



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

## Citation

Zhan, X., D. E. Larson, C. Wang, D. C. Koboldt, Y. V. Sergeev, R. S. Fulton, L. L. Fulton, et al. 2013. "Identification of a Rare Coding Variant in Complement 3 Associated with Age-related Macular Degeneration." *Nature genetics* 45 (11): 10.1038/ng.2758. doi:10.1038/ng.2758. <http://dx.doi.org/10.1038/ng.2758>.

## Published version

<https://doi.org/10.1038/ng.2758>

## Link

<http://nrs.harvard.edu/urn-3:HUL.InstRepos:12406840>

## Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material (LAA), as set forth at

<https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFIOWI2M2EwYTg>

## Accessibility

<https://accessibility.huit.harvard.edu/digital-accessibility-policy>

## Share Your Story

The Harvard community has made this article openly available. Please share how this access benefits you. [Submit a story](#)

Published in final edited form as:

*Nat Genet.* 2013 November ; 45(11): . doi:10.1038/ng.2758.

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

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

<sup>1</sup>Center for Statistical Genetics, Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI <sup>2</sup>The Genome Institute, Washington University School of Medicine, St. Louis, MO <sup>3</sup>Department of Biostatistics, Harvard School of Public Health, Boston, MA <sup>4</sup>Ophthalmic Genetics and Visual Function Branch, National Eye Institute, Bethesda, MD <sup>5</sup>Department of Ophthalmology and Visual Sciences, University of Michigan Kellogg Eye Center, Ann Arbor, MI <sup>6</sup>Neurobiology-Neurodegeneration and Repair Laboratory, National Eye Institute,

Correspondence: goncalo.umich.edu (G.R.A.).

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

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

### Computer Software

LASER software for estimation of genetic ancestry can be obtained from

<http://genome.sph.umich.edu/wiki/LASER>

UMAKE / GotCloud tools for variant calling can be obtained from

<http://genome.sph.umich.edu/wiki/GotCloud>

<http://genome.sph.umich.edu/wiki/UMAKE>

### Author Contribution Statement

R.K.W., J.R.H., E.Y.C., D.S., E.R.M., A.S., G.R.A. conceived, designed and supervised the experiments, X.Z., G.R.A. wrote the initial version of the paper, X.Z., D.E.L., C.W., D.C.K. analyzed the data, D.E.L., D.C.K., R.S.F., L.L.F., 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., D.L. contributed data and analysis tools, M.B., R.R., 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 AMD family members sporadic cases and controls and collected peripheral blood samples for the Regensburg study, L.M.O., M.A.P.-V., J.L.H. provided results and analysis for the Vanderbilt/Miami samples, G.H.S.B., A.H., C.M.v.D., C.C.W.K. provided results and analysis for samples from the Rotterdam Study, Erasmus Medical Center, V.C., A.T.M., H.S., J.R.W.Y. provided results and analysis for the Cambridge AMD Study samples, Y.J., Y.P.C., D.E.W., M.B.G. provided results and analysis for the UCLA/University of Pittsburgh samples, D.J.M., L.A.F., M.M.D. provided results and analysis for the Utah samples, M.P.J., J.B., M.L.K. provided results and analysis for the Oregon Health Sciences Center samples, S.C., A.J.R., R.H.G., P.N.B. provided results and analysis for the University of Melbourne samples, H.L., H.O., M.M.Z., K.Z. provided results and analysis for the University of California, San Diego samples, C.L., F.G.P. provided results and analysis for a cohort of Atypical Hemolytic Uremic Syndrome patients, B.H.F.W. involved in design and planning of 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.

