Chapter 4

Integrating healthcare and research genetic data empowers the discovery of 28 novel developmental disorders

4.1 Introduction

It has previously been estimated that 42-48% of patients with a severe developmental disorder (DD) have a pathogenic *de novo* mutation (DNM) in a protein coding gene [41, 140]. However, over half of these patients remain undiagnosed despite the identification of hundreds of dominant and X-linked DD-associated genes. This implies that there are more DD-associated genes left to find.

Whole genome and exome sequencing allows for direct identification of DNMs and statistical assessment of their contribution to DD. Associated genes are typically identified by observing a gene-specific enrichment of DNMs over what is expected by chance. Initially, when cohorts comprised of fewer than 200 trios, candidate DD-associated genes were proposed by identifying those with multiple non-synonymous DNMs across the cohort [157, 163, 58]. Improvements in the modelling of the mutation rate and increased sample sizes allowed for a more statistical approach. The most recent approaches taken in DD and autism cohorts have generally identified associated genes by separately evaluating the enrichment of protein truncating and missense DNMs compared to gene-specific mutation rates that account for gene length and sequence context[187, 215, 41, 31, 162]. This enrichment was calculated in some cases by assuming a Poisson model for the distribution of mutations and calculating this enrichment analytically[41, 215]. Other approaches were simulation based where a null distribution was calculated by simulating the observed number of mutations across all genes

using a multinomial distribution[31, 162]. McRae et al, in a previous paper from the Hurles group evaluating the role of DNMs in ~4,000 trios from the DDD study, also combined the missense enrichment test with a missense clustering test within genes. Clustering of mutations within protein structures is frequently observed for DNMs acting via activating or dominant negative mechanisms. Statistical testing for clustered mutations was initially developed in the study of somatic mutation enrichment in cancer[119, 172]. Somatic mutation rate variability can lead to false positive associations if locus-specific mutation rates are higher than expected. Methods have been developed to protect against this which combine the local observed synonymous mutation rate with a model including variables that predict the somatic mutation rate across the genome[141]. Existing methods to detect gene-specific enrichments of damaging DNMs typically ignore much prior information about which variants and genes are more likely to be disease-associated. Missense variants and protein-truncating variants (PTVs) vary in their impact on protein function [107, 186, 102, 112]. A study in a cohort of neurodevelopmental disorders observed that *de novo* PTVs that do not appear as standing variation in healthy population cohorts and fall in genes that exhibit patterns of strong selective constraint on heterozygous PTVs in the general population contribute to the majority of the enrichment of *de novo* PTVs in the cohort[112]. Known dominant DD-associated genes are also strongly enriched in this set of PTV constrained genes[125]. To identify the remaining DD genes, we need to increase our power to detect gene-specific enrichments for damaging DNMs by both increasing sample sizes and improving our statistical methods. In previous studies of pathogenic Copy Number Variation (CNV), utilising healthcare-generated data has been key to achieve much larger sample sizes than would be possible in a research setting alone [36, 32].

4.1.1 Chapter overview

To identify novel DD-associated genes, I integrated healthcare and research exome sequences on 31,058 DD parent-offspring trios. These were pooled from the Deciphering Developmental Disorders study, GeneDx (a US-based genetic diagnostic company) and Radboud University Medical Center. I developed a simulation-based statistical test to identify gene-specific enrichments of DNMs. Applying this to the dataset I identified 285 significantly DD-associated genes, including 28 not previously robustly associated with DDs. I then explored how these genes differed to those previously known to the field and examined the remaining burden of DNMs in genes yet to be associated with DD. Despite detecting more DD-associated genes than in any previous study, much of the excess of DNMs of protein-coding genes remains unaccounted for. To address this I built a model to estimate how many genes are left to be discovered. The model suggests that over 1,000 novel DD-associated genes

await discovery, many of which are likely to be less penetrant than the currently known genes.

4.1.2 Publication and contributions

The results described in this chapter have been submitted to the biorXiv preprint server and has recently been accepted to *Nature* [101]. This work was conducted as a collaboration between our group at the Wellcome Sanger Institute, GeneDx and Radboud University Medical Center (RUMC). I briefly summarise the various contributions to this project. Zhancheng Zhang called *de novo* mutations in GeneDx data, Kevin J. Arvai and Rebecca Torene performed the phenotypic comparison work, Stefan H. Lelieveld called and filtered *de novo* mutations in RUMC. Kaitlin Samocha (post-doc in the Hurles group) and I worked jointly on many aspects of this project. In this chapter I have mostly included analyses that I performed and otherwise have been explicit about who performed them. All of this work was done under the supervision of Christian Gillissen (RUMC), Kyle Retterer (GeneDx) and Matthew E. Hurles.

4.2 Methods

4.2.1 Sample collection and individual quality control

DDD

Patients with severe, undiagnosed developmental disorders were recruited from 24 regional genetics services within the United Kingdom National Health Service and the Republic of Ireland. Families gave informed consent to participate, and the study was approved by the UK Research Ethics Committee (10/H0305/83 granted by the Cambridge South Research Ethics Committee, and GEN/284/12 granted by the Republic of Ireland Research Ethics Committee). Additional details on sample collection, exome sequencing, alignment, variant calling (inherited and *de novo*) and variant annotation have been described previously [41]. These analyses involve 9,858 trios from 9,307 families, a subset of whom have been analyzed in previous publications [215, 41].

GeneDx

Patients were referred to GeneDx for clinical whole-exome sequencing for diagnosis of suspected Mendelian disorders as previously described[177]. Patient medical records were abstracted into HPO terms using Neji concept recognition[21] with manual review by laboratory genetic counselors or clinicians. Patients were selected for inclusion in this study based

on having one or more HPO phenotypes overlapping the inclusion criteria for the DDD study [215]. The study was conducted in accordance with all guidelines set forth by the Western Institutional Review Board, Puyallup, WA (WIRB 20162523). Informed consent for genetic testing was obtained from all individuals undergoing testing, and WIRB waived authorization for use of de-identified aggregate data. Individuals or institutions who opted out of this type of data use were excluded.

The sequencing and variant calling was done by collaborators at GeneDx. The exomes were sequenced and aligned as previously described [177] with either SureSelect Human All Exon v4 (Agilent Technologies, Santa Clara, CA), Clinical Research Exome (Agilent Technologies, Santa Clara, CA), or xGen Exome Research Panel v1.0 (IDT, Coralville, IA) and sequenced with either 2x100 or 2x150bp reads on HiSeq 2000, 2500, 4000, or NovaSeq 6000 (Illumina, San Diego, CA). Alignment BAM files were then converted to CRAM format with Samtools version 1.3.1 and indexed. Individual GVCF files were called with GATK v3.7-0 HaplotypeCaller [143, 43] in GVCF mode by restricting output regions to plus/minus 50bp of the RefGene primary coding regions. Single-sample GVCF files were then combined into multi-sample GVCF files with each combined file contained 200 samples. These multi-sample GVCF files were then joint-genotyped using GATK GenotypeGVCFs. The cohort of 18,789 trios was joint-genotyped in two separate batches, one with 10,138 trios and the other 8651 trios. GATK VariantRecalibrator (VQSR) was applied for both SNPs and INDELs, with known SNPs from 1000 Genomes phase 1 high confidence set and "gold standard" INDELs from Mills et al [149].

Variants in VQSR VCF files were annotated with Ensembl Variant Effect Predictor (VEP)[144] using RefSeq transcripts. The transcript with the most severe consequence was selected, and all associated VEP annotations were based on the predicted effect of the variant on that particular transcript. Variants called in the proband and not in the parents were selected as potential *de novo* mutations. Filtering of these *de novo* mutations is described below.

Radboud University Medical Center

The Department of Human Genetics from the Radboud University Medical Center (RUMC) is a tertiary referral center for clinical genetics. Approximately 350 individuals with unexplained intellectual disability (ID) are referred annually to the clinic for diagnostic evaluation. Since September 2011 whole exome sequencing (WES) is part of the routine diagnostic work-up aimed at the identification of the genetic cause underlying disease[158]. For individuals with unexplained ID, a family-based WES approach is used which allows the identification of *de novo* mutations (DNMs) as well as variants segregating according to

other types of inheritance, including recessive mutations and maternally inherited X-linked recessive mutations in males [39]. For this study, RUMC selected all individuals with ID who had family-based WES using the Agilent SureSelect v4 and v5 enrichment kit combined with sequencing on the Illumina HiSeq platform in the time period 2013-2018. This selection yielded a set of 2418 individual probands, including 1040 females and 1378 males across 2387 different families. The level of ID ranged between mild (IQ 50-70) and severe-profound (IQ<30).

