Supplementary Data

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

Jing Guo, Owen J.L. Rackham, Bing He, Anne-May Österholm, Niina Sandholm, Erkka Valo, Valma Harjutsalo, Carol Forsblom, Iiro Toppila, Maikki Parkkonen, Qibin Li, Wenjuan Zhu, Nathan Harmston, Sonia Chothani, Miina K. Öhman, Eudora Eng, Yang Sun, Enrico Petretto, Per-Henrik Groop, Karl Tryggvason

Table of contents

SUPPLEMENTARY METHODS

Whole Genome Sequencing (WGS)

Sample preparation and DNA sequencing. DNA was extracted from blood samples using DNA extraction kit (Qiagen). DNA amount was measured using NanoDrop, only samples displaying an intact band on gel electrophoresis and an OD260/280 1.8-2.0 were used for WGS. 38 samples (19 sib pairs) were sequenced at Complete Genomics (CG) (Mountain View, CA) and 131 samples (61 sib pairs and 3 multiple sib families) were sequenced at BGI Genomics (Shenzhen, China), respectively. Eight samples (4 pairs) were sequenced by both CG and Illumina platforms for quality control purpose.

Alignment and QC of WGS data. DNA samples analysed at BGI were sequenced using Illumina HighSeq 2000. For each sample, 1.12 billion pair-end reads with average length of 100 bp were obtained. Reads containing adapter sequences or high rate of low quality bases were removed. Cleaned data was aligned to a human reference genome (hg19) using Burrows-Wheeler Aligner (BWA)[1]. Sequencing depth and coverage were calculated against the reference genome (**Table S1a**). Recalibration of base cycle, original quality score and dinucleotide context were applied by GATK[2,3] best practice protocol[61]. PCR duplicates were removed using the Picard tools (http://broadinstitute.github.io/picard/). Sequencing of DNA at Complete Genomics was carried out using the CG platform with 35+35 bp mate-pair reads. The average coverage of our 38 samples was >30x. Quality control and variants callings were completed by the standard CG pipeline (**Table S1**). However, to be able to compare with results from BGI, we converted the CG masterVar files to standard vcf files.

Genotype Calling and annotation of single nucleotide variants (SNVs). For Illumina sequences, GATK[2,3] was used for SNPs (Single Nucleotide Variants) and Indels (small Insertion and Deletions). ANNOVAR[4] was used for functional annotation, which includes the variant location (RefSeq and Ensembl), deleterious prediction (SIFT and Polyphen2), frequency in sub-population of 1000 Genomics, and ExAC[5]. We used a relatively conservative strategy for variants filtration: variants with genotyping rate <90% across all samples were removed from further analysis; variants at the same location but different functional changes (protein altering).

Quality control of comparison between two sequencing platforms. In total, 8 samples were sequenced on both CG and Illumina (by BGI). In general, Illumina calls 350 thousand (15%) more SNVs than CG for each sample. Among SNVs called by both platforms, the rate of difference is around 0.12%, as shown in **Table S1b**. For these 8 samples, conflicting results were removed from the downstream analysis.

Quality control of comparison between WGS and genotyping array data. A total of 144 samples in the WGS cohort were also genotyped by HumanCoreExome Bead arrays (Illumina) in FinnDiane. SNVs that genotyped by both WGS and Array are compared as quality control.

Annotation for Genome-wide Analysis

Enrichment of mutations in transcription factor binding sites (TFBS). In order to identify transcription factor binding sites we used the collated ENCODE set provided by ReMap[6]. These regions were then tested for over/under representation by counting the number of mutations found in DN cases or controls and carried out Fisher's Exact Test (FET) to assess mutation overrepresentation in either cases or controls and applying a Benjamini-Hochberg false discovery rate (FDR) correction in the 76 DSP.

Enrichment of mutations in annotated regulatory regions (promoters and enhancers). Promoter and Enhancer regions were defined using the permissive set from the FANTOM5 promoterome[7] and enhancerome[8]. The locations were then cross-matched with chromHMM[9] to provide additional support for these being true promoters/enhancers. The regions were then extended by 1Kb up and downstream in order to capture any regulatory regions. Following this the regions were tested for association to disease as described above. Differential expression analysis of transcriptomics in DN glomeruli and tubuli were done using data downloaded from GEO with accession number GSE3012210, using limma package from R. False Discovery Rate (FDR)<0.05 was used as criteria for significance.

Analysis of genes linked with the enhancers. Genes within the same topologically associated domains (TADs) as the enhancer were prioritized based on (*i*) number of Hi-C datasets it is found in the enhancer's TAD, (*ii*) differential expression in the glomeruli or tubuli in diabetic kidney[10-12], and (*iii*) chromatin marks found in

Roadmap epigenome data[10] . Human Hi-C datasets were downloaded from GEO (GSE52457) for five distinct lineages; H1-ESC (H1), mesenchymal stem cells, mesendoderm, neural progenitor cells and trophoblast-like[13]. Reads were iteratively aligned using bowtie[14] against hg19. Reads mapping to chrM and chrY were removed. A bin size of 20kb and a window size of 40kb were used to generate contact matrices to identify TADs. TADs were identified using both HOMER[15] and the TAD calling pipeline (HMM_calls) proposed by Dixon *et al.*[13].

ChromHMM[9] based annotations of chromatin states on the Roadmap Epigenome[16] dataset including 127 epigenomes were used to further prioritize the genes. Presence of the enhancer marks (EnhG1 Genic enhancer1, EnhG2 Genic enhancer2, EnhA1 Active Enhancer 1, EnhA2 Active Enhancer 2, EnhWk Weak Enhancer, EnhBiv Bivalent Enhancer) and gene marks (TssA: Active TSS, TssFlnk: Flanking TSS, TssFlnkU: Flanking TSS Upstream, TssFlnkD: Flanking TSS Downstream, Tx: Strong transcription, TxWk: Weak transcription, TssBiv: Bivalent/Poised TSS) in the same epigenome were given a higher similarity score. The similarity score was computed using the Jaccard distance of the presence or absence of the marks.

Annotation for gene-level analysis.

eQTL annotation of F-SKAT genes. We downloaded the human nephrotic syndrome expression QTL (eQTL) data for both glomerulus and tubulointerstitium from nephqtl.org[17]. We mapped the eQTL locations with our F-SKAT SNVs with direct mapping (exact loci), and indirect mapping (SNVs within the same linkage disequilibrium (LD) block in the Finnish population). LD blocks were defined using the 'strong LD' criteria by PLINK1.9[18] (bottom of the 90% LD confidence interval is greater than 0.70, and the top of the confidence interval is at least 0.98). We estimated the LD blocks in the discovery cohort by Haploview/PLINK1.9, assigning 1,000 kb maximum estimated block length (--blocks-max-kb 1000) and ignoring MAF<0.05 (--blocks-min-maf 0.05).

Studies in zebrafish

Zebrafish (Danio rerio) and their embryos were handled according to standard protocols at the Karolinska Institutet zebrafish core facility. The *Tg* (podocin:GFP) line[19] was used for knockdown experiments.

Morpholino-mediated knockdown. We designed two MOs targeting two splicing regions, (the intron3-exon4; the exon3-intron3), of zebrafish abtb1 with a dose of 4 ng/embryo. The sequences of the MOs used are: abtb1-I3E4-MO1, 5′-TCGCCTCACACTTTGCTCCTGTCAC; abtb1-E3I5-MO2, 5′-AAGTGAAACAGCGCCTACCAGTGGA. To exclude potential p53 activation, a p53 MO was included in injection as described[20]. Typically, 2 nl of MO solutions including 200 μM p53 MO were normally injected into the yolk of 1- or 2-cell embryos. Efficacy of the splicing inhibition MOs was confirmed by RT-PCR assay, where MO1 led to a 145-bp deletion and MO2 led to 45-bp deletion. As a negative control, an unrelated standard control MO, referred to as controls provided by the manufacturer, was used. A combination of pericardial edema and glomerular GFP expression at 4 dpf was used as readout for judging kidney phenotype. For the rescue assay, the MOs with in vitro synthesized human wt or mutant ABTB1 mRNA (100 ng/μl) were co-injected.

