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], Rosario I. Corona[5], George de V. Carvalho Neto[5], Julian Willett[6], Michele Berselli[1], William Ronchetti[1], Stanley F. Nelson[5], Julian A. Martinez-Agosto[5], Richard Sherwood[7], Joel Krier[8], 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 Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA
6 Department of Pathology and Laboratory Medicine, NewYork-Presbyterian Weill Cornell Medical Center, New York, NY
7 Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA
8 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 with ultra-rare diseases. The increasing sizes of ultra-rare 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 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 further release a software package, RaMeDiES, enabling automated cross-analysis of deidentified sequenced cohorts for new diagnostic and research discoveries. Gene-level findings and variant-level information across the cohort are 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] This opens the prospect of international cross-cohort analyses, leveraging parallel efforts in many countries, and appreciating that rare diseases know no borders.

Undiagnosed Diseases Network dataset

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 rare disease 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 UDN clinical sites.[15] Both broad Human Phenotype Ontology

(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" analyses, where probands are evaluated on a per-case basis. 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 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]

Clinical Evaluation of Computational Findings

Here, we generate candidate gene–patient matches via a series of statistical genomic analyses implemented in our software suite, **Ra**re **Me**ndelian **Di**sease **E**nrichment **S**tatistics (RaMeDiES, Figure 1g,h). We focus on the model of monogenic, autosomal inheritance in *de novo* and compound heterozygous cases to prioritize candidates via a genotype-first approach, with no clinical input or phenotypic information used. Each candidate is then evaluated with respect to the patient's clinical presentation and 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—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 S3). 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 non-causative candidate genes as controls. Specifically, non-causative genes were selected with identical criteria to true candidate genes except biallelic variants were in *cis* rather than in *trans* or had low predicted pathogenicity scores. The clinical team applying the protocol consistently scored true candidate genes higher than control genes (Wilcoxon one-sided rank-sum p-value =

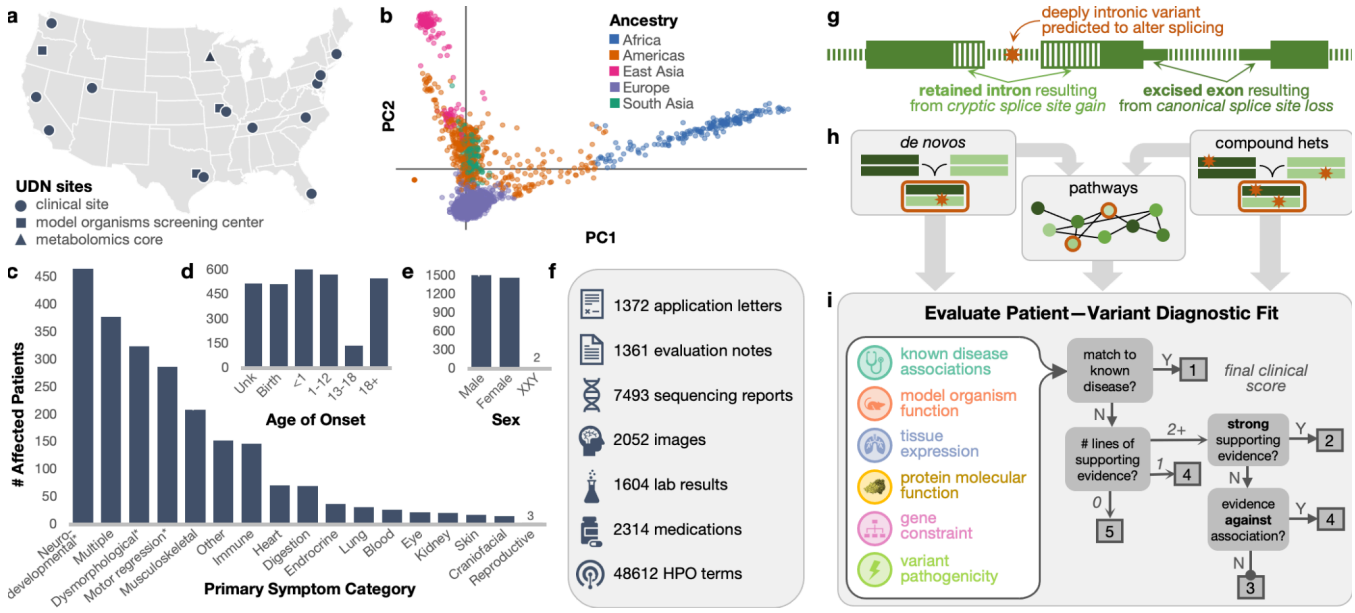

Figure 1. Undiagnosed Diseases Network cohort analysis. (a) Map of clinical and research sites within the Undiagnosed Diseases Network (UDN) for evaluating patients and candidate variant functionality. **(b)** Genetic ancestry across the sequenced patient cohort. **(c)** Clinician-recorded primary symptom 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) with a predicted splice-altering impact are considered alongside exonic variants in our statistical framework; these variants may result in retained introns or excised exons in processed transcripts. **(h)** We consider genes and gene pathways harboring *de novo* and compound heterozygous variants in sequenced trios (72% of cases). Complete case count by family structure (e.g., proband-only, duo) is in Supplementary Figure S2. Other inheritance modes (e.g., homozygous, uniparental disomy) are not considered in our cohort-based framework. **(i)** Depiction of clinical framework to uniformly evaluate how well a patient's phenotypes are concordant with a candidate gene or variant.

0.0171, Methods, Supplementary Table S1), suggesting that the scores generated by the clinicians' protocol may be used to prioritize candidates.

Results

De novo analysis

Several highly penetrant, extreme phenotypic presentations underlying Mendelian and other congenital, complex human diseases have been linked to *de novo* mutations.[1,18,19] We began by evaluating 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.[20] Mutation count

showed expected dependency on parental ages with Poisson-distributed adjusted counts, attesting to the quality of *de novo* calling (Figure 2a, Supplementary Figure S4).

We then sought to identify genes enriched for deleterious *de novo* mutations across our patient cohort. The power of this enrichment calculation increases with better models of underlying mutation rates and estimates of variant deleteriousness. Recently, the rate of *de novo* emergence has been estimated at basepair resolution with a high degree of accuracy.[21] Newly developed deep learning models for predicting the pathogenicity of *de novo* and other variants also now exhibit unprecedented accuracy in distinguishing disease-relevant variants.[22,23] We leverage these recent advances to build an accurate, unbiased statistical procedure called RaMeDiES-DN to detect genes enriched for deleterious *de novos*.

