Genome-wide Association Scan for Diabetic Nephropathy Susceptibility Genes in Type 1 Diabetes Mellitus


1 Research Division, Joslin Diabetes Center, and Department of Medicine, Harvard Medical School, Boston, MA 02215
2 Center for Public Health Genomics, University of Virginia School of Medicine, Charlottesville, VA 22908
3 Program in Genetics and Genome Biology, Hospital for Sick Children, University of Toronto, Toronto, Canada M5G 1L7
4 Dalla Lana School of Public Health, University of Toronto, Toronto, Canada M5T 3M7
5 Kidney Disease Program, University of Louisville, Louisville, KY 40202
6 Department of Community, Occupational and Family Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 117597
7 Department of Hypertension, Medical University of Warsaw, Warsaw, Poland 02-097
8 Department of Endocrinology, Hospital de Clínicas de Porto Alegre, Porto Alegre, Brazil 90035-903
9 Samuel Lunenfeld Research Institute of Mount Sinai Hospital, Prosserman Centre for Health Research, Toronto, Canada M5T 3L9
10 Department of Pharmacology, Jagiellonian University, School of Medicine, Krakow, Poland 31-531
11 Department of Medicine, University Health Network, University of Toronto, Canada M5G 2C4

Corresponding author: Andrzej S. Krolewski MD, PhD
Section on Genetics and Epidemiology, Joslin Diabetes Center, One Joslin Place
Boston, MA 02215
e-mail: Andrzej.Krolewski@joslin.harvard.edu

Submitted 31 October 2008 and accepted 23 February 2009

Additional information for this article can be found in an online appendix at http://diabetes.diabetesjournals.org

This is an uncopyedited electronic version of an article accepted for publication in Diabetes. The American Diabetes Association, publisher of Diabetes, is not responsible for any errors or omissions in this version of the manuscript or any version derived from it by third parties. The definitive publisher-authenticated version will be available in a future issue of Diabetes in print and online at http://diabetes.diabetesjournals.org.
Genome-wide Association Scan in Diabetic Nephropathy

ABSTRACT

Objective: Despite extensive evidence for genetic susceptibility to diabetic nephropathy (DN), the identification of susceptibility genes and their variants has had limited success. To search for genes that contribute to DN, a genome-wide association scan was implemented on the Genetics of Kidneys in Diabetes collection.

Research Design and Methods: We genotyped ~360,000 single nucleotide polymorphisms (SNPs) in 820 cases (284 with proteinuria and 536 with ESRD) and 885 controls with type 1 diabetes (T1D). Confirmation of implicated SNPs was sought in 1,304 participants of the DCCT/EDIC study, a long-term, prospective investigation of the development of diabetes-associated complications.

Results: A total of 13 SNPs located in four genomic loci were associated with DN with \( P < 1 \times 10^{-5} \). The strongest association was at the \( FRMD3 \) (FERM domain containing 3) locus (OR=1.45, \( P = 5.0 \times 10^{-7} \)). A strong association was also identified at the \( CARS \) (cysteinyl-tRNA synthetase) locus (OR=1.36, \( P = 3.1 \times 10^{-6} \)). Associations between both loci and time to onset of DN were supported in the DCCT/EDIC study (HR=1.33, \( P = 0.02 \) and HR=1.32, \( P = 0.01 \), respectively). We demonstrated expression of both \( FRMD3 \) and \( CARS \) genes in human kidney.

Conclusions: We identified genetic associations for susceptibility to DN at two novel candidate loci near the \( FRMD3 \) and \( CARS \) genes. Their identification implicates previously unsuspected pathways in the pathogenesis of this important late complication of T1D.
Diabetic nephropathy (DN) is the leading contributor to end-stage renal disease (ESRD) in the United States (1). Clinically, DN is manifest as a progressive disease process that advances through characteristic stages. It begins with microalbuminuria (leakage of small amounts of albumin into the urine) and progresses to overt proteinuria. In a large proportion of these patients, renal function declines and continues to deteriorate until ESRD is reached and replacement therapy is indicated (2-4). Overall, ESRD develops in approximately 20% of all patients with type 1 diabetes (T1D) (5,6).

Despite evidence that genetic susceptibility plays a role in the development of DN in T1D (7-9), success in identifying the responsible genetic variants has been limited (10, 11). This has been due, in part, to the small size of the DNA collections available to individual research groups and the narrow focus of the searches on candidate genes. Another challenge that has received little attention in previous studies is the possibility that successive stages of DN are influenced by different genetic factors (12,13).

In order to conduct a statistically robust study that provides genome-wide coverage for detection of common variants that may have relatively small, but pathogenically significant, effect on risk of DN in T1D, the Genetics of Kidneys in Diabetes (GoKinD) collection was established (14). A genome-wide scan of this collection was supported by the Genetic Association Information Network (GAIN) initiative (15). This report presents: 1) results of this genome-wide association (GWA) scan in the GoKinD collection; 2) replication of the significant associations in this scan with time to onset of diabetes-associated complications (severe nephropathy) in the Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications (DCCT/EDIC) study; and 3) characterization of expression of the identified candidate DN genes in normal human cell lines.

