Supporting Information Appendix: Recurrent structural variation, clustered sites of selection and disease risk for the complement factor H (*CFH*) gene family

Stuart Cantsilieris[1], Bradley J. Nelson[1], John Huddleston[1,2], Carl Baker[1], Lana Harshman[1], Kelsi Penewit[1], Katherine M. Munson[1], Melanie Sorensen[1], AnneMarie E. Welch[1], Vy Dang[1], Felix Grassmann[6], Andrea J. Richardson[7], Robyn H. Guymer[7], Tina A. Graves-Lindsay[3], Richard K. Wilson[8,9], Bernhard H.F. Weber[6], Paul N. Baird[7], Rando Allikmets[4,5], and Evan E. Eichler[1,2,*]

[1]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA 98195, USA
[2]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA
[3]McDonnell Genome Institute at Washington University, St. Louis, MO 63018, USA
[4]Department of Ophthalmology, Columbia University, New York, NY 10027, USA
[5]Department of Pathology and Cell Biology, Columbia University, New York, NY 10027, USA
[6]Institute of Human Genetics, University of Regensburg, Regensburg 93053, Germany
[7]Centre for Eye Research Australia, Department of Surgery (Ophthalmology) University of Melbourne, Royal Victorian Eye and Ear Hospital, East Melbourne, Victoria 3002, Australia
[8]Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, OH 43205, USA
[9]Department of Pediatrics, The Ohio State University College of Medicine, Columbus, OH 93053, USA

Corresponding author[*]: Evan E. Eichler, Ph.D.
University of Washington School of Medicine
Howard Hughes Medical Institute
3720 15th Ave NE, S413C
Box 355065
Seattle, WA 98195-5065
Phone: (206) 543-9526
E-mail: eee@gs.washington.edu

www.pnas.org/cgi/doi/10.1073/pnas.1717600115

SI MATERIALS AND METHODS

Copy number genotyping using Illumina sequence data. Whole-genome Illumina sequencing data from 224 diverse human genomes (HGDP (1)), 2,143 human individuals through Phase 3 of the 1KG (2), and 86 NHP individuals from the Great Ape Genome Project (including bonobo (N = 14), chimpanzee (N = 23), gorilla (N = 32), and orangutan (N = 17)) (3) were mapped to the human reference genome (GRCh37) using mrsFAST (4). Overall read-depth (whole-genome shotgun sequence detection or WSSD) and paralog-specific read-depth (SUNK) approaches were performed genome-wide across 500 bp sliding windows in 100 bp increments using previously described methods and visualized as heatmaps using bigBed tracks within the UCSC Genome Browser (5, 6). We used the Vst statistic (7) (calculated using a custom python script) to measure copy number stratification between populations.

Sequence and assembly of BAC clones. High-quality sequence and assembly of large-insert clones was performed as previously described (8). In brief, DNA from human (CH17, ABC7, ABC9, ABC13, VMRC65) and NHP (CH251, CH276, CH277, CH250, CH259) BAC and fosmid clone libraries were isolated, prepped into barcoded genomic libraries, and sequenced (PE101) on a MiSeq using a Nextera protocol as previously described (9). Sequence data were mapped with mrsFAST (4) to the GRCh37 reference genome and SUN (6) identifiers were used to discriminate between highly identical SDs. PacBio (Pacific Biosciences, Inc., Menlo Park, CA) SMRTbell™ libraries were prepared and sequenced using RS II P6-C4 chemistry. Inserts were assembled using Quiver and HGAP (10). Contig assembly was performed using Sequencher (Gene Codes Corporation, Ann Arbor, MI) and compared to the human reference genome (GRCh37) using Miropeats (11) and BLAST (12). SDs were annotated within individual contigs using a modified version of whole-genome assembly comparison (WGAC) (13), WSSD (6) and DupMasker (14). Gene annotation was performed using full-length transcripts obtained from RefSeq based on the GRCh37 reference assembly and mapped to individual contigs using GMAP (15) and BLAT (16). Comparative sequence analysis between NHP reference assemblies and large-insert clone-based assemblies were performed using BLASR (17), with parameters fine-tuned for contig alignments (-bestn 1 -minAlignLength 1000 -m 1 -alignContigs –piecewise).

Phylogenetic analyses. We estimated the evolutionary timing of SD events by generating MSAs representative of human, chimpanzee, gorilla, orangutan, macaque and marmoset orthologous and paralogous sequences using MAFFT (18). We constructed an unrooted phylogenetic tree using the neighbor-joining method (MEGA5) (19). Genetic distances were computed using the Kimura two-parameter method with standard error estimates and interior branch test of phylogeny (n = 500 bootstrap replicates). Tajima's relative rate test (MEGA5) was used to assess branch length neutrality (20). We estimated the coalescence of time using the equation R = K/2T assuming a chimpanzee–human divergence time (T) of 6-7 mya for chimpanzee, 15 mya for the orangutan, and 25 mya for the macaque as previously described (21, 22).

Breakpoint refinement using HMMSeg. We mapped NAHR-mediated breakpoints by generating MSAs of sequences corresponding to deletion/duplication-mediating SDs. PSV positions were compared and variant positions showing signatures of transition between corresponding SDs were used to narrow breakpoint regions. To objectively identify these breakpoint transition regions, we performed a three-state Viterbi segmentation using HMMSeg (23) on PSVs for each base of the alignment with an additional state representing uninformative bases.

IGC detection. We implemented two approaches to detect signatures consistent with recent IGC among *CFH* paralogs. First, we created a series of pairwise alignments between all WGAC (13) annotated duplications. We next calculated the identity of two aligned sequences over 2 kbp sliding windows across the alignment with a stepwise increment of 100 bp. Pairwise sequence identity was plotted against the length of the alignment and potential IGC events were identified by the presence of sharp sequence identity transitions from low <99% to high >99%. We next used the program GENECONV (24), which identifies pairs of sequences with longer than expected tracks of 100% sequence identity, conditioning on the overall pattern of variable sites in the alignment (25). The program was run using default parameters, and tracks with a global P values <0.05 were considered significant for follow-up analysis.

Tests for selection. We obtained full-length transcript sequences from the Ensembl genome browser (26) (release 86) for all *CFHR* paralogs, including the ancestral *CFH* gene (Dataset S13). Coding exons corresponding to the largest conical transcript were used as BLAST (12) queries to obtain full-length transcript sequences from NHP and human PacBio-assembled contigs. We generated MSAs using MAFFT (18) and from these MSAs constructed a series of unrooted phylogenetic trees (MEGA5) using the neighbor-joining method with complete-deletion option. The MSAs and phylogenetic trees were used as inputs to test for signals of positive selection using the CODEML package of PAML software v4.7 (27). The substitution rate ratio of nonsynonymous/synonymous (dN/dS) variation (also referred to as omega) was used as a measure of selective pressure. CODEML site model tests were performed by allowing omega to vary among sites and performing a likelihood ratio tests for positive selection. P values were calculated by performing a Chi-square test (df = 2) on twice the difference between the log-likelihood values for different models considered. The Bayes empirical Bayes (BEB) statistic was used to calculate the posterior probabilities for site classes and identify sites under positive selection if the likelihood ratio was significant (28). To detect signatures of a recent selective sweep, we performed an eHH analysis to characterize long-range linkage disequilibrium patterns among four super populations (African [AFR], Americas [AMR], Europeans [EUR], and East Asians [EAS]) generated as part of the 1KG. Using phased single-nucleotide variant and indel calls from >2,000 individuals, we computed the eHH statistic in 50-100 single-nucleotide polymorphism windows (minor allele >5%) restricting our analysis to unique target regions.

PacBio isoform sequencing (Iso-Seq) and GTEx expression quantification. We used both full-length and non-full-length cDNA sequence data, generated from SMRT sequencing of PolyA+ RNA obtained from human liver source material (http://datasets.pacb.com.s3.amazonaws.com/2014/Iso-seq_Human_Tissues/list.html). SMRT cDNA sequence data was mapped to our CHM1 clone-based sequence assembly using GMAP (15). Kallisto (v. 0.42.4) (29) was used to estimate the expression levels of the 24 transcripts detected from *CFH* and *CFHR* paralogs (Dataset S12). We added these transcript sequences to the GENCODE reference transcriptome (release 25) and generated a new index using Kallisto. Transcripts per million values were then calculated using Kallisto with default parameters for all of the GTEx RNA-seq samples (dbGaP version phs000424.v3.p1) from liver source tissue.

Exon sequencing using MIPs. Single-molecule MIPs (smMIPs) were designed to CDS exons (±20 bp) annotated in the GRCh37 human reference assembly using MIPgen (30). Each MIP 70-

mer is designed to capture 112 bp of genomic sequence—this included 40 bp unique to the target of a region (split between a ligation and an extension arm of the MIP), a universal 30 bp backbone, and a degenerate 5 bp single-molecule unique tag (31) included on the extension arm. MIP libraries were prepared as previously described (32, 33). In brief, MIP oligonucleotides (Dataset S14) were pooled together at equal molar concentrations (100 µM) to generate a MIP megapool. MIP megapools were phosphorylated using T4 polynucleotide kinase (1U) at 37ºC for 45 minutes, with a final denaturation of 65ºC for 20 minutes. Capture reactions were performed using 150 nanograms total DNA and a ratio of 800 MIP copies to 1 DNA copy (800:1). Capture reactions (10X Ampligase reaction buffer, 0.006 mM dNTPs, Klentaq (0.32U) Ampligase (1U) MIP megapool and DH2O) were performed using an initial denaturation time of 95ºC for 10 minutes, followed by 60ºC for 20 hours. Exonuclease treatment (ExoI and ExoIII) was performed at 37ºC for 45 minutes with a final denaturation of 95ºC for 2 minutes. PCR (2X iProof high fidelity master mix, barcode primer (10 µM), DH2O) was performed using an initial denaturation time of 98ºC for 30 seconds, 22 cycles at 98ºC for 10 seconds, 60ºC for 30 seconds and 72ºC for 30 seconds and a final extension time of 72ºC for 2 minutes. Finished libraries were pooled together and sequenced using either MiSeq (2 x 150 bp) or HiSeq2000 (2 x 101 bp).

Variant calling and validation. We used the MIPgen (30) data analysis pipeline to filter and map reads in fastq format to a hard-masked GRCh37 human reference assembly. Discovery variant calling was performed per pooled set of 384 samples using FreeBayes/v1.0.2(34) (https://github.com/ekg/freebayes). All samples that met a cutoff of 20X sequence coverage for >80% of the MIPs targeting unique space were used in the final analysis (*SI Appendix*, Fig. S16). Variant calls were filtered for trinucleotide or homopolymer repeat sequences, read-depth ≤10, quality score ≤20, or no alleles as previously described (35). We annotated variant calls using the Ensembl Variant Effect Predictor (Assembly: GRCh37.p13) (36) against the canonical transcript for each gene. We scored single-nucleotide variants for deleteriousness using CADD (Combined Annotation Dependent Depletion) v1.3 (http://cadd.gs.washington.edu/score) and compared frequencies of variants using ExAC (http://exac.broadinstitute.org/) (37). Missense and loss-of-function variants were validated by Sanger sequencing.

Genotyping of *CFHR* paralogs using MIPs. We designed smMIPs (Dataset S14) to variants that distinguish individual *CFHR* paralogs (SUNs). Briefly, we aligned *CFHR* paralogous sequences using MAFFT and used a series of custom scripts to select and design MIPs to SUN identifiers as described previously (38, 39). We performed MIP capture, library preparation, and massively parallel sequencing as described above. This allowed quantification of reads derived from each individual paralog across each individual patient. Paralog-specific copy number calls were visualized by plotting paralog-specific read-depth across the length of shared sequence between paralogs. These estimates were calculated at each MIP target by multiplying paralog-specific *CFHR* read count relative frequencies by corresponding aggregate *CFHR* copy number estimates called by the genotyping program. The algorithmic details of this program have been previously described (38).

AMD patient cohorts. We sequenced 2,546 advanced AMD cases and controls from three separate cohorts (Columbia, Melbourne and Regensburg) of European ancestry. Details on ophthalmological grading and inclusion/exclusion criteria have been published previously (40). In brief, we sequenced 535 advanced AMD cases and 534 controls from Columbia University, 688 advanced AMD cases and 163 controls from the Centre for Eye Research Australia (Melbourne), and 450 advanced AMD cases and 176 controls from the University of Regensburg (41). All controls were age and ethnicity matched. All groups collected data according to Declaration of Helsinki principles. At the Columbia center, the study was reviewed and approved by Columbia University Human Research protection Office Institutional Review Board. In Melbourne, the study was approved by the Human Research and Ethics Committee of the Royal Victorian Eye and Ear Hospital (RVEEH). In Regensburg, the study was approved by the Ethics Committees at the University Eye Clinics of Würzburg (Study No. 78/01) and München (Study No. 226/02). Written informed consent was obtained for all study participants before participation.

Statistical analysis. We applied standard tests for association using count data (one- and two-sided Fisher's exact tests). For common variants, we applied Chi Square tests for association and Bonferroni correction based on the number of single-nucleotide variants analyzed per gene, implemented in the PLINK software package (version 1.07) (42). Logistic regression analysis adjusting for age, gender and the effect of the Y402H polymorphism was also performed. Genotyping completeness rates were >0.99 for 16/18 nonsynonymous events in *CFH* and *CFHR* paralogs (Datasets S17 and S19). Departure from Hardy-Weinberg Equilibrium was considered significant for P values <0.001 as previously described (43).

