Nat. Genet. doi:10.1038/ng.2758; published online 15 September 2013; corrected online 9 October 2013

Identification of a rare coding variant in complement 3 associated with age-related macular degeneration

Xiaowei Zhan, David E Larson, Chaolong Wang, Daniel C Koboldt, Yuri V Sergeev, Robert S Fulton, Lucinda L Fulton, Catrina C Fronick, Kari E Branham, Jennifer Bragg-Gresham, Goo Jun, Youna Hu, Hyun Min Kang, Dajiang Liu, Mohammad Othman, Matthew Brooks, Rinki Ratnapriya, Alexis Boleda, Felix Grassmann, Claudia von Strachwitz, Lana M Olson, Gabriëlle H S Buitendijk, Albert Hofman, Cornelia M van Duijn, Valentina Cipriani, Anthony T Moore, Humma Shahid, Yingda Jiang, Yvette P Conley, Denise J Morgan, Ivana K Kim, Matthew P Johnson, Stuart Cantsilieris, Andrea J Richardson, Robyn H Guymer, Hongrong Luo, Hong Ouyang, Christoph Licht, Fred G Pluthero, Mindy M Zhang, Kang Zhang, Paul N Baird, John Blangero, Michael L Klein, Lindsay A Farrer, Margaret M DeAngelis, Daniel E Weeks, Michael B Gorin, John R W Yates, Caroline C W Klaver, Margaret A Pericak-Vance, Jonathan L Haines, Bernhard H F Weber, Richard K Wilson, John R Heckenlively, Emily Y Chew, Dwight Stambolian, Elaine R Mardis, Anand Swaroop & Goncalo R Abecasis

In the version of this supplementary file originally posted online, panels were missing from Supplementary Figures 1 and 2. The errors have been corrected in this file as of 9 October 2013.

Identification of a Rare Coding Variant in Complement 3 Associated with Age-related Macular Degeneration

Xiaowei Zhan[1,*], David E. Larson[2,*], Chaolong Wang[1,3,*], Daniel C. Koboldt[2], Yuri V. Sergeev[4], Robert S. Fulton[2], Lucinda L. Fulton[2], Catrina C. Fronick[2], Kari E. Branham[5], Jennifer Bragg-Gresham[1], Goo Jun[1], Youna Hu[1], Hyun Min Kang[1], Dajiang Liu[1], Mohammad Othman[5], Matthew Brooks[6], Rinki Ratnapriya[6], Alexis Boleda[6], Felix Grassmann[7], Claudia von Strachwitz[8], Lana M. Olson[9,10], Gabriëlle H.S. Buitendijk[11,12], Albert Hofman[12,13], Cornelia M. van Duijn[12], Valentina Cipriani[14,15], Anthony T. Moore[14,15], Humma Shahid[16,17], Yingda Jiang[18], Yvette P. Conley[19], Denise J. Morgan[20], Ivana K. Kim[21], Matthew P. Johnson[22], Stuart Cantsilieris[23], Andrea J. Richardson[23], Robyn H. Guymer[23], Hongrong Luo[24,25], Hong Ouyang[24,25], Christoph Licht[26], Fred G. Pluthero[27], Mindy M. Zhang[24,25], Kang Zhang[24,25], Paul N. Baird[23], John Blangero[22], Michael L. Klein[28], Lindsay A. Farrer[29,30,31,32,33], Margaret M. DeAngelis[20], Daniel E. Weeks[18,34], Michael B. Gorin[35], John R.W. Yates[14,15,16], Caroline C.W. Klaver[11, 12], Margaret A. Pericak-Vance[36], Jonathan L. Haines[9,10], Bernhard H.F. Weber[7], Richard K. Wilson[2], John R. Heckenlively[5], Emily Y. Chew[37], Dwight Stambolian[38], Elaine R. Mardis[2,+], Anand Swaroop[6,+], Goncalo R. Abecasis[1,+]

* X.Z., D.L. and C.W. are joint first authors.

+ E.M., A.S. and G.R.A. jointly directed the project.

