Whole genome sequencing of Finnish type 1 diabetic siblings discordant for kidney disease reveals DNA variants associated with diabetic nephropathy

Jing Guo,[1,2] Owen J.L. Rackham,[2] Niina Sandholm,[3,4,5] Bing He,[1] Anne-May Österholm,[1,2] Erkka Valo, [3,4,5] Valma Harjutsalo,[3,4,5,6] Carol Forsblom,[3,4,5] Iiro Toppila,[3,4,5] Maikki Parkkonen,[3,4,5] Qibin Li,[7] Wenjuan Zhu,[7] Nathan Harmston,[2,11] Sonia Chothani,[2] Miina K. Öhman,[2] Eudora Eng,[2] Yang Sun,[2] Enrico Petretto,[2,8]* Per-Henrik Groop,[3,4,5,9]* Karl Tryggvason[1,2,10]*

[1]Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden

[2]Cardiovascular and Metabolic Disorders Programme, Duke-NUS Medical School, Singapore

[3]Folkhälsan Institute of Genetics, Folkhälsan Research Centre, Helsinki, Finland

[4]Abdominal Center, Nephrology, University of Helsinki and Helsinki University Hospital, 00290, Helsinki, Finland

[5]Research Programs Unit, Diabetes and Obesity, University of Helsinki, 00290, Helsinki, Finland

[6]National Institute for Health and Welfare, The Chronic Disease Prevention Unit, Helsinki, Finland

[7]BGI Genomics, BGI-Shenzhen, China

[8]MRC London Institute of Medical Sciences (LMS), Imperial College London, London, UK

[9]Department of Diabetes, Central Clinical School, Monash University, Melbourne, Victoria, Australia

[10]Division of Nephrology, Department of Medicine, Duke University, Durham, North Carolina, USA

[11]Yale-NUS College, National University of Singapore, Singapore

*Corresponding author, enrico.petretto@duke-nus.edu.sg

*Corresponding author, per-henrik.groop@helsinki.fi

*Corresponding author, karl.tryggvason@duke-nus.edu.sg

216 words in abstract, 3561 words in main text.

ABSTRACT

Background

Diabetic nephropathy (DN) is a major cause of morbidity and premature mortality of diabetic patients. Several genetic susceptibility loci have been documented, but no causative variants implying novel pathogenetic mechanisms have been elucidated.

Methods

We carried out whole-genome sequencing of a cohort of Finnish type 1 diabetes (T1D) siblings discordant for the presence (case) or absence (control) of DN, where the controls have had diabetes without complications for 15-37 years. We analyzed and annotated variants at genome, gene, and single nucleotide variant levels. We then replicated the associated variants, genes and regions in the FinnDiane replication cohort which includes 3,531 unrelated Finns with T1D.

Results

We observed protein-altering variants and an enrichment of variants in regions associated with the presence or absence of DN. Replication in FinnDiane confirmed variants in both regulatory and protein-coding regions. We also observed that DN associated variants, when clustered at the gene level, are enriched in a core protein interaction network of podocyte. These genes include protein kinases, i.e. protein kinase C isoforms epsilon and iota, and protein tyrosine kinase 2.

Conclusion

We carried out a comprehensive analysis of a DN cohort with T1D patients discordant for kidney disease and report our findings on a website http://dnc.systems-genetics.net/. The results shed light on variants and genes potentially causative or protective for DN. The results may facilitate analyses of other cohorts with DN.

INTRODUCTION

With the increase in the incidence of diabetes worldwide, complications like diabetic nephropathy, retinopathy, neuropathy, skin ulcers and amputations, have become a major global health and socio-economic threat. In addition to intensive blood glucose control[1], the only drugs providing a significant delay in progression of diabetic nephropathy (DN) are angiotensin-converting enzyme inhibitors (ACE-I) or angiotensin receptor blockers (ARB) that reduce intraglomerular pressure and efferent arteriolar vasoconstriction[2]. The molecular pathogenesis of DN is still poorly understood. Hyperglycemia, a major risk factor for complications, causes accumulation of toxic glucose derivatives, such as methylglyoxal, that bind covalently to the side chains of amino acids, particularly arginine and lysine, and also methionine and cysteine[3; 4]. Hyperglycemia alone is not sufficient to trigger the development of complications, as only 30-40 % of T1D individuals develop diabetic microangiopathy[1; 5; 6]. Independent familial studies have shown a trend of family aggregation of DN in different populations[7; 8], suggesting a genetic predisposition to DN. At least four metabolic pathways have been implicated in the development of complications: polyol flux, increased the formation of advanced glycation end products, hyperactivity of the hexosamine pathway and activation of protein kinase C (PKC) isoforms[4; 9; 10].

Genome-wide association studies (GWAS) and candidate gene approaches have identified several potential genomic loci for DN susceptibility[11], but no variants with a major effect on the risk of complications have been found, suggesting that DN is modulated by a number of variants in genes that cooperate within complex pathways. It is intriguing, however, that several independent genome-wide linkage analysis studies carried out on American Caucasians, Pima Indians, African Americans, and Finns, have identified the same DN susceptibility locus on chromosome 3q[12-15]. The complex interaction between genetics, risk factors such as hyperglycemia and environmental components makes it more challenging to find specific genes for DN using genetic association studies. To that end, it could be advantageous to search for DN susceptibility genes in populations such as Finns, a uniquely homogeneous European population[16] with the world's highest incidence of T1D[17; 18]. With a combination of founder effects and genetic isolation, the population has accumulated rare genetic traits referred to as the "Finnish Disease Heritage"[19]. In addition, Finland

has a good public health care system, including nationwide disease and treatment registries, which facilitates identification of patients and follow-up of their clinical records.

CONCISE METHODS

Experimental design

In order to search for DN susceptibility genes, we have assembled a cohort of Finnish T1D siblings with extreme phenotypes regarding the presence (case) or absence (control) of DN. This discovery cohort contained 76 T1D sibling pairs discordant (DSP) for DN, and three T1D families with three siblings (in total 80 cases and 81 controls). The samples came from two sources: the Finnish National Institute of Health and Welfare diabetes collections, as described elsewhere[15], and the Finnish Diabetic Nephropathy (FinnDiane) study[20]. Furthermore, 3,531 unrelated T1D individuals (1,344 cases and 2,187 controls) (**Figure 1a**) from FinnDiane were used for replication of findings made in the discovery cohort. The main clinical characteristics of patients in the discovery cohort are summarized in **Table 1**.

Study subjects

The discovery cohort consisted of sib-pairs and small families, whereas the replication cohort consisted of unrelated individuals, all having T1D. The renal status was based on the albumin excretion rate (AER) in a 24 hr urine collection or the albumin/creatinine ratio (ACR) in a random, spot urine collection. The presence of end-stage renal disease (ESRD) was defined according to whether patients were undergoing dialysis or had received a kidney transplant. DN was defined by (1) persistent macroalbuminuria (AER$\geq$300 mg/24 hr or ACR>30 mg/mmol) in two of three consecutive measurements or the presence of end-stage renal disease; and (2) the absence of clinical or laboratory evidence of nondiabetic renal or urinary-tract disease. Control status was defined by normoalbuminuria (AER < 30 mg/24 hr or ACR < 3 mg/mmol) despite duration of diabetes for at least 15 years [range 15-37]. In the discovery cohort, all study subjects had been diagnosed with T1D for at least 15 years, with the age at onset<30 years; in the replication cohort, age at diabetes onset was $\leq$40 years, with insulin dependence within one year after the diagnosis of diabetes (or age at diabetes onset $\leq$15 years). Controls in the replication cohort had minimum diabetes

duration of 15 years. The replication cohort included 2,187 controls with normal AER and 1,344 cases with macroalbuminuria and ESRD.