SI FIGURES

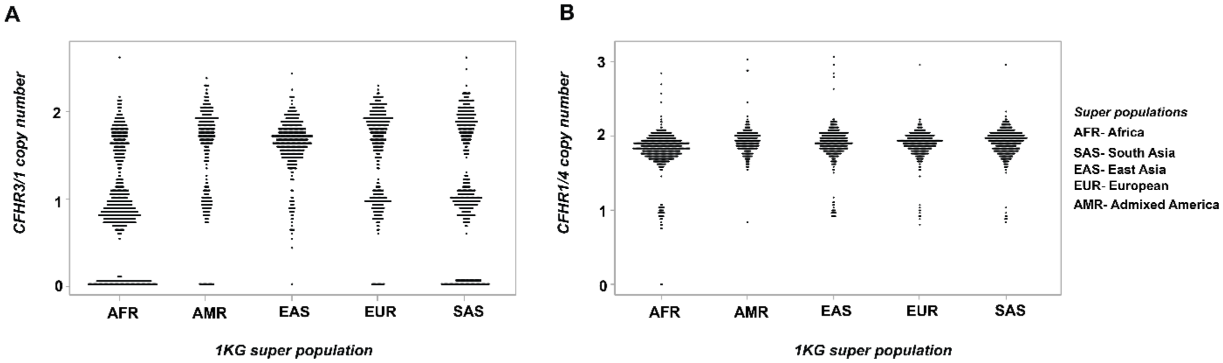


Fig. S1. *CFHR* copy number diversity in 2,143 humans from 1KG. Scatter plots depicting copy number estimates obtained using whole-genome sequencing read-depth approach among five super populations representing 2,143 individuals from 1KG. **A)** *CFHR3-1* copy number estimates range from 0-2. Africans and South Asians are enriched for the deletion allele. **B)** *CFHR1-4* copy number estimates range from 0-3 with Africans demonstrating an increased frequency of the deletion allele relative to other populations.

	5093	5124	5135	5139	5145	5214	5245	5273	5288	5298	5331	5345	5368	5369	5370	5371	5374	5387	5430	5628	5639	5641	5730	6047	6048	6186
GRCh37-1	c	c	a	t	g	c	t	g	a	a	t	g	a	g	a	g	g	a	c	a	c	a	c	t	g	t
NA18517	c	c	a	t	g	c	t	g	a	a	t	g	a	g	a	g	g	a	c	a	c	a	c	t	g	t
GRCh37-2	t	t	t	c	a	t	a	a	t	g	a	a	-	-	-	-	a	t	t	g	g	c	t	c	a	a

	6197	6200	6209	6240	6523	6530	6551	6558	6562	6644	6685	6717	6726	6727	6728	6729	6730	6731	6774	6841	6844	6845	6846	6935	6942	7052
GRCh37-1	t	g	c	a	c	c	c	c	t	a	t	t	-	-	-	-	-	-	g	g	a	a	a	a	a	g
NA18517	t	g	c	a	c	c	c	c	t	a	t	t	-	-	-	-	-	-	g	g	a	a	a	a	a	g
GRCh37-2	a	a	t	g	t	t	t	t	a	c	c	c	a	g	t	a	a	t	a	a	-	-	-	g	g	t

	7128	7195	7276	7321	7452	7625	7642	7683	7690	7694	7720	7811	7813	7816	7818	7821	7823	7835	7846	7847	7881	7882	7893	7944	7953	7979
GRCh37-1	c	g	t	t	a	c	g	g	a	t	c	a	c	a	c	a	c	a	t	g	c	a	t	t	a	c
NA18517	c	g	t	t	a	c	g	g	a	t	c	a	c	a	c	a	c	a	t	g	c	a	t	t	a	c
GRCh37-2	t	t	a	c	g	a	a	a	g	c	t	c	a	c	a	c	a	g	c	a	t	c	c	g	g	t

	7984	7997	8026	8047	8059	8073	8080	8095	8097	8098	8128	8308	8368	8395	8408	8444	8469	8493	8731	8743	8759	8894	9070	9071	9072	9073
GRCh37-1	c	t	c	t	g	g	t	a	t	a	t	g	-	g	a	g	a	c	t	g	a	g	a	g	a	c
NA18517	c	t	c	t	g	g	t	a	t	a	t	g	-	g	a	g	a	c	t	g	a	g	-	-	-	-
GRCh37-2	a	g	t	c	c	a	c	g	g	g	c	a	a	a	c	a	g	a	a	a	-	a	-	-	-	-

	9266	9352	9610	9664	9954	10274	10312	10586	11960	12204	12216	12309	12318	12338	12531	12551	12628	12766	12893	12930	13315	13316	13365	13366	13450	13604
GRCh37-1	a	g	t	t	-	a	a	c	c	t	t	g	g	t	g	a	g	c	a	a	a	a	a	c	a	a
NA18517	a	g	t	t	-	a	t	c	t	c	c	t	t	t	a	g	a	a	c	g	-	-	a	-	g	a
GRCh37-2	g	t	c	c	t	g	a	t	c	c	c	t	t	a	g	g	g	a	a	g	a	a	t	c	g	g

	13668	13787	13906	13928	13938	14200	14320	14478	14519	14764	14994	15068	15078	15079	15080	15100	15564	15590	15672	15837	16013	16211	16233	16244	16643	16916
GRCh37-1	t	a	g	g	c	g	a	t	c	g	c	g	a	t	t	a	t	c	g	-	t	c	g	t	t	g
NA18517	t	a	t	g	t	a	g	g	c	a	t	g	a	t	t	a	c	a	a	-	t	t	g	g	c	a
GRCh37-2	c	g	t	a	t	a	a	g	t	g	c	t	t	c	c	t	t	a	a	t	c	t	a	a	t	g

	16974	16983	16990	16991	17003	17007	17058	17107	17205	17381	17399	17423	17547	17586	17595	17596	17597	17598	17599	17690	17696	17838	17842	18033	18115	18310
GRCh37-1	c	a	t	g	a	c	c	g	g	a	a	g	a	t	t	t	t	g	t	t	g	t	g	a	t	t
NA18517	t	c	-	-	g	t	t	g	g	t	a	c	g	c	-	-	-	-	c	g	a	a	t	g	-	t
GRCh37-2	t	c	-	-	g	t	t	c	t	-	t	c	g	c	-	-	-	-	c	g	a	a	t	g	-	c

	18399	18491	18528	18562	18651	18665	18788	18956	19014	19020	19041	19054	19055	19056	19089	19105	19165	19328	19329	19330	19331	19349	19350	19423	19426	19501
GRCh37-1	t	t	c	t	a	c	t	t	g	a	t	a	c	t	a	c	g	-	-	-	-	a	a	g	t	t
NA18517	t	c	t	a	g	c	a	t	a	t	a	-	-	-	a	t	g	t	t	a	t	c	g	g	c	t
GRCh37-2	a	t	t	t	g	g	a	a	a	t	a	-	-	-	t	t	t	t	t	a	t	c	g	c	t	c

	19588	19580	19703	19716	19721	19732	19830	19995	20436	20463	20472	20509	20650	20749	20997	21070	21073	21465	21472	21482	21556	21567	21610	21678	21691	21713
GRCh37-1	c	a	c	a	c	a	t	g	c	g	c	a	g	t	t	t	c	t	g	g	a	t	a	g	g	a
NA18517	t	a	t	t	t	c	t	g	g	a	g	g	g	g	t	t	c	c	t	c	g	c	-	a	g	a
GRCh37-2	c	g	t	t	t	c	c	a	g	a	g	g	t	g	c	c	a	c	g	g	g	c	a	g	a	g

	21754	21772	21848	21898	21930	22042	22112	22167	22190	22236	22238	22268	22270	22414	22575	22657	22740	22767	22774	22775	22825	23057	23155	23195	23303	23397
GRCh37-1	c	c	t	c	t	g	g	g	g	g	t	a	a	g	g	g	a	g	c	a	g	t	a	a	c	t
NA18517	t	t	c	t	a	a	a	t	c	a	c	c	a	a	a	a	t	a	t	g	g	t	c	g	t	t
GRCh37-2	t	c	c	t	a	a	a	g	c	a	c	c	g	g	a	a	t	a	t	g	a	c	c	a	t	c

	23604	23796	23854	23855	23856	23857	23938	24003	24010	24011	24026	24052	24070	24218	24230	24231	24278	24289	24373	24404	24424	24471	24473	24524	24541	24581
GRCh37-1	t	a	-	-	-	-	c	g	t	g	a	c	a	a	a	t	t	a	g	c	a	-	g	g	c	g
NA18517	c	g	a	t	a	a	t	g	a	a	g	t	g	g	a	t	c	g	a	t	g	a	a	g	t	a
GRCh37-2	c	a	a	t	a	a	t	a	a	a	g	t	g	a	-	-	c	g	a	t	g	a	a	t	t	a

	24628	24634	24684	24787	24793	24814	24834	24889	25090	25098	25111	25176	25374	25408	25427	25497	25527	25535	25540	25584	25586	25626	25637	25638	25682	25685
GRCh37-1	t	a	t	c	a	g	g	a	t	t	a	c	a	g	t	c	a	g	t	c	t	a	a	c	-	a
NA18517	t	c	c	t	a	a	a	a	c	g	g	t	a	a	c	t	g	a	g	t	c	g	a	c	t	t
GRCh37-2	c	c	c	t	g	a	a	g	c	t	g	t	t	a	c	t	g	g	g	c	t	g	g	t	t	t

	25693	25732	25736	25749	25764	25771	25813	25824	25840	25855	25870	25874	25886	25900	25927	25949	25978	25889	25990	25993	26004	26007	26021	26050	26053	26059
GRCh37-1	a	t	g	c	a	c	c	a	c	g	c	t	c	g	c	c	t	c	a	a	a	a	c	g	a	-
NA18517	c	c	c	a	c	t	a	g	g	c	g	c	a	a	t	t	a	t	g	g	c	c	g	a	g	t
GRCh37-2	c	t	c	a	c	t	a	g	g	c	g	t	c	a	t	t	a	t	g	g	c	c	g	a	g	t

Fig. S2. Detailed alignment of *CFHR3/1* segmental duplications (SDs) to refine deletion breakpoints. Pictured are the consensus positions from the MSA, which include sequences corresponding to GRCh37 chr1:196711705-196740354 and chr1:196796320-196825045 and a breakpoint spanning clone from the NA18517 fosmid library (ABC7). 303 PSVs are annotated as being shared with the centromeric (blue) or telomeric (green) copies of the duplication. The 6,380 bp high-confidence breakpoint transition region is highlighted using a yellow box. The putative breakpoint signature identified by Hughes et al. 2006 is highlighted using a red box. The putative breakpoint signature identified by HMMSeg is highlighted using a purple box.

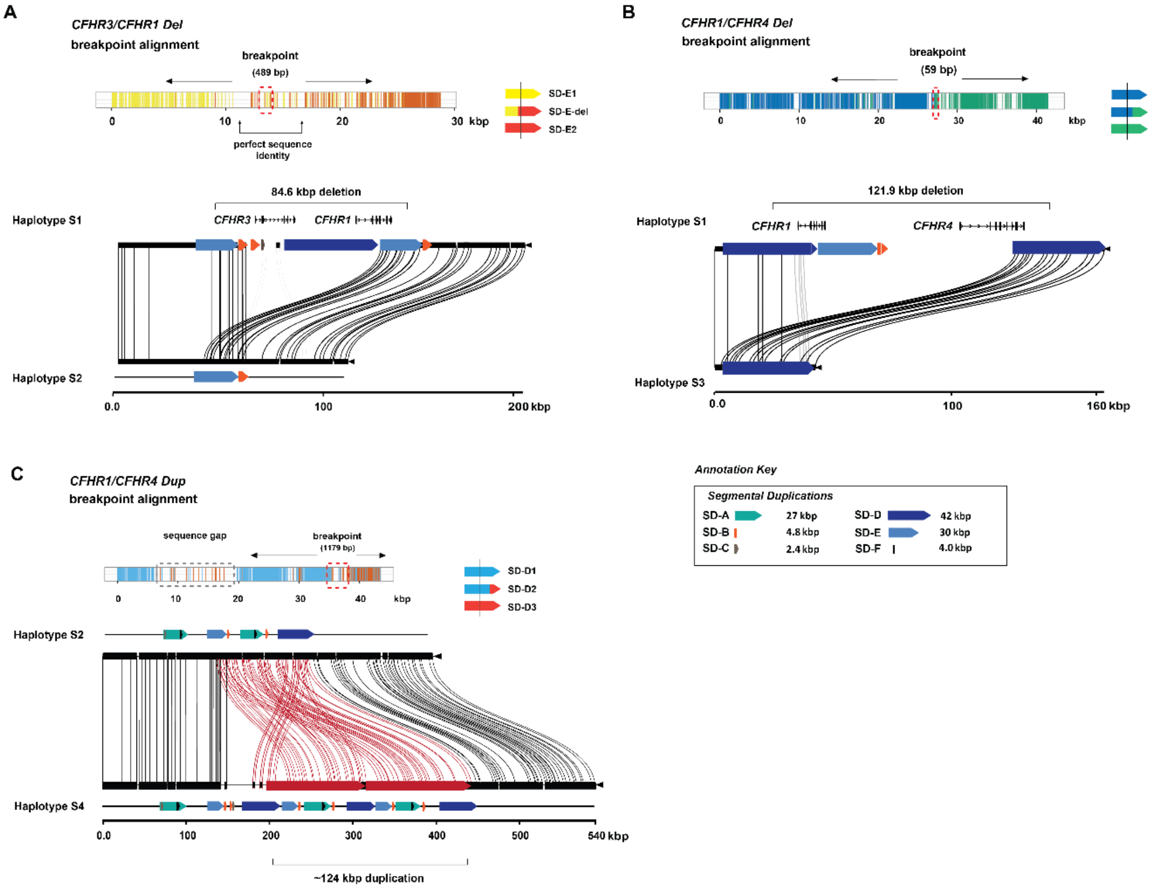


Fig. S3. Structural variation and breakpoint sequence refinement at 1q31.3. A Miropeats comparison between structurally diverse haplotypes is shown. Lines connect stretches of homologous sequence between haplotypes based on a chosen threshold (s), defined as the number of matching bases minus the number of mismatching bases. Sequences from flanking SDs were aligned between haplotypes, and variant sites were compared to map breakpoint positions. A) The breakpoint interval of the 84.6 kbp *CFHR3/CFHR1* deletion (Miropeats comparison between S1 and S2; bottom) is narrowed to a 489 bp region (dashed red box) based on unique variants that distinguish SD-E1 (yellow) from SD-E2 (red) as defined by HMMSeg (top). The breakpoint is embedded in a larger stretch of predicted IGC. B) A Miropeats comparison between two human haplotypes (S1 and S3) depicts a 121.9 kbp deletion that includes *CFHR1* and *CFHR4*. The deletion breakpoint highlighted by the blue to green transition region (dashed red box) is narrowed to a 59 bp sequence interval associated with SD-D paralogs. C) A Miropeats comparison between two human haplotypes (S2 and S4) shows a large tandem

duplication of ~124.9 kbp in the S4 haplotype, including *CFHR1* and *CFHR4* (red lines). The duplication breakpoint interval (dashed red box of blue to red transition) is narrowed to a 1,179 bp sequence interval distinct from the S3 haplotype breakpoint (based on a 42 kbp MSA alignment of SD-D1, SD-D2 and SD-D3).

