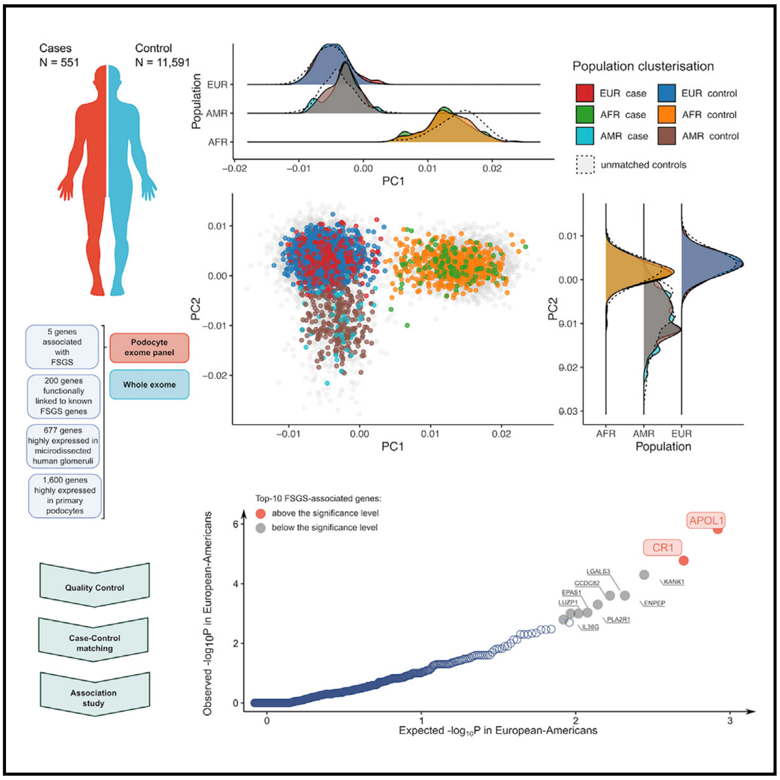
iScience

Article

# *CR1* variants contribute to FSGS susceptibility across multiple populations

## Graphical abstract




## Highlights

- Analysis of the podocyte-related gene panel across 726 FSGS cases and 13,994 controls
- Rare variant association tests identified a significant association with the CR1 gene
- rs17047661 linked to malaria protection is identified as a risk variant for FSGS


## Authors

Rostislav Skitchenko, Zora Modrusan, Alexander Loboda, ..., Mark J. Daly, Andrey Shaw, Mykyta Artomov

## Correspondence

shaw.andrey@gene.com (A.S.), mykyta.artomov@nationwidechildrens.org (M.A.)


## In brief

Association analysis; Clinical syndrome; Human Genetics; Quantitative genetics

Skitchenko et al., 2025, iScience *28*, 112234
April 18, 2025 
https://doi.org/10.1016/j.isci.2025.112234

CellPress

iScience

CellPress
OPEN ACCESS

Article

# *CR1* variants contribute to FSGS susceptibility across multiple populations

Rostislav Skitchenko,[1,2] Zora Modrusan,[3] Alexander Loboda,[1,2,4] Jeffrey B. Kopp,[5] Cheryl A. Winkler,[6] Alexey Sergushichev,[1] Namrata Gupta,[4] Christine Stevens,[4] Mark J. Daly,[4,7,8] Andrey Shaw,[3,*] and Mykyta Artomov[4,9,10,11,*]
[1]ITMO University, St. Petersburg, Russia
[2]Almazov National Medical Research Centre, St. Petersburg, Russia
[3]Research Biology, Genentech Inc., San Francisco, CA, USA
[4]Broad Institute, Cambridge, MA, USA
[5]Kidney Disease Section, Kidney Diseases Branch, National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), NIH, Bethesda, MD, USA
[6]Molecular Genetic Epidemiology Studies Section, National Cancer Institute (NCI), Frederick, MD, USA
[7]Massachusetts General Hospital, Boston, MA, USA
[8]Institute for Molecular Medicine Finland, Helsinki, Finland
[9]Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, OH, USA
[10]Department of Pediatrics, The Ohio State University College of Medicine, Columbus, OH, USA
[11]Lead contact
*Correspondence: shaw.andrey@gene.com (A.S.), mykyta.artomov@nationwidechildrens.org (M.A.)
https://doi.org/10.1016/j.isci.2025.112234

## SUMMARY

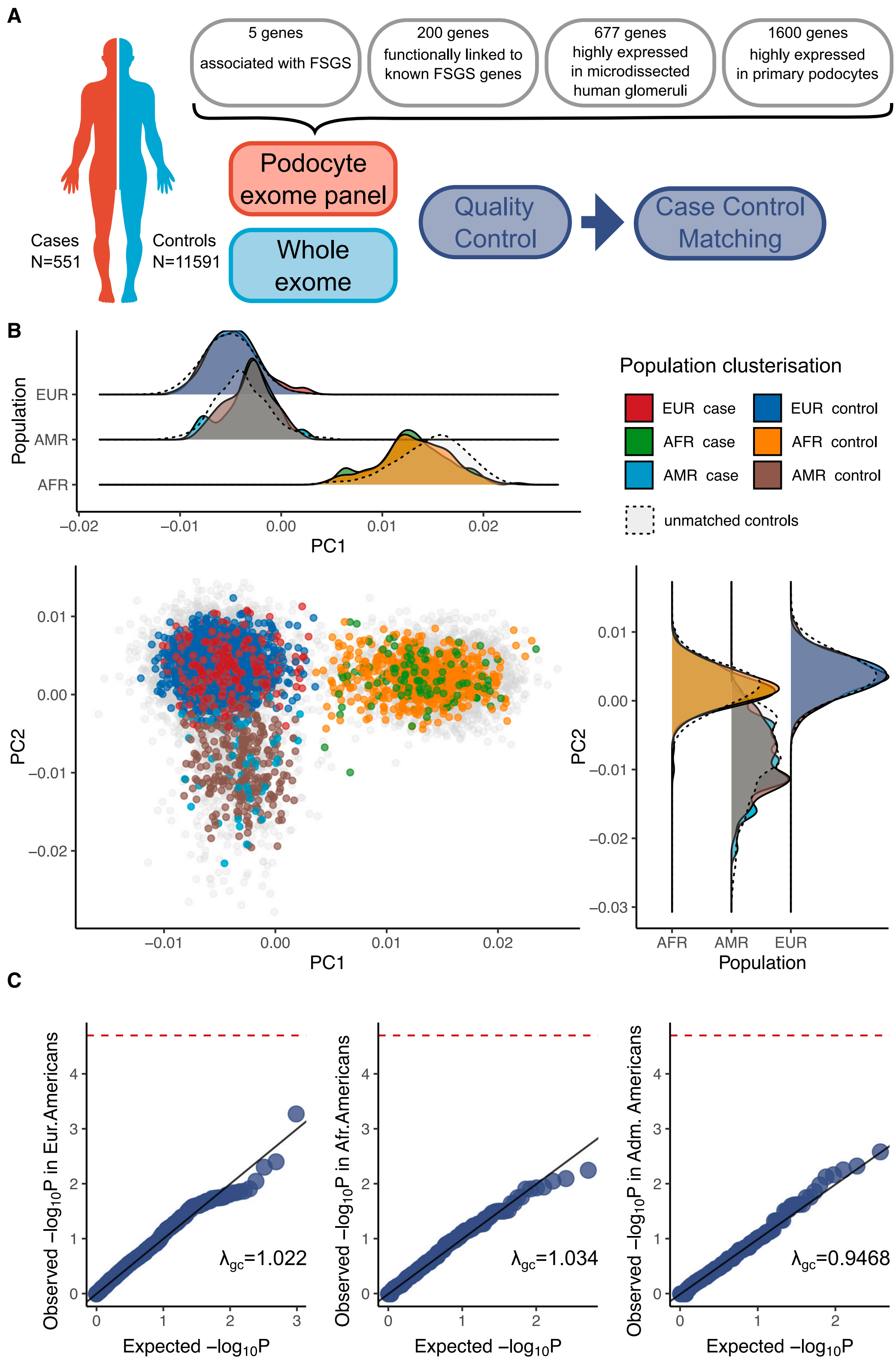
Focal segmental glomerulosclerosis (FSGS) is a leading cause of nephrotic syndrome, with an annual incidence of 24 cases per million among African-Americans and 5 per million among European-Americans in the United States. It ranks as the second most common glomerular disease in Europe and Latin America and the fifth in Asia. We conducted a case-control study involving 726 FSGS cases and 13,994 controls from diverse ethnic backgrounds, using panel sequencing of ∼2,500 podocyte-expressed genes. Rare variant association tests confirmed known risk genes (*KANK1*, *COLAPOL1*) and identified a significant association with the *CR1* gene. The *CR1* variant rs17047661, which encodes the SI1/SI2 (R1601G) allele, was previously linked to cerebral malaria protection and is now identified as a risk variant for FSGS. This highlights an evolutionary trade-off between infectious disease resistance and kidney disease susceptibility, emphasizing the role of adaptive immunity in FSGS pathogenesis and potential therapeutic targets.