Unlike the earliest generation of *de novo* recurrence approaches which leveraged Poisson approximations for runtime efficiency but could not take advantage of improved deleteriousness scores and mutation rate models,[18] RaMeDiES methods seamlessly incorporate per-variant deleteriousness scores and mutation rates without sacrificing runtime. Briefly, 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 possible variants with as high or higher a deleteriousness score. By construction, this per-variant mutational target is expected to be a uniformly distributed statistic (Supplementary Note S3). Our framework naturally combines different variant types including SNVs and indels with a distinct mutation rate model, and can interchangeably utilize various deleteriousness scores (Figure 2b, Methods). Although current state-of-the-art *de novo* recurrence approaches also incorporate relevant variant-level information, they rely on a complex, permutation procedure.[1] RaMeDiES' analytical approach eliminates the need for permutation-based significance calculations and can process large datasets in mere seconds while maintaining well-calibrated p-values (Supplementary Figure S5). Furthermore, RaMeDiES' operation at the level of mutational targets enables sharing of intermediate statistics across cohorts without revealing patients' individual variants.

We first focus on the subset of missense variants, which comprise a sizable proportion of known Mendelian disease-causing variants and for which new, specialized pathogenicity predictions exist (e.g., PrimateAI-3D and AlphaMissense).[22–24] We find one significant gene, *KIF21A*, corresponding to the correct, complete diagnosis in one patient and a strong partial diagnosis in one other (Bonferroni-adjusted Cauchy-combined p-value < 0.05, Figure 2c). Notably, disease genes with a *de novo* mode of inheritance are expected to be under strong selection against heterozygous loss-of-function variants. We further refine our method to incorporate this intuition by prioritizing genes by their GeneBayes values, which indicate selection against heterozygous protein-truncating variants, using a weighted false discovery rate (FDR) procedure.[25–27] With this correction, we obtain three

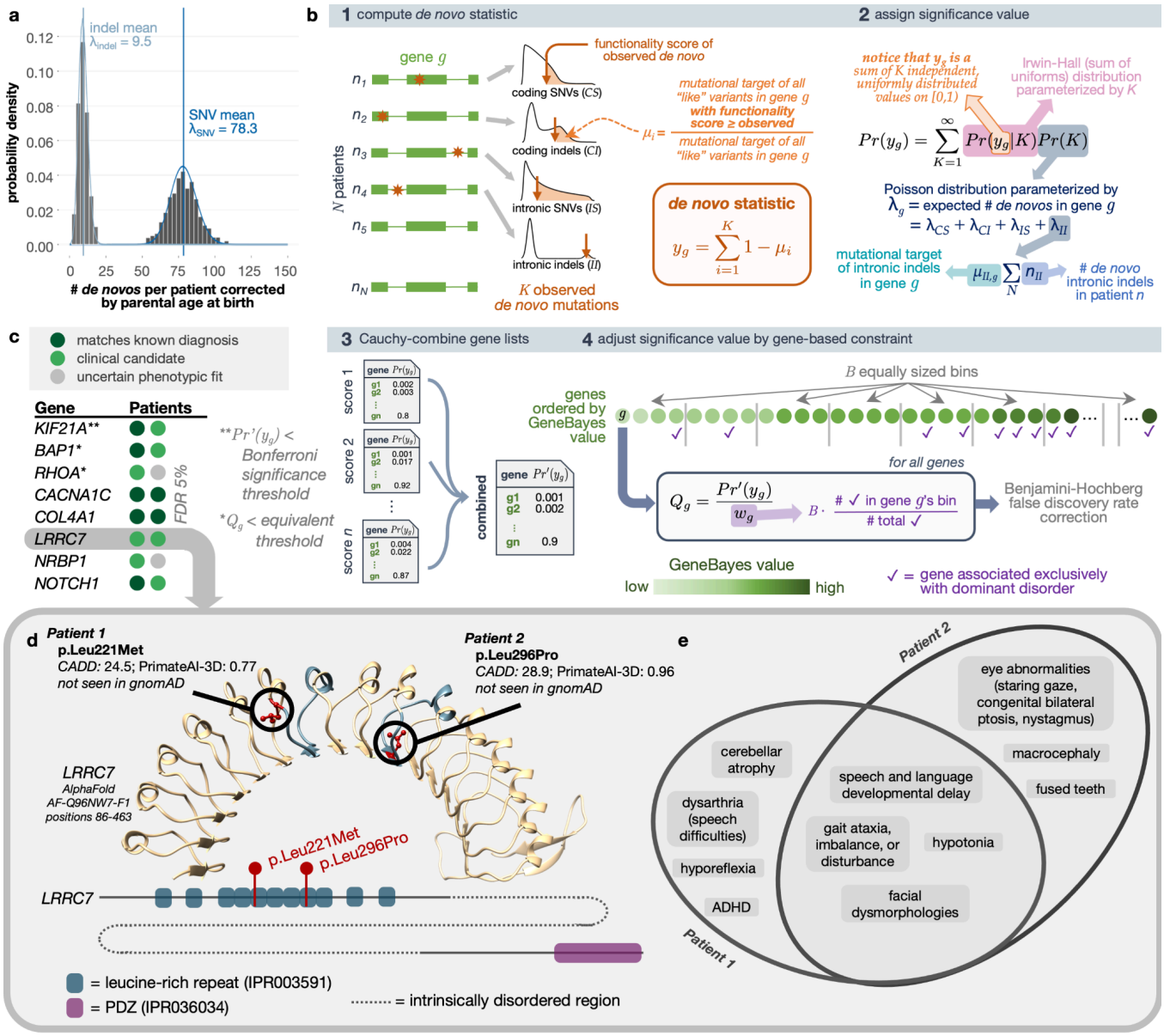

Figure 2. *De novo* recurrence. (a) *De novo* mutation counts per proband adjusted for parental ages. Blue vertical lines show the mean values of the distributions, and curves represent the Poisson fits. **(b)** Schematic of analytical test for the recurrence of *de novos* that considers distal splice-altering and exonic SNV and indel variants, their variant functionality scores, a genome-wide mutation rate model Roulette, and per-gene GeneBayes constraint values. "Like" variants refer to those of the same variant class (i.e., coding SNVs [*CS*], coding indels [*CI*], intronic SNVs [*IS*], intronic indels [*II*]) and within the same functionality score and minor allele frequency thresholds. **(c)** Genes with highest significance values for *de novo* recurrence across the cohort when focusing on missense variants with AlphaMissense and PrimateAI-3D scores; patients are represented as colored circles. Complete gene list can be found in Supplementary Table S2. **(d)** AlphaFold-predicted human *LRRC7* protein structure (AF-Q96NW7-F1) covering the leucine-rich repeat region with high predicted structural confidence (amino acid positions 86-463). The fifth and eighth LRR domains where missense *de novos* were found are highlighted in blue. Reference alleles for missense *de novo* variants observed across two UDN patients (red) are shown in circles. A depiction of *LRRC7*'s linear protein sequence (Ensembl ID ENSP00000498937) with InterPro predicted domains shown in colored boxes is below. **(e)** Overlap of phenotype terms annotated to each patient.

