The American Journal of Human Genetics, Volume *108*

Supplemental information

Exome variant discrepancies due to reference-genome differences

He Li, Moez Dawood, Michael M. Khayat, Jesse R. Farek, Shalini N. Jhangiani, Ziad M. Khan, Tadahiro Mitani, Zeynep Coban-Akdemir, James R. Lupski, Eric Venner, Jennifer E. Posey, Aniko Sabo, and Richard A. Gibbs

Supplemental Figures

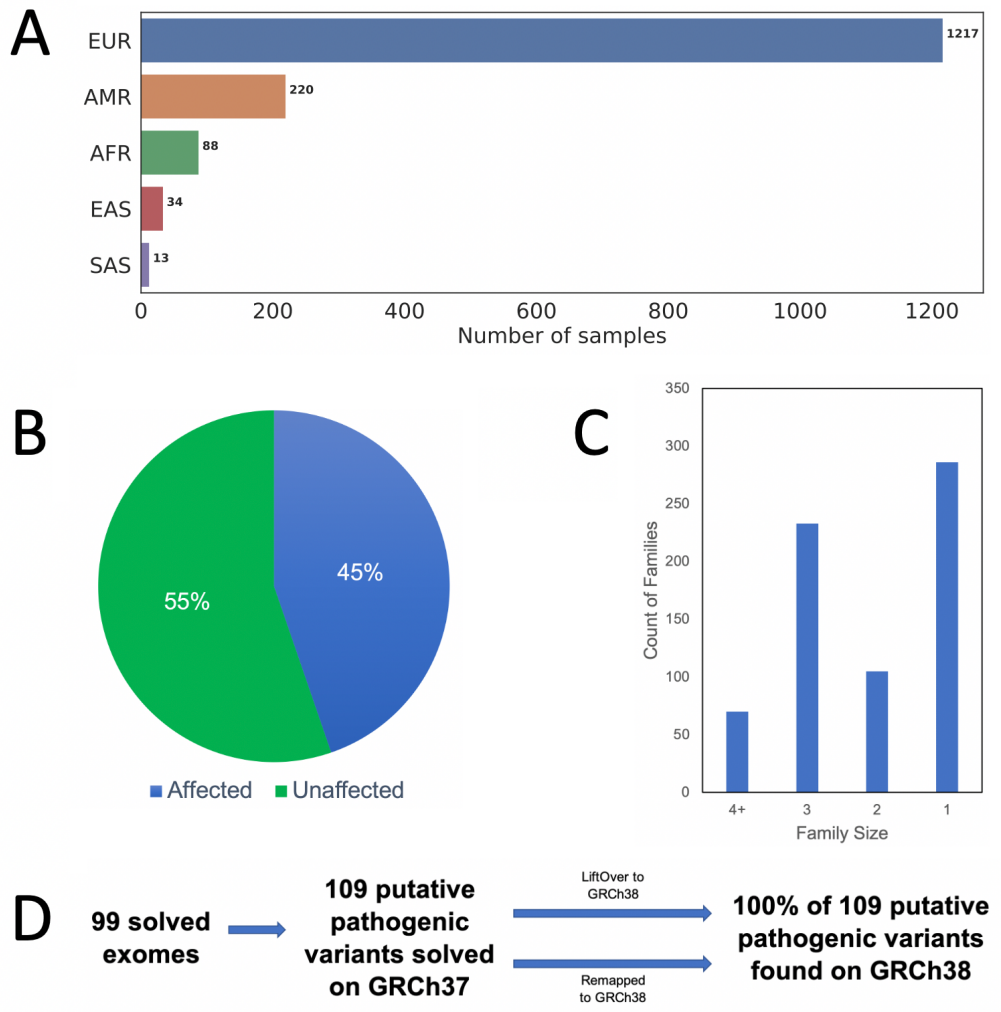

Figure S1. Demographics of 1,572 individuals in this study. (A) shows the genetic ancestry, and (B) shows disease status of the individuals in our study. The majority of the individuals were recruited as families in this study, and the family size is shown in (C). (D) Overall, 127 probands have had their exomes rigorously analyzed of which 99 have been assigned a molecular diagnosis and are considered putatively solved. From these 99 solved exomes, a total of 109 putative pathogenic variants were assigned as molecular diagnoses. These analyses were done on the GRCh37 reference. All these 109 putative pathogenic variants were still discoverable when lifting-over to GRCh38 or remapping the exomes to GRCh38.

