Contributions of Rare Gene Variants to Familial and Sporadic FSGS

Minxian Wang,1,2,3 Justin Chun,1,2,4 Giulio Genovese,5 Andrea U. Knob,1 Ava Benjamin,1 Maris S. Wilkins,1 David J. Friedman,1,2 Gerald B. Appel,6 Richard P. Lifton,7 Shrikant Mane,8 and Martin R. Pollak1,2,3

Due to the number of contributing authors, the affiliations are listed at the end of this article.

ABSTRACT

Background Over the past two decades, the importance of genetic factors in the development of FSGS has become increasingly clear. However, despite many known monogenic causes of FSGS, single gene defects explain only 30% of cases.

Methods To investigate mutations underlying FSGS, we sequenced 662 whole exomes from individuals with sporadic or familial FSGS. After quality control, we analyzed the exome data from 363 unrelated family units with sporadic or familial FSGS and compared this to data from 363 ancestry-matched controls. We used rare variant burden tests to evaluate known disease-associated genes and potential new genes.

Results We validated several FSGS-associated genes that show a marked enrichment of deleterious rare variants among the cases. However, for some genes previously reported as FSGS related, we identified rare variants at similar or higher frequencies in controls. After excluding such genes, 122 of 363 cases (33.6%) had rare variants in known disease-associated genes, but 30 of 363 controls (8.3%) also harbored rare variants that would be classified as “causal” if detected in cases; applying American College of Medical Genetics filtering guidelines (to reduce the rate of false-positive claims that a variant is disease related) yielded rates of 24.2% in cases and 5.5% in controls. Highly ranked new genes include SCAF1, SETD2, and LY9. Network analysis showed that top-ranked new genes were located closer than a random set of genes to known FSGS genes.

Conclusions Although our analysis validated many known FSGS-causing genes, we detected a nontrivial number of purported “disease-causing” variants in controls, implying that filtering is inadequate to allow clinical diagnosis and decision making. Genetic diagnosis in patients with FSGS is complicated by the nontrivial rate of variants in known FSGS genes among people without kidney disease.

JASN 30: 1625–1640, 2019. doi: https://doi.org/10.1681/ASN.2019020152

Focal and segmental glomerulosclerosis (FSGS) is a common histologically defined pattern of kidney injury. The clinical syndrome associated with FSGS varies in severity, age of onset, progression of disease, responsiveness to steroids, and degree of proteinuria.1 Over the past two decades, the importance of genetic factors in the development of FSGS has become increasingly clear. More than 50 single gene forms of the pathologically defined entity FSGS or the related clinically defined phenotype steroid-resistant nephrotic syndrome (SRNS) have been identified.2 As a general rule, the autosomal dominant forms of FSGS tend to be late onset, associated with progressive kidney disease, and often manifest with subnephrotic proteinuria and the absence of the full set of features that define the nephrotic syndrome. This is typical of kidney
disease associated with mutations in ACTN4, INF2, and TRPC6.\textsuperscript{3–5} In contrast, recessive forms of FSGS and SRNS tend to be earlier in onset and associated with more severe proteinuria and overt nephrotic syndrome.\textsuperscript{6} Among the most severe form of inherited SRNS is congenital nephrotic syndrome caused by mutations in the nephrin gene, NPHS1. Mutations in NPHS1 typically cause neonatal nephrosis and diffuse podocyte foot process effacement, rather than FSGS. However, at least one report has suggested increased frequency of rare NPHS1 variants in patients with FSGS.\textsuperscript{7} It is reasonable to view early onset nephrotic syndrome and later onset FSGS as part of a spectrum of phenotypes caused by a variety of mutations in genes that primarily affect podocyte function.\textsuperscript{8} In addition, it is increasingly clear that mutations in genes not typically thought of as glomerular disease genes can lead to FSGS as a secondary response to primary injury. As examples, mutations in type 4 collagen genes and nephronophthisis genes have been reported in association with FSGS.\textsuperscript{9–12}

The majority of known FSGS and SRNS genes have been identified through linkage analysis (including homozygosity mapping) and subsequent sequence analysis of the candidate regions.\textsuperscript{6} Here, we report on the results of an exome–sequencing study aimed at understanding the genetic landscape of mutations and variants in patients and families carrying the diagnosis of FSGS. We examined the frequencies of rare variants in known FSGS-causing genes, other known kidney disease genes, and the rest of the human exome in a cohort of patients with familial or sporadic FSGS. We performed the same analysis in a similar number of control individuals. We analyzed the burden of such variants in cases versus controls. We identified several new genes that may be possible contributors to the risk of FSGS. The top 50 new genes in our analysis are significantly closer to known FSGS genes in a gene interaction network analysis than are randomly chosen sets of genes, suggesting that several of these genes are in fact true contributors to FSGS. Importantly, we find that for many genes previously described as FSGS or kidney disease genes, nonsynonymous and loss-of-function (LOF) variants are as common in matched controls as they are in cases. These results have important implications for understanding the landscape of genetic variation in FSGS, as well as for the interpretation of genetic testing in patients with FSGS.

METHODS

Human Subjects

The study was approved by the institutional review board at Beth Israel Deaconess Medical Center. Individuals belonging to 395 families with FSGS were included in this study. Familial FSGS–affected status was defined as having either a reported history of proteinuria with urine albumin-to-creatinine ratio of 250 mg/g, nephrotic syndrome, or biopsy-proven FSGS in a family with at least one other case of documented FSGS or nephrotic syndrome. We obtained blood or saliva for DNA isolation as well as clinical information after receiving informed consent. Control exome data were obtained from the Yale Center for Mendelian Genetics. DNA for sequencing was extracted from blood or saliva using standard methods. Of 1284 exomes, 1079 (84%) were captured by Roche’s MedExome target enrichment kit. The other 205 exomes were captured by different capture technologies. This study analyzed 1284 exomes. Excluding the repeated samples, this represents 1241 unique samples, 622 controls and 619 affected individuals or their family members. The number of sequenced individuals per family ranged from one to seven (Supplemental Table 1).

