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

Shilpa Nadimpalli Kobren[1,*], Mikhail A. Moldovan[1,*], Rebecca Reimers[2], Daniel Traviglia[1], Xinyun Li[3], Danielle Barnum[4], Alexander Veit[1], Julian Willett[5], Michele Berselli[1], William Ronchetti[1], Richard Sherwood[6], Joel Krier[7], Isaac S. Kohane[1], 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 Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT
4 Access to Medicine Foundation, Amsterdam, The Netherlands
5 Department of Neurology, Massachusetts General Hospital, Boston, MA
6 Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA
7 Department of Genetics, Atrius Health, Boston, MA

Abstract

Genomics for rare disease diagnosis has advanced at a rapid pace due to our ability to perform "N-of-1" analyses on individual patients. The increasing sizes of ultra-rare, "N-of-1" disease cohorts internationally newly enables cohort-wide analyses for new discoveries, but well-calibrated statistical genetics approaches for jointly analyzing these patients are still under development.[1,2] The Undiagnosed Diseases Network (UDN) brings multiple clinical, research and experimental centers under the same umbrella across the United States to facilitate and scale N-of-1 analyses. Here, we present the first joint analysis of whole genome sequencing data of UDN patients across the network. We apply existing and introduce new, well-calibrated statistical methods for prioritizing disease genes with *de novo* recurrence and compound heterozygosity. We also detect pathways enriched with candidate and known diagnostic genes. Our computational analysis, coupled with a systematic clinical review, recapitulated known diagnoses and revealed new disease associations. We make our gene-level findings and variant-level information across the cohort available in a public-facing browser (https://dbmi-bgm.github.io/udn-browser/). These results show that N-of-1 efforts should be supplemented by a joint genomic analysis across cohorts.

Introduction

For decades preceding the widespread application of DNA sequencing, identifying the genetic etiology of rare monogenic phenotypes including human diseases relied on segregation in pedigrees.[3] DNA sequencing enabled the analysis of sporadic cases with no segregation data.[4] Early studies analyzed small cohorts of phenotypically similar cases,[5,6] a highly successful approach that is, however, limited to diseases with multiple known patients with fairly homogeneous presentations. In the absence of such phenotypically matched case cohorts, N-of-1 studies of undiagnosed patients are gaining popularity.[7–10] By design, these studies cannot attain statistical power from the shared genotypes of unrelated patients and require extensive clinical and biological inquiry to prove the causal involvement of the genotype in disease.[11–13] The most recent phase of human Mendelian genetics employs a "data science" approach to gene discovery propelled by the joint genomic analysis of phenotypically broad cohorts. Recent studies by the Deciphering Developmental Disorders and 100,000 Genomes consortia have demonstrated the power of this approach to identify new diagnoses and disease genes.[1,14]

Here, we apply existing and newly developed statistical genetics methods to the Undiagnosed Diseases Network (UDN) cohort that includes extremely difficult-to-solve, likely genetic cases (Figure 1a-e). The unique, diagnostically elusive presentation is the only criterion for inclusion, and patients have varied presentations including neurological, musculoskeletal, immune, endocrine, cardiac, and other disorders. Symptom onset ranges from neonatal through late adulthood. In contrast to most existing cohorts, individuals accepted to the UDN have already undergone lengthy but ultimately unfruitful diagnostic odysseys prior to enrollment. These patients subsequently undergo extensive phenotypic characterization at clinical UDN sites.[15] Both broad HPO terms and highly detailed clinical notes are collected and made available for all UDN researchers. Phenotypic information includes laboratory evaluations, dysmorphology examinations, specialist assessments, surgical records, and imaging (Figure 1f).

There is a similar emphasis on collecting sequencing data, with whole genomes sequenced for probands and their immediate or otherwise relevant family members. Although smaller than some other rare disease cohorts,[2] the UDN—with a design bridging clinical, research and functional validation teams and a focus on extreme patient presentations—was thought to be optimized for N-of-1 analysis. Patients' detailed phenotypic information, ongoing confirmation of new diagnoses, and the potential enrichment for novel genetic disorders make for an ideal data space to validate and develop statistical approaches. We

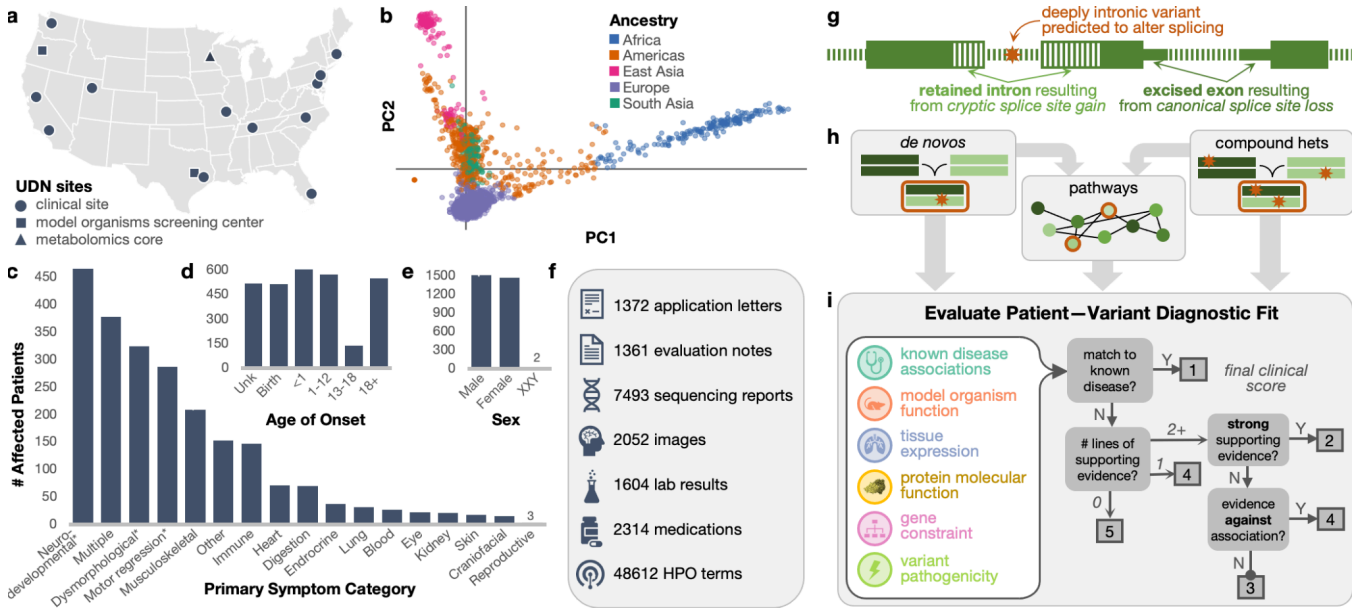

Figure 1. Undiagnosed Diseases Network cohort analysis. (a) Map of clinical and research sites within the Undiagnosed Diseases Network (UDN) for evaluating patients and candidate variant functionality. **(b)** Genetic ancestry across the patient cohort. **(c)** Clinician-recorded primary symptom categories of patients. "Multiple" indicates 2+ categories could be considered primary and "other" indicates an unlisted category. Categories marked with an asterisk (*) are neurological subtypes (Supplementary Note S1). **(d)** Patient-reported age of first symptom onset. **(e)** Patient sex. **(f)** Categories and quantity of phenotype information collected for patients and made available to all UDN researchers (icons are from Microsoft PowerPoint). **(g)** Intronic variants detectable from genome sequencing (orange star) may alter splicing, resulting in retained introns or excised exons in processed transcripts. **(h)** We consider genes and gene pathways harboring *de novo* and compound heterozygous variants. **(i)** Depiction of clinical framework to uniformly evaluate how well a patient's phenotypes are concordant with a candidate gene or variant.

hypothesized that despite the modest size and high phenotypic diversity of the UDN cohort, a joint statistical analysis of all cases would reveal new diagnoses. We harmonized and jointly called single nucleotide (SNV) and insertion/deletion (indel) variants across 4,236 individuals with whole genome sequencing in the UDN dataset and additionally called *de novo* mutations from aligned reads across complete trios (Methods, Supplementary Figure S1).[16]

Here, we generate candidate gene–patient matches via a series of statistical whole genomic analyses with no clinical input or phenotypic information used (Figure 1g,h). Each candidate is then evaluated with respect to the gene and variant's putative role in disease—based on known disease associations, functionality in model organisms, tissue expression, molecular function, evolutionary constraint, and *in silico* predicted pathogenicity—and the patient's clinical presentation to assess phenotypic match (Figure 1i). For genes or gene pathways harboring deleterious variants across multiple individuals, phenotypic similarity between patients is also assessed. To scale clinical evaluation to the cohort level, we developed a semi-quantitative protocol guided by the ClinGen framework[17] that uses hierarchical

decision models to increase efficiency and enables consistent and comparable evaluations of a gene–patient diagnostic fit by independent experts (Supplementary Note S2, Supplementary Figure S2). We calibrated the protocol during development by testing whether the resulting clinical scores assigned by different experts on the clinical team were in agreement. We validated the protocol in a blind test using decoy genes as controls. The clinical team applying the protocol consistently scored candidate genes higher than decoy genes (Wilcoxon rank-sum p-value = 0.0171, Methods, Supplementary Table S1), suggesting that the scores generated by the clinicians' protocol may be used to prioritize candidates.

We focus on the model of monogenic, autosomal inheritance in *de novo* and compound heterozygous cases, but do not limit the analysis to protein-coding exons, leveraging the availability of whole genome sequencing data in UDN. On the one hand, it remains challenging to identify non-coding regulatory variants involved in rare Mendelian diseases,[18] and the overall role of such variants in congenital disorders is still a subject of debate.[19] On the other hand, distal gain-of-splice site mutations creating new acceptor or donor splicing sites deep in the intronic sequences of genes are now a well-recognized cause of monogenic disease.[20] Identification of these variants in whole genome sequencing data is especially appealing due to their potential targetability for antisense oligonucleotide therapies.[21] We used a combined computational–experimental approach to prioritize distal splice variants using SpliceAI and an *in vitro* massively parallel splicing reporter assay (Methods, Supplementary Figure S3).[22,23] We found the per-variant *in silico* SpliceAI predictions to be mostly concordant with the *in vitro* assay readouts. Variants assigned higher SpliceAI scores are more frequently supported by the experimental assay, and those with relatively lower SpliceAI scores (< 0.5) have a non-negligible validation rate as well (Supplementary Figure S4). This prompted us to incorporate the full range of continuous SpliceAI scores, disregarding only the lowest scoring variants, in our statistics. We found this approach to consider distal splice-site variants attractive because it lends itself to a statistical analysis alongside exonic variants and does not rely on RNA sequencing. RNA sequencing has limitations for diagnosis because it depends on the availability of relevant tissue material that is especially challenging to obtain for neurodevelopmental patients, and it may miss lowly-expressed isoforms and those targeted by nonsense mediated decay.[24]

Analogously, for the analysis of exonic variants, we use the full continuous scale of *in silico* Combined Annotation Dependent Depletion (CADD) deleteriousness scores, which are based on variant impact on protein function and evolutionary conservation (Supplementary Figure S5).

De novo analysis

First, we tested all independent sporadic trios with complete sequencing data for *de novo* mutation etiologies. We detected 78.3 *de novo* point mutations and 9.5 *de novo* indels on average per proband genome concordant with the expectation.[25] Mutation count showed expected dependency on parental ages with Poisson-distributed adjusted counts, attesting to the quality of *de novo* calling (Figure 2b, Supplementary Figure S6).

To detect genes harboring *de novo* mutations across patients in the cohort beyond expectation, we first employed the previously published DeNovoWEST algorithm, which tests separate models for loss-of-function and gain-of-function mutations and was originally developed to analyze a cohort of over 30,000 exomes.[1] Clustering of missense mutations is considered in the gain-of-function model alongside mutation recurrence, whereas only mutation recurrence is considered in the loss-of-function model. We extended DeNovoWEST to incorporate deeply intronic variants with predicted splicing impact. In addition, we equipped the algorithm with the new base-pair resolution mutation rate model Roulette to replace the previously incorporated trinucleotide model and updated estimates of selection coefficients for heterozygous protein truncating variants (s_{het})[26,27] (Methods, Supplementary Figure S7).

This approach yielded two statistically significant genes (Bonferroni-adjusted p-value < 0.05) in seven probands (ranging from two to five patients per gene) with an additional enrichment for correct or highly plausible diagnoses among the top 200 hits (Fisher's exact test P = 2.6e-10, Methods, Supplementary Figure S8). This enrichment below the Bonferroni cutoff calls for a false discovery rate (FDR)-based approach. However, FDR correction is not immediately applicable because DeNovoWEST p-values are a construct over three sometimes dependent tests. This is generally not a problem when DeNovoWEST is applied to large cohorts because of the application of a conservative multiple hypothesis correction. We also noticed that some highly ranked false positive findings were driven by indel variants (e.g., *CSMD1*, a highly indel-prone fragile gene[28]).

This prompted us to derive and apply a simple closed-form statistic largely inspired by the DeNovoWEST enrichment procedure (Methods, Figure 2a),[29] while ignoring the distinction between loss-of-function and gain-of-function models. A key feature distinguishing our method from existing approaches is the way in which we incorporate variant-level functionality scores (e.g., SpliceAI and CADD). For a given observed variant in a gene, we define its "mutational target" as the sum of per-variant *de novo* mutation rates for all

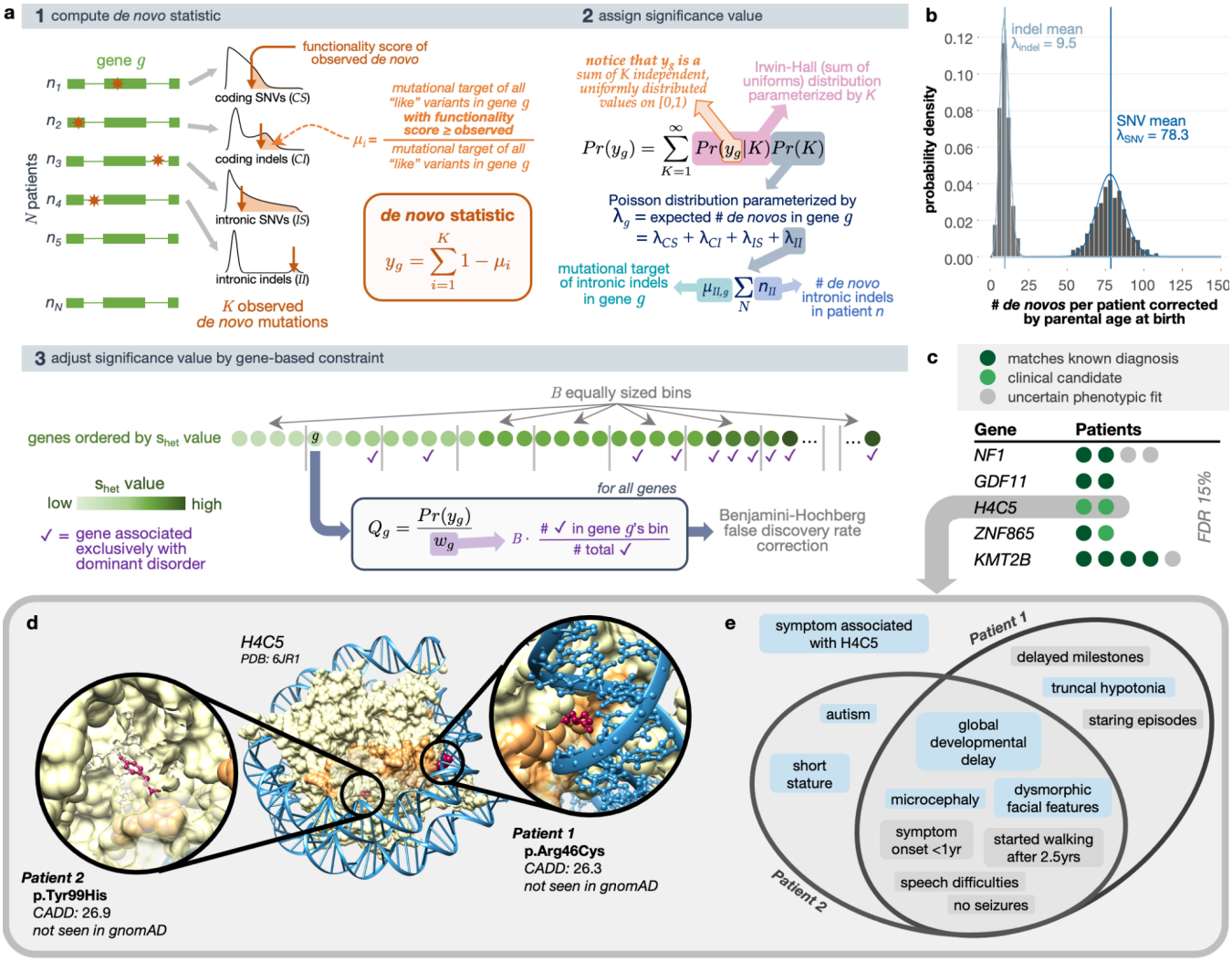

Figure 2. *De novo* recurrence. (a) Schematic of analytical test for the recurrence of *de novos* that considers distal splice-altering and exonic SNV and indel variants, their variant functionality scores, a genome-wide mutation rate model Roulette, and gene-based s_{het} constraint values. "Like" variants refer to those of the same variant class (i.e., *CS*, *CI*, *IS*, *II*) and within the same functionality score and minor allele frequency thresholds. **(b)** *De novo* mutation counts per proband adjusted for parental ages. Blue vertical lines show the mean values of the distributions, and curves represent the Poisson fits. **(c)** Genes with highest significance values for *de novo* recurrence across the cohort; patients are represented as colored circles. Complete gene list can be found in Supplementary Table S2. **(d)** 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. **(e)** 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.

possible variants with as high or higher a functionality score. By construction, this per-variant mutational target is expected to be a uniformly distributed statistic. Our framework naturally combines different variant types, including SNV and indel variants falling into coding and intronic space. We further include s_{het} gene constraint values, which indicate selection against heterozygous protein-truncating variants, by applying a weighted

FDR procedure.[30] Two added benefits of our analytical approach, which performs comparably to DeNovoWEST (Supplementary Figure S10), are its substantially decreased runtime due to the elimination of permutation-based significance calculations and its operation at the level of mutational targets, which can be shared across cohorts without revealing patients' individual variants.