National Institutes of Health, Bethesda, MD <sup>7</sup>Institute of Human Genetics, University of Regensburg, Regensburg, Germany <sup>8</sup>Southwest Eye Center, Stuttgart, Germany <sup>9</sup>Center for Human Genetics Research, Vanderbilt University Medical School, Nashville, TN <sup>10</sup>Department of Molecular Physiology and Biophysics, Vanderbilt University Medical School, Nashville, TN <sup>11</sup>Department of Ophthalmology, Erasmus Medical Center, Rotterdam, The Netherlands <sup>12</sup>Department of Epidemiology, Erasmus Medical Center, Rotterdam, The Netherlands <sup>13</sup>Netherlands Consortium for Healthy Aging, Netherlands Genomics Initiative, the Hague, the Netherlands <sup>14</sup>UCL Institute of Ophthalmology, University College London, London, United Kingdom <sup>15</sup>Moorfields Eye Hospital, London, United Kingdom <sup>16</sup>Department of Medical Genetics, Cambridge Institute for Medical Research, University of Cambridge, Cambridge <sup>17</sup>Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK <sup>18</sup>Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA <sup>19</sup>Department of Health Promotion and Development, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA <sup>20</sup>Department of Ophthalmology and Visual Sciences, John A Moran Eye Center, University of Utah, Salt Lake City, UT <sup>21</sup>Retina Service and Ophthalmology, Harvard Medical School, Massachusetts Eye and Ear Infirmary, Boston, MA <sup>22</sup>Texas Biomedical Research Institute, San Antonio, TX <sup>23</sup>Centre for Eye Research Australia, University of Melbourne, Royal Victorian Eye and Ear Hospital, East Melbourne, Victoria, Australia <sup>24</sup>Department of Ophthalmology, Shiley Eye Center, University of California, San Diego, La Jolla, CA <sup>25</sup>Institute for Genomic Medicine, University of California, San Diego, La Jolla, CA <sup>26</sup>Department of Pediatrics, The Hospital for Sick Children, Toronto, Ontario, Canada <sup>27</sup>Program in Cell Biology, The Hospital for Sick Children, Toronto, Ontario, Canada <sup>28</sup>Macular Degeneration Center, Casey Eye Institute, Oregon Health & Science University, Portland, OR <sup>29</sup>Section on Biomedical Genetics, Department of Medicine, Boston University Schools of Medicine and Public Health, Boston, MA <sup>30</sup>Departments of Epidemiology, Boston University Schools of Public Health, Boston, MA <sup>31</sup>Departments of Biostatistics, Boston University Schools of Public Health, Boston, MA <sup>32</sup>Departments of Neurology, Boston University Schools of Medicine, Boston, MA <sup>33</sup>Departments of Ophthalmology, Boston University Schools of Medicine, Boston, MA <sup>34</sup>Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA <sup>35</sup>Department of Ophthalmology, Jules Stein Eye Institute, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA <sup>36</sup>John P. Hussman Institute for Human Genomics, University of Miami Miller School of Medicine, Miami, FL <sup>37</sup>Division of Epidemiology and Clinical Applications, National Eye Institute/National Institutes of Health, Bethesda, MD <sup>38</sup>Department of Ophthalmology and Human Genetics, University of Pennsylvania Medical School, Philadelphia, PA

## Abstract

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 revealed two large-effect rare variants; previously described R1210C in the *CFH* gene ( $f_{\text{case}} = 0.51\%$ ,  $f_{\text{control}} = 0.02\%$ , OR = 23.11), and newly identified K155Q in the *C3* gene ( $f_{\text{case}} = 1.06\%$ ,  $f_{\text{control}} = 0.39\%$ , OR = 2.68). The variants suggest decreased inhibition of C3 by Factor H, resulting in increased activation of the alternative complement pathway, as a key component of disease biology.

---

Genetic and environmental factors contribute to age-related macular degeneration (AMD)<sup>1,2</sup>, a major cause of vision loss in elderly individuals<sup>3</sup>. Pioneering discovery of association of AMD with complement factor H (*CFH*<sup>4-6</sup>) was quickly followed by the

identification of additional susceptibility loci that now include *ARMS2/HTRA1*<sup>7,8</sup> and complement genes *C3*, *C2/CFB* and *CFP*<sup>12</sup>. Genome-wide association studies (GWAS) of AMD cases and controls have now revealed common susceptibility variants at ~20 different loci<sup>13,14</sup> and begun to uncover specific cellular pathways involved in AMD biology.

While common variants tag the associated genomic region, rare coding variants can provide more specific clues about the underlying disease mechanism<sup>15</sup>. For example, rare variant R1210C in the *CFH* gene was recently associated with a large increase in AMD risk using targeted sequencing of rare *CFH* risk haplotypes<sup>16</sup>. The resulting altered protein has decreased binding to C3b, C3d, heparin and endothelial cells<sup>17–19</sup>. A reduction in *CFH*'s ability to inactivate C3, leading to increased cell killing activity by the complement pathway, could contribute to AMD – 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<sup>20</sup> (near *CFH*, *ARMS2*, *C3*, *C2/CFB*, *CFI*, *CETP*, *LIPC* and *TIMP3/SYN3*) and two candidate regions (*LPL* and *ABCA1*) (Supplementary Table 1). We re-sequenced these regions in 3,124 individuals (2,335 cases and 789 controls) recruited in ophthalmology clinics at the University of Michigan and at the University of Pennsylvania and among Age-Related Eye Disease Study (AREDS) participants<sup>20,21</sup> ENREF\_17. Genomic targets were enriched using a set of 150-bp probes designed by Agilent Technologies, and sequence data was generated on Illumina Genome Analyzer and HiSeq instruments. The ten 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 bases of on-target sequence per individual (127.5× average depth counting bases with quality >20 in reads with mapping quality >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 NHLBI Exome Sequencing Project data<sup>22</sup> (Supplementary Table 2). We identified an average of 1,714 non-reference sites in each sequenced individual. In total, this resulted in 31,527 single nucleotide variants of which 18,956 were not in dbSNP 135. Discovered sites included 834 synonymous variants, 1,379 nonsynonymous variants and 43 nonsense variants, most of which were extremely rare (see Supplementary Table 3). Among 13 samples sequenced in duplicate, genotype concordance was 99.82% (when depth >10×). Among 908 samples previously examined with GWAS arrays<sup>20</sup>, sequence-based genotypes were 98.99% concordant with array-based calls (again, when depth >10×).