**RESEARCH DESIGN AND METHODS**

**Study Populations.** Subjects for the GoKinD collection were recruited through two centers with different methods of ascertainment and recruitment (14). The George Washington University (GWU) Biostatistics Center coordinated the recruitment of volunteers (through mass media advertisement) living throughout the USA (excluding New England) and Canada to one of 27 clinical centers located across the USA and Canada. The Section of Genetics and Epidemiology at the Joslin Diabetes Center (JDC) recruited and examined patients of the Joslin Clinic from New England who were already enrolled in the Joslin Kidney Study on the Genetics of Diabetic Nephropathy, a clinic-based cohort study in which cases with DN and a random sampling of eligible controls were identified and recruited (16).

A detailed description of the GoKinD collection has been published (14). Briefly, subjects enrolled in GoKinD had T1D diagnosed before age 31, began insulin treatment within one year of their diagnosis, and were between 18 and 59 years of age at the time of enrollment. Participation in the DCCT/EDIC study was an exclusion criterion so that the two study populations would be independent. Cases with DN had either persistent proteinuria, defined by a urinary albumin to creatinine ratio (ACR) $\geq$300 $\mu$g/mg in two of the last three measurements taken at least one month apart, or ESRD (dialysis or renal transplant). Controls had T1D for at least 15 years and normoalbuminuria, defined by an ACR $<20\mu$g/mg in two of the last three measurements taken at least one month apart (if a third measurement was required, a value...
<40μg/mg was necessary for inclusion), without ever having been treated with angiotensin-converting enzyme inhibitors or angiotensin receptor blockers, and were not being treated with antihypertensive medication at the time of recruitment into the study. For additional information regarding the definition of cases and controls used in this analysis refer to the publication by Mueller et al. (14). In total, 1,879 subjects (935 cases and 944 controls) were recruited into GoKinD. The GWU panel included 437 cases with DN (58 with proteinuria and 379 with ESRD) and 446 controls; the JDC panel included 498 cases with DN (268 with proteinuria and 230 with ESRD) and 498 controls. Further details are also provided in Supplementary Information.

Confirmation of our findings in the GoKinD collection was sought in GWA data from the DCCT/EDIC study (Paterson et al., in preparation), a long-term, prospective investigation of the development of diabetes-associated complications (17,18). Of the original DCCT cohort recruited between 1983 and 1989, 1,375 subjects (95%) were retained in the EDIC follow-up study. Participants in EDIC underwent baseline examinations between 1994 and 1995 and have since participated in annual follow-up examinations to assess the development or progression of complications. As of EDIC year 12 (2005), this cohort had 16 to 22 years of follow-up and 132 cases of severe nephropathy (proteinuria or ESRD) had been documented in 1,304 Caucasian DCCT/EDIC participants. This phenotype is the closest to the phenotype used in the GoKinD collection. Detailed clinical characteristics of the DCCT/EDIC study have been published (13,17,18). Additional details are also provided in Supplementary Information.

Genotyping. The GoKinD collection was genotyped on the Affymetrix 5.0 500K SNP Array by the GAIN genotyping laboratory at the Eli and Edythe L. Broad Institute. A description of study genotyping is available in Supplementary Information. Additionally, two single nucleotide polymorphisms (SNPs), rs39075 and rs1888747, were genotyped in the GoKinD collection using Taqman (Applied Biosystems, Foster City, CA) technology by the Genetics Core of the Diabetes and Endocrinology Research Center at the JDC in accordance with the manufacturer’s protocols. DNA samples used for the genotyping of these SNPs in the GoKinD collection were obtained through the National Institute of Diabetes and Digestive and Kidney Diseases Central Repository (www.niddkrepository.org/).

SNP Quality Control. After internal quality control, the GAIN genotyping laboratory released genotypes for 467,144 SNPs. Several quality control metrics, including filters for minor allele frequency (MAF) < 0.01, rejection of Hardy-Weinberg assumptions ($P \leq 10^{-5}$) and differential rates of missing data (by case/control status) were applied to these data. After reconciliation of SNPs eliminated by these analyses, the resulting data contained 359,193 autosomal SNPs. More details are available in Supplementary Information and Supplementary Table 1.

Sample Quality Control and SNP Imputation. The application of quality control criteria reduced the GoKinD population from 1,879 to 1,705 individuals of European ancestry. Samples from the two GoKinD panels that constitute this sample are 379 GWU cases (49 with proteinuria and 330 with ESRD), 413 GWU controls, 441 JDC cases (235 with proteinuria and 206 with ESRD), and 472 JDC controls. The measured genotypes for these individuals were augmented by imputation of un-genotyped SNPs across two mega-basepair regions flanking each of the lead SNPs in the GoKinD collection (where linkage disequilibrium (LD), as measured by $r^2$, decayed to <0.20 for all lead SNPs). A total of 8,245 imputed
HapMap SNPs across these four loci were included in our association analysis. Further details of sample quality control and imputation procedures are available in Supplementary Information and Supplementary Table 2.

Statistical Analysis. GoKinD samples were collected under two separate ascertainment protocols (JDC and GWU panels) so that tests for latent residual stratification were performed. Cochran-Armitage tests of trend for the JDC versus GWU controls, and JDC versus GWU cases, revealed an over-dispersion in the test statistics for both controls and cases compared to the complete null. The median genomic control parameters were estimated at $\lambda=1.13$ for controls and $\lambda=1.097$ for cases. Permutation analysis within the controls and within the cases resulted in a stratification significance of $P < 10^{-3}$ for both cases and controls. Therefore, the primary association analysis used in the study was a stratified test of association combining case-control tests of allele frequencies in JDC and GWU strata. Combined P-values and odds ratios (ORs) were calculated using a Cochran-Mantel-Haenszel procedure. Homogeneity across strata was assessed using the Breslow-Day statistic.

