features of hands, makes it difficult to design reliable *in vivo* functional studies. So far, the culture of primary cells derived from DD tissues (DDcell) served as the *in vitro* experimental model for DD to study gene function. However, one of the significant differences of cultured cells compared to *in vivo* tissues is the lack of ECM structural support and growth factor secretion.

In addition, although microarrays have enabled rapid and cost-effective analyses of gene expression profiling in DD tissues, there are still several weaknesses of this technique, such as the limited detection of low-abundance transcripts and identification of novel transcripts.

In contrast, the recent development of RNA-seq offers a great power to analyze gene expression at a high resolution and investigate different layers of transcriptome complexity[42]. RNA-seq analyzes the complementary DNA (cDNA) by means of NGS and maps sequence reads onto the reference genome. The massive capacity of RNA-seq allows not only the measurement of low abundance transcripts, but also the identification of novel transcripts and alternative splicing patterns associated with traits, which may be crucial in disease pathogenesis.

1.3.3.2 The unknown role of alternative splicing (AS) in DD

In humans, a gene has approximately 9 exons and 8 introns on average[43]. Genes need to be transcribed over their full length (including both introns and exons) to generate primary mRNAs or precursor mRNAs (pre-mRNAs). The introns are then removed from the pre-mRNAs and exons are joined together to form the functional mRNA. This process is catalyzed around splice sites at the ends of each intron by spliceosome, a highly dynamic ribonucleoprotein (RNP) complex comprised of five small nuclear ribonucleoproteins (U1, U2, U4/U6 and U5) and numerous other proteins[44]. From a given pre-mRNA to mature mRNAs, some constitutive exons are spliced into every mRNA produced, whereas many exons are alternatively spliced into various mRNA isoforms leading to the synthesis of different protein isoforms with diverse functions[45]. This process is defined as alternative splicing (AS).

There are six common types of AS events (Figure 1-3), which include exon skipping/inclusion (ESI), intron skipping/inclusion (ISI), alterative 5' splice sites (A5), alternative 3' splice sites (A3), multiple exon skipping (MESI) and mutually exclusive exons (MEE) (Figure 1-3)[46]. ESI is known as the most common splicing phenomenon in eukaryotes, describing that an exon is spliced in or out of the transcript thereby leading to extended or shortened mRNA isoforms[47]. A5 and A3 events occur when two or more splice sites are recognized at one end of an exon/intron. They account for 7.9% and 18.4% and of all AS events in higher eukaryotes, respectively[47]. The ISI event is defined by the retention of an intron in the mature mRNA transcript, accounting for less than 5% of known events in vertebrates and invertebrates[47]. MESI or MEE are complex AS events, which are subject to specific regulation[48].

In general, AS is tightly regulated in a tissue- or cell-specific manner. The splicing regulatory sequences and RNA-binding splicing factors, which recognize and bind to these sites, compose a common mechanism for setting up and maintaining AS patterns[45]. Tissue-specific splicing factors may activate or inhibit the use of splice sites or act in both ways depending on the sequence and position of the target site in the pre-mRNA.

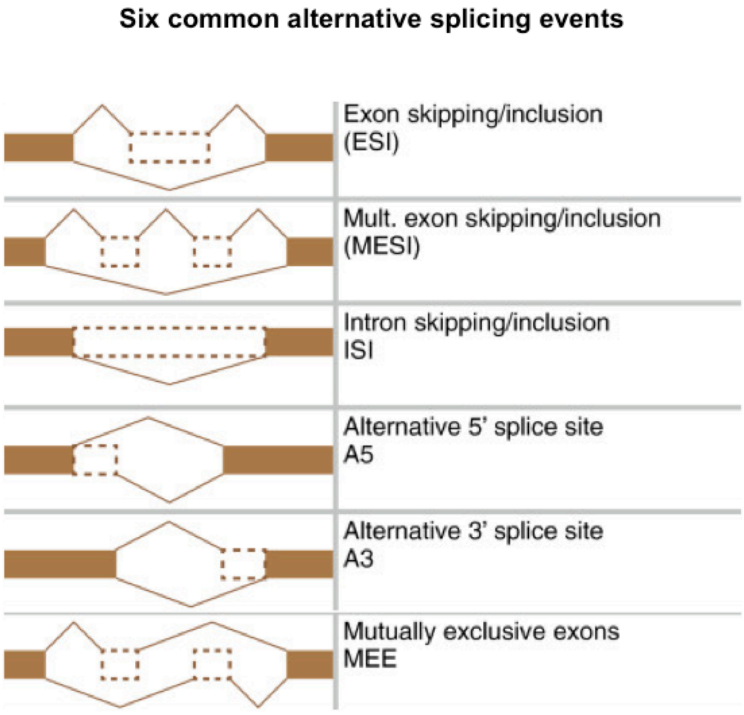


Figure 1-3 A schematic representation of common alternative splicing events

Six major splicing events were represented, which are exon skipping /inclusion (ESI), intron skipping/inclusion (ISI), alterative 5' splice sites (A5), alternative 3' splice sites (A3), multiple exon skipping (MESI) and mutually exclusive exons (MEE). The figure is modified from Vitting-Seerup et al.[46]

Unraveling AS of genes at the molecular level is important for understanding not only gene expression, but also disease causation as aberrant pre-mRNA splicing is the basis of many complex diseases. Here are three examples: first, the imbalanced ratio of 4R/3R tau isoform ratio, due to inclusion/exclusion of exon 10, is linked to tau-related neurodegeneration, for example Alzheimer disease. Second, the lack of SMN1 isoform (survival of motor neuron), which is required for assembly of macromolecular splicing factors, causes spinal muscular atrophy. Additionally, the dysregulation of the splicing process is common in cancer. This may include deregulated genome-wide splicing events, differentially expressed splicing factor genes, and aberrantly spliced cancer-critical genes. All this may contribute to tumorigenesis and tumor severity. One well-documented example is the association between aberrantly spliced *CD44* isoforms and metastasis[49].

So far, studies on the presence and regulation of AS in DD are completely lacking. Therefore, investigations of AS and their contribution to the DD phenotype is in great need and will broaden our understanding of the transcriptional misregulation in DD.

1.4 The aim of this project

DD represents an ideal, however not well studied, disease model to study aging-related diseases with high genetic susceptibility. Investigations to unravel the genetic architecture and molecular etiology of DD may offer an opportunity to develop biomarkers for prognosis and the selection of non-operative treatment strategies. To date, the genetic architecture of DD is mostly unknown due to the complex interactions between genes and environmental factors. So far a few risk loci explaining only about 1% of the genetic heritability have been identified by previous GWAS studies, however, many more genetic components are expected to exist.

Therefore, this thesis aimed to

1) identify functional variants at a DD risk locus using targeted NGS

2) prioritize DD phenotype-related genes carrying rare variants using WES

3) characterize the transcriptional deregulation in DD tissues/cells using RNA-seq.

Chapter 2 Method

2.1 Study design, subjects and ethical approval

The study design of this project is shown in Figure 2-1. The samples from participants were collected after written informed consent according to the protocols approved by the participating institutions. All participants are of European origin. Patients with DD were recruited. DNA was extracted from peripheral blood samples for targeted NGS and WES. RNA was isolated from disease relevant tissues/cells of patients undergoing a standard fasciectomy. Control palmar connective tissue/cell samples were obtained from patients undergoing carpal tunnel (CT) release. The Ethics Commission of the Faculty of Medicine of the University of Cologne fully approved the study.

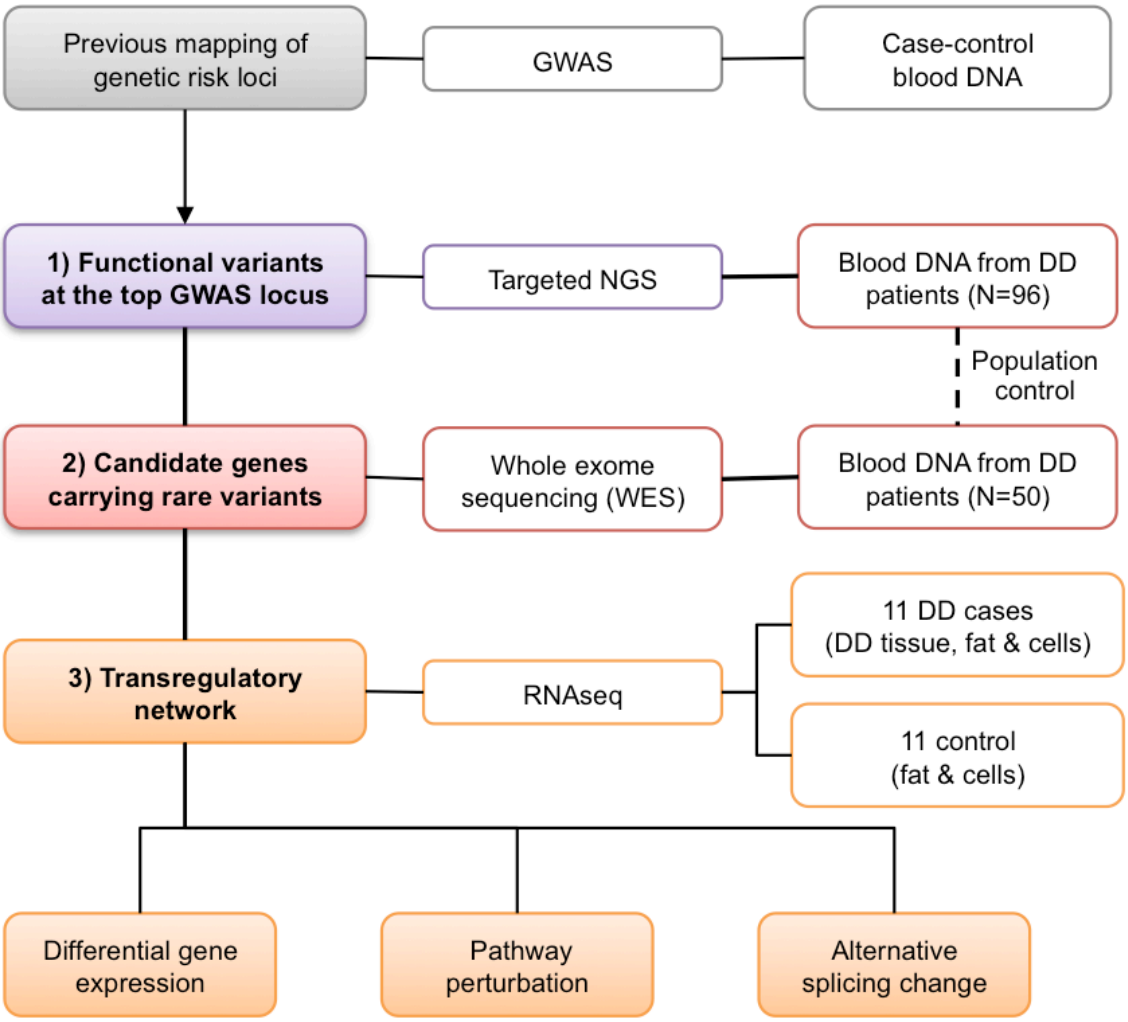


Figure 2-1 The study design of the project

2.2 Targeted NGS and data analysis

2.2.1 Target NGS and variant calling

Guided by LD structure (Figure 3-1), we selected a 500kb region (chr7: 37.77-38.27kb) containing the leading SNP rs16879765 on 7p14.1 for targeted sequencing. DNA was isolated from peripheral blood of 96 DD patients. The DD-associated locus was enriched using the SureSelect XT2 kit (Agilent). Libraries were prepared and labeled with barcodes for 100bp paired-end sequencing on the HiSeq 2000 device (Illumina) at the Cologne Center for Genomics (CCG). Sequence reads in FastQ format were mapped to the hg19/GRC37 reference genome using Varbank pipeline 2.16 (https://varbank.ccg.uni-koeln.de). Variant calling was performed using GATK v1.6[50] (in Varbank 2.16 pipeline) and 91.5% bases were called with calling accuracy more than 99.99% (Phred quality score 30). The mean target coverage was 77x. At 10x/30x coverage, 96%/91% of the region was covered. Annotation of variants was performed using Varbank 2.16 and Ensembl VEP 2.7[51].

2.2.2 Validation and replication using Sanger sequencing

Sanger sequencing was performed on the fifth exon of *EPDR1* gene. For standard PCR reactions, A 10 ul reaction was pipetted with 10 ng of genomic DNA, 0.25 ul BigDye terminator v3.1 (Applied Biosystems), 1.125 µM BigDye sequencing buffer and 0.25 µM primers (Eurofins). Cycling conditions were as follows: 96 °C for 10 seconds, 55 °C for 5s, 60 °C for 4 minutes, and the cycle was repeated 32 times. The quality of the PCR product was examined on an electrophoresis gel and only high quality PCR products were further cleaned. For a 10 µl PCR product cleanup reaction, 3 U Exonuclease I (Neo Lab), 0.9 U SAP (Shrimps Phosphatase Alkali, Promega), 2-10 ng PCR product and H_2O was mixed and incubated at 37˚C for 20 minutes and then at 85˚C for 15 minutes. Samples were stored at 4 ˚C until needed.

2.2.3 Pyrosequencing

Template DNA (6ng genomic DNA or external cDNA products) was amplified using HotStarTaq Plus DNA Polymerase kit (Qiagen) using stand reaction (2 U HotStarTaq Plus DNA polymerase, 1x PCR buffer, 200 μM of each dNTP and 0.1 μM of each primer) and amplification conditions (95 °C for 5 minutes, 30 cycle of 3 step cycling including 94 °C for 30 seconds, 60 °C for 30 seconds and 72 °C for 1 minute as well as final extension at 72 °C for 10 minutes). Reverse PCR primers were biotinylated for subsequent pyrosequencing analysis. Pyrosequencing reactions were carried on PSQ HS96A instrument and pyrosequencing SNP analysis software (PyroMarkTMQ96MD, Qiagen). Pyrosequencing signals for alternative alleles were normalized to signals for known reference alleles. Significant allele specific expression was normalized from experimental noise using T-test ($p<0.05$).

2.2.4 Protein model

Gremlin global statistical model method[52] was used to generate a multiple sequence alignment and construct the protein model for EPDR1. The EPDR1 protein sequence was searched against UniProt and the significant hits were added to an alignment, which was further searched against Uniprot. This procedure was repeated 4-8 times and a diverse alignment for the entire family of EPDR1 -like protein was identified. This alignment was then compared to a database of alignments (protein data bank, PDB) using HHsearh Jackhmmer (EBI). The alignment-to-alignment search eventually revealed the best hit to 3bmz in PDB database with 94% confidence that 3bmz shared the same folding with EPDR1. After threading and refining with co-evolution data and Rosetta v3.2, the best predicted model is shown in Figure 3-3A.

2.2.5 Transcription factor binding sites

We performed the prediction of transcription factor (TF) binding sites for rs149095633 using two bioinformatics tools: the Genomatix MatInspector[53] and TRANSFAC databases[54]. Genomatix MatInspector is a software tool that utilizes a large library of matrix descriptions for transcription factor binding sites (also named as motif) to locate matches in DNA sequences[54]. It uses two scores including the matrix similarity score and core similarity score to measure the quality of a match between the sequences and the consensus TF binding site matrix, which ranges from 0 to 1 with 1 denoting an exact match[53]. The core of each matrix is defined as the first five most conserved consecutive positions of a matrix[53]. Using TRANSFAC database, the consensus TF binding site matrix and the core of each matrix can be visualized in motif logos (for example, in Figure 3-4B), which scales each nucleotide by the total bits of information multiplied by the relative occurrence of the nucleotide at the position[55].

2.3 Whole exome sequencing (WES) and data analysis

2.3.1 Exome sequencing and variant calling

Blood DNA was isolated from 40 DD patients. Library construction and exome capture were performed with the SeqCap EZ Human Exome Library v2.0 (NimbleGen). The prepared library was sequenced on Illumina HiSeq 2000 using paired-end 100bp sequencing at CCG. Variants were called using using GATK HaplotypeCaller[50] under standard hard filtering parameter and variant quality score recalibration according to GATK Best Practices recommendations[56,57]. This allowed variant calling simultaneously using local de novo assembly and a Bayesian statistical model.

In total, 3.48 billion sequencing reads were aligned to the human genome (hg19/GRC37), with a 99.5% mean coverage across the exome region in 40 samples. In each sample, at least 20x/30x coverage of 95%/90% exome region of was achieved. A total of 4,214,948 loci in VCF file passed QC threshold, representing 3,933,980 SNPs and 301,509 indels/substitutions. Only high-confidence variants (99,323 variants with "High Qual") were further annotated by Annovar tool.

2.3.2 Annotation of variants by Annovar

ANNOVAR (2016Feb01)[58] was used to annotate high-confidence variants (99,323 High Qual variants) called from GATK. Variants were annotated and filtered for function priority (missense, nonsense and splice variants), conservation, rare variants (MAF ≤ 1% in large population database) and functional consequences (SIFT<0.05 applied, which predicts an amino acid substitution affects protein function based on the degree of conservation of the amino acid residues[59]) using annotate_variation.pl function. In total, 3919 variants in 3088 genes were considered as fictional rare variant candidates (Figure 3-6A). Variants were validated by read alignments visualization using Varbank 2.16 'Browse Reads' and IGV[60].

2.3.3 Phenotype-based gene prioritization

The Phenolyzer (Phenotype Based Gene Analyzer)[29] tool was used to prioritize genes based on DD related phenotypes, which includes 'flexion contracture of finger', 'connective tissue' and 'fibrosis'. In total, the three phenotypes and 3088 genes carrying rare variants (from Annovar) were used as input in the Phenolyzer. As a result, 88 genes were identified as highly related to three DD phenotypes (phenolyzer score ≥ 1). An overlapping list of 30 genes between 88 DD phenotype-related genes and 1774 genes with gene burden greater than two were prioritized (Figure 3-7A).

2.3.4 GESA of pathogenicity

The WGPA-GSEA tool[61,62] calculated if the genic intolerance scores EvolTol[63] (genome-wide score in the palmar part of the hand) of a list of genes occupy higher positions in the ranked gene list than it

would be expected by chance. Gene set enrichment scores and significance level of the enrichment (FDR adjusted q-value) suggests the list of 30 genes from Figure 3-7A was significantly enriched for intolerant genes in the palm (Figure 3-7B), which included 12 genes predicted as the top 25% genes that are intolerant to mutations in the palm tissue (Figure 3-7C, Table 3-5).

2.3.5 GO and pathway analysis using Enrichr

The gene ontology (GO)[64] covers three domains: biological processes, cellular components and molecular functions. The Enrichr[65] tool was used to detect GO features of 12 candidate genes (Table 3-5) beyond that which would be expected by chance (Fisher's exact test, Bonferroni adjusted p-value ≤ 0.05 is applied) (Table 3-6). Enrichr pathway analysis was used to map genes to KEGG pathways[66]. The p-value denotes the significance of the pathway correlation (Bonferroni adjusted p-value ≤ 0.05 was applied) (Table 3-7).

2.4 RNA-seq and whole transcriptome analysis

In this project, we propose that an appropriate transcriptome study design should involve not only the disease tissue, but also internal and external controls to interpret the origins of the observed changes in gene expression in disease tissues.

By comparing DDtis (palmar nodule biopsy from DD patients) to external tissue controls CTtis (palmar fat from carpal tunnel patients, which are considered as CT healthy controls without DD), the general DD-related changes in gene expression could be unraveled, though, the changes represent a mix of differences including the difference in genetic background between DD patients and CT controls and the difference in tissue types between DDtis (the nodule or cord tissues) and CTtis (mainly fat tissue). However, by comparing gene expression profiling between two fat controls, including internal control DDfat (perinodular fat from DD patients) and external fat control CTtis, the contribution of genetic components might be observed since both tissues are the same tissue type. Moreover, by comparing DDtis to DDfat — the fat tissue adjacent to DDtis, disease specific features could be identified.

In addition, by including DDcell (DDtis derived *in vitro* cells) and control CTcell (CTtis derived *in vitro* cells), an evaluation of the true *in vivo* relevance of the *in vitro* cell model could be conducted by comparing the gene expression profiling of DDtis/CTtis and DDcell/CTcell.

A flowchart illustrating the different analyses of RNA-seq data performed in this project is provided in Figure 2-2.

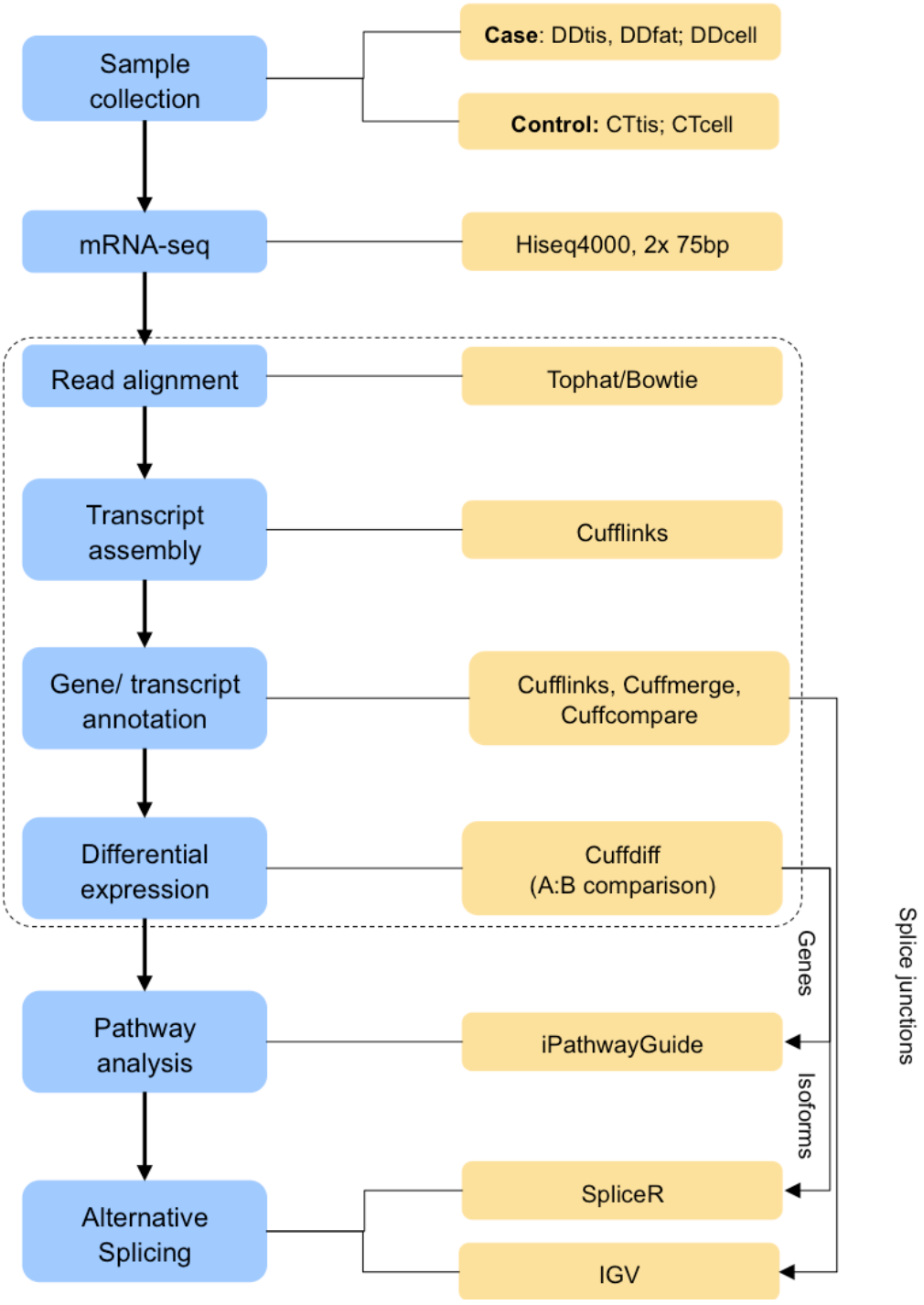


Figure 2-2 A schematic representation of the RNA-seq pipeline

2.4.1 RNA isolation

Using InviTrap Spin Universal RNA kit (Stratec Biomedical), total mRNA was extracted from DDtis (n=10), DDfat (n=9) and DDcell (n=10) from 11 DD patients as well as from CTtis (n=10) and CTcell (n=10) from 11 CT control people (Table 3-8). The RNA concentration was assessed via NanoDrop ND-8000 spectrophotometry (Thermo Scientific). The quality of the total RNA was evaluated using both agarose gel electrophoresis and Bioanalyzer 2100 (Agilent).

2.4.2 Library preparation and sequencing

For mRNA-seq sample preparation, the TruSeq stranded mRNA library prep kit (Illumina) was used. First, 1µg of each total RNA sample was used for polyA mRNA selection using streptavidin-coated magnetic beads. The polyA selected mRNA was fragmented and amplified for cDNA synthesis using reverse transcriptase and random hexamer priming. In addition, the amplified cDNA underwent double stranded cDNA conversion, end repair and adaptor ligation. The gel purification (2% agarose gel) was used for size selection and cDNA libraries ranging in size from 200–250 bp were generated. Finally, the libraries were amplified using PCR (15 cycles) and quantified using Bioanalyzer 2100 (Agilent). Each library was run at a concentration of 7pmol using paired-end 75 bp sequencing on a HiSeq 4000 device (Illumina).

2.4.3 RNA-seq read mapping, transcript assembly and abundance estimation

In total, 32 billion reads were sequenced on HiSeq 4000 device. An average of 64 million reads (75bp paired-end) per sample were sequenced and aligned to the hg19/GRC37 (Ensembl 75) human genome using TopHat 2.1.1 (based on Bowtie)[42], which truncated each read of the pair to 25nt and aligned each end of the pair separately under mammalian default parameters, leading to an average mapping of 92.8% with a standard deviation of 5.3%.

The aligned reads were then assembled by Cufflinks package (including Cufflinks, Cuffmerge and Cuffcompare)[42] to reveal novel transcripts and genes as well as low abundance transcripts. The relative abundance of each transcript was normalized to Fragments Per Kilobase of exon per Million fragments (FPKM)[42].

Cuffdiff program was used for differential analysis of each transcript and gene between two conditions. With multiple replicates, Cuffdiff learns how read counts vary for each gene across the replicates and uses the variance estimates to calculate the significance of observed changes in expression[42]. A t-test was used on ratio of log transformed FPKM between two condition, which approximately follows a normal distribution[42]. Limma (moderated t-statistic) was used to calculate p-values using empirical Bayes estimates for standard error and degrees of freedom[42].

In each comparison, genes with minimum FPKM values of expression higher than 1 (minFPKM≥1) in either condition were first selected. Then, genes with |fold change| ≥ 1.5 (that is |log2 old change| ≥ 0.6) and FDR adjusted p-value ≤ 0.05 after Benjamini-Hochberg correction for multiple-testing with significance level "yes" were considered as significantly DEGs . The gene expression volcano plots were derived from iPathwayGuide tool.

2.4.4 GO overrepresentation and pathway perturbation

The Cuffdiff results of differential gene expression analysis were uploaded to the iPathwayGuide tool (www.advaitabio.com/ipathwayguide.html)[67,68]. The DEGs were annotated for overrepresented GO features[64] and perturbed KEGG pathways[66] using iPathwayGuide.

For the large DEG sets, to reduce the false positive rate caused by GO term 'inheritance problem' — the higher level (more general) GO terms inherit annotations from the lower level (more specific) descendant terms, a minimum elim adjusted p value 0.05 was used a cuts-off to screen the GO term enrichment[69]. The elim (elimination gene) method first investigates the GO terms from bottom to — top, the more specific to more general and then removes the genes mapped to significant GO terms from higher level (the more general level)[69]. The pathway perturbation was analyzed on DEG sets.

The iPathwayGuide employs a third generation pathway topology (PT)-based pathway perturbation analysis[68] by evaluating both pathway over-representation and accumulation. For over-representation analysis, the number of DEGs involved in a pathway was compared between two comparisons. For pathway accumulation analysis, the significance of a particular DEG to a pathway was considered in determining the overall impact on the pathway by examining all annotated functions/interactions of the gene in KEGG pathway databases[66].

The results are shown in scatter plots. For example, in Figure 3-11A, each round circle represents a pathway. The x-axis represents the significance of pathway overrepresentation (–log10OVA p-value). The y-axis represents the significance of pathway accumulation (–log10Acc p-value). The pathway perturbation was further calculated as an additive measurement of pathway overrepresentation and accumulation. The red circle indicates a pathway is significantly perturbed (Bonferroni adjusted p-value ≤ 0.05), whereas the grey circle indicates non-significance.

2.4.6 Alternative splicing analysis

Using Cuffdiff, pair-wise comparisons of differentially expressed transcripts (isoforms) between three tissue types (DDfat/CTtis, DDtis/DDfat and DDtis/CTtis) were first conducted. The results were further used as input in R package SpliceR[46], which enabled us to perform an elaborate genome-wide analysis of alternative splicing (AS) including two major aspects: 1) analysis of the AS events in disease tissues and controls; 2) identification of isoform switching in disease and controls.

First, for each gene annotated by Cufflinks, spliceR constructs the hypothetical pre-RNA based on the exon information from all transcripts originating from that gene. Subsequently, all transcripts are compared to this hypothetical pre-RNA in a pairwise manner, and AS events are classified and annotated.

Second, for each significant differentially expressed isoform identified by Cuffdiff, spliceR calculates an isoform fraction (IF) value, which is calculated as (transcript FPKM/gene FPKM)% to represent the contribution of a transcript to the expression of the parent gene. The isoform fraction change dIF (dIF = $IF_{condition\ A}$ / $IF_{condition\ B}$), which measures the ratio of IF values between two conditions, was also calculated. The isoform switch is defined by a large positive or negative dIF between two conditions (|dIf|≥1.5, the same as |log2 dIF| ≥0.6).

However, to avoid overestimation of the number of functionally relevant isoform switching, the coding potential and nonsense mediated decay (NMD) sensitivity were also predicated for each isoform. Isoforms were marked NMD-sensitive if the stop codon falls more than 50 nt upstream of the final exon-exon junction indicating a pre-mature stop codon (PTC)[46].

Additionally, the IGV tool was used to visualize the exon coverage and splice junctions using the aligned reads from Tophat and Cufflinks package[60]. Alternative splicing events were observed on IGV Sashimi plot[70]. To filter out low-count splicing events, the minimum junction coverage was set to 30 when Sashimi plots were generated unless elsewhere specified.

Overall, to identify biologically relevant isoforms in DD pathogenesis, isoform switches in three tissue comparisons (DDtis/DDfat, DDtis/CTtis and DDfat/CTtis) were filtered by 5 steps: a) the minimal isoform expression cutoff: isoform expression ≥1 FPKM, which supports either tissue type in one comparison; b) a minimal threshold for isoform expression fold change: |log2 (fold change)| ≥ 0.6; c) a minimal threshold for isoform fraction change: |log2 dIF| ≥0.6; d) the isoform is marked NMD-insensitive; 5) the AS events of the isoform can be visualized (or validated) on IGV.

2.4.7 Heatmap and dendrogram visualization

For heatmap generation in R (Figure 3-9, 3.19 and 3.20B, 3.26), due to the variable levels of expression of individual genes, Z-score normalized gene expression was used[71]. The Z score was determined for any row n by the formula Z = (Sn− Smean)/SD, where Smean is the average signal intensity for the gene (across the row) and SD is the standard deviation. The row Z-score normalized gene expression was plotted in two color scale with the expression higher than average represented in red (Z > +1), the expression lower than average in blue (Z < −1), and average in orange or white (Z = 0).

The dendrograms added to the heatmaps were used to measure the hierarchy clustering of samples and gene expression profiles using one-minus Spearman's rank. Genes with similar expression or

samples with similar gene expression profiles were grouped together and connected by a short line between two gene nodes.

For example, in Figure 3-9, the gene expression clustering of all DEGs in tissue groups is shown in a heatmap. Genes with significant expression change in at least one of the three tissue comparisons (DDtis/DDfat, DDfat/CTtis and DDtis/CTtis) were first selected. Then the Cufflinks normalized gene expression FPKM value for individual tissue sample was Z-score transformed. The Z score normalization expresses each gene expression profile as a deviation from the mean in standard-deviation units and allows the comparison of gene expression patterns whose absolute expression levels may differ by orders of magnitude. In summary, this scaled row Z-score value was plotted in red–blue color scale with the expression higher than average represented in red ($Z > +1$), the expression lower than average in blue ($Z < -1$), and the average expression represented in orange ($Z = 0$) (Figure 3-9A).

Chapter 3 Results

3.1 Identification of functional variants at the top GWAS locus at 7p14.1

The strongest SNP associated with DD is rs16879765, which lies in an intron of *EPDR1* and upstream of *SFRP4* ($p=5.6x10^{-39}$, Odds Ratio 1.98) at 7p14.1 locus[19]. In addition, the rs16879765 showed a stronger association in the subgroup of patients with a familial predisposition of DD ($p=4.7x10^{-5}$, Odds Ratio 2.08, $n_1 = 184$) compared to the patients without known family history ($n_2 = 281$) suggesting a pivotal role of this locus in the genetically caused pathogenesis of DD[72].

Therefore, we performed a fine mapping study of the 7p14.1 locus by targeted NGS to capture the unrecognized functional variants that are supposed to be in LD with the GWAS tag SNP rs16879765. The region for target enrichment included a 500 kb haploblock region containing the entire genes *EPDR1, SFRP4, TXNDC3* and *GPR141* as well as part of *STARD3NL* (Figure 3-1). To identify causal variants represented by GWAS SNP rs16879765, we adopted a 'risk haplotype block' guided strategy under the hypothesis that the causal variants lie in the risk haplotype tagged by the risk allele of rs16879765 (risk allele A, non-risk allele G). In the discovery set (Table 3-1), 96 cases from the aforementioned GWAS cohort (Germany and Switzerland)[20] with available (or imputed) genotype data were selected, including 23 cases homozygous for the risk allele (AA), 51 cases heterozygous (AG) and 22 cases homozygous for the non-risk allele (GG).