Families gave informed consent for both the diagnostic procedure as well as for forthcoming research that could result in the identification of new genes underlying ID by meta-analysis, as presented here.

The sequencing and variant calling was done by collaborators at RUMC. The exomes of 2418 patient-parent trios were sequenced, using DNA isolated from blood, at the Beijing Genomics Institute (BGI) in Copenhagen. Exome capture was performed using Agilent SureSelect v4 and v5 and samples were sequenced on an Illumina HiSeq 4000 instrument with paired-end reads to a median target coverage of 112x. Sequence reads were aligned to the hg19 reference genome using BWA version v0.7.12 and duplicate marking by Picard v1.90. Variants were subsequently called by the GATK haplotypecaller (version v3.4-46).

The diagnostic WES process as outlined above only reports (*de novo*) variants that can be linked to the individuals' phenotype. In this study, we systematically collected all DNMs in regions in or close to (200 bp) a capture target. DNMs were called as described previously [39]. Briefly, variants called within parental samples were removed from the variants called in the child. For the remaining variants pileups were generated from the alignments of the child and both parents. Based on pileup results variants were then classified into the following categories: "maternal (for identified in the mother only)", "paternal (for identified in the father only)", "low coverage" (for insufficient read depth in either parent), "shared" (for identified in both parents)", and "possibly *de novo*" (for absent in the parents). Variants classified as possibly *de novo* were included in this study.

Various quality filters were applied to ensure that only the most reliable calls were included in the study, these are described in the quality control section below.

4.2.2 Definition of diagnostic lists

For various analyses in this work, a list of genes already known to be associated with developmental disorders was needed. In order to define this list, diagnostic gene lists were collected from each center to create sets of "consensus" and "discordant" genes.

For the DDD cohort, the Developmental Disorders Genotype-Phenotype Database (DDG2P) list was used. This is a curated list of genes specifically associated with de-

velopmental disorders. For every gene on the list, DDG2P provides the level of certainty, consequence of the mutation, and allelic status of variants associated with developmental disorders. We downloaded the DDG2P list on 22 September 2019 from https://decipher.sanger.ac.uk/info/ddg2p. In order to define diagnostic genes that act in a dominant fashion, the genes were subsetted to include only genes that were considered "probable", "confirmed", or "both RD and IF" (i.e. high levels of certainty of being a true DD-associated gene) and had an allelic status of "monoallelic", "x-linked dominant", "hemizygous", or "imprinted".

GeneDx maintains a continually curated list of genes, used to define reporting categories for clinical exome and genome testing, which have been definitively or putatively implicated in human Mendelian disease, with modes of inheritance noted for each gene. Starting with the January 2020 curation list, those genes with dominant modes of inheritance and definitive implications in disease were manually reviewed to remove any genes with no association to developmental disorders either because of no phenotypic overlap with the inclusion criteria for this study or because the relevant phenotypes were adult onset.

For the list from RUMC, gene panels for intellectual disability, epilepsy, and craniofacial anomalies/Multiple congenital anomalies were designed by multidisciplinary expert teams consisting of a clinical laboratory geneticist, a molecular geneticist, and a clinical specialist. Each set contained all genes known to be associated with the disease. The gene panel version from December 2019 (DGD-2.17) was used. From each of the three gene lists, the genes were subsetted to those with a reported inheritance of “AD”, “AD,AR”, “AD,IMP”, “AD/AR”, “xl”, “XL”, “XLD”, “XLR,XLD”, or “XLR/XLD”.

After mapping to HGNC IDs and symbols, any gene that was considered diagnostic by all three centers was designated as a “consensus” gene (n=380). For genes on one or two of the diagnostic lists, we considered them “discordant” genes (n=607).

4.2.3 Joint quality control of datasets

De novo mutation filtering

I applied the following filters specifically to each center. I chose these to minimise the number of false positive DNMs. I evaluated this by looking at various plots and metrics. These included plots of the variant allele fraction across each cohort to ensure this was centred symmetrically around 0.5, the distribution of the number of exome DNMs per person, the mutational spectra of the DNMs, the distribution of the size of indels and the ratios of insertion/deletion, frameshift/nonsense and frameshift/inframe variants. I compared these across each cohort to ensure they were comparable. The VAF distributions pre and post

filtering VAF distirbutions are shown in Figure 4.1 . The DNMs from the DDD dataset in the 'pre-filtering' had already undergone a basic set of the filters described below such as read depth, strand bias and number of parental alt alleles. The DNMs from RUMC and GeneDx had undergone prior filtering which are specified below. The filters were applied in this specific order.

DDD:

- Autosomes
  - I applied the following filters as base filters before attempting to harmonize the three datasets (these are applied to the 'pre-filtering' set as shown in Figure 4.1).
    - The minor allele frequency (MAF) < 0.01 across all DDD samples, Exome Aggregation Consortium (ExAC)[125], and 1000 Genomes[34] populations
    - Read depth (RD) of child > 7, mother RD > 5, father RD > 5
    - Fisher exact test on strand bias p-value > 10^{-3}
    - Remove DNM if any two of the following conditions are met:
    - Both parents had ≥ 1 supporting the alternative allele
    - There is an excess of parental alternative allele within the cohort at the DNMs position. This is defined as p-value $< 10^{-3}$ under a one-sided binomial test given an expected site error rate of 0.002
    - There is an excess of alternative alleles within the cohort for DNMs in a gene. This is defined as p-value < 10^{-3} under a one-sided binomial test given an expected site error rate of 0.002
  - Filter only applied to indels; remove indel if all three conditions are met:
    - Variant allele frequency (VAF) in child < 0.2
    - MAF > 0 for any of the following cohorts: across all DDD samples, ExAC, 1000 Genomes populations
    - Size of indel < 5 bp
  - Posterior probability of being a *de novo* mutation (output from DeNovo Gear) > 0.00781 for autosomal DNMs. These thresholds have been determined through earlier work such that the observed number of synonymous DNMs match the expected number [41, 140].
  - Filter out mutations in sites with more than one mutation with VAF < 0.3
- X chromosome: DeNovoGear was run as previously described, but with a different set of hard filters to account for the lower coverage in males and to maximise sensitivity

and specificity. These were developed by Hilary Martin. All candidate DNMs in males and a large subset of those in females were inspected manually in IGV[182], and this was used to settled on the following set of filters:

- Removed DNMs in the pseudoautosomal regions.
- The variant had to be called heterozygous or, for males, hemizygous in the child in the original GATK calls, and called homozygous reference in the parents.
- Removed variants in segmental duplications.
- For male probands, the depth requirements were as follows: in the child, alternate allele depth > 2 and RD > 2; in the mother, RD > 5
- For female probands,the depth requirements were as follows: in the child, alternate allele depth > 2, RD > 7; in the mother, RD > 5; in the father, RD > 1
- For single nucleotide variants, a p>10^{-3} was required on a Fisher's exact test for strand bias, pooling across trios (ignoring fathers of male probands) where a *de novo* was called at the same site by DeNovoGear.
- For female probands, indels < 5 bp were removed if they had VAF < 0.3 or MAF > 0, since these were vastly over-represented and seemed to be a common error mode.
- Removed DNMs if any two of the following conditions were met (these conditions were applied separately for males and females):
  - Lowest alternative read count for the parents (or, for males, the mothers) is higher than the maximum allowable given the depth, an error rate of 0.002, and a probability threshold of 0.98 (using the mindepth function in DeNovoFilter)
  - An excess of parental (or, for males, maternal) alternative alleles with a putative *de novo* at that site, defined as p-value < 10^{-3} under a one-sided binomial test given an expected error rate of 0.002
  - An excess of parental (or, for males, maternal) alternative alleles with a putative *de novo* in the same gene, defined as p-value > 10^{-3} under a one-sided binomial test given an expected error rate of 0.002
- Implemented a cutoff for the ppDNM from DeNovoGear to > 0.00085 based on matching the observed number of synonymous DNMs in females to the expected number

GeneDx:

- The following filters were initially applied by collaborators at GeneDx:
  - RD > 10 for child, mother, and father
  - VAF > 0.15 for child for SNVs and VAF > 0.25 for indels
  - More than 3 reads supporting the alternative allele
  - Genotype Quality (GQ) score > 40
  - Phred-scaled p-value using Fisher's exact test to detect strand bias < 30
  - Log odds of being a true variant versus being false from VQSR > -10 outputted from GATK
  - Any variant with general population frequency above 0.01 was also excluded based on 1000 Genomes and Exome Aggregation Consortium (ExAC) variant population frequency data
  - Filtered out *de novo* variants called > 4 times in the parental samples in the cohort
- Filter out DNMs with VAF < 0.3 and VQSLOD < 7 where VQSLOD is the log odds ratio of being a true variant outputted from GATK
- Filter out *de novo* indels > 100 bp
- Filter out DNMs, not on chromosome X, with a VAF of 1
- Filter out 5 individuals with more than 10 coding DNMs. These appeared to be due to relatively poor sample quality.