Ethical permits

All diabetic patients gave written, informed consent to participate in the study and the Ethical Committee of the Finnish National Public Institute, the Ethical Committee of the Helsinki and Uusimaa Health District, and Karolinska Institutet approved the protocol for the study. The transgene manipulation in zebrafish was approved by the local ethical committee (the North Stockholm district court).

Whole Genome Sequencing (WGS)

WGS was carried out on the discovery cohort using both Illumina HiSeq 2000 and Complete Genomics platforms. In order to evaluate the quality of the two different sequencing methods, we sequenced four discordant sib pairs with both platforms and compared the difference of the called variants across different platforms. The methods used for sequence alignment, quality control, variant calling and single nucleotide variant (SNV) annotation can be found in the Supplementary Methods.

Bioinformatics approaches for Whole Genome Sequencing (WGS) analysis

To fully utilize WGS data, we performed the association analysis with DN at three levels (**Figure 1b**): (A) genome-level analyses to study hot-spots of mutations and SNVs impacting regulatory regions; (B) gene-level aggregation tests to identify genes with DN-predisposing (or protecting) variants; and (C) SNV-level focusing on the PAVs (Protein-Altering Variants) present only in cases or only in controls and therefore, potentially causal or protective for DN. Each level of analysis uses different criteria for statistical significance; a brief summary of the statistical models and criteria used in each analysis is reported in **Table 2**. A global snapshot of all DN-associated variants and replicated in the FinnDiane cohort is provided in **Figure 2.**

Association test for single nucleotide variants (SNVs)

For each SNV, we tested the association with DN using four genetic models: (1) case-dominant, (2) case-recessive, (3) control-dominant and (4) control-recessive[21]. To this aim, we employed the Firth logistic regression method that accounts for rare variants and provides bias-reduction in case of small sample size analysis[22; 23] to assess the

significance of the association (P-value) in the discovery, replication and combined cohorts. Odds ratio (OR) and P-values for association were calculated using Firth's bias-reduced penalized-likelihood logistic regression method, and was implemented in the *R* package logistf[24]. The association test results were used to select SNVs for (B) gene-level test, and (C) SNV level test. The criteria for selection are different in (B) and (C), see details below.

(A) Genome level analysis

To identify genomic regions with frequent variants associated with DN in the 76 DSP, we set out to (1) identify regions that are significantly recurrently mutated (recurrently mutated regions or RMRs) compared to the distribution of mutations across the genome and (2) test each region for significant overrepresentation of mutations in DN cases or controls. For (1) we followed the method proposed by Weinhold et al[25]. Briefly, all mutations located within 50 base pairs (bp) of each other were merged using BEDTools[26] into hot-spot clusters and this procedure was repeated until no cluster was found within 50 bp of another cluster. The optimal cluster size was determined empirically given the observed distribution of mutations and their distance in the genome (data not shown). Clusters with less than three mutations were removed. For each cluster, a P-value was calculated using the negative binomial distribution, taking into account the length of the candidate hot-spot region, the number of mutations in the cluster and the background mutation rate (average mutation rate per sample) for the cluster that was estimated using the genome-wide expectation. The candidate hot-spot regions were selected for further analyses based on their P-value for significance and using a stringent Bonferroni correction for the number of regions tested (**Figure S1**). To identify RMR associated with DN (called DN-RMR), for each region we counted the number of mutations found in DN cases or controls and carried out Fisher's Exact Test (FET) to assess whether a mutation was overrepresented in either cases or controls. The Benjamini-Hochberg false discovery rate (FDR) correction to account for the number of regions tested by FET was applied to identify DN-RMR at the genome-wide level. For details of the analyses performed on transcription factor binding sites (TFBS), promoters and enhancers, please see Supplementary Methods.

(B) Gene-level analysis

We applied the adjusted SKAT test for familial data of dichotomous traits (F-SKAT[27]) on the multi-sib cohort (N=161). SNVs within a gene region were clustered together for the analysis. The gene region included variants in upstream 1000 bp, downstream 1000 bp, 3' UTR, 5' UTR, intron and exon. Only the protein-altering variants in the exonic region were included, i.e. nonsynonymous, stop-gain, stop-loss and splice site variants in RefSeq. We performed the gene-level aggregation test on three different sets of variants: (*i*) SNVs nominally associated with a case-control phenotype in the discovery cohort (OR>1.5 and nominal $P<0.05$, Firth test) irrespective of their MAF (Minor Allele Frequency); (*ii*) all SNVs with MAF<0.01 irrespective of their association with DN in discovery cohort; (*iii*) all SNVs with MAF<0.05 irrespective of their association with DN in discovery cohort. Genes that reached significance in the F-SKAT analysis (nominal F-SKAT $P<0.01$) have been annotated for functional enrichment test using Enrichr[28].

(C) SNV level analysis

To select significant SNVs for replication, we focused on the SNVs that are PAV (missense, nonsense, stop-loss, splicing site) or located in an exonic region in non-coding RNAs (ncRNAs). SNVs present only in cases or only in controls in dominant (≥3 individuals) or recessive (≥1 individual) were selected for replication in the FinnDiane cohort. For analysis of replication, an association test, as described above, was carried out on the discovery (T1D sib-pairs discordant for nephropathy, N=152), replication (FinnDiane, N=3,531) and combined cohorts (discovery plus replication, N=3,683). Methods for power calculation for SNVs association tests are described in Supplementary Methods.

Analysis of replication cohort

Genome-wide genotyping was performed on the Illumina HumanCoreExome Bead arrays 12-1.0, 12-1.1, and 24-1.0. The arrays include a core set of genome-wide variants plus an extensive set of exome variants. Data processing and quality control have been described earlier[29]. The genotype data were imputed with Micmac3 using the 1000 Genomes reference panel (Phase 3, version 5). SNVs with poor quality ($R^2<0.3$) were removed from analysis. Samples overlapping with the discovery cohort were excluded. Candidate SNVs were extracted from the GWAS imputation data and the number of

genotypes were counted for controls and cases based on the most likely genotypes using SNPTest.

To evaluate the false positive rate of replication at the SNV level, we performed an empirical test in 3 steps: (1) select a random set of SNVs from discovery PAVs; (2) test association on this random set, and count the number of significant variants (OR>1.5, P<0.05); (3) repeat the steps 10,000 times to assess the false positive rate. For gene level replication, we could not apply F-SKAT to FinnDiane data, since that replication cohort does not contain familial data. Instead, we used SKAT[30] (Sequence Kernel Association Test) on the same SNV set (if found in FinnDiane) as we used for F-SKAT in the discovery cohort. For replication in the genome level DN-RMR test, we extracted all SNVs within the RMR regions defined by tests of the discovery cohort, and tested enrichment of variants in cases or controls for each region using two-tailed FET, then corrected by Bonferroni P<0.01.

RESULTS

Variants detected by whole genome sequencing (WGS)

We evaluated the sequencing quality by sequencing four sib pairs with both Complete Genomics and Illumina HiSeq 2000 platforms. The concordance rate across the two platforms for all eight individuals was 98.8% (**Table S1a-c**).

WGS of the discovery cohort revealed 12 million SNVs (**Table S2**) and >6 million short insertions and deletions (indels) (**Table S3**). Here, we focused on genetic variants functionally associated with the DN phenotype, *i.e.* variants affecting gene regulatory elements and/or coding regions.

