

1 ***CR1* variants contribute to FSGS susceptibility across multiple populations.**

2 Rostislav Skitchenko^{1,2}, Zora Modrusan³, Alexander Loboda^{1,2,4}, Jeffrey B. Kopp⁵,
3 Cheryl A. Winkler⁶, Alexey Sergushichev¹, Namrata Gupta⁴, Christine Stevens⁴,
4 Mark J. Daly^{4,7,8}, Andrey Shaw^{3,#}, Mykyta Artomov^{4,9,10,#}

5
6 ¹ – ITMO University, St. Petersburg, Russia

7 ² – Almazov National Medical Research Centre, St. Petersburg, Russia

8 ³ – Research Biology, Genentech Inc., San Francisco, CA, USA

9 ⁴ – Broad Institute, Cambridge, MA, USA

10 ⁵ – Kidney Disease Section, Kidney Diseases Branch, National Institute of Diabetes and
11 Digestive and Kidney Diseases (NIDDK), NIH, Bethesda, Maryland, USA

12 ⁶ – Molecular Genetic Epidemiology Studies Section, National Cancer Institute (NCI),
13 Frederick, Maryland, USA

14 ⁷ – Massachusetts General Hospital, Boston, MA, USA

15 ⁸ – Institute for Molecular Medicine Finland, Helsinki, Finland

16 ⁹ – Institute for Genomic Medicine, Nationwide Children’s Hospital, Columbus, OH, USA

17 ¹⁰ – Department of Pediatrics, The Ohio State University College of Medicine, Columbus,
18 OH, USA

19

20

21 # – correspondence: shaw.andrey@gene.com, mykyta.artomov@nationwidechildrens.org

22

23 **A conflict of interest statement is at the end of the manuscript.**

24

25

26 **Abstract**

27

28 Focal segmental glomerulosclerosis (FSGS) is a common cause of nephrotic syndrome
29 with an annual incidence in the United States in African-Americans compared to European-
30 Americans of 24 cases and 5 cases per million, respectively. Among glomerular diseases in
31 Europe and Latin-America, FSGS was the second most frequent diagnosis, and in Asia the fifth.
32 We expand previous efforts in understanding genetics of FSGS by performing a case-control
33 study involving ethnically-diverse groups FSGS cases (726) and a pool of controls (13,994),
34 using panel sequencing of approximately 2,500 podocyte-expressed genes. Through rare variant
35 association tests, we replicated known risk genes – *KANK1*, *COL4A4*, and *APOL1*. A novel
36 significant association was observed for the gene encoding complement receptor 1 (*CR1*). High-
37 risk rare variants in *CR1* in the European-American cohort were commonly observed in Latin-
38 and African-Americans. Therefore, a combined rare and common variant analysis was used to
39 replicate the *CR1* association in non-European populations. The *CR1* risk variant, rs17047661,
40 gives rise to the S11/S12 (R1601G) allele that was previously associated with protection against
41 cerebral malaria. Pleiotropic effects of rs17047661 may explain the difference in allele
42 frequencies across continental ancestries and suggest a possible role for genetically-driven
43 alterations of adaptive immunity in the pathogenesis of FSGS.

44 **Introduction**

45 Focal segmental glomerulosclerosis (FSGS) is a common cause of primary nephrotic
46 syndrome among both adults and children in the USA, and its incidence is increasing.^{1,2} The
47 incidence and prevalence of FSGS are not precisely known due to the requirement of a kidney
48 biopsy for diagnosis and the lack of a central registry. Estimates for incidence range from 1.4 to
49 21 cases per million population.³ The incidence of FSGS in the U.S. is about 4 times higher in
50 African Americans (6.8 patients per million) and 2 times higher in Latin-Americans
51 (3.7 patients per million) compared to European-Americans (1.9 patients per million).⁴

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

62 In this study, a large-scale genetic database was assembled from biopsy-confirmed
63 cases of Focal Segmental Glomerulosclerosis (FSGS) and ethnically matched controls. The study
64 used a panel of approximately ~2,500 genes associated with podocytes, which play a crucial
65 role in the formation and maintenance of the glomerular filtration barrier. The purpose of the
66 study was to investigate the genetic basis of FSGS and identify novel susceptibility genes. This
67 study is a significant extension of previous work conducted by Yu et al⁵, with increased power
68 and a more diverse multi-ethnic cohort with greater sample size.

69

70

71 **Methods**

72 DNA samples were obtained from patients participating in a multicenter NIDDK study of
73 biopsy-confirmed FSGS⁶ and from patients diagnosed at Washington University. The research
74 protocols were approved in advance and all subjects provided informed consent or assent. As
75 all samples were de-identified, the Washington University in St. Louis Institutional Review
76 Board (IRB) deemed that these studies did not require IRB approval. A total of 726 samples
77 were collected in a multicenter NIH study of biopsy-confirmed FSGS⁶ and from patients
78 similarly diagnosed at Washington University, the latter inherited from Yu et al⁵. Genetic data
79 for cases were obtained using a “podocyte exome” sequencing approach, consisting of a panel of
80 2,482 genes, as described in Yu et al⁵. Of the selected genes, mutations in five genes cause
81 familial FSGS and 200 genes are functionally related to these five. Additional genes were
82 selected based on expression profiles, as 677 genes are highly expressed in human micro-
83 dissected glomeruli, and 1600 genes are human orthologs of highly-expressed mouse podocyte
84 genes (**Figure 1A**). In the present study using this panel, only the 2,482 genes constituting the
85 “podocyte exome” were analyzed, with 1.12% of the sequencing capture nucleotides located in
86 non-coding DNA.

87 The raw data files with sequencing reads (FASTQ files) were obtained for control
88 subjects from dbGAP general population cohorts not ascertained for kidney disease history
89 (**Sup. Table S1**). We extracted the regions sequenced in the podocyte exome from the full-
90 exome data of the control cohort. There were 333,239 variants in the raw dataset of 726 cases
91 and 13,994 controls. We performed joint variant calling according to GATK best practices,¹⁰ to
92 construct a case-control dataset. To confirm the absence of insufficient coverage biases
93 between cases and controls, we crossed the intervals common to both panels and then we
94 calculated the fraction of sequencing intervals that were well-covered (>10X) in cases and
95 controls; this was found to be 89% for both groups (**Figure S1**). In this calculation, only

96 variants that passed initial GATK hard-filtering¹⁰ were used. Next, the case-control dataset was
97 subjected to quality filtration using the Hail 0.2 open source software library¹¹ (**Figure S2**).