All genome-wide statistical association analyses were performed using PLINK and R (19). Further details of quality control procedures, software, statistical analysis and adjustments, and cluster plots are available in Supplementary Information.

Genotyping and Analysis of the DCCT/EDIC Data. Genotypes of the DCCT/EDIC study participants were generated with the Illumina Human1M Beadchip (Paterson et al., in preparation). Briefly, quality control measures resulted in 840,354 SNPs suitable for statistical analysis. Population substructure was assessed to ensure that all included samples were of European ancestry. Multivariate Cox proportional hazard analyses were performed on data from 1,304 Caucasian subjects using time to onset of severe nephropathy, defined by an albumin excretion rate (AER) >300 mg/24h on at least two consecutive examinations or dialysis/renal transplant with prior persistent microalbuminuria (two consecutive AERs >30 mg/24h) as the outcome phenotype (n=132). Among those with severe nephropathy, 116 subjects developed only proteinuria (AER >300 mg/24h), while 16 progressed to ESRD. The DCCT cohort (primary prevention versus secondary intervention), treatment (intensive versus conventional), and interaction between cohort and treatment were used as covariates in the analysis of the effect of an independent additive SNP genetic factor. This model was examined for all associated loci in GoKinD and subsequently tested for both statistical significance and the same direction of effect for associated alleles (20-22).

Gene Expression. The expression of candidate genes was examined in four primary human cell lines derived from cells that have been implicated in the pathogenesis of kidney complications (endothelial cells from the iliac artery, adult dermal fibroblasts, mesangial cells, and epithelial cells from proximal tubules) by quantitative real-time PCR. Sources of these cells, cell culture conditions, and protocols used in these experiments are available in Supplementary Information.

RESULTS

Genome-wide Association Scan for Genes Associated with DN in T1D. The application of metrics for SNP and sample quality resulted in the analysis of 359,193 autosomal SNPs and 1,705 GoKinD samples of European ancestry (885 controls and 820 cases) (see Methods and Supplementary Information). Clinical characteristics of the JDC and GWU panels are summarized in Table 1. Because different ascertainment
protocols were used by the JDC and GWU, the resulting data were found to exhibit significant stratification. As a result, the primary association analyses were conducted using a stratified test of association.

Although no SNP achieved genome-wide significance ($0.05/359,193=1.4\times10^{-7}$), the primary association analysis identified 11 SNPs representing four distinct chromosomal regions with $P<1\times10^{-5}$ (Figure 1 and Table 2), which were considered for replication. The strongest association with DN occurred on chromosome 9q with rs10868025 (OR=1.45, $P=5.0\times10^{-7}$). This SNP is located near the 5’-end of the FERM domain containing 3 (FRMD3) gene.

Three additional genomic regions located on chromosomes 7p, 11p, and 13q were also associated with DN. The rs39059 SNP (OR=1.39, $P=5.0\times10^{-6}$) localizes to the first intron of CHN2 (beta chimerin) isoform 2 and upstream of an alternatively spliced CPVL (serine carboxypeptidase vitellogenic-like) transcript on chromosome 7p. The rs451041 SNP (OR=1.36, $P=3.1\times10^{-6}$) is located on chromosome 11p in an intronic region of the CARS (cysteinyl-tRNA synthetase) gene. And finally, the region bounded by rs1411766/rs1742858 (OR=1.41, $P=1.8\times10^{-6}$) is located in a 42 kilo-basepair (kb) intergenic region on chromosome 13q.

Analyses of the imputed SNPs in our lead loci identified 11 additional SNPs that were highly correlated with the original associations ($P<1\times10^{-5}$). Of these, two were more strongly associated with DN than our lead genotyped SNPs. Imputed SNP rs1888747 (chromosome 9q), which is in partial LD ($r^2=0.81$) with rs10868025, was more strongly associated with DN than the original SNP ($P=4.7\times10^{-7}$) (Figure 2B). Similarly, two imputed SNPs in the 7p region (rs39075 and rs39076) were also more strongly associated than the original SNP in that region (rs39059) (Figure 2A). Both imputed SNPs were genotyped in the GoKinD samples and the associations with the imputed data were confirmed (rs39075, $P=6.5\times10^{-7}$ and rs1888747, $P=6.3\times10^{-7}$, Table 2).

If the etiology of DN involves the interaction of a locus with the cumulative effect of hyperglycemia, the association of the locus with DN can vary according to diabetes duration at DN onset, such that it is strongest in early-appearing cases and diminishes in later ones – even reversing in direction in very late-appearing cases (23). We examined the SNPs in Table 2 according to diabetes duration by stratifying cases and controls across tertiles of diabetes duration (at the onset of ESRD or at enrollment into GoKinD for proteinurics and controls). The strength of the associations was consistent across these strata (data not shown).

Additionally, if a locus influences mortality risk, the high mortality experienced by patients with ESRD would alter its association with DN according to the duration of survival with ESRD and may mask the effect of a DN risk allele or produce a false association. For this reason, we also analyzed the lead SNPs in Table 2 according to duration of ESRD. For each of these SNPs, the odds ratios were consistent across tertiles of ESRD duration (Supplementary Table 4), a pattern consistent with the absence of survival bias. However, the current study is underpowered to formally exclude the presence of such effects.