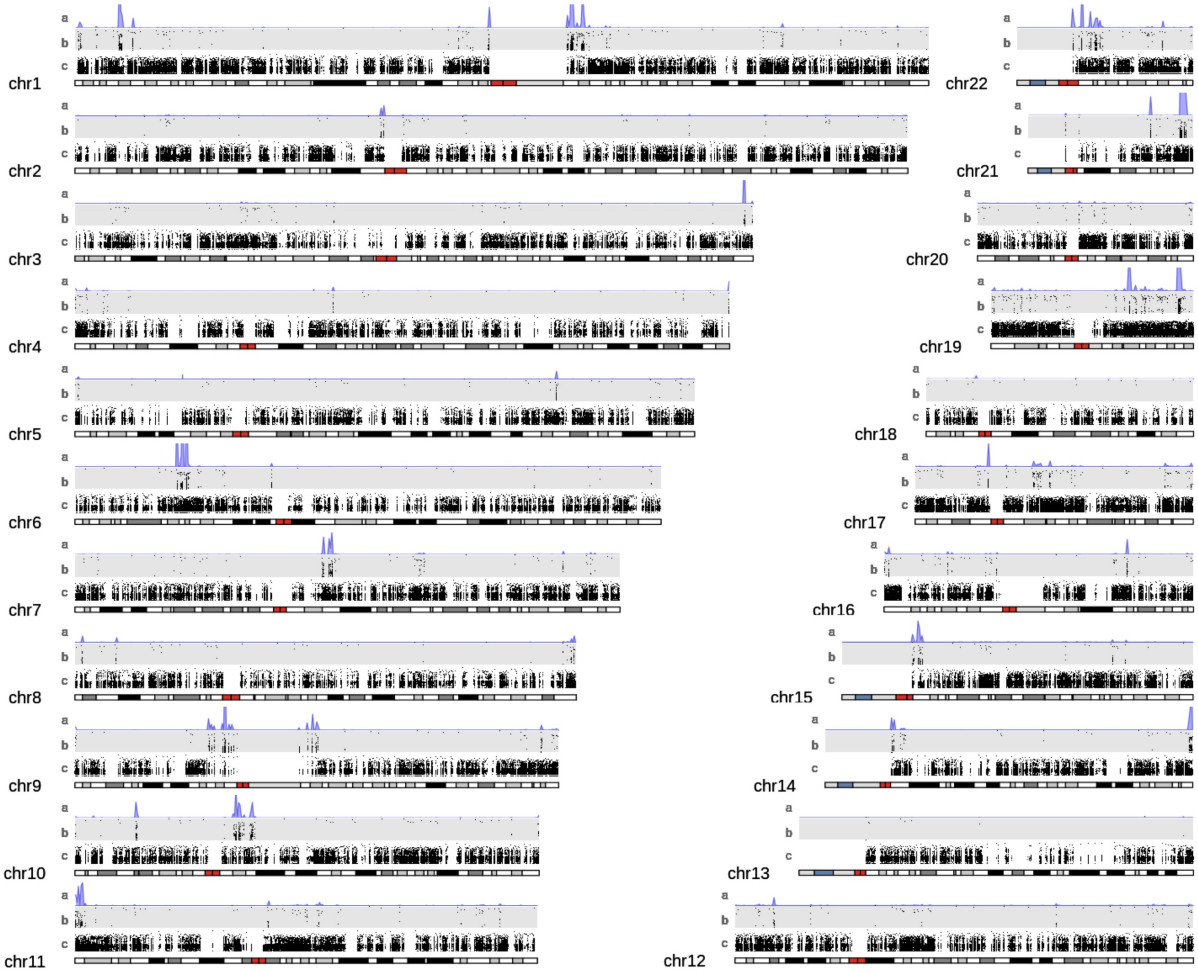

Figure S2. Genomic location of discordant variants found on GRCh37. On each chromosome, Panel (a) shows the density of all the discordant variants; Panel (b) shows all the discordant variants in rainfall plots (y-axis indicates distances between consecutive variants in a $\log_{10}$ scale); and Panel (c) shows all the variants found across all samples in rainfall plots.