At 15% FDR we obtain five genes covering 15 patients, nine of which corresponded to established diagnoses and three to highly plausible diagnoses following clinical review (Figure 2c, Supplementary Table S2). Both genes previously inferred by DeNovoWEST were recapitulated by our approach (i.e., *KMT2B* and *H4C5*). Two patients had high impact missense *de novo* variants impacting the *H4C5* gene; both patients had infantile-onset gross motor developmental delays, dysmorphic facial features, and speech difficulties (Figure 2d-e). These and other phenotypes exhibited by each patient were recently found to be linked to missense variants in histone H4 genes.[31] For one of the patients, the *de novo* variant was contemporaneously interpreted by UDN clinical experts to be causal.[32] Due to the findings of our analysis and the connection of UDN clinical sites through this project, the second patient's *de novo* variant has now been reclassified as "pathogenic" and resulted in a new diagnosis for this participant. Two other patients with sporadic neurodevelopmental delay each harbor truncating *de novo* variants in *ZNF865*. Both patients have phenotypic overlap with a series of 10+ other patients with *ZNF865* mutations, which makes a compelling case for pathogenicity.[33] Subsequent to the publication of the case series, we anticipate this gene–disease relationship will be established as causal and both variants to be reclassified as likely pathogenic. One high-ranking gene of interest, discovered uniquely by the DeNovoWEST gain-of-function clustering procedure, is *LRRC7* (Supplementary Figure S8b). The two impacted patients had phenotypic overlap of hypotonia and developmental delay; one patient additionally experienced nystagmus, staring spells, and balance problems and the second had ataxic gait. These findings and *LRRC7*'s expression in the brain further support its link to an emerging neurodevelopmental disorder.[14]

Both DeNovoWEST and our procedure are gene-based models, meaning not all patients with variants in highly-ranking genes are expected to be good phenotypic fits. In *NF1* and *KMT2B*, some patients are expected to harbor *de novo* variants by random chance that are unrelated to their condition (Methods, binomial p-value > 0.05), which is what we observe across three patients with poor phenotypic fits.

Compound heterozygous variant analysis

We next evaluate compound heterozygous (comphet) variants, which are the most likely cause of rare recessive disorders in populations with low degrees of consanguinity, as is largely the case in the United States.[34,35] Comphet variants are defined as a pair of distinct alleles landing within the same gene and inherited *in trans* from unaffected parents who are also heterozygous at these loci. These inherited disease-causing variants tend to be rare in the population. Directly selecting for highly deleterious comphet variants, even restricting to those falling into known disease-causing genes, results in numerous false positive findings at the cohort level, motivating a statistical approach. Developing a statistical framework analogous to *de novo* recurrence requires modeling the distribution of rare inherited alleles per individual. The distribution of the total number of derived alleles per haploid genome is not dependent on demographic history of the population and therefore straightforward to model.[36,37] In contrast, the distribution of rare alleles is highly dependent on population structure, even at a very fine scale, which is notoriously difficult to account for.[38,39] Some previous successful approaches ignore population structure while modeling the number of rare variants.[35] This approach results in power loss, which can be acceptable for very large cohorts but potentially intolerable in the case of smaller cohorts.

In our model, we sidestep directly modeling the distribution of rare variants and instead condition on the number of rare variants inherited from each parent. We compute the probabilities of comphet variants landing in high-scoring positions in the same gene across the cohort. Although the positional distribution of inherited variants is influenced in part by direct and background selection and biased gene conversion, for very rare variants, the effect of these factors is negligible compared to the effect of the variation in mutation rate along the genome and the overall gene target size.[27,40] Here we present a method that relies on the comphet mutational target, defined similarly as before as the total squared mutation rate of all possible variants with higher functionality scores. Our method uses variant-level functionality scores but not gene constraint scores, which do not exist for recessive selection.[41] Our approach computes well-calibrated per-gene p-values for comphet variants in a cohort (Methods, Supplementary Figure S9).

Across the set of non-consanguineous UDN families, we did not find significant recurrent comphet occurrences (Supplementary Figures S9, S11). However, our individual-based statistic yielded a ranked list of patient–gene pairs that was significantly enriched for correct diagnoses among the top seven genes (permutation p-value = 0.001, Methods, Figure 3a-b, Supplementary Table S3). Among these genes, we recapitulated three known diagnoses (i.e., *NEUROG3*, *COX20* and *NDUFAF8*)[42–45] and newly identified the genomic

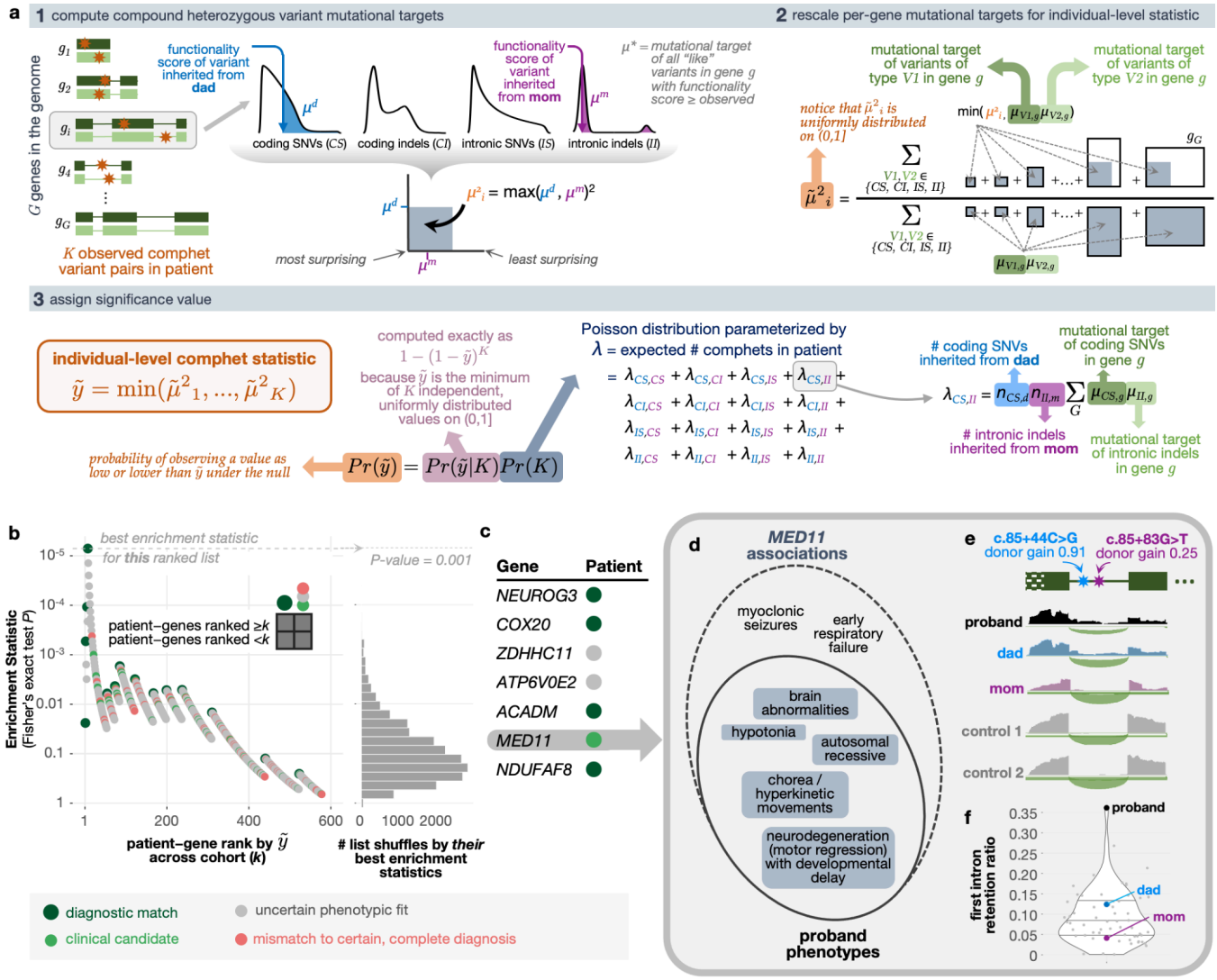

Figure 3. Compound heterozygous variants. (a) Illustration of comphet variants observed in an individual and simplified graphics depicting the calculation of a comphet mutational target, the rescaling of that mutational target with respect to all *G* genes in a genome to compute an individual-level statistic, and the significance calculation for that individual-level comphet statistic. "Like" variants refer to those of the same variant class (i.e., *CS*, *CI*, *IS*, *II*) and within the same functionality score and minor allele frequency thresholds. **(b)** Fisher's exact test *P* values indicate an enrichment of correct diagnoses (dark green circles) among top-ranked genes (left); this enrichment is significant based on minimum Fisher's exact test *P*s achieved across 10,000 random shuffles of the gene list (right). **(c)** Top seven genes resulting in minimum Fisher's exact test *P*; dark green indicates recapitulated diagnoses and light green indicates compelling phenotypic fit. **(d)** Overlap between phenotypes associated with *MED11* and those exhibited by the affected patient. **(e)** RNA-Seq reads from whole blood samples aligned to first two exons and first intron of *MED11* for proband (black), mom (purple), dad (blue) and two tissue-matched control samples (gray). Thin green line represents the intron, solid boxes represent protein-coding exonic regions, and the dotted box represents the 5' untranslated region of *MED11*. **(f)** Proband exhibits significant retention of the first intron relative to parents and fifty tissue-matched control samples. Intron retention ratio is calculated as the (median read depth of first intron) / (number of reads spanning first and second exons + median read depth of first intron).

cause of a known biochemical diagnosis (i.e., ACADM in a patient with MCAD deficiency). Two genes (i.e., ZDHHC11 and ATP6V0E2) remain candidates for pathogenicity due to their

phenotypically-relevant tissue expression. However, there is not enough functional evidence or published gene–disease relationships to establish causality at this time (Figure 3c). Finally, we identified comphet variants in *MED11* which are now leading diagnostic candidates in an undiagnosed patient experiencing neurodegeneration, developmental delay, brain abnormalities, chorea, and hypotonia (Figure 3d). *MED11* is associated with epilepsy and intellectual disability, and this patient's presentation could represent a phenotypic expansion of this known disorder.[46] Both inherited variants occur in the first intron of *MED11* and are predicted to cause cryptic splice donor gains. Transcriptome (RNA) sequencing of blood samples from the affected patient and both parents highlighted a significantly higher rate of first intron retention in the affected patient relative to both parents and to fifty unrelated blood control samples (Figure 3e-f, Supplementary Figure S12).[47]

Our comphet models do not generalize to rare homozygous variants (Supplementary Note S4). However, due to low levels of consanguinity in the UDN cohort, we do not expect homozygous recessive variants to underlie a substantial portion of diagnoses in this dataset.[35]

Pathway analysis

Genes involved in the same pathway are frequently involved in similar phenotypic presentations.[48–51] This provides an enticing possibility of drawing statistical power from multiple independent occurrences of deleterious variants in the same functional units, rather than just in the same genes. Moreover, therapeutics for disorders of the same functional unit that are individually too rare to meet minimal participant requirements for clinical trials may be evaluated together within the same umbrella or basket trial for more efficient approval.[52] However, such an approach should be pursued with caution, as the phenotypes stemming from perturbations of different genes in the same functional unit may vary to a great extent. Such differences in patient presentations may render the clinical evaluations and therapeutic potential of statistically significant findings virtually impossible. To mitigate this issue, we first initially consider groups of patients with similar phenotypes, and then within each of these groups, assess the overrepresentation of deleterious mutations across established biological pathways (Figure 4a).

We start by clustering 2,662 affected patients—with or without sequencing data—into 120 groups (median = 17, min = 2, max = 97 patients per cluster) based on the semantic similarity of their phenotype terms. Within each cluster, we then combine our *de novo* candidates, compound heterozygous candidates and known UDN diagnoses and perform gene set enrichment analysis (Methods). We focus our attention on undiagnosed cases with *de novo*

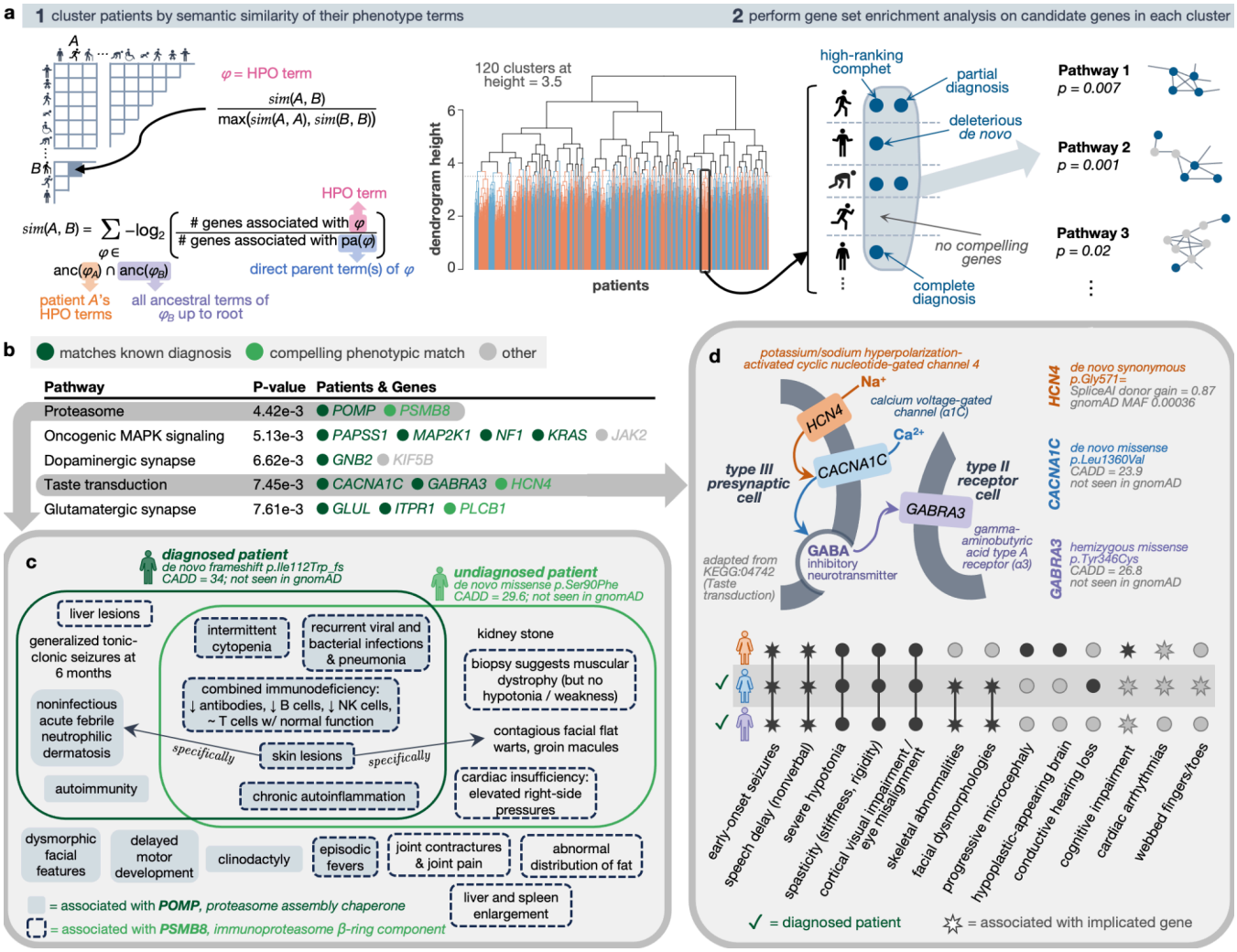

Figure 4. Biological pathways enriched within phenotypically-similar patient subgroups. (a) Schematic illustrating the two-step process of first clustering patients according to the semantic similarity of their phenotype terms and second finding enriched biological pathways among the genes within each patient cluster. **(b)** The most significant pathways per cluster (adjusted *p*-value < 0.01) with 1+ genes from 1+ undiagnosed patients; complete list in Supplementary Table S4. **(c)** Two patients with primarily immune-related symptoms each harbored a compelling *de novo* variant in genes involved in immunoproteasome assembly (*POMP*) and structure (*PSMB8*). Their symptoms strongly overlap, and a subset of these symptoms were also known to be associated with either gene in OMIM. **(d)** Three neurological patients had variants in transmembrane genes involved in the same pathway. These patients had substantial phenotypic overlap with each other, as expected, and with the phenotypes associated with each of their genes (depicted as star shapes in the upset plot).

or compound heterozygous candidates within enriched pathways in each cluster (Figure 4b). We also report all enriched pathways including those with only diagnosed patients for potential therapeutic grouping (Supplementary Table S4).