gene findings at an equivalent significance threshold (Q-value < 3e-6) and eight gene findings at FDR 5% (Supplementary Table S2). Our second and third gene hits, *BAP1* and *RHOA*, correspond to a known correct diagnosis in one patient and strong clinical matches in two other patients. Among the five remaining genes at FDR 5%, three genes (*CACNA1C*, *COL4A1* and *NOTCH1*) correspond to known diagnoses in five patients and the top clinical candidate in one patient. Two impacted patients with *de novo* missense variants in the leucine-rich repeat region of *LRRC7*, a gene not yet known to be disease-associated, had phenotypic overlap of hypotonia and developmental delay; one patient additionally experienced nystagmus, staring spells, and balance problems and the second had ataxic gait (Figure 2d-e). These findings and *LRRC7*'s expression in the brain further support its link to an emerging neurodevelopmental disorder.[14] Another gene, *NRBP1*, remains a strong candidate in two patients due to their neurological phenotype overlap and *NRBP1*'s expression in the brain. An initial functional study in fly through the UDN Model Organism Screening Core was inconclusive. This gene has been submitted to Matchmaker Exchange.

We next consider all exonic variants, including nonsense variants and indels, and further incorporate additional well-established deleteriousness predictors, CADD and REVEL.[28,29] Different mutagenesis processes lead to indel mutations, so SNV mutation rate models can be inappropriate for modeling this mutation type for some genes.[30] We therefore constructed a separate per-gene mutation rate approximation for indels (see Methods for details). When we reran RaMeDiES-DN on all exonic variants using four deleteriousness predictors, we additionally identified *KMT2B* (Bonferroni-adjusted Cauchy-combined p-value < 0.05), corresponding to a correct diagnosis in four patients due to *de novo* indel variants (Supplementary Table S3, Supplementary Figure S6a). The next seven gene findings at FDR 5% were all identified when assessing recurrence of missense variants. At FDR 10%, we identify five new putative diagnoses. For instance, two patients had high impact missense *de novo* variants impacting *H4C5*, a histone gene that was not detected with significance in our missense-only enrichment test due to its lack of precomputed AlphaMissense scores. Both patients had infantile-onset gross motor developmental delays, dysmorphic facial features, and speech difficulties (Supplementary Figure S6b,c). 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] 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.

Inclusion of deep intronic splice variants

Next, we demonstrate how RaMeDiES-DN can be extended to additionally consider non-exonic variants uncovered uniquely from whole genome sequencing using the same methodological infrastructure. On the one hand, it remains challenging to identify non-coding regulatory variants involved in rare Mendelian diseases,[34] and the overall role of such variants in congenital disorders is still a subject of debate.[35] 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.[36] Identification of splice-altering variants directly from genome sequencing data is recently possible using newly-developed *in silico* predictive scores without relying 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.[37] Moreover, identifying disease-causal intronic splice variants is especially appealing due to their potential targetability using antisense oligonucleotide therapies.[38]

Unlike functional predictions for exonic variants, which have been extensively validated for consistency and accuracy via decades of experimental *in vitro* and *in vivo* studies, functional predictions of splice-altering intronic variants are relatively new and still require experimental confirmation. We used a combined computational–experimental approach to prioritize distal splice variants using *in silico* predicted scores and an *in vitro* massively parallel splicing reporter assay (Methods, Supplementary Figure S7).[39,40] We found the per-variant *in silico* predictions to be mostly concordant with the *in vitro* assay readouts. Variants assigned higher *in silico* scores are more frequently supported by the experimental, *in vitro* assay, and those with relatively lower *in silico* scores (SpliceAI < 0.5) have a non-negligible validation rate as well (Supplementary Figure S8). 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. Once genome-wide functionality score tracks are released for the next generation of splice predictors as well (e.g., Pangolin),[41] they can be integrated into RaMeDiES using the same methodology leveraged for exonic variant predictors.

No new candidate genes with a significant recurrence of intronic *de novos* were found in the UDN dataset. However, by seamlessly incorporating non-exonic variants within the

same statistical test, our approach enables a more complete, automated analysis of the growing volume of whole genome sequencing data across rare disease consortia.

We also ran the state-of-the-art *de novo* enrichment approach, DeNovoWEST.[1] Unlike our approach, DeNovoWEST incorporates a gain-of-function model alongside a loss-of-function model, which has the potential to yield additional findings. We equipped the DeNovoWEST algorithm with the Roulette mutation rate model, up-to-date CADD variant deleteriousness and s_{het} gene constraint scores,[26] and further incorporated deep intronic variants with predicted splice-altering impact (Supplementary Figure S9). This approach yielded two Bonferroni-significant genes, one of which was also uncovered by RaMeDiES-DN at Bonferroni significance and the second at a FDR of 6% (*KMT2B* and *H4C5*, Supplementary Figure S10). We did not apply an FDR-based approach to DeNovoWEST's results to consider additional gene findings, because DeNovoWEST p-values are a construct over three sometimes dependent tests, rendering an FDR adjustment inappropriate. We also find *CSMD1*, a highly indel-prone gene, within DeNovoWEST's top-ranked five genes, likely because indels and SNVs are not distinguished in the mutation rate model.[42]

