Jung *et al. BMC Medical Genomics* 2019, **12**(Suppl 5):98
https://doi.org/10.1186/s12920-019-0518-3

BMC Medical Genomics

RESEARCH Open Access

Integrative genomic and transcriptomic analysis of genetic markers in Dupuytren's disease

Junghyun Jung[1], Go Woon Kim[2], Byungjo Lee[1], Jong Wha J. Joo[3*†] and Wonhee Jang[1*†]

From The 8th Annual Translational Bioinformatics Conference
Seoul, South Korea. 31 October - 2 November 2018

Abstract

Background: Dupuytren's disease (DD) is a fibroproliferative disorder characterized by thickening and contracting palmar fascia. The exact pathogenesis of DD remains unknown.

Results: In this study, we identified co-expressed gene set (DD signature) consisting of 753 genes via weighted gene co-expression network analysis. To confirm the robustness of DD signature, module enrichment analysis and meta-analysis were performed. Moreover, this signature effectively classified DD disease samples. The DD signature were significantly enriched in unfolded protein response (UPR) related to endoplasmic reticulum (ER) stress. Next, we conducted multiple-phenotype regression analysis to identify *trans*-regulatory hotspots regulating expression levels of DD signature using Genotype-Tissue Expression data. Finally, 10 *trans*-regulatory hotspots and 16 eGenes genes that are significantly associated with at least one *cis*-eQTL were identified.

Conclusions: Among these eGenes, major histocompatibility complex class II genes and ZFP57 zinc finger protein were closely related to ER stress and UPR, suggesting that these genetic markers might be potential therapeutic targets for DD.

Keywords: Dupuytren's disease, Unfolded protein response (UPR), Endoplasmic reticulum (ER) stress, Multiple-phenotype analysis, *trans*-regulatory hotspots, ZFP57 zinc finger protein, Major histocompatibility complex class II

Background

Dupuytren's disease (DD) is a fibroproliferative disorder characterized by palmar fascia hypertrophy that often results in thickening and contracting palmar fascia [1]. DD mostly occurs in ring finger, followed by little and middle fingers, where affected fingers become permanently and irreversibly bent in a flexed position [2]. The prevalence of DD rises with increasing age and DD is most commonly seen in Europe. It has higher prevalence in northern Europe than that in southern Europe [3]. Even though Lee et al. have recently shown that DD is not a disease limited to European descent anymore, it is still classified as a rare and hard-to-care disease in Korea [4]. Alcoholism, smoking, dyslipidemia, and diabetes are regarded as risk factors of DD; however, the exact etiopathogenesis of DD remains unclear [5].

With rapid growth of high-throughput technology, previous studies have reported that several genes are associated with the progression of DD based on differentially expressed gene (DEG) analysis using microarray [6–11]. Most DEG studies were focused on single genes without considering interconnections between genes, especially those with high number of connections (edges) in a network. Co-expression network analysis based on similar expression patterns can be effectively used for identifying a set of genes that are simultaneously active in the same functional processes [12]. Together with transcriptomic data, several risk genetic loci related to

* Correspondence: jwjjoo@dongguk.edu; wany@dongguk.edu
†Jong Wha J. Joo and Wonhee Jang contributed equally to this work.
[3]Department of Computer Science and Engineering, Dongguk University-Seoul, Seoul 04620, South Korea
[1]Department of Life science, Dongguk University-Seoul, Seoul 04620, Republic of Korea
Full list of author information is available at the end of the article

© The Author(s). 2019 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

the pathogenesis of DD have been identified by genome-wide association studies (GWAS) to inspect the association of a single phenotype and each single nucleotide polymorphism (SNP) [13–15]. However, SNPs cannot solely explain biological processes because most of such variants reside in noncoding regions of the genome [16].

Recently, large quantities of genomic data alongside with expression data per individual have been gathered from GWAS cohorts. The Genotype-Tissue Expression (GTEx) project was founded for expression quantitative trait locus (eQTL) mapping, which investigates effects of genetic variation on gene expression in extensively diverse primary tissues from human [17]. eQTL analysis indeed is an approach to explain genetic variation underlying altered gene expression [18]. Recent studies have described tissue-specific eQTLs because gene expression patterns are different across tissues [19, 20]. *cis*-eQTLs or *trans*-eQTLs usually refers to eQTLs that regulate nearby or distal genes, respectively [18], and genes that are significantly associated with at least one *cis*-eQTL are referred to as eGenes [21]. Notably, previous eQTL studies showed that a small number of genomic regions referred to as *trans*-regulatory hotspots can regulate expression levels of hundreds of genes [22–24], suggesting the existence of master regulators of transcription. Typically, eQTL approaches analyzing independent phenotypes have low statistical power [25]. On the other hand, multi-variate methods analyzing many phenotypes simultaneously can increase the power to identify underlying regulatory hotspots in a complex biological system [26].

In this study, we used weighted gene co-expression network analysis (WGCNA) to find co-expression gene set (module) of highly correlated genes for DD. Using independent datasets as validation sets, we confirmed the reliability of selected gene set via module enrichment analysis based on gene set enrichment analysis (GSEA) and disease classification. Finally, a multiple-phenotype regression analysis was performed using GTEx muscle data to identify regulatory hotspots related to gene set of DD.

Methods

Microarray preprocessing and meta-analysis

Affymetrix microarray datasets were preprocessed and normalized following Jung et al. [27]. Other platform datasets including CodeLink, Stanford, and Illumina platform were preprocessed using limma R package [28]. Two-color microarray dataset (GSE2688) comparing relative expression levels between a sample RNA and a universal RNA in a single microarray was adjusted for batch effects after independent normalization because the dataset was from two types of array platforms. Meta-analysis was conducted using the one-color microarray datasets measuring expression levels from each sample separately (Table 1). These datasets were combined using unique Entrez IDs. Meta-analysis was carried out according to SVA R package [29] after adjusting for batch effects using Combat [30].

Weighted gene correlation network analysis

GSE75152 dataset contained mRNA expression profiles of 12 DD patients and 12 control subjects with total RNA extracted from the connective tissue from the hand [11]. Top 5000-most expressed probes were selected for computational cost and simplicity after normalization. Multiple probes representing one gene were collapsed using collapseRows function [31]. A signed WGCNA was used to identify co-expression modules comprised of positively correlated genes based on Pearson correlation coefficient [32]. In detail, a similarity matrix based on Pearson correlation of all pairs of genes was converted into an adjacency matrix via a power function. A suitable soft-thresholding power (β) of the power function was selected via analysis of scale-free topology. Next, the adjacency matrix was transformed into a topological overlap matrix (TOM) to reflect topological information of a network. Modules were defined by a hybrid tree cut method when cutting a hierarchical cluster trees [32]. Expression patterns in modules were

Table 1 Characteristics of NCBI GEO datasets used for WGCNA analysis

	GEO series ID	Array type	Array platform	No. of arrays (DD: Control)	Source of tissue	PMID
1	GSE2688 (GSE4457)	Two-color	Stanford Microarray	11 (4: 7)	DD tissues	18694919 [6], 16473681 [7]
2	GSE21221	One-color	GE Healthcare CodeLink Human Whole Genome Bioarray	12 (6: 6)	Fibroblasts derived from DD tissues	18433489 [40]
3	GSE31356	One-color	Affymetrix Human Genome U133A Array	6 (3: 3)	DD tissues	22965824 [8]
4	GSE41524	One-color	Affymetrix Human Exon 1.0 ST Array	10 (4: 6)	Fibroblasts derived from DD tissues	23554969 [9]
5	GSE59746	One-color	Affymetrix Human Genome U133 Plus 2.0 Array	4 (2: 2)	DD tissues	25379672 [10]
6	GSE75152	One-color	Illumina HumanWG-6 v3.0 expression beadchip	24 (12: 12)	DD tissues	27467239 [11]

summarized by module eigengene. The threshold of minimum size of modules was 50 genes and pairs of modules with high ME correlations ($r > 0.85$) were merged.