98 The final data set included 577 cases (including 179 from Yu et al⁵), (including 378 from
99 Yu et al⁵), and 131,179 variants. The drop-out rate was 60.64% for variants and 6.85% for
100 samples. The high drop-out rate for variants is explained by exome sequencing being joined
101 with panel sequencing in a single dataset, requiring exclusion of many variants detected in the
102 exome sequences of controls and not sequenced in the case panel, due to broader DNA region
103 coverage in controls.

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

110 First, we partitioned the dataset into clusters representing global population groups.
111 Clustering was performed using mixed Gaussian models AutoGMM package¹². The data were
112 stratified into eight clusters according to the Gaussian mixture model. Agnostic clusters
113 modeled by the AutoGMM algorithm were mapped to known clusters of the 1000 Genomes
114 Project and labeled accordingly (**Figure S3**). Two minor clusters of individuals belonging to
115 South Asian and East Asian populations and admixed ancestry were excluded from the analysis
116 due to small case count in each cluster, resulting in low statistical power. Of the six clusters
117 retained for further study, three included individuals of European descent and reflected
118 different local-population origins; these minor-clusters were combined into a single major-
119 European cluster. The fourth cluster represented the individuals with African ancestry. Two
120 other clusters belonged to the Latin-American population. After filtering out the low power
121 clusters, the dataset had 551 cases and 11,591 controls (**Figure S4**).

122 Further case-control matching was conducted using the MatchIt¹³ package (**Figure 1B**,
123 **Figure S4**). 16 cases were excluded from further consideration because it was not possible to
124 select appropriate population controls for them. The final dataset included 358 cases and 1,488
125 controls for the European-American cluster, 125 cases and 137 controls for the African-
126 American cluster and 52 cases and 288 controls for the Latin-American cluster (**Figure S4**). The
127 Weir and Cockerham F-statistic for analysis of population structure showed that the European-
128 American cluster and the African-American cluster were sufficiently isolated from each other
129 (weighted mean fixation index $F_{st}=0.0878$ for case cohorts), demonstrating distinct population
130 differences. The Latin-American cluster is also sufficiently isolated from the other clusters
131 (weighted mean fixation index $F_{st}=0.0148$). The power analysis for the European-American
132 dataset showed many-fold power superiority over previous FSGS cohort studies ¹⁴, and that an
133 exponential power increase threshold has been reached with the current number of cases, *i.e.*, a
134 significant power increase could not be achieved with a larger number of controls (**Figure S5**).
135 The values of significant power for the African and Latin-American clusters are comparable to
136 those of similar studies on European cohorts, but they may not be sufficient to detect genome-
137 wide significant associations of frequent variants because of their small effect size (**Figure S6**).

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

143 **Results**

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

147 For each cluster separately, we conducted a variant-based association study using linear
148 regression with no additional covariates. 3,777 variants with missense and PTV (stop_gained,
149 frameshift_variant, splice_acceptor_variant, splice_donor_variant) effects on protein were
150 included in this analysis. In the European-American cluster, two variants were significantly
151 associated with FSGS: (1) rs601314 - $p=8.1 \times 10^{-9}$, reference allele is a minor allele;
152 $OR_{\text{minor allele}}=13.24$ ($CI_{95\%}=[3.996, 56.51]$), missense, *EFEMP2* (EGF-containing fibulin
153 extracellular matrix protein 2); and (2) rs117071588 - $p=4.0 \times 10^{-6}$, alternative allele is a minor
154 allele; $OR_{\text{minor allele}}=11.66$ ($CI_{95\%}=[2.790, 68.35]$), missense, *CCDC82* (coiled-coiled domain
155 containing 82), Significance threshold was determined with Bonferroni correction -
156 $p < 0.05/3,777 = 1.32 \times 10^{-5}$; **Figure S7, Sup. Table S2**). Replication analysis of these two variants
157 in the African-American cohort showed that the rs601314 variant (*EFFMP2*) was not
158 significantly associated with FSGS ($p=0.062$, reference allele is a minor allele,
159 $OR_{\text{minor allele}}=0.7418$, $CI_{95\%}=[0.5370, 1.015]$) and that rs117071588 (*CCDC82*) was absent from
160 the African-American dataset. A similar situation was observed when replication was attempted
161 in the Latin-American cohort: rs601314 was not significantly associated ($p=0.097$, reference
162 allele is a minor allele, $OR_{\text{minor allele}}=2.99$, $CI_{95\%}=[0.6891, 14.76]$), rs117071588 was absent from
163 the data. Despite the significant association statistics of *EFEMP2* and *CCDC82* in the European-
164 American group, the lack of replication of the *EFEMP2* variant FSGS association in the other
165 populations makes this a less robust finding but might serve as a starting point for future
166 studies.

167 The rs601314 in *EFEMP2* (NC_000011.9:g.65636053T>C, ENSP00000434151:p.I259V)
168 and rs117071588 in *CCDC82* (NC_000011.9:g.96117537A>C, ENSP00000278520:p.D125E)
169 variants are defined by most common *in silico* pathogenicity predictors as benign¹⁵. FATHMM
170 classified rs601314 as “damaging” (Fathmm Score Converted = 0.48)¹⁵. The specific predictors
171 of missense deleteriousness classified rs601314 and rs117071588 as benign (MISTIC<0.5)¹⁶.
172 Variant rs601314 affects the von Willebrand factor type A (vWA) domain of *EFFMP2* and
173 variant rs117071588 affects the domain of unknown function (DUF4196) of *CCDC82*¹⁵. Both

174 variants have a missense effect on protein function and come from non conservative parts of
175 proteins (MPC<2.0)¹⁷.

176 We carried out rare variant burden analyses in the European-American cohort, focusing
177 on missense variants and PTV with a population frequency below 0.01. This cutoff was chosen
178 according to the presence of a signal in each of the quartiles of allele frequency distribution in
179 the interval [0; 0.01] (**Figure S8**). A rare variant association study (RVAS) was performed using
180 five tests representing different statistical classes of methods for each gene, in order to identify
181 all potential risk patterns. If most variants are causal and have unidirectional effects, classical
182 burden tests are useful, due to their high power. However, adaptive burden tests are considered
183 more reliable than those using fixed weights or thresholds.¹⁸ In addition, some tests can
184 improve understanding of the results. Tests of variance components are effective when there
185 are variants that both increase and decrease a trait, or have a limited number of causal variants
186 ¹⁸. These tests included Fisher's exact test, C-alpha, adaptive sum statistic (ASUM), weighted
187 sum statistics (WSS) and kernel-based adaptive clustering (KBAC) (**Figure S9**). The resulting p-
188 values were combined using the Simes method for multiple hypothesis testing, which is suitable
189 for merging dependent test statistics (**Sup. Table S3**). The top 10 associated genes included
190 four genes, *APOL1*, *KANK1*, *COL4A4*, *IL36G*, that were previously identified in FSGS association
191 studies. Of these, the top two genes reached significance after Bonferroni correction
192 ($p=0.05/2,482=2.015 \times 10^{-5}$): *APOL1* ($p=1.47 \times 10^{-6}$), a known FSGS susceptibility gene, and *CR1*
193 ($p=1.67 \times 10^{-5}$), a novel candidate gene (**Figure 2A, Sup. Figure S10**).

