nature genetics

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

Xiaowei Zhan[1,39], David E Larson[2,39], Chaolong Wang[1,3,39], 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[29–33], Margaret M DeAngelis[20], Daniel E Weeks[18,34], Michael B Gorin[35], John R W Yates[14–16], 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[2,40], Anand Swaroop[6,40] & Goncalo R Abecasis[1,40]

Macular degeneration is a common cause of blindness in the elderly. To identify rare coding variants associated with a large increase in risk of age-related macular degeneration (AMD), we sequenced 2,335 cases and 789 controls in 10 candidate loci (57 genes). To increase power, we augmented our control set with ancestry-matched exome-sequenced controls. An analysis of coding variation in 2,268 AMD cases and 2,268 ancestry-matched controls identified 2 large-effect rare variants: previously described p.Arg1210Cys encoded in the *CFH* gene (case frequency (f_{case}) = 0.51%; control frequency ($f_{control}$) = 0.02%; odds ratio (OR) = 23.11) and newly identified p.Lys155Gln encoded in the *C3* gene (f_{case} = 1.06%; $f_{control}$ = 0.39%; OR = 2.68). The variants suggest decreased inhibition of C3 by complement factor H, resulting in increased activation of the alternative complement pathway, as a key component of disease biology.

Genetic and environmental factors contribute to AMD[1,2], a major cause of vision loss in elderly individuals[3]. Pioneering discovery of association of AMD with complement factor H (encoded by *CFH*[4–6]) was quickly followed by the identification of additional susceptibility loci that now include *ARMS2-HTRA1* (refs. 7,8) and complement genes *C3*, *C2-CFB* and *CFI*[9–12]. Genome-wide association studies (GWAS) of AMD cases and controls have now identified common susceptibility variants at ~20 different loci[13,14] and have begun to uncover specific cellular pathways involved in AMD biology.

Whereas common variants tag an associated genomic region, rare coding variants can provide more specific clues about the underlying disease mechanism[15]. For example, rare variant p.Arg1210Cys encoded in the *CFH* gene was recently associated with a large increase in AMD risk using targeted sequencing of rare *CFH* risk haplotypes[16]. The resulting altered protein has decreased binding to C3b, C3d, heparin and endothelial cells[17–19]. A reduction in the ability of CFH to inactivate C3, leading to increased cell killing activity of the complement pathway, could contribute to AMD, representing a much more specific and testable hypothesis about disease mechanism than provided by common *CFH* variants whose mechanistic consequences are unclear.

To systematically identify rare, large-effect variants, we carried out targeted sequencing of eight AMD risk loci identified in GWAS[20] (near *CFH*, *ARMS2*, *C3*, *C2-CFB*, *CFI*, *CETP*, *LIPC* and *TIMP3-SYN3*) and two candidate regions (*LPL* and *ABCA1*) (**Supplementary Table 1**). We resequenced these regions in 3,124 individuals (2,335 cases and 789 controls) recruited in ophthalmology clinics at the University of Michigan and the University of Pennsylvania and in Age-Related Eye Disease Study (AREDS) participants[20,21]. We enriched genomic targets using a set of 150-bp probes designed by Agilent Technologies and generated sequence data on Illumina Genome Analyzer and HiSeq instruments. The 10 loci comprised 115,596 nucleotides of protein-coding sequence and totaled 2,757,914 nucleotides overall. We designed probes to capture 111,592 protein-coding nucleotides (96.5% of coding sequence) and 966,607 nucleotides overall (35.1% of the locus sequence), generating an average of 123,221,974 mapped

A full list of author affiliations appears at the end of the paper.

Received 26 January; accepted 19 August; published online 15 September 2013; doi:10.1038/ng.2758

bases of on-target sequence per individual (127.5× average depth when counting bases with quality of >20 in reads with mapping quality of >30 after duplicate read removal); 98.49% of sites with designed probes were covered at >10× depth. We applied variant calling tools and quality control filters similar to those used to analyze National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project (ESP) data[22,23] (**Supplementary Table 2**). We identified an average of 1,714 non-reference sites in each sequenced individual. In total, we identified 31,527 single-nucleotide variants, of which 18,956 were not in dbSNP135. Discovered sites included 834 synonymous variants, 1,379 nonsynonymous variants and 43 nonsense variants, most of which were extremely rare (**Supplementary Table 3**). For 13 samples sequenced in duplicate, genotype concordance was 99.82% (when depth was >10×). For 908 samples previously examined with GWAS arrays[20], sequencing-based genotypes were 98.99% concordant with array-based calls (again, when depth was >10×).

In an initial comparison of AMD cases and controls (**Supplementary Table 4**), no rare coding variants with frequency of <1% reached experiment-wide significance ($P < 0.05/31{,}527 = 1.6 \times 10^{-6}$ when including all discovered variants or $P < 0.05/1{,}422 = 3.5 \times 10^{-5}$ when considering only protein-altering variants), although several showed encouraging patterns of association. For example, the rare variant p.Arg1210Cys encoded in the *CFH* gene was observed in 23 of the 2,335 sequenced cases but in none of the 789 sequenced controls (exact test $P = 0.0025$). Common variants in several loci exhibited strong evidence of association, including in *CFH* (peak variant rs9427642: $f_{case} = 12\%$; $f_{control} = 27\%$; P value $= 2.52 \times 10^{-48}$), *ARMS2* (rs10490924: $f_{case} = 33\%$; $f_{control} = 18\%$; P value $= 5.48 \times 10^{-27}$), *C3* (rs2230199: $f_{case} = 25\%$; $f_{control} = 17\%$; P value $= 3.94 \times 10^{-9}$) and *C2-CFB* (rs556679: $f_{case} = 7\%$; $f_{control} = 12\%$; P value $= 1.32 \times 10^{-10}$).