Module enrichment analysis and functional annotation

The fast preranked gene set enrichment analysis (fgsea) R package was used for GSEA [33]. Microarrays of log fold change (log FC) values were regarded as a pre-ranked list. Modules derived from WGCNA were used as gene sets for module enrichment analysis. Database for Annotation, Visualization and Integrated Discovery (DAVID) [34] was used to conduct functional enrichment analysis.

Disease classification

Random forest [35] classification was conducted using classification and regression training (caret) r package [36]. Combined one-color microarray dataset and two-color microarray dataset were transformed to mean 0 and variance 1 for each gene. Model training for DD sample classification was performed with the one-color microarray dataset using LOOCV. Model validation was carried out with the two-color microarray dataset.

Regulatory hotspot analysis

To identify *trans*-regulatory hotspots, we performed GAMMA, one of multiple-phenotype analysis approaches to examine an association between a number of phenotypes or gene expression levels and each SNP [25]. GAMMA is preferred over other multiple-phenotype approaches as it is scalable to high dimensional data, containing hundreds to thousands number of genes which is often the case with eQTL data. In addition, utilizing linear mixed model, it considers widely known genetic relatedness referred to as population structure in the data. The population structure complicates association analysis by inducing spurious signals. Especially, in multiple-phenotype analysis, these problems may compound as bias because population structure accumulates from each phenotype [25]. Skeletal muscle data from 361 samples in GTEx project (version 6) data in dbGaP database (accession phs000424.v6.p1) were used for this analysis. The top three principal components of covariates were regressed out in expression data. GAMMA was performed by an adaptive permutation which increased the number of permutations from 10^2 to 10^6, increasing by 10 folds each time. A summary data of SNP-gene associations was obtained from the GTEx Portal (http://gtexportal.org). SNPs within ±1 Mb of the transcriptional start site of each gene were used to identify *cis*-eQTL.

Cross-species mapping using ER stress gene expression data

In this study, gene expression microarray dataset GSE35209 obtained from MEFs treated with an ER stress-inducing agent (tunicamycin) was used. The pre-processed data were obtained using GEOquery r package [37]. Cross-species mapping between human and mouse genes was performed using National Center for Biotechnology Information HomoloGene database (Build68) [38].

Results

Identification of co-expression module for DD

We constructed co-expression networks using the WGCNA r package describing correlation patterns among genes across DD patients and normal subjects to identify a representative set of genes for DD. Among the top 5000-most expressed microarray probes in GSE75152, we identified unique 4141 representative genes based on collapseRows function in WGCNA r package [31]. Selected genes with similar expression patterns were clustered into gene set modules via average linkage hierarchical cluster analysis. The power of $\beta = 12$ (scale-free $R^2 = 0.82$) was selected as the soft-thresholding power for co-expression network construction (Additional file 1). We identified 16 co-expression modules representing genes that shared highly similar expression patterns (Fig. 1a). Among these modules, the red and blue modules were enriched in up- and down-regulated genes, respectively (Fig. 1b). Next, we carried out meta-analysis using one-color microarray datasets to identify representative module for DD. A total of 67 microarray data in five independent studies were used, consisting of 31 DD samples and 36 normal samples (Table 1). The results showed that DEGs derived from meta-analysis were significantly enriched in red (Two-sided Fisher's exact test: odds ratio = 2.98 and P value = 2.64e-23) and blue (Two-sided Fisher's exact test: odds ratio = 3.96 and P value = 1.23e-14) modules (Fig. 1c and Additional file 2A). Additionally, DEGs of two-color microarray datasets were also significantly enriched in red (Two-sided Fisher's exact test: odds ratio = 2.50 and P value = 6.35e-19) and blue modules (Two-sided Fisher's exact test: 0.35, odds ratio = 0.35 and P value = 6.49e-4) (Fig. 1d and Additional file 2B).

DD signature has the power to classify DD samples

To test whether these identified modules were replicated in other 5 independent datasets (Table 1), GSEA were performed for module enrichment analysis. The results showed that only the red module (753 genes) was significantly and positively enriched in all 5 independent datasets (FDR < 0.05) (Fig. 2a and b). Consistently, the expression levels of red module genes were distinct and discriminative according to the DD patient or the normal subject data in GSE75152 (Fig. 1e and f). We then asked whether red module genes could sufficiently classify disease state of individual samples related to DD. In order to apply a sample classification approach of DD samples, all datasets including CodeLink, Affymetrix, and Illumina