Genome-level analysis

We analyzed the complete genome sequences of the 76 T1D discordant sibling pairs to systematically identify genomic regions that are recurrently mutated and overrepresented in the DN cases or controls (*i.e.*, individuals with T1D but without DN). A similar approach has been employed in studies of genome-wide noncoding regulatory mutations in cancer[25], and is based on (*i*) genome-wide "hot-spot" mutation analysis to identify small regions with frequent (recurrent) mutations and analysis of clusters of recurrent mutations, (*ii*) DNA variants impacting TFBS, and (*iii*) annotated regulatory regions (e.g., promoters and enhancers).

For the genome-wide "hot-spot" mutation analysis, we identified a total of 850,137 RMR. Each RMR represents a genomic locus enclosing a cluster of variants within 50 bp of each other, and genome-wide RMRs have a median size of 436 (min 4 to max 37,433) bp. Each identified RMR is significantly recurrently mutated compared to a random distribution of mutations across the genome (Bonferroni corrected $P<3.7\times10^{-5}$, **Figure S1**).

We first tested whether these RMRs are significantly over-represented in DN cases or in controls. After correcting for the number of total RMRs analyzed, we detected 732 RMRs that are over-represented in either DN cases or in controls at FDR <5%, thereby identifying a set of RMRs associated with diabetic nephropathy (hereafter, DN-RMR) (**Figure 3a** and **Table S4**). 141 of these DN-RMRs (19.26%) were replicated in the FinnDiane cohort (Bonferroni $P<0.01$). 458 (63%) DN-RMRs are intergenic, whereas 274 (37%) overlap with 194 annotated genes. When compared with the whole set of RMRs identified at the genome-wide level in the discovery cohort, the DN-RMRs more frequently overlap with exons, introns, 3' and 5' UTRs, enhancers and gene promoter regions (**Figure 3a**). This suggests that DN-associated clusters of mutations are more likely to impact exons and regulatory regions than the RMRs that are not associated with DN. The genes overlapping with DN-RMRs are significantly enriched for several canonical KEGG pathways relevant to the pathobiology of DN ($P<0.01$), including ECM-receptor interaction, focal adhesion and type I diabetes (**Figure 3b**). These pathways have several genes in common, suggesting that the identified DN-RMRs affect multiple genes interacting across overlapping functional pathways (**Figure 3b**). Interestingly, *COL4A1* and *COL4A2*, which encode the most prominent non-GBM collagens were shown to be associated with DN, as previously reported[31,32]. Both genes were enriched for variants in intronic regulatory regions, but their possible role in the pathogenesis of DN remains obscure, especially as no exonic mutations were different between cases and controls in these genes in the discovery cohort.

As the second genome-level approach, in order to investigate the potential regulatory impact of DN-associated variants, we retrieved and annotated experimentally derived TFBS data from a large repository of ChIP-seq data representing DNA binding data for 237 transcription factors (TFs)[33]. Within each TFBS region, we tested whether there was a significant over-representation of variants in DN-ascertained cases or in controls (**Figure 3c**). Overall, we found more variants impacting

TFBS in controls than in cases, and in some instances these variants are present only in controls and across multiple families. By pooling results for TFs over their corresponding TFBSs, we identified 40 TFs with significantly different variant frequencies between cases and controls (Benjamini-Hochberg corrected $P<0.05$) and 6 (out of 20 TFs for which genotype data was available in the replication cohort) were replicated in FinnDiane (Bonferroni $P<0.01$) (**Table S5**). The 40 TFs were enriched for pathways relevant to the pathophysiology of DN (**Figure 3d**). These include the epidermal growth factor receptor (EGFR)-dependent endothelin signaling (implicated in the development and progression of renal fibrosis and hypertrophy of the glomerular basement membrane), which has been proposed for targeting by endothelin antagonist therapy in DN[34]. We also found the structurally related transmembrane receptors belonging to the receptor tyrosine kinase superfamily (e.g. ErbB1) that are involved in the development and progression of DN[35]. Of note, variants in *ERBB4* have previously been suggested to be associated with DN[18; 36], even though the causal variants were not identified.

The third genome-level analysis approach, was to study annotated regulatory regions in the genome (gene promoters and enhancers) which are derived from the FANTOM5 database[37] and were further supported by ENCODE[38] histone modification data, and to test whether variants in these regions were significantly overrepresented in DN cases or controls. We found significant enrichment (FDR<0.05) for DN-associated variants in 270 promoter regions (±1kb around the annotated gene transcription start site (TSS)), 68 (25.2%) were replicated in the FinnDiane cohort (Bonferroni $P<0.01$) (**Table S6a**). We also found significant enrichment (FDR<0.05) for DN-associated variants within ±1kb of 44 predicted enhancers (**Table S6b**). DN-associated variants in five enhancers were replicated in the larger FinnDiane cohort (Bonferroni $P<0.01$). We further prioritized candidate genes within these replicated enhancers using data related to topologically associated domains, epigenetic regulation, and transcriptome analysis of DN in human[39] (**Table S7**).

Not surprisingly, in a few cases distinct genome-level analyses prioritized the same gene locus. For instance, *ALOX5*, encoding arachidonate 5-lipoxygenase (a member of the lipoxygenase gene family regulating metabolites of arachidonic acid), was found to overlap with an intragenic DN-RMR spanning 4,724 bp and has DN-associated variants in two predicted enhancers and in its annotated promoter region, suggesting potential enhancer-promoter interaction[40] (**Figure 3e**). A role for

lipoxygenase inhibitors in DN has been proposed in the rat[41] and 12-lipoxygenase is increased in glucose-stimulated cultured mesangial cells and in kidney of rat DN model[42]. Furthermore, it has been shown that 5-lipoxygenase contributes to degeneration of retinal capillaries in a mouse model of diabetic retinopathy, suggesting a proinflammatory role of 5-lipoxygenase in the pathogenesis of DN[43].

Gene-level analysis

To investigate the aggregated gene-level contribution of multiple SNVs, we used the F-SKAT framework (Sequence Kernel Association Test adjusted for familial data of dichotomous traits)[27]. We tested different sets of SNVs that were aggregated at the gene level (see **Methods**). We only found a few genes that reached the nominal significance level of $P<0.05$ by testing on the rare variants (**Table S8**), and found no associations with any relevant functional pathways or networks. Alternatively, we first identified 28,237 SNVs (within 3,745 genes) that were nominally associated with DN susceptibility or protection (OR >1.5, $P<0.05$). Then we gathered all DN-associated SNVs that were within upstream 1000 bp, downstream 1000 bp, UTR regions, intronic and PAVs, and tested their accumulative effect on each gene. We found 206 genes that reach a significance level of $P<0.01$ in the F-SKAT analysis (**Table S9a-b**).

To investigate the potential function of the SNVs in the 206 genes detected by F-SKAT, we analyzed these SNVs using a recent expression quantitative trait locus (eQTL)[44] dataset from the glomerulus and tubulointerstitium of subjects with nephrotic syndrome. We found that these F-SKAT significant genes are more likely to be under *cis*-acting regulation in the glomeruli of nephrotic syndrome patients than genes with non-significant F-SKAT (OR=3.84, $P=2.2\times10^{-16}$). This suggests that the SNVs contributing to the gene-level association with DN (detected by F-SKAT) may exert their pathological function by regulating gene expression in the kidney. We then used Enrichr[28] to test for functional enrichment in the 206 genes identified by F-SKAT, and observed the only significant enrichment for protein-protein interactions in the podocyte network expanded by STRING (XPodNet[45]), (22/808 genes, enrichment $P=0.0045$, Wikipathways 2016). The F-SKAT associated genes within the core XPodNet are shown in **Figure 4a** and **Table S10**. The genes in this sub-network of XPodNet are enriched for several pathways, including focal adhesion and insulin signaling (**Figure 4b** and **Table S11**). The top candidate gene from the F-SKAT test is the protein kinase C epsilon gene (*PRKCE*) (F-SKAT $P=0.0004$), with multiple intronic

DN-associated SNVs that overlap with predicted regulatory regions (**Figure 4c, Table S9b**). Protein kinases *PRKCE*, *PTK2* (F-SKAT *P*=0.0037) and *PRKCI* (F-SKAT *P*=0.0085) are part of a "core protein-interaction network" representing proteins essential for podocyte function. These genes are particularly interesting as PKCs have been implicated in the pathogenesis of DN[10]. However, specific inhibitors for those three PKCs have not yet been developed to our knowledge.