Cell culture, transfection and Western blotting. HEK293 cells were maintained in DMEM with 10% foetal bovine serum (FBS), 10 units/ml penicillin and 0.1μg/ml streptomycin at 37ºC. The cells with 80-90% confluence in 60-mm plates were transfected with plasmids using Lipofectamine 2000 (Invitrogen). We transfected 3 μg of the wild-type or mutant ABTB1 plasmid in serum-free medium containing 12 μl of Lipofectamine 2000 following the manufacturer's protocol. After 2 days post-transfection, transfected cells were collected by trypsinisation and harvested in RIPA buffer as whole cell lysates for Western blotting. The primary antibody to ABTB1 (Acris AP22527PU-N 1:500) was used together with a loading control antibody to ß-actin (Abcam ab8227; 1:3000) and the HRP conjugated secondary antibody to rabbit IgG (Amersham NA934; 1:10,000). The local ethical committee (the North Stockholm district court) approved studies in mice.

Immunofluorescence Staining. The kidney from C57BL/6 mice was snap-frozen and embedded in OCT. Cryosections (8 μm) were post-fixed with cold acetone for 10 min followed by blocking in 5% normal donkey serum. For double immunofluorescence staining, the primary antibodies to ABTB1 (Acris AP22527PU-N 1:100) and nephrin (1:200, Acris GmbH; guinea pig) were incubated at 37°C for 1 h, followed by 45 min

incubation with corresponding Alexa fluor secondary antibodies (Invitrogen). Confocal imaging was performed using Zeiss LSM 700 (magnification 63×).

Quantitative PCR. Total RNA was isolated from a mixture of 5 whole zebrafish embryos using the RNeasy Mini Kit (Qiagen). The first-strand cDNA synthesis was carried out using the iScript cDNA synthesis kit (Bio-Rad). qPCR was performed on the ABI PRISM 7300 Sequence Detection System using the SYBR Green method (Applied Biosystems). Triplicate for each sample was carried out. The relative quantification of gene expression was analysed using the comparative threshold (Ct) method. Data was presented as mean ± S.E.M $2^{-\Delta Ct}$.

Transmission Electron Microscopy (TEM). Larvae were fixed in the fixation solution buffer (2% glutaraldehyde, 0.5% paraformaldehyde, 0.1 M cacodylate, 0.1M sucrose, 3 mM $CaCl_2$) and washed in 0.1 M cacodylate buffer pH 7.4 prior to staining in 2% OsO4 in cacodylate buffer for 1 h at room temperature. Samples were dehydrated and *en bloc* staining was performed in 2% uranyl acetate in absolute ethanol for 1 h at room temperature; then samples were taken through an Epon 812/acetone series and embedded at 60°C in pure Epon 812. Thin sections of 70 nm thickness were made on a Leica EM UC6 ultratome and mounted on formvar coated copper slot grids. Post-staining was done with 2% aqueous acetate pH 3.5 and Venable and Cogglesall's lead citrate. Grids were analysed on a FEI TECNAI electron microscopy.

Power Calculation

Power calculation in discovery cohort. To estimate the power we used the generalized linear model proposed by Li Z. et al.[21] The calculation was based on dominant and recessive model, and we estimated the power taking penetrances into consideration. The power was to estimate variants detection with significance level Bonferroni p-value $<4.11\times10^{-9}$ (for 12,165,600 variants).

MAF (Minor Allele Frequency) range adjusted according to actual test results in our study (Table 2 and Table S12). As we only used case/control only variants, odds ratio is infinite in the recessive or dominant model respectively. Prevalence of diabetic nephropathy in diabetic patients, here assumed to be 30%.[22] Relative Risk (RR) was calculated based on P_1/P_0, where P_1=probability of DN given exposure of the genotype in the respective model, and P_0=probability of DN given no exposure of the

genotype in the respective model. Estimated within family correlation was set to 0.4 (see discussion in original publication Li Z. et al.[21]).

Power calculation in replication cohort.

Estimation of statistical power for replication cohort was performed using GAS Power Calculator in additive and recessive models.
MAF (Minor Allele Frequency) range adjusted according to Table 2 and Table S12.
Odds Ratio and MAF selection based on the replicated SNVs reported in Table S12.
Relative risk was calculated as RR = OR/((1-prev)+(prev×OR)), where prev is the prevalence of diabetic nephropathy in in diabetic patients, here assumed to be 30%.[22]
The significance level was determined by Bonferroni corrected p-value = 0.05/6821 = 7.3×10^{-6}.

SUPPLEMENTARY RESULTS

Analysis of ABTB1. *ABTB1* is located at the chromosome 3q21 locus that has been linked with DN in previous linkage studies[23-27] (**Figure S5**). Here, we found a homozygous stop-gain (Arg164Ter) in the *ABTB1* gene in one DN case, whose heterozygotic twin (control) is homozygous for the arginine codon CGA. *ABTB1* (also called *BPOZ*) encodes a protein with an ankyrin repeat region and two BTB/POZ domains. *ABTB1* is reported to be a positive mediator of the PTEN growth-suppressive signalling pathway[28], and loss of PTEN *in vivo* promotes podocyte cytoskeleton rearrangement and aggravates DN[29]. Thus *ABTB1* may be a positive regulator of the PI3-kinase pathway that affects cellular growth, survival and proliferation. Analysis of the *ABTB1* variant in the FinnDiane cohort led to the identification of the second homozygous case patient with a mild form of DN. In the entire discovery and replication cohorts, a total of 65 individuals are heterozygotes for this SNV (32 cases and 33 controls), respectively. The *ABTB1* nonsense mutation seems to be highly Finn-specific, since in the 1,000 genomes study[30] the frequency of this mutation in Finns is 1.1%, while it is only 0.1% in all non-Finnish individuals in the ExAC exome database[5].

The ABTB1 protein is reported to be ubiquitously expressed in all human foetal tissues examined, including brain, liver, and kidney. It is also expressed at low levels in both adult heart and hypertrophic heart. Our *in vitro* studies showed that the stop-gain mutation (Arg164Ter) in exon 12 leads to a truncated ABTB1 in cultured cells transfected with plasmid encoding mutant or wild-type (wt) *ABTB1* cDNA (**Fig. 5a-b**). Further, immunofluorescence staining of wt mouse kidney cryosection revealed expression of *Abtb1* in glomerular podocytes co-localizing with the podocyte marker podocin (**Fig. 5c**). To gain insight into *ABTB1* biological role, we knocked down the expression of *abtb1* in embryos from the Podocin-GFP zebrafish[19] that express GFP driven by the podocin promoter using two morpholinos (MOs) targeting different sites. As shown in **Fig. 5d-h**, 4 day-post-fertilization (dpf) morphants exhibited severe pericardial edema, frequently observed after morpholinos knockdown of glomerulus-associated genes[31]. Pericardial edema concurrent with loss or decline of glomerular GFP expression in *abtb1* morphants demonstrates the glomerular origin injury, suggesting that the protein plays a role in glomerular development or function.

Interestingly, human wt mRNA significantly rescued the edema, but mutant mRNA did not (**Fig. 5g**), suggesting a loss-of-function mutation. Transmission electron microscopy analysis revealed glomerular capillary wall damage with deteriorated filtration barrier consisting of endothelial cells, basement membrane and podocytes (**Fig. 5h**).