In an initial comparison of AMD cases and controls (see Supplementary Table 4), no rare coding variants with frequency <1% reached experiment wide significance ( $p < 0.05 / 31,527 = 1.6 \times 10^{-6}$ , including all discovered variants, or  $p < 0.05 / 1,422 = 3.5 \times 10^{-5}$  considering only protein altering variants), although several showed encouraging patterns. For example, rare variant R1210C 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 with case frequency  $f_{\text{case}} = 12\%$ , control frequency  $f_{\text{control}} = 27\%$ , P-value =  $2.52 \times 10^{-48}$ ), *ARMS2* (rs10490924,  $f_{\text{case}} = 33\%$ ,  $f_{\text{control}} = 18\%$ , P-value =  $5.48 \times 10^{-27}$ ), *C3* (rs2230199,  $f_{\text{case}} = 25\%$ ,  $f_{\text{control}} = 17\%$ , P-value =  $3.94 \times 10^{-9}$ ) and *C2/CFB* (rs556679,  $f_{\text{case}} = 7\%$ ,  $f_{\text{control}} = 12\%$ , 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 the NHLBI Exome Sequencing Project (ESP)<sup>23</sup>, which sequenced 15,336 genes across 6,515 individuals. Sequence data for our samples and the NHLBI Exome Sequencing Project samples were analyzed with the same analysis pipeline, which minimized potential differences due to heterogeneity in analysis tools and parameters. To further avoid sequencing and variant calling artifacts, 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<sup>24</sup>. Since careful matching of genetic ancestry is critical for rare variant association studies<sup>24,25</sup>, we selected an ancestry-matched subset of our samples and of samples from the NHLBI Exome Sequencing Project. We used principal component analysis to construct a genetic ancestry map of the world with samples from the Human Genome Diversity Project, each genotyped at 632,958 SNPs<sup>26</sup>. If GWAS array genotypes were available for our samples and for the NHLBI Exome Sequencing Project samples, it would be straightforward to place them directly in 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 principal component analysis. Thus, we relied on the new LASER algorithm (Wang and Abecasis, personal communication) to place each sequenced sample in a pre-defined 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 Exome Sequence data, which have average off-target coverage of ~0.23× and ~0.90×, respectively (see Supplementary Figures 1A, 1B, 1E and 1F, which show that 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; see **Online Methods**) and used a greedy algorithm to match cases and controls based on estimated genetic ancestry. As shown in the **Online Methods**, alternative matching algorithms do not alter our conclusions. After matching, we focused on a set of 2,268 AMD cases and 2,268 controls, ancestry-matched one-to-one (Supplementary Figure 1C and 1G). Since AMD phenotype information was not available for most controls, we expect that a small proportion may eventually develop disease; however, this should not impact power substantially<sup>27</sup>. After matching case-control samples, we excluded 1 variant with Hardy-Weinberg Equilibrium test p-value <10<sup>-6</sup> and focused our analysis on 430 protein changing variants in regions that were targeted and deeply sequenced in both experiments as well as far away from insertion deletion polymorphisms.

In this expanded analysis (see Table 1), common variant signals at all loci increased in significance (in comparison to Supplementary Table 4). In addition, two rare coding variants exhibited association with  $p < 0.01$ . The first was R1210C in the *CFH* gene (observed in one control and 23 cases, OR = 23.11,  $p_{\text{exact}} = 2.9 \times 10^{-6}$ ), providing strong support for the original report<sup>16</sup>. The second variant was K155Q in the *C3* gene (18 controls, 48 cases, OR = 2.68,  $p_{\text{exact}} = 2.7 \times 10^{-4}$ ; Supplementary Figure 1D and 1H for carrier ancestry distribution). When controlling for a previously described common variant signal nearby, rs2230199 ( $f_{\text{control}} = 20.63\%$ ,  $f_{\text{case}} = 25.26\%$ , marginal  $p_{\text{exact}} = 1.8 \times 10^{-7}$ , OR = 1.31), the evidence for association with K155Q increased slightly (conditional OR = 2.91,  $p_{\text{exact}} = 2.8 \times 10^{-5}$ ). Inspection of the raw read data shows the variant is well supported and is unlikely to be a sequencing or alignment artifact, a result further confirmed by Sanger sequencing (see Supplementary Figures 2, 3 and 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 K155Q 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 Figure 5)<sup>28–30</sup>.