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

201 We identified rare variants in the European-American cohort that drove the association
202 signals in *APOL1* and *CR1* and four variants, consisting of pair-locus G1 in *APOL1* (rs60910145
203 [NC_000022.10:g.36662034T>G, ENSP00000317674.4:p.Ile400Met] and rs73885319
204 [NC_000022.10:g.36661906A>G, ENSP00000317674.4:p.Ser358Gly]⁹) and two closely-adjacent
205 variants in *CR1* (rs17047661 [NC_000001.10:g.207782889A>G,
206 ENSP00000356016.4:p.Arg2051Gly] and rs17047660 [NC_000001.10:g.207782856A>G,
207 ENSP00000356016.4:p.Lys2040Glu]) were selected for replication in other ancestries
208 (**Figure 2B**). Additionally, we eliminated the possibility of this result being a false positive due
209 to coverage imbalance in the associated genes (**Sup. Figure S11**). *APOL1* variants were
210 successfully replicated (rs73885319: p=0.001779, alternative allele is a minor allele,
211 $OR_{\text{minor allele}}=1.59$, $CI_{95\%}=[1.18, 2.14]$; rs60910145: p=0.002271, alternative allele is a minor
212 allele, $OR_{\text{minor allele}}=1.58$, $CI_{95\%}=[1.17, 2.12]$). The variants in *CR1* were more common and had
213 smaller effect size, therefore, we lacked the statistical power to see the significant replication
214 (rs17047661: p=0.28, reference allele is a minor allele, $OR_{\text{minor allele}}=1.22$, $CI_{95\%}=[0.84, 1.76]$;
215 rs17047660: p=0.74, alternative allele is a minor allele, $OR_{\text{minor allele}}=1.09$, $CI_{95\%}=[0.69, 1.71]$).

216 Second replication was attempted in the Latin-American cohort because the variant
217 frequencies for the variants of interest are more similar to the original European-American
218 cohort. Variants rs17047661 in *CR1* and rs60910145 and rs73885319 in *APOL1* surpassed the
219 replication significance threshold (p=0.05/4=0.0125) (**Sup. Table S5**). Analysis of the
220 statistical power for identified effect sizes in Latin-American and African-American cohorts
221 indicated that the lack of replication in the latter is most likely driven by the statistical power
222 limitations (**Sup. Figure S6**).

223 Meta-estimates of p_{METAL} ¹⁸ for all 4 variants were also calculated for the European-
224 American and Latin-American cohorts: rs60910145 ($p_{\text{METAL}}=9.706 \times 10^{-6}$), rs73885319
225 ($p_{\text{METAL}}=1.420 \times 10^{-4}$), rs17047660 ($p_{\text{METAL}}=0.6359$), rs17047661 ($p_{\text{METAL}}=9.314 \times 10^{-3}$).
226 Interestingly, the variants in *CR1*: rs17047661 and rs17047660 are linked with only three out
227 of four possible haplotypes observed in African subpopulations in 1000 genomes

228 (AFR:YRI+LWK+GWD+MSL+ESN+ASW+ACB: $r^2=0.15$, $D'=1$; YRI: $r^2=0.15$, $D'=1$;
229 ASW: $r^2=0.18$, $D'=1$; ACB: $r^2=0.13$, $D'=1$) and observed in global Latin-American population
230 (AMR:MXL+PUR+CLM+PEL: $r^2=0.46$, $D'=1$).

231 Next, the normalized integral haplotype score (iHS) was directly estimated in the
232 discovery cohort for the variants included in the replication analysis. For the African American
233 cohort, selection pressure analysis confirmed positive selection (iHS<-2.0) for the G1 APOL1
234 alleles: rs73885319 (iHS=-2.16), rs60910145 (iHS=-2.21) and revealed positive selection for
235 rs17047660 (iHS=-2.71) in CR1, whereas no such selection was detected for rs17047661 (iHS=-
236 1.04). The following results were obtained for the Latin American cohort: rs73885319 (iHS=-
237 1.99), rs60910145 (iHS=-2.01), rs17047660 (iHS=-0.142) and rs17047661 (iHS=1.13).

238 Allele frequencies for all variants included in the replication analyses are significantly
239 different between population groups, which can nominally indicate either positive selection or
240 genetic drift. These included the following variants: rs60910145: gnomAD EUR AF=8.6x10⁻⁵,
241 gnomAD AFR AF=0.23, $p=2.2 \times 10^{-16}$; rs73885319: gnomAD EUR AF=1.1x10⁻⁴,
242 gnomAD AFR AF=0.23, $p=2.2 \times 10^{-16}$; rs17047661: gnomAD EUR AF=3.0x10⁻³,
243 gnomAD AFR AF=0.62, $p=2.2 \times 10^{-16}$; and rs17047660: gnomAD EUR AF=1.0x10⁻³,
244 gnomAD AFR AF=0.24, $p=2.2 \times 10^{-16}$). It is likely that iHS estimates can be skewed by complex
245 population structure or demographic variables such as population growth, bottleneck events,
246 and changes in recombination and mutation frequencies. Notably, the allele frequencies within
247 African and African-American populations in 1000 genomes significantly vary for
248 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,
249 AF_{ACB}=0.66).

250 We sought evidence of co-evolving changes in allele frequencies between (a) the G1 and
251 G2 variants in *APOL1* and (b) replicated rs17047661 in *CR1*. We estimated the number of
252 individuals who carry both rs17047661 and either one or both G1 and G2 alleles in the African-
253 American case cohort and compared this with the expectation of random assortment. There

254 were no signs of linkage between these variants ($p=0.93$, binomial test), which suggests that the
255 effects of rs17047661 are fully independent of those of *APOL1*.

256 Most common in silico variant effect predictors do not categorize rs17047661
257 (NC_000001.10:g.207782889A>G, ENSP00000383744:p.R1601G) as pathogenic¹⁵. Exceptions
258 are, for example, PolyPhen2 HVAR and MutationAssessor, which classify rs17047661 as
259 "probably damaging" (Polyphen 2 Hvar Score = 0.964) and "medium functional effect"
260 (Mutationassessor Score Converted = 0.70), respectively¹⁵. The rs17047661 variant of CR1
261 induces a missense effect (p.R1601G) in the protein domains common to secreted complement
262 fixation protein (PHA02927) and complement control protein (CCP) modules, also known as
263 consensus short repeats (SCR) or SUSHI repeats. Specifically, it affects the one of the four Long
264 Homologous Repeats (LHRs), LHR-D, which is responsible for binding C1q, Mannose Binding
265 lectin (MBL) and ficolin^{15,19}. However, the specific predictors of missense deleteriousness
266 classified rs17047661 as benign (MISTIC<0.5)¹⁶, which is likely due to the nonconservative
267 nature of the affected region (MPC<2.0)¹⁷.