RUMC:

- The following filters were initially applied by collaborators at RUMC:
  - Minimal number of variant reads: 10
  - Minimal number of total reads: 20
  - Minimal percentage of variant reads: 20%
  - Frequency in dbSNP < 0.1%
  - Coverage in parents of at least 10 reads
  - 15 complex variants were discarded after manual inspection in IGV
- GATK Quality score > 450

The following filters were applied across all three datasets:

- Removed DNMs outside coding regions
- Removed mutations that fell within known segmental duplication regions as defined by UCSC (http://humanparalogy.gs.washington.edu/build37/data/GRCh37GenomicSuperDup.tab)
- Keep only the most severe DNM in each gene per individual according to the consequence severity from VEP[144]

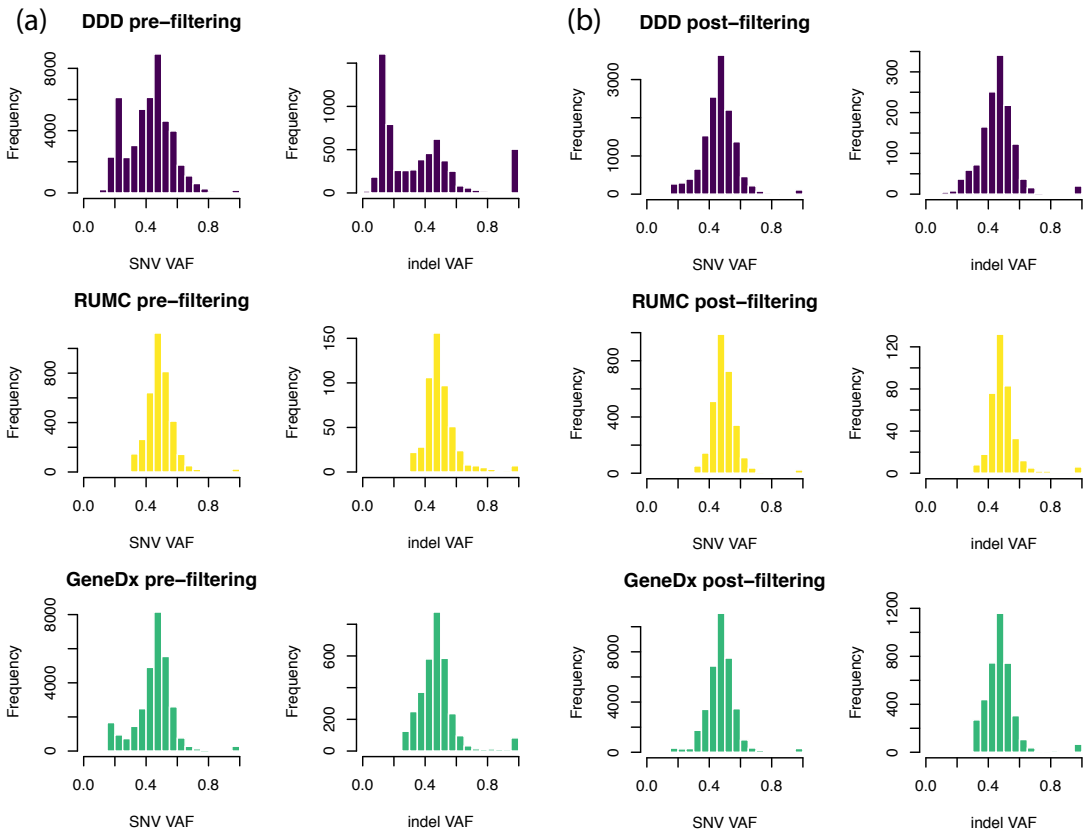

Fig. 4.1 Variant allele fraction (VAF) for all *de novo* coding mutations, split by center pre and post filtering. The pre-filtering sets did include base filters as specified in the Methods. Note that mutations with a VAF of 1 are hemizygous *de novo* mutations in males. DDD = Deciphering Developmental Disorders. RUMC = Radboud University Medical Center.

Duplicate Samples

50 common exonic SNPs were selected (only 47 of which could be evaluated across all three centers) and collected genotypes at these SNPs for every sample with a *de novo* mutation

found in another individual in the joint set (n=781 DDD, 1307 GeneDx, and 164 RUMC). Kaitlin Samocha then used the gtcheck function from bcftools (https://samtools.github.io/bcftools/bcftools.htm to find discordance between each pair of samples. Pairs with low discordance were manually confirmed, leading to a total of 8 duplicate samples identified. One individual from each duplicate pair was removed from the analyses, leaving 31,058 samples for downstream analyses.

Removing variants from siblings

Siblings will sometimes share DNMs that arose as the same mutational event in one of their parental germlines. To avoid double counting these shared DNMs, Kaitlin identified siblings in each cohort and one of each pair of shared variants was randomly removed. In total, 11 DNMs found in siblings were removed.

After filtering I had a total set of 45,221 coding DNMs (these are available online as supplementary table 1 with the bioRxiv preprint [101]). The filtered VAFs across the three datasets are displayed in Figure 4.1b.

4.2.4 DeNovoWEST framework

DeNovoWEST (*De novo* Weighted Enrichment Simulation Test) is the testing framework I developed and applied to assess gene-wise *de novo* mutation enrichment. Each observed DNM in our dataset was assigned a mutation severity score. This severity score is a proxy for how deleterious we expect the mutation to be. Details of how these were calculated are given below. For each gene I then calculated a gene severity score which is the sum of severity scores for all mutations that fall into that gene. There are two modes of enrichment testing within DeNovoWEST: the overall enrichment test which includes all variant consequences and the 'altered-function' specific test which assesses enrichment and clustering of missense variants only. An overview of the method is given in Figure 4.2.

Enrichment Test

I used a simulation-based approach to evaluate whether these observed gene severity scores are higher than what we would expect under the null hypothesis of no *de novo* mutation enrichment. To calculate the probability of observing a gene severity score that is as or more extreme than the one that we observe for this gene (the p-value) we considered the case of observing k number of DNMs in the gene where k ranged from 0 to 250. This upper limit was chosen as it is far above the number of DNMs seen in any individual gene in the dataset and the probability of observing more than that number of DNMs for our cohort in a single gene