Nature Genetics: doi:10.1038/ng.2758

Supplementary Materials

Supplementary Table 1. List of Regions Selected for Sequencing. For each region, we list genomic coordinates for the locus (based on the Genome Reference Consortium build 37 assembly) together with a summary of designed probes and targeted bases. A list of the protein coding genes in each locus, which are the focus of our analysis, is also provided.

Interval					Target Information					Protein Coding Genes		
Chr	Start Position	End Position	Length	Protein Coding Bases	# Probes	# Bases	% Interval	Protein Coding Bases	% Protein Coding Bases	Locus Name	# Genes	Gene Names
1	196,341,101	196,994,612	653,511	11,359	1,520	226,684	34.69	11,007	96.90	*CFH*	7	*CFH, CFHR1, CFHR2, CFHR3, CFHR4, CFHR5, KCNT2*
4	110,547,457	110,733,347	185,890	4,116	891	132,950	71.52	4,087	99.30	*CFI*	4	*CASP6, CCDC109B, CFI, PLA2G12A*
6	31,720,915	32,087,186	366,271	66,023	1,393	207,700	56.71	63,090	95.56	*C2/CFB*	29	** See legend **
8	19,786,532	19,938,633	152,101	1,428	737	109,963	72.30	1,418	99.30	*LPL*	1	*LPL*
9	107,533,234	107,700,286	167,052	10,408	860	128,141	76.71	10,341	99.36	*ABCA1*	3	*ABCA1, LOC286367, NIPSNAP3B*
10	124,113,939	124,412,943	299,004	10,432	388	57,812	19.33	10,146	97.26	*ARMS2*	5	*ARMS2, DMBT1, HTRA1, MIR3941, PLEKHA1*
15	58,555,986	58,870,773	314,787	1,500	197	29,453	9.36	1,488	99.20	*LIPC*	1	*LIPC*
16	56,980,401	57,026,900	46,499	1,482	61	9,089	19.55	1,451	97.91	*CETP*	2	*CETP, NLRC5*
19	6,669,795	6,734,343	64,548	6,469	122	18,178	28.16	6,204	95.90	*C3*	3	*C3, GPR108, TNFSF14*
22	32,904,490	33,412,741	508,251	2,379	313	46,637	9.18	2,360	99.20	*SYN3/TIMP3*	2	*SYN3, TIMP3*
Total			2,757,914	115,596	6,482	966,607	35.05	111,592	96.54		57	

** For the chromosome 6 region, the following 29 genes were sequenced: *ATF6B, C2, C4A, C4B, C6orf26, C6orf27, C6orf48, CFB, CYP21A2, DOM3Z, EHMT2, HSPA1A, HSPA1B, HSPA1L, LSM2, MIR1236, MSH5, MSH5-C6ORF26, NEU1, RDBP, SKIV2L, SLC44A4, SNORD48, SNORD52, STK19, TNXA, TNXB, VARS, ZBTB12.*

Nature Genetics: doi:10.1038/ng.2758

Supplementary Table 2: Summary of Sequencing Results and Analyzed Variants. Variants were called using UMAKE with standard filters (See **Online Methods** for details). Comparisons to ESP were restricted to regions targeted in both ESP and our experiment, where depth of coverage >10x for 90% of samples, and that were >5-bp away from an insertion-deletion polymorphism (as noted in text).

	Initial Call Set	Protein Coding Regions	Sites Compared To ESP
Target Summary			
Targeted nucleotides	2,757,914	115,596	-
Examined nucleotides	966,607	111,592	97,196
Mean coverage	106.8	128.6	133.0
Fraction >10x[#]	.95 (.92-.99)	.98 (.98-1.00)	.98 (.98-1.00)
Overall			
No. sites	31,527	2,368	1,148
No. in 1000 Genomes Phase I	11,721	750	707
No. in dbSNP 135	12,571	1,017	797
Fraction Novel[*]	59.82%	55.03%	25.78%
No. synonymous	834	834	280
No. nonsynonymous	1,379	1,379	416
No. nonsense	43	43	10
Ts/Tv ratio	2.09	2.88	2.73
Variation Per Sample			
No. sites	1,714	78	89
No. in 1000 Genomes Phase I	1,650	75	88
No. in dbSNP 135	1,691	76	87
Fraction Novel[*]	1%	0%	0%
No. synonymous	40	40	24
No. nonsynonymous	34	34	19
No. nonsense	1	1	1

Fraction of variant sites covered. We showed average values and quartile ranges are shown within parentheses.
* Fraction novel denotes the fractions of variants that not reported in 1000 Genomes Project Phase 1 or dbSNP 135.