268

269 **Discussion**

270 The complement system is a complex network of proteins that play an important role in
271 protecting the body against microbial infections, which are activated either through the
272 classical immune pathway in response to binding to Fc-fragments of IgM or IgG, or through an
273 alternative pathway of non-specific binding to antigens on membranes or to mannose residues
274 through the lectin pathway.²⁰ Each protein in the cascade is activated by proteolysis, splitting
275 the original proenzyme into "a" and "b" structures (the exception is C1, which splits into q, r, s
276 molecules). The large molecule "b" is directly involved in the sequential activation of the
277 complement system and the small molecule "a" is an anaphylatoxin, which causes degranulation
278 of mast cells and chemotaxis of other immune cells such as neutrophils, eosinophils, monocytes,
279 and T lymphocytes. All of these factors have the potential to contribute to either innate immune

280 functions or tissue injury. C3 is the central element of the complement system, which is
281 activated by C3-convertase, a complex composed of the preceding elements of the cascade
282 (classical/lectin pathway: C4bC2b complex, alternative pathway: C3bBb complex). Upon
283 activation, the complement system can affect cells in two ways: (1) by forming the membrane
284 attack complex (MAC, sC5b-9 complex), resulting in osmotic lysis of the targeted cell, and (2)
285 indirect opsonization through the deposition of C3b on the surface of microbes, which
286 facilitates phagocytosis by immune cells.

287 *CR1* acts as a negative complement regulator, reducing C3 activation and tissue
288 deposition, by processing and bounding immune complexes, which then facilitates their
289 transfer to the liver or spleen where macrophages ingest and eliminate them. In both the
290 classical and lectin pathways, *CR1* has decay-activating activity, in that its binding C4b prevents
291 the formation of C3-convertase. In the alternative pathway, *CR1* then acts as a cofactor for the
292 cleavage of active C3b fragments (on C3c and C3dg fragments), significantly reducing
293 deposition of C3b fragments, which could activate C3.²¹ *CR1* significantly reduces C3b
294 deposition by ~80% over the classical pathway, but *CR1* is most potent when the alternative
295 pathway is activated (> 95% reduction in C3b deposition).²¹ By stopping the activation of the
296 complement system at the stage of C3-convertase formation, *CR1* is also indirectly involved in
297 reducing the deposition of sC5b-9, which would have formed afterwards if the complement
298 system had been subsequently activated.

299 *CR1* is expressed by several cell types, including red blood cells (RBC), leukocytes, and
300 among specialized renal cells, *CR1* is localized exclusively on glomerular podocytes. The FSGS
301 related variants in *CR1* - rs17047661 which was replicated in Latin-American descent
302 individuals and its pair - rs17047660 are known as Knops group polymorphisms and are a part
303 of the Red Cell Surface Antigens, which give rise to the Sl2 and McC^b alleles in the Swain-Langley
304 1 and 2 allele pairs (Sl1/Sl2) and McCoy a and b (McC^a/McC^b), respectively. ²². The K1590E
305 substitution and the R1601G substitution in *CR1* are located just 11 amino acids away from
306 each other and are in strong LD for both the 1000 genomes for the global African population

307 ($r^2=0.15$, $D'=1$) and for African-American descent from study data ($r^2=0.24$, $D'=1$). They are
308 located in the Long Homologous Repeats (LHRs), motif D, which is responsible for C1q and MBL
309 binding.²²

310 No genetic alterations in the adaptive immune system have been identified in FSGS
311 patients to date, and the role of immune and complement systems genetic variants remains
312 unknown. Recent studies have described the role of the complement system in various
313 glomerulopathies.²³ Typically, the reduction in *CR1* expression is linked to the severity of the
314 disease, as indicated by the degree of inflammation or tissue damage.²⁴ Autoantibodies directed
315 against kidney-expressed autoantigens or antibody/antigen complexes deposited in the kidney
316 are causative agents of various human kidney diseases. There are cases of C3-mediated
317 inflammation and deposition.^{25,26} Further, inhibition of C3 reduces proteinuria in animal
318 models.²⁷ With regard to FSGS, IgG and C3 deposits are often observed in the affected glomeruli,
319 but the pathogenic role of these deposits remains still unclear, and therapy against the
320 complement system has not been studied in FSGS.²⁸

321 Another study has demonstrated elevated levels of Ba, Bb, C4a, and sC5b-9 in the
322 plasma and urine of patients afflicted with primary FSGS. The detection of these protein
323 deposits in the blood signifies that the complement cascade is activated at a site where
324 fragments can gain entry into the vascular space, likely in the mesangial and sclerotic regions.
325 Conversely, the rise in urine Ba, C4a, and sC5b-9 levels in some patients may reflect
326 complement activation in the glomerulus, or alternatively, activation of filtered proteins in the
327 tubular lumen or downstream in the urinary collection system.²⁹ While C5b-9 complexes
328 generally form directly on the membranes of microorganisms, particularly Gram-negative ones,
329 they can also affect adjacent cells, resulting in "bystander" harm.

330 The relationship between activation of the classical and alternative pathways in
331 response to the presence of Knops antigens in *CR1* has been investigated previously, and a
332 lack of correlation has been noted. However, there is a reasonable discrepancy between

333 the results of serological studies of human samples and those obtained from parts of
334 recombinant proteins. Prior investigations have challenged the conjecture that SI2 and
335 McC^b influence the phenotype by modulating the activity of the cofactor implicated in the
336 cleavage of C3b and C4b or the C1q binding activity.³⁰ Nevertheless, these findings
337 warrant future exploration of the involvement of the lectin pathway in the activation of
338 the complement system.³¹ For example, the contribution of the lectin pathway in the
339 activation of the complement system may play a crucial role in the development of
340 progressive glomerular damage and long-term urinary abnormalities in patients with
341 Henoch-Schönlein purpura nephritis (HSPN).³² In the case of FSGS, the presence of MBL
342 deposits that are focal and segmental has been observed, as reported in previous studies,
343 and this can also result in tissue damage, MBL deficiency can lead to autoimmune
344 diseases.^{33,34}