Furthermore, we tested the 206 genes which were found to be significant using F-SKAT in the replication cohort. This replication is limited by the less numerous SNVs in FinnDiane compared with the discovery cohort (2,316 out of 3,755 SNVs), which also does not include family data. Therefore, we applied SKAT using only the same SNVs used by F-SKAT in the discovery cohort. This is a rather stringent replication approach, as it tests for both the genes and the specific SNVs that were found to be associated with DN in the discovery cohort. Out of the 206 genes tested, only 120 genes were found with at least one F-SKAT SNV, and nine genes passed the nominal criteria $P<0.05$, including a protein kinase gene *PTK2* (**Table S9c**). The replicated genes are highlighted in **Figure 2**.

Analyses of protein-altering variants (PAVs)

It has been estimated that about 85% of mutations underlying Mendelian diseases reside in coding sequences or at exon-intron borders[46; 47]. Numerous reports have described rare but highly penetrant exon mutations in Mendelian disease[48,49], and it is likely that such mutations also frequently contribute to complex disease phenotypes. Our initial exon variant analyses have focused on 53,449 PAVs (nonsynonymous, stop-gain, stop-loss and splice site variants **Table S2**) that were exclusively found in cases or controls in the 76 T1D DSP and are associated with DN-susceptibility or DN-protection. The PAVs were tested for association with DN in the FinnDiane cohort using a recessive disease model for the homozygous variants detected in ≥1 cases/controls in the discovery cohort and by a dominant model for the heterozygous SNVs detected in ≥3 cases/controls in the discovery cohort. The 47 PAVs identified in the recessive model were replicated in FinnDiane ($P<0.05$, OR>1.5). By using a permutation based-strategy (see **Methods**), we estimated the probability that these 47 PAVs are replicable by chance alone is only 2.3%. However, the false positive rate in the dominant model is estimated to be high (**Figure S2**). Therefore, only candidate SNVs that were replicated

in the recessive model are reported (top SNVs in **Table 3**, and in full in **Table S12)**. Some of the top-replicated PAVs are within genes that have previously been linked to renal disease, implying a potential role in DN, e.g. mutations in WDR73 have been reported to be responsible for late-onset steroid-resistant nephrotic syndrome[50]. We also studied the gene function of ABTB1, where we found the only case-only homozygous mutation that is truncating the protein. Zebrafish knockout of the gene displayed a phenotype that is specific for kidney damage (Supplementary Method, Supplementary Results, **Figure S3**).

Hyperglycemia causes an increase in intracellular ROS that leads to increase in glucose derivatives, such as methylglyoxal, that readily react with amino groups of protein amino acid residues, particularly arginine, lysine, cysteine and methionine[51]. Here, PAVs altering amino acid codons to arginine were found to be significantly less represented in the set of mutations detected in controls only as compared with all PAVs (OR = 0.66, 95% CI [0.43-0.97], *P*=0.03, **Figure S4**). No other classes of mutations leading to individual amino acid(s) substitution showed significant over-representation/depletion in either cases or controls.

Power Calculation

To estimate the statistical power for detecting association in our sibship discovery cohort, we used a method described by Li et.al[52]. We estimated the power assuming different levels of penetrance (**Table S14a**). Our sample size of 76 DSP reaches >80% power to detect significant associations ($P<4.11x10^{-9}$) for rare variants with high penetrance (penetrance=90%, MAF=0.01). Furthermore, we estimate the power for the replication study. Similar to a previous report[18], our replication cohort (N=3,531) reaches at least 80% power detect common variants with high OR (OR=2, MAF=0.05 in the dominant model; OR=5, MAF=0.2 in the recessive model).

DISCUSSION

To the best of our knowledge, this is the first study where WGS has been applied in a search for genomic variants specifically associated with the presence or absence of DN in T1D patients. The challenge with finding susceptibility genes for diabetes complications is that one searches for mutations that only cause complications if the individual has hyperglycemia. We assembled a unique discovery cohort of T1D siblings

from the highly homogeneous Finnish population and replicated key findings in a larger cohort of unrelated T1D Finns. This enabled a direct comparison of whole-genome sequences in individuals with extreme phenotypes, *i.e.* T1D with progressive DN on one hand, and siblings with no complications for at least 15 years [range 15-37] on the other. The results provide a unique catalogue of DNA variants in Finns.

We have developed a comprehensive panel of multiple bioinformatic approaches to detect genetic pre-deposition of DN in the discovery sib cohort. The SNVs approach, which evaluates PAVs that are present only in cases or controls, focuses on the potential protein function in DN. The kernel test (F-SKAT) prioritizes genes with multiple associated variants within the gene region, and hypothesizes that the accumulated burden leads to malfunction of the gene. The genomic approach includes variants in other genome regions and could potentially detect functionally important regions. These approaches identified different individual variants, genes and regulatory regions that are potentially involved in DN susceptibility.

Although the discovery cohort only consisted of 161 individuals with T1D, together with the FinnDiane replication cohort, we show that they can provide enough power to identify and replicate potential causative and protective mutations for DN. Here, the use of discordant T1D sib-pairs for DN has been pivotal to increase power to identify variants associated with DN susceptibility of protection.

We have also studied the replication of candidates and report candidates with robust signals for each analysis approach. However, while the replication of SNVs is commonly used for GWAS where it applies on the same loci, the replication for statistical tests which involve multiple loci, i.e. RMR, F-SKAT and TFBS have limitations that need to be taken into consideration. For replication of F-SKAT in FinnDiane, about one-third of F-SKAT SNVs cannot be found by array genotyping plus imputation. Thus, the number of replicable genes is limited (120/206), and within each gene, SNVs are also less represented. Additionally, the use of a different statistical model (SKAT versus F-SKAT) might also introduce a bias in the replication test. The constraints caused by limited genotypes in the replication cohort also apply to the RMR and TFBS replications. The data-driven detection of RMR requires comprehensive SNV data (i.e. WGS data). Using a panel of predefined genotyped SNPs (i.e. SNP array data), even if the panel is large and supported by imputation, might introduce a considerable bias in the replication of DN-associated RMRs.

The analyses of the discovery cohort led to the identification of several novel DN candidate genes in Finns, including *PRKCE, PTK1, PRKCI, ABTB1,* and *ALOX5* as discussed above. The significant association of three protein kinase genes with DN is intriguing, as the large PKC protein family has long been associated with diabetes complications[4; 10]. Several clinical trials have been carried out for the treatment of DN with Ruboxistaurine, a compound that inhibits PRKC-β[53]. This suggests that hyperglycemia-driven PKC activation, particularly that of the β isoform, may underlie endothelial dysfunction. In the present study, we identified two novel isoforms of protein kinase C family (i.e. epsilon and iota) that have not been previously linked to DN. The results strongly support and extend previous hypotheses that protein kinases, especially protein kinase C family, play a role in the pathogenesis of DN, and could be attractive novel targets for the development of PKC inhibitors for DN treatment.