A key requirement for establishing significance of rare disease-associated variants is the availability of sufficient numbers of control samples. To increase power, we sought to identify additional controls and focused on samples from NHLBI ESP[23], which sequenced 15,336 genes across 6,515 individuals. Sequence data for our samples and the NHLBI ESP samples were analyzed with the same analysis pipeline, which minimized potential differences due to heterogeneity in analysis tools and parameters. To further avoid artifacts from sequencing and variant calling, we restricted our analysis to sites within regions targeted in both sequencing experiments, genotyped and covered with >10 reads in >90% of the samples examined in each project and >5 bp away from insertion-deletion polymorphisms catalogued by the 1000 Genomes Project[24]. Because careful matching of genetic ancestry is critical for rare variant association studies[24,25], we selected an ancestry-matched subset of our samples and of samples from NHLBI ESP. We used principal-component analysis (PCA) to construct a genetic ancestry map of the world with samples from the Human Genome Diversity Project, each genotyped at 632,958 SNPs[26]. If GWAS array genotypes were available for our samples and for the NHLBI ESP samples, it would be straightforward to place the samples directly on this genetic ancestry map. Using targeted sequence data, however, the analysis is more challenging: targeted regions include too few variants to accurately represent global ancestry, and off-target regions are covered too poorly, precluding estimation of the accurate genotypes needed for standard PCA. Thus, we relied on the new LASER algorithm (C.W., X.Z., J.B.-G., H.M.K., D.S. *et al.*, unpublished data) to localize each sequenced sample on a predefined genetic ancestry map of the world. The method can accurately place individuals on this worldwide ancestry map with <0.05× average coverage of the genome and is thus ideal for targeted sequence data, such as ours and the NHLBI ESP data, which have average off-target coverage of ~0.23× and ~0.90×, respectively (see **Supplementary Fig. 1a**,**b**,**e**,**f**, which show that the PCA coordinates inferred using 0.10× genome coverage or using GWAS array genotypes are highly similar). We focused on samples where PCA coordinates could be estimated confidently (Procrustes similarity larger than 0.95; Online Methods) and used a greedy algorithm to match cases and controls on the basis of estimated genetic ancestry. As shown in the Online Methods, alternative matching algorithms did not alter our conclusions. After matching, we focused on a set of 2,268 AMD cases and 2,268 controls that were ancestry-matched one to one (**Supplementary Fig. 1c**,**g**). Because AMD phenotype information was not available for most controls, we expect that a small proportion may eventually develop disease; however, this should not affect power substantially[27]. After matching case-control samples, we excluded 1 variant with Hardy-Weinberg equilibrium test P value $< 1 \times 10^{-6}$ and focused our analysis on 430 protein-changing variants in regions that were targeted and deeply sequenced in both experiments as well as being far away from insertion-deletion polymorphisms.

In this expanded analysis (**Table 1**), common variant signals at all loci increased in significance (in comparison to what is shown in **Supplementary Table 4**). In addition, two rare coding variants exhibited association with $P < 0.01$. The first variant was p.Arg1210Cys encoded in the *CFH* gene (observed in 1 control and 23 cases; OR = 23.11; exact $P = 2.9 \times 10^{6}$), providing strong support for the original report[16].

Table 1 Summary association results for 2,268 sequenced AMD cases and 2,268 sequenced controls

SNP	Chromosome	Position (bp)	Nearest gene	Consequence	Alleles (ref/alt)	Frequency (alt allele)		OR	*P* value	Conditional *P* value[a]
						Cases	Controls			
Common variant hits										
rs1061170	1	196659237	*CFH*	p.His402Tyr	C/T	0.478	0.623	0.555	1.01×10^{-43}	
rs438999	6	31928306	*SKIV2L*	p.Gln151Arg	A/G	0.058	0.098	0.566	1.26×10^{-12}	
rs10490924	10	124214448	ARMS2	p.Ala69Ser	G/T	0.329	0.197	1.990	1.04×10^{-45}	
rs2230199	19	6718387	C3	p.Arg102Gly	G/C	0.253	0.206	1.300	1.58×10^{-7}	
Rare variant hits (MAF < 1%; marginal and conditional *P* < 0.01 after conditioning on nearby common variants)										
rs121913059	1	196716375	*CFH*	p.Arg1210Cys	C/T	0.005	0.000	23.11	2.9×10^{-6}	6.0×10^{-4} (rs1061170)
rs147859257	19	6718146	*C3*	p.Lys155Gln	T/G	0.011	0.004	2.68	2.7×10^{-4}	2.8×10^{-5} (rs2230199)

Samples in this expanded analysis include our sequenced AMD samples and genetically matched controls, sequenced by us or by the NHLBI ESP. The top coding variant in each locus is included in this table when $P < 1 \times 10^{-6}$. Rare coding variants are included when the corresponding P value for conditional or marginal analysis was less than 1×10^{-4}. All P values were calculated using exact logistic regression. Ref, reference; alt, alternative; MAF, minor allele frequency.