**Confirmation of Associated T1D DN SNPs in the DCCT/EDIC Study.** Data from a GWA scan of the DCCT/EDIC study (Paterson et al., in preparation) were used to assess whether genome regions identified in the GoKinD collection were associated with advanced DN in an independent collection. Among the 11 SNPs identified in GoKinD, 8 were included on the Illumina array used in the DCCT/EDIC study (Table 3). The 3 SNPs not included on this platform, rs39059, rs739401, and rs9521445, were in strong LD ($r^2\geq0.87$) with rs39075, rs451041, and
rs7989848, respectively. Analysis of time to onset of severe nephropathy confirmed the significant associations with DN in GoKinD for rs1888746 (FRMD3, \( P = 0.02 \)), rs13289150 (FRMD3, \( P = 0.05 \)), and rs451041 (CARS, \( P = 0.01 \)).

**Analysis of Candidate DN Gene Expression.** Previous studies, as well as publicly available gene expression data (www.ncbi.nlm.nih.gov/geo), have shown that genes closest to the lead SNPs identified in GoKinD are expressed in a variety of human tissues, including kidney (24-26). To further test whether these candidate genes may be involved in the development of DN, we examined their expression in cell lines relevant to this disease. The expression of CHN2, CPVL, FRMD3, and CARS was examined in four primary human cell lines: iliac artery endothelial cells, adult dermal fibroblasts, mesangial cells, and renal proximal tubule cells. Our data show that CARS expression was high in all four of the cell lines that we examined (Table 4). FRMD3 expression was also detected in each cell type, with its highest expression being observed in renal proximal tubule cells. Of the two candidate DN genes located in chromosome 7p region, neither was detected in mesangial cells, while CPVL expression was greatest in proximal tubule cells.

**DISCUSSION**

In this report, we describe the results of a GWA scan in the GoKinD collection to identify loci associated with risk of DN in T1D. The most significant associations were identified with variants located within four distinct chromosomal regions. Although the biology underlying these associations remains to be elucidated, they implicate CHN2/CPVL, FRMD3, CARS, and an intergenic region on chromosome 13q as novel genes/genetic regions involved in the pathogenesis of DN. None of these loci overlap with previously reported associations between candidate genes and the development of any stage of DN (10,11). Importantly, replication in a Cox proportional hazard analysis of the associations at the FRMD3 and CARS loci with time to the onset of severe nephropathy in the DCCT/EDIC study bolsters the significance of these two findings; that two studies having such different designs (one a case-control study and the other a prospective cohort study) yielded similar odds ratios strengthens confidence in this conclusion.

FRMD3 encodes the 4.1O protein, a structural protein with unknown function and a member of the 4.1 family of proteins (26). Members of the 4.1 protein family have well characterized roles as cytoskeletal proteins, maintaining both cellular shape and form, in a variety of cell types, including mouse nephron (27,28). Although membership of the 4.1O protein in this family has recently been questioned, it does contain a FERM (four one protein ezrin, radixin, moesin) domain; a module that is integral in maintaining cell integrity through its interactions with transmembrane proteins and actin filaments (29,30). FRMD3 is detectable in adult ovary, as well as in fetal skeletal muscle, brain, and thymus (26). Our data extend the expression profile of FRMD3 to specifically include mesangial and proximal tubular cells. Interestingly, among 18 genes that contain FERM domains, including several members of the 4.1 protein family, we identified nominally significant associations with DN for SNPs located in eight of these genes (Supplementary Table 5), including FARP2 (FERM, RhoGEF and pleckstrin domain protein 2, \( P = 3.0 \times 10^{-4} \)) and EPB41L2 (erythrocyte membrane protein band 4.1-like 2, \( P = 2.3 \times 10^{-4} \)). Although these findings require further study, including replication in additional collections, it is interesting to speculate that these data may point to the involvement of new, previously unsuspected, pathways in the pathogenesis of DN.
The *CARS* gene encodes cysteinyl-tRNA synthetase, one of several aminoacyl-tRNA synthetases (ARSs) that have been identified in humans (31,32). ARSs are important regulators of intracellular amino acid concentrations and protein biosynthesis in both the cytoplasm and mitochondria (a process facilitated by specialized mitochondria-specific and bifunctional ARSs). In the initial steps of protein translation, the function of these enzymes is to attach amino acids to their cognate tRNA molecules. To date, both autosomal dominant and recessive mutations in ARS encoding genes have been identified only in neurodegenerative disease, including missense changes in glycyl-tRNA synthetase (*GARS*) and both missense mutations and in-frame deletions in tyrosyl-tRNA synthetase (*YARS*) in Charcot-Marie-Tooth disease (32).

*CARS* has been implicated in cystinosis, an autosomal recessive renal tubule disorder caused by the accumulation of free cystine in cellular lysosomes (33,34). A recent study identified defects in lysosomal cystine transport as the primary cause of the disease (35). However, ESRD is prominent in this disorder, and such an outcome may be due to vulnerability of specific renal cells to damage by excess cystine. Interesting in this light, is the observation that, of all the associated SNPs, only those in the *CARS* locus were associated primarily with ESRD (Supplementary Table 4). *CARS* is expressed in mesangial and proximal tubule cells. Further work is needed to characterize the role of *CARS* in the pathway that is involved in the development of ESRD in diabetes. Similar to the set of genes containing FERM domains, analysis of 21 ARS genes identified nominally significant associations with DN for SNPs located in four members of this class of genes (Supplementary Table 6), with the most significant association ($P=9.1\times10^{-3}$) occurring at the *TARS* (threonyl-tRNA synthetase) locus.