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.[43] 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, due to the effect of selection against biallelic variant occurrences or against slightly deleterious phenotypes of heterozygous variants.[44] Despite the expected low frequency of individual alleles comprising a comphet pair, directly selecting for highly deleterious comphet variants still results in numerous false positive findings at the cohort level, motivating a statistical approach for cohort-level comphet prioritization. Developing a statistical framework analogous to *de novo* recurrence requires modeling the distribution of rare inherited alleles per individual. *De novo* mutations arise through the universal process of mutagenesis and are therefore straightforward to model. Similarly, the distribution of the total number of all derived alleles per haploid genome (i.e., all non-ancestral variants inherited from one parent without any imposed frequency constraints) are also not dependent on the demographic history of the population and therefore are straightforward to model.[45,46] In contrast, however, the distribution of the total number of *rare* alleles per individual is highly dependent on population structure, which is notoriously difficult to account for. Some previous approaches for determining

cohort-level significance of comphet variants ignore population structure when modeling the number of rare variants. Although this may be an accurate statistical test in controlled model organism cross experiments, it is inappropriate for natural human populations, where population structure is present even at a very fine scale.[47] In the Genome of the Netherlands (GoNL) dataset for instance, the number of synonymous singletons across unrelated individuals still reflects geographic structure along a south-north cline.[48]

In our framework, we sidestep directly modeling the distribution of rare variant counts per individual and instead condition on the observed number of rare variants inherited from each parent using trio-level data. Given the number of rare variants inherited from each parent per individual, we then compute the probabilities of comphet variants landing in high-scoring positions in the same gene across the cohort. Although the positions where inherited variants land 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.[21,49] We therefore model the positional distribution of rare inherited variants using the same Roulette basepair-resolution *de novo* mutation rate model leveraged in our *de novo* recurrence model. Our comphet recurrence model, called RaMeDiES-CH, relies on the comphet mutational target, computed for each comphet variant pair and defined similarly as the *de novo* mutational target previously introduced. Specifically, the comphet mutational target is computed as the total *squared* mutation rate of all possible variants with higher functionality scores (Figure 3a). RaMeDiES-CH applies the Cauchy p-value combination approach as before to leverage multiple variant-level functionality scores while considering exonic and intronic variants, but does not incorporate gene constraint scores, which do not exist for recessive selection (Methods, Supplementary Figure S11).[50] RaMeDiES-CH computes well-calibrated per-gene p-values for comphet variants in a cohort (Supplementary Figure S5).

Across the set of non-consanguineous UDN families, we did not find significant recurrent comphet occurrences across genes. This result is unsurprising, as previous estimates suggest that in panmictic disease populations, only one deleterious comphet variant is expected for every five dominant *de novos*.[47] Nevertheless, RaMeDiES-CH represents an accurate and unbiased statistical test for the recurrence of comphet variants in human populations, which can be applied to reveal new diagnoses as sequenced rare disease datasets expand.

We suspected that singleton disease-causing comphet variants were still present in the cohort. We adapted our statistical framework to compute an individual-based statistic, RaMeDiES-IND, that normalizes each observed comphet variant mutational target across all genes in the genome rather than across all individuals in a cohort (Supplementary Figure

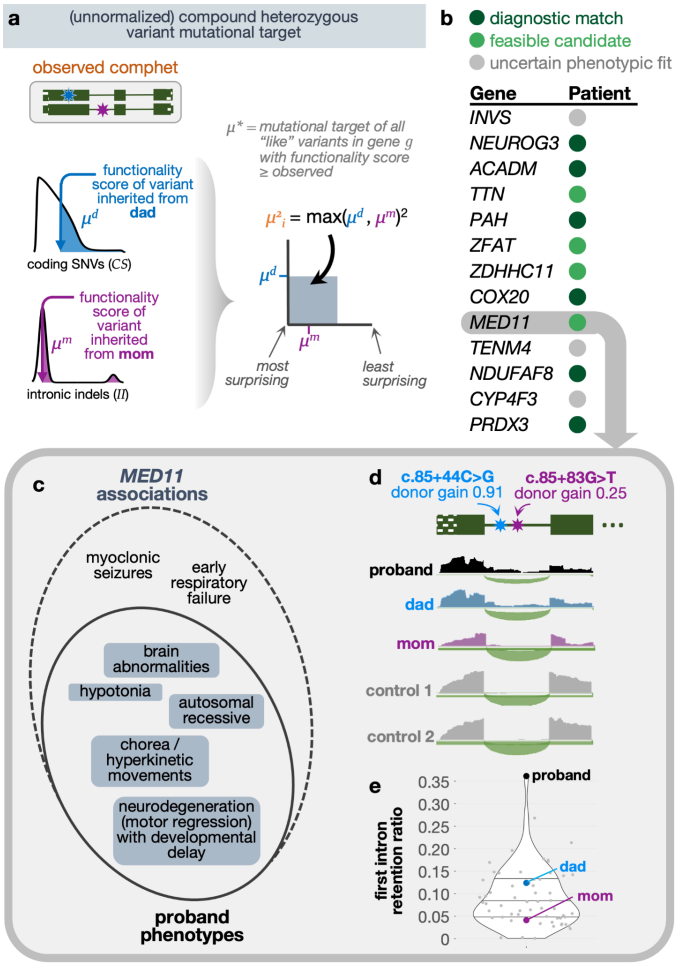

Figure 3. Compound heterozygous variants. (a) Illustration of the unnormalized squared mutational target computed for each observed comphet variant in a gene across the cohort (RaMeDiES-CH, Supplementary Figure S11) or in an individual across the genome (RaMeDiES-IND, Supplementary Figure S12). "Like" variants refer to those of the same variant class (i.e., coding SNVs [*CS*], coding indels [*CI*], intronic SNVs [*IS*], intronic indels [*II*]) and within the same functionality score and minor allele frequency thresholds. **(b)** Top ranked genes resulting in the best enrichment statistic computed for RaMeDiES-IND. Putative candidates refer to genes that remain candidates for pathogenicity due to their phenotypically-relevant tissue expression, but where there is not enough functional evidence or published gene–disease relationships to establish causality at this time. **(c)** Overlap between phenotypes associated with *MED11* and those exhibited by the affected patient. **(d)** RNA-Seq reads from whole blood samples aligned to first two exons and first intron of *MED11* for proband (black), dad (blue), mom (purple) 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*. **(e)** Proband exhibits significant retention of the first intron relative to parents and fifty-three 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).