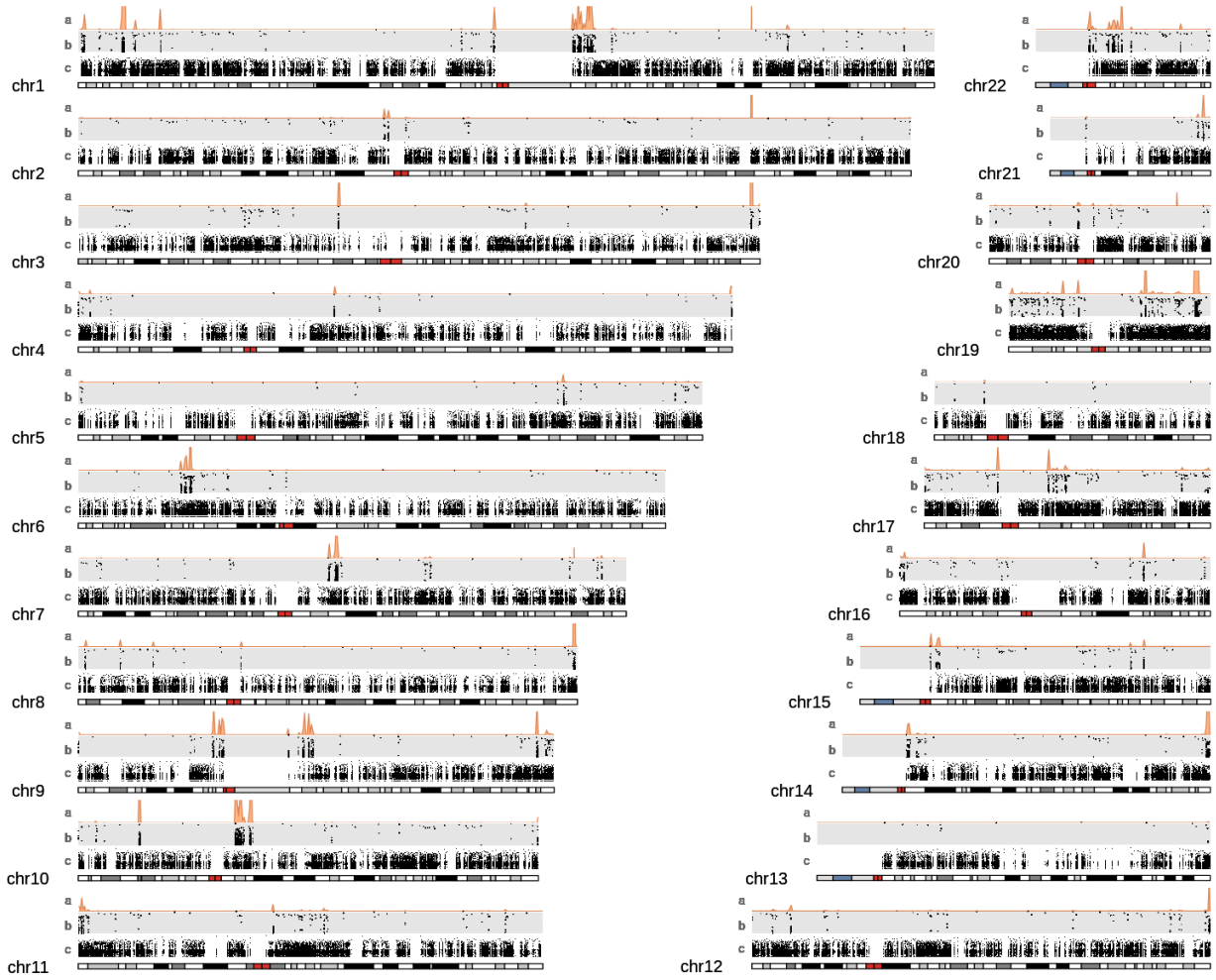

Figure S3. Genomic location of discordant variants found on GRCh38. On each chromosome, Panel (a) shows the density of all the discordant variants; Panel (b) shows all the discordant variants in rainfall plots (y-axis indicate distances between consecutive variants in a $\log_{10}$ scale); and Panel (c) shows all the variants found across all samples in rainfall plots.