DN is a disorder characterized by hyperglycemia, which can lead to non-enzymatic glycation of amino acids and formation of advanced glycation end products in both intracellular and extracellular proteins[4; 9; 54]. It can be speculated that glycation of amino acids in functionally important regions of the protein can affect functionality of the protein or promote their degradation[3]. Amino acids that are most prone to become non-enzymatically glycated by methylglyoxal and other carbonyls, are arginine, and to a lesser extent lysine[55], cysteine and methionine[4; 9]. Our study highlighted mutated arginine codons as being of special interest when considering mutations that can cause pathogenic non-enzymatic glycation of proteins and consequent development of DN.

Previously reported genes/regions associated with DN were not strongly replicated in our discovery cohort (**Table S15**), suggesting that different sets of loci/variants contribute to the pathogenesis of DN. However, despite the scarce replication of previous loci in our cohort, we report the identification of variants/genes in functional pathways relevant to the pathobiology of DN, many of which have been previously reported (e.g. EGFR-dependent endothelin signaling[34] and PodNet[45]).

Overall, we have performed a comprehensive study on the genetics of a unique T1D Finnish cohort of siblings discordant for nephropathy using WGS data. Although the sample size is relatively small and the association test for SNV cannot reach the genome-wide significance ($P<4\times10^{-9}$), efforts were made to optimize the test model to fit for the specific sibship cohort, and the top-listed SNVs were replicated (when applicable) in larger Finnish cohort. Novel potential DN susceptibility genes and

regulatory variants are promoted in hope to merit further investigation in other populations and animal models.

Availability of genome analysis results

All genetic association data presented here are made freely accessible via http://dnc.systems-genetics.net.

DESCRIPTION OF SUPPLEMENTAL DATA

Supplemental Data include supplementary methods, results, web resources used in the study, five figures and fifteen tables.

DISCLOSURES

P-H G has received investigator-initiated research grants from Eli Lilly and Roche, is an advisory board member for AbbVie, Astellas, AstraZeneca, Boehringer Ingelheim, Cebix, Eli Lilly, Janssen, Medscape, Merck Sharp & Dohme, Mundipharma, Nestle, Novartis, Novo Nordisk and Sanofi; and has received lecture fees from AstraZeneca, Boehringer Ingelheim, Eli Lilly, Elo Water, Genzyme, Merck Sharp & Dohme, Medscape, Novartis, Novo Nordisk, PeerVoice and Sanofi.

ACKNOWLEDGEMENTS

The work was supported by grants from the Novo Nordisk Foundation, Knut and Alice Wallenberg, Söderberg's, and Hedlund's Foundations, the Swedish Medical Research Council, the Swedish Foundation for Strategic Research, Sigrid Juselius Foundation, Folkhälsan Research Foundation, the Wilhelm and Else Stockmann Foundation, the Liv och Hälsa Foundation, Helsinki University Central Hospital Research Funds (EVO), JDRF (17-2013-7 (DNCRI)), European Foundation for the Study of Diabetes (EFSD) Young Investigator Research Award funds, and the Academy of Finland (Grants 38387, 46558, 275614, 299200, and 316664). This work has also been supported by Singapore grants from NMRC grant (NMRC/STaR/0010/2012) and the Singapore National Medical Research Council (NMRC/OFLCG/001/2017). We are grateful to the physicians, nurses, and researchers in the FinnDiane study group and at each center participating in the collection of patients. We thank Leena Ollitervo and Maire Jarva for technical assistance on DNA extraction. We also acknowledge Dr. Jaakko Tuomilehto for the original collection of samples. The computational analyses were

493 performed on resources provided by SNIC through Uppsala Multidisciplinary Center
494 for Advanced Computational Science (UPPMAX) under Project b2013027, and the
495 High-Performance Computing Cluster in Duke-NUS computational center.
496

Supplementary Table of Content

Supplementary Figures

Supplementary Figure 1. Manhattan plot of the recurrently mutated regions (RMR) identified genome-wide in the 76 T1D discordant sibling pairs.

Supplementary Figure 2. Estimation of replication false positive rate on protein-altering variants (PAV) in FinnDiane cohort.

Supplementary Figure 3. Expression and functional analysis of Abtb1.

Supplementary Figure 4. Forest plots showing that protein-altering variants (PAVs) altering amino acid codons for arginine (Arg) are less represented in the set of mutations detected in controls as compared with all protein altering mutations (indicated with ◆).

Supplementary Figure 5. Chromosome 3q21 locus for DN susceptibility that was previously identified.

Web Resources

Reference

References

1. Group, U.P.D.S.U. (1998). Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). UK Prospective Diabetes Study (UKPDS) Group. Lancet 352, 837-853.
2. Braam, B., and Koomans, H.A. (1996). Renal responses to antagonism of the renin-angiotensin system. Curr Opin Nephrol Hypertens 5, 89-96.
3. Westwood, M.E., Argirov, O.K., Abordo, E.A., and Thornalley, P.J. (1997). Methylglyoxal-modified arginine residues--a signal for receptor-mediated endocytosis and degradation of proteins by monocytic THP-1 cells. Biochim Biophys Acta 1356, 84-94.
4. Brownlee, M. (2001). Biochemistry and molecular cell biology of diabetic complications. Nature 414, 813-820.
5. Thomas, M.C., Groop, P.H., and Tryggvason, K. (2012). Towards understanding the inherited susceptibility for nephropathy in diabetes. Current opinion in nephrology and hypertension 21, 195-202.
6. Thomas, M.C., Brownlee, M., Susztak, K., Sharma, K., Jandeleit-Dahm, K.A., Zoungas, S., Rossing, P., Groop, P.H., and Cooper, M.E. (2015). Diabetic kidney disease. Nat Rev Dis Primers 1, 15018.
7. Borch-Johnsen, K., Norgaard, K., Hommel, E., Mathiesen, E.R., Jensen, J.S., Deckert, T., and Parving, H.H. (1992). Is diabetic nephropathy an inherited complication? Kidney Int 41, 719-722.
8. Seaquist, E.R., Goetz, F.C., Rich, S., and Barbosa, J. (1989). Familial clustering of diabetic kidney disease. Evidence for genetic susceptibility to diabetic nephropathy. N Engl J Med 320, 1161-1165.
9. Giacco, F., and Brownlee, M. (2010). Oxidative stress and diabetic complications. Circ Res 107, 1058-1070.
10. Geraldes, P., and King, G.L. (2010). Activation of protein kinase C isoforms and its impact on diabetic complications. Circ Res 106, 1319-1331.
11. Dahlstrom, E., and Sandholm, N. (2017). Progress in Defining the Genetic Basis of Diabetic Complications. Curr Diab Rep 17, 80.
12. Moczulski, D.K., Rogus, J.J., Antonellis, A., Warram, J.H., and Krolewski, A.S. (1998). Major susceptibility locus for nephropathy in type 1 diabetes on chromosome 3q: results of novel discordant sib-pair analysis. Diabetes 47, 1164-1169.
13. Imperatore, G., Hanson, R.L., Pettitt, D.J., Kobes, S., Bennett, P.H., and Knowler, W.C. (1998). Sib-pair linkage analysis for susceptibility genes for microvascular complications among Pima Indians with type 2 diabetes. Pima Diabetes Genes Group. Diabetes 47, 821-830.
14. Bowden, D.W., Colicigno, C.J., Langefeld, C.D., Sale, M.M., Williams, A., Anderson, P.J., Rich, S.S., and Freedman, B.I. (2004). A genome scan for diabetic nephropathy in African Americans. Kidney Int 66, 1517-1526.
15. Österholm, A.M., He, B., Pitkäniemi, J., Albinsson, L., Berg, T., Sarti, C., Tuomilehto, J., and Tryggvason, K. (2007). Genome-wide scan for type 1 diabetic nephropathy in the Finnish population reveals suggestive linkage to a single locus on chromosome 3q. Kidney Int 71, 140-145.

16. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285-291.
17. Harjutsalo, V., Sund, R., Knip, M., and Groop, P.H. (2013). Incidence of type 1 diabetes in Finland. JAMA 310, 427-428.
18. Sandholm, N., Van Zuydam, N., Ahlqvist, E., Juliusdottir, T., Deshmukh, H.A., Rayner, N.W., Di Camillo, B., Forsblom, C., Fadista, J., Ziemek, D., et al. (2017). The Genetic Landscape of Renal Complications in Type 1 Diabetes. J Am Soc Nephrol 28, 557-574.
19. Peltonen, L., Jalanko, A., and Varilo, T. (1999). Molecular genetics of the Finnish disease heritage. Hum Mol Genet 8, 1913-1923.
20. Thorn, L.M., Forsblom, C., Fagerudd, J., Thomas, M.C., Pettersson-Fernholm, K., Saraheimo, M., Waden, J., Ronnback, M., Rosengard-Barlund, M., Bjorkesten, C.G., et al. (2005). Metabolic syndrome in type 1 diabetes: association with diabetic nephropathy and glycemic control (the FinnDiane study). Diabetes Care 28, 2019-2024.
21. Clarke, G.M., Anderson, C.A., Pettersson, F.H., Cardon, L.R., Morris, A.P., and Zondervan, K.T. (2011). Basic statistical analysis in genetic case-control studies. Nat Protoc 6, 121-133.
22. Wang, X. (2014). Firth logistic regression for rare variant association tests. Front Genet 5, 187.
23. Ma, C., Blackwell, T., Boehnke, M., Scott, L.J., and Go, T.D.i. (2013). Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. Genet Epidemiol 37, 539-550.
24. Georg Heinze, M.P. (2016). logistf: Firth's Bias-Reduced Logistic Regression. R package version 1.22.
25. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C., and Lee, W. (2014). Genome-wide analysis of noncoding regulatory mutations in cancer. Nat Genet 46, 1160-1165.
26. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841-842.
27. Yan, Q., Tiwari, H.K., Yi, N., Gao, G., Zhang, K., Lin, W.Y., Lou, X.Y., Cui, X., and Liu, N. (2015). A Sequence Kernel Association Test for Dichotomous Traits in Family Samples under a Generalized Linear Mixed Model. Hum Hered 79, 60-68.
28. Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res 44, W90-97.
29. Syreeni, A., Sandholm, N., Cao, J., Toppila, I., Maahs, D.M., Rewers, M.J., Snell-Bergeon, J.K., Costacou, T., Orchard, T.J., Caramori, M.L., et al. (2019). Genetic Determinants of Glycated Hemoglobin in Type 1 Diabetes. Diabetes 68, 858-867.
30. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 89, 82-93.

31. Pihlajaniemi, T., Myllyla, R., Kivirikko, K.I., and Tryggvason, K. (1982). Effects of streptozotocin diabetes, glucose, and insulin on the metabolism of type IV collagen and proteoglycan in murine basement membrane-forming EHS tumor tissue. J Biol Chem 257, 14914-14920.
32. Mason, R.M., and Wahab, N.A. (2003). Extracellular matrix metabolism in diabetic nephropathy. J Am Soc Nephrol 14, 1358-1373.
33. Griffon, A., Barbier, Q., Dalino, J., van Helden, J., Spicuglia, S., and Ballester, B. (2015). Integrative analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory landscape. Nucleic Acids Res 43, e27.
34. Barton, M. (2010). Therapeutic potential of endothelin receptor antagonists for chronic proteinuric renal disease in humans. Biochim Biophys Acta 1802, 1203-1213.
35. Zhang, M.Z., Wang, Y., Paueksakon, P., and Harris, R.C. (2014). Epidermal growth factor receptor inhibition slows progression of diabetic nephropathy in association with a decrease in endoplasmic reticulum stress and an increase in autophagy. Diabetes 63, 2063-2072.
36. Sandholm, N., Salem, R.M., McKnight, A.J., Brennan, E.P., Forsblom, C., Isakova, T., McKay, G.J., Williams, W.W., Sadlier, D.M., Makinen, V.P., et al. (2012). New susceptibility loci associated with kidney disease in type 1 diabetes. PLoS Genet 8, e1002921.
37. Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al. (2014). An atlas of active enhancers across human cell types and tissues. Nature 507, 455-461.
38. Consortium, E.P. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57-74.
39. Woroniecka, K.I., Park, A.S., Mohtat, D., Thomas, D.B., Pullman, J.M., and Susztak, K. (2011). Transcriptome analysis of human diabetic kidney disease. Diabetes 60, 2354-2369.
40. Nolis, I.K., McKay, D.J., Mantouvalou, E., Lomvardas, S., Merika, M., and Thanos, D. (2009). Transcription factors mediate long-range enhancer-promoter interactions. Proc Natl Acad Sci U S A 106, 20222-20227.
41. Ma, J., Natarajan, R., LaPage, J., Lanting, L., Kim, N., Becerra, D., Clemmons, B., Nast, C.C., Surya Prakash, G.K., Mandal, M., et al. (2005). 12/15-lipoxygenase inhibitors in diabetic nephropathy in the rat. Prostaglandins Leukot Essent Fatty Acids 72, 13-20.
42. Kang, S.W., Adler, S.G., Nast, C.C., LaPage, J., Gu, J.L., Nadler, J.L., and Natarajan, R. (2001). 12-lipoxygenase is increased in glucose-stimulated mesangial cells and in experimental diabetic nephropathy. Kidney Int 59, 1354-1362.
43. Gubitosi-Klug, R.A., Talahalli, R., Du, Y., Nadler, J.L., and Kern, T.S. (2008). 5-Lipoxygenase, but not 12/15-lipoxygenase, contributes to degeneration of retinal capillaries in a mouse model of diabetic retinopathy. Diabetes 57, 1387-1393.
44. Gillies, C.E., Putler, R., Menon, R., Otto, E., Yasutake, K., Nair, V., Fermin, D., Eddy, S., Vega-Warner, V., Hacohen, N., et al. (2018). An eQTL landscape of kidney tissue in human nephrotic syndrome. bioRxiv.
45. Warsow, G., Endlich, N., Schordan, E., Schordan, S., Chilukoti, R.K., Homuth, G., Moeller, M.J., Fuellen, G., and Endlich, K. (2013). PodNet, a protein-protein interaction network of the podocyte. Kidney Int 84, 104-115.

