Supplemental Materials

Table of contents

Extended Methods........2

Variant calling and filtering........2

Logistic regression........3

Pseudo-sibling genotype generation........4

Single-gene rare variant transmission test and multi-locus generalization........4

Oligogenic network discovery and depiction........5

Obtaining unconditional p-values........6

Supplemental Figures........7

Supplemental Figure 1: Calculation of highest-order gene sets........7

Supplemental Figure 2: Counts of oligogenic pairs from synonymous mutations........7

Supplemental Figure 3: Expression patterns of GCOD pair genes........8

Supplemental Figure 4: Canonical gene set enrichment (extended)........9

Supplemental Figure 5: Conservation and structural impact of disease-associated mutations in *GATA6* and *POR*........9

Supplemental Figure 6: Mouse histology and echocardiography........10

Extended Methods

Variant calling and filtering

Whole Exome Sequencing data from the 3382 CHD trios were sequenced using the Illumina HiSeq 2000 and mapped to the GRCh37 reference genome using Eland, as described previously (Jin et al. 2017; Morton et al. 2021; Zaidi et al. 2013). Protein-coding mutations were filtered at the site level using SAMtools (H. Li et al. 2009) with Mapping Quality score > 59 and Genotype Quality (GQ) Mean ≥ 85, and individual variant calls were filtered at GQ ≥ 60. Variants with non-Mendelian transmission patterns were excluded from analysis unless determined to be *de novo,* or a result of uniparental disomy or heterozygous deletion. Uniparental disomy and structural deletions were determined using UPDio software (King et al. 2014) with default parameters. *De novo* variants were called using the TrioDeNovo program (Wei et al. 2015) and accepted if the minor allele frequency (MAF) of the variant is below 4x10^{-4}, with a minimum of 5 alternative reads and 10 total reads in the proband, and a minimum of 10 reference reads in the parents (with a maximum alternate allele ratio of 3.5%), as has been done previously (Homsy et al. 2015). All variants were called and are reported in GRCh37 coordinates.

For both datasets, we annotated variants using ANNOVAR version date 2021-06-08 (K. Wang, Li, and Hakonarson 2010). We restricted our analysis to variants annotated as MAF<5% in either the gnomAD (Karczewski et al. 2020) or ExAC non-psychiatric (Lek et al. 2016) databases, and below an in-cohort MAF of 10%. GCOD users can change these cutoffs where desired. Variants predicted to be damaging by at least one of MetaSVM, FatHMM, and SIFT models were considered in the Strict and Base tiers of variant severity (Liu et al. 2016; Rogers et al. 2017). Variants with a scaled (phred-like) CADD score greater than or equal to 25 were considered in the CADD-based tier of variant severity (Rentzsch et al. 2019). Additional filtering criteria of variant severity is specified in **Supplemental Table 3**. We used gnomAD observed/expected (Karczewski et al. 2020), S_{het} (Zhu, Zhang, and Sha 2018), CADD scores (Rentzsch et al. 2019), and a minimum expression of transcripts per million (tpm) > 0.5 in any cardiac cell type included in the DESCARTES developmental gene expression database (Cao et al. 2020). Cell types considered are listed below.

Cardiac cell types: 'Heart-Cardiomyocytes', 'Heart-CLC_IL5RA positive cells', 'Heart-ELF3_AGBL2 positive cells', 'Heart-Endocardial cells', 'Heart-Epicardial fat cells', 'Heart-Erythroblasts', 'Heart-Lymphatic endothelial cells', 'Heart-Lymphoid cells', 'Heart-Megakaryocytes', 'Heart-Myeloid cells', 'Heart-SATB2_LRRC7 positive cells', 'Heart-Schwann cells', 'Heart-Smooth muscle cells', 'Heart-Stromal cells', 'Heart-Vascular endothelial cells', 'Heart-Visceral neurons.'

Logistic regression

Before simulation analysis, candidate sets are filtered based on their interaction coefficient in a logistic regression model. Disease status is predicted by n+1 variables, where n is the number of genes in an oligogenic candidate set. For all parents and probands (and sequenced siblings where applicable), the presence/absence of a qualifying variant in each individual gene, as well as whether all genes in the combination harbor a variant in that individual, are denoted by 0/1. We use the binomial glm() function in R (version 3.6.1,(R Core Team 2020)) with 50 maximum iterations to test whether the coefficient of the gene interaction is greater than or equal to 1. Candidates with an interaction coefficient < 1 are removed, as this result indicates a spurious combination whose disease association is driven by a single gene or smaller subset of genes within the set. In other words, this step restricts our analysis to sets in which a higher proportion of individuals with damaging mutations in the full oligogenic set have CHD compared to individuals with damaging mutations in each gene separately.