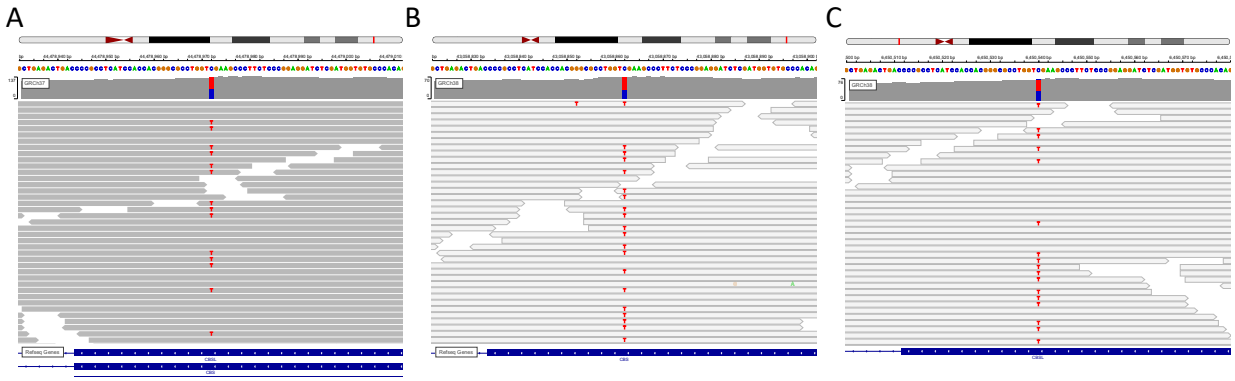

Figure S4. Example IGV screenshots of the read alignment in *CBS* and *CBSL* genes on GRCh37 and GRCh38. (A) shows the read alignment on GRCh37 that contains both *CBS* and *CBSL* genes where a variant was called; (B) shows the read alignment on GRCh38 that contains the gene *CBS*, and (C) shows the read alignment on GRCh38 that contains the gene *CBSL* (different loci from *CBS*). Neither of the variant on GRCh38 was called. Solid reads indicate mapping score > 30, whereas blank reads indicate mapping score of zero.

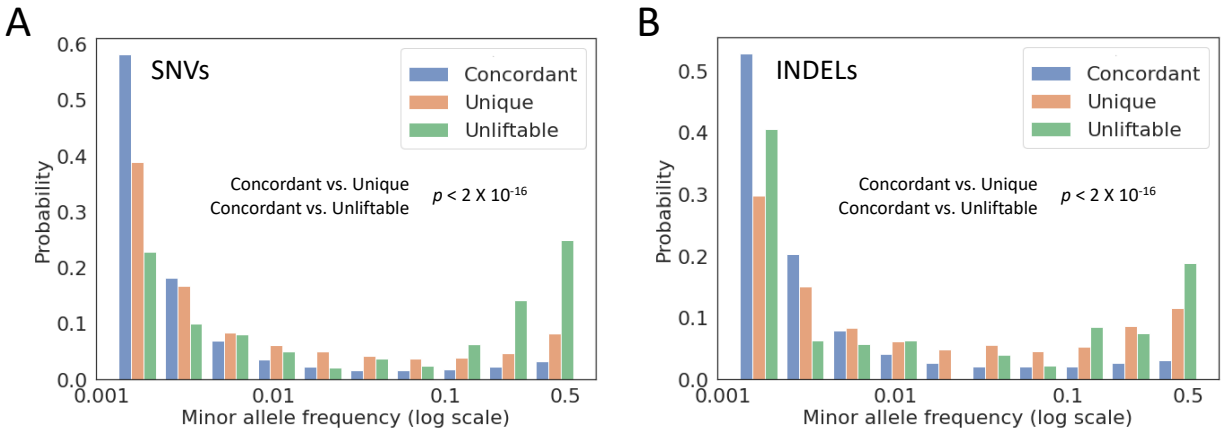

Figure S5. Comparison of minor allele frequency (MAF) distribution between different variant sets. The probably of MAF for Concordant variants, Unique variants (including both GRCh37 unique and GRCh38 unique variants), and Unliftable variants (including both GRCh37 unliftable and GRCh38 unliftable variants) were plotted for SNVs (A) and INDELs (B) separately. The statistical tests were performed using the Mann-Whitney U test.