46. Ng, S.B., Buckingham, K.J., Lee, C., Bigham, A.W., Tabor, H.K., Dent, K.M., Huff, C.D., Shannon, P.T., Jabs, E.W., Nickerson, D.A., et al. (2010). Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 42, 30-35.
47. Lalonde, E., Albrecht, S., Ha, K.C., Jacob, K., Bolduc, N., Polychronakos, C., Dechelotte, P., Majewski, J., and Jabado, N. (2010). Unexpected allelic heterogeneity and spectrum of mutations in Fowler syndrome revealed by next-generation exome sequencing. Hum Mutat 31, 918-923.
48. Tsui, L.C., and Dorfman, R. (2013). The cystic fibrosis gene: a molecular genetic perspective. Cold Spring Harb Perspect Med 3, a009472.
49. Sulem, P., Helgason, H., Oddson, A., Stefansson, H., Gudjonsson, S.A., Zink, F., Hjartarson, E., Sigurdsson, G.T., Jonasdottir, A., Jonasdottir, A., et al. (2015). Identification of a large set of rare complete human knockouts. Nat Genet 47, 448-452.
50. Colin, E., Huynh Cong, E., Mollet, G., Guichet, A., Gribouval, O., Arrondel, C., Boyer, O., Daniel, L., Gubler, M.C., Ekinci, Z., et al. (2014). Loss-of-function mutations in WDR73 are responsible for microcephaly and steroid-resistant nephrotic syndrome: Galloway-Mowat syndrome. Am J Hum Genet 95, 637-648.
51. Thao, M.T., and Gaillard, E.R. (2016). The glycation of fibronectin by glycolaldehyde and methylglyoxal as a model for aging in Bruch's membrane. Amino Acids 48, 1631-1639.
52. Li, Z., McKeague, I.W., and Lumey, L.H. (2014). Optimal design strategies for sibling studies with binary exposures. Int J Biostat 10, 185-196.
53. Bansal, D., Badhan, Y., Gudala, K., and Schifano, F. (2013). Ruboxistaurin for the treatment of diabetic peripheral neuropathy: a systematic review of randomized clinical trials. Diabetes Metab J 37, 375-384.
54. Qi, W., Keenan, H.A., Li, Q., Ishikado, A., Kannt, A., Sadowski, T., Yorek, M.A., Wu, I.H., Lockhart, S., Coppey, L.J., et al. (2017). Pyruvate kinase M2 activation may protect against the progression of diabetic glomerular pathology and mitochondrial dysfunction. Nat Med 23, 753-762.
55. Dhar, I., Dhar, A., Wu, L., and Desai, K. (2012). Arginine attenuates methylglyoxal- and high glucose-induced endothelial dysfunction and oxidative stress by an endothelial nitric-oxide synthase-independent mechanism. J Pharmacol Exp Ther 342, 196-204.

Figures and Tables

Figure 1. Cohorts and study design. (**a**) Cohorts used in the search for DN susceptibility genes in Finnish type 1 diabetes (T1D) patients: the genomes of a total of 76 sib pairs concordant for T1D but discordant for diabetic nephropathy (DSPs) were subjected to whole genome sequencing (WGS). Additionally, T1D siblings from three families with three siblings (Multiple Siblings, MS) with or without diabetic nephropathy (DN) were included in the sequencing analyses. The control siblings (81) have had diabetes for at least 15 years [range 15-37] without developing DN, and have never been on ACE-I or ARB medication for kidney disease. The case siblings (80) have had overt proteinuria, been on dialysis, received a kidney transplant or have died from kidney complications. (**b**) Multi-level strategy used to analyze the WGS data from Finnish T1D individuals with or without diabetes complications.

Figure 2. Schematic view of DNA variants and regions in Finnish T1D sib pairs discordant for diabetic nephropathy (DN). The circos plot consists of multiple layers, each of which represents a bioinformatic analysis approach and its significant outcomes in the discovery cohort. From the outside to the center; Cytoband as a genome location reference. DN-associated protein altering variants (PAVs) that are replicated in FinnDiane are highlighted. PAVs that are highly enriched in cases are marked in red while green in controls. In the second layer, genes with highly enriched cluster of DN-associated variants that has been prioritized by F-SKAT are depicted in the orange circle, and those passing the stringent replication are marked by their names. From the third layer, regions of recurrent mutations that are associated with case or control (DN-RMR) are shown in the light green circle, followed by promoters (±500 bp from a promoter annotated CAGE cluster according to FANTOM5) in light blue, enhancers (±500 bp from an enhancer annotated CAGE cluster according to FANTOM5) in light purple, and Transcription Factor Binding Sites (TFBS) in light red. The details of the statistical models and the call of significance of association for each approach are listed in **Table 2**.

Figure 3. Genome-wide analysis of variants in recurrently mutated regions and transcription factor binding sites associated with diabetic nephropathy in the discovery cohort. (**a**) Annotation of recurrently mutated regions (RMR) with respect to overlapping gene regulatory elements; relative frequencies have been calculated with respect to each group: all RMR (white) and DN-associated RMR (red). (**b**) Significantly over-represented KEGG pathways comprise common genes overlapping with DN-RMR. The relationships between genes overlapping with DN-RMR and KEGG pathways is depicted as a network graph, wherein the outer circle comprises genes and inner circle comprises the pathways. (**c**) Schematic representation of genome-wide analysis of variants occurring in transcription factor binding sites (TFBSs), that were derived from 668 ChIP-Seq datasets (see **Methods**). (**d**) We identified 40 transcription factors (TFs) with significantly different variant frequencies between cases and

controls, in the TFBSs, which were significantly enriched for pathways relevant to the pathophysiology of DN. For the top ten enriched KEGG pathways, the known relationships (edges) between transcription factors (inner circle) and the KEGG pathways (outer circle) are depicted as a network graph. (**e**) We found an enrichment of variants in cases in the promoter and enhancer regions (±1kb) of the *ALOX5* gene locus. Enhancers and promoter regions were retrieved from FANTOM5 and cross-checked with chromHMM, whereas other gene annotations were obtained from RefSeq (see **Methods**).

Figure 4. Genes identified by F-SKAT analysis within the podocyte network. (**a**) Graphical representation of the core podocyte network that includes the genes associated with DN by F-SKAT analysis **in the discovery cohort**. Node color indicates the statistical significance (P-value) of the F-SKAT test. White color nodes indicates podocyte network genes not detected in the current study. (**b**) The F-SKAT associated genes within the podocyte network are enriched (adjusted $P<0.05$) for several pathways; top six pathways and contributing genes are reported. Full functional enrichment results are reported in **Table S6**. (**c**) Details on the protein kinase C epsilon (*PRKCE*) gene that showed the highest association with DN (by F-SKAT) and location of the intronic SNVs associated with DN. For each SNV, the association with DN is reported by odds ratio tested in either a recessive or dominant model. Full statistics and regulatory information on the SNVs are reported in **Table S4b**.

Table 1. Clinical characteristics of the Finnish type 1 diabetes patient discovery cohort. Data are reported as range or mean ± standard deviation.

	Cases	**Controls**
N[1] (male %)	80 (61.3)	81 (46.9)
T1D		
Duration[2] (years)	Range 21-38	Range 15-37
Age at onset (years)	11.6±8.1	16.6±11.3
Blood pressure (mmHg)		
Systolic	149.2±23.1 (n=60)	135.4±15.3 (n=59)
Diastolic	82.1±11.3 (n=60)	79.2±8.0 (n=59)
Antihypertensive medication (%)		
At baseline	83.8	25.9
During follow-up	98.0	74.0
HbA_{1c} (%)	9.0±2.0 (n=70)	8.4±1.4 (n=57)
BMI (kg/m^2)	26.3±5.0 (n=63)	26.4±3.9 (n=57)
Total cholesterol (mmol/L)	5.5±1.2 (n=69)	5.1±1.0 (n=77)
Lipid-lowering medication (%)		
At baseline	22.5	9.9
During follow-up	82.5	69.1
ESRD[3] (%)	46.3	0

[1] N, number of subjects
[2] Duration till year 2017
[3] ESRD, end-stage renal disease

Table 2. Summary of test results from the genomic, gene and single variants levels of data analysis on 76 DSP.

