Supplemental information

Table of Contents

SUPPLEMENTAL METHODS

Theoretical Heterozygous SNP Sensitivity (THS)

THS is a quality metric that estimates the theoretical sensitivity to detect heterozygous variants based on coverage distribution and base quality distribution from massively parallel sequencing data.[1,2] Under the assumptions: 1) DNA is diploid; 2) at a HET site with genotype AB, the only possible calls are A and B; 3) there is no reference bias; and 4) coverage distribution *P(n)* and base quality distribution *P(q)* are known and statistically independent, the model of HET detection is based on Bernoulli distribution as:

$$\frac{\binom{n}{m}\left(\frac{1}{2}\right)^n}{\binom{n}{m}\prod_{j=1}^{m} e_j \prod_{j=m+1}^{n}(1-e_j)} > T$$

where, *n* is the depth from the coverage distribution *P(n)*; *m* is the number of true alternate alleles from *m ~ binomial (n, 0.5)* covering the HET site; $e_j = 10^{-qj/10}$ is the probability of error, and q_j is from the base quality distribution *P(q)*.

Variant Pathogenicity Interpretation

Pathogenicity of variant was based on absence in large health populations, presence/enrichment in aHUS patients, and functional association.

1. Large health populations refer to the gnomAD database (138,632 subjects).
2. Presence in aHUS patients is defined as: 1) reported in the literature, 2) reported in an aHUS disease mutation database, or 3) observed in our patient cohort.
3. Enrichment in aHUS patients is determined by association analysis in patients and controls with adjustment for population stratification.
4. Functional association is defined as: 1) well-studied functional changes that contribute to aHUS development, 2) truncating protein where loss of function is a known disease mechanism, 3) known disruption of protein structure (e.g. cysteine-related missense variants in SCRs of *CFH* and *CD46*), or 4) localization in well-defined aHUS-related domains.

Pathogenic is defined as: 1) absent in gnomAD *and* reported at least once in the literature or an aHUS database *and* observed at least once in our patient cohort; OR 2) absent in gnomAD *and* observed at least twice in our patient cohort *and* with a functional impact.

Likely pathogenic is defined as: 1) absent in gnomAD *and* observed at least twice in our patient cohort; OR 2) absent in gnomAD *and* observed at least once in our patient cohort *and* with a functional impact; OR 3) significantly enriched in our patient cohort compare to the corresponding gnomAD population *and* with a functional impact.

Likely benign is defined as: frequency > 0.1% in any gnomAD population *and* not enriched in patients *and* lacking a functional effect.

Benign is defined as: frequency > 1% in any gnomAD population *and* not enriched in patients.

Reference

1. Yossi Farjoun Jon Bloom: Theoretical HET Sensitivity. https://www.broadinstitute.org/files/shared/mia/theoretical_HET_sensitivity.pdf

2. Kylee Degatano, David Benjamin, Jonathan M. Bloom, Maura Costello, Jason Rose, Kathleen TibbeFs, CharloFe Tolonen, Yossi Farjoun: Optimizing Delivered Sequencing Data with a Theoretical Sensitivity to Heterozygous SNPs. ASHG, 2016. http://www.genomics.broadinstitute.org/data-sheets/POS_OptimizingDeliveredSequencingDataTheoreticalSensitivityHeteroSNPs_ASHG_2016.pdf

SUPPLEMENTAL TABLES

Supplement Table S1. Genes sequenced

	Gene	Full Name	RefSeq ID
1	*A2M*	Alpha-2-Macroglobulin	NM_000014
2	*ABCD4*	ATP Binding Cassette Subfamily D Member 4	NM_005050
3	*ADAMTS13*	ADAM Metallopeptidase With Thrombospondin Type 1 Motif 13	NM_139025
4	*ADM*	Adrenomedullin	NM_001124
5	*ADM2*	Adrenomedullin 2	NM_001253845
6	*APCS*	Amyloid P Component, Serum	NM_001639
7	*C1QA*	Complement C1q A Chain	NM_015991
8	*C1QB*	Complement C1q B Chain	NM_000491
9	*C1QC*	Complement C1q C Chain	NM_172369
10	*C1R*	Complement Component 1, R Subcomponent	NM_001733
11	*C1S*	Complement Component 1, S Subcomponent	NM_201442
12	**C2*	Complement C2	NM_000063
13	*C3*	Complement C3	NM_000064
14	*C3AR1*	Complement C3a Receptor 1	NM_004054
15	**C4A*	Complement C4A	NM_007293
16	**C4B*	Complement C4B	NM_000715
17	*C4BPA*	Complement Component 4 Binding Protein Alpha	NM_000715
18	*C4BPB*	Complement Component 4 Binding Protein Beta	NM_000716
19	*C5*	Complement C5	NM_001735
20	*C5AR1*	Complement C5a Receptor 1	NM_001736
21	*C5AR2*	Complement C5a Receptor 2	NM_018485
22	*C6*	Complement C6	NM_000065
23	*C7*	Complement C7	NM_000587
24	*C8A*	Complement C8 Alpha Chain	NM_000562
25	*C8B*	Complement C8 Beta Chain	NM_000066
26	*C8G*	Complement C8 Gamma Chain	NM_000606
27	*C9*	Complement C9	NM_001737
28	*CD46*	Membrane Cofactor Protein	NM_002389
29	*CD55*	Decay Accelerating Factor For Complement	NM_000574
30	*CD59*	Membrane Attack Complex Inhibition Factor	NM_000611
31	*CFB*	Complement Factor B	NM_001710
32	*CFD*	Complement Factor D	NM_001928
33	*CFH*	Complement Factor H	NM_000186
34	**CFHR1*	Complement Factor H Related 1	NM_002113
35	*CFHR2*	Complement Factor H Related 2	NM_005666
36	**CFHR3*	Complement Factor H Related 3	NM_021023
37	*CFHR4*	Complement Factor H Related 4	NM_006684
38	*CFHR5*	Complement Factor H Related 5	NM_030787
39	*CFI*	Complement Factor I	NM_000204
40	*CFP*	Complement Factor Properdin	NM_002621
41	*CLU*	Clusterin	NM_001831
42	*COLEC11*	Collectin Subfamily Member 11	NM_199235
43	*CPN1*	Anaphylatoxin Inactivator	NM_001308
44	**CR1*	Complement C3b/C4b Receptor 1	NM_000651
45	*CR2*	Complement C3b/C4b Receptor 2	NM_001006658
46	*CRP*	C-Reactive Protein	NM_000567
47	*DGKE*	Diacylglycerol Kinase Epsilon	NM_003647
48	*F10*	Coagulation Factor X	NM_000504
49	*F11*	Coagulation Factor XI	NM_000128
50	*F12*	Coagulation Factor XII	NM_000505
51	*F2*	Coagulation Factor II, Thrombin	NM_000506
52	*F2RL2*	Coagulation Factor II Thrombin Receptor Like 2	NM_004101
53	*F3*	Coagulation Factor III, Tissue Factor	NM_001993
54	*F5*	Coagulation Factor V	NM_000130
55	*F7*	Coagulation Factor VII	NM_000131
56	*F8*	Coagulation Factor VIII, Procoagulant Component	NM_000132