[a]For rare variants, we re-evaluated statistical significance after adjusting for the top common variant in the locus to avoid shadow signals driven by linkage disequilibrium. The variant used for conditioning is named (in parentheses).

Table 2 Follow-up genotyping summary and meta-analysis summary

	Controls		Cases		
Sample set	*N*	MAF	*N*	MAF	*P* value
Discovery sample					
Sequenced samples (*N* = 4,536)	2,268	0.004	2,268	0.011	2.7×10^{-4}
Follow-up samples					
Germany: University of Regensburg (*N* = 2,976)	1,147	0.006	1,829	0.016	1.7×10^{-3}
United States: Vanderbilt/Miami (*N* = 1,819)	726	0.004	1,093	0.007	3.5×10^{-1}
Netherlands: Rotterdam Study (*N* = 1,409)	1,280	0.005	129	0.031	1.5×10^{-4}
UK: Cambridge AMD Study (*N* = 1,279)	423	0.006	856	0.015	6.2×10^{-2}
United States: University of California, Los Angeles/University of Pittsburgh (*N* = 830)	211	0.004	619	0.017	8.3×10^{-4}
deCODE study					
deCODE discovery sample (*N* = 52,578)	51,435	0.005	1,143	–[a]	1.1×10^{-7}
Meta-analysis					
All follow-up samples (*N* = 8,313)	3,787	0.005	4,526	0.013	7.7×10^{-7}
Discovery and all follow-up samples (*N* = 12,849)	6,055	0.005	6,794	0.013	1.1×10^{-9}
Discovery, all follow-up and deCODE samples (*N* = 65,427)	57,490	0.005	7,937	–[a]	1.6×10^{-15}

The table includes the number of cases and controls in each comparison, the corresponding allele frequency for the allele encoding p.Lys155Gln in each set of samples and the *P* value for a comparison of allele frequencies in cases and controls. Meta-analysis *P* values were calculated using Stouffer's method.
[a]MAF values are unavailable for imputed cases from the deCODE study.

The second variant was p.Lys155Gln encoded in the *C3* gene (observed in 18 controls and 48 cases; OR = 2.68; exact $P = 2.7 \times 10^{4}$; see **Supplementary Fig. 1d,h** for carrier ancestry distribution). When controlling for a previously described common variant signal nearby, rs2230199 ($f_{control}$ = 20.63%; f_{case} = 25.26%; marginal exact $P = 1.8 \times 10^{-7}$; OR = 1.31), the evidence for association with p.Lys155Gln increased slightly (conditional OR = 2.91; exact $P = 2.8 \times 10^{-5}$). Inspection of the raw read data showed that the variant was well supported and was unlikely to be an artifact of sequencing or alignment, a result further confirmed by Sanger sequencing (**Supplementary Figs. 2–4**). Finally, in an examination of our sequenced samples and available whole-genome sequences (Online Methods), we observed no additional variants in strong linkage disequilibrium with the mutation encoding p.Lys155Gln that might account for the association signal. Analysis with burden tests, which jointly evaluate evidence for association with rare variants at each gene, identified no significant association signals (**Supplementary Fig. 5**)[28–30].

To confirm the signal corresponding to p.Lys155Gln, we genotyped additional samples totaling 4,526 cases and 3,787 controls and, again, observed strong association ($f_{control}$ = 0.5%; f_{case} = 1.3%; follow-up $P = 7.7 \times 10^{-7}$; combined $P = 1.1 \times 10^{-9}$; **Table 2**). In addition, we genotyped 471 families with multiple AMD cases to identify 18 nuclear families where the mutation encoding p.Lys155Gln segregates. These families included 49 affected individuals, with at least 1 individual carrying an allele encoding p.Lys155Gln, and, adjusting for ascertainment, we estimated that 75% of the first-degree relatives of a p.Lys155Gln carrier who also had AMD would carry the variant, consistent with an OR of ~3 (Online Methods and **Supplementary Table 5**). Further strong evidence for association of this variant with macular degeneration is provided in independent work by deCODE Genetics[31] examining 1,143 Icelandic macular degeneration cases and 51,435 Icelandic controls ($f_{control}$ = 0.55%; OR = 3.45; deCODE $P = 1.1 \times 10^{-7}$; combined $P = 1.6 \times 10^{-15}$). In 1,606 directly genotyped cases of macular degeneration from AREDS2 (ref. 32), the variant had a frequency of 1.77%, similar to our sequenced AMD cases (1.10%) and our follow-up AMD cases (1.30%) and notably higher than our sequenced controls (0.30%), our genotyped controls (0.50%), NHLBI ESP participants with primarily European ancestry (0.40%) and deCODE controls (0.55%). We found no evidence of the p.Lys155Gln variant in a small sample of individuals with atypical hemolytic uremic syndrome (aHUS; *n* = 53), a rare disorder whose genetic risk factors partially overlap with those of macular degeneration.