Table 3-1 The sample set for targeted NGS at 7p14.1

Sample set	Method	No. of DD blood DNA samples			
		No. of patients	Genotype for GWAS leading SNP rs16879765 (risk allele A; non-risk allele G)		
			AA	AG	GG
Discovery set	Targeted NGS	96	23	51	22
Validation set	Sanger sequencing	280			

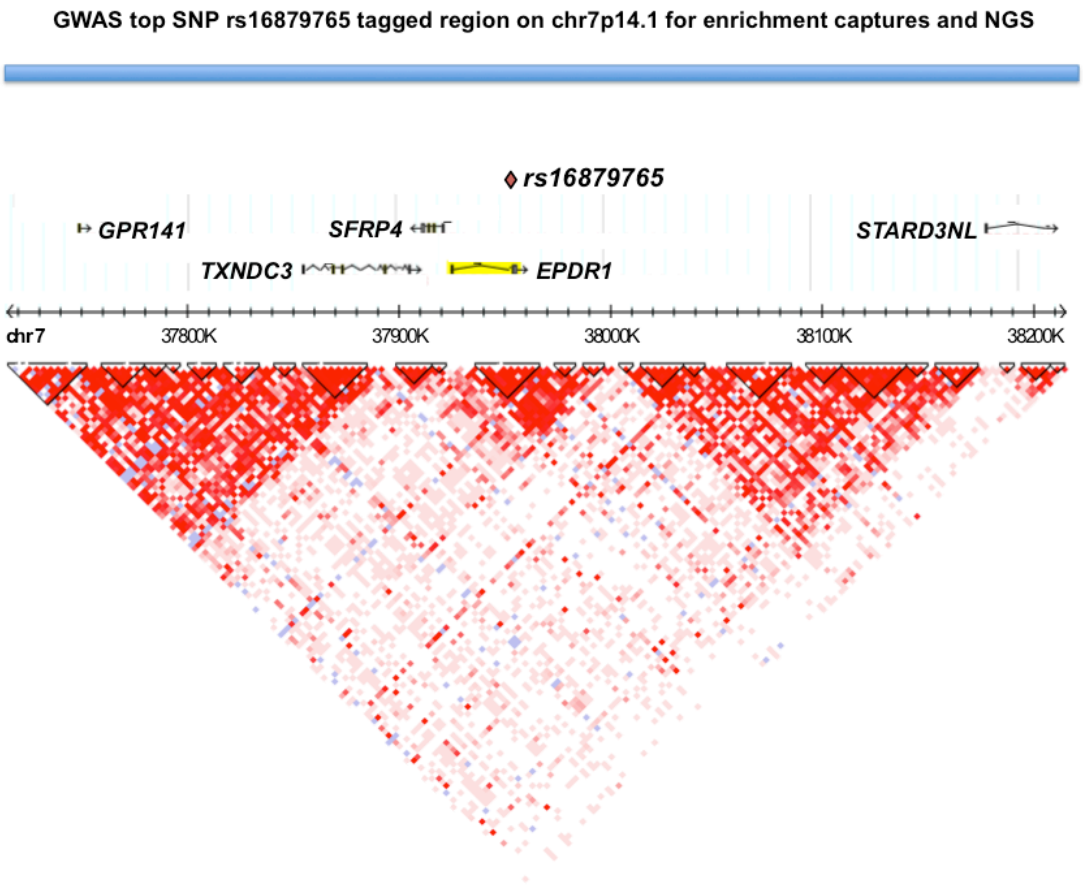


Figure 3-1 Targeted NGS of a 500 kb region at 7p14.1

The region for target enrichment included a 500 kb haploblock region containing the entire EPDR1, SFRP4, TXNDC3 and GPR141. The GWAS tag SNP rs16879765 is indicated in red diamond. The haplomap was constructed based on the pairwise r^2 values from Hapmap data for CEU (Central European ancestry) population.

3.1.1 Targeted NGS and variant calling in the discovery dataset

DNA was isolated from peripheral blood of 96 DD patients. A 500kb DD-associated locus (chr7: 37.77-38.27kb) was enriched in these samples and further sequenced by NGS. In total, 41.7 million sequencing reads were aligned to the human genome (hg19/GRC37), providing a 77x mean coverage across the 500kb targeted region in 96 samples. Variant calling was performed using GATK v1.650 (in Varbank 2.16 pipeline) under standard hard filtering parameters and variant quality score recalibration according to GATK Best Practices recommendations

As a result, 12,308 variants were called and 7,242 variants remained after quality control, which were then annotated to examine if variants overlapped with bioinformatic features using Varbank 2.16 and Ensembl VEP 2.7[51]. The majority of variants were annotated as intronic variants (73%) in the target region (Figure 3-2A). Only less than 1% of all variants were coding variants.

A

Category	No. of variants
Variants processed	12038
Variants remained after filtering	7242
Genes	7
Transcripts	28
Regulatory features	51

B

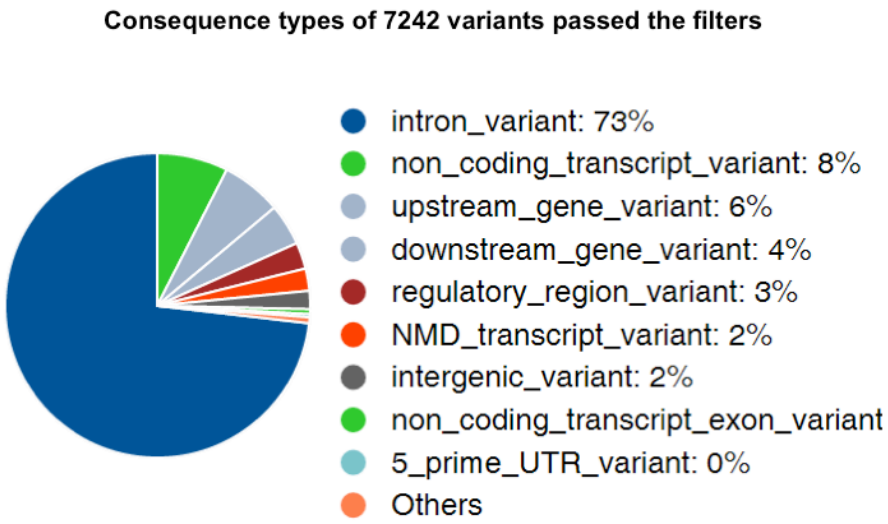


Figure 3-2 Variant calling and annotation in the 500 kb region at 7p14.1

(A) A total number of 12,038 variants were called by GATKv1.6 in Varbank pipeline 2.16. After quality control and Ensembl VEP 2.7 annotation, 7242 variants remained and were distributed in 7 genes, 28 transcripts and 51 regulatory features.

(B) Functional consequences of 7242 variants annotated by Ensembl VEP 2.7 based on Sequence Ontology (SO) consequence terms[73]. Coding variants were included in 'Others'.

3.1.2 Functional rare coding variants at 7p1.4

3.1.2.1 Identification of two rare coding variants on GWAS risk haplotype

Of the coding variants identified in the discovery dataset, two rare nonsynonymous single nucleotide variants (SNVs) in *EPDR1*, defined as variants with MAF less than 1% in 1KG[74] and ExAC[75] large population database, were predicted to be damaging to protein function by multiple annotation tools, such as SIFT[59], MutationTaster[76], Fathmm[77], CADD[78] and DANN[79] (Table 3-2).

Table 3-2 Nonsynonymous variants in *EPDR1* identified by targeted NGS at 7p14.1

dbSNP ID	Codon;	Variant annotation					Discovery set n=96	Validation dataset n=280		ExAC NFE n=33,368	Fisher's exact test (Validation dataset vs. ExAC NFE)		
	Amino acid change	SIFT	Mutation Taster	Fathm mMKL	CADD	DANN	GWAS SNP rs16879765 genotype	GWAS SNP rs16879765 genotype	MAF	MAF	p-value	Odds ratio	95%CI
rs149095633	cCc>c**T**C	D	D	D	31	0.78	4 cases (3AG, 1AA)	8 cases (7AG, 1AA)	1.40%	0.04%	2.6E-07	37.5	17.7-79.7
	p.P121**L**												
rs374633177	gTg>g**C**g	T	D	D	22	0.91	1 case (AG)	1 case (AG)	0.17%	0.03%	0.14	6.6	0.8-49.7
	p.V102**A**												

D: deleterious; T: tolerated
CADD score > 10: suggetes deleterious
DANN score (0-1): the closer to 1, the higher the probability of being deleterious

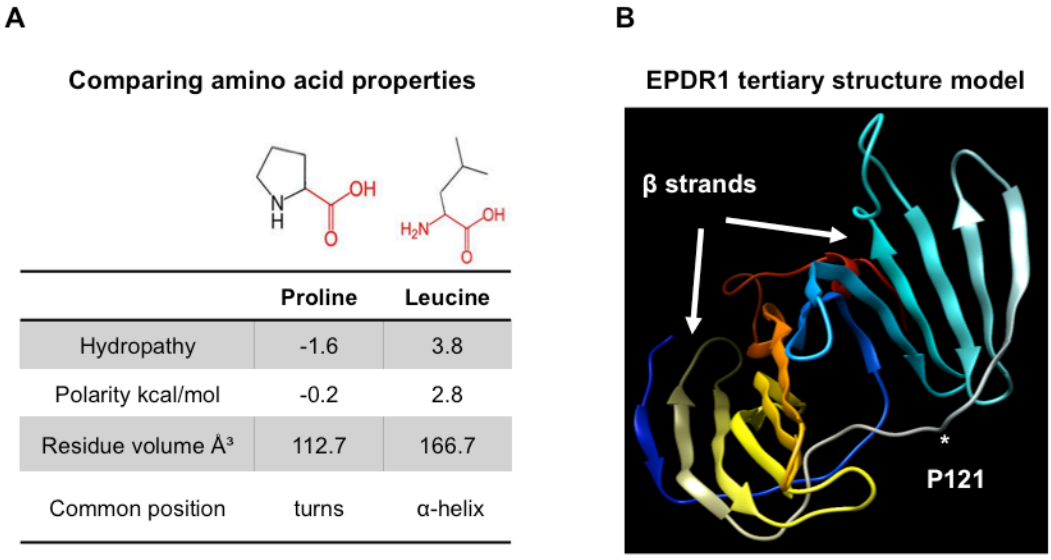
In the discovery cohort including 96 cases, rs149095633 (C>T, p.P121L) was detected in 4 cases and rs37463317 (T>C, p.V102A) was identified in only 1 case. By Sanger sequencing, we successfully validated the presence of these two variants and further replicated the findings in a random and independent set containing 280 DD blood DNA samples, which were genotyped in a recently reported GWAS[20]. No extra case carrying rs37463317 (T>C, p.V102A) was identified. However, additional 4 DD patients with the rare variant rs149095633 (C>T, p.P121L) were identified. In total, 8 out of 280 DD cases carried the heterozygous T allele of rs149095633 (C>T, p.P121L), which leads to an enrichment of variant allele frequency in DD patients (n=280) compared to ancestry matched non-Finnish European (NFE) population in ExAC database[75] (n=33,368, Fisher's exact p-value= 2.63E-07, Odds Ratio 37.5) (Table 3-2). In addition, all the 9 patients in the validation set with either the rs149095633 or rs37463317 rare variant, carried the risk allele A of the GWAS tag SNP rs16879765 (genotype AA or AG, the LD r2 between rs149095633) (Table 3-2). And both coding SNVs are within 0.62 kb distance from the GWAS tag SNP rs16879765.

Taken together, by targeted NGS of a 500kb region at the GWAS risk locus 7p14.1, we identified two rare coding variants related to the GWAS risk haplotype (tagged by rs16879765) in DD patients. Both coding variants were predicted to be deleterious. In particular, the SNV rs149095633 (P>L) was significantly enriched in the DD cohort (n=280).

3.1.2.2 rs149095633 is a functional rare coding variant in DD

The SNV rs149095633 (C>T) leads to an amino acid substitution from proline (C**C**C) to leucine (C**T**C) at position 121 (p.P121L, NP_060019 classical isoform)**.** Leucine is substantially different from proline regarding its amino acid properties, for instance, leucine is more hydrophobic[80] and polar[81] than proline (Figure 3-3A). To further assess the in silico significance of the enriched rare variant rs149095633 (p.P121L) on EPDR1, we first built a 3-D structure model for the EPDR1 classical protein isoform (NP_060019, 224 amino acids) using Gremlin method[52,82]. The best-predicted

structural model with a high confidence score of 94% is shown in Figure 3-3B. Based on this model, the EPDR1 is suggested to be a monomer composed of two β-strands and a connective loop called β-turn between them. The β-turn conformation is known to be crucial in protein structure by actively facilitating cooperative formation of β-strands[83]. Additionally, β-turns are mostly surface-exposed, which makes them well suited for ligand binding and protein-protein or protein-nucleic acid interactions, which in turn may modulate protein functions and intermolecular interactions[83].



	Proline	Leucine
Hydropathy	-1.6	3.8
Polarity kcal/mol	-0.2	2.8
Residue volume Å³	112.7	166.7
Common position	turns	α-helix

Figure 3-3 The impact of rs149095633 (p.P121L) on EPDR1 protein

(A) A comparison of the amino acid features between proline and leucine[81]. Leucine is more hydrophobic, and polar than proline. The residual volume of leucine is larger than proline. Leucine is commonly found in α-helix, whereas proline is commonly found in turns.

(B) A 3-D structure model for EPDR1 classical protein isoform (NP_060019) using Gremlin method[52,82] with a high confidence score of 94%. According to this model, EPDR1 protein is suggested to be composed of two β strands and a connective loop called β-turn between them. The SNV rs149095633 leads to an amino acid substitution from proline (CCC) to leucine (C**T**C) at position 121 (p.P121L) in the β-turn domain of EPDR1.

Additionally, EPDR1 (NP_060019) is a 224 amino acid protein including one highly conserved Ependymin domain (PF00811) spanning residues 40-223 (http://www.ebi.ac.uk/interpro/protein/-Q9UM22). The molecular function of ependymins appears to be related to cell contact phenomena involving the extracellular matrix[84]. The SNV rs149095633 (p.P121L) lies in the Epedymin domain of EPDR1. Taken together, the rare coding variant rs149095633 (p.P121L) is likely to affect the protein structure and function of EPDR1.

Recently, studies suggested that about 15% of human codons are dual-use codons, which specify both amino acids and transcription factor (TF) recognition sites[85]. This suggests the potential for coding exons to accommodate regulatory code[85].

Therefore, using Genomatix MatInspector tool[53] and TRANSFAC databases[54], we examined the effect of the rs149095633 missense mutation on TF binding. The MatInspector analysis revealed that the minor allele T of rs149095633 (C>T) and DNA sequences around it created a perfect match to the core of consensus binding motif of AP-4 (with a high core similarity score 1 and matrix similarity score 0.864) (Figure 3-4A)[53]. On the contrary, the same region containing the non-risk allele C of rs149095633 did not correspond to any known TF binding motif. The core of the AP-4 consensus binding motif (5'-CAGC**T**G-3', the position of rs149095633*T is shown in bold) is represented in a motif logo (Figure 3-4B)[54]. AP-4 (activating enhancer-binding protein 4) is a highly conserved member of the basic helix-loop-helix-zipper family of TFs[86]. AP-4 has been identified as both a transcriptional activator[87,88] as well as a transcriptional repressor of gene expression[89,90]. Therefore, it is possible that the allele-specific binding of AP-4 on *EPDR1* due to rs149095633 may influence *EPDR1* expression.

To capture whether there is a direct cis-regulatory effect of rs149095633, we performed Pyrosequencing, a highly sensitive sequencing method using luminometric detection of released pyrophosphate during nucleotide incorporation[91] which enables high-accuracy quantification of alleles at a position of interest in DNA or RNA samples. In brief, we isolated high quality RNA from available DD tissues (n=2) and DD cells (matched DD tissue-derived primary cells, n=2) from two patients heterozygous for rs149095633. Blood DNA samples from these two DD patients and four random DD patients were used as internal and external controls.

The RNA from DD tissues and DD cells was further converted in cDNA, which was further used for Pyrosequencing. The luminometric signal intensity of allele T and C of rs149095633 in each cDNA (or blood DNA) sample was quantified and determined in allele percentage. On average, we observed an allelic imbalance at position rs149095633 with a 20% increased expression of risk allele T compared to the reference allele C in DD cells (n=2, Figure 3-4C), but not in DD tissues or control blood DNA samples. Thus, the rs149095633 missense variant is not only expressed in DD cells, but also appears to cause a suggestive allelic imbalance favoring the expression of the risk allele T in DD cells.

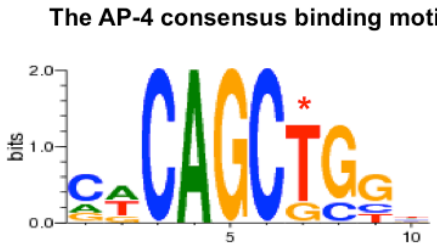
In summary, the rare variant rs149095633 (C>T) in *EPDR1* is enriched in the DD cohort (1.4%, n=280) compared to a large European population control (0.04%, ExAC NFE, n=33,368). It is annotated as a deleterious missense variant (p.P121L). Based on the predicted protein model, the substitution of leucine for proline is suggested to affect the EPDR1 protein structure, stability and function. Moreover, the cell-type specific allele expression was observed for rs149095633 with higher expression of the risk allele (T) compared to the non-risk allele (C), suggesting a contribution to higher expression of *EPDR1.* In addition, the allelic imbalance of rs149095633 may be relevant to the function of AP-4 since it creates a consensus-binding site for AP-4. In the future, studies to clarify

whether higher expression of rs149095633*T has a direct consequence on *EPDR1* expression and function and whether AP-4 is a transcriptional activator of *EPDR1* are required.

A

SNP	Codon	Loss/ Gain	TF matrix	Core similarity	Matrix similarity
rs149095633	cCc<cTC	Gain	AP-4	1	0.864

B



C

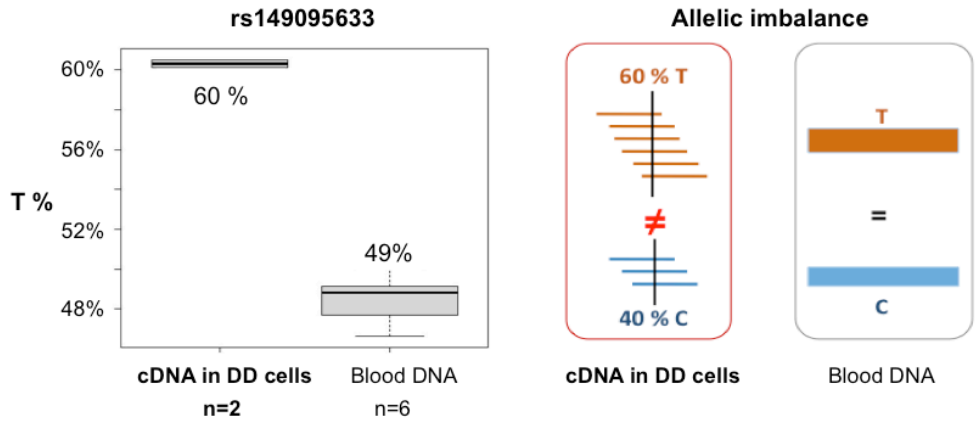


Figure 3-4 The allele specific expression of rs149095633

(A) The SNV rs149095633 changes C to T. Based on Genomatix MatInspector prediction, the minor allele T and the DNA sequences around it creates a perfect match to the core of the AP-4 consensus binding motif with a high core similarity score (=1) and matrix similarity score (=0.864).

(B) The core of the AP-4 motif (5'-CAGC**T**G-3') is shown in a motif logo, which scales each nucleotide by the total bits of information multiplied by the relative occurrence of the nucleotide at the position. The T marked by * is the position of rs149095633 T*.

(C) Left: An allelic imbalance at rs149095633 position was detected by Pyrosequencing, which displayed 60% allele expression for risk allele T and 40% for the reference allele C in available DD cells (n=2). The allelic imbalance on rs149095633 was not observed in blood DNA control samples (n=6, 49% expression for allele T and 51% expression for allele C).

Right: A cartoon representation of allelic imbalance at rs149095633 in DD cells and allelic balance in blood DNA controls.

3.1.3 A functional common variant in *EPDR1* at 7p14.1

To test whether common variants in LD with GWAS tag SNP are associated with gene expression, we manually searched whether SNPs, which are in LD ($r^2 \geq 0.2$) with rs16879765 in Hapmap CEU population[23,92], are eQTLs in GTEx database. In eight tissue types (subcutaneous adipose, lung, stomach, pancreas, thyroid etc) in GTEx, rs2044831, a common synonymous variant for *EPDR1* (Table 3-3), was identified as a significant eQTL for *EPDR1* expression. The allele C of rs2044831 is related to increased expression of *EPDR1* in all the 8 tissue types (an example of rs2044831 as an eQTL for *EPDR1* in subcutaneous adipose tissue is shown in Figure 3-5A).

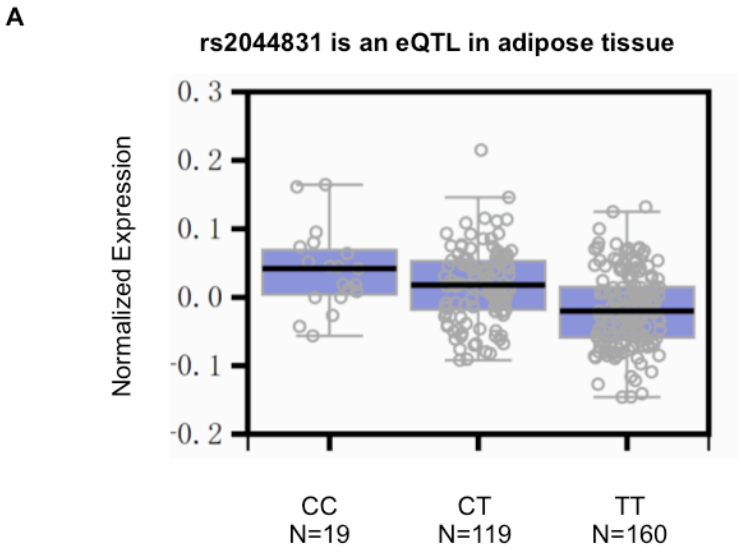
The common variant rs2044831 is in moderate LD with the GWAS tag SNP rs16879765 in DD (r^2=0.38 in Hapmap European population)[23,92]. To test whether rs2044831 is an eQTL for *EPDR1* expression in DD tissues/cells by expression analysis, a large sample size of 200 is required based on statistical power analysis[25]. However, such a sample size of DD tissues/cells was not available in this project.

Therefore, we adopted another method — Pyrosequencing, which allowed us to determine and measure the cis effect of rs2044831 by analyzing the allelic expression using a small sample size of DD cells. The RNA of DD cells (n=17) from DD patients heterozygous for rs2044831 (genotype CT) was converted in cDNA and used for Pyrosequencing. As shown in Figure 3-5B, a significantly higher expression of the DD-risk allele C (54%) was detected compared to non-risk allele T (47%), suggesting rs2044831 is a cis-eQTL for *EPDR1* expression in DD cells. The risk allele C is likely related to increased *EPDR1* expression.

Table 3-3 A common regulatory variant identified at 7p14.1

dbSNP ID	Uniprot	Risk allele in DD	Non risk allele in DD	LD r^2 with GWAS tag SNP rs16879765	Variant annotation			Allele frequency for risk allele C in DD	
					Consequence	CADD	GTEx	Discovery set (n=96)	ExAC NFE (n=33,368)
rs2044831	EPDR1	**C**	T	0.38 (Hapmap CEU)	Synonymous p.I=	18.8	Cis-eQTL in 8 tissue types	30%	20%

CADD score > 10: suggests deleterious



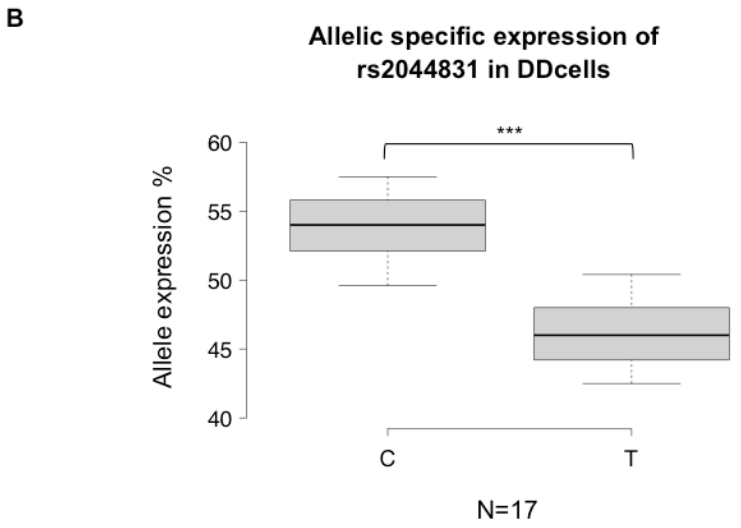


Figure 3-5 rs2044831 is an eQTL candidate for *EPDR1* in DD cells

(A) rs2044831 is a an eQTL for *EPDR1* in subcutaneous adipose in GTEx database. The rs2044831*C allele is associated with increased expression of *EPDR1* compared to the rs2044831*T allele in subcutaneous adipose as well as other 7 tissue types.

(B) Using Pyrosequencing, a significant allelic imbalance was observed for rs2044831 in DDcells (n=17, student's t-test p-value=1.08E-05, normality was tested by Shapiro-Wilk test). Higher expression of DD-risk allele C (54%) was detected compared to non-risk allele T (47%).

3.2 Prioritization of candidate genes carrying rare variants in DD

To test whether genome-wide rare variants contribute to the high genetic predisposition of DD, we performed whole exome sequencing (WES) of DD patients in a pilot study.

Low-frequency coding variants have long been known to contribute to family-based complex diseases including early-onset forms[93] or Mendelian forms[94] of complex diseases. Family-based WES of affected individuals and their unaffected family members has become a crucial tool to evaluate the contribution of rare variants to the disease trait in the family and unravel new candidate genes and pathways with relevance to molecular pathogenesis of common disorders[95]. Such study design is powerful and cost-effective but challenging to be implemented for DD. Because most of the DD cases are late-onset and consequently multi-generation pedigrees for DD were hardly available.

This leaves us to choose another strategy — the population-based strategy for which non-related cases were sequenced to identify rare variants in DD with effect on population level. However, the statistical power of rare variant association tests is usually very low unless the sample size or variant effect size are very large[96]. Moreover, like many complex disorders, DD susceptibility is likely to be associated with the effects of multiple variants and gene-gene interactions in contrast to monogenic disorders.

Therefore, in this small-size WES pilot study, instead of purely prioritizing variants based on their variant score (frequency and deleteriousness), I used a phenotype-driven method to prioritize functional candidate genes using a combination of disease phenotype ontology, gene intolerance scores and variant scores. In total, 12 candidate genes carrying rare deleterious variants were identified in DD patients.

3.2.1 Characteristics of WES study cohort

Two major criteria were used to select DD patients for WES. The first criterion was to select DD patients with an early age of first operation. The second was to select patients with a known family history of DD (at least one family member affected within two generations).

Accordingly, 40 DD patients were selected with high genetic predisposition possibilities (Table 3-4). The average age of these 40 patients is 44 years, which is much lower than the typical manifestation age 55-64 years[97]. Moreover, 39 patients have a family history of DD and 30 patients have both hands affected. Blood DNA was isolated from 40 DD patients and sequenced on Illumina HiSeq 2000.

Table 3-4 The DD cohort for WES

Patient ID	Gender	Age at 1st DD operation	Family history	Hand affected	Country of origin
DDpatient1	m	46	both parental lines	both	DE
DDpatient2	f	40	fatherside related	both	DE
DDpatient3	m	44	fatherside related	both	DE
DDpatient4	m	48	fatherside related	both	DE
DDpatient5	f	68	motherside related	both	DE
DDpatient6	m	41	motherside related	both	DE
DDpatient7	m	42	motherside related	both	DE
DDpatient8	m	47	motherside related	both	DE
DDpatient9	m	48	motherside related	both	DE
DDpatient10	m	45	positive	both	DE
DDpatient11	m	41	uncle	both	DE
DDpatient12	m	37	father	both	DE
DDpatient13	m	44	father	both	DE
DDpatient14	f	48	father	both	DE
DDpatient15	m	44	father	both	DE
DDpatient16	m	44	father	left	DE
DDpatient17	m	49	father	right	DE
DDpatient18	m	50	father	both	DE
DDpatient19	m	51	father, brother	right	DE
DDpatient20	m	43	father, brother, sister	both	DE

Continued on the next page

Patient ID	Gender	1st Operation age	Family history	Hand affected	Country of origin
DDpatient21	m	45	father, grandmother	both	DE
DDpatient22	f	47	father, mother	left	DE
DDpatient23	m	35	father, mother, brother, uncle (fatherside)	both	DE
DDpatient24	m	30	father, mother, sister	both	DE
DDpatient25	m	43	fatherside related	both	DE
DDpatient26	m	44	grandfather (fatherside)	both	DE
DDpatient27	m	47	grandfather (fatherside)	right	DE
DDpatient28	f	41	grandfather (motherside)	left	DE
DDpatient29	m	45	grandfather (motherside)	both	DE
DDpatient30	m	48	grandmother	both	DE
DDpatient31	m	40	granduncle	left	DE
DDpatient32	f	34	mother	both	DE
DDpatient33	m	50	mother	both	DE
DDpatient34	f	48	mother, brother	left	DE
DDpatient35	m	48	mother, grandfather (motherside)	both	DE
DDpatient36	m	34	motherside related	both	DE
DDpatient37	m	39	positive	both	DE
DDpatient38	m	51	positive	both	DE
DDpatient39	m	44	uncle (motherside)	right	DE/PL
DDpatient40	m	28	not reported	right	DE

3.2.2 Functional annotation of variants in WES data

A two-step integrative approach was applied to annotate variants and prioritize disease-causing candidate genes (see Figure 3-6 and Figure 3-7).

The high-confidence variants (n=99,323) called by GATK were annotated using Annovar (2016Feb01)[58]. Variants were classified based on their functional consequences. Missense, nonsense, and splicing variants were selected among the identified variants and further filtered for conservation using conservation information in 46 vertebrate species[98]. To reduce the false-discovery rate, only variants not in segmental duplication regions were selected since reads mapped to these regions can match to other regions of the genome[99].

Of the remained 20,750 variants, 9263 variants were identified as rare variants, which were defined as MAF less than 1% in both 1KG and large population databases including ExAC121K[75] and NHLBI-ESP6500 (http://evs.gs.washington.edu/EVS).

The functional effect of rare variants was further assessed by SIFT prediction score[59]. In total, 3919 rare variants (42%) in 3088 genes were considered as deleterious variants according to their SIFT scores (SIFT < 0.05) (Figure 3-6A). As shown in Figure 3-6B, 94% deleterious variants were nonsynonymous variants. About 86% of deleterious variants were found only once (Allele count =1) in 40 samples (Figure 3-6C), suggesting not only that they were heterozygous variants, but also the difficulty to associate a single genetic variant with DD.

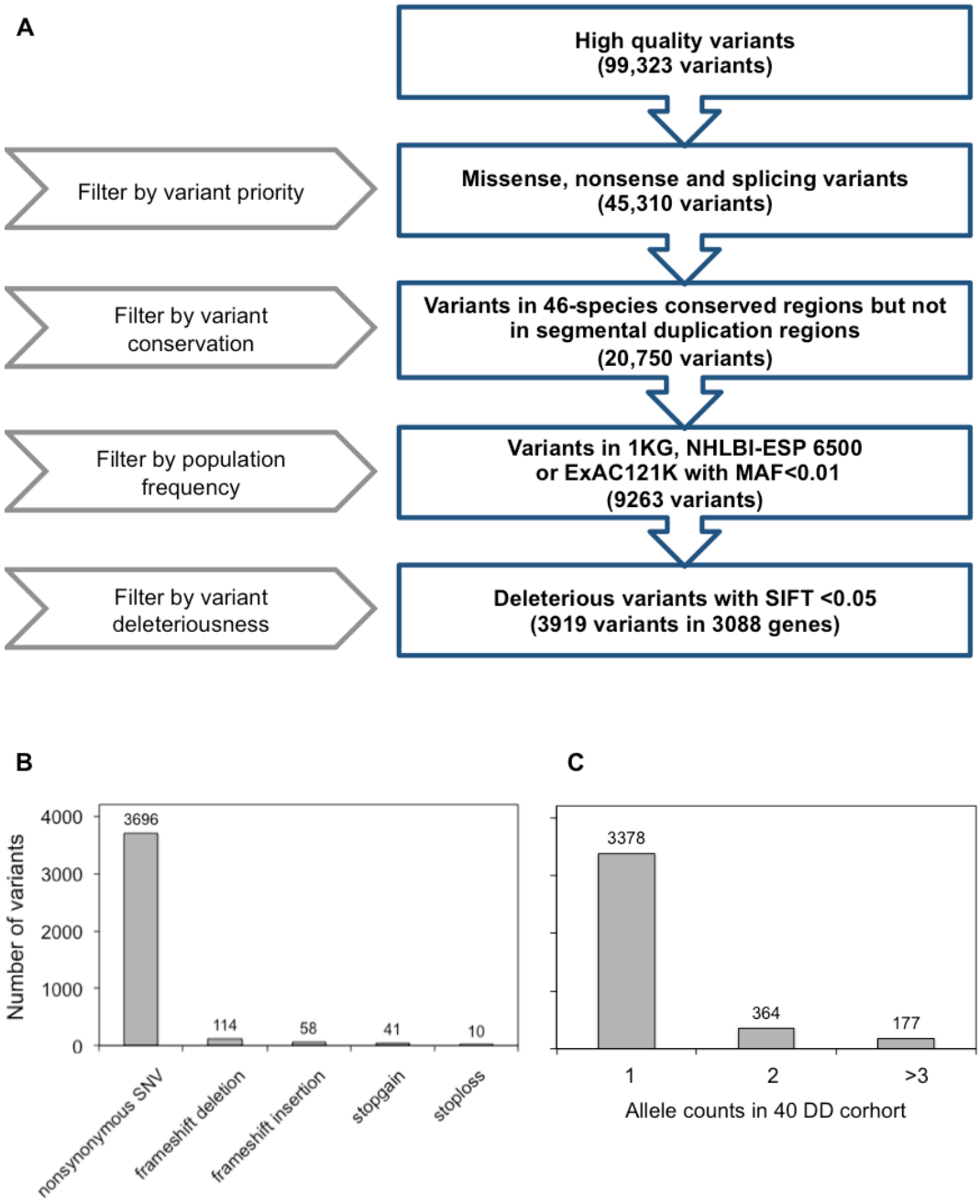


Figure 3-6 Annotation of functional variants in WES data of 40 DD patients

(A) Variant annotation steps using Annovar (2016Feb01).