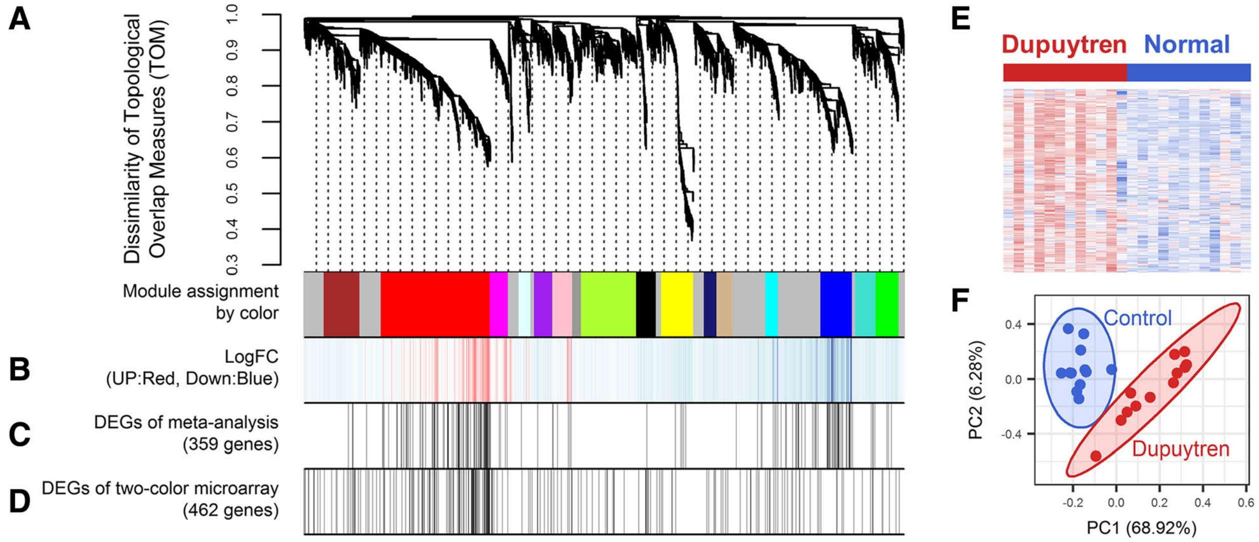

Fig. 1 Identification of modules associated with gene expression of DD. **a** Dendrogram showing modules based on the dissimilarity of TOM (1-TOM). Color bars below show assignment of modules. **b** Heatmap showing the gene expression pattern of modules. Red and blue color lines indicate up- and down-regulation, respectively. **c** A Heatmap showing the results of the meta-analysis derived from one-color microarray dataset. Black indicates DEGs (FDR < 0.01). **d** A Heatmap showing results of DEGs analysis of two-color microarray dataset. Black color lines indicate DEGs (FDR < 0.01). **e** A Heatmap showing expression patterns of genes for red module. Black color lines indicate DEGs (FDR < 0.01). **f** A scatter plot of principal component analysis (PCA) showing a distinct separation of the expression level of red module genes between patients with DD and normal subjects

were merged (Table 1). Among the 753 red module genes, only 255 genes remained after merging the datasets because genes contained by each microarray platforms are different. Model training using random forest method was carried out with one-color microarray datasets using Leave-one-out cross-validation (LOOCV) while model validation was performed using two-color microarray dataset. The Classification performance showed that area under the curve (AUC) value in receiver operating characteristic (ROC) analysis curve was higher for red module genes than that of two randomly selected genes having the same number of red module gene (Fig. 2c). Collectively, these results strongly suggested that the red module genes (so-called DD signature in this paper) (Additional file 3) were a robust set of genes representing DD.

Functional enrichment analysis of DD signature revealed dysregulated functions in DD

To obtain insights into the biological process of DD, we performed functional enrichment analysis to define dysregulated Gene Ontology (GO) categories and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway of the DD signature. We found terms for extracellular matrix (ECM) or collagen were significantly enriched in extracellular matrix organization (GO:0030198 and FDR = 5.97e-10), ECM-receptor interaction (hsa04512 and FDR = 7.60e-04), and collagen catabolic process (GO:0030574 and FDR = 4.62e-06), while the enriched terms for adhesion were cell-cell adhesion (GO:0098609 and FDR = 1.41e-03) and focal adhesion (hsa04510 and FDR = 2.01e-02) (Fig. 2d and Additional file 4). Previous results from microarray studies indicated that several proteins for extracellular matrix (ECM) are correlated with mRNA dysregulation [6, 39–41] and that pathological collagen deposition is related to DD [42]. Notably, unfolded protein response (UPR) related to endoplasmic reticulum (ER) stress terms were significantly enriched in protein processing in endoplasmic reticulum (hsa04141 and FDR = 1.34e-04) and IRE1-mediated unfolded protein response (GO:0036498 and FDR = 2.06e-05) (Fig. 2d and Additional file 4). UPR and ER stress have been detected in many disease including neurodegenerative disease, cancer, diabetes, liver disorders, and obesity [43]. Our results suggest that unfolded protein-induced ER stress can also be involved in the pathogenesis of DD.

Identification of regulatory hotspot related to DD signature

To identify *trans*-regulatory hotspots regulating expression levels of DD signature, multiple-phenotype regression analysis was carried out using GAMMA (Generalized analysis of molecular variance for mixed-model analysis) [25] in eQTL datasets of GTEx version 6 dataset. The presence of abnormal myofibroblasts in DD palmar fascia plays a causative role in digital contracture of DD [44]. Therefore, muscle tissues were used in this analysis. First, we found 512 loci using GAMMA (*P* value <5e-05) (Fig. 3a). We then examined *cis*-acting SNPs within ±1 Mb region of

