Sequence and expression of complement factor H gene cluster variants and their roles in age-related macular degeneration risk

Anne E. Hughes, PhD; Stephen Bridgett, PhD; Weihua Meng, PhD; Mingyao Li, PhD; Christine A. Curcio, PhD; Dwight Stambolian, MD, PhD; Declan T. Bradley, PhD.

Supplementary methods

BLAST study of regions of homology between *CFH* and *CFH*-related genes

We anticipated that correct mapping of genomic sequence in our massively-parallel sequencing experiment would be complicated by homologous sequences in the *CFH* gene region, so we conducted analyses to identify homologous sequences so that we could apply caution when interpreting sequence that was mapped to these areas.

We compared all exons of *CFH* and *CFH*-related genes by NCBI Basic Local Alignment Search Tool (BLAST) analysis, using default *megablast* parameters to identify highly similar sequences.[1] Several exons of *CFHR3*, *CFHR1* or *CFH* showed more than 94% homology (Figure S1). We also found strong homology between exon 1 of *CFHR3* and an upstream intergenic region between *CFH* and *CFHR3*. A putative gene, *LOC100996886*, (located between *CFHR1* and *CFHR4*) showed homology with exons of *CFHR3* and *CFH*. Of note was the exceptionally strong homology between *CFHR1*, *CFHR2* and the final exons of *CFH*. All genetic locations are numbered according to NCBI build 37 (GRCh37), which is identical to hg19 for all autosomes. Variants in genes are numbered according to the following Genbank accession numbers: *CFH* NM_000186; *CFHR3* NM_021023; *CFHR1* NM_002113; *CFHR4* NM_001201551; *CFHR2* NM_005666 and *CFHR5* NM_030787.[2]

Sequence capture and massively parallel sequencing

We aimed to achieve enrichment of all exons of *CFH* and all *CFH*-related genes for DNA samples of four patients with neovascular AMD. As part of a larger experiment with multiple purposes, we also prepared separate pools of DNA from patients with AMD and from patients with two other complex diseases. We quantified DNA by taking the mean of three readings using a NanoDrop spectrophotometer (Thermo Scientific). The final concentration was 30 ng/µl. We fragmented DNA by sonication using Bioruptor (Diagenode), size-selected to approximately 350 bp and indexed for Illumina sequencing using the Truseq indexing system and protocol.[3] We designed a customised Nimblegen SeqCap EZ capture library of long probes using the supplier's NimbleDesign software.[4] We targeted a total of 5 Mb in the capture, including the entire genomic region from 196,619,961 to 196,979,838, spanning *CFH* and *CFH*-related

genes. We followed Nimblegen's capture protocol for enrichment, with the exception of adding 1 extra cycle of ligation mediated-PCR (total 8 cycles). We used picogreen for DNA quantitation during library preparation, capture and evaluation. We merged five differently indexed libraries of DNA, including 4 of this study for a single capture reaction (15 µl), to which the each library contributed 20% of the total. We assessed capture efficiency by quantitative PCR using primers for a sequence included in the SeqCap EZ capture library. We performed post-capture PCR (18 cycles) before merging with additional libraries from other projects prepared with 8 different indices from two further captures. GATC Biotech (Konstanz, Germany) performed NGS on a HiSeq 2000 (Illumina, San Diego, CA) with 100 bp paired-end reads. Each *CFH* haplotype-specific library contributed 1% of the total reads.

Sequence analysis

We aligned genomic reads to hg19 human reference sequence with the Burroughs Wheeler Aligner (BWA) 0.5.9 *aln* algorithm,[5] and used SAMtools 0.1.14[6] for sorting, indexing, and removal of duplicate reads. We employed Genome Analysis Toolkit (GATK) to recalibrate, realign and to call polymorphisms.[7] All exons were covered adequately. Eleven kb of intronic regions and 16.5 kb of intergenic DNA were not covered (7.6%), most of which represented extremely low complexity or repetitive DNA. We viewed variants in *CFH* and related genes using IGV with an allelic threshold of 0.015.[8] In these analyses, homozygous differences from the reference sequence reflected haplotype-tagging polymorphisms, and unexpected heterozygosity indicated either rare SNPs, regions susceptible to incorrect mapping from a closely related sequence, or breakdown of homozygosity towards the 3` end of the captured region. We also mapped reads using NovoAlign (http://www.novocraft.com), allowing each read to map to all good matches within the cluster of *CFH* and *CFH*-related genes, in which case 'heterozygosity' often indicated differences between highly homologous sequences, and variation of allelic ratios allowed deletions or conversion events to be interpreted. Short regions of extraordinarily high read depth indicated low complexity repetitive elements.