(B) The distribution of 9263 rare variants based on functional consequences or (C) allele counts.

3.2.3 Identification of genes related to the DD phenotype

Though the genes that contribute to DD etiology are largely unknown, DD patients display a similar and well-defined phenotype with progressive fibrosis in the palmar connective tissue, which further leads to finger contracture.

Therefore, here I applied a phenotype-based strategy to identify candidate genes for DD using Phenolyzer[29], which involves computation tools for three key steps 1) associate the input human phenotypic ontology (HPO) terms[100] with known human diseases; 2) associate genes causing (or predicted as disease-causing) known diseases with the HPO terms; 3) integrate multiple features to score and prioritize all candidate genes.

Here, three HPO terms were used to represent the DD phenotype, which are 'flexion contracture of finger', 'connective tissue', and 'fibrosis'. Only 88 genes predicated as highly related to three DD phenotype terms (defined as a raw Phenolyzer score ≥ 1) were considered in the following analysis. Then the set of 1774 genes with gene burden greater than 2 (carrying at least 2 deleterious rare variants or variant allele accounts appeared more than 2 times) was compared with 88 genes related to DD phenotype. As a result, an overlapping set of 30 genes was considered as DD phenotype-related genes carrying rare genetic variants (Figure 3-7A).

3.2.4 Identification of genes intolerant to mutations in the palmar tissue

To assess whether there is an enrichment of genes predicted to be involved in DD pathogenesis, gene set enrichment analysis (GESA) on gene intolerance was applied to the above 30 genes using the gene intolerance score EvoTol[63], which identifies an intolerant gene as a gene containing an excess of mutations that, on the protein space, are not favored by evolution compared to other genes with the same number of mutations. To faithfully evaluate the tissue-specific impact, the intolerance ranking was limited to genes expressed in the palmar part of hand — the tissue affected by DD. Accordingly, an enrichment of pathogenicity in palmar tissue was determined (FDR q-value <0.001) by WGPA-GSEA tool[61,62], suggesting the overrepresented pathogenicity of the 30 genes (Figure 3-7B) in palmar part of hands. The class of pathogenic genes includes 12 genes, which were identified as the top 25% intolerant genes in the palm, suggesting a functional role of these genes in the hand palm (Figure 3-7C).

In conclusion, in the DD cohort for WES, 12 genes carrying rare deleterious variants were suggested as DD phenotype-related candidate genes (Table 3-5). Mutations in these genes may contribute to the pathogenicity in hand palmar part. Of note, this WES pilot study was an exploratory study. Functional effects of the candidate genes in DD pathogenesis should be tested systematically in the future.

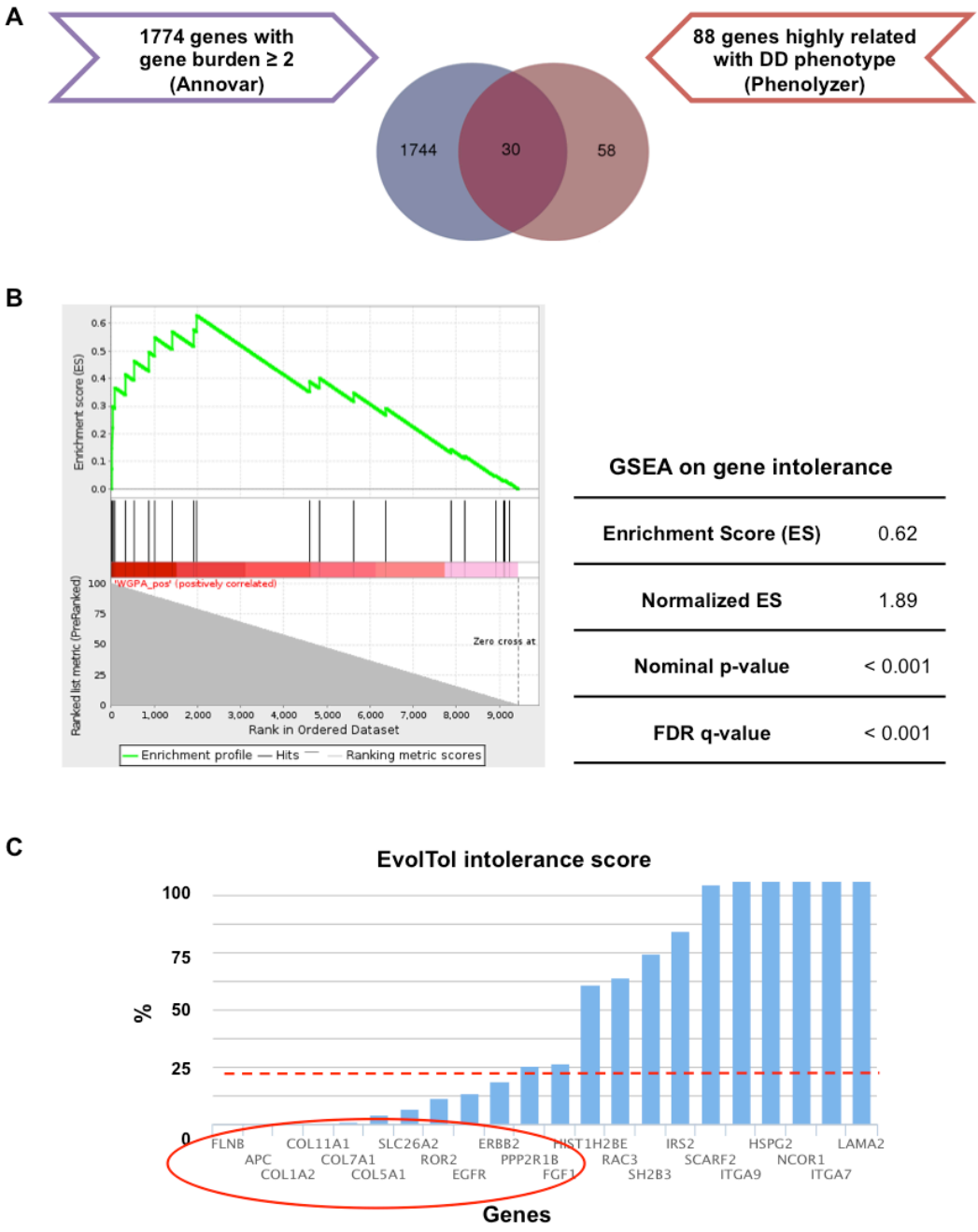


Figure 3-7 The prioritization of pathogenic candidate genes related to DD phenotypes

(A) 1774 genes with gene burden more than 2 were compared with 88 genes highly related to DD phenotype (Phenolyzer score ≥ 1). An overlapping set of 30 genes was identified.

(B) An enrichment of pathogenicity in palmar tissue for the above 30 genes was determined by WGPA-GSEA tool.

(C) Among 30 DD phenotype-related genes, 12 genes were predicted as the top 25% genes, which are intolerant to mutations in hand palmar part based on EvolTol score.

Table 3-5 12 candidate genes related to the DD phenotype

Chr	Exome data									
	Gene annotation				Rare variant annotation					
	Gene	EvoTol	Phenolyzer score * (0.0001 - 5.7)	Gene burden	dbSNP138	Start	End	Ref	Alt	Allele count
3	COL7A1	0.86%	1.3	4	.	48608296	48608296	G	A	1
					rs139434755	48629340	48629340	G	A	1
					rs35623035	48630252	48630252	G	A	2
1	COL11A1	0.32%	1.2	3	rs139064549	103354135	103354135	G	C	2
					rs141978499	103544374	103544374	C	G	1
9	COL5A1	3.48%	1.0	4	rs41306397	137591755	137591755	C	T	1
					rs199735010	137694750	137694750	C	T	1
					rs368305377	137702117	137702117	C	T	1
					rs61739195	137708884	137708884	C	T	1
7	COL1A2	0.27%	1.2	2	.	94055143	94055143	G	A	1
					.	94057101	94057101	A	C	1
9	ROR2	9.28%	1.2	3	rs35852786	94487187	94487187	C	T	3
5	SLC26A2	5.58%	3.6	2	rs114260147	149359938	149359938	C	G	1
					rs104893915	149359991	149359991	C	T	1
3	FLNB	0.04%	1.4	3	.	58141747	58141747	T	C	1
					rs116826041	58145348	58145348	T	C	1
					rs149638325	58148895	58148895	C	T	1
7	EGFR	10.76%	1.4	2	rs373336251	55240795	55240795	G	A	1
					rs201830126	55268023	55268023	G	A	1
11	PPP2R1B	20.26%	1.4	2	rs61756429	111608216	111608216	T	A	1
					rs115287852	111612783	111612783	T	C	1
5	APC	0.24%	1.2	2	rs1801166	112175240	112175240	G	C	1
					rs141010008	112178781	112178781	C	T	1
5	FGF1	21.06%	1.2	2	rs17223632	141993631	141993631	C	T	2
17	ERBB2	15.02%	1.4	2	.	37879585	37879585	A	C	1
					rs55943169	37884176	37884176	C	A	1

* Phenolyzer score: the higher, the more evidence for the association between a candiate gene and a phenotype

3.2.5 Pathway overrepresentation analysis of candidate genes

To interpret the biological signatures of the above 12 candidate genes, gene ontology (GO)[101] and pathway overrepresentation was analyzed by Enrichr tool[65].

The GO is a structured, controlled vocabulary of terms providing computational or experimental knowledge on gene product properties. For the set of 12 candidate genes, three main GO categories were analyzed including overrepresented biological process, cellular component and molecular function. [101]

GO analysis revealed that the top overrepresented term ranked by lowest p-value in the cellular components is 'fibrillar collagen trimmer'. The affected molecular functions were mostly involved in 'growth factor binding' and 'cell-extracellular matrix (ECM) interactions' (including 'ECM structural constituent', 'cell adhesion molecule binding', 'glycoprotein binding' and 'integrin binding'). The significantly affected biological processes were related to metabolic catabolic processes of macromolecules, including collagen, and the regulation of epithelial cell proliferation (Table 3-6).

Using the KEGG pathway resource[66], 18 significantly overrepresented pathways (Bonferroni adjusted p-value ≤ 0.05) were identified. The top 10 pathways include the 'Hippo signaling network' (including the child Hippo pathway, Wnt/ß-catenin pathway and TGFß pathway), the 'PI3K-Akt pathway', 'Focal Adhesion', 'Regulation of Actin Cytoskeleton', 'Protein Digestion and Absorption' and 'Pathways in Cancer' etc (Table 3-7).

Table 3-6 Overrepresented GO terms for 12 candidate genes

Index	GO Molecular Function 2015	Overlap*	Adjusted p-value	Z-score	Combined score
1	growth factor binding	4	2.0E-04	-2.4	20.2
2	platelet-derived growth factor binding	2	6.1E-04	-2.4	17.9
3	receptor signaling protein tyrosine kinase activity	2	6.1E-04	-2.3	17.2
4	transmembrane receptor protein kinase activity	3	6.1E-04	-2.2	16.4
5	extracellular matrix structural constituent	3	6.1E-04	-2.2	16.2
6	transmembrane receptor protein tyrosine kinase activity	3	6.1E-04	-2.2	16.1
7	protein tyrosine kinase activity	3	2.2E-03	-2.3	14.0
8	cell adhesion molecule binding	3	3.2E-03	-2.3	13.1
9	glycoprotein binding	2	1.7E-02	-2.2	9.0
10	integrin binding	2	2.2E-02	-2.2	8.4
Index	**GO Biological Process 2015**	**Overlap**	**Adjusted p-value**	**Z-score**	**Combined score**
1	multicellular organismal macromolecule metabolic process	4	1.2E-04	-2.2	19.8
2	collagen metabolic process	4	1.2E-04	-2.2	19.7
3	collagen catabolic process	4	1.2E-04	-2.2	19.7
4	multicellular organismal catabolic process	4	1.2E-04	-2.2	19.6
5	multicellular organismal metabolic process	4	1.2E-04	-2.2	19.6
6	collagen fibril organization	3	4.7E-04	-2.5	19.3
7	extracellular matrix disassembly	4	2.2E-04	-2.2	18.3
8	cellular component disassembly	5	4.7E-04	-2.3	17.9
9	regulation of epithelial cell proliferation	4	2.9E-03	-2.4	14.3
10	Fc-epsilon receptor signaling pathway	3	2.4E-02	-3.8	14.0
Index	**GO Cellular Component 2015**	**Overlap**	**Adjusted p-value**	**Z-score**	**Combined score**
1	fibrillar collagen trimer	4	5.7E-08	-2.5	42.0
2	collagen trimer	4	3.7E-05	-2.2	22.2
3	extracellular matrix part	4	8.9E-05	-2.1	19.7
4	endoplasmic reticulum lumen	4	1.4E-04	-2.2	19.2
5	extracellular matrix	4	2.6E-03	-2.2	13.0
6	basement membrane	2	2.3E-02	-2.2	8.2
7	extracellular region	5	5.1E-02	-2.5	7.3
8	adherens junction	3	4.5E-02	-2.3	7.0
9	anchoring junction	3	4.5E-02	-2.2	6.9
10	basolateral plasma membrane	2	4.8E-02	-2.0	6.1

* Overlap: Number of genes overlapped between 12 candidate genes and genes in a GO term

Table 3-7 Overrepresented pathways for 12 candidate genes

Index	KEGG pathway 2016	Overlap*	Adjusted p-value	Z-score	Combined score
1	Protein digestion and absorption	4	3.3E-04	-1.7	13.8
2	Endometrial cancer	3	1.4E-03	-2.0	12.8
3	Focal adhesion	4	2.4E-03	-1.9	11.3
4	PI3K-Akt signaling pathway	4	1.3E-02	-2.0	8.7
5	Pathways in cancer	5	1.3E-02	-2.0	8.5
6	Proteoglycans in cancer	3	1.6E-02	-1.8	7.5
7	Non-small cell lung cancer	2	1.6E-02	-1.8	7.3
8	Bladder cancer	2	1.3E-02	-1.7	7.2
9	Hippo signaling pathway	3	1.3E-02	-1.6	7.2
10	Regulation of actin cytoskeleton	3	1.6E-02	-1.7	7.0

* Overlap: Number of genes overlapped between 12 candidate genes and genes in a KEGG pathway

3.3. Global gene expression profiling in DD

3.3.1 Study subjects and sampling for RNA-seq

To obtain an in-depth understanding of the transcriptional regulation in DD, we performed a comprehensive transcriptomic analysis using DD biopsy samples from 11 DD patients in the WES cohort (described in session 3.2.1).

Two types of tissue samples were collected from a single DD patient including the affected palmar nodule tissue (DDtis) and the matched perinodular fat (DDfat). Additionally, the primary cells derived from DDtis were also collected and used as the *in vitro* cell model for DD (DDcell).

The control samples were obtained from 11 patients affected with carpal tunnel (CT) syndrome. The CT syndrome is a peripheral neuropathy due to compression of the median nerve as it travels through the wrist at the CT. All 11 CT patients involved in this study have no diagnosis of DD and no reported familial history of DD, therefore, are considered as healthy controls without DD. Palmar connective tissues (CTtis), mainly composed of fat, were collected during carpal release and used as external healthy tissue control for DDtis. The cells collected from CTtis (CTcell) were used as external healthy cell control for DDcell.

Additionally, in order to minimize the cofounding bias between case and control, 11 DD patients and 11 CT controls were carefully selected by matched age at sample collection (mean age: 60 and 61 separately) and country of origin (Germany) (Table 3-8). In total, we analyzed 50 transcriptomes using high quality RNA isolated from DD relevant samples (DDtis, DDfat and DDcell) and control samples (CTtis and CTcell).

Table 3-8 Study subjects and samples for RNA-seq

Subjects	**Genetic susceptibility to DD**			**Exome data**	**No. of samples collected for RNAseq**			**Matched cofounding factors**		
	Family histroy of DD	**Age at 1st DD operation**	**Hands affected with DD**		**Tissues**		**Primary cells**	**Mean age at sample collection**	**Country of origin**	**Gender**
11 DD Cases	**yes**	46	both	**yes**	9 DDfat	10 DDtis	10 DDcell	60	DE	9 male 2 female
11 CT Controls	not reported	negative		-	-	10 CTtis	11 CTcell	61	DE	11 male 0 female

3.3.2 Differentially expressed genes (DEGs) in DD related tissues

We first conducted pair-wise comparisons of differentially expressed genes (DEGs) between three tissue types, including DDfat compared to CTtis (abbreviated as DDfat/CTtis), DDtis vs. DDfat (DDtis/DDfat) and DDtis vs. CTtis (DDtis/CTtis).

To determine significantly DEGs, the following filtering approach was applied (Figure 3-8A). On a global scale, a) 769 significant DEGs were identified in DDfat/CTtis; b) 2777 significant DEGs in DDtis/DDfat; c) 3597 significant DEGs in DDtis/CTtis. These DEG sets were used for downstream GO features and pathway analyses. The distribution of gene expression fold change in each comparison is shown in a volcano plot (expression fold change vs. p-values) in Figure 3-8B.

The gene expression clustering of all DEGs is shown in a heatmap in Figure 3-9A. Overall, more genes were significantly upregulated in DDtis compared to two controls (DDfat and CTtis). Additionally, the gene expression profiling in DDtis classified the DD samples into two subgroups: one subgroup of DDtis samples from 4 patients and the other subgroup of DDtis from 6 patients.

A

Three filters for differentially expressed genes (DEGs)

	DDfat/CTtis	DDtis/DDfat	DDtis/CTtis
1. Gene expression FPKM	In each comparison min FPKM≥1 in either condition		
No. of Genes	12,170	12,333	12,497
2. FDR adjusted q-value	FDRqvalue ≤ 0.05, -log10 (FDRqval) ≥ 1.3		
3. Mean difference cut-off	|Fold change| ≥ 1.5, |log2 (Fold change)| ≥ 0.6		
No. of DEGs	**769**	**2777**	**3597**

B

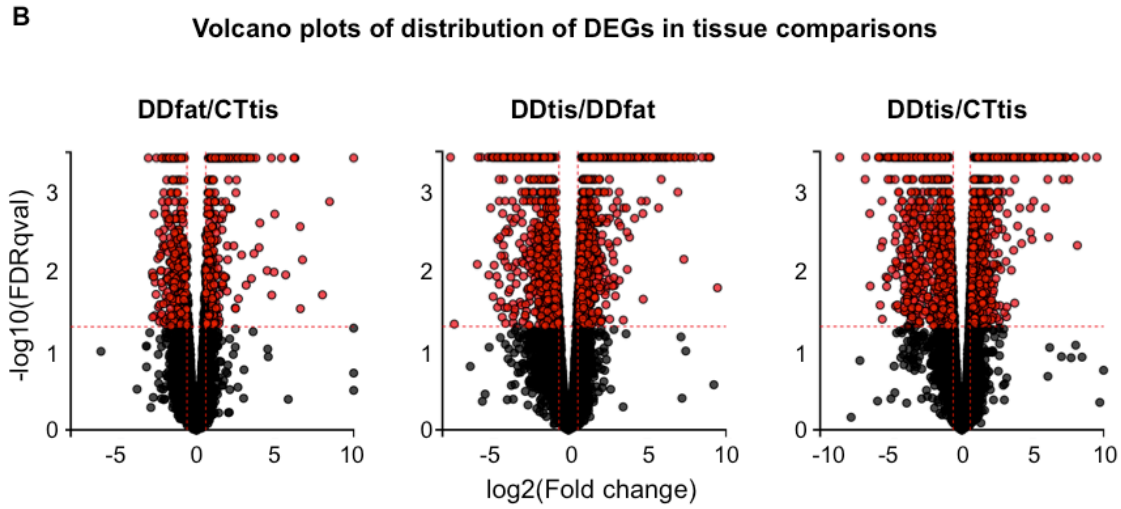


Figure 3-8 The identification of DEGs between tissue groups

(A) Filters applied on gene expression data in 3 comparisons: (a) DDfat vs CTtis; (b) DDtis vs. DDfat; (c) DDtis vs CTtis.

(B) Volcano plots of distribution of DEGs in each comparison, where log2 fold change (x-axis) is plotted against −log10 FDRadjusted q-value (y-axis). Significant DEGs are colored in red, while the rest are colored in black.

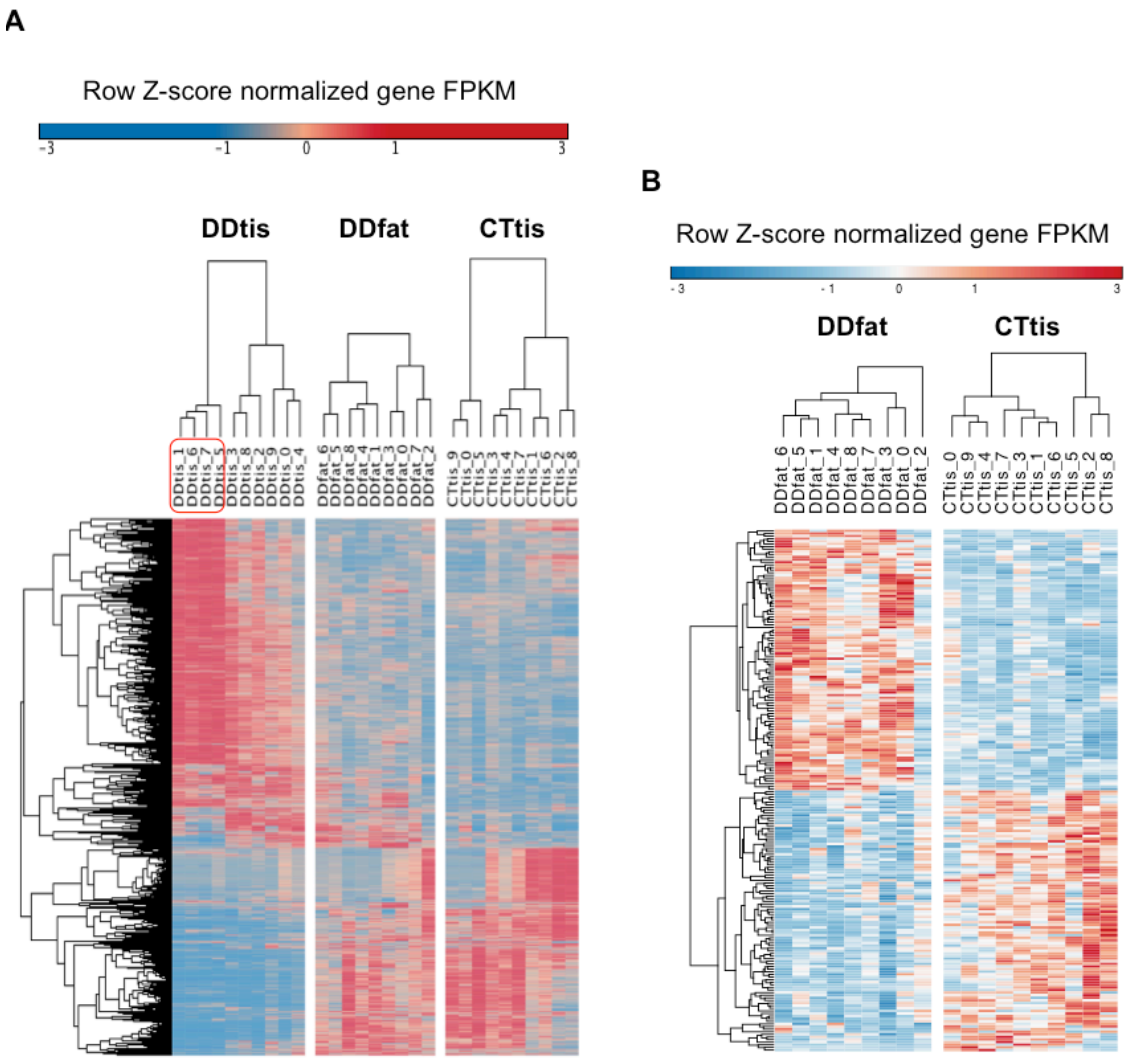


Figure 3-9 A heatmap representation of DEGs in tissue groups

(A) Gene expression profilings of DDtis, DDfat and CTtis. Each row represents a gene that is significant differentially expressed in any of the 3 comparisons (DDtis/DDfat, DDtis/CTtis and DDfat/CTtis). Each column represents an individual tissue sample.

The expression of each gene (FPKM value) was normalized by the Z-score to gain insight into the gene expression patterns in three tissue groups (DDtis, DDfat and CTtis). This scaled row Z-score value is plotted in red–blue color scale with expression higher than average represented in red ($Z > +1$), expression lower than average in blue ($Z < -1$), and average in orange ($Z = 0$).

(B) A zoom in to the gene expression profiling of DDfat and CTtis. The gene expression (FPKM value) higher than average is represented in red ($Z > +1$), expression lower than average in blue ($Z < -1$), and average in white ($Z = 0$).

3.3.3 Gene Ontology enrichment of DEGs in DD related tissues

To interpret the function of the DEGs, we first queried enriched GO terms[101] using iPathwayGuide in three tissue comparisons: a) DDfat/CTtis; b) DDtis/DDfat and c) DDtis/CTtis. The common enriched GO terms among three comparisons are shown in Venn diagrams (Figure 3-10).

For example, in GO category biological processes, there were 265 GO terms only shared between DDtis/DDfat and DDtis/CTtis, which are likely related to distinctive biological processes in DDtis (Figure 3-10A). The GO biological processes shared in both DDtis/CTtis and DDtis/DDfat were ranked by GO p-values in DDtis/CTtis (from low to high). The top-10 GO terms represented in a heatmap, which reveals significant involvement of ECM in DD, e.g. 'ECM organization', 'ECM disassembly', 'collagen catabolic processes', and 'collagen fibril organization' (Figure 3-10A and Table 3-9).

The same analysis was performed on GO molecular function and GO cellular components (Figure 3-10B and C, Table 3-9). A remarkable feature in the GO molecular function analysis was that the 9 top GO terms in both DDtis/CTtis and DDtis/DDfat (ranked by GO p-values in DDtis/CTtis from low to high) were involved in either extracellular structural constituent or molecule binding (including 'collagen binding', 'integrin binding', 'actin binding', 'actin filament binding', 'calcium ion binding', 'heparin binding', 'growth factor binding', and 'phosphatidylserine binding'). With respect to GO cellular component, the most significant GO term in both DDtis/CTtis and DDtis/DDfat was 'extracellular space' supporting the pivotal involvement of ECM in DD pathogenesis.

Taken together, this GO term analysis in all three comparisons (DDfat/CTtis, DDtis/DDfat and DDtis/CTtis) reveals significant functional alterations of genes involved in ECM components, dynamics and function in DDtis, which is in line with the important role of ECM in connective tissue.

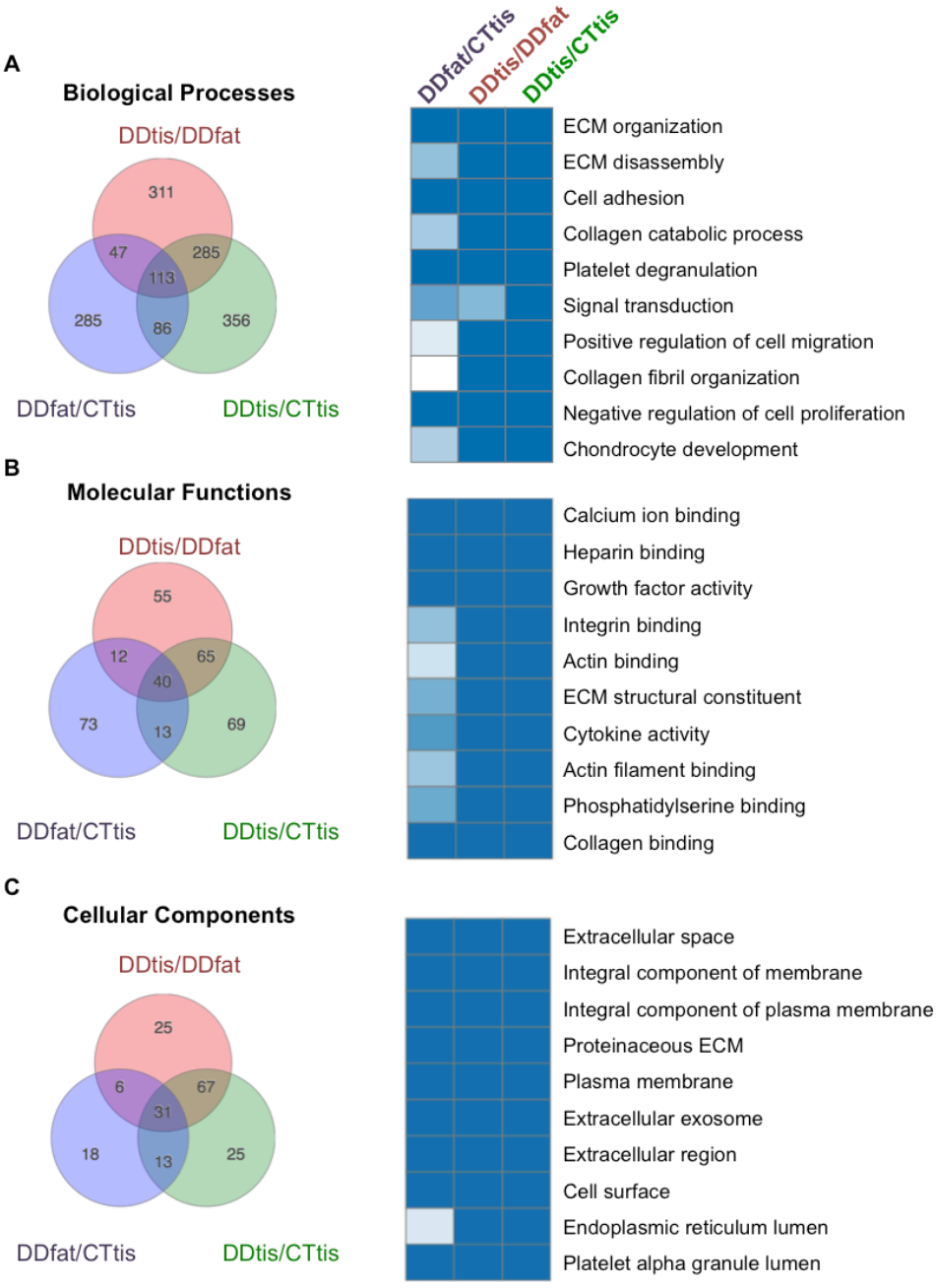


Figure 3-10 Overrepresented GO terms for DEGs in tissue comparisons

The Venn diagrams highlight the common enriched GO terms among three comparisons. The most significant 10 GO terms only shared between DDtis/DDfat and DDtis/CTtis were represented in heatmaps (ranked by the –log10pvalue in DDtis/CTtis, the most significant on top). Each row represents a GO term. Each column represents a comparison (DDfat/CTtis, DDtis/DDfat and DDtis/CTtis). The intensity of the shading indicates the significance of the GO term. Each blue box has a p-value of ≤0.05; white box has a p-value >0.05 and is not significantly altered. See details in Table 3-9.

Table 3-9 Overrepresented GO terms for DEGs in tissue comparisons

	GO term Biological Processes	DDtis/DDfat		DDtis/CTtis		DDfat/CTtis	
		#genes (DE/All)	p-value	#genes (DE/All)	p-value	#genes (DE/All)	p-value
1	ECM organization	162/298	4.20E-14	186/301	1.10E-15	47/284	1.20E-05
2	ECM disassembly	55/96	1.80E-11	64/97	1.40E-11	13/87	3.20E-02
3	Cell adhesion	348/927	2.60E-11	454/937	8.60E-11	135/908	1.00E-03
4	Collagen catabolic process	38/56	5.10E-12	40/57	6.80E-10	7/49	5.00E-02
5	Platelet degranulation	33/67	6.70E-06	43/66	5.80E-09	15/66	2.90E-04
6	Signal transduction	1028/3482	2.30E-02	1331/3501	4.00E-08	327/3407	1.00E-02
7	Positive regulation of cell migration	114/298	8.50E-05	155/303	1.10E-07	32/291	.
8	Collagen fibril organization	24/34	1.30E-08	25/34	2.70E-07	2/30	
9	Negative regulation of cell proliferation	161/472	8.00E-05	209/476	3.60E-07	68/459	1.90E-05
10	Chondrocyte development	16/22	2.00E-04	18/22	9.20E-07	4/22	.