Two of three total candidate genes in one cluster with 19 immunological disorder patients are both involved in the immunoproteasome complex (KEGG:03050, n = 46, adjusted p-value = 4.42e-3). One patient's genome contained a known diagnostic, *de novo* frameshift

variant in *POMP*, an immunoproteasome chaperone protein.[53] An undiagnosed patient with evidence of chronic inflammation, recurrent infections, and skin lesions had a missense *de novo* in *PSMB8*, a component of the immunoproteasome □-ring with overlapping phenotypic associations (OMIM:256040). Both patients had similar combined immunodeficiency beyond what was captured in their standardized phenotype terms, including decreased global antibodies, decreased B cells and natural killer cells, and retained T cell functionality (Figure 4c). Disruptions to immunoproteasome assembly and structure have been shown to lead to an accumulation of precursor intermediates, impaired proteolytic activity and subsequent uncontrolled inflammation.[54]

In another cluster of 15 similarly presenting neurological patients, three candidate transmembrane genes were represented in the same functional pathway named for some genes' known involvement in taste transduction (KEGG:04742, n = 85, adjusted *p*-value = 7.45e-3). Two of these genes, *CACNA1C* and *GABRA3*, harbored high impact *de novo* and hemizygous missense variants respectively, corresponding to known patient diagnoses.[55,56] The genome of another, undiagnosed, now deceased patient from this cluster with no prior candidate variants contained a synonymous *de novo* variant predicted to alter splicing in another gene in the same functional pathway, *HCN4* (Figure 4d). All three patients exhibited seizures at a young age, speech delays, severe hypotonia, spasticity and visual impairment. Mouse knockouts of *HCN4* demonstrate neurological phenotypes.[57,58] In humans, *HCN4* is expressed in the visual and nervous systems and has recently been associated with infantile epilepsy, suggesting that this patient's undiagnosed disorder plausibly represents a phenotypic expansion of this gene.[57,58]

Discussion

In total, we analyze 886 sporadic or suspected recessive cases with complete trio or quad genome sequencing alongside an additional 463 phenotyped, diagnosed individuals using computational methods to identify *de novo* recurrence, compound heterozygosity, and pathway enrichment. We establish five new diagnoses and three new putative diagnoses in known disease-causing genes or genes previously unlinked to these patients' exact presentations. Our prioritization framework for pathway analysis further recapitulates 70 known *de novo* and 10 known comphet diagnoses and suggests 82 *de novo* and eight comphet candidates for follow-up.

In the field of common disease genetics, statistical inference of disease-associated genomic loci is confidently regarded as primary evidence for their causality. Rare disease genetics, in contrast, is in a transition state. Due to a lack of large disease-matched cohorts, N-of-1

analyses relying heavily on detailed patient phenotyping and clinical intuition have typically been used to generate candidate variant hypotheses. Evidence required to shift these variants from uncertain significance to known pathogenic status comes from experimental, functional studies and by identifying additional, unrelated, genotype-matched individuals with similar phenotypes through variant matchmaking services such as MatchMaker Exchange.[59,60] Recently, analyses of large, broadly-phenotyped cohorts of N-of-1 patients have demonstrated the potential for statistical approaches to reveal diagnoses and generate new gene discoveries in the rare disease space as well.[1,2,61]

Because the genome is a finite place, the human population is effectively a natural mutation saturation experiment, where all variants that are compatible with life are represented. This suggests that statistical methods applied to sufficiently large sequenced cohorts of rare disease patients, even those with diverse phenotypic presentations like the UDN, will enable the eventual discovery of all causes of monogenic disease. In order to take statistical discoveries as primary evidence, as is the case for common diseases, we need accurate, well-calibrated statistical methods. Even slight model misspecification may propagate and exacerbate the rate of false discoveries. The rapid growth of genomic datasets on which these models may be applied, coupled with an ongoing difficulty in phenotyping patients at scale to confirm findings,[62] further increases the urgency for more rigorous models.

Here we show that well-calibrated statistical models can be built for both *de novo* and compound heterozygous modes of inheritance. We show that analysis of a diverse, moderately-sized disease cohort at the gene and the pathway level shows promise.

Even though genomic sequencing has been liberalized, currently many analyses are still restricted to individual programs, and regulatory and technical barriers prevent sharing individual-level variant data broadly. In contrast, there are avenues for sharing some variant-level data in a way that is easily accessible to clinical geneticists. MatchMaker Exchange, for instance, enables the sharing of specific variants prioritized through N-of-1 analyses with the goal of finding new genotype- and phenotype-matched patients. Broadening the success of MatchMaker Exchange to include variants that may not have risen to the level of strong candidates in N-of-1 analyses is desirable. We developed a browser containing our gene-level findings and variant-level information about rare genetic variation in UDN patients (https://dbmi-bgm.github.io/udn-browser/). In addition, we developed RaMeDiES (Rare Mendelian Disease Enrichment Statistics, pronounced "remedies"), a software package implementing closed-form statistics for *de novo* recurrence and deleterious compound heterozygous inference proposed here.

Code Availability

Our software package RaMeDiES is available at https://github.com/hms-dbmi/RaMeDiES.

Acknowledgments
Thank you to the following individuals and organizations: Feruza Abraamyan for her contribution in the initial stages of developing the clinical evaluation protocol, Tian Yu for formatting RNA library reads for the MPSA analysis, the Undiagnosed Diseases Network Tool Building Coalition working group for advice on variant calling and sequencing quality metrics, Cecilia Esteves for sequencing file management, Amazon Web Services for complimentary data processing cloud credits, Rafael Aldana and members of the Harvard University Research Computing team for advice in optimizing joint variant calling, Logan Blaine for the initial local run of DeNovoWEST, Stan Nelson, Rosario Ivetth Corona and George Carvalho Neto for transcriptome outlier analysis for *MED11*, Vladimir Seplyarskiy and Ryan McGinty for advice regarding mutation rate models, Kaitlin Samocha for critically reviewing our manuscript, and members of the Sunyaev and Kohane research groups for helpful feedback on the manuscript text and figures. This work was funded by the National Institutes of Health (NIH) Common Fund grant U01HG007530, NIH National Institute of Neurological Disorders and Stroke (NINDS) grant U2CNS132415, and NIH National Human Genome Research Institute (NHGRI) grant R01HG012286.

Author Contributions
S.N.K., M.A.M. and S.R.S. designed the study, developed and applied the statistical models, and wrote the manuscript. I.S.K. provided guidance and feedback. S.N.K. and D.T. harmonized the sequencing data and M.B. and W.R. provided assistance. R.R. and J.K. developed the clinical evaluation protocol. R.R., J.K. and J.W. applied the clinical protocol. X.L. ran DeNovoWEST. D.B. and R.S. designed and performed the MPSA experiment, S.N.K. analyzed the data. A.V. developed the variant browser. All authors read and approved the final manuscript.

References

1. Kaplanis, J. *et al.* Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature* **586**, 757–762 (2020).
2. 100,000 Genomes Project Pilot Investigators *et al.* 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. *N. Engl. J. Med.* **385**, 1868–1880 (2021).
3. Marx, J. L. The cystic fibrosis gene is found. *Science* **245**, 923–925 (1989).
4. Roach, J. C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636–639 (2010).
5. O'Roak, B. J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat. Genet.* **43**, 585–589 (2011).
6. Jin, S. C. *et al.* Contribution of rare inherited and de novo variants in 2,871 congenital

heart disease probands. *Nat. Genet.* **49**, 1593–1601 (2017).
7. Vissers, L. E. L. M. *et al.* A de novo paradigm for mental retardation. *Nat. Genet.* **42**, 1109–1112 (2010).
8. Chong, J. X. *et al.* The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am. J. Hum. Genet.* **97**, 199–215 (2015).
9. Zurek, B. *et al.* Solve-RD: systematic pan-European data sharing and collaborative analysis to solve rare diseases. *Eur. J. Hum. Genet.* **29**, 1325–1331 (2021).
10. Boycott, K. M. *et al.* Care4Rare Canada: Outcomes from a decade of network science for rare disease gene discovery. *Am. J. Hum. Genet.* **109**, 1947–1959 (2022).
11. Hartley, T. *et al.* New Diagnostic Approaches for Undiagnosed Rare Genetic Diseases. *Annu. Rev. Genomics Hum. Genet.* **21**, 351–372 (2020).
12. Kobren, S. N. *et al.* Commonalities across computational workflows for uncovering explanatory variants in undiagnosed cases. *Genet. Med.* **23**, 1075–1085 (2021).
13. Chung, H.-L. *et al.* Loss- or Gain-of-Function Mutations in ACOX1 Cause Axonal Loss via Different Mechanisms. *Neuron* **106**, 589–606.e6 (2020).
14. Greene, D. *et al.* Genetic association analysis of 77,539 genomes reveals rare disease etiologies. *Nat. Med.* **29**, 679–688 (2023).
15. Splinter, K. *et al.* Effect of Genetic Diagnosis on Patients with Previously Undiagnosed Disease. *N. Engl. J. Med.* **379**, 2131–2139 (2018).
16. Mohanty, A. K. *et al.* novoCaller: a Bayesian network approach for de novo variant calling from pedigree and population sequence data. *Bioinformatics* **35**, 1174–1180 (2019).
17. 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).
18. Albert, F. W. & Kruglyak, L. The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.* **16**, 197–212 (2015).
19. Umans, B. D., Battle, A. & Gilad, Y. Where Are the Disease-Associated eQTLs? *Trends Genet.* **37**, 109–124 (2021).
20. Rowlands, C. *et al.* Comparison of in silico strategies to prioritize rare genomic variants impacting RNA splicing for the diagnosis of genomic disorders. *Sci. Rep.* **11**, 20607 (2021).
21. Crooke, S. T., Baker, B. F., Crooke, R. M. & Liang, X.-H. Antisense technology: an overview and prospectus. *Nat. Rev. Drug Discov.* **20**, 427–453 (2021).
22. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535–548.e24 (2019).
23. Rhine, C. L. *et al.* Massively parallel reporter assays discover de novo exonic splicing mutants in paralogs of Autism genes. *PLoS Genet.* **18**, e1009884 (2022).
24. Lindeboom, R. G. H., Supek, F. & Lehner, B. The rules and impact of nonsense-mediated mRNA decay in human cancers. *Nat. Genet.* **48**, 1112–1118 (2016).
25. Veltman, J. A. & Brunner, H. G. De novo mutations in human genetic disease. *Nat. Rev.*

Genet. **13**, 565–575 (2012).
26. Cassa, C. A. *et al.* Estimating the selective effects of heterozygous protein-truncating variants from human exome data. *Nat. Genet.* **49**, 806–810 (2017).
27. 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).
28. Gao, G. *et al.* Common fragile sites (CFS) and extremely large CFS genes are targets for human papillomavirus integrations and chromosome rearrangements in oropharyngeal squamous cell carcinoma. *Genes Chromosomes Cancer* **56**, 59–74 (2017).
29. Jiang, Y. *et al.* Incorporating Functional Information in Tests of Excess De Novo Mutational Load. *Am. J. Hum. Genet.* **97**, 272–283 (2015).
30. Ignatiadis, N., Klaus, B., Zaugg, J. B. & Huber, W. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nat. Methods* **13**, 577–580 (2016).
31. Tessadori, F. *et al.* Recurrent de novo missense variants across multiple histone H4 genes underlie a neurodevelopmental syndrome. *Am. J. Hum. Genet.* **109**, 750–758 (2022).
32. Borja, N. *et al.* H4C5 missense variant leads to a neurodevelopmental phenotype overlapping with Angelman syndrome. *Am. J. Med. Genet.* A **191**, 1911–1916 (2023).
33. Program Planner. https://www.abstractsonline.com/pp8/#!/9070/presentation/2575.
34. Wright, C. F. *et al.* Evaluating variants classified as pathogenic in ClinVar in the DDD Study. *Genet. Med.* **23**, 571–575 (2021).
35. Martin, H. C. *et al.* Quantifying the contribution of recessive coding variation to developmental disorders. *Science* **362**, 1161–1164 (2018).
36. Simons, Y. B., Turchin, M. C., Pritchard, J. K. & Sella, G. The deleterious mutation load is insensitive to recent population history. *Nat. Genet.* **46**, 220–224 (2014).
37. Do, R. *et al.* No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nat. Genet.* **47**, 126–131 (2015).
38. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
39. Sohail, M. *et al.* Negative selection in humans and fruit flies involves synergistic epistasis. *Science* **356**, 539–542 (2017).
40. Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).
41. Balick, D. J., Jordan, D. M., Sunyaev, S. & Do, R. Overcoming constraints on the detection of recessive selection in human genes from population frequency data. *Am. J. Hum. Genet.* **109**, 33–49 (2022).
42. Wang, J. *et al.* Mutant neurogenin-3 in congenital malabsorptive diarrhea. *N. Engl. J. Med.* **355**, 270–280 (2006).
43. Xu, H. *et al.* Observation of novel COX20 mutations related to autosomal recessive

axonal neuropathy and static encephalopathy. *Hum. Genet.* **138**, 749–756 (2019).
44. Nohara, F., Tajima, G., Sasai, H. & Makita, Y. MCAD deficiency caused by compound heterozygous pathogenic variants in ACADM. *Hum Genome Var* **9**, 2 (2022).
45. Alston, C. L. *et al.* Pathogenic Bi-allelic Mutations in NDUFAF8 Cause Leigh Syndrome with an Isolated Complex I Deficiency. *Am. J. Hum. Genet.* **106**, 92–101 (2020).
46. Calì, E. *et al.* A homozygous MED11 C-terminal variant causes a lethal neurodegenerative disease. *Genet. Med.* **24**, 2194–2203 (2022).
47. Middleton, R. *et al.* IRFinder: assessing the impact of intron retention on mammalian gene expression. *Genome Biol.* **18**, 51 (2017).
48. Rual, J.-F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173–1178 (2005).
49. Costanzo, M. *et al.* The genetic landscape of a cell. *Science* **327**, 425–431 (2010).
50. Rolland, T. *et al.* A proteome-scale map of the human interactome network. *Cell* **159**, 1212–1226 (2014).
51. Ferrari, S. *et al.* Retinitis pigmentosa: genes and disease mechanisms. *Curr. Genomics* **12**, 238–249 (2011).
52. Park, J. J. H. *et al.* Systematic review of basket trials, umbrella trials, and platform trials: a landscape analysis of master protocols. *Trials* **20**, 572 (2019).
53. Poli, M. C. *et al.* Heterozygous Truncating Variants in POMP Escape Nonsense-Mediated Decay and Cause a Unique Immune Dysregulatory Syndrome. *Am. J. Hum. Genet.* **102**, 1126–1142 (2018).
54. Brehm, A. *et al.* Additive loss-of-function proteasome subunit mutations in CANDLE/PRAAS patients promote type I IFN production. *J. Clin. Invest.* **125**, 4196–4211 (2015).
55. Rodan, L. H. *et al.* Phenotypic expansion of CACNA1C-associated disorders to include isolated neurological manifestations. *Genet. Med.* **23**, 1922–1932 (2021).
56. Syed, P., Durisic, N., Harvey, R. J., Sah, P. & Lynch, J. W. Effects of GABAA Receptor α3 Subunit Epilepsy Mutations on Inhibitory Synaptic Signaling. *Front. Mol. Neurosci.* **13**, 602559 (2020).
57. Campostrini, G. *et al.* A Loss-of-Function HCN4 Mutation Associated With Familial Benign Myoclonic Epilepsy in Infancy Causes Increased Neuronal Excitability. *Front. Mol. Neurosci.* **11**, 269 (2018).
58. Blake, J. A. *et al.* Mouse Genome Database (MGD): Knowledgebase for mouse-human comparative biology. *Nucleic Acids Res.* **49**, D981–D987 (2021).
59. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).
60. Osmond, M. *et al.* Outcome of over 1500 matches through the Matchmaker Exchange for rare disease gene discovery: The 2-year experience of Care4Rare Canada. *Genet. Med.* **24**, 100–108 (2022).

61. Coe, B. P. *et al.* Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. *Nat. Genet.* **51**, 106–116 (2019).
62. Chopra, M. & Duan, T. Rare genetic disease in China: a call to improve clinical services. *Orphanet J. Rare Dis.* **10**, 140 (2015).

Online Methods for

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

Shilpa Nadimpalli Kobren[1,*], Mikhail A. Moldovan[1,*], Rebecca Reimers[2], Daniel Traviglia[1], Xinyun Li[3], Danielle Barnum[4], Alexander Veit[1], Julian Willett[5], Michele Berselli[1], William Ronchetti[1], Richard Sherwood[6], Joel Krier[7], Isaac S. Kohane[1], 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 Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT
4 Access to Medicine Foundation, Amsterdam, The Netherlands
5 Department of Neurology, Massachusetts General Hospital, Boston, MA
6 Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA
7 Department of Genetics, Atrius Health, Boston, MA

Undiagnosed Diseases Network (UDN) structure **2**
Harmonization of whole genome sequencing data **2**
Clinical evaluation framework **3**
Protocol overview 3
Clinical score calibration 4
Validation 4
Massively Parallel Splicing Reporter Assay (MPSA) **4**
Assay design 4
Library cloning and experimental protocol 5
Barcode mapping 5
MPSA validation rate 5
Identification of de novo variants **6**
DeNovoWEST gene-specific enrichment of de novo variants **6**
Analytical test for de novo cohort-level recurrence **7**
Basic statistic definition 7

Different deleteriousness scores for coding and intronic variants....9
Different mutation rate models for SNV and indel variants.... 9
Incorporation of different variant types.... 9
Incorporation of shet values....10
Analytical test for compound heterozygous cohort-level recurrence.... 11
Modeling false positive diagnoses.... 12
Analytical test for individual-level compound heterozygous configuration....12
Enrichment for correct diagnoses....14
Pathway enrichment analysis....14
Phenotypically-similar patient groupings....14
Selecting genes per patient cluster....15
Gene Set Enrichment Analysis (GSEA).... 15
References.... 15

Undiagnosed Diseases Network (UDN) structure

The Undiagnosed Diseases Network (UDN) was established in 2014 with the goal of uncovering clinical diagnoses and novel disease-causing genetic variants and their molecular functionalities. In its current phase, the UDN is composed of 12 clinical research centers across the United States and a CLIA-certified sequencing core at Baylor Genetics. Typical UDN patients have already endured a multiyear “diagnostic odyssey” of extensive prior testing by multiple medical specialists and often inconclusive targeted, whole exome and even whole genome sequencing at the time of their application to the UDN.