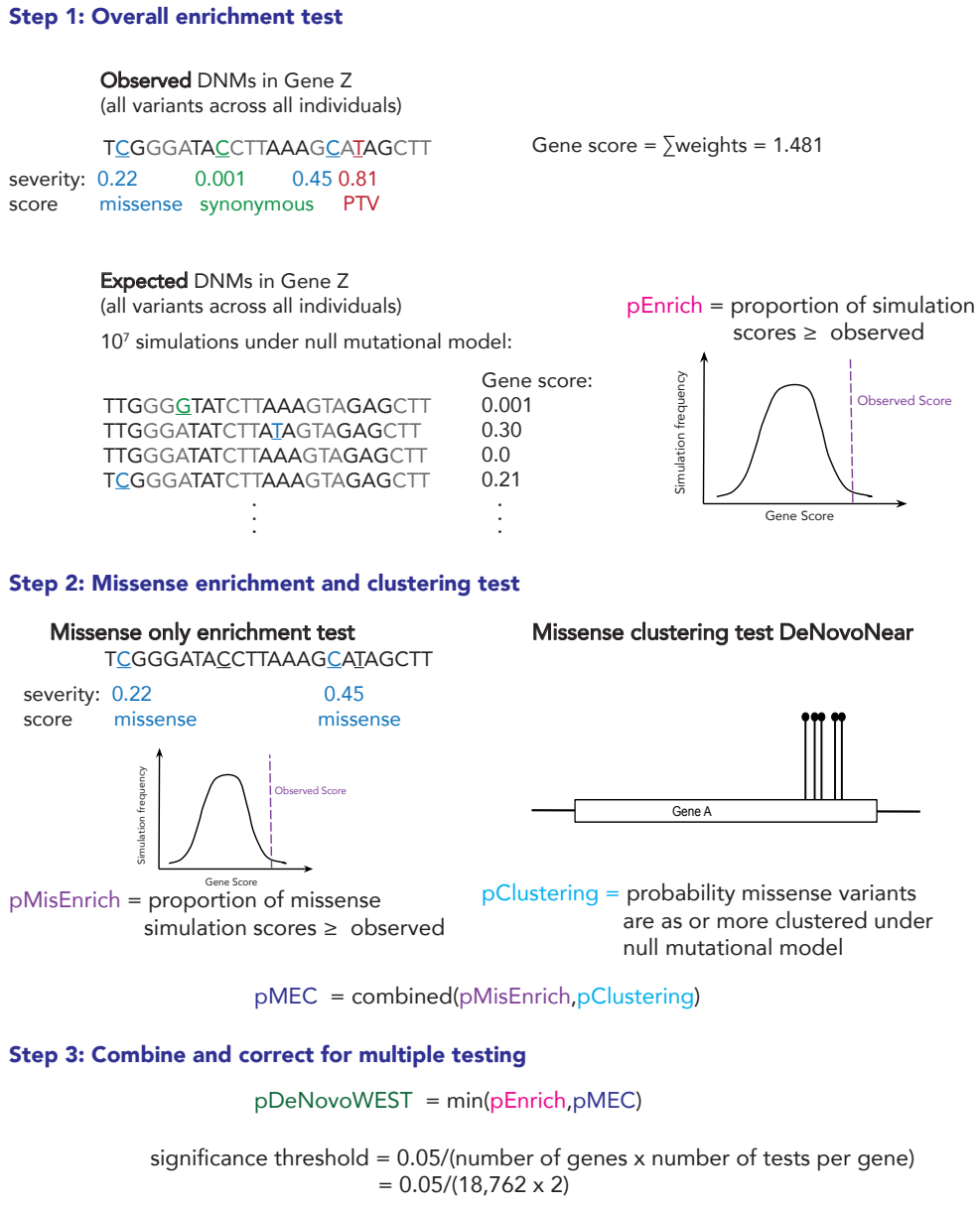

Fig. 4.2 Overview of DeNovoWEST method

is negligible with respect to our significance threshold. The enrichment p-value, pEnrich, was then calculated as the sum across all k of the product of the probability of observing k mutations and the probability of obtaining a gene severity score greater than the one observed in our data. These probabilities are summarized by the following equation where S denotes the gene score, s is the observed gene score and K is the number of DNMs in the gene:

$$P(S \geq s) \approx \sum_{k=0}^{250} P(S \geq s|k)P(K = k)$$

$$= P(K = 0)P(S \geq s|K = 0) + P(S \geq s|K = 1)P(K = 1) + \sum_{k=2}^{250} P(S \geq s|k)P(K = k)$$

These probabilities were calculated as follows:

- $P(S \geq s|K = 0)$ = 0 (unless $s = 0$ in which case we would not be testing this gene as it would have no observed DNMs).
- I also analytically calculated $P(S \geq s|K = 1)P(K = 1)$, the probability that the severity score of 1 mutation was greater than what we observed. This was calculated using the mutability of each position and the annotated score for that mutation.
- I then used a simulation-based approach to calculate $\sum_{k=2}^{250} P(S \geq s|k)P(K = k)$, the probability of observing an as, or more, extreme gene score if we see 2 to 250 DNMs in this gene. I calculated $P(K = k)$ analytically assuming the DNMs followed a poisson distribution. This was calculated based on our sample size (N) and the mutation rate of the gene assuming mutations follow a Poisson distribution with $\lambda = 2\mu_{gene}N$. This was adjusted to $\lambda = \mu_{gene}(N_{males} + 2N_{females})$ for genes that fall in the X chromosome. I then multiplied this with a simulation based estimate for $P(S \geq s|k)$, the probability of observing a gene score greater or equal to the one we observe. To calculate the latter probability I simulated the distribution of gene scores as follows:

    1. Simulate k DNMs across the gene. The probability of mutation was weighted by the trinucleotide sequence context at every base position in that gene. These probabilities were taken from Samocha et al [187].
    2. Assign the simulated *de novo* mutations a mutation severity score
    3. Sum the simulated mutation severity scores to get the simulated gene severity score
    4. I performed $10^9 \times P(K = k)$ simulations for every k. This number was chosen as I wanted to run the smallest number of simulations possible to obtain a robust p-value. By distributing these simulations across the mutations this was the equivalent of 10^9 simulations per gene and meant that the p-value was robust to stochasticity far below the p-value threshold.

Determination of Weights

I calculated the weights used in the DeNovoWEST test from observed enrichments across mutation consequence classes, s_{het} values, missense constraint information and, for some, CADD score bins (version 1.0)[107]. s_{het} refers to the estimated selection coefficient of heterozygous PTVs [24]. Genes were stratified into two groups of 'high' s_{het} and 'low' s_{het}. High s_{het} genes were defined as those with a $s_{het} \geq 0.15$ and low s_{het} genes as those with an estimate <0.15. This threshold was suggested by Cassa et al. Enrichments were calculated by dividing the number of observed *de novo* mutations by the number of expected *de novo* mutations across all sites in the exome that fell into a specific strata. I calculated the number of expected mutations given our sample size and the triplet context mutation rates at the sites [187]. Details for the weight calibration for each consequence class are given below:

- Missense mutations were stratified based on whether they fell in a low or high s_{het} gene [24], whether or not they fell into a region of missense constraint [186], and finally into CADD score bins of size 6[107]. I fit four LOESS lines on the enrichments for mutations that were in high s_{het} genes + missense constrained regions, high s_{het} genes + not in a missense constrained region, low s_{het} genes + missense constrained regions, low s_{het} genes + not in a missense constrained region.

- Nonsense mutations were stratified based on whether they fell in a low or high s_{het} gene, and then into CADD score bins of size 15 for the high s_{het} genes and CADD score bins of 7.5 for low s_{het} genes. Two LOESS lines were fit on the enrichments for mutations that were in high s_{het} genes vs those in low s_{het} genes.

- Synonymous mutations were stratified based on whether they fell in a low or high s_{het} gene.

- Canonical splice site mutations were stratified based on whether they fell in a low or high s_{het} gene.

- Inframe indels were assigned weights based on the overall enrichment of missense mutations as an appropriate approximation for their deleteriousness. These were stratified by whether they fell in a low or high s_{het} gene but not stratified by CADD score bins.

- Frameshift indels were assigned the same weights as nonsense mutations with a CADD score ≥ 45 and whether they fell in a low or high s_{het} gene.

These enrichments are depicted in Figure 4.3. The enrichment values for each stratum were normalised by the level of synonymous enrichment and converted into a positive predictive value (PPV) using the following formula:

$$PPV = \frac{OR - 1}{OR} \tag{4.1}$$

The synonymous variants were artificially given a PPV of 0.001 as we would expect 1 in 1000 synonymous DNMs in our cohort to be pathogenic according to how many are estimated to be cryptic splice sites[91]. The PPV weights are depicted in Figure 4.3.