Continued on the next page

	GO Molecular Function	DDtis/DDfat		DDtis/CTtis		DDfat/CTtis	
		#genes (DE/All)	p-value	#genes (DE/All)	p-value	#genes (DE/All)	p-value
1	Calcium ion binding	140/371	4.20E-10	177/377	2.40E-12	62/372	6.50E-12
2	Heparin binding	55/105	1.50E-10	67/107	3.30E-12	29/98	2.60E-12
3	Growth factor activity	40/81	3.80E-07	52/84	1.60E-09	14/74	2.90E-04
4	Integrin binding	42/81	3.20E-08	50/81	3.80E-09	10/79	3.40E-02
5	Actin binding	119/280	1.70E-07	146/284	3.70E-08	27/277	.
6	ECM structural constituent	30/45	6.30E-08	31/46	2.00E-07	7/41	1.70E-02
7	Cytokine activity	37/88	9.60E-05	50/89	2.80E-07	17/78	6.00E-03
8	Actin filament binding	38/82	5.20E-06	47/83	4.60E-07	10/82	4.30E-02
9	Phosphatidylserine binding	14/22	7.30E-05	17/20	5.80E-07	5/22	1.30E-02
10	Collagen binding	34/57	5.30E-08	36/59	8.70E-07	13/57	6.50E-04

	GO Cellular Components	DDtis/DDfat		DDtis/CTtis		DDfat/CTtis	
		#genes (DE/All)	p-value	#genes (DE/All)	p-value	#genes (DE/All)	p-value
1	Extracellular space	295/698	1.00E-24	378/709	1.00E-24	136/678	4.10E-23
2	Integral component of membrane	881/2696	1.80E-16	1132/2737	2.30E-22	265/2634	5.50E-04
3	Integral component of plasma membrane	284/763	3.40E-13	370/772	1.80E-20	110/739	2.70E-13
4	Proteinaceous ECM	155/260	1.00E-24	165/262	1.80E-20	52/237	1.50E-10
5	Plasma membrane	874/2580	1.20E-12	1116/2613	9.50E-18	304/2529	7.90E-10
6	Extracellular exosome	571/1900	6.20E-11	730/1919	6.20E-15	187/1903	3.40E-08
7	Extracellular region	908/2669	4.30E-10	1144/2704	5.70E-14	316/2623	2.30E-06
8	Cell surface	162/445	5.20E-07	227/451	3.80E-13	77/439	4.40E-08
9	Endoplasmic reticulum lumen	76/137	4.70E-14	81/136	1.90E-12	12/130	.
10	Platelet alpha granule lumen	20/35	3.00E-05	27/35	1.70E-08	10/34	6.20E-05

3.3.4 Pathway perturbation analysis using DEGs in DD related tissues

To study the specific pathways associated with DD, a third generation pathway topology (PT)-based pathway perturbation analysis[68] via a combination of over-representation analysis and pathway accumulation was performed using iPathwayGuide on aforementioned DEG sets (Figure 3-11). For over-representation analysis, the number of DEGs involved in a pathway was compared between two tissue groups. For pathway accumulation analysis, the significance of a particular DEG to a pathway was considered in determining the overall impact on the pathway by examining all annotated functions/interactions of the gene in KEGG pathway databases[66]. The calculated pathway perturbation is an additive measurement of pathway overrepresentation and accumulation. Bonferroni adjusted p-value 0.05 on pathway perturbation was used as a cut-off to identify significantly perturbed pathways.

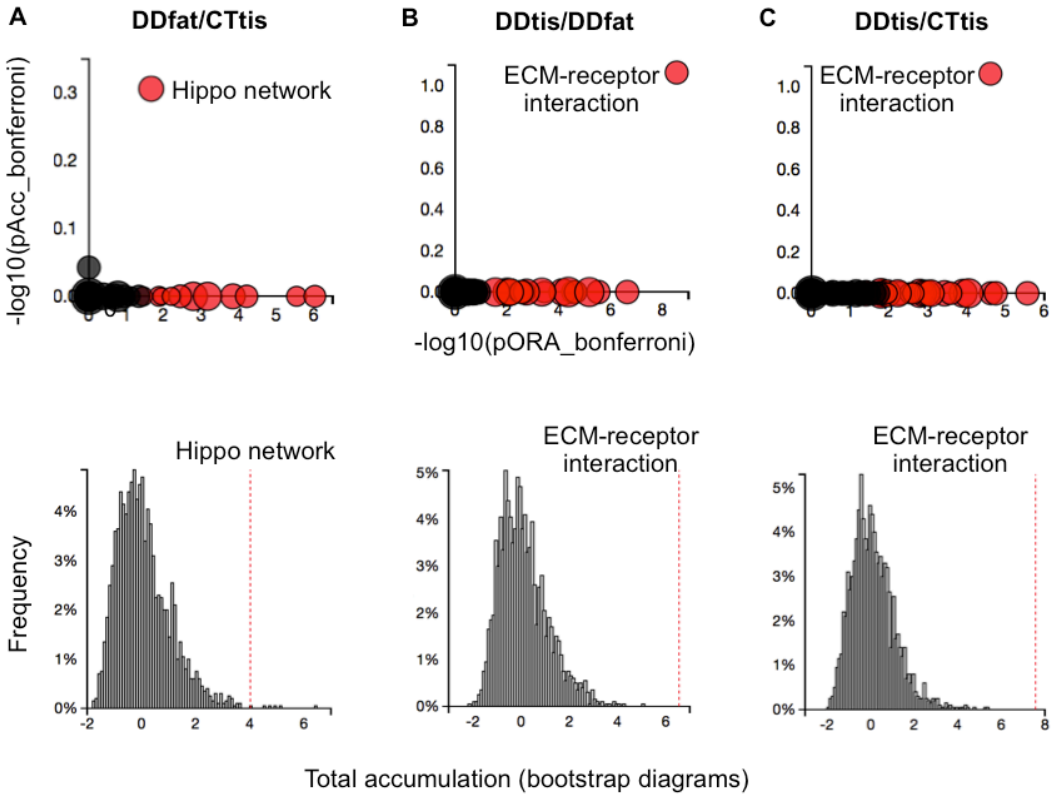


Figure 3-11 Perturbed pathways in tissue comparisons

The scatter plots represent the pathways (in red) significantly perturbed in three comparisons: (A) DDfat/CTtis; (B) DDtis/DDfat; (C) DDtis/CTtis. The pathway perturbation is an additive measurement of pathway overrepresentation and accumulation. The x-axis represents the significance of pathway overrepresentation (–log10OVA p-value). The y-axis represents the significance of pathway accumulation (–log10Acc p-value). Bonferroni adjusted p-value 0.05 on pathway overrepresentation and accumulation was used as a cut-off for identifying significant perturbed pathways. (Continued on the next page)

The bootstrap diagram shows the distribution (mean=0, stdev=1) of expected accumulated perturbation values based on the observed data for the selected pathway. The red line indicates where the actual observed values lie in the distribution of expected results. The further away from the mean, the less likely the observed values are due to random chance.

3.3.4.1 Perturbation of the Hippo signaling pathway in DDfat/CTtis

In the comparison DDfat/CTtis, the most significantly perturbed pathway (with the lowest p-value for pathway overrepresentation and accumulation) was the Hippo network (Figure 3-11A, x- and y-axis represent the significance of overrepresentation and accumulation separately). A bootstrap diagram was also shown separately, which displays the distribution (mean=0, stdev=1) of expected accumulation values based on the observed data. 2,000 iterations were used to construct the distribution. The red line indicates where the actual observed values lie in the distribution of expected results. The further away from the mean, the more likely the observed value is not due to random chance.

Of note, the KEGG Hippo network is designated as a 'parent' network composed of the 'child' Hippo signaling pathway (also named as the YAP/TAZ pathway since YAP/TAZ are the major effectors of the Hippo signaling pathway) and two interactive pathways — the Wnt/ß-catenin signaling pathway and TGFß pathway (shown in Figure 3-12B).

In DDfat/CTtis, four genes (*LLGL2*, *RASSF6*, *WWC1* and *INADL*) involved in the 'child' Hippo pathway were downregulated (Figure 3-12A), suggesting an inactivation of the 'child' Hippo pathway. Inhibition of downstream transcriptional activators, YAP/TAZ, is the major target of the 'child' Hippo signaling pathway. Therefore, the downregulation of these four genes may activate YAP/TAZ, as suggested by iPahtwayGuide perturbation computation, which represents the propagation of the effect of upstream gene expression change by considering interactions between different genes (Figure 3-12B, genes in red indicates activation or increased expression, while genes in blue indicates inhibition or decreased expression.

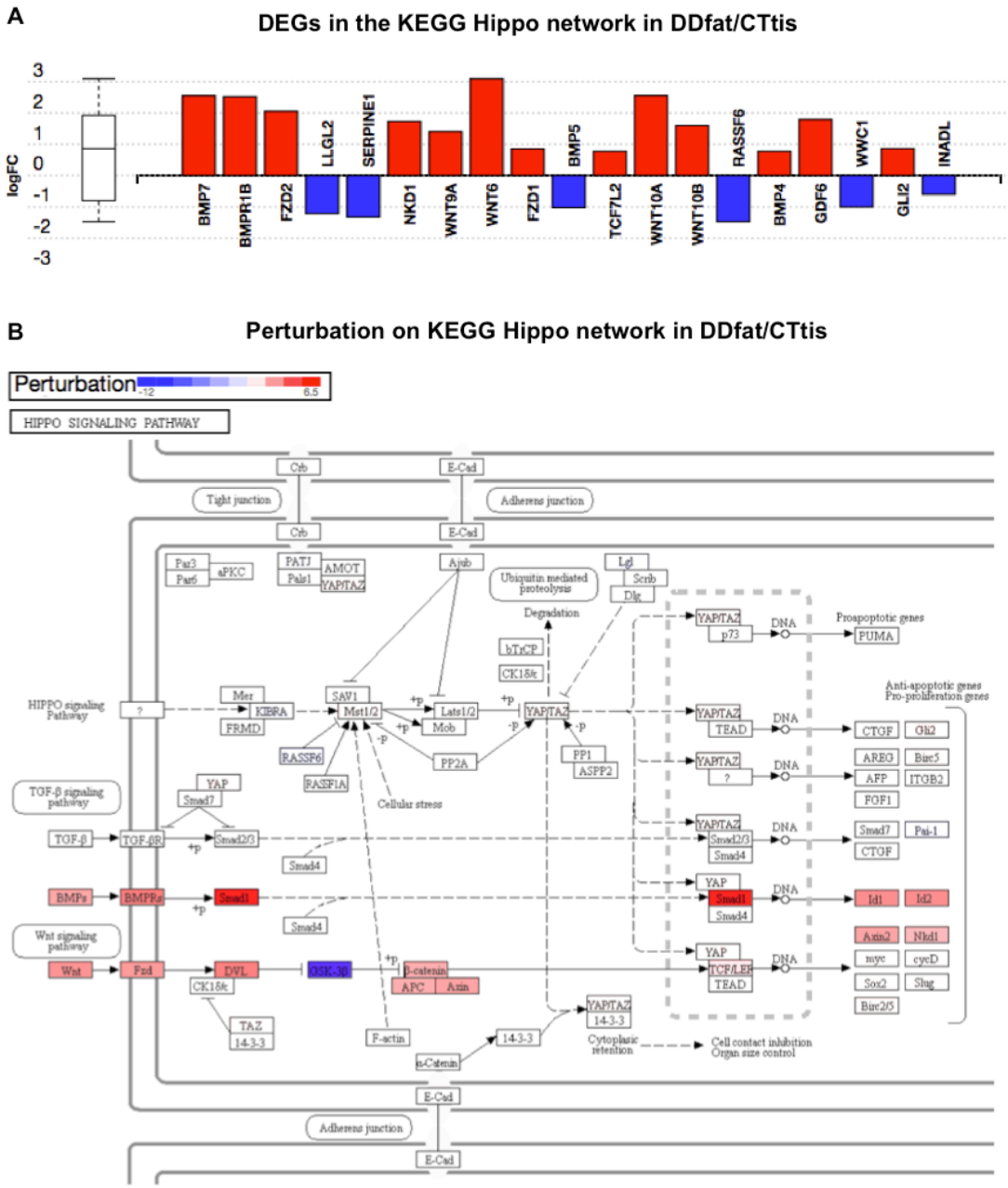


Figure 3-12 The Hippo network is the most significantly perturbed pathway in DDfat/CTtis

(A) The expression fold change of DEGs in the KEGG Hippo network in DDfat/CTtis.

(B) The perturbation on genes involved in the KEGG Hippo network due to DEGs in DDfat/CTtis. Genes in red indicates activation or increased expression, while genes in blue indicates inhibition or decreased expression. The darker color indicates the stronger perturbation.

Moreover, the expression four Wnt ligand genes (*WNT6*, *WNT10A*, *WNT10B* and *WNT9A*), two Frizzled receptor genes (*FZD1* and *FZD2*) involved in the upstream of the Wnt signaling pathway was also increased in DDfat compared to CTtis (Figure 3-12A), leading to perturbation on the downstream effector ß-catenin and transcription factors TCF/LEF (Figure 3-12B).

In addition, the expression of three genes in the Bone Morphogenetic Protein (BMP) family and a BMP receptor (*BMP7*, *BMP4*, *GDF6* and *BMP1B*) was increased, while, the expression *BMP5* was decreased in DDfat compared to CTtis. The increased expression of BMPs, which are upstream regulators to the TGFβ, is likely to induce the TGFβ pathway.

In summary, though the DDfat and CTtis were considered as non-disease adipose tissue for DD in the past, significant DEGs in DDfat/CTtis were identified and suggested to be involved in three interactive pathways (including the YAP/TAZ signaling pathway, the Wnt/ß-catenin signaling pathway and the TGFß pathways). By pathway perturbation analysis on all DEGs in the DDfat/CTtis comparison, the KEGG parent network — the Hippo network was identified as the most perturbed pathway with the lowest p-value (Bonferroni adjusted p-value = 0.001). These findings reveal, for the first time, perturbation on pathways in perinodule fat tissue (DDfat) and the potential role of the Hippo network in DD pathogenesis.

3.3.4.2 Perturbation of the ECM-receptor interactions in DDtis

The pathway perturbation analysis was also conducted using DEGs in DDtis/DDfat and DDtis/CTtis. The ECM-receptor interaction was identified as the most significant perturbed pathway in both comparisons (Figure 3-11B and C).

About 80% of DEGs involved in the ECM-receptor interactions pathway showed increased expression in DDtis compared to either internal control (DDfat) or external control (CTtis) (Figure 3-13A and 3.14A). The common upregulated DEGs in the ECM-receptor interactions included genes encoding the ECM structural and functional macromolecules (COL1A1, COL1A2, COL6A1, COL6A2, COL6A3, FN1 and THBS2 etc., which maintain cell/tissue structure and function), transmembrane integrins (ITGB1, ITGA8, ITGA10 etc, which function as mechanoreceptors and provide force-transmitting physical links between the ECM and cells) as well as the cytoskeleton proteoglycans (such as SDC2 and SV2A) and other cell-surface-associated components (for instance, CD44). The increased expression of these genes can strongly influence interactions between cells and ECM, directly/indirectly leading to cellular activities such as adhesion, migration, differentiation, proliferation, and apoptosis[66].

The most significant downregulated DEG in both comparisons (DDtis/DDfat and DDtis/CTtis) is *CD36* (log2FC value -3.4 and -3.6 separately). Recent studies have revealed a causative role of *CD36* downregulation in suppressing the ECM deposition and in promoting the protumorigenic phenotypes of tumor stromal microenvironment[102]. Therefore, the reduced expression of *CD36* in DDtis (vs. DDfat or CTtis) opens up the question whether there is also an influence of the DDtis on its microenvironment — DDfat, the adipose tissue adjacent to DDtis.

Overall, these results suggest the ECM-receptors interactions pathway is strongly related to the biochemical and functional characteristics of DD nodules.

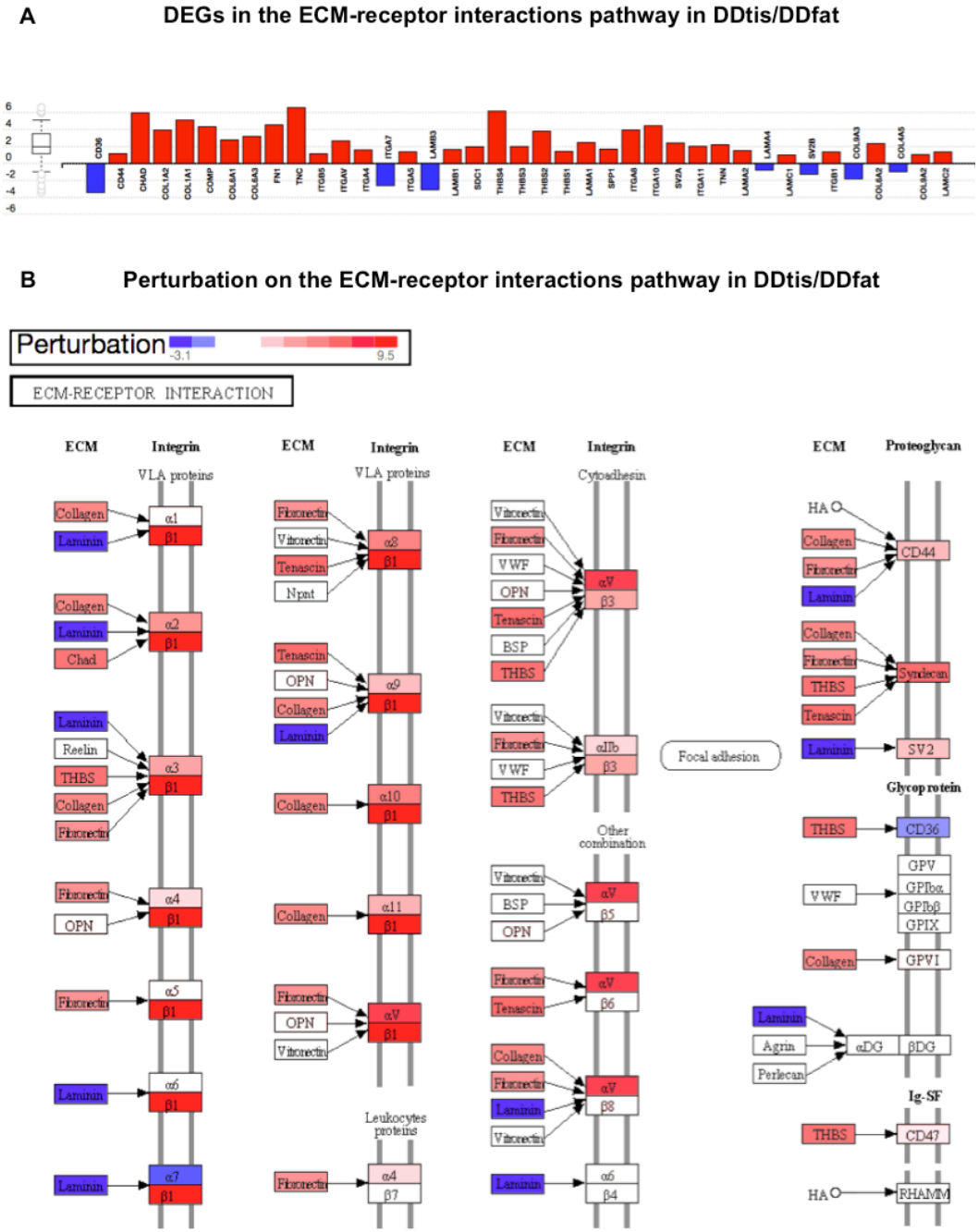


Figure 3-13 The ECM-receptor interactions pathway is the most significantly perturbed pathway in DDtis/DDfat

(A) The gene expression fold change of DEGs in the KEGG ECM-receptor interactions pathway

(B) The perturbation calculation on genes involved in the KEGG ECM-receptor interactions pathway due to DEGs in DDtis/DDfat. Genes in red indicates activation or increased expression, while genes in blue indicates inhibition or decreased expression. The darker color indicates the stronger perturbation.

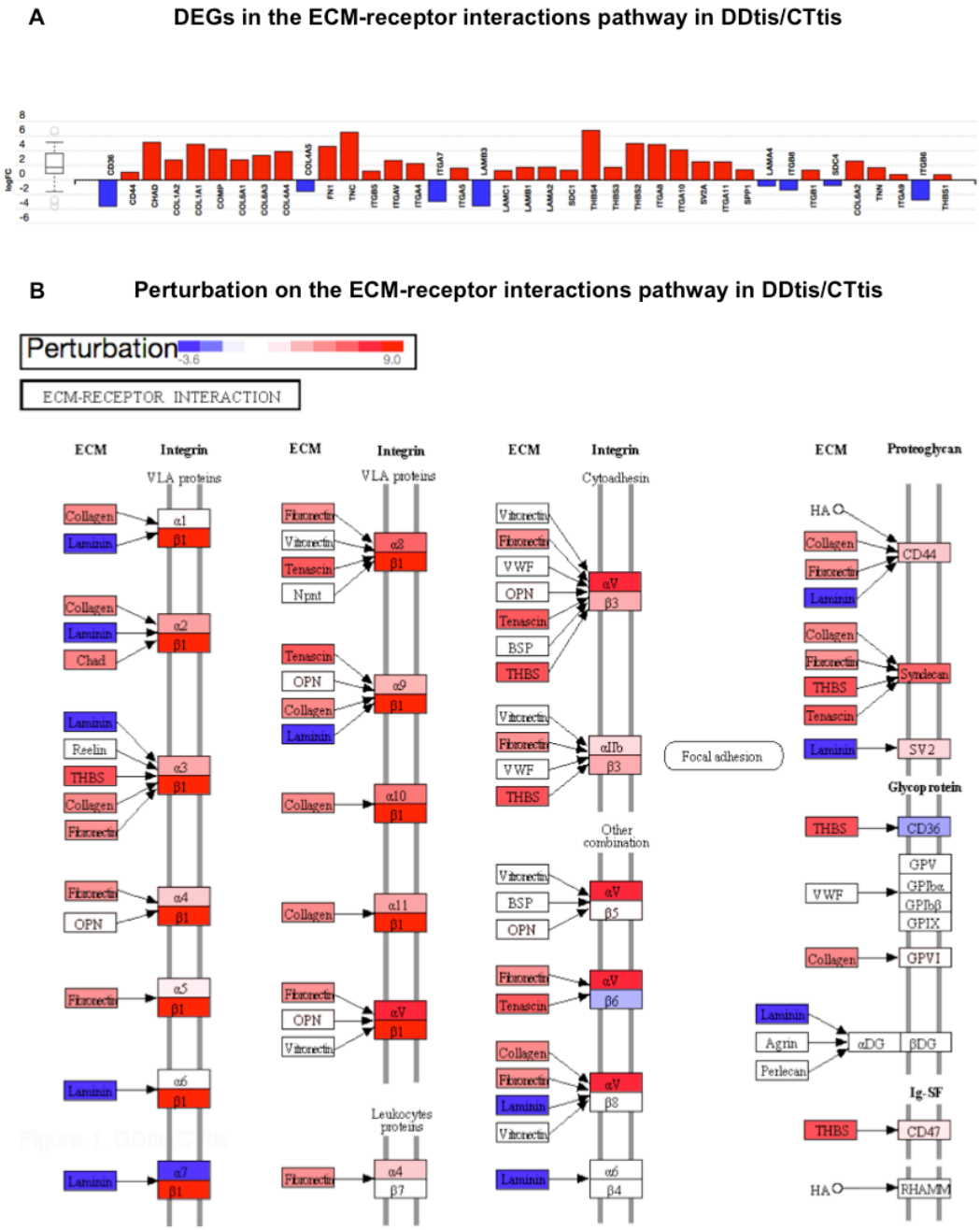


Figure 3-14 The ECM-receptor interactions pathway is the most significantly perturbed pathway in DDtis/CTtis

(A) The gene expression fold change of DEGs in the KEGG ECM-receptor interactions pathway

(B) The perturbation calculation on genes involved in the KEGG ECM-receptor interactions pathway due DEGs in DDtis/CTtis. Genes in red indicates activation or increased expression, while genes in blue indicates inhibition or decreased expression. The darker color indicates the stronger perturbation.

3.3.4.3 Summary of all overrepresented KEGG pathways in three tissue comparisons

As shown in Figure 3-11, besides the perturbed pathways (both significantly overrepresented and accumulated), a few pathways are only strongly overrepresented (represented by the red dots, with high -log10pvalue on x-axis). In Table 3-10 and Figure 3-15, all significant overrepresented pathways (without regard to accumulated or not) in three comparisons (DDfat/CTtis, DDtis/DDfat and DDtis/CTtis) were listed.

Two common overrepresented pathways in DDfat/CTtis and DDtis/CTtis are the Hippo network and the Cell adhesion molecules (CAM). Seven pathways were significantly overrepresented in both DDtis/DDfat and DDtis/CTtis including the 'Focal Adhesion', 'Pathways in Cancer', 'Protein Digestion and Absorption' and three myopathies (Hypertrophic Cardiomyopathy, Dilated Cardiomyopathy and Arrhythmogenic Right Ventricular Cardiomyopathy).

Table 3-10 Overrepresented pathways for DEGs in tissue comparisons

Overrepresented pathways (KEGG)			
	Combined Bonferroni p-value		
Pathway name	**DDfat/CTtis**	**DDtis/DDfat**	**DDtis/CTtis**
Cell adhesion molecules (CAMs)	0.000		0.000
Salivary secretion	0.000		0.003
Staphylococcus aureus infection	0.001		0.015
Hippo signaling pathway	0.001		0.022
Complement and coagulation cascades	0.001		
Basal cell carcinoma	0.001		
Neuroactive ligand-receptor interaction	0.005	0.004	0.001
Cytokine-cytokine receptor interaction	0.008	0.003	0.000
Asthma	0.012		
Pancreatic secretion	0.043		
Intestinal immune network for IgA production	0.044		
Aldosterone-regulated sodium reabsorption	0.047		
Protein digestion and absorption		0.000	0.000
cAMP signaling pathway			0.012
ECM-receptor interaction		0.000	0.000
Hypertrophic cardiomyopathy (HCM)		0.000	0.000
Dilated cardiomyopathy		0.000	0.000
PI3K-Akt signaling pathway		0.004	0.000
Focal adhesion		0.000	0.001
Pathways in cancer		0.002	0.002
Arrhythmogenic right ventricular cardiomyopathy		0.001	0.003
Axon guidance			0.004
cGMP-PKG signaling pathway		0.015	0.007
Regulation of lipolysis in adipocytes		0.003	0.029
Starch and sucrose metabolism			0.047
PPAR signaling pathway		0.012	
Morphine addiction		0.010	
Rap1 signaling pathway		0.008	
GABAergic synapse		0.020	

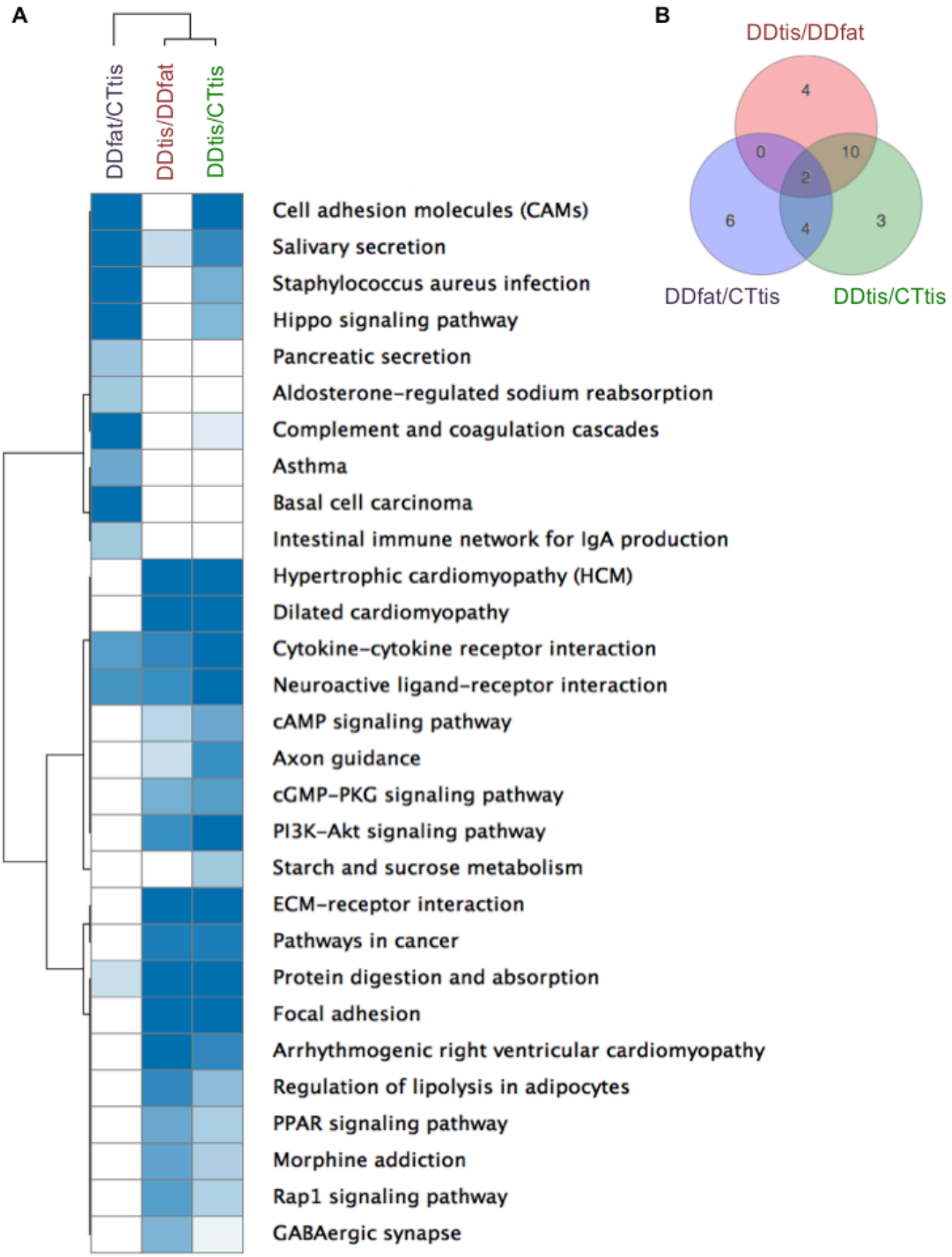


Figure 3-15 Summary of overrepresented KEGG pathways in three tissue comparisons

(A) A clustering of overrepresented pathways in 3 comparisons (DDfat/CTtis, DDtis/DDfat and DDtis/CTtis). The intensity of the shading indicates the significance of the pathway overrepresentation. All blue boxes have a Bonferroni adjusted p-value of ≤0.05, white boxes have a p-value of >0.05. The –log10(Bonferroni p-value) was used in the heatmap.

(B) The Venn diagram depicts the number of overrepresented pathways in three comparisons.

3.3.5 Pathway overrepresentation analysis using DEGs in DD cells

3.3.5.1 DEGs in DD cells

DDcells cultured from DDtis are often used as *in vitro* cell models to study cellular mechanisms involved in DD. To gain an in depth understanding on the gene expression profiling and pathway perturbation of DDcells, analysis of DEGs was first conducted in three comparisons DDtis/CTtis, DDcell/CTcell and DDtis/DDcell (Figure 3-16A). These filtered DEGs were used for pathway analysis. The distribution of gene expression fold change in each comparison is shown in volcano plots in Figure 3-16B.

3.3.5.2 Enrichment of perturbed KEGG pathways in DD cells

Pathway perturbation analysis combining pathway overrepresentation and pathway accumulation was performed using DEGs in DDtis/CTtis, DDcell/CTcell and DDtis/DDcell. In DDtis/DDcell, the most significantly perturbed pathway is the 'Cytokine-cytokine receptor interactions pathway' (see Figure 3-17A), which is likely to be the main molecular difference between DD tissue and cell models. In DDcell/CTcell, no candidate pathway was predicted to obtain strong accumulation due to the DEGs (see Figure 3-17A).

The significantly overrepresented pathways among three comparisons (DDtis/CTtis, DDcell/CTcell and DDtis/DDcell) were also compared. The Venn diagram in Figure 3-17B illustrates the number of common and individual pathways among three comparisons. By comparing the similarities between DDcell/CTcell and the DDtis/CTtis, common pathways contributing to DD pathogenesis in both tissue and cell models might be identified if the tissue-cell difference is removed (represented by DDtis/DDcell). Therefore, as highlighted in the Venn diagram, 5 overrepresented pathways were uniquely shared between DDtis/CTtis and DDcell/CTcell, which include the 'Focal Adhesion', three myopathies, and 'Straphylococcus aureus Infection' which is related to inflammation[66].

A

Three filters for DEGs

	DDtis/CTtis	DDcell/CTcell	DDtis/DDcell
1. Gene expression FPKM	In each comparison min FPKM≥1 in either condition		
No. of Genes	12,497	11,118	12,160
2. FDR adjusted q-value	FDRqvalue ≤ 0.05, -log10 (FDRqval) ≥ 1.3		
3. Mean difference cut-off	|Fold change| ≥ 1.5, |log2 (Fold change)| ≥ 0.6		
No. of DEGs	3597	770	3879

B

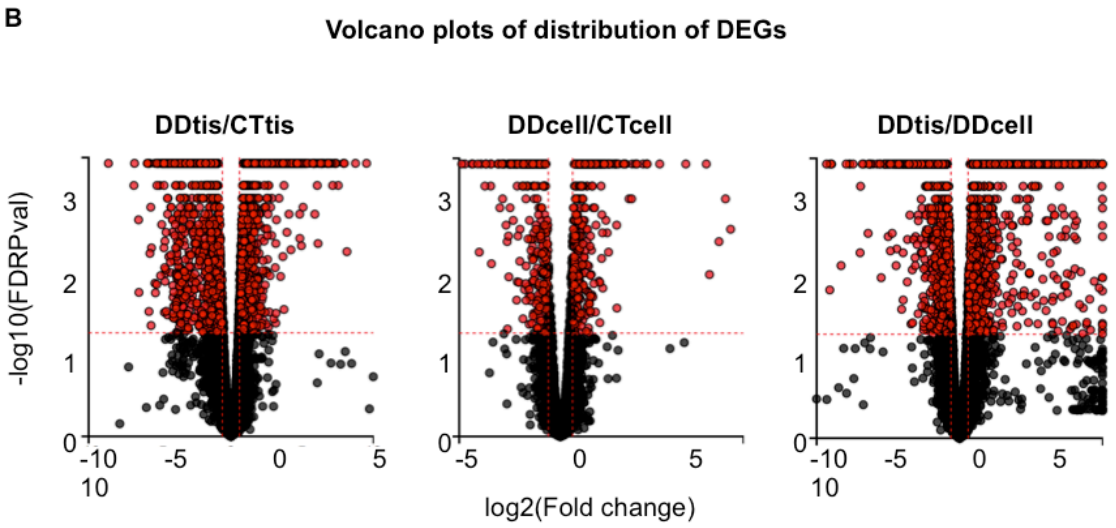
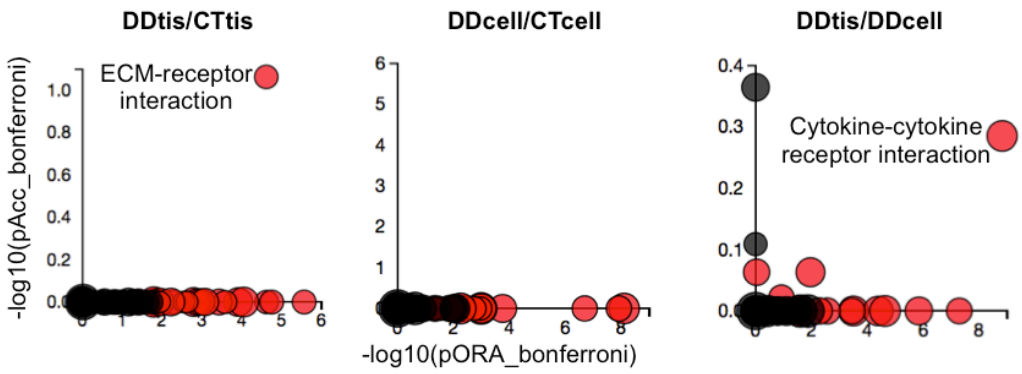


Figure 3-16 Volcano plots for DEGs in *in vitro* cell models

(A) Filters applied on gene expression change under 3 conditions: (a) DDtis vs CTtis; (b) DDcell vs. CTcell; (c) DDtis vs DDcell.