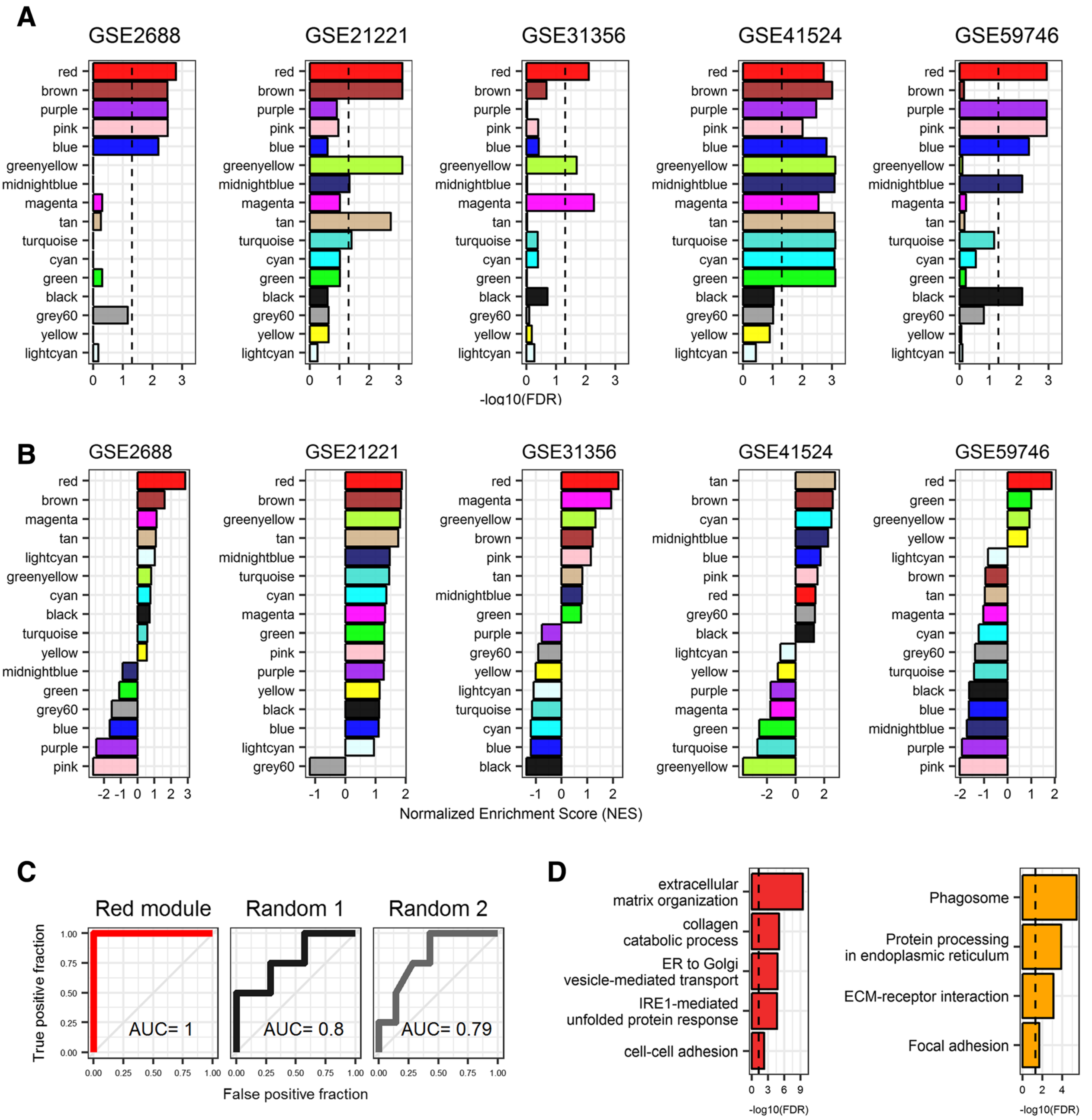

Fig. 2 Robustness of DD signature and its functional annotations. **a** Bar plots showing the results of module enrichment analysis using GSEA. Black dotted lines indicate significant threshold (FDR < 0.05). **b** Bar plots showing normalized enrichment score (NES) for results of module concentration analysis using GSEA. **c** ROC analysis with AUC showing the performance of classification using red module in (**a**). Randomly selected genes consist of the same number of the red module gene. **d** Bar plots showing the results of functional enrichment analysis. Red and orange color bars represent GO biological process and KEGG pathway, respectively. Black dotted lines indicate significant threshold (FDR < 0.05)

the transcription start site (TSS) of each gene among GAMMA loci because recent studies indicated that expression change of a *trans*-acting factor by a *cis*-eQTL was another possible causal mechanism [45, 46]. Finally, we identified 10 GAMMA loci and 16 eGenes (q value < 0.05) that were significantly associated with at least one *cis*-acting SNP [21] (Table 2), including 5 protein-encoding eGenes: major histocompatibility complex class (MHC) II DQ alpha 1 (*HLA-DQA1*), DQ beta 1 (*HLA-DQB1*), DQ beta 2 (*HLA-DQB2*), DR beta 1 (*HLA-DRB1*), and DR beta 5 (*HLA-DRB5*). Their expression levels were significantly associated with one rs2269423 on chromosome 6 (Fig. 3b and Table 2). Previous results showed that human leukocyte alleles (HLA) encoding MHC proteins

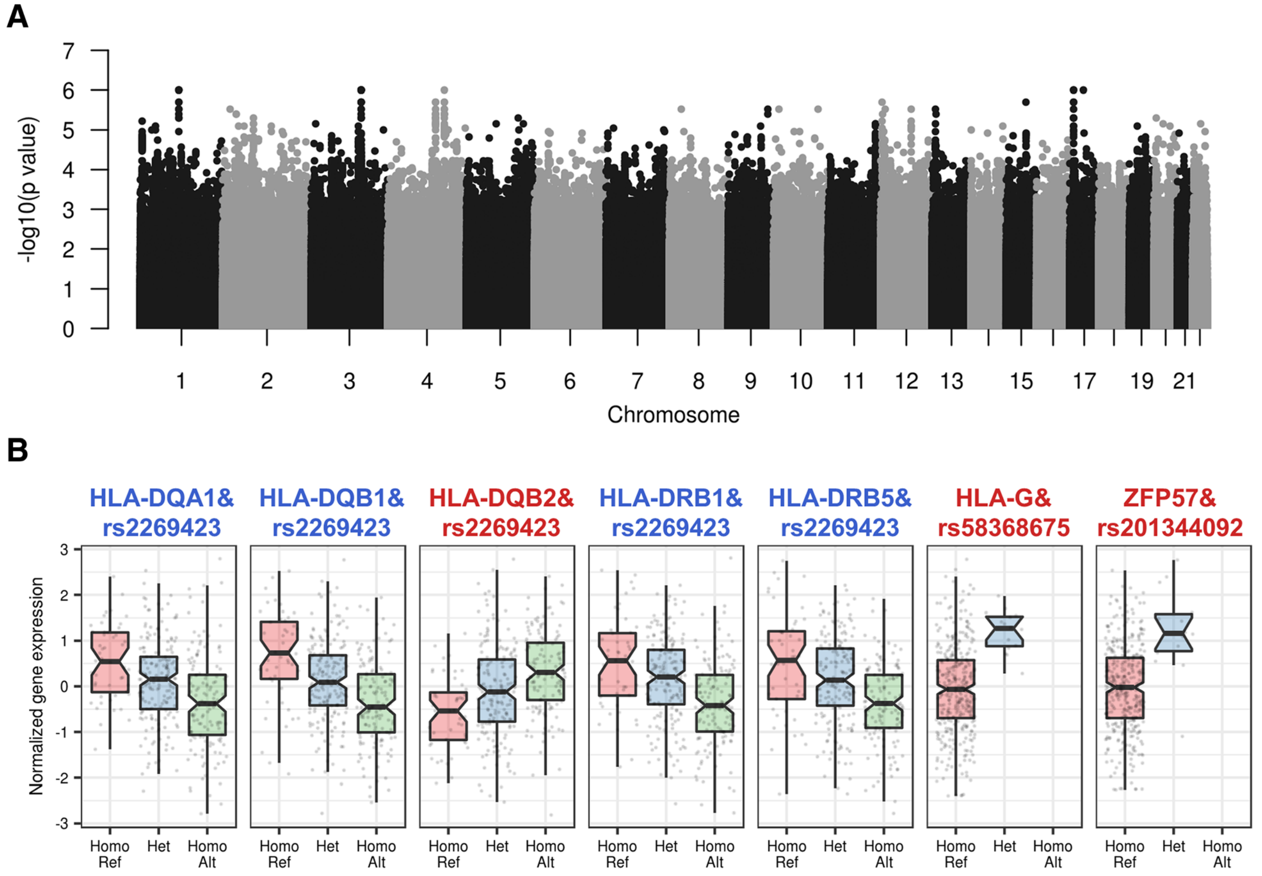

Fig. 3 Identification of *trans*-regulatory hotspots associated with DD signature. **a** The GAMMA results applied to GTEx dataset using DD signature. **b** Box plots of gene expression levels of eGenes by each *trans*-regulatory hotspot between the different genotype groups

in humans, especially HLA-DR alleles, are related to DD [47, 48]. Among the identified 5 HLA genes related to rs2269423, 4 genes were also identified to have significant variant-gene associations in GTEx version 7 dataset, and 3 genes including HLA-DR beta 6 (*HLA-DRB6*), *NOTCH4*, and activating transcription factor 6 beta (*ATF6B*) were additionally identified (Additional file 5 and Fig. 4). ATF6 is a key transcription factor for unfolded protein response (UPR) pathway during ER stress [49]. rs201344092 allele was associated with an increase in gene expression of *ZFP57* (Fig. 3b, c and Table 2). Importantly, a mutation of *ZFP57* is associated with neonatal diabetes type 1 [50]. Previous studies indicated that *Zfp57* expression was up-regulated by ER stress condition in mice [51, 52], suggesting that *Zfp57* might be secondary UPR-regulated transcriptional repressor [53].

