

Supplementary Information for

Joint, multifaceted genomic analysis enables diagnosis of diverse, ultra-rare monogenic presentations

Shilpa Nadimpalli Kobren^{1,*}, Mikhail A. Moldovan^{1,*}, Rebecca Reimers^{2,3}, Daniel Traviglia¹, Xinyun Li^{4,5}, Danielle Barnum^{6,7}, Alexander Veit¹, Rosario I. Corona⁸, George de V. Carvalho Neto⁸, Julian Willett⁹, Michele Berselli¹, William Ronchetti¹, Stanley F. Nelson⁸, Julian A. Martinez-Agosto⁸, Richard Sherwood⁷, Joel Krier¹⁰, Isaac S. Kohane¹, Undiagnosed Diseases Network, Shamil R. Sunyaev^{1,†}

*Indicates equal contribution

†Email: shamil_sunyaev@hms.harvard.edu

1 Department of Biomedical Informatics, Harvard Medical School, Boston, MA

2 Scripps Research Translational Institute, La Jolla, CA

3 Division of Pediatric Genetics, University of California, San Diego, CA

4 Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA

5 Interdepartmental Neuroscience Program, Yale University, New Haven, CT

6 Vrije Universiteit Amsterdam, Medical School of V, De Boelelaan 1105, 1081 HV, Amsterdam, Netherlands

7 Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA

8 Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA

9 Department of Pathology and Laboratory Medicine, NewYork-Presbyterian Weill Cornell Medical Center, New York, NY

10 Department of Genetics, Atrius Health, Boston, MA

Supplementary Figures..... 3

Supplementary Figure 1. UDN cohort pairwise relatedness and sex quality control checks.....3

Supplementary Figure 2. Family structure of sequenced individuals..... 4

Supplementary Figure 3. Hierarchical decision trees for clinically assessing candidates. 5

Supplementary Figure 4. NovoCaller selects for confident de novo variants.....6

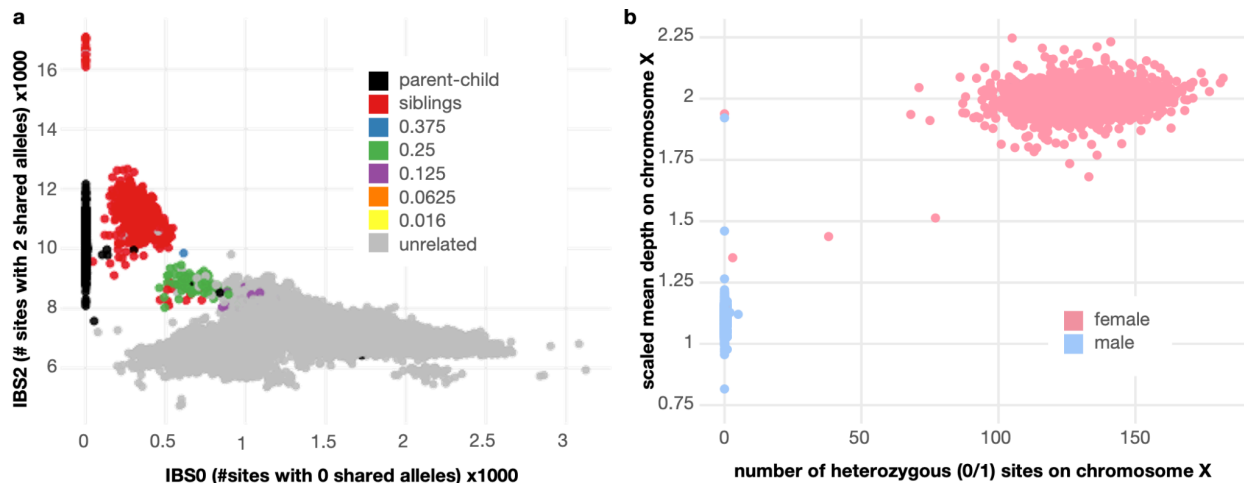
Supplementary Figure 5. QQ-plots for closed-form de novo and compound heterozygous statistics.....7

Supplementary Figure 6. Genes recurrently hit with exonic variants highlight H4C5..... 8

Supplementary Figure 7. Massively Parallel Splicing Reporter Assay (MPSA) design..... 9

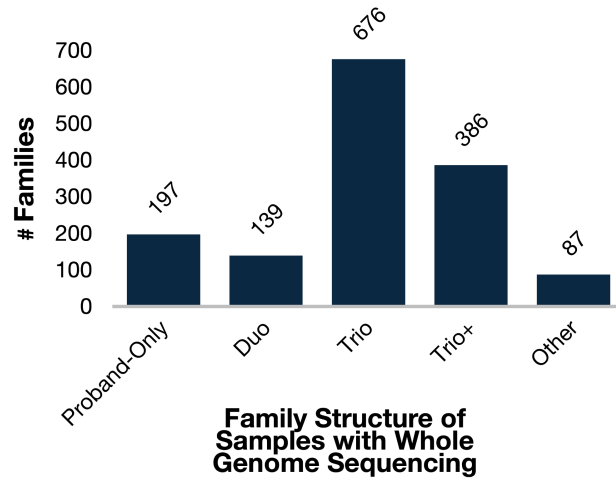
Supplementary Figure 8. MPSA validation rates by SpliceAI scores and across replicates..	10
Supplementary Figure 9. DeNovoWEST modifications and variant distributions used to calculate weights.....	11
Supplementary Figure 10. DeNovoWEST results are enriched for correct diagnoses.....	12
Supplementary Figure 11. Schematic of cohort-level compound heterozygous variant recurrence.....	13
Supplementary Figure 12. Schematic of “goodness-of-fit” test for individual-level compound heterozygous variants.....	14
Supplementary Figure 13. Enrichment of correct diagnoses in RaMeDiES-IND’s patient–gene ranking.....	15
Supplementary Figure 14. MED11 transcriptome sequencing.....	16
Supplementary Figure 15. Correlation between CADD and SpliceAI scores in intronic regions.....	16
Supplementary Figure 16. Comparison of DeNovoWEST and RaMeDiES-DN (CADD only)..	17
Supplementary Figure 17. UDN patients do not fall into well-defined phenotypic clusters.....	18
Supplementary Note 1. Subdividing neurological symptom category.....	19
Supplementary Note 2. Clinical evaluation protocol.....	19
Preparation: Collecting information prior to evaluating diagnostic fit of candidate.....	19
Information collected about the affected patient:.....	19
Information collected about each variant:.....	20
Tip for planning for literature review.....	20
Evaluation of variant–patient phenotypic fit by category.....	20
Known disease associations.....	21
Model organism studies.....	22
Expression of gene in phenotypically-relevant tissues.....	22
Protein molecular function.....	23
Gene evolutionary constraint.....	23
Variant frequency, pathogenicity, and position.....	24
Summarizing information for final phenotypic fit score.....	25
Supplementary Note 3. Compound heterozygous mutational target is uniformly distributed.....	26
Supplementary Note 4. Biallelic variant independence is violated in consanguineous cases.....	27
References.....	28

Supplementary Figures



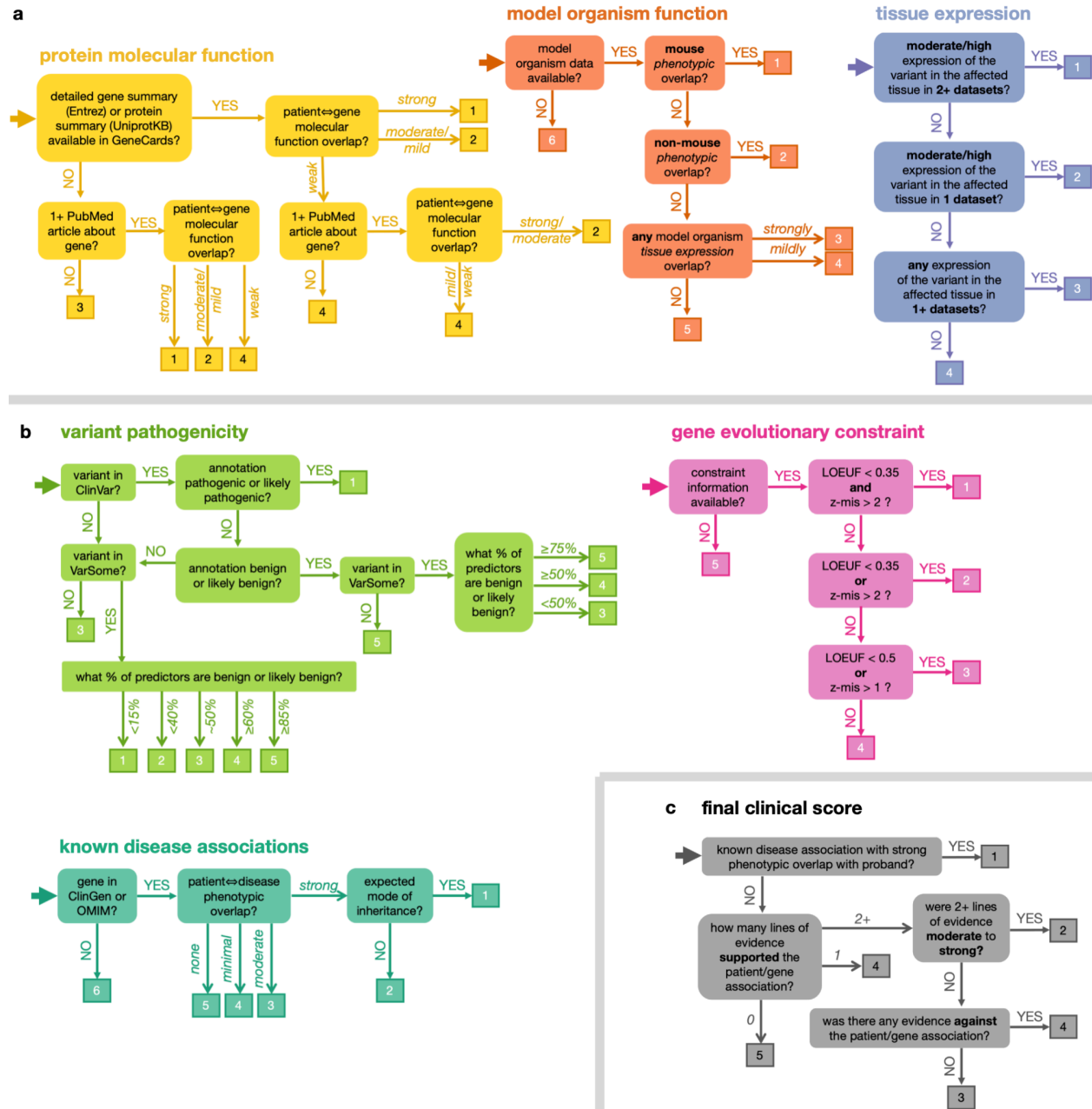
Supplementary Figure 1. UDN cohort pairwise relatedness and sex quality control checks.