Analysis of ncRNAs

Non-coding RNAs (ncRNAs) play important roles in disease, including DN[32]. We therefore tested whether mutations impacting ncRNAs were unequally distributed amongst the case and control individuals. We found a total of 3,259 SNVs within exon regions of ncRNAs, which were represented as heterozygous or homozygous either in cases only (1,592 SNVs) or in controls only (1,667 SNVs). Of these SNVs, 122 were also nominally replicated in the FinnDiane cohort (**Table S13**). Among these, the antisense non-coding RNA CTBP1-AS was previously identified by GWAS to be associated with type 2 diabetes and diabetic retinopathy[33].

SUPPLEMENTARY TABLES

Supplementary Table 1. Comparison of DNA sequencing quality using the Illumina and Complete Genomics platforms.

a. Sequencing depth and mapping rates of samples sequenced by Illumina and Complete Genomics

	Illumina		Complete Genomics	
	Mean (s.d.)	*Range*	*Mean (s.d.)*	*Range*
Raw bases (Gb)	112.5 (8.59)	110.19-151.78	152.0 (9.96)	145.7-209.7
Mapped bases (Gb)	108.6 (8.35)	96.11-146.13	123.0 (5.42)	114.8-150.4
Fraction covered ≥ 5x (%)	99.6 (0.002)	98.01-99.79	99.3 (4e-4)	99.2-99.4
Fraction covered ≥ 10x (%)	99.0 (0.006)	92.82-99.56	98.3 (0.002)	97.6-98.5
Fraction covered ≥ 20x (%)	93.4 (0.03)	72.34-07.72	92.2 (0.01)	88.5-94.1
Fraction covered ≥ 30x (%)	68.5 (0.09)	46.81-89.32	78.4 (0.02)	70.5-82.6
Fraction covered ≥ 40x (%)	27.5 (0.11)	11.27-69.36	56.3 (0.03)	47.8-64.2

b. Sequencing coverage of 8 individuals' DNA sequenced by both Illumina and Complete Genomics platforms

Sample ID	Illumina				Complete Genomics			
	Total*	**Overlap† %**	**Uniq‡ %**	**Dif§ %**	**Total**	**Overlap %**	**Uniq %**	**Dif %**
Sib a-1	3,802,578	83.64%	16.24%	0.12%	3,368,113	97.53%	2.34%	0.14%
Sib a-2	3,775,720	84.38%	15.50%	0.12%	3,368,206	97.56%	2.31%	0.14%
Sib b-1	3,715,856	84.51%	15.37%	0.12%	3,335,549	97.44%	2.42%	0.13%
Sib b-2	3,703,362	85.19%	14.69%	0.12%	3,345,894	97.39%	2.47%	0.14%
Sib c-1	3,724,286	85.63%	14.25%	0.13%	3,377,400	97.34%	2.52%	0.14%
Sib c-2	3,716,935	85.83%	14.04%	0.13%	3,378,031	97.33%	2.54%	0.14%
Sib d-1	3,722,928	85.76%	14.11%	0.13%	3,378,699	97.31%	2.55%	0.14%
Sib d-2	3,725,357	85.55%	14.33%	0.12%	3,370,070	97.38%	2.48%	0.14%

*Total: Total number of SNVs in each individual; †Overlap: percentage of SNVs genotyped by both Illumina and CG; ‡Uniq: percentage of SNVs only genotyped by respective platform; §Dif: percentage of SNVs that has different genotypes in two platforms.

c. Comparison of Whole Genome Sequencing results with Illumina SNParray genotyping results

	ALL	**MAF <0.01**	**MAF 0.01-0.05**	**MAF >0.05**	**Complete Genomics**	**Illumina**
Sample No.	144				19	120
SNPs	241,212	8,795	21,237	209,055	241,212	241,212
Tested Genotype	34,734,528 (100%)	1,266,480 (100%)	3,058,128 (100%)	30,103,920 (100%)	4,583,028 (100%)	28,945,440 (100%)
Both Genotyped	34,634,942 (99.71%)	1,263,630 (99.77%)	3,050,796 (99.76%)	30,016,318 (99.71%)	4,529,476 (98.83%)	28,899,859 (99.84%)
Concordant (%)	34,331,252 (99.12%)	1,263,136 (99.96%)	3,044,851 (99.81%)	29,723,134 (99.02%)	4,418,723 (97.55%)	20,707,242 (99.33%)

Supplementary Table 2. Annotation of SNVs and indels identified in 161 genomes in the Discovery cohort by RefSeq.

SNVs		Indels	
Intergenic	6,672,250	**Intergenic**	3,278,601
Intronic	4,417,565	**Intronic**	2,257,481
Withihn non coding RNA	703,931	**Withhn non coding RNA**	345,024
Downstream / upstream	158,668	**Downstream / upstream**	81,758
5' UTR / 3' UTR	117,077	**5' UTR / 3' UTR**	53,095
Splicing¶	547	**Splicing**¶	579
Exonic	96,562	**Exonic**	3,509
Functional annotation of exonic SNVs		***Functional annotation of exonic idels***	
Nonsynonymous (missense)¶	52,022	Frameshift insertion/deletion/substitution	1,877
Stop-gain (nonsense)¶	817	Nonframeshift insertion/deletion/substitution	1,471
Stop-loss¶	63	Unknown	161
Synonymous	42,116		
Unknown	1,544		
Total SNVs	**12,165,600**	**Total indels**	**6,021,219**

¶ Splicing variants, nonsynonymous, stop-gain and stop-loss are considered protein-altering variants (PAVs)

Supplementary Table 3. Frameshift-causing small insertions and deletions (indels) found in DN cases-only or controls-only individuals of the Finnish T1D DSP discovery cohort

	Location (chromosome-bp) or dbSNP ID	Gene	Case|Control*	Type	Model
DN-susceptibility	5-1414885	*SLC6A3*	3|0	ins	Dominant
	9-113148267	*SVEP1*	3|0	ins	
	9-131709581	*DOLK*	4|0	ins	
	11-58919922	*FAM111A*	4|0	ins	
	12-58019428	*SLC26A10*	3|0	ins	
	16-29998896	*TAOK2*	3|0	del	
	rs200056085	*DSC2*	3|0	ins	
	18-29867027	*GAREM*	4|0	ins	
	rs201063949	*NUDT17*	2|0	del	Recessive
	rs199513201	*NUDT17*	2|0	del	
	6-18134021	*TPMT*	2|0	del	
	15-42111753	*MAPKBP1*	3|0	ins	
	18-40096275	*LINC00907*	2|0	ins	
	20-1559024	*SIRPB1*	2|0	del	
DN-protection	1-75038842	*C1orf173*	0|3	del	Dominant
	rs35715260	*C4orf47*	0|3	del	
	6-30136137	*TRIM15*	0|4	del	
	6-42074829	*C6orf132*	0|3	ins	
	6-121615780	*C6orf170*	0|3	ins	
	16-55361767	*IRX6*	0|3	ins	
	16-83933199	*MLYCD*	0|3	ins	
	11-118939941	*VPS11*	0|3	ins	Recessive
	rs113768780	*DDX55*	0|2	del	
	15-78587744	*WDR61*	0|2	del	

*For the recessive model, number of homozygous carriers of the variant; for the dominant model, number of heterozygous carriers of the variant

Supplementary Table 4. Recurrently mutated regions (RMR) significantly overrepresented in DN cases or controls (FDR<5% in discovery cohort) and replication in FinnDiane cohort.
Table 4 as a separate excel file (Supplementary Table 4.xlsx)

Supplementary Table 5. Transcription factor binding site (TFBS) impacted by DN-mutations.
Please find Supplementary Table 5 as a separate excel file (Supplementary Table 5.xlsx)

Supplementary Table 6. Enhancer (S6a) and promoter (S6b) region with mutations overrepresented in DN cases or controls (FDR<0.05 in discovery cohort) and replication statistics in FinnDiane cohort.
Please find Supplementary Table 6 as a separate excel file (Supplementary Table 6.xlsx)