Nature Genetics: doi:10.1038/ng.2758

Supplementary Table 3. Protein Coding Variants Observed in Sequenced Samples, Categorized by Frequency and Functional Consequence.

	Single Copy	Two Copies	Two Copies to 0.1%	0.1-1%	1-5%	5-10%	>10%	Total
Synonymous	431	82	102	110	34	19	56	834
Nonsynonymous	821	148	135	151	62	20	42	1379
Nonsense	27	6	3	4	1	0	2	43
Total	1279	236	240	265	97	39	100	2256

Nature Genetics: doi:10.1038/ng.2758

Supplementary Table 4: Initial Statistical Association Analysis of 2,335 Sequenced AMD Cases and 789 Sequenced Controls. A total of 1,422 coding variants (see **Supplementary Table 3**) were tested for association and the top coding variant association signal in each locus is listed. Common variants are tabulated when $p < 1x10^{-6}$ and rare variants are listed when $p < .01$. All p-values were calculated using exact logistic regression. For rare variants, we re-evaluated statistical significance after adjusting for the top common variant at the locus to avoid shadow signals driven by linkage disequilibrium.

SNP	Chromosome	Position(bp)	Nearest Gene	Consequence	Alleles (ref/alt)	Frequency (alt allele)		OR	P-value	Conditional P-value*
						Cases	Controls			
Common variant hits										
rs1061170	1	196,659,237	CFH	H402Y	C/T	0.481	0.662	0.47	4.48×10^{-36}	
rs641153	6	31,914,180	C2	R32Q	G/A	0.060	0.105	0.55	1.26×10^{-8}	
rs10490924	10	124,214,448	ARMS2	A69S	G/T	0.326	0.184	2.15	1.85×10^{-28}	
rs2230199	19	6,718,387	C3	R102G	G/C	0.247	0.175	1.55	2.31×10^{-9}	
Rare variant hits MAF < 1% , marginal and conditional P <.01 (after conditioning on nearby common variants)										
rs121913059	1	196,716,375	CFH	R1210C	C/T	0.005	0.000	∞	2.57×10^{-3}	2.00×10^{-4} (rs1061170)
rs143667999	6	31,922,453	RDBP	D208E	G/C	0.001	0.005	0.21	5.99×10^{-3}	6.70×10^{-3} (rs641153)
rs147859257	19	6,718,146	C3	K155Q	T/G	0.010	0.003	3.27	6.30×10^{-3}	2.50×10^{-3} (rs2230199)

*: Conditional P-value of rare variants is calculated by adjusting for the top common SNPs in the parenthesis. When analysis was not restricted to coding variants, an additional common variant signal was found at rs255, in an intron of the *LPL* gene ($p = 3.6 \times 10^{-20}$).

Nature Genetics: doi:10.1038/ng.2758

Supplementary Table 5. Results for Follow-up Genotyping of K155Q in 471 Families with Multiple Affected Individuals.

Sample Set	Pedigrees Screened	Nuclear Families with a Carrier*		
		Number of Families	Number of Affected Individuals	Number of Affected Carriers
United States / University of Utah	109	3	6	3
United States / Oregon Health Sciences Center	91	1	4	4
United Kingdom / University College London	87	2	5	5
Germany / University of Regensburg	78	3	9	5
Australia / University of Melbourne	56	4	13	7
United States / University of California, San Diego	27	2	4	4
United States / University of Wisconsin – Case Western Reserve University	23	3	8	7
Total	471	18	49	35

* Each pedigree contributed no more than one nuclear family to this analysis.