(a) Between-sample pairwise relatedness computed using Somalier, which operates on a predefined, informative subset of genomic positions.¹ Unrelated individuals are expected to have a nonzero number of genomic positions with no shared alleles (IBS0), whereas parent-child pairs should have at least one shared allele at every genomic site. **(b)** Within-sample metrics for checking the sex of samples; females (XX, pink) are expected to have heterozygous sites on chromosome X and a scaled mean depth (calculated as two times the read depth on chromosome X / mean read depth across the genome) of about 2.



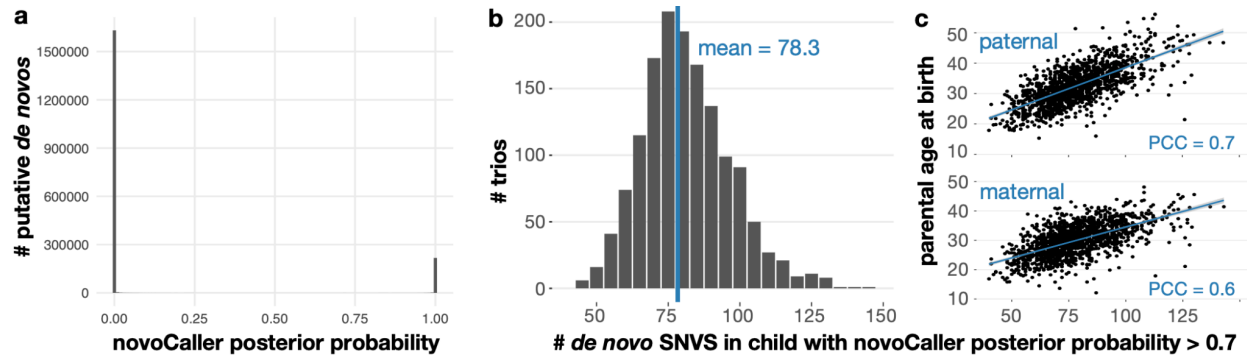
Supplementary Figure 2. Family structure of sequenced individuals.

Duo cases are not necessarily always parent–child relationships; some sibling pairs are recorded as duos. Trios (and trio+) families require at least one child with both parents sequenced. Families with three sequenced individuals who do not share this relationship pattern are included in the “other” category.



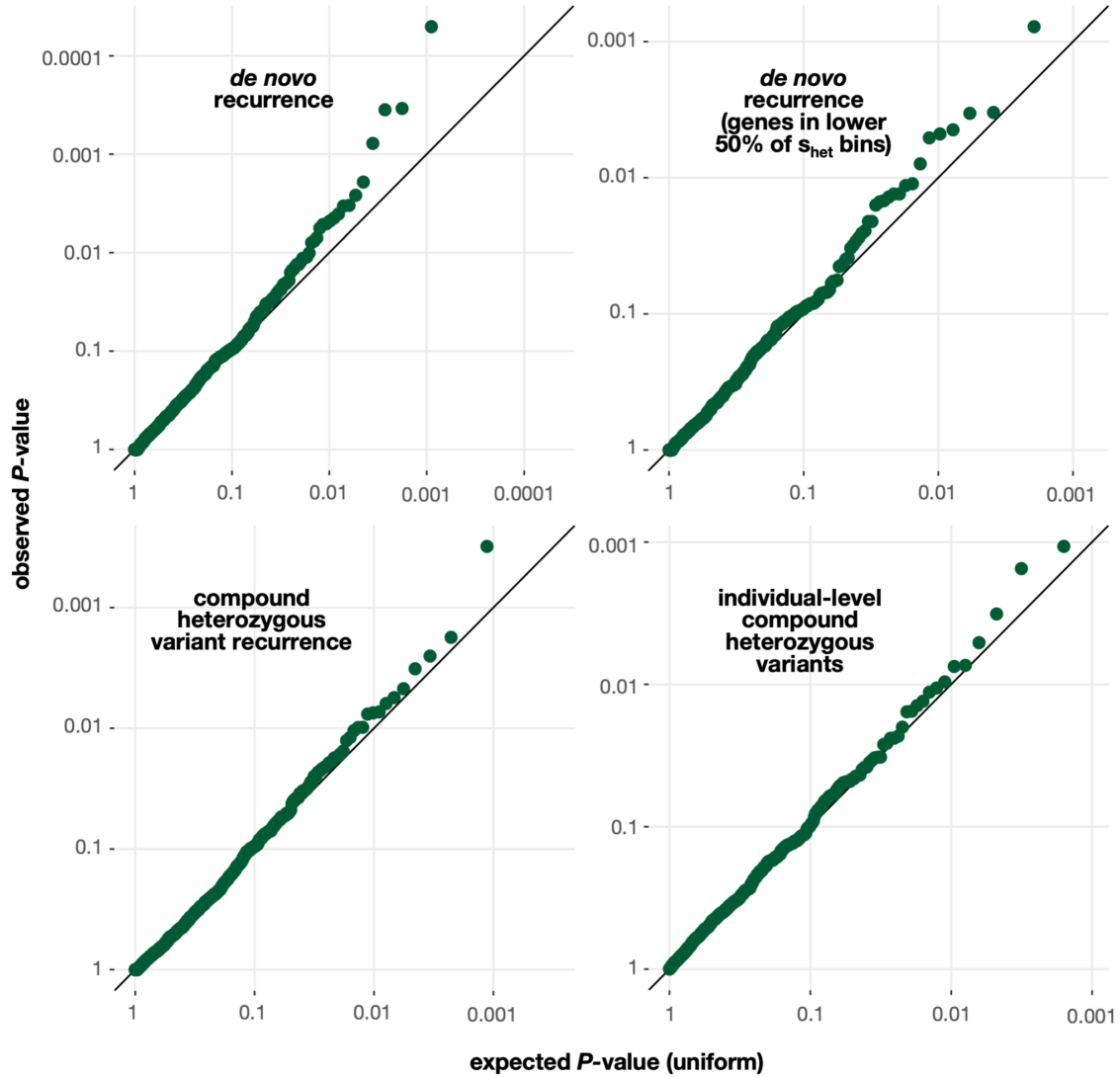
Supplementary Figure 3. Hierarchical decision trees for clinically assessing candidates.

(a) Three categories of evidence used in the clinical evaluation protocol for assessing gene-disease fit for each participant phenotype that were not used in any statistical candidate prioritization methods. **(b)** Three categories of evidence used where clinicians are cognizant of how evidence assessed may be reflective of or correlated with features that were used in the statistical candidate prioritization. **(c)** Final clinical score assigned based on the six categories of evidence.