Variant Calling, Quality Control, and Inbreeding Coefficient Estimation

Read quality was checked and preprocessed by FastQC\textsuperscript{13} and Trimmomatic (version 0.36). Quality-checked FASTQ files were aligned against the human reference genome (GRCh38, hg38) by BWA (version 0.7.13-r1126).\textsuperscript{14} Variants were called according to the best practices for use of the Genome Analysis Toolkit (version 3.5–0–g36282e4).\textsuperscript{15,16} Only variants located in exon regions and the 100-base flanking region of each exon were called. The gene exon annotations were from refSeq (hg38).\textsuperscript{17} CrossMap (v0.2.2)\textsuperscript{18} was used to convert the genome coordinates from hg38 to hg19 for the VCF file. Three additional filters were also applied at each individual call, read depth ≥4, genotype quality ≥20, and minor read ratio ≥0.2 for heterozygous sites. The average read depth was approximately 20–30× after removing individuals with read depth ≤10×. Three individuals with an excess number of Mendelian errors were also removed. Individuals were also removed if genetic relationship was not consistent with pedigree structure. Additionally, the genome-wide inbreeding coefficient was estimated by IBDLD software.\textsuperscript{19}

Variant Annotation

Variants were annotated by Variant Effect Predictor (VEP) software,\textsuperscript{20} with annotated information including: minor allele

Significance Statement

Despite many known monogenic causes of FSGS, single gene defects explain only 30% of cases. In this study, sequencing of 662 exomes from families with FSGS and 622 control exomes validated many known FSGS-causing genes. However, for some genes previously reported as FSGS related, they identified a number of purported “disease-causing” variants in controls at similar or higher frequencies. They also identified multiple additional candidate FSGS genes in which rare variants were more common among cases. Network analysis showed that their top-ranked genes were located significantly closer to known FSGS genes compared with a random gene set. These findings imply that genetic diagnosis in patients with FSGS is complicated by the nontrivial rate of variants in known FSGS genes among people without kidney disease.
frequency [MAF] of continental populations (in the GnomAD project\textsuperscript{21} and in the Exome Aggregation Consortium [ExAC] project\textsuperscript{21}); the Genomic Evolutionary Rate Profiling (GERP) score\textsuperscript{22}; CADD score\textsuperscript{23}; and deleterious effect prediction by PolyPhen2,\textsuperscript{24} SIFT,\textsuperscript{25} M-CAP,\textsuperscript{26} LRT,\textsuperscript{27} and the MutationTaster method.\textsuperscript{28} The confidences of LOF effects of splice sites, nonsense sites, and indels were annotated using the LOFTEE plugin of the VEP software.\textsuperscript{20}

**Rare Variant Spectrum in Known Disease-Causing Genes**

In order to catalog the spectrum of rare variants in known disease-causing genes, we performed rare variant filtering according to the suspected genetic inheritance model, the allele frequency, and the genetic prediction scores from \textit{in silico} genetic score prediction software. Variant cosegregation with phenotype was checked whenever possible. Only the splicing sites, stop gain/loss sites, in-frame/frame-shift indels, and nonsynonymous sites were analyzed. Note that for the compound heterozygous model we did not have phasing information available. In order to comply with the American College of Medical Genetics (ACMG) standards for gaining at least a supporting level of pathogenic classification by computational prediction, we required that putative causal variants should be predicted as damaging by at least three algorithms of five prediction methods used for nonsynonymous rare variants, or as LOF variants classified by the LOTTEE plugin of VEP\textsuperscript{20} as high-confidence LOF sites. Additional filters specifically for each inheritance model were as follows:

\textit{Dominant Model}

(1) The maximum population-level allele frequency (pop_max) was assumed to be \(<0.0001\) in the general population. (2) The identified rare variant was absent from control samples. (3) CADD score \(\geq 10\).

\textit{Recessive Model (Homozgyous)}

(1) Maximum population frequency \(<1\%\). (2) No homozygous alternative allele genotype in GnomAD and ExAC databases, or in our in-house control samples. (3) CADD score \(\geq 10\).

\textit{Recessive Model (Compound Heterozygous)}

(1) Maximum population frequency \(<8\%\). Note: the allele frequency of R229Q in NPHS2 was 0.06893 in the European Finnish population in the GnomAD database. (2) At least two rare heterozygous genotypes in a single gene, such that at least one of these heterozygous sites did not have a homozygous genotype found in the GnomAD and ExAC databases.

**Ancestry Labeling and Case/Control Matching**

Our samples were mixed with the samples from the 1000 Genomes Project\textsuperscript{29} for the principal component analysis (PCA). The population label was consistent with the label from the 1000 Genomes Project; in total, six ancestry labels were used (EUR, European; EAS, East Asian; AMR, admixed American [MXL, Mexican Ancestry from Los Angeles; PUR, Puerto Ricans from Puerto Rico; CLM, Colombians from Medellin, Colombia; and PEL, Peruvians from Lima, Peru]; SAS, South Asian; AA, African American; AFR, African). The ancestry of our samples was inferred by the \(k\)-nearest neighbors (k-NN) algorithm on the basis of the distance between individuals measured in the top ten principal components space. Specifically, we set \(k=1\), and the distance between two individuals \((i\ and \ j)\) was measured as \(D_{ij} = \sum_{m=1}^{10} V_{m} \times (P_{mi} - P_{mj})^2\), where \(V_{m}\) is the proportion of variance explained by the \(m_{th}\) principal component, and \(P_{ij}\) is the \(k_{th}\) projection value of individual \(i\). We matched each case by one control with the closest distance measured by \(D\).

**Association Test for Common Variants and Gene-Based Burden Test for Moderately Rare Variants**

The genomic inflation factor \((\hat{\lambda})\textsuperscript{30,31} and the data for QQ plot were estimated by PLINK2 software\textsuperscript{32} for the case/control data of before and after case/control matching. Only common variants (MAF\(\geq0.05\)) were used in this analysis. In order to deal with heterogeneous variant effects and suboptimal allele frequency cut-offs, three methods (variable threshold [VT] test, C-\(\alpha\) test, and SKATO test) were applied for moderately rare variants (MAF\(\leq5\%\)). These tests were performed using PLINK/seq (http://atgu.mgh.harvard.edu/plinkseq/) and EPACTS software (http://genome.sph.umich.edu/wiki/EPACTS).