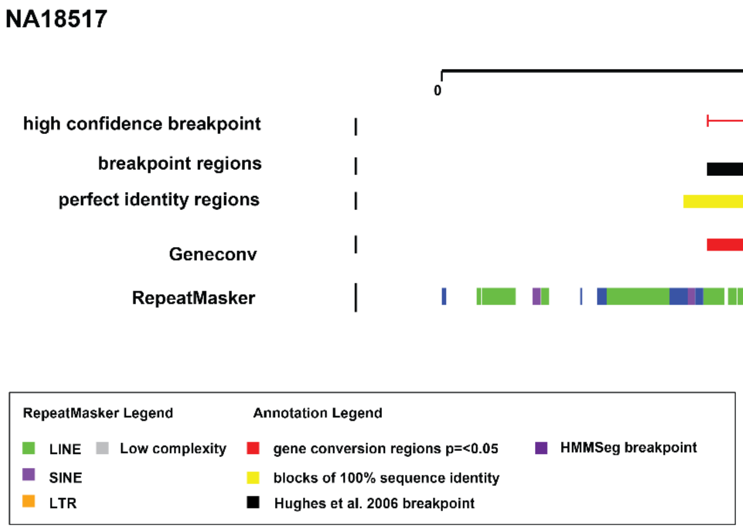


Fig. S4. Sequence analysis of the *CFHR3/1* breakpoint transition region. The schematic shows the high-confidence 6,380 bp breakpoint transition region identified from mapping 303 PSVs across the length of the alignment. The breakpoint region identified by Hughes et al. 2006 is annotated in black and the breakpoint region annotated using HMMSeg is annotated in purple. Regions of perfect sequence identity are highlighted in yellow based on comparison between four sequenced haplotypes (CHM1, NA19129, NA18956 and the GRCH37 reference assembly). Regions annotated as being globally significant for IGC are in red. RepeatMasker annotations show the high density of LINE elements, particularly LINE/L1 across the duplication breakpoint.

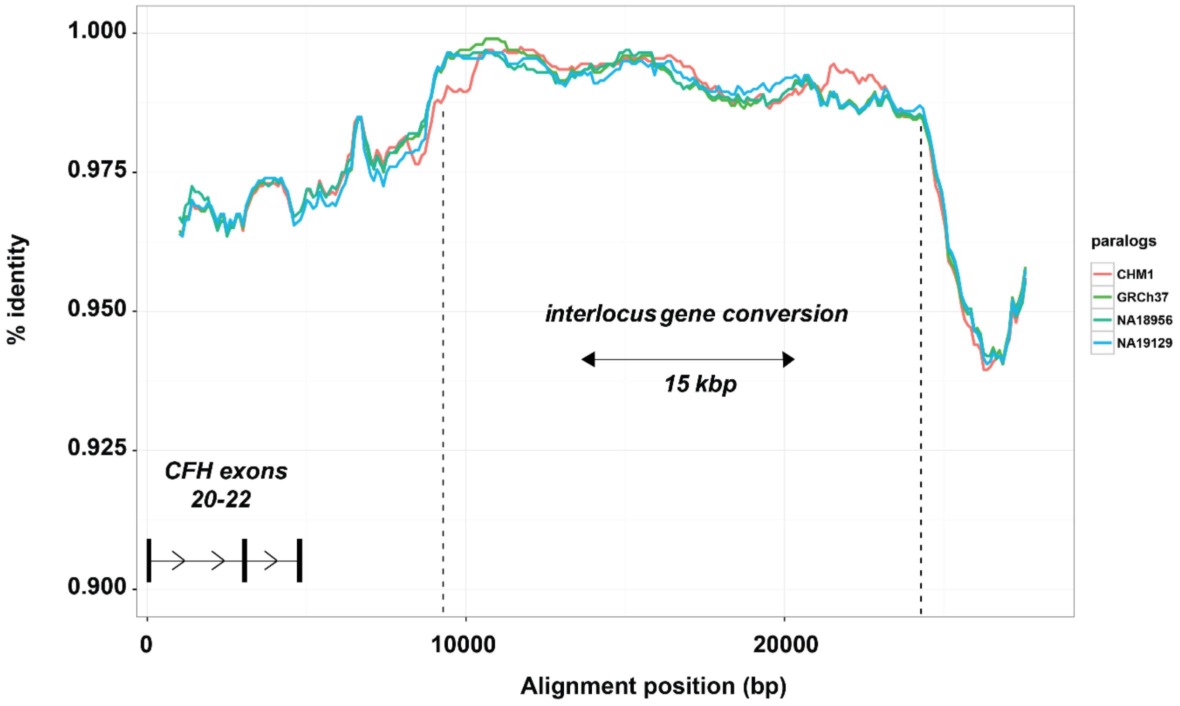
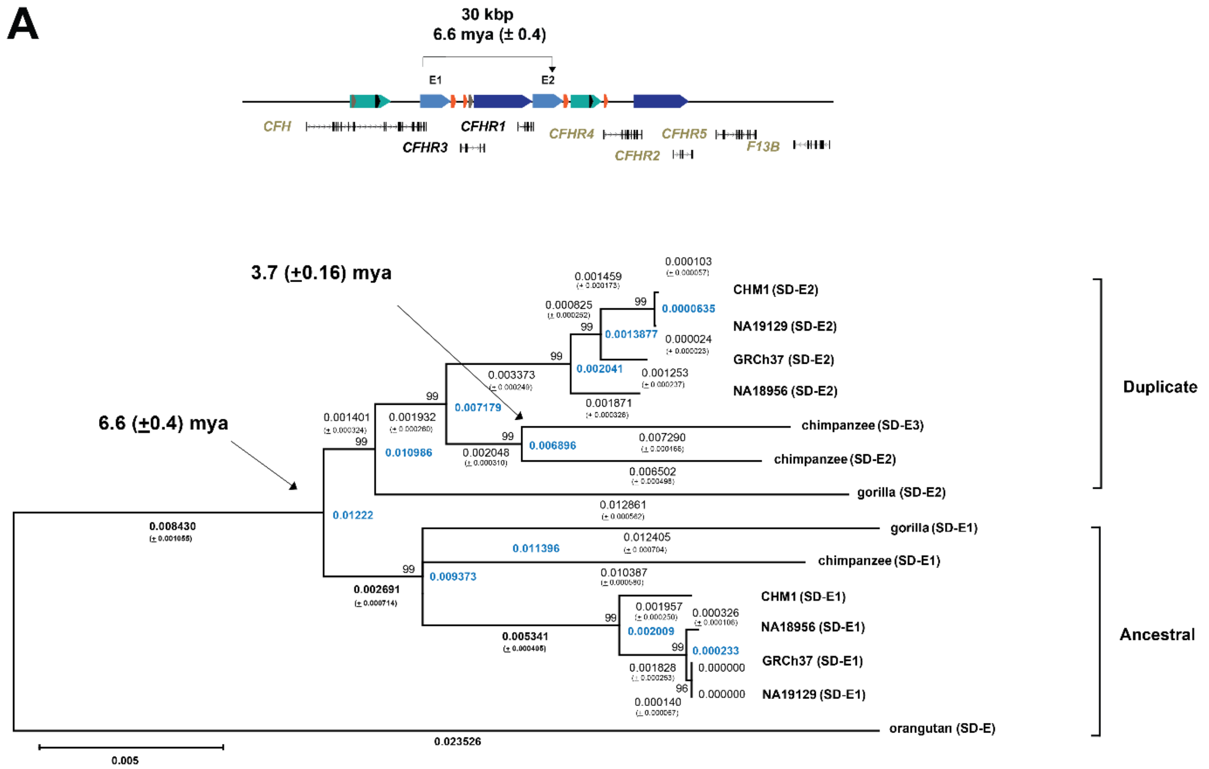


Fig. S5. Percent sequence identity calculated in sliding windows between *CFHR3/1* SDs. Pairwise sequence alignments between CHM1 (red), GRCh37 (green), NA18956 (aqua) and NA19129 (blue) *CFHR3/1* SDs are performed. The sequence identity for each alignment was computed and plotted over 2 kbp windows, sliding by 100 bp. The pattern of sequence identity shows a sharp transition from 0.975 to >0.99 identity between 9,500-10,000 bp (dashed lines) to 25,000 indicative of a 15 kbp region of IGC.

A



GRCh37 coordinates
chr1:196711705-196740354 and chr1:196796320-196825045

branch length

average branch length for all sequences in a clade

Timing Estimate SD-E1 and SD-E2

T=((0.01222 subs/site)/(0.01222 +0.008430+0.023526))*(12 MYA) = 6.6 MYA

T=((0.01222 subs/site)/(0.01222 +0.008430+0.023526))*(14 MYA) = 7.7 MYA

T=((0.01222 subs/site)/(0.01222 +0.008430+0.023526))*(15 MYA) = 8.3 MYA

Timing Estimate SD-E2 and SD-E3

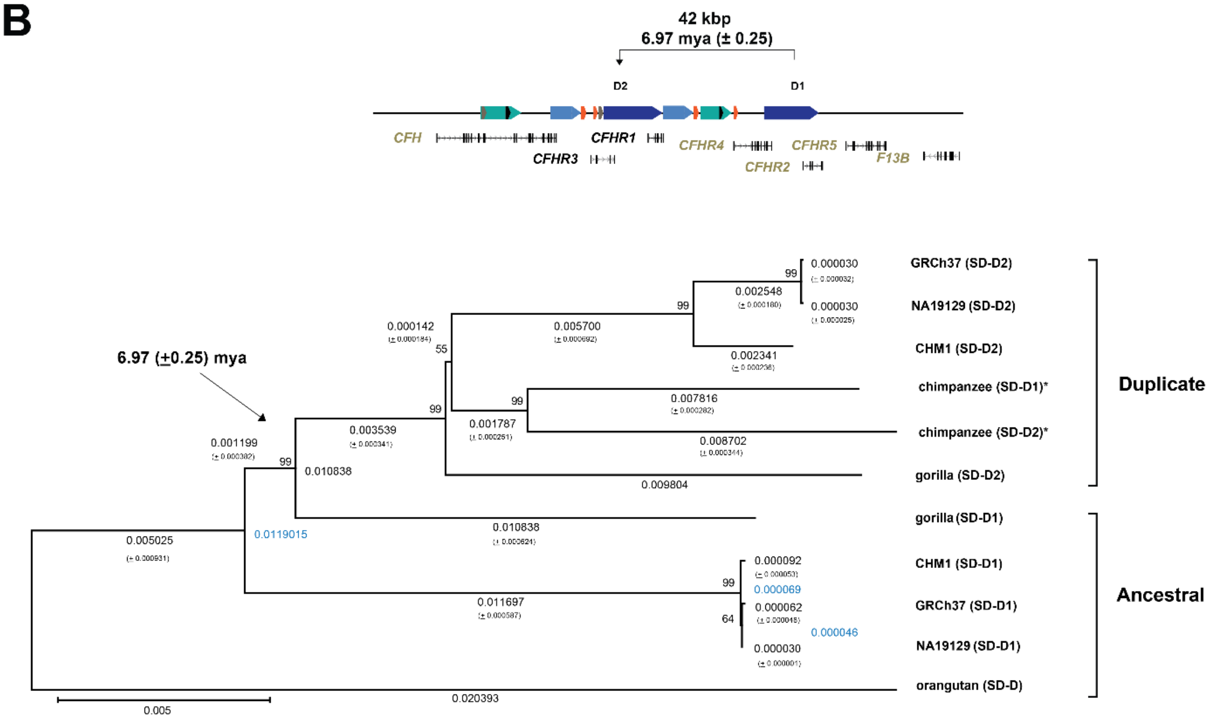
T=((0.006896 subs/site)/(0.012667 +0.008430+0.023526))*(12 MYA) = 3.7 MYA

T=((0.006896 subs/site)/(0.012667 +0.008430+0.023526))*(14 MYA) = 4.3 MYA

T=((0.006896 subs/site)/(0.012667 +0.008430+0.023526))*(15 MYA) = 4.6 MYA

Annotation Key

SD-A	🟩	27 kbp	SD-E	🟦	30 kbp
SD-B	🟥	4.8 kbp	SD-F	▮	4.0 kbp
SD-C	▮	2.4 kbp			
SD-D	🟦	42 kbp			