Pseudo-sibling genotype generation

Pseudo-sibling genotypes were generated from the two parental alleles not transmitted to the proband as described in (Z. Yu and Deng 2011). In the rare case in which a proband *de novo* mutation occurs at a locus where a parent carries a qualifying variant, GCOD encodes the pseudo-sibling to inherit the parental rare allele if the proband did not. *De novo* variants in pseudo-siblings were randomly assigned to genes based on the previously-derived protein-coding non-synonymous ('prot') mutability (Samocha et al. 2014). Since all such variants are automatically included at the strictest variant tier, GCOD does not predict specific amino acid substitutions to further qualify the *de novo* variant's CADD, s_{het}, or predicted-damaging status.

Single-gene rare variant transmission test and multi-locus generalization

To identify individual genes in which rare variants are transmitted to probands more often than expected by chance, we used a rare variant aggregation extension of the Combined Multivariate and Collapsing (CMC) transmission disequilibrium test (B. Li and Leal 2008; He et al. 2014). Briefly, for each gene in the dataset containing rare variants of interest, the number of trios, *b*, in which the minor allele was transmitted by a parent heterozygous for a variant is summed across the dataset, as is the number of trios, *c*, in which the major allele was transmitted in this scenario. When parents in a trio carry more than one heterozygous variant, the fraction of major/minor transmissions is added to the respective variable (e.g. if the parents collectively carried 4 heterozygous loci of interest and the proband inherited the minor allele at 3 of the loci, then the trio contributes 0.75 to *b* and 0.25 to *c*). The null hypothesis holds when the proportions $b/(b + c)$ and $c/(b + c)$ are comparable with probabilities 0.5 and 0.5 (i.e., $b = c$). The hypothesis is tested using a 1-degree of freedom asymptotical χ^2 (McNemar's test), with the Edwards correction for continuity. We ran this single-gene transmission test on variants at each variant severity cutoff for the full cohort of 3377 CHD trios. However, with Benjamini-Hochberg FDR correction, no genes were significantly over-transmitted; in **Table 1**, we therefore report information for genes at $p \leq 0.01$ without multiple hypothesis correction.

We also compared GCOD's performance to that of a multi-locus rare-variant generalization of the Transmission Disequilibrium Test (mTDT). Due to the nature of our variant dataset (i.e. rare variants with MAF < 0.05), we aggregated variants across the coding region of a gene using the Burden of Rare Variants (BRV) method as done previously (B. Li and Leal 2008; He et al. 2014). We further generalized this framework to test whether the joint transmission of minor alleles at multiple loci (rather than a single gene locus) occurs more often than expected, using a 2x2 table to tally all possible informative transmission events. In an informative event, a trio's parent genomes collectively carry at least one variant of interest in each gene in a tested pair, and/or a proband carries a *de novo* mutation in one of the genes (this is considered a minor allele transmission per (He et al. 2014)). Based on the proband's genotype, a count of 1 is added to the appropriate scenario in a 2x2 contingency table, where *A* will ultimately indicate the number of times a minor allele in Gene1 was jointly transmitted with a minor allele in Gene2 to the proband. We use a χ^2-test with one degree of freedom to determine whether *A* is significantly different from the expected value of *N*/4, where *N* is the total number of informative events for a given gene pair. We used the Benjamini-Hochberg method to correct for multiple hypotheses.