To confirm the K155Q signal, we genotyped additional samples totaling 4,526 cases and 3,787 controls and, again, observed strong association ( $f_{\text{control}} = 0.5\%$ ,  $f_{\text{case}} = 1.3\%$ ,  $p_{\text{follow-up}} = 7.7 \times 10^{-7}$ ,  $p_{\text{combined}} = 1.1 \times 10^{-9}$ , Table 2). In addition, we genotyped 471 families with multiple AMD cases to identify 18 nuclear families where K155Q segregates. These families included 49 affected individuals, where at least one individual carries K155Q and, adjusting for ascertainment, we estimate that 75% of first degree relatives of a K155Q carrier who also have AMD will carry the variant, consistent with an OR of  $\sim 3$  (Supplementary Table 5 and **Online Material**). Further strong evidence for association of this variant with macular degeneration is provided in independent work by deCODE Genetics<sup>31</sup>, examining 1,143 Icelandic macular degeneration cases and 51,435 Icelandic controls (control frequency 0.55%, OR = 3.45,  $p_{\text{deCODE}} = 1.1 \times 10^{-7}$ ,  $p_{\text{combined}} = 1.6 \times 10^{-15}$ ). In 1,606 directly genotyped cases of macular degeneration from the Age Related Disease Study II<sup>32</sup> the variant has frequency 1.77%, similar to our sequenced AMD cases (frequency 1.10%) and our follow-up AMD cases (1.30%) and is notably higher than in our sequenced controls (0.30%), our genotyped controls (0.50%), in NHLBI Exome Sequencing Project participants with primarily European Ancestry (0.40%) and in deCODE controls (0.55%). We found no evidence of the K155Q variant in a small sample of patients with atypical haemolytic-uremic syndrome (aHUS, n=53), a rare disorder whose genetic risk factors partially overlap with macular degeneration.

We next investigated the potential functional consequences of the K155Q variant *in silico*. Based on protein crystallography, the model in Figure 1 shows that CFH variant R1210C (OR=23.11), C3 variant K155Q (OR=2.91) and C3 variant R102G (OR=1.31) all map near the surface where CFH and C3b interact and suggests they might affect binding of complement factor H to C3b. Factor H inhibits C3b and limits immune responses mediated by the alternative complement pathway. We hypothesize that K155Q and R102G affect binding of the first macro-globular domain of C3 to CFH and thus interferes with inactivation of the alternative complement pathway\_ENREF\_31, a hypothesis that must be confirmed experimentally<sup>33</sup>. Interestingly, the three variants (R102G and K155Q in C3 and R1210C in CFH) all are associated with replacement of a positively charged residue.

In summary, our work and the companion paper identify K155Q as a rare C3 variant associated with a  $\sim 2.91$ -fold increased risk of macular degeneration. Together with rare CFH variant R1210C and previously described common C3 variant R102G, K155Q may reduce binding of CFH to C3b, inhibiting the ability of Factor H to inactivate the alternative complement pathway. Clarifying the mechanistic impact of K155Q is likely to be challenging, as illustrated by contradictory results of previous functional follow-up of AMD loci<sup>34–36</sup>, but functional studies of complement activity suggest potential next steps<sup>33,37</sup>. Our work relied on targeted sequencing of GWAS 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 to many targeted sequencing studies, but the strictness of matching and variant filtering required for preventing 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 utility of the approach will grow – making it possible to match multiple controls to each case and to focus on

progressively finer ancestry matches. Our results also emphasize the challenges and the large sample sizes will be required for rare variant studies of complex human traits, as well as the promise of these studies to highlight disease biology, as illustrated by the interaction between Factor H and C3b that is suggested as a key factor in AMD biology here.

## 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 Age Related Eye Diseases Study, as previously described. For replication, we contacted members of the International AMD Genetics Consortium; their samples are described in Fritsche et al<sup>13</sup>. All participants provided informed consent allowing for collection of genetic data and all data contributors obtained approval from their local Institutional Review Boards before generating genetic data. Our discovery sample, with ~2350 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 CFH R1210C) at significance level  $\alpha = 0.00005$ , which corresponds to an adjustment for analysis of 1,000 independent coding variants.