## INTRODUCTION

Focal segmental glomerulosclerosis (FSGS) is a common cause of primary nephrotic syndrome among both adults and children in the USA, and its incidence is increasing.[1,2] The incidence and prevalence of FSGS are not precisely known due to the requirement of a kidney biopsy for diagnosis and the lack of a central registry. Estimates for incidence range from 1.4 to 21 cases per million population.[3] The incidence of FSGS in the U.S. is about 4 times higher in African Americans (6.8 patients per million) and 2 times higher in Latin-Americans (3.7 patients per million) compared to European-Americans (1.9 patients per million).[4]

Genetic studies of FSGS, conducted using both pedigree analyses and cohort-based association studies, have identified a number of susceptibility genes, yet explaining only a fraction of family-history enriched cases.[5] The first genetic studies identifying the chromosome (chr) 22 region with FSGS were prompted by the observed higher prevalence of FSGS in African and African American populations, suggesting that one or more FSGS susceptibility gene variants would be enriched on African-derived haplotypes.[6–8] The subsequent discovery of association of G1 and G2 coding variants in *APOL1* with FSGS provided an explanation for increased prevalence of the disease among African-descent populations. Pleiotropic properties of these variants resulted in protection against trypanosomiasis but at the cost of increased FSGS risk.[9]

In this study, a large-scale genetic database was assembled from biopsy-confirmed cases of focal segmental glomerulosclerosis (FSGS) and ethnically matched controls. The study used a panel of approximately ∼2,500 genes associated with podocytes, which play a crucial role in the formation and maintenance of the glomerular filtration barrier. The purpose of the study was to investigate the genetic basis of FSGS and identify novel susceptibility genes. This study is a significant extension of previous work conducted by Yu et al.[5] with increased power and a more diverse multi-ethnic cohort with greater sample size.




*(legend on next page)*

## RESULTS

### Population stratification and quality control analysis

To investigate the genetic basis of FSGS across multiple populations, we designed a case-control study examining variants in a podocyte exome panel of approximately 2,500 genes (Figure 1A; STAR Methods, method details). Control samples were obtained from general population cohorts not ascertained for kidney disease (Table S1). After quality control and coverage analysis (Figure S1), we performed population stratification and case-control matching (Figure 1B). Population clustering analysis identified eight initial clusters, with minor clusters being excluded due to limited sample size (Figures S2–S24). Power analysis demonstrated our dataset's superiority over previous FSGS studies (Figures S5–S56). Using common synonymous variants, we confirmed the absence of systematic bias between cases and controls in all population clusters (Figures 1C, Figure S7).

### Association analysis of common variants in European-American cohort

Using a dataset of matched cases and controls after quality filtration, we performed several association studies. Because analyses of this dataset were limited to the "podocyte exome", we focused on missense variants and protein truncation variants (PTV).

For each cluster separately, we conducted a variant-based association study using linear regression (from the "base" library R) with no additional covariates. 3,777 variants with missense and PTV (stop_gained, frameshift_variant, splice_acceptor_variant, splice_donor_variant) effects on protein were included in this analysis. In the European-American cluster, two variants were significantly associated with FSGS: (1) rs601314 – $p = 8.1x10^{-9}$, reference allele is a minor allele; $OR_{minor\ allele} = 13.24$ ($CI_{95\%}$ = [3.996, 56.51]), missense, *EFEMP2* (EGF-containing fibulin extracellular matrix protein 2); and (2) rs117071588 – $p = 4.0x10^{-6}$, alternative allele is a minor allele; $OR_{minor\ allele} = 11.66$ ($CI_{95\%}$ = [2.790, 68.35]), missense, *CCDC82* (coiled-coiled domain containing 82), Significance threshold was determined with Bonferroni correction – $p < 0.05/3,777 = 1.32x10^{-5}$; Figure S8, Table S2). Replication analysis of these two variants in the African American cohort showed that the rs601314 variant (*EFFMP2*) was not significantly associated with FSGS ($p$ = 0.062, reference allele is a minor allele, $OR_{minor\ allele} = 0.7418$, $CI_{95\%}$ = [0.5370, 1.015]) and that rs117071588 (*CCDC82*) was absent from the African American dataset. A similar situation was observed when replication was attempted in the Latin-American cohort: rs601314 was not significantly associated ($p$ = 0.097, reference allele is a minor allele, $OR_{minor\ allele}$ = 2.99, $CI_{95\%}$ = [0.6891, 14.76]), rs117071588 was absent from the data. Despite the significant association statistics of *EFEMP2* and *CCDC82* in the European-American group, the lack of replication of the *EFEMP2* variant FSGS association in the other populations makes this a less robust finding but might serve as a starting point for future studies.

### In silico pathogenicity analysis of EFEMP2 and CCDC82 variants

The rs601314 in *EFEMP2* (NC_000011.9:g.65636053T>C, ENSP00000434151:p.I259V) and rs117071588 in *CCDC82* (NC_000011.9:g.96117537A>C, ENSP00000278520:p.D125E) variants are defined by most common *in silico* pathogenicity predictors as benign.[10] FATHMM classified rs601314 as "damaging" (Fathmm score converted = 0.48).[10] The specific predictors of missense deleteriousness classified rs601314 and rs117071588 as benign (MISTIC<0.5).[11] Variant rs601314 affects the von Willebrand factor type A (vWA) domain of *EFFMP2* and variant rs117071588 affects the domain of unknown function (DUF4196) of *CCDC82*.[10] Both variants have a missense effect on protein function and come from non-conservative parts of proteins (MPC<2.0).[12]

### *Rare variant burden analysis in European-American population*

We carried out rare variant burden analyses in the European-American cohort, focusing on missense variants and PTV with a population frequency below 0.01. This cutoff was chosen according to the presence of a signal in each of the quartiles of allele frequency distribution in the interval [0; 0.01] (Figure S9). To identify multiple possible risk patterns, a rare variant association study (RVAS) was performed using five tests representing different statistical classes of methods for each gene. If most variants are causal and have unidirectional effects, classical burden tests are optimal, due to their high power. However, adaptive burden tests are considered more reliable than those using fixed weights or thresholds in the scenario when there is no clear evidence for selecting specific allele frequency threshold to define "rare" variant.[13] In addition, some tests can improve interpretability of the results. Tests of variance components are effective when there are multiple variants with opposite directionality of the effect on trait susceptibility, or a very limited number of causal variants.[13] To cover the variety of possible genetic effects, we used Fisher's exact test, C-alpha, adaptive sum

**Figure 1. Study design, principal component analysis, and quantile-quantile plots to identify calibration of synonymous variants between case and control cohorts**

(A) Case-control study design. DNA samples from FSGS cases and controls were examined for coding variants in a podocyte exome panel gene panel composed of 2482 genes and also was subjected to whole exome analysis.

(B) Principal-component analysis illustrates case-control matching in European-derived, African-derived and Latin-American-derived-populations and demonstrates genetic segregation of these three populations.