Expression of eGenes under ER stress condition

To determine whether expression levels of the identified genes are regulated by ER stress condition, we examined the gene expression levels using microarray data derived from mouse embryonic fibroblasts (MEFs) treated with ER stress-inducing agent (GSE35209). 10 genes including 2 up-regulated and 4 down-regulated DEGs remained after cross-species mapping procedure. Remarkably, *Zfp57* (*ZFP57* in human) was significantly up-regulated in ER stress condition (Fig. 4) and *ZFP57* is positively regulated by the *trans*-regulatory hotspot (rs201344092) associated with DD signature in human (Fig. 3b). *H2-Eb1* (*HLA-DRB5* in human) and *H2-Ab1* (*HLA-DQB1* in human) were significantly down-regulated in ER stress condition (Fig. 4) and *HLA-DRB5* and *HLA-DQB1* are negatively regulated by the *trans*-regulatory hotspot (rs2269423) associated with DD signature in human (Fig. 3b). Together, our data suggest that the regulatory hotspot nearby *ZFP57* and MHC class II genes may develop ER stress condition by regulating ER stress-s and UPR-related genes.

Discussion

The number of patients with DD has continued to increase from 1118 in 2007 to 3280 in 2014 in the United States [54]. A surgical treatment including fasciotomy with or without dermofasciectomy is the current treatment option for DD [47]. However, surgical treatment has a high recurrence rate [54]. It has been reported that

Table 2 List of significant associations with DD signature. The lowest GAMMA *P* for SNP was listed among multiple SNP related to one eGene (See also Additional file 4)

Chr	Position	rs ID	GAMMA *P* value	eGenes (q value < 0.05)			
				Gene Symbol	Gene Name	Effect size (beta)	Gene Type
3	49,769,419	rs62262722	1.4E-05	AMT	aminomethyltransferase	−0.16	protein coding
				FAM212A	family with sequence similarity 212 member A	0.26	protein coding
	49,858,661	rs55997059	1.7E-05	WDR6	WD repeat domain 6	0.20	protein coding
	49,898,318	rs62260755	2.4E-05	RBM6	RNA binding motif protein 6	−0.28	protein coding
6	30,500,730	rs148983519	3.8E-05	HCG9	HLA complex group 9 (non-protein coding)	1.11	lincRNA
	30,642,417	rs201344092	1.1E-05	ZFP57	ZFP57 zinc finger protein	1.44	protein coding
	30,664,568	rs58368675	2.2E-05	HLA-G	major histocompatibility complex, class I, G	1.34	protein coding
	32,145,707	rs2269423	1.2E-05	HLA-DQA1	major histocompatibility complex, class II, DQ alpha 1	−0.44	protein coding
				HLA-DRB5	major histocompatibility complex, class II, DR beta 5	−0.41	protein coding
				HLA-DQB2	major histocompatibility complex, class II, DQ beta 2	0.46	protein coding
				HLA-DQB1	major histocompatibility complex, class II, DQ beta 1	−0.51	protein coding
				HLA-DRB1	major histocompatibility complex, class II, DR beta 1	−0.42	protein coding
				HLA-DQB1-AS1	HLA-DQB1 antisense RNA 1	−0.54	antisense
12	9,799,363	rs58481733	4.4E-05	DDX12P	DEAD/H-box helicase 12, pseudogene	−0.57	pseudogene
15	40,374,582	rs73390668	2.9E-05	SRP14-AS1	SRP14 antisense RNA1 (head to head)	−0.31	lincRNA
	77,312,826	rs16968627	4.4E-05	RP11-797A18.4	RP11-797A18.4	−0.55	lincRNA

collagenase clostridium histolyticum (CCH) injection approved by Food and Drug Administration in 2010 is more effective and safer treatment option than surgical strategy, although the long-term recurrence rate of this CCH injection approach has not been fully determined yet [47]. Understanding the pathogenesis of DD is important to find out novel nonsurgical approach for DD treatment.

Although a recent study showed that DD is not a disease limited to European descent anymore [4], DD is still not commonly found in African Americans and Asians than in Europeans [55]. A total of 6 datasets were found in the public database (Table 1), due to the rareness of the disease. We carried out WGCNA to find co-expression modules based on Pearson correlation for gene expression similarities of DD. To construct the co-expression

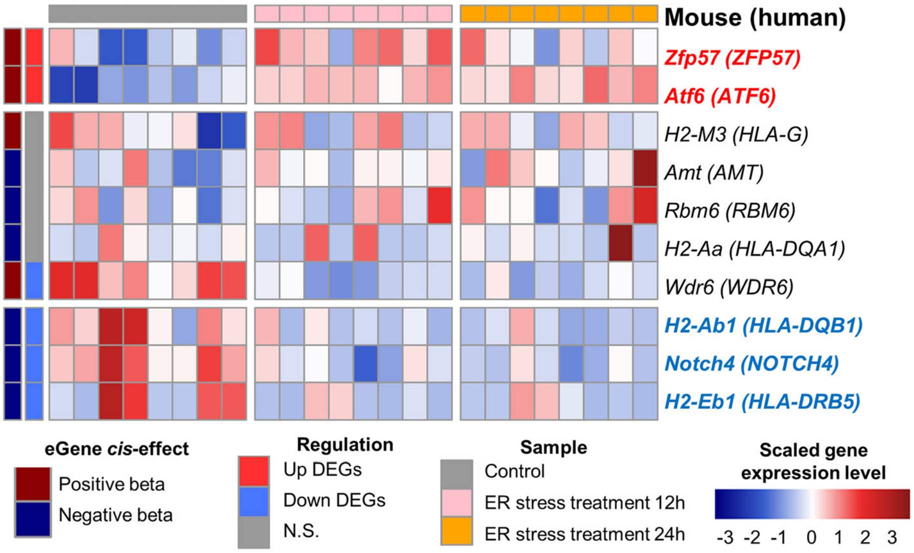

Fig. 4 A Heatmap showing gene expression levels of eGene using microarray data derived from ER stress-inducing agent (tunicamycin) treated MEFs. Color bars give information on tunicamycin treatment condition, DEGs, and direction of effect of SNP. Mouse gene symbol (human gene symbol) was represented by row names of the heatmap

network, we used GSE75152 which is the largest dataset used in this study (Table 1). Because WGCNA was performed based on Pearson correlation, more samples could lead to more robust the results [56]. We used signed WGCNA that created biologically meaningful modules than unsigned WGCNA [12]. After identifying co-expressed modules, module enrichment analysis was carried out. The red module genes were enriched in all 5 datasets (Fig. 2a and b) although the datasets were heterogeneous, consisting of two types of tissue sources and different microarray platforms with small sample size. The results suggested that red module genes were representative and robust genes for DD. Moreover, the WGCNA may tend to classify DEGs into certain modules due to the higher between-group variance which can translate to stronger correlations. That is why only red module was significant in the five independent datasets. The pre-filtering based on top 5000-most expressed microarray probes and limitation of 50 gene size of the modules were used for WGNCA. The prefiltering for reducing noise and the minimizing module size for module functional annotation both seemed to be appropriate because the red module genes were enriched in the biological process of unfolded protein-induced ER stress, which was not emphasized in previous studies.