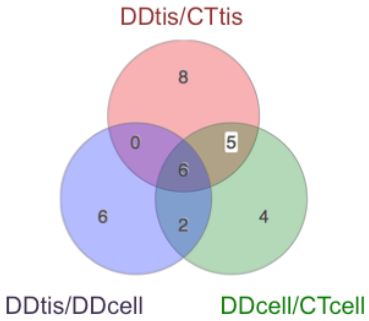
(B) Volcano plots of distribution of DEGs in each comparison, where log2 fold change (x-axis) is plotted against −log10 FDRadjusted q-value (y-axis). Significant DEGs are colored in red, while the rest are colored in black.

A



B

Overrepresented pathways between DDtis/CTtis and DDcell/CTcell



Pathway	Bonferroni adjusted p-value	
	DDtis/CTtis	DDcell/CTcell
Hypertrophic cardiomyopathy (HCM)	2.4E-05	0.004
Dilated cardiomyopathy	3.5E-04	0.010
Focal adhesion	0.001	0.007
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	0.003	0.003
Staphylococcus aureus infection	0.016	0.013

Figure 3-17 Perturbed pathways in *in vitro* cell models

(A) The most significantly perturbed pathways in DDtis vs CTtis, DDcell vs. CTcell and DDtis vs DDcell. Bonferroni adjusted p-value ≤ 0.05 was applied to select significantly perturbed pathways.

(B) The left Venn diagram depicts the overrepresented pathways uniquely shared in DDtis/CTtis and DDcell/CTcell. The 5 pathways and their p-values are listed in the right table.

3.4 Characterization of alternative splicing (AS) in DD

3.4.1 Comparison of the AS classes in disease tissues and controls

To explore the extent of AS between disease and control tissues, we first classified the differential AS events for each gene using spliceR[46]. SpliceR first constructs the hypothetical pre-RNA based on the exon information from all transcripts originating from that gene[46]. Subsequently, all transcripts are compared to this hypothetical pre-RNA in a pairwise manner, and AS events are classified and annotated[46].

Here, six common types of gene splicing events were considered including the exon skipping/inclusion (ESI), intron skipping/inclusion (ISI), alterative 5' splice sites (A5), alternative 3' splice sites (A3), multiple exon skipping (MESI) and mutually exclusive exons (MEE)[46]. In all 3 comparisons (DDfat/CTtis, DDtis/DDfat and DDtis/CTtis), different classes of AS vary in relative abundance but maintain constant relative ratios among all AS events (Figure 3-18). The most frequent AS class was ESI in all comparisons (>30%), followed by A3 and A5 (in total ~50%), MESI (~10%) and ISI (~9%). The most rare AS form was MEE.

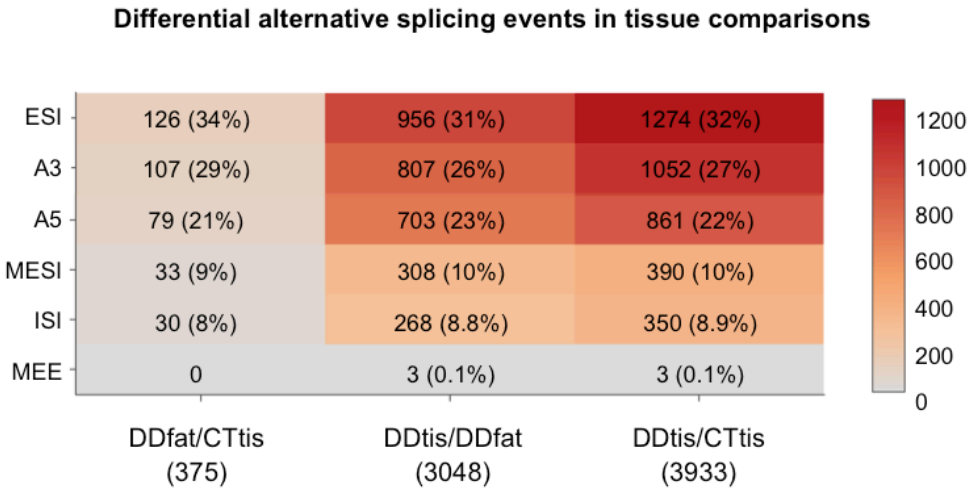


Figure 3-18 The classification of AS events in tissue comparisons

The relative abundance of individual AS event in each tissue comparison: DDfat/CTtis, DDtis/DDfat and DDtis/CTtis. Six major splicing events were analyzed including exon skipping /inclusion (ESI), intron skipping/inclusion (ISI), alterative 5' splice sites (A5), alternative 3' splice sites (A3), multiple exon skipping (MESI) and mutually exclusive exons (MEE).

Notably, a large difference in the number of AS events was observed in DDtis compared to DDfat or CTtis (Figure 3-18), whereas in DDfat compared to CTtis, only 375 AS events were detected. The same holds true for each of the AS event type (ESI, AD, A5, MESI and IR). The high number of differential AS events suggests a distinct set of transcripts exist in DD tissue, which may contribute to the DD nodule disease phenotype. Moreover, a DD tissue-specific mechanism regulating AS might underline the observed AS difference.

3.4.2 Characterization of isoform switching in DDtis

To understand the impact of AS on individual isoforms, we focused on isoform switching with in a gene in DDtis compared to DDfat or CTtis (described in Method 2.4.6).

Using spliceR[46], a set of 30 high confidence isoform switches, defined by a large change of isoform fraction (IF) between two conditions, was identified between tissue comparisons. In DDtis/CTtis, 21 isoform switches were identified. Similarly, in DDtis/DDfat, 17 isoform switches were identified (Table 3-11). Between DDtis/CTtis and DDtis/DDfat, common isoform switches were identified, which were isoforms for 7 genes including *PLOD2*, *TPM1*, *SPON2*, *MYO1D*, *GOLT1B*, *FAM198B* and *ACSL1*. In DDfat/CTtis, only 4 isoform switches remained after the first four filters, however, by checking junction reads on IGV, none of them can be validated.

The IF value for each of the 30 isoforms in each sample was plotted in a heatmap in Figure 3-19. The majority of isoform switches displayed a higher IF in DDtis compared to both DDfat and CTtis. Interestingly, in DDfat, a high value of IF for isoforms in some genes (such as *CD44*, *ITGA7* and *ACTN1*) was also observed (in more than 5 out of 9 DDfat samples). However, the IF change (dIF value) did not meet the minimal threshold for dIF (Method 2.4.6) if mean IF was compared across pooled samples. This suggests isoform switches identified in DDtis may be also present in DDfat, although not statistically significant. In sum, we identified isoforms switches in DDtis compared to either internal control DDfat or external control CTtis. This suggests these isoforms may have specific functions contributing to the development of disease tissues.

Table 3-11 Isoform switches in tissue groups

Group	Isoform Name	Ensembl Transcript ID	Cufflinks Transcript	Code	Start Exon	Stop Exon	Exon No.	ESI	MESI	ISI	A5	A3	log2 Isoform dIF	log2 Isoform FC	log2 Gene FC	Gene
DDtis/CTtis	*ACSL1-201*	*ENST00000437665*	*TCONS_00245614*	j	2	21	19	2	0	0	1	1	-0.8	-3.9	-3.1	*ACSL1*
	ACTN1-001	*ENST00000193403*	*TCONS_00095839*	=	1	21	21	5	1	0	2	3	1.0	1.7	0.8	*ACTN1*
	ATRN-001	*ENST00000446916*	*TCONS_00194582*	=	1	25	25	1	0	0	0	0	0.6	1.8	1.1	*ATRN*
	BMP1-010	*ENST00000306349*	*TCONS_00293516*	=	1	16	16	2	1	0	0	4	1.2	3.2	1.8	*BMP1*
	CD44-003	*ENST00000263398*	*TCONS_00045757*	=	1	9	9	1	1	0	2	3	1.0	2.0	1.1	*CD44*
	COL16A1-001	*ENST00000373672*	*TCONS_00018932*	=	2	71	70	6	1	0	5	5	1.0	3.2	2.2	*COL16A1*
	DNAJB11-201	*ENST00000265028*	*TCONS_00221564*	=	1	10	10	0	0	0	1	1	0.7	1.1	0.4	*DNAJB11*
	FAM198B-004	*ENST00000296530*	*TCONS_00244905*	=	2	5	3	1	0	0	0	0	1.1	2.6	1.4	*FAM198B*
	FBLN2-002	*ENST00000295760*	*TCONS_00213535*	=	2	17	16	2	0	2	0	1	1.0	1.8	0.7	*FBLN2*
	FOS-001	*ENST00000303562*	*TCONS_00090589*	=	1	4	4	1	0	0	1	1	-0.7	1.5	2.3	*FOS*
	GOLT1B-001	*ENST00000229314*	*TCONS_00064583*	=	1	5	4	0	1	0	0	0	0.9	2.2	1.2	*GOLT1B*
	ITGA7-002	*ENST00000553804*	*TCONS_00076251*	=	1	25	25	0	1	1	0	1	0.7	-2.2	-2.9	*ITGA7*
	KIF1B-003	*ENST00000377093*	*TCONS_00001027*	=	2	21	20	2	0	0	0	1	0.9	1.7	0.8	*KIF1B*
	LEPRE1-001	*ENST00000296388*	*TCONS_00020151*	=	1	15	15	1	0	0	2	5	0.7	2.7	2.0	*LEPRE1*
	*MYO1D-004**		*TCONS_00138059*	j	2	22	20	3	1	0	0	1	0.7	3.5	3.0	*MYO1D*
	PALMD-001	*ENST00000263174*	*TCONS_00007250*	=	1	8	8	0	1	0	2	2	0.6	-1.3	-1.9	*PALMD*
	*PLOD2-001**		*TCONS_00229911*	j	2	20	18	2	0	0	1	4	0.6	2.2	1.5	*PLOD2*
	PLXDC1-008	*ENST00000444911*	*TCONS_00138773*	=	1	13	13	0	2	0	1	1	0.7	3.1	2.2	*PLXDC1*
	SPON2-002	*ENST00000290902*	*TCONS_00239444*	=	2	6	4	1	0	0	2	1	0.8	4.0	3.2	*SPON2*
	SSC5D-005	*ENST00000594321*	*TCONS_00159739*	=	1	3	3	0	0	0	0	0	0.6	3.7	2.9	*SSC5D*
	TPM1-028	*ENST00000559556*	*TCONS_00101180*	=	1	9	8	3	1	2	3	2	2.0	3.7	1.8	*TPM1*
DDtis/DDfat	*ACSL1-201*	*ENST00000437665*	*TCONS_00245614*	j	2	21	19	2	0	0	1	1	-0.7	-3.6	-2.9	*ACSL1*
	*COL1A2-001**		*TCONS_00280877*	j	2	53	52	0	0	0	4	7	-0.6	3.4	3.9	*COL1A2*
	COL3A1-001	*ENST00000304636*	*TCONS_00178764*	=	1	51	51	1	0	0	2	1	-1.1	3.9	5.0	*COL3A1*
	FAM198B-004	*ENST00000296530*	*TCONS_00244905*	=	2	5	3	1	0	0	0	0	0.6	2.0	1.4	*FAM198B*
	GOLT1B-001	*ENST00000229314*	*TCONS_00064583*	=	1	5	4	0	1	0	0	0	1.0	2.2	1.2	*GOLT1B*
	*MYO1D-004**		*TCONS_00138059*	j	2	22	20	3	1	0	0	1	0.7	3.6	2.8	*MYO1D*
	*PALLD-002**		*TCONS_00238553*	j	3	11	9	1	1	0	0	2	1.1	3.2	2.1	*PALLD*
	PFKFB3-002	*ENST00000379775*	*TCONS_00031452*	=	1	15	15	2	0	0	1	2	-0.6	-3.3	-2.6	*PFKFB3*
	*PLOD2-001**		*TCONS_00229911*	j	2	20	18	2	0	0	1	4	0.7	2.5	1.8	*PLOD2*
	PYGL-001	*ENST00000216392*	*TCONS_00094963*	=	1	20	20	1	0	0	2	1	-0.7	-2.3	-1.6	*PYGL*
	RBPMS-003	*ENST00000339877*	*TCONS_00294164*	=	1	7	7	2	1	0	1	0	-0.8	-1.7	-0.8	*RBPMS*
	RCN1-003	*ENST00000381132*	*TCONS_00204262*	=	1	4	4	1	0	0	1	1	0.7	2.4	1.8	*RCN1*
	*SEC23A-001**		*TCONS_00094550*	j	3	21	19	2	1	0	1	1	0.9	2.3	1.4	*SEC23A*
	*SLC39A6-002**	*ENST00000269187*	*TCONS_00149029*	j	2	10	9	0	0	0	1	0	0.9	2.0	1.1	*SLC39A6*
	SPON2-002	*ENST00000290902*	*TCONS_00239444*	=	2	6	4	1	0	0	2	1	0.7	3.7	3.0	*SPON2*
	TPM1-028	*ENST00000559556*	*TCONS_00101180*	=	1	9	8	3	1	2	3	2	1.5	3.0	1.5	*TPM1*

Code j: represents "novel" isoforms not found in Ensembl b75

Code =: represents known isoforms in Ensembl b75

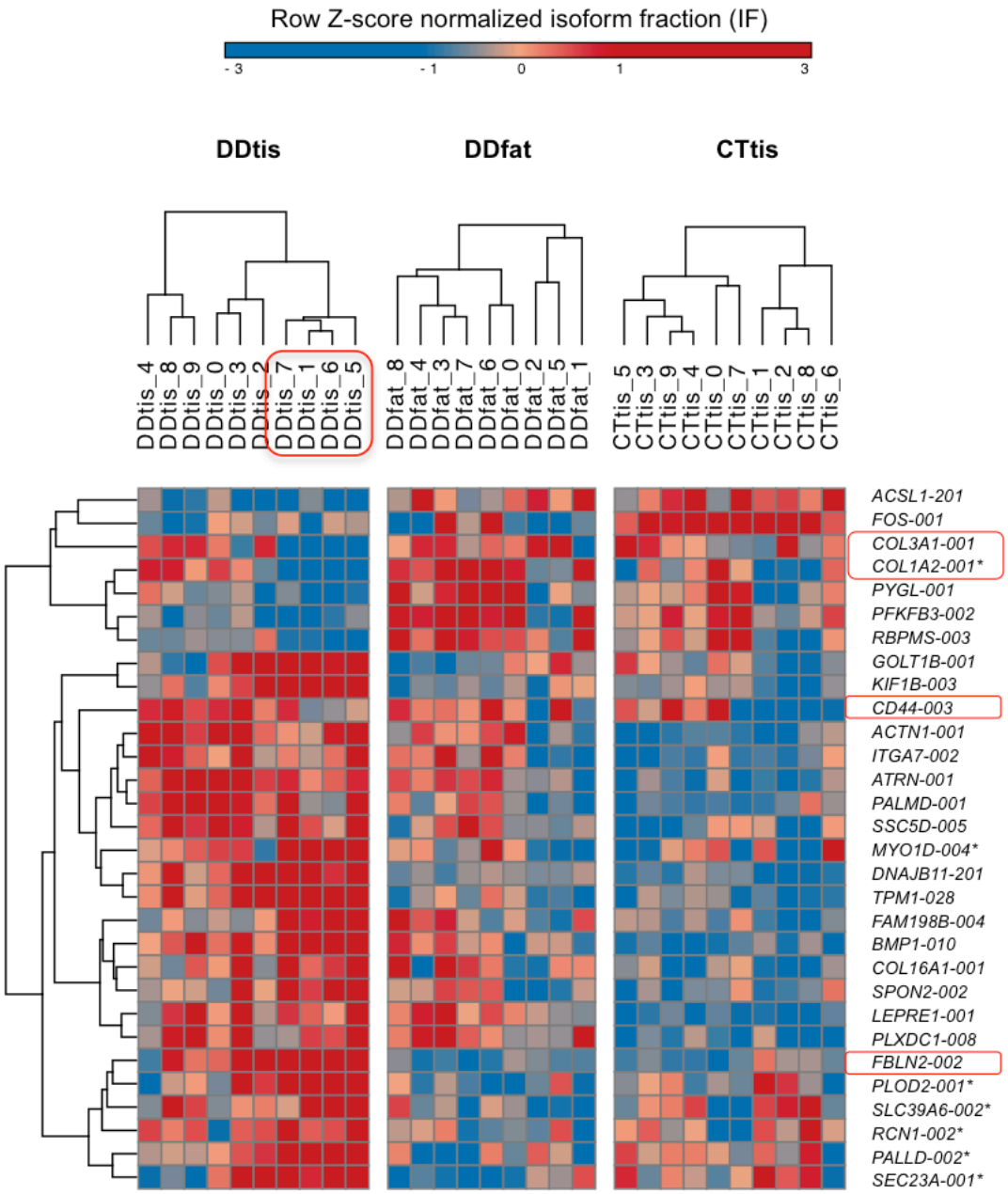


Figure 3-19 A heatmap representation of IF of 30 isoforms in individual tissue groups

In any of the 3 comparisons (DDtis/DDfat, DDtis/CTtis and DDfat/CTtis), transcripts defined as isoform switches were selected. Each box (in row) represents the calculated IF value for a given gene in each sample. The Z-score normalized IF of each isoform was used in the heatmap, in which the red color indicates the high expression, whereas blue indicates the low expression. Spearman's rank correlation was applied. The symbol * represents isoforms identified as novel isoforms.

3.4.3 Altered expression of five splicing factors in DDtis

An emerging concept proposes that tissue-specific splicing factors can coordinate AS involving isoforms encoding proteins to function in biologically coherent pathways and contribute to tissue specification[103]. Therefore, to understand the cause of the observed pattern of isoform usage in DD relevant tissues, I analyzed the expression change of 71 known splicing factors[104] using Cuffdiff results. In total, five splicing factors showed significant gene expression change in DDtis compared to CTtis. *RBFOX2* (RNA Binding Protein, Fox-1 Homolog 2) *and PTBP2* (Polypyrimidine Tract Binding Protein 2) were significantly increased in DDtis, whereas *NOVA1* (Neuro-Oncological Ventral Antigen 1), *ESRP1* (Epithelial Splicing Regulatory Protein 1) and *ESRP2* (Epithelial Splicing Regulatory Protein 2) were significantly decreased in DDtis (Figure 3-20A). The expression of these splicing factors in individual samples was shown in a heatmap in Figure 3-20B.

To test whether there is any correlation among the co-expressed splicing factors, Spearman's rank correlation test was performed using all 50 tissue samples (DDtis, DDfat and CTtis). As shown by the dendrogram (Figure 3-20B), a positive correlation was identified between *RBFOX2* and *PTBP2* expression (Spearman ρ=0.61, Bonferroni adjusted p-value < 0.05). Additionally, a strong positive correlation was observed between *ESRP1* and *ESRP2* gene expression (ρ=0.93, Bonferroni adjusted p-value < 0.05). In summary, the gene expression of five splicing factors was significantly altered in DDtis and it may represent one of the many expression profiles that characterize the intrinsic properties of DD tissues and their tissue-specific regulation of AS.

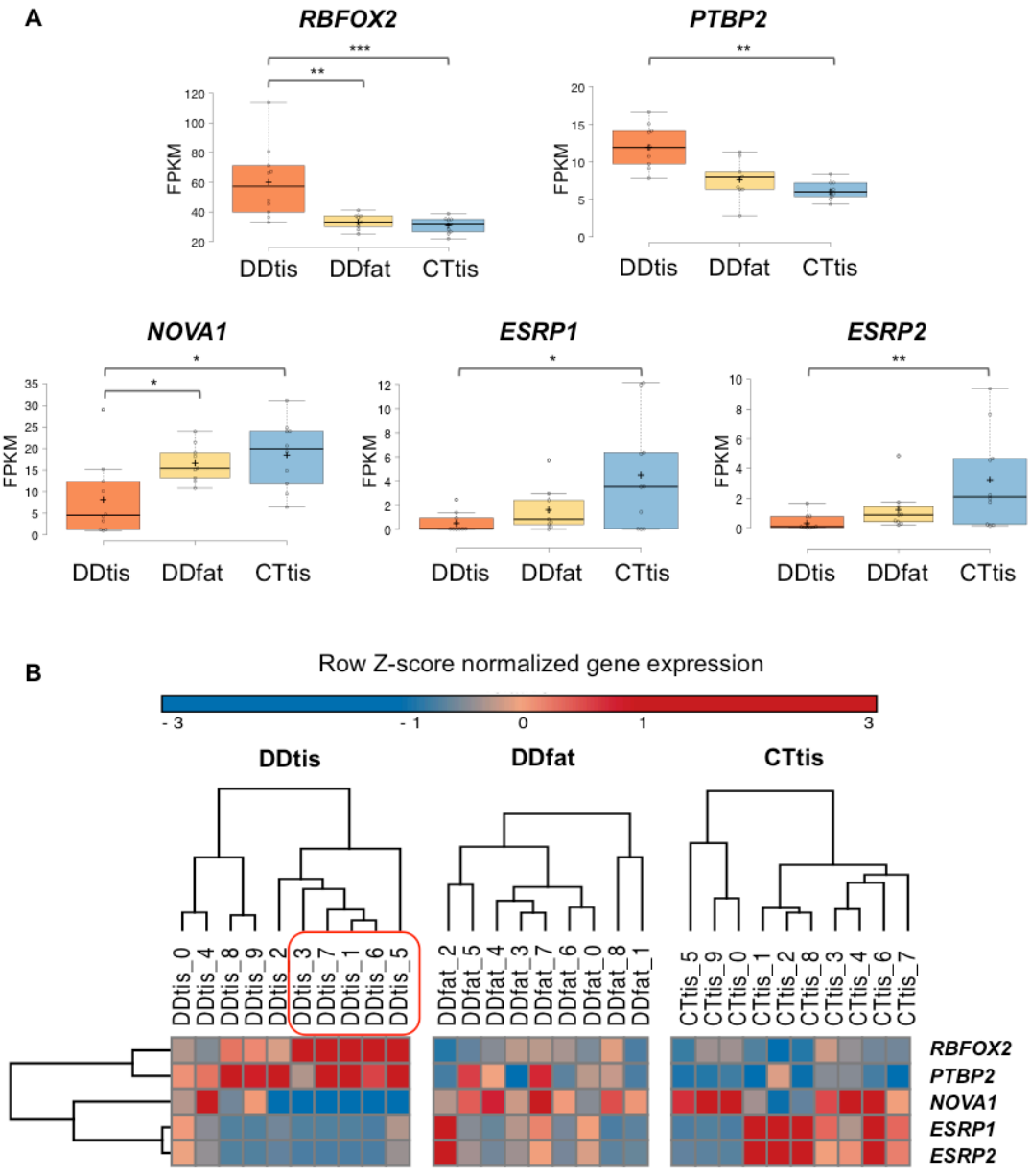


Figure 3-20 The significant gene expression change of 5 splicing factors in DDtis/CTtis

(A) Box plots of 5 splicing factors. Compared to CTtis, *RBFOX2* and *PTBP2* were significantly increased, whereas *NOVA1*, E*SRP1* and *ESRP2* were significantly decreased in DDtis. The y-axis is the FPKM of of a gene from Cuffdiff result. FDR adjusted q-value was applied. (*q ≤0.05, **q-value≤0.01, ***q-value ≤0.001).

(B) A heatmap representation of Z-score normalized expression of five splicing factors in individual tissue samples. The distance measure for hierarchy clustering of samples and gene expression was one-minus Spearman's rank correlation. A positive correlation was observed between *RBFOX2* and *PTBP2* gene expression (Spearman's rank correlation coefficient ρ=0.61, p-value < 0.001). A strong positive correlation was also observed between *ESRP1* and *ESRP2* gene expression (ρ=0.93, p-value < 0.001).

3.4.4 Specific correlation patterns between isoform switches and splicing factors in DD

To test whether there is a tissue-specific association between isoform switches and the gene expression of splicing factors, I took each isoform exhibited isoform switching (30 isoforms in Figure 3-19) and calculated the correlation coefficient between the IF of each isoform and the gene expression of each of the 5 splicing factors in each tissue type (DDtis, DDfat and CTtis). Spearman's rank correlation coefficient was used to assess the association and Bonferroni adjustment was applied.

In DDtis, positive correlations between *RBFOX2* expression and IF of two isoforms were observed (Figure3-21A). *RBFOX2* expression was strongly associated with the IF of *PLOD2-001** (Spearman's rank correlation coefficient ρ= 0.92, Bonferroni adjusted p-value < 0.05) in DDtis, in contrast to no correlation in DDfat and CTtis. Similarly, a gain-of correlation was observed between *RBFOX2* expression and IF of *SPON2-002* (ρ= 0.72, Bonferroni adjusted p-value < 0.05) in DDtis.

In DDtis compared to both DDfat and CTtis, *RBFOX2* was significantly increased as shown in Figure 3-20A. This suggests increased *RBFOX2* expression may be causal for the increased isoform usage of *PLOD2-001** and *SPON2-002*. *RBFOX* was known to regulate AS of *PLOD2* (procollagen-lysine, 2-oxoglutarate 5-dioxygenase 2)[105], which encodes a regulator of collagen stiffness[106]. Enhanced inclusion of the last exon of *PLOD2* by *RBFOX2* can lead to elevated expression of *PLOD2-001,* which is associated with TGF β-induced fibrosis[106,107]. However the *PLOD2-001** is a 'novel' isoform which is not in the Ensembl transcript database, though it is only different from *PLOD2-001* in A3 and A5 sites. Therefore, the presence of *PLOD2-001** isoform needs to validated and the regulatory association between *PLOD2-001** and *RBFOX2* needs to be tested *in vitro*.

In contrary to *RBFOX2*, *ESRP1/2* expression was significantly decreased in DDtis compared to healthy control CTtis. In CTtis, *ESRP1/2* expression was negatively correlated with *CD44-003* (ρ = -0.95, -0.93 separately, p-values < 0.05). However this correlation was lost in DDtis as well as in DDfat (Figure3-21B). *CD44* isoform switching represents a very conserved splicing event regulated by *ESRP1/2* in EMT and malignant cancer progression[49], which is used as an example for reduced AS usage in DDtis in the next session.

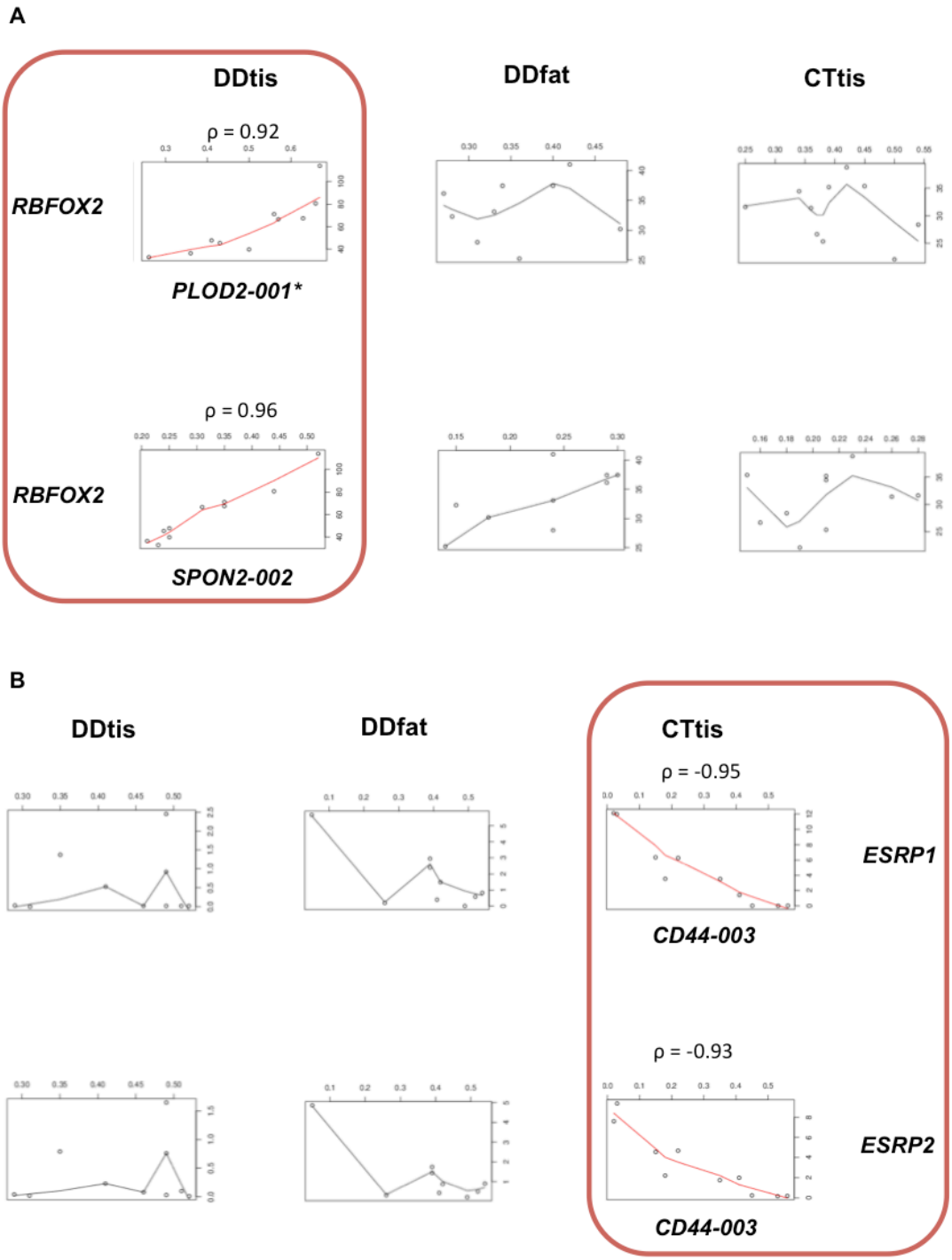


Figure 3-21 The tissue specific correlation between gene expression of splicing factors and IF of isoforms

The Spearman's rank correlation coefficient ρ for gene expression of 5 splicing factors and IF of 30 candidate isoforms in each tissue type were calculated. Bonferroni adjusted p-value ≤ 0.05 was applied. In total, significant correlation was identified in 4 pairs, including 2 isoforms (*PLOD2-001** and *SPON2-002*) correlated with *RBFOX2* (A). and one isoform *CD44-003* correlated with *ESRP1* and *ESRP2*.

3.4.5 Examples of tissue-specific AS in DDtis

3.4.5.1 Example 1: reduced *CD44* AS usage in DDtis and DDfat

As illustrated in Figure 3-22A, human *CD44* has 19 expressed exons, of which 10 exons are constitutive exons (C1-C10), whereas 9 exons are variant exons (V2-V9). The standard isoform *CD44s*, is composed of constitutive exons C1-C5 at the 5' end and C6-C10 at the 3' end. Between exon C5 and C6, there are nine variable exons (V2-V10) which are alternatively spliced to produce a plethora of isoforms[108]. *CD44s* and *CD44v* protein isoforms are all transmembrane proteins that function primarily to maintain tissue structure by mediating cell–cell and cell–matrix adhesion[109] (Figure 3-22B). Inclusion of the variable exons lengthens the extracellular membrane-proximal region by forming a heavily glycosylated stalk-like structure that provides interaction sites for additional molecules.

Here, the *CD44* AS events in DDtis were first visualized in Sashimi plot (Figure 3-22C). In both DDtis (n=9, out of 10) and DDfat (n=9), reduced junctions between exon C5 and exon C6 were observed suggesting less use of *CD44* variant exons compared to CTtis. Moreover, increased expression of the standard isoform *CD44s* was detected in DDtis compared to either DDfat or CTtis (Figure 3-22D). The gene expression of *CD44*, which represents the sum of all *CD44* isoforms, was also increased in DDtis compared to either DDfat or CTtis (Figure 3-22E). This suggests *CD44s* is the dominant isoform in DDtis, which contributes to the most of *CD44* expression observed in DDtis. In addition, in both DDtis and DDfat, a loss of correlation between *ESRP1/2* and *CD44*s isoform fraction was observed (Figure 3-21B).

Taken together, the reduced *CD44* AS and its loss-of-correlation by *ESRP1/2* were not only observed in DDtis but also in DDfat, suggesting the altered *CD44* AS and *ESRP1/2* expression may modulate both disease tissue (DDtis) and its niche (DDfat).