As part of the application process, a team of clinicians and genetic counselors at one of the UDN clinical sites reviews the patient’s medical records, referral letters and lab data and creates an abstracted case review document. If the team concludes that a UDN evaluation may aid in the identification of a diagnosis, the patient is accepted to the program and undergoes a thorough in-person evaluation at their assigned clinical site. Most patients and available affected and unaffected family members receive whole genome sequencing (GS) as well. All genomic sequencing data, clinical sequencing reports prepared in accordance with the American College of Medical Genetics and Genomics (ACMG) variant classification guidelines, structured phenotyping in the form of Human Phenotype Ontology (HPO) terms, lab results, imaging data, medication data, referral letters and clinical notes, the abstracted case review document, and candidate variants and genes are uploaded to the UDN Data Management and Coordinating Center. All of this information is available to researchers when evaluating gene–phenotype fit for a specific patient and candidate gene.

Moreover, UDN patients have consented to follow-up if additional tests or information are deemed useful.

Harmonization of whole genome sequencing data

Short-read whole genome sequencing was performed between 2014 and 2022 in accordance with the UDN Manual of Operations, which specifies that the average coverage across the genome must be >40X, and >97.5% of all coding and noncoding genes (UTRs, coding regions, and intronic regions) must be covered at >20X. Paired-end FASTQs were retrieved in June 2022 for 4268 samples collected from 4236 unique individuals. Six individuals subsequently dropped out of the UDN program and are excluded from the analyses presented here. All FASTQ pairs were within expected parameters (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and were aligned to human reference hg38 (with decoys and all alt contigs) using the Sentieon[1] bwa-mem implementation via the Clinical Genome Analysis Pipeline (CGAP, https://cgap.hms.harvard.edu/). Read groups were added via a custom CGAP script, multiple FASTQ pairs corresponding to the same sample were merged, and resulting BAMs were sorted, deduped, and recalibrated using a Sentieon implementation. GVCFs were produced using CGAP's implementation of GATK's HaplotypeCaller. All processing steps from FASTQ to GVCF were deployed on spot instances in Amazon Web Services (AWS). GVCFs were then egressed to the Harvard Medical School institutional cluster. SNVs/indels were jointly called across genomic shards then merged using Sentieon tools. Per-sample sex and cross-sample relatedness were confirmed using Somalier (Supplementary Figure S1).[2] We required that all trios under consideration in our analysis had two parents reported as “unaffected”, a child reported as “affected”, parent–child relatedness 0.5±0.075, parent–parent relatedness <0.15, mothers had heterozygous variants present and a scaled mean depth of ~2 on chromosome X, and fathers had a scaled mean depth of ~1 on chromosome Y. All variants were annotated using Ensembl VEP (version 108) and slivar for TOPMed and per-population gnomAD variant frequencies and homozygote counts.[3,4]

For our compound heterozygote analysis, we inferred within-family regions of identity by descent (IBD) using KING.[5] We required at least one IBD region between the child and each parent to further confirm relatedness (in addition to kinship coefficient filtering) and no IBD regions of length >3Mb between parents to confirm non-consanguinity between parents. In families with multiple affected siblings, we select one sibling as the proband and disregard the other siblings during initial analyses. Variants in other affected siblings were then used to check segregation during validation of our findings. This process resulted in 846 non-consanguineous trios with an affected child and two unaffected parents for our analyses.

Clinical evaluation framework

Protocol overview

We developed a clinical analysis protocol to reduce subjectivity in the assessment of diagnostic candidates. We used the case evaluation process implemented at Brigham Genomic Medicine as a foundation.[6] We then transformed this process into a systematic and structured protocol with inspiration from the gene–disease association criteria developed by the Clinical Genome Resource (ClinGen) group.[7,8] Evidence in support of or against a candidate variant–participant match was evaluated by a team of clinical geneticists according to three categories for experimental evidence not taken into account by our statistical analyses: (i) model organism or cell line studies, (ii) tissue expression, and (iii) protein molecular function. Clinicians also took into account case-level data and published literature with case-control data including (iv) known disease associations, (v) gene evolutionary constraint, and (vi) variant pathogenicity. Discrepancies in opinion were mediated by joint discussion until a consensus decision was reached. A detailed description of the protocol and scoring scheme is available in Supplementary Note S2 and hierarchical decision trees to streamline the scoring process are provided in Supplementary Figure S2.

Clinical score calibration

We ensured that the protocol was specific and detailed to the extent that different clinicians with access to the same patient data would independently assign equal clinical scores to the same candidates. Over the course of two months, at least two clinicians each evaluated 2–3 compound heterozygous candidates per week and independently recorded their notes, final clinical scores, and score rationale in a REDCap database.[9] At weekly joint discussions, they iteratively updated the protocol to improve specificity and reduce discrepancies in scoring. The final two joint discussions confirmed that categorical and final clinical scores assigned by different clinicians were consistently in agreement.

Validation

The clinical team was provided with ten “candidates” and ten “decoys” from real UDN patients in random order to evaluate. The team was blinded to gene labels, variant inheritance and SpliceAI score information during evaluation. “Candidate” genes had two rare variants (gnomAD popmax AF < 0.001) inherited in *trans* where one variant was exonic with CADD > 23 and the second variant was intronic with a max SpliceAI > 0.3. “Decoy” genes were selected with identical criteria except that variants were actually inherited in *cis* or the intronic variant had a max SpliceAI score of 0. After assigning final clinical scores

to each of the 20 genes, the candidate/decoy labels were revealed to the clinical team (Supplementary Table S1).

Massively Parallel Splicing Reporter Assay (MPSA)

Assay design

We designed oligonucleotides to evaluate the impact of a variant predicted to cause a cryptic splice site gain or a canonical splice site loss. For each variant with a predicted splice-altering impact, we extracted the surrounding genomic sequence from the UDN patient harboring the variant (alternate) as well as a paired version with the variant of interest replaced with the reference allele (reference). We centered the candidate sequence on the variant of interest, noting that the impacted splice site junction could be up to 50 nucleotides away from the variant. For a subset of variants, we also generated candidate sequences that were centered on the predicted site of the altered splice junction rather than on the variant itself. We embedded each candidate sequence in an oligonucleotide template containing a 4-nt barcode and flanking primers as follows:

Splice donor library structure
GCACGGACAAAGTACTAGCC [155-nt candidate sequence][4-nt SD-associated barcode] GGAAGATCGACGCAGgtaagt

Splice acceptor library structure
TGCTCTTATGCGAACGTGTTAAC [4-nt SA-associated barcode] [151-nt candidate sequence] GGAAGATCGACGCAGgtaagtt

The final oligonucleotide library contained 6,000 200-nt oligonucleotides, encompassing 1,920 alternate/reference pairs, which we ordered from Twist Bioscience.

Library cloning and experimental protocol

The oligonucleotide library was cloned separately using PCR amplification and NEBuilder assembly into lentiviral splice acceptor (pLenti-MPSA-acceptor) and splice donor (pLenti-MPSA-donor) vectors. These vectors consisted of an EF1A promoter and an mCherry open reading frame (ORF) followed by splicing reporter modules based off of prior published massively parallel splicing reporter constructs[10,11] (Supplementary Figure S3) as well as a separate Puromycin selection cassette.

Lentiviral particles for each library were produced and titrated. Each library was transduced at a multiplicity of infection (MOI) of 0.3 in three biological replicates into $6.25*10^6$ cells/replicate of HepG2 (liver) and SK-N-SH (neural-like) cells, both acquired from American Type Culture Collection (ATCC). Cells were selected with Puromycin to completion, and genomic DNA and RNA were harvested one week after transduction.

PCR-based nextgen sequencing (NGS) library preparation was performed on all 12 genomic DNA and RNA samples. Libraries were sequenced with 75-nt paired-end reads using an Illumina NextSeq 500 sequencer, ensuring an average of >1,000 reads per library member from all libraries.

Barcode mapping

Over ~75% of all RNA reads could be mapped back to a 15-nt barcode found in our starting dictionary. This resulted in ~6–15 million mapped RNA reads per MPSA replicate, yielding a median of 1,170 mapped reads per alternate/reference library pair per replicate. Results from Tapestation, an automated electrophoresis system for sizing and quantifying nucleic acid samples, showed that 49.6% of mapped reads from splice donor MPSA experiments utilized some library splice donor site and 50.4% utilized the experimentally fixed site. Across splice acceptor MPSA experiments, 58.3% of mapped reads utilized some library splice acceptor site and 41.7% utilized the fixed site.

MPSA validation rate

We considered all alternate/reference library pairs with at least 10 barcode-disambiguated mapped reads each in one or more MPSA experiments; 99.4% of pairs met this requirement. Each read was then categorized as (1) using the experimentally fixed splice site, (2) using a splice site corresponding to a known intron/exon junction as annotated in Ensembl, (3) using the SpliceAI-predicted cryptic splice gain site, (4) using a cryptic splice site at a different location, (5) malformed where the read did not begin with the correct fixed sequence due to a next-generation sequencing error, or (6) recombined where the read did not align to the expected oligo sequence at all. The percent of malformed and recombined reads per alternate/reference pair was 7.5% (SD=1.9%) and 6.2% (SD=10.6%) respectively on average. The position of SpliceAI-predicted cryptic splice sites often did not correspond to the expected splice junction based on manual inspection or to the splice sites observed in MPSA experiments (55.4% of splice acceptor and 5.4% of splice donor predicted positions matched). We instead considered the most common cryptic splice site position observed in each alternate library sequence to be the predicted site. MPSA validation rate is computed per alternate/reference library pair as the difference in percentages of total reads supporting the predicted cryptic splice site between oligos

containing the alternate variant and the corresponding reference oligonucleotides (Supplementary Figure S4a).

We compared the MPSA validation rates across the three biological replicates and two cell types using Pearson's correlation (Supplementary Figure S4b).

Identification of *de novo* variants

For each of the 1463 sequenced trios in our harmonized UDN dataset, including trios with unaffected offspring, we select the subset of variants with read depth >10 and genotype quality (GQ) >20 across proband, mother and father. We further subset to variants with a "high" Roulette quality score, gnomAD population maximum allele frequency < 0.01, TOPMed[12] allele frequency < 0.01, proband alternate allele read depth >4 and frequency >0.2, and alternate read depth <2 in both parents.

We then utilize observed aligned reads across each trio and across thirty unrelated individuals to assign posterior probabilities to each putative *de novo* variant on autosomes using the CGAP reimplementation of novoCaller (https://cgap.hms.harvard.edu/).[13] We consider all *de novos* with a novoCaller posterior probability >0.7 to be high confidence, noting that thresholding the novoCaller posterior probability from 0.5 to 0.95 has negligible impact on the number of passing variants overall and per-proband (Supplementary Figure S6a). We further exclude probands with over 150 high confidence *de novo* calls, as these patients frequently had "suspected parental mosaicism" mentioned in their clinical records. Finally, because clonal sperm mosaicism may lead to siblings inheriting identical *de novo* variants, we exclude duplicate *de novo* variants within each family from downstream recurrence analyses.[14] This process resulted in 1072 trios with an affected proband and unaffected parents for further analysis.

DeNovoWEST gene-specific enrichment of de novo variants

We modified the DeNovoWEST weighted permutation test by first augmenting the set of variants under consideration beyond exonic variants to include all possible intronic variants in protein-coding genes with a SpliceAI score >0.4, resulting in ~400k additional possible variants under consideration.[15] To this end, we modified the codebase to consider these intronic putatively splice-altering variants to have the same functional consequence as canonical splice site variants if they had a VEP annotation of "splice_acceptor" or "splice_donor" or the same functional consequence as missense variants otherwise. We then updated the required precomputed values, including per-variant mutation rates,

minor allele frequencies, deleteriousness scores and per-region constraint values as detailed below, for all exonic and intronic variants under consideration (Supplementary Figure S7a). The underlying triplet-context mutational model was replaced with genome-wide, per-SNV Roulette mutation rate estimates.[16] Each variant's minor allele frequency was set to the maximum gnomAD-v3 population or TOPMed allele frequency. Per-variant Phred-scaled and unscaled CADD values were obtained from https://cadd.gs.washington.edu/ (version 1.6 for GRCh38/hg38). Updated per-gene s_{het} values were obtained from http://genetics.bwh.harvard.edu/genescores/selection.html and binned into a "low" category if mean s_{het} was below 0.15 and a "high" category otherwise.[17] Notably, some stable Ensembl gene IDs in GRCh37/hg19 are not present in GRCh38/hg38 and vice versa; all variants from the 894 GRCh38/hg38 genes without s_{het} values are binned into the "low" category. Regional missense constraint values, defined for adjacent windows covering the full genomic region of each protein-coding gene were obtained from https://gnomad.broadinstitute.org/downloads#exac-regional-missense-constraint. We translated these genomic region coordinates from GRCh37/hg19 to GRCh38/hg38 using UCSC's LiftOver tool and then assigned a constraint value to exonic and intronic variants corresponding to the genomic region they fell into. We recomputed the weights assigned to each variant type using the union of all *de novo* variants in our cohort and the *de novo* variants released with DeNovoWEST (encompassing ~31,000 exome-only trios), because the distribution of *de novo* variant classes in UDN data was very similar to the distribution of *de novo* variant classes in the dataset used by DeNovoWEST (Supplementary Figure S7b-c) and because the authors warn that weights generated from smaller datasets alone may be unreliable. Gene severity scores were then computed for every gene harboring one or more *de novo* variants across our cohort. We adjust DeNovoWEST assigned p-values using Bonferroni correction for twice the total number of genes evaluated as suggested by the authors.

Analytical test for *de novo* cohort-level recurrence

Basic statistic definition

We define a cohort as a set of N genomes (i.e., collections of genes) each with sets of *de novo* variants arising independently but based on the same background *de novo* mutation rate. Let μ_i denote the *de novo* mutation rate of a specific variant i. The mutational target of a gene g is

$$\mu_g = \sum_{i \in g} \mu_i \quad .$$

The mutational target of a variant v in gene g is

$$\mu_{g,v} = \sum_{i \in g} \mu_i 1_{score_i \geq score_v} \qquad \text{(Equation 1)}$$

where $score_i$ is the deleteriousness score of variant i. Intuitively, the more surprising and/or deleterious a variant, the smaller its mutational target. By definition, variant mutational targets are uniformly distributed from 0 to μ_g, so

$$\frac{\mu_{g,v}}{\mu_g} \sim U_{[0,1]}.$$

Suppose there are K *de novo* variants falling within gene g across the cohort, where $K \geq 1$. We define a statistic y as

$$y = \sum_{i=1}^{K} \left(1 - \frac{\mu_{g,v_i}}{\mu_g}\right). \qquad \text{(Equation 2)}$$

Note that y is a sum of K uniformly distributed variables on [0,1] under the null. The distribution of y given parameter K can thus be modeled by the Irwin-Hall (IH) "sum of uniforms" distribution, which has a closed form for its cumulative density function (CDF) and thus also for its survival function (SF), where $SF = 1 - CDF$. This enables us to replace permutation-based significance evaluations and instead analytically compute the probability of achieving a y as high or higher than observed with K variants using the IH survival function as $Pr(y|K) = IH_{SF}(y|K)$. We note that there are many other constructions over a set of uniformly-distributed random variables (such as *p*-values).[18,19] We further note that as the cohort size dramatically increases, the Irwin-Hall distribution can be replaced with the normal distribution.

Finally, we also model $Pr(K)$, the probability of K independent *de novo* variants to land in gene g given this cohort of size N, to assign an overall significance value to our statistic y as

$$Pr(y) = \sum_{K=1}^{\infty} Pr(y|K)Pr(K) \quad \text{(Equation 3)}$$

Because neither y nor $Pr(y|K)$ are defined for $K = 0$, we do not expect $Pr(y)$ to be uniformly distributed. Instead, only $Pr(y|K \geq 1) = Pr(y)/Pr(K \geq 1)$ is expected to be uniformly distributed (Supplementary Figure S9).

In a single genome with n total observed *de novo* variants, the number of *de novo* variants to land in a particular gene g, given that $\mu_g \ll 1$, is Poisson distributed, parameterized by the expected number of *de novos* $\lambda = n\mu_g$. In a cohort of N genomes, the number of *de novo* variants to land in gene g is therefore a sum of N Poisson-distributed random variables, which itself is also Poisson distributed. We thus compute $Pr(K) = Pois(K|\lambda)$, where λ is given by

$$\mathbb{E}[K] \equiv \lambda = \mu_g \sum_{j=1}^{N} n_j .$$

Different deleteriousness scores for coding and intronic variants

We use continuous, per-variant CADD and SpliceAI deleteriousness scores for coding and intronic variants respectively in our computations. Precomputed scores are publicly-available for all possible variants genome-wide, enabling the calculation of comprehensive, basepair-resolution mutational targets as described above. We use different variant functionality scores for exonic and intronic variants because we found that these values are poorly correlated with each other in intronic space (Supplementary Figure S5). Clinical sequencing centers also regularly report these scores, suggesting their relevance in rare disorders.[8,20]

Different mutation rate models for SNV and indel variants

We use Roulette *de novo* mutation rates for SNVs genome-wide. Different mutational processes lead to indel mutations, so Roulette values cannot necessarily be adapted to model this mutation type.[21] We approximated per-gene joint distributions of indel mutation rates and deleteriousness scores as follows. First, we considered all possible exonic indels of length ≤10nt for which precomputed CADD scores were available for download and all possible intronic insertions of length 1nt and deletions of length ≤4nt for which precomputed SpliceAI scores were available for download. Although SpliceAI provides predictions exhaustively for all possible indels, CADD provides scores for the subset of

indels observed in gnomAD-v2. We excluded all indels that overlapped with any SNVs assigned a Roulette "low quality" filter, which are based on gnomAD quality metrics, abnormal density of segregating sites, and suspicious patterns of recurrence. We further excluded indels with a gnomAD popmax MAF > 0.1%and/or a number of alleles in gnomAD (AN) in the bottom decile. For exonic and intronic variants separately, we binned all indels by their precomputed CADD or SpliceAI score rounded to the nearest hundredth. The total number of indels within a deleteriousness score bin and all bins corresponding to higher deleteriousness scores was used as an approximation to the mutational target associated with that score.