Supplementary Table 7. Enhancers replicated in FinnDiane cohort and gene prioritization

Enhancers							*Prioritized genes within enhancer*						
Chr.	**Start (bp)**	**End (bp)**	**Mutation frequency**		**Bonferroni corrected P-value**		**Gene Symbol**	**Gene description**	**No. of TADs[1]**	**Epigenetic mark co-occurrence[2]**	**Differential gene expression in DN[3]**		
			Cases	**Controls**	**Discovery**	**Replication**					**Log_2FC[3]**	**FDR[4]**	**Tissue**
6	81,661,295	81,663,532	0.543	0.570	0.032	1.45e-10	*ELOVL4*	ELOVL fatty acid elongase 4	2	0.018	-1.86	6.86e-05	Glomeruli
4	107,947,370	107,949,896	0.268	0.287	6.55e-06	7.36e-07	*PAPSS1*	3'-Phosphoadenosine 5'-Phosphosulfate Synthase 1	10	N.A.	-0.88	6.53e-03	Glomeruli
											-1.03	2.91e-03	Tubuli
17	49,512,112	49,514,244	0.238	0.223	1.81e-05	5.67E-05	*CA10*	carbonic anhydrase 10	10	0.275	-2.55	5.77e-07	Glomeruli
20	19,796,665	19,799,070	0.241	0.262	0.018	1.79e-13	*DZANK1*	double zinc ribbon and ankyrin repeat domains 1	1	0.378	-0.46	0.02	Glomeruli
4	84,171,444	84,173,539	0.186	0.196	2.5E-05	9.08E-03	*HNRNPD*	heterogeneous nuclear ribonucleoprotein D	1	0.425	-0.75	2.19e-03	Glomeruli
							HNRNPDL	heterogeneous nuclear ribonucleoprotein D like	1	0.425	-0.74	6.98e-03	Glomeruli
							ENOPH1	enolase-phosphatase 1	2	0.425	-0.86	2.18e-03	Glomeruli
							SEC31A	SEC31 homolog A, COPII coat complex component	9	0.425	-0.56	0.0315	Glomeruli
							COPS4	COP9 signalosome subunit 4	10	0.425	-0.59	3.90e-03	Glomeruli
							PLAC8	placenta specific 8	10	0.425	1.26	0.0362	Glomeruli
											-1.72	6.6e-03	Tubuli

[1] Number of supporting TADs: out of the 10 predicted TADs, using 5 cell types and 2 TAD callers (see **Methods and Materials**), we report the number of TADs where the gene and enhancer co-occur within the same TAD. [2] Epigenetic mark co-occurrence was assessed by the Jaccard similarity score, where the presence of the enhancer marks and gene transcription start site marks in the same epigenome determines a higher score. [3] Microarray data in tubular and glomerular samples from patients with DN as well as from control "healthy" kidneys samples[10] were retrieved from Gene Expression Omnibus under the following accession number: GSE30122. Differential gene expression analysis was carried out using the *limma* package in R and fold changes (FCs) in gene expression ($expression_{DN}/expression_{controls}$) were derived from tubular and glomerular samples, respectively. Microarray probes were assigned to genes using the Affymetrix Human Genome U133A 2.0 Array annotation file. FDR, false discovery rate was calculated using the Benjamini Hochberg method to account for the number of probes tested on the microarray. Only tissues with significant (FDR<0.05) differential gene expression are reported in the table. N.A. Data not available in the corresponding dataset.

Supplementary Table 8a. F-SKAT test results on rare SNVs with MAF<0.01. Only top genes with *P*<0.1 are reported.

Gene	F-SKAT P-value	Number of SNVs with MAF<0.01 in gene region
DKFZP434K028	0.0521164	3
TPM2	0.0616317	6
MINCR	0.0736739	2
APOA1-AS	0.0750578	2
MRGPRX4	0.0753246	5
MIR527	0.0763543	2
OR2T1	0.0763543	2
SAA2	0.0786862	10
RBP5	0.0811927	3
MIR3611	0.0819603	1
MIR520B	0.0819603	1
NPM3	0.0819603	1
TSPYL5	0.0819603	1
ZP4	0.0819603	1
PCDHGB7	0.0819993	3
KLK11	0.0820293	12
KLK14	0.082369	11
LOC339298	0.0828625	31
AMT	0.083445	4
KLK9	0.0852478	15
STATH	0.0862749	8
MOGAT1	0.0863741	43
NMI	0.0869195	18
XIRP2-AS1	0.0872685	15
ENO1-AS1	0.0876638	4
LOC101927164	0.0885387	9
SH2D5	0.0891928	7
SIX2	0.0893007	4
EIF5A	0.0895823	11
CTU1	0.0897272	10
SEH1L	0.0900792	37
PARP15	0.0907272	63
FDCSP	0.0908689	16
NPBWR2	0.0908724	2
OR5B2	0.0908724	2
TNKS1BP1;SSRP1	0.0908724	2
GPS2	0.0912829	2
MIR3591;MIR122	0.0912829	2
KLK10	0.0919056	11
DAGLA	0.0923201	93
PLA2G4E	0.0935282	62
GSDMA	0.0937141	20
AQP9	0.0942802	61
KRT17P5	0.0946153	2
UQCRQ	0.0946153	2
PAFAH1B2	0.0951019	34
ACSM4	0.0953823	23
ARHGAP9	0.0954757	10
RPL26	0.0958218	5
MED18	0.0958508	8
HIST1H4E	0.0958845	10
KLK13	0.0961522	8
RUNDC1	0.0974302	23
RPS14	0.0978619	8
C1RL	0.0990689	16

Supplementary Table 8b. F-SKAT test on SNVs with MAF<0.05. Only top genes P-value<0.1 are reported.

Gene	F-SKAT P-value	Number of SNVs with MAF<0.05 in gene region
MIR4417	0.00507944	2
MIR3909	0.00712075	1
OR4K15	0.0171208	1
PRDM15	0.0262447	213
ACER1	0.0275954	114
CABP7	0.0280719	21
GNRHR	0.0291421	34
JHDM1D-AS1	0.0326767	4
LINC01544	0.0337478	24
NMRAL1	0.0384935	43
COL28A1	0.0415321	270
C17orf49	0.0417821	1
ZNF730	0.042109	73
DENND4C	0.0430949	293
MYEOV	0.0445732	9
AP3M1	0.0448869	41
OVGP1	0.0451543	31
HMOX2	0.047205	81
UBA6	0.0473561	102
OR52B2	0.0482733	9
ART1	0.0498913	56
PAIP2B	0.0507372	130
PLIN3	0.0512861	82
HOXC-AS3	0.0522921	3
CDIP1	0.0540348	84
GRAPL	0.0552494	1
INO80B	0.0552494	2
ZNF593	0.0557399	2
GDF3	0.0572279	17
THEMIS2	0.06043	28
CA5A	0.0608856	163
MIR365B	0.0609248	4
LOC102723377	0.0611037	6
LINC01571	0.0635391	17
ATF7	0.0637729	145
CYP4Z2P	0.0642709	66
LOC101927391	0.0656373	46
LINC00375	0.0666329	133
PET100	0.0677635	7
TINAGL1	0.0680027	19
TPBGL	0.0683699	12
PDHX	0.0688548	118
MBD3L1	0.0704893	9
CYP4A11	0.0706853	22
DRAIC	0.0709673	26
COMMD5	0.0718596	14
NSMCE1	0.0723662	105
MAD1L1	0.0726898	1588
CASP10	0.0755509	108
GRPEL2	0.0762364	3
CYP4Z1	0.0770908	53
ACTL9	0.0779564	5
PCP2	0.0781209	16
MIR7846	0.0788554	1
MIR138-2	0.0819603	1
MIR550A1	0.0819603	1
MST1	0.0819603	1
USP43	0.082603	162
ZNF250	0.084042	35
DEFB108B	0.0846808	8