We next investigated the potential functional consequences of the p.Lys155Gln variant *in silico*. On the basis of protein crystallography, the model in **Figure 1** shows that CFH variant p.Arg1210Cys (OR = 23.11), C3 variant p.Lys155Gln (OR = 2.91) and C3 variant p.Arg102Gly (OR = 1.31) all map near the surface where CFH and C3b interact, suggesting that they might affect binding of complement factor H to C3b. CFH inhibits C3b and limits the immune responses mediated by the alternative complement pathway. We hypothesize that p.Lys155Gln and p.Arg102Gly affect binding of the first macroglobular domain of C3 to CFH and thus interfere with inactivation of the alternative complement pathway, a hypothesis that must be confirmed

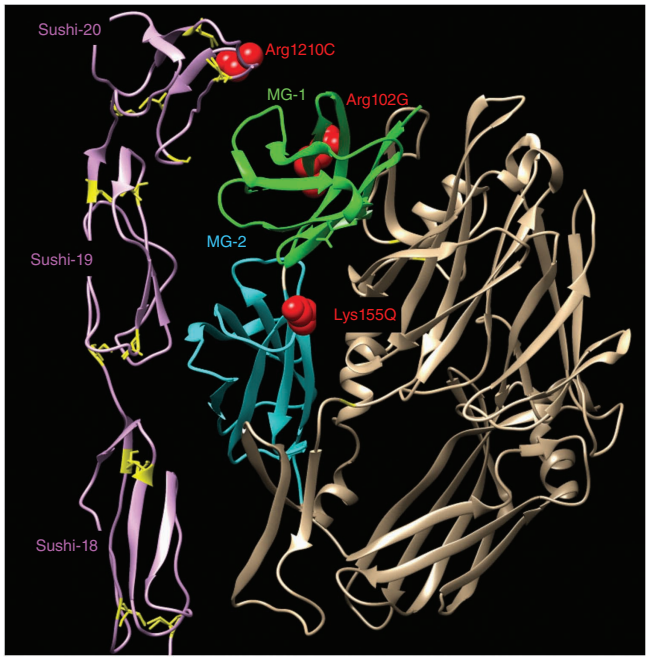

Figure 1 C3 variants p.Arg102Gly and p.Lys155Gln and CFH variant p.Arg1210Cys are in the interaction domains of the first α-macroglobular domains of C3b and CFH, respectively. A fragment of the crystal structure of the four Sushi domains of CFH (purple; one not shown for clarity) in a complex with complement fragment C3b (Protein Data Bank (PDB) 2wii) was used to explore the effect of disease-associated nonsynonymous changes. CFH residues 987–1230 were used to generate the structure with the first four Sushi domains from 2wii serving as a structural template (light purple, with cysteine residue side chains in yellow). The C-terminal Sushi domains were docked to the binding site in C3b. The first two α-macroglobulin domains of C3b, MG-1 and MG-2, are shown in green and cyan, respectively. The locations of the p.Arg102Gly, p.Lys155Gln and p.Arg1210Cys alterations are marked in red.

experimentally[33]. Interestingly, the three variants (p.Arg102Gly and p.Lys155Gln in C3 and p.Arg1210Cys in CFH) all involve the replacement of a positively charged residue.

In summary, our work and that described in the companion paper identify p.Lys155Gln as a rare C3 variant associated with ~2.91-fold increased risk of macular degeneration. Together with rare CFH variant p.Arg1210Cys and previously described common C3 variant p.Arg102Gly, p.Lys155Gln may reduce binding of CFH to C3b, inhibiting the ability of CFH to inactivate the alternative complement pathway. Clarifying the mechanistic impact of p.Lys155Gln is likely to be challenging, as illustrated by contradictory results from previous functional follow-up studies of AMD-associated loci[34–36], but functional studies of complement activity suggest potential next steps[33,37]. Our work relied on targeted sequencing of GWAS-identified loci, genetic ancestry matching of our sequenced samples to additional sequenced controls analyzed with the same variant calling and filtering tools, focused analysis of regions deeply sequenced in both our project and previously sequenced controls, and avoidance of common calling artifacts near insertion-deletion polymorphisms. The use of publicly available samples to augment control sets may be useful in many targeted sequencing studies, but the strictness of matching and variant filtering required to prevent false positive findings due to population stratification and/or sequence analysis artifacts are areas deserving of further study. As the number of sequenced human genomes and exomes grows, we expect that the usefulness of the approach will grow, making it possible to match multiple controls to each case and to focus on progressively finer ancestry matches. Although our results emphasize that large sample sizes will be required for rare variant studies of complex human traits, they also show the promise of these studies for clarifying disease biology.