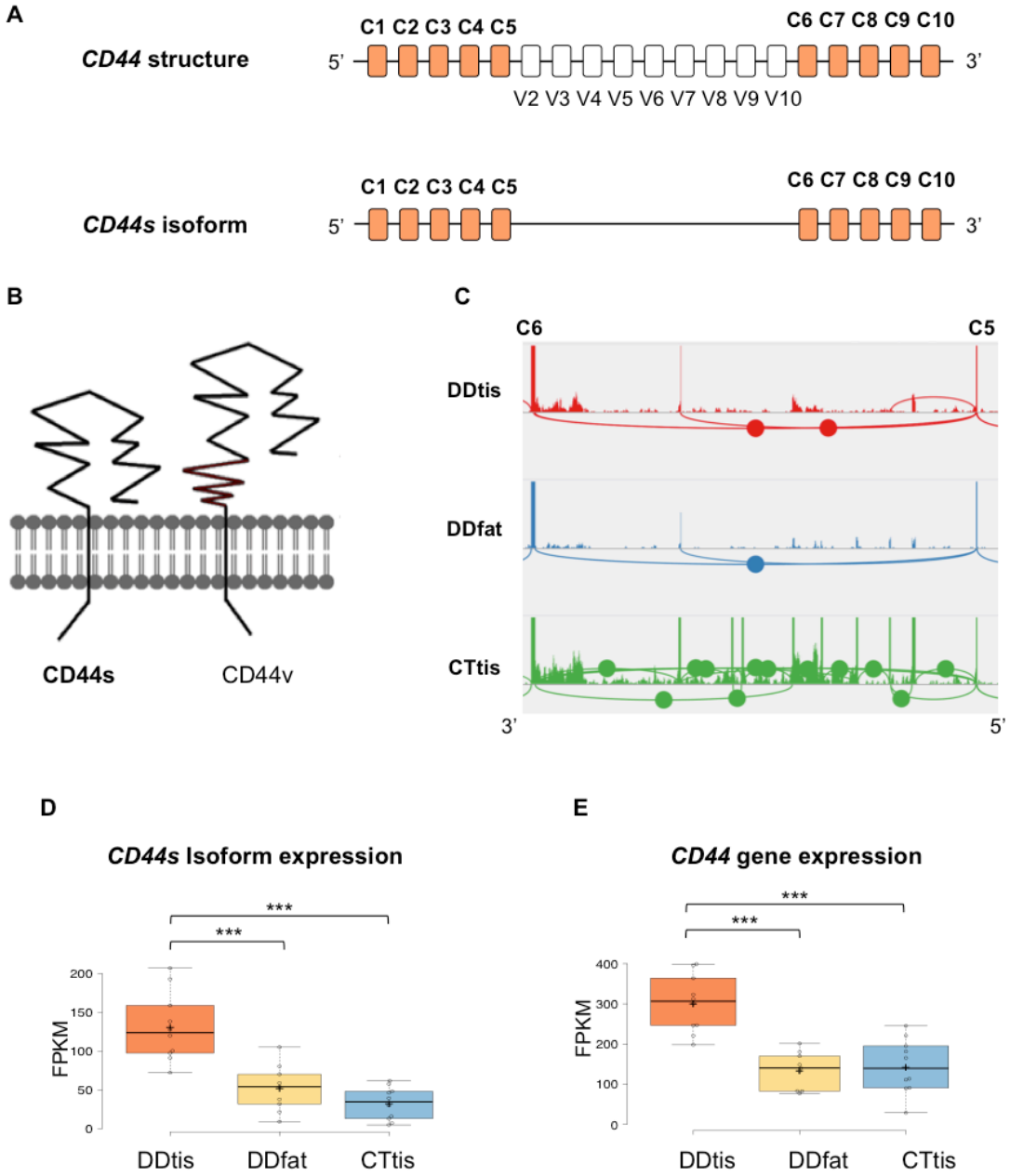


Figure 3-22 The significant increased IF of *CD44s* and *CD44* gene expression in DDtis

(A) *CD44* gene structure consists of 10 constitutive exons (C1-C10) and 9 variant exons (V2-V10) in human, The standard isoform *CD44s* is transcribed from constitutive exons only (C1-C10).

(B) The *CD44s* isoform has basic transmembrane domains, whereas *CD44v* has additional longer stems, which contains the variant exon(s). The figure was adapted from atlasgeneticsoncology.org/-Genes/CD44ID980CH11p13.html.

(C) Loss of alterative spliced *CD44* isoforms in DDtis and DDfat compared to CTtis.

(D) *CD44s* isoform expression, and (E) *CD44* gene expression were increased in DDtis compared to both DDfat and CTtis. FDR-adjusted p-value was calculated (***p-value ≤0.001).

3.4.5.2 Example 2: an imbalance of *FBLN2* isoform ratio in DDtis

Another interesting candidate is an isoform of Fibulin-2 (*FBLN2*). *FBLN2* gene contains 18 exons and encodes an ECM glycoprotein[110] (Figure 3-23A). The exon 9 can be alternative spliced and exclusion of exon 9 leads to a short protein isoform *FBLN2-002*, which is also named as *FBLN2*-Δexon9[110].

In line with the increased IF of *FBLN2*-Δexon9 detected by splice R (Figure 3-19), reduced expression of exon 9 was observed on Sashimi plot in 50% DDtis compared to DDfat or CTtis (Figure 3-23B). The overall *FBLN2*-Δexon9 isoform expression was clearly increased in DDtis compared to either DDfat or CTtis (Figure 3-23C, FDR q-value < 0.01), but the overall gene expression was not changed, leading to increased IF of *FBLN2*-Δexon9 in DDtis, suggesting a differential regulation of *FBLN2*-Δexon9 compared to other *FBLN2* isoforms.

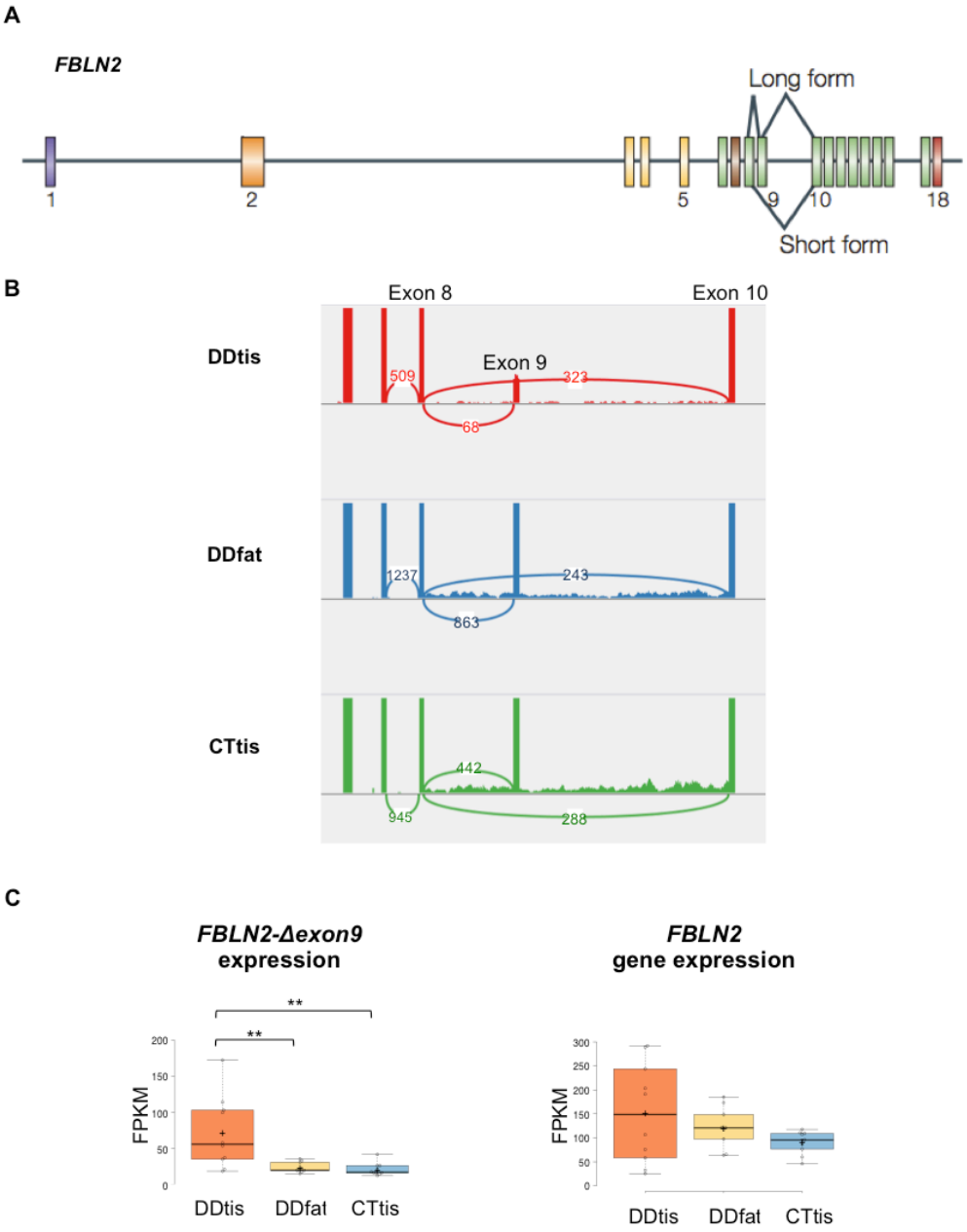


Figure 3-23 The significantly increased isoform fraction of *FBLN2-Δexon9* in DDtis

A) The gene structure of *FBLN2.* Figure was adapted from Timl et. al [110].

(B) The exon 9 of *FBLN2* in DDtis has reduced coverage corresponding to reduced expression. Increased Exon 9 skipping was observed in DDtis, as indicated by the higher number above the junction spanning exon9.

(C) Significantly increased isoform fraction of *FBLN2-Δexon9* was detected in DDtis and DDfat compared to CTtis (FDR adjusted q-value <0.01). However the gene expression of *FBLN2* was not significantly increased.

3.4.5.3 Example 3: Increased *COL1A2* AS usage as a feature in DDtis

As observed in Figure 3-19, a few isoforms displayed a lower IF in DDtis compared to either DDfat or CTtis, including two collagen isoforms (*COL1A2-001** and *COL3A1-001*). However the expression of both isoforms and their gene expression were significantly increased in DDtis compared to two controls. Therefore, the likely explanation is that the number of isoforms in DDtis is increased.

As shown in Figure 3-24, Sashimi plot was used to visualize all the isoforms in *COL1A2* in DDtis. Increased number of reads were mapped across exon-exon junctions in 9 DDtis compared to DDfat and CTtis suggesting increased AS of *COL1A2*. Therefore, although the isoform expression of *COL1A2-001** was increased in DDtis, its contribution to *COL1A2* gene expression (measured by IF in Figure 3-19) was reduced due to a larger number of alternatively spliced *COL1A2* isoforms. This indicates a specific set of *COL1A2* isoforms were produced in most of DDtis, which might be a general feature of DD affected tissues.

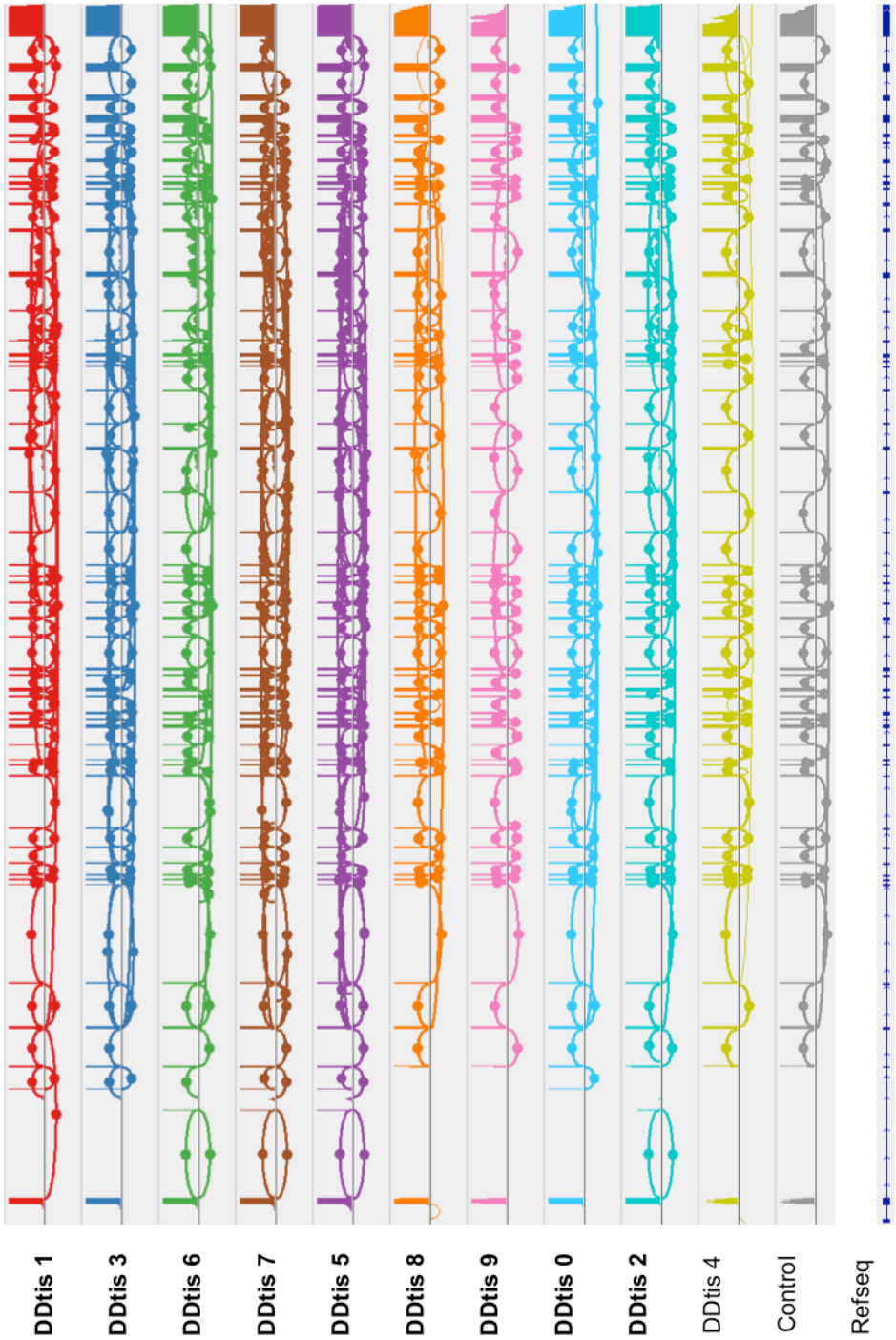


Figure 3-24 The increased use of AS of *COL1A2* in DDtis

Increased alterative spliced *COL1A2* isoforms in 9 out of 10 DDtis compared to a representative sample for both DDfat and CTtis.

3.4.5.4 Example 4: Extensive *COL3A1* AS usage as a subgroup feature of DDtis

In accordance to *COL1A2*, increased use of AS of *COL3A1* was also observed in the majority of DDtis (70%, shown in Figure 3-25), compared to both DDfat and CTtis. However, a remarkable variance within DDtis samples was noted. Extensive AS usage indicated by a large number of splicing junctions was observed in 4 DDtis samples (DDtis 1, 6, 7 and 5), which were referred as DD subgroup 1. In DDtis 3, *COL3A1* also exhibited elevated AS but to a lesser extent compared to DDtis samples in DD subgroup 1.

This variance with in DDtis was also observed in *RBFOX2* expression (Figure 3-20B). *RBFOX2* expression was about 2-fold higher in DD subgroup 1 compared to the rest of DDtis (Welch's t-test p=0.01). However, *RBFOX2* expression was not correlated with *COL3A1-001* IF (Spearman's rank correlation test, p-value > 0.05). So it is likely the *COL3A1* AS is regulated by other regulatory mechanisms instead of by RBFOX2.

Collectively, a larger number of *COL3A1* isoforms was observed in the majority of DDtis compared to either DDfat or CTtis. Particularly, a massive increase in *COL3A1* AS was observed in a subset of DDtis (4 samples in DD subgroup 1), which might represent a differential clinical profile of the patients or molecular mechanism of pathogenesis.

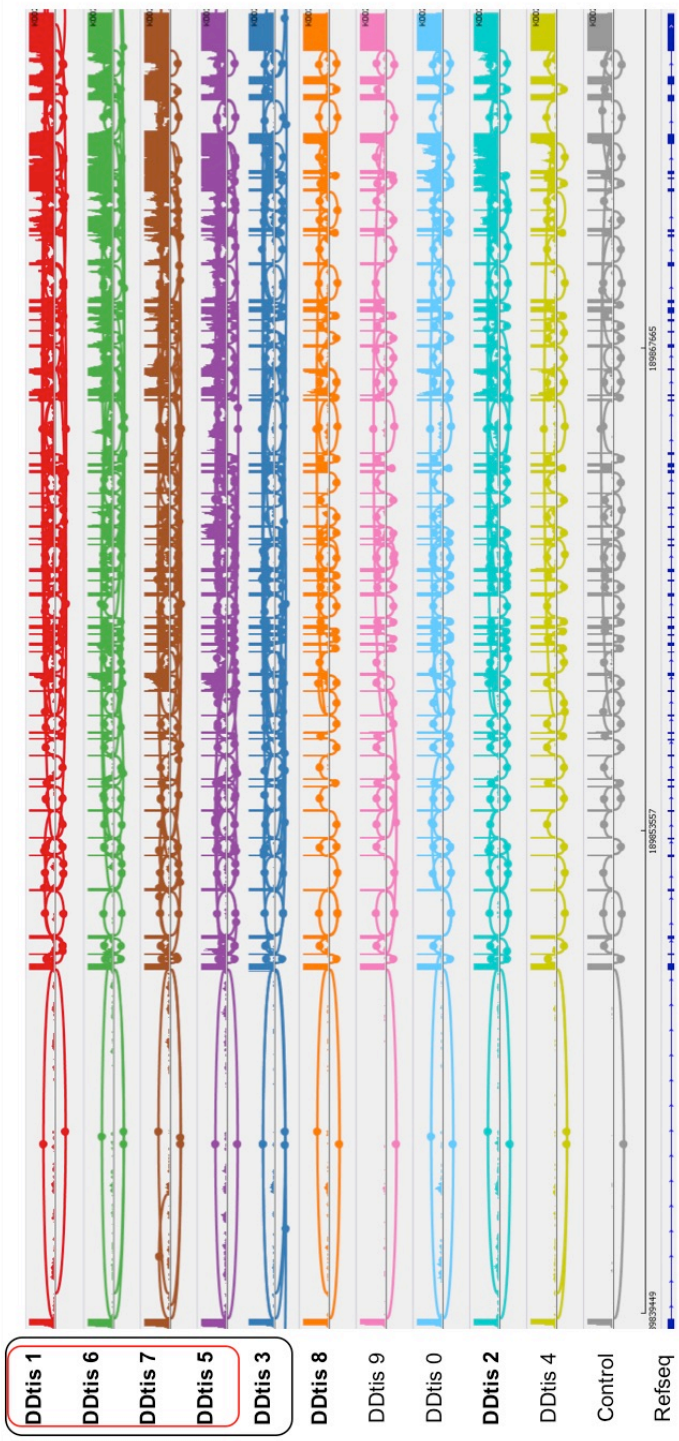


Figure 3-25 Extensive AS of *COL3A1* in a subset of DDtis

Increased use of AS for *COL3A1* in 6 out of 10 DD tissues compared to a representative DDfat was observed in the Sashimi plot. A circle above an arc indicates the number of reads that span an exon–exon junction (minimal number of 40 mapped reads). The blue transcript at the bottom represents the longest transcript of *COL3A1* in Refseq.

3.4.6 Characterization of two subgroups in DD based on *COL3A1* and gene expression profiling

DD is considered as a progressive fibromatosis disease that may be heterogeneous in many aspects including the severity and clinical appearance[111], treatment effect and recurrence rates (27% to 66%)[112]. These may arise from differential mechanisms of pathogenesis. Identification of biomarkers, which show homogeneity within one subgroup but difference in other subgroups, may help identifying subgroups of patients and reveal pathophysiological insights.

DDtis samples can be divided into two subgroups based on substantial or limited AS of *COL3A1*, as shown in Figure 3-25. To quantify the difference of the effect on gene level, the total expression of all isoforms, which means *COL3A1* gene expression, was examined. In DD subgroup 1 (DDtis 1, 6, 7 and 5) with extensive AS, expression of *COL3A1* was more than 10-fold higher than DD subgroup 2 (10960 FPKM vs. 925 FPKM) (Figure 3-26A).

The same classification of these two subgroups was determined using unsupervised hierarchical clustering of gene expression profiling, as previously showed in Figure 3-9A, suggesting a consistent molecular difference between subgroup 1 and 2. To capture the fundamental differences between two subgroups, I used Cuffdiff results to identify differentially expressed genes (DEGs) between two subgroups (defined as 1.5 fold change of gene expression, FDR adjusted q-value ≤ 0.05) (Figure 3-26A).

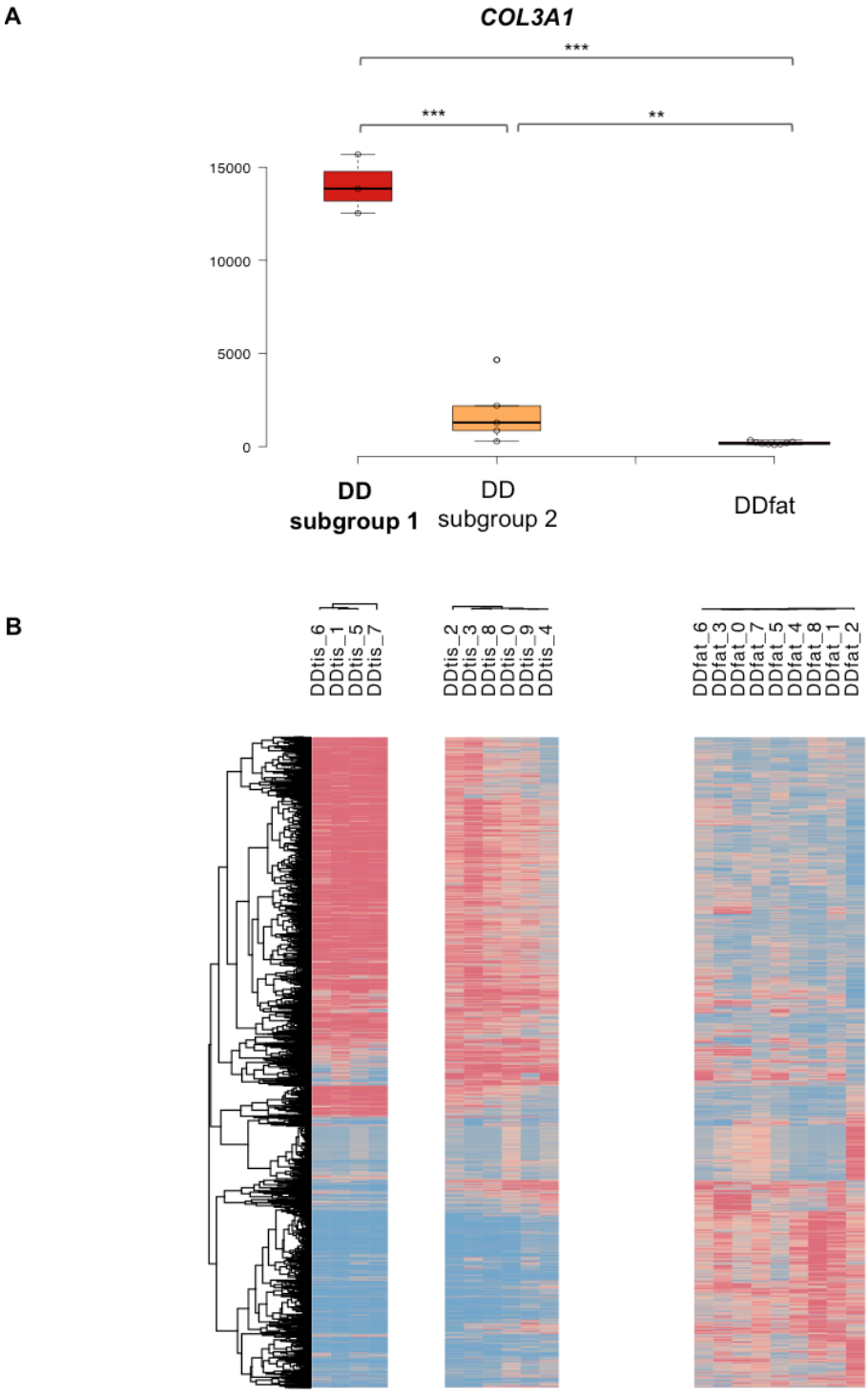


Figure 3-26 The stratification of two subgroups of DD patients

(A) *COL3A1* expression was significantly increased in DD subgroup 1 compared to subgroup 2. Subgroup 1 was referred to 4 DDtis samples exhibited extensive *COL3A1* AS

(B) Gene expression profiles in DD subgroup 1, DD subgroup 2 and DDfat.

Then GO terms and pathway overrepresentation analysis using significant DEGs (478 genes) were conduced using Enrichr tool[65]. As shown in Table 3-12, the most significant overrepresented GO terms were the 'ECM organization', 'ECM structure organization and disassembly' and 'collagen fibril organization'. The majority of these genes exhibited higher expression in DD subgroup 1 compared to subgroup 2, suggesting increased ECM accumulation and enhanced ECM stiffness in DDtis samples in subgroup 1.

Using Enrichr KEGG pathway analysis[66], four pathways were overrepresented (Table 3-13) including 'ECM-receptor interactions' (12 DEGs) and 'PI3-Akt Signaling pathway' (22DEGs). The 24 DEGs overlapped with 'Pathways in Cancer' include genes encoding protein in the TGFβ pathway (TGFβ2, TGFβ3 and LAMB1), Wnt pathway (WNT11, LEF1 and FZD4), transcription factors (CEBPA and RUNX1), cytokines (CDKN1 and CDKN2A), and fibroblast growth factors (FGF2 and FGF16). The 'Protein Digestion and Abortion' pathway was only considered as associated but not functional in DDtis since all the 13 overlapped DEGs with this pathway are genes encoding ECM.

Table 3-12 Overrepresented GO biological processes in DD subgroup 1/subgroup 2

	GO Biological Process	Adjusted p-value	Z-score	Combined score
1	extracellular matrix organization	1.0E-13	-2.4	71.2
2	extracellular structure organization	1.0E-13	-2.4	71.2
3	extracellular matrix disassembly	6.2E-09	-2.2	41.3
4	collagen fibril organization	5.2E-07	-2.5	36.6
5	multicellular organismal metabolic process	1.8E-06	-2.2	28.6
6	collagen metabolic process	2.0E-06	-2.2	28.4
7	multicellular organismal macromolecule metabolic process	4.1E-06	-2.2	27.1
8	collagen catabolic process	1.7E-05	-2.2	23.6
9	response to wounding	9.1E-05	-2.4	22.0
10	multicellular organismal catabolic process	3.6E-05	-2.2	22.0

Table 3-13 Overrepresented pathways in DD subgroup 1/subgroup 2

KEGG pathway 2016	Overlap	Adjusted p-value	Z-score	Combined score
Protein digestion and absorption	13	0.000	-1.72	14
ECM-receptor interactions	12	0.000	-1.68	14
Pathways in cancer	24	0.004	-2.03	11
PI3K-Akt signaling pathway	22	0.004	-2.02	11

3.5 Integrative analysis combining exome and RNA-seq data

3.5.1 The overlapping cohort for WES and RNA-seq design

The 11 patients in RNA-seq cohort were also included in the WES study. RNA-seq on a larger sample size, for example RNA-seq for DDtis from the rest of 39 patients in WES study, was not conducted mainly owing to our stringent case-control sample collection criteria, such as a) matched age between case and control when samples were collected, b) the availability and the high quality of RNA from tissue samples and cells. Although the cohorts for RNA-seq and WES studies do not completely match, a comparison of both studies may provide insight to the true candidate genes and pathways.

3.5.2 Functional candidate genes with genetic predisposition

In the whole exome study, 12 genes carrying rare variants were identified as DD phenotype-related genes. To assess whether these genes have transcriptomic consequences, I examined the gene expression of these 12 genes in the RNA-seq gene expression data. As shown in Table 3-14, 8 of the 12 candidate genes displayed significantly differential expression in at least one comparison group in tissue comparison (DDfat/CTtis, DDtis/DDfat and DDtis/CTtis) or cell comparison (DDcell/CTcell). The expression of 6 genes was significantly changed in more than two comparisons, suggesting functional roles of these genes in transcriptome. They are genes encoding collagen (*COL7A1*, *COL11A1*, *COL5A1* and *COL1A2*), *ROR2* (receptor tyrosine kinase-like orphan receptor 2 — a single-span transmembrane receptor regulating both canonical[113,114] and noncanonical Wnt signaling pathway[115,116]), and *FLNB* (Filamin B). Besides increased gene expression, increased AS usage for *COL1A2* was also identified in DDtis (Figure 3-24), suggesting *COL1A2* as a functional candidate for DD.

In summary, combining exome and RNA-seq data, 6 candidate genes carrying rare variants were suggested to play a functional role in DD transcriptome. These candidate genes are *COL7A1*, *COL11A1*, *COL5A1*, *COL1A2*, *ROR2* and *FLNB*.

Table 3-14 Functional candidate genes with genetic predisposition

Exome data										RNAseq data					
Gene annotation				Rare variant annotation						RNAseq Gene expression					Allternative splicing
Chr	EvoTol	Phenolyzer (0.0001 - 5.7)	Gene burden	dbSNP138	Start	End	Ref	Alt	Allele count	Gene	DDfat /CTtis	DDtis /DDfat	DDtis /CTtis	DDcell/ CTcell	
9	9.28%	1.2	3	rs35852786	94487187	94487187	C	T	3	***ROR2***	up	up	up		
3	0.86%	1.3	4	.	48608296	48608296	G	A	1	***COL7A1***	up	up	up		
				rs139434755	48629340	48629340	G	A	1						
				rs35623035	48630252	48630252	G	A	2						
1	0.32%	1.2	3	rs139064549	103354135	103354135	G	C	2	***COL11A1***	up	up	up		
				rs141978499	103544374	103544374	C	G	1						
7	0.27%	1.2	2	.	94055143	94055143	G	A	1	***COL1A2***		up	up		**increase**
				.	94057101	94057101	A	C	1						
9	3.48%	1.0	4	rs41306397	137591755	137591755	C	T	1	***COL5A1***		up	up		
				rs199735010	137694750	137694750	C	T	1						
				rs368305377	137702117	137702117	C	T	1						
				rs61739195	137708884	137708884	C	T	1						
3	0.04%	1.4	3	.	58141747	58141747	T	C	1	***FLNB***		down	down	down	
				rs116826041	58145348	58145348	T	C	1						
				rs149638325	58148895	58148895	C	T	1						
5	5.58%	3.6	2	rs114260147	149359938	149359938	C	G	1	*SLC26A2*	up				
				rs104893915	149359991	149359991	C	T	1						
7	10.76%	1.4	2	rs373336251	55240795	55240795	G	A	1	*EGFR*			down		
				rs201830126	55268023	55268023	G	A	1						
11	20.26%	1.4	2	rs61756429	111608216	111608216	T	A	1	*PPP2R1B*					
				rs115287852	111612783	111612783	T	C	1						
5	0.24%	1.2	2	rs1801166	112175240	112175240	G	C	1	*APC*					
				rs141010008	112178781	112178781	C	T	1						
5	21.06%	1.2	2	rs17223632	141993631	141993631	C	T	2	*FGF1*					
17	15.02%	1.4	2	.	37879585	37879585	A	C	1	*ERBB2*					
				rs55943169	37884176	37884176	C	A	1						

3.5.3 Shared overrepresented pathways in the exome and transcriptome data

To further define the pathogenic contribution of the genetic-relevant pathways identified in the exome study, I compared the overrepresented pathways in the exome and transcriptome analysis. The common pathways were shown in Figure 3-27 with Bonferroni adjusted p-value (-log10 transformed) plotted (The blue boxes represent the significant pathway overrepresentation).

Among the overlapped 5 pathways, the KEGG Hippo network was overrepresented in exome analysis (using 12 candidate genes) and transcriptome analysis of DD/DDfat and DDtis/CTtis. This suggests the genetic perturbation of this pathway might have a high penetrance leading to disturbed activity of this pathway in the development of disease tissue and its niche.

Pathways specifically related to DDtis were also identified, which include the 'Focal Adhesion' and 'PI3-Akt pathway'. Similarly, the 'Pathways in Cancer' was also associated with DDtis, which might suggest some common features shared by tumor and the localized benign tumor of DD (DD nodule tissue). However this pathway designated by KEGG involves more than ten pathways including the 'Wnt signaling pathway', 'PI3K-Akt pathway' and 'ECM-receptor interactions' etc (http://www.kegg.jp/keggbin/highlight_pathway?scale=1.0&map=map05200&keyword=pathways%20in%20cancer). Therefore, the 'Pathways in Cancer' cannot be attributed as a specific functional candidate pathway in DD. In addition, genes overlapped with the 'Protein Digestion and Absorption pathway' are all ECM genes, which suggest this pathway is only relevant instead of functional.

In sum, three pathways were suggested to be genetically and functionally overrepresented in DD, including the 'Hippo network' contributing to the niche environment, the 'Focal Adhesion' and the 'PI3-Akt pathway' contributing to the DD phenotype development.

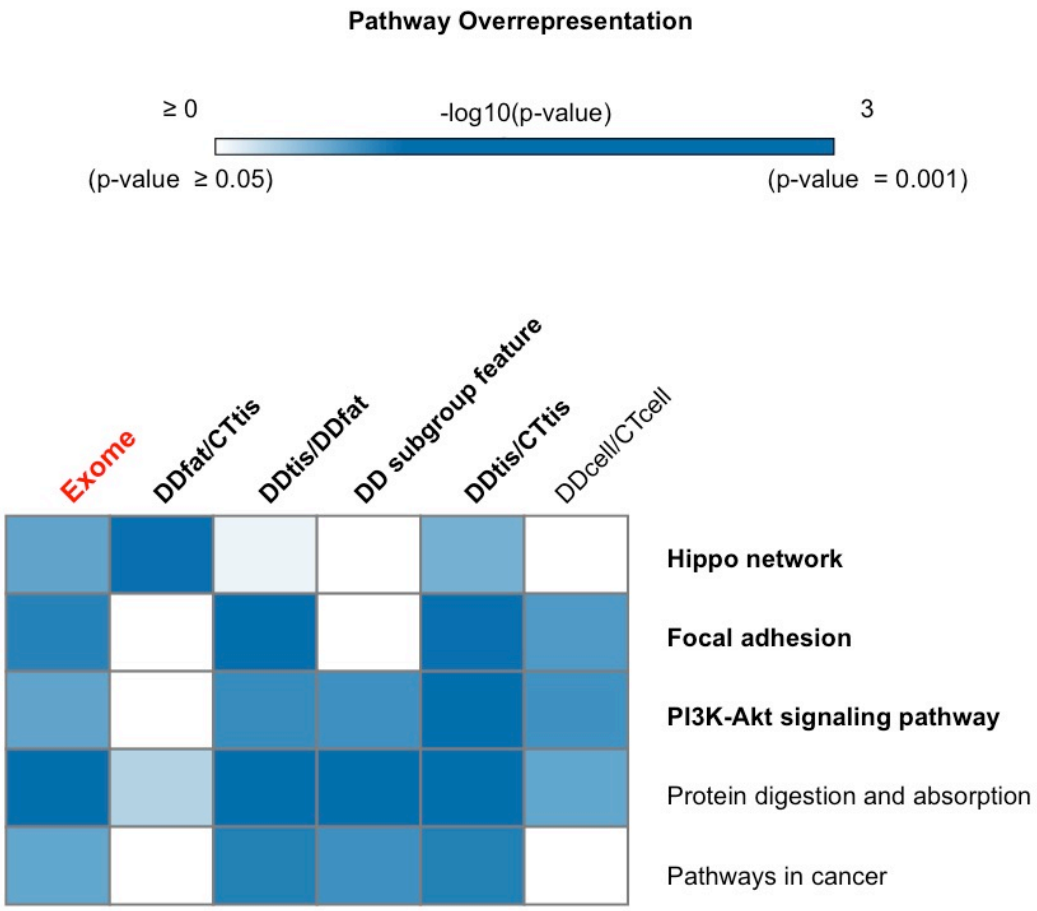


Figure 3-27 Overrepresented pathways shared in exome and transcriptome data

A summary of common pathways overrepresented by exome candidate genes (12 genes) and DEGs in transcriptome data from 5 comparisons is shown. The DD subgroup feature represents the pathways enriched using DEGs between DD subgroup1 and DD subgroup 2. The intensity of the shading indicates the significance of the KEGG pathway overrepresentation using Enrichr analysis. All blue boxes have a Bonferroni p-value of <0.05, white boxes have a p-value of >0.05. The negative log10(Bonferroni p-value) was used in the heatmap.