57	*F9*	Coagulation Factor IX	NM_000133
58	*FCN1*	Ficolin 1	NM_002003
59	*FCN2*	Ficolin 2	NM_004108
60	*FCN3*	Ficolin 3	NM_003665
61	*FGL2*	Fibrinogen Like 2	NM_006682
62	*IFNG*	Interferon Gamma	NM_000619
63	*INF2*	Inverted Formin, FH2 And WH2 Domain Containing	NM_022489
64	*ITGAM*	Integrin Subunit Alpha M	NM_000632
65	*KLKB1*	Kallikrein B1	NM_000892
66	*LMBRD1*	LMBR1 Domain Containing 1	NM_018368
67	*MAP3K5*	Mitogen-Activated Protein Kinase Kinase Kinase 5	NM_005923
68	*MASP1*	Mannan Binding Lectin Serine Peptidase 1	NM_001879
69	*MASP2*	Mannan Binding Lectin Serine Peptidase 2	NM_006610
70	*MBL2*	Mannose Binding Lectin 2	NM_000242
71	*MBTPS1*	Membrane Bound Transcription Factor Peptidase, Site 1	NM_003791
72	*MMACHC*	Methylmalonic Aciduria CblC Type, With Homocystinuria	NM_015506
73	*MMADHC*	Methylmalonic Aciduria CblD Type, With Homocystinuria	NM_015702
74	*MTR*	5-Methyltetrahydrofolate-Homocysteine Methyltransferase	NM_000254
75	*MTRR*	5-Methyltetrahydrofolate-Homocysteine Methyltransferase Reductase	NM_002454
76	*PHB*	Prohibin	NM_002634
77	*PLAT*	Plasminogen Activator, Tissue	NM_000930
78	*PLAU*	Plasminogen Activator, Urokinase	NM_002658
79	*PLG*	Plasminogen	NM_000301
80	*PROC*	Protein C, Inactivator Of Coagulation Factors Va And VIIIa	NM_000312
81	*PROS1*	Protein S alpha	NM_000313
82	*PTX3*	Pentraxin 3	NM_002852
83	*SERPINA1*	Serpin Family A Member 1	NM_000295
84	*SERPINA5*	Serpin Family A Member 5	NM_000624
85	*SERPINC1*	Serpin Family C Member 1	NM_000488
86	*SERPIND1*	Serpin Family D Member 1	NM_000185
87	*SERPINE1*	Serpin Family E Member 1	NM_000602
88	*SERPINF2*	Serpin Family F Member 2	NM_000934
89	*SERPING1*	Serpin Family G Member 1	NM_000062
90	*THBD*	Thrombomodulin	NM_000361
91	*VSIG4*	V-Set And Immunoglobulin Domain Containing 4	NM_007268
92	*VTN*	Vitronectin	NM_000638
93	*VWF*	von Willebrand Factor	NM_000552

Genes marked with asterisk were not included in burden analysis due to ambiguous read mapping

Supplement Table S2. Sequence ontology terms and functional impact definition

SO accession	SO term	IMPACT
SO:0001893	transcript_ablation	**HIGH**
SO:0001574	splice_acceptor_variant	**HIGH**
SO:0001575	splice_donor_variant	**HIGH**
SO:0001587	stop_gained	**HIGH**
SO:0001589	frameshift_variant	**HIGH**
SO:0001578	stop_lost	**HIGH**
SO:0002012	start_lost	**HIGH**
SO:0001889	transcript_amplification	**HIGH**
SO:0001821	inframe_insertion	**MODERATE**
SO:0001822	inframe_deletion	**MODERATE**
SO:0001583	missense_variant	**MODERATE**
SO:0001818	protein_altering_variant	**MODERATE**
SO:0001630	splice_region_variant	LOW
SO:0001626	incomplete_terminal_codon_variant	LOW
SO:0001567	stop_retained_variant	LOW
SO:0001819	synonymous_variant	LOW
SO:0001580	coding_sequence_variant	MODIFIER
SO:0001620	mature_miRNA_variant	MODIFIER
SO:0001623	5_prime_UTR_variant	MODIFIER
SO:0001624	3_prime_UTR_variant	MODIFIER
SO:0001792	non_coding_transcript_exon_variant	MODIFIER
SO:0001627	intron_variant	MODIFIER
SO:0001621	NMD_transcript_variant	MODIFIER
SO:0001619	non_coding_transcript_variant	MODIFIER
SO:0001631	upstream_gene_variant	MODIFIER
SO:0001632	downstream_gene_variant	MODIFIER
SO:0001895	TFBS_ablation	MODIFIER
SO:0001892	TFBS_amplification	MODIFIER
SO:0001782	TF_binding_site_variant	MODIFIER
SO:0001894	regulatory_region_ablation	MODERATE
SO:0001891	regulatory_region_amplification	MODIFIER
SO:0001907	feature_elongation	MODIFIER
SO:0001566	regulatory_region_variant	MODIFIER
SO:0001906	feature_truncation	MODIFIER
SO:0001628	intergenic_variant	MODIFIER