(C) Quantile-quantile (QQ)-plots for the association study of the common synonymous variants with gnomAD population specific allele frequency $\geq$0.01). These plots illustrate case-control matching quality for the European-derived (left), African-derived (middle), Latin-American-derived (right) populations. The test lambda-GC (genomic inflation factor) for genome-wide association studies (GWASs) compares the median test statistic against the expected median test statistic under the null hypothesis, in which there is no association for each variant. This test identifies systemic biases and significant associations. Here, most of the points fall along the diagonal, indicating the absence of systemic bias. Abbreviations. EUR, European-Americans. AFR, African-Americans. AMR, Admixture Americans.

statistic (ASUM), weighted sum statistics (WSS) and kernel-based adaptive clustering (KBAC) (Figure S10). All of the tests successfully identified previously known FSGS susceptibility genes – APOL1, KANK1, COL4A4, IL36G among the top associations. Top associations also included the significant signal for *CR1*, mutations in which have not been previously reported in FSGS. We found that KBAC and ASUM were somewhat under-powered compared to the other tests. While capturing the expected top candidates, statistical power was not always sufficient for experiment-wide significance. This observation suggests that FSGS risk variants do not follow distinct co-inheritance patterns in affected individuals, and their contribution to disease risk may not be fully captured by adaptive weighting or specific multi-site genotype configurations.[13,14]

The resulting *p* values were combined using the Simes method for multiple hypothesis testing, which is suitable for merging dependent test statistics (Table S3). The top associated genes included four genes, *APOL1*, *KANK1*, *COL4A4*, *IL36G*, that were previously identified in FSGS association studies and had their functionality confirmed using murine studies.[5] The top two genes reached significance after Bonferroni correction ($p = 0.05/2{,}482 = 2.015\text{x}10^{-5}$): *APOL1* ($p = 1.47\text{x}10^{-6}$), a known FSGS susceptibility gene, and *CR1* ($p = 1.67\text{x}10^{-5}$), a gene with no previously reported germline variants in FSGS (Figures 2A and S11).

## Multi-population replication analysis of APOL1 and CR1 variants

Significantly associated genes in the European-American cohort were further examined in a replication study of the African American and Latin-American cohorts. Neither *APOL1* nor *CR1* were replicated using rare (MAF<0.01) variant analysis (Table S4). Previously observed positive selection acting on the *APOL1*[9] variants in the African American population suggests that FSGS risk variants might be too common to be detected by an RVAS in non-European populations. Therefore, we used the variant-based tests to replicate the FSGS-association signals in *APOL1* and *CR1*.

We identified rare variants in the European-American cohort that drove the association signals in *APOL1* and *CR1* and four variants, consisting of pair-locus G1 in *APOL1* (rs60910145 [NC_000022.10:g.36662034T>G, ENSP00000317674.4:p.Ile400Met] and rs73885319 [NC_000022.10:g.36661906A>G, ENSP00000317674.4:p.Ser358Gly][9]) and two closely adjacent variants in *CR1* (rs17047661 [NC_000001.10:g.207782889A>G, ENSP00000356016.4:p.Arg2051Gly] and rs17047660 [NC_000001.10:g.207782856A>G, ENSP00000356016.4:p.Lys2040Glu]) were selected for replication in other ancestries (Figure 2B). Additionally, we eliminated the possibility of this result being a false positive due to coverage imbalance in the associated genes (Figure S12). *APOL1* variants were successfully replicated (rs73885319: $p = 0.001779$, alternative allele is a minor allele, $OR_{\text{minor allele}} = 1.59$, $CI_{95\%} = [1.18, 2.14]$; rs60910145: $p = 0.002271$, alternative allele is a minor allele, $OR_{\text{minor allele}} = 1.58$, $CI_{95\%} = [1.17, 2.12]$). The variants in *CR1* were more common and had smaller effect size, therefore, we lacked the statistical power to see the significant replication (rs17047661: $p = 0.28$, reference allele is a minor allele, $OR_{\text{minor allele}} = 1.22$, $CI_{95\%} = [0.84, 1.76]$; rs17047660: $p = 0.74$, alternative allele is a minor allele, $OR_{\text{minor allele}} = 1.09$, $CI_{95\%} = [0.69, 1.71]$).

Second replication was attempted in the Latin-American cohort because the variant frequencies for the variants of interest are more similar to the original European-American cohort. Variants rs17047661 in *CR1* and rs60910145 and rs73885319 in *APOL1* surpassed the replication significance threshold ($p = 0.05/4 = 0.0125$) (Table S5). Analysis of the statistical power for identified effect sizes in Latin-American and African American cohorts indicated that the lack of replication in the latter is most likely driven by the statistical power limitations (Figure S7).

## Meta- and selection analyses of associated variants

Meta-estimates of pMETAL[13] for all 4 variants were also calculated for the European-American and Latin-American cohorts: rs60910145 ($p_{METAL} = 9.706\text{x}10^{-6}$), rs73885319 ($p_{METAL} = 1.420\text{x}10^{-4}$), rs17047660 ($p_{METAL} = 0.6359$), rs17047661 ($p_{METAL} = 9.314\text{x}10^{-3}$). Interestingly, the variants in *CR1*: rs17047661 and rs17047660 are linked with only three out of four possible haplotypes observed in African subpopulations in 1000 genomes (AFR:YRI+LWK+GWD+MSL+ESN+ASW+ACB: $r^2 = 0.15$, D' = 1; YRI: $r^2 = 0.15$, D' = 1; ASW: $r^2 = 0.18$, D' = 1; ACB: $r^2 = 0.13$, D' = 1) and observed in global Latin-American population (AMR:MXL+PUR+CLM+PEL: $r^2 = 0.46$, D' = 1).

Next, the normalized integral haplotype score (iHS) was directly estimated in the discovery cohort for the variants included in the replication analysis. For the African American cohort, selection pressure analysis confirmed positive selection (iHS<−2.0) for the G1 APOL1 alleles: rs73885319 (iHS = −2.16), rs60910145 (iHS = −2.21) and revealed positive selection for rs17047660 (iHS = −2.71) in CR1, whereas no such selection was detected for rs17047661 (iHS = −1.04). The following results were obtained for the Latin American cohort: rs73885319 (iHS = −1.99), rs60910145 (iHS = −2.01), rs17047660 (iHS = −0.142) and rs17047661 iHS = 1.13).

Allele frequencies for all variants included in the replication analyses are significantly different between population groups, which can nominally indicate either positive selection or genetic drift. These included the following variants: rs60910145 (gnomAD EUR AF = $8.6\text{x}10^{-5}$, gnomAD AFR AF = 0.23, $p = 2.2\text{x}10^{-16}$); rs73885319 (gnomAD EUR AF = $1.1\text{x}10^{-4}$, gnomAD AFR AF = 0.23, $p = 2.2\text{x}10^{-16}$); rs17047661 (gnomAD EUR AF = $3.0\text{x}10^{-3}$, gnomAD AFR AF = 0.62, $p = 2.2\text{x}10^{-16}$); and rs17047660 (gnomAD EUR AF = $1.0\text{x}10^{-3}$, gnomAD AFR AF = 0.24, $p = 2.2\text{x}10^{-16}$). It is likely that iHS estimates can be skewed by complex population structure or demographic variables such as population growth, bottleneck events, and changes in recombination and mutation frequencies. Notably, the allele frequencies within African and African American populations in 1000 genomes significantly vary for rs17047661 ($AF_{YRI} = 0.69$, $AF_{LWK} = 0.70$, $AF_{GWD} = 0.79$, $AF_{MSL} = 0.79$, $AF_{ESN} = 0.72$, $AF_{ASW} = 0.58$, $AF_{ACB} = 0.66$).