345 There are many different pathogens that use *CR1* as a receptor for cell entry, for
346 example: *Leishmania major*³⁵, *Legionella pneumophila*³⁶, *Leishmania panamensis*³⁷ and
347 *Mycobacterium tuberculosis*³⁸. It also has been shown that *CR1* is a RBC receptor used by
348 *Plasmodium falciparum* for cell invasion, independent of sialic acid.³⁹ Consistently with this
349 hypothesis, SI2 (rs17047661) was previously associated with protection against cerebral
350 malaria in sub-Saharan African populations, which resulted in much higher prevalence of SI2 in
351 African-descent individuals compared to Europeans.¹⁹ In the study conducted by Opi et al, it
352 was noted by the authors that the opposite concomitant effect of the McC^b (rs17047660) allele
353 on the development of severe malaria was of only nominal borderline significance, despite
354 being under significant strong positive selection ($iHS < -2.0$)¹⁹. At first glance, the association
355 with severe malaria and significant positive McC^b selection are discouraging, but this may
356 explain the linkage of negative and protection haplotypes to each other. Moreover, the authors
357 of the original article tested some haplotypes of SI and McC combinations and found that the
358 combination of SI2/McC^a alleles has an additive protective effect against malaria, which may
359 explain the lack of replication signal for rs17047660 in the African-American descent.¹⁹

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

367

368 **Data availability**

369 FSGS cohort allele frequencies and gene burden rare allele counts are available in the
370 Supplementary Tables. Raw sequencing data for control cohort subjects are available through
371 the dbGAP, accession numbers are available in the Supplementary Table 1.

372

373 **Author contributions**

374 R.S., Z.M., J.B.K., C.A.W., M.J.D., A.S., M.A. designed and conceived the study.

375 R.S., Z.M., A.L., J.B.K., C.A.W., A.L.S., A.S., M.J.D., M.A. analyzed the data

376 J.B.K., C.A.W., M.J.D., A.S., M.A. acquired funding

377 N.G., C.S. managed control cohorts

378 M.J.D., A.S., M.A. supervised the study

379 R.S., J.B.K., C.A.W., M.J.D., A.S., M.A. wrote the manuscript

380 All authors reviewed and approved the manuscript

381

382 **Acknowledgments.** This work was supported in part by the Intramural Research Program,
383 NIDDK, NIH (JBK). This project has been funded in part with federal funds from the National

384 Cancer Institute, National Institutes of Health, under contract 75N91019D00024. The content
385 of this publication does not necessarily reflect the views or policies of the Department of Health
386 and Human Services, nor does mention of trade names, commercial products, or organizations
387 imply endorsement by the U.S. Government. This Research was supported [in part] by the
388 Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research
389 and NIDDK. The authors acknowledge the contributions of the following investigators who
390 recruited subjects for FSGS genetic studies, published in Kopp et al, Nature Genetics, 2008, as
391 DNA from those subjects was used in the present study. These investigators include Kopp JB,
392 Freedman BI, Ahuja TS, Berns JS, Briggs W, Cho ME, Dart RA, Kimmel PL, Korbet SM, Michel DM,
393 Mokrzycki MH, Schelling JR, Simon E, Trachtman H.

394 R.S., A.L.S. were supported by the Ministry of Science and Higher Education of the Russian
395 Federation (Priority 2030 Federal Academic Leadership Program).

396 A.L. was supported by Ministry of Science and Higher Education of the Russian Federation
397 (Agreement # 075-15-2022-301).

398 M.A. was in part supported by Nationwide Foundation Pediatric Innovation Fund.

399

400 **Disclosures**

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

402 Other authors have no competing interests to disclose.

403 **References**

404

- 405 1. O'Shaughnessy, M.M., Hogan, S.L., Thompson, B.D., Coppo, R., Fogo, A.B., and Jennette, J.C.
406 (2018). Glomerular disease frequencies by race, sex and region: results from the International
407 Kidney Biopsy Survey. *Nephrol. Dial. Transplant.* *33*, 661–669.
- 408 2. Dragovic, D., Rosenstock, J.L., Wahl, S.J., Panagopoulos, G., DeVita, M.V., and Michelis, M.F.
409 (2005). Increasing incidence of focal segmental glomerulosclerosis and an examination of
410 demographic patterns. *Clin. Nephrol.* *63*, 1–7.
- 411 3. Shabaka, A., Tato Ribera, A., and Fernández-Juárez, G. (2020). Focal Segmental
412 Glomerulosclerosis: State-of-the-Art and Clinical Perspective. *Nephron* *144*, 413–427.
- 413 4. Kitiyakara, C., Kopp, J.B., and Eggers, P. (2003). Trends in the epidemiology of focal segmental
414 glomerulosclerosis. *Semin. Nephrol.* *23*, 172–182.
- 415 5. Yu, H., Artomov, M., Brähler, S., Stander, M.C., Shamsan, G., Sampson, M.G., White, J.M., Kretzler,
416 M., Miner, J.H., Jain, S., et al. (2016). A role for genetic susceptibility in sporadic focal segmental
417 glomerulosclerosis. *J. Clin. Invest.* *126*, 1067–1078.
- 418 6. Kopp, J.B., Nelson, G.W., Sampath, K., Johnson, R.C., Genovese, G., An, P., Friedman, D., Briggs,
419 W., Dart, R., Korbet, S., et al. (2011). APOL1 genetic variants in focal segmental
420 glomerulosclerosis and HIV-associated nephropathy. *J. Am. Soc. Nephrol.* *22*, 2129–2137.
- 421 7. Kopp, J.B., Smith, M.W., Nelson, G.W., Johnson, R.C., Freedman, B.I., Bowden, D.W., Oleksyk, T.,
422 McKenzie, L.M., Kajiyama, H., Ahuja, T.S., et al. (2008). MYH9 is a major-effect risk gene for focal
423 segmental glomerulosclerosis. *Nat. Genet.* *40*, 1175–1184.
- 424 8. Kao, W.H.L., Klag, M.J., Meoni, L.A., Reich, D., Berthier-Schaad, Y., Li, M., Coresh, J., Patterson, N.,
425 Tandon, A., Powe, N.R., et al. (2008). MYH9 is associated with nondiabetic end-stage renal
426 disease in African Americans. *Nat. Genet.* *40*, 1185–1192.
- 427 9. Genovese, G., Friedman, D.J., Ross, M.D., Lecordier, L., Uzureau, P., Freedman, B.I., Bowden,
428 D.W., Langefeld, C.D., Oleksyk, T.K., Uscinski Knob, A.L., et al. (2010). Association of trypanolytic
429 ApoL1 variants with kidney disease in African Americans. *Science* *329*, 841–845.
- 430 10. Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A.,
431 Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ data to high confidence
432 variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics*
433 *11*, 11.10.1-11.10.33.
- 434 11. Hail Team. Hail 0.2.
- 435 12. Athey, T.L., Liu, T., Pedigo, B.D., and Vogelstein, J.T. (2019). AutoGMM: Automatic and
436 Hierarchical Gaussian Mixture Modeling in Python. ArXiv.
- 437 13. Ho, D.E., Imai, K., King, G., and Stuart, E.A. (2011). MatchIt: Nonparametric Preprocessing
438 for Parametric Causal Inference. *J. Stat. Softw.* *42*.
- 439 14. Wang, M., Chun, J., Genovese, G., Knob, A.U., Benjamin, A., Wilkins, M.S., Friedman, D.J., Appel,
440 G.B., Lifton, R.P., Mane, S., et al. (2019). Contributions of rare gene variants to familial and
441 sporadic FSGS. *J. Am. Soc. Nephrol.* *30*, 1625–1640.
- 442 15. Edmonson, M.N., Patel, A.N., Hedges, D.J., Wang, Z., Rampersaud, E., Kesserwan, C.A., Zhou, X.,