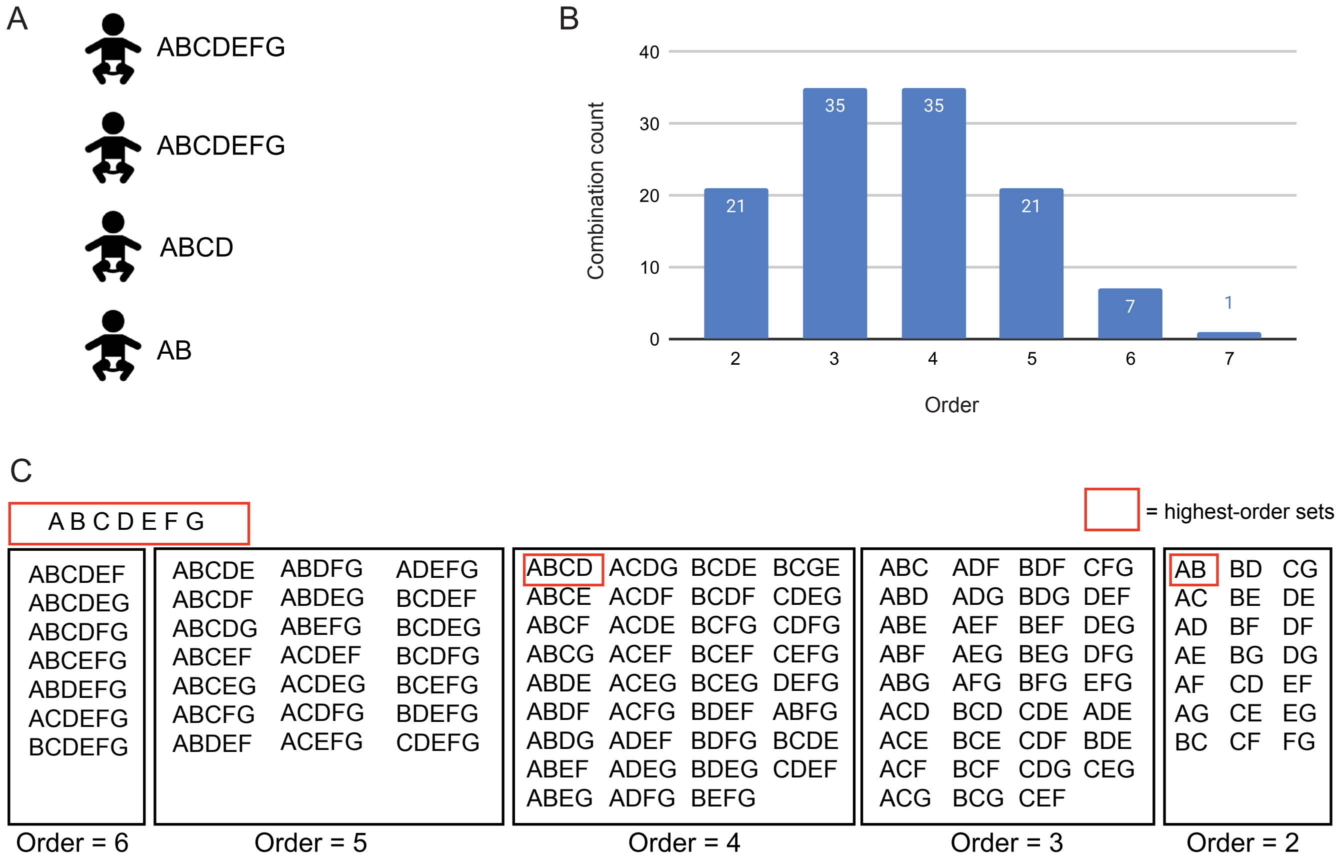
Oligogenic network discovery and depiction

For **Figure 4B**, we identified two oligogenic sets in which genes are known to physically interact in a canonical protein complex. We selected all other significant oligogenic sets containing at least one gene in these complexes, and visualized them using genes for nodes and co-occurrence as edges. For **Figure 4D**, we sought an oligogenic set with several counts of oligogenic transmissions but rarely any variant combinations seen in unaffected parents, discovering the *MYO18B-SACS* combination transmitted oligogenically 10 out of 12 times. We additionally incorporated genes appearing in significant oligogenic sets in any of these 12 probands, reporting co-occurrence counts across the entire PCGC dataset.

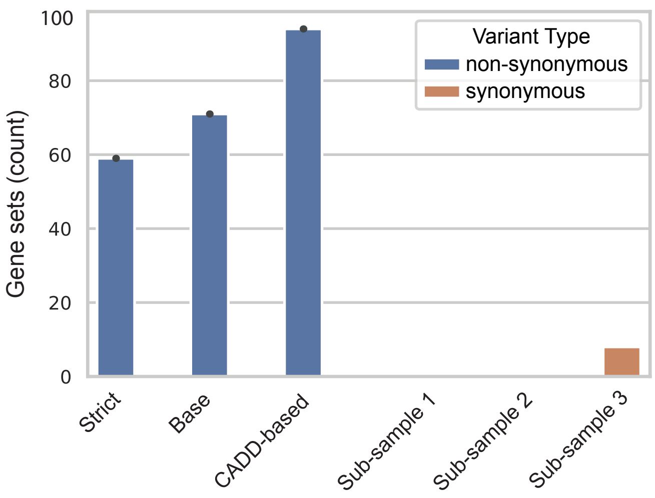
Genes are visualized as nodes and colored according to functional annotations. Edges indicate that the connected genes appear together in a significantly over-transmitted oligogenic gene set, with edge width representing the number of probands with co-occurring mutations in that gene pair. Note that the edge counts can include probands in which the gene pair was inherited from one parent (non-oligogenic transmission), under the condition that a significantly over-transmitted higher-order gene set includes that transmission. Edges were not drawn between nodes unless the two genes appeared together in at least one significant oligogenic set. Oligogenic sets are reported in **Supplemental Table 4** (full cohort pairs), **Supplemental Table 5** (highest order sets from the full cohort), and **Supplemental Table 6** (sets over-transmitted in sub-diagnoses of CHD)**.**