URLs. LASER software for estimation of genetic ancestry can be obtained from http://genome.sph.umich.edu/wiki/LASER. UMAKE and GotCloud tools for variant calling can be obtained from http://genome.sph.umich.edu/wiki/UMAKE and http://genome.sph.umich.edu/wiki/GotCloud, respectively. The QPLOT tool for assessing sequence quality can be obtained from http://genome.sph.umich.edu/wiki/QPLOT.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank all study participants for their generous volunteering. We thank B. Li, W. Chen, C. Sidore, T. Teslovich, L. Fritsche and M. Boehnke for useful discussion and suggestions. This project was supported by grants from the US National Institutes of Health (National Eye Institute, National Human Genome Research Institute; grants EY022005, HG007022, HG005552, EY016862, U54HG003079 and EY09859); the Medical Research Council, UK (grant G0000067); the Deutsche Forschungsgemeinschaft (grant WE1259/19-2); the Alcon Research Institute; The UK Department of Health's National Institute for Health Research (NIHR) Biomedical Research Centre for Ophthalmology at Moorfields Eye Hospital and the UCL Institute of Ophthalmology; Research to Prevent Blindness (New York); the Thome Memorial Foundation; the Harold and Pauline Price Foundation; and the National Health and Medical Research Council of Australia (NHMRC) Clinical Research Excellence (grant 529923, NHMRC practitioner fellowship 529905 and NHMRC Senior Research Fellowship 1028444). The study was also supported by the Intramural Research Program (Computational Medicine Initiative) of the National Eye Institute. The Centre for Eye Research Australia (CERA) receives operational infrastructure support from the Victorian Government. The views expressed in the publication are those of the authors and not necessarily those of their employers or the funders.

AUTHOR CONTRIBUTIONS

R.K.W., J.R.H., E.Y.C., D.S., E.R.M., A.S. and G.R.A. conceived, designed and supervised the experiments. X.Z. and G.R.A. wrote the initial version of the manuscript. X.Z., D.E.L., C.W. and D.C.K. analyzed the data. D.E.L., D.C.K., R.S.F., L.L.F. and C.C.F. supervised data generation. C.W. developed statistical methodology. Y.V.S. analyzed protein structures. K.E.B. supervised sample and data collection. J.B.-G., G.J., Y.H., H.M.K. and D.L. contributed data and analysis tools. M.B., R.R. and A.B. assisted in laboratory experiments. M.O. and F.G. carried out experimental studies (genotyping and data analysis) for the Michigan and Regensburg samples, respectively. C.v.S. recruited the family members of sporadic AMD cases and controls and collected peripheral blood samples for the Regensburg study. L.M.O., M.A.P.-V. and J.L.H. provided results and analysis for the Vanderbilt/Miami samples. G.H.S.B., A.H., C.M.v.D. and C.C.W.K. provided results and analysis for samples from the Rotterdam Study, Erasmus Medical Center. V.C., A.T.M., H.S. and J.R.W.Y. provided results and analysis for the Cambridge AMD Study samples. Y.J., Y.P.C., D.E.W. and M.B.G. provided results and analysis for the University of California, Los Angeles/University of Pittsburgh samples. D.J.M., I.K.K., L.A.F. and M.M.D. provided results and analysis for the Utah samples. M.P.J., J.B. and M.L.K. provided results and analysis for the Oregon Health Sciences Center samples. S.C., A.J.R., R.H.G. and P.N.B. provided results and analysis for the University of Melbourne samples. H.L., H.O., M.M.Z. and K.Z. provided results and analysis for the University of California, San Diego samples. C.L. and F.G.P. provided results and analysis for a cohort of individuals with aHUS. B.H.F.W. was involved in the design and planning of the Southern Germany AMD Study. B.H.F.W. participated in study coordination and critically read the manuscript. All authors have critically commented on this manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Priya, R.R., Chew, E.Y. & Swaroop, A. Genetic studies of age-related macular degeneration: lessons, challenges, and opportunities for disease management. *Ophthalmology* **119**, 2526–2536 (2012).
2. Swaroop, A., Chew, E.Y., Rickman, C.B. & Abecasis, G.R. Unraveling a multifactorial late-onset disease: from genetic susceptibility to disease mechanisms for age-related macular degeneration. *Annu. Rev. Genomics Hum. Genet.* **10**, 19–43 (2009).
3. Friedman, D.S. *et al.* Prevalence of age-related macular degeneration in the United States. *Arch. Ophthalmol.* **122**, 564–572 (2004).
4. Haines, J.L. *et al.* Complement factor H variant increases the risk of age-related macular degeneration. *Science* **308**, 419–421 (2005).
5. Edwards, A.O. *et al.* Complement factor H polymorphism and age-related macular degeneration. *Science* **308**, 421–424 (2005).
6. Klein, R.J. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385–389 (2005).
7. Jakobsdottir, J. *et al.* Susceptibility genes for age-related maculopathy on chromosome 10q26. *Am. J. Hum. Genet.* **77**, 389–407 (2005).
8. Rivera, A. *et al.* Hypothetical *LOC387715* is a second major susceptibility gene for age-related macular degeneration, contributing independently of complement factor H to disease risk. *Hum. Mol. Genet.* **14**, 3227–3236 (2005).
9. Yates, J.R. *et al.* Complement C3 variant and the risk of age-related macular degeneration. *N. Engl. J. Med.* **357**, 553–561 (2007).
10. Gold, B. *et al.* Variation in factor B (*BF*) and complement component 2 (*C2*) genes is associated with age-related macular degeneration. *Nat. Genet.* **38**, 458–462 (2006).
11. Fagerness, J.A. *et al.* Variation near complement factor I is associated with risk of advanced AMD. *Eur. J. Hum. Genet.* **17**, 100–104 (2009).
12. Maller, J.B. *et al.* Variation in complement factor 3 is associated with risk of age-related macular degeneration. *Nat. Genet.* **39**, 1200–1201 (2007).
13. Fritsche, L.G. *et al.* Seven new loci associated with age-related macular degeneration. *Nat. Genet.* **45**, 433–439 (2013).
14. Arakawa, S. *et al.* Genome-wide association study identifies two susceptibility loci for exudative age-related macular degeneration in the Japanese population. *Nat. Genet.* **43**, 1001–1004 (2011).
15. Nejentsev, S., Walker, N., Riches, D., Egholm, M. & Todd, J.A. Rare variants of *IFIH1*, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* **324**, 387–389 (2009).
16. Raychaudhuri, S. *et al.* A rare penetrant mutation in *CFH* confers high risk of age-related macular degeneration. *Nat. Genet.* **43**, 1232–1236 (2011).
17. Józsi, M. *et al.* Factor H and atypical hemolytic uremic syndrome: mutations in the C-terminus cause structural changes and defective recognition functions. *J. Am. Soc. Nephrol.* **17**, 170–177 (2006).