### Sequence production and quality control

Illumina multiplexed libraries were constructed according to the manufacturer's protocol (Illumina Inc, San Diego, CA) with modifications: 1) DNA was fragmented using a Covaris E220 DNA Sonicator (Covaris, Inc. Woburn, MA) to range in size between 100 and 400bp. 2) Illumina adapter-ligated library fragments were amplified in four 50 $\mu$ L PCR reactions for eighteen cycles. 3) Solid Phase Reversible Immobilization (SPRI) bead cleanup was used for enzymatic purification and final library size selection targeting 300–500bp fragments. Samples were pooled in groups of 4–24 before hybridization. A custom targeted probe set of 150bp probes was designed (Agilent Technologies, Santa Clara, CA) and captured 0.97 Mb of sequence. The concentration of each captured library pool was determined through qPCR (Kapa Biosystems, Inc, Woburn, MA) to produce cluster counts appropriate for the Illumina GAIIx and HiSeq 2000 platforms. We generated approximately 1.7Gb of sequence per sample, covering 80% of the targeted space at depth  $>20\times$ . Reads were aligned to the NCBI37/hg19 reference sequence using BWA<sup>38</sup>. 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<sup>39</sup>; generation of diagnostic graphs and evaluation of sequencing quality<sup>40</sup>; checks for DNA contamination<sup>41</sup>. After removing samples with high contamination, unexpected relatedness or with 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  $>30$  and bases with quality  $>20$ . Across the 966,607 base pair target region, we retained an average 123,221,974 bases per individual ( $127.5\times$  average coverage). Within targeted regions, 98.49% of the protein coding exons had coverage  $>10\times$ .

We performed variant calling step using UMAKE<sup>23</sup>. Genotype calling and polymorphism discovery were attempted across the original target  $\pm 50$  basepairs. To remove low quality variants, we excluded: 1) sites with average depth  $<0.5$  or  $>500$ ; 2) sites with evidence of strand bias or cycle bias; 3) sites within 5 basepairs of a 1000 Genomes Project indel; 4) sites with excess heterozygosity. These filters excluded 15,219 low quality variants. The

transition-transversion ratio (Ts/Tv) for the remaining 31,527 site was 2.10. Concordance rates between sequence-based genotypes in 13 duplicates were 99.82% when depth >10×. Concordance with array-based genotypes<sup>20</sup> was 98.99% when depth >10×.

59.8% of discovered variants are novel (versus dbSNP 135 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 *CFH*, *C2*, *ARMS2* and *C3* genes. An initial examination of rare variants suggested some signals were shadows of common variants with larger effects, so we focused on those where association remained significant after accounting for nearby common variants. Conditional signals were evaluated by exact logistic regression<sup>42,43</sup>. Three coding variants had conditional exact P-value less than 0.01 (all also had marginal p-values < 0.01).

### Augmenting our sample

We sought ancestry matched controls among samples sequenced in the ESP project. 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 that are calibrated to the exact sequencing depth and coverage pattern of each sample and include the reference samples together with a single sequenced sample. Finally, we transformed sample specific PCA coordinates to the original map using Procrustes analysis. This procedure generates a metric (the Procrustes similarity) that summarizes similarity of reference sample placements using array genotypes to placements using sequence data and we only considered samples where this metric was >0.95 as candidates for matching. Second, we used a procedure inspired on propensity score matching to pair cases and controls<sup>44</sup>. 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 <.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 K155Q had OR=2.68, exact p-value  $4.5 \times 10^{-5}$  using Fisher's exact test).

To avoid variant calling artifacts, 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 >90%, Phred-scaled variant quality scores >30, passing all study specific quality control filters, with depth >10× for >90% of the samples in the AMD or ESP callsets, and >5-bp from a 1000 Genomes Project indel. Primers used to confirm the presence of K155Q by Sanger Sequencing are given in Supplementary Table 6.

### Analyses using the combined AMD and ESP data set

Similar to our initial analysis, we first applied Fisher's exact test for association to all variants. Next, we examined variants with frequency <1% for which signal remained significant after adjusting for common variants. This analysis highlighted R1210C in *CFH* and K155Q in *C3* (Figure 1).

## Linkage disequilibrium analysis

To search for variants that might explain the signal at K155Q, we evaluated linkage disequilibrium between K155Q 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, 300 controls; Swaroop, Stambolian and Abecasis, personal communication). This analysis did not find variants in strong linkage disequilibrium in the nearby region. The variant is only present in one 1000 Genomes Project sample, which does not allow for reliable estimates of linkage disequilibrium.