Incorporation of different variant types

Because there are different deleteriousness scores for coding and intronic variants and different mutational targets for SNV and indel variants, we expand our basic test statistic to accommodate different variant types $t \in$ {coding SNV, coding indel, intronic SNV, intronic indel}. We redefine a gene and variant mutational target with respect to each variant type as

$$\mu_{g,t} = \sum_{i \in g,t} \mu_i$$

and

$$\mu_{g,v,t} = \sum_{i \in g,t} \mu_i 1_{score_i \geq score_v}$$

where g, t refers to the subset of all possible variants in gene g of type t. We define y' as

$$y' = \sum_t \sum_{i=1}^{K} \left(1 - \frac{\mu_{g,v_i,t}}{\mu_{g,t}}\right) \sim IH\big(y' | \sum_t K_t\big)$$

where K_t is the number of observed *de novo* mutations of variant type t landing in gene g. The expected number of *de novos* to land in gene g when considering different variant types is

$$\mathbb{E}[K] \equiv \lambda' = \sum_t \mu_{g,t} \sum_{j=1}^{N} n_{j,t}$$

where $n_{j,t}$ denotes the total number of observed *de novo* variants of type t in an individual j. For each variant type t, we scale $\mu_{g,t}$ such that $\sum_g \mu_{g,t} = 1$. We compute $Pr(K) = Pois(K|\lambda')$.

Incorporation of s_{het} values

We incorporate s_{het} values, which estimate the selection against heterozygous protein-truncating variants per gene, as weights in a weighted false discovery rate (FDR) procedure. We sort all genes in ascending order by their s_{het} values. We then separate these sorted genes into 10 equally sized decile bins. For each gene g in each bin $b[g]$, we compute a weight w_g as

$$w_g = 10 \cdot \frac{|DD \in b[g]|}{|DD|}$$

where DD is the set of exclusively dominant disease-causing genes as annotated in OMIM (accessed December 2023). Genes without s_{het} values are assigned a weight $w_g = 1$. Note that $\mathbb{E}[w_g] = 1$ and that s_{het} values, which are constant for all variants within a given gene, are independent from y and y' values, which vary for variants within a gene based on variant mutational targets and deleteriousness scores. This enables us to perform Benjamini-Hochberg false discovery rate correction on weighted Q-values computed for each $Pr(y')$ as $Q = Pr(y')/w_g$.[22]

Analytical test for compound heterozygous cohort-level recurrence

A compound heterozygous configuration is an independent occurrence of two variants: one maternally (M) and the other paternally (D) inherited. The mutational target of a compound heterozygous configuration should therefore lie in a space of *squared* mutational targets. We define the mutational target of a compound heterozygous configuration as

$$\mu_{g,v_M,v_D} = \max(\mu_{g,v_M}, \mu_{g,v_D})^2 \qquad \text{(Equation 4)}$$

where v_M and v_D are maternally and paternally inherited variants comprising a compound heterozygous configuration, and μ_{g,v_M} and μ_{g,v_D} are computed as in Equation 1. To prioritize compound heterozygous configurations with both deleterious variants, we use the maximum over per-variant mutational targets. A single deleterious variant in a compound heterozygous configuration may indicate carrier status rather than a compelling candidate for a rare disorder. By this definition, μ_{g,v_M,v_D} is uniformly distributed at null (Supplementary Note S4). This enables us to define a similarly constructed statistic y^c modelable by the Irwin-Hall distribution as in the case of recurrent *de novos* (Equation 2):

$$y^c = \sum_{j=1}^{K} \left(1 - \frac{\mu_{g,v_{M,i},v_{D,i}}}{{\mu_g}^2}\right) \sim IH(y^c|K)$$

where K is the number of compound heterozygous configurations independently landing in gene g across the cohort, and $v_{M,j}$ and $v_{D,j}$ are the maternally and paternally inherited variants in gene g in individual j. As before, K is approximately Poisson distributed, and parameter λ^c, the expected number of compound heterozygous configurations to land in gene g, is given by

$$\mathbb{E}[K] \equiv \lambda^c = {\mu_g}^2 \sum_{j=1}^{N} n_{M,j} n_{D,j}$$

where $n_{M,j}$ and $n_{D,j}$ are the numbers of maternally and paternally inherited rare variants in an individual j, respectively. We compute $Pr(K) = Pois(K|\lambda^c)$ as before.

Finally, we extend this basic test statistic to accommodate 16 compound heterozygous configuration types as $(t_M, t_D) \in$ {coding SNV, coding indel, intronic SNV, intronic indel}^{2} and define $y^{c\prime}$ and Poisson parameter $\lambda^{c\prime}$ accordingly as

$$y^{c\prime} = \sum_{t_M,t_D} \sum_{i=1}^{K_{t_M,t_D}} \left(1 - \frac{\mu_{g,v_{M,i},t_M,v_{D,i},t_D}}{\max(\mu_{g,t_M}, \mu_{g,t_D})^2}\right)$$

and

$$\mathbb{E}[K] \equiv \lambda^{c'} = \sum_{t_M, t_D} \mu_{g,t_M} \mu_{g,t_D} \sum_{j=1}^{N} n_{M,t_M,j} n_{D,t_D,j}$$

where K_{t_M,t_D} is the number of compound heterozygous configurations in gene g across the cohort where the maternally inherited variant is of type t_M and the paternally inherited variant is of type t_D. Instances where $K_{t_M,t_D} = 0$ are excluded from the above sums. For each variant type t, we scale $\mu_{g,t}$ such that $\sum_g \mu_{g,t} = 1$. We compute the probability of $y^{c'}$ as in Equation 3. Note that homozygous recessive variants violate the assumptions of our approach and are excluded (Supplementary Note S4).

Modeling false positive diagnoses

For any gene where the observed number of variants $K > \mathbb{E}[K]$ across the cohort, we suspect that there are some true diagnoses in specific patients as well as some "false positives" where a randomly occurring variant in a patient is unrelated to the patient's condition. We use the binomial distribution parameterized by K independent trials and probability of success per trial $\mathbb{E}[K]/N$ to estimate the proportion of false positive diagnoses for each gene.

Analytical test for individual-level compound heterozygous configuration

Given a set of independent compound heterozygous configurations across genes in a single individual's genome, we construct a test for the hypothesis of a monogenic, recessive disorder caused by *one* of these compound heterozygous configurations against the null. We assume up to one compound heterozygous configuration per gene, i.e., for each gene g, $n_M n_D {\mu_g}^2 \ll 1$, where $\sum_g \mu_g = 1$ and n_M and n_D are the numbers of maternally and paternally inherited rare variants in this individual's genome.

We now rescale the mutational target of a compound heterozygous configuration (Equation 4) with respect to all genes G in the genome as

$$\tilde{\mu}_{g,v_M,v_D} = \frac{\sum_{i \in G} \min({\mu_i}^2, \mu_{g,v_M,v_D})}{\sum_{i \in G} {\mu_i}^2}.$$

Intuitively, this corresponds to the probability of observing a compound heterozygous configuration with an equal or smaller (i.e., more surprising) mutational target occurring in any gene in the genome. Thus, $\tilde{\mu} \sim U_{[0,1]}$. We precompute each gene's compound heterozygous mutational target ${\mu_i}^2$ for all genes in the genome in order to quickly compute $\tilde{\mu}$ values for each observed compound heterozygous configuration in an individual.

Next, we define our statistic $\tilde{y}^c$ per individual as the minimal observed rescaled compound heterozygous mutational target:

$$\tilde{y}^c = \min(\tilde{\mu}_1, ..., \tilde{\mu}_K). \qquad \text{(Equation 5)}$$

We compute the probability of observing a $\tilde{y}^c$ value this low or lower given K total genes with observed compound heterozygous configurations in an individual's genome as

$$Pr(\tilde{y}^c|K) = 1 - \prod_{i=1}^{K} Pr(Y > \tilde{y}^c) = 1 - Pr(Y > \tilde{y}^c)^K.$$

where Y is a dummy variable. Because $\tilde{y}^c$ is uniformly distributed on [0,1], $Pr(Y > \tilde{y}^c) = (1 - \tilde{y}^c)$, so we simplify this calculation as

$$Pr(\tilde{y}^c|K) = 1 - (1 - \tilde{y}^c)^K.$$

We also model the distribution of K observed compound heterozygous configurations across an individual's genome in order to compute the overall probability of our statistic $\tilde{y}^c$ using the same formulation as before (Equation 3). The distribution of K, given our prior assumption of at most one compound heterozygous configuration per gene, has an exact solution as the number of double events in a bivariate binomial distribution with correlation parameter ρ capturing the effect of different gene lengths on K. However, due to the complexity in calculations of the exact solution, here we use the Poisson approximation instead because, for each gene g, $\sum_g \mu_g = 1$ and $n_M n_D {\mu_g}^2 \ll 1$. The $\tilde{\lambda}^c$ parameter for the Poisson approximation in this case is

$$\mathbb{E}[K] \equiv \tilde{\lambda}^c = n_M n_D \sum_{i=1}^{G} {\mu_i}^2. \qquad \text{(Equation 6)}$$

Finally, we accommodate the 16 compound heterozygous configuration types as $(t_M, t_D) \in$ {coding SNV, coding indel, intronic SNV, intronic indel}^{2} and redefine $\tilde{\mu}'$, $\tilde{y}^{c'}$ and Poisson parameter $\tilde{\lambda}^{c'}$ accordingly as

$$\tilde{\mu}'_{g,v_M,v_D} = \frac{\sum_{t_M,t_D} \sum_{i \in G} \min(\mu_{i,t_M}\mu_{i,t_D}, \mu_{g,v_{M,t_M},v_{D,t_D}})}{\sum_{t_M,t_D} \sum_{i \in G} \mu_{i,t_M}\mu_{i,t_D}}$$

and

$$\tilde{y}^{c'} = \min(\tilde{\mu}'_1, ..., \tilde{\mu}'_K)$$

and

$$\mathbb{E}[K] \equiv \tilde{\lambda}^{c'} = \sum_{t_M,t_D} n_{M,t_M} n_{D,t_D} \sum_{i \in G} \mu_{i,t_M}\mu_{i,t_D} \quad .$$

Enrichment for correct diagnoses

Given a ranked list of genes across a cohort of patients, where each gene may be diagnostic for the given patient, we can compute enrichment for correct diagnoses at each gene rank. We use Fisher's exact test to compare the proportion of complete, certain diagnoses in all genes up to and including rank *k* compared to the proportion of correct diagnoses at genes ranked *k*+1 through the end of the list. We consider the minimum Fisher's exact test *P* across all *k* to be our overall enrichment. We assign a permutation-based *P*-value to this enrichment value by randomly permuting the initial gene list 10,000 times and recomputing the minimum Fisher's exact test *P* for each permuted list.

Pathway enrichment analysis

Phenotypically-similar patient groupings

Phrank was used to compute all-against-all pairwise phenotype similarity scores between all affected patients' sets of standardized HPO terms. We normalized these scores by dividing by the maximum self-similarity score in each pair.[23] UDN patients experience a spectrum of symptoms across overlapping biological categories and therefore cannot be easily separated into distinct, well-defined clusters (Supplementary Figure S13). We

iteratively grouped similar patient pairs using complete-linkage hierarchical clustering with the agnes function from R's cluster package, which allows for patient groups of different sizes while minimizing the maximum distance between any two patients in the same cluster. We assigned patients to clusters by cutting the resulting dendrogram at height=3.5, resulting in 120 clusters of 2–97 patients per cluster (mean=22, median=17).

Selecting genes per patient cluster

We identify genes per patient cluster as follows. First, we consider known diagnoses for all UDN patients in that cluster. For patients with a diagnosis that was "complete" (i.e., explained all symptoms including asserted phenotypes), no further genes are considered. For patients with no diagnosis or at most one "partial" diagnosis, we then consider genes with an exonic (or intronic with a SpliceAI score >0.4) *de novo* variant and assign each gene its variant's severity weight (s) from our modified DeNovoWEST procedure. Recall that weights are assigned per variant class based on functional impact (e.g., frameshift, nonsense, missense), variant deleteriousness, and gene constraint. Autosomal *de novos* are considered as before in addition to *de novos* on chromosome X with gnomAD population maximum allele frequency < 0.0001, TOPMed allele frequency < 0.0001, proband alternate allele read depth >20 and frequency >0.2 (for females) and alternate read depth >20 in both parents. We consider the most significant gene per patient with $Pr(s) < 0.0005$, where $Pr(s) = Pr(S \geq s \mid K{=}1)Pr(K{=}1)$. $Pr(S \geq s \mid K{=}1)$ is computed exactly using precomputed per-variant Roulette mutation rates and variant weights per gene. $Pr(K{=}1)$ is computed assuming mutations follow a Poisson distribution with $\lambda{=}\mu_{gene}$ for genes falling on chromosome X in males and $\lambda{=}2\mu_{gene}$ otherwise. Finally, in patients who still have fewer than two genes at this point and no complete diagnoses, we include up to one additional gene harboring a compound heterozygous variant pair that ranked in the top 100 in our cohort-wide per-individual analysis, as there was significant enrichment for correct diagnoses in this set (Fisher's exact test p-value = 5.23e-4). Across all patient clusters, we considered 70 genes (6.15%) with *de novo* variants corresponding to known diagnoses, 10 genes (0.88%) with compound heterozygous variants corresponding to known diagnoses, 562 (49.38%) other known diagnostic genes, 434 genes (38.14%) with new *de novo* candidates, and 62 genes (5.45%) with new compound heterozygous candidates.

Gene Set Enrichment Analysis (GSEA)

The genes found across all patients in each patient cluster were used as a query set for gene set enrichment analysis (GSEA) using g:Profiler.[24] We considered Reactome and KEGG biological pathway gene sets of size <150 genes and set our background gene set to all human genes annotated in Ensembl. Enrichment p-values are adjusted using g:Profiler's g:SCS approach.[25] Briefly, for every query gene set size, 2000 random gene sets of the same size are used as queries for GSEA with the same parameters, and the lowest pathway

enrichment p-value is recorded for each random query set. A threshold t is selected for each query gene set size as the 5% quantile of these random minimum p-values. Enrichment p-values resulting from the true gene query are then adjusted by multiplying by $0.05/t$.

References

1. Freed, D., Aldana, R., Weber, J. A. & Edwards, J. S. The Sentieon Genomics Tools - A fast and accurate solution to variant calling from next-generation sequence data. *bioRxiv* (2017) doi:10.1101/115717.
2. Pedersen, B. S. *et al.* Somalier: rapid relatedness estimation for cancer and germline studies using efficient genome sketches. *Genome Med.* **12**, 62 (2020).
3. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
4. Pedersen, B. S. *et al.* Effective variant filtering and expected candidate variant yield in studies of rare human disease. *NPJ Genom Med* **6**, 60 (2021).
5. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
6. Haghighi, A. *et al.* An integrated clinical program and crowdsourcing strategy for genomic sequencing and Mendelian disease gene discovery. *NPJ Genom Med* **3**, 21 (2018).
7. 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).
8. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).
9. Harris, P. A. *et al.* The REDCap consortium: Building an international community of software platform partners. *J. Biomed. Inform.* **95**, 103208 (2019).
10. Rosenberg, A. B., Patwardhan, R. P., Shendure, J. & Seelig, G. Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell* **163**, 698–711 (2015).
11. Soemedi, R. *et al.* Pathogenic variants that alter protein code often disrupt splicing. *Nat. Genet.* **49**, 848–855 (2017).
12. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
13. Mohanty, A. K. *et al.* novoCaller: a Bayesian network approach for de novo variant calling from pedigree and population sequence data. *Bioinformatics* **35**, 1174–1180 (2019).
14. Yang, X. *et al.* Developmental and temporal characteristics of clonal sperm mosaicism. *Cell* **184**, 4772–4783.e15 (2021).

15. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535–548.e24 (2019).
16. Seplyarskiy, V. *et al.* A mutation rate model at the basepair resolution identifies the mutagenic effect of Polymerase III transcription. *bioRxiv* 2022.08.20.504670 (2023) doi:10.1101/2022.08.20.504670.
17. Cassa, C. A. *et al.* Estimating the selective effects of heterozygous protein-truncating variants from human exome data. *Nat. Genet.* **49**, 806–810 (2017).
18. Heard, N. A. & Rubin-Delanchy, P. Choosing between methods of combining |p|-values. *Biometrika* **105**, 239–246 (2018).
19. Liu, Y. & Xie, J. Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *J. Am. Stat. Assoc.* **115**, 393–402 (2020).
20. Niehaus, A. *et al.* A survey assessing adoption of the ACMG-AMP guidelines for interpreting sequence variants and identification of areas for continued improvement. *Genetics in medicine*: *official journal of the American College of Medical Genetics* vol. 21 1699–1701 (2019).
21. Bethune, J., Kleppe, A. & Besenbacher, S. A method to build extended sequence context models of point mutations and indels. *Nat. Commun.* **13**, 7884 (2022).
22. Genovese, C. R., Roeder, K. & Wasserman, L. False Discovery Control with p-Value Weighting. *Biometrika* **93**, 509–524 (2006).
23. Jagadeesh, K. A. *et al.* Phrank measures phenotype sets similarity to greatly improve Mendelian diagnostic disease prioritization. *Genet. Med.* **21**, 464–470 (2019).
24. Raudvere, U. *et al.* g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **47**, W191–W198 (2019).
25. Reimand, J., Kull, M., Peterson, H., Hansen, J. & Vilo, J. g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.* **35**, W193–200 (2007).