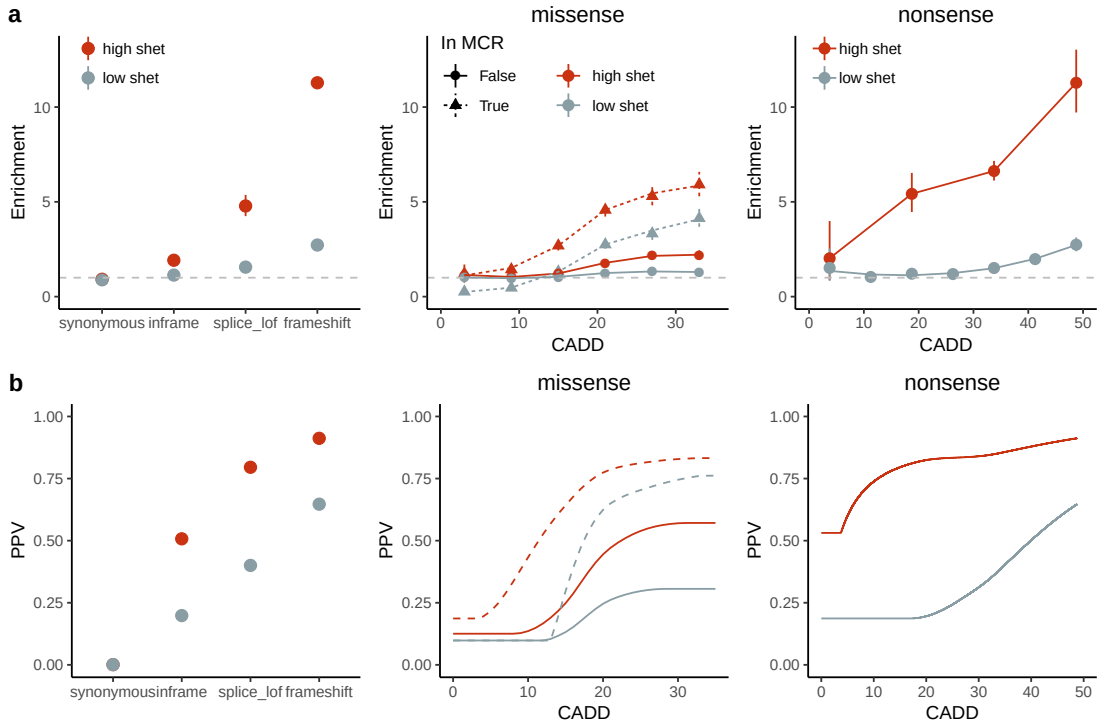

Fig. 4.3 Enrichment of consequence classes and corresponding PPV weights used for DeNovoWEST test. (a) Depicts the observed enrichment in each consequence class with 95% confidence intervals. The lines fit in the missense and nonsense class are LOESS fits. (b) Depicts the PPV derived from these observed enrichments. In all these plots points and lines are colored red if the variants occurred in a gene with a high s_{het} value (≥ 0.15)[24] or gray if the gene had a s_{het} value < 0.15 ("low s_{het}" genes). For missense variants, dashed lines indicate that the variants fell within missense constrained regions (MCR) while solid lines fell outside of MCRs.

Missense enrichment and clustering test

This test is geared to detect genes that may be acting via an altered-function mechanism, such as a gain-of-function. The test consists of two parts. The first is implementing the enrichment test (as described above) but only considering missense variants to obtain a pMisEnrich p-value. The second part consists of a missense clustering test. I assessed clustering of missense *de novo* mutations within genes to identify genes where DNMs may be acting through dominant negative or activating mechanisms. For this I used DeNovoNear, which has been described previously [41] and is available on Github (https://github.com/jeremymcrae/denovonear). I refer to this clustering p-value as pClustering. I combined pMisEnrich and pClustering using Fisher's method to obtain pMEC. To ensure Fisher's method was appropriate I confirmed that these two p-values were independent by simulating DNMs under the null for ~60,000 genes and found a nonsignificant correlation between pMissenseEnrich and pCluster (ρ = -0.01, p-value 0.08).

Combining Tests

I combined the results from the overall enrichment test and the missense enrichment/clustering test by taking the minimum of the two p-values as follows:

$$pDenovoWEST = min(pEnrich, pMEC) \tag{4.2}$$

To correct for multiple testing, I used a Bonferonni corrected significance threshold of 0.05/(18762 * 2) for pDeNovoWest. This accounts for testing all 18,762 genes that I was able to conduct tests for and for 2 tests per gene: the overall enrichment test and the specific gain-of-function test. The final set of DeNovoWEST p-values are listed online as supplementary table 2 of the bioRxiv preprint [101].

4.2.5 Functional similarity between new and known genes

To compare the functional similarity between the consensus and novel genes I looked across various properties that have been known to be important in classifying haploinsufficiency[83]. The details of each variable we have used are detailed below:

- Somatic driver gene: a binary variable of whether the gene is a known somatic driver gene[141].
- Median reads per kilobase million (RPKM) fetal brain: the median RPKM in the fetal brain taken from BrainSpan[148].

- Relevant GO term: a binary variable of whether the gene was annotated with one of twenty GO terms that were enriched in consensus DD genes. To select these terms I annotated all genes with GO terms and looked at the enrichment of each GO term between consensus DD genes and non-DD genes (genes that are not significant in our analysis and are not on either the discordant or consensus genes lists). I defined relevant GO terms as the top 20 most enriched terms that appear in at least 20 of the 380 consensus genes. This was to ensure that I was picking terms that were generalisable to the entire set were not specific to only a few genes. The terms selected are detailed in Table 4.1. At least one of these 20 terms were present in 237 (71%) of consensus genes and 2,874 (16%) of non-DD genes.
- Network distance to consensus genes: As in Huang et al[83], Kaitlin created a protein-protein interaction network by integrating information from the Human Protein Reference Database[104], STRING[211], and Reactome[38]. Kaitlin then calculated the shortest path distance (a measure of proximity) between each gene and consensus genes.
- Network degree and betweenness: Kaitlin used MCL[220] (version 14-137; https://micans.org/mcl/) to determine network degree and betweenness (both measures of centrality).
- Promoter GERP and Coding GERP: These were calculated as described in Huang et al. [83]
- Macaque dN/dS: downloaded from Ensembl.

I compared the mean of these variables across consensus and novel genes compared to non-DD genes (Figure 4.7b).

4.2.6 DNM enrichment in non-significant genes

I calculated the remaining DNM burden in the genes that were not significantly associated with developmental disorders (DD) in our analysis and were not consensus DD genes. There were 2,172 genes that were not associated with DD and had a pLI $\geq$ 0.9 (high pLI) and 10,472 genes with a pLI < 0.9 (low pLI)[102]. The burden was calculated by calculating the observed/expected number of DNMs across the four groups of genes categorised by both missense/PTV mutations and low/high pLI. I then repeated the analysis removing nominally significant genes (unadjusted p-value > 0.05).

GO ID	GO name	Number of non-DD genes with GO term	Number of consensus DD genes with GO term	Enrichment of GO term in consensus vs non-DD genes
GO:0001501	skeletal system development	119	31	12.577748
GO:0007507	heart development	161	33	9.896377
GO:0007605	sensory perception of sound	107	21	9.47597
GO:0008543	fibroblast growth factor receptor signaling pathway	138	24	8.396926
GO:0001701	in utero embryonic development	217	37	8.23247
GO:0044212	transcription regulatory region DNA binding	168	28	8.047054
GO:0003682	chromatin binding	326	53	7.84958
GO:0010628	positive regulation of gene expression	155	24	7.475972
GO:0007411	axon guidance	295	44	7.201431
GO:0000122	negative regulation of transcription from RNA polymerase II promoter	535	76	6.858797
GO:0043234	protein complex	323	42	6.278197
GO:0045893	positive regulation of transcription, DNA-dependent	512	66	6.223893
GO:0007268	synaptic transmission	359	45	6.052102
GO:0007420	brain development	200	25	6.03529
GO:0005667	transcription factor complex	220	27	5.925558
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	711	85	5.772148
GO:0007399	nervous system development	285	33	5.590585
GO:0003713	transcription coactivator activity	223	25	5.412816
GO:0008134	transcription factor binding	281	30	5.154696

Table 4.1 Table showing the GO terms selected as being relevant to consensus DD genes

4.2.7 Modelling remaining PTV DNM burden

I modelled the remaining DNM burden separately for PTV and missense mutations.