**Gene-Based Burden Test for Extremely Rare Variants**

For different inheritance models, the qualified variants were classified similarly to how variants in known disease-causing genes were classified, but only case/control-matched data were used. After variant filtering and classification, two tests were performed. The “single-group” burden test\textsuperscript{33,34} (Supplemental Figure 1A) compares the observed rare variant rate to the expected rate of each gene. The “case-control” burden test (Supplemental Figure 1B) compares the observed rare variant rate between case and control samples. To validate the performance of this method, we explored the model performance by different parameter combinations. The performance was assessed by the total information content \(I(I = - \sum \log(2(P_i)))\) and by the ranking of these genes (Supplemental Figure 2). To test the asymmetric distribution of the top genes between case and control samples, the rank sums of genes from each group were compared using the Wilcoxon rank-sum test (\(U\) test) to check the overall rank orders between these groups.\textsuperscript{35}

**Enrichment Test for Top-Ranked Genes in Renal Glomerulus–Enriched Genes**

Nine glomerular expression datasets were collected by Ding et \textit{et al.}\textsuperscript{36} A high-confidence set of glomerulus-enriched genes was defined by those genes that have significant differential expression between glomerular and nonglomerular compartments in at least two datasets (see Table 1 of Ding, et \textit{al.}\textsuperscript{36}). Fisher’s exact test was used to evaluate the significance of the enrichment of renal glomerulus–enriched genes in top-ranked genes.
Table 1. Summary for familial and sporadic cases

<table>
<thead>
<tr>
<th>Group</th>
<th>Patients</th>
<th>Age of Onset</th>
<th>Sex</th>
<th>Ancestry</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Male</td>
<td>EUR</td>
</tr>
<tr>
<td>Familial cases</td>
<td>337</td>
<td>22.6 (15.8)</td>
<td>179</td>
<td>158</td>
</tr>
<tr>
<td>Sporadic cases</td>
<td>187</td>
<td>20.1 (14.9)</td>
<td>77</td>
<td>110</td>
</tr>
</tbody>
</table>

EUR, European; EAS, East Asian; SAS, South Asian; AA, African American; AMR, admixed American (MXL, Mexican Ancestry from Los Angeles, PUR, Puerto Ricans from Puerto Rico, CLM, Colombians from Medellin, Colombia, and PEL, Peruvians from Lima, Peru); AFR, African.

aMean in years (SD).

Network Distance Analysis
We define the distance \((d_{ij})\) of two genes \((i\) and \(j)\) as the shortest path between gene \(i\) and \(j\) of all of the possible paths between them on a given network. This shortest path problem was solved by Dijkstra’s algorithm.\(^{37}\) Given two gene sets \(K\) and \(T\), we define the distance between a gene \(f\) from \(T\) to gene set \(K\) as \(D_{K} = \min\{d_{ik}, \text{for } k \in K\}\), the minimal distance of \(t\) with all of the genes from \(K\). Then, the distance \((G_{TK})\) of a candidate gene set \(T\) with a known gene set \(K\) was defined as \(G_{TK} = \text{average}\{D_{K}, \text{for } t \in T\}\), the average of the distance of all of the candidate genes from \(T\) with its minimal distance to genes from \(K\). Depending on the comparisons, the known gene set \(K\) and the novel gene set \(T\) were set appropriately. Statistical significance was estimated by repeating the random process 1000 times. Each time, \(T\) was replaced by 50 random genes from a gene pool that was covered by the exome-sequencing target. We also performed another analysis by fixing the top 50–ranked new genes, but replacing the known genes by randomly picked genes for the estimation of the background distribution, repeating this process 1000 times. The network analysis was implemented in Java on the basis of the Java version of Dijkstra’s algorithm from Princeton University (https://algs4.cs.princeton.edu/44sp/DijkstraUndirectedSP.java.html).

Two gene-gene interaction networks were used as back-bones for checking and comparing distances, the STRING\(^{38}\) and inBioMap networks.\(^{39}\) The connection score \((C_i)\) in these networks was converted to \(C_{\text{max}}\) \(-\) \(C_i\) in order to apply the shortest path algorithm. \(C_{\text{max}}\) is the maximum edge score in a network. In order to check the robustness of the gene set distance between top-ranked new genes and known disease-associated genes, we performed “leave one out” and “leave three out” analyses (details in Supplemental Material).

RESULTS

Study Design, Samples, Sequencing, and Quality Control
To identify genes that contribute to FSGS, we analyzed a total of 1284 whole exomes including 662 exomes from individuals diagnosed with, or with family members diagnosed with, FSGS (Figure 1), either sporadic or familial, and 622 exomes from unrelated control individuals (all adults). None of the control individuals were reported to have kidney disease, although, in most cases, they were not extensively studied for kidney-related phenotypes. The number of sequenced individuals per family in our cohort ranged from one to seven (Supplemental Table 1). In total, we sequenced samples from 395 distinct family units, including 337 cases coming from 208 families with multiple affecteds and 187 from sporadic cases (Supplemental Table 1, Table 1). Variants from our cohort were called using the Genome Analysis Toolkit.\(^{15,16}\) To estimate individual-level ancestry and perform quality control, we performed PCA and genetic relationship estimation using the SNPRelate R package.\(^{40}\) Our case and control samples had a world-wide ancestry distribution, with the majority having predominantly recent European ancestry (Figure 2, Table 1). As revealed by PCA-based clustering, the breakdown for families was 73.4%, 21.7%, 4.5%, and 0.25% for European, African American and Admixed American, Asian, and African ancestry, respectively (see Methods).

Next, given the diverse ancestry distribution, we applied a case/control matching approach to control for population structure in our case/control statistical testing.\(^{41,42}\) After the removal of low-call-rate samples, 363 family units, either unrelated FSGS families or sporadic FSGS cases, with high-quality sequence were used for downstream analysis (Supplemental Table 1). PCA plots for the matched case and control samples are shown in Figure 2, C and D. As shown in Figure 2, E and F, the genomic inflation factor \((\lambda)\) for a common variant association test dropped from 1.21 to 1.06 after case/control matching, indicative of an improvement in our ability to control for population structure.

Exome Sequence Analysis for Genetic Diagnosis and Identifying Candidate Genes for FSGS
We hypothesized that in our FSGS cohort, approximately 30% of samples would be explained by mutations in previously reported genes on the basis of earlier reports, with a higher rate of mutation identification in the familial cases.\(^{43,44}\) We also hypothesized that mutations in genes not yet identified as FSGS-causing contributed to the development of disease in a nontrivial fraction of these subjects. We therefore performed several analyses: (1) characterizing the rare variant distribution spectrum in known disease-causing genes \((i.e.,\) genetic diagnosis \((2)\) common variant association test; and \((3)\) burden test for moderately rare variants and extremely rare variants. For each class of variants, different kinds of statistical methods were applied to test the association of variants/genes with the disease status.

Rare Variants Identified in Known Disease-Causing Genes
In order to define the distribution of rare variants in known FSGS-associated genes, we performed variant filtering on the basis of each family’s compatibility with the different inheritance models. Variants were filtered on the basis of the annotation of
allele frequency and the predicted effects of several in silico genetic prediction tools (see Methods).20 We checked the distribution of these variants in 165 known kidney disease–related genes (excluding kidney stone genes, Supplemental Table 2)6.