EFCAB14-AS1	0.085222	30
HACD4	0.0858244	48
SNORD116-28	0.0858579	2
MIR146B	0.0860974	3
OR11L1	0.0872414	4
TMIGD3	0.0872797	153
CACNA1C-IT2	0.0873785	6
MED11	0.088024	4
ABCB8	0.0882607	39
ZNF572	0.089265	23
LINC00597	0.0896801	2
MIR8063	0.0896801	2
LINC01533	0.0910499	75
PPME1	0.0911636	249
CORO7	0.0912197	170
CYP2G1P	0.0912539	31
SNORA92	0.0912829	2
MPP6	0.0917127	342
LCE1F	0.0923003	3
ADRM1	0.0940163	20
FAM157A	0.0946153	2
GPR162;CD4	0.0946153	2
VPS53	0.0946679	472
CST7	0.0949256	47
ATPAF1	0.0955582	79
ACER2	0.0966951	83
TBC1D3P1-DHX40P1	0.0973601	9
TLR5	0.0983784	92
LRP3	0.0988775	37
AQP7P1	0.0989532	1
DPYD-AS2	0.0989532	1

Supplementary Table 9a. Genes associated with DN by F-SKAT analysis (*P*<0.01).

Supplementary Table 9b. Details on the DN-associated SNVs used in the F-SKAT analysis.

Supplementary Table 9c. Replication of F-SKAT significant (*P*<0.01) genes in FinnDiane cohort.

Please find Supplementary Table 9 as a separate excel file (Supplementary Table 9.xlsx)

Supplementary Table 10. PodNet genes detected by F-SKAT (*P*<0.01)

Gene symbol	Gene name	DN-associated SNVs[1]	F-SKAT P-value	SNV localization	eQTL in NS[2]
PRKCE	Protein Kinase C Epsilon	48	0.0004	intronic	1
RBFOX1	RNA Binding Protein, Fox-1 Homolog 1	158	0.0004	intronic	79
CTNNA3	Catenin Alpha 3	126	0.0008	intronic	27
CDK5RAP2	CDK5 Regulatory Subunit Associated Protein 2	5	0.0019	intronic	3
PALLD	Palladin, Cytoskeletal Associated Protein	49	0.0025	intronic	15
CDH4	Cadherin 4	52	0.0029	intronic	4
PTK2	Protein Tyrosine Kinase 2	40	0.0037	intronic, 3'UTR	39
SORBS1	Sorbin And SH3 Domain Containing 1	45	0.0047	intronic, 3'UTR	1
INPP5D	Inositol Polyphosphate-5-Phosphatase D	10	0.0047	intronic, upstream	1
LRP1B	LDL Receptor Related Protein 1B	46	0.0051	intronic	44
ARRB1	Arrestin Beta 1	3	0.0052	intronic	0
DOCK4	Dedicator Of Cytokinesis 4	3	0.0053	intronic	1
NEO1	Neogenin 1	5	0.0068	intronic	1
GRID2	Glutamate Ionotropic Receptor Delta Type	32	0.0069	intronic	29
CDC42EP4	CDC42 Effector Protein 4	1	0.0071	intronic	0
ROBO1	Roundabout Guidance Receptor 1	13	0.0072	intronic	8
CAV1	Caveolin 1	17	0.0077	intronic, upstream,	2
PRKCI	Protein Kinase C Iota	3	0.0085	intronic	2
MYO1E	Myosin IE	3	0.0091	intronic	1
TBC1D4	TBC1 Domain Family Member 4	6	0.0092	intronic	1
KIF2A	Kinesin Family Member 2A	4	0.0099	intronic	4
CTNNA2	Catenin Alpha 2	74	0.0099	intronic	24

[1] Number DN-associated SNVs (OR>1.5, *P*<0.05) used in the F-SKAT analysis

[2] Number of SNVs that are significant *cis*-eQTLs in the glomeruli from patients with nephrotic syndrome (NS)

Supplementary Table 11. Functional enrichment test (KEGG pathways) on the core genes within the XPodNet network in Figure 4.
Please find Supplementary Table 11 as a separate excel file (Supplementary Table 11.xlsx).

Supplementary Table 12. Protein-altering SNVs replicated in FinnDiane cohort (combined P-value < 0.05, OR>1.5), in each genetic model.
Please find Supplementary Table 12 as a separate excel file (Supplementary Table 12.xlsx).

Supplementary Table 13. Non-coding RNA SNVs replicated in FinnDiane cohort (combined P-value < 0.05, OR>1.5), in each genetic model.
Please find Supplementary Table 13 as a separate excel file (Supplementary Table 13.xlsx).

Supplementary Table 14a. Power estimation of discovery cohort (76 discordant sibling pairs) on the whole genome level of significance (12 million, $P<4.11\times10^{-9}$) of case-only and control-only variants. Power estimation based different penetrance.

Penetrance	Dominant Model			Recessive Model		
	MAF	Relative Risk (AA+Aa)	Power (%)	MAF	Relative Risk (AA)	Power (%)
0.8	**0.005**	1.62	10.56	**0.005**	1.60	9.23
	0.01	1.63	12.05	**0.05**	1.60	9.54
	0.05	1.79	31.31	**0.1**	1.62	10.56
	0.1	2.09	71.45	**0.2**	1.67	15.67
0.9	**0.005**	1.82	78.63	**0.005**	1.80	76.13
	0.01	1.83	81.02	**0.05**	1.80	76.76
	0.05	2.02	94.90	**0.1**	1.82	78.63
	0.1	2.35	99.77	**0.2**	1.88	85.50
1.0	**0.005**	2.02	99.83	**0.005**	2.00	99.77
	0.01	2.04	99.87	**0.05**	2.01	99.79
	0.05	2.24	99.99	**0.1**	2.02	99.83
	0.1	2.61	100	**0.2**	2.09	99.93

Estimation of statistical power (presented as %) for sibship cohort using method proposed by method adapted from Li Z. et al.[21]
MAF (Minor Allele Frequency) range adjusted according to Table 2 and Table S12. Odds Ratio is infinite for all case-only and control-only in recessive and dominant model.
Prevalence of diabetic nephropathy in diabetic patients, here assumed to be 30%.[22]
Relative Risk (RR) was calculated based on P_1/P_0, where P_1=probability of DN given exposure of the genotype in the respective model, and P_0=probability of DN given no exposure of the genotype in the respective model.
Estimated within family correlation was set to 0.4 (see discussion in original publication Li Z. et al.[21]).
Significance level, Bonferroni corrected p-value = 0.05/12,165,600 = 4.11×10^{-9}

Supplementary Table 14b. Power estimation of replication cohort (2,187 controls and 1,344 cases) with genome wide significance level ($P<7.3\times10^{-6}$) with one-stage study design.