Two additional loci were strongly associated with DN in both panels of the GoKinD collection. Of the two genes located on chromosome 7p, *CPVL*, a carboxypeptidase that is highly expressed in the kidney and, more specifically, in proximal tubules, is a particularly interesting candidate gene. Other carboxypeptidases, such as angiotensin converting enzyme and bradykinin, are important regulators of renal hemodynamics and have previously been implicated in the pathogenesis of DN (36,37). The last DN associated locus involves multiple SNPs within a 33 kb haplotype block on chromosome 13q. Previously, genomic deletions of this locus have been linked to congenital renal abnormalities (38). The two genes closest to the associated SNPs, *MYO16* (myosin heavy chain Myr 8) and *IRS2* (insulin receptor substrate 2), are located approximately 384 kb centromeric and 120 kb telomeric of this region, respectively. Although there is little LD between the variants within this block and those in the vicinity of either *MYO16* or *IRS2*, the multiple signals identified in this region give credence to the association detected in our analysis. Additional experiments are needed to characterize the nature of these associations further.

The findings presented in our study contribute to understanding the genetic susceptibility of DN in T1D. As has been reported for other complex genetic disorders, no single major gene that contributes to an increased risk of disease emerged (20,39). However, given the incomplete coverage of the genome by the genotyping platform and the sub-optimal study design (prevalent rather than incident cases of ESRD), detection of any existing major gene effect was not guaranteed. For example, because most of the cases with ESRD had survived many years on dialysis or with a kidney transplant, a disease allele that not only increased susceptibility to DN but also increased mortality in patients...
Genome-wide Association Scan in Diabetic Nephropathy

with ESRD could go undetected. Appreciably, the SNPs that we identified in the GoKinD collection were mortality neutral (Supplementary Table 4). The optimal study design for detecting all disease loci, regardless of their effect on mortality, would be a large cohort of incident ESRD cases. Such a data set is presently unavailable.

There are other limitations to this study as well. The GoKinD collection is heavily weighted with cases with ESRD; thus, the small number of cases with proteinuria limited our ability to detect variants primarily associated with the risk of proteinuria. Second, because of the limited power of the DCCT/EDIC study and the need to contain inflation of the alpha error in seeking replication for multiple SNPs in this dataset, our replication efforts refrained from considering SNPs less significant than $P=1 \times 10^{-5}$. It is certainly possible that additional variants among those not meeting this threshold may truly be associated with DN; however, given these limitations, these variants remain to be identified. Similarly, despite replication in the DCCT/EDIC cohort, we acknowledge that positive associations at both the $FRMD3$ and $CARS$ loci require additional study to be certain of these findings. Thirdly, while the locations of the variants confirmed in this study implicate both $FRMD3$ and $CARS$ as novel genes involved in the pathogenesis of DN, the underlying mechanisms of disease of these associations need to be elucidated. And lastly, although confirmation in DCCT/EDIC has been achieved for variations near $FRMD3$ and $CARS$, additional cohorts, particularly non-Caucasian, would be useful to further characterize the pathogenic role of these, and other, candidate genes identified in the GoKinD collection.

ACKNOWLEDGEMENTS

We acknowledge the following grant support from the National Institutes of Health (NIH): DK77532 (to ASK), DK36836 (to the Genetics Core of the Diabetes and Endocrinology Research Center at the JDC), DK-62204 (to ADP) and DK-077510 (to ADP); and from the Foundation for NIH (FNIH) grant 06GAIN0 (to JHW). We acknowledge the JDC's NIH Training Grant T32 DK007260-31 (to MGP) and the Canadian Institutes of Health Research (CIHR) Scholarship (to LM) and Senior Investigator Award (to SBB). We also acknowledge support from the Canadian Network of Centres of Excellence in Mathematics (to SBB). The GoKinD study was conducted by the GoKinD Investigators and supported by the Juvenile Diabetes Research Foundation (JDRF), and by funding from the Centers for Disease Control (CDC) (PL 105-33, 106-554, and 107-360 administered by the National Institute of Diabetes and Digestive and Kidney Diseases, NIDDK). The GoKinD collection of DNA was genotyped through the Genetic Association Information Network (GAIN) program, with the support of the FNIH and NIDDK. The GAIN database was accessed through National Center for Biotechnology Information (NCBI). The DCCT/EDIC study is supported by contracts with the Division of Diabetes, Endocrinology, and Metabolic Diseases of the NIDDK and the General Clinical Research Centers Program, National Center for Research Resources. The content and conclusions presented in this manuscript do not necessarily reflect the opinions or views of the JDRF, CDC, FNIH, NIDDK or NCBI.
REFERENCES
19. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC: PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559-575, 2007
Genome-wide Association Scan in Diabetic Nephropathy