Known Dominantly Acting Genes
First, dominant inheritance genes were analyzed in the 363 case family units (including a total of 483 affected individuals) and 363 matched controls. The rare variants in each individual were filtered by consistency with a dominant inheritance model (i.e., shared by all affecteds in a family if multiple affecteds were available), allele frequency ≤0.0001 in ExAC and GnomAD databases, as well as CADD≥10. Figure 3A and Supplemental Table 3 illustrate the numbers of families (cases) and controls carrying rare variant genotypes in known disease-associated genes that passed these filters. COL4A5, WT1, UMOD, and INF2 were the genes identified most frequently as harboring likely causal mutations, with only a small number of similar variants identified in controls. However, we identified several known disease genes in which a nontrivial number of rare variant genotypes passing these same filtering conditions in the control samples (Figure 3A). For example, genes MYH9, FN1, SALL1, PODXL, TBX18, ROBO2, COL4A6, and SRGAP1 were all found to have a higher burden of such variants in control samples than in cases. There are also seven kidney disease genes in which suspicious variants were detected only in controls, namely LMX1B, SIX2, UPK3A, BMP4, DSTYK, HNF1B, and SIX5.

Given this observation, we explored whether a stricter filter using ACMG guidelines could help us better distinguish the causal rare variants from noise. We applied filters to include only those rare variants with multiple lines of computational evidence supporting pathogenicity following these ACMG guidelines. Specially, we included only nonsynonymous rare variants predicted as damaging by at least three of five algorithms (PolyPhen2,24 SIFT,25 M-CAP,26 LRT,27 and MutationTaster28), as well as splice sites, nonsense sites, and indels that were classified as high-confidence LOF by the LOTTEE plugin of VEP.20 We refer to such variants as LOF+3D. After application of these stricter filtering guidelines, most of the known dominant disease genes in which we observed a high variant burden in controls still retained this trend (Figure 3D). Thus, applying a stricter in silico classification is not sufficient to distinguish causal variants from noise.

We further partitioned the rare variants seen in cases according to whether they were observed in patients with familial or sporadic FSGS (Supplemental Figure 3). The rare variants in COL4A5 and WT1 were evenly distributed between familial and sporadic patients. In contrast, rare variants in UMOD and COL4A3 were significantly enriched in those patients with familial inheritance (P value <0.05, Fisher’s exact test). For INF2, TRPC6, and PAX2, we observed trends toward detection in familial cases, but the statistical significance was limited by the smaller numbers.

Figure 1. Data analysis diagram. After quality control, 363 unrelated case families with ancestry-matched controls were included in this study. Rare variants in known disease-causing genes were profiled (genetic diagnosis); common variants association test and gene burden test for moderately rare and extremely rare variants were conducted thereafter.
Figure 2. Quality control for study samples. (A and B) PCA plot for case and control samples mixed with 1000 Genomes Project samples. EUR, European; EAS, East Asian; AMR, admixed American; SAS, South Asian; AA, African American; AFR, African. (C and D) PCA plot for case and control matched samples. (E) QQ plot for the genome-wide association study before case/control matching. (F) QQ plot for the genome-wide association study after case/control matching, where $\lambda$ is the genomic inflation factor.
Known Recessive Disease Genes

For genes following recessive inheritance, 266 affected cases from 236 unrelated family units (one case per family plus sporadic cases) with an inheritance pattern compatible with a recessive model and 236 matched controls were analyzed. For the compound heterozygous model, the rare variants in each family were filtered by: (1) allele frequency $\geq$0.08 in ExAC and GnomAD databases (a number slightly greater than the most common recessive monogenic FSGS variant NPHS2 p.R229Q, with allele frequency as high as 6.8% in some populations per the GnomAD database); (2) the presence of at least two heterozygous rare variants in the same gene; and (3) at least one of the heterozygous variants lacking any corresponding homozygous genotype found in either the ExAC or GnomAD databases (for details see the Methods section). We also required that this genotype be shared by all of the affecteds in a family. We identified FRAS1, FAT1, and PKHD1 as the three genes most likely to have a genotype consistent with disease contribution under this model (Figure 3B). However, we detected a similar burden of such genotypes in control samples for FRAS1, FAT1, PKHD1, ALMS1, DYNC2H1, CUBN, AH1, NPHP4, WDR35, WDR19, TTC21B, and RPRGIP1L, as well as C5orf42 (Figure 3B, Supplemental Table 4 illustrated by orange bars). There are also several genes in which suspicious variants were detected only in controls, including ZNF423, PDE6D, NUP93, NPHP3, NEK1, KANK4, ITGA4, ITGA3, IFT140, GRIP1, GRHPR, EMP2, COL4A6, CEP41, CC2D2A, BBS9, and ACE genes. Even after applying stricter in silico thresholds for classification of LOF (LOF+3D sites only), we still find a high burden of variants in these genes in control individuals, including FRAS1, FAT1, FRME2, ADCK4, RPRGIP1L, PKHD1, NPHP4, NPHP3, DYNC2H1, and CUBN genes (Figure 3E).

To assess the hypothesis that these genes are likely to harbor false-positive “disease-causing” genotypes, we used the gene damage index score$^{45}$ and frequently mutated genes (FLAGS) classification (rare variants in gene “Frequently Mutated in Public Exomes”).$^{46,47}$ On the basis of the gene damage index prediction, FRAS1, FAT1, ALMS1, DYNC2H1, and CUBN are all significantly enriched for rare variants in the general populations.$^{45}$ These five genes have also been labeled as FLAGS, as all are observed to be frequently mutated in public exomes, irrespective of the phenotypes studied.$^{46,47}$ By contrast, for several other genes, the burden of disease-consistent genotypes in cases is much greater than in controls. For example, we saw no candidate disease-causing genotypes in NPHS2 or COL4A4 among controls, but a significant burden in cases. As shown in Supplemental Table 5, the distribution of variant NPHS2 genotypes is consistent with previous reports, with the disease-causing variants seen in combination with the relatively common p.Arg229Gln variant all located in exon 7 or 8.$^{48}$

Under a homozygous recessive model, the rare variants were filtered by CADD$\geq$10, allele frequency $\leq$0.01, and the absence of any homozygous genotype in ExAC and GnomAD databases (details in Methods section). Eleven families have rare homozygous variants in the known disease-causing genes C5orf42, COL4A3, IFT80, NPHP3, NPHP2, PLCE1, NUP93, and OFD1 (Figure 3C, Supplemental Table 6). All of these families show a high level of inbreeding (Supplemental Table 6); ten of 11 families had a genome-wide inbreeding coefficient ranked in the top 12% of all of the families tested, and four ranked in the top 1%. All of the presumed causal variants detected in these families are very rare, with the maximum population allele frequency $<0.0066$ as estimated from the ExAC database.$^{21}$ However, attributing a causal relationship to variants in C5orf42, IFT80, and OFD1 is difficult, because these genes contained candidate variants that were not predicted to be damaging and were detected in control samples. Except for the variants in these three genes, all of the other missense variants were predicted to be either very strongly damaging or LOF and all were located in evolutionarily conserved regions$^{22}$ (Supplemental Table 6). All of the rare variant genotypes identified in NPHS2, COL4A3, NPHP3, NUP93, and PLCE1 show compelling genetic evidence for being disease causing (Figure 3, Supplemental Table 6).