Model for PTV DNM burden

I simulated the number of *de novo* PTVs across a range of numbers of remaining haploinsufficient genes and the PTV enrichment observed in these genes. The PTV model made the following assumptions:

- PTV enrichment was the same across all remaining undiscovered HI DD-associated genes.
- All undiscovered HI DD-associated genes have the same level of penetrance
- The likelihood of being an undiscovered HI DD-associated gene was $1 - power$, where power was calculated as described below
- The probability of a currently unassociated gene being selected as a HI DD gene was higher if the gene was intolerant of loss-of-function variation. Specifically the likelihood was multiplied by the observed relative likelihood of being a DD-associated gene (significant in our analysis or a consensus gene) for genes with pLI ≥ 0.9. This was defined as:

$$\frac{P(DD\ associated\ gene|pLI \geq 0.9\ \&\ power > 0.8)}{P(DD\ associated\ gene|pLI < 0.9\ \&\ power > 0.8)} \tag{4.3}$$

- The PTV enrichment for known DD-associated genes (significant in the analysis or consensus) was taken as the observed enrichment in our cohort
- I ignore missense variants that may act as loss-of-function

Power as referred to in the above assumptions was calculated as follows. Calculating power tailored specifically according to our *de novo* enrichment test was challenging as I would have had to make assumptions about the distribution of mutation consequences according to undiscovered DD-associated genes. Even with these assumptions, the calculation would be computationally intensive given the simulation based framework. I decided to calculate power in the context of the enrichment of PTV mutations. Therefore this power measure is specifically the power to detect haploinsufficient genes. Power was defined as the power to detect the median PTV enrichment in known monoallelic DD-associated genes. The set of known monoallelic DD-associated genes was defined as 163 genes in the consensus

gene list that had at least one observed *de novo* PTV in our dataset. I calculated the PTV enrichment by dividing the observed number of *de novo* PTVs in each gene by the expected number of *de novo* PTVs as defined by the gene-specific mutation rate. The median of the PTV enrichment distribution was 34.0. Power was then calculated as the probability of observing a significant p-value under the Poisson test assuming the median PTV enrichment.

To assess the likelihood of the model I calculated the following across the distribution of all 200 simulations per scenario:

$$\begin{aligned} Likelihood = & P(6,861\ PTV\ DNMs)P(147\ significantly\ PTV\ enriched\ genes) \\ & \times P(2,929\ genes\ with\ PTV\ enrichment > 2) \end{aligned}$$

This captures three essential parts of the distribution of observed *de novo* PTVs: the total number of *de novo* PTVs, the number of genes that are currently significantly enriched for PTVs in our cohort and the number of genes that are not significant but have an elevated PTV enrichment (> 2). I explored other properties to characterise the distribution such as the number of genes with an enrichment p-value below a larger nominal threshold (eg 0.05) however the PTV enrichment appeared to be perform better. I evaluated this by examining distributions of PTV counts per gene and seeing if the most likely scenario using different metrics was similar to the observed distribution. Using this approach I also explored altering the threshold of PTV enrichment from a range of 1.5 to 5.

Model for missense DNM burden

The set up for the model for missense mutations was very similar to the PTV model with a few key differences. I simulated the number of *de novo* missense variants across a range of numbers of genes with a pathogenic DD-associated variant, mean missense enrichments in these genes and distributions of missense enrichment. I included a third dimension to the likelihood space which allowed me to model the distribution of missense enrichment across the genes with a pathogenic DD-associated missense variant.

I modelled the distribution of missense enrichment using the gamma distribution. I used 6 different shape parameters to represent different scenarios (0.1,0.5,1,510,20) (Figure 4.13 a). Missense mutations acting via gain-of-function mechanisms are likely to act on a small mutational target within a gene and have a small gene-wise missense enrichment. A shape parameter of 0.1 represents a model where most genes with pathogenic missense DNMs are acting via gain-of-function mechanisms and have generally smaller missense enrichment values. Missense mutations acting via loss-of-function mechanisms in haploinsufficient genes will have a larger mutational target and so the enrichment values will tend to be

larger. A shape parameter of 20 represents the model where most genes with pathogenic missense variants are acting via loss-of-function and have larger enrichments. The other major difference from the PTV model was that the likelihood of being selected as a DD-associated gene in the simulations was not a function of pLI. It was only proportional to $1 - power$.

To assess the likelihood of the model I calculated the following across the distribution of all 200 simulations per scenario:

$$\begin{aligned} Likelihood = & P(27,139 \textit{ missense DNMs}) P(130 \textit{ significantly missense enriched genes}) \\ & \times P(3,764 \textit{ genes with missense enrichment} > 2) \end{aligned}$$

4.2.8 Expression in fetal brain

I defined expression in the fetal brain for a gene as having a median RPKM > 0 in the BrainSpan dataset[19]. I then compared the proportion of genes expressed in the fetal brain between the genes that were significant and not significant in our analysis . I then subsetted genes into those with a high pLI (pLI $\geq$ 0.9) and repeated this analysis.

4.3 Results

4.3.1 Improved statistical enrichment test identifies 300 significant DD-associated genes

Following clear consent practices and only using aggregate, de-identified data, DNMs in patients with severe developmental disorders were pooled from three centres: GeneDx (a US-based diagnostic testing company), the Deciphering Developmental Disorders study, and Radboud University Medical Center. I performed stringent quality control on variants and samples. I adjusted filters on read depth, VAF and *de novo* calling quality scores individually for each cohort. These were chosen to ensure the distribution of DNMs per person, mutational spectra and VAF distribution were consistent across the three datasets. After filtering I obtained 45,221 coding and splicing DNMs in 31,058 individuals (Figure 4.4), which includes data on over 24,000 trios not previously published. These DNMs included 40,992 single nucleotide variants (SNVs) and 4,229 indels.

All three cohorts are comprised of individuals with severe developmental disorders. The cohorts had comparable rates of male to female probands (55-57% male cohorts; Figure 4.4a) as well as similar DNM rates (average 1.81-1.96 per individual exome). The DDD study

has a significantly higher rate of synonymous DNMs (0.31 *de novo* synonymous DNMs per exome) compared to individuals from GeneDx (GDX; 0.28 per exome; Poisson rate test p = 5.2×10^{-7}) or Radboud University Medical Center (RUMC; 0.28 per exome; Poisson rate test p = 0.0132), which is likely due to differences in *de novo* identification pipelines (Figure 4.4b). Specifically, as described in McRae et al[41] and mentioned in the methods, the DDD study selected a ppDNM (posterior probability of a *de novo* mutation) threshold such that the observed number of synonymous DNMs matched the expected number under a null germline mutation model.

All three cohorts have far more carriers of nonsynonymous DNMs in the consensus genes than expected based on a null mutational model (Figure 4.1c). The specific rate of such carriers differs between cohorts, with GeneDx showing the lowest fraction of such cases, which can be explained by the varying ascertainment between centers.

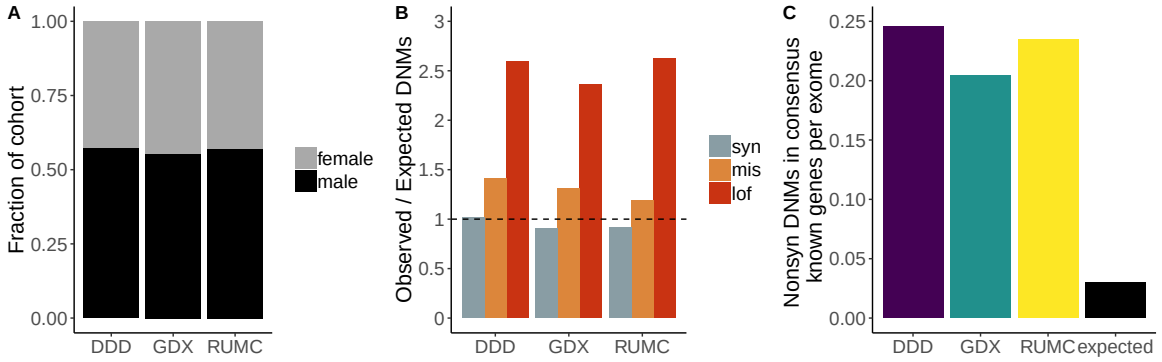

Fig. 4.4 Comparing cohorts from the three centers. A) Fraction of each cohort that is female (gray) vs male (black). B) Enrichment of observed *de novo* mutations compared to the expected number from a sequence-context based mutational model [187]. C) Rate of nonsynonymous *de novo* mutations (excluding inframe indels) in consensus genes in each cohort as well as the expected rate based on the aforementioned mutational model. DDD = Deciphering Developmental Disorders. GDX = GeneDx. RUMC = Radboud University Medical Center. syn = synonymous. mis = missense. lof = loss-of-function (including nonsense/stop gained, essential splice site, and frameshift variants).