High and moderate variants were included in the analysis

Source: http://www.ensembl.org/info/genome/variation/predicted_data.html

Supplement Table S3. Association analysis for common variants

Chr	dbSNP	position	Ref	Alt	Gene	Function	MAF aHUS	MAF UI Control	P_1	P_1 adj	OR_1	MAF gnomAD	P_2	OR_2
1	rs9287090	169510380	A	G	*F5*	p.Leu1316Leu	15.58%	28.00%	1.13E-10	6.41E-08	0.47	21.50%	3.85E-05	0.68
1	rs3753396	196695742	G	A	*CFH*	p.Gln672Gln	28.88%	15.33%	4.32E-13	6.12E-10	2.24	16.71%	1.86E-17	2.02
1	rs1065489	196709774	T	G	*CFH*	p.Glu936Asp	28.75%	15.33%	8.82E-13	6.25E-10	2.23	16.84%	8.34E-17	1.99
1	rs3828032	196920178	T	C	*CFHR2*	intronic	39.38%	28.17%	1.87E-07	5.30E-05	1.66	30.00%	1.85E-08	1.52
1	rs11118580	207959070	C	T	*CD46*	intronic	30.00%	21.17%	8.57E-06	1.87E-03	1.60	21.37%	1.29E-08	1.58
3	rs3733001	186938956	T	C	*MASP1*	intronic	30.50%	22.83%	1.42E-04	2.36E-02	1.48	25.18%	6.90E-04	1.30

MAF: minor allele frequency

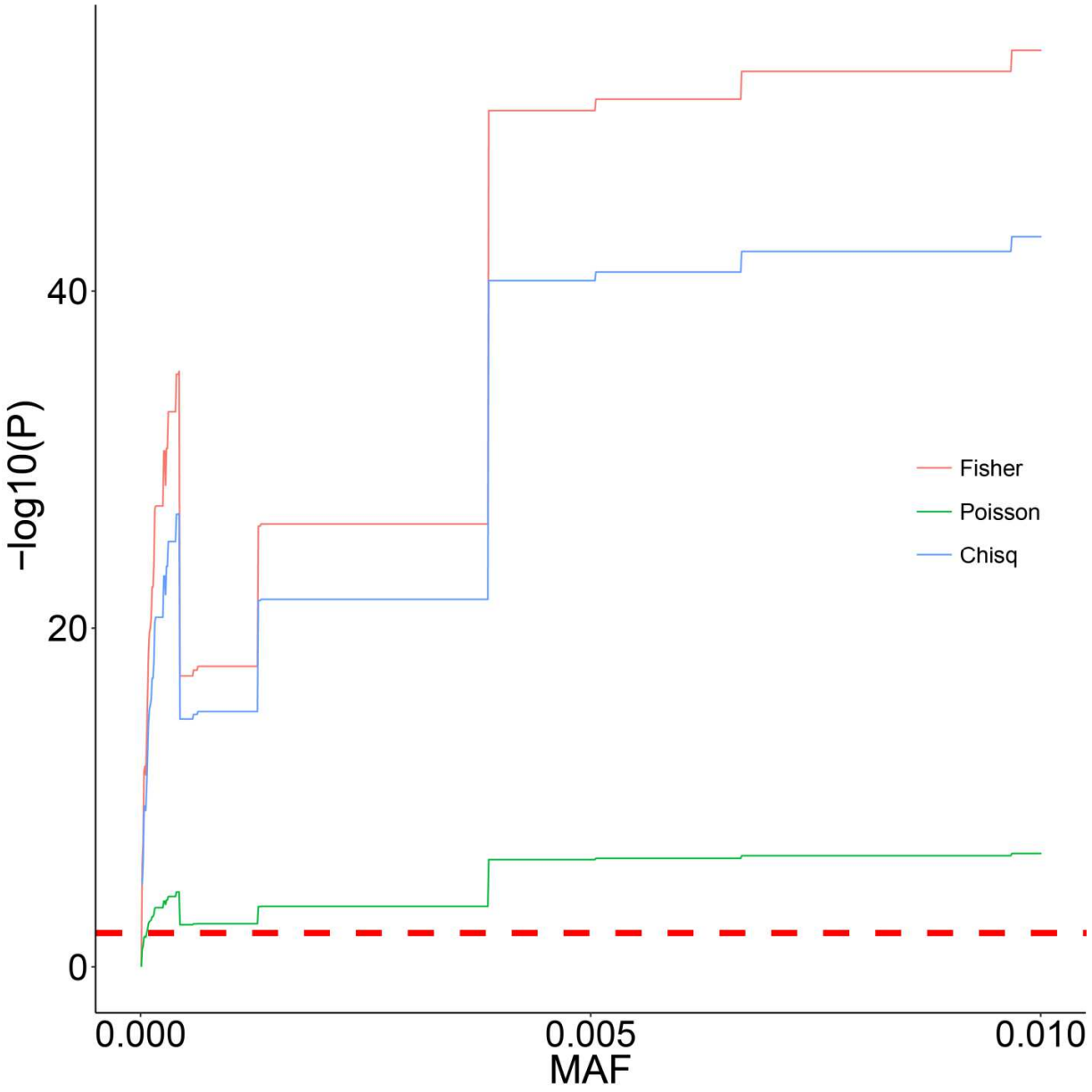
Supplement Table S4. FH-FHRs fusion proteins identified in 400 aHUS patients