Obtaining unconditional p-values

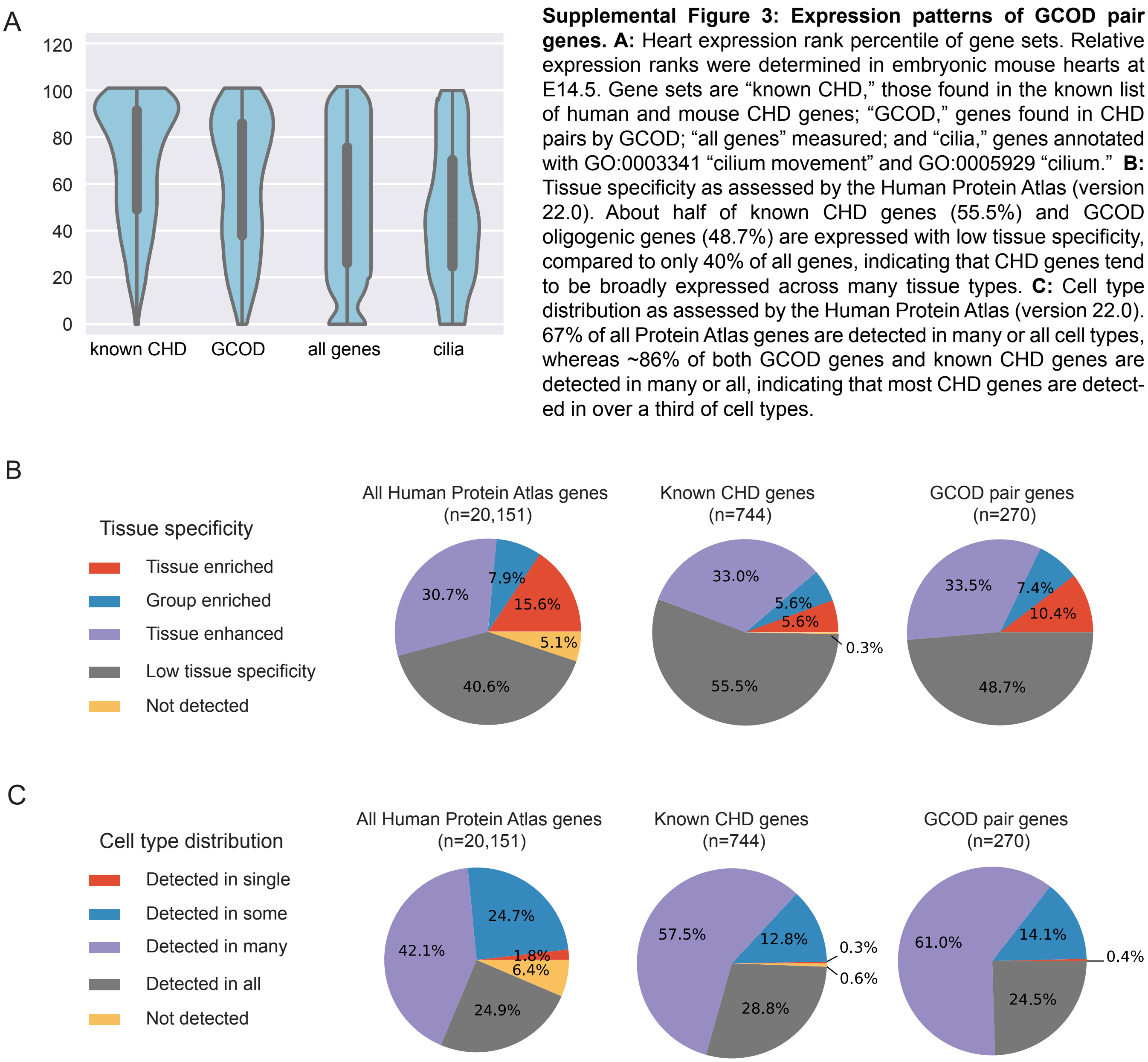
While our marginal p-values are accurate and valid for prioritizing gene sets, this pre-filtering does affect the interpretation of multiple testing corrected p-values (FDR or Bonferroni), which are the error rate conditional on only testing pairs observed at least twice. To obtain unconditional adjusted p-values, users can remove this filter or they can enumerate all the untested sets and assign a large marginal p-value (e.g., p=1) to them before including them in the multiple testing correction procedure. The number of simulations will need to be increased in order to estimate p-values with sufficient significant digits for distinguishing significant from non-significant adjusted p-values in the context of these larger sets of tests.