Prediction of functional effect of coding changes

We predicted the effect of coding polymorphisms with PolyPhen-2 (Polymorphism Phenotyping v2) (software version 2.2.2; protein sequences from UniProtKB/UniRef100 Release 2011_12; structures from PDB/DSSP Snapshot 03-Jan-2012; UCSC MultiZ multiple alignments of 45 vertebrate genomes with hg19/GRCh37 human genome using the HumDiv Model and canonical transcripts).[9]

Retinal RNA-seq data

We aligned retinal RNA-seq reads from our (ML, CAC, DS) previous studies[10] to the human reference hg19 genome using GSNAP with default settings.[11] To reduce mapping errors, we used RNA-SeQC to remove reads were that had mapping quality <30, or where the pairs mapped to different chromosomes, had unexpected orientations, or distance between pairs was >500 kbp.[12] We used Samtools (version 0.1.19)[6] to sort reads; picard (version 1.121)[13] to mark duplicate reads; and the Samtools[6] 'depth' command to obtain read depths. A custom Perl script (available on request) summarised the minimum, maximum and average read depths in each exon for each sample. We calculated the proportion of full length *CFH* transcripts (FH) relative to total *CFH* (for both FH and FHL-1) combined by DerSimonian-Laird estimate random-effects meta-analysis using *metaprop* from the *meta* package in *R* 3.2.1.

Liver RNA-seq data

We accessed liver RNA-seq reads from three individuals from the EBI Illumina body map[14] and EBI Expression Atlas[15] and aligned them to the human reference hg19 chromosome 1 using STAR aligner (version 2.4.0f1).[16] We sorted the resulting alignments with Samtools before extracting the regions of chromosome 1 between 190 and 200 Mb from the BAMs for viewing in IGV.[6, 8]

Secondary analysis of AMD genome-wide associations study

To investigate the effect of polymorphisms on progression from drusen (which are common and impair sight minimally) to neovascular AMD, we conducted a genome-wide case-case study analysis to compare individuals with neovascular AMD to individuals who had drusen only from the Chen *et al.* AMD study[17] .

To investigate their roles in progression of the four *CFH* haplotypes and a representative polymorphism from each of the other known major AMD loci (*CFB, C3 and HTRA1*)[18-24] we also conducted a candidate gene study. For reference, we also compared these individuals to individuals designated as disease-free.

The Chen et al study included of 2,157 cases with AMD (neovascular AMD, geographic atrophy or drusen) and 1,150 disease-free control patients, recruited in four centres in the USA (Table S1). Study participants were described fully.[17] Data from Illumina Human370 chip experiments were provided having already undergone quality control steps as described. [17] We re-validated and expanded quality control measures in the present analysis: We excluded 7 participants with apparently misattributed sex. There were no samples with an inbreeding coefficient >0.08 or <-0.08 (suggesting contamination or a sample problem), no pairs with relatedness greater than 12.5%. We conducted principal components analysis using Eigensoft v3.0 with reference to the 11 populations in the HapMap phase 3 data

