

Whole-exome and Whole-genome Sequencing of 1097 Individuals with Type 1 Diabetes Reveals Novel Genes for Diabetic Kidney Disease

Jani K Haukka^{1,2,3}, Anni A Antikainen^{1,2,3}, Erkka Valo^{1,2,3}, Anna Syreeni^{1,2,3}, Emma H Dahlström^{1,2,3},
Bridget M Lin⁴, Nora Franceschini⁴, Valma Harjutsalo^{1,2,3}, Per-Henrik Groop^{1,2,3,5}, Niina Sandholm^{1,2,3}
on behalf of the FinnDiane Study Group

¹ Folkhälsan Institute of Genetics, Folkhälsan Research Center, Helsinki

² Department of Nephrology, University of Helsinki and Helsinki University Hospital, Helsinki, Finland

³ Research Program for Clinical and Molecular Metabolism, Faculty of Medicine, University of Helsinki, Helsinki, Finland

⁴ Department of Epidemiology, University of North Carolina, Chapel Hill, NC, United States

⁵ Department of Diabetes, Central Clinical School, Monash University, Melbourne, Victoria, Australia

Corresponding authors:

Niina Sandholm, DSc
Folkhälsan Research Center and
University of Helsinki and
Helsinki University Hospital
Biomedicum Helsinki
Haartmaninkatu 8
00290 Helsinki, Finland
e-mail: niina.sandholm@helsinki.fi

Per-Henrik Groop, MD, DMSc, FRCPE
Folkhälsan Research Center and
University of Helsinki and
Helsinki University Hospital
Biomedicum Helsinki (C318b)
Haartmaninkatu 8
00290 Helsinki, Finland
e-mail: per-henrik.groop@helsinki.fi

Word count: 3480

Figures: 2

Tables: 4

Supplementary Figures: 4

Supplementary Tables: 10

Abstract

Background and hypothesis: Diabetic kidney disease (DKD) is a severe diabetic complication affecting one third of individuals with type 1 diabetes. Although several genes and common variants have been associated with DKD, much of the predicted inheritance remain unexplained. Here, we performed next-generation sequencing to assess whether low-frequency variants — single or aggregated — contribute to the missing heritability in DKD.

Methods: We performed whole-exome sequencing (WES) of 498 individuals and whole-genome sequencing (WGS) of 599 individuals with type 1 diabetes. After quality control, we had next-generation sequencing data available for altogether 1064 individuals, of whom 546 had developed either severe albuminuria or end-stage kidney disease, and 528 had retained normal albumin excretion despite a long duration of type 1 diabetes. Single variants and gene aggregate tests were performed separately for WES and WGS data and combined with meta-analysis. Furthermore, we performed genome-wide aggregate analyses on genomic windows (sliding-window), promoters, and enhancers with the WGS data set.

Results: In single variant meta-analysis, no variant reached genome-wide significance, but a suggestively associated *THAP7* rs369250 variant ($P=1.50\times 10^{-5}$) was replicated in the FinnGen general population GWAS data for chronic kidney disease (CKD) and DKD phenotypes. Gene-aggregate meta-analysis identified suggestive evidence ($P<4.0\times 10^{-4}$) at four genes for DKD, of which *NAT16* and *LTA* (TNB- β) replicated in FinnGen. Of the intergenic regions suggestively associated with DKD, the enhancer on chromosome 18q12.3 ($P=3.94\times 10^{-5}$) showed interaction with the *METTL4* gene; the lead variant was replicated, and predicted to alter Maf β binding.

Conclusions: Our sequencing-based meta-analysis revealed multiple genes, variants and regulatory regions suggestively associated with DKD. However, as no variant or gene reached genome-wide significance, further studies are needed to validate the findings.

What was known:

- Genetics is an important factor in the development and progression of diabetic kidney disease (DKD) in individuals with type 1 diabetes.
- Previously identified genetic associations have mostly been common variants as they originated from GWAS studies. Based on inheritance estimates, the current findings only explain a fraction of the predicted disease risk.

This study adds:

- Our study with 1097 sequenced individuals with type 1 diabetes is to date one of the largest sequencing studies on DKD in type 1 diabetes.
- The study reveals several suggestive variants, genes and intergenic regulatory regions associated with DKD. Low-frequency protein-altering variants inside *NAT16* and *LTA* (encoding for TNF- β), and chromosome 18q12.3 enhancer variant linking to *METTL4* were also replicated in FinnGen kidney disease phenotypes.

Potential impact:

- The results suggest novel genes that may be important for the onset and development of serious DKD in individuals with type 1 diabetes. In addition to revealing novel biological mechanisms leading to DKD, they may reveal novel treatment targets for DKD. However, further validation and functional studies are still needed.

1 INTRODUCTION

2 Type 1 diabetes (T1D) is an autoimmune disease caused by the destruction of the insulin-secreting beta cells
3 in the islets of Langerhans in the pancreas. Long-term insulin irregularity leads to complications in several
4 organs for a large proportion of individuals with T1D [1]. In particular, prolonged hyperglycaemia leads to a
5 decline of kidney function and diabetic kidney disease (DKD) for approximately 30% of individuals with T1D
6 [1, 2]. In the Western world, DKD is the most common cause of kidney failure, which can be treated only with
7 dialysis or kidney transplantation [3]. In addition, DKD predisposes the individuals to cardiovascular disease
8 (CVD), and already early-stage DKD (moderate albuminuria) elevates the risk of myocardial infarction and
9 stroke two to three-fold [4, 5].