Patient ID	Fusion Protein (inferred based on MLPA)
1	FH SCR 1-18 + FHR1 SCR 4-5
2	FH SCR 1-18 + FHR1 SCR 4-5
3	FH SCR 1-18 + FHR1 SCR 4-5
4	FH SCR 1-19 + FHR1 SCR 5
5	FH SCR 1-19 + FHR1 SCR 5
6	FHR1 SCR 1-2 + FH SCR 18-20
7	FHR1 SCR 1-2 + FH SCR 18-20
8	FHR1 SCR 1-3 + FH SCR 19-20
9	FHR3 SCR 1-4 + FHR4 SCR 4-5
10	FHR3 SCR 1-4 + FHR4 SCR 4-5
11	FHR3 SCR 1-4 + FHR4 SCR 4-5
12	FHR3 SCR 1-4 + FHR4 SCR 4-5
13	FHR3 SCR 1-4 + FHR4 SCR 4-5
14	Complex (FH and FHR1 involved)
15	Complex (FH, FHR3, FHR1, FHR4 and FHR2 involved)

Supplement Table S5. Rare variants with 0.1% < MAF (NFE) < 1% identified in *CFH*, *CD46*, *C3*, *CFI* and *CFB* in 400 aHUS patients

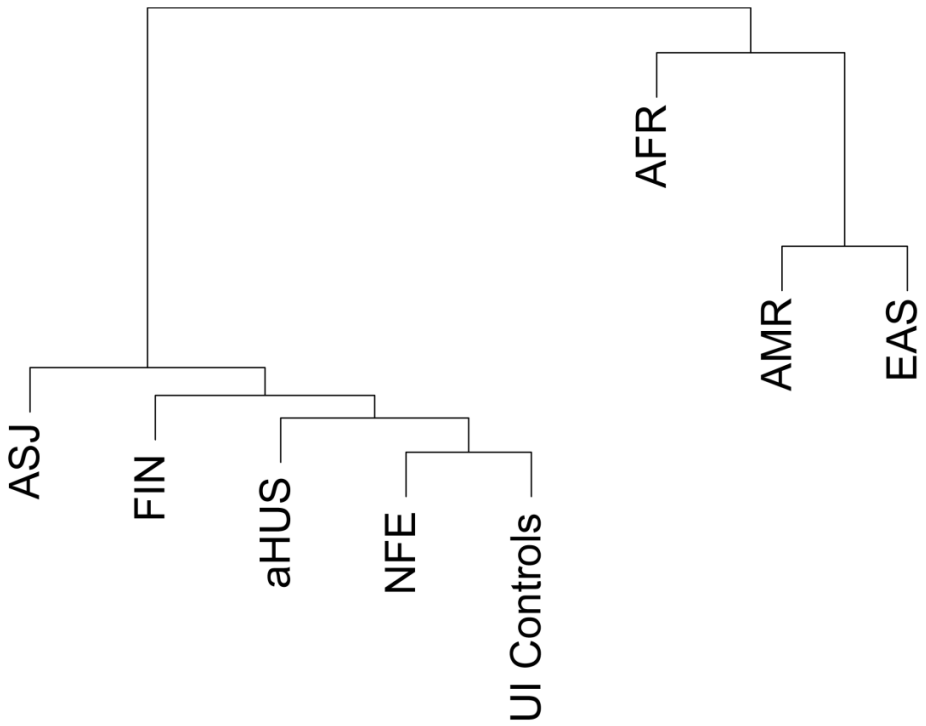
Gene	HGVS	dbSNP	aHUS MAF	Control MAF	NFE MAF	Max MAF	Max Pop	Pathogenicity
CFH	c.2850G>T; p.Gln950His	rs149474608	0.63%	0.75%	0.59%	1.78%	ASJ	B
CFH	c.2867C>T; p.Thr956Met	rs145975787	0.13%	0.33%	0.17%	0.17%	NFE	LB
CFI	c.1657C>T; p.Pro553Ser	rs113460688	0.38%	0.33%	0.27%	0.27%	NFE	LB
CFI	c.1322A>G; .Lys441Arg	rs41278047	0.75%	0.50%	0.24%	4.77%	ASJ	B
CFI	c.782G>A; p.Gly261Asp	rs112534524	0.13%	0.00%	0.19%	0.46%	ASJ	LB
CFI	c.782G>A; p.Gly261Asp	rs112534524	0.13%	0.00%	0.19%	0.46%	ASJ	LB
CFB	c.1697A>C; p.Glu566Ala	rs45484591	3.00%	0.00%	1.00%	2.10%	ASJ	VUS
C3	c.4855A>C; p.Ser1619Arg	rs2230210	0.63%	0.25%	0.22%	0.22%	NFE	VUS
C3	c.2203C>T; p.Arg735Trp	rs117793540	0.13%	0.17%	0.25%	1.24%	ASJ	B
C3	c.463A>C; p.Lys155Gln	rs147859257	0.75%	0.58%	0.54%	0.54%	NFE	LB