## Segregation analysis

In a segregation analysis, one identifies probands who carry K155Q and then evaluates the probability that they transmit the variant to affected relatives (under the null, we would expect to find the variant in 50% of first degree relatives of a carrier). We genotyped 471 pedigrees with multiple affected individuals. In each pedigree where K155Q was found in >1 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 K155Q carriers (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 K155Q) and then scored the number of affected first degree relatives (N-1) and of carriers among those (C-1). The estimated fraction of carriers among affected first degree relatives of a proband is then calculated by summing C/N \* (C-1) and C/N \* (N-1) over families and taking the ratio of the two quantities.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

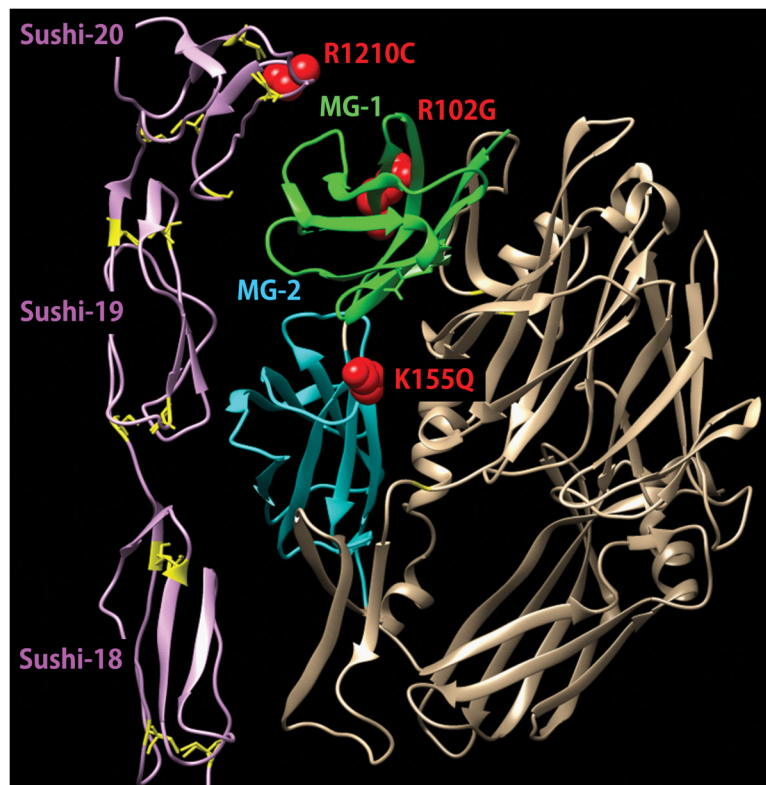
We thank all study participants for their generous volunteering. We thank Drs. Bingshan Li, Wei Chen, Carlo Sidore, Tanya Teslovich, Lars Fritsche and Michael Boehnke for useful discussion and suggestions. This project was supported by grants from the National Institute of Health Grant (National Eye Institute, National Human Genome Research Institute, grants EY022005, HG007022, HG005552, EY016862, U54HG003079, EY09859), The Medical Research Council, UK (grant G0000067); the Deutsche Forschungsgemeinschaft (grant WE1259/19-2), the Alcon Research Institute, The Department of Health's NIHR Biomedical Research Centre for Ophthalmology at Moorfields Eye Hospital and UCL Institute of Ophthalmology, Research to Prevent Blindness (New York, NY), the Thome Memorial Foundation, the Harold and Pauline Price Foundation, the National Health and Medical Research Council of Australia (NHMRC) Clinical Research Excellence (grant 529923, NHMRC practitioner fellowship 529905, 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.

## References

1. Priya RR, Chew EY, Swaroop A. Genetic Studies of Age-related Macular Degeneration: Lessons, Challenges, and Opportunities for Disease Management. *Ophthalmology*. 2012; 119:2526–36. [PubMed: 23009893]
2. Swaroop A, Chew EY, Rickman CB, Abecasis GR. Unraveling a multifactorial late-onset disease: from genetic susceptibility to disease mechanisms for age-related macular degeneration. *Annu Rev Genomics Hum Genet*. 2009; 10:19–43. [PubMed: 19405847]
3. Friedman DS, et al. Prevalence of age-related macular degeneration in the United States. *Archives of Ophthalmology*. 2004; 122:564–72. [PubMed: 15078675]
4. Haines JL, et al. Complement factor H variant increases the risk of age-related macular degeneration. *Science*. 2005; 308:419–21. [PubMed: 15761120]