Fraction of Families Explained by Rare Variants in Known Genes

By grouping together all of the rare variants detected under the above models, rare variants or rare variant genotypes in known FSGS and FSGS-like disease-causing genes were detected in 49.6% (180 of 363) of unrelated families and sporadics, although 35.5% (129 of 363) of control individuals showed a similar rare variant genotype. This observation in cases is clearly an overestimation of the disease explanation rate, because a nontrivial proportion of genes harbor a similar or higher level of burden in control samples (Figure 3). In order to give a better estimate of the identification of true causal genotypes, we removed those genes that were among the top 100 FLAGS genes$^{46,47}$ or with a similar or higher level of rare variant burden in controls (with enrichment of rare variants in cases versus controls $\leq$1.5 in our data set). We refer to these genes as “high control burden genes” hereafter. When we do this, 33.6% (122 of 363) of unrelated families or sporadic cases have rare variants in known kidney disease genes (Figure 4A). By comparison, only 8.3% (30 of 363) of the control individuals have rare variant genotypes that pass the same filtering procedure as used for case samples.

We observed a higher proportion of rare variant rate in familial cases (37.2%, 70 of 188) than sporadic cases (29.7%, 52 of 175). These numbers are similar to previous findings in SRNS, where NPHS1, NPHS2, and WT1 are the major contributors.$^{45,44}$ In our cohort, COL4A5, WT1, and COL4A4 are the top three contributors under a dominant or X-linked model and NPHS2 is the top contributor under a recessive model (Figure 3). The differences in the relative contributions of the major genes reflect differences in subject ascertainment.$^{43,44}$ COL4A5, WT1, NPHS2, COL4A4, UMOD, COL4A3, INF2, ACTN4, PLCE1, CEP290, PAX2, and
Figure 3. Rare variants discovered in known genes. Variants were filtered by allele frequency, CADD score, and genotype-phenotype cosegregation. (A) Variants were filtered by dominant model/X-linked model. (B) Variants were filtered by recessive compound heterozygous model. (C) Variants were filtered by single site homozygous recessive model. (D–F) Variants filtered by the same conditions as A, B, and C, respectively, with further variant filtering following ACMG guidelines, where only variants with multiple lines of computational evidence supporting pathogenicity were included. *Variants outside of the known disease-causing domain were not included.
TRPC6 accounted for 62.8% of the families with likely causal rare variant genotypes (Figure 4B). For these 12 genes, together with known disease-causing protein domain information, we have very high confidence in assigning disease causality to the rare variants identified.3–5 However, we are not able to use in silico prediction alone to attribute disease causality for high control burden genes (Figure 3). The in silico predictions of CADD and GERP scores of the extremely rare variants (frequency, 1/10,000) seen in case and control samples are indistinguishable (P value >0.05, t test; Supplemental Figure 4). Even by applying stricter filtering (LOF+3D sites), an equal or higher level of burden in control samples remained (Figure 3, lower panel). This highlights the difficulty in using simple sequencing and filtering for classifying variants as disease-causing even in those genes previously reported as disease-associated. Previously reported data as well as future data regarding these genes and their variants therefore need to be interpreted cautiously.

**APOL1**

Two specific variants in APOL1 (termed G1 and G2) greatly influence the risk of FSGS (with an odds ratio of approximately 10 (95% confidence interval 6.0 to 18.4) under a recessive model), and are common in people of recent African ancestry.49–51 We therefore looked specifically at APOL1. We counted the number of samples carrying two copies of the so-called G1 and G2 risk variants (genotypes G1G1, G2G2, and G1G2, referred as high-risk APOL1 genotypes). Of the 363 unrelated families studied, 34 families (9.4%) carried a high-risk APOL1 genotype. All of the members tested from these families were found to be admixed individuals (labeled as AA or AMR) by PCA-based ancestry inference (see Methods section), which is consistent with the African origin of the G1 and G2 variants.49–51 None of the control samples studied were found to have a high-risk APOL1 genotype. We note that two sporadic cases and one control are compound heterozygous for rare APOL1 variants (Supplemental Tables 4 and 7). When including APOL1 high-risk genotypes, 41.3% (150 of 363) of families or sporadic cases can be explained by a genetic cause or genetic susceptibility. We note that ten families with high-risk APOL1 genotypes also had rare variants in known kidney disease genes, of which eight families qualify for LOF+3D criteria (Supplemental Table 8). The severity of the phenotype, as assessed by age of disease onset, was not significantly different in the individuals with mutations in a known disease gene with a high-risk APOL1 genotype compared with those with a high-risk APOL1 genotype alone. In addition, we found no evidence to support a genetic interaction between APOL1 and these other genes, as the number of co-occurrences of these events is less than the expected value of independence (specifically: 10<363×0.094×0.336≈11).

**Age of Disease Onset in the Different Inheritance Models**

We looked at the age of disease onset in families with rare variants detected in known disease-associated genes (excluding high control burden genes) after partitioning by inheritance model. Age of onset is significantly lower for those families with...
the homozygous recessive inheritance model when compared with the other two models (P value < 0.05, U test\textsuperscript{35}; Supplemental Figure 5). The age of disease onset in families with dominant inheritance is also slightly higher than that of families with the compound heterozygous model, but not significantly (P value = 0.14, U test\textsuperscript{35}).

Analysis for the Effect of Common Variants and Moderately Rare Variants
We performed an association study to try to identify coding variants associated with increased risk of FSGS, given the high-density coverage of common coding variants in the whole-exome-sequencing data. The only locus that reached genome-wide significance was APOL1, specifically the G1 (rs73885319 and rs60910145) variant (P value = 1.69E-7; Supplemental Figure 6). This implies that larger sample sizes are required to detect the contribution of common coding variants other than those in APOL1 to FSGS risk.