10 Both genetic and environmental factors affect the occurrence of T1D and its complications. Heritability
11 estimates from genome-wide association studies (GWAS) suggest that genetic factors explain approximately
12 one-third of the DKD risk [6, 7]. Microarray-based chips used in GWASs include hundreds of thousands of
13 common human-variable loci, excellent for the study of common variants having modest effects on the
14 disease risk [8, 9]. Indeed, GWASs have shed light upon DKD mechanisms, but these earlier findings explain
15 only a minority of the predicted genetic risk of DKD [10-13]. Our recent family-based linkage study performed
16 on GWAS data suggested a role for rare genetic variants as risk factors for the development of DKD as well
17 [14]. Recently, whole-exome and whole-genome sequencing (WES and WGS, respectively) have enabled the
18 study of low-frequency and rare variants that are expected to have larger effects on the disease risk. WES
19 offers a computationally simpler way of studying protein-altering or truncating variants (PAV and PTV,
20 respectively) [15]. The WGS studies additionally enable exploration of the intronic and intergenic regions,
21 which may affect the gene expression levels through transcription-factor binding-site activity or other
22 regulatory processes [16]. Recently, a WGS in mainly non-diabetic individuals from multiple ancestries
23 identified three novel loci for estimated glomerular filtration rate (eGFR) [17]. However, there are currently
24 only a few WES or WGS-based studies for DKD. Our previous WES study on DKD in T1D yielded no significant
25 findings, whereas a recent WES study identified four exome-wide significant loci for DKD [6, 18].

26 Furthermore, our previous WGS of 74 sibling pairs proposed involvement of protein kinase C family members
27 with DKD [14, 16]. Here, we study the effect of low-frequency and rare variants on DKD by a WES and WGS
28 meta-analysis design, and perform genome-wide and regulome-wide scans for the non-coding regions using
29 the WGS data (**Figure 1**).

30

31 **MATERIALS AND METHODS**

32 **Study population and phenotypes**

33 All participants were recruited from the Finnish Diabetic Nephropathy Study (FinnDiane). Participants were
34 diagnosed with T1D by their attending physician, had an onset of diabetes before the age of 40, and initiated
35 permanent insulin treatment within the first year after the diagnosis (See **Supplementary Methods**, and [19]
36 for more detail). In brief, data on diabetic complications, history of cardiovascular events, and prescribed
37 medications were registered using standardized questionnaires, and blood and urine samples were collected
38 during a standard visit to the attending physician. DKD was defined based on albuminuria status and cases
39 had either severe albuminuria (≥ 200 $\mu\text{g}/\text{min}$) in two out of three consecutive urine collections, or kidney
40 failure requiring dialysis or a kidney transplant, whereas controls had retained normal albumin excretion rate
41 (AER < 20 $\mu\text{g}/\text{min}$) throughout the follow-up.

42 **Sequencing and data analysis**

43 WES and WGS were performed for 498 and 599 individuals, respectively. WES was performed at the
44 University of Oxford with Illumina HiSeq2000 platform, with an average requirement of 20 \times target capture
45 with $\geq 80\%$ coverage as described earlier [6]. WGS was performed with Illumina HiSeq X platform (Macrogen
46 Inc., Rockville, MD, USA), with a requirement of $> 30\times$ average coverage for mapped reads. Based on the initial
47 quality control (QC) performed at Oxford for WES, or by Macrogen Inc. for WGS, 27 WES samples and 15
48 WGS samples were excluded due to high homozygosity-heterozygosity ratio, abnormally low mapping depth,
49 or mapped PCR reads.

50 The WES and WGS samples were processed with the Genome Analysis Toolkit (GATK) 4 Golden Standard
51 pipeline (**Supplementary Methods**) and annotated with SnpEff v5.0e [20, 21]. Post-pipeline variants with
52 <98% call rate and Hardy–Weinberg Equilibrium (HWE) P -value $<1 \times 10^{-10}$ were excluded. Principal component
53 analysis with PLINK indicated no population outliers (**Supplementary Figure 1**).

54 **Variant-based tests**

55 We performed single variant genome-wide association testing for the WGS data with Firth regression, using
56 age of diabetes onset, sex, and the two first genetic principal components as covariates. For the meta-analysis
57 of the WES and WGS data, we performed association testing with inverse variant weighted score test statistic
58 meta-analysis. In WGS single variant analysis, variants with P -value $<5 \times 10^{-8}$ were considered genome-wide
59 significant and $P < 5 \times 10^{-5}$ suggestive. In WES-WGS meta-analysis, $P < 3.5 \times 10^{-7}$ was considered exome-wide
60 significance and $P < 3.5 \times 10^{-5}$ as suggestive. All single variant tests were performed with RVTESTS (v2.1.0) and
61 the single variant meta-analysis with METAL v2011-03-25 [22, 23].

62 **Gene-aggregation tests**

63 We performed gene-based tests with WES and WGS data, separately for protein-truncating variants (PTVs,
64 including frameshift and nonsense variants, and loss or gain of start or stop codons) and protein-altering
65 variants (PAVs, including missense variants, in-frame insertions and deletions (indels), and PTVs), designated
66 to specific genes within RefSeq exomes (**Supplementary Table 1**). Variants were further filtered by their
67 minor allele frequency (MAF) to those with $MAF \leq 1\%$, $\leq 5\%$, and $\leq 10\%$. P -values $< 4 \times 10^{-6}$ were considered
68 significant (adjusted for 18,226 genes with PAVs with $MAF \leq 10\%$) and $P < 4 \times 10^{-4}$ suggestive. Gene-based SKAT-
69 O meta-analysis was performed for WES and WGS cohorts with MetaSKAT v0.8.1 R-package [24].

70 **Lookup of monogenic kidney disease genes**

71 We studied whether known monogenic kidney disease genes were significantly enriched in our gene-based
72 meta-analysis for DKD. We considered altogether 464 unique genes causing syndromic or monogenic kidney
73 diseases as listed by Connaughton *et al.* [25].

74 **Whole-genome sliding window and regulome-wide analysis**

75 We performed the genome-wide sliding window tests and aggregation tests for promoters and enhancers
76 for WGS data using STAAR R-package v0.9.6 using the omnibus STAAR-O test [26]. The genome-wide window
77 scan tests were done with window size of 4,000 base pairs (bp) with a gap of 2,000 bps between the windows,
78 and a minimum of 5 variants within the region. As enhancer and promoter regions, we considered the
79 FANTOM5 CAGE profiles (lifted over to GRCh38) (**Supplementary Methods**). At least two variants were
80 required for these regulatory region tests. MAF<5% threshold was used for all the tests. Altogether 174,576
81 promoters and 24,240 enhancers were tested, resulting in Bonferroni corrected significance levels of
82 $P < 2.86 \times 10^{-7}$ and 2.12×10^{-6} .