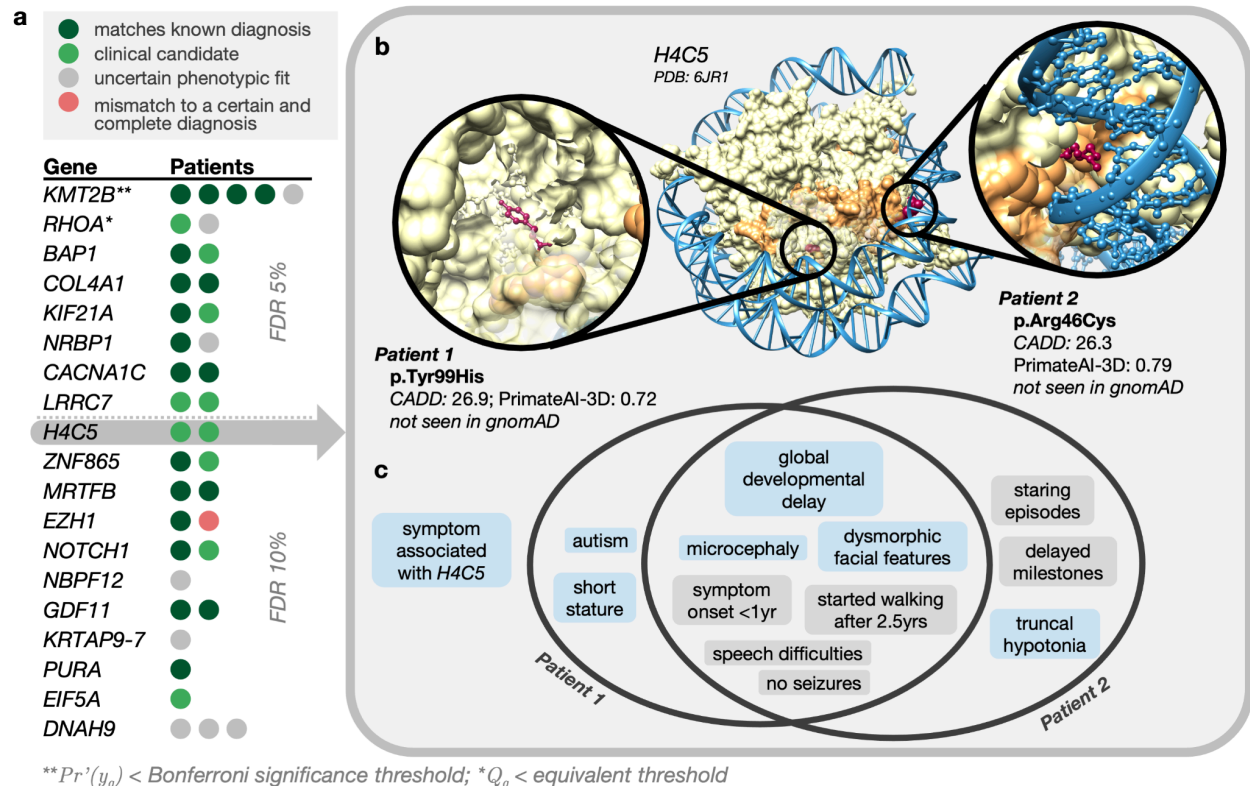
Supplementary Figure 4. NovoCaller selects for confident *de novo* variants.

Putative *de novo* variants were selected in each trio as described in Methods. **(a)** Posterior probabilities assigned by novoCaller are bimodal; 40.4% of putative *de novo*s were assigned a posterior probability of 0 and are excluded from the histogram. **(b)** Counts of *de novo* SNVs per child in each trio uncorrected by parental ages. **(c)** Counts of *de novo* SNVs show expected correlation with paternal (top) and maternal (bottom) age.



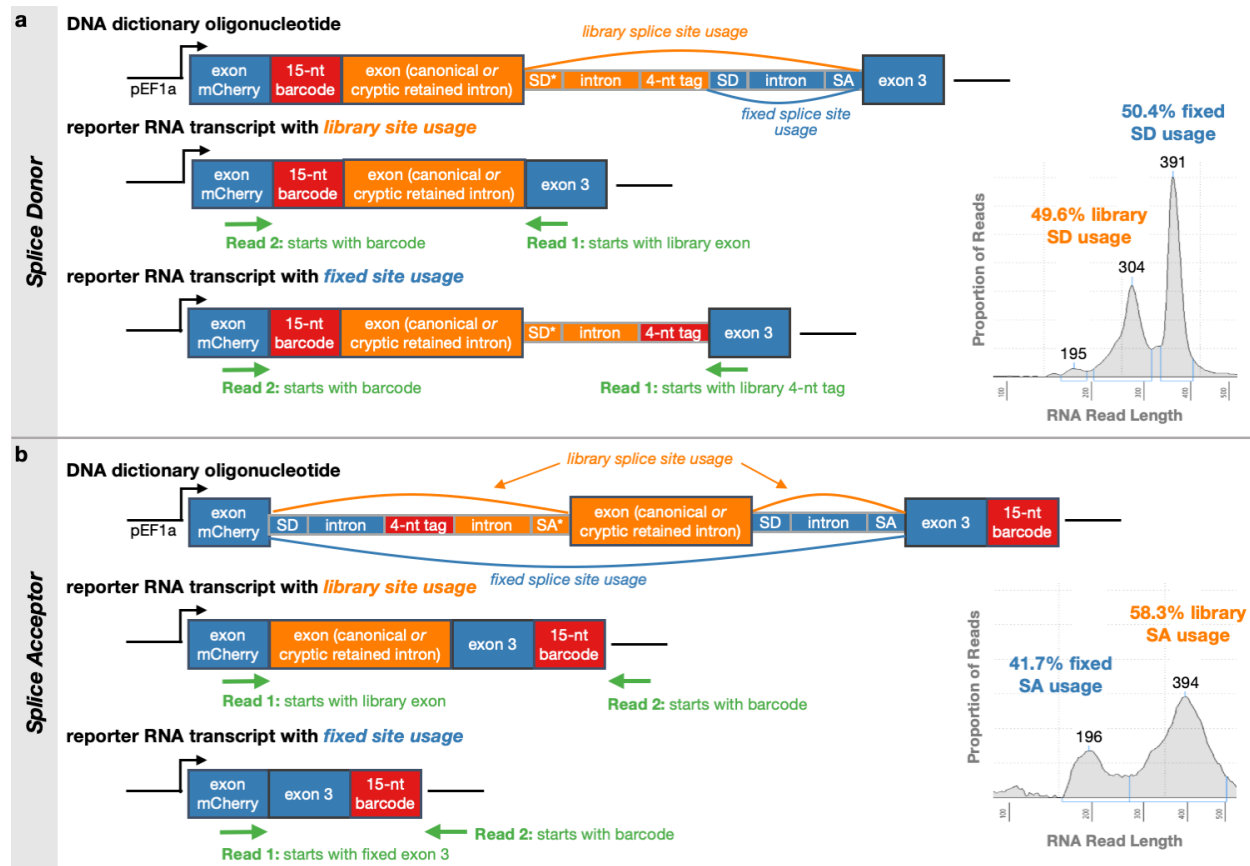
Supplementary Figure 5. QQ-plots for closed-form *de novo* and compound heterozygous statistics.

Top row depicts QQ-plots for *de novo* recurrence. The full curve is shown on the left and the subset of genes falling into the lower half of s_{het} bins are shown on the right. The bottom row depicts QQ-plots for compound heterozygous recurrence (left) and individual-level (right) statistics. Because all of our closed-form statistics y and $\Pr(y|K)$ are not defined for $K=0$ and therefore not expected to be uniformly-distributed (see Methods), the P-values plotted here are $\Pr(y|K \geq 1) = \Pr(y)/\Pr(K \geq 1)$.



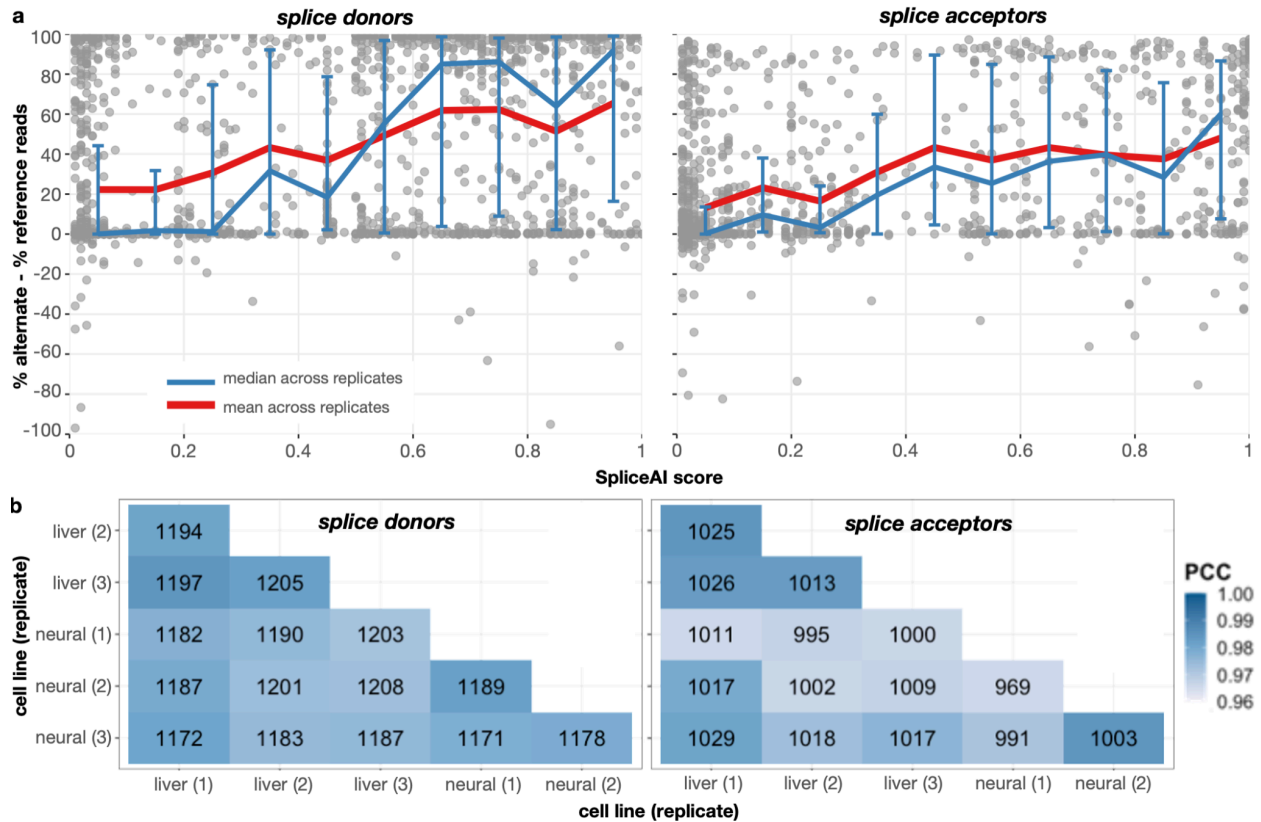
Supplementary Figure 6. Genes recurrently hit with exonic variants highlight H4C5.