GRCh37 coordinates
chr1:196756102-196796319 and chr1:196880627-196920352

branch length

average branch length for all sequences in a clade

Timing Estimate

T=((0.010838 subs/site)/(0.0119015 + 0.005025 + 0.020393)) * (12 MYA) = 6.97 MYA

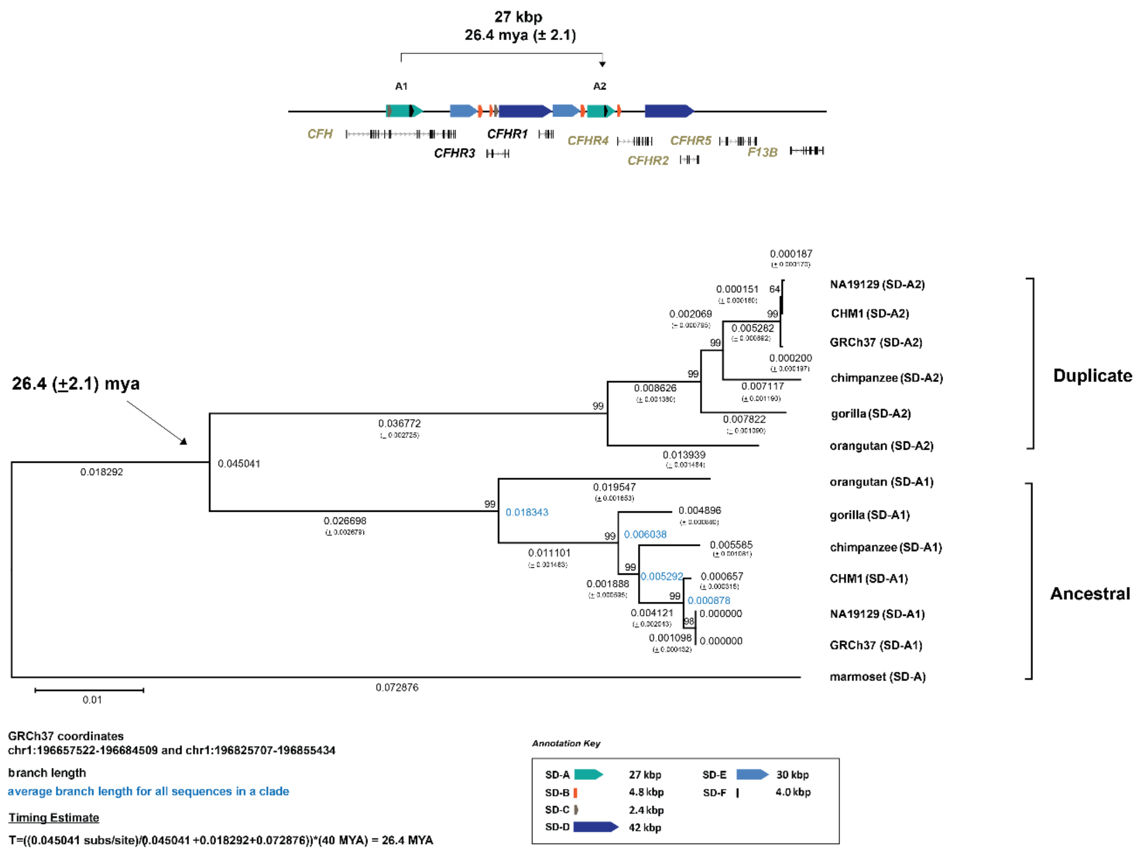
T=((0.010838 subs/site)/(0.0119015 + 0.005025 + 0.020393)) * (14 MYA) = 8.13 MYA

T=((0.010838 subs/site)/(0.0119015 + 0.005025 + 0.020393)) * (15 MYA) = 8.71 MYA

Annotation Key

SD-A	🟩	27 kbp	SD-E	🟦	30 kbp
SD-B	🟥	4.8 kbp	SD-F	▮	4.0 kbp
SD-C	▶	2.4 kbp			
SD-D	🟦	42 kbp			

C



GRCh37 coordinates
chr1:196657522-196684509 and chr1:196825707-196855434

branch length

average branch length for all sequences in a clade

Timing Estimate

T=((0.045041 subs/site)/(0.045041 +0.018292+0.072876))*(40 MYA) = 26.4 MYA

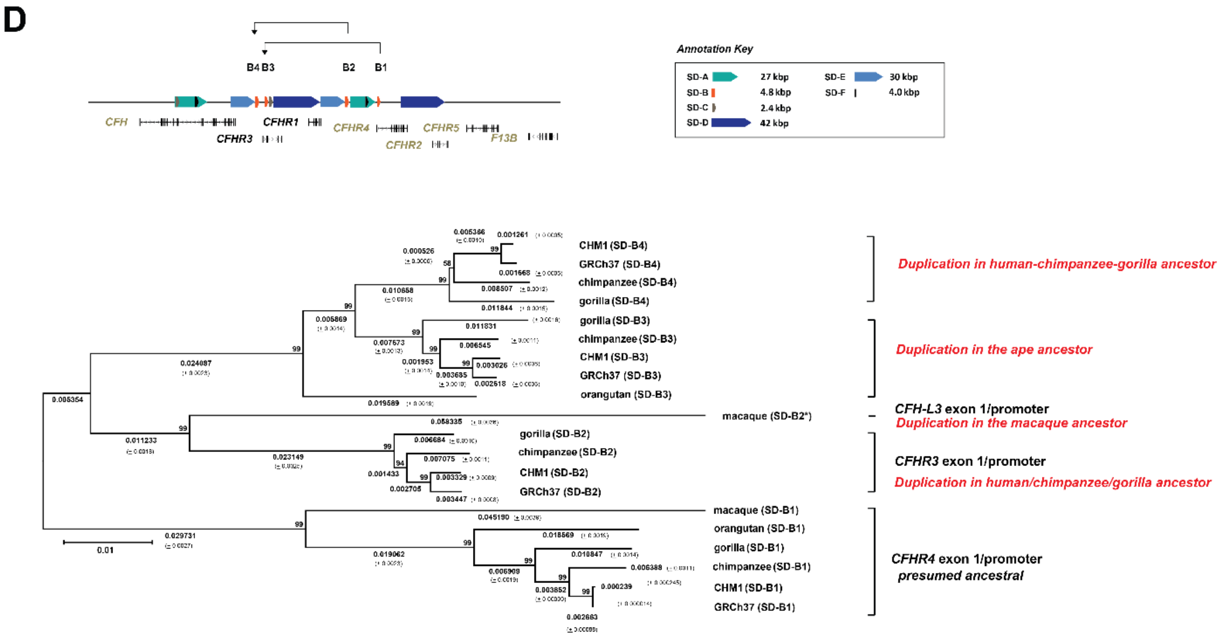

D
B4 B3
B2
B1
CFH
CFHR3
CFHR1
CFHR4
CFHR2
CFHR5
F13B
Annotation Key
SD-A 27 kbp
SD-B 4.8 kbp
SD-C 2.4 kbp
SD-D 42 kbp
SD-E 30 kbp
SD-F 4.0 kbp
CHM1 (SD-B4)
GRCh37 (SD-B4)
chimpanzee (SD-B4)
gorilla (SD-B4)
gorilla (SD-B3)
chimpanzee (SD-B3)
CHM1 (SD-B3)
GRCh37 (SD-B3)
orangutan (SD-B3)
macaque (SD-B2*)
gorilla (SD-B2)
chimpanzee (SD-B2)
CHM1 (SD-B2)
GRCh37 (SD-B2)
macaque (SD-B1)
orangutan (SD-B1)
gorilla (SD-B1)
chimpanzee (SD-B1)
CHM1 (SD-B1)
GRCh37 (SD-B1)
0.005366
0.001261
0.000526
0.001668
0.010658
0.008507
0.011844
0.005869
0.011831
0.007673
0.006545
0.001953
0.003026
0.003685
0.002618
0.024087
0.019589
0.005354
0.058335
0.011233
0.006684
0.023149
0.007075
0.001433
0.003329
0.002705
0.003447
0.045190
0.029731
0.018569
0.019062
0.010847
0.006909
0.006388
0.003852
0.000239
0.002663
99
58
99
99
99
99
99
99
99
99
94
99
99
99
99
99
99
0.01
Duplication in human-chimpanzee-gorilla ancestor
Duplication in the ape ancestor
CFH-L3 exon 1/promoter
Duplication in the macaque ancestor
CFHR3 exon 1/promoter
Duplication in human/chimpanzee/gorilla ancestor
CFHR4 exon 1/promoter
presumed ancestral

Fig. S6. Phylogenetic analysis of *CFHR* duplications. MSAs were generated using MAFFT from human and NHP haplotype sequences. An unrooted neighbor-joining tree was constructed using MEGA5 and all positions containing gaps and missing data were eliminated from the final analysis. Genetic distances were computed using the Kimura two-parameter method with standard error estimates and an interior branch test of phylogeny (n = 500 bootstrap replicates). We determined that these sequences evolved at the same rate as orthologous counterparts in the chimpanzee, gorilla and orangutan using Tajima's relative rate test (MEGA5), and timing estimates were performed taking into account uncertainty in the orangutan–human divergence time. **A)** Timing estimate for the 28.6 kbp SD designated SD-E. The duplication is estimated to have occurred 6.6 (±0.4) mya and includes exons 19-22 from the ancestral *CFH* gene. The duplicate copy SD-E2 forms the C-terminus of the *CFHR1* gene paralog. Similarly, an additional duplication of SD-E specific to the chimpanzee is estimated to have occurred 3.7 (±0.16) mya. **B)** Timing estimate for the 40.2 kbp SD designated SD-D. The duplication is estimated to have occurred 6.9 (±0.25) mya and includes exons 1-3 from *CFHR2* and exons 1-4 from *CFHR4*. The resulting duplication SD-D2 forms the C-terminus of *CFHR3* and the N-terminus of *CFHR1* gene paralogs. **C)** Timing estimate for the 26.9 kbp SD designated SD-A. The duplication is estimated to have occurred 26.4 (±2.1) mya and includes exons 8, 9 and 10 from the ancestral *CFH*. The SD-A2 duplication does not result in the formation of transcripts containing an ORF. **D)** Phylogeny of the 4.8 kbp SD designated SD-B, which includes exon 1 and the promoter of the *CFHR4* gene paralog. This is the presumed ancestral location for the duplication, which has undergone several independent duplication events in primate lineages. SD-B1 has duplicated twice independently during primate evolution. The first was in the rhesus macaque, which forms the N-terminus of the *CFH*-like 3 gene. This event was subsequently deleted after the divergence from Old World monkeys to great apes and duplicated again after the split between the orangutan and the chimpanzee–gorilla ancestor. This secondary event forms the N-terminus of *CFHR3* in the chimpanzee, gorilla and human lineages.

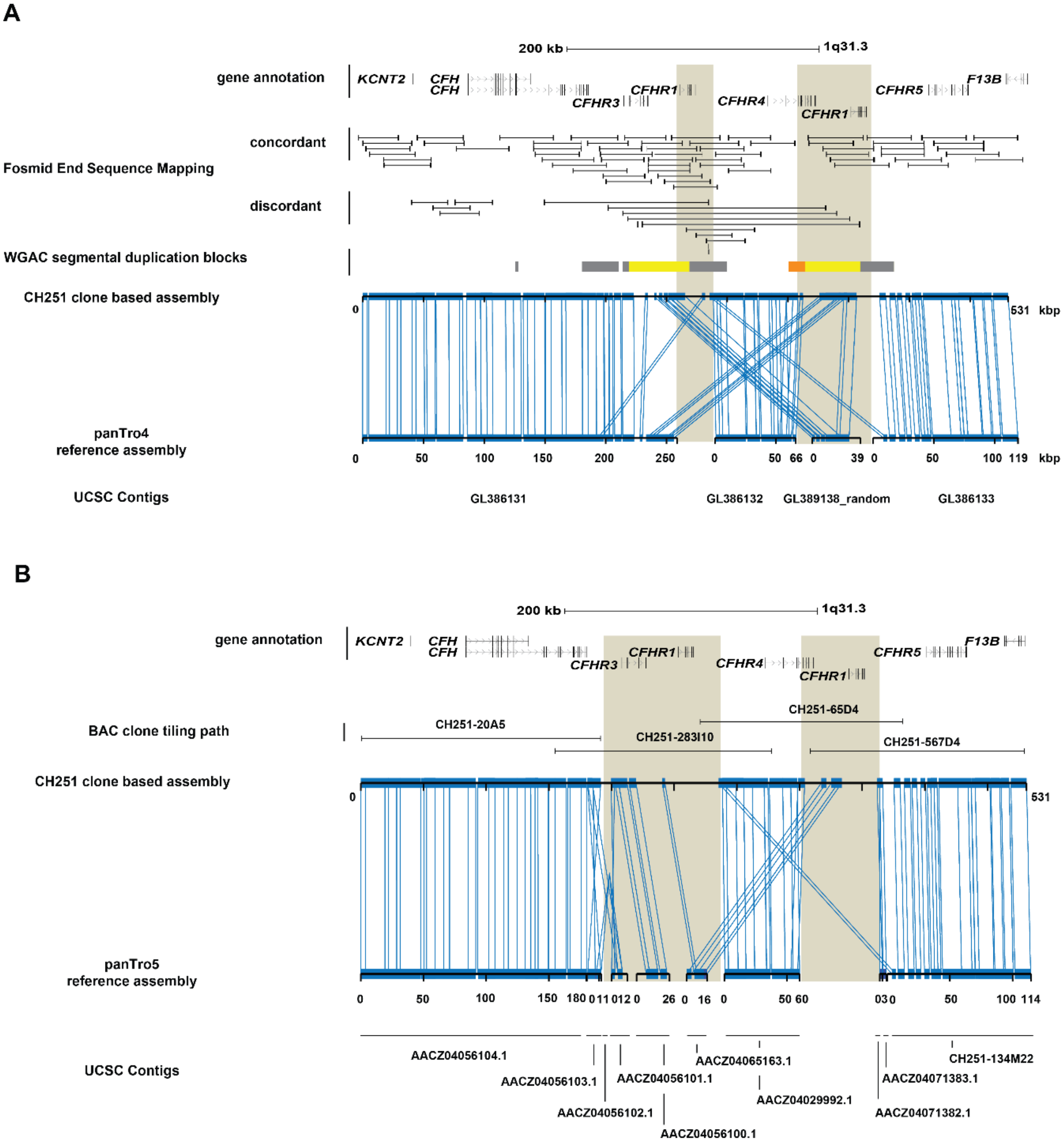

A
200 kb
1q31.3
gene annotation
KCNT2
CFH
CFH
CFHR3
CFHR1
CFHR4
CFHR1
CFHR5
F13B
concordant
Fosmid End Sequence Mapping
discordant
WGAC segmental duplication blocks
CH251 clone based assembly
0
531 kbp
panTro4
reference assembly
0 50 100 150 200 250 0 50 66 0 39 0 50 100 119 kbp
UCSC Contigs
GL386131
GL386132
GL389138_random
GL386133
B
200 kb
1q31.3
gene annotation
KCNT2
CFH
CFH
CFHR3
CFHR1
CFHR4
CFHR1
CFHR5
F13B
BAC clone tiling path
CH251-20A5
CH251-283I10
CH251-65D4
CH251-567D4
CH251 clone based assembly
0
531
panTro5
reference assembly
0 50 100 150 180 0 11 0 12 0 26 0 16 0 50 60 0 3 0 50 100 114
UCSC Contigs
AACZ04056104.1
AACZ04056103.1
AACZ04056102.1
AACZ04056101.1
AACZ04056100.1
AACZ04065163.1
AACZ04029992.1
AACZ04071383.1
AACZ04071382.1
CH251-134M22

