13 Network Analysis and Fine-Mapping GWAS Loci to Identify Genes and Functional Variants Involved in the Development of Dupuytren Disease

Kerstin Becker, Juanjiangmeng Du, Peter Nürnberg, and Hans Christian Hennies

Contents

13.1 Introduction

Many DD patients have a positive family history, and genetic factors play a far greater role in the etiology of this disease than is often acknowledged. In 1963, Ling already showed that the rate of patients with a positive family history increased from 16 % reported by the patients themselves to 68 % when the patient's relatives were examined by the author (Ling 1963). Studies have determined varying family predisposition rates in patients, between 12.5 and 44 % (Brenner et al. 2001; Coert et al. 2006; Early 1962; Hakstian 1966; Hindocha et al. 2006a, b; Lanting et al. 2013; Makela et al. 1991). The sibling recurrence risk λ_s has been determined as 2.9 based on a prevalence of 3.5 % in north-western England (Capstick et al. 2013; Hindocha et al. 2006a). We have shown that DD patients who had a known family history for this disease are significantly younger at the time of first surgery (Becker et al. 2015). Moreover, a positive family history had a

K. Becker (✉) • J. Du • P. Nürnberg
Cologne Center for Genomics, University of Cologne, Weyertal 115b, 50931 Cologne, Germany

Cologne Excellence Cluster on Cellular Stress Responses in Aging-associated Diseases, University of Cologne, Cologne, Germany
e-mail: k.becker@uni-koeln.de; j.du@uni-koeln.de; nuernberg@uni-koeln.de

H.C. Hennies
Cologne Center for Genomics, University of Cologne, Weyertal 115b, 50931 Cologne, Germany

Cologne Excellence Cluster on Cellular Stress Responses in Aging-associated Diseases, University of Cologne, Cologne, Germany

Division of Human Genetics and Department of Dermatology, Medical University of Innsbruck, Innsbruck, Austria
e-mail: h.hennies@uni-koeln.de

© Springer International Publishing Switzerland 2017
P.M.N. Werker et al. (eds.), *Dupuytren Disease and Related Diseases – The Cutting Edge*,
DOI 10.1007/978-3-319-32199-8_13

info@dupuytrens.org

far greater influence on the mean age of first surgery than other risk factors, namely, heavy smoking. We clearly showed in this study that a positive family history, and with it the underlying genetic risk factors, strongly contributes to disease severity (time of first surgical intervention). In a recent population-based twin study, the heritability of DD was calculated to be 80 % (Larsen et al. 2015). Therefore we aim to identify the causal variants in the genome to understanding the genetic basis of this multifactorial disease.

13.2 Previous GWAS Findings

The first GWAS in DD (Dolmans et al. 2011) identified nine genomic loci associated with this disease on genome-wide significant level (p-value $< 5 * 10^{-8}$). Six of these nine loci harbor one gene each that codes for an upstream modulator of the Wnt signaling pathway, e.g., Wnt ligands or inhibitors. This intriguing overrepresentation of Wnt signaling-related genes in the GWAS loci led to the first valuable insight into the possible genetic factors underlying DD. None of these genes have so far been further investigated as the true culprit at a given loci. The complex nature of common diseases makes it a difficult task to identify truly causative genetic variants by linking them to the disease phenotype. This is a well-known problem in complex genetic diseases and presents one of the major challenges of our age.

The majority of GWAS loci that have been identified for complex diseases fall outside of coding genes, and they are supposed to reflect alterations in regulatory features (Schaub et al. 2012). Many of these regulatory effects will be small and difficult to detect individually (Civelek and Lusis 2014). On top of that, GWASs for common, complex diseases (or traits) only explain a small proportion of the genetic basis, and a lot true positives may be hidden in the statistical noise produced by GWAS. This necessitates a systems approach to analyze the pathways and networks involved in DD. In particular network modeling may help to uncover relationships between genes from top GWAS loci, allowing for the inclusion of susceptibility loci with more subtle effects and increasing statistical power to detect them as they are viewed in context of each other.