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], Daniel Traviglia[1], Xinyun Li[3], Danielle Barnum[4], Alexander Veit[1], Julian Willett[5], Michele Berselli[1], William Ronchetti[1], Richard Sherwood[6], Joel Krier[7], Isaac S. Kohane[1], 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 Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT
4 Access to Medicine Foundation, Amsterdam, The Netherlands
5 Department of Neurology, Massachusetts General Hospital, Boston, MA
6 Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA
7 Department of Genetics, Atrius Health, Boston, MA

Figures.....3
Figure S1. UDN cohort pairwise relatedness and sex quality control checks.....3
Figure S2. Hierarchical decision trees for clinically assessing candidates.....4
Figure S3. Massively Parallel Splicing Reporter Assay (MPSA) design.....5
Figure S4. MPSA validation rates by SpliceAI scores and across replicates.....6
Figure S5. Correlation between CADD and SpliceAI scores in intronic regions.....7
Figure S6. NovoCaller selects for confident de novo variants.....7
Figure S7. DeNovoWEST modifications and variant distributions used to calculate weights.....8
Figure S8. DeNovoWEST results are enriched for correct diagnoses.....9
Figure S9. QQ-plots for closed-form de novo and compound heterozygous statistics....10
Figure S10. Comparison of DeNovoWEST and de novo recurrence method.....11
Figure S11. Schematic of cohort-level compound heterozygous variant recurrence.....12
Figure S12. MED11 transcriptome sequencing.....13
Figure S13. UDN patients do not fall into well-defined phenotypic clusters.....13
Tables.....13

Table S1. Candidate vs. Decoy Genes and Final Clinical Scores....14
Table S2. De novo recurrence findings.... 15
Table S3. Individual-level compound heterozygous findings....15
Table S4. All enriched biological pathways across patient clusters....16
Note S1. Subdividing neurological symptom category.... 16
Note S2. Clinical evaluation protocol....16
Preparation: Collecting information prior to evaluating diagnostic fit of candidate....16
Information collected about the affected patient:.... 16
Information collected about each variant:....17
Tip for planning for literature review....17
Evaluation of variant–patient phenotypic fit by category....18
Known disease associations.... 18
Model organism studies.... 19
Expression of gene in phenotypically-relevant tissues....20
Protein molecular function....20
Gene evolutionary constraint.... 21
Variant frequency, pathogenicity, and position....21
Summarizing information for final phenotypic fit score.... 22
Note S3. Compound heterozygous mutational target is uniformly distributed.... 23
Note S4. Independence assumption is violated in homozygous recessive case....24
Note S5. Undiagnosed Diseases Network Consortium Members.... 25
References.... 27

Figures

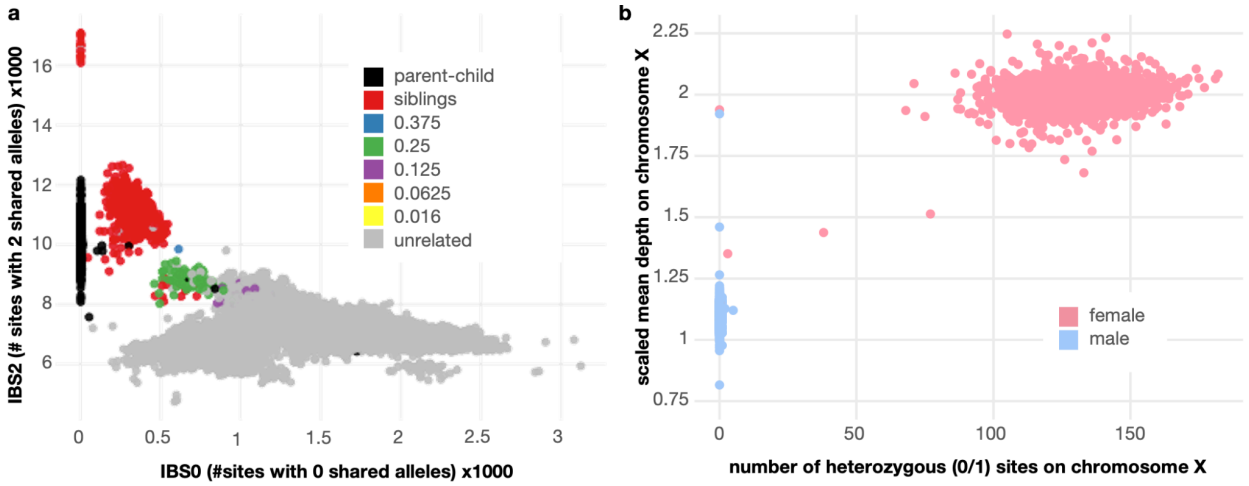

Figure S1. UDN cohort pairwise relatedness and sex quality control checks.
(a) Between-sample pairwise relatedness computed using Somalier, which operates on a predefined, informative subset of genomic positions.[1] Unrelated individuals are expected to have a nonzero number of genomic positions with 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.

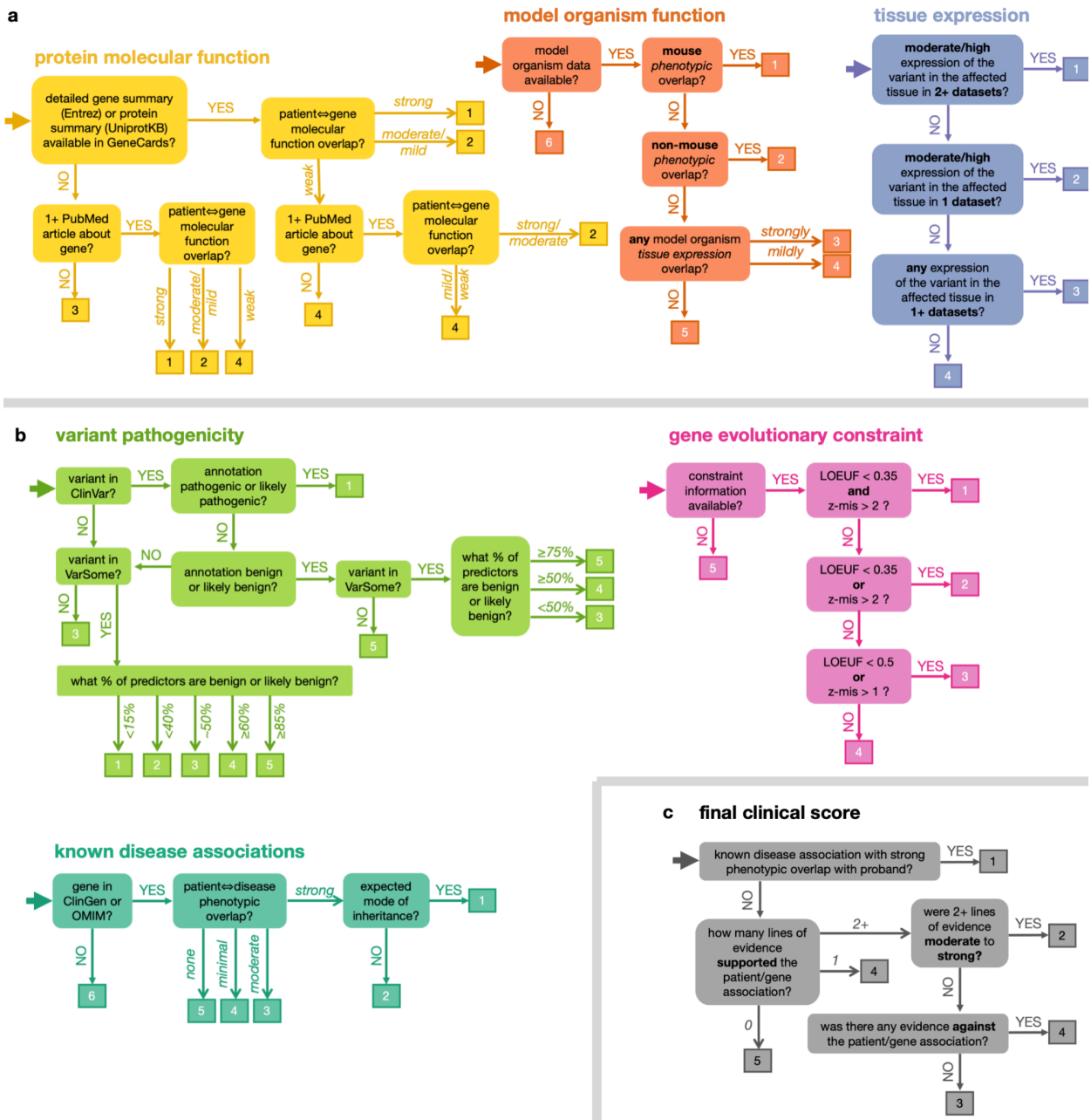

Figure S2. Hierarchical decision trees for clinically assessing candidates.
(a) Three categories of evidence used in the clinical evaluation protocol for assessing gene-disease fit for each participant phenotype that were not used in any statistical candidate prioritization methods. **(b)** Three categories 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.

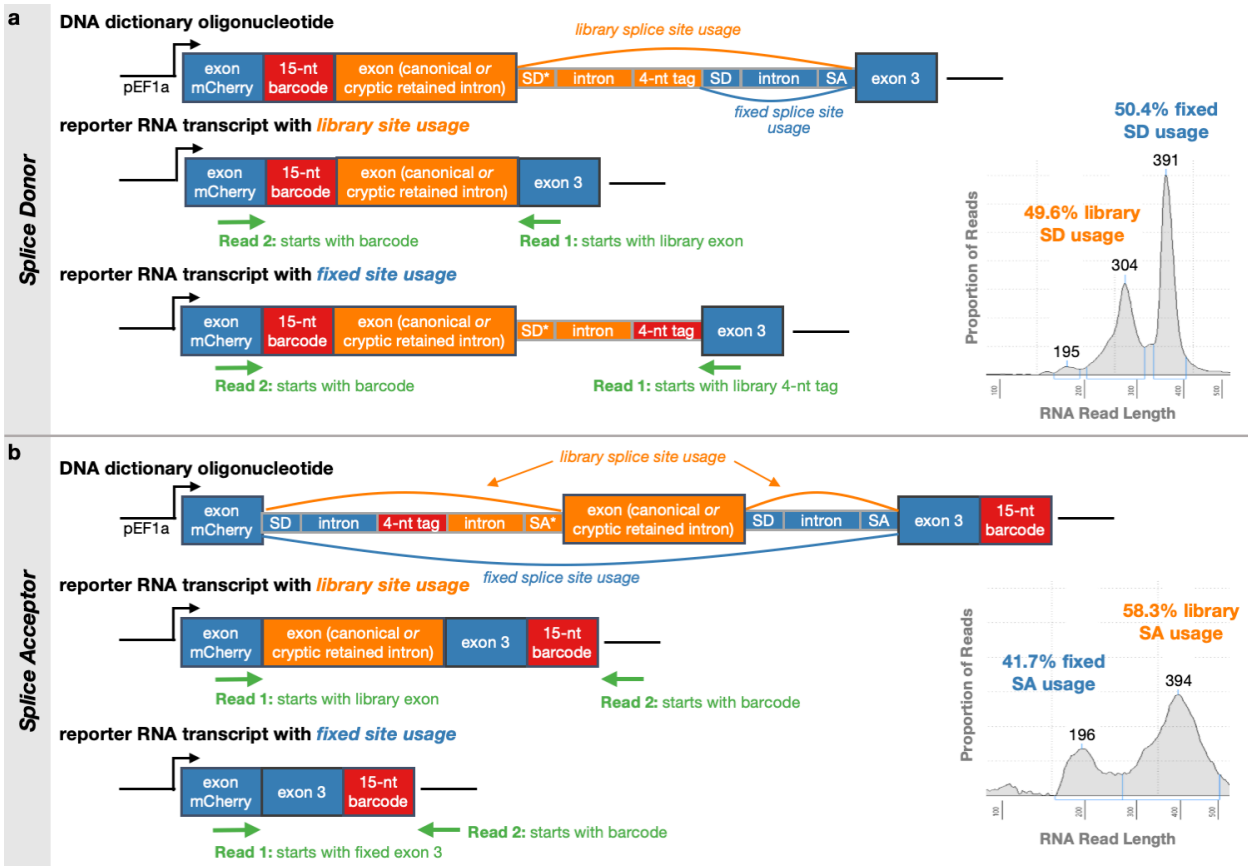

Figure S3. Massively Parallel Splicing Reporter Assay (MPSA) design
Oligonucleotides containing a 155-nucleotide section of genomic DNA from a UDN patient, called a 'library' (orange) were inserted into lentiviral plasmids (pEF1a). Exons are represented as tall boxes and introns as short boxes in the 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.

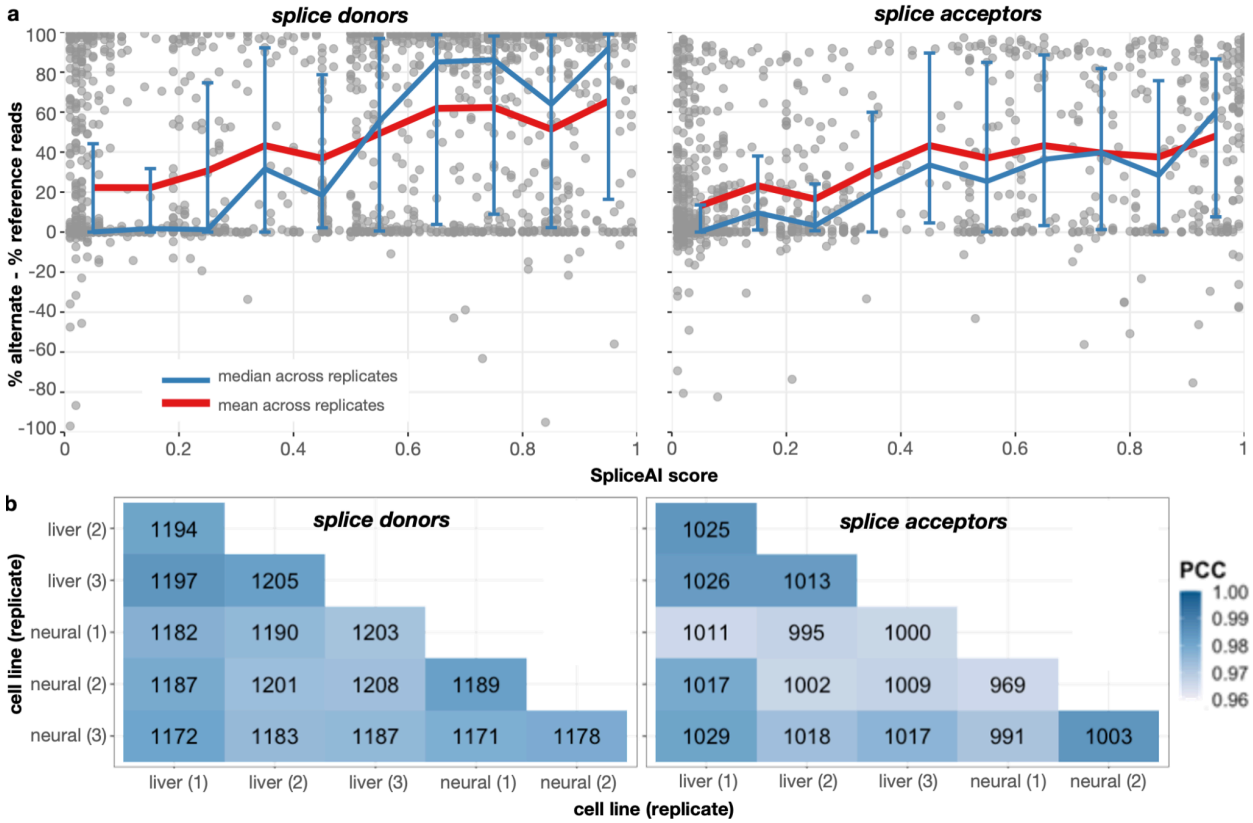

Figure S4. MPSA validation rates by SpliceAI scores and across replicates.
A "validation rate" is computed for every variant tested in the MPSA as described in Methods. **(a)** Each point represents the MPSA validation rate for a specific variant that was tested. Blue lines correspond to the median validation 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.

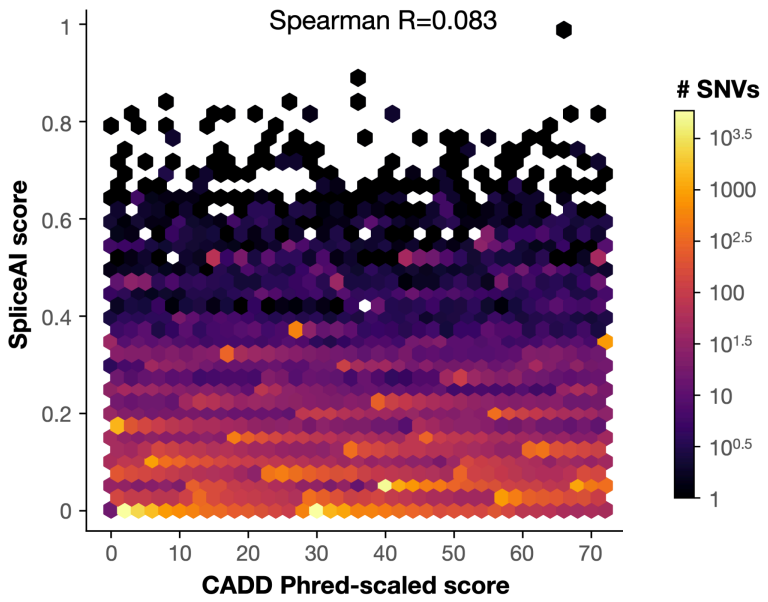

Figure S5. Correlation between CADD and SpliceAI scores in intronic regions.
We assessed the correlation between precomputed *in silico* CADD deleteriousness and SpliceAI functionality scores within all intronic regions between protein-coding exons on chromosome 21. We observed poor correlation between these scores, even though CADD uses SpliceAI as an input feature, supporting our distinction between these score types in introns.