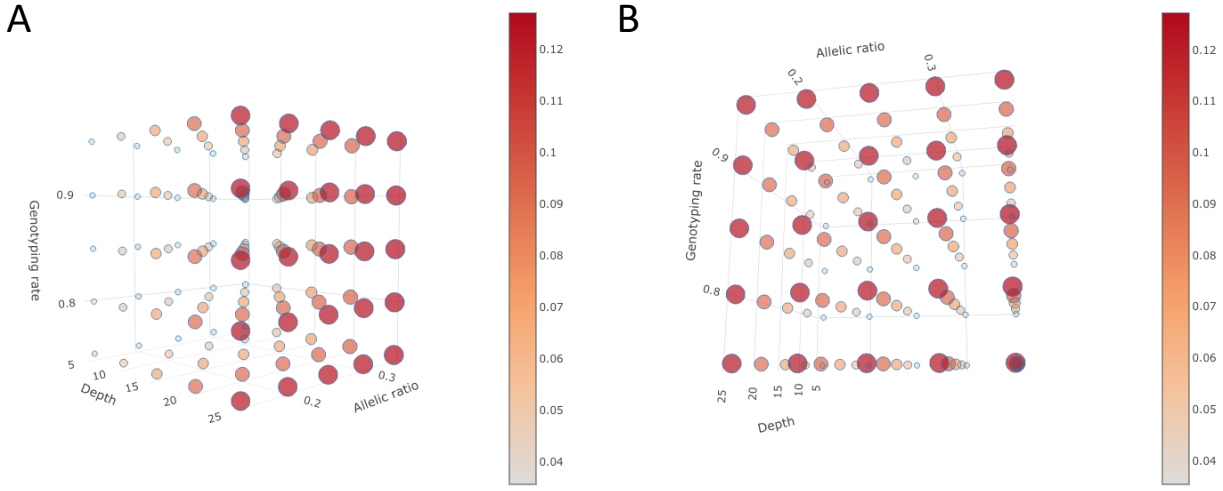

Figure S6. The rates at which individual genotypes were under different criteria. The quality control criteria from the following filtering ranges were selected: depth greater than [5, 10, 15, 20, 25]; allelic ratio above [0.15, 0.20, 0.25, 0.30, 0.35]; and genotyping rate above [0.75, 0.8, 0.85, 0.9, 0.95]. The rate at which individual genotypes were plotted for variants called on (A) GRCh37 and (B) GRCh38. The color and size of each dot are proportional to the associated rates.