SUPPLEMENTAL FIGURES



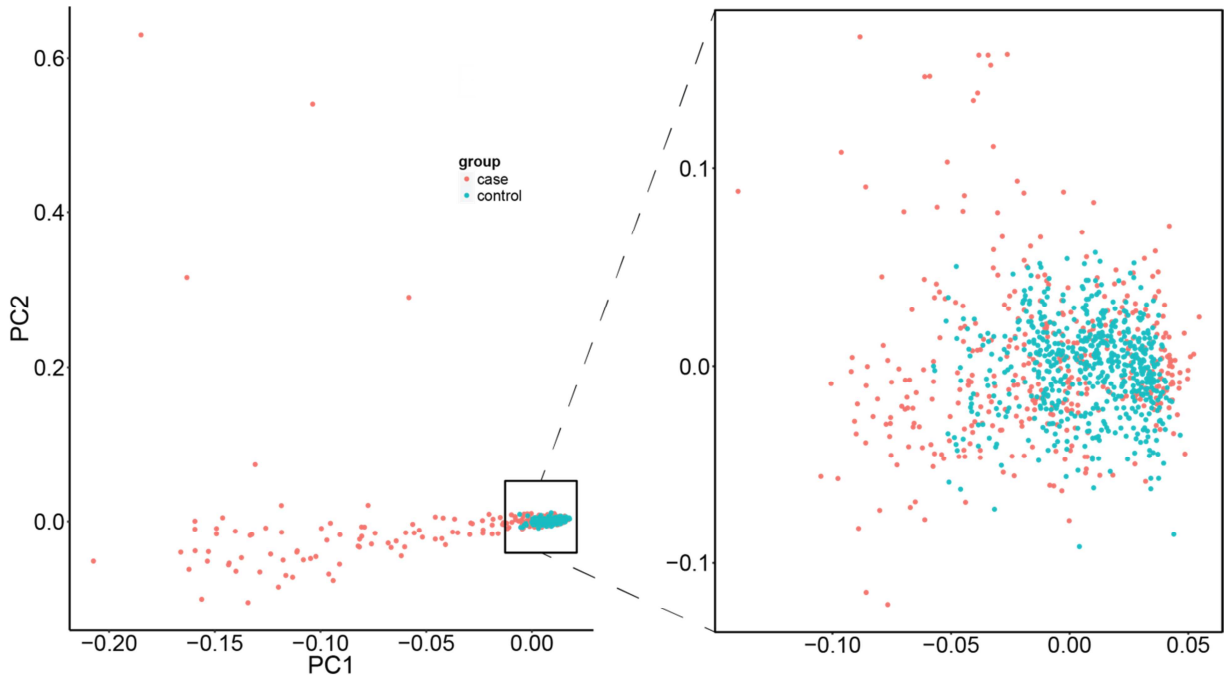
Supplement Figure S1. Rare variants in the *CFH* gene are significantly more abundant in the non-Finnish European (NFE) subpopulation as compared to the Finnish (FIN) subpopulation from gnomAD due to population stratification.

Three algorithms are used to test for enrichment and demonstrate that the modified Poisson exact test is least sensitive to population stratification. P values are shown as curves: red curve, Fisher's exact test; green curve, Poisson exact test; blue curve, Chi-square test; red dashed line, P value=0.05.



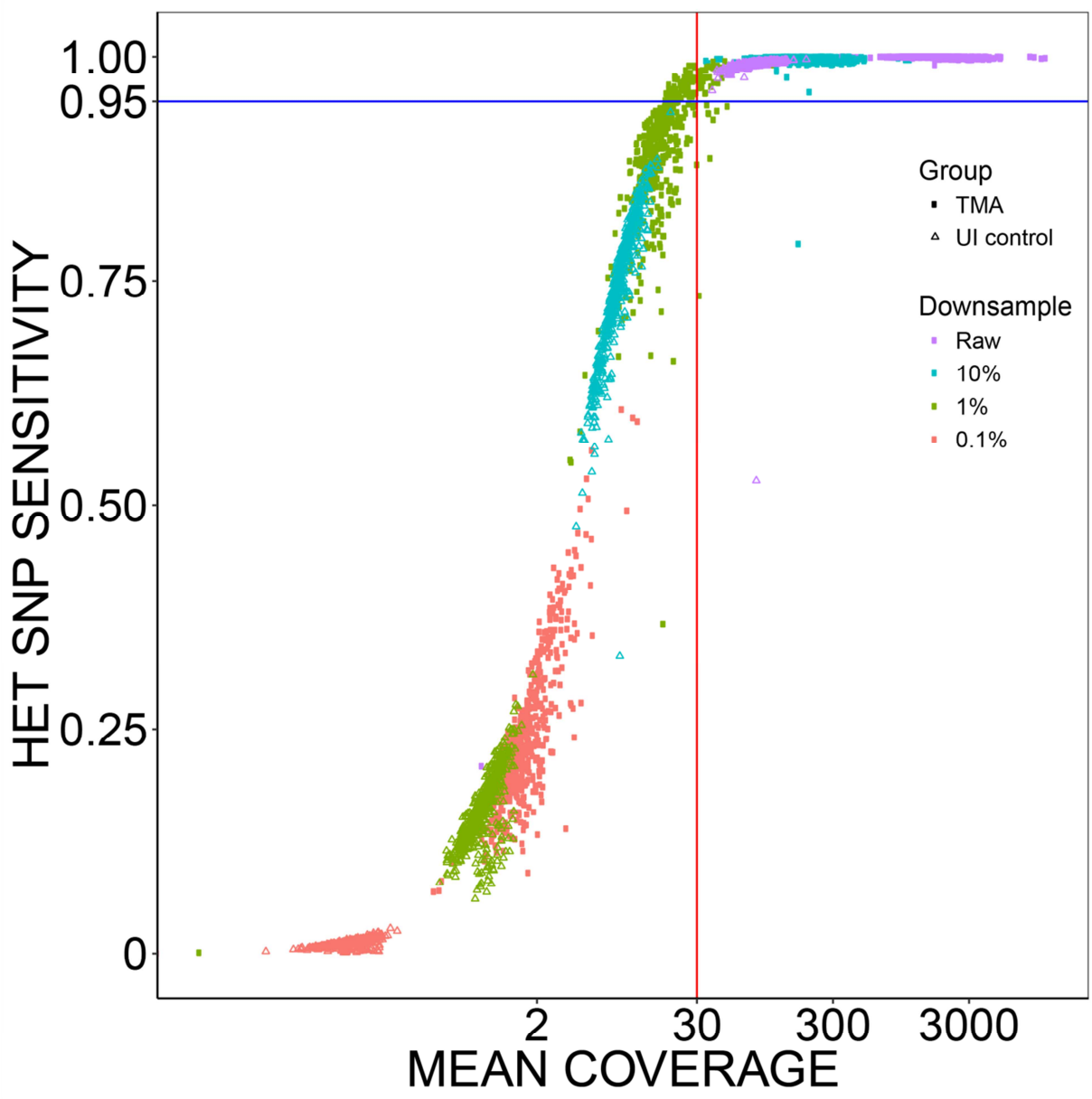
Supplement Figure S2. Cluster analysis shows that UI controls and aHUS cases are mostly similar to NFE.