To look for the possible contribution of moderately rare variants, we used three other burden tests: (1) variant threshold (VT) test, (2) C-\(\alpha\) test, and (3) SKAT-O, in order to deal with the unknown optimal cut-off of allele frequency (with the VT test) and heterogeneity of variant effects (with C-\(\alpha\) test and SKAT-O). We applied these three methods to our ancestry-matched case/control data for those rare variants with MAF < 5%. APOL1, WT1, and COL4A5 were ranked in the top ten by the P values from the VT test (Supplemental Table 9), but none of them reached a genome-wide level of significance. The APOL1 signal was driven by the G1 and G2 variants, and the WT1 and COL4A5 signals were driven by the very rare variants as discussed earlier (Figure 3, Supplemental Table 3).

Gene-Based Burden Tests for Extremely Rare Variants
Next, we wanted to determine whether additional new genes could also be identified as likely contributors to the development of FSGS, given that in only 33.6% of families with disease be explained by rare variants in known Mendelian disease genes. We performed a single-group burden test and a case-control burden test. In the single-group burden test, the observed number of rare variants was compared with the expected number estimated from the whole-genome background (Supplemental Figure 1A).\textsuperscript{33,34} In the case-control burden test, the number of observed rare variants in cases was compared with the observed number in controls (Supplemental Figure 1B). The statistical power of the single-group test was tuned by testing several combinations of different parameters. As shown in Supplemental Figure 2, we found that the best combination of parameters for the single-group analysis was to use missense and LOF variants (splice-altering, nonsense, frameshift, and indels) with zero allele frequency as estimated from the GnomAD database.\textsuperscript{21} In this setting, the positive control genes WT1, COL4A5, ACTN4, INF2, TRPC6, and PAX2 had the largest total information content \(I\), (where \(I = \sum \log_2(P_i)\), the smallest overall P values of these genes), and the best overall ranking (summation of the rank of each gene, Supplemental Figure 2, A and B).

To help visualize and distinguish the distribution of disease-associated genes, we compiled the information from the single-group burden test with that of the case-control burden test in a single plot. We subtracted the number of rare variants seen in each gene in controls from that of cases (x axis of Figure 5) and plotted this difference against the P value of single-group analysis (y axis of Figure 5). Using this graphing method, the known disease-causing genes PAX2, NPHS2, TRPC6, WT1, COL4A5, INF2, and ACTN4 are all located in the top-right quadrant of the graph (Figure 5), as expected.

In order to quantitatively determine whether the distribution of candidate genes in cases compared with controls was asymmetric, we compared the rank order of genes from the single-group analysis. We expected that the genes from the case group would have a smaller rank sum (with smaller \(P\) value) than those from the control group, because the case group should include rare variants that are in fact causally related to the disease in addition to statistical noise. As shown in Table 2, we observed a significantly smaller rank sum in the case group when either the top 50 or 75 genes were chosen for the analysis (P value = 0.02 and 0.01, respectively). The same trend was also found when the top 100 genes were compared (P value = 0.08). As a negative control, we analyzed rare synonymous variants (zero allele frequency in GnomAD database) in the same manner as we did for the rare missense and LOF variants. In the case of the synonymous variants, a smaller rank sum from the case group was not observed (Supplemental Table 10). In fact, there was a trend toward lower rank sum in controls (Supplemental Table 10). Therefore, despite a limited number of genes reaching genome-wide significance, we conclude that there is an enrichment of true disease-causing rare variants in the case group, where enrichment of disease-causing rare missense and LOF variants is driving genes to have smaller rank sum in the case group.

Ranking FSGS Genes Using the Rare Variant Burden Test
We next combined the burden information from the y axis (single-group analysis) and x axis (case-control analysis) shown in Figure 5 as a single value using Fisher’s method. The combination of the \(P\) values as a single value provides a simpler gene rank list. As shown in Table 3, the top four genes in this analysis were WT1, COL4A5, TRPC6, and NPHS2. The known disease-causing genes INF2, ACTN4, and PAX2 also rank near the top of this list, with ranks of 6, 20, and 56, respectively. As expected, we found a significant enrichment of our top-ranked genes toward those genes enriched for glomerular expression (Supplemental Figure 7, Supplemental Table 11).\textsuperscript{36} When the top 20 genes were considered, the enrichment odds ratio was 8.51 (\(P\) value < 0.001; 95% confidence interval, 3.21 to 22.09). The top 60 genes ranked by this method (and the associated rare variants) are listed in Supplemental Tables 12 and 13.
Network-Based Analysis of Top Candidate Genes

Next, we wanted a better sense as to whether it is likely that some of our highly ranked new genes are true contributors to kidney disease and FSGS. To test this, we compared the gene set distance between the set of our top-ranked new genes and previously identified genes associated with kidney disease or with FSGS (Supplemental Tables 2 and 14), because previous studies have shown that disease-causing genes tend to cluster together.52,53 We curated two sets of disease-causing genes, $K_{fgs}$ (Supplemental Table 14) and $K_{expanded}$ (Supplemental Table 2). The first gene set, $K_{fgs}$, comprises genes with compelling genetic evidence as causative of FSGS or FSGS-like phenotypes when mutated; this set is called the FSGS panel.6 The other set, $K_{expanded}$, is a mixture of genes that are associated with a broader range of kidney phenotypes (excluding kidney stones).6 We referred to this set as the EXPANDED panel. The shorter distance between disease-causing genes was validated by a shorter distance between the $K_{expanded}$ gene panel (excluding genes in $K_{fgs}$) and the $K_{fgs}$ gene panel than the distance between disease-causing genes and the randomly picked genes ($P$ value $\leq 0.002$; Figure 6, Supplemental Table 15). Next, we computed the gene set distance between the 50 top-ranked new genes ($T$, Supplemental Table 12, with known kidney disease risk genes excluded if any) and the $K_{fgs}$ or $K_{expanded}$ gene panel, respectively. As shown in Figure 6 and Supplemental Table 15, compared with a randomly chosen set of 50 genes, the top 50 genes ranked from the control group, or the top 50 genes ranked by synonymous rare variants from the case group, our top 50 new genes had shorter distances to the known disease-causing genes. This held true for both the STRING and the inBio Map networks (Figure 6, Supplemental Table 15).38,39 Fixing these 50 new genes and bootstrapping the known genes 1000 times by choosing random genes, we saw that the distance between the top 50 new genes was also significantly closer to known disease-associated genes than to genes chosen by random bootstrapping ($Z$ score $\leq -1.87$, $P$ value $=0.02$).