Nature Genetics: doi:10.1038/ng.2758

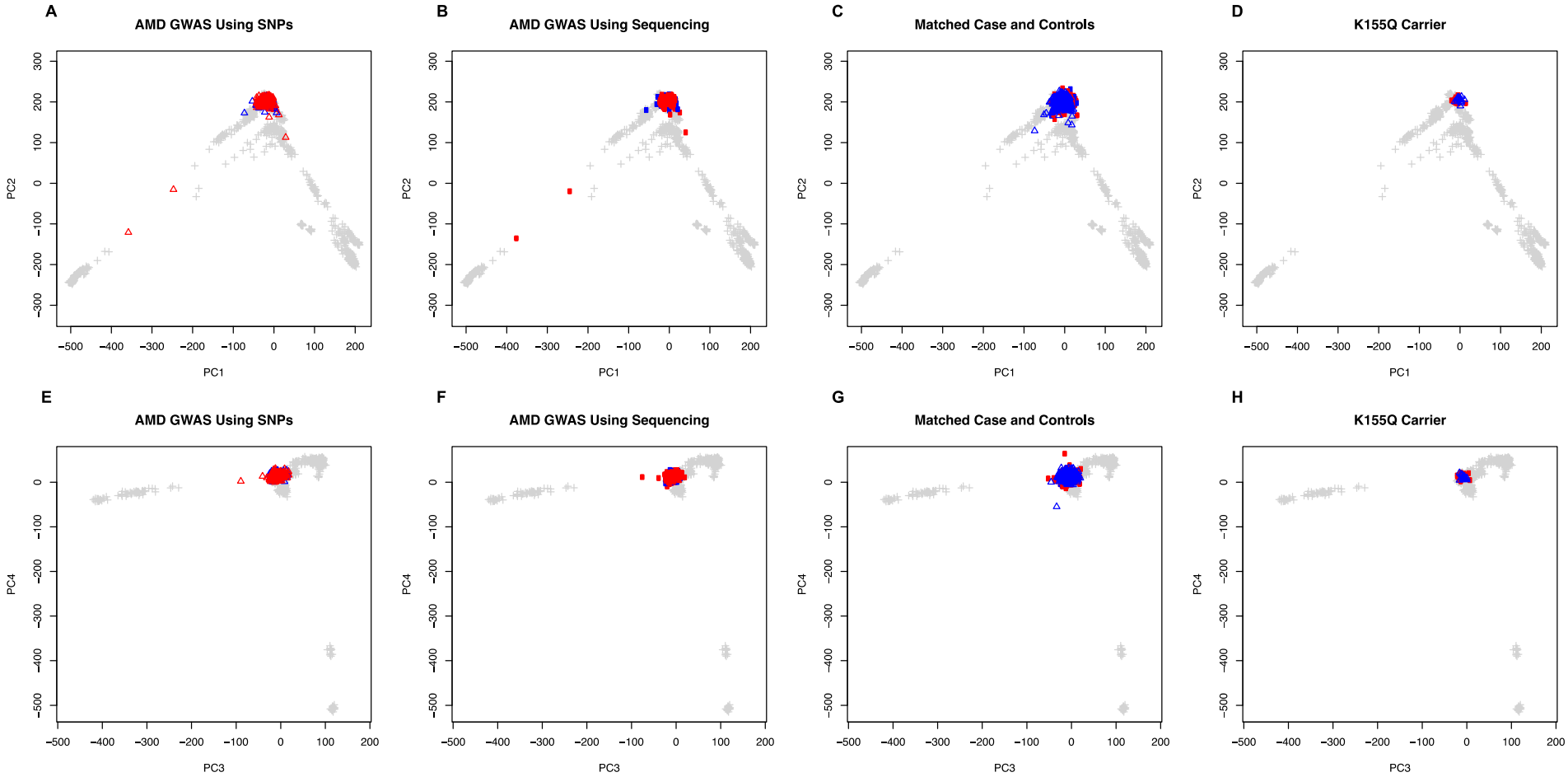
Supplementary Table 6. Primers for Sanger Sequencing of K155Q.