Test Name	Test Model	Multi-test Correction and threshold (discovery)	Functional Annotation	Results	Replicated in FinnDiane	Result Report
Genome Level						
Recurrently Mutated Regions (RMR)	Hot-spot* clustering, negative binomial distribution	Bonferroni $p<3.7\times10^{-5}$	N.A.	850,137 RMR	N.A. Only DN-RMR are replicated	Figure 3, S1
Diabetic Nephropathy associated RMR (DN-RMR)	Fisher's Exact Test	FDR**<0.05	Genome location, Pathway overrepresentation (KEGG), Protein-protein interaction	732 DN-RMR and the pathways involved	Bonferroni $p<0.01$ 141 DN-RMR replicated	Figure 3, Table S4
Promoters, Enhancer, Transcription Factor Binding Sites (TFBS)	Fisher's Exact Test	FDR<0.05	Functional enrichment test, Protein-protein interaction	270 promoters, 44 enhancers, 40 TFBS	Bonferroni $p<0.01$, 68 promoters, 5 enhancers, 6 TFBS replicated	Figure 3, Table S5-S7
Gene Level						
F-SKAT*** (76 pairs plus 3 multi-sib families)	F-SKAT on DN-associated SNVs (OR>1.5 & $P<0.05$)	N.A. nominal $P<0.01$	Functional enrichment test, Protein-protein interaction	206 F-SKAT significant genes	9 genes using strict replication approach	Figure 4, Table S9,S10,S11
	F-SKAT on rare SNVs, (MAF#<0.05, MAF<0.01)			N.A.		Table S8
Single Variant Level						
Single variant association test	Odds Ratio (OR) in dominant and recessive model	N.A. Case-only or control- only§, PAV¥ or ncRNA exonic	SNV location, SIFT, Polyphen2	3562 PAVs, 3259 variants in ncRNA exonic	OR>1.5 & $P<0.05$, 47 recessive PAVs replicated, 86 recessive ncRNA variants replicated	Table 3, S12, S13

*Variant clustering method proposed by Weinhold *et al.*
**FDR: Benjamini-Hochberg False Discovery Rate.
***F-SKAT: Sequence kernel association test for familial data with dichotomous traits.
§Case-only or control-only: ≥3 heterozygous individuals in only case/control in dominant model; ≥1 homozygous individual in only cases/control in recessive model.
¥PAV: Protein-Altering Variants, i.e. nonsynonym, stopgain, and stoploss.
#MAF: Minor Allele Frequency
N.A. Not Available.

Table 3. Top protein-altering variants replicated cohort with criteria P-value < 0.05, OR>1.5 in FinnDiane cohort.

Gene symbol	Gene description	dbSNP ID	MAF		SIFT | PP2	AA change	Discovery	Replication N=3,531		Combined N=3,683	
			1000G	ExAC (All|Finns)			Case|Control	Case|Control* (Odds ratio)	P-value	Case|Control* (Odds ratio)	P-value
WDR73	WD repeat domain 73	rs72750868	0.044	0.076	T|B	D->G	2|0	2.516	0.002	2.635	0.001
TPPP2	tubulin polymerization promoting protein family member 2	rs9624	0.160	0.148	D|D	R->L	1|0	3.284	0.003	3.395	0.002
UBR7	ubiquitin protein ligase E3 component n-recognin 7	rs2286653	0.113	0.147	T|B	A->T	1|0	3.549	0.008	3.744	0.005
ATP10D	ATPase phospholipid transporting 10D	rs34208443	0.077	0.141	T|B	P->T	1|0	1.654	0.009	1.648	0.009
ANO9	anoctamin 9	rs114405390	0.015	0.027	T|B	T->A	1|0	4.089	0.012	4.407	0.007
SIGIRR	single Ig and TIR domain containing	rs117739035	0.016	0.029	D|D	S->Y	1|0	3.600	0.013	3.847	0.008
SFT2D1	SFT2 domain containing 1	rs11551053	0.111	0.077	T|B	I->V	1|0	3.041	0.015	3.208	0.009
HKR1	HKR1, GLI-Kruppel zinc finger family member	rs2921563	0.098	0.054	T|D	R->H	1|0	5.717	0.017	6.402	0.009
KRT32	keratin 32	rs2604956	0.046	0.071	T|D	D->E	1|0	2.153	0.020	2.210	0.015
C6orf118	chromosome 6 open reading frame 118	rs17852379	0.103	0.073	T|D	G->E	1|0	2.547	0.026	2.673	0.018
PPP4R1	protein phosphatase 4 regulatory subunit 1	rs329003	0.041	0.073	.|B	I->V	2|0	2.999	0.027	3.474	0.009
ANKRD26	ankyrin repeat domain 26	rs12572862	0.067	0.036	T|B	V->L	1|0	8.157	0.029	9.595	0.013
PKHD1L1	polycystic kidney and hepatic disease 1	rs117037399	0.005	0.019	T|P	G->V	1|0	8.157	0.029	9.592	0.013
CSMD1	CUB and Sushi multiple domains 1	rs34337712	0.021	0.069	T|B	Q->H	1|0	1.865	0.033	1.899	0.027
C6orf10	chromosome 6 open reading frame 10	rs7775397	0.019	0.060	T|P	K->Q	1|0	1.546	0.036	1.546	0.035
TMEM176A	transmembrane protein 176A	rs10378	0.128	0.139	D|D	L->F	0|1	0.383	0.004	0.366	0.002
C4orf51	chromosome 4 open reading frame 51	rs10008599	0.077	0.098	D|B	D->N	0|1	0.180	0.007	0.167	0.004
SIGMAR1	sigma non-opioid intracellular receptor 1	rs1800866	0.217	0.184	T|B	Q->P	0|2	0.432	0.010	0.403	0.005
CPTP	ceramide-1-phosphate transfer protein	rs150672559	0.005	0.007	T|B	R->H	0|1	0|9	0.013	0.000	0.008
NEFH	neurofilament heavy polypeptide	rs5763269	0.151	0.182	D|B	P->L	0|1	0.493	0.014	0.473	0.009
TNFRSF14	TNF receptor superfamily member 14	rs2234167	0.114	0.130	T|B	V->I	0|1	0.472	0.018	0.451	0.012
TBC1D9	TBC1 domain family member 9	rs13118702	0.010	0.020	T|B	E->K	0|1	0.135	0.021	0.122	0.012
UNC93A	unc-93 homolog A	rs2235197	0.110	0.109	.|.	W->*	0|4	0.514	0.022	0.463	0.007
TYR	tyrosinase	rs1042602	0.123	0.252	.|D	S->Y	0|5	0.630	0.025	0.581	0.007
ATAD3B	ATPase family, AAA domain containing 3B	rs139902189	0.078	0.076	D|P	C->T	0|1	0.324	0.028	0.302	0.018
TEX101	testis expressed 101	rs35033974	0.041	0.084	D|D	G->T	0|3	0.511	0.032	0.466	0.013
AVEN	apoptosis and caspase activation inhibitor	rs61729120	0.007	0.016	D|D	G->T	0|2	0.231	0.033	0.198	0.014
ZNF844	zinc finger protein 844	rs76842919	0.026	0.060	D|B	A->G	0|1	0.231	0.033	0.211	0.020
ZNF844	zinc finger protein 844	rs8102258	0.119	0.095	T|B	T->C	0|1	0.231	0.033	0.211	0.020
OR6X1	olfactory receptor family 6 subfamily X member 1	rs12364099	0.077	0.122	D|B	C->A	0|1	0.537	0.035	0.515	0.023

Case/control only protein-altering SNVs that remain significant (odds ratio >1.5, p<0.05) after replication in the FinnDiane cohort (1,344 cases, 2,187 controls). Only the top 15 protein-altering SNVs detected in the recessive model (case or control only) are listed here (full results are reported in **Table S12**).
Minor Allele Frequency (MAF) in general population is annotated from 1000 Genome (1000G) project and ExAC. Odds ratios and p-values were assessed using the Firth's Penalized Likelihood logistic regression (see **Methods**). *Odds ratio, or number of homozygous carriers of the variant. The potential effect of a variant in the protein is predicted by SIFT and Polyphen2 (PP2).