To determine whether this shorter distance of our top-ranked new genes was driven by a few genes or driven by at least several genes, we performed “leave one out” and “leave three out” analyses. As shown in Supplemental Figure 8, the shorter distance was robust to these analyses, confirming that the shorter distance was driven by at least several genes ranked among the top. Taken together, these results strongly suggest that among the variants identified in the most highly ranked genes (Supplemental Table 12, Table 3), a nontrivial subset are in fact true contributors to the development of FSGS.

**DISCUSSION**

In this study, we performed exome sequencing in a large set of individuals diagnosed with familial or sporadic FSGS and
compared this data with exome sequence data from control individuals. We identified rare nonsynonymous variants and LOF variants in Mendelian disease genes that are plausible contributors to disease in 49.6% of cases (not including APOL1, a non-Mendelian FSGS risk gene). This is almost certainly an overestimate of the true percentage of families explained by known disease genes, because (1) rare variant genotypes were detected in 35.5% of controls and (2) for a number of known kidney disease genes (specifically MYH9, FN1, SALL1, PODXL, TBX18, ROBO2, COL4A6, SRGAP1, FRAS1, FAT1, PKHD1, ALMS1, DYN2C2H1, CUBN, AH1, NPHP4, WDR35, WDR19, RGRP11, and C5orf42) we identified rare variant genotypes at similar or higher frequencies in controls as in cases. A more conservative estimate, excluding those genes with an equal or higher variant burden in controls, puts the percentage at 33.6% for cases compared with 8.3% in controls. In our analysis, a higher explanation rate was found in familial cases: 37.2% of familial cases and 29.7% of sporadic cases could be attributed to variants in known FSGS genes. This is consistent with the notion of a third of individuals with the histologic diagnosis of FSGS; (2) in a significant subset of these cases, the genotype is of unclear significance, because genetic variants indistinguishable from putative disease-causing variants are found in several Mendelian disease genes in control individuals at rates similar to that in cases. Even by applying stricter genetic score–based filtering according to ACMG guidelines in which multiple lines of in silico mutation effect prediction were used, a similar or higher rare variant burden in controls was still present in several genes. This demonstrates the difficulty in using filtering and in silico prediction alone for judging the pathogenicity of rare variants in many genes reported to cause FSGS (or kidney diseases that can phenocopy FSGS). After filtering by ACMG guidelines (and for consistency with observed inheritance pattern in familial cases), the only genes in which the remaining variants were limited to cases but not controls were COL4A5, WT1, INF2, ACTN4 (but only when restricting to the known disease-associated domain), and TRPC6, under a dominant model; and NPHS2, PLCE1, OFD1, NPHP3, COL4A3, C5orf42, CEP290, TTC8, TTC21B, TMEM67, PLCE1, NPHS1, LRRG2, LAM2B, ITGA8, COQ6, and ALMS1, under a recessive model. However, attributing disease to genes with a high control variant burden, such as C5orf42, TTC21B, and ALMS1, should be done with extreme caution.

The difficulty of assigning disease causality to rare variants was also demonstrated by the large proportion of cases with candidate rare variants detected in more than one disease-associated gene. In 14.9% (54 of 363) of familial cases or sporadic cases, we detected two or more genes with a candidate rare variant genotype. After removing those genes with a similar or
higher burden in control samples, 5.9% (21 of 363) of the familial and sporadic cases still remain with more than one candidate mutation.

Among the merits of this study is the analysis of case samples together with a set of well matched controls. On the basis of the observation of many putative disease genes in which the rare variant burden in cases was similar to that in control samples, we conclude that special care must be taken to assign disease causality in those genes based solely on in silico genetic prediction. When we used a larger set of candidate genes for analysis, rare variant genotypes were detected in 49.6% (180 of 363) of cases, although a surprisingly high 35.5% (129 of 363) of control individuals showed similar rare variant genotypes. The high rare variant burden in known kidney disease genes in controls was also highlighted recently in a large cohort with 7974 self-declared healthy adults in which such genes were analyzed.55 Panels of genes for analysis in individuals with kidney disease are being used with increasing frequency in the clinic. Our results emphasize that substantial caution should be taken in both the design and interpretation of such panels.

ACMG guidelines are widely used for reducing the false-positive rate in claiming a variant to be disease related. After applying ACMG guidelines in our cohort, 30.9% (112 of 363) of case families have a rare variant genotype in a plausible kidney disease gene, compared with 18.7% (68 of 363) in controls (Figure 1). Thus, a nontrivial proportion of rare variant genotypes remained in control samples after application of this strict filter. Removing high control burden genes gives the values 24.2% and 5.5%, respectively (Figure 1). The high proportion in control samples illustrates the limited help of ACMG guidelines in this setting. Reasons for this may include incomplete penetrance of disease-causing variants and incorrect labeling of some genes as disease-causing in the existing literature.

After filtering by ACMG guidelines and removal of genes with high variant burden in controls, there are still 5.5% of control samples with rare variants in previously reported disease genes that appear similar to what are presumed to be disease-causing variants in cases. This may be because of several reasons: (1) incomplete penetrance, (2) the existence of “neutral” or unknown significance rare variants in our genome, (3) false-positive calls from sequencing errors or unknown phasing information under a compound heterozygous model, and (4) limited power to filter out “neutral” rare variants and false-positive genes. This nontrivial number cannot be attributed to undetected kidney phenotypes in the control, because rare variants were observed in cases but almost none in controls for several well established disease-causing genes, including COL4A5, WT1, INF2, ACTN4, TRPC6, NPHS2, NPHP3,

Figure 6. Closer network distance of top ranked genes with known kidney disease–associated genes than a random set of genes. We computed the network distance between the 50 top-ranked genes (with known kidney disease–associated genes excluded) identified from the rare variant burden test under a dominant model and known disease-associated genes. “fsgs” refers to genes known to cause FSGS when mutated; “expanded” means an expanded set of genes that are associated with various kidney disease phenotypes. “expanded.fsgs” indicates the normalized network distance between the “fsgs” panel and the “expanded” gene panel (excluding overlaps with “fsgs” panel) as a positive control. “fsgs.CASE” indicates the network distance of the 50 top-ranked genes from case samples on the basis of rare missense variants and LOF variants compared with the FSGS panel. “expanded.CASE” indicates the distance from the top 50 genes to the “expanded” panel. “fsgs.CTRL” and “expanded.CTRL” are control analyses that indicate the network distance between the 50 top-ranked genes from the control samples chosen by the burden of rare missense variants and LOF variants and the “fsgs” or “expanded” panels. “fsgs.CASE.syn” and “expanded.CASE.syn” indicate the network distance between the 50 top-ranked genes from case samples by synonymous rare variants and the “fsgs” or “expanded” panels. The background distribution was estimated by bootstrapping 1000 times, and randomly picking 50 genes from the pool of all genes to replace the 50 top-ranked genes. (A) Network distance measured on the basis of the structure from the STRING network. (B) Network distance measured on the basis of the structure from the inBio Map network.
and APOL1. Unlike other genes with very high control variant burdens, these genes are not “contaminated” by controls in this study. The sample with variants that are “neutral” or of unknown significance in cases is expected to be <5.5%, because we have >30% of families in which we sequenced two or more members (and 17% in which we sequenced three or more members). Therefore, we are better able to filter out false-positive variants by checking cosegregation in case families.