83 Putative enhancer target genes were inspected with Zenbu promoter capture Hi-C data from the YUE Lab
84 database (<http://3dgenome.fsm.northwestern.edu/chic.php>) [27]. As no Hi-C data were available for
85 kidneys, we searched for bladder as a related tissue.

86 **Replication**

87 Variants and genes were replicated with FinnDiane GWAS data (**Supplementary Methods**, [28]). Altogether
88 6,449 individuals and 15.21M variants with imputation quality $r^2 > 0.7$ passed the QC. After exclusion of
89 individuals included in WES or WGS, the replication was tested in 3,724 individuals with T1D, whereby cases
90 had severe albuminuria or ESRD, and controls had normal AER with minimum of 26 years diabetes duration.
91 Additionally, replication was attempted in GWAS data for 2,356 non-overlapping Finnish individuals with T1D
92 from the THL Biobank (THLBB) diabetes studies collection (thl.fi/biobank), with registry data for ESRD based
93 on ICD codes available for 2,131 individuals. This GWAS data was imputed with the same pipeline as the main
94 data, using SISu v4 reference panel, and included 16.94M variants after the QC. As albuminuria data were
95 not available for these individuals, they were grouped based on ESRD occurrence — altogether there were
96 70 individuals with ESRD, and 2,061 individuals without ESRD.

97 Single variant replication was further tested with Finnish general population GWAS data from the FinnGen
98 project r9 release (<https://r9.finnngen.fi/>) with four kidney disease phenotypes. Finally, we requested
99 replication of the three non-HLA region genes and 16 SNPs in the Trans-Omic for Precision Medicine
100 (TopMED) eGFR WGS study on multi-ancestry general population with 23,732 individuals [17]. P-values
101 <0.0029 were considered Bonferroni significant after correction for altogether 17 tested variants.

102 ***in silico* annotation of the lead genes and variants**

103 We queried the Human Kidney expression quantitative trait locus (eQTL) Atlas to identify the variants
104 associated with gene expression [29]. Differential gene expression was investigated in kidney conditions data
105 sets in the Nephroseq classic portal v4. Transcription factor Affinity Prediction (TRAP) tool
106 (http://trap.molgen.mpg.de/cgi-bin/trap_two_seq_form.cgi) was used to predict transcription factors with
107 differential binding affinity to the lead regulatory variant reference and alternative allele sequences by
108 searching the JASPAR and TRANSFAC vertebrate motifs with the human background model [30]. Linkage
109 disequilibrium (LD) was evaluated with NIH LDlink tools using the Finnish population (<https://ldlink.nih.gov/>).

RESULTS

Study cohorts

In both the WES and WGS cohorts the cases and controls had similar age of T1D onset and BMI, whereby controls had longer duration of T1D and, by selection, higher baseline eGFR (**Table 1**). All the following analyses were adjusted with sex, age of T1D onset and the two most important genetic principal components.

Single variant associations

The meta-analysis of WES and WGS data resulted in six variants, including two PAVs, with a suggestive P -value of 3.5×10^{-5} (**Table 2** and **Figure 2**). Among these, rs369250 in *THAP7*, and rs1048365 in *AP1S1* showed significant eQTL activity in human kidneys with *THAP7-AS1* ($P=8.488 \times 10^{-45}$) and *AP1S1* ($P=4.98 \times 10^{-175}$), respectively (**Supplementary Table 2A**). Extending the analysis outside the exons and their flanking regions, the single variant association test for WGS data identified variants on two intergenic regions 14p12 (close to *CPSF2*) and 16p11.1 (near a group of RNA-genes and pseudogenes), and two intronic variants in *MYO9B* suggestively associated with DKD ($P\text{-value} < 5 \times 10^{-6}$; **Supplementary Table 3**).

We attempted replication of the suggestive single variant DKD associations with the non-overlapping FinnDiane and THLBB GWAS data for individuals with T1D (**Supplementary Table 4**), general population FinnGen GWAS data using four kidney disease definitions (**Supplementary Table 5**), and in the TOPMed WGS data for CKD. The *THAP7* promoter rs369250 was replicated for CKD ($P=2.7 \times 10^{-4}$, Bonferroni significant) and DKD ($P=0.012$) in the FinnGen GWAS data (**Supplementary Table 5**).

Gene aggregate tests

In gene-aggregate meta-analysis, *NAT16*, *LTA*, *SLC10A6*, and *TSEN54* reached a suggestive P -value $< 4.0 \times 10^{-4}$ for analyses with low-frequency PAVs (**Table 3**). *NAT16* contained six PAVs with $\text{MAF} < 0.1$, and the association was driven by the rs34985488 (A>C p.Phe63Cys, $P=5.82 \times 10^{-5}$) (**Supplementary Table 6**). The variant was predicted deleterious or possibly damaging by the SIFT and PolyPhen algorithms, respectively.

Replication of the gene aggregate results was sought in FinnDiane GWAS, THLBB GWAS, TOPMed WGS, and UKBB WES gene aggregate data. Furthermore, for the single variants that were nominally significant ($P < 0.05$) in the FinnDiane WES+WGS meta-analysis, we tested for replication in the FinnDiane, THLBB, and FinnGen GWAS, and TOPMed WGS data. Of the 4 tested suggestive genes, *NAT16* rs34985488 was replicated in the FinnGen for the CKD phenotype ($P = 0.0028$; Bonferroni significant) and *LTA* rs2229092 for the T1D with kidney complications phenotype ($P = 0.0044$; rs2229092 was the only identified PAV in *LTA*, **Supplementary Table 5**). No replication was observed in the gene aggregate tests (**Supplementary Table 7**).

Enrichment of Monogenic or syndromic kidney disease genes

We further tested whether monogenic kidney disease genes are enriched among DKD-associated genes (**Supplementary Table 8**). The strongest evidence of enrichment occurred among the “cystic kidney disease or nephronophthisis” class, as 11% (11 out of 96) of the monogenic genes were associated ($P < 0.05$) with DKD in our WES-WGS meta-analysis (Binomial test $P = 0.004$).