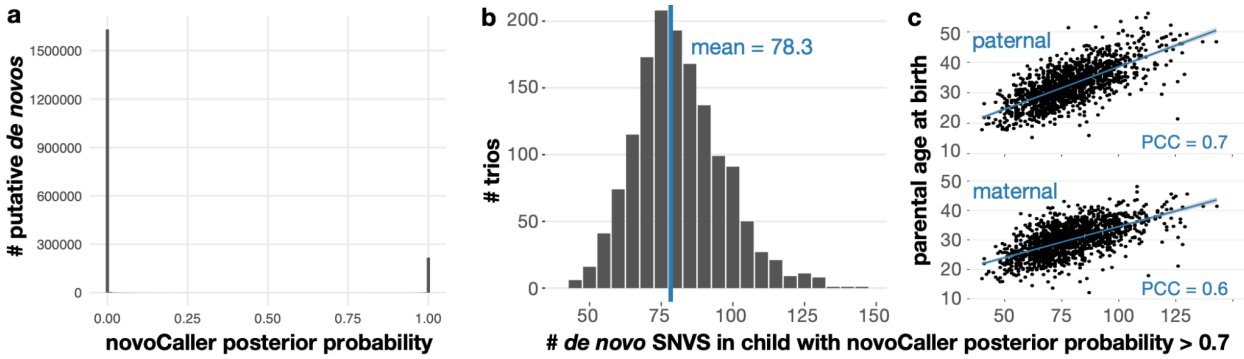

Figure S6. NovoCaller selects for confident *de novo* variants.
Putative *de novo* variants were selected in each trio as described in Methods. **(a)** Posterior probabilities assigned by novoCaller are bimodal; 40.4% of putative *de novos* were assigned a posterior probability of 0 and are excluded 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.

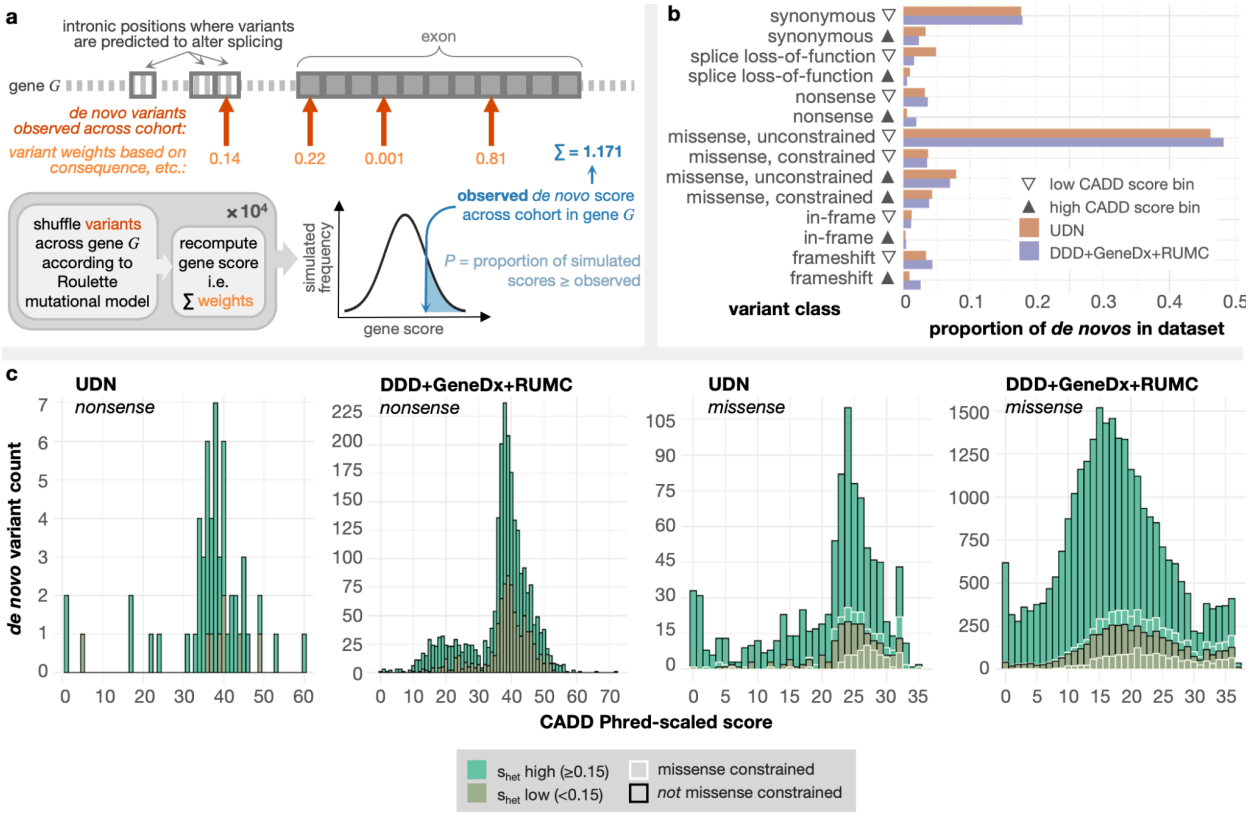

Figure S7. DeNovoWEST modifications and variant distributions used to calculate weights.

(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.

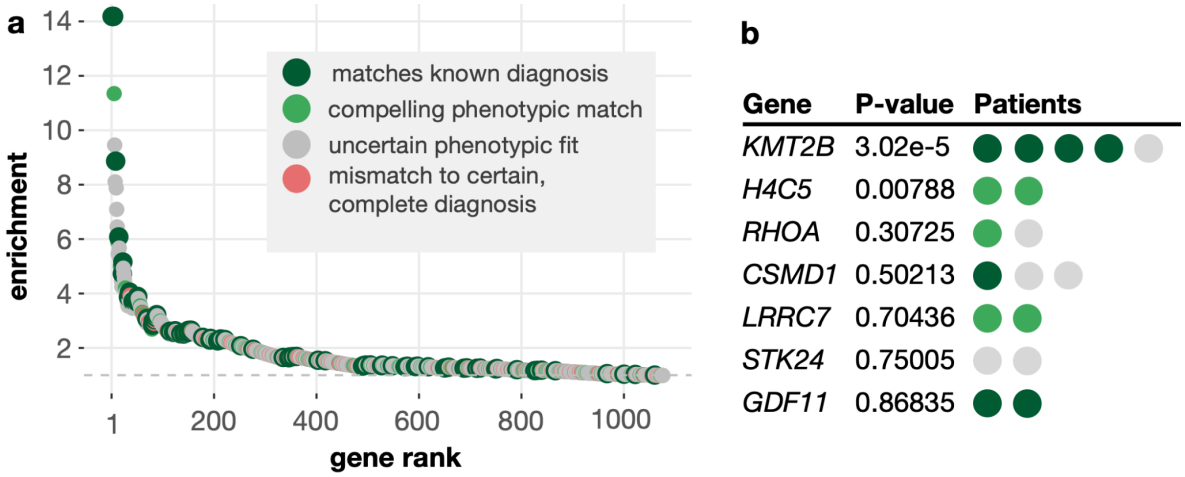

Figure S8. DeNovoWEST results are enriched for correct diagnoses.
Genes with observed *de novo* variants across the UDN cohort were evaluated by DeNovoWEST for significant *de novo* recurrence. **(a)** Enrichment for correct diagnoses (green) is calculated as the proportion of correct diagnoses across the top *k* genes divided by the total proportion of correct diagnoses across all genes with observed *de novo* variants. Light green corresponds to an annotated candidate gene. **(b)** Top genes found by DeNovoWEST with Bonferroni-corrected *P*-value. Here, light green corresponds to either a previously annotated candidate gene or a gene that our clinical team found to be a compelling phenotypic match.

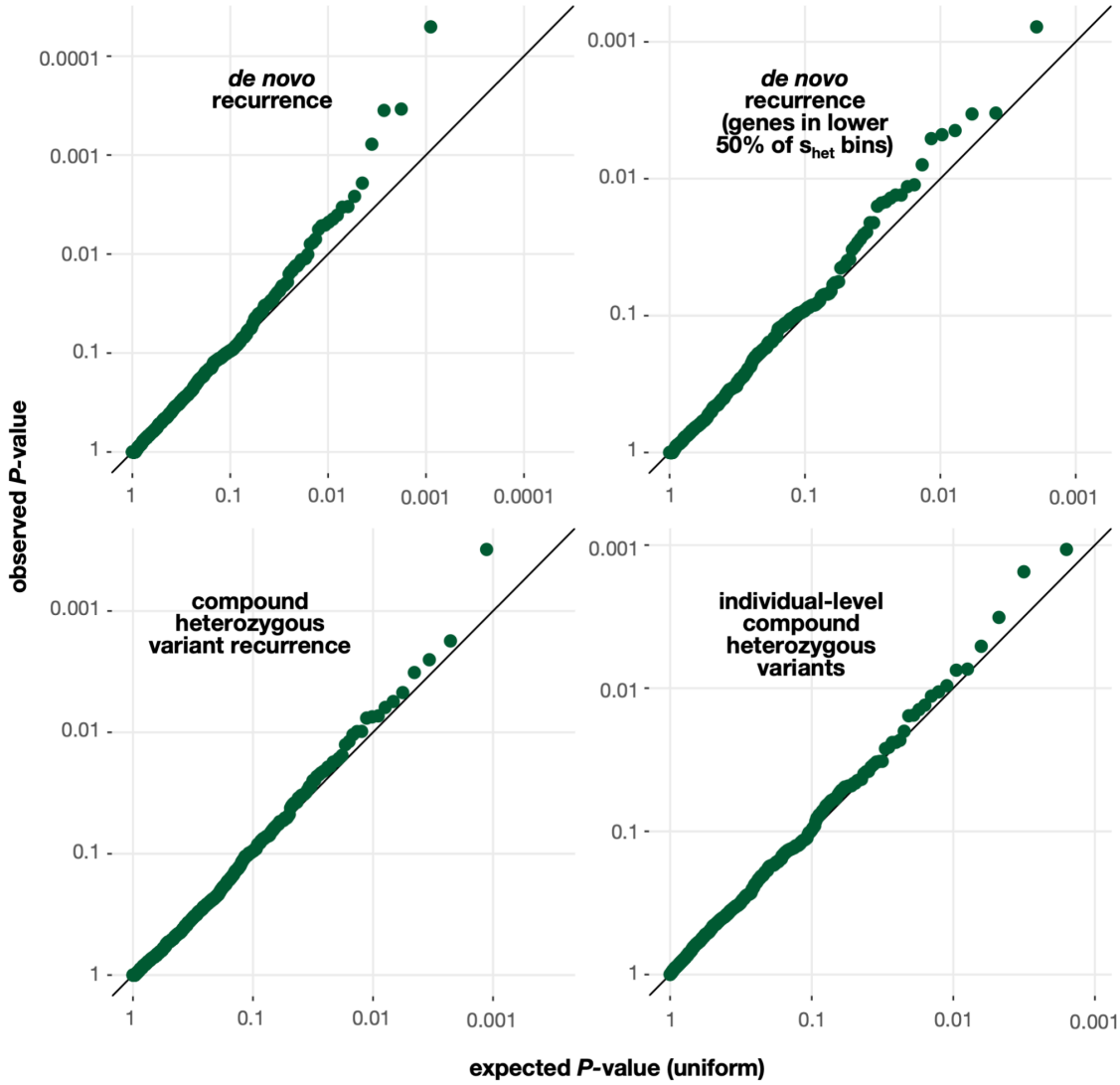

Figure S9. QQ-plots for closed-form *de novo* and compound heterozygous statistics. Top row depicts QQ-plots for *de novo* recurrence. The full curve is shown on the left and the subset of genes falling into the lower half of s_{het} bins are shown on the right. The bottom row depicts QQ-plots for compound 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 $Pr(y|K\geq1) = Pr(y)/Pr(K\geq1)$.

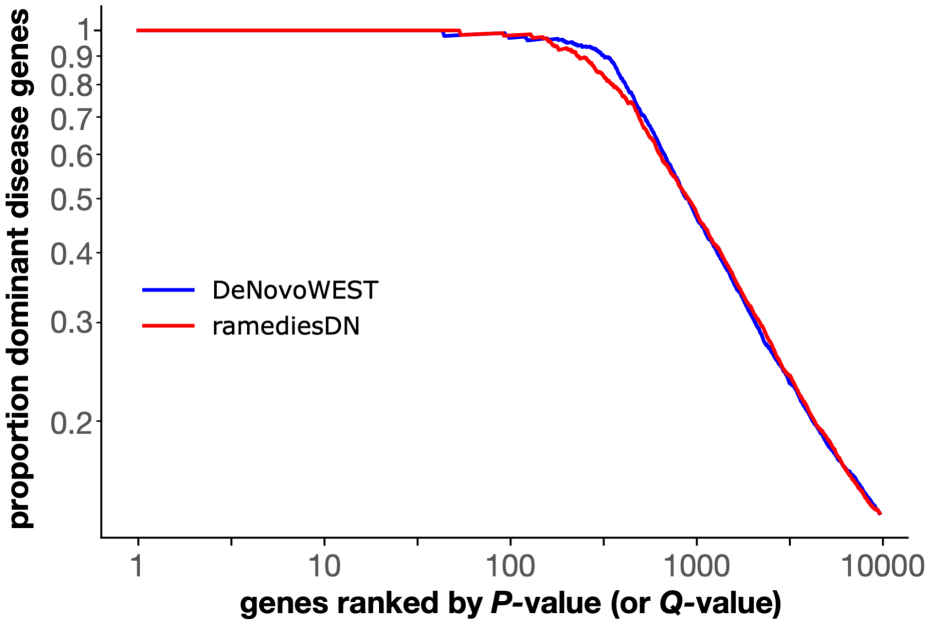

Figure S10. Comparison of DeNovoWEST and *de novo* recurrence method
Curves depict the ability of the two *de novo* recurrence methods to recover known autosomal dominant disease genes (reported in OMIM) from the DDD+GeneDx+RUMC dataset. Around 4000 putative *de novos* from the DDD+GeneDx+RUMC were removed 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. Both axes are log-scaled for visual clarity.

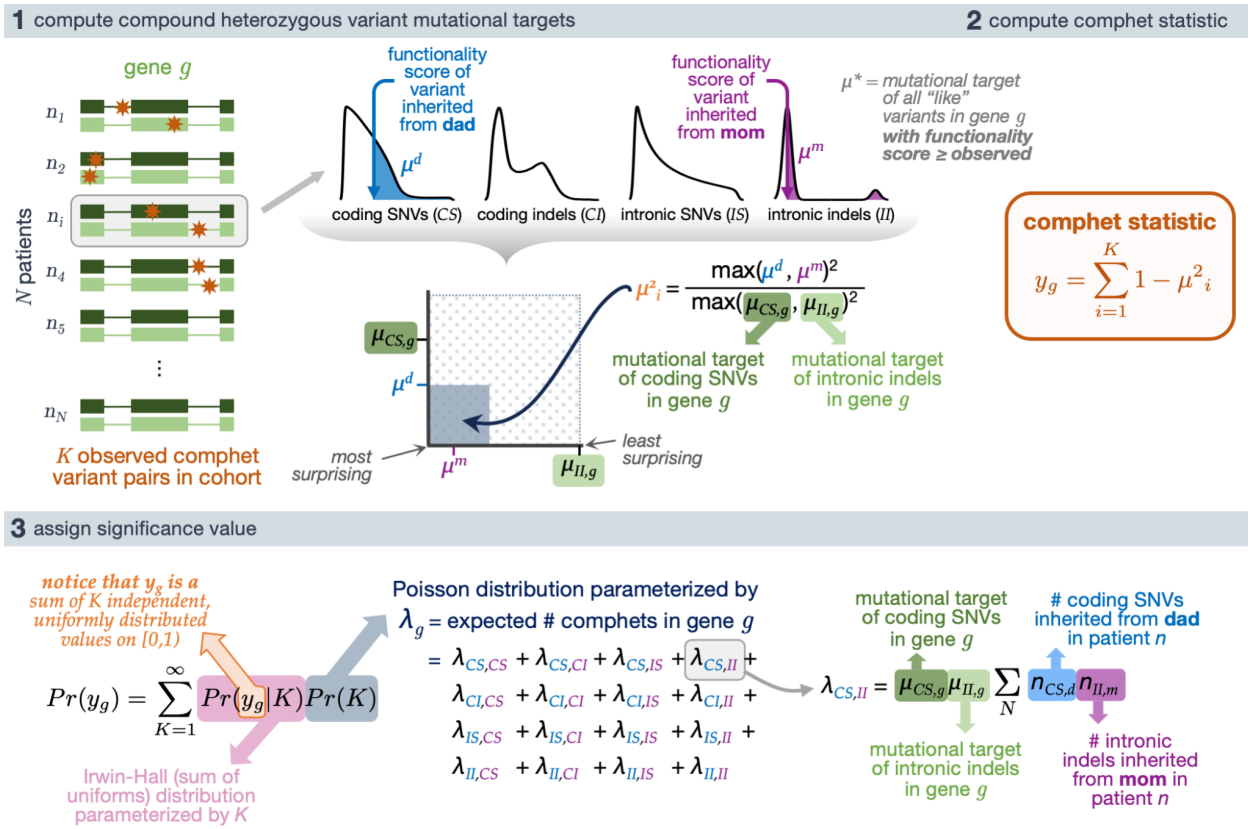

Figure S11. Schematic of cohort-level compound heterozygous variant recurrence
Schematic of analytical test for the recurrence of compound heterozygous variant pairs (comphets) that considers distal splice-altering and exonic SNV and indel variants, their variant functionality scores, and a genome-wide mutation rate model Roulette. "Like" variants refer to those of the same variant class (i.e., *CS*, *CI*, *IS*, *II*) and within the same functionality score and minor allele frequency thresholds.