A
ABCDEFG
ABCDEFG
ABCD
AB
B
Combination count
40
30
20
10
0
21
35
35
21
7
1
2
3
4
5
6
7
Order
C
A B C D E F G
= highest-order sets
ABCDEF ABCDEG ABCDFG ABCEFG ABDEFG ACDEFG BCDEFG
ABCDE ABCDF ABCDG ABCEF ABCEG ABCFG ABDEF
ABDFG ABDEG ABEFG ACDEF ACDEG ACDFG ACEFG
ADEFG BCDEF BCDEG BCDFG BCEFG BDEFG CDEFG
ABCD ABCE ABCF ABCG ABDE ABDF ABDG ABEF ABEG
ACDG ACDF ACDE ACEF ACEG ACFG ADEF ADEG ADFG
BCDE BCDF BCFG BCEF BCEG BDEF BDFG BDEG BEFG
BCGE CDEG CDFG CEFG DEFG ABFG BCDE CDEF
ABC ABD ABE ABF ABG ACD ACE ACF ACG
ADF ADG AEF AEG AFG BCD BCE BCF BCG
BDF BDG BEF BEG BFG CDE CDF CDG CEF
CFG DEF DEG DFG EFG ADE BDE CEG
AB AC AD AE AF AG BC
BD BE BF BG CD CE CF
CG DE DF DG EF EG FG
Order = 6
Order = 5
Order = 4
Order = 3
Order = 2

Supplemental Figure 1: Calculation of highest-order gene sets. A: Genes with observed variants across four probands. In this theoretical cohort, two individuals share seven genes, one proband carries mutations in four of those seven genes, and the final proband carries an oligogenic transmission of the AB digenic pair. **B:** Number of unique combinations for seven genes. Combination counts are visualized by order, i.e. the number of genes comprising the set. **C:** All possible unique gene combinations for gene set ABCDEFG, with the three "highest-order" sets theoretically tested by GCOD highlighted in red.

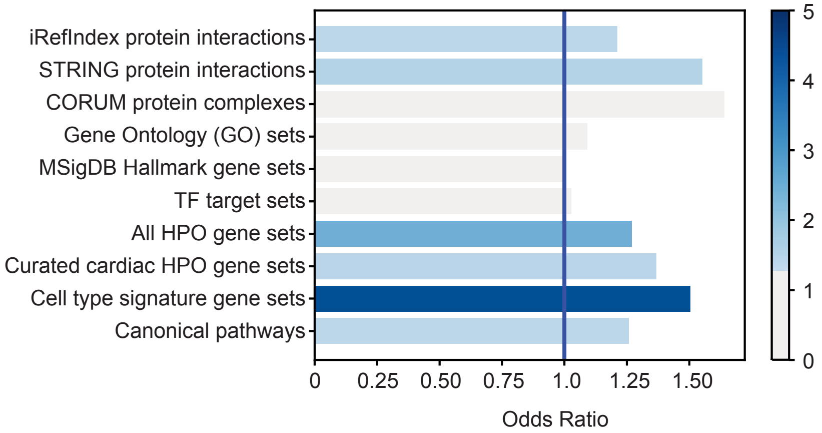


Supplemental Figure 2: Counts of oligogenic pairs for proband non-synonymous and synonymous mutations. Due to the vastly higher count of rare synonymous variants compared to rare damaging non-synonymous variants, GCOD was run on subsets of synonymous variants (Sub-samples 1-3). These subsets each comprise a random set of 200,000 proband synonymous variants that occur within the list of genes qualifying for strict analysis (see severity criteria in Supplemental Table 1). These sub-samples are of similar size to the rare non-synonymous variants analyzed (178859 variants). Three replicates were created using random selection with replacement. Two sub-samples found no significant pairs.