(a) Gene findings at FDR 5% and 10% when running RaMeDiES-DN on all exonic variants (including indels and nonsense variants) using four deleteriousness predictors (AlphaMissense, PrimateAI-3D, CADD and REVEL). (b) Human histone protein structure (PDB ID: 6JR1) where chains B and F, sequence matches to the H4C5 gene, are highlighted in orange. Reference alleles for missense *de novo* variants observed across two UDN patients (red) are shown in zoomed-in circles. (c) Overlap of phenotype terms annotated to each patient. Phenotypes associated with histone genes or H4C5 specifically are highlighted in blue, and phenotypes not yet associated with histone genes are highlighted in gray.



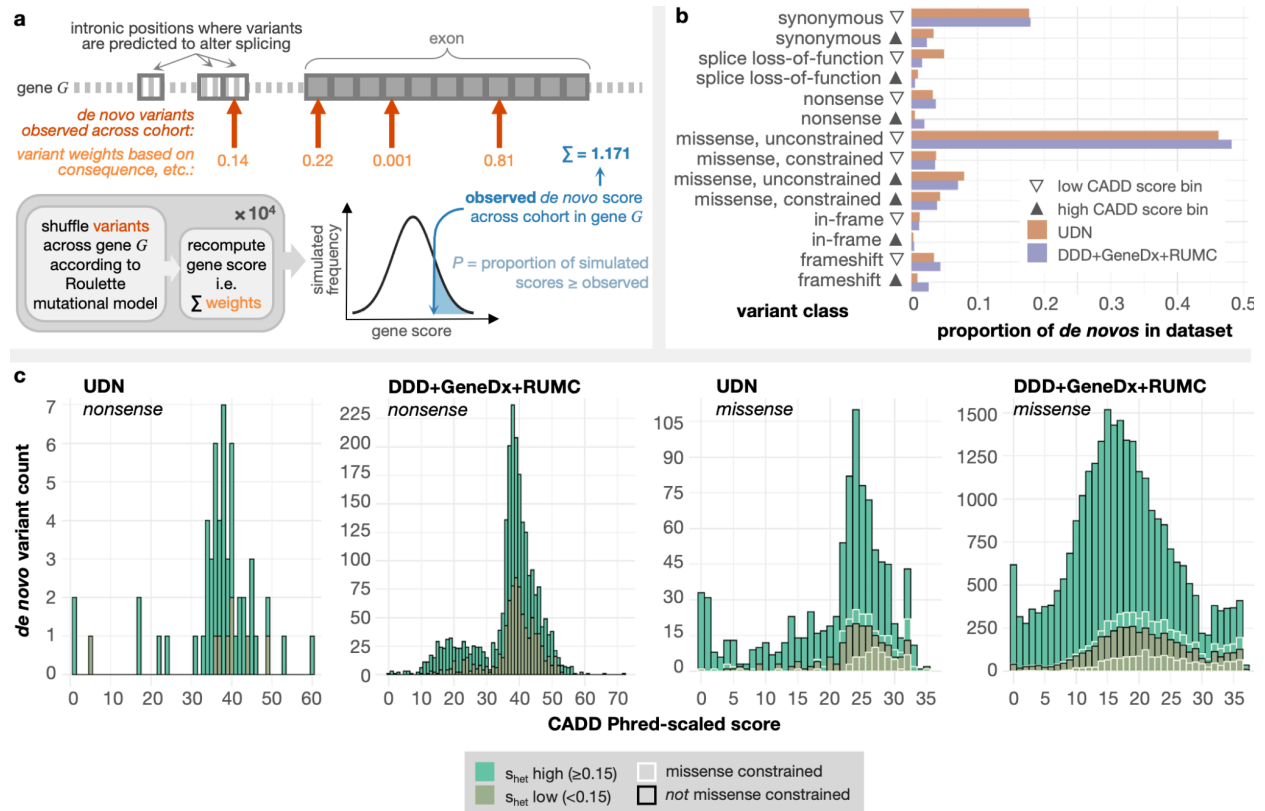
Supplementary Figure 7. Massively Parallel Splicing Reporter Assay (MPSA) design

Oligonucleotides containing a 155-nucleotide section of genomic DNA from a UDN patient, called a 'library' (orange) were inserted into lentiviral plasmids (pEF1a). Exons are represented as tall boxes and introns as short boxes in the DNA depictions. These genomic DNA library segments were uniquely tagged with barcodes detectable in DNA- or RNA-sequencing (red). Libraries contained either canonical splice sites that a variant was predicted to cause the loss of, or cryptic splice sites that a variant was predicted to cause the gain of (denoted with an '*'). Fixed exonic and intronic sequence with fixed splice sites surround the inserted library (blue). Expected paired RNA-sequencing reads are depicted in green. **(a)** Usage of a library splice donor site versus the fixed splice donor site can be detected in reporter RNA transcripts. Inset shows the proportions of RNA reads of different lengths produced through Tapestation, an automated electrophoresis system for sizing and quantifying nucleic acid sequence reads. **(b)** Similarly, usage of a library splice acceptor site versus a fixed splice acceptor site can be detected in reporter RNA transcripts.



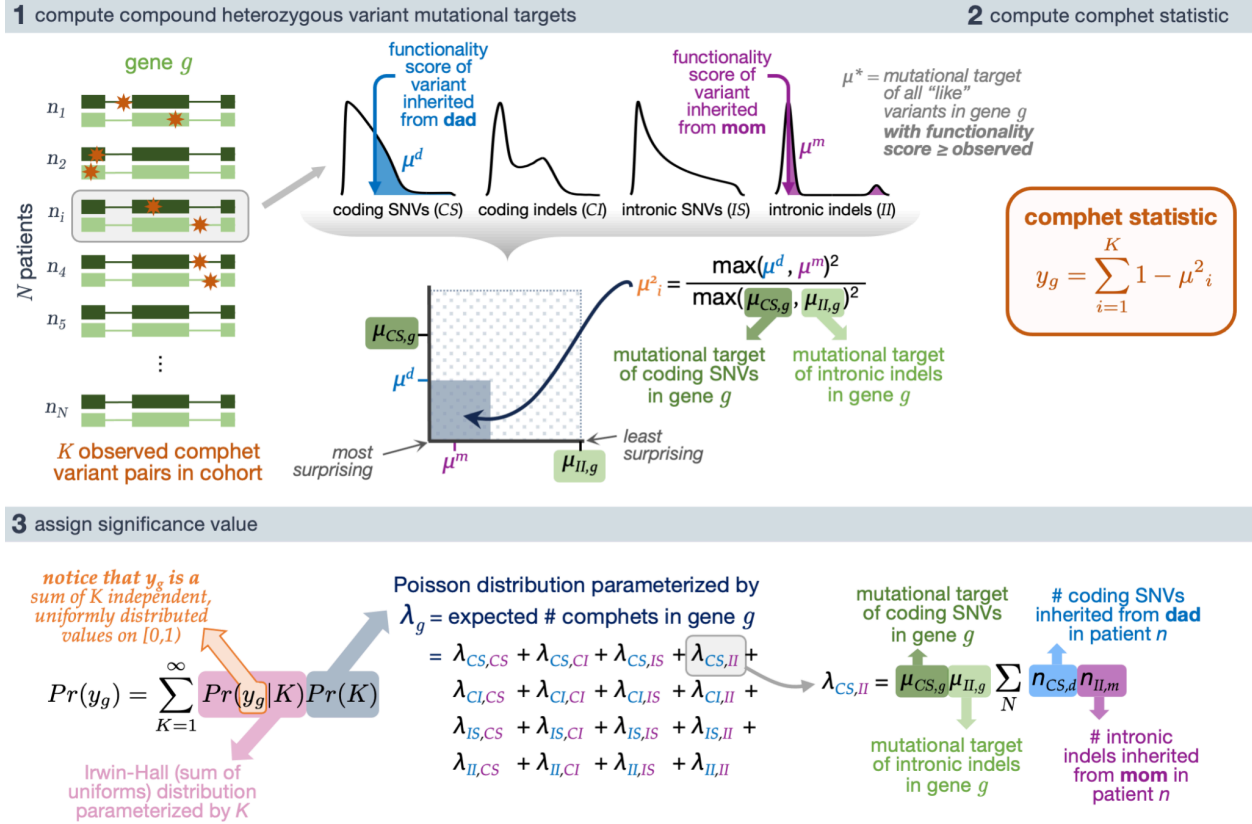
Supplementary Figure 8. MPSA validation rates by SpliceAI scores and across replicates.