GAMMA is a multiple-phenotype analysis method that examines an association between each SNP and multiple phenotypes or gene expression levels, while simultaneously correcting for population stratification utilizing linear mixed model [25]. We identified a total of 10 regulatory hotspots in muscle eQTL dataset of GTEx project using gene expression levels of DD signature. The identified regulatory hotspots included *HLA-DQA1*, *HLA-DQB1*, *HLA-DQB2*, *HLA-DRB1*, *HLA-DRB5*, and *ZFP57*. HLA is known to be associated with various human diseases, including rheumatoid arthritis, multiple sclerosis, Crohn's disease, type 1 diabetes (T1D), and HIV [57]. Intriguingly, HLA regions are known as the strongest genetic determinants in T1D, contributing up to 50% of the genetic risk to T1D susceptibility [58]. Diabetes is one of the known risk factors in DD and has been reported that approximately 20% of diabetic patients have DD [59]. DD is also involved in a constellation of musculoskeletal diseases affecting hand associated with diabetes [60]. T1D and DD seem to be inherited together rather than diabetes being an etiological risk factor for DD. Along with a previous report on association of HLA-DRB1 and HLA-DQB1 with T1D patients [61], it can be considered that having altered SNPs in HLA region might lead to genetic susceptibility in both DD and T1D, thus explaining the association of these two conditions.

Recently, *ZFP57* has been identified as a candidate gene contributing to HLA associated diseases including cancers, autoimmune diseases, and HIV [62]. *ZFP57* is located in HLA region of chromosome 6 and acts as a transcriptional factor that *trans*-regulates genomic imprinting, especially during development [63, 64]. Notably, transient neonatal diabetes (TND), an early onset T1D, are primarily caused by aberrant expression of imprinted genes due to mutations in *ZFP57* [65].

To date, alterations in extracellular matrix proteoglycan organization and collagen overproduction are the two main mechanisms proposed for the development of DD [9, 66]. Apart from these known pathological factors, we newly found the involvement of ER stress in DD. Accumulation of unfolded or misfolded proteins in ER under various pathophysiological conditions is defined as ER stress. ER stress and ER stress-responsive genes have been implicated in numerous diseases including neurological diseases, cancers, and diabetes [43]. It has been reported that expression of HLA molecules is decreased under ER stress conditions such as palmitate or glucose starvation and tunicamycin treatment [67]. Previous microarray results from MEFs identified that expression of *ZFP57* is altered after treatment with ER stress inducer [51]. Based on previous studies and our results, it can be concluded that ER stress-induced transcriptional changes in *ZFP57* and HLA molecules are implicated in the disease phenotype of DD. Thus, an in-depth investigation on the connection of ER stress with DD, possibly in relation to *ZFP57* and HLA, is required in order to understand pathophysiology of DD.

Because DD was classified as a rare disease amongst Eastern Asians, there were substantial difficulties in studying DD until recently. In such cases, an integrative analysis using genomic and transcriptomic data can serve as a powerful tool to study the pathogenesis, individual susceptibility, and progression of the disease. This study has some limitations. First, a further experimental step is needed to validate the identified targets. Second, we only focused on skeletal muscle data when identifying regulatory hotspots although subcutaneous fat and fibroblast are also known to be related to DD [47]. Despite these limitations, we successfully identified robust genetic markers of DD, suggesting that they may be potential therapeutic targets.

Conclusions

DD is a fibroproliferative disorder in thickening and contracting palmar fascia with unknown etiopathogenesis. In this study, we identified DD signature and potential cause of regulatory hotspots for DD based on integrative genomic and transcriptomic analysis using multiple phenotype regression analysis and WGCNA. Module enrichment analysis and classification analysis was used to determine the robustness of the identified markers. Finally, we identified MHC class II genes and ZFP57 were closely related to ER stress and UPR, suggesting that these genetic markers might be potential therapeutic targets for DD.

Additional files

Additional file 1: Identification of soft-thresholding power for co-expression network construction. (A) An analysis of scale free topology for picking an appropriate soft-thresholding power. (B) An analysis of the mean connectivity for picking an appropriate soft-thresholding power. (PPTX 277 kb)

Additional file 2: Point range plots showing the results of the enrichment of DEGs about each module via Fisher's exact test. (A) Odds ratios for the enrichment of DEGs compared with non-DEGs by one-color meta-analysis with 95% confidence intervals of odds. Because tan and midnight blue modules have no DEGs, the odds ratio is zero. (B) Odds ratios for the enrichment of DEGs compared with non-DEGs by two-color microarray analysis and 95% confidence intervals of the odds. (PPTX 242 kb)

Additional file 3: The identified red module genes (DD signature). (XLS 120 kb)

Additional file 4: The results of functional enrichment analysis in Fig. 2d. (XLS 38 kb)

Additional file 5: Violin plots showing the results of significant variant-gene associations related to rs2269423 in GTEx version 7. (PPTX 155 kb)

Abbreviations

AUC: Area under the curve; DD: Dupuytren's disease; ECM: Extracellular matrix; eQTL: Expression quantitative trait locus; ER: Endoplasmic reticulum; GO: Gene Ontology; GWAS: Genome-wide association studies; HLA: Human leukocyte alleles; KEGG: Kyoto Encyclopedia of Genes and Genomes; LOOCV: Leave-one-out cross-validation; MEF: Mouse embryonic fibroblast; MHC: Major histocompatibility complex; ROC: Receiver operating characteristic; SNP: Single nucleotide polymorphism; T1D: Type 1 diabetes; TOM: Topological overlap matrix; TSS: Transcription start site; UPR: Unfolded protein response; WGCNA: Weighted gene co-expression network analysis; ZFP57: ZFP57 zinc finger protein

Acknowledgements

Not applicable.

Funding

Publication of this article has been funded by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science and ICT) (Grant No. 2015R1C1A2A01054400).

Availability of data and materials

R code for the WGCNA analysis in this study is available at https://github.com/junghyunJJ/WGCNA_for_dupuytren.

About this supplement

This article has been published as part of *BMC Medical Genomics Volume 12 Supplement 5, 2019: Selected articles from the 8th Translational Bioinformatics Conference: Medical Genomics*. The full contents of the supplement are available online at https://bmcmedgenomics.biomedcentral.com/articles/supplements/volume-12-supplement-5.

Authors' contributions

JJ made substantial contributions to the conception and design of the acquisition, analysis, and interpretation of bioinformatics data and to drafting the manuscript. GWK contributed the acquisition, analysis, and interpretation of bioinformatics data and to drafting the manuscript. BL contributed the acquisition and analysis of bioinformatics. JWJ made contributions to the acquisition of GTEx data, interpretation of data, critical revisions for intellectual content, and gave final approval of the version to be published. WJ made contributions to the interpretation of data, critical revisions for intellectual content, and gave final approval of the version to be published. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Department of Life science, Dongguk University-Seoul, Seoul 04620, Republic of Korea. [2]Department of Pharmacology, College of Pharmacy, Kyung Hee University, 26 Kyungheedae-ro, Seoul 02447, South Korea. [3]Department of Computer Science and Engineering, Dongguk University-Seoul, Seoul 04620, South Korea.