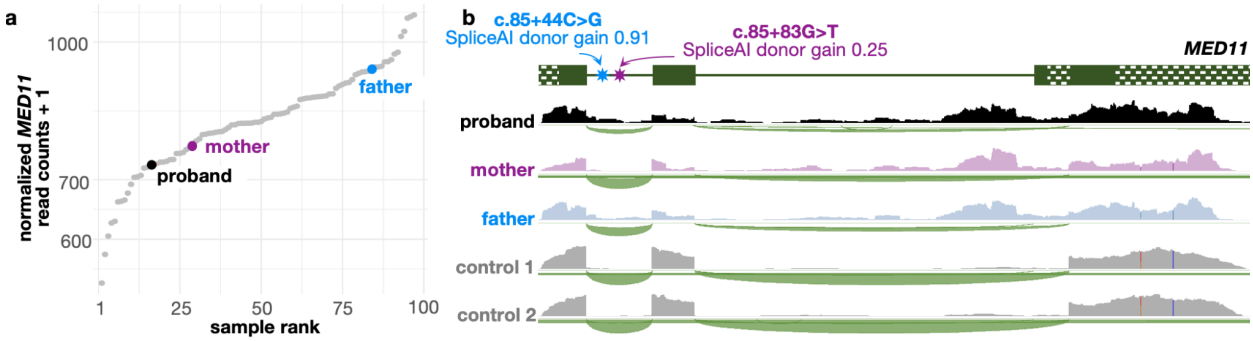

Figure S12. *MED11* transcriptome sequencing.
(a) Overall gene expression of *MED11* in blood samples from affected proband (black), mother (purple) and father (blue) are within the normal range of expression when compared to 50 sequenced control samples (gray) and would not have been flagged as differentially expressed. **(b)** RNA-sequencing read depth over the full *MED11* gene region for proband, mother, father, and two control samples.

Figure S13. 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.

Tables

Table S1. Candidate vs. Decoy Genes and Final Clinical Scores

Clinical team applying protocol to assign final clinical scores (descending values from most to least compelling) was blinded to gene labels (i.e., decoy or candidate) and *in silico* variant pathogenicity predictions.

gene	patient phenotype	score	label
AHR	global brain atrophy, history of pituitary gland inflammation (causing hormone deficiency), hypothalamus-related obesity, reduced bone density, history lung blood clots (treated with Eliquis)	4	decoy
MYO18B	suspected neurodegeneration with brain iron accumulation	5	decoy
LTBP4	chronic face/lip swelling, soft tissue thickening (on MRI), inflamed lips with lumps or nodules, tongue with cracks or grooves, high blood pressure, heart mitral valve prolapse	4	decoy
LRPPRC	adult-onset sudden blurry vision, history of demyelinating disease (white matter, gray matter, cerebellum), double vision, lack of coordination and balance (ataxia), tremor, difficulty swallowing (dysphagia), peripheral sensory neuropathy, motor speech disorder (dysarthria), heart atrial fibrillation, lower extremity weakness (proximal > distal and left > right)	4	decoy
PNPT1	connective tissue problems (cutis laxa), pouch-like protrusions (diverticula) in gastrointestinal tract from esophagus to colon, diverticula in bladder, left abdominal hernia (repaired 3x), weakened heart muscles not due to blocked arteries (non-ischemic cardiomyopathy), thickened heart mitral valve with calcium deposits, normal heart aortic valve	3	decoy
DYSF	severely high blood pressure (refractory hypertension), vision (ophthalmologic) disturbances, inability to eat solid foods, asymmetric tremor, chronic rash, connective tissue problems (cutis laxa), skin thickening (hyperkeratosis), mild heart aortic dilation, narrowing of renal arteries, low level of white blood cells (lymphopenia)	5	decoy
OTOGL	autoimmune skin condition (discoid lupus), rheumatoid arthritis, weakened heart with cardioverter-defibrillator device implanted, cerebellar stroke with severe tremor, anemia, low platelets, abnormal protrusion of shoulder blades, scoliosis requiring surgical intervention, progressive muscle loss (cachexia), severe muscle pain, fatigue with lack of endurance	4	decoy
CEP164	adolescent-onset stiffness, weakness and cramping after exercise lasting up to a week accompanied by dark urine	5	decoy
CLCN1	reduced muscle tone, failure to thrive (improved with gastrostomy tube), gastrointestinal issues (impaired colon movement, extra tissue pushing anus forward, misaligned intestines), lung issues (high blood pressure in arteries leading to lungs, widening of bronchial airways, abnormalities in interstitial tissue)	3	decoy
PRG4	enlarged lymph nodes in abdomen with small clusters of immune cells (non-caseating granulomas), elevated white blood cells (eosinophils) in peripheral blood, inflammation around portal veins in the liver	5	decoy

KCNH8	feet turning inward (progressively worsening), knee contractures, normal upper extremities, streak-like decreased skin pigmentation with rough patches on left side of neck and arm, delayed muscle relaxation after contraction, decreased reflexes	2	candidate
NEK8	refractory epilepsy, movement disorder (dystonia, choreoathetosis), cortical visual impairment, intellectual disability, concern for mitochondrial dysfunction, adrenal insufficiency	5	candidate
FAN1	Turner Syndrome, high blood pressure with unknown cause, generalized epilepsy, central auditory processing disorder, intellectual disability, concern for regression.	3	candidate
ITPR3	global developmental delay, impaired coordination and balance (ataxia), abnormally large head, autism, cerebral palsy	2	candidate
SPTBN5	global developmental delay, absent speech, low muscle tone, failure to thrive, significantly reduced reflexes	2	candidate
PPP6R1	severe adult-onset sensory polyneuropathy beginning with numbness and tingling in feet and hands, dizziness with vertigo, diminished and distorted sense of taste	3	candidate
UNC5C	slowly progressive spinal cord dysfunction (myelopathy), compressed nerve roots (radiculopathy), peripheral nerve damage (neuropathy)	1	candidate
TPP2	elevated inflammatory markers, elevated IgM and IgG, neuropathy, chronic pain, cardiomyopathy, cognitive impairment, dysmorphic features, short stature, excessive growth of coarse, dark hair in patterns typical in males (hirsutism), abnormally small red blood cells (microcytic anemia)	2	candidate
WDR96	impaired coordination and balance (ataxia), speech disorder due to weak speech production muscles (dysarthria), difficulty swallowing (dysphagia), progressive cerebellar volume loss shown on MRI	5	candidate
COL3A1	adult-onset oral and genital ulcers, extensive skin rash with itchy, small reddish-brown bumps lasting over a year	4	candidate

Table S2. *De novo* recurrence findings

We computed our *de novo* recurrence statistic across confident *de novo* variants from 872 affected individuals as described in Methods. Columns (from left to right) correspond to: gene name (HGNC ID), Ensembl gene ID, Q-value (asterisk indicates FDR < 0.15), patient diagnosis status (correct match, strong candidate, or blank), unadjusted patient–gene HPO term similarity score (computed with Phrank), patient sex, variant consequence, variant functionality score(s).

Table S3. Individual-level compound heterozygous findings
We computed our individual-level compound heterozygous statistic across rare inherited variants from 854 affected individuals as described in Methods. Columns (from left to right) correspond to: unadjusted p-value, gene name (HGNC ID), Ensembl gene ID, patient diagnosis status, unadjusted patient–gene HPO term similarity score (computed with Phrank), patient sex, maternal variant consequence (if known), maternal variant functionality scores (if known), paternal variant consequence (if known), paternal variant functionality scores (if known) .

Table S4. All enriched biological pathways across patient clusters
We performed gene set enrichment analysis using Reactome and KEGG biological pathways as described in Methods. Columns (from left to right) correspond to: pathway ID (asterisk indicates a patient–gene pair that also appears earlier in the list), pathway name, pathway size, GSEA adjusted p-value, cluster ID and effective gene set query size, gene name (HGNC ID), Ensembl gene ID, patient diagnosis status, unadjusted patient–gene HPO term similarity score (computed with Phrank), patient sex, variant consequence (if known), variant functionality scores (if known). Note that the effective gene set query size can be different for the same cluster ID between Reactome and KEGG enriched pathways. The original gene set query for each patient cluster is intersected with the set of annotated genes across all Reactome or all KEGG pathways respectively; some genes may only be present in one of the two data sources.

Notes

Note S1. 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). We created a text corpus for each cluster from the concatenated list of patients' HPO terms within each cluster using the R tm package and visualized these text corpuses as comparative word clouds using the R quanteda package. Subgroup names were determined based on manual inspection and comparison of the word clouds.

Note S2. 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 (ClinGen[3]) diseases & expert panel, Gene Curation Coalition (GenCC[4]), and Genomics England PanelApp[5]
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 OMIM[6] (obtained via manual review of OMIM pages)
6. Associated traits in gene from GWAS catalog[7] / 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)*

Do any known disease associations overlap with proband's key phenotypic features?	
1	ClinGen curated disease association with correct inheritance mode (i.e., AD for *de novos* 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)

Does tissue-level expression in model organisms or null models suggest correlation with proband phenotypes?	
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 Map[11]
2. The Genotype-Tissue Expression (GTEx) Portal[12] – search for gene, scroll down to "Bulk tissue gene expression" to see relevant tissues

Does available tissue-level expression data align with proband's key phenotypic features?	
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 GeneCards[14])
3. Published reports of gene/phenotype associations from OMIM[6,14]

Does the protein function relate to the potential model of molecular pathophysiology?	
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).

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 DECIPHER[18] 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)

Does the variant seem pathogenic?	
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)

Summary Classification	
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

Note S3. 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.

$$Pr(\mu_{g,v_M,v_D}) = Pr(\mu_{g,v_M,v_D} \geq \max(\mu_1, \mu_2)^2)$$

$$= Pr(\mu_{g,v_M,v_D} \geq \max(\mu_1^2, \mu_2^2))$$

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

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

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

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

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

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 1, 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

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

Note S4. Independence assumption is violated in homozygous recessive case

Our individual-level compound heterozygous configuration statistic $\tilde{y}^c$ (Equation 5, Methods) and the expected number of genes harboring compound heterozygous configurations per individual genome $\tilde{\lambda}^c$ (Equation 6, Methods) both assume that the two inherited variants comprising a compound heterozygous configuration arose as a result of independent mutations. In the case of homozygous recessive biallelic variants, however, where the exact same variant is inherited from each parent, this independence assumption is violated because in most cases a single mutational event resulted in the presence of a variant in both parents.

Note S5. Undiagnosed Diseases Network Consortium Members

Maria T. Acosta[1], David R. Adams[1], Ben Afzali[1], Ali Al-Beshri[2], Eric Allenspach[3], Aimee Allworth[3], Raquel L. Alvarez[4], Justin Alvey[5], Ashley Andrews[5], Euan A. Ashley[6], Carlos A. Bacino[7], Guney Bademci[8], Ashok Balasubramanyam[7], Dustin Baldridge[9], Jim Bale[5], Michael Bamshad[3], Deborah Barbouth[8], Pinar Bayrak-Toydemir[5], Anita Beck[3], Alan H. Beggs[10], Edward Behrens[11], Gill Bejerano[4], Hugo J. Bellen[12], Jimmy Bennett[3], Jonathan A. Bernstein[4], Gerard T. Berry[10], Anna Bican[13], Stephanie Bivona[8], Elizabeth Blue[3], John Bohnsack[5], Devon Bonner[4], Nicholas Borja[8], Lorenzo Botto[5], Lauren C. Briere[10], Elizabeth A. Burke[1], Lindsay C. Burrage[7], Manish J. Butte[14], Peter Byers[3], William E. Byrd[2], Kaitlin Callaway[2], John Carey[5], George Carvalho[14], Thomas Cassini[13], Sirisak Chanprasert[3], Hsiao-Tuan Chao[7], Ivan Chinn[7], Gary D. Clark[7], Terra R. Coakley[4], Laurel A. Cobban[10], Joy D. Cogan[13], Matthew Coggins[10], F. Sessions Cole[15], Brian Corner[13], Rosario I. Corona[14], William J. Craigen[7], Andrew B. Crouse[2], Vishnu Cuddapah[11], Precilla D'Souza[1], Hongzheng Dai[7], Kahlen Darr[16], Surendra Dasari[16], Joie

Davis[1], Margaret Delgado[1], Esteban C. Dell'Angelica[14], Katrina Dipple[3], Daniel Doherty[3], Naghmeh Dorrani[14], Jessica Douglas[10], Emilie D. Douine[14], Dawn Earl[3], Lisa T. Emrick[7], Christine M. Eng[17], Cecilia Esteves[18], Kimberly Ezell[13], Elizabeth L. Fieg[10], Paul G. Fisher[4], Brent L. Fogel[14], Jiayu Fu[1], William A. Gahl[1], Rebecca Ganetzky[11], Emily Glanton[18], Ian Glass[3], Page C. Goddard[4], Joanna M. Gonzalez[8], Andrea Gropman[1], Meghan C. Halley[4], Rizwan Hamid[13], Neil Hanchard[1], Kelly Hassey[11], Nichole Hayes[9], Frances High[10], Anne Hing[3], Fuki M. Hisama[3], Ingrid A. Holm[10], Jason Hom[4], Martha Horike-Pyne[3], Alden Huang[14], Yan Huang[1], Anna Hurst[2], Wendy Introne[1], Gail P. Jarvik[3], Suman Jayadev[3], Orpa Jean-Marie[1], Vaidehi Jobanputra[19], Oguz Kanca[12], Yigit Karasozen[14], Shamika Ketkar[7], Dana Kiley[9], Gonench Kilich[11], Eric Klee[16], Shilpa N. Kobren[18], Isaac S. Kohane[18], Jennefer N. Kohler[4], Bruce Korf[2], Susan Korrick[10], Deborah Krakow[14], Elijah Kravets[4], Seema R. Lalani[7], Christina Lam[3], Brendan C. Lanpher[16], Ian R. Lanza[16], Kumarie Latchman[8], Kimberly LeBlanc[18], Brendan H. Lee[7], Kathleen A. Leppig[3], Richard A. Lewis[7], Pengfei Liu[17], Nicola Longo[5], Joseph Loscalzo[10], Richard L. Maas[10], Calum A. MacRae[10], Ellen F. Macnamara[1], Valerie V. Maduro1, AudreyStephannie C. Maghiro[18], Rachel Mahoney[18], May Christine V. Malicdan[1], Rong Mao[5], Ronit Marom[7], Gabor Marth[20], Beth A. Martin[4], Martin G. Martin[14], Julian A. Martínez-Agosto[14], Shruti Marwaha[4], Allyn McConkie-Rosell[21], Ashley McMinn[13], Matthew Might[2], Mohamad Mikati[21], Danny Miller[3], Ghayda Mirzaa[3], Breanna Mitchell[16], Paolo Moretti[5], Marie Morimoto[1], John J. Mulvihill[1], Lindsay Mulvihill[16], Mariko Nakano-Okuno[2], Stanley F. Nelson[14], Serena Neumann[13], Dargie Nitsuh[3], Donna Novacic[1], Devin Oglesbee[16], James P. Orengo[7], Laura Pace[5], Stephen Pak[9], J. Carl Pallais[10], Neil H. Parker[14], LéShon Peart[8], Leoyklang Petcharet[1], John A. Phillips III[13], Filippo Pinto e Vairo[16], Jennifer E. Posey[7], Lorraine Potocki[7], Barbara N. Pusey Swerdzewski[1], Aaron Quinlan[5], Daniel J. Rader[11], Ramakrishnan Rajagopalan[11], Deepak A. Rao[10], Anna Raper[11], Wendy Raskind[3], Adriana Rebelo[8], Chloe M. Reuter[4], Lynette Rives[13], Lance H. Rodan[10], Martin Rodriguez[2], Jill A. Rosenfeld[7], Elizabeth Rosenthal[3], Francis Rossignol[1], Maura Ruzhnikov[4], Marla Sabaii[1], Jacinda B. Sampson[4], Timothy Schedl[9], Lisa Schimmenti[16], Kelly Schoch[21], Daryl A. Scott[7], Elaine Seto[7], Vandana Shashi[21], Emily Shelkowitz[3], Sam Sheppeard[3], Jimann Shin[9], Edwin K. Silverman[10], Giorgio Sirugo[11], Kathy Sisco[9], Tammi Skelton[2], Cara Skraban[11], Carson A. Smith[8], Kevin S. Smith[4], Lilianna Solnica-Krezel[9], Ben Solomon[1], Rebecca C. Spillmann[21], Andrew Stergachis[3], Joan M. Stoler[10], Kathleen Sullivan[11], Shirley Sutton[4], David A. Sweetser[10], Virginia Sybert[3], Holly K. Tabor[4], Queenie Tan[16], Arjun Tarakad[7], Herman Taylor[22], Mustafa Tekin[8], Willa Thorson[8], Cynthia J. Tifft[1], Camilo Toro[1], Alyssa A. Tran[7], Rachel A. Ungar[4], Adeline Vanderver[11], Matt Velinder[5], Dave Viskochil5, Tiphanie P. Vogel7, Colleen E. Wahl1, Melissa Walker10, Nicole M. Walley21, Jennifer Wambach[9], Michael F. Wangler[12], Patricia A. Ward[17], Daniel Wegner[9], Monika Weisz Hubshman[7], Mark Wener[3], Tara Wenger[3], Monte Westerfield[23], Matthew T. Wheeler[4], Jordan Whitlock[2], Lynne A. Wolfe[1], Heidi Wood[1], Kim Worley[7], Shinya Yamamoto[12], Stephan Zuchner[8]

1 National Institutes of Health (NIH) Undiagnosed Diseases Program (UDP) clinical site

2 University of Alabama at Birmingham, Data Management and Coordinating Center
3 Pacific Northwest clinical site
4 Stanford University clinical site
5 University of Utah clinical site
6 Stanford University Data Management and Coordinating Center
7 Baylor College of Medicine clinical site
8 University of Miami clinical site
9 Washington University clinical site, Model Organisms Screening Center
10 Harvard University clinical site
11 Children's Hospital of Philadelphia and University of Pennsylvania clinical site
12 Baylor College of Medicine Model Organisms Screening Center
13 Vanderbilt University clinical site
14 University of California at Los Angeles clinical site
15 Washington University Data Management and Coordinating Center
16 Mayo Clinic
17 Baylor College of Medicine Sequencing Core
18 Harvard University Data Management and Coordinating Center
19 Columbia University clinical site
20 University of Utah Data Management and Coordinating Center
21 Duke University clinical site
22 Morehouse Data Management and Coordinating Center
23 University of Oregon Model Organisms Screening Center

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.