- 443 Liu, Y., Newman, S., Rusch, M.C., et al. (2019). Pediatric Cancer Variant Pathogenicity
444 Information Exchange (PeCanPIE): a cloud-based platform for curating and classifying germline
445 variants. *Genome Res.* 29, 1555–1565.
- 446 16. Chennen, K., Weber, T., Lornage, X., Kress, A., Böhm, J., Thompson, J., Laporte, J., and Poch, O.
447 (2020). MISTIC: A prediction tool to reveal disease-relevant deleterious missense variants. *PLoS*
448 *ONE* 15, e0236962.
- 449 17. Samocha, K.E., Kosmicki, J.A., Karczewski, K.J., O’Donnell-Luria, A.H., Pierce-Hoffman, E.,
450 MacArthur, D.G., Neale, B.M., and Daly, M.J. (2017). Regional missense constraint improves
451 variant deleteriousness prediction. *BioRxiv*.
- 452 18. Lee, S., Abecasis, G.R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis:
453 study designs and statistical tests. *Am. J. Hum. Genet.* 95, 5–23.
- 454 19. Opi, D.H., Swann, O., Macharia, A., Uyoga, S., Band, G., Ndila, C.M., Harrison, E.M., Thera, M.A.,
455 Kone, A.K., Diallo, D.A., et al. (2018). Two complement receptor one alleles have opposing
456 associations with cerebral malaria and interact with α -thalassaemia. *ELife* 7, .
- 457 20. Freiwald, T., and Afzali, B. (2021). Renal diseases and the role of complement: Linking
458 complement to immune effector pathways and therapeutics. *Adv. Immunol.* 152, 1–81.
- 459 21. Poppelaars, F., and Thurman, J.M. (2020). Complement-mediated kidney diseases. *Mol.*
460 *Immunol.* 128, 175–187.
- 461 22. Moulds, J.M. (2010). The Knops blood-group system: a review. *Immunohematology* 26, 2–7.
- 462 23. Mathern, D.R., and Heeger, P.S. (2015). Molecules great and small: the complement system.
463 *Clin. J. Am. Soc. Nephrol.* 10, 1636–1650.
- 464 24. Moll, S., Miot, S., Sadallah, S., Gudat, F., Mihatsch, M.J., and Schifferli, J.A. (2001). No
465 complement receptor 1 stumps on podocytes in human glomerulopathies. *Kidney Int.* 59, 160–
466 168.
- 467 25. Willows, J., Wood, K., Bourne, H., and Sayer, J.A. (2019). Acquired C1-inhibitor deficiency
468 presenting with nephrotic syndrome. *BMJ Case Rep.* 12, .
- 469 26. Sethi, S., Fervenza, F.C., Zhang, Y., Zand, L., Vrana, J.A., Nasr, S.H., Theis, J.D., Dogan, A., and
470 Smith, R.J.H. (2012). C3 glomerulonephritis: clinicopathological findings, complement
471 abnormalities, glomerular proteomic profile, treatment, and follow-up. *Kidney Int.* 82, 465–473.
- 472 27. Salant, D.J., Belok, S., Madaio, M.P., and Couser, W.G. (1980). A new role for complement in
473 experimental membranous nephropathy in rats. *J. Clin. Invest.* 66, 1339–1350.
- 474 28. Strassheim, D., Renner, B., Panzer, S., Fuquay, R., Kulik, L., Ljubanović, D., Holers, V.M., and
475 Thurman, J.M. (2013). IgM contributes to glomerular injury in FSGS. *J. Am. Soc. Nephrol.* 24,
476 393–406.
- 477 29. Thurman, J.M., Wong, M., Renner, B., Frazer-Abel, A., Giclas, P.C., Joy, M.S., Jalal, D., Radeva,
478 M.K., Gassman, J., Gipson, D.S., et al. (2015). Complement Activation in Patients with Focal
479 Segmental Glomerulosclerosis. *PLoS ONE* 10, e0136558.
- 480 30. Tetteh-Quarcoo, P.B., Schmidt, C.Q., Tham, W.-H., Hauhart, R., Mertens, H.D.T., Rowe, A.,
481 Atkinson, J.P., Cowman, A.F., Rowe, J.A., and Barlow, P.N. (2012). Lack of evidence from studies
482 of soluble protein fragments that Knops blood group polymorphisms in complement receptor-
483 type 1 are driven by malaria. *PLoS ONE* 7, e34820.

- 484 31. Ghiran, I., Barbashov, S.F., Klickstein, L.B., Tas, S.W., Jensenius, J.C., and Nicholson-Weller, A.
485 (2000). Complement receptor 1/CD35 is a receptor for mannan-binding lectin. *J. Exp. Med.* *192*,
486 1797–1808.
- 487 32. Roos, A., Rastaldi, M.P., Calvaresi, N., Oortwijn, B.D., Schlagwein, N., van Gijlswijk-Janssen,
488 D.J., Stahl, G.L., Matsushita, M., Fujita, T., van Kooten, C., et al. (2006). Glomerular activation of
489 the lectin pathway of complement in IgA nephropathy is associated with more severe renal
490 disease. *J. Am. Soc. Nephrol.* *17*, 1724–1734.
- 491 33. Lhotta, K., Würzner, R., and König, P. (1999). Glomerular deposition of mannose-binding
492 lectin in human glomerulonephritis. *Nephrol. Dial. Transplant.* *14*, 881–886.
- 493 34. Tsutsumi, A., Takahashi, R., and Sumida, T. (2005). Mannose binding lectin: genetics and
494 autoimmune disease. *Autoimmun. Rev.* *4*, 364–372.
- 495 35. Da Silva, R.P., Hall, B.F., Joiner, K.A., and Sacks, D.L. (1989). CR1, the C3b receptor, mediates
496 binding of infective *Leishmania major* metacyclic promastigotes to human macrophages. *J.*
497 *Immunol.* *143*, 617–622.
- 498 36. Payne, N.R., and Horwitz, M.A. (1987). Phagocytosis of *Legionella pneumophila* is mediated
499 by human monocyte complement receptors. *J. Exp. Med.* *166*, 1377–1389.
- 500 37. Robledo, S., Wozencraft, A., Valencia, A.Z., and Saravia, N. (1994). Human monocyte infection
501 by *Leishmania (Viannia) panamensis*. Role of complement receptors and correlation of
502 susceptibility in vitro with clinical phenotype. *J. Immunol.* *152*, 1265–1276.
- 503 38. Schlesinger, L.S., Bellinger-Kawahara, C.G., Payne, N.R., and Horwitz, M.A. (1990).
504 Phagocytosis of *Mycobacterium tuberculosis* is mediated by human monocyte complement
505 receptors and complement component C3. *J. Immunol.* *144*, 2771–2780.
- 506 39. Tham, W.-H., Wilson, D.W., Lopaticki, S., Schmidt, C.Q., Tetteh-Quarcoop, P.B., Barlow, P.N.,
507 Richard, D., Corbin, J.E., Beeson, J.G., and Cowman, A.F. (2010). Complement receptor 1 is the
508 host erythrocyte receptor for *Plasmodium falciparum* PfRh4 invasion ligand. *Proc Natl Acad Sci*
509 *USA* *107*, 17327–17332.