Published: 11 July 2019

References

1. Black EM, Blazar PE. Dupuytren disease: an evolving understanding of an age-old disease. J Am Acad Orthop Surg. 2011;19(12):746–57.
2. Zerajic D, Finsen V. Dupuytren's disease in Bosnia and Herzegovina. An epidemiological study. BMC Musculoskelet Disord. 2004;5:10.
3. Mansur HG, Oliveira ER, Goncalves CB. Epidemiological analysis of patients with Dupuytren's disease. Rev Bras Ortop. 2018;53(1):10–4.
4. Lee KH, Kim JH, Lee CH, Kim SJ, Jo YH, Lee M, Choi WS. The epidemiology of Dupuytren's disease in Korea: a Nationwide population-based study. J Korean Med Sci. 2018;33(31):e204.
5. Burke FD, Proud G, Lawson IJ, McGeoch KL, Miles JN. An assessment of the effects of exposure to vibration, smoking, alcohol and diabetes on the prevalence of Dupuytren's disease in 97,537 miners. J Hand Surg Eur Vol. 2007;32(4):400–6.
6. Zhang AY, Fong KD, Pham H, Nacamuli RP, Longaker MT, Chang J. Gene expression analysis of Dupuytren's disease: the role of TGF-beta2. J Hand Surg Eur. 2008;33(6):783–90.
7. Lee LC, Zhang AY, Chong AK, Pham H, Longaker MT, Chang J. Expression of a novel gene, MafB, in Dupuytren's disease. J Hand Surg Am. 2006;31(2):211–8.
8. Ratkaj I, Bujak M, Jurisic D, Baus Loncar M, Bendelja K, Pavelic K, Kraljevic Pavelic S. Microarray analysis of Dupuytren's disease cells: the profibrogenic role of the TGF-beta inducible p38 MAPK pathway. Cell Physiol Biochem. 2012;30(4):927–42.
9. Forrester HB, Temple-Smith P, Ham S, de Kretser D, Southwick G, Sprung CN. Genome-wide analysis using exon arrays demonstrates an important role for expression of extra-cellular matrix, fibrotic control and tissue remodelling genes in Dupuytren's disease. PLoS One. 2013;8(3):e59056.
10. Alfonso-Rodriguez CA, Garzon I, Garrido-Gomez J, Oliveira AC, Martin-Piedra MA, Scionti G, Carriel V, Hernandez-Cortes P, Campos A, Alaminos M. Identification of histological patterns in clinically affected and unaffected palm regions in dupuytren's disease. PLoS One. 2014;9(11):e112457.
11. Becker K, Siegert S, Toliat MR, Du J, Casper R, Dolmans GH, Werker PM, Tinschert S, Franke A, Gieger C, et al. Meta-analysis of genome-wide association studies and network analysis-based integration with gene expression data identify new suggestive loci and unravel a Wnt-centric network associated with Dupuytren's disease. PLoS One. 2016;11(7):e0158101.
12. van Dam S, Vosa U, van der Graaf A, Franke L, de Magalhaes JP. Gene co-expression analysis for functional classification and gene-disease predictions. Brief Bioinform. 2018;19(4):575–92.
13. Dolmans GH, Werker PM, Hennies HC, Furniss D, Festen EA, Franke L, Becker K, van der Vlies P, Wolffenbuttel BH, Tinschert S, et al. Wnt signaling and Dupuytren's disease. N Engl J Med. 2011;365(4):307–17.
14. Ojwang JO, Adrianto I, Gray-McGuire C, Nath SK, Sun C, Kaufman KM, Harley JB, Rayan GM. Genome-wide association scan of Dupuytren's disease. J Hand Surg Am. 2010;35(12):2039–45.
15. Ng M, Thakkar D, Southam L, Werker P, Ophoff R, Becker K, Nothnagel M, Franke A, Nurnberg P, Espirito-Santo AI, et al. A genome-wide association study of Dupuytren disease reveals 17 additional variants implicated in fibrosis. Am J Hum Genet. 2017;101(3):417–27.
16. Nica AC, Dermitzakis ET. Expression quantitative trait loci: present and future. Philos Trans R Soc Lond Ser B Biol Sci. 2013;368(1620):20120362.
17. Consortium GT. The genotype-tissue expression (GTEx) project. Nat Genet. 2013;45(6):580–5.