A “validation rate” is computed for every variant tested in the MPSA as described in Methods. **(a)** Each point represents the MPSA validation rate for a specific variant that was tested. Blue lines correspond to the median validation rate for all variants within each SpliceAI score decile across biological replicates; error bars depict the interquartile range of validation rates for variants within each decile across replicates. Red lines correspond to the mean validation rate for all variants within each SpliceAI score decile. **(b)** We computed all-against-all Pearson correlations between MPSA validation rates corresponding to the same alternate/reference pairs across experiments. The numbers in the center of each cell correspond to the number of alternate/reference pairs with at least 10 supporting reads each in both experiments being compared. The number in parentheses in the axis labels indicates the biological replicate number. Liver cells are HepG2 and neural-like cells are SK-N-SH.



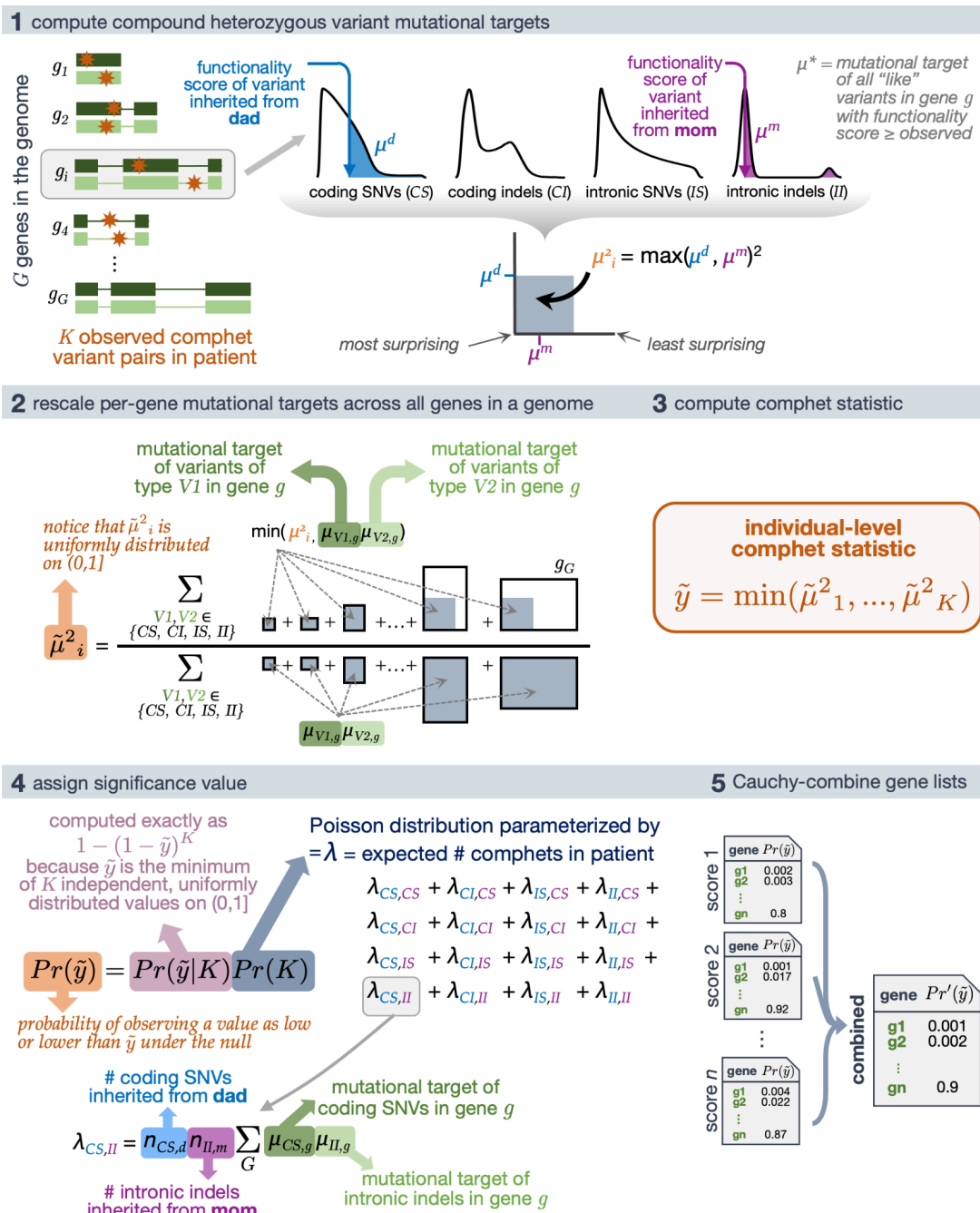
Supplementary Figure 9. DeNovoWEST modifications and variant distributions used to calculate weights.

(a) DeNovoWEST incorporates a weighted enrichment test in its loss-of-function (all variants) and gain-of-function (missense variants only) models. We modified this portion of the software to incorporate deeply-intronic variants with a predicted splice-altering impact and replaced the trinucleotide mutation rate model with Roulette as described in Methods.² Weights assigned to each variant class were recomputed using the variant type distributions across the UDN and existing DDD+GeneDx+RUMC datasets together. **(b)** Distributions of variants across each variant type bin are largely similar between the UDN and the DDD+GeneDx+RUMC datasets. **(c)** In the nonsense and missense variant type categories, CADD score distribution is similar. We note that the full continuous scale of CADD scores is not used in DeNovoWEST, and variants are binned according to their CADD score instead.



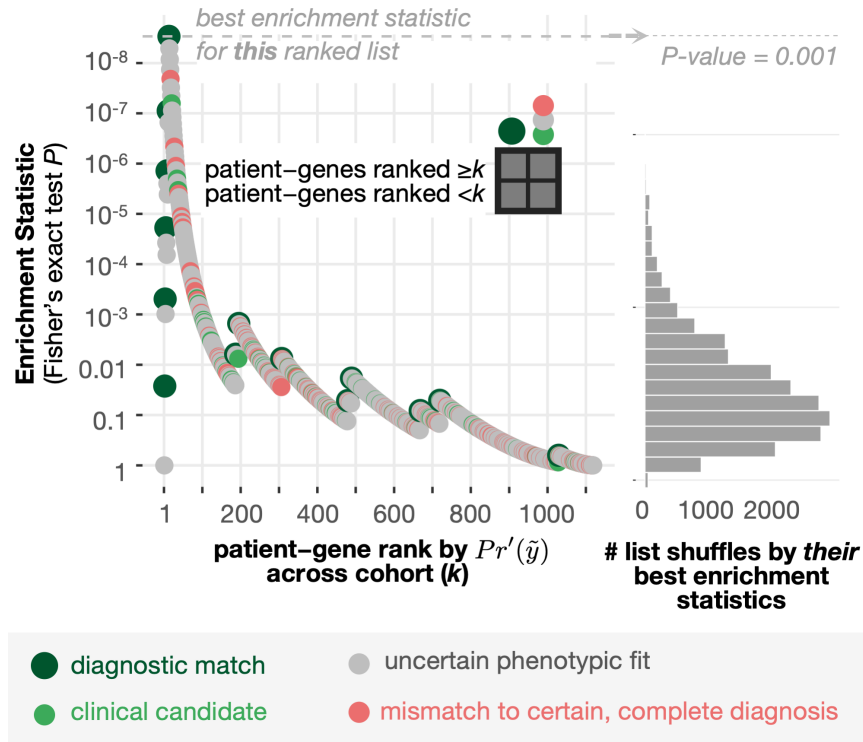
Supplementary Figure 11. Schematic of cohort-level compound heterozygous variant recurrence

Schematic of analytical test for the recurrence of compound heterozygous variant pairs (comphets) that considers distal splice-altering and exonic SNV and indel variants, their variant functionality scores, and a genome-wide mutation rate model Roulette. “Like” variants refer to those of the same variant class (i.e., coding SNVs [CS], coding indels [CI], intronic SNVs [IS], intronic indels [II]) and within the same functionality score and minor allele frequency thresholds. Step 4, Cauchy-combination of significance values achieved using different functionality prediction scores, is not pictured here (Methods).