13.3 Network Analysis

Pathway and network analysis have been extensively used in the analysis of expression data. But they are also useful tools to investigate complex genetic diseases. In complex genetic diseases, different genetic variants within an individual and different genetic variants between individuals contribute to the disease. Genetic variants can act additively and in concert with environmental factors. The same genetic variant in two individuals does not necessarily lead to the disease in both individuals, depending on the genomic background of each individual and the environment the individual was exposed to. Individual genetic variants in complex diseases can cover the whole spectrum of pathogenicity from fully penetrant to slight effects. By design only few of these variants are picked up by GWAS because the multiple testing of thousands of variants (SNPs, markers) requires a strict significance threshold, in order to reduce the number of false positive findings. Consequently only the very top loci are considered in a traditional GWAS approach. But although the genetic basis of a complex disease can be spread out over many genetic loci and genes, these genetic alterations are not randomly distributed but affect a limited number of cellular functions and pathways. This consideration makes it possible to search for functional connections between genes in GWAS loci. In contrast to pathway analysis, network analysis does not require prior knowledge about the function of a gene product or its affiliation to a pathway but relies on a network constructed from protein-protein interaction (PPI) data. The nature of the PPI data can either be physical interactions, classically generated by yeast two-hybrid screens, or other data sources, e.g., co-expression data. A network in this context constitutes interaction data consisting of molecules (e.g., proteins), termed nodes, and their relationships with each other (e.g., physical

info@dupuytrens.org

interactions, co-expression), termed edges. The aim in network analysis of GWAS data is to identify modules – groups of connected proteins/genes that share characteristics under study – that are enriched in small *p*-values. In this process GWAS data can be integrated with whole transcriptome expression data, for instance, for the disease tissue. Considering genes that are co-expressed in the affected tissue increases the power to detect true associations.

As a complex disease, DD is well suited for a network-based approach (Fig. 13.1). DD has a strong genetic basis, disease tissue and healthy tissue are readily accessible, and some prior information about pathways likely involved in the development of this disease is available (in particular Wnt signaling but other pathways may also be similarly important). Moreover, the phenotype is clearly defined, although the severity of the disease differs between individuals.

13.3.1 Network Analysis Workflow

In the first step SNP-based *p*-values are translated into gene-based *p*-values. For this, SNPs must be assigned to genes. The simplest method to do this is to define a window around each gene and assign all SNPs within this window to this gene. But this is no trivial task as SNPs not necessarily act on the nearest gene and long-range interactions are possible. We used the software VEGAS2 (Mishra and Macgregor 2014), which also takes into account linkage information (e.g., from the 1000 Genomes Project reference population) and gene sizes. VEGAS2 combines the test statistics of all SNPs within ±50 kb of each gene. Based on SNP association *p*-values, the software calculates empirical gene-based *p*-values by a simulation procedure.

The next step is to search for modules enriched in small *p*-values within a protein-protein interaction (PPI) dataset. The assumption behind is that in complex genetic settings, many different variants affecting several different genes may contribute to the disease, but these genes are assumed to act in a limited number of pathways or cellular functions. Because of the limited number of affected pathways/cellular functions, truly associated genes are expected to be more functionally connected to each other than random genes. The search for modules instead of individual genes increases statistical power since association does not rely on individual genes but a module of functionally connected genes.

To further increase the power to detect true associations in the statistical noise of GWAS, one can combine the GWAS data with tissue-specific whole-genome transcription data by considering only genes that are expressed or co-expressed in the tissue of interest when searching for connections between genes with small *p*-values.

Network analysis results in lists of genes. The next logical step to do is to look for enrichment of functional annotations in these lists of genes (e.g., canonical pathways or gene ontology (GO) terms). Although the function of all genes in a sub-network may not be known, this constitutes the first insight into which pathways may be affected by genetic alterations in DD. The ultimate aim is to unravel which roles the specific genes in the detected sub-networks play in the pathway in the context of the disease and how genetic alterations change these functions in DD. For these more functional experimental, studies are necessary.

13.4 Targeted Sequencing

13.4.1 Ongoing Studies to Identify Genetic Variants in GWAS Loci by Targeted NGS

The SNPs tested in GWAS are selected as a set of informative single SNPs able to tag common haplotype blocks. To explicitly capture causative variants in GWAS-identified DD susceptibility loci, it will be critical to sequence each candidate locus using targeted next-generation sequencing (NGS). By mapping NGS data to the human genome reference sequence, the variability of the entire locus can be exhaustively identified, including both coding and noncoding regions and comprising all common and rare variants (Udler et al. 2010).