Intergenic variant aggregate tests

To improve the statistical power to discover non-coding genetic factors behind DKD, we performed promoter and enhancer aggregate analyses, and functionally-informed genome-wide sliding-window analyses. Two enhancers located at 18q12.3 ($P = 6.78 \times 10^{-5}$) and 9q21.11 ($P = 2.17 \times 10^{-4}$) were suggestively associated with DKD (**Table 4**). Based on the Hi-C database, the chromosome 18q12.3 region showed enhancer activity with the closest protein-coding gene *METTL4* (located 186kb upstream) in the bladder. The 9q21.11 locus, close to *CTSL*, is a promoter-enhancer region, which showed enhancer activity with *CTSL* (**Supplementary Table 2C**).

In the sliding window analysis, the strongest associations were obtained for a chromosomal window on 12q14.3, 21kbps upstream to *WSCD2* ($P = 4.66 \times 10^{-6}$) and an intergenic region 4q22.3 ($P = 9.46 \times 10^{-6}$). However, none of the genomic windows (1,318,834 regions), promoter (174,609 regions) or enhancer regions (N=24,240) remained significant after Bonferroni correction (**Table 4**).

Replication for the genome-wide sliding-window, promoter- and enhancer-wide analyses were also performed by testing the nominally significant variants (**Supplementary Table 9**) in the replication data. We observed replication for rs183413211 of the *2q14.2* region 106kb from *SLC8A1* (TOPMed $P_{\text{CKD}} = 0.042$; **Supplementary Table 4**), and rs16943099 in the *METTL4* enhancer region in 18q12.3 (FinnGen $P_{\text{T2D_with_renal_compl.}} = 8.6 \times 10^{-4}$, Bonferroni significant; and $P_{\text{DKD}} = 0.036$; **Supplementary Table 5**).

To assess the potential functional effect of the variants within the identified enhancer regions, we applied the TRAP prediction tool to estimate whether the variants affect the transcription-factor binding probability of the sequence. For the *METTL4*/18q12.3 enhancer region lead SNP, rs1694309, only the reference allele was predicted to provide a binding site for a podocyte-specific transcription factor MafB (TRANSFAC $V\$\text{MAFB_01}$, $p_{\text{REF}} < 1.75 \times 10^{-6}$, $p_{\text{ALT}} = 0.239$; **Supplementary Figure 4, Supplementary Table 2D**).

DISCUSSION

Here, we studied 1,064 Finnish individuals with T1D, representing the extreme phenotypes for DKD, to identify rare and low-frequency variants associated with DKD. We included 546 individuals with severe DKD (severe albuminuria or ESRD) and 528 with T1D and normal albumin excretion rate despite at least 25 years of diabetes duration. We performed single variant meta-analysis for the whole genome including the HLA-region, and we utilized both gene aggregate tests and the regulatory region aggregate test to identify low-frequency variants associated with DKD susceptibility, revealing several putative associations with novel and functionally plausible genes for DKD.

The gene aggregate meta-analysis for low-frequency and rare PAVs and PTVs resulted in four genes suggestively associated with DKD: *NAT16*, *LTA*, *SLC10A6*, and *TSEN54*. Through look-ups of the lead SNPs behind the gene aggregate results, replication was found for *NAT16* and *LTA* in the FinnGen cohort (**Supplementary Table 4**). *NAT16* putatively encodes N-Acetyltransferase 16. N-Acetyltransferases transfer acetyl groups from acetyl-CoA to molecules such as arylamines. According to GTEx RNA-seq data, *NAT16* is expressed in kidney cortex and medulla, along with most other tissues, and *NAT16* was previously observed

to show significantly higher expression in individuals with idiopathic nodular glomerulosclerosis compared with healthy controls (Log_2 fold change=7.11, $P=1.67\times 10^{-8}$) [31]. The lead SNP rs34985488 was classified as deleterious by SIFT and was replicated in the FinnGen CKD phenotype data set ($P=0.0028$). Interestingly, the *NAT16* missense variant rs34985488, and one of the single variant lead SNPs, rs1048365 in the *AP1S1* 3' untranslated region (UTR), are in moderate LD, and they are both eQTLs for *AP1S1*, *VGF* and *MOGAT3* expression in kidney [29] (**Supplementary Figure 2A**); *MOGAT3* encodes Monoacylglycerol O-acyltransferase 3 that catalyses the synthesis of diacylglycerol from 2-monoacylglycerol and fatty acyl-CoA [32].

The HLA-region gene *LTA* encodes lymphotoxin-alpha (LT- α), also known as tumor necrosis factor β (TNF- β), which plays an important role in the immune response, inflammation, and apoptosis. LT- α binds to TNF receptors, which have been associated with progression of DKD [33, 34]. The *LTA* association with DKD was driven by rs2229092 (OR=0.39, $P=1.513\times 10^{-5}$), and the association was replicated in FinnGen (T1D with kidney complications data set $P=0.0044$). Interestingly, in FinnGen the rs2229092 was protective of T1D (OR=0.74, $P=3.4\times 10^{-16}$), other autoimmune diseases, and diabetic complications, ophthalmic complications being the most significant OR=0.70, $P=4.5\times 10^{-13}$ (**Supplementary Table 4**). However, others have noted that assessing causal effects of rs2229092 is difficult due to pleiotropy [35]. Of note, an association between DKD and *LTA* p.T60N (rs1041981) was suggested by candidate gene studies already 15 years ago, but the association has not been replicated in more recent large GWASs [36]. This could be partly due to leaving out the HLA-region from GWAS imputation or sequencing due to complexity of the region.

The WGS data allowed us to investigate low-frequency variants also in the non-coding regions. On chromosome 18q12.3, enhancer interacting with *METTL4* gene was suggestively associated with DKD ($P=6.78\times 10^{-5}$). The association was led by rs16943099, and it replicated for "T2D with kidney complications" ($P=8.6\times 10^{-4}$) in the FinnGen GWAS data. *In silico* prediction suggested that the rs16943099 minor C allele, associated with lower risk of DKD, disrupted a transcription-factor binding site for a podocyte-specific transcription factor Mafb (**Supplementary Figure 3, Supplementary Table 2D**); forced *maf*b expression was recently shown to prevent CKD in mice [37]. *METTL4* encodes Mettl4 methyltransferase, which has been

shown to mediate m⁶A methylation on U2 snRNA *in vitro* [38]. An intergenic variant rs185299109 in the *LINC00470/METTL4* locus was previously associated with DKD (eGFR based CKD phenotype, $P=1.3\times 10^{-8}$) [10]. Moreover, methyl adenosine modification of paralogous *METTL3* was observed to promote podocyte injury in DKD [39], and a rare intronic variant in *METTL8* was one of the novel findings for eGFR in the TOPMed WGS [17].