Strand	Sequence
Forward primer	5′ GTCAG AAAAG GGGCG CAACA A
Reverse primer	5′ TCTGG CTGGC ACCTC AATGT T

Nature Genetics: doi:10.1038/ng.2758

Supplementary Figure 1: Ancestry Based Matching Using the HGDP Reference Panel.

We defined a 4 dimensional genetic ancestry map using genotypes for the Human Genome Diversity Panel (HGDP). We label cases in red and controls in blue. Ancestry of sequenced samples with GWAS genotypes is summarized in panel A (PC1 and PC2) and in panel E (PC3 and PC4). Ancestry for the same samples estimated using off-target sequence reads is displayed in panel B (PC1 and PC2) and in panel F (PC3 and PC4). Comparison of these panels shows that ancestry information can be inferred from either GWAS genotypes or data from targeted sequencing experiments. Ancestry for 2,268 cases and 2,268 controls matched according to estimated ancestry and drawn from our AMD study and the NHLBI exome sequencing study is summarized in Panel C (PC1 and PC2) and in Panel G (PC3 and PC4). Finally, ancestry of K155Q carriers is summarized in Panel D (PC1 and PC2) and Panel H (PC3 and PC4).



Nature Genetics: doi:10.1038/ng.2758

Supplementary Figure 2: Quality Control of K155Q Variant Calls in Our Samples and Those from the NHLBI Exome Sequencing Project. The figure shows that deep sequencing data was obtained for K155Q and that depth of sequencing at K155Q (and at K155Q heterozygotes) was similar to that at other sequenced sites.

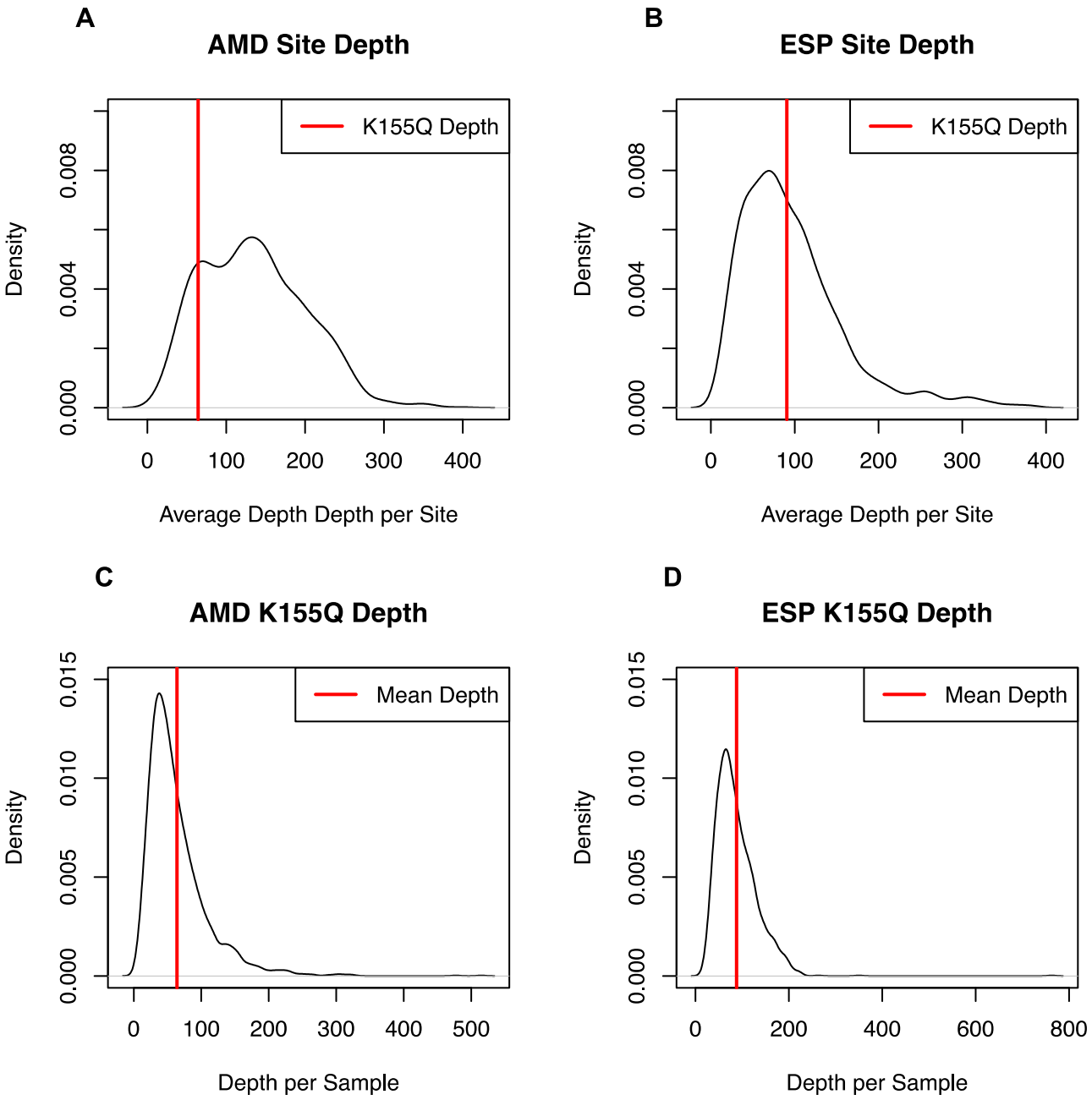
Panel A: Density plot comparing average sequencing depth at K155Q in our targeted sequence data (red line) to that in the remaining sites examined in the comparison between 2,268 cases and 2,268 ancestry matched controls (histogram). The average sequencing depth at K155Q was 63.73 in our data.