Chapter 4 Discussion

DD is an aging-related common disease with a high degree of genetic predisposition. Identifying genetic components contributing to the disease etiology could greatly improve the pre-symptomatic diagnosis and treatment of the disease. Therefore, in this study, I designed three sets of experiments to identify genetic components in DD including:

1) identification of functional variants contributing to a strong GWAS association signal using targeted NGS

2) prioritization of DD phenotype-related genes carrying a mutational burden as revealed by whole exome sequencing (WES)

3) characterization of the transcriptional deregulation in disease tissues/cells using RNA-seq.

4.1 A candidate gene carries functional variants at the 7p14.1 GWAS locus

In this study, we applied a 'risk haplotype block' strategy in which the risk allele of the top GWAS association lead (ranked by lowest p-value) haplotype block at 7p14.1 was determined by interrogating the tagging SNP rs16879765. Using targeted NGS in a discovery cohort (n=96) selected according to the 'risk haplotype block' information, we identified a rare deleterious coding variant, rs149095633 (C>T, p.P121L), and a common regulatory variant, rs2044831, both situated in *EPDR1* at 7p14.1.

In the subsequent randomly selected, independent validation cohort, the rs149095633 minor allele T was 35-fold enriched compared to large European control populations, suggesting an association between the rare allele of rs149095633 and DD.

The SNV rs149095633 (C>T) leads to an amino acid substitution from proline (CCC) to leucine (CTC) at position 121 (p.P121L) in the β-turn domain of EPDR1. Proline is a unique amino acid where the side chain is cyclized onto the backbone twice, forming a five-membered ring[81]. This unique feature makes proline unable to occupy many of the main-chain conformations which are easily adopted by all other amino acids[81]. Instead, proline is statistically preferred in β-turns because its cyclic structure is ideally suited for the tight β-turns, making it so-called β-proline[117]. Despite being aliphatic and preferred for β-turn structures, prolines are also found on the protein surface[117]. In contrast, leucine prefers to be buried in protein hydrophobic cores and it plays an important role in stabilizing α-helices[81]. Therefore, it is likely that this P121L substitution disturbs the local β-turn structure, which might lead to a change of EPDR1 function or its intermolecular interactions.

Moreover, allele specific expression (ASE) analysis in DD cells suggests that the risk allele T of rs149095633 might lead to an increased expression of *EPDR1.* A higher expression of the DD risk allele C of rs2044831 was also observed in DD cells suggesting rs2044831 to be an eQTL for *EPDR1.* Notably, both variants were identified in the *EPDR1* gene and related to *EPDR1* expression, suggesting that *EPDR1* is the right candidate gene at the risk locus 7p14.1 associated with DD.

EPDR1 encodes Ependymin Related 1 Protein. An ortholog of EPDR1 was first identified in reticular shaped fibroblasts in brain extracellular fluid in zebrafish and reported to play a role in cell adhesion[118]. The DNA and protein sequence of EPDR1 is highly conserved among vertebrates[119]. Yet the precise function of human EPDR1 is unknown. Higher expression of *EPDR1* was observed in colon cancer a long time ago hence *EPDR1* is also known as *UCC1* (Upregulated In Colorectal Cancer Gene 1)[120]. The high expression of *EPDR1* was particularly linked to the stemness phenotype of colon cancer[121] and high-risk myeloma[122].

A recent *in vitro* knockdown study revealed that decreased *EPDR1* expression is associated to decreased cellular contractility of DD cells[123]. This suggests EPDR1 is involved in cell-matrix interactions in DD cells and expression of *EPDR1* might regulate cellular contractility. Thus, abnormal *EPDR1* overexpression is likely to increase the cellular contractile activity, which is a hallmark of DD cells. If *EPDR1* indeed contributes to both cancer and DD etiology, a different or weaker effect of *EPDR1* is expected in DD, as DD is defined as an aging related benign disease. In our data, the changes in ASE of both genetic variants (rs149095633 and rs2044831) observed *in vitro* may translate approximately into an increase of total *EPDR1* expression of 1.1-fold in DD cells. Given that a 3 fold change in *EPDR1* expression was sufficient to affect the contractile phenotype *in vitro*[123], our data supports the notion that a modest life-long increase of *EPDR1* expression may represent the molecular basis of an increased risk to develop DD*. In vitro* functional studies of the effect on *EPDR1* expression of both variants are to corroborate this hypothesis.

4.2 Functional candidate genes contributing to the DD phenotype

Over the past few years, advances in NGS and its affordable pricing have accelerated genomic studies in complex diseases to systematically investigate genetic variants and involved genes using WES. RNA-seq data are helpful to reasonably interpret the functional elements of the genome and reveal the molecular mechanisms. Here we analyzed whole exomes of 50 DD patients using WES and whole transcriptomes in 11 of these patients using RNA-seq.

Using a phenotype-driven strategy, 12 genes with rare variants were identified as DD phenotype-related candidate genes. By EvolTol prediction, mutations in these genes are likely to be pathogenic in the palmar part of the hand. By incorporating RNA-seq data, 6 of the 12 candidate genes exhibited significant differential expression in more than two comparisons involving disease relevant tissues/cells (Table 3-14).

Three genes exhibited a signature related to DDtis or DDcell, of which *COL1A2* and *COL5A1* showed increased expression in DDtis compared to either its matched DDfat or external control CTtis. A further gene, *FLNB*, displayed decreased expression in DDtis and DDcell compared to CTtis and CTcell, respectively. Another three genes (*ROR2*, *COL7A1* and *COL11A1*) exhibited an upregulation in a pattern DDtis >DDfat >CTtis. This suggests not only a change of gene expression in the disease tissue (DDtis) but also in the niche of the disease tissue (DDfat) compared to healthy CTtis.

Among the investigated four collagen genes, three (*COL1A2*, *COL5A1* and *COL11A1*) encode fibril-forming collagens (type I, V and XI), which function as the principal source of tensile strength in tissues[124]. *COL7A1* encodes type VII collagen, the major component of the anchoring fibrils beneath the basal lamina which functions as an important adhesion molecule at the dermal-epidermal junction[125]. Overexpression of these collagen genes has been linked to various TGFβ-mediated fibrosis conditions[126-131].

Additionally, the altered expression of the other two genes, *FLNB* and *ROR2*, was also associated to tissue fibrosis. FLNB is an actin-binding protein, which regulates cytoskeleton-dependent cell proliferation, differentiation and migration[132]. A recent study has linked the absence of FLNB to increased activity of TGFβ signaling both *in vitro* and *in vivo*[133]. ROR2, a transmembrane protein[134] has been identified as an upregulated cell surface marker for human mesenchymal progenitor cells and it is suggested to play a role in cell proliferation and migration[135]. Moreover, *ROR2* overexpression was also identified to result in a partially activated state for the Wnt/β-catenin signaling and an enhancement of downstream target genes following Wnt3a stimulation in renal cell carcinoma cells and HEK293T cells[136]. Therefore, in the DDfat/CTtis comparison, the increased expression of both *ROR2* and 6 other genes, which are acting upstream of the Wnt/β-catenin signaling pathway[136] (shown in Figure 3-12A, described in session 3.3.4.1), suggests that the niche of disease tissue (DDfat) harbors a partially activated state of the canonical Wnt signaling pathway, which might be necessary for TGFβ-mediated fibrosis[137] in DDtis.

Overall, two clusters of DD candidate genes seem to be related to either disease tissue or its niche. Deleterious variants in these genes are likely to affect their functions and to contribute to fibrosis in the palm.

4.3 The mechanisms involved in DD pathogenesis

With the goal of finding genetic causes, linkage analysis in a pedigree[138] and GWAS studies in DD cases and controls[19,20] have been performed. As a result, we begin to better understand the genetic basis of DD. However, the contributions to DD of particular genetic variants and functional candidate genes are still elusive, which might be due to genetic heterogeneity. By GWAS analysis, six DD-associated loci were identified which harbor genes involved in the Wnt pathway, suggesting that genetic variants of these genes belonging to this common pathway are responsible for DD. Recently, the development of RNA-seq transcriptome analysis provides an effective way to link genes and their products into functional pathways. Therefore, integrating the pathways overrepresented by genes carrying genetic variants with pathways affected in transcriptome analysis may help to identify the key mechanisms contributing to DD.

In exome and transcriptome analyses, three common pathways (Hippo network, PI3K-Akt pathway and Focal Adhesion) were overrepresented (Figure 3-27), which supports the hypothesis that the genetic variants in critical functional pathways may affect pathogenesis in this disease.

4.3.1 The Hippo network and the altered niche

The role of adipose tissue around DD tissue had been largely ignored for a long time. In our DD study cohort, we constantly observed DD patients having a lean phenotype. This association was also independently observed in an epidemiology study of DD in an Icelandic population[97], which supports a significant low adiposity association with male DD patients. Therefore, it is important to include the matched unaffected adipose tissue from DD patients (DDfat) and characterize the possible changes in gene expression and function.

By comparing DDfat with healthy adipose tissue (CTtis), the KEGG Hippo network was suggested to be the most significantly perturbed pathway (Figure 3-12B). It represents a 'parent' network including the 'child' Hippo signaling pathway (also named as the YAP/TAZ pathway), the Wnt/ß-catenin signaling pathway and TGFß pathway (shown in Figure 3-12B).

By perturbation analysis, YAP/TAZ and Mst1/2, the major effectors of the Hippo signaling pathway[139], were predicted as accumulated and permanently activated. Recent studies have shown that YAP/TAZ can serve as important mechanosensors[140] between cells and their microenvironment.

Under *in* vivo conditions, cells are exposed to a signaling microenvironment including specific ECM, secreted proteins, growth factors and ions, resulting in a soft, topographically featured substrate. This microenvironment does not only control the biochemistry of the substrate, but also maintain the mechanical properties generated by the substrate. By definition, the mechanical properties indicate the elastic or inelastic behavior of a substrate under force, such as the stiffness (resistance to deformation), elongation and tensile strength[141]. The mechanical properties are critical for tissue

morphology and function but vary, depending on tissue types. For instance, adipose tissue has a stiffness of 2 kPasal (kPa)[142] and muscles have more than 12 kPa[143]. During fibrosis, the tissue stiffness is highly increased. For example, during bleomycin-related lung injuries, the localized stiffness was increased 5-fold (from 3kPa to 15kPa)[144].

In the DD exome study, genetic rare variants were identified in two collagen genes (*COL7A1*, a anchoring fibril collagen gene and *COL11A1,* a fibril-forming collagen[124] gene), which form fibril structures that increase ECM stiffness. In the following whole-transcriptome gene expression profiling studies, the expression of both genes was increased not only in DDtis/CTtis, but also in DDfat/CTtis comparisons (Table 3-12). This suggests that the tissue stiffness of both disease tissue (DDtis) and its niche (DDfat) is increased compared to healthy adipose tissue (CTtis). Recently, a significant correlation between increased adipose tissue stiffness and lower BMI was observed[145]. Taken together, we propose a first hypothesis according to which the stiffness of DD palmar fasia (DDfat) is likely to increase due to loss of subcutaneous fat (low BMI as a marker).

Multiple studies have revealed that the increased stiffness of substrate can inhibit the Hippo signaling pathway and induce YAP/TAZ activity[146]. Furthermore, the stiffness of substrate has been recently reported to promote adipocyte progenitors (with low cellular stiffness) to differentiate into fibroblasts (higher cellular stiffness), which positively feedback to further increase tissue stiffness[147]. This leads to a second hypothesis, which suggests an extensive source of fibroblasts in the early stage of DD.

As RNA-seq only captures the transcriptome at a given time point, it is difficult to interpret whether the inactivation of the Hippo pathway and an increased YAP/TAZ activity is causal or a functional consequence. However, in the exome study, the KEGG Hippo network was one of the overrepresented pathways based on 12 candidate genes. This suggests a third hypothesis, according to which the genetic predisposition in the Hippo network component may lead to the altered YAP/TAZ activity in the niche adjacent to DD.

Overall, in DDfat (the palmar adipose tissue surrounding DDtis), the Hippo pathway is suggested to be inactive, which leads to activation of YAP/TAZ effectors. Moreover, the increased expression of fibrillar collagen genes in DDfat may increase tissue stiffness, which further activate the YAP/TAZ pathway and might lead to adipocyte-fibroblast transition. In addition, a genetic burden on the Hippo network or collagen genes may enhance the activity of the YAP/TAZ pathway.

4.3.2 Fibrosis signatures of DD

The activation of YAP/TAZ is known to promote TGFβ production[148,149]. In the gene expression profiling analysis, TGFβ expression was highly increased in DDtis compared to DDfat or CTtis, which is in accordance with previous findings. TGFβ release induces fibroblast proliferation and myofibroblast transition[12]. Transition to myofibroblasts is characterized by α-SMA expression, high contractile activity and enhanced ECM deposition[14]. The extensive deposition of ECM increases the

tissue stiffness, which further amplifies the activity of the TGFβ pathway and eventually leads to fibrosis[141].

In this study, several signatures related to DD tissues were characterized including prominent ECM-receptor interactions and fibrosis tissue-specific splicing. The enhanced ECM-receptor interactions are general features of fibrosis and have been associated with DD in other expression profiling studies[33]. However, we here provided for the first time a basic understanding of fibrosis-specific splicing signatures in DDtis.

Alternative splicing (AS) is often regulated in a tissue-specific manner[103]. As a result, a gene undergoing AS can encode distinct protein isoforms involved in different biological processes in specific tissues[150]. The changes in AS have been linked to translation of aberrant proteins that can contribute to many diseases, such as Alzheimer's disease[151,152], Parkinson's disease[153], cystic fibrosis[154] and cancer[155,156]. The effect of AS in DD has not yet been addressed. Therefore, taking advantage of RNA-seq's potential to detect AS, I adopted an isoform-fraction (IF) guided strategy to identify changes of AS in DDtis.

First, DDtis displayed the highest use of AS compared to its matched internal control DDfat and CTtis. As a result, we found most of the genes to show AS, generating a large amount of differentially expressed isoforms between DDtis and two control tissue (DDfat or CTtis) in the Cuffdiff result.

Second, a small set of 30 isoforms displayed isoform-switching events in DDtis (compared to either DDfat or CTtis), suggesting potential differential functions of the isoforms from the same gene or an altered regulation mechanism of AS. Genes exhibiting isoform switching in DDtis were found to be associated with fibrosis progression, such as acquired mesenchymal features (*CD44v* to *CD44s*[157]), cellular abnormalities (Increased isoform ratio of *FBLN2*-Δexon9[158,159]), and contraction and organization of stress fibers (increased *TPM1-028*[160]).

Third, among all collagen genes, only *COL1A2* and *COL3A1* displayed extensive increase in the number of isoforms, the majority of which were generated by exon-skipping events. Both COL1A2 and COL3A1 are fibrillar proteins[161], which provide tensile strength and stiffness to the ECM[161]. Distinct *COL1A2*/*COL3A1* isoforms may lead to expression of protein isoforms with different structural and functional properties. Recently, AS of *COL1A2* was associated with the malignancy of colon cancer progression[162], in which ECM stiffness clearly plays a causative role to tumor formation and progression. Therefore, it is likely that increased AS and expression of the fibrillar collagen genes, *COL1A2* and *COL3A1* can contribute to the enhanced ECM stiffness and cellular growth in DDtis.

Thus, the observed fibrosis tissue specific AS may not only represent a consequence of cellular/extracellular abnormalities in DDtis, but also underline the acquisition of a myofibroblast phenotype and contractile function[163,164].

Moreover, in the expression profiling study, the GO term 'collagen fibril formation' and KEGG 'PI3K-Akt pathway' was constantly overrepresented when DDtis was compared to controls (DDfat and CTtis), suggesting increased ECM stiffness and activated PI3K-Akt as two specific features related to DDtis. Recent studies proposed that the PI3K-Akt pathway is a key regulator linking the ECM stiffness and AS changes[165]. Future experiments on possible effects of both ECM stiffness and the PI3K-AKT pathway on DD-specific AS are needed.

Taken together, this study provides a first insight into altered AS and the possible physiological pathways involved in AS regulation in DD tissues.

4.4 Therapeutic potential

The expression profiling study suggests an activation of the YAP/TAZ mechanotransduction pathway in DDfat and DDtis compared to CTtis, which is likely to contribute to the activation of adipo-fibrogenesis and continuous induction of the TGFβ fibrosis pathway. YAP/TAZ pathway is considered as a reader of ECM stiffness. Recently Dupont et al. have proposed a promising strategy to inactivate YAP/TAZ activity by culturing cells on soft matrices, which limited YAP/TAZ activity in cells and reduced growth[140]. Therefore, targeting the YAP/TAZ pathway represents a promising mechano-interference strategy to block the link between ECM stiffness and pro-fibro induction and prevent fibrosis progression.

The AS analysis suggests a fibrosis-specific feature related to DD tissue. This immediately incites to explore the potential to target AS in DD tissues. Therapeutic strategies can be designed to target the DD specific isoforms (for example, the dominant *CD44*s isoform expressed in DDtis and DDfat), the splicing factor expression or to use antisense-mediated splicing modulation[166].

In addition, the distinct AS usage of *COL3A1* and distinct expression profiling within DD samples supports the presence of DD subgroups (DD subgroup 1 and 2) in the DD cohort used for RNA-seq. Stratification of patients into more homogeneous subgroups, for example using *COL3A1* AS pattern, are likely to improve therapeutic efficacy of certain pharmaceutical agents. For example, DD tissues in DD subgroup 1 exhibited enhanced *COL3A1* AS usage. By GO term and pathway overrepresentation analysis on DEGs in DDtis from subgroup 1 compared to subgroup 2, increased ECM stiffness and overrepresented PI3-Akt signaling pathway were suggested. A recent study suggested that in the presence of cell contractility, the ECM stiffness could regulate AS via PI3K-AKT pathway[165]. If this mechanism proves to be true in DDtis, modulating PI3K-Akt activity using small molecules might represent an effective treatment of DD for DD subgroup 1 patients.

4.5 Concluding remarks and a model of DD development

DD is a common complex disease with a major genetic component (~ 80%). Large pedigrees even suggest the existence of monogenic forms with an autosomal dominant mode of inheritance[21]. DD is typically observed in adults with a manifestation age of 55-64[97]. It is rarely seen in adolescents. The late-onset manifestation of DD suggests that the effects of genetic predispositions develop over time or interact with enviormental factors[167]. Consequently, it is difficult to ascertain the causal variants contributing to DD. Additionally, variant identification captures only one layer of the genomic features of DD. Thus there is a need to integrate genomic approaches to fully characterize the multi-layered genomic features of the DD-associated traits.

This study investigated the genetic architecture and molecular basis of DD at different levels including a GWAS locus, genome wide rare variants and candidate genes, gene expression and alternative splicing as well as pathways.

First, as a follow-up of GWAS, targeted NGS of the most significant locus at 7p14.1 revealed two GWAS risk haplotype-related variants, a rare coding SNV and a common eQTL candidate, both in *EPDR1.* This suggests the observed association on the short arm of chromosome 7 may represent both rare and common variants in *EPDR1*, a verified functional candidate gene involved in ECM-receptor interactions, which contributes to the contractile phenotype of DD primary cells[123].

Second, by whole exome sequencing in combination with expression analysis, 6 functional candidate genes carrying rare variants were prioritized and they may represent genes underlying the genetic vunerability to develop DD.

Third, by whole transcriptome analysis, the potential key mechanisms in DD pathogenesis were revealed, which included mechanotransdution pathways (in particular the Hippo signaling pathway) that provide a profibrotic microenvironment, followed by induction of major fibrogenic mediators (enhanced ECM-receptor interactions and alternative splicing) that induce fibroblasts to acquire a fibrotic phenotype and promote DD progression. In addition, common pathways were shared among different levels suggesting a genetic network involving interactions among the Hippo network (including the YAP/TAZ pathway, WNT/β-catenin pathway, TGFβ pathway), the PI3K-Akt pathway and ECM-receptor interactions.

Taken together, here I proposed a preliminary model of DD development. The myofibroblasts in DDtis are potentially derived from the palmar fat tissue (DDfat). The adipocyte-fibroblast differentiation is first induced by the Hippo network to establish a profibrotic niche. The niche components further induce fibrotic factors (including enhanced ECM-receptor interactions and alternative splicing etc.) to mediate the fibroblast-myofibroblast transformation. The deregulated activity of myofibroblasts increases the ECM stiffness and and exert contractile forces on the

ECM. A mechano-feedback response between myofibroblasts and their stiff ECM is continuously mediated by increased ECM-receptor interactions, the YAP/TAZ pathway and PI3K-Akt pathway. Moreover, a genetic burden on the Hippo network or ECM genes can potentially increase the risk to develop DD. Overall, this study sheds light on likely mechanistic links between the genetic predisposition and the development of DD.

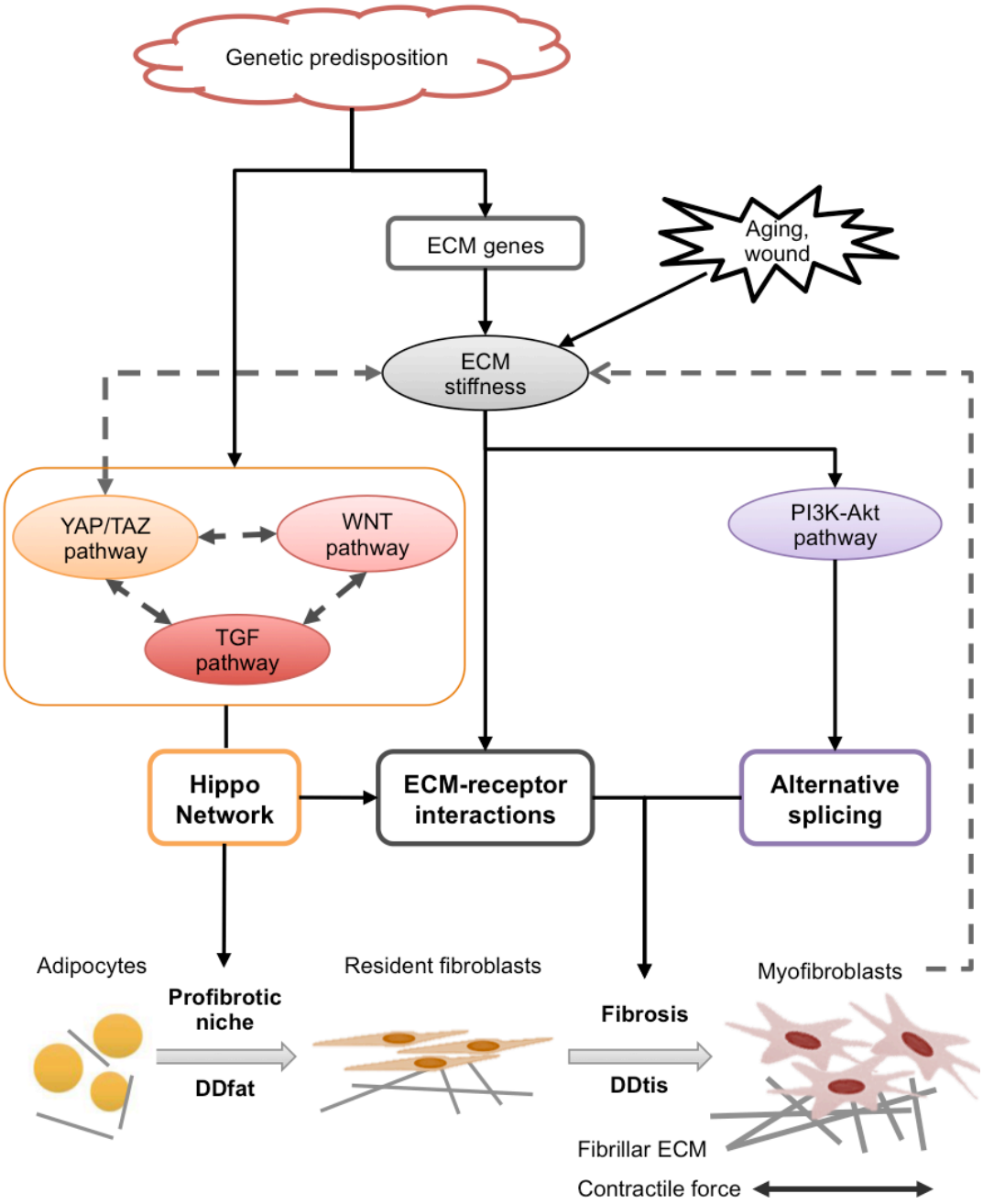


Figure 4-1 A preliminary model of DD development

In the palmar adipose tissue, the Hippo network first activates the adipocyte-fibroblast differentiation to establish a profibrotic niche. Factors released by the niche further induce fibrotic mediators (enhanced ECM-receptor interactions and alternative splicing etc.), which lead to transformation of fibroblasts into myofibroblasts with prodigious fibrillar ECM accumulation and excessive cell contraction. A mechano-feedback response between cells and their stiff ECM is mediated by enhanced ECM-receptor interactions, the YAP/TAZ pathway and PI3K-Akt pathway. A genetic predisposition in the Hippo network or ECM genes can increase the risk of developing DD.

References

1 Rodriguez-Rodero, S. *et al.* Aging genetics and aging. *Aging Dis* **2**, 186-195 (2011).

2 Brenner, P., Krause-Bergmann, A. & Van, V. H. [Dupuytren contracture in North Germany. Epidemiological study of 500 cases]. *Der Unfallchirurg* **104**, 303-311 (2001).

3 McFarlane, R. M. On the origin and spread of Dupuytren's disease. *The Journal of hand surgery* **27**, 385-390 (2002).

4 Zerajic, D. & Finsen, V. Dupuytren's disease in Bosnia and Herzegovina. An epidemiological study. *BMC musculoskeletal disorders* **5**, 10, doi:10.1186/1471-2474-5-10 (2004).

5 Saboeiro, A. P., Porkorny, J. J., Shehadi, S. I., Virgo, K. S. & Johnson, F. E. Racial distribution of Dupuytren's disease in Department of Veterans Affairs patients. *Plastic and reconstructive surgery* **106**, 71-75 (2000).

6 Shih, B. & Bayat, A. Scientific understanding and clinical management of Dupuytren disease. *Nature reviews. Rheumatology* **6**, 715-726, doi:10.1038/nrrheum.2010.180 (2010).

7 Dolmans, G. H. C. G. & Hennies, H. C. in *Dupuytren's Disease and Related Hyperproliferative Disorders: Principles, Research, and Clinical Perspectives* (eds Charles Eaton *et al.*) 87-91 (Springer Berlin Heidelberg, 2012).

8 Henry, M. Dupuytren's disease: current state of the art. *Hand* **9**, 1-8, doi:10.1007/s11552-013-9563-0 (2014).

9 Lam, W. L., Rawlins, J. M., Karoo, R. O., Naylor, I. & Sharpe, D. T. Re-visiting Luck's classification: a histological analysis of Dupuytren's disease. *The Journal of hand surgery, European volume* **35**, 312-317, doi:10.1177/1753193410362848 (2010).

10 Worrell, M. Dupuytren's Disease. *Orthopedics* **35**, 52-60, doi:10.3928/01477447-20111122-23 (2012).

11 Van De Water, L., Varney, S. & Tomasek, J. J. Mechanoregulation of the Myofibroblast in Wound Contraction, Scarring, and Fibrosis: Opportunities for New Therapeutic Intervention. *Advances in Wound Care* **2**, 122-141, doi:10.1089/wound.2012.0393 (2013).

12 Vaughan, M. B., Howard, E. W. & Tomasek, J. J. Transforming growth factor-beta1 promotes the morphological and functional differentiation of the myofibroblast. *Experimental cell research* **257**, 180-189, doi:10.1006/excr.2000.4869 (2000).

13 Tomasek, J. J., Gabbiani, G., Hinz, B., Chaponnier, C. & Brown, R. A. Myofibroblasts and mechano-regulation of connective tissue remodelling. *Nat Rev Mol Cell Biol* **3**, 349-363, doi:10.1038/nrm809 (2002).

14 Hinz, B., Dugina, V., Ballestrem, C., Wehrle-Haller, B. & Chaponnier, C. alpha-smooth muscle actin is crucial for focal adhesion maturation in myofibroblasts. *Mol Biol Cell* **14**, 2508-2519, doi:10.1091/mbc.E02-11-0729 (2003).

15 Dugina, V., Fontao, L., Chaponnier, C., Vasiliev, J. & Gabbiani, G. Focal adhesion features during myofibroblastic differentiation are controlled by intracellular and extracellular factors. *Journal of cell science* **114**, 3285-3296 (2001).

16 Shiga, K. *et al.* Cancer-Associated Fibroblasts: Their Characteristics and Their Roles in Tumor Growth. *Cancers* **7**, 2443-2458, doi:10.3390/cancers7040902 (2015).

17 Godtfredsen, N. S., Lucht, H., Prescott, E., Sorensen, T. I. A. & Gronbaek, M. A prospective study linked both alcohol and tobacco to Dupuytren's disease. *J Clin Epidemiol* **57**, 858-863, doi:10.1016/j.jclinepi.2003.11.015 (2004).

18 Gudmundsson, K. G., Arngrimsson, R., Sigfusson, N., Bjornsson, A. & Jonsson, T. Epidemiology of Dupuytren's disease - Clinical, serological, and social assessment. The Reykjavik Study. *J Clin Epidemiol* **53**, 291-296, doi:Doi 10.1016/S0895-4356(99)00145-6 (2000).

19 Dolmans, G. H. *et al.* Wnt signaling and Dupuytren's disease. *The New England journal of medicine* **365**, 307-317, doi:10.1056/NEJMoa1101029 (2011).

20 Becker, K. *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).

21 Larsen, S. *et al.* Genetic and environmental influences in Dupuytren's disease: a study of 30,330 Danish twin pairs. *The Journal of hand surgery, European volume* **40**, 171-176, doi:10.1177/1753193414535720 (2015).

22 Gibson, G. Rare and common variants: twenty arguments. *Nature reviews. Genetics* **13**, 135-145, doi:10.1038/nrg3118 (2011).

23 Altshuler, D. *et al.* A haplotype map of the human genome. *Nature* **437**, 1299-1320, doi:10.1038/nature04226 (2005).

24 Steinberg, S. *et al.* Loss-of-function variants in ABCA7 confer risk of Alzheimer's disease. *Nature genetics* **47**, 445-U424, doi:10.1038/ng.3246 (2015).

25 Consortium, G. T. The Genotype-Tissue Expression (GTEx) project. *Nature genetics* **45**, 580-585, doi:10.1038/ng.2653 (2013).

26 Keen, J. C. & Moore, H. M. The Genotype-Tissue Expression (GTEx) Project: Linking Clinical Data with Molecular Analysis to Advance Personalized Medicine. *Journal of personalized medicine* **5**, 22-29, doi:10.3390/jpm5010022 (2015).

27 Pritchard, J. K. Are rare variants responsible for susceptibility to complex diseases? *American journal of human genetics* **69**, 124-137, doi:10.1086/321272 (2001).

28 Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics* **12**, 745-755, doi:10.1038/nrg3031 (2011).

29 Yang, H., Robinson, P. N. & Wang, K. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nature methods* **12**, 841-843, doi:10.1038/nmeth.3484 (2015).

30 Kishore, A. *et al.* Association Study for 26 Candidate Loci in Idiopathic Pulmonary Fibrosis Patients from Four European Populations. *Frontiers in immunology* **7**, 274, doi:10.3389/fimmu.2016.00274 (2016).

31 Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747-753, doi:10.1038/nature08494 (2009).

32 Moutsianas, L. *et al.* The Power of Gene-Based Rare Variant Methods to Detect Disease-Associated Variation and Test Hypotheses About Complex Disease. *PLoS Genet* **11**, e1005165, doi:10.1371/journal.pgen.1005165 (2015).

33 Shih, B., Watson, S. & Bayat, A. Whole genome and global expression profiling of Dupuytren's disease: systematic review of current findings and future perspectives. *Annals of the rheumatic diseases* **71**, 1440-1447, doi:10.1136/annrheumdis-2012-201295 (2012).

34 Varallo, V. M. *et al.* Beta-catenin expression in Dupuytren's disease: potential role for cell-matrix interactions in modulating beta-catenin levels in vivo and in vitro. *Oncogene* **22**, 3680-3684, doi:10.1038/sj.onc.1206415 (2003).

35 Bridgewater, D. *et al.* beta-catenin causes renal dysplasia via upregulation of Tgfbeta2 and Dkk1. *Journal of the American Society of Nephrology : JASN* **22**, 718-731, doi:10.1681/ASN.2010050562 (2011).

36 Edlund, S. *et al.* Interaction between smad7 and beta-catenin: Importance for transforming growth factor beta-induced apoptosis. *Mol Cell Biol* **25**, 1475-1488, doi:10.1128/MCB.25.4.1475-1488.2005 (2005).