We sought evidence of co-evolving changes in allele frequencies between (a) the G1 and G2 variants in *APOL1* and (b) replicated rs17047661 in *CR1*. We estimated the number of individuals who carry both rs17047661 and either one or both G1 and G2 alleles in the African American case cohort and compared this with the expectation of random assortment. There were no signs of linkage between these variants ($p = 0.93$, binomial

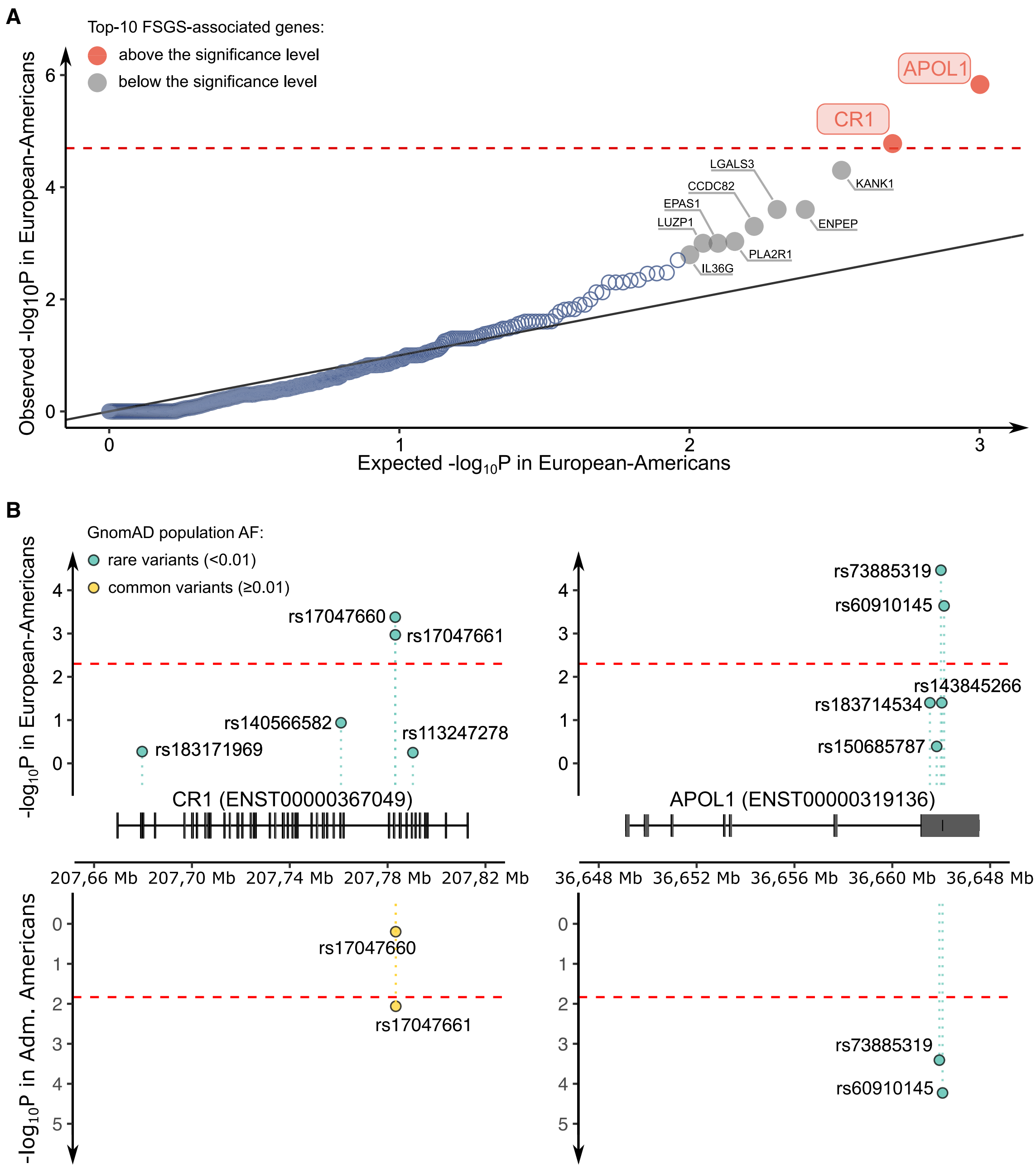



**Figure 2. Rare variant association study in the European-American cohort and replication of *CR1* and *APOL1* variants in the Admixed American cohort**

(A) Shown graphically are the results of rare-variant association study involving the European-American subject cluster (gnomAD EUR AF <0.01; missense and PTV (protein truncating variants). Statistical approaches included the following: the Simes method for multiple hypothesis testing, the Fisher exact test for testing two groups, C-alpha test for comparing the variance of each group against the expected, adaptive sum statistic (ASUM) for testing variants; weighted sum statistics (WSS) for testing variants and kernel-based adaptive clustering (KBAC) for variant classification and association testing. Of the top 10 (most significant) genes, four with known FSGS-associated variants. *CR1*, complement C3b/C4B receptor 1; *KANK1*, KN motif and ankyrin repeat domains 1; *COL4A4*, collagen type 4, alpha 4 chain; IL-36G, interleukin 36G.

(B) Shown are associations of FSGS-related SNPs in *CR1* and *APOL1* in the GnomAD aggregation database population among European-Americans (top panels) and Admixed-Americans (bottom panels). Left Panels. Two rare variants in *CR1* have been associated with FSGS in Euro-Americans, and one of these variants in *CR1* has also been associated with FSGS in Latin-Americans (both variants in *CR1* are common in Latin-Americans). Right panels. Two rare variants in *APOL1* are associated with FSGS in European-Americans and in Latin-Americans.

test), which suggests that the effects of rs17047661 are fully independent of those of *APOL1*.

### Functional prediction and domain analysis of CR1 variant

Most common *in silico* variant effect predictors do not categorize rs17047661 (NC_000001.10:g.207782889A>G, ENSP0000038 3744:p.R1601G) as pathogenic.[10] SNPred assigned pathogenicity scores of 4.1x$10^{-4}$ for rs17047661, indicating predicted weak effect on protein function.[15] Exceptions are, for example, PolyPhen2 HVAR and MutationAssessor, which classify rs17047661 as "probably damaging" (PolyPhen 2 Hvar Score = 0.964) and "medium functional effect" (Mutationassessor Score Converted = 0.70), respectively.[10] The rs17047661 variant of CR1 induces a missense effect (p.R1601G) in the protein domains common to secreted complement fixation protein (PHA02927) and complement control protein (CCP) modules, also known as consensus short repeats (SCR) or SUSHI repeats. Specifically, it affects the one of the four long homologous repeats (LHRs), LHR-D, which is responsible for binding C1q, Mannose binding lectin (MBL) and ficolin.[10,16] However, the specific predictors of missense deleteriousness classified rs17047661 as benign (MISTIC<0.5),[11] which is likely due to the nonconservative nature of the affected region (MPC<2.0).[12]

## DISCUSSION

The complement system is a complex network of proteins that play an important role in protecting the body against microbial infections, which are activated either through the classical immune pathway in response to binding to Fc-fragments of IgM or IgG, or through an alternative pathway of non-specific binding to antigens on membranes or to mannose residues through the lectin pathway.[17] Each protein in the cascade is activated by proteolysis, splitting the original proenzyme into "a" and "b" structures (the exception is C1, which splits into q, r, s molecules). The large molecule "b" is directly involved in the sequential activation of the complement system and the small molecule "a" is an anaphylatoxin, which causes degranulation of mast cells and chemotaxis of other immune cells, such as neutrophils, eosinophils, monocytes, and T lymphocytes. All of these factors have the potential to contribute to either innate immune functions or tissue injury. C3 is the central element of the complement system, which is activated by C3-convertase, a complex composed of the preceding elements of the cascade (classical/lectin pathway: C4bC2b complex, alternative pathway: C3bBb complex). Upon activation, the complement system can affect cells in two ways: (1) by forming the membrane attack complex (MAC, sC5b-9 complex), resulting in osmotic lysis of the targeted cell, and (2) indirect opsonization through the deposition of C3b on the surface of microbes, which facilitates phagocytosis by immune cells.