30. Baines AJ: A FERM-adjacent (FA) region defines a subset of the 4.1 superfamily and is a potential regulator of FERM domain function. BMC Genomics 7:85, 2006
<table>
<thead>
<tr>
<th>Clinical Characteristic</th>
<th>JDC GoKinD Controls</th>
<th>JDC GoKinD Cases</th>
<th>GWU GoKinD Controls</th>
<th>GWU GoKinD Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>472</td>
<td>441</td>
<td>413</td>
<td>379</td>
</tr>
<tr>
<td>Men/Women</td>
<td>220/252</td>
<td>233/208</td>
<td>143/270</td>
<td>190/189</td>
</tr>
<tr>
<td>Age at DM Diagnosis (years)</td>
<td>12.9±7.5</td>
<td>11.5±6.7</td>
<td>12.9±7.2</td>
<td>12.1±6.6</td>
</tr>
<tr>
<td>Duration of T1D (years)†</td>
<td>26.4±8.0</td>
<td>27.0±7.8</td>
<td>24.2±7.3</td>
<td>25.9±7.6</td>
</tr>
<tr>
<td>Age at Examination (years)</td>
<td>39.4±6.5</td>
<td>42.0±7.1</td>
<td>37.1±8.8</td>
<td>44.4±6.5</td>
</tr>
<tr>
<td>Laser Treatment (%)</td>
<td>22%</td>
<td>84%</td>
<td>10%</td>
<td>85%</td>
</tr>
<tr>
<td>HbA(_1c) (%)‡</td>
<td>7.5±1.7</td>
<td>8.5±1.6</td>
<td>7.4±1.1</td>
<td>7.8±1.4</td>
</tr>
<tr>
<td>ACR in Controls (µg/mg)</td>
<td>7.0±3.7</td>
<td>---</td>
<td>6.3±3.6</td>
<td>---</td>
</tr>
<tr>
<td>Cases with Proteinuria/ESRD</td>
<td>---</td>
<td>235/206</td>
<td>---</td>
<td>49/330</td>
</tr>
<tr>
<td>ACR in Proteinurics (µg/mg)</td>
<td>---</td>
<td>1571±1534</td>
<td>---</td>
<td>1282±1096</td>
</tr>
<tr>
<td>ESRD Duration (years)</td>
<td>---</td>
<td>7.5±5.4</td>
<td>---</td>
<td>7.2±5.6</td>
</tr>
<tr>
<td>Kidney Transplant (%)§</td>
<td>---</td>
<td>86%</td>
<td>---</td>
<td>95%</td>
</tr>
</tbody>
</table>

JDC –Joslin Diabetes Center; GWU – George Washington University.

* Clinical characteristics are presented as mean values ± standard deviation for all Caucasians patients (n=1,705) included in the present analysis following the application of quality control metrics (see supplemental information regarding Sample Quality Control Analysis and Population Substructure and Ancestry Analysis).

† The duration of T1D in controls and proteinurics is based on the duration at examination. Among ESRD cases, this is based on the duration of T1D at the onset of ESRD. All other clinical characteristics are based on measurements performed at examination.

‡ HbA\(_1c\), glycosylated hemoglobin. Mean HbA\(_1c\) values do not include data from cases that have undergone pancreas transplantation (11% of JDC cases and 58% of GWU cases).

§ Percentages are of ESRD group.
### Table 2. Summary of SNPs Associated with DN in the GoKinD Collection*

<table>
<thead>
<tr>
<th>SNP</th>
<th>Chr.</th>
<th>Position (Mb)</th>
<th>Nearest Gene(s)</th>
<th>Risk Allele (Non-Risk Allele)</th>
<th>GWU GoKinD Controls (n=413)</th>
<th>Cases (n=379)</th>
<th>Controls (n=472)</th>
<th>Cases (n=441)</th>
<th>P-value</th>
<th>OR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs39059</td>
<td>7p</td>
<td>29.2</td>
<td>CPVL/CHN2</td>
<td>A(G)</td>
<td>0.61</td>
<td>0.69</td>
<td>0.60</td>
<td>0.67</td>
<td>8.8x10^{-4}</td>
<td>1.7x10^{-3}</td>
</tr>
<tr>
<td>rs39075†</td>
<td>7p</td>
<td>29.2</td>
<td>CPVL/CHN2</td>
<td>G(A)</td>
<td>0.57</td>
<td>0.66</td>
<td>0.57</td>
<td>0.64</td>
<td>2.0x10^{-4}</td>
<td>8.2x10^{-4}</td>
</tr>
<tr>
<td>rs1888747†</td>
<td>9q</td>
<td>85.3</td>
<td>FRMD3</td>
<td>G(C)</td>
<td>0.68</td>
<td>0.73</td>
<td>0.66</td>
<td>0.74</td>
<td>3.6x10^{-3}</td>
<td>4.4x10^{-5}</td>
</tr>
<tr>
<td>rs10868025</td>
<td>9q</td>
<td>85.4</td>
<td>FRMD3</td>
<td>A(G)</td>
<td>0.59</td>
<td>0.66</td>
<td>0.56</td>
<td>0.66</td>
<td>1.9x10^{-3}</td>
<td>7.2x10^{-5}</td>
</tr>
<tr>
<td>rs739401</td>
<td>11p</td>
<td>3.0</td>
<td>CARS</td>
<td>C(T)</td>
<td>0.46</td>
<td>0.54</td>
<td>0.49</td>
<td>0.55</td>
<td>4.7x10^{-4}</td>
<td>3.6x10^{-3}</td>
</tr>
<tr>
<td>rs451041</td>
<td>11p</td>
<td>3.0</td>
<td>CARS</td>
<td>A(G)</td>
<td>0.46</td>
<td>0.54</td>
<td>0.48</td>
<td>0.56</td>
<td>6.9x10^{-4}</td>
<td>1.3x10^{-3}</td>
</tr>
<tr>
<td>rs1041466</td>
<td>13q</td>
<td>109.0</td>
<td>no gene</td>
<td>G(A)</td>
<td>0.39</td>
<td>0.47</td>
<td>0.43</td>
<td>0.51</td>
<td>3.6x10^{-3}</td>
<td>2.7x10^{-4}</td>
</tr>
<tr>
<td>rs1411766/</td>
<td>13q</td>
<td>109.1</td>
<td>no gene</td>
<td>A(G) G(A)</td>
<td>0.31</td>
<td>0.39</td>
<td>0.32</td>
<td>0.40</td>
<td>8.5x10^{-4}</td>
<td>6.4x10^{-4}</td>
</tr>
<tr>
<td>rs6492208/</td>
<td>13q</td>
<td>109.1</td>
<td>no gene</td>
<td>T(C) G(A)</td>
<td>0.55</td>
<td>0.62</td>
<td>0.56</td>
<td>0.65</td>
<td>8.7x10^{-3}</td>
<td>1.9x10^{-4}</td>
</tr>
<tr>
<td>rs7989848</td>
<td>13q</td>
<td>109.1</td>
<td>no gene</td>
<td>A(G)</td>
<td>0.49</td>
<td>0.56</td>
<td>0.50</td>
<td>0.57</td>
<td>2.0x10^{-3}</td>
<td>1.1x10^{-3}</td>
</tr>
<tr>
<td>rs9521445</td>
<td>13q</td>
<td>109.1</td>
<td>no gene</td>
<td>A(C)</td>
<td>0.47</td>
<td>0.54</td>
<td>0.47</td>
<td>0.55</td>
<td>2.1x10^{-3}</td>
<td>4.2x10^{-4}</td>
</tr>
</tbody>
</table>