S12). This approach yielded a ranked list of patient–gene pairs across the UDN cohort, where each patient–gene pair could be annotated as corresponding to a correct diagnosis or otherwise (Supplementary Table S4). We computed a single enrichment statistic for this overall patient–gene ranking, which simultaneously suggested a threshold for clinical consideration of findings, as the best Fisher's exact test P achieved across all positions in

the list. This enrichment statistic was significant when compared to the distribution of the same statistic computed across 10,000 random shuffles of the patient–gene list (permutation p-value = 0.001, Methods, Supplementary Figure S13). Among the top thirteen hits yielding this best enrichment statistic, we recapitulated five known diagnoses (i.e., *NEUROG3*, *PAH*, *COX20*, *NDUFAF8*, *PRDX3*)[51,52] and newly identified the genomic cause of a known biochemical diagnosis (i.e., *ACADM* in a patient with MCAD deficiency). We also 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 3c). *MED11* is associated with epilepsy and intellectual disability, and this patient's presentation could represent a phenotypic expansion of this known disorder.[53] Both inherited variants occur deep in the first intron of *MED11*, a region that would be missed by exome-only sequencing or analysis, 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 3d-e, Supplementary Figure S14).[54]

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.[47]

Pathway analysis

Genes involved in the same pathway are frequently involved in similar phenotypic presentations.[55–58] 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.[59] 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

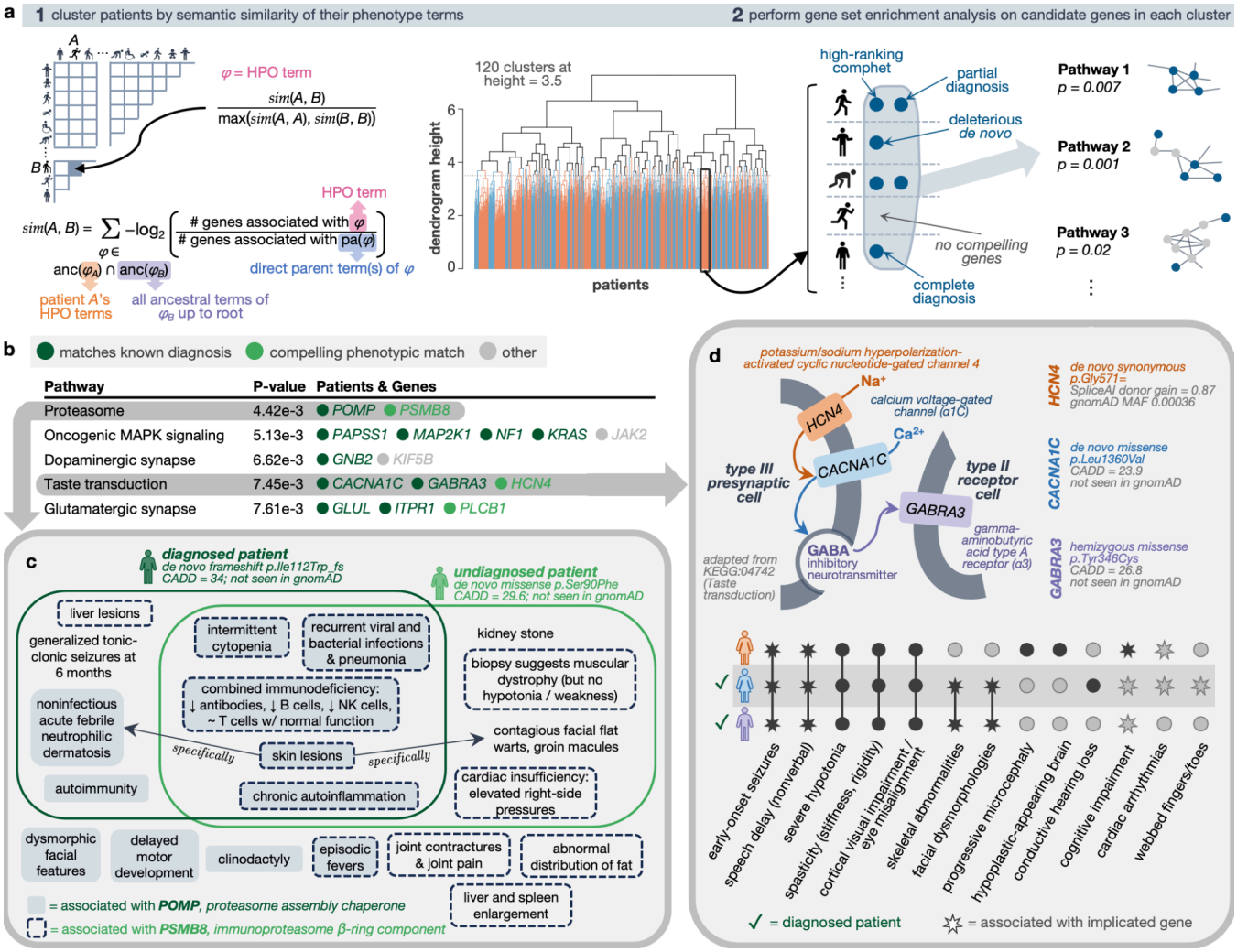

Figure 4. Biological pathways enriched within phenotypically-similar patient subgroups. (a) Schematic illustrating the two-step process of first clustering patients according to the semantic similarity of their phenotype terms and second finding enriched biological pathways among the genes within each 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 S6. **(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).

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, Supplementary Table S5). We focus our attention on undiagnosed cases with *de novo* 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 S6).

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.[60] 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.[61]

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.[62,63] 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.[64,65] 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.[64,65]

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 (Methods, Supplementary Table S5).

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.[66,67] 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.[1,2,68]

Although the genome is a big place, it is also a finite space with respect to gene regions impacted by simple variants such as SNVs and short (<10 basepairs) indels. This suggests that, in theory, recurrence-based 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 prenatally viable monogenic disease stemming from these variant types. In order to take statistical discoveries as primary evidence, as is the case for common diseases, we need accurate, well-calibrated statistical methods.[69] 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,[70] 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. Although novel disease–gene discovery from large, phenotypically- and genetically-homogenous cohorts has been demonstrated, we show here that rigorous analysis of a diverse, moderately-sized disease cohort at the gene and the pathway level shows promise.