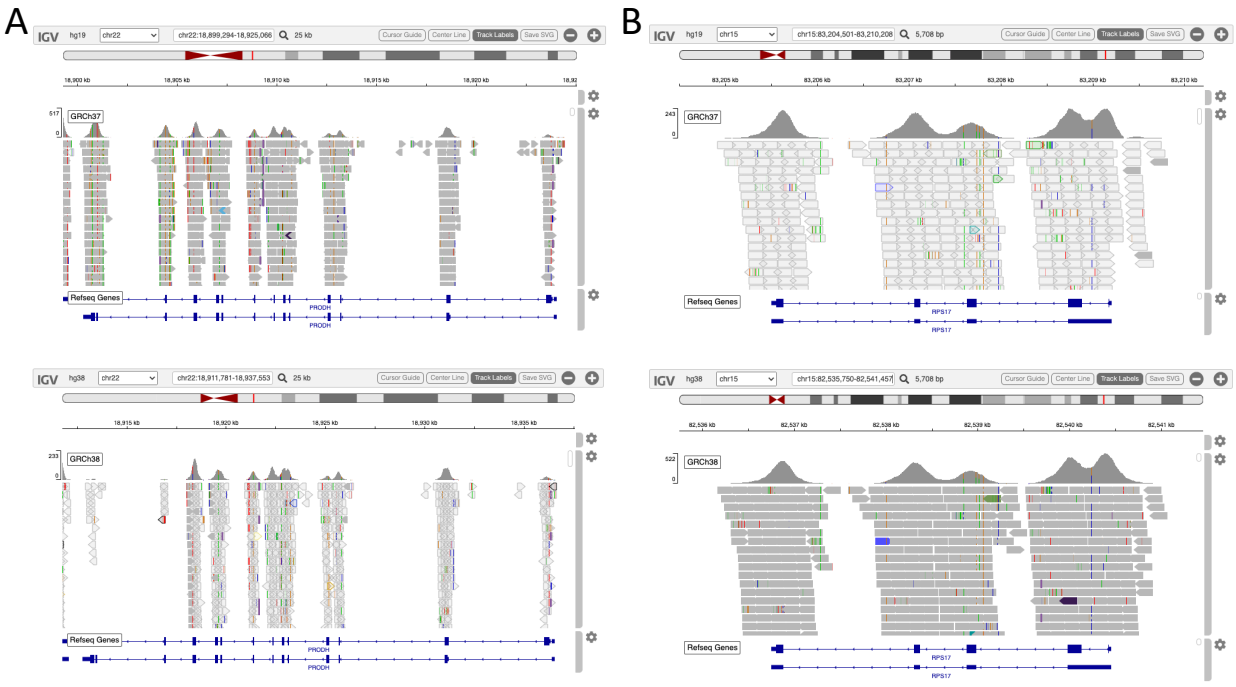

Figure S7. Example IGV screenshots of the read alignment. Read alignment in the gene regions of (A) PRODH and (B) RPS17 are shown. Top figures show alignment on GRCh37, and bottom figures show alignment on GRCh38.

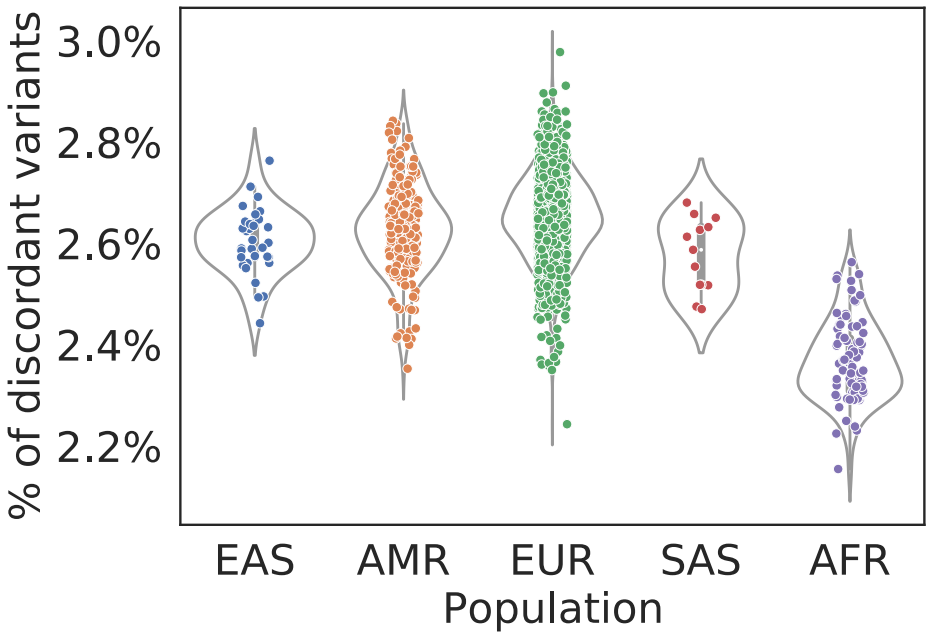

Figure S8. Percentage of discordant variants in different genetic ancestries. (EAS: East Asian; AMR: Hispanic genetic ancestry; EUR: European; SAS: South Asian; AFR: African American)

Supplemental Tables (Excel Spreadsheets)

Table S1. Genomic windows on GRCh37 enriched for discordant variant calls

Table S2. Genomic windows on GRCh38 enriched for discordant variant calls

Table S3. Results of enrichment analyses of genomic features within GHOST regions

Table S4. Genes enriched for unique variants called only by the GRCh37 or GRCh38 reference

Table S5. Discordant variants with potential deleterious effect

Table S6. Genes enriched for discordant variant calls and their associated phenotype / trait from previous GWAS