Fig. S7. Pairwise sequence comparison between the panTro4/panTro5 chimpanzee reference assemblies and the newly constructed contig using CH251 large-insert BAC clones. A) A Miropeats comparison between panTro4 and the CH251 BAC contig shows pairwise sequence differences between orthologous regions. Lines connect stretches of homologous sequence (threshold s = 2000) between the two assemblies with large breaks in sequence contiguity highlighted by tan-colored blocks. Gene annotations are shown based on mapping human RefSeq annotations from GRCh37 to the new alternate reference assembly. Fosmid end-sequence (FES) mapping using the fosmid library constructed from the same source material (CH1251) shows complete FES concordance with the exception of a 7.5 kbp collapsed duplication corresponding to *CFHR4*. In comparison, panTro4 contains four sequence contigs of which one is a randomly assigned chromosome and there are three sequence gaps corresponding to 91.8 kbp. The missing sequence corresponds to high-identity SDs arranged in tandem across the locus. panTro4 is missing annotations for 3/5 *CFH* gene paralogs. **B)** A Miropeats comparison between panTro5 and the CH251 BAC contig shows pairwise sequence differences between orthologous regions. Lines connect stretches of homologous regions (threshold s = 2000) between the two assemblies with large breaks in sequence contiguity highlighted by tan-colored blocks. panTro5 contains 10 sequence contigs that range from 1.6–183 kbp in size. panTro5 is missing annotations for 3/5 *CFH* gene paralogs and we estimate that 133.3 kbp of sequence is missing from the final assembly.

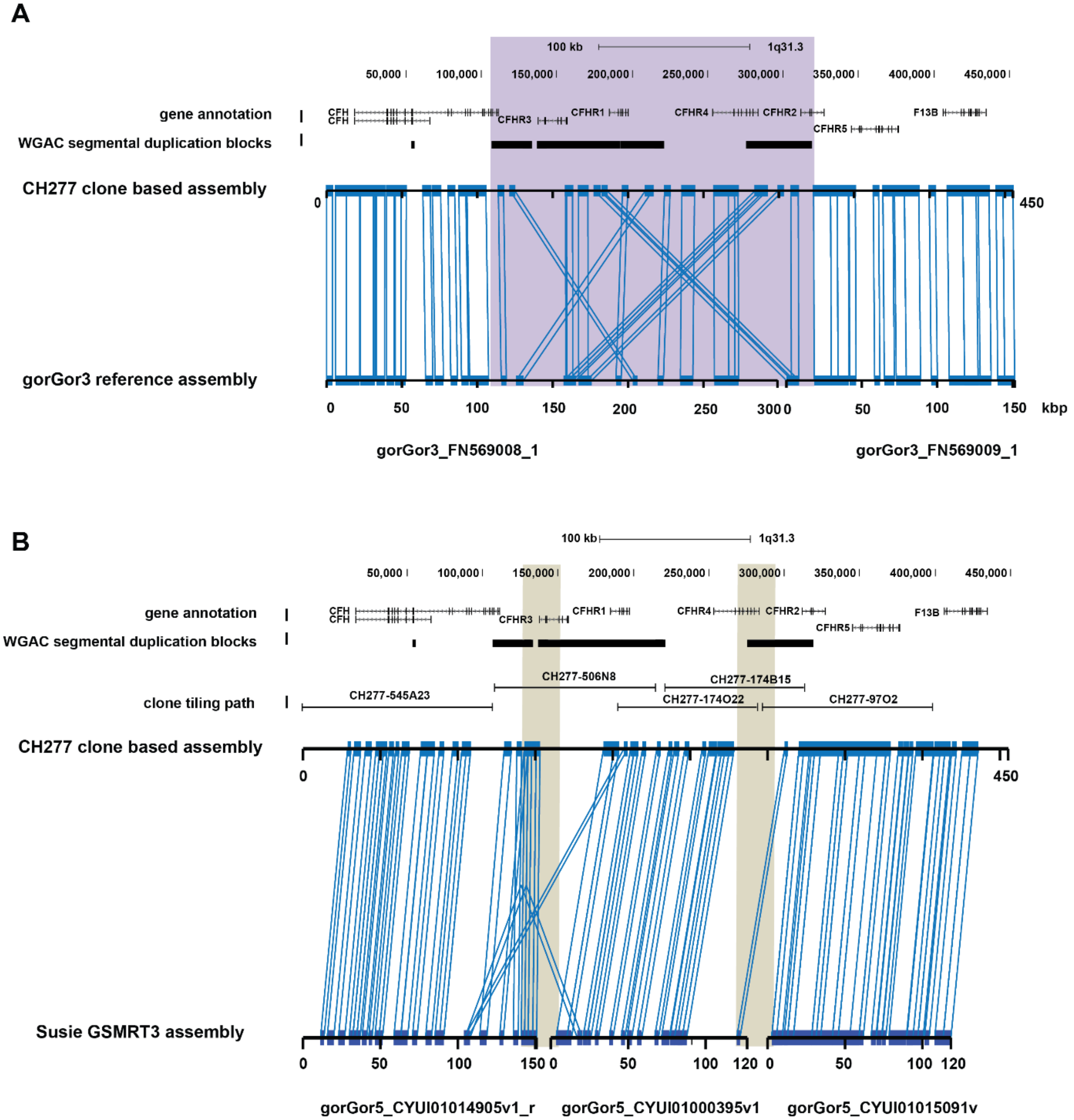

A
100 kb
1q31.3
50,000
100,000
150,000
200,000
250,000
300,000
350,000
400,000
450,000
gene annotation
WGAC segmental duplication blocks
CFH
CFH
CFHR3
CFHR1
CFHR4
CFHR2
CFHR5
F13B
CH277 clone based assembly
0
450
gorGor3 reference assembly
0
50
100
150
200
250
300
0
50
100
150
kbp
gorGor3_FN569008_1
gorGor3_FN569009_1
B
100 kb
1q31.3
50,000
100,000
150,000
200,000
250,000
300,000
350,000
400,000
450,000
gene annotation
WGAC segmental duplication blocks
CFH
CFH
CFHR3
CFHR1
CFHR4
CFHR2
CFHR5
F13B
clone tiling path
CH277-545A23
CH277-506N8
CH277-174B15
CH277-174O22
CH277-97O2
CH277 clone based assembly
0
450
Susie GSMRT3 assembly
0
50
100
150
0
50
100
120
0
50
100
120
gorGor5_CYUI01014905v1_r
gorGor5_CYUI01000395v1
gorGor5_CYUI01015091v

Fig. S8. Pairwise sequence comparison between the gorGor3/GSMRT3 gorilla reference assemblies and the newly constructed contig using CH277 large-insert BAC clones. A) A Miropeats comparison between gorGor3 and the CH277 BAC contig shows pairwise sequence differences between orthologous regions. Lines connect stretches of homologous sequence (threshold s = 2000) between the two assemblies with large breaks in sequence contiguity highlighted by purple-colored blocks. Gene annotations are shown based on mapping human RefSeq annotations from GRCh37 to the new alternate reference assembly. gorGor3 contains 22 sequence contigs that range from 288 bp – 182 kbp in size and is missing annotations for 4/5 *CFH* gene paralogs. Contrary to the 61 kbp of missing sequence reported at this locus in the gorGor3 assembly, we estimate that the number is approximately 12 kbp of sequence when compared to our CH277 contig. **B)** A Miropeats comparison between GSMRT3 and the CH277 BAC contig shows pairwise sequence differences between orthologous regions. Lines connect stretches of homologous sequence (threshold s = 2000) between the two assemblies with large breaks in sequence contiguity highlighted by tan-colored blocks. GSMRT3 contains three sequence contigs that range from 118–150 kbp in size and represents an improvement in sequence contiguity relative to gorGor3; however, 39.4 kbp are still missing from the final assembly.

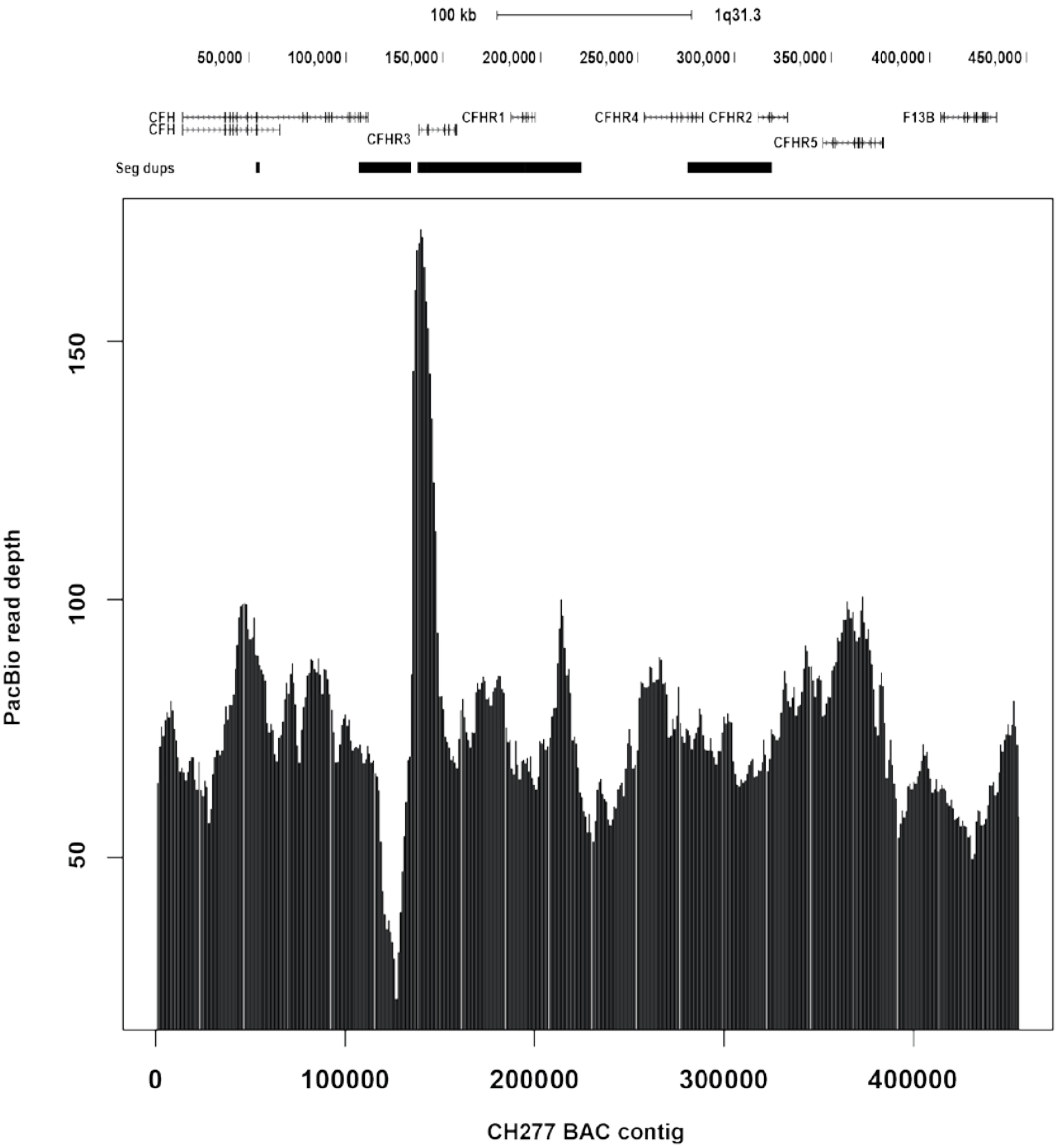
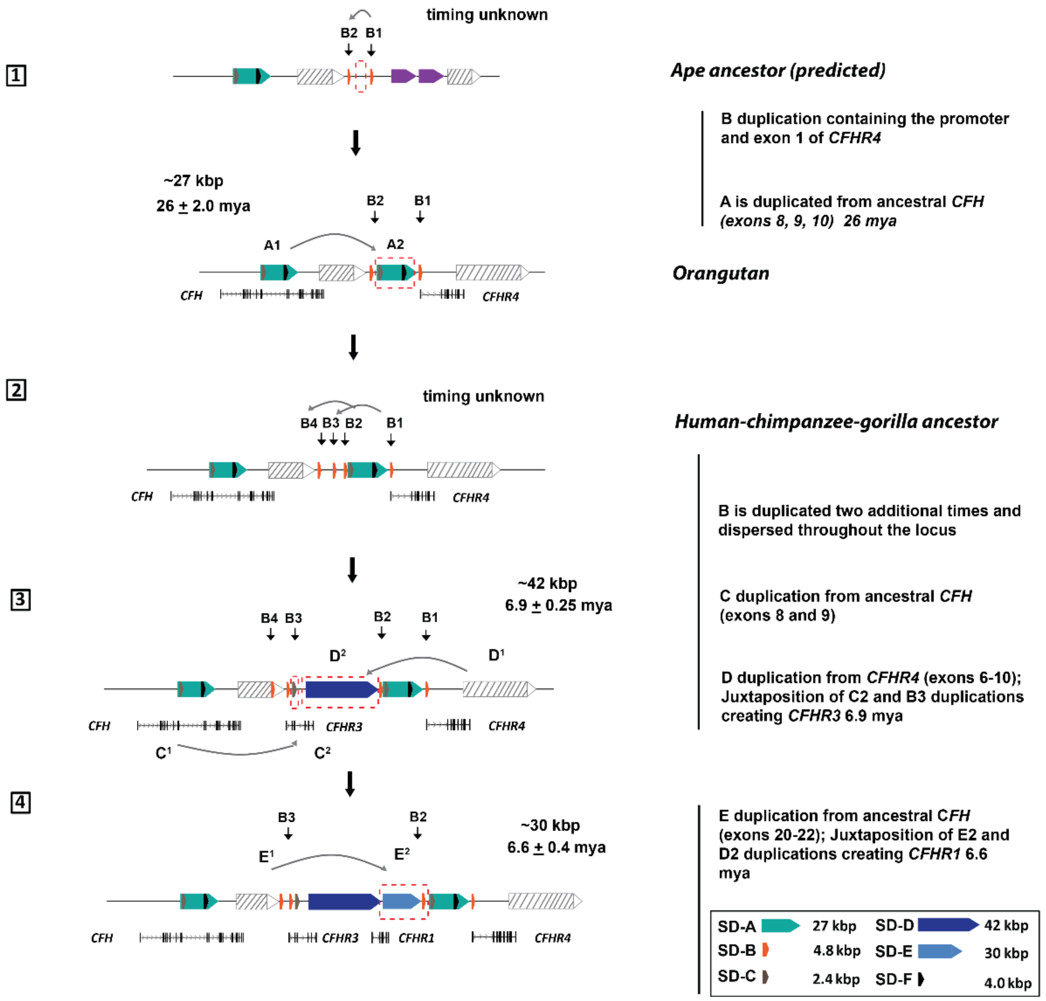