info@dupuytrens.org

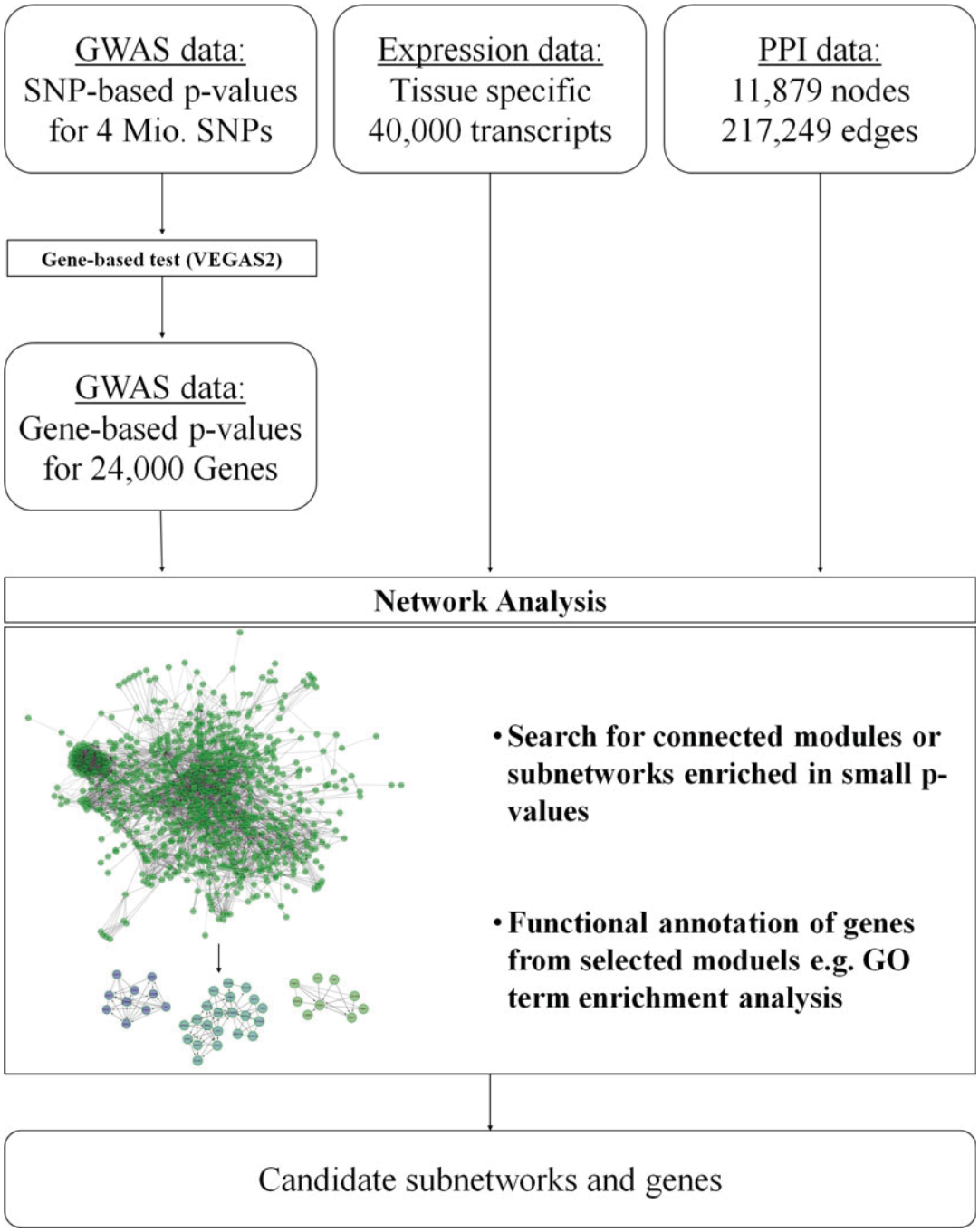

Fig. 13.1 Overview network analysis. Simplified workflow of the network analysis integrating GWAS and transcriptome data to search for disease-specific sub-networks in DD. In the first step SNP-based *p*-values are translated into gene-based *p*-values. Genes with *p*-values are then imposed on protein-protein interaction (*PPI*) data, and a search for connections between genes with small *p*-values is conducted. Once sub-networks (modules) enriched for genes with small *p*-values are identified, these can be validated and further analyzed for functional annotation in the context of the disease. To further increase the power to detect relevant sub-networks, tissue-specific expression data can also be integrated in the network analysis approach

info@dupuytrens.org

As a first step, we have selected a 500 kb region containing the lead SNP rs16879765 (chromosome 7p14.1) for targeted sequencing (Fig. 13.2). DNA was isolated from peripheral blood of 96 DD patients. The DD-associated locus was enriched in these samples using a custom designed Agilent SureSelect XT2 kit and sequenced on the Illumina HiSeq 2000 platform. Sequencing data are analyzed with the Varbank pipeline (v2.13) (CCG, Cologne) and Ensembl Variant Effect Predictor (http://www.ensembl.org/info/docs/variation/vep/index.html).