Statistical power (presented as %) in Dominant Model

Odds Ratio	Relative Risk (AA+Aa)	Power (%) for different MAF			
		0.01	0.05	0.1	0.2
1.4	**1.25**	0.1%	4.8%	19.4%	38.1%
1.5	**1.30**	0.2%	12.6%	41.9%	66.8%
2	**1.54**	4.2%	88.2%	99.7%	100%
2.6	**1.76**	21.3%	99.9%	100%	100%

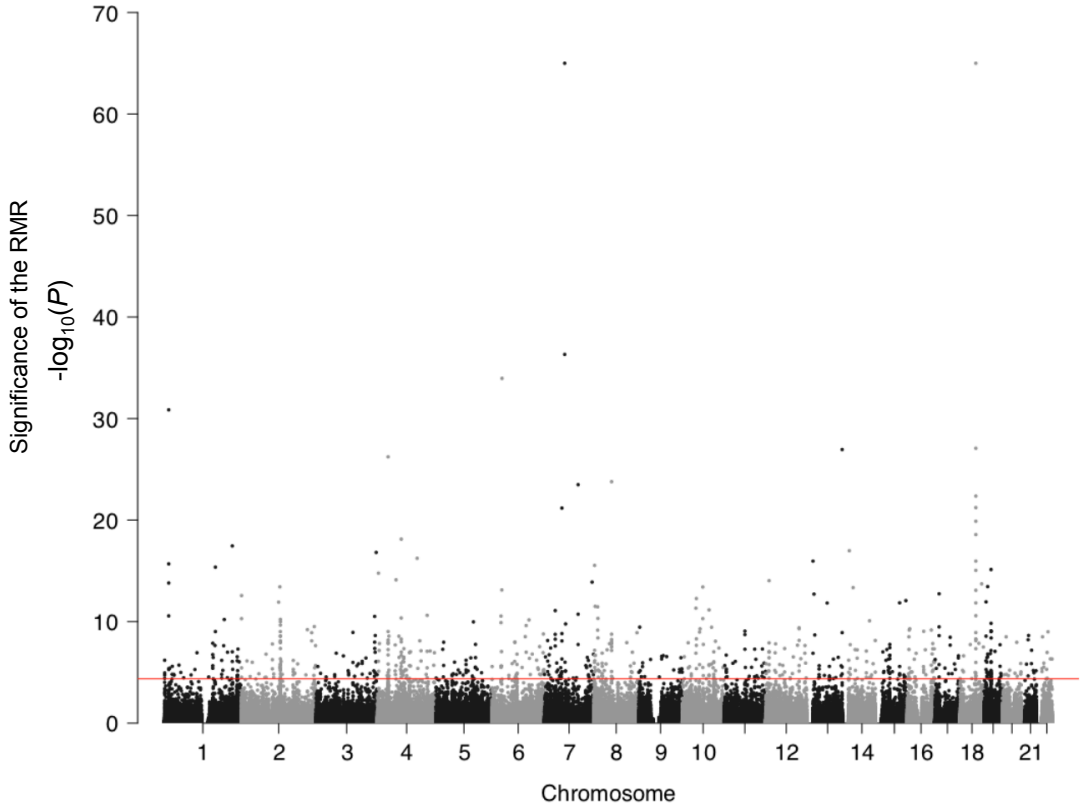
Statistical power (presented as %) in Recessive Model

Odds Ratio	Relative Risk (Aa)	Power (%) for different MAF			
		0.05	0.1	0.2	0.3
1.4	**1.25**	0	0	0	0.6%
1.5	**1.30**	0	0	0	1.5%
3	**1.88**	0	0.1%	27.3%	97.4%
5	**2.27**	0	1%	83.9%	100%
8	**2.58**	0	3.3%	98.6%	100%

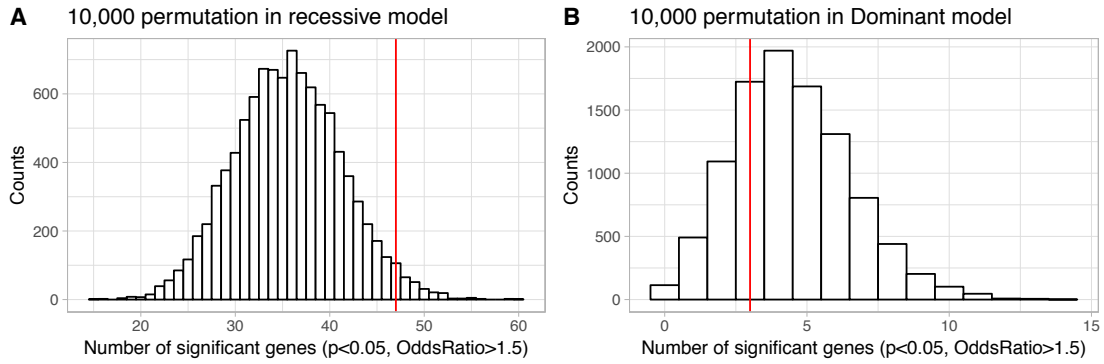
Estimation of statistical power (presented as %) by GAS Power Calculator using additive and recessive models. http://csg.sph.umich.edu/abecasis/cats/gas_power_calculaor/
MAF (Minor Allele Frequency) range adjusted according to Table 2 and Table S12.
Odds Ratio and Minor Allele Frequency (MAF) selection based on the replicated SNVs reported in Table S12.
RR: Relative risk, calculated as RR = OR/((1-prev)+(prev×OR)), where prev is the prevalence of diabetic nephropathy in in diabetic patients, here assumed to be 30%.[22]
Significance level, Bonferroni corrected p-value = 0.05/6821 = 7.3×10^{-6}

Supplementary Table 15. Test variants association in discovery cohort for the previously reported SNVs (Single Nucleotide Variants). SNVs were downloaded from GWAS Catalog by trait "Diabetic Nephropathy".
Please find Supplementary Table 15 as a separate excel file (Supplementary Table 15.xlsx).