To detect gene-specific enrichments of damaging DNMs, I developed a method named DeNovoWEST (*De Novo* Weighted Enrichment Simulation Test, described in detail in Methods; https://github.com/queenjobo/DeNovoWEST). DeNovoWEST assigns each observed DNM in our dataset a mutation severity score. This severity score is based on the empirically estimated positive predictive value of being pathogenic (Figures 4.2,4.3). For each gene the observed severity scores are then summed to obtain a gene score. Enrichment tests are then performed by comparing this observed gene score to a simulated null distribution. I

performed two tests per gene: the first is an enrichment test on all nonsynonymous DNMs and the second is a test designed to detect genes likely acting via an altered-function mechanism. This second test combines an enrichment test on missense DNMs with a test of linear clustering of missense DNMs within the gene. I then applied a Bonferroni multiple testing correction accounting for 18,762 x 2 tests, which takes into account the number of genes and two tests per gene.

I first applied DeNovoWEST to all individuals in our cohort and identified 281 significant genes, 18 more than when using the method described in McRae et al (Figure 4.5a). I also ran DeNovoWEST on the DNMs from the ~4k DDD trios from the 2017 publication[41] and found an increase of 9 additional significant genes which again demonstrates a ~10% increase in power. (Figure 4.6a). The previous method consisted of a PTV enrichment test, a missense enrichment test and a missense clustering test. This previous method tested enrichment on counts of DNMs and variant consequence classes were treated separately.

As a negative control analysis, I applied DeNovoWEST to only synonymous DNMs . While synonymous mutations can be pathogenic, it is expected that, as a class, they will not be significantly enriched in any gene. There were 6,029 genes with a *de novo* synonymous mutation, but none of these genes was significantly enriched (enrichment p $< 2.66 \times 10^{-6}$, Bonferroni corrected for 18,762 tests; Figure 4.6c). Of note, the gene with the highest synonymous enrichment p-value from DeNovoWEST is *KAT6B* (synonymous enrichment p = 3.1×10^{-5}), which contains 9 *de novo* synonymous mutations. Six of those 9 synonymous variants are the known pathogenic synonymous variant (p.Pro1049Pro) that causes Say-Barber-Biesecker/Young-Simpson syndrome via aberrant splicing of *KAT6B* [243].

The majority (196/281; 70%) of these DeNovoWEST significant genes already had sufficient evidence of DD-association to be considered of diagnostic utility (as of late 2019) by all three centres, and I refer to them as "consensus" genes. 54/281 of these significant genes were previously considered diagnostic by one or two centres ("discordant" genes).

To discover novel DD-associated genes with greater power, I then applied DeNovoWEST only to DNMs in patients without damaging DNMs in consensus genes (I refer to this subset as 'undiagnosed' patients) and identified 94 significant genes (Figure 4.5c). There is a strong correlation between DeNovoWEST p-values in the full dataset compared to those in the undiagnosed-only analysis (Figure 4.6b; $\rho = 0.729$ for all genes with non-NA DeNovoWEST p-values in both analyses). While 61 of these genes were discordant known genes, I identified 33 putative 'novel' DD-associated genes. To further ensure robustness to potential mutation rate variation between genes, Kaitlin determined whether any of the putative novel DD-associated genes had significantly more synonymous variants in the Genome Aggregation

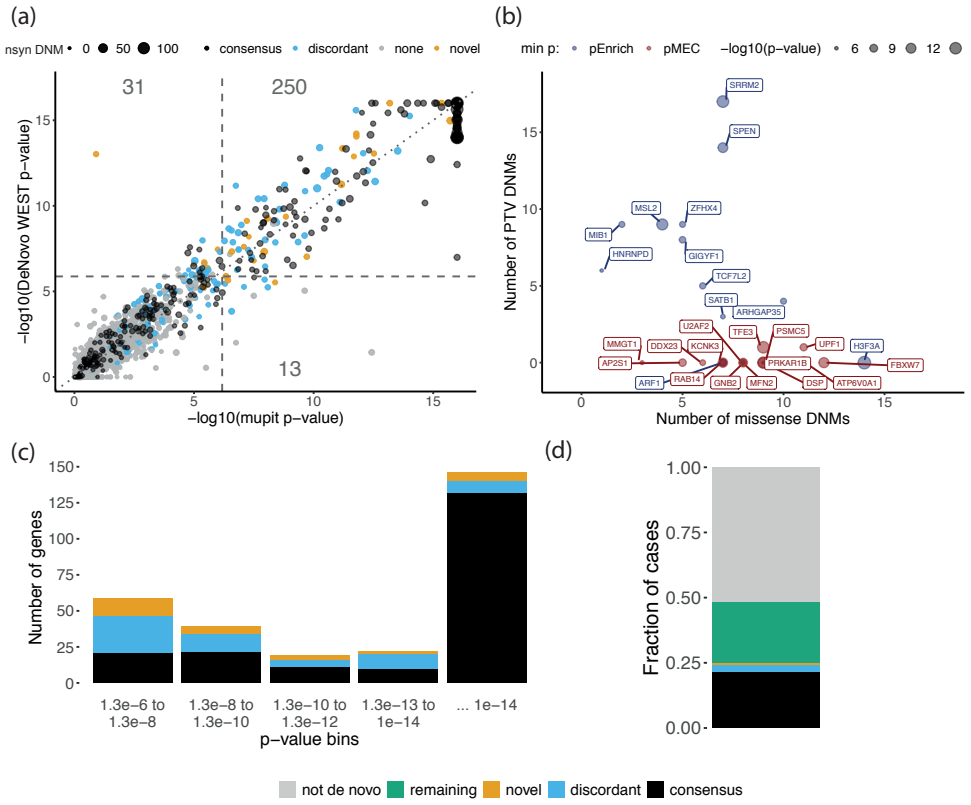

Fig. 4.5 Results of DeNovoWEST analysis. (a) Comparison of p-values generated using the new method (DeNovoWEST) versus the previous method (mupit). These are results from DeNovoWEST run on the full cohort. The dashed lines indicate the threshold for genome-wide significance. The size of the points is proportional to the number of nonsynonymous DNMs in our cohort (nsyn). The numbers describe the number of genes that fall into each quadrant. (b) The number of missense and PTV DNMs in our cohort in the 3249 novel genes. The size of the points are proportional to the -log10(p-value) from the analysis on the undiagnosed subset. The colour corresponds to which test p-value was the minimum (more significant) for these genes: pEnrich in blue, which corresponds to the overall enrichment test, or pMEC in red, which refers to the missense enrichment and clustering test. (c) The histogram depicts the distribution of p-values from the analysis on the undiagnosed subset for discordant and novel genes; p-values for consensus genes come from the full analysis. The number of genes in each bin is coloured by diagnostic gene group. (d) The fraction of cases with a nonsynonymous mutation in each diagnostic gene group. The green represents the remaining fraction of cases expected to have a pathogenic *de novo* coding mutation ("remaining") and grey is the fraction of cases that are likely to be explained by other genetic or nongenetic factors ("not *de novo*"). Figures (c) and (d) have been made by Kaitlin Samocha.