Once the potential candidate variants are discovered and validated, the next step will be to prioritize the candidates based on the following criteria: (1) exclude known, assumed harmless variations present in dbSNP databases (http://www.ncbi.nlm.nih.gov/SNP) and published studies; (2) select variants causing changes in protein-coding sequences and likely to compromise protein structure, function, or stability; and (3) select noncoding variants that may affect regulation of gene expression.

13.4.2 Functional Studies of Variants Within Coding Regions

All coding variants identified in DD patients are validated by Sanger sequencing, in particular variants in regions that contain multiple and/or recurrent variants in patients as compared to controls. Then, replication of the results in an independent cohort is needed. The identified variations are analyzed to predict the structure of the gene carrying variations and the function of the resulting protein by using tools such as SIFT (http://sift.jcvi.org/www/SIFT_dbSNP.html), PolyPhen-2 (http://genetics.bwh.harvard.edu/pph2), Mutation Profiling (http://profile.mutdb.org), and ModBase (http://modbase.compbio.ucsf.edu). After identification of a set of DD-predisposing gene variants, in vitro studies to test the functional consequences of these candidates are crucial.

The primary functional studies will focus on the key molecular events in DD development – the aberrant proliferation of fibroblasts and their differentiation into myofibroblasts. After cloning and expression of mutated genes in DD and control cells, cell proliferation is assessed using, for instance, the CyQUANT® Cell Proliferation kits (Thermo), which measures the cellular DNA content. Furthermore, as the expression and organization of α-SMA are hallmarks of myofibroblast differentiation (Tomasek et al. 2002), LightCycler® (Roche) qRT-PCR and Western blotting are employed to detect α-SMA expression. Additionally, myofibroblast contractile activity and migration will be investigated by collagen matrix contraction and in vitro wound healing assays separately.

13.4.3 Functional Studies of Noncoding Regulatory Variants

Many variants associated with GWAS were identified in noncoding regions of the genome, and this has increased the interest in the effect of genetic variants on regulation of gene expression. Recently, expression quantitative trait loci (eQTL) mapping has become a powerful tool to understand how noncoding variants in GWAS loci influence disease risk (Conde et al. 2013; Li et al. 2013). Identification of an eQTL, a genomic locus which regulates transcript expression levels, involves association analysis between genetic markers and gene expression levels typically measured in hundreds of individuals. Microarrays or RNA-seq are often used to measure the expression levels of genes in a genome simultaneously and map these phenotypes to genomic regions represented by genetic markers captured in GWAS. One important advantage of such an approach is that it allows identification of regulators of expression of disease-associated genes if there are variants affecting expression of that regulator (Steiling et al. 2013).

To identify noncoding variants that affect gene expression in DD-associated loci, we use published eQTL and ENCODE data to prioritize genomic variants found by targeted NGS. Using

info@dupuytrens.org

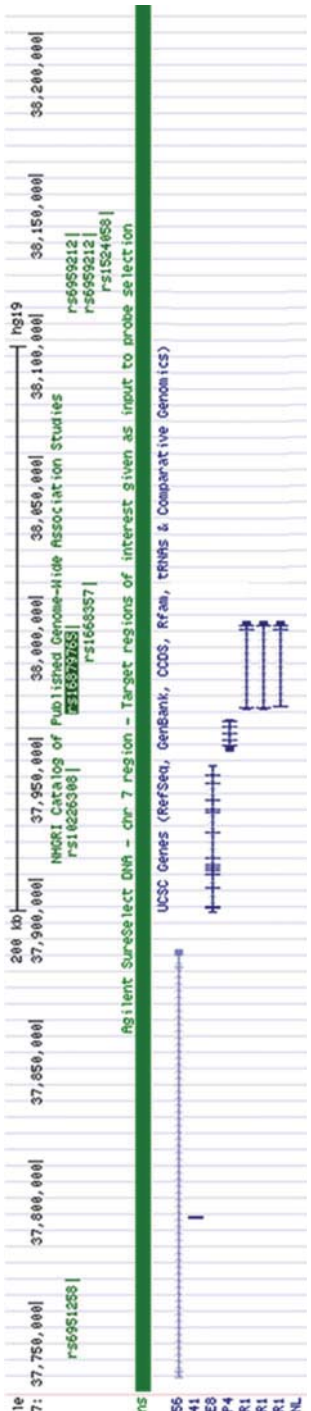

Fig. 13.2 UCSC Genome Browser plot of target region containing rs16879798 for enrichment capturing and sequencing. UCSC Genome Browser plot of target region containing rs16879798 for enrichment captures and sequencing. Target region: chr7:37,714,869 – 38,214,857. Predicted Agilent SureSelect XT2 coverage: 98.3 %

info@dupuytrens.org

qPCR assays, we test the candidate eQTLs in DD tissues and primary cells derived from DD tissues. A number of bioinformatics tools will then be used to predict possible activities of noncoding variants using, for instance, Genomatix (http://www.genomatix.de), Transfac (http://www.gene-regulation.com/pub/databases.html), and Human Splicing Finder (http://www.umd.be/HSF). Taken together, we expect to identify noncoding variations that underlie inherited differences in expression levels of genes, which is supposed to lead to the identification of genes involved in the susceptibility to DD.