Fig. S9. PacBio sequence read-depth profile for the CH277 clone-based assembly of the *CFHR* locus. PacBio subreads from gorilla PacBio assembly (GSMRT3) are mapped to the CH277 BAC contig using BLASR. Raw read-depth was calculated for every 1,000 bp sliding windows across the assembly and the read-depth profile was created by plotting the average depth for each window. The sharp spike in read-depth corresponding to coordinates 120,000-140,000 is indicative of a duplication that is not resolved within our finally gorilla assembly. This position of increased read-depth maps partially to the SD-E and SD-B duplications as identified by Illumina-based read-depth estimates mapped to the GRCh37 assembly (Fig. 1D).



B duplication containing the promoter and exon 1 of *CFHR4*

A is duplicated from ancestral *CFH* *(exons 8, 9, 10) 26 mya*

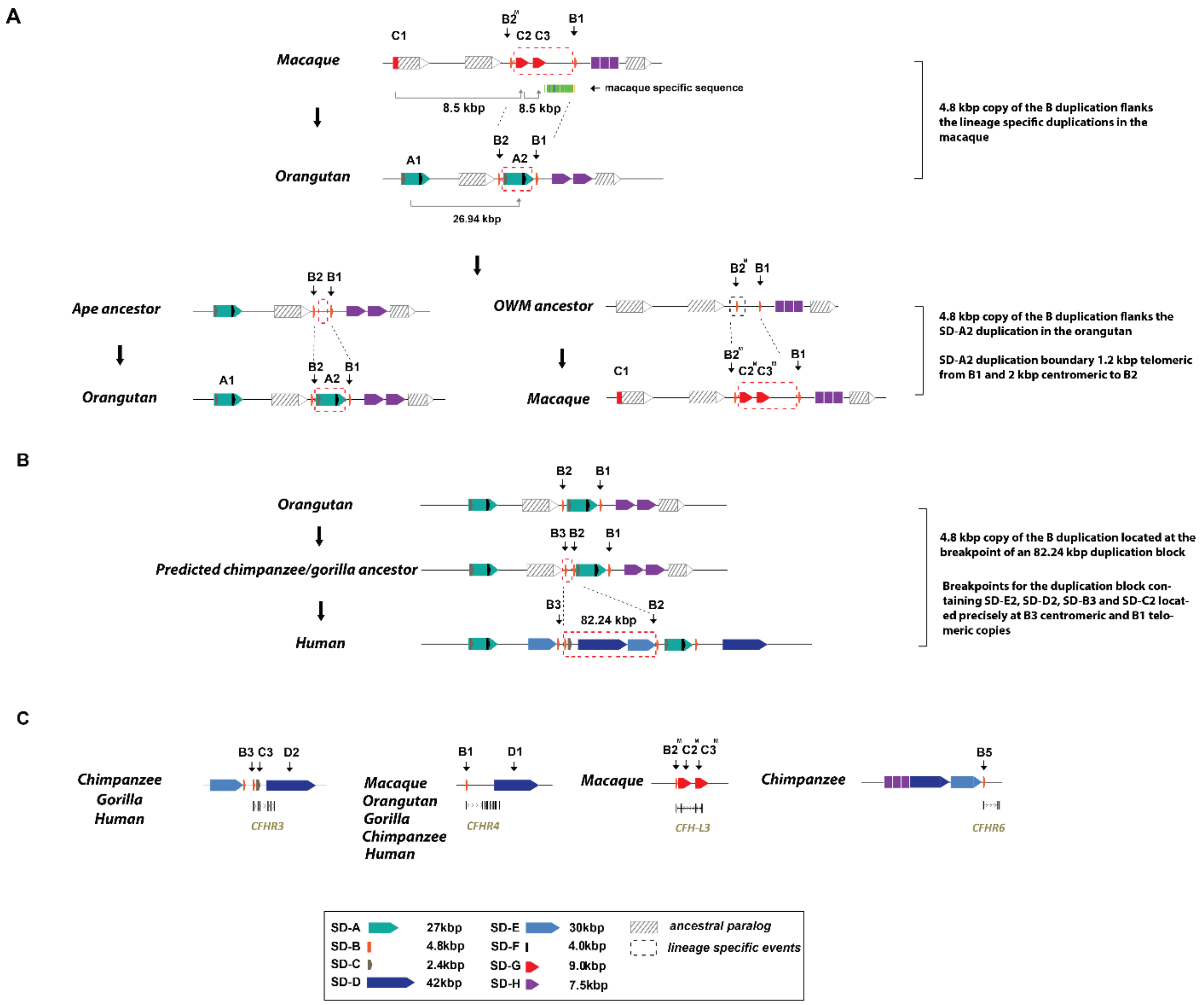
B is duplicated two additional times and dispersed throughout the locus

C duplication from ancestral *CFH* (exons 8 and 9)

D duplication from *CFHR4* (exons 6-10); Juxtaposition of C2 and B3 duplications creating *CFHR3* 6.9 mya

E duplication from ancestral *CFH* (exons 20-22); Juxtaposition of E2 and D2 duplications creating *CFHR1* 6.6 mya

Fig. S10. Proposed model for the evolution of *CFHR* SDs. The schematic depicts the extent of the *CFHR* duplication blocks (colored arrows), estimates of size (*SI Appendix*, Table S6) and evolutionary timing of events between each predicted intermediate genomic structure. The model shows the predicated evolutionary history of the SDs, beginning with the predicted structure of the ape ancestor to the most common haplotype present in modern-day humans. Gene annotations depict the formation of *CFHR* gene paralogs in conjunction with the corresponding duplications.



4.8 kbp copy of the B duplication flanks the lineage specific duplications in the macaque

4.8 kbp copy of the B duplication flanks the SD-A2 duplication in the orangutan

SD-A2 duplication boundary 1.2 kbp telomeric from B1 and 2 kbp centromeric to B2

4.8 kbp copy of the B duplication located at the breakpoint of an 82.24 kbp duplication block

Breakpoints for the duplication block containing SD-E2, SD-D2, SD-B3 and SD-C2 located precisely at B3 centromeric and B1 telomeric copies

Fig. S11. Evolutionary breakpoints and fusion transcripts mapping to the promoter duplication. A) *CFHR4* promoter duplications mapping to the boundaries of evolutionary breakpoints in macaque and orangutan. **B)** An 82.24 kbp duplication transposition flanked by *CFHR4* promoter duplications at the breakpoints. **C)** *CFHR4* promoter duplications compose the 5' UTR and exon 1 of four *CFHR* fusion transcripts.

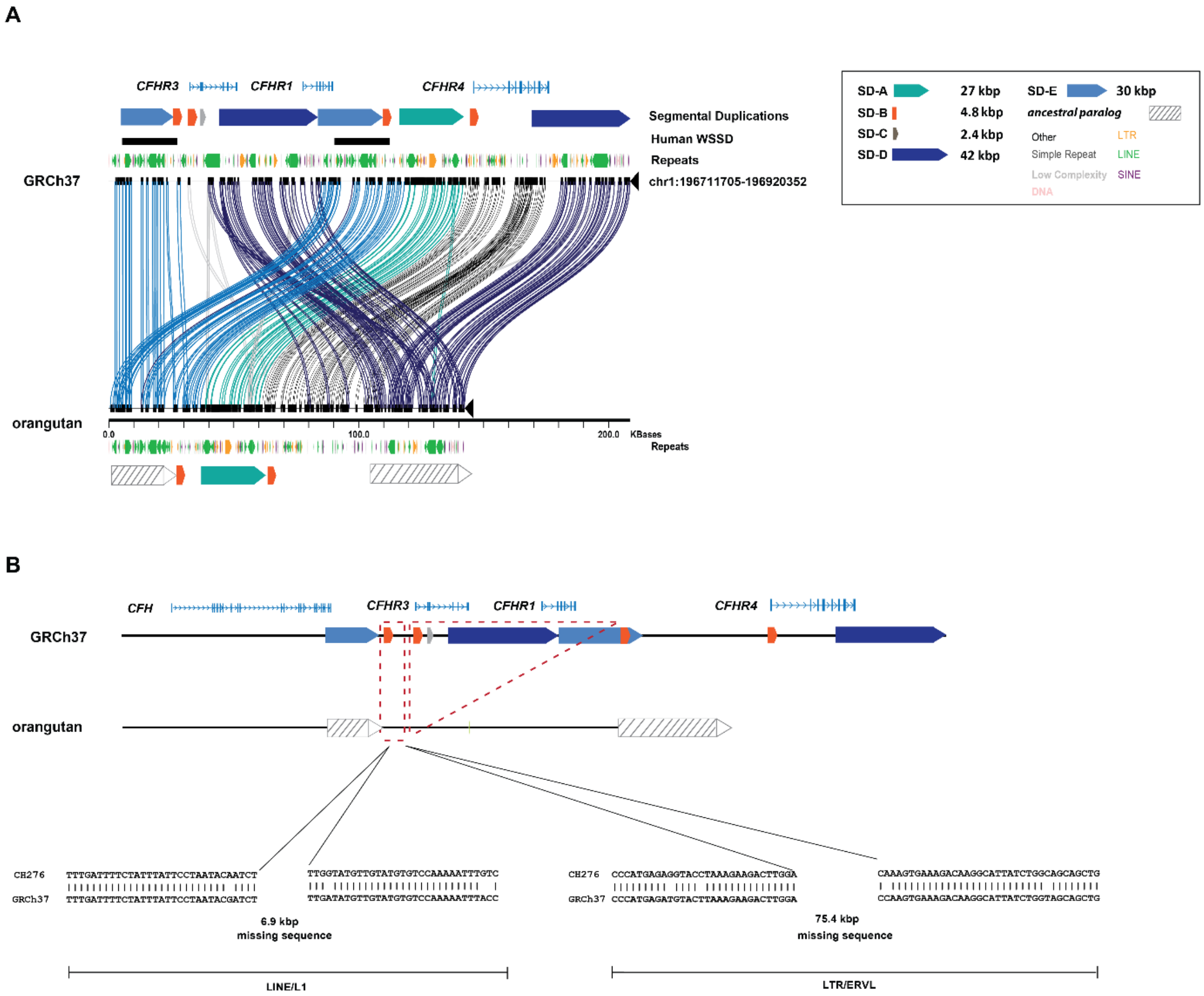


Fig. S12. Sequence refinement of *CFHR* duplication integration sites in orangutan. A pairwise alignment between human and orangutan shows two breakpoints resulting from two independent duplicative transpositions. Both breakpoints boundaries map to SD-B duplications.

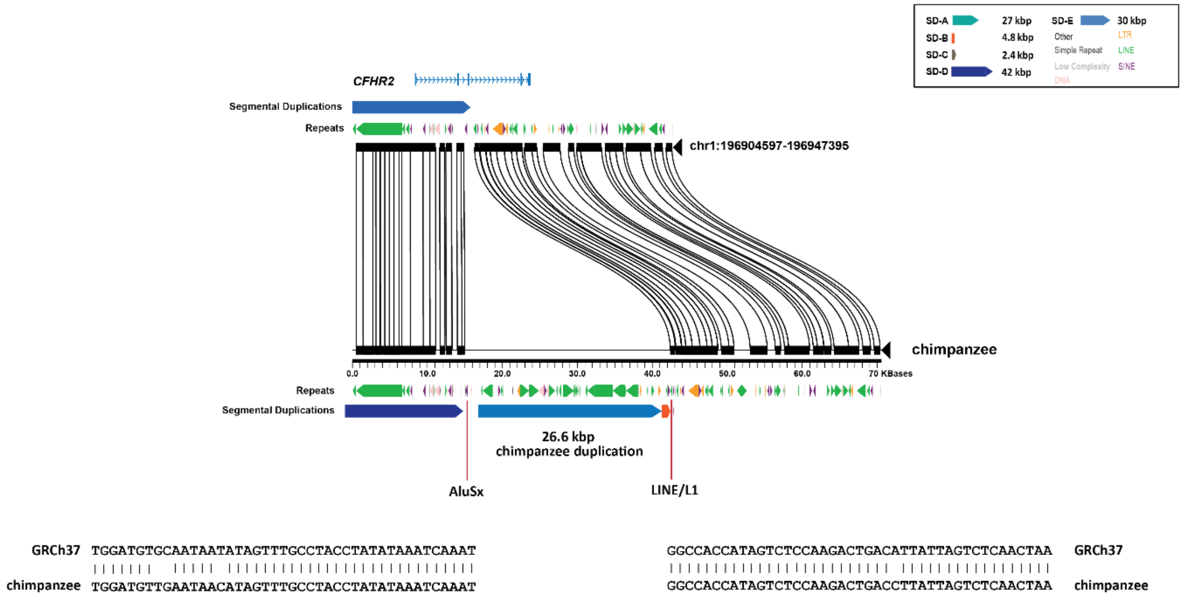


Fig. S13. Sequence refinement of SD-E and SD-B duplication integration sites in chimpanzee. The duplication integration site occurred 455 bp upstream of an AluSx element and at the boundary of an LINE/L1, with coordinated loss of 538 bp of unique sequence. At the telomeric boundary of the SD-E3 duplication (blue arrows) lies the SD-B5 promoter duplication creating the chimpanzee-specific *CFHR6*.