Panel B: Density plot comparing total sequencing depth at K155Q in ESP (red line) to that in the remaining sites examined in the comparison between 2,268 cases and 2,268 ancestry matched controls (histogram). The average sequencing depth at K155Q was 90.53 in ESP samples.

Panel C: Density plot comparing depth at K155Q in heterozygote carriers in our targeted sequencing sample to that in the remaining samples. The red line marks average depth (64.11) at K155Q for carriers, the histogram summarizes depth distribution across all genotyped sites.

Panel D: Density plot examining depth at K155Q in heterozygote carriers in the ESP sample. The red line marks average depth (87.93) at K155Q for carriers, the histogram summarizes depth distribution across all genotyped sites.

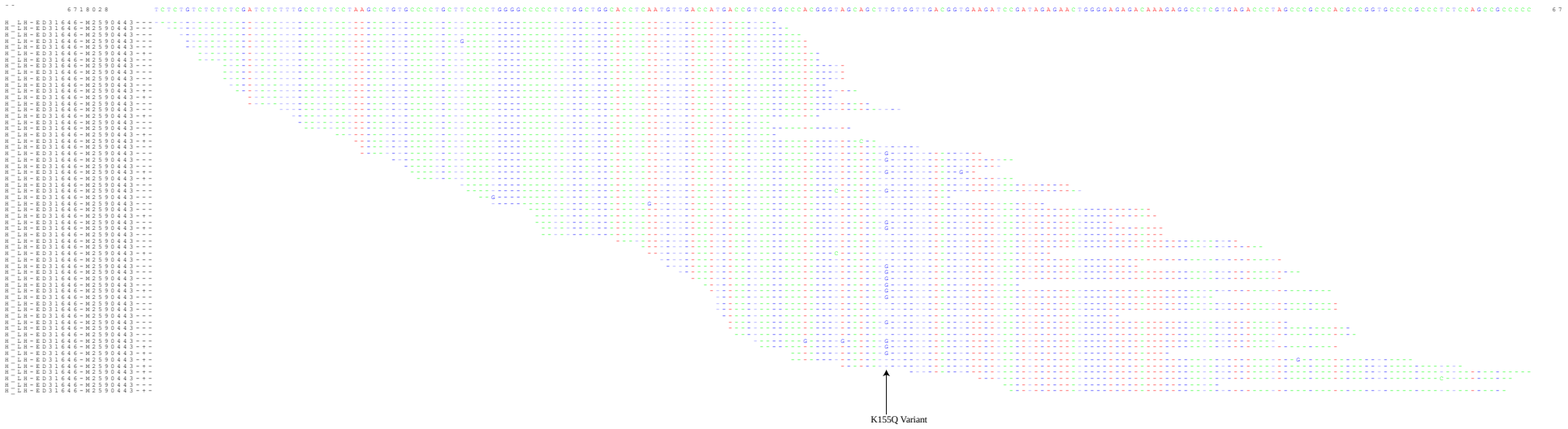
Nature Genetics: doi:10.1038/ng.2758



Nature Genetics: doi:10.1038/ng.2758

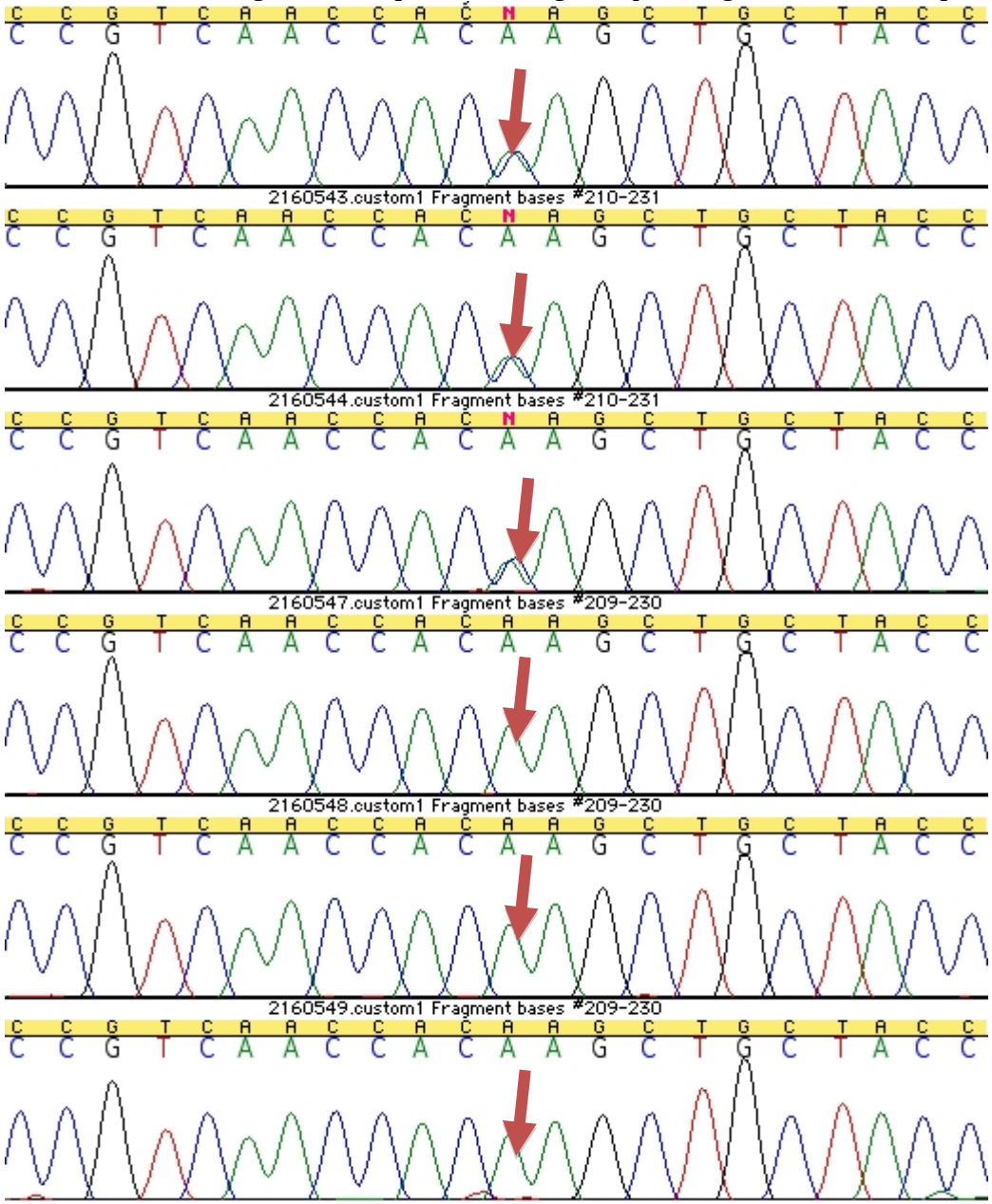
Supplementary Figure 3: Raw Sequence Reads in One Putative K155Q Variant Carrier.