Conclusions

- Dupuytren Disease has a strong genetic basis and unraveling this basis is challenging.
- Network analysis integrating GWAS and differential transcriptome expression data with protein-protein interactions facilitates the identification of modules and pathways perturbed by genetic alterations in DD.
- Targeted sequencing of GWAS loci aims to identify the underling causative genetic variants.
- Most causative genetic variants in DD are expected to lie outside coding regions, and efforts both in computational methods and functional study design must be undertaken to address them.

Acknowledgments and Conflict of Interest Declaration We are particularly grateful to all patients and control persons who participated in the study. The study was approved by the institutional review board and participants provided written informed consent. The project is supported in part by grants from the DFG through the Cologne Cluster of Excellence on Cellular Stress Responses in Aging-Associated Diseases and the Köln Fortune Program of the Faculty of Medicine, University of Cologne. The authors have no conflict of interest to declare.

References

Becker K et al (2015) The importance of genetic susceptibility in Dupuytren's disease. Clin Genet 87: 483–487

Brenner P, Krause-Bergmann A, Van VH (2001) Dupuytren contracture in North Germany. Epidemiological study of 500 cases. Unfallchir 104:303–311

Capstick R, Bragg T, Giele H, Furniss D (2013) Sibling recurrence risk in Dupuytren's disease. J Hand Surg Eur 38(4):424–429

Civelek M, Lusis AJ (2014) Systems genetics approaches to understand complex traits. Nat Rev Genet 15:34–48

Coert JH, Nerin JP, Meek MF (2006) Results of partial fasciectomy for Dupuytren disease in 261 consecutive patients. Ann Plast Surg 57:13–17

Conde L et al (2013) Integrating GWAS and expression data for functional characterization of disease-associated SNPs: an application to follicular lymphoma. Am J Hum Genet 92:126–130

Dolmans GH et al (2011) Wnt signaling and Dupuytren's disease. N Engl J Med 365:307–317

Early PF (1962) Population studies in Dupuytren's contracture. J Bone Joint Surg 44B:602–613

Hakstian RW (1966) Long-term results of extensive fasciectomy. Br J Plast Surg 19:140–149

Hindocha S, John S, Stanley JK et al (2006a) The heritability of Dupuytren's disease: familial aggregation and its clinical significance. J Hand Surg 31:204–210

Hindocha S, Stanley JK, Watson S, Bayat A (2006b) Dupuytren's diathesis revisited: evaluation of prognostic indicators for risk of disease recurrence. J Hand Surg 31:1626–1634

Lanting R et al (2013) Prevalence of Dupuytren disease in The Netherlands. Plast Reconstr Surg 132:394–403

Larsen S et al (2015) Genetic and environmental influences in Dupuytren's disease: a study of 30,330 Danish twin pairs. J Hand Surg Eur 40:171–176

Li L et al (2013) Using eQTL weights to improve power for genome-wide association studies: a genetic study of childhood asthma. Front Genet 4:103

Ling RS (1963) The genetic factor in Dupuytren's disease. J Bone Joint Surg Br 45:709–718

Makela EA, Jaroma H, Harju A, Anttila S, Vainio J (1991) Dupuytren's contracture: the long-term results after day surgery. J Hand Surg Br 16:272–274

Mishra A, Macgregor S (2014) VEGAS2: software for more flexible gene-based testing. Twin Res Hum Genet 18(1):86–91

Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M (2012) Linking disease associations with regulatory information in the human genome. Genome Res 22:1748–1759

Steiling K, Lenburg ME, Spira A (2013) Personalized management of chronic obstructive pulmonary disease via transcriptomic profiling of the airway and lung. Ann Am Thorac Soc 10(Suppl):S190–S196

Tomasek JJ et al (2002) Myofibroblasts and mechano-regulation of connective tissue remodelling. Nat Rev Mol Cell Biol 3:349–363

Udler MS, Ahmed S, Healey CS et al (2010) Fine scale mapping of the breast cancer 16q12 locus. Hum Mol Genet 19:2507–2515