For clustering, Euclidean distance was calculated based on allele frequencies of variants in each population. Hierarchical cluster analysis was applied using Ward's clustering criterion. Based on this cluster analysis, we used the NFE subpopulation as an additional control.



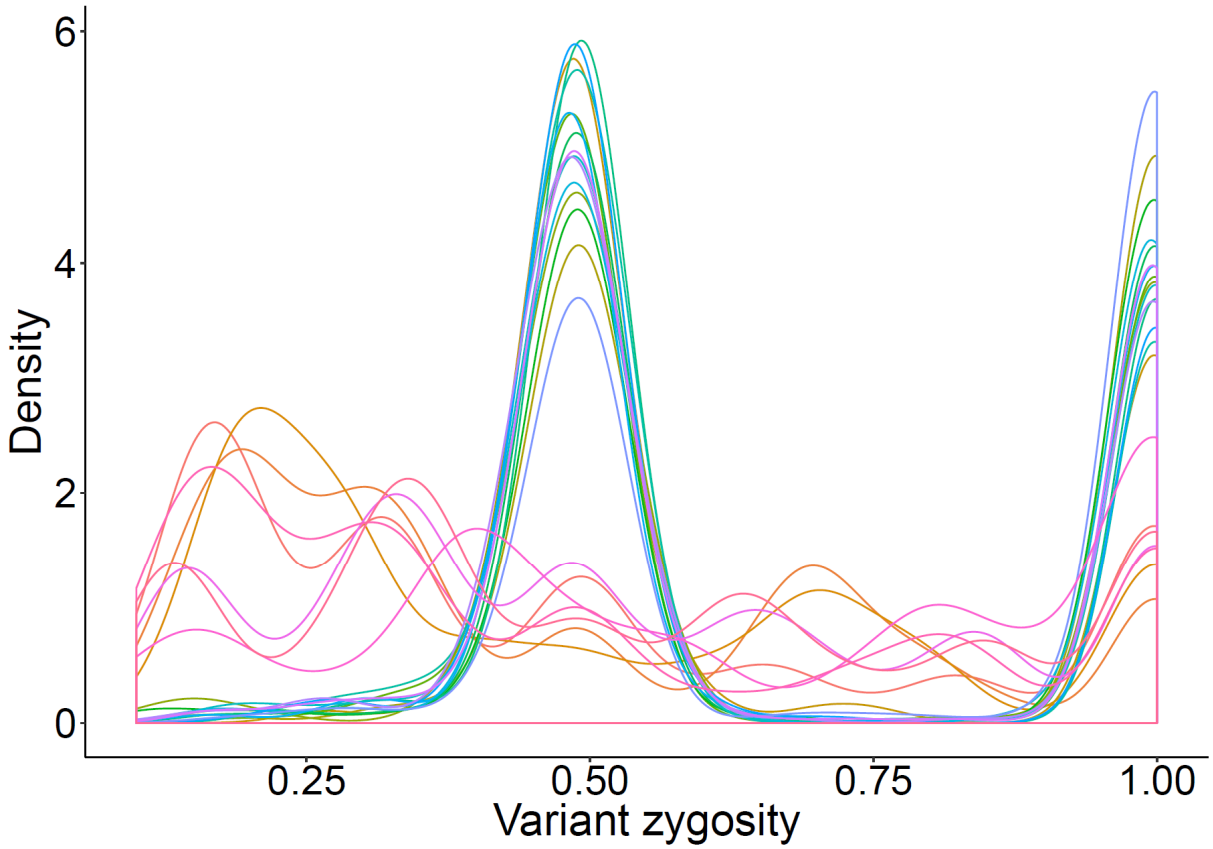
Supplement Figure S3. Population stratification within aHUS cases and UI controls was used to remove outliers.

Left panel, distribution of patients and controls prior to sample removal; right panel, distribution of cases and controls after removing outliers.