*CR1* acts as a negative complement regulator, reducing C3 activation and tissue deposition, by processing and bounding immune complexes, which then facilitates their transfer to the liver or spleen where macrophages ingest and eliminate them. In both the classical and lectin pathways, *CR1* has decay-activating activity, in that its binding C4b prevents the formation of C3-convertase. In the alternative pathway, *CR1* then acts as a cofactor for the cleavage of active C3b fragments (on C3c and C3dg fragments), significantly reducing deposition of C3b fragments, which could activate C3.[18] CR1 significantly reduces C3b deposition by ~80% over the classical pathway, but *CR1* is most potent when the alternative pathway is activated (>95% reduction in C3b deposition).[18] By stopping the activation of the complement system at the stage of C3-convertase formation, *CR1* is also indirectly involved in reducing the deposition of sC5b-9, which would have formed afterward if the complement system had been subsequently activated.

*CR1* is expressed by several cell types, including red blood cells (RBC), leukocytes, and among specialized renal cells, *CR1* is localized exclusively on glomerular podocytes. The FSGS related variants in *CR1*—rs17047661 which was replicated in Latin-American descent individuals and its pair—rs17047660 are known as Knops group polymorphisms and are a part of the red cell surface antigens, which give rise to the Sl2 and McC$^b$ alleles in the Swain-Langley 1 and 2 allele pairs (Sl1/Sl2) and McCoy a and b (McC$^a$/McC$^b$), respectively.[19] The K1590E substitution and the R1601G substitution in *CR1* are located just 11 amino acids away from each other and are in strong LD for both the 1000 genomes for the global African population ($r^2$ = 0.15, D' = 1) and for African American descent from study data ($r^2$ = 0.24, D' = 1). They are located in the long homologous repeats (LHRs), motif D, which is responsible for C1q and MBL binding.[19]

While significant changes in complement regulatory proteins, particularly reduced CR1 expression correlating with proteinuria levels, have been observed in FSGS patients, the comprehensive role of immune and complement systems genetic variants remains to be fully elucidated.[20] Recent studies have described the role of the complement system in various glomerulopathies.[21] Typically, the reduction in *CR1* expression is linked to the severity of the disease, as indicated by the degree of inflammation or tissue damage.[22] Autoantibodies directed against kidney-expressed autoantigens or antibody/antigen complexes deposited in the kidney are causative agents of various human kidney diseases. There are cases of C3-mediated inflammation and deposition.[23,24] Further, inhibition of C3 reduces proteinuria in animal models.[25] With regard to FSGS, IgG, and C3 deposits are often observed in the affected glomeruli, but the pathogenic role of these deposits remains still unclear, and therapy against the complement system has not been studied in FSGS.[26]

Another study has demonstrated elevated levels of Ba, Bb, C4a, and sC5b-9 in the plasma and urine of patients afflicted with primary FSGS. The detection of these protein deposits in the blood signifies that the complement cascade is activated at a site where fragments can gain entry into the vascular space, likely in the mesangial and sclerotic regions. Conversely, the rise in urine Ba, C4a, and sC5b-9 levels in some patients may reflect complement activation in the glomerulus, or alternatively, activation of filtered proteins in the tubular lumen or downstream in the urinary collection system.[27] While C5b-9 complexes generally form directly on the membranes of microorganisms, particularly Gram-negative ones, they can also affect adjacent cells, resulting in "bystander" harm.

The relationship between activation of the classical and alternative pathways in response to the presence of Knops antigens in *CR1* has been investigated previously, and a lack of correlation has been noted. However, there is a reasonable discrepancy between the results of serological studies of human samples and those obtained from parts of recombinant proteins. Prior investigations have challenged the conjecture that Sl2 and McC$^b$ influence the phenotype by modulating the activity of the cofactor implicated in the cleavage of C3b and C4b or the C1q binding activity.[28] Nevertheless, these findings warrant future exploration of the involvement of the lectin pathway in the activation of the complement system.[29] For example, the contribution of the lectin pathway in the activation of the complement system may play a crucial role in the development of progressive glomerular damage and long-term urinary abnormalities in patients with Henoch-Schönlein purpura nephritis (HSPN).[30] In the case of FSGS, the presence of MBL deposits that are focal and segmental has been observed, as reported in previous studies, and this can also result in tissue damage, MBL deficiency can lead to autoimmune diseases.[31,32]

There are many different pathogens that use *CR1* as a receptor for cell entry, for example: *Leishmania major*,[33] *Legionella pneumophila*,[34] *Leishmania panamensis*[35] and *Mycobacterium tuberculosis*.[36] It also has been shown that *CR1* is an RBC receptor used by *Plasmodium falciparum* for cell invasion, independent of sialic acid.[37] Consistently with this hypothesis, Sl2 (rs17047661) was previously associated with protection against cerebral malaria in sub-Saharan African populations, which resulted in much higher prevalence of Sl2 in African-descent individuals compared to Europeans.[16] In the study conducted by Opi et al., it was noted by the authors that the opposite concomitant effect of the McC$^b$ (rs17047660) allele on the development of severe malaria was of only nominal borderline significance, despite being under significant strong positive selection (iHS$<-2.0$).[16] At first glance, the association with severe malaria and significant positive McC$^b$ selection are discouraging, but this may explain the linkage of negative and protection haplotypes to each other. Moreover, the authors of the original article tested some haplotypes of Sl and McC combinations and found that the combination of Sl2/McC$^a$ alleles has an additive protective effect against malaria, which may explain the lack of replication signal for rs17047660 in the African American descent.[16]

In conclusion, the findings reported here establish the role of DNA variants in *CR1* in FSGS, involving an autoimmune disease component. Significant alterations in allele frequencies among populations suggest that environmental factors that induce selection pressure, might be responsible for an adaptive benefit, at the cost of kidney disease. These results, together with other evidence of the polygenic nature with many potential mechanisms of FSGS, could be used as a motivation for future GWAS, which would enhance understanding of the molecular genetic mechanisms underlying the disease.

### Limitations of the study

Several limitations should be considered when interpreting our findings. First, although our study included ethnically diverse populations, the sample size for some ethnic groups remained relatively small, potentially limiting our ability to detect population-specific genetic variants. While we achieved maximum statistical power for the given number of cases by optimizing our control sample size, detecting variants with smaller effects would require an increased number of cases. Second, our targeted sequencing approach focused on podocyte-expressed genes in primary FSGS cases; however, genetic predisposition to secondary FSGS might involve genes expressed in other cell types and tissues. This suggests that additional genetic factors contributing to secondary FSGS pathogenesis may exist beyond the podocyte-specific gene set examined in this study.

## RESOURCE AVAILABILITY

### Lead contact

Further information and enquiries should be directed to Mykyta Artomov (mykyta.artomov@nationwidechildrens.org).

### Materials availability

This study did not generate new unique reagents or materials.

### Data and code availability

- Data
  - FSGS cohort allele frequencies and gene burden rare allele counts are available in the Table S2.
  - Raw sequencing data for control cohort subjects are available through the database of genotypes and phenotypes (dbGaP); all relevant accession numbers are listed in the key resources table.
  - All raw data accession numbers and database identifiers used in this study are provided in the key resources table.
  - Any additional data that support the findings of this study are available from the corresponding author upon request.
- Code
  - Custom analysis codes have been deposited at Zenodo with https://doi.org/10.5281/zenodo.14883396 (listed in the key resources table).
  - The code repository is publicly accessible as of the date of publication and will be shared indefinitely.
  - Original GitHub repository link (https://github.com/rostkick/fsgs_meta) is also provided in the key resources table.