We also acknowledge the limitations of our models and of statistical approaches in general for comprehensive rare disease diagnosis using short-read sequencing data. First, although our models integrate non-coding variants with predicted splice-altering impacts, they do not consider potentially functional variants within whole genome data that fall into untranslated gene regions, RNA-coding genes or between genes, as genome-wide tracks of

verifiable deleteriousness scores do not exist for these variant types. Improvements to and precomputed scores for these variants will be beneficial for interpretation efforts in general and can be leveraged in future iterations of RaMeDiES. Our statistical analysis also does not consider structural, large indel, copy number, or tandem repeat variants, as their identification from short-read sequencing data is computationally expensive and often inaccurate. Investing in the detection of these variants from available data is difficult to justify given the advent of affordable long-read sequencing technologies and ongoing efforts to generate this data within the UDN and elsewhere, which should enable improved identification and analysis of pathogenic complex variants.[71,72] Developing a statistical model for these variants will still require accurate mutation rate estimates for these variant types, which is lacking. GnomAD-SV represents a promising iteration toward this goal, but is still highly dependent on their specific variant calling pipeline and data rather than biological mutagenic processes.[73]

The presented method considers only autosomal *de novo* and compound heterozygous inheritance patterns due to complications in modeling other disease-relevant inheritance patterns. First, it is difficult to propose a statistical model for biallelic variant counts in consanguineous and founder populations, including homozygous variants, because these counts strongly depend on the ancestral population history and inbreeding patterns. A more appropriate statistical approach for assessing recurrence of these variants would be the extension of parametric linkage applied to very large cohorts.[74] Second, inclusion of hemizygous or other X-chromosome variants requires accurate sex-chromosome variant calling, which is notoriously error prone, as well as an accurate mutational model of the X chromosome, which is complicated due to sex-dependent selection and random X-inactivation. Finally, although we do not model parental mosaicism or uniparental disomy in our recurrence statistics, these inheritance patterns and events are regularly assessed via complementary, traditional "N-of-1" case-based approaches.[12]

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 provide an open-source software package, RaMeDiES, implementing the efficient and well-calibrated statistics for *de novo* recurrence and deleterious compound heterozygous

inference proposed here. RaMeDiES' operation on shareable summary statistics rather than on variant-level data enables automated, deidentified cross-analysis of substantial existing yet siloed sequenced cohorts for new diagnostic discoveries. As the Mendelian genomics field continues the transition to this new data science phase, the methods we present here should facilitate the exciting prospect of international cross-cohort analyses, resulting in new findings and a vastly improved rare disease diagnostic rate globally.

Data Availability

Deidentified genome data, transcriptome data, and corresponding phenotype data in the form of HPO terms are regularly deposited in dbGaP (accession phs001232.v5.p2). Genome-wide, rare SNV and indel variants and HPO codes for UDN participants included in this study are queryable in our public-facing browser. Standardized phenotype data and candidate genes and variants are submitted to MatchmakerExchange. Variant-level data, clinical significance and supporting evidence, demographic information, and phenotype information for all diagnostic variants are regularly submitted to ClinVar. Identifiable patient data is under controlled access to protect patient privacy. Other relevant, deidentified patient-specific clinical information may be shared on a case-by-case basis at the discretion of the corresponding clinical team if it is directly related to diagnosing or potentially treating the patient.

Code Availability

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

Ethics & Inclusion Statement

The authors declare no competing interests. This work was performed in accordance with all ethical guidelines outlined in the NIH IRB #15HG0130. The study proposal and manuscript were approved by the UDN Publications and Research Committee. Local researchers from each UDN site supplying data to the UDN, including clinical team members and bioinformaticians participating in the UDN Tool Building Coalition working group, were included. This research is relevant to individual patients and their clinical teams.

Acknowledgments

The authors would like to thank 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, 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. The authors acknowledge funding from 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, NIH National Human Genome Research Institute (NHGRI) grants U01HG012009, R01HG012286 and R21HG010391, NIH National Institute of General Medical Sciences (NIGMS) grants R35GM127131 and 5T32GM007748, NIH National Institute of Mental Health (NIHM) grant R01MH101244, NIH National Center for Advancing Translational Sciences (NCATS) grant KL2TR002552, and NIH National Heart Lung and Blood Institute (NHLBI) grant 1R01HL164409-01.

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. Neale, B. M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242–245 (2012).
19. Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
20. Veltman, J. A. & Brunner, H. G. De novo mutations in human genetic disease. *Nat. Rev. Genet.* **13**, 565–575 (2012).
21. 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).
22. Gao, H. *et al.* The landscape of tolerated genetic variation in humans and primates. *Science* **380**, eabn8153 (2023).
23. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
24. Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433–438 (2017).
25. Genovese, C. R., Roeder, K. & Wasserman, L. False Discovery Control with p-Value Weighting. *Biometrika* **93**, 509–524 (2006).
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. Zeng, T., Spence, J. P., Mostafavi, H. & Pritchard, J. K. Bayesian estimation of gene constraint from an evolutionary model with gene features. *Res Sq* (2023)

doi:10.21203/rs.3.rs-3012879/v1.

28. Ioannidis, N. M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am. J. Hum. Genet.* **99**, 877–885 (2016).
29. Rentzsch, P., Schubach, M., Shendure, J. & Kircher, M. CADD-Splice-improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Med.* **13**, 31 (2021).
30. Bethune, J., Kleppe, A. & Besenbacher, S. A method to build extended sequence context models of point mutations and indels. *Nat. Commun.* **13**, 7884 (2022).
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. Albert, F. W. & Kruglyak, L. The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.* **16**, 197–212 (2015).
35. Umans, B. D., Battle, A. & Gilad, Y. Where Are the Disease-Associated eQTLs? *Trends Genet.* **37**, 109–124 (2021).
36. 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).
37. 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).

38. 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).
39. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535–548.e24 (2019).
40. 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).
41. Zeng, T. & Li, Y. I. Predicting RNA splicing from DNA sequence using Pangolin. *Genome Biol.* **23**, 103 (2022).
42. 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).
43. Temaj, G., Nuhii, N. & Sayer, J. A. The impact of consanguinity on human health and disease with an emphasis on rare diseases. *Orphanet J. Rare Dis.* **1**, 2 (2022).
44. Loreau, M. & Hector, A. Partitioning selection and complementarity in biodiversity experiments. *Nature* **412**, 72–76 (2001).
45. 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).
46. 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).
47. Martin, H. C. *et al.* Quantifying the contribution of recessive coding variation to developmental disorders. *Science* **362**, 1161–1164 (2018).
48. Sohail, M. *et al.* Negative selection in humans and fruit flies involves synergistic

epistasis. *Science* **356**, 539–542 (2017).

49. Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).
50. 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).
51. Wang, J. *et al.* Mutant neurogenin-3 in congenital malabsorptive diarrhea. *N. Engl. J. Med.* **355**, 270–280 (2006).
52. Hillert, A. *et al.* The Genetic Landscape and Epidemiology of Phenylketonuria. *Am. J. Hum. Genet.* **107**, 234–250 (2020).
53. Calì, E. *et al.* A homozygous MED11 C-terminal variant causes a lethal neurodegenerative disease. *Genet. Med.* **24**, 2194–2203 (2022).
54. Middleton, R. *et al.* IRFinder: assessing the impact of intron retention on mammalian gene expression. *Genome Biol.* **18**, 51 (2017).
55. Rual, J.-F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173–1178 (2005).
56. Costanzo, M. *et al.* The genetic landscape of a cell. *Science* **327**, 425–431 (2010).
57. Rolland, T. *et al.* A proteome-scale map of the human interactome network. *Cell* **159**, 1212–1226 (2014).
58. Ferrari, S. *et al.* Retinitis pigmentosa: genes and disease mechanisms. *Curr. Genomics* **12**, 238–249 (2011).
59. 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).

60. 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).
61. 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).
62. Rodan, L. H. *et al.* Phenotypic expansion of CACNA1C-associated disorders to include isolated neurological manifestations. *Genet. Med.* **23**, 1922–1932 (2021).
63. 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).
64. 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).
65. Blake, J. A. *et al.* Mouse Genome Database (MGD): Knowledgebase for mouse-human comparative biology. *Nucleic Acids Res.* **49**, D981–D987 (2021).
66. 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).
67. 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).
68. Coe, B. P. *et al.* Neurodevelopmental disease genes implicated by de novo mutation and

copy number variation morbidity. *Nat. Genet.* **51**, 106–116 (2019).

69. Uffelmann, E. *et al.* Genome-wide association studies. *Nature Reviews Methods Primers* **1**, 1–21 (2021).
70. Chopra, M. & Duan, T. Rare genetic disease in China: a call to improve clinical services. *Orphanet J. Rare Dis.* **10**, 140 (2015).
71. Steyaert, W. *et al.* Unravelling undiagnosed rare disease cases by HiFi long-read genome sequencing. *medRxiv* (2024) doi:10.1101/2024.05.03.24305331.
72. Gorzynski, J. E. *et al.* Clinical application of Complete Long Read genome sequencing identifies a 16kb intragenic duplication in EHMT1 in a patient with suspected Kleefstra syndrome. *bioRxiv* (2024) doi:10.1101/2024.03.28.24304304.
73. Collins, R. L. *et al.* A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020).
74. Andres, E. M. *et al.* A genome-wide analysis in consanguineous families reveals new chromosomal loci in specific language impairment (SLI). *Eur. J. Hum. Genet.* **27**, 1274–1285 (2019).

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], Rosario I. Corona[5], George de V. Carvalho Neto[5], Julian Willett[6], Michele Berselli[1], William Ronchetti[1], Stanley F. Nelson[5], Julian A. Martinez-Agosto[5], Richard Sherwood[7], Joel Krier[8], 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 Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA
6 Department of Pathology and Laboratory Medicine, NewYork-Presbyterian Weill Cornell Medical Center, New York, NY
7 Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA
8 Department of Genetics, Atrius Health, Boston, MA

Undiagnosed Diseases Network (UDN) structure **2**
Harmonization of whole genome sequencing data **3**
Clinical evaluation framework **4**
Protocol overview 4
Clinical score calibration 4
Validation 5
Identification of de novo variants **5**
Analytical test for de novo cohort-level recurrence **6**
Basic statistic definition 6
Different deleteriousness scores for coding and intronic variants 7
Different mutation rate models for SNV and indel variants 8

Incorporation of different variant types....8
Cauchy-combination of p-values computed with different deleteriousness predictors...9
Incorporation of GeneBayes values....9
Massively Parallel Splicing Reporter Assay (MPSA)....10
Assay design....10
Library cloning and experimental protocol....10
Barcode mapping....11
MPSA validation rate....11
DeNovoWEST gene-specific enrichment of de novo variants....12
Analytical test for compound heterozygous cohort-level recurrence....13
Modeling false positive diagnoses....14
Analytical test for individual-level compound heterozygous configuration....15
Enrichment for correct diagnoses....17
Transcriptome sequencing analysis for MED11....17
RNA extraction, sequencing and quality control....17
Intron retention outlier analysis....18
Pathway enrichment analysis....18
Phenotypically-similar patient groupings....18
Selecting genes per patient cluster....18
Gene Set Enrichment Analysis (GSEA)....19
References....19

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 patients enrolled in the UDN have consented to the broad sharing of all their genomic, phenotypic and clinical data with researchers network-wide for use in research projects and 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 (versions v2.1.1 and v3.1.2) 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. We chose to stringently exclude individuals with evidence of familial consanguinity (i.e., by imposing a parental relatedness and IBD region length constraints) rather than excluding patients based on their relative recessive burden because an assumption of our statistical models is violated in consanguineous cases (Supplementary Note S4).

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 maximum 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).

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[10] 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/).[11] 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.[12] This process resulted in 1072 trios with an affected proband and unaffected parents for further analysis.

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

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

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 $SF = 1 - CDF$.[13] 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).[14,15] 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 deleteriousness scores that are precomputed and publicly-available for all possible variants genome-wide in our computations. Precomputed scores are required for the calculation of comprehensive, basepair-resolution mutational targets as described above. For missense variants, we interchangeably use AlphaMissense (version hg38 released with their 2023 publication), PrimateAI-3D (academic license, accessed May 2024), CADD (version 1.6), and REVEL (accessed May 2024).[16–19] CADD is also used for scoring all other exonic variants, including nonsense and indel variants. For intronic variants, we use SpliceAI (academic license, accessed May 2021).[20] 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 S15). Clinical sequencing centers also regularly report these scores, suggesting their relevance in rare disorders.[8,21]

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.[22] 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

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

where K_t is the number of observed *de novo* mutations of variant type t landing in gene g. The expected number of *de novo*s 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 $Pr(K) = Pois(K|\lambda')$.

Cauchy-combination of *p*-values computed with different deleteriousness predictors

We can run our method using different deleteriousness score predictions for coding SNVs (i.e., AlphaMissense, PrimateAI-3D, CADD, or REVEL), resulting in slightly different lists of genes with corresponding *p*-values when incorporating this variant type. We combine these lists using the Cauchy combination test, an analytic calculation that is applicable under arbitrary dependence structures.[15]

Incorporation of GeneBayes values

We incorporate GeneBayes values, which estimate the selection against heterozygous protein-truncating variants per gene, as weights in a weighted false discovery rate (FDR) procedure.[23,24] We sort all genes in ascending order by their GeneBayes 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 GeneBayes values are assigned a weight $w_g = 1$.