The 9q21.11 enhancer ($P=1.17\times 10^{-4}$) and the partly overlapping promoter ($P=2.67\times 10^{-5}$) are adjacent to *CTSL*, encoding Cathepsin L. Cathepsins are proteases which degrade and detoxify advanced glycation end product (AGE)-modified proteins. Higher concentrations of urinary cathepsin L were predictive of improvement in albuminuria after four years of follow-up, and positively correlated with urinary glucosepane, a cross-linking AGE derived from D-glucose, in individuals with type 2 diabetes [40].

To date, our WES/WGS of 1,097 individuals with T1D is one of the largest sequencing studies for DKD. However, previous studies have shown that a larger sample size is often needed to discover variants with modest effect size, and thus, the main limitation of this study is the sample size. However, the study participants were carefully selected and characterized for their phenotype, and had either advanced DKD or long duration of T1D without DKD. We could not replicate common variants previously identified for DKD, such as the reported *COL4A3* missense variant rs55703767 ($P=0.056$), suggesting that we may have missed variants relevant for DKD due to the limited sample size. However, our focus here was on discovery of rare variants with functional relevance.

We used two different sequencing platforms in the study. Even though the datasets were analysed using the same pipeline, there were differing read lengths of 150 bps for WGS, and 100 bps for WES, and the mean amount of low-frequency variants per gene was significantly greater in the WGS (5.88) compared to the WES data set (5.31). Due to the limitation of databases and tools, our study included only transcribed enhancers, and the promoters were defined with an arbitrarily selected 1,000 bp extension downstream TSS, although the promoter lengths vary.

Finally, no other WGS data for DKD were available for replication of our findings, but we attempted replication in multiple data sets, including studies with imputed GWAS data for DKD in T1D, and GWAS, WES, and WGS data sets of phenotypes gathered from the general population for kidney disease. Due to these limitations, we here report replication at nominal significance ($P < 0.05$). After correcting for the total number of tested SNPs ($n=17$), 18q12.3 (*METTL4*) rs16943099, *THAP7* rs369250, and *NAT16* rs34985488 remained significant. In addition, evidence from eQTL and differential gene expression in kidney tissue supports the relevance of the novel loci identified here, especially for the 18q12.3 (*METTL4*) enhancer region, and *NAT16* and *LTA* genes.

CONCLUSIONS

In conclusion, this WES-WGS meta-analysis resulted in several plausible DKD genes that may be important for the development of DKD in individuals with T1D. However, further validation is needed to confirm their role in DKD.

Acknowledgements

We are indebted to the late Carol Forsblom (1964–2022), the international coordinator of the FinnDiane Study Group, for his considerable contribution. The skilled technical assistance of H. Krigsman, H Olanne, M. Parkkonen, M. Korolainen, A. Sandelin, J. Tuomikangas, and Kirsi Uljala (Folkhälsan Research Center, Finland) are gratefully acknowledged. The authors also acknowledge all the physicians and nurses at each center taking part in the enrolment and clinical characterization of the participants (see **Supplementary table 10** for a list of study centres and investigators involved in the FinnDiane study). The THLBB GWAS data used for the research were obtained from the THL Biobank (study number THLBB2021_10). We thank all study participants for their generous participation in the THL Biobank and THL Diabetes Studies. We acknowledge

that the ELIXIR Finland node, hosted at the CSC – IT Center for Science for ICT resources, enabled the WES and WGS data processing.

Funding

This study was supported by funding from Folkhälsan Research Foundation, Wilhelm and Else Stockmann Foundation, Liv och Hälsa Society, Munuaissäätiö, Helsinki University Hospital Research Funds (EVO TYH2018207), Academy of Finland (299200, and 316664), Novo Nordisk Foundation (NNF OC0013659, NNF23OC0082732), the Sigrid Jusélius Foundation, and the Finnish Diabetes Research Foundation.

Author contributions

JKH performed data processing and statistical analysis, and wrote the original draft. AAA performed data processing and did intergenic variant analysis pipeline. NS contributed to data analysis and investigation, edited the manuscript, and supervised the work. AS contributed to GWAS data processing and edited the manuscript. EV contributed to WGS patient selection and reviewed the manuscript, ED performed the THLBB data processing, P-HG supervised the work, BLM and NF performed the replication in TOPMed study. All authors approve the version to be submitted published, and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Data availability statement

The FinnDiane WES and WGS datasets generated and/or analyzed during the current study are not publicly available as the participants' written consent does not allow data sharing. The Readers may propose collaboration to research the individual level data with correspondence with the lead investigator.

Disclosures

P-HG has served on advisory boards for AbbVie, Astellas, AstraZeneca, Bayer, Boehringer Ingelheim, Eli Lilly, Janssen, Medscape, MSD, Mundipharma, Novartis, Novo Nordisk, Sanofi, and has received lecture honoraria from Astellas, AstraZeneca, Bayer, Boehringer Ingelheim, Eli Lilly, Elo Water, Medscape, MSD, Mundipharma, Novartis, Novo Nordisk and Sanofi. P-H G has also received investigator-initiated grants from Eli Lilly and Roche.