510

511 **Figure Captions**

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

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

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

520 **(C) Quantile-quantile (QQ)-plots for the association study of the common**
521 **synonymous variants with gnomAD population specific allele frequency ≥ 0.01).**

522 These plots illustrate case-control matching quality for the European-derived (left),
523 African-derived (middle), Latin-American-derived (right) populations. The test lambda-
524 GC (genomic inflation factor) for genome-wide association studies (GWAS) compares the
525 median test statistic against the expected median test statistic under the null hypothesis,
526 in which there is no association for each variant. This test identifies systemic biases and
527 significant associations. Here, most of the points fall along the diagonal, indicating the
528 absence of systemic bias.

529 Abbreviations. EUR, European-Americans. AFR, African-Americans. AMR, Admixture
530 Americans

531

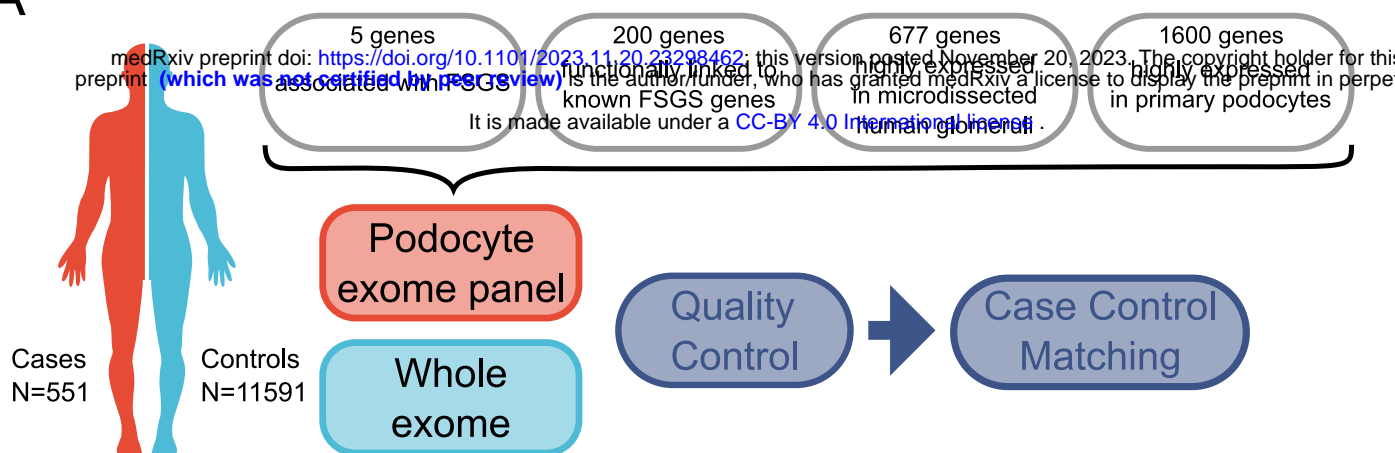
532 **Figure 2. Rare variant association study in the European-American cohort and**
533 **replication of *CR1* and *APOL1* variants in the Admixed American cohort.**