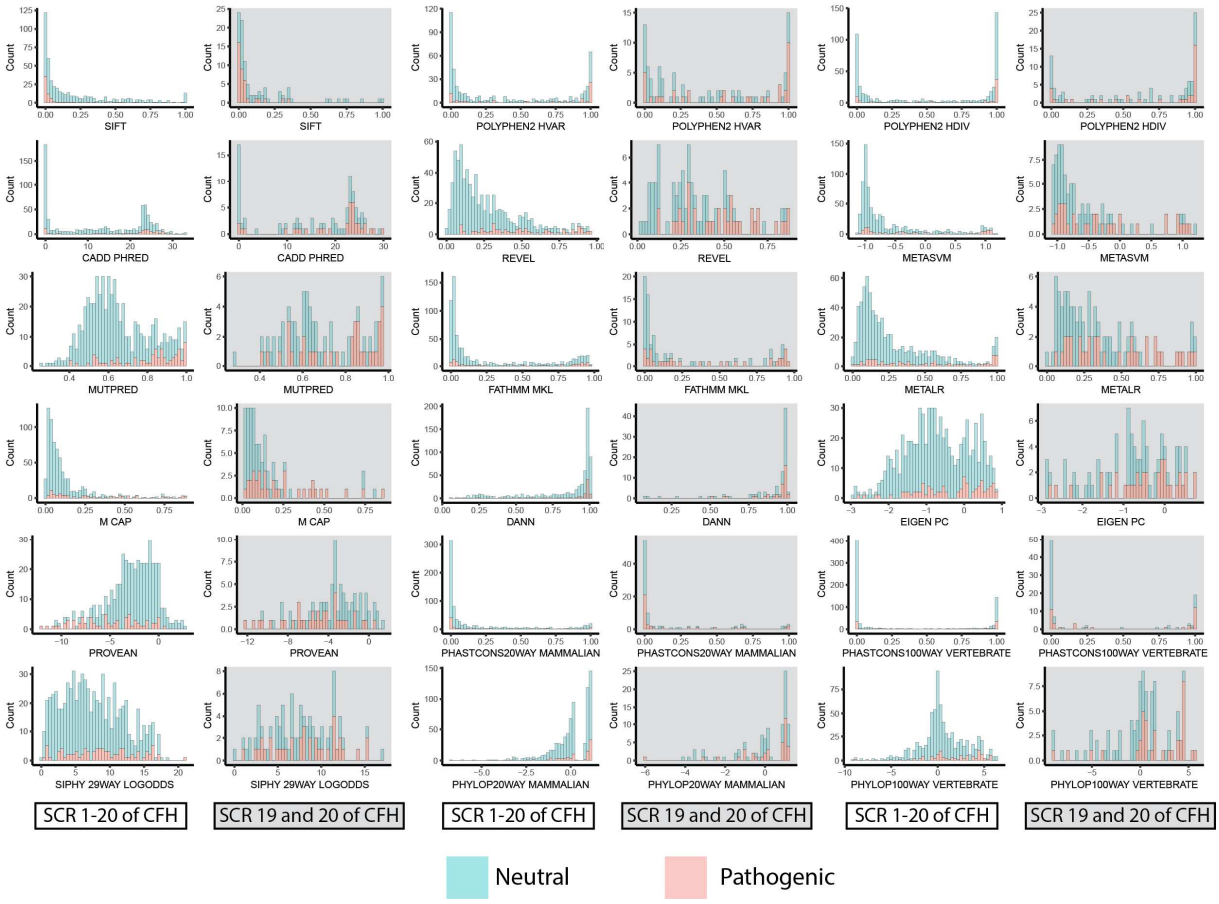
Supplement Figure S4. Mean coverage is correlated with theoretical heterozygous SNP sensitivity (HET SNP sensitivity) in patients and controls.

Random down sampling demonstrated a quick drop of HET SNP sensitivity when mean coverage is below 30X. Most raw sequencing data (purple) from patients (dots) and controls (triangles) is good. Two low quality samples were excluded. Blue line, 95% of HET SNP sensitivity, Red line, 30X of mean coverage.



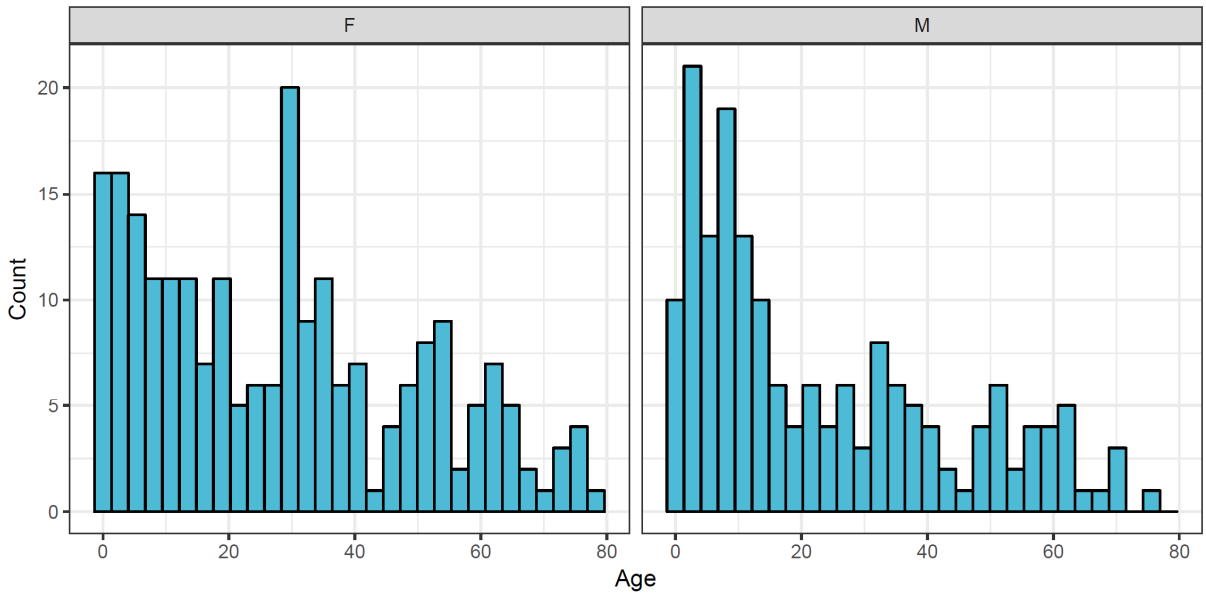
Supplement Figure S5. Six samples with a shifted ratio of ref/alt reads were excluded from the study cohort.