References

1. Forbes JM, Cooper ME. Mechanisms of diabetic complications. *2013*;93(1):137-188
2. Sigfrids FJ, Groop P-H, Harjutsalo V. Incidence rate patterns, cumulative incidence, and time trends for moderate and severe albuminuria in individuals diagnosed with type 1 diabetes aged 0–14 years: a population-based retrospective cohort study. *The Lancet diabetes & endocrinology* 2022;10(7):489-498
3. Fineberg D, Jandeleit-Dahm KA, Cooper ME. Diabetic nephropathy: diagnosis and treatment. *Nature Reviews Endocrinology* 2013;9(12):713-723
4. Jansson FJ, Forsblom C, Harjutsalo V, *et al.* Regression of albuminuria and its association with incident cardiovascular outcomes and mortality in type 1 diabetes: the FinnDiane Study. *Diabetologia* 2018;61(5):1203-1211
5. Harjutsalo V, Thomas MC, Forsblom C, *et al.* Risk of coronary artery disease and stroke according to sex and presence of diabetic nephropathy in type 1 diabetes. *Diabetes, Obesity and Metabolism* 2018;20(12):2759-2767
6. Sandholm N, Van Zuydam N, Ahlqvist E, *et al.* The genetic landscape of renal complications in type 1 diabetes. *2017*;28(2):557-574
7. Dahlström E, Sandholm N. Progress in defining the genetic basis of diabetic complications. *Current diabetes reports* 2017;17(9):80
8. Mahajan A, Spracklen CN, Zhang W, *et al.* Multi-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation. *Nature genetics* 2022;54(5):560-572
9. Germain M, Pezzolesi MG, Sandholm N, *et al.* SORBS1 gene, a new candidate for diabetic nephropathy: results from a multi-stage genome-wide association study in patients with type 1 diabetes. *Diabetologia* 2015;58:543-548
10. Salem RM, Todd JN, Sandholm N, *et al.* Genome-wide association study of diabetic kidney disease highlights biology involved in glomerular basement membrane collagen. *Journal of the American Society of Nephrology* 2019;30(10):2000-2016
11. Sandholm N, Dahlström EH, Groop P-H. Genetic and epigenetic background of diabetic kidney disease. *Frontiers in Endocrinology* 2023;14:1163001
12. Khattab A, Torkamani A. Nidogen-1 could play a role in diabetic kidney disease development in type 2 diabetes: a genome-wide association meta-analysis. *Human Genomics* 2022;16(1):47
13. Iyengar SK, Sedor JR, Freedman BI, *et al.* Genome-wide association and trans-ethnic meta-analysis for advanced diabetic kidney disease: Family Investigation of Nephropathy and Diabetes (FIND). *PLoS genetics* 2015;11(8):e1005352
14. Haukka J, Sandholm N, Valo E, *et al.* Novel linkage peaks discovered for diabetic nephropathy in individuals with type 1 diabetes. *Diabetes* 2021;70(4):986-995
15. Locke AE, Steinberg KM, Chiang CW, *et al.* Exome sequencing of Finnish isolates enhances rare-variant association power. *Nature* 2019;572(7769):323-328
16. Guo J, Rackham OJ, Sandholm N, *et al.* Whole-genome sequencing of Finnish type 1 diabetic siblings discordant for kidney disease reveals DNA variants associated with diabetic nephropathy. *Journal of the American Society of Nephrology: JASN* 2020;31(2):309
17. Lin BM, Grinde KE, Brody JA, *et al.* Whole genome sequence analyses of eGFR in 23,732 people representing multiple ancestries in the NHLBI trans-omics for precision medicine (TOPMed) consortium. *EBioMedicine* 2021;63:103157
18. Pan Y, Sun X, Mi X, *et al.* Whole-exome sequencing study identifies four novel gene loci associated with diabetic kidney disease. *Human molecular genetics* 2023;32(6):1048-1060
19. Thorn LM, Forsblom C, Fagerudd J, *et al.* Metabolic syndrome in type 1 diabetes: association with diabetic nephropathy and glycemic control (the FinnDiane study). *Diabetes care* 2005;28(8):2019-2024
20. Cingolani P, Platts A, Wang LL, *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *fly* 2012;6(2):80-92
21. Poplin R, Ruano-Rubio V, DePristo MA, *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv* 2017:201178

22. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 2010;26(17):2190-2191
23. Zhan X, Hu Y, Li B, *et al.* RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. *Bioinformatics* 2016;32(9):1423-1426
24. Lee S, Teslovich TM, Boehnke M, Lin X. General framework for meta-analysis of rare variants in sequencing association studies. *The American Journal of Human Genetics* 2013;93(1):42-53
25. Connaughton DM, Kennedy C, Shril S, *et al.* Monogenic causes of chronic kidney disease in adults. *Kidney international* 2019;95(4):914-928
26. Li X, Li Z, Zhou H, *et al.* Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nature genetics* 2020;52(9):969-983
27. Wang Y, Song F, Zhang B, *et al.* The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. *Genome biology* 2018;19(1):1-12
28. Sandholm N, Hotakainen R, Haukka JK, *et al.* Whole-exome sequencing identifies novel protein-altering variants associated with serum apolipoprotein and lipid concentrations. *Genome medicine* 2022;14(1):1-18
29. Liu H, Doke T, Guo D, *et al.* Epigenomic and transcriptomic analyses define core cell types, genes and targetable mechanisms for kidney disease. *Nature Genetics* 2022;54(7):950-962
30. Thomas-Chollier M, Hufton A, Heinig M, *et al.* Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs. *Nature protocols* 2011;6(12):1860-1869
31. Eadon MT, Lampe S, Baig MM, *et al.* Clinical, histopathologic and molecular features of idiopathic and diabetic nodular mesangial sclerosis in humans. *Nephrology Dialysis Transplantation* 2022;37(1):72-84
32. Cheng D, Nelson TC, Chen J, *et al.* Identification of acyl coenzyme A: monoacylglycerol acyltransferase 3, an intestinal specific enzyme implicated in dietary fat absorption. *Journal of Biological Chemistry* 2003;278(16):13611-13614
33. Niewczas MA, Gohda T, Skupien J, *et al.* Circulating TNF receptors 1 and 2 predict ESRD in type 2 diabetes. *Journal of the American Society of Nephrology: JASN* 2012;23(3):507
34. Niewczas MA, Pavkov ME, Skupien J, *et al.* A signature of circulating inflammatory proteins and development of end-stage renal disease in diabetes. *Nature medicine* 2019;25(5):805-813
35. Ek WE, Karlsson T, Höglund J, *et al.* Causal effects of inflammatory protein biomarkers on inflammatory diseases. *Science Advances* 2021;7(50):eabl4359
36. Lindholm E, Bakhtadze E, Cilio C, *et al.* Association between LTA, TNF and AGER polymorphisms and late diabetic complications. *PLoS one* 2008;3(6):e2546
37. Morito N, Usui T, Ishibashi S, Yamagata K. Podocyte-specific Transcription Factors: Could MafB become a Therapeutic Target for Kidney Disease? *Internal Medicine* 2023;62(1):11-19
38. Chen H, Gu L, Orellana EA, *et al.* METTL4 is an snRNA m6Am methyltransferase that regulates RNA splicing. *Cell research* 2020;30(6):544-547
39. Jiang L, Liu X, Hu X, *et al.* METTL3-mediated m6A modification of TIMP2 mRNA promotes podocyte injury in diabetic nephropathy. *Molecular Therapy* 2022;30(4):1721-1740
40. Brings S, Fleming T, Herzig S, *et al.* Urinary cathepsin L is predictive of changes in albuminuria and correlates with glucosepane in patients with type 2 diabetes in a closed-cohort study. *Journal of Diabetes and its Complications* 2020;34(9):107648
41. Canpolat N, Liu D, Atayar E, *et al.* A splice site mutation in the TSEN2 causes a new syndrome with craniofacial and central nervous system malformations, and atypical hemolytic uremic syndrome. *Clinical genetics* 2022;101(3):346-358