* The most strongly associated SNPs from the combined analysis of the GWU and JDC GoKinD panels are presented along with the risk allele frequencies and P-values (calculated using the Cochran-Mantel-Haenszel method, adjusting for gender, between cases and controls within each collection) for each separate collection. Combined P-values and odds ratios (ORs) were calculated using the
Cochran-Mantel-Haenszel method. Chromosomal locations, SNP positions, and gene annotations are in reference to NCBI Build 36.1. A summary of the genotype frequencies for the most strongly associated SNPs in the GoKinD collection are presented in Supplementary Table 3.
† rs39075 and rs1888747 were identified through imputation and genotyped using Taqman assays in the GoKinD collection.
‡ rs1411766 and rs17412858 were both genotyped on the Affymetrix array and are in complete LD ($r^2 = 1.0$).
§ rs6492208 and rs2391777 were both genotyped on the Affymetrix array and are in complete LD ($r^2 = 1.0$).
Table 3. Hazard Ratios for the Development of Severe Nephropathy During 16 to 22 Years of Follow-up in the DCCT/EDIC Study for SNPs Associated with DN in GoKinD*

<table>
<thead>
<tr>
<th>SNP</th>
<th>Chr.</th>
<th>Position (Mb)</th>
<th>Nearest Gene(s)</th>
<th>Risk Allele†</th>
<th>Frequency of Risk Allele</th>
<th>P-value (one-sided) ‡</th>
<th>HR</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs39075</td>
<td>7p</td>
<td>29.2</td>
<td>CPVL/CHN2</td>
<td>G</td>
<td>0.60</td>
<td>n.s.</td>
<td>0.85</td>
</tr>
<tr>
<td>rs1888746§</td>
<td>9q</td>
<td>85.3</td>
<td>FRMD3</td>
<td>C</td>
<td>0.70</td>
<td><strong>0.02</strong></td>
<td>1.33</td>
</tr>
<tr>
<td>rs13289150</td>
<td></td>
<td></td>
<td>85.4</td>
<td>FRMD3</td>
<td>A</td>
<td>0.62</td>
<td><strong>0.05</strong></td>
</tr>
<tr>
<td>rs451041</td>
<td>11p</td>
<td>3.0</td>
<td>CARS</td>
<td>A</td>
<td>0.51</td>
<td><strong>0.01</strong></td>
<td>1.32</td>
</tr>
<tr>
<td>rs1041466</td>
<td>13q</td>
<td>109.0</td>
<td>no gene</td>
<td>G</td>
<td>0.47</td>
<td>0.11</td>
<td>1.22</td>
</tr>
<tr>
<td>rs1411766</td>
<td>13q</td>
<td>109.1</td>
<td>no gene</td>
<td>A</td>
<td>0.36</td>
<td>0.11</td>
<td>1.17</td>
</tr>
<tr>
<td>rs6492208</td>
<td>13q</td>
<td>109.1</td>
<td>no gene</td>
<td>T</td>
<td>0.61</td>
<td>n.s.</td>
<td>0.90</td>
</tr>
<tr>
<td>rs7989848</td>
<td>13q</td>
<td>109.1</td>
<td>no gene</td>
<td>A</td>
<td>0.53</td>
<td>n.s.</td>
<td>0.93</td>
</tr>
</tbody>
</table>

* Data are from multivariate Cox proportional hazard analysis of time to onset of severe nephropathy. As of 2005, the number of severe nephropathy cases = 132 (versus 1172 censored). Chromosomal locations, SNP positions, and gene annotations are in reference to NCBI Build 36.1.
† The risk alleles that are presented are in reference to those identified in the GoKinD collection.
‡ One-sided P-values (consistent with the current ‘best practices’ for replication in GWA scans) (20-22) are used to test for the same direction of effect as in the GoKinD collection. n.s. = not significant.
§ rs1888746 was genotyped on the Illumina array in DCCT/EDIC and is in complete LD ($r^2 = 1.0$) with rs1888747 (genotyped using a Taqman assay in GoKinD).
|| rs13289150 was genotyped on the Illumina array in DCCT/EDIC and is in complete LD ($r^2 = 1.0$) with rs10868025 (genotyped on the Affymetrix array in GoKinD).
Table 4. Relative Gene Expression of Novel Candidate DN Susceptibility Genes in Primary Human Cell Lines*