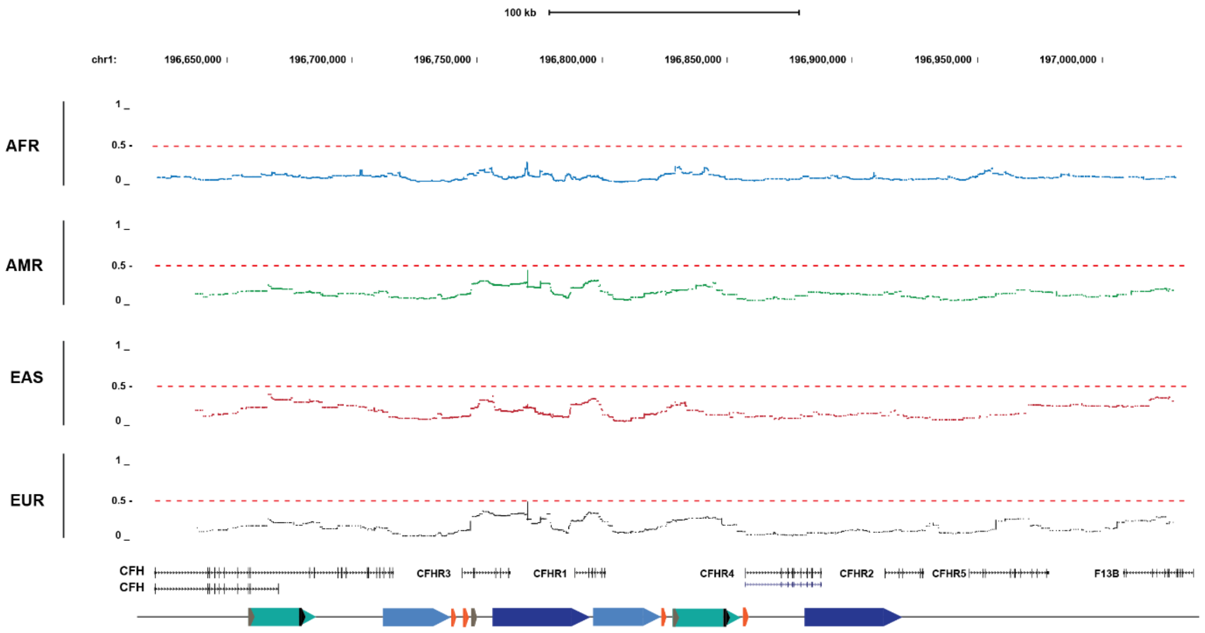


Fig. S14. Extended haplotype homozygosity analysis performed on four super populations from 1KG in 50 single-nucleotide polymorphism (SNP) windows. The eHH metric is plotted in windows of 50 SNPs (minor allele >5%) across the 1q31.3 locus in >2,000 individuals from four super populations (African [AFR], Americas [AMR], Europeans [EUR], and East Asians [EAS]). No significant loss of sequence diversity (eHH >0.5) indicative of long-range linkage disequilibrium is observed.

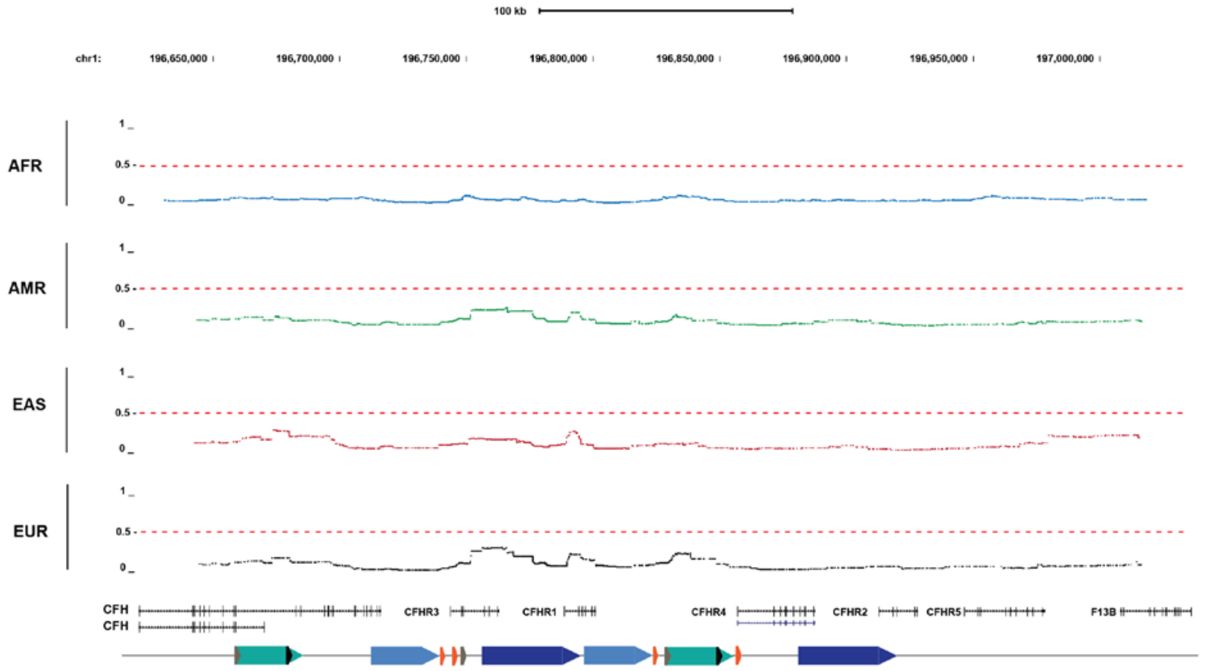


Fig. S15. Extended haplotype homozygosity analysis performed on four super populations from 1KG in 100 SNP windows. The eHH metric is plotted in windows of 100 SNPs (minor allele >5%) across the 1q31.3 locus in >2,000 individuals from four super populations (African [AFR], Americas [AMR], Europeans [EUR], and East Asians [EAS]). No significant loss of sequence diversity (eHH >0.5) indicative of long-range linkage disequilibrium is observed.

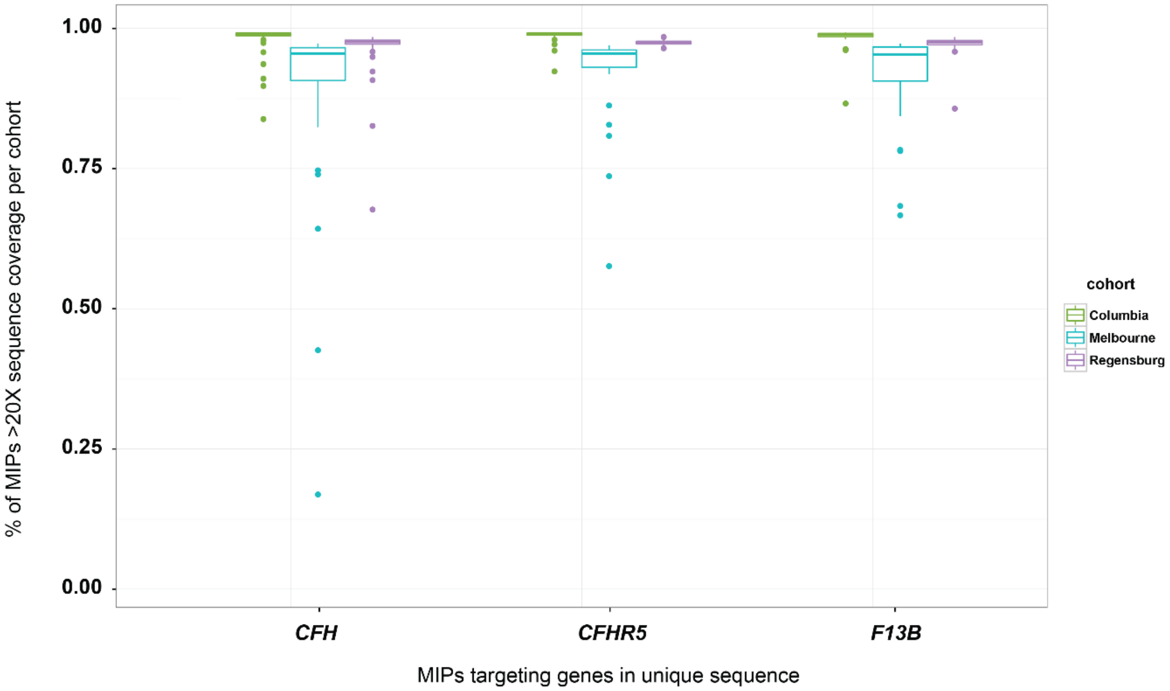
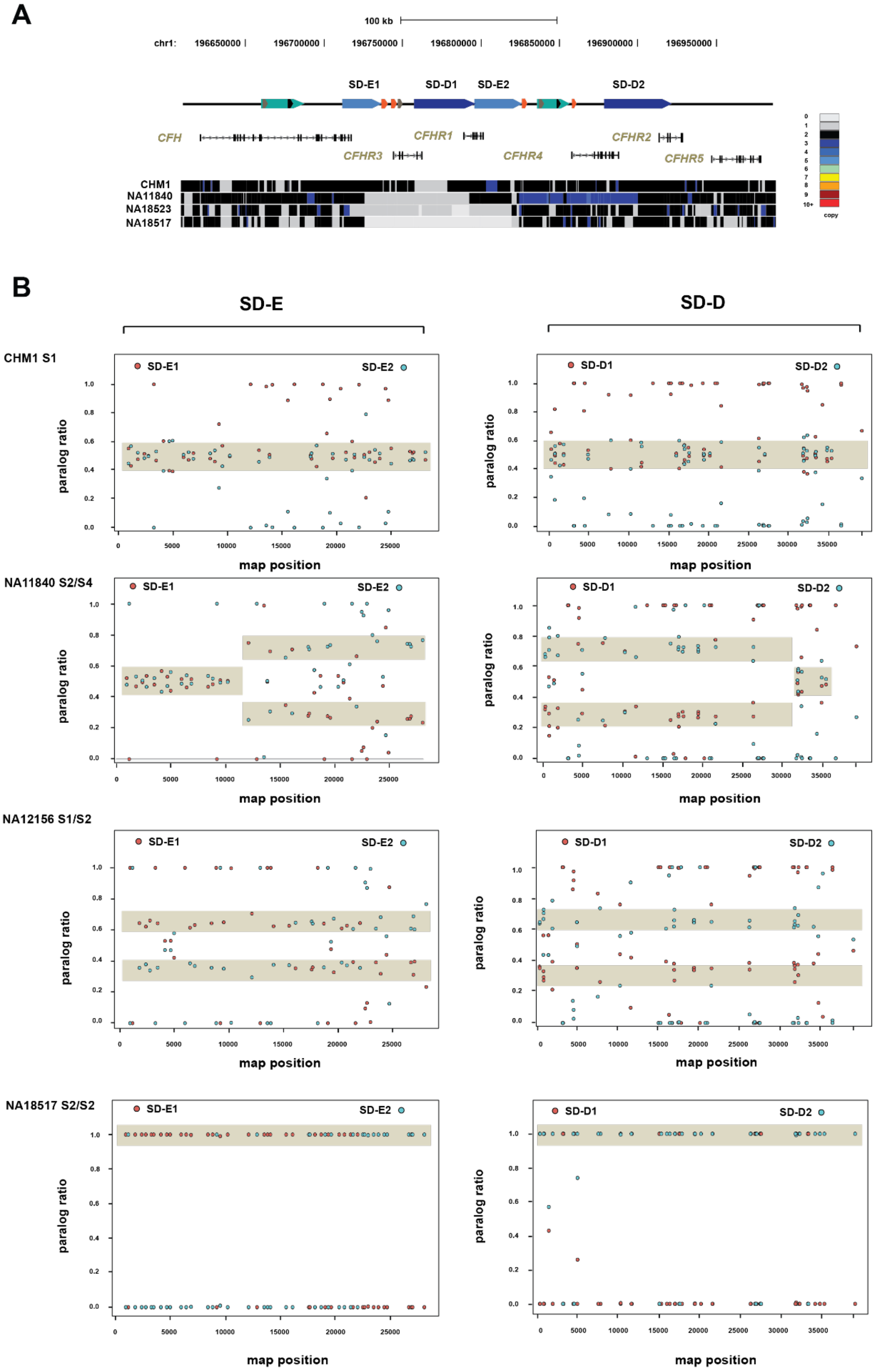


Fig. S16. A boxplot depicting the percentage of MIPs at >20X sequence coverage per cohort targeting unique genes at the *CFHR* locus. Coverage statistics were assessed based on 142 MIPs from three genes, *CFH*, *CFHR5* and *F13B*, primarily anchored in unique sequence.