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

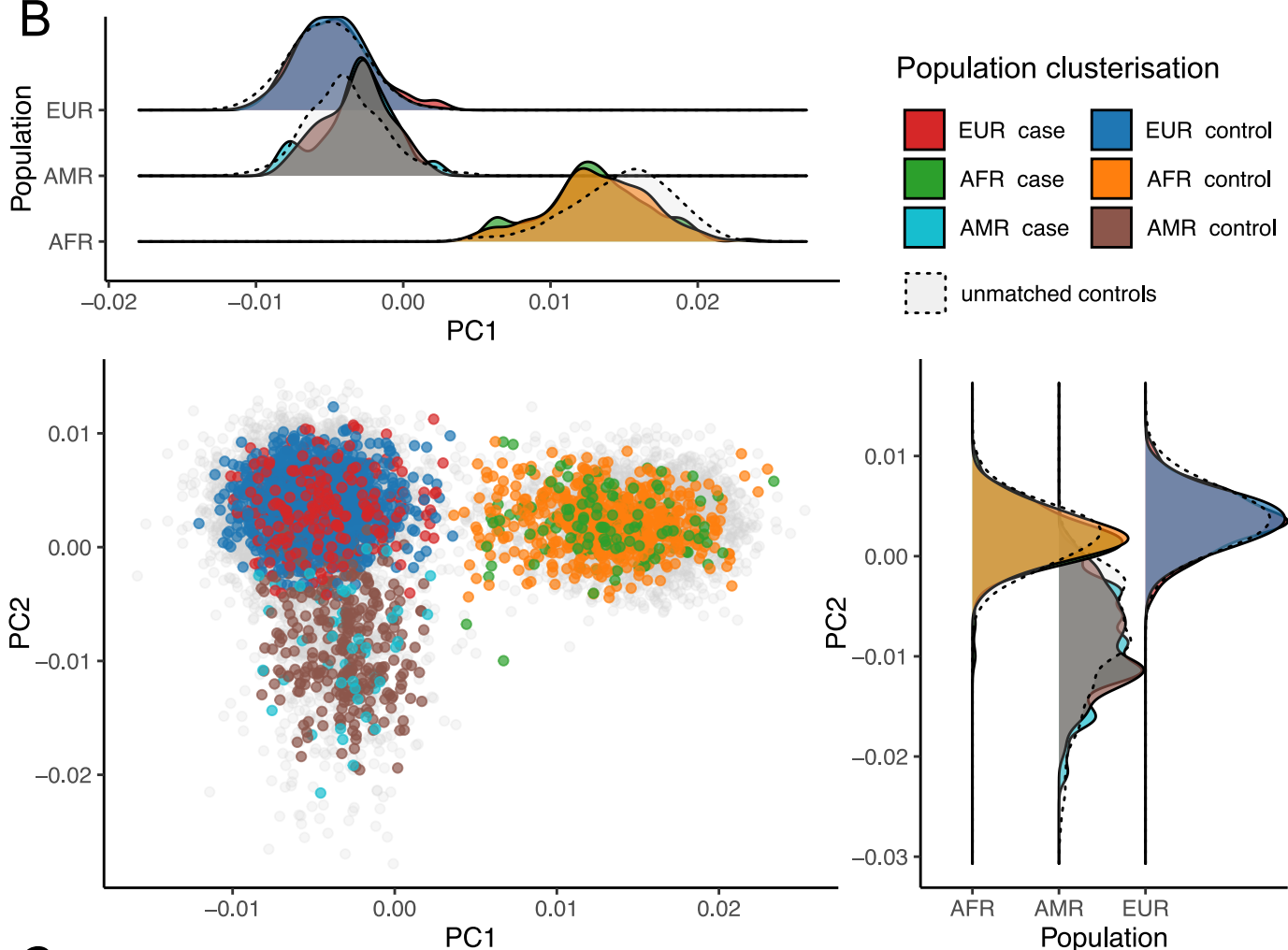
544

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

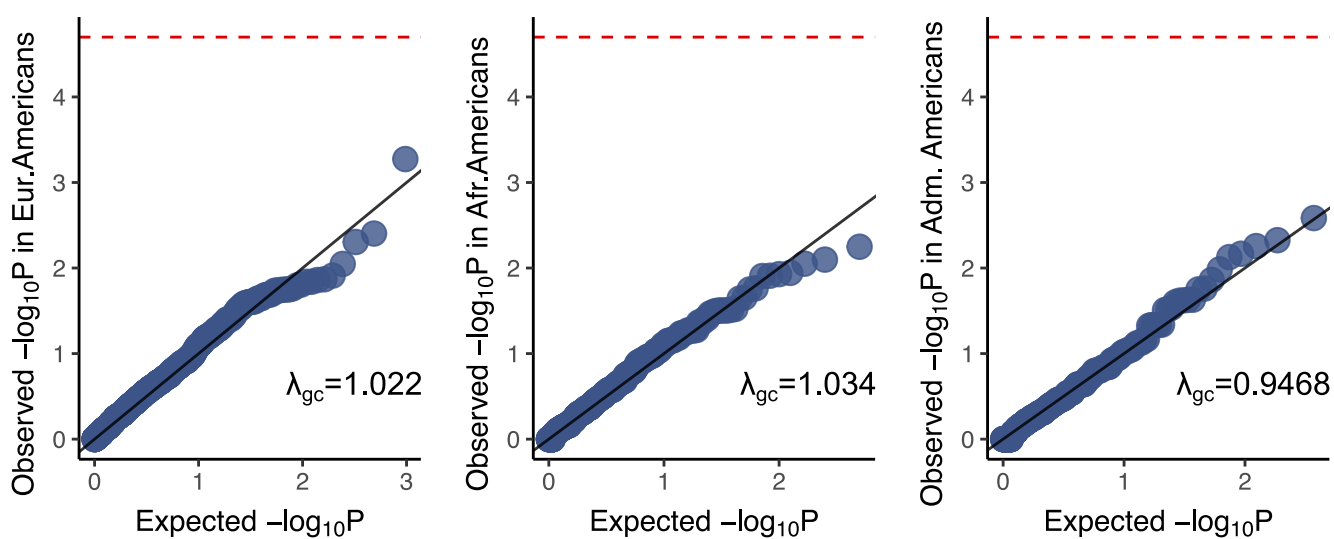
A

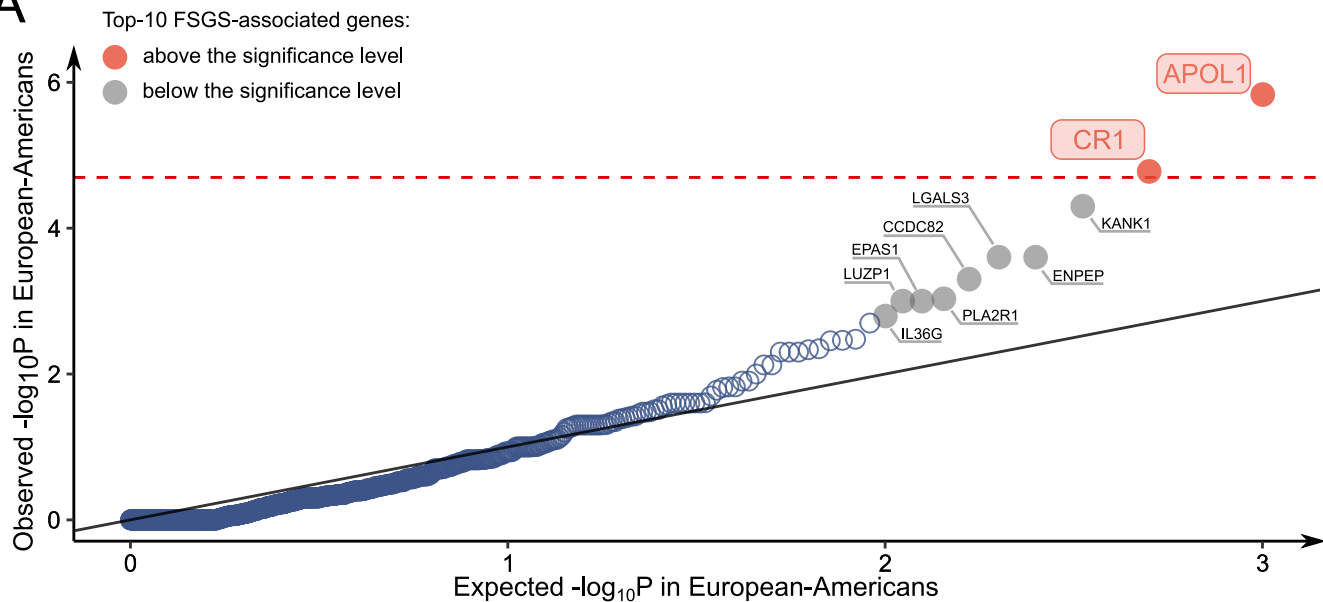


B



C



A**B**