On the other hand, this strict filtering removed variant genotypes in several case families which are likely to be in fact disease-causing (for example, in WT and NPHS2). There are several possible reasons for this. For example, there is discordance between different genetic prediction algorithms. The availability of increasing numbers of large next-generation sequencing cohorts, together with better resolution of regional burdens, these genes are not known.

This study confirms the observation that mutations in type IV collagen genes are a frequent contributor to FSGS.11,57 It also confirms that WT1 mutations, which cause a variety of syndromes of abnormal urogenital development with glomerulopathy, are a frequent cause of inherited FSGS.44,58,59 In addition, UMOD variants, a known cause of CKD characterized by tubule dysfunction, can present with a histologic lesion of FSGS (presumably developing secondary to the tubulopathy).60 Although we were not able to identify new FSGS genes in this study, our results make us confident that our top-ranking candidate genes include some that do in fact contribute to FSGS risk. Specifically, in a network analysis, we find that the top 50 genes on our list (after eliminating the known genes) are significantly closer to known FSGS genes as well as a longer list of kidney disease genes than they are to a randomly chosen set of genes. By comparing the rare variant distribution in cases with controls and by comparing the distribution pattern of missense and LOF variants with synonymous rare variants, we observe that there is enrichment of missense and LOF rare variants in case samples (Figure 5, Supplemental Table 10, Table 2). We conclude, therefore, that some of these genes and their variants are almost certainly contributing to FSGS. However, without either analyses of larger families (e.g., large families segregating these variants) or larger sample sizes (more individuals with familial and/or sporadic FSGS), we cannot at present determine which of these genes are most likely to be true contributors to disease.

ACKNOWLEDGMENTS

We thank the study participants.

Dr. Wang, Dr. Chun, and Dr. Pollak designed the study. Ms. Wilkins, Ms. Benjamin, Ms. Knob, and Dr. Appel carried out patient ascertainment and managed the DNA sequencing. Dr. Wang analyzed the data with the help of Dr. Chun, Dr. Genovese, Dr. Friedman, and Dr. Pollak. Dr. Lifton and Dr. Mane provided control sample data and performed exome sequencing in both cases and controls. Dr. Wang, Dr. Chun, and Dr. Pollak drafted and revised the paper.

DISCLOSURES

Dr. Friedman and Dr. Pollak are co-inventors on patents related to APOL1 diagnostics and therapeutics. Dr. Friedman and Dr. Pollak receive research support from and have consulted for Vertex Pharmaceuticals. Dr. Friedman and Dr. Pollak have equity in Apolo1Bio.

FUNDING

This work was supported by National Institutes of Health grant R01DK54931 (to Dr. Pollak) and by National Institutes of Health award UM1HG006504-08 (Yale Center for Mendelian Genomics, to Dr. Lifton and Dr. Mane). Dr. Chun was supported by an Alberta Innovates Health Solutions Clinician Fellowship and a KRESCENT (Kidney Research Scientist Core Education and National Training Program) postdoctoral fellowship.

SUPPLEMENTAL MATERIAL

This article contains the following supplemental material online at http://jasn.asnjournals.org/lookup/suppl/doi:10.1681/ASN.2019020152/-/DCSupplemental.

Detailed methods.

Supplemental Figures:

Figure S1: Illustration figures for the “single-group” analysis (A) and “case-control” analysis (B).

Figure S2: “Single-group” burden test for extreme rare variants of dominant model, showing the statistics for known FSGS genes WT1, COL4A5, ACTN4, INF2, TRPC6, and PAX2.

Figure S3: The rare variant distribution in familial and sporadic families for each dominant disease-causing gene.

Figure S4: The comparison of rare variant CADD and GERP score between the case and control samples for those dominant genes with high burden in control samples.

Figure S5: Diagnosis age distribution for affected individuals with rare variants detected in known disease-causing genes (without genes with high level of burden in control samples), partitioned by inheritance model.

Figure S6: Manhattan plot for the results of the association test of the exonic common variants.
Figure S7: The enrichment of renal glomerulus expression enriched genes in top ranked genes.

Figure S8: Robustness analysis of gene set distance comparison.

Supplemental Tables:
Table S1. Summary of study samples.
Table S2. The gene list of kidney disease-associated genes.
Table S3. The rare variants in dominant/X-linked kidney disease-associated genes.
Table S4. The rare variants in recessive (compound-heterozygous) kidney disease-associated genes.
Table S5. The rare variants in NPHS2 gene.
Table S6. The rare variants in NPHS1 gene.
Table S7: Individuals of compound heterozygous for rare APOL1 variants.
Table S8. Rare variants detected in APOL1 high risk genotype families.
Table S9. Top 100 genes for moderate and rare variant burden test.
Table S10. Comparison of gene rank sum between case and control group, genes were ranked by synonymous sites.
Table S11. The enrichment of glomerulus-expression enriched genes in top ranked genes.
Table S12. Top ranked 60 genes, known genes and new genes.
Table S13. The rare variants in top ranked 60 genes (Table S12).
Table S14. FSGS disease associated genes.
Table S15. Network distance of top ranked new genes with known genes.

REFERENCES


AFFILIATIONS

1Division of Nephrology, Department of Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts; 2Department of Medicine, Harvard Medical School, Boston, Massachusetts; 3Program in Medical and Population Genetics, Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts; 4Division of Nephrology, Department of Medicine, University of Calgary, Cumming School of Medicine, Calgary, Alberta, Canada; 5Stanley Center for Psychiatric Research, Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts; 6Division of Nephrology, Department of Medicine, Columbia University College of Physicians and Surgeons, New York, New York; 7Laboratory of Human Genetics and Genomics, The Rockefeller University, New York, New York; and 8Department of Genetics, Yale University School of Medicine, New Haven, Connecticut