18. Manuelian, T. *et al.* Mutations in factor H reduce binding affinity to C3b and heparin and surface attachment to endothelial cells in hemolytic uremic syndrome. *J. Clin. Invest.* **111**, 1181–1190 (2003).
19. Ferreira, V.P. *et al.* The binding of factor H to a complex of physiological polyanions and C3b on cells is impaired in atypical hemolytic uremic syndrome. *J. Immunol.* **182**, 7009–7018 (2009).
20. Chen, W. *et al.* Genetic variants near *TIMP3* and high-density lipoprotein–associated loci influence susceptibility to age-related macular degeneration. *Proc. Natl. Acad. Sci. USA* **107**, 7401–7406 (2010).
21. Age-Related Eye Disease Study Research Group. Risk factors associated with age-related macular degeneration. A case-control study in the age-related eye disease study: Age-Related Eye Disease Study Report Number 3. *Ophthalmology* **107**, 2224–2232 (2000).
22. Tennessen, J.A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).
23. Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216–220 (2013).
24. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
25. Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.* **44**, 243–246 (2012).
26. Li, J.Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100–1104 (2008).
27. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
28. Li, B. & Leal, S.M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* **83**, 311–321 (2008).
29. Price, A.L. *et al.* Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.* **86**, 832–838 (2010).
30. Wu, M.C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89**, 82–93 (2011).
31. Helgason, H. *et al.* A rare nonsynonymous sequence variant in *C3* confers high risk of age-related macular degeneration. *Nat. Genet.* doi:10.1038/ng.2740 (15 September 2013).
32. AREDS2 Research Group. *et al.* The Age-Related Eye Disease Study 2 (AREDS2): study design and baseline characteristics (AREDS2 report number 1). *Ophthalmology* **119**, 2282–2289 (2012).
33. Heurich, M. *et al.* Common polymorphisms in C3, factor B, and factor H collaborate to determine systemic complement activity and disease risk. *Proc. Natl. Acad. Sci. USA* **108**, 8761–8766 (2011).
34. Fritsche, L.G. *et al.* Age-related macular degeneration is associated with an unstable *ARMS2* (*LOC387715*) mRNA. *Nat. Genet.* **40**, 892–896 (2008).
35. Kanda, A. *et al.* A variant of mitochondrial protein LOC387715/ARMS2, not HTRA1, is strongly associated with age-related macular degeneration. *Proc. Natl. Acad. Sci. USA* **104**, 16227–16232 (2007).
36. Dewan, A. *et al. HTRA1* promoter polymorphism in wet age-related macular degeneration. *Science* **314**, 989–992 (2006).
37. Sánchez-Corral, P. *et al.* Structural and functional characterization of factor H mutations associated with atypical hemolytic uremic syndrome. *Am. J. Hum. Genet.* **71**, 1285–1295 (2002).