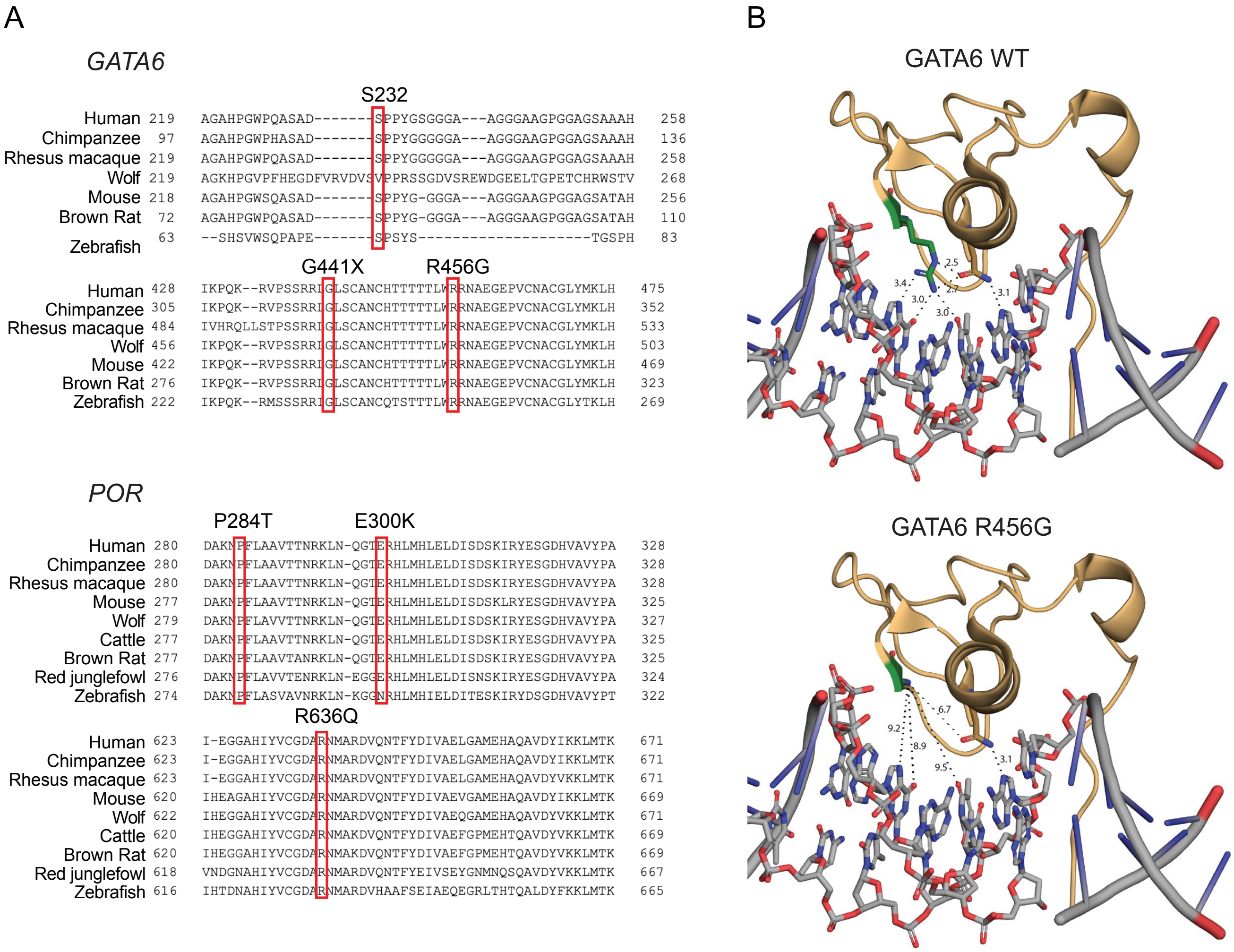


Supplemental Figure 3: Expression patterns of GCOD pair genes. A: Heart expression rank percentile of gene sets. Relative expression ranks were determined in embryonic mouse hearts at E14.5. Gene sets are “known CHD,” those found in the known list of human and mouse CHD genes; “GCOD,” genes found in CHD pairs by GCOD; “all genes” measured; and “cilia,” genes annotated with GO:0003341 “cilium movement” and GO:0005929 “cilium.” **B:** Tissue specificity as assessed by the Human Protein Atlas (version 22.0). About half of known CHD genes (55.5%) and GCOD oligogenic genes (48.7%) are expressed with low tissue specificity, compared to only 40% of all genes, indicating that CHD genes tend to be broadly expressed across many tissue types. **C:** Cell type distribution as assessed by the Human Protein Atlas (version 22.0). 67% of all Protein Atlas genes are detected in many or all cell types, whereas ~86% of both GCOD genes and known CHD genes are detected in many or all, indicating that most CHD genes are detected in over a third of cell types.

Supplemental Figure 4: Canonical gene set enrichment (extended). The length of the bars indicates the odds that at least two genes in an oligogenic set occurred in CHD probands (as opposed to pseudo-siblings) and co-occur in a canonical gene set, pathway, or protein-protein interaction. Color indicates the inverse log p-value of the odds ratio (Fisher exact test), where categories with p-values greater than 0.05 are shown in grey. TF = Transcription Factor, HPO = Human Phenotype Ontology. While co-occurrence of oligogenic set genes in CORUM protein complexes show a relatively high odds ratio suggesting a potential association, oligogenic sets are clearly not enriched for co-occurrence in MSigDB Hallmark gene sets, and show only weak tendency to co-occur in the full list of GO sets and TF target sets.



Supplemental Figure 5: Conservation and structural impact of disease-associated mutations in *GATA6* and *POR*. A: Multiple sequence alignments of GATA6 and POR proteins across vertebrate species highlight the evolutionary conservation of amino acid residues affected by disease-associated variants (S232fs, G441X, R456G in GATA6; P284T, E300K, R636Q in POR), marked by red boxes. High conservation of these residues suggests functional importance. **B:** Structural modeling of the GATA6 DNA-binding domain in the wild-type (WT, top) and R456G mutant (bottom) forms. In the WT model, Arg456 forms multiple hydrogen bonds with the DNA backbone (dashed lines), which are disrupted or lost in the Arg456Gly variant, potentially impairing DNA-binding affinity and transcriptional function.