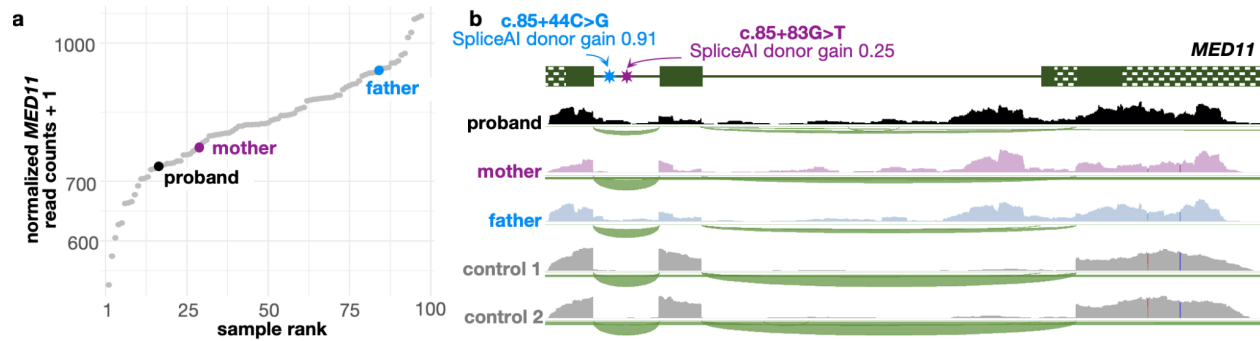
Supplementary Figure 12. Schematic of “goodness-of-fit” test for individual-level compound heterozygous variants

Illustration of compound heterozygous (comphet) variants observed in an individual and simplified graphics depicting the calculation of a comphet mutational target, the rescaling of that mutational target with respect to all G genes in a genome to compute an individual-level statistic, and the significance calculation for that individual-level comphet statistic. “Like” variants refer to those of the same variant class (i.e., coding SNVs [CS], coding indels [CI], intronic SNVs [IS], intronic indels [II]) and within the same functionality score and minor allele frequency thresholds.



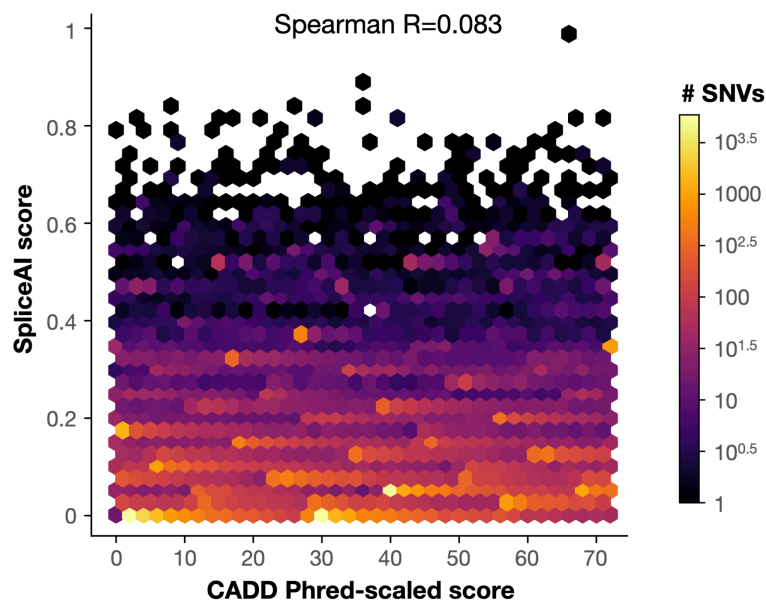
Supplementary Figure 13. Enrichment of correct diagnoses in RaMeDiES-IND's patient-gene ranking

Fisher's exact test P values indicate an enrichment of correct diagnoses (dark green circles) among top-ranked genes in a run of RaMeDiES-IND using AlphaMissense, PrimateAI-3D, CADD, and REVEL (left). The 2x2 grid indicates how patient-gene values were counted at each position k in the ranked list to compute Fisher's exact test P . This enrichment is significant based on the minimum Fisher's exact test P s achieved across 10,000 random shuffles of the gene list (right).



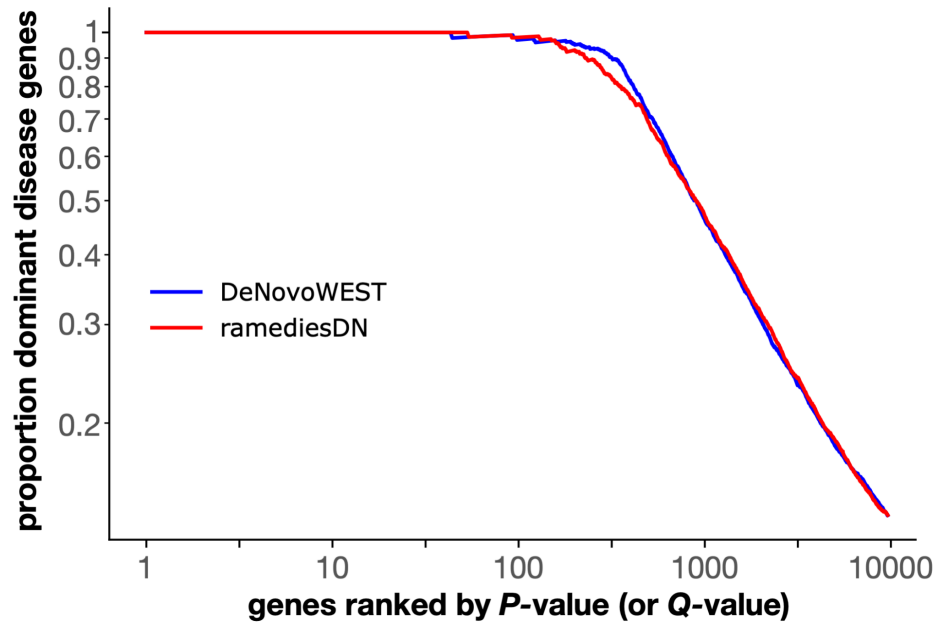
Supplementary Figure 14. MED11 transcriptome sequencing.

(a) Overall gene expression of MED11 in blood samples from affected proband (black), mother (purple) and father (blue) are within the normal range of expression when compared to 50 sequenced control samples (gray) and would not have been flagged as differentially expressed. (b) RNA-sequencing read depth over the full MED11 gene region for proband, mother, father, and two control samples.

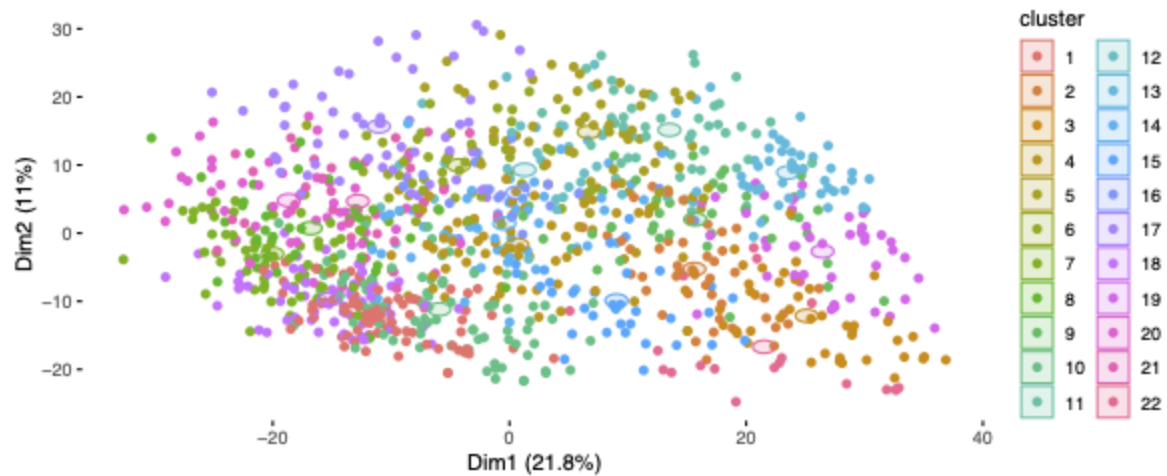


Supplementary Figure 15. Correlation between CADD and SpliceAI scores in intronic regions.

We assessed the correlation between precomputed *in silico* CADD deleteriousness and SpliceAI functionality scores within all intronic regions between protein-coding exons on chromosome 21. We observed poor correlation between these scores, even though CADD uses SpliceAI as an input feature, supporting our distinction between these score types in introns.



Supplementary Figure 16. Comparison of DeNovoWEST and RaMeDiES-DN (CADD only)
 Curves depict the ability of the two *de novo* enrichment methods to recover known autosomal dominant disease genes (reported in OMIM) from the DDD+GeneDx+RUMC dataset. Around 4000 putative *de novos* from the DDD+GeneDx+RUMC were removed because they were on chromosome X, were not in standard VCF format (i.e., non-normalized or left-aligned indels), or could not be mapped to GRCh38. In total, *de novos* from 6,913 probands from DDD, 12,772 from GeneDx, and 1,609 from RUMC were included. Both axes are log-scaled for visual clarity. A version of RaMeDiES-DN using only CADD scores in exonic regions was used here. DeNovoWEST genes are ranked by *P*-value, and RaMeDiES-DN genes are ranked by *Q*-value (Supplementary Data 5).



Supplementary Figure 17. UDN patients do not fall into well-defined phenotypic clusters.

UDN patients were clustered using K-medoids clustering based on their pairwise phenotypic similarity score; 22 clusters were selected for visual clarity. Patients separated by pairwise similarity and colored by cluster membership are visualized with a Uniform Manifold Approximation and Projection (UMAP) to show that patients do not easily separate into equally-sized and well-defined clusters.

Supplementary Notes

Supplementary Note 1. Subdividing neurological symptom category

To identify subgroups within the largest “neurological” primary symptom category, we used the all-against-all pairwise phenotypic similarity scores previously computed (Methods) to cluster the subset of neurological patients into three equivalently-sized groups (using Ward-Linkage hierarchical clustering from the R cluster package, version v.2.1.4). We created a text corpus for each cluster from the concatenated list of patients’ HPO terms within each cluster using the R tm package (version v0.7.11) and visualized these text corpuses as comparative word clouds using the R quanteda package (version v3.3.1). Subgroup names were determined based on manual inspection and comparison of the word clouds.