<table>
<thead>
<tr>
<th>Cell Line</th>
<th>CPVL Ct values</th>
<th>CHN2 Ct values</th>
<th>FRMD3 Ct values</th>
<th>CARS Ct values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Endothelial cells, iliac</td>
<td>17.3±0.7</td>
<td>12.5±0.1</td>
<td>12.5±0.3</td>
<td>8.4±0.4</td>
</tr>
<tr>
<td>Fibroblasts</td>
<td>15.7±0.1</td>
<td>n.e.</td>
<td>11.9±0.7</td>
<td>6.5±0.4</td>
</tr>
<tr>
<td>Mesangial cells</td>
<td>n.e.</td>
<td>n.e.</td>
<td>12.9±0.5</td>
<td>7.2±0.1</td>
</tr>
<tr>
<td>Proximal tubules</td>
<td>8.6±0.1</td>
<td>13.1±0.1</td>
<td>10.2±0.2</td>
<td>8.0±0.1</td>
</tr>
</tbody>
</table>

*The relative gene expression of CPVL, CHN2, FRMD3, and CARS was determined in four primary human cell lines using real-time PCR. Relative gene expression of each gene was calculated in reference to a normalization control (β-actin) and is presented as the mean delta Ct (Ct value from each gene minus Ct value from the normalization control) and standard deviation from three separate experiments. n.e. = not expressed (mean delta Ct > 40.0). Mean delta Ct = 0 equals high relative gene expression (i.e. expression similar to β-actin).
FIGURE LEGENDS

Figure 1. Summary of GWA scan results in the GoKinD collection. The \(-\log_{10} P\)-values calculated using the Cochran-Mantel-Haenszel method (adjusting for gender and GoKinD sub-collection (JDC/GWU)) across the entire genome are shown for the combined GoKinD collection. The horizontal dashed line corresponds to a \(-\log_{10} P\)-value = 5.0 (\(P\)-value = 1x10\(^{-5}\)). SNPs shown in green (n=11) exceed this threshold (due to the resolution of this image, some of the SNPs located on chromosome 13 (n=7) appear indistinguishable).
Figure 2. Summary of GWA results for the chromosome 7p, 9q, 11p, and 13q loci.

A) GWA scan and imputed data for the chromosome 7p locus. Solid triangles represent SNPs genotyped on the Affymetrix array (n = 163). Unshaded triangles represent imputed SNPs (n = 694). rs39059 (solid red triangle) is located at position IVS1+21350 relative to exon 1 of CHN2 isoform 2 and is in tight LD with rs39075 ($r^2 = 0.96$), located at position IVS1+42572. rs39059 and rs39075 reside -69318 kb and -90540 kb, respectively, upstream of CPVL isoforms 1 and 2. A third alternate transcript (isoform 3) is predicted for CPVL and contains an exon that extends to intron 1 of CHN2. rs39059 and rs39075 are located at positions -20579 and -41801, respectively, relative to this transcript. *Imputed SNP rs39075 was genotyped in the GoKinD samples to confirm the imputation.

B) GWA scan and imputed data for the chromosome 9q locus. One hundred genotyped SNPs from the Affymetrix array data and 450 imputed SNPs are shown. rs10868025 (solid red triangle) is located at position -10829 relative to FRMD3’s transcription start site. rs10868025 is in complete LD ($r^2 = 1.0$) and only 253bp from imputed SNP rs13289150 (the unshaded blue triangle superimposed on rs10868025). rs1888747, located at position -2204, is in partial LD ($r^2 = 0.81$) with rs10868025. *Imputed SNP rs1888747 was genotyped in the GoKinD samples to confirm the imputation.

C) GWA scan and imputed data for the chromosome 11p locus. Thirty-three genotyped SNPs from the Affymetrix array data and 190 imputed SNPs are shown. rs739401 and rs451041 (solid red triangles) are in strong LD ($r^2 = 0.97$). rs739401 is located in intron 16 (isoforms a and c)/17 (isoforms b and d) of the CARS gene (position IVS16+687/IVS17+687). rs451041 is located in intron 4 (isoforms a and c)/5 (isoforms b and d), position IVS4-203/IVS5-203).

D) GWA scan and imputed data for the chromosome 13q locus. Sixty-eight genotyped SNPs from the Affymetrix array data and 268 imputed SNPs are shown. Seven lead SNPs (rs1041466, rs1411766, rs17412858, rs6492208, rs2391777, rs7989848, and rs9521445) from this region are indicated in red. rs1411766 and rs17412858 are in complete LD ($r^2 = 1.0$). Similarly, rs6492208 and rs2391777 are in complete LD ($r^2 = 1.0$). rs7989848, and rs9521445 are in strong LD ($r^2 = 0.87$), while only modest LD exists between all other SNP pairs ($r^2 = 0.30-0.65$). The two nearest genes are MYO16 and IRS2, located approximately 384 kb centromeric and 120 kb telomeric of this region, respectively.
Figure 2.