High quality samples are expected to have peaks at 0.5 and 1.0.



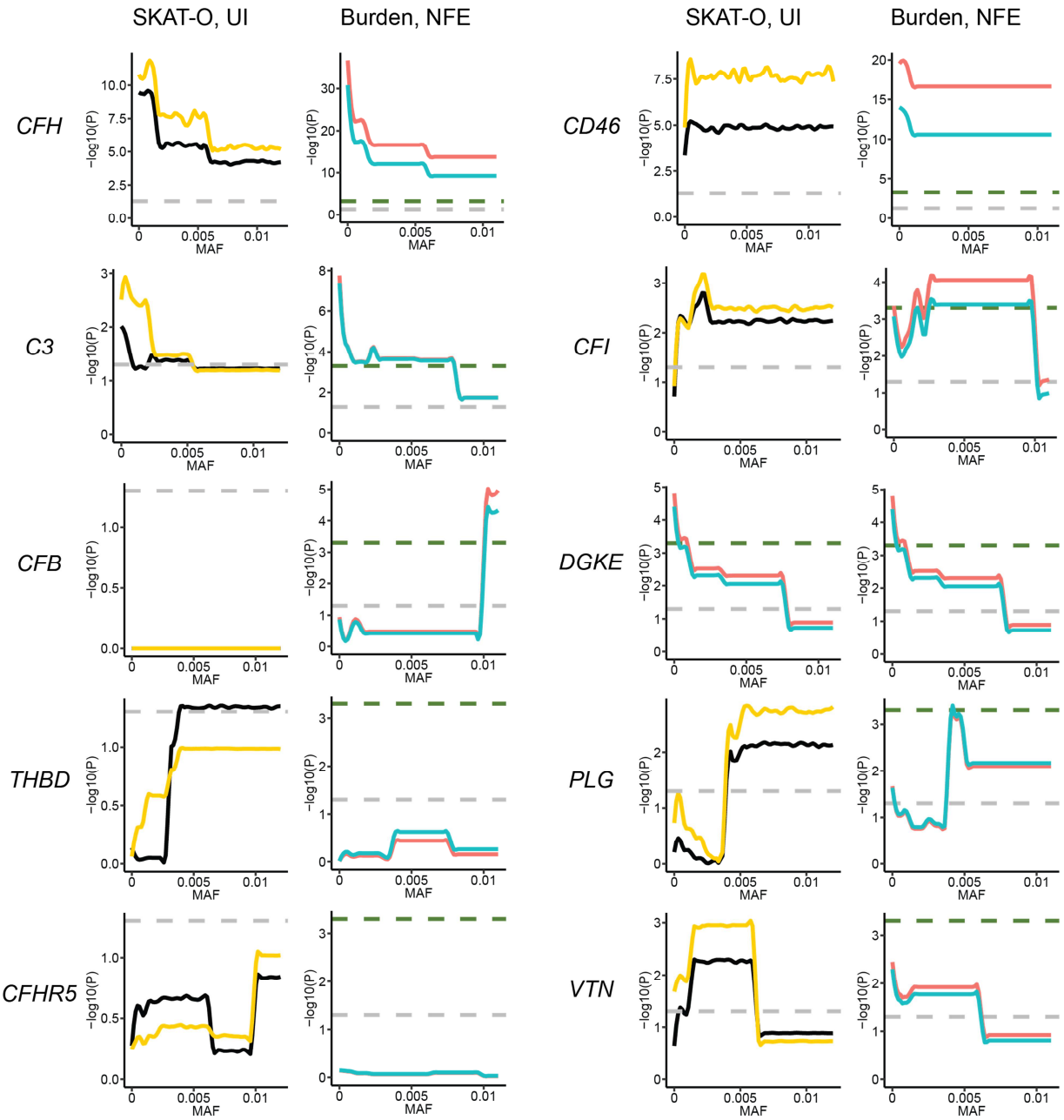
Supplement Figure S6. Prediction score distribution of neutral and pathogenic variants in *CFH*.

Neutral variants from gnomAD and pathogenic variants from the literature were both filtered by MAF < 0.1%. 18 different tools were used to perform in-silico prediction on these variants to compare neutral and pathogenic variants across the entire gene (white background) or restricting the analysis to SCR19-20 (grey background). Note the heavily mixed distribution of neutral and pathogenic variants for all tools, indicating their ineffectiveness in predicting variant effect in *CFH* for aHUS.



Supplement Figure S7. Median age of female patients is significantly lower than that of male patients

Female: left panel, 27.3 years; Male: right panel, 14.8 years; Mann-Whitney U testP = 0.011



Supplement Figure S8: Enrichment of ultra-rare variants 'contaminates' the result of the association analysis when MAF thresholds are set higher.

The minor allele frequency threshold (cut off) was increased in a stepwise fashion to select variants for the analyses. Sets of p values are shown as curves: black curve, SKAT-O test adjusting for population stratification in UI controls; yellow curve, SKAT-O test without adjusting in UI controls; red curve, Fisher's exact test in NEF controls; blue curve, Poisson exact test in NEF controls; grey dashed line, P<0.05; green dashed line, P<0.0005.