37 Liu, F. Y., Li, X. Z., Peng, Y. M., Liu, H. & Liu, Y. H. Arkadia-Smad7-mediated positive regulation of TGF-beta signaling in a rat model of tubulointerstitial fibrosis. *American journal of nephrology* **27**, 176-183, doi:10.1159/000100518 (2007).

38 Meng, X. M., Nikolic-Paterson, D. J. & Lan, H. Y. TGF-beta: the master regulator of fibrosis. *Nature reviews. Nephrology* **12**, 325-338, doi:10.1038/nrneph.2016.48 (2016).

39 Rehman, S. *et al.* Molecular phenotypic descriptors of Dupuytren's disease defined using informatics analysis of the transcriptome. *The Journal of hand surgery* **33**, 359-372, doi:10.1016/j.jhsa.2007.11.010 (2008).

40 Johnston, P., Larson, D., Clark, I. M. & Chojnowski, A. J. Metalloproteinase gene expression correlates with clinical outcome in Dupuytren's disease. *The Journal of hand surgery* **33**, 1160-1167, doi:10.1016/j.jhsa.2008.04.002 (2008).

41 Shih, B., Brown, J. J., Armstrong, D. J., Lindau, T. & Bayat, A. Differential gene expression analysis of subcutaneous fat, fascia, and skin overlying a Dupuytren's disease nodule in comparison to control tissue. *Hand* **4**, 294-301, doi:10.1007/s11552-009-9164-0 (2009).

42 Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols* **7**, 562-578, doi:10.1038/nprot.2012.016 (2012).

43 Sakharkar, M. K., Chow, V. T. & Kangueane, P. Distributions of exons and introns in the human genome. *In silico biology* **4**, 387-393 (2004).

44 Will, C. L. & Luhrmann, R. Spliceosome structure and function. *Cold Spring Harbor perspectives in biology* **3**, doi:10.1101/cshperspect.a003707 (2011).

45 Lee, Y. & Rio, D. C. Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annu Rev Biochem* **84**, 291-323, doi:10.1146/annurev-biochem-060614-034316 (2015).

46 Vitting-Seerup, K., Porse, B. T., Sandelin, A. & Waage, J. spliceR: an R package for classification of alternative splicing and prediction of coding potential from RNA-seq data. *BMC bioinformatics* **15**, 81, doi:10.1186/1471-2105-15-81 (2014).

47 Alekseyenko, A. V., Kim, N. & Lee, C. J. Global analysis of exon creation versus loss and the role of alternative splicing in 17 vertebrate genomes. *Rna* **13**, 661-670, doi:10.1261/rna.325107 (2007).

48 Sammeth, M., Foissac, S. & Guigo, R. A General Definition and Nomenclature for Alternative Splicing Events. *Plos Comput Biol* **4**, doi:ARTN e100014710.1371/journal.pcbi.1000147 (2008).

49 Bebee, T. W. *et al.* The splicing regulators Esrp1 and Esrp2 direct an epithelial splicing program essential for mammalian development. *Elife* **4**, doi:10.7554/eLife.08954 (2015).

50 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).

51 McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome biology* **17**, 122, doi:10.1186/s13059-016-0974-4 (2016).

52 Ovchinnikov, S. *et al.* Large-scale determination of previously unsolved protein structures using evolutionary information. *Elife* **4**, doi:ARTN e0924810.7554/eLife.09248 (2015).

53 Cartharius, K. *et al.* MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics* **21**, 2933-2942, doi:10.1093/bioinformatics/bti473 (2005).

54 Matys, V. *et al.* TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* **31**, 374-378 (2003).

55 Wasserman, W. W. & Sandelin, A. Applied bioinformatics for the identification of regulatory elements. *Nature reviews. Genetics* **5**, 276-287, doi:10.1038/nrg1315 (2004).

56 Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* **43**, 11 10 11-33, doi:10.1002/0471250953.bi1110s43 (2013).

57 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**, 491-+, doi:10.1038/ng.806 (2011).

58 Wang, K., Li, M. Y. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, doi:ARTN e16410.1093/nar/gkq603 (2010).

59 Ng, P. C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome research* **11**, 863-874, doi:Doi 10.1101/Gr.176601 (2001).

60 Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics* **14**, 178-192, doi:10.1093/bib/bbs017 (2013).

61 Diaz-Montana, J. J., Rackham, O. J., Diaz-Diaz, N. & Petretto, E. Web-based Gene Pathogenicity Analysis (WGPA): a web platform to interpret gene pathogenicity from personal genome data. *Bioinformatics* **32**, 635-637, doi:10.1093/bioinformatics/btv598 (2016).

62 Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545-15550, doi:10.1073/pnas.0506580102 (2005).

63 Rackham, O. J., Shihab, H. A., Johnson, M. R. & Petretto, E. EvoTol: a protein-sequence based evolutionary intolerance framework for disease-gene prioritization. *Nucleic Acids Res* **43**, e33, doi:10.1093/nar/gku1322 (2015).

64 Gene Ontology, C. The Gene Ontology project in 2008. *Nucleic Acids Res* **36**, D440-444, doi:10.1093/nar/gkm883 (2008).

65 Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* **44**, W90-97, doi:10.1093/nar/gkw377 (2016).

66 Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* **44**, D457-D462, doi:10.1093/nar/gkv1070 (2016).

67 Draghici, S. *et al.* A systems biology approach for pathway level analysis. *Genome research* **17**, 1537-1545, doi:10.1101/gr.6202607 (2007).

68 Donato, M. *et al.* Analysis and correction of crosstalk effects in pathway analysis. *Genome research* **23**, 1885-1893, doi:10.1101/gr.153551.112 (2013).

69 Alexa, A., Rahnenfuhrer, J. & Lengauer, T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* **22**, 1600-1607, doi:10.1093/bioinformatics/btl140 (2006).

70 Katz, Y. *et al.* Quantitative visualization of alternative exon expression from RNA-seq data. *Bioinformatics* **31**, 2400-2402, doi:10.1093/bioinformatics/btv034 (2015).

71 Cheadle, C., Vawter, M. P., Freed, W. J. & Becker, K. G. Analysis of microarray data using Z score transformation. *J Mol Diagn* **5**, 73-81, doi:Doi 10.1016/S1525-1578(10)60455-2 (2003).

72 Becker, K. Molecular Genetics of Dupuytren's Disease. *University of Cologne PhD Thesis* (2012).

73 Eilbeck, K. *et al.* The Sequence Ontology: a tool for the unification of genome annotations. *Genome biology* **6**, R44, doi:10.1186/gb-2005-6-5-r44 (2005).

74 Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).

75 Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291, doi:10.1038/nature19057 (2016).

76 Schwarz, J. M., Rodelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nature methods* **7**, 575-576, doi:10.1038/nmeth0810-575 (2010).

77 Shihab, H. A. *et al.* Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum Mutat* **34**, 57-65, doi:10.1002/humu.22225 (2013).

78 Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics* **46**, 310-315, doi:10.1038/ng.2892 (2014).

79 Quang, D., Chen, Y. & Xie, X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **31**, 761-763, doi:10.1093/bioinformatics/btu703 (2015).

80 Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology* **157**, 105-132 (1982).

81 Betts, M. J. & Russell, R. B. in *Bioinformatics for Geneticists* 289-316 (John Wiley & Sons, Ltd, 2003).

82 Balakrishnan, S., Kamisetty, H., Carbonell, J. G., Lee, S. I. & Langmead, C. J. Learning generative models for protein fold families. *Proteins* **79**, 1061-1078, doi:10.1002/prot.22934 (2011).

83 Marcelino, A. M. & Gierasch, L. M. Roles of beta-turns in protein folding: from peptide models to protein engineering. *Biopolymers* **89**, 380-391, doi:10.1002/bip.20960 (2008).

84 Hoffmann, W. Ependymins and their potential role in neuroplasticity and regeneration: Calcium-binding meningeal glycoproteins of the cerebrospinal fluid and extracellular matrix. *International Journal of Biochemistry* **26**, 607-619, doi:http://dx.doi.org/10.1016/0020-711X(94)90160-0 (1994).

85 Stergachis, A. B. et al. Exonic transcription factor binding directs codon choice and affects protein evolution. *Science* **342**, 1367-1372, doi:10.1126/science.1243490 (2013).

86 Atchley, W. R. & Fitch, W. M. A natural classification of the basic helix–loop–helix class of transcription factors. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 5172-5176 (1997).

87 Mermod, N., Williams, T. J. & Tjian, R. Enhancer binding factors AP-4 and AP-1 act in concert to activate SV40 late transcription in vitro. *Nature* **332**, 557-561, doi:10.1038/332557a0 (1988).

88 Hu, Y. F., Luscher B Fau - Admon, A., Admon A Fau - Mermod, N., Mermod N Fau - Tjian, R. & Tjian, R. Transcription factor AP-4 contains multiple dimerization domains that regulate dimer specificity. *Genes & development* (1990).

89 Egawa, T. & Littman, D. R. Transcription factor AP4 modulates reversible and epigenetic silencing of the Cd4 gene. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 14873-14878, doi:10.1073/pnas.1112293108 (2011).

90 Imai, K. & Okamoto, T. Transcriptional repression of human immunodeficiency virus type 1 by AP-4. *The Journal of biological chemistry* **281**, 12495-12505, doi:10.1074/jbc.M511773200 (2006).

91 Ronaghi, M., Uhlen, M. & Nyren, P. A sequencing method based on real-time pyrophosphate. *Science* **281**, 363, 365 (1998).

92 Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* **40**, D930-934, doi:10.1093/nar/gkr917 (2012).

93 Bonnefond, A. *et al.* Rare MTNR1B variants impairing melatonin receptor 1B function contribute to type 2 diabetes. *Nature genetics* **44**, 297-301, doi:10.1038/ng.1053 (2012).

94 Peltonen, L., Perola, M., Naukkarinen, J. & Palotie, A. Lessons from studying monogenic disease for common disease. *Hum Mol Genet* **15 Spec No 1**, R67-74, doi:10.1093/hmg/ddl060 (2006).

95 Kazma, R. & Bailey, J. N. Population-based and family-based designs to analyze rare variants in complex diseases. *Genetic epidemiology* **35 Suppl 1**, S41-47, doi:10.1002/gepi.20648 (2011).

96 Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *American journal of human genetics* **95**, 5-23, doi:10.1016/j.ajhg.2014.06.009 (2014).

97 Ross, D. C. Epidemiology of Dupuytren's disease. *Hand clinics* **15**, 53-62, vi (1999).

98 Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research* **15**, 1034-1050, doi:10.1101/gr.3715005 (2005).

99 Kelley, D. R. & Salzberg, S. L. Detection and correction of false segmental duplications caused by genome mis-assembly. *Genome biology* **11**, R28, doi:10.1186/gb-2010-11-3-r28 (2010).

100 Robinson, P. N. & Mundlos, S. The human phenotype ontology. *Clinical genetics* **77**, 525-534, doi:10.1111/j.1399-0004.2010.01436.x (2010).

101 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* **25**, 25-29, doi:10.1038/75556 (2000).

102 DeFilippis, R. A. *et al.* CD36 repression activates a multicellular stromal program shared by high mammographic density and tumor tissues. *Cancer discovery* **2**, 826-839, doi:10.1158/2159-8290.CD-12-0107 (2012).

103 Cieply, B. & Carstens, R. P. Functional roles of alternative splicing factors in human disease. *Wiley interdisciplinary reviews. RNA* **6**, 311-326, doi:10.1002/wrna.1276 (2015).

104 Piva, F., Giulietti, M., Burini, A. B. & Principato, G. SpliceAid 2: a database of human splicing factors expression data and RNA target motifs. *Hum Mutat* **33**, 81-85, doi:10.1002/humu.21609 (2012).

105 Zhang, C. *et al.* Defining the regulatory network of the tissue-specific splicing factors Fox-1 and Fox-2. *Genes & development* **22**, 2550-2563, doi:10.1101/gad.1703108 (2008).

106 van der Slot, A. J. *et al.* Identification of PLOD2 as telopeptide lysyl hydroxylase, an important enzyme in fibrosis. *Journal of Biological Chemistry* **278**, 40967-40972, doi:10.1074/jbc.M307380200 (2003).

107 Remst, D. F. G. *et al.* Osteoarthritis-related fibrosis is associated with both elevated pyridinoline cross-link formation and lysyl hydroxylase 2b expression. *Osteoarthr Cartilage* **21**, 157-164, doi:10.1016/j.joca.2012.10.002 (2013).

108 Warzecha, C. C. & Carstens, R. P. Complex changes in alternative pre-mRNA splicing play a central role in the epithelial-to-mesenchymal transition (EMT). *Semin Cancer Biol* **22**, 417-427, doi:10.1016/j.semcancer.2012.04.003 (2012).

109 Terpe, H. J., Stark, H., Prehm, P. & Gunthert, U. CD44 variant isoforms are preferentially expressed in basal epithelial of non-malignant human fetal and adult tissues. *Histochemistry* **101**, 79-89 (1994).

110 Timpl, R., Sasaki, T., Kostka, G. & Chu, M. L. Fibulins: a versatile family of extracellular matrix proteins. *Nat Rev Mol Cell Biol* **4**, 479-489, doi:10.1038/nrm1130 (2003).

111 Dias, J. J., Singh, H. P., Ullah, A., Bhowal, B. & Thompson, J. R. Patterns of recontracture after surgical correction of Dupuytren disease. *The Journal of hand surgery* **38**, 1987-1993, doi:10.1016/j.jhsa.2013.05.038 (2013).

112 Peimer, C. A. *et al.* Dupuytren Contracture Recurrence Following Treatment With Collagenase Clostridium histolyticum (CORDLESS [Collagenase Option for Reduction of Dupuytren Long-Term Evaluation of Safety Study]): 5-Year Data. *The Journal of hand surgery* **40**, 1597-1605, doi:10.1016/j.jhsa.2015.04.036 (2015).

113 Li, C. G. *et al.* Ror2 modulates the canonical Wnt signaling in lung epithelial cells through cooperation with Fzd2. *Bmc Mol Biol* **9**, doi:Artn 1110.1186/1471-2199-9-11 (2008).

114 Billiard, J. *et al.* The orphan receptor tyrosine kinase Ror2 modulates canonical Wnt signaling in osteoblastic cells. *Mol Endocrinol* **19**, 90-101, doi:10.1210/me.2004-0153 (2005).

115 Oishi, I. *et al.* The receptor tyrosine kinase Ror2 is involved in non-canonical Wnt5a/JNK signalling pathway. *Genes Cells* **8**, 645-654, doi:Doi 10.1046/J.1365-2443.2003.00662.X (2003).

116 Grumolato, L. *et al.* Canonical and noncanonical Wnts use a common mechanism to activate completely unrelated coreceptors. *Genes & development* **24**, 2517-2530, doi:10.1101/gad.1957710 (2010).

117 Kay, B. K., Williamson, M. P. & Sudol, P. The importance of being proline: the interaction of proline-rich motifs in signaling proteins with their cognate domains. *Faseb Journal* **14**, 231-241 (2000).

118 Shashoua, V. E. Ependymin, a brain extracellular glycoprotein, and CNS plasticity. *Annals of the New York Academy of Sciences* **627**, 94-114 (1991).

119 Geer, L. Y. *et al.* The NCBI BioSystems database. *Nucleic Acids Res* **38**, D492-496, doi:10.1093/nar/gkp858 (2010).

120 Nimmrich, I. *et al.* The novel ependymin related gene UCC1 is highly expressed in colorectal tumor cells. *Cancer Lett* **165**, 71-79, doi:Doi 10.1016/S0304-3835(01)00390-1 (2001).

121 Tang, J. *et al.* Cancer driver-passenger distinction via sporadic human and dog cancer comparison: a proof-of-principle study with colorectal cancer. *Oncogene* **33**, 814-822, doi:10.1038/onc.2013.17 (2014).

122 Wu, P. *et al.* A gene expression based predictor for high risk myeloma treated with intensive therapy and autologous stem cell rescue. *Leukemia & lymphoma* **56**, 594-601, doi:10.3109/10428194.2014.911863 (2015).

123 Staats, K. A., Wu, T., Gan, B. S., O'Gorman, D. B. & Ophoff, R. A. Dupuytren's disease susceptibility gene, EPDR1, is involved in myofibroblast contractility. *J Dermatol Sci* **83**, 131-137, doi:10.1016/j.jdermsci.2016.04.015 (2016).

124 Kadler, K. E., Baldock, C., Bella, J. & Boot-Handford, R. P. Collagens at a glance. *Journal of cell science* **120**, 1955-1958, doi:10.1242/jcs.03453 (2007).

125 Bruckner-Tuderman, L., Hopfner, B. & Hammami-Hauasli, N. Biology of anchoring fibrils: lessons from dystrophic epidermolysis bullosa. *Matrix Biol* **18**, 43-54, doi:Doi 10.1016/S0945-053x(98)00007-9 (1999).

126 Avouac, J. *et al.* The Nuclear Receptor Constitutive Androstane Receptor/NR1I3 Enhances the Profibrotic Effects of Transforming Growth Factor beta and Contributes to the Development of Experimental Dermal Fibrosis. *Arthritis & rheumatology* **66**, 3140-3150, doi:10.1002/art.38819 (2014).

127 Rozen-Zvi, B. *et al.* TGF-beta/Smad3 activates mammalian target of rapamycin complex-1 to promote collagen production by increasing HIF-1 alpha expression. *Am J Physiol-Renal* **305**, F485-F494, doi:10.1152/ajprenal.00215.2013 (2013).

128 Remst, D. F. G. *et al.* Gene Expression Analysis of Murine and Human Osteoarthritis Synovium Reveals Elevation of Transforming Growth Factor beta-Responsive Genes in Osteoarthritis-Related Fibrosis. *Arthritis & rheumatology* **66**, 647-656, doi:10.1002/art.38266 (2014).

129 Zhao, J. S. *et al.* Smad3 deficiency attenuates bleomycin-induced pulmonary fibrosis in mice. *Am J Physiol-Lung C* **282**, L585-L593 (2002).

130 Zhou, X. D., Xiong, M. M., Tan, F. K., Guo, X. J. & Arnett, F. C. SPARC, an upstream regulator of connective tissue growth factor in response to transforming growth factor ss stimulation. *Arthritis and rheumatism* **54**, 3885-3889, doi:10.1002/art.22249 (2006).

131 Vazquez-Villa, F. *et al.* COL11A1/(pro)collagen 11A1 expression is a remarkable biomarker of human invasive carcinoma-associated stromal cells and carcinoma progression. *Tumor Biol* **36**, 2213-2222, doi:10.1007/s13277-015-3295-4 (2015).

132 Hu, J. *et al.* Formin 1 and filamin B physically interact to coordinate chondrocyte proliferation and differentiation in the growth plate. *Hum Mol Genet* **23**, 4663-4673, doi:10.1093/hmg/ddu186 (2014).

133 Zieba, J. *et al.* TGF beta and BMP Dependent Cell Fate Changes Due to Loss of Filamin B Produces Disc Degeneration and Progressive Vertebral Fusions. *Plos Genetics* **12**, doi:ARTN e100593610.1371/journal.pgen.1005936 (2016).

134 Masiakowski, P. & Carroll, R. D. A Novel Family of Cell-Surface Receptors with Tyrosine Kinase-Like Domain. *Journal of Biological Chemistry* **267**, 26181-26190 (1992).

135 Holley, R. J. *et al.* Comparative Quantification of the Surfaceome of Human Multipotent Mesenchymal Progenitor Cells. *Stem Cell Rep* **4**, 473-488, doi:10.1016/j.stemcr.2015.01.007 (2015).

136 Rasmussen, N. R. *et al.* Receptor Tyrosine Kinase-like Orphan Receptor 2 (Ror2) Expression Creates a Poised State of Wnt Signaling in Renal Cancer. *Journal of Biological Chemistry* **288**, 26301-26310, doi:10.1074/jbc.M113.466086 (2013).

137 Akhmetshina, A. *et al.* Activation of canonical Wnt signalling is required for TGF-beta-mediated fibrosis. *Nat Commun* **3**, doi:Artn 73510.1038/Ncomms1734 (2012).

138 Hu, F. Z. *et al.* Mapping of an autosomal dominant gene for Dupuytren's contracture to chromosome 16q in a Swedish family. *Clinical genetics* **68**, 424-429, doi:10.1111/j.1399-0004.2005.00504.x (2005).

139 Zhao, B., Lei, Q. Y. & Guan, K. L. The Hippo-YAP pathway: new connections between regulation of organ size and cancer. *Curr Opin Cell Biol* **20**, 638-646, doi:10.1016/j.ceb.2008.10.001 (2008).

140 Dupont, S. *et al.* Role of YAP/TAZ in mechanotransduction. *Nature* **474**, 179-U212, doi:10.1038/nature10137 (2011).

141 Humphrey, J. D., Dufresne, E. R. & Schwartz, M. A. Mechanotransduction and extracellular matrix homeostasis. *Nat Rev Mol Cell Biol* **15**, 802-812, doi:10.1038/nrm3896 (2014).

142 Young, D. A., Choi, Y. S., Engler, A. J. & Christman, K. L. Stimulation of adipogenesis of adult adipose-derived stem cells using substrates that mimic the stiffness of adipose tissue. *Biomaterials* **34**, 8581-8588, doi:10.1016/j.biomaterials.2013.07.103 (2013).

143 Wells, R. G. The role of matrix stiffness in regulating cell behavior. *Hepatology* **47**, 1394-1400, doi:10.1002/hep.22193 (2008).

144 Liu, F. *et al.* Feedback amplification of fibrosis through matrix stiffening and COX-2 suppression. *Journal of Cell Biology* **190**, 693-706, doi:10.1083/jcb.201004082 (2010).

145 Abdennour, M. *et al.* Association of adipose tissue and liver fibrosis with tissue stiffness in morbid obesity: links with diabetes and BxMI loss after gastric bypass. *The Journal of clinical endocrinology and metabolism* **99**, 898-907, doi:10.1210/jc.2013-3253 (2014).

146 Yu, F. X. & Guan, K. L. The Hippo pathway: regulators and regulations. *Genes & development* **27**, 355-371, doi:10.1101/gad.210773.112 (2013).

147 Chandler, E. M. *et al.* Implanted adipose progenitor cells as physicochemical regulators of breast cancer. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 9786-9791, doi:10.1073/pnas.1121160109 (2012).

148 Hiemer, S. E., Szymaniak, A. D. & Varelas, X. The transcriptional regulators TAZ and YAP direct transforming growth factor beta-induced tumorigenic phenotypes in breast cancer cells. *The Journal of biological chemistry* **289**, 13461-13474, doi:10.1074/jbc.M113.529115 (2014).

149 Lee, D. H. *et al.* LATS-YAP/TAZ controls lineage specification by regulating TGFbeta signaling and Hnf4alpha expression during liver development. *Nat Commun* **7**, 11961, doi:10.1038/ncomms11961 (2016).

150 Xu, Q., Modrek, B. & Lee, C. Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res* **30**, 3754-3766 (2002).

151 Shi, J. *et al.* Cyclic AMP-dependent protein kinase regulates the alternative splicing of tau exon 10: a mechanism involved in tau pathology of Alzheimer disease. *The Journal of biological chemistry* **286**, 14639-14648, doi:10.1074/jbc.M110.204453 (2011).

152 Zhu, H. & Ding, Q. Lower expression level of two RAGE alternative splicing isoforms in Alzheimer's disease. *Neuroscience letters* **597**, 66-70, doi:10.1016/j.neulet.2015.04.032 (2015).

153 Fu, R. H. *et al.* Aberrant alternative splicing events in Parkinson's disease. *Cell transplantation* **22**, 653-661, doi:10.3727/096368912X655154 (2013).

154 Beck, S. *et al.* Cystic fibrosis patients with the 3272-26A-->G mutation have mild disease, leaky alternative mRNA splicing, and CFTR protein at the cell membrane. *Hum Mutat* **14**, 133-144, doi:10.1002/(SICI)1098-1004(1999)14:2<133::AID-HUMU5>3.0.CO;2-T (1999).

155 Alsafadi, S. *et al.* Cancer-associated SF3B1 mutations affect alternative splicing by promoting alternative branchpoint usage. *Nat Commun* **7**, 10615, doi:10.1038/ncomms10615 (2016).

156 Sumithra, B., Saxena, U. & Das, A. B. Alternative splicing within the Wnt signaling pathway: role in cancer development. *Cellular oncology* **39**, 1-13, doi:10.1007/s13402-015-0266-0 (2016).

157 Brown, R. L. *et al.* CD44 splice isoform switching in human and mouse epithelium is essential for epithelial-mesenchymal transition and breast cancer progression. *Journal of Clinical Investigation* **121**, 1064-1074, doi:10.1172/JCI44540 (2011).

158 Danan-Gotthold, M. *et al.* Identification of recurrent regulated alternative splicing events across human solid tumors. *Nucleic Acids Res* **43**, 5130-5144, doi:10.1093/nar/gkv210 (2015).

159 Gu, Y. C., Nilsson, K., Eng, H. & Ekblom, M. Association of extracellular matrix proteins fibulin-1 and fibulin-2 with fibronectin in bone marrow stroma. *Brit J Haematol* **109**, 305-313, doi:Doi 10.1046/J.1365-2141.2000.02011.X (2000).

160 Gunning, P. W., Schevzov, G., Kee, A. J. & Hardeman, E. C. Tropomyosin isoforms: divining rods for actin cytoskeleton function. *Trends Cell Biol* **15**, 333-341, doi:10.1016/j.tcb.2005.04.007 (2005).

161 Bonnans, C., Chou, J. & Werb, Z. Remodelling the extracellular matrix in development and disease. *Nat Rev Mol Cell Biol* **15**, 786-801, doi:10.1038/nrm3904 (2014).

162 Bisognin, A. *et al.* An integrative framework identifies alternative splicing events in colorectal cancer development. *Molecular oncology* **8**, 129-141, doi:10.1016/j.molonc.2013.10.004 (2014).

163 Llorian, M. *et al.* The alternative splicing program of differentiated smooth muscle cells involves concerted non-productive splicing of post-transcriptional regulators. *Nucleic Acids Res*, doi:10.1093/nar/gkw560 (2016).

164 Davis, J. *et al.* MBNL1-mediated regulation of differentiation RNAs promotes myofibroblast transformation and the fibrotic response. *Nat Commun* **6**, 10084, doi:10.1038/ncomms10084 (2015).

165 Bordeleau, F. *et al.* Tissue stiffness regulates serine/arginine-rich protein-mediated splicing of the extra domain B-fibronectin isoform in tumors. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 8314-8319, doi:10.1073/pnas.1505421112 (2015).

166 Arechavala-Gomeza, V., Khoo, B. & Aartsma-Rus, A. Splicing modulation therapy in the treatment of genetic diseases. *The application of clinical genetics* **7**, 245-252, doi:10.2147/TACG.S71506 (2014).

167 Salvatore, J. E. *et al.* Polygenic risk for externalizing disorders: Gene-by-development and gene-by-environment effects in adolescents and young adults. *Clinical psychological science : a journal of the Association for Psychological Science* **3**, 189-201, doi:10.1177/2167702614534211 (2015).

Resources

1k Genomes	www.1000genomes.org
ANNOVAR	www.openbioinformatics.org/annovar
CADD	cadd.gs.washington.edu
Ensembl	grch37.ensembl.org/index.html
ExAC	exac.broadinstitute.org_
GATK	www.broadinstitute.org/gatk
Hapmap	ftp://ftp.ncbi.nlm.nih.gov/hapmap
HPO	human-phenotype-ontology.github.io
IGV	software.broadinstitute.org/software/igv
MutationTaster	www.mutationtaster.org
NHLBI ESP	evs.gs.washington.edu/EVS
Phenolyzer	phenolyzer.usc.edu
PolyPhen-2	genetics.bwh.harvard.edu/pph2
SIFT	sift.jcvi.org
Splice R	bioconductor.org/packages/release/bioc/html/spliceR.html
SpliceAid2	193.206.120.249/splicing_tissue.html
UniProt	www.uniprot.org/uniprot
Varbank	varbank.ccg.uni-koeln.de
WGPA	wgpa.systems-genetics.net

Declaration

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit − einschließlich Tabellen, Karten und Abbildungen −, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie − abgesehen von unten angegebenen Teilpublikationen − noch nicht veröffentlicht worden ist, sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde.

Die Bestimmungen der Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von **Prof. Dr. Peter Nürnberg** betreut worden.

Köln, den 14.11.2016

_____

Juanjiangmeng Du

TeilPublikationen:

Becker, K., **Du**, J., Nürnberg, P. & Hennies, H. C. in Dupuytren Disease and Related Diseases - The Cutting Edge (eds Paul M. N. Werker et al.) 105-111 (Springer International Publishing, 2017).

Becker, K., Siegert, S., Toliat, M., R. **Du, J.** 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).

Acknowledgement

Three years time is like a flying arrow. In October 2013, I met Prof. Peter Nürnberg for the first time in the CCG seminar given by Prof. Dennis Lo. That was exactly the moment I decided to have my PhD in genetics/genomics field. Thanks to Peter, I was able to join CCG later and started my exciting PhD journey.

As the director of CCG, Peter takes a lot of responsibilities and works more than 12 hours a day. But he never complains about his workload, instead, he always tries his best to help other people. Working with Peter made me realize what integrity, leadership and wisdom are. As my PhD supervisor, Peter has also given me professional guidance and supported my personal development. Thanks, Peter, for helping me tremendously during my PhD study.

I would also like to express my sincere thanks to my other thesis committee members, Prof. Angelika A. Noegel and Prof. Wolfgang Werr. Prof. Noegel has helped me since the beginning of my projects. She has been actively involved in advising me on my work. My special thanks go to Prof. Werr who has been so kind to be in my committee and willing to provide expertise input from the developmental biological point of view.

I'm grateful to Dr. Hans Christian Hennies who has given much help on my project especially in biopsy collection, laboratory work and thesis proofreading. I also benefited a lot from his expertise in the area of connective tissue disease. I'm particularly thankful to Prof. Michael Nothnagel who has helped me greatly on statistical analysis and proofreading. I strongly recommend two courses I have taken from him — 'Epidemiology and statistical genetics' and 'Complex trait analysis of NGS data'. I would also like to express my heartfelt gratitude to Dr. Herbert Schulz, Lisa-Marie Neupert and Dr. Ann-Kathrin Ruppert for spending time to proofread my thesis.

The completion of my projects is a result of teamwork with my wonderful CCG colleagues. The targeted NGS was an extension of Dr. Kerstin Becker's previous GWAS project. She has been greatly helpful in my PhD especially in sample collection, Pyrosequencing and Sanger sequencing. My special thanks go to Ramona Casper for isolating blood DNA and technical support. My hearty gratitude goes to Dr. Janine Altmüller, Elisabeth Kirst, Christian Becker and Marek Franitza for their outstanding work on library preparations and NGS. My sincere appreciation goes to Dr. Holger Thiele, Dr. Amit Kawalia and Dr. Susanne Motameny for their help on genomic and exome variant calling. I'm particularly thankful to Dr. Dmitriy Drichel, who has contributed exceptionally on Cufflinks pipeline running, and Eberechi Innocentia Uwakah, who did a good job in running SpliceR package. I would also like to thank Dr. Muhammad Reza. Toliat and Nina Dalibor for their technical assistance in Pyrosequencing and Sanger sequencing; thank Wilfried Gunia and Heinrich Rohde for their numerous IT work; thank Nicole Riedel and Gabriele Thorn for their constant administrative help; and thank Dr. Kamel Jabbari, Dr. Sajid M. Hussain, Dr. Birgit Budde, Prof. Michal-Ruth Schweiger and George Kanoungi for their fruitful discussions. (My major work includes study designs and data analyses of

three projects, DNA/RNA sample preparation for targeted NGS, WES and RNA-seq, library preparation for targeted NGS, tissue/cell culture and relative *in vitro* experiments.)

I'm truly grateful to our collaborators who have contributed to my work. My profound gratitude goes to Prof. Riccardo Fodde (Erasmus University Rotterdam), who has given valuable support and suggestions on signaling pathways and functional studies. He has always been a good mentor and a good friend to me. My sincere thanks goes to Sergey Ovchinnikov (University of Washington, Seattle), a talented scientist who has shown great interest in my work and contributed on EPDR1 protein model building in this project.

I would also like to thank CECAD (The Cologne Cluster of Excellence in Cellular Stress Responses in Aging-associated Diseases) for providing funding and various training opportunities. My sincere appreciation goes to Prof. Carien Niessen, Jenny Ostermann, Dr. Doris Birker, Dr. Justin Lorek, Dr. Katharina Costa Rodrigues and Dr. Maria Vilgertshofer for their great support in the Cologne Graduate School of Aging Research. My big thanks go to Dr. Isabell Witt for her kindness and help in the Graduate School for Biological Sciences.

I'm very thankful to all patients, all surgeons (including Dr. Frank Staub, Center for Peripheral Neurosurgery, Dossenheim) and German Dupuytren Society for their participation and cooperation in this study. I hope our research can gradually contribute to the understanding of this common but rarely studied disease — Dupuytren's Disease. The application of research might take long time, but our passion and commitment in studying complex diseases always outlast.

For me, the advantage of doing a PhD was not only about intellectual development but also interactions with different people. I would like to thank Dr. Dennis Lal, who has deeply influenced me by sharing scientific knowledge and helping me with my PhD/career planning. My sincere gratitude goes to Dr. Ruth Willmott, who has been an excellent coach in scientific presentation/writing and career development. She has a great impact on me in creative thinking and soft skills training. I strongly encourage PhD students to attend her training courses. My special thanks go to my six other classmates in our graduate school for their support during my PhD study. We had so many stories and so much fun together. I'd also like to thank the CECAD PhD/PostDoc committee members who have organized activities together with me. I learned a lot from each of them. Thanks to my other friends who have encouraged and motivated me in the past years.

Finally, I would like to express my deepest gratitude to all my family members who always love me, in particular, to my parents who have always been my role models and provided a happy environment for me, and to my husband who has taken care of me in the past three years and makes every day with him worth looking forward to.

Looking back, my 3-year PhD was a journey of excitement, discovery and growth. By learning from all of you in this journey, I gradually become a better person. I do hope we keep in touch in the future and share our stories in the next 3 years and 3 decades.