Figure 1 – Study setup – The figure illustrates the study setup for single variant and gene- and intergenic region aggregate analyses.

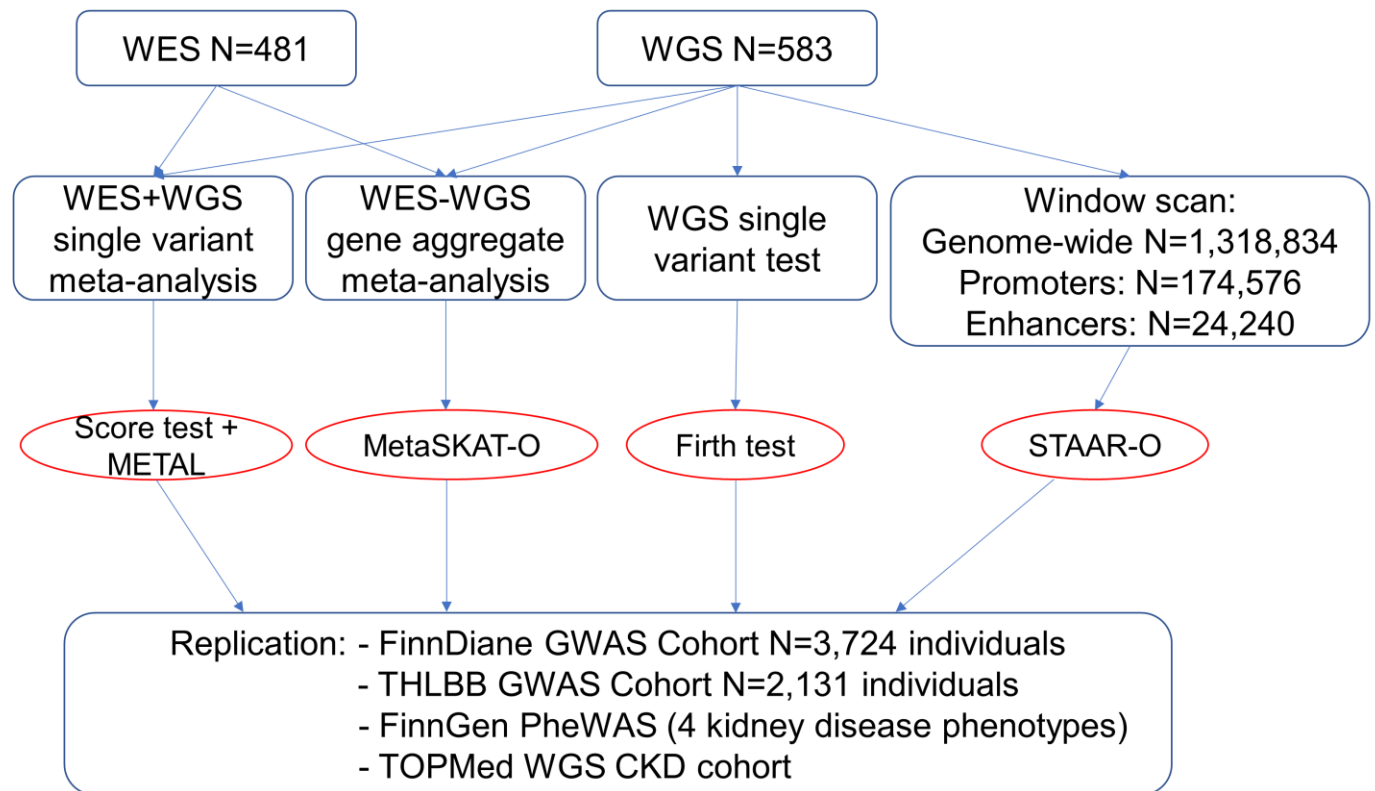


Figure 2 - Manhattan plot for DKD WES-WGS single-variant meta-analysis.