SUPPLEMENTARY FIGURES

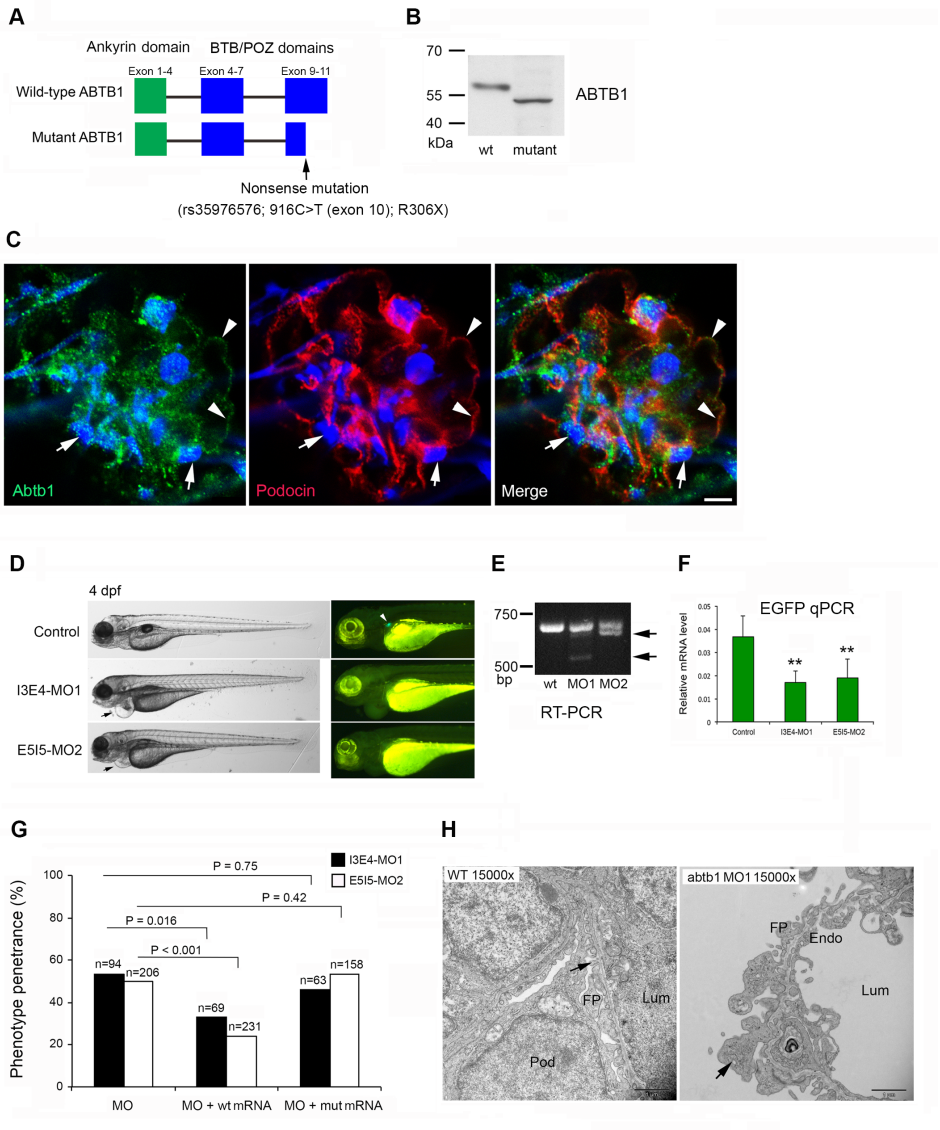


Supplementary Figure 1. Manhattan plot of the recurrently mutated regions (RMR) identified genome-wide in the 76 T1D discordant sibling pairs using the method proposed by Weinhold *et al*[34].
Each dot in the graph represents a RMR and for each RMR a P-value of significance is calculated using the negative binomial distribution, taking into account the length of the candidate mutated region, the number of mutations in the region and the background mutation rate for a similar sized region (estimated using the genome-wide expectation). The red line represents the Bonferroni corrected $P=3.7 \times 10^{-5}$ threshold used to identify the RMRs that are significantly recurrently mutated compared to a random distribution of mutations across the genome.



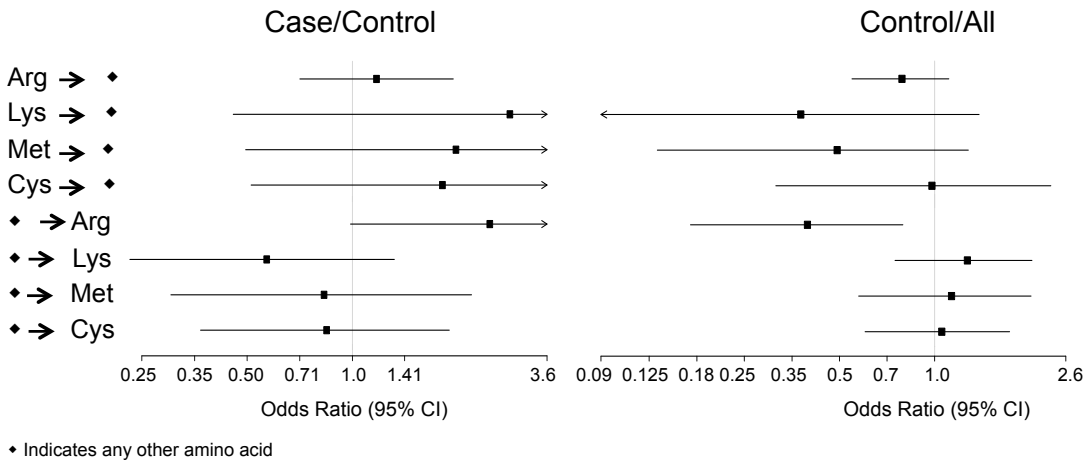
Supplementary Figure 2. Estimation of replication false positive rate on protein-altering variants (PAV) in FinnDiane cohort.

To evaluate the false positive rate of PAV replication, we performed an empirical permutation test in replication cohort using 10,000 random sets of PAVs with the exact number for recessive (3256) and dominant (306) SNVs used in our study. We found that only 2.3% of the PAV sets result in more than 47 replicated variants by chance alone using the recessive model (OR>1.5, *P*<0.05), while this proportion rises to 65.8% for the PAVs replicated using the dominant model (OR>1.5, *P*<0.05). Red line indicates the number of replicated SNVs in discovery cohort using recessive and dominant model.



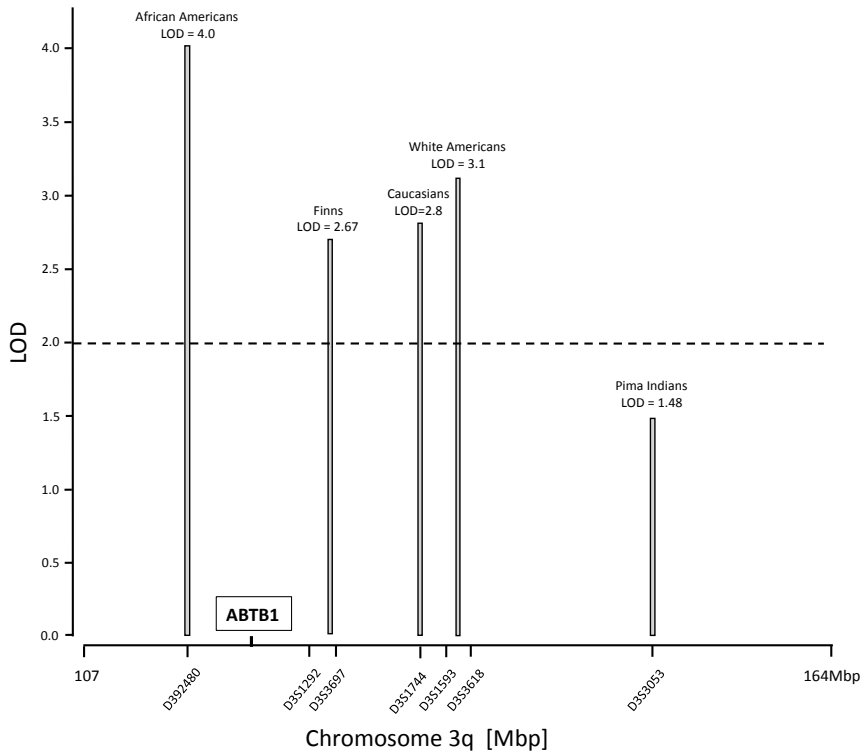
Supplementary Figure 3. Expression and functional analysis of Abtb1.
(**a**) Schematic structure of wild-type (wt) and mutant ABTB1 protein. The mutation position is indicated (arrow). (**b**) Western blot of ABTB1. The full-length ABTB1 cDNA plasmid and the mutant ABTB1 cDNA (R306X) were transfected into cultured podocytes and whole lysates were extracted after 2 days post-transfection. (**c**) Abtb1 immunofluorescence staining of the adult mouse kidney section. An ABTB1 antibody (green) and a podocin antibody (Red), a podocyte marker, were used for staining. Podocyte foot processes and podocyte nuclei are indicated with arrow and arrowheads, respectively. Scale bar = 5 µm. (**d**) Morpholinos (MO) knockdown of abtb1 in zebrafish. Injection of two-splicing inhibition MOs, I3E4-MO1 and E5I5-MO2 leads to similar kidney phenotype, pericardial edema in bright-field imaging (arrow) and loss of glomerular GFP expression in dark-field imaging (arrowhead) at 4 dpf (days post-fertilisation). The control is the control-MO morphant. (**e**) Efficacy of MO knockdown of abtb1 gene detected by RT-PCR. I3E4-MO1 (MO1) and E5I5-MO2 (MO2) resulted in in-frame deletion of 145-bp in exon 4 and deletion of 45-bp in exon 6, respectively. Arrows indicate size of deleted bands compared to wt one. (**f**) Quantitative evaluation of glomerular EGFP expression by real-time qPCR. Relative

mRNA level (mean $2^{-\Delta Ct} \pm$ SD) is presented after normalized with ß-actin in wt and the two morphants. ** $P<0.01$ in comparison with wt. (**g**) Rescue assay of human wt and mutant ABTB1 mRNA. Bar graphs show penetrance of pericardial edema of morphants at 4 dpf (y-axel). Penetrance of pericardial edema caused by injection of two MOs is significantly reduced by co-injection of wt human ABTB1 mRNA (100 pg/embryo), but not by mutant ABTB1 mRNA (100 pg/embryo). The number of morphants injected is indicated above corresponding bars. (**h**) Ultrastructural analysis of 4 dpf larval glomerulus. Compared to the wt larva, glomerulus in MO1 morphant displays overt damage including massive foot process effacement and abnormality; uneven and disorganized glomerular basement membrane, severely distorted endothelium. Arrow indicates podocyte foot processes. Scale bar = 1 µm.