## ACKNOWLEDGMENTS

This work was supported in part by the Intramural Research Program, NIDDK, NIH. This project has been funded in part with federal funds from the National Cancer Institute at the National Institutes of Health, under contract 75N91019D00024. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. This Research was supported (in part) by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research and NIDDK. The authors acknowledge the contributions of the following investigators who recruited subjects for FSGS genetic studies, published in Kopp et al., Nature Genetics, 2008, as DNA from those subjects was used in the present study. These investigators include Kopp JB, Freedman BI, Ahuja TS, Berns JS, Briggs W, Cho ME, Dart RA, Kimmel PL, Korbet SM, Michel DM, Mokrzycki MH, Schelling JR, Simon E, Trachtman H. R.S. and A. Sergushichev. were supported by the Ministry of Science and Higher Education of the Russian Federation (Priority 2030 Federal Academic Leadership Program). A.L. was supported by Ministry of Science and Higher Education of the Russian Federation (Agreement # 075-15-2022-301).M.A. was in part supported by Nationwide Foundation Pediatric Innovation Fund.

## AUTHOR CONTRIBUTIONS

R.S., Z.M., J.B.K., C.A.W., M.J.D., A.S., and M.A. designed and conceived the study. R.S., Z.M., A.L., J.B.K., C.A.W., A. Sergushichev., A.S., M.J.D., and M.A. analyzed the data, J.B.K., C.A.W., M.J.D., A.S., and M.A. acquired funding, N.G. and C.S. managed control cohorts; M.J.D., A.S., and M.A. supervised the study R.S., J.B.K., C.A.W., M.J.D., A.S., and M.A. wrote the manuscript. All authors reviewed and approved the manuscript.

## DECLARATION OF INTERESTS

M.J.D. is a founder of Maze Therapeutics; A.S. and Z.M. are employees of Genentech Inc.

## STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Sample size and how subjects/samples were allocated to experimental groups
- METHOD DETAILS
- QUANTIFICATION AND STATISTICAL ANALYSIS

## SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2025.112234.

Received: February 25, 2024
Revised: November 25, 2024
Accepted: March 13, 2025
Published: March 18, 2025

## REFERENCES

1. O'Shaughnessy, M.M., Hogan, S.L., Thompson, B.D., Coppo, R., Fogo, A.B., and Jennette, J.C. (2018). Glomerular disease frequencies by race, sex and region: results from the International Kidney Biopsy Survey. Nephrol. Dial. Transplant. *33*, 661–669. https://doi.org/10.1093/ndt/gfx189.
2. Dragovic, D., Rosenstock, J.L., Wahl, S.J., Panagopoulos, G., DeVita, M.V., and Michelis, M.F. (2005). Increasing incidence of focal segmental glomerulosclerosis and an examination of demographic patterns. Clin. Nephrol. *63*, 1–7. https://doi.org/10.5414/cnp63001.
3. Shabaka, A., Tato Ribera, A., and Fernández-Juárez, G. (2020). Focal Segmental Glomerulosclerosis: State-of-the-Art and Clinical Perspective. Nephron *144*, 413–427. https://doi.org/10.1159/000508099.
4. Kitiyakara, C., Kopp, J.B., and Eggers, P. (2003). Trends in the epidemiology of focal segmental glomerulosclerosis. Semin. Nephrol. *23*, 172–182. https://doi.org/10.1053/snep.2003.50025.
5. Yu, H., Artomov, M., Brähler, S., Stander, M.C., Shamsan, G., Sampson, M.G., White, J.M., Kretzler, M., Miner, J.H., Jain, S., et al. (2016). A role for genetic susceptibility in sporadic focal segmental glomerulosclerosis. J. Clin. Investig. *126*, 1067–1078. https://doi.org/10.1172/JCI82592.
6. Kopp, J.B., Nelson, G.W., Sampath, K., Johnson, R.C., Genovese, G., An, P., Friedman, D., Briggs, W., Dart, R., Korbet, S., et al. (2011). APOL1 genetic variants in focal segmental glomerulosclerosis and HIV-associated nephropathy. J. Am. Soc. Nephrol. *22*, 2129–2137. https://doi.org/10.1681/ASN.2011040388.
7. Kopp, J.B., Smith, M.W., Nelson, G.W., Johnson, R.C., Freedman, B.I., Bowden, D.W., Oleksyk, T., McKenzie, L.M., Kajiyama, H., Ahuja, T.S., et al. (2008). MYH9 is a major-effect risk gene for focal segmental glomerulosclerosis. Nat. Genet. *40*, 1175–1184. https://doi.org/10.1038/ng.226.
8. Kao, W.H.L., Klag, M.J., Meoni, L.A., Reich, D., Berthier-Schaad, Y., Li, M., Coresh, J., Patterson, N., Tandon, A., Powe, N.R., et al. (2008). MYH9 is associated with nondiabetic end-stage renal disease in African Americans. Nat. Genet. *40*, 1185–1192. https://doi.org/10.1038/ng.232.
9. Genovese, G., Friedman, D.J., Ross, M.D., Lecordier, L., Uzureau, P., Freedman, B.I., Bowden, D.W., Langefeld, C.D., Oleksyk, T.K., Uscinski Knob, A.L., et al. (2010). Association of trypanolytic ApoL1 variants with kidney disease in African Americans. Science *329*, 841–845. https://doi.org/10.1126/science.1193032.
10. Edmonson, M.N., Patel, A.N., Hedges, D.J., Wang, Z., Rampersaud, E., Kesserwan, C.A., Zhou, X., Liu, Y., Newman, S., Rusch, M.C., et al. (2019). Pediatric Cancer Variant Pathogenicity Information Exchange (PeCanPIE): a cloud-based platform for curating and classifying germline variants. Genome Res. *29*, 1555–1565. https://doi.org/10.1101/gr.250357.119.
11. Chennen, K., Weber, T., Lornage, X., Kress, A., Böhm, J., Thompson, J., Laporte, J., and Poch, O. (2020). MISTIC: A prediction tool to reveal disease-relevant deleterious missense variants. PLoS One *15*, e0236962. https://doi.org/10.1371/journal.pone.0236962.
12. Samocha, K.E., Kosmicki, J.A., Karczewski, K.J., O'Donnell-Luria, A.H., Pierce-Hoffman, E., MacArthur, D.G., Neale, B.M., and Daly, M.J. (2017). Regional missense constraint improves variant deleteriousness prediction. Preprint at bioRxiv. https://doi.org/10.1101/148353.
13. Lee, S., Abecasis, G.R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. Am. J. Hum. Genet. *95*, 5–23. https://doi.org/10.1016/j.ajhg.2014.06.009.
14. Liu, D.J., and Leal, S.M. (2010). A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. PLoS Genet. *6*, e1001156. https://doi.org/10.1371/journal.pgen.1001156.
15. Molotkov, I., Koboldt, D.C., and Artomov, M. (2023). SNPred outperforms other ensemble-based SNV pathogenicity predictors and elucidates the challenges of using ClinVar for evaluation of variant classification quality. Preprint at medRxiv. https://doi.org/10.1101/2023.09.07.23295192.
16. Opi, D.H., Swann, O., Macharia, A., Uyoga, S., Band, G., Ndila, C.M., Harrison, E.M., Thera, M.A., Kone, A.K., Diallo, D.A., et al. (2018). Two complement receptor one alleles have opposing associations with cerebral malaria and interact with α+thalassaemia. Elife *7*, e31579. https://doi.org/10.7554/eLife.31579.
17. Freiwald, T., and Afzali, B. (2021). Renal diseases and the role of complement: Linking complement to immune effector pathways and therapeutics. Adv. Immunol. *152*, 1–81. https://doi.org/10.1016/bs.ai.2021.09.001.
18. Poppelaars, F., and Thurman, J.M. (2020). Complement-mediated kidney diseases. Mol. Immunol. *128*, 175–187. https://doi.org/10.1016/j.molimm.2020.10.015.
19. Moulds, J.M. (2010). The Knops blood-group system: a review. Immunohematol *26*, 2–7. https://doi.org/10.21307/immunohematology-2019-193.
20. Arora, M., Arora, R., Tiwari, S.C., DAS1, N., and SRIVASTAVA1, L.M. (2001). Expression of complement regulatory proteins on erythrocytes from patients with idiopathic focal segmental glomerulosclerosis. Nephrology *6*, 37–41. https://doi.org/10.1046/j.1440-1797.2001.00020.x.
21. Mathern, D.R., and Heeger, P.S. (2015). Molecules great and small: the complement system. Clin. J. Am. Soc. Nephrol. *10*, 1636–1650. https://doi.org/10.2215/CJN.06230614.
22. Moll, S., Miot, S., Sadallah, S., Gudat, F., Mihatsch, M.J., and Schifferli, J.A. (2001). No complement receptor 1 stumps on podocytes in human glomerulopathies. Kidney Int. *59*, 160–168. https://doi.org/10.1046/j.1523-1755.2001.00476.x.
23. Willows, J., Wood, K., Bourne, H., and Sayer, J.A. (2019). Acquired C1-inhibitor deficiency presenting with nephrotic syndrome. BMJ Case Rep. *12*, e230388. https://doi.org/10.1136/bcr-2019-230388.
24. Sethi, S., Fervenza, F.C., Zhang, Y., Zand, L., Vrana, J.A., Nasr, S.H., Theis, J.D., Dogan, A., and Smith, R.J.H. (2012). C3 glomerulonephritis: clinicopathological findings, complement abnormalities, glomerular proteomic profile, treatment, and follow-up. Kidney Int. *82*, 465–473. https://doi.org/10.1038/ki.2012.212.