release,[25] and excluded 40 individuals with outlying ancestry on the basis of visualisation of plots of the first three axes. We chose not to adjust analyses for residual population stratification or ancestry on the basis of a genomic inflation factor of 1 in the primary analysis (individuals with neovascular AMD compared to those with drusen only; a Q-Q plot is shown in Figure S2). The genomic inflation factor for all AMD phenotypes combined (neovascular AMD, geographic atrophy and drusen) compared to healthy controls was 1.04. We retained 2,131 cases and 1,131 controls in our final study dataset (1,313 males; 1,949 females). This dataset included cases with three phenotypes (neovascular AMD, geographic atrophy and drusen) and as healthy controls. For the genome-wide association study of neovascular AMD compared to drusen, we compared 867 participants with neovascular AMD to 519 participants with drusen with >97% genotyping success. We applied filters of 97% for SNP genotyping (excluding 3941 SNPs) and a minimum minor allele frequency of 5% (excluding 17560 polymorphic and 25450 monomorphic SNPs) to this subset of participants, resulting in a study of 323,867 SNPs and a genotyping rate of 99.8%. We analysed data using additive model univariate binary logistic regression in PLINK v1.07.[26, 27] Statistical significance was accepted at two-sided $P<0.05$ after Bonferroni correction for multiple testing in the genome-wide association study. We did not apply correction to the candidate polymorphisms because their selection was based on prior hypotheses due to their known association with the AMD phenotype. We phased haplotypes of *CFH* in PLINK using four SNPs that tag the haplotypes described: rs10801555, as a proxy for rs1061170 (r^2=0.96), tagging haplotype A; rs11582939, as a proxy for rs1065489 (r^2=1), tagging haplotype B; rs800292 tagging haplotype C; rs6677604 tagging haplotype D. We included haplotypes with frequency >5% in analyses. We used SNP Annotation and Proxy Search (Broad Institute)[28] to identify proxy SNPs.

References for Supplementary Methods

1. Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. J Mol Biol 1990;215:403-10.

2. Benson DA, Cavanaugh M, Clark K, et al. GenBank. Nucleic Acids Research 2013;41:D36-42.

3. Anonymous. TruSeq® DNA Sample Preparation Guide (Rev. C). San Diego, California, USA: Illumina; 2012.

4. Anonymous. NimbleDesign .

5. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics 2009;25:1754-60.

6. Li H, Handsaker B, Wysoker A, et al. The sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25:2078-9.

7. McKenna A, Hanna M, Banks E, et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20:1297-303.

8. Robinson JT, Thorvaldsdottir H, Winckler W, et al. Integrative genomics viewer. Nat Biotechnol 2011;29:24-6.

9. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. Nature methods 2010;7:248-9.

10. Li M, Jia C, Kazmierkiewicz KL, et al. Comprehensive analysis of gene expression in human retina and supporting tissues. Hum Mol Genet 2014;23:4001-14.

11. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 2010;26:873-81.

12. DeLuca DS, Levin JZ, Sivachenko A, et al. RNA-SeQC: RNA-seq metrics for quality control and process optimization. Bioinformatics 2012;28:1530-2.

13. Broad Institute. Picard Tools. http://broadinstitute.github.io/picard/.

14. Schroth G. RNA-seq of human individual tissues and mixture of 16 tissues (illumina body map).accession number E-MTAB-513: Array express database; 2011 .

15. Anonymous. E-MTAB-2836 - RNA-seq of coding RNA from tissue samples of 122 human individuals representing 32 different tissues .

16. Dobin A, Davis CA, Schlesinger F, et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics 2013;29:15-21.

17. Chen W, Stambolian D, Edwards AO, 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.

18. Gold B, Merriam JE, Zernant J, 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.

19. Klein RJ, Zeiss C, Chew EY, et al. Complement factor H polymorphism in age-related macular degeneration. Science 2005;308:385-9.

20. Haines JL, Hauser MA, Schmidt S, et al. Complement factor H variant increases the risk of age-related macular degeneration. Science 2005;308:419-21.

21. Edwards AO, Ritter R,3rd, Abel KJ, et al. Complement factor H polymorphism and age-related macular degeneration. Science 2005;308:421-4.

22. Yates JR, Sepp T, Matharu BK, et al. Complement C3 variant and the risk of age-related macular degeneration. N Engl J Med 2007;357:553-61.

23. Yang Z, Camp NJ, Sun H, et al. A variant of the HTRA1 gene increases susceptibility to age-related macular degeneration. Science 2006;314:992-3.

24. Dewan A, Liu M, Hartman S, et al. HTRA1 promoter polymorphism in wet age-related macular degeneration. Science 2006;314:989-92.

25. Price AL, Patterson NJ, Plenge RM, et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006;38:904-9.

26. Purcell S, Neale B, Todd-Brown K, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559-75.

27. Purcell S. PLINK. ; 2009. Available at http://pngu.mgh.harvard.edu/purcell/plink/

28. Johnson AD, Handsaker RE, Pulit SL, et al. SNAP: A web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics 2008;24:2938-9.