[1]Department of Biostatistics, Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, Michigan, USA. [2]The Genome Institute, Washington University School of Medicine, St. Louis, Missouri, USA. [3]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, USA. [4]Ophthalmic Genetics and Visual Function Branch, National Eye Institute, Bethesda, Maryland, USA. [5]Department of Ophthalmology and Visual Sciences, University of Michigan Kellogg Eye Center, Ann Arbor, Michigan, USA. [6]Neurobiology–Neurodegeneration and Repair Laboratory, National Eye Institute, US National Institutes of Health, Bethesda, Maryland, USA. [7]Institute of Human Genetics, University of Regensburg, Regensburg, Germany. [8]Southwest Eye Center, Stuttgart, Germany. [9]Center for Human Genetics Research, Vanderbilt University Medical School, Nashville, Tennessee, USA. [10]Department of Molecular Physiology and Biophysics, Vanderbilt University Medical School, Nashville, Tennessee, USA. [11]Department of Ophthalmology, Erasmus Medical Center, Rotterdam, The Netherlands. [12]Department of Epidemiology, Erasmus Medical Center, Rotterdam, The Netherlands. [13]Netherlands Consortium for Healthy Aging, Netherlands Genomics Initiative, The Hague, The Netherlands. [14]UCL Institute of Ophthalmology, University College London, London, UK. [15]Moorfields Eye Hospital, London, UK. [16]Department of Medical Genetics, Cambridge Institute for Medical Research, University of Cambridge, Cambridge, UK. [17]Cambridge University Hospitals National Health Service (NHS) Foundation Trust, Cambridge, UK. [18]Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania, USA. [19]Department of Health Promotion and Development, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania, USA. [20]Department of Ophthalmology and Visual Sciences, John A. Moran Eye Center, University of Utah, Salt Lake City, Utah, USA. [21]Retina Service and Ophthalmology, Harvard Medical School, Massachusetts Eye and Ear Infirmary, Boston, Massachusetts, USA. [22]Texas Biomedical Research Institute, San Antonio, Texas, USA. [23]Centre for Eye Research Australia, University of Melbourne, Royal Victorian Eye and Ear Hospital, East Melbourne, Victoria, Australia. [24]Department of Ophthalmology, Shiley Eye Center, University of California, San Diego, La Jolla, California, USA. [25]Institute for Genomic Medicine, University of California, San Diego, La Jolla, California, USA. [26]Department of Pediatrics, The Hospital for Sick Children, Toronto, Ontario, Canada. [27]Program in Cell Biology, The Hospital for Sick Children, Toronto, Ontario, Canada. [28]Macular Degeneration Center, Casey Eye Institute, Oregon Health & Science University, Portland, Oregon, USA. [29]Section on Biomedical Genetics, Department of Medicine, Boston University Schools of Medicine and Public Health, Boston, Massachusetts, USA. [30]Department of Epidemiology, Boston University School of Public Health, Boston, Massachusetts, USA. [31]Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts, USA. [32]Department of Neurology, Boston University School of Medicine, Boston, Massachusetts, USA. [33]Department of Ophthalmology, Boston University School of Medicine, Boston, Massachusetts, USA. [34]Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania, USA. [35]Department of Ophthalmology, Jules Stein Eye Institute, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California, USA. [36]John P. Hussman Institute for Human Genomics, University of Miami Miller School of Medicine, Miami, Florida, USA. [37]Division of Epidemiology and Clinical Applications, National Eye Institute, US National Institutes of Health, Bethesda, Maryland, USA. [38]Department of Ophthalmology and Human Genetics, University of Pennsylvania Medical School, Philadelphia, Pennsylvania, USA. [39]These authors contributed equally to this work. [40]These authors jointly directed this work. Correspondence should be addressed to G.R.A. (goncalo@umich.edu).

ONLINE METHODS

Study samples. Macular degeneration cases and controls were recruited at ophthalmology clinics at the University of Michigan and the University of Pennsylvania and through the AREDS, as previously described. For replication, we contacted members of the International AMD Genetics Consortium; their samples are described in Fritsche *et al.*[13]. All participants provided informed consent allowing for the collection of genetic data, and all data contributors obtained approval from their local institutional review boards before generating genetic data. Our discovery sample, with ~2,350 sequenced cases and ~750 sequenced controls, provides 90% power to discover variants with a frequency of 0.1% and an associated relative risk of 19.2 or greater (similar to the p.Arg1210Cys variant in CFH) at significance level $\alpha = 0.00005$, which corresponds to an adjustment for the analysis of 1,000 independent coding variants.

Sequence production and quality control. Illumina multiplexed libraries were constructed according to the manufacturer's protocol with modifications: (i) DNA was fragmented using a Covaris E220 DNA Sonicator to range in size between 100 and 400 bp, (ii) Illumina adaptor-ligated library fragments were amplified in 4 50-µl PCR runs for 18 cycles, and (iii) Solid-Phase Reversible Immobilization (SPRI) bead cleanup was used for enzymatic purification and final library size selection targeting 300- to 500-bp fragments. Samples were pooled in groups of 4–24 before hybridization. A custom targeted probe set of 150-bp probes was designed (Agilent Technologies) and captured 0.97 Mb of sequence. The concentration of each captured library pool was determined through quantitative PCR (Kapa Biosystems) to produce cluster counts appropriate for the Illumina Genome Analyzer IIx and HiSeq 2000 platforms. We generated approximately 1.7 Gb of sequence per sample, covering 80% of the targeted space at a depth of >20×. Reads were aligned to the NCBI37/hg19 reference sequence using Burrows-Wheeler Aligner (BWA)[38]. Where pre-existing genotype information was available, sample identity was confirmed by comparing sequence data with pre-existing array data.

Quality control and variant calling. Quality control steps for all BAM files included removal of duplicated reads; recalibration of base qualities[39]; generation of diagnostic graphs and evaluation of sequencing quality (QPLOT; see URLs); and checks for DNA contamination[40]. After removing samples with high contamination, unexpected relatedness or high discordance rate, we retained 2,335 cases and 789 controls for an initial round of analysis. We calculated the sequencing depth using reads with mapping quality of >30 and bases with quality of >20. Across the 966,607-bp target region, we retained an average of 123,221,974 bases per individual (127.5× average coverage). Within targeted regions, 98.49% of the protein-coding exons had coverage of >10×.

We performed the variant calling step using UMAKE[23]. Genotype calling and polymorphism discovery were attempted across the original target ±50 bp. To remove low-quality variants, we excluded (i) sites with average depth of <0.5 or >500; (ii) sites with evidence of strand bias or cycle bias; (iii) sites within 5 bp of a 1000 Genomes Project indel; and (iv) sites with excess heterozygosity. These filters excluded 15,219 low-quality variants. The transition-transversion ratio (Ts/Tv) for the remaining 31,527 sites was 2.10. Concordance rates between sequencing-based genotypes in 13 duplicates were 99.82% when depth was >10×. Concordance with array-based genotypes[20] was 98.99% when depth was >10×.