25. Salant, D.J., Belok, S., Madaio, M.P., and Couser, W.G. (1980). A new role for complement in experimental membranous nephropathy in rats. J. Clin. Investig. *66*, 1339–1350. https://doi.org/10.1172/JCI109987.

26. Strassheim, D., Renner, B., Panzer, S., Fuquay, R., Kulik, L., Ljubanović, D., Holers, V.M., and Thurman, J.M. (2013). IgM contributes to glomerular injury in FSGS. J. Am. Soc. Nephrol. *24*, 393–406. https://doi.org/10.1681/ASN.2012020187.

27. Thurman, J.M., Wong, M., Renner, B., Frazer-Abel, A., Giclas, P.C., Joy, M.S., Jalal, D., Radeva, M.K., Gassman, J., Gipson, D.S., et al. (2015). Complement Activation in Patients with Focal Segmental Glomerulosclerosis. PLoS One *10*, e0136558. https://doi.org/10.1371/journal.pone.0136558.

28. Tetteh-Quarcoo, P.B., Schmidt, C.Q., Tham, W.-H., Hauhart, R., Mertens, H.D.T., Rowe, A., Atkinson, J.P., Cowman, A.F., Rowe, J.A., and Barlow, P.N. (2012). Lack of evidence from studies of soluble protein fragments that Knops blood group polymorphisms in complement receptor-type 1 are driven by malaria. PLoS One *7*, e34820. https://doi.org/10.1371/journal.pone.0034820.

29. Ghiran, I., Barbashov, S.F., Klickstein, L.B., Tas, S.W., Jensenius, J.C., and Nicholson-Weller, A. (2000). Complement receptor 1/CD35 is a receptor for mannan-binding lectin. J. Exp. Med. *192*, 1797–1808. https://doi.org/10.1084/jem.192.12.1797.

30. Roos, A., Rastaldi, M.P., Calvaresi, N., Oortwijn, B.D., Schlagwein, N., van Gijlswijk-Janssen, D.J., Stahl, G.L., Matsushita, M., Fujita, T., van Kooten, C., and Daha, M.R. (2006). Glomerular activation of the lectin pathway of complement in IgA nephropathy is associated with more severe renal disease. J. Am. Soc. Nephrol. *17*, 1724–1734. https://doi.org/10.1681/ASN.2005090923.

31. Lhotta, K., Würzner, R., and König, P. (1999). Glomerular deposition of mannose-binding lectin in human glomerulonephritis. Nephrol. Dial. Transplant. *14*, 881–886. https://doi.org/10.1093/ndt/14.4.881.

32. Tsutsumi, A., Takahashi, R., and Sumida, T. (2005). Mannose binding lectin: genetics and autoimmune disease. Autoimmun. Rev. *4*, 364–372. https://doi.org/10.1016/j.autrev.2005.02.004.

33. Da Silva, R.P., Hall, B.F., Joiner, K.A., and Sacks, D.L. (1989). CR1, the C3b receptor, mediates binding of infective Leishmania major metacyclic promastigotes to human macrophages. J. Immunol. *143*, 617–622.

34. Payne, N.R., and Horwitz, M.A. (1987). Phagocytosis of Legionella pneumophila is mediated by human monocyte complement receptors. J. Exp. Med. *166*, 1377–1389. https://doi.org/10.1084/jem.166.5.1377.

35. Robledo, S., Wozencraft, A., Valencia, A.Z., and Saravia, N. (1994). Human monocyte infection by Leishmania (Viannia) panamensis. Role of complement receptors and correlation of susceptibility in vitro with clinical phenotype. J. Immunol. *152*, 1265–1276.

36. Schlesinger, L.S., Bellinger-Kawahara, C.G., Payne, N.R., and Horwitz, M.A. (1990). Phagocytosis of Mycobacterium tuberculosis is mediated by human monocyte complement receptors and complement component C3. J. Immunol. *144*, 2771–2780.

37. Tham, W.-H., Wilson, D.W., Lopaticki, S., Schmidt, C.Q., Tetteh-Quarcoo, P.B., Barlow, P.N., Richard, D., Corbin, J.E., Beeson, J.G., and Cowman, A.F. (2010). Complement receptor 1 is the host erythrocyte receptor for Plasmodium falciparum PfRh4 invasion ligand. Proc. Natl. Acad. Sci. USA *107*, 17327–17332. https://doi.org/10.1073/pnas.1008151107.

38. Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinformatics *43*, bi1110s43. https://doi.org/10.1002/0471250953.bi1110s43.

39. Hail Team (2021). Hail 0.2. https://github.com/hail-is/hail.

40. Athey, T.L., Liu, T., Pedigo, B.D., and Vogelstein, J.T. (2019). AutoGMM: Automatic and Hierarchical Gaussian Mixture Modeling in Python. Preprint at arXiv. https://doi.org/10.48550/arxiv.1909.02688.

41. Ho, D.E., Imai, K., King, G., and Stuart, E.A. (2011). MatchIt : Nonparametric Preprocessing for Parametric Causal Inference. J. Stat. Softw. *42*, 1–28. https://doi.org/10.18637/jss.v042.i08.

42. Willer, C.J., Li, Y., and Abecasis, G.R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics *26*, 2190–2191. https://doi.org/10.1093/bioinformatics/btq340.

43. Karssen, L.C., van Duijn, C.M., and Aulchenko, Y.S. (2016). The GenABEL Project for statistical genomics. F1000Res. *5*, 914, [version 1; peer review: 2 approved]. https://doi.org/10.12688/f1000research.8733.1.

44. Sanchez, G. (2012). AssotesteR: Statistical Tests for Genetic Association Studies. R package version 0.1.9. https://github.com/gastonstat/AssotesteR.

45. Zhbannikov, I.Y. (2015). Vartools: A toolkit for Rare Variant Analysis. R package version 1.2.0. https://izhbannikov.github.io/vartools/.

46. Szpiech, Z.A., and Hernandez, R.D. (2014). selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. Mol. Biol. Evol. *31*, 2824–2827. https://doi.org/10.1093/molbev/msu211.

47. Artomov, M., Loboda, A.A., Artyomov, M.N., and Daly, M.J. (2024). Public platform with 39,472 exome control samples enables association studies without genotype sharing. Nat. Genet. *56*, 327–335. https://doi.org/10.1038/s41588-023-01637-y.

48. Wang, M., Chun, J., Genovese, G., Knob, A.U., Benjamin, A., Wilkins, M.S., Friedman, D.J., Appel, G.B., Lifton, R.P., Mane, S., and Pollak, M.R. (2019). Contributions of rare gene variants to familial and sporadic FSGS. J. Am. Soc. Nephrol. *30*, 1625–1640. https://doi.org/10.1681/ASN.2019020152.