Supplementary Note 2. Clinical evaluation protocol

In all of the following sections, an asterisk (*) indicates fields that were extracted computationally in our implementation.

Preparation: Collecting information prior to evaluating diagnostic fit of candidate

Compile relevant information in your database of choice; Microsoft Excel spreadsheets and REDCap databases have both been used effectively in following this protocol.

Information collected about the affected patient:

1. Patient ID*
2. Family ID*
3. Clinical site of evaluation*
4. Sequencing lab where genome sequencing was completed*
5. RNA-Seq available? If so: sequence ID and tissue*
6. Known diagnosis available? If so: gene name, diagnosis certainty level, diagnosis completeness (complete or partial)*
7. Other candidate genes available? If so: gene name*
8. Heritage/demography*
9. Age at symptom onset*
10. Age at application to UDN and age at clinical evaluation*
11. Sex*
12. Affected family members? If so: their patient IDs*
13. Human Phenotype Ontology (HPO) codes and terms*

14. Abstracted case review of patient presentation (based on Application Letter, Wrap Up Form/Letter, Photography and Imaging Reports, see Methods)*
15. Clinician-determined primary symptom category*
16. Affected tissue(s) - manually determined, and used to assess relevant tissue-specific expression later

Information collected about each variant:

1. Variant classification (exonic or intronic)*
2. Genome build*
3. Variant location = chromosome, position, reference DNA allele(s), alternate DNA allele(s)*
4. Variant read depth across trio*
5. Variant impact from Ensembl's Variant Effect Predictor*
6. Variant transcript = transcript name and HGVS cDNA impact*
7. Variant protein = protein name and HGVS protein impact*
8. CADD score (exonic and intronic)*
9. SpliceAI score (exonic and intronic)*
10. Confirmed in IGV? (i.e., present on a clinical sequencing report or in a research variant table)

Tip for planning for literature review

Before beginning, note that there are **three** separate fields where PubMed IDs and brief take-home summaries should be recorded; in practice the PubMed search happens *once* and you will record relevant information in the different corresponding fields:

- (1) Articles associating gene to known human diseases (matching patient phenotype or not)[†]
- (2) Articles about protein molecular function[†]
- (3) Articles about experiments (engineered cell lines or animal models) relating gene variants to phenotype changes[‡]

[†]These articles are sometimes curated in OMIM, but thorough review of PubMed is necessary for the most up-to-date evidence review.

[‡]This information is sometimes found in MGI (Mouse Genome Informatics); record PubMed IDs where possible.

Evaluation of variant-patient phenotypic fit by category

External databases are consulted to compile information that will then be used to evaluate the diagnostic fit of the candidate variant(s) for the patient. Here we provide example databases used by our clinical team to execute this protocol, but newer databases not

explicitly listed here that compile the same information can be used as well. There are six categories of information to be collected for review.

Known disease associations

1. Notes from Clinical Genome Resource (ClinGen³) diseases & expert panel, Gene Curation Coalition (GenCC⁴), and Genomics England PanelApp⁵
2. PubMed articles (PMID and brief take-home summary) relating gene to phenotype or gene to patient presentation
 - a. **Note:** keep track of supporting as well as contradictory evidence
3. Manually-written summary of interesting information from the above step 2
4. OMIM disease associations and their IDs*
5. Published reports of disease/gene associations from OMIM⁶ (obtained via manual review of OMIM pages)
6. Associated traits in gene from GWAS catalog⁷ / linkage associations
 - a. Sometimes there are hundreds of traits, so we take the top/most relevant 15-20
7. GeneMatcher (MatchMaker Exchange)
8. Phrank score relating gene to patient's phenotype terms (low scores are unreliable but high scores are interesting)*

<i>Do any known disease associations overlap with proband's key phenotypic features?</i>	
1	ClinGen curated disease association with correct inheritance mode (i.e., AD for <i>de novos</i> or AR for compound heterozygous variants) with strong phenotypic overlap with patient's primary phenotype(s)/pathophysiology
2	Published report or OMIM annotation with correct inheritance mode supporting disease monogenic association with strong phenotypic overlap
3	Any published report/ClinGen curation with mild-moderate phenotypic overlap (regardless of inheritance mode)
4	Published report/OMIM annotation supporting disease polygenic/GWAS association with phenotypic overlap
5	No evidence supporting any gene/disease association
6	Known gene-disease association with minimal phenotypic overlap (regardless of inheritance mode)
7	Known disease association with correct inheritance mode with no phenotypic overlap

Model organism studies

1. Mouse Genome Informatics (MGI)⁸ > search for gene name (and click symbol in results table) > Mutations, Alleles and Phenotypes section > review summary at bottom
2. International Mouse Phenotyping Consortium (IMPC)⁹ can also be consulted, similar to information available in MGI.
3. Review PubMed articles about experimental evidence (engineered cell lines or animal models) relating variants to phenotypes.
4. Models and Rescue score, max 4, (based on information from MGI, IMPC, and PubMed):¹⁰
 - a. +2 = Animal model rescue: Was the variant introduced and did it result in a matching phenotype in the animal, and then was the variant reversed/blocked via a CRISPR/Cas9 or silencing RNA intervention that reversed the phenotype? (1 point for cell culture model or engineered equivalent)
 - b. +1 = Functional alteration in cell model (0.5 engineered, 1 for cell line from affected individual)

<i>Does tissue-level expression in model organisms or null models suggest correlation with proband phenotypes?</i>	
1	Null mouse shows phenotypic overlap with proband
2	Other model organism data demonstrate overlap with proband phenotypic features
3	Expression in other organisms overlaps with key phenotypic features/tissues of interest
4	Expression in other organisms overlaps mildly with key phenotypic features/tissues of interest
5	Expression in other organisms does not overlap with key phenotypic features
6	No model organism data available

Expression of gene in phenotypically-relevant tissues

1. Check expression in the Allen Brain Institute's Brain Map¹¹
2. The Genotype-Tissue Expression (GTEx) Portal¹² – search for gene, scroll down to “Bulk tissue gene expression” to see relevant tissues

<i>Does available tissue-level expression data align with proband's key phenotypic features?</i>	
1	Multiple datasets show tissue expression / phenotypic feature overlap
2	At least one dataset shows tissue expression / phenotypic features overlap
3	Available data shows minimal tissue expression of gene
4	Datasets show expression in tissues other than those related to proband's phenotypic features

Protein molecular function

1. Entrez gene summary (reviewed via GeneCards^{12,13})
2. UniprotKB/SwissProt protein summary (reviewed via GeneCards¹⁴)
3. Published reports of gene/phenotype associations from OMIM^{6,14}

<i>Does the protein function relate to the potential model of molecular pathophysiology?</i>	
1	Well-annotated gene product function with strong association with patient's primary phenotypic feature(s)/pathophysiology
2	Published report/OMIM annotation supporting monogenic disease association with mild phenotypic overlap
3	Minimal or no gene product function annotation
4	Well-annotated gene product function with no obvious association with patient's primary phenotype(s)/pathophysiology

Gene evolutionary constraint

1. Loss-of-function (nonsense, splice acceptor, splice donor) scores from gnomAD*¹⁵
 - a. pLI (ranges 0-1) where 1 means *intolerant* to LoF variants
 - b. Observed/expected ratio (ranges 0-1) and Poisson-derived confidence interval; where 0 means *intolerant* to LoF variants; note that the upper bound of the Poisson confidence interval is referred to as "LOEUF" for "loss-of-function observed/expected upper bound fraction"
2. Missense scores from gnomAD* browser: Z-score (positive value = more constraint) and o/e ratio and confidence interval (where upper bound of the Poisson-derived confidence interval is called MOEUF) from gnomAD* browser

- a. Missense Z-score > 3.09 is more likely to be intolerant of missense changes
3. UCSC PhastCons100Way score (higher means more conserved) and manual notes from visual inspection (e.g., “conserved in all but zebrafish”)
 - a. PhyloP ranges from -14 to 3, with higher values indicating more conserved
 - b. GERP++ is recommended for use in ClinGen (again, higher is more conserved).

*In our implementation, gnomAD v2.1.1 was used for exomes and v3.1.2 for genomes, but any subsequent gnomAD releases should be applicable for these steps as well.

Is there evidence of evolutionary constraint at the gene level?	
1	Constrained for LOF variation (LOEUF<0.35) and constrained for missense variation (z-mis>+2)
2	Constrained for LOF variation (LOEUF<0.35)
3	Constrained for missense variation (z-mis>+2)
5	Moderate evidence for constraint (LOEUF<0.5*** or z-mis>+1)
4	Minimal to no evidence of constraint

Variant frequency, pathogenicity, and position

Note: This section is only completed if the gene seems to be a good phenotypic match based on the previous sections.