5. Edwards AO, et al. Complement factor H polymorphism and age-related macular degeneration. *Science*. 2005; 308:421–4. [PubMed: 15761121]
6. Klein RJ, et al. Complement factor H polymorphism in age-related macular degeneration. *Science*. 2005; 308:385–9. [PubMed: 15761122]
7. Jakobsdottir J, et al. Susceptibility genes for age-related maculopathy on chromosome 10q26. *Am J Hum Genet*. 2005; 77:389–407. [PubMed: 16080115]
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*. 2005; 14:3227–36. [PubMed: 16174643]
9. Yates JR, et al. Complement C3 variant and the risk of age-related macular degeneration. *N Engl J Med*. 2007; 357:553–61. [PubMed: 17634448]
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*. 2006; 38:458–62. [PubMed: 16518403]
11. Fagerness JA, et al. Variation near complement factor I is associated with risk of advanced AMD. *Eur J Hum Genet*. 2009; 17:100–4. [PubMed: 18685559]
12. Maller JB, et al. Variation in complement factor 3 is associated with risk of age-related macular degeneration. *Nat Genet*. 2007; 39:1200–1. [PubMed: 17767156]
13. Fritsche LG, et al. Seven new loci associated with age-related macular degeneration. *Nat Genet*. 2013; 45:433–9. 439e1–2. [PubMed: 23455636]
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*. 2011; 43:1001–4. [PubMed: 21909106]
15. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science*. 2009; 324:387–9. [PubMed: 19264985]
16. Raychaudhuri S, et al. A rare penetrant mutation in CFH confers high risk of age-related macular degeneration. *Nature Genetics*. 2011; 43:1232–6. [PubMed: 22019782]
17. Jozsi 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*. 2006; 17:170–7. [PubMed: 16338962]
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*. 2003; 111:1181–90. [PubMed: 12697737]
19. Ferreira VP, 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*. 2009; 182:7009–18. [PubMed: 19454698]
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 U S A*. 2010; 107:7401–6. [PubMed: 20385819]
21. Age-Related Eye Disease Study Research G. 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*. 2000; 107:2224–32. [PubMed: 11097601]
22. Tennessen JA, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*. 2012; 337:64–9. [PubMed: 22604720]
23. Fu W, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*. 2012
24. The 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012; 491:56–65. [PubMed: 23128226]
25. Mathieson I, McVean G. Differential confounding of rare and common variants in spatially structured populations. *Nat Genet*. 2012; 44:243–6. [PubMed: 22306651]
26. Li JZ, et al. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*. 2008; 319:1100–4. [PubMed: 18292342]

27. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007; 447:661–78. [PubMed: 17554300]
28. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *American Journal of Human Genetics*. 2008; 83:311–21. [PubMed: 18691683]
29. Price AL, et al. Pooled association tests for rare variants in exon-resequencing studies. *American Journal of Human Genetics*. 2010; 86:832–8. [PubMed: 20471002]
30. Wu MC, et al. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet*. 2011; 89:82–93. [PubMed: 21737059]
31. Helgason H, et al. A rare nonsynonymous sequence variant in C3 confers high risk of age-related macular degeneration. *Nat Genet*. 2013
32. Group AR, et al. The Age-Related Eye Disease Study 2 (AREDS2): study design and baseline characteristics (AREDS2 report number 1). *Ophthalmology*. 2012; 119:2282–9. [PubMed: 22840421]
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 U S A*. 2011; 108:8761–6. [PubMed: 21555552]
34. Fritsche LG, et al. Age-related macular degeneration is associated with an unstable ARMS2 (LOC387715) mRNA. *Nature Genetics*. 2008; 40:892–6. [PubMed: 18511946]
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 U S A*. 2007; 104:16227–32. [PubMed: 17884985]
36. Dewan A, et al. HTRA1 promoter polymorphism in wet age-related macular degeneration. *Science*. 2006; 314:989–92. [PubMed: 17053108]
37. Sanchez-Corral P, et al. Structural and functional characterization of factor H mutations associated with atypical hemolytic uremic syndrome. *Am J Hum Genet*. 2002; 71:1285–95. [PubMed: 12424708]
38. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009; 25:1754–60. [PubMed: 19451168]
39. McKenna A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010; 20:1297–303. [PubMed: 20644199]
40. Li B, et al. qplot: Quality Assessment Plots for Next Generation Sequence Runs. 2012
41. Jun G, et al. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am J Hum Genet*. 2012; 91:839–48. [PubMed: 23103226]
42. Cox DR, Shell EJ. *Analysis of Binary Data*. 1970
43. Hirji KF, Mehta CR, Patel NR. Computing Distributions for Exact Logistic-Regression. *Journal of the American Statistical Association*. 1987; 82:1110–1117.
44. Rosenbaum PR, Rubin DB. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*. 1983; 70:41–55.