Overall, 59.8% of discovered variants were newly identified (compared to dbSNP135 and the 1000 Genomes Project). On average, each sample carried 40 synonymous variants, 34 nonsynonymous variants and 1 nonsense variant.

Initial analyses. We first performed single-variant association tests using Fisher's exact test. This analysis confirmed strong association for common variants near the *CFH*, *C2*, *ARMS2* and *C3* genes. An initial examination of rare variants suggested that some signals were shadows of common variants with larger effects, so we focused on those signals where association remained significant after accounting for nearby common variants. Conditional signals were evaluated by exact logistic regression[41,42]. Three coding variants had conditional exact *P* values < 0.01 (all also had marginal *P* values < 0.01).

Augmenting our sample. We sought ancestry-matched controls among samples sequenced in ESP. First, we used genome-wide reads to infer sample ancestries on a worldwide population map. Briefly, we first generated a genetic ancestry PCA space using genotyped reference samples (such as those from the Human Genome Diversity Panel). Then, we generated a series of sample-specific genetic ancestry PCA data that were calibrated to the exact sequencing depth and coverage pattern of each sample and included the reference samples together with a single sequenced sample. Finally, we transformed sample-specific PCA coordinates onto the original map using Procrustes analysis. This procedure generates a metric (Procrustes similarity) that summarizes the similarity of reference sample placements using array genotypes to placements using sequencing data, and we only considered samples where this metric was >0.95 as candidates for matching. Second, we used a procedure inspired by propensity score matching to pair cases and controls[43]. Briefly, this procedure uses logistic regression to predict the probability that an individual is a case using the four principal components of ancestry as predictors and disease status as the outcome. This estimated probability of being a case for each sample is a propensity score and can be used to match cases and controls. For matching, we used a greedy algorithm to match cases and controls, allowing matches when the respective propensity scores differed by <0.0001. An alternative matching algorithm that matched cases and controls mapping close together in principal-component space according to the Euclidean distance between them gave similar results (association at p.Lys155Gln had OR = 2.68; exact $P = 4.5 \times 10^{-5}$ using Fisher's exact test).

To avoid artifacts from variant calling, we applied very stringent filters to both the AMD study and ESP study call sets. For both studies, we examined only sites with call rates of >90% and Phred-scaled variant quality scores of >30 that passed all study-specific quality control filters, had depth of >10× for >90% of the samples in the AMD or ESP call sets and were >5 bp from a 1000 Genomes Project indel. Primers used to confirm the presence of the mutation encoding p.Lys155Gln by Sanger sequencing are given in **Supplementary Table 6**.

Analyses using the combined AMD and ESP data set. As in our initial analysis, we first applied Fisher's exact test for association with all variants. Next, we examined variants with frequency of <1% for which signal remained significant after adjusting for common variants. This analysis highlighted p.Arg1210Cys encoded by *CFH* and p.Lys155Gln encoded by *C3* (**Fig. 1**).

Linkage disequilibrium analysis. To search for variants that might explain the signal encoding p.Lys155Gln, we evaluated linkage disequilibrium between the variant encoding p.Lys155Gln and all variants within 1 Mb, both within the samples sequenced for this experiment and also in preliminary whole-genome sequence data for 600 individuals (300 macular degeneration cases and 300 controls; A.S., D.S. and G.R.A., unpublished data). This analysis did not find variants in strong linkage disequilibrium in the nearby region. The variant was only present in one 1000 Genomes Project sample, which did not allow for reliable estimates of linkage disequilibrium.

Segregation analysis. In a segregation analysis, one identifies probands who carry p.Lys155Gln and then evaluates the probability that they transmit the variant to affected relatives (under the null hypothesis, we would expect to find the variant in 50% of the first-degree relatives of a carrier). We genotyped 471 pedigrees with multiple affected individuals. In each pedigree where p.Lys155Gln was found in more than one affected individual, we selected the nuclear family with the largest number of affected individuals. We recorded the number of affected individuals (N) and the number of carriers of p.Lys155Gln (C). Then, to average over possible choices of proband, we assigned each family a weight of C/N (this is the probability that a randomly selected proband in the family carries p.Lys155Gln) and then scored the number of affected first-degree relatives ($N - 1$) and carriers among those ($C - 1$). The estimated fraction of carriers among affected first-degree relatives of a proband was then calculated by summing $C/N \times (C - 1)$ and $C/N \times (N - 1)$ over families and taking the ratio of the two quantities.

doi:10.1038/ng.2758

38. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
39. McKenna, A. *et al*. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
40. Jun, G. *et al.* Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am. J. Hum. Genet.* **91**, 839–848 (2012).
41. Cox, D.R. & Shell, E.J. *Analysis of Binary Data* 2nd edn. (CRC Press, New York, 1989).
42. Hirji, K.F., Mehta, C.R. & Patel, N.R. Computing distributions for exact logistic-regression. *J. Am. Stat. Assoc.* **82**, 1110–1117 (1987).
43. Rosenbaum, P.R. & Rubin, D.B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55 (1983).