18. Gilad Y, Rifkin SA, Pritchard JK. Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet. 2008;24(8):408–15.
19. Petretto E, Bottolo L, Langley SR, Heinig M, McDermott-Roe C, Sarwar R, Pravenec M, Hubner N, Aitman TJ, Cook SA, et al. New insights into the genetic control of gene expression using a Bayesian multi-tissue approach. PLoS Comput Biol. 2010;6(4):e1000737.
20. Grundberg E, Small KS, Hedman AK, Nica AC, Buil A, Keildson S, Bell JT, Yang TP, Meduri E, Barrett A, et al. Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat Genet. 2012;44(10):1084–9.
21. Consortium GT. Human genomics. The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015;348(6235):648–60.
22. Wang X, Korstanje R, Higgins D, Paigen B. Haplotype analysis in multiple crosses to identify a QTL gene. Genome Res. 2004;14(9):1767–72.
23. Cervino AC, Li G, Edwards S, Zhu J, Laurie C, Tokiwa G, Lum PY, Wang S, Castellani LW, Lusis AJ, et al. Integrating QTL and high-density SNP analyses in mice to identify Insig2 as a susceptibility gene for plasma cholesterol levels. Genomics. 2005;86(5):505–17.
24. Hillebrandt S, Wasmuth HE, Weiskirchen R, Hellerbrand C, Keppeler H, Werth A, Schirin-Sokhan R, Wilkens G, Geier A, Lorenzen J, et al. Complement factor 5 is a quantitative trait gene that modifies liver fibrogenesis in mice and humans. Nat Genet. 2005;37(8):835–43.
25. Joo JW, Kang EY, Org E, Furlotte N, Parks B, Hormozdiari F, Lusis AJ, Eskin E. Efficient and accurate multiple-phenotype regression method for high dimensional data considering population structure. Genetics. 2016;204(4):1379–90.
26. O'Reilly PF, Hoggart CJ, Pomyen Y, Calboli FC, Elliott P, Jarvelin MR, Coin LJ. MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. PLoS One. 2012;7(5):e34861.
27. Jung J, Mok C, Lee W, Jang W. Meta-analysis of microarray and RNA-Seq gene expression datasets for carcinogenic risk: an assessment of bisphenol a. Mol Cell Toxicol. 2017;13(2):239–49.
28. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.
29. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28(6):882–3.
30. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118–27.
31. Miller JA, Cai C, Langfelder P, Geschwind DH, Kurian SM, Salomon DR, Horvath S. Strategies for aggregating gene expression data: the collapseRows R function. BMC Bioinformatics. 2011;12:322.
32. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.
33. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545–50.
34. Jiao X, Sherman BT, Huang d W, Stephens R, Baseler MW, Lane HC, Lempicki RA. DAVID-WS: a stateful web service to facilitate gene/protein list analysis. Bioinformatics. 2012;28(13):1805–6.
35. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
36. Kuhn M. Building predictive models in R using the caret package. J Stat Softw. 2008;28(5):1–26.
37. Davis S, Meltzer PS. GEOquery: a bridge between the gene expression omnibus (GEO) and BioConductor. Bioinformatics. 2007;23(14):1846–7.
38. Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Tatusova TA, et al. Database resources of the National Center for biotechnology. Nucleic Acids Res. 2003;31(1):28–33.
39. Rehman S, Salway F, Stanley JK, Ollier WE, Day P, Bayat A. Molecular phenotypic descriptors of Dupuytren's disease defined using informatics analysis of the transcriptome. J Hand Surg Am. 2008;33(3):359–72.
40. Satish L, LaFramboise WA, O'Gorman DB, Johnson S, Janto B, Gan BS, Baratz ME, Hu FZ, Post JC, Ehrlich GD, et al. Identification of differentially expressed genes in fibroblasts derived from patients with Dupuytren's contracture. BMC Med Genet. 2008;1:10.
41. Vi L, Feng L, Zhu RD, Wu Y, Satish L, Gan BS, O'Gorman DB. Periostin differentially induces proliferation, contraction and apoptosis of primary Dupuytren's disease and adjacent palmar fascia cells. Exp Cell Res. 2009;315(20):3574–86.
42. Hurst LC, Badalamente MA, Hentz VR, Hotchkiss RN, Kaplan FT, Meals RA, Smith TM, Rodzvilla J, Group CIS. Injectable collagenase clostridium histolyticum for Dupuytren's contracture. N Engl J Med. 2009;361(10):968–79.
43. Hetz C, Chevet E, Harding HP. Targeting the unfolded protein response in disease. Nat Rev Drug Discov. 2013;12(9):703–19.
44. Tomasek JJ, Gabbiani G, Hinz B, Chaponnier C, Brown RA. Myofibroblasts and mechano-regulation of connective tissue remodelling. Nat Rev Mol Cell Biol. 2002;3(5):349–63.
45. Sudarsanam P, Cohen BA. Single nucleotide variants in transcription factors associate more tightly with phenotype than with gene expression. PLoS Genet. 2014;10(5):e1004325.
46. Yao C, Joehanes R, Johnson AD, Huan T, Liu C, Freedman JE, Munson PJ, Hill DE, Vidal M, Levy D. Dynamic role of trans regulation of gene expression in relation to complex traits. Am J Hum Genet. 2017;100(4):571–80.
47. Shih B, Bayat A. Scientific understanding and clinical management of Dupuytren disease. Nat Rev Rheumatol. 2010;6(12):715–26.
48. Brown JJ, Ollier W, Thomson W, Bayat A. Positive association of HLA-DRB1*15 with Dupuytren's disease in Caucasians. Tissue Antigens. 2008;72(2):166–70.
49. Walter P, Ron D. The unfolded protein response: from stress pathway to homeostatic regulation. Science. 2011;334(6059):1081–6.
50. Bak M, Boonen SE, Dahl C, Hahnemann JM, Mackay DJ, Tumer Z, Gronskov K, Temple IK, Guldberg P, Tommerup N. Genome-wide DNA methylation analysis of transient neonatal diabetes type 1 patients with mutations in ZFP57. BMC Med Genet. 2016;17:29.
51. Wu J, Rutkowski DT, Dubois M, Swathirajan J, Saunders T, Wang J, Song B, Yau GD, Kaufman RJ. ATF6alpha optimizes long-term endoplasmic reticulum function to protect cells from chronic stress. Dev Cell. 2007;13(3):351–64.
52. Rutkowski DT, Wu J, Back SH, Callaghan MU, Ferris SP, Iqbal J, Clark R, Miao H, Hassler JR, Fornek J, et al. UPR pathways combine to prevent hepatic steatosis caused by ER stress-mediated suppression of transcriptional master regulators. Dev Cell. 2008;15(6):829–40.
53. Arensdorf AM, Diedrichs D, Rutkowski DT. Regulation of the transcriptome by ER stress: non-canonical mechanisms and physiological consequences. Front Genet. 2013;4:256.
54. Lipman MD, Carstensen SE, Deal DN. Trends in the treatment of Dupuytren disease in the United States between 2007 and 2014. Hand (N Y). 2017; 12(1):13–20.
55. Izadpanah A, Viezel-Mathieu A, Izadpanah A, Luc M. Dupuytren contracture in the pediatric population: a systematic review. Eur J Pediatr Surg. 2015; 25(2):151–4.
56. Hardin J, Mitani A, Hicks L, VanKoten B. A robust measure of correlation between two genes on a microarray. BMC Bioinformatics. 2007;8:220.
57. Dendrou CA, Petersen J, Rossjohn J, Fugger L. HLA variation and disease. Nat Rev Immunol. 2018;18(5):325–39.
58. Ounissi-Benkalha H, Polychronakos C. The molecular genetics of type 1 diabetes: new genes and emerging mechanisms. Trends Mol Med. 2008; 14(6):268–75.
59. Noble J, Heathcote JG, Cohen H. Diabetes mellitus in the aetiology of Dupuytren's disease. J Bone Joint Surg Br. 1984;66(3):322–5.
60. Papanas N, Maltezos E. The diabetic hand: a forgotten complication? J Diabetes Complicat. 2010;24(3):154–62.
61. Thomson G, Valdes AM, Noble JA, Kockum I, Grote MN, Najman J, Erlich HA, Cucca F, Pugliese A, Steenkiste A, et al. Relative predispositional effects of HLA class II DRB1-DQB1 haplotypes and genotypes on type 1 diabetes: a meta-analysis. Tissue Antigens. 2007;70(2):110–27.
62. Plant K, Fairfax BP, Makino S, Vandiedonck C, Radhakrishnan J, Knight JC. Fine mapping genetic determinants of the highly variably expressed MHC gene ZFP57. Eur J Hum Genet. 2014;22(4):568–71.
63. Moalla M, Boussen K, Meddeb S, Gorgi Y, Khalfallah T, Hentati F, Ben Ayed H. Fatal acute pancreatitis in systemic lupus erythematosus. Ann Med Interne (Paris). 1991;142(3):230–2.
64. Takikawa S, Wang X, Ray C, Vakulenko M, Bell FT, Li X. Human and mouse ZFP57 proteins are functionally interchangeable in maintaining genomic imprinting at multiple imprinted regions in mouse ES cells. Epigenetics. 2013;8(12):1268–79.
65. Mackay DJ, Callaway JL, Marks SM, White HE, Acerini CL, Boonen SE, Dayanikli P, Firth HV, Goodship JA, Haemers AP, et al. Hypomethylation of multiple imprinted loci in individuals with transient neonatal diabetes is associated with mutations in ZFP57. Nat Genet. 2008;40(8):949–51.
66. Kozma EM, Olczyk K, Wisowski G, Glowacki A, Bobinski R. Alterations in the extracellular matrix proteoglycan profile in Dupuytren's contracture affect the palmar fascia. J Biochem. 2005;137(4):463–76.
67. Granados DP, Tanguay PL, Hardy MP, Caron E, de Verteuil D, Meloche S, Perreault C. ER stress affects processing of MHC class I-associated peptides. BMC Immunol. 2009;10:10.