**Figure 1.**

*C3* variants R102G and K155Q and *CFH* variant R1210C are in the interaction domains of the first alpha-macro-globular domain of C3b and CFH, respectively. The fragment of the crystal structure of the four Sushi domains (purple in figure, one not shown for clarity) of CFH in a complex with complement fragment C3b (PDB file: 2wii) was used to explore the effect of disease associated nonsynonymous changes. The CFH residues 987–1230 were used to generate the structure using the first four Sushi domains from 2wii as a structural template (shown in pink, with cysteine residue side chains in yellow). The C-terminal Sushi domains were docked to the binding site in C3b. The first two alpha-macro-globulin domains of C3b, MG-1 and MG-2, are shown in green and cyan, respectively. The location of mutations R102G, K155Q, and R1210C are marked in red.

Table 1

Summary association results for 2,268 sequenced AMD cases and 2,268 sequenced controls. Samples in this expanded analysis include our sequenced AMD samples and genetically matched controls, sequenced by us or by the NHLBI Exome Sequencing Project. The top coding variant in each locus is included in this table when  $p < 10^{-6}$ . Rare coding variants are included when the corresponding p-values for the conditional or marginal analysis are less than  $1 \times 10^{-4}$ .

SNP	Chromosome	Position (bp)	Nearest Gene	Consequence	Alleles (ref/alt)	Frequency (alt allele)		OR	P-value	Conditional P-value*
						Cases	Controls			
<b>Common variant hits</b>										
rs1061170	1	196,659,237	<i>CFH</i>	H402Y	C/T	0.478	0.623	0.555	$1.01 \times 10^{-43}$	
rs438999	6	31,928,306	<i>SKIV2L</i>	Q151R	A/G	0.058	0.098	0.566	$1.26 \times 10^{-12}$	
rs10490924	10	124,214,448	<i>ARMS2</i>	A69S	G/T	0.329	0.197	1.990	$1.04 \times 10^{-45}$	
rs2230199	19	6,718,387	<i>C3</i>	R102G	G/C	0.253	0.206	1.300	$1.58 \times 10^{-7}$	
<b>Rare variant hits MAF &lt; 1%, marginal and conditional P &lt; 0.01 (after conditioning on nearby common variants)</b>										
rs121913059	1	196,716,375	<i>CFH</i>	R1210C	C/T	0.005	0.000	23.11	$2.9 \times 10^{-6}$	$6.0 \times 10^{-4}$ (rs1061170)
rs147859257	19	6,718,146	<i>C3</i>	K155Q	T/G	0.011	0.004	2.68	$2.7 \times 10^{-4}$	$2.8 \times 10^{-5}$ (rs2230199)

All p-values were calculated using exact logistic regression. 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 (parenthesis).

Table 2

Follow-up Genotyping Summary and Meta-Analysis Summary.

SAMPLE SET	CONTROLS			CASES			P-value
	N	MAF	MAF	N	MAF	MAF	
<b>Discovery Sample</b>							
Sequenced Samples (N=4,536)	2,268	.004	.011	2,268	.011		$2.7 \times 10^{-4}$
<b>Follow-up Samples</b>							
Germany / University of Regensburg (N=2,976)	1,147	.006	.016	1,829	.016		$1.7 \times 10^{-3}$
United States / Vanderbilt/Miami (N=1,819)	726	.004	.007	1,093	.007		$3.5 \times 10^{-1}$
Netherlands / Rotterdam Study (N=1,409)	1,280	.005	.031	129	.031		$1.5 \times 10^{-4}$
United Kingdom / Cambridge AMD Study(N=1,279)	423	.006	.015	856	.015		$6.2 \times 10^{-2}$
United States / UCLA/University of Pittsburgh (N=830)	211	.004	.017	619	.017		$8.3 \times 10^{-4}$
<b>deCODE Study</b>							
deCODE Discovery Sample (N=52,578)	51,435	.005	*	1,143	*		$1.1 \times 10^{-7}$
<b>Meta-Analysis</b>							
All Follow-Up Samples (N=8,313)	3,787	.005	.013	4,526	.013		$7.7 \times 10^{-7}$
Discovery and All Follow-Up Samples (N=12,849)	6,055	.005	.013	6,794	.013		$1.1 \times 10^{-9}$
Discovery, All Follow-Up Samples, and deCODE (N=65,427)	57,490	.005	*	7,937	*		$1.6 \times 10^{-15}$

The table includes the number of cases and controls in each comparison, the corresponding allele frequency for the K155Q allele in each set of samples, and the p-value for a comparison of allele frequencies in the two groups. Meta-analysis p-values are calculated using Stouffer's method.