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 ◆).

For each test, odds ratios and their 95% confidence intervals (CI) are reported.



Supplementary Figure 5. Chromosome 3q21 locus for DN susceptibility that was previously identified.
Previous reports by four independent genome-wide linkage studies[23-26]. For each study, the genetic markers that showed the most significance linkage are reported, together with the logarithm (base 10) of the odds (LOD score) and the population where linkage was detected. The dotted line (LOD score = 2.0) indicates the suggestive significance threshold for positive linkage. In the graph we also indicate the location of *ABTB1* that was identified in this study as a candidate gene for DN susceptibility in Finns.

WEB RESOURCES

1000 Genomes, http://www.internationalgenome.org/
ANNOVAR, http://annovar.openbioinformatics.org/
Bedtools, https://bedtools.readthedocs.io/en/latest/
BWA, http://bio-bwa.sourceforge.net/
ChromHMM, http://compbio.mit.edu/ChromHMM/
ENCODE, https://www.encodeproject.org/
ExAC, http://exac.broadinstitute.org/
Enrichr, http://amp.pharm.mssm.edu/Enrichr/
FANTOM5, http://fantom.gsc.riken.jp/5/
F-SKAT, http://www.soph.uab.edu/ssg/software
GATK, https://software.broadinstitute.org/gatk/
HOMER, http://homer.ucsd.edu/homer/
Logistf, https://www.rdocumentation.org/packages/logistf/versions/1.23/topics/logistf
Micmac3, https://github.com/guirudave/micmac3
Picard toolkit, https://broadinstitute.github.io/picard/
Plink, http://zzz.bwh.harvard.edu/plink/
PolyPhen-2, http://genetics.bwh.harvard.edu/pph2/
RefSeq, https://www.ncbi.nlm.nih.gov/RefSeq
Remap, http://tagc.univ-mrs.fr/remap/
Roadmap Epigenome, http://www.roadmapepigenomics.org/
SIFT, http://sift.bii.a-star.edu.sg/
GAS power calculator, http://csg.sph.umich.edu/abecasis/cats/gas_power_calculaor/

DNC browser, http://dnc.systems-genetics.net
Temporary access (username: review; password: reviewer) open for review purpose.

REFERENCE:

1 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

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

3 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-498, doi:10.1038/ng.806 (2011).

4 Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164, doi:10.1093/nar/gkq603 (2010).

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

6 Griffon, A. *et al.* Integrative analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory landscape. *Nucleic Acids Res* **43**, e27, doi:10.1093/nar/gku1280 (2015).

7 Consortium, F. *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462-470, doi:10.1038/nature13182 (2014).

8 Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455-461, doi:10.1038/nature12787 (2014).

9 Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**, 215-216, doi:10.1038/nmeth.1906 (2012).

10 Woroniecka, K. I. *et al.* Transcriptome analysis of human diabetic kidney disease. *Diabetes* **60**, 2354-2369, doi:10.2337/db10-1181 (2011).

11 Fagerstrom, S., Pahlman, S., Gestblom, C. & Nanberg, E. Protein kinase C-epsilon is implicated in neurite outgrowth in differentiating human neuroblastoma cells. *Cell Growth Differ* **7**, 775-785 (1996).

12 Geraldes, P. & King, G. L. Activation of protein kinase C isoforms and its impact on diabetic complications. *Circ Res* **106**, 1319-1331, doi:10.1161/CIRCRESAHA.110.217117 (2010).

13 Dixon, J. R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331-336, doi:10.1038/nature14222 (2015).

14 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25, doi:10.1186/gb-2009-10-3-r25 (2009).

15 Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-589, doi:10.1016/j.molcel.2010.05.004 (2010).

16 Satterlee, J. S. *et al.* Community resources and technologies developed through the NIH Roadmap Epigenomics Program. *Methods Mol Biol* **1238**, 27-49, doi:10.1007/978-1-4939-1804-1_2 (2015).

17 Gillies, C. E. *et al.* An eQTL landscape of kidney tissue in human nephrotic syndrome. *bioRxiv*, doi:10.1101/281162 (2018).

18 Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7, doi:10.1186/s13742-015-0047-8 (2015).

19 He, B., Ebarasi, L., Hultenby, K., Tryggvason, K. & Betsholtz, C. Podocin-green fluorescence protein allows visualization and functional analysis of podocytes. *Journal of the American Society of Nephrology : JASN* **22**, 1019-1023, doi:10.1681/ASN.2010121291 (2011).
20 Robu, M. E. *et al.* p53 activation by knockdown technologies. *PLoS Genet* **3**, e78, doi:10.1371/journal.pgen.0030078 (2007).
21 Li, Z., McKeague, I. W. & Lumey, L. H. Optimal design strategies for sibling studies with binary exposures. *Int J Biostat* **10**, 185-196, doi:10.1515/ijb-2014-0015 (2014).
22 Gheith, O., Farouk, N., Nampoory, N., Halim, M. A. & Al-Otaibi, T. Diabetic kidney disease: world wide difference of prevalence and risk factors. *J Nephropharmacol* **5**, 49-56 (2016).
23 Moczulski, D. K., Rogus, J. J., Antonellis, A., Warram, J. H. & Krolewski, A. S. Major susceptibility locus for nephropathy in type 1 diabetes on chromosome 3q: results of novel discordant sib-pair analysis. *Diabetes* **47**, 1164-1169 (1998).
24 Bowden, D. W. *et al.* A genome scan for diabetic nephropathy in African Americans. *Kidney Int* **66**, 1517-1526, doi:10.1111/j.1523-1755.2004.00915.x KID915 [pii] (2004).
25 Imperatore, G. *et al.* 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 (1998).
26 Österholm, A. M. *et al.* 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, doi:5001933 [pii] 10.1038/sj.ki.5001933 (2007).
27 He, B. *et al.* Association of genetic variants at 3q22 with nephropathy in patients with type 1 diabetes mellitus. *Am J Hum Genet* **84**, 5-13, doi:S0002-9297(08)00597-1 [pii] 10.1016/j.ajhg.2008.11.012 (2009).
28 Unoki, M. & Nakamura, Y. Growth-suppressive effects of BPOZ and EGR2, two genes involved in the PTEN signaling pathway. *Oncogene* **20**, 4457-4465, doi:10.1038/sj.onc.1204608 (2001).
29 Lin, J. *et al.* Loss of PTEN promotes podocyte cytoskeletal rearrangement, aggravating diabetic nephropathy. *J Pathol* **236**, 30-40, doi:10.1002/path.4508 (2015).
30 Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).
31 Majumdar, A. & Drummond, I. A. Podocyte differentiation in the absence of endothelial cells as revealed in the zebrafish avascular mutant, cloche. *Dev Genet* **24**, 220-229, doi:10.1002/(SICI)1520-6408(1999)24:3/4<220::AID-DVG5>3.0.CO;2-1 (1999).
32 Tang, W., Zhang, D. & Ma, X. RNA-sequencing reveals genome-wide long non-coding RNAs profiling associated with early development of diabetic nephropathy. *Oncotarget* **8**, 105832-105847, doi:10.18632/oncotarget.22405 (2017).
33 Sheu, W. H. *et al.* Genome-wide association study in a Chinese population with diabetic retinopathy. *Hum Mol Genet* **22**, 3165-3173, doi:10.1093/hmg/ddt161 (2013).

34 Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat Genet* **46**, 1160-1165, doi:10.1038/ng.3101 (2014).