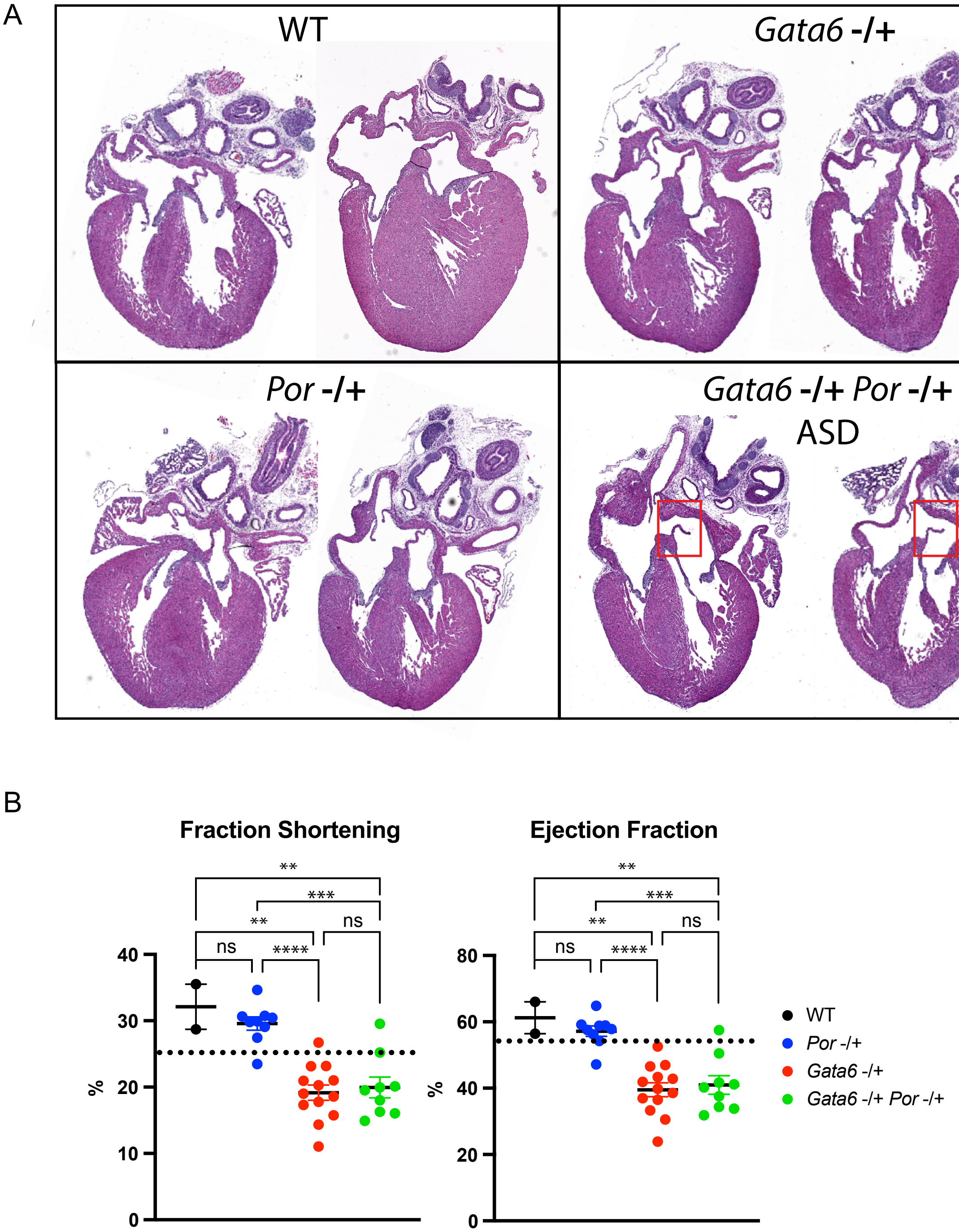


```
Human 219            AGAHPGWPQASAD-------SPPYGSGGGA---AGGGAAGPGGAGSAAAH    258
Chimpanzee 97        AGAHPGWPHASAD-------SPPYGGGGGA---AGGGAAGPGGAGSAAAH    136
Rhesus macaque 219   AGAHPGWPQASAD-------SPPYGGGGGA---AGGGAAGPGGAGSAAAH    258
Wolf 219             AGKHPGVPFHEGDFVRVDVSVPPRSSGDVSREWDGEELTGPETCHRWSTV    268
Mouse 218            AGAHPGWSQASAD-------SPPYG-GGGA---AGGGAAGPGGAGSATAH    256
Brown Rat 72         AGAHPGWPQASAD-------SPPYG-GGGA---AGGGAAGPGGAGSATAH    110
Zebrafish 63         --SHSVWSQPAPE-------SPSYS--------------------TGSPH    83

Human 428            IKPQK--RVPSSRRIGLSCANCHTTTTTLWRRNAEGEPVCNACGLYMKLH    475
Chimpanzee 305       IKPQK--RVPSSRRIGLSCANCHTTTTTLWRRNAEGEPVCNACGLYMKLH    352
Rhesus macaque 484   IVHRQLLSTPSSRRIGLSCANCHTTTTTLWRRNAEGEPVCNACGLYMKLH    533
Wolf 456             IKPQK--RVPSSRRIGLSCANCHTTTTTLWRRNAEGEPVCNACGLYMKLH    503
Mouse 422            IKPQK--RVPSSRRIGLSCANCHTTTTTLWRRNAEGEPVCNACGLYMKLH    469
Brown Rat 276        IKPQK--RVPSSRRIGLSCANCHTTTTTLWRRNAEGEPVCNACGLYMKLH    323
Zebrafish 222        IKPQK--RMSSSRRIGLSCANCQTSTTTLWRRNAEGEPVCNACGLYTKLH    269

Human 280            DAKNPFLAAVTTNRKLN-QGTERHLMHLELDISDSKIRYESGDHVAVYPA    328
Chimpanzee 280       DAKNPFLAAVTTNRKLN-QGTERHLMHLELDISDSKIRYESGDHVAVYPA    328
Rhesus macaque 280   DAKNPFLAAVTTNRKLN-QGTERHLMHLELDISDSKIRYESGDHVAVYPA    328
Mouse 277            DAKNPFLAAVTTNRKLN-QGTERHLMHLELDISDSKLRYESGDHVAVYPA    325
Wolf 279             DAKNPFLAVVTTNRKLN-QGTERHLMHLELDISDSKIRYESGDHVAVYPA    327
Cattle 277           DAKNPFLAAVTTNRKLN-QGTERHLMHLELDISDSKIRYESGDHVAVYPA    325
Brown Rat 277        DAKNPFLAAVTANRKLN-QGTERHLMHLELDISDSKIRYESGDHVAVYPA    325
Red junglefowl 276   DAKNPFLAVVTENRKLN-EGGERHLMHLELDISNSKIRYESGDHVAVYPA    324
Zebrafish 274        DAKNPFLASVAVNRKLN-KGGNRHLMHIELDITESKIRYDSGDHVAVYPT    322

Human 623            I-EGGAHIYVCGDARNMARDVQNTFYDIVAELGAMEHAQAVDYIKKLMTK    671
Chimpanzee 623       I-EGGAHIYVCGDARNMARDVQNTFYDIVAELGAMEHAQAVDYIKKLMTK    671
Rhesus macaque 623   I-EGGAHIYVCGDARNMARDVQNTFYDIVAELGAMEHAQAVDYVKKLMTK    671
Mouse 620            IHEAGAHIYVCGDARNMARDVQNTFYDIVAEVGAMEHAQAVDYIKKLMTK    669
Wolf 622             IHEGGAHIYVCGDARNMARDVQNTFYDIVAEQGAMEHAQAVDYVKKLMTK    671
Cattle 620           IHEGGAHIYVCGDARNMAKDVQNTFYDIVAEFGPMEHTQAVDYVKKLMTK    669
Brown Rat 620        IHEGGAHIYVCGDARNMAKDVQNTFYDIVAEFGPMEHTQAVDYVKKLMTK    669
Red junglefowl 618   VNDGNAHIYVCGDARNMARDVQNTFYEIVSEYGNMNQSQAVDYVKKLMTK    667
Zebrafish 616        IHTDNAHIYVCGDARNMARDVHAAFSEIAEQEGRLTHTQALDYFKKLMTK    665
```