A
100 kb
chr1:
196650000
196700000
196750000
196800000
196850000
196900000
196950000
SD-E1
SD-D1
SD-E2
SD-D2
CFH
CFHR1
CFHR2
CFHR3
CFHR4
CFHR5
CHM1
NA11840
NA18523
NA18517
0
1
2
3
4
5
6
7
8
9
10+
copy
B
SD-E
SD-D
CHM1 S1
NA11840 S2/S4
NA12156 S1/S2
NA18517 S2/S2
SD-E1
SD-E2
SD-D1
SD-D2
paralog ratio
map position
0.0
0.2
0.4
0.6
0.8
1.0
0
5000
10000
15000
20000
25000
30000
35000

Fig. S17. MIP sequencing refines *CFHR* rearrangement breakpoints. A comparison between whole-genome Illumina sequencing **(A)** and MIP-based copy number typing **(B)** shows concordant copy number estimates among four individuals from 1KG. **A)** SUNK-based copy number heatmaps over *CFHR* SDs from four 1KG individuals depict diversity of *CFHR* structural haplotypes: CHM1 (S1 haplotype), NA11840 (S2 and S4 haplotype), NA18523 (S1 and S2 haplotype) and NA18517 (S2/S2 haplotype). **B)** Resequencing of duplicate loci using MIPs targeted towards sequence differences that distinguish duplicate copies (red dots vs. blue dots). The ratio of each paralog distinguishing variant (>160 unique sequence differences) is plotted over the length of the duplication (SD-D and SD-E). Paralog-specific copy number and breakpoint refinement (tan shading) can be estimated in each individual and shows concordance with both Illumina whole-genome sequencing and BAC-based sequence assembly.

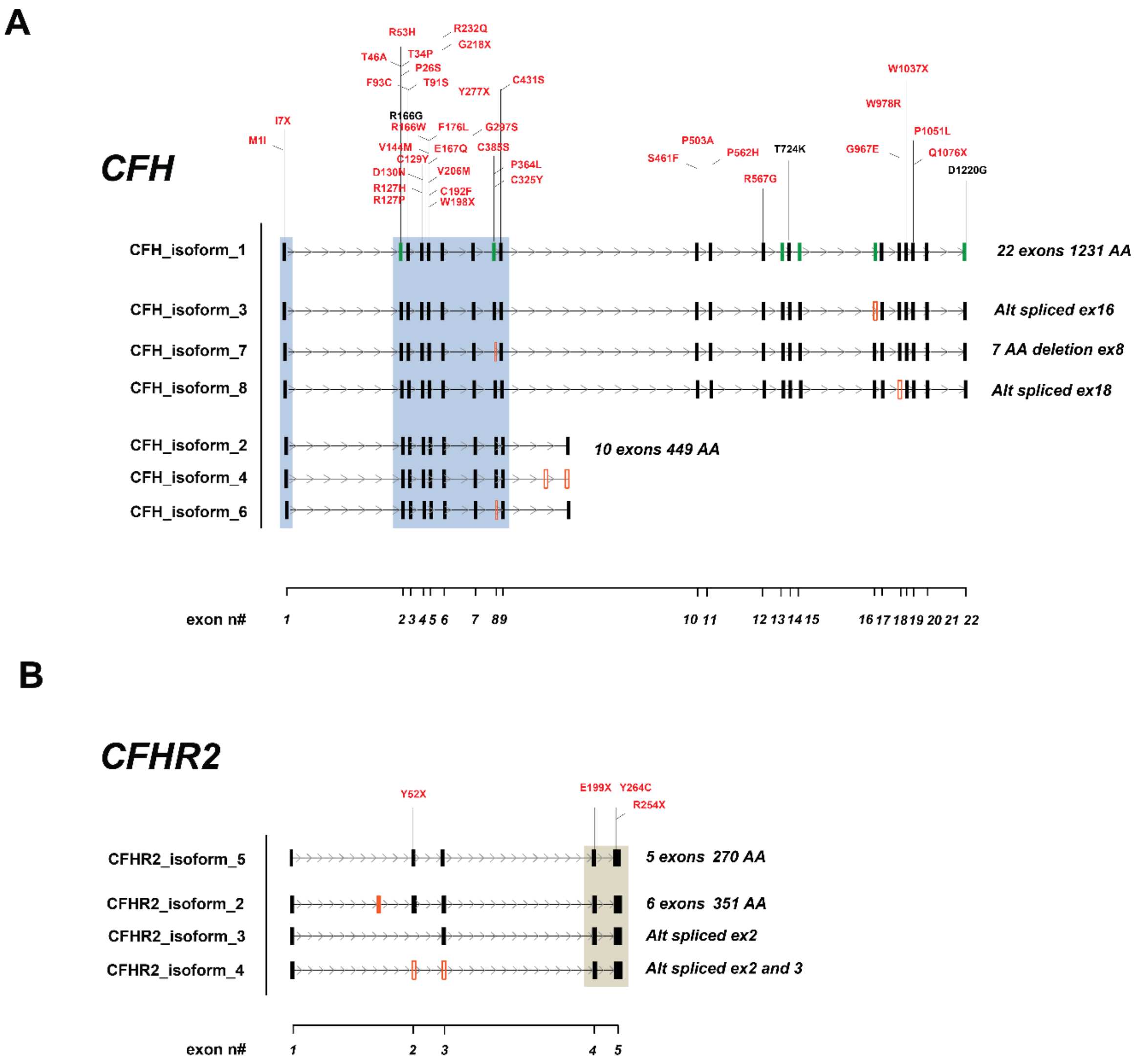


Fig. S18. Pattern of missense and LGD mutations, plotted against the most common isoforms detected for *CFH* and *CFHR2*. A) *CFH* nonsynonymous mutations in cases (red) and controls (black) are plotted against a schematic of the transcript isoforms detected using PacBio Iso-Seq. Exons annotated in orange represent alternative splicing observed from long-read isoform data and exons annotated in green harbor amino acids showing signals of natural selection. The burden of nonsynonymous variation is clustered at the N-terminus (blue shading) and maps to canonical exons. **B)** *CFHR2* nonsynonymous mutations associated with AMD map to unique sequence and canonical exons defined by PacBio Iso-Seq (tan shading).

SI DATASETS

Dataset S1: Copy number diversity for the *CFHR3/CFHR1* and *CFHR1/CFHR4* CNPs among 1KG and HGDP individuals

Dataset S2: Vst analysis calculated for the *CFHR3-1* and *CFHR1-4* CNPs in 1KG and HGDP individuals

Dataset S3: List of SMRT-sequenced large-insert clones from humans and NHPs

Dataset S4: BAC and fosmid libraries sequenced as part of this study

Dataset S5: Human and NHP *CFHR* contigs assembled from large-insert clones

Dataset S6: List of human and NHP SDs at 1q31.3

Dataset S7: Breakpoint location and sequence characterization among human and NHP haplotypes

Dataset S8: Regions of IGC identified amongst *CFHR* paralog sequences

Dataset S9: Quality metrics for the assembly of the 1q31.3 locus in NHP reference assemblies

Dataset S10: ORF annotation among *CFHR* gene paralogs for humans and NHPs

Dataset S11: RepeatMasker annotation for *CFHR* alternate reference assemblies

Dataset S12: PacBio Iso-Seq transcripts identified by SMRT of liver cDNA

Dataset S13: Assessment of the dN/dS ratio between *CFH* and its gene paralogs

Dataset S14: MIP sequences used to analyze *CFH* and *CFHR* gene paralogs

Dataset S15: *CFH* missense mutations identified in the UW AMD case-control cohort

Dataset S16: Association analysis of rare, nonsynonymous variation in AMD cases and controls

Dataset S17: Association analysis for common missense variants in the *CFH* gene with AMD

Dataset S18: *CFHR* gene paralog missense mutations identified in the UW AMD case-control cohort

Dataset S19: Association analysis between common missense variation of *CFHR* paralogs and AMD

Dataset S20: Combined set of *CFH* missense mutations among five sequencing studies

Dataset S21: Frequencies of *CFHR* structural haplotypes in AMD cases and controls

SI REFERENCES

1. Sudmant PH, *et al.* (2015) Global diversity, population stratification, and selection of human copy-number variation. *Science* 349(6253).
2. Sudmant PH, *et al.* (2015) An integrated map of structural variation in 2,504 human genomes. *Nature* 526(7571):75-81.
3. Prado-Martinez J, *et al.* (2013) Great ape genetic diversity and population history. *Nature* 499(7459):471-475.
4. Hach F, *et al.* (2014) mrsFAST-Ultra: a compact, SNP-aware mapper for high performance sequencing applications. *Nucleic Acids Res* 42(W1):W494-W500.
5. Sudmant PH, *et al.* (2013) Evolution and diversity of copy number variation in the great ape lineage. *Genome Res* 23(9):1373-1382.
6. Sudmant PH, *et al.* (2010) Diversity of Human Copy Number Variation and Multicopy Genes. *Science* 330(6004):641-646.
7. Redon R, *et al.* (2006) Global variation in copy number in the human genome. *Nature* 444(7118):444-454.
8. Huddleston J, *et al.* (2014) Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res* 24(4):688-696.
9. Steinberg KM, *et al.* (2012) Structural diversity and African origin of the 17q21.31 inversion polymorphism. *Nat Genet* 44(8):872-880.
10. Chin C-S, *et al.* (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Meth* 10(6):563-569.
11. Parsons JD (1995) Miropeats: graphical DNA sequence comparisons. *Computer applications in the biosciences : CABIOS* 11(6):615-619.
12. Altschul SF, Gish W, Miller W, Myers EW, & Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215(3):403-410.
13. Bailey JA, Yavor AM, Massa HF, Trask BJ, & Eichler EE (2001) Segmental Duplications: Organization and Impact Within the Current Human Genome Project Assembly. *Genome Res* 11(6):1005-1017.
14. Jiang Z, Hubley R, Smit A, & Eichler EE (2008) DupMasker: A tool for annotating primate segmental duplications. *Genome Res* 18(8):1362-1368.
15. Wu TD & Watanabe CK (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21(9):1859-1875.
16. Kent WJ (2002) BLAT:The BLAST-Like Alignment Tool. *Genome Res* 12(4):656-664.
17. Chaisson MJ & Tesler G (2012) Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* 13(1):238.
18. Katoh K & Standley DM (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* 30(4):772-780.
19. Tamura K, *et al.* (2011) MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Mol Biol Evol* 28(10):2731-2739.
20. Tajima F (1993) Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics* 135(2):599-607.
21. Mohajeri K, *et al.* (2016) Interchromosomal core duplicons drive both evolutionary instability and disease susceptibility of the Chromosome 8p23.1 region. *Genome Res* 26(11):1453-1467.
22. Dennis Megan Y, *et al.* (2012) Evolution of Human-Specific Neural SRGAP2 Genes by Incomplete Segmental Duplication. *Cell* 149(4):912-922.
23. Day N, Hemmaplardh A, Thurman RE, Stamatoyannopoulos JA, & Noble WS (2007) Unsupervised segmentation of continuous genomic data. *Bioinformatics* 23(11):1424-1426.
24. Sawyer S (1989) Statistical tests for detecting gene conversion. *Mol. Biol. Evol.* 6(5):526-538.

25. Dumont BL & Eichler EE (2013) Signals of Historical Interlocus Gene Conversion in Human Segmental Duplications. *PLOS ONE* 8(10):e75949.
26. Aken BL*, et al.* (2016) The Ensembl gene annotation system. *Database* 2016.
27. Yang Z (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol* 24(8):1586-1591.
28. Yang Z, Wong WSW, & Nielsen R (2005) Bayes Empirical Bayes Inference of Amino Acid Sites Under Positive Selection. *Mol Biol Evol* 22(4):1107-1118.
29. Bray NL, Pimentel H, Melsted P, & Pachter L (2016) Near-optimal probabilistic RNA-seq quantification. *Nat Biotech* 34(5):525-527.
30. Boyle EA, O'Roak BJ, Martin BK, Kumar A, & Shendure J (2014) MIPgen: optimized modeling and design of molecular inversion probes for targeted resequencing. *Bioinformatics* 30(18):2670-2672.
31. Hiatt JB, Pritchard CC, Salipante SJ, O'Roak BJ, & Shendure J (2013) Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Res* 23(5):843-854.
32. Cantsilieris S, Stessman HA, Shendure J, & Eichler EE (2017) Targeted Capture and High-Throughput Sequencing Using Molecular Inversion Probes (MIPs). *Genotyping: Methods and Protocols*, (Springer New York, New York, NY), pp 95-106.
33. O'Roak BJ*, et al.* (2012) Multiplex Targeted Sequencing Identifies Recurrently Mutated Genes in Autism Spectrum Disorders. *Science* 338(6114):1619-1622.
34. Garrison E & Marth G (2012) Haplotype-based variant detection from short-read sequencing. *arXiv* arXiv:1207.3907.
35. Coe BP*, et al.* (2014) Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat Genet* 46(10):1063-1071.
36. McLaren W*, et al.* (2016) The Ensembl Variant Effect Predictor. *Genome Biol* 17(1):122.
37. Lek M*, et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536(7616):285-291.
38. Nuttle X*, et al.* (2013) Rapid and accurate large-scale genotyping of duplicated genes and discovery of interlocus gene conversions. *Nat Meth* 10(9):903-909.
39. Nuttle X, Itsara A, Shendure J, & Eichler EE (2014) Resolving genomic disorder–associated breakpoints within segmental DNA duplications using massively parallel sequencing. *Nat protocols* 9(6):1496-1513.
40. Fritsche LG*, et al.* (2016) A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nat Genet* 48(2):134-143.
41. Grassmann F*, et al.* (2015) A Candidate Gene Association Study Identifies DAPL1 as a Female-Specific Susceptibility Locus for Age-Related Macular Degeneration (AMD). *Neuromolecular Med* 17(2):111-120.
42. Purcell S*, et al.* (2007) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet* 81(3):559-575.
43. Li M*, et al.* (2006) CFH haplotypes without the Y402H coding variant show strong association with susceptibility to age-related macular degeneration. *Nat Genet* 38(9):1049-1054.