1. Maximum allele frequency (gnomAD* population maximum)*
2. GnomAD* population group corresponding to popmax
3. Number of homozygotes (for compound heterozygous candidates) and number of heterozygotes (for *de novo* dominant candidates) (gnomAD*)
4. Computational pathogenicity predictors (# deleterious / # tolerated), reviewed from VarSome^{16,17}
5. ClinVar categorization and star category*
6. Proximity to other ClinVar variants (from visual inspection in DECIPHER¹⁸ database Protein/ Genomic track view)
7. Variant location relative to protein functional domains (from visual inspection in DECIPHER and UniProt)
8. Missense depleted area (from visual inspection in DECIPHER gnomAD* Missense track)

- a. DECIPHER > search for gene > Protein/Genomic tab > scroll down to tracks > gnomAD* missense track
9. Tissue expression of exon containing variant (gnomAD*)
 - a. From the gnomAD* browser, search for the gene name > copy the HGVS protein position into the field under “gnomAD* variants” > hover over location where (coding) variant is then scroll up through the tissues > hover over exon in each tissue and note the “pext” score (ranges 0=low to 1=high)

*In our implementation, gnomAD v2.1.1 was used for exomes and v3.1.2 for genomes, but any subsequent gnomAD releases should be applicable for these steps as well.

<i>Does the variant seem pathogenic?</i>	
1	Known/likely pathogenic variant
2	VUS - suspect likely pathogenic
3	VUS
4	VUS - suspect likely benign
5	Known benign variant

Summarizing information for final phenotypic fit score

Once information has been filled in and scores assigned for the intermediate steps, at least two clinicians will meet to go over the scores. We recommend keeping track of per-clinician notes and the group consensus.

1. Notes from individual clinicians (if any)
2. Evidence supporting association
3. Evidence against association
4. Reviewed as a group? (yes/no)
5. Notes from group meeting (if any)

<i>Summary Classification</i>	
1	Known disease association with strong phenotypic overlap with proband
2	Multiple lines of supporting evidence for association with proband phenotype
5	Multiple weak lines of evidence for association with proband phenotype

3	One line of supporting evidence for association with proband phenotype or conflicting evidence
4	No evidence for gene-phenotype association

Supplementary Note 3. Compound heterozygous mutational target is uniformly distributed

Recall that we define the mutational target of a compound heterozygous variant configuration as

$$\mu_{g,v_M,v_D} = \max(\mu_{g,v_M}, \mu_{g,v_D})^2.$$

We will show here that, by this definition, μ_{g,v_M,v_D} is uniformly distributed. Let μ_1 and μ_2 be dummy variables corresponding to some maternal and paternal variant mutational targets in a compound heterozygous configuration in gene g , and let μ_g be the mutational target of gene g .

$$\begin{aligned} Pr(\mu_{g,v_M,v_D}) &= Pr(\mu_{g,v_M,v_D} \geq \max(\mu_1, \mu_2)^2) \\ &= Pr(\mu_{g,v_M,v_D} \geq \max(\mu_1^2, \mu_2^2)) \\ &= Pr\left(\frac{\mu_{g,v_M,v_D}}{\mu_g^2} \geq \max\left(\frac{\mu_1^2}{\mu_g^2}, \frac{\mu_2^2}{\mu_g^2}\right)\right) \end{aligned}$$

Suppose $\max(\mu_1, \mu_2) = \mu_1$.

$$\begin{aligned} &= Pr\left(\frac{\mu_{g,v_M,v_D}}{\mu_g^2} \geq \frac{\mu_1^2}{\mu_g^2}\right) Pr\left(\frac{\mu_{g,v_M,v_D}}{\mu_g^2} \geq \frac{\mu_1^2}{\mu_g^2}\right) \\ &= Pr\left(\frac{\mu_{g,v_M,v_D}}{\mu_g^2} \geq \frac{\mu_1^2}{\mu_g^2}\right)^2 \\ &= Pr\left(\sqrt{\frac{\mu_{g,v_M,v_D}}{\mu_g^2}} \geq \frac{\mu_1}{\mu_g}\right)^2 \end{aligned}$$

Let $x = \sqrt{\mu_{g,v_M,v_D}/\mu_g^2}$. Recall that, by construction, variant mutational targets μ_{g,v_M} and

μ_{g,v_D} are both $\leq \mu_g$ and so $\mu_{g,v_M,v_D} \leq \mu_g^2$, and $\mu_{g,v_M,v_D}/\mu_g^2$ and its square root must range from 0 to 1 (Equation 3, Methods). So, $x \in [0, 1]$.

$$= Pr\left(x \geq \frac{\mu_1}{\mu_g}\right)^2$$

We have already established that μ_1/μ_g is uniformly distributed on $[0, 1]$. Since the cumulative distribution function of a uniformly distributed variable on $[a, b]$ is $(x - a)/(b - a)$ for $x \in [a, b]$ and therefore just x when $x \in [0, 1]$, we simplify the above expression as

$$= x^2 = \left(\sqrt{\frac{\mu_{g,v_M,v_D}}{\mu_g^2}}\right)^2 = \frac{\mu_{g,v_M,v_D}}{\mu_g^2}$$

Since we have shown that $Pr(\mu_{g,v_M,v_D}) = \mu_{g,v_M,v_D}/\mu_g^2$, by the same logic applied above, it follows that

$$\frac{\mu_{g,v_M,v_D}}{\mu_g^2} \sim \mathcal{U}_{[0,1]}$$

Supplementary Note 4. Biallelic variant independence is violated in consanguineous cases

Our individual-level compound heterozygous configuration statistic $\tilde{y}^c$ (Equation 18, Methods) and the expected number of genes harboring compound heterozygous configurations per individual genome $\tilde{\lambda}^c$ (Equation 21, Methods) both assume that the two inherited alleles comprising a compound heterozygous configuration arose as a result of independent mutations in the ancestral lineages of each parent. Homozygous recessive biallelic variants, where the exact same variant is inherited from each parent, are included in our framework and handled like compound heterozygous configurations so long as this variant independence assumption holds. In probands with consanguineous parentage, however, this assumption is violated because in most cases, a single ancestral mutational event resulted in the presence of a variant in both parents. Given this violation, we directly and stringently exclude individuals with evidence of consanguinity (i.e., requiring parental relatedness < 0.15 and maximum IBD length $< 3\text{Mb}$) rather than apply a simple test to filter individuals with relatively high recessive burden.

A large fraction of a proband's genome is autozygous in consanguineous cases, with up to 6.3% composed of identical-by-descent (IBD) regions in offspring of first cousin marriages. Pinpointing disease-causing homozygous variants based on statistical analysis alone, that is, by identifying overlapping IBD regions across affected individuals, is extremely difficult in a moderately-sized cohort. One approach for handling homozygous recessive biallelic variants in consanguineous cohorts has been proposed by HC Martin et al., 2018 and again by the same group in 2024.^{19,20} Briefly, rather than considering only the product of the expected rates for each allele in a biallelic configuration (i.e., as in our squared mutational target), homozygous recessive variants are considered linearly alongside compound heterozygous variants considered quadratically. These two terms are each scaled to the proportion of individuals with and without a region of autozygosity spanning that gene in the population, and then summed together. However, this approach does not model a critical source of uncertainty regarding the presence of recessively-acting variants due to genealogical history of these variants.

References

1. Pedersen, B. S. *et al.* Somalier: rapid relatedness estimation for cancer and germline studies using efficient genome sketches. *Genome Med.* **12**, 62 (2020).
2. Seplyarskiy, V. *et al.* A mutation rate model at the basepair resolution identifies the mutagenic effect of polymerase III transcription. *Nat. Genet.* **55**, 2235–2242 (2023).
3. Clinical Genome Resource (ClinGen). <https://clinicalgenome.org/>.
4. Gene Curation Coalition (GenCC). <https://thegencc.org/>.
5. Genomics England PanelApp. <https://panelapp.genomicsengland.co.uk/>.
6. Online Mendelian Inheritance in Man (OMIM). <https://www.omim.org/>.
7. NHGRI-EBI Genome-Wide Association Studies (GWAS) Catalog. <https://www.ebi.ac.uk/gwas>.
8. Mouse Genome Informatics (MGI). <http://www.informatics.jax.org/>.
9. International Mouse Phenotyping Consortium (IMPC). <https://www.mousephenotype.org/> (2022).
10. Strande, N. T. *et al.* Evaluating the Clinical Validity of Gene-Disease Associations: An Evidence-Based Framework Developed by the Clinical Genome Resource. *Am. J. Hum. Genet.* **100**, 895–906 (2017).
11. Allen Brain Institute | Brain Map Portal. <https://portal.brain-map.org/>.
12. Genotype-Tissue Expression (GTEx) Portal. <https://gtexportal.org/home/>.
13. GeneCards | The Human Gene Database. <https://www.genecards.org/>.
14. Website. <https://www.genecards.org/>.
15. Genome Aggregation Database (gnomAD) Browser. <https://gnomad.broadinstitute.org/>.
16. VarSome. *VarSome* <https://varsome.com/>.
17. Kopanos, C. *et al.* VarSome: the human genomic variant search engine. *Bioinformatics* **35**, 1978–1980 (2018).
18. DECIPHER v11.24: Mapping the clinical genome.

<https://www.deciphergenomics.org/browser>.

19. Martin, H. C. *et al.* Quantifying the contribution of recessive coding variation to developmental disorders. *Science* (2018) doi:10.1126/science.aar6731.
20. Chundru, V. K. *et al.* Federated analysis of autosomal recessive coding variants in 29,745 developmental disorder patients from diverse populations. *Nature Genetics* **56**, 2046–2053 (2024).