Supplemental Figure 6: Cardiac structural abnormalities and impaired cardiac function in *Gata6* and *Por* mutant mice. A: Histological analysis of cardiac morphology in neonatal hearts. Representative hematoxylin and eosin (H&E)-stained heart sections from postnatal day 0 (P0) mice illustrate the cardiac morphology in littermate controls (WT), *Gata6*−/+, *Por*−/+, and *Gata6*−/+;*Por*−/+ compound heterozygous animals. While littermate controls and single heterozygous *Gata6*−/+ and *Por*−/+ mice display normal cardiac structure, compound heterozygous *Gata6*−/+;*Por*−/+ mice exhibit atrial septal defects (ASDs, red boxes), indicating a synergistic effect of combined haploinsufficiency on cardiac development. **B:** Quantitative echocardiographic analysis in adult mutant mice. Both fraction shortening (FS) and ejection fraction (EF) are significantly reduced in *Gata6*−/+ and *Gata6*−/+;*Por*−/+ mice compared to littermate controls (WT) and *Por*−/+ mice, indicating compromised cardiac contractile function associated with *Gata6* but not *Por* haploinsufficiency. Data are shown as mean ± SEM; p-values determined by one-way ANOVA with multiple comparisons: p<0.05 (*), p<0.01(**), p<0.001 (***), p<0.0001 (****); ns, not significant.