# STAR★METHODS

## KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| **Biological samples** | | |
| FSGS patient DNA samples | This study | N/A |
| Control DNA samples | dbGAP | Table S1 |
| **Software and algorithms** | | |
| Custom code | https://github.com/rostkick/fsgs_meta | https://doi.org/10.5281/zenodo.14883396 |
| GATK | Van der Auwera et al.[38] | https://gatk.broadinstitute.org/hc/en-us |
| Hail 0.2 | Hail Team[39] | https://github.com/hail-is/hail |
| AutoGMM | Athey et al.[40] | https://github.com/tathey1/autogmm |
| MatchIt | Ho et al.[41] | https://cran.r-project.org/web/packages/MatchIt/vignettes/MatchIt.html |
| METAL | Willer et al.[42] | https://github.com/statgen/METAL |
| GenABEL | Karssen et al.[43] | https://github.com/GenABEL-Project |
| AssotesteR | Sanchez[44] | https://github.com/gastonstat/AssotesteR |
| vartools | Zhbannikov[45] | https://izhbannikov.github.io/vartools/ |
| selscan | Szpiech and Hernandez[46] | https://github.com/szpiech/selscan |
| **Deposited data** | | |
| Raw sequencing data | This study | N/A |
| Control cohort data | dbGAP | Table S1 |

## EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

### Sample size and how subjects/samples were allocated to experimental groups

DNA samples were obtained from patients participating in a multicenter NIDDK study of biopsy-confirmed FSGS[6] and from patients diagnosed at Washington University. The research protocols were approved in advance and all subjects provided informed consent or assent. As all samples were de-identified, the Washington University in St. Louis Institutional Review Board (IRB) deemed that these studies did not require IRB approval. A total of 726 samples were collected in a multicenter NIH study of biopsy-confirmed FSGS[6] and from patients similarly diagnosed at Washington University, the latter inherited from Yu et al.[7] Control dataset was obtained from dbGAP datasets listed in Table S1 – population based datasets without known kidney disease.

A limitation of this study is the inability to provide detailed demographic information on study participants, including specific data on genotype, age/stage of development and sex. Together with the uneven subpopulation composition of our cohort, this limitation affects the generalisability of our findings and prevents us from conducting more fine-grained analyses within global population clusters (e.g., analysis of HLA-types). However, the impact of sex on the study results is not expected to be significant due to the lack of sex differences in allele frequency of the pathogenic variant. Readers should interpret our current results with these demographic data limitations in mind.

## METHOD DETAILS

Genetic data for cases were obtained using a "podocyte exome" sequencing approach, consisting of a panel of 2,482 genes, as described in Yu et al.[7] Of the selected genes, mutations in five genes cause familial FSGS and 200 genes are functionally related to these five. Additional genes were selected based on expression profiles, as 677 genes are highly expressed in human microdissected glomeruli, and 1600 genes are human orthologs of highly-expressed mouse podocyte genes (Figure 1A). In the present study using this panel, only the 2,482 genes constituting the "podocyte exome" were analyzed, with 1.12% of the sequencing capture nucleotides located in non-coding DNA.

The raw data files with sequencing reads (FASTQ files) were obtained for control subjects from dbGAP general population cohorts not ascertained for kidney disease history (Table S1). We extracted the regions sequenced in the podocyte exome from the full-exome data of the control cohort. There were 333,239 variants in the raw dataset of 726 cases and 13,994 controls. We performed joint variant calling according to GATK best practices,[38] to construct a case-control dataset. To confirm the absence of insufficient coverage biases between cases and controls, we crossed the intervals common to both panels and then we calculated the fraction of sequencing intervals that were well-covered (>10X) in cases and controls; this was found to be 89% for both groups (Figure S1). In this calculation, only variants that passed initial GATK hard-filtering[38] were used. Next, the case-control dataset was subjected to quality filtration using the Hail 0.2 open source software library[39] (Figure S2).

The final dataset included 577 cases (including 179 from Yu et al.[5]), and 13,994 controls (including 378 from Yu et al. [5]), and 131,179 variants. The drop-out rate was 60.64% for variants and 6.85% for samples. The high drop-out rate for variants is explained by exome sequencing being joined with panel sequencing in a single dataset, requiring exclusion of many variants detected in the exome sequences of controls and not sequenced in the case panel, due to broader DNA region coverage in controls.

To account for population stratification, a joint principal components analysis (PCA) was performed for case and control genotypes. Uncorrelated common variants (linkage disequilibrium pruning: $r^2 < 0.9$; minor allele frequency – MAF>0.05) were used to cluster the samples in PCA space. To reduce the risk of false associations in the rare variant analysis due to population stratification, we used principal components to subset the control cohort to match the genetic background of cases.

First, we partitioned the dataset into clusters representing global population groups. Given that genetic ancestry can be viewed as a polygenic trait influenced by many genetic variants, it is reasonable to assume that the distribution of genetic ancestry follows a multivariate Gaussian distribution.[47] This is particularly true for both continental and small-scale ancestry. As an implementation of such an algorithm, we performed clustering using AutoGMM package (10 PCs, affinity='euclidean', linkage='all', covariance_type='all').[40] The data were stratified into eight clusters according to the Gaussian mixture model. Agnostic clusters modeled by the AutoGMM algorithm were mapped to known clusters of the 1000 Genomes Project and labeled accordingly (Figure S3). Two minor clusters of individuals belonging to South Asian and East Asian populations and admixed ancestry were excluded from the analysis due to small case count in each cluster, resulting in low statistical power. Of the six clusters retained for further study, three included individuals of European descent and reflected different local-population origins; these minor-clusters were combined into a single major-European cluster. The fourth cluster represented the individuals with African ancestry. Two other clusters belonged to the Latin-American population. After filtering out the low power clusters, the dataset had 551 cases and 11,591 controls (Figure S4).

Further case-control matching was conducted using the MatchIt package.[41] The top three principal components with the greatest contribution to the genetic ancestry imbalance between cases and controls were used for each population (Figure S5). 16 cases were excluded from further consideration because it was not possible to select appropriate population controls for them. The final dataset included 358 cases and 1,488 controls for the European-American cluster, 125 cases and 137 controls for the African-American cluster and 52 cases and 288 controls for the Latin-American cluster (Figure S4).

## QUANTIFICATION AND STATISTICAL ANALYSIS

The majority of genetic analyses were performed using Hail framework (v2.0).[39] Statistical analyses were performed using multiple specialized software tools. Meta-analysis of genetic associations was conducted using METAL software (version from 2020-05-05) for case-control study comprising 726 FSGS cases and 13,994 controls across diverse ethnic populations.[42] Genomic control inflation factors ($\lambda$GC) were calculated using GenABEL package (v1.3.8).[43] For rare variant analysis of approximately 2,500 podocyte-expressed genes, we employed multiple statistical tests including C-Alpha, ASUM, and WSS implemented in AssotesteR (v0.1.9), while KBAC testing was performed using vartools (v1.2.0).[44,45] Selection signatures were assessed using the integrated haplotype score (iHS) calculated with selscan (v2.0.3).[46]

The Weir and Cockerham F-statistic for analysis of population structure showed that the European-American cluster and the African-American cluster were sufficiently isolated from each other (weighted mean fixation index Fst=0.0878 for case cohorts), demonstrating distinct population differences. The Latin-American cluster is also sufficiently isolated from the other clusters (weighted mean fixation index Fst=0.0148). The power analysis for the European-American dataset showed many-fold power superiority over previous FSGS cohort studies,[48] and that an exponential power increase threshold has been reached with the current number of cases, i.e., a significant power increase could not be achieved with a larger number of controls (Figure S6). The values of significant power for the African and Latin-American clusters are comparable to those of similar studies on European cohorts, but they may not be sufficient to detect genome-wide significant associations of frequent variants because of their small effect size (Figure S7).

We performed an association study using common synonymous variants (gnomAD population specific AF $\geq$ 0.01), as these are unlikely to contribute to a phenotype and yet reflect possible ancestral bias between case and matched controls cohorts. For all three European-American, African-American and Latin-American post-matching datasets we confirmed the absence of systematic bias between cases and controls (Figure 1C).