We examined the reads supporting K155Q variant calls. The top line summarizes the reference genome sequence and is followed by all the sequence reads overlapping K155Q in a predicted heterozygote. For ease of visualization, bases are colored and we use a '-' (dash) to represent bases that match the reference genome in each sample. The figure shows that the alignment strongly supports the variant, with no evidence that (for example) reads with the variant carry an excess of other differences in relation to the reference genome.



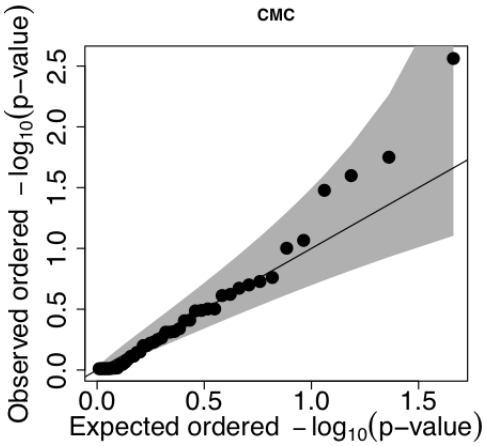
Nature Genetics: doi:10.1038/ng.2758

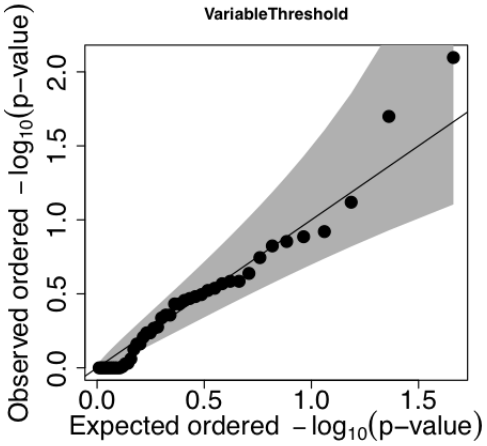
Supplementary Figure 4. Sanger Validation of K155Q Variant Status for Three Carriers and Three Non-carriers. We used Sanger sequencing to confirm our original genotype assignments for 100 individuals (50 carriers, 50 non-carriers). All of the original genotype assignments were confirmed and the figure below provides Sanger sequencing results for six representative samples.

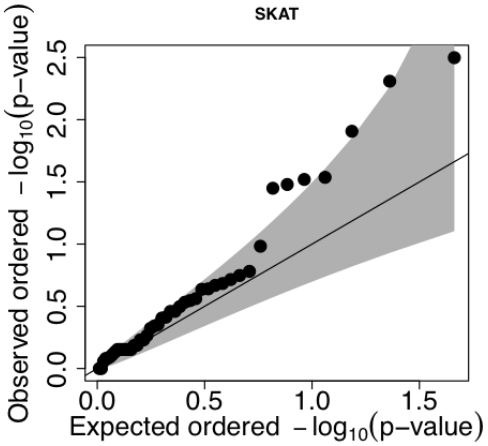


Nature Genetics: doi:10.1038/ng.2758

Supplementary Figure 5. Results for Gene-based Burden Tests. The three figures summarize results for gene-based burden tests that aggregate evidence for association across all rare variants. We calculated a simple burden test (grouping all rare variants), a variable threshold test (that automatically selects the optimal frequency threshold for defining rare variants) and a sequence kernel association test (that allows for variants with opposite effects to reside in the same locus). For all tests, results did not deviate significantly from null expectations (the gray shaded region in QQ plots).







Nature Genetics: doi:10.1038/ng.2758