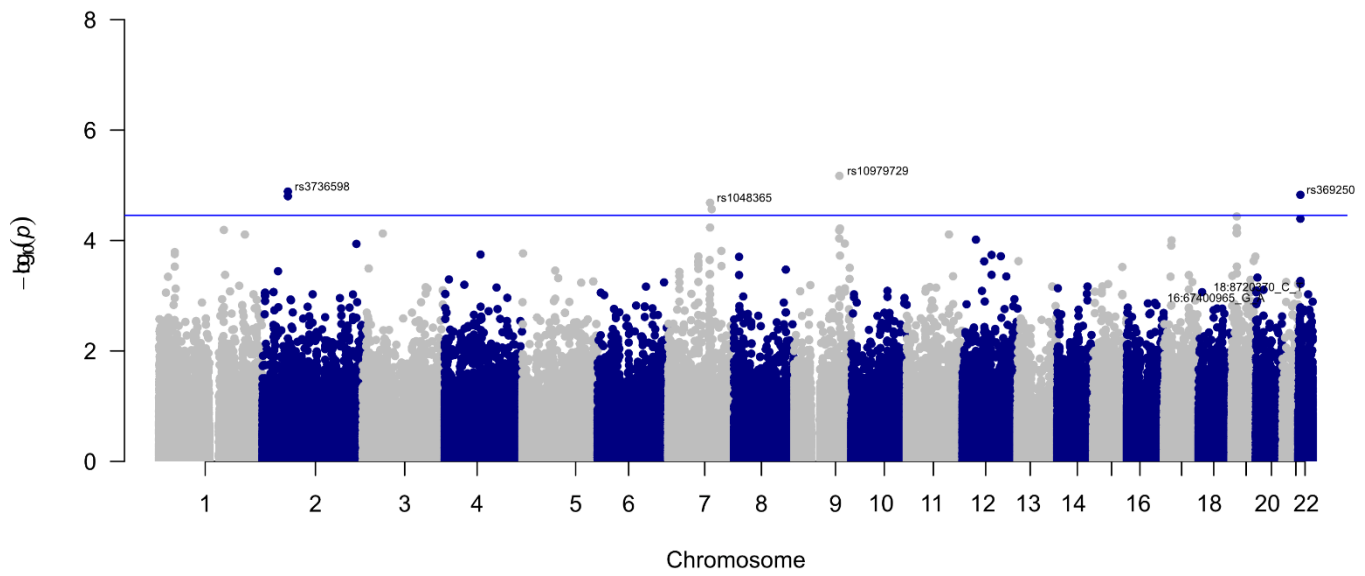


Table 1: Clinical characteristics for WES and WGS cohorts

	WGS				WES			
	Cases		Controls		Cases		Controls	
n	291		292		257		236	
Sex (% Women)	28.9		60.6		47.5		61.8	
Age at T1D onset	11.8	[6.8, 17.0]	12.3	[7.3, 16.9]	12.6	[6.8, 17.0]	12.1	[7.7, 18.0]
eGFR	10	[10, 10]	94	[83, 102]	10	[10, 50]	89	[77, 99]
Duration of T1D (years)	29	[25, 35]	40	[38, 45]	24	[20, 33]	43	[40, 48]
HbA1c (%)	9.0	[8.0, 9.9]	8.0	[7.5, 8.4]	8.6	[7.7, 9.9]	8.3	[7.7, 8.9]
HbA1c mmol/mol	75	[64,85]	64	[58,68]	70	[61,85]	67	[61,74]
BMI	23.8	[21.2, 27.8]	24.5	[22.8, 26.5]	25.1	[21.9, 27.1]	25.1	[22.9, 27.9]
SBP	148	[134, 165]	135	[123, 145]	141	[129, 154]	138	[128, 149]
DBP	84	[76, 92]	77	[70, 82]	85	[79, 92]	76	[71, 82]

Values are given as median and [inter-quartile range].

Table 2: Suggestive associations ($p < 3.5 \times 10^{-5}$) in the WES-WGS single variant-based meta-analysis for DKD

Position	Variant	Gene	Annotation	MAF	OR	P-value
9:109176646	rs10979729	<i>EPB41L4B</i>	C>T (Pro846Pro)	0.07	0.49	6.76×10^{-6}
2:61825245	rs3736598	<i>FAM161A</i>	non-coding	0.40	0.67	1.30×10^{-5}
22:21002277	rs369250	<i>THAP7</i>	promoter-TSS	0.51	0.68	1.50×10^{-5}
2:61826155	rs6748320	<i>FAM161A</i>	non-coding	0.40	0.67	1.58×10^{-5}
7:101161149	rs1048365	<i>AP1S1</i>	3' UTR	0.12	0.58	2.07×10^{-5}
7:105107452	rs117986340	<i>KMT2E</i>	G>T (Gly999Cys)	0.14	1.81	2.71×10^{-5}

Table 3 - Genes with suggestive $P < 4 \times 10^{-4}$ in SKAT-O gene aggregate test meta-analysis.

Gene	P-value	MAF threshold	Variants tested	Variant type	Previous findings
<i>NAT16</i>	1.4×10^{-4}	0.1	6	PAV+PTV	Significant difference in gene expression between ING patients and controls [31]
<i>LTA</i>	1.5×10^{-4}	0.05	1	PAV+PTV	DKD candidate gene [36]
<i>SLC10A6</i>	2.7×10^{-4}	0.1	3	PAV+PTV	-
<i>TSEN54</i>	3.7×10^{-4}	0.01	14	PAV+PTV	<i>TSEN2</i> splice site mutation associated with atypical hemolytic uremic syndrome [41]

* *LTA* is localized on HLA-region for which a different variant calling pipeline was used (Please see the Methods for further details).

Table 4 - STAAR-O aggregate analysis results for genome-wide sliding windows (n=1,318,834 regions), promoters (n=184,609) and enhancers (n=24,240)

WINDOW	VARIANT COUNT	CUMULATIVE MAC	P-VALUE	MAF CLASS	CLOSEST GENE
SLIDING WINDOWS					
chr12:108104001-108108000 (12q14.3)	14	22	4.66×10 ⁻⁶	0.01	<i>WSCD2</i>
chr4:75122001-75126000 (4q13.3)	21	61	9.46×10 ⁻⁶	0.01	-
PROMOTERS					
chr9:87792126-87793126 (9q21.11)	6	79	2.67×10 ⁻⁵	0.05	<i>CTSL3P (Pseudogene)</i>
chr2:40755215-40756215 (2q14.2)	5	14	3.17×10 ⁻⁵	0.05, 0.01	<i>LINC01794 (lncRNA)</i>
ENHANCERS					
chr18:2350865-2351244 (18q12.3)	2	24	3.94×10 ⁻⁵	0.05	<i>METTL4; MYO11</i>
chr9:87792785-87793201 (9q21.11)	4	27	1.17×10 ⁻⁴	0.05	<i>CTSL</i>

Cumulative MAC: cumulative minor allele count across all the included variants within the given minor allele frequency (MAF) class.