Note that $\mathbb{E}[w_g] = 1$ and that GeneBayes 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 $Q = Pr(y')/w_g$.[23]

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[25,26] (Supplementary Figure S7) as well as a separate Puromycin selection cassette. Plasmids have been deposited to Addgene.

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 S8a).

We compared the MPSA validation rates across the three biological replicates and two cell types using Pearson's correlation (Supplementary Figure S8b).

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.[20] 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 S9a). The underlying triplet-context mutational model was replaced with genome-wide, per-SNV Roulette mutation rate estimates.[27] 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.[28] 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 S9b-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. We find that DeNovoWEST and RaMeDiES-DN (using only CADD in exonic regions as a closer comparison to DeNovoWEST) recovered known autosomal dominant disease genes at a comparable rate across *de novo* variants provided in the original DeNovoWEST paper (Supplementary Figure S16).

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

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

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 $y^{c\prime}$ 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 $\tilde{\mu}'$, $\tilde{y}^{c\prime}$ and Poisson parameter $\tilde{\lambda}^{c\prime}$ 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

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

and

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

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.

Transcriptome sequencing analysis for *MED11*

RNA extraction, sequencing and quality control

RNA was extracted from UDN patients' whole blood samples received at UCLA between 2018 and 2019 using PAXgene Blood RNA extraction kits from Qiagen. Concentration of RNA in each sample was quantified using the Qubit 3.0 Fluorometer. RNA integrity numbers (RINs), a quality control measure, were assessed per sample using the Agilent bioanalyzer. RNA libraries were prepared for each sample using either the NuGEN Universal Plus mRNA-Seq kit or the Illumina TruSeq mRNA + Globin Minus kit. Sequencing was then performed on the Illumina NovaSeq 6000 to generate ~50-100 million 100-150bp paired-end reads per sample. Library preparation and sequencing were performed at the UCLA Neuroscience Genomics Core and the UCLA Technology Center for Genomics and Bioinformatics Core. Sequenced reads in FASTQ format were aligned to human reference genome GRCh37 using STAR v2.5.2b with default parameters and Gencode v19 annotations.[29,30] To increase sensitivity to novel splice junctions, reads were mapped using the STAR 2-pass mode, where novel splice junctions detected during the first pass alignment are indexed and used alongside known splice junctions in the second pass remapping. We confirmed effective ribosomal RNA (rRNA) depletion per sample by aligning all paired-end reads to the complete sequences for nuclear and mitochondrial rRNAs using BWA-mem v0.7.17[31] and ensured that the proportion of aligned reads did not suggest excessive rRNA contamination. Duplicate reads were marked using PicardTools v4.2.4.0 and post-alignment sequencing quality was assessed using RNA-SeQC v1.1.8 to ensure adequate library complexity.[32,33] RNA sample identity was confirmed by comparing single nucleotide variant (SNV) calls from RNA sequencing to SNV calls from corresponding exome or genome sequencing data per sample.

Intron retention outlier analysis

Fifty-three tissue-matched control samples from UDN participants unrelated to the proband, mother and father were selected for outlier analysis. IRFinder v1.2.4[34] was run in BAM mode using the same human reference GRCh37 and Gencode v19 annotations on aligned BAM files to measure the level of intron retention (i.e., "IRratio") in *MED11* across the proband, mother, father, and control samples. The IRratio is computed per sample as (median read depth of first intron) / (number of reads spanning first and second exons + median read depth of first intron). Aligned reads covering the *MED11* gene region (i.e. chr17:4,634,723-4,636,903) from the proband, mother, father and two control samples were viewed using a local installation of the Integrative Genomics Viewer (IGV) v2.16.0.[35]

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.[36] UDN patients experience a spectrum of symptoms across overlapping biological categories and therefore cannot be easily separated into distinct, well-defined clusters (Supplementary Figure S17). 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 RaMeDiES-IND 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.[37] 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.[38] 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. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
11. 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).
12. Yang, X. *et al.* Developmental and temporal characteristics of clonal sperm mosaicism. *Cell* **184**, 4772–4783.e15 (2021).
13. Philip, H. The distribution of means for samples of size N drawn from a population in which the variate takes values between 0 and 1, all such values being equally probable. *Biometrika* **19**, 240–244 (1927).

14. Heard, N. A. & Rubin-Delanchy, P. Choosing between methods of combining |p|-values. *Biometrika* **105**, 239–246 (2018).

15. 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).

16. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).

17. Gao, H. *et al.* The landscape of tolerated genetic variation in humans and primates. *Science* **380**, eabn8153 (2023).

18. Rentzsch, P., Schubach, M., Shendure, J. & Kircher, M. CADD-Splice-improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Med.* **13**, 31 (2021).

19. Ioannidis, N. M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am. J. Hum. Genet.* **99**, 877–885 (2016).

20. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535–548.e24 (2019).

21. 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).

22. Bethune, J., Kleppe, A. & Besenbacher, S. A method to build extended sequence context models of point mutations and indels. *Nat. Commun.* **13**, 7884 (2022).

23. Genovese, C. R., Roeder, K. & Wasserman, L. False Discovery Control with p-Value

Weighting. *Biometrika* **93**, 509–524 (2006).

24. Zeng, T., Spence, J. P., Mostafavi, H. & Pritchard, J. K. Bayesian estimation of gene constraint from an evolutionary model with gene features. *bioRxiv* (2024) doi:10.1101/2023.05.19.541520.
25. 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).
26. Soemedi, R. *et al.* Pathogenic variants that alter protein code often disrupt splicing. *Nat. Genet.* **49**, 848–855 (2017).
27. 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.
28. Cassa, C. A. *et al.* Estimating the selective effects of heterozygous protein-truncating variants from human exome data. *Nat. Genet.* **49**, 806–810 (2017).
29. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
30. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).
31. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
32. Picard. http://broadinstitute.github.io/picard/.
33. Graubert, A., Aguet, F., Ravi, A., Ardlie, K. G. & Getz, G. RNA-SeQC 2: efficient RNA-seq quality control and quantification for large cohorts. *Bioinformatics* **37**, 3048–3050

(2021).

34. Middleton, R. *et al.* IRFinder: assessing the impact of intron retention on mammalian gene expression. *Genome Biol.* **18**, 51 (2017).
35. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
36. Jagadeesh, K. A. *et al.* Phrank measures phenotype sets similarity to greatly improve Mendelian diagnostic disease prioritization. *Genet. Med.* **21**, 464–470 (2019).
37. 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).
38. 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).