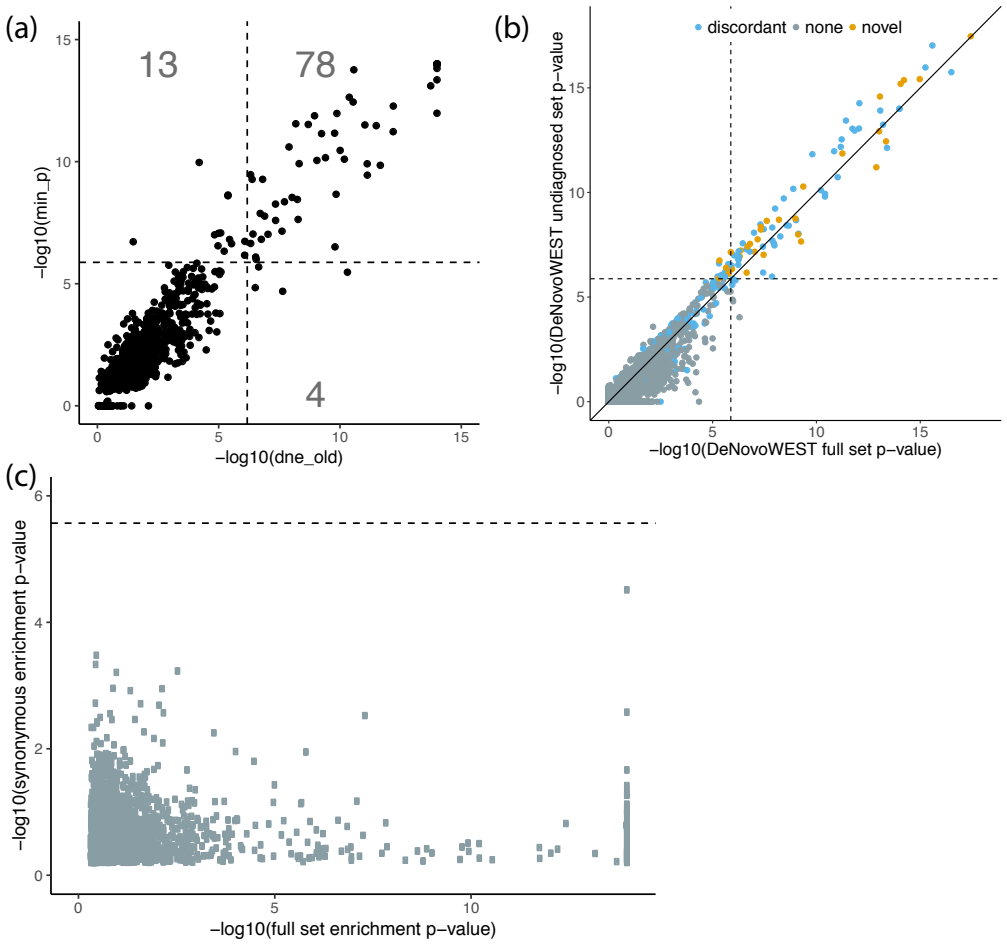

Fig. 4.6 Quality Control analyses for DeNovoWEST (a) Figure comparing p-values from published results on 4k DD trios to re-analysis of these data with the new method (DeNovoWEST). Due to constraints on the number of simulations we do not achieve p-values $< 10^{-14}$, therefore the old results are capped at this value for appropriate comparison. (b) Comparison between DeNovoWEST p-values for the full analysis vs undiagnosed-only analysis. Note that consensus genes have been removed since individuals with *de novo* nonsynonymous mutations in those genes were considered diagnosed and removed from the undiagnosed-only analysis. Genes are colored by their diagnostic list (discordant = blue; novel = orange; no list / none = gray) (c) Comparison of enrichment p-values from the full analysis vs the synonymous-only analysis. Genes with a p-value of 0 have been removed from the plot for clarity.

Database (gnomAD)[102] of population variation than expected under the null mutation model (see methods). Kaitlin identified 11/33 genes with a significant excess of synonymous variants. For these 11 genes I then repeated the DeNovoWEST test, increasing the null mutation rate by the ratio of observed to expected synonymous variants in gnomAD. Five of these genes then fell below the exome-wide significance threshold and were removed, leaving

28 novel genes, with a median of 10 nonsynonymous DNMs in our dataset (Figure 4.5b). There were 314 patients with nonsynonymous DNMs in these 28 genes (1.0% of our cohort); all DNMs in these genes were inspected in IGV[182] and, of 198 for which experimental validation was attempted, all were confirmed as DNMs in the proband. The DNMs in these novel genes were distributed approximately randomly across the three datasets (no genes with $p < 0.001$, heterogeneity test). Six of the 28 novel DD-associated genes are further corroborated by OMIM entries or publications, including *TFE3* for which patients were described in two recent publications [224, 44]. Taken together, 25.0% of individuals in the combined cohort have a nonsynonymous DNM in one of the consensus or significant DD-associated genes (Figure 4.5d).

4.3.2 Characteristics of the novel DD-associated genes and disorders

Based on an analysis from our collaborators at GeneDx of semantic similarity between Human Phenotype Ontology terms, patients with DNMs in the same novel DD-associated gene were less phenotypically similar to each other, on average, than patients with DNMs in a consensus gene ($p = 2.3\times10^{-11}$, Wilcoxon rank-sum test; Figure 4.7a) [238]. Patients with DNMs is the same novel DD-associated genes were more phenotypically similar compared to the null (pairs of random patients in the cohort) ($p = 2.0\times10^{-30}$, Wilcoxon rank-sum test). This suggests that these novel disorders less often result in distinctive and consistent clinical presentations, which may have made these disorders harder to discover via a phenotype-driven analysis or recognise by clinical presentation alone. Each of these novel disorders requires a detailed genotype-phenotype characterisation.

Overall, novel DD-associated genes encode proteins that have very similar functional and evolutionary properties to consensus genes, e.g. developmental expression patterns, network properties and biological functions (Figure 4.7b; Table 4.1). Across the properties that were considered, the only significant difference between novel and consensus genes was found in the network distance to another consensus DD gene ($p = 0.002$, Wilcoxon test). Despite the high-level functional similarity between known and novel DD-associated genes, the nonsynonymous DNMs in the more recently discovered DD-associated genes are much more likely to be missense DNMs, and less likely to be PTVs (discordant and novel; $p = 1.2\times10^{-25}$, χ^2 test). Fifteen of the 28 (54%) of the novel genes only had missense DNMs, and only a minority had more PTVs than missense DNMs. Consequently, we expect that a greater proportion of the novel genes will act via altered-function mechanisms (e.g. dominant negative or gain-of-function). For example, the novel gene *PSMC5* (DeNovoWEST $p = 2.6\times10^{-15}$) had one inframe deletion and nine missense DNMs, eight of which altered two structurally important amino acids within the 3D protein structure: p.Pro320Arg and

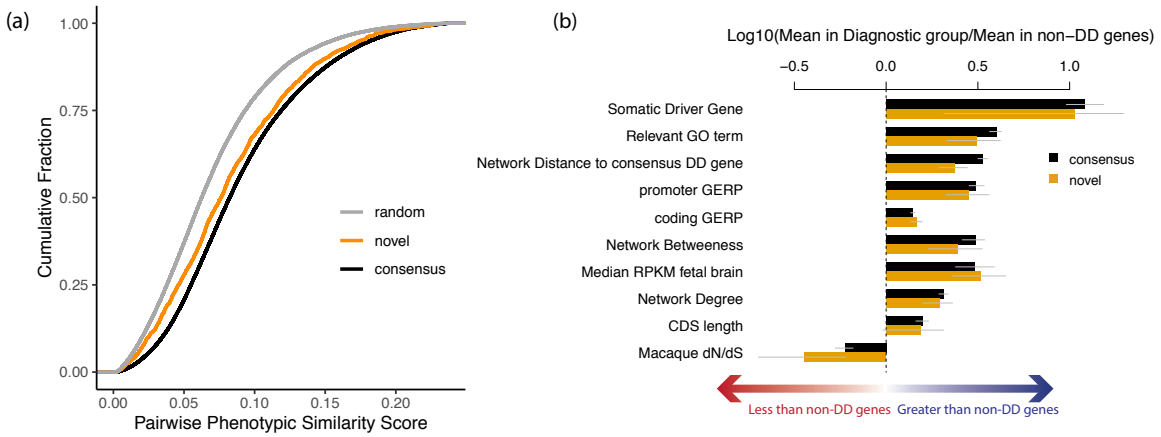

Fig. 4.7 Functional properties and mechanisms of novel genes. (a) Comparing the phenotypic similarity of patients with DNMs in novel and consensus genes. Random phenotypic similarity was calculated from random pairs of patients. Patients with DNMs in the same novel DD-associated gene were less phenotypically similar than patients with DNMs in a known DD-associated gene ($p = 1.29.5 \times 10^{-1338}$, Wilcoxon rank-sum test). Figure made by Kevin Arvai (b) Comparison of functional properties of consensus known and novel DD genes. Properties were chosen as those known to be differential between consensus and non-DD genes

p.Arg325Trp and so is likely to operate via an altered-function mechanism. None of the novel genes exhibited significant clustering of *de novo* PTVs.

4.3.3 Recurrent mutations and potential new germline selection genes

I identified 773 recurrent DNMs (736 SNVs and 37 indels), ranging from 2-36 independent observations per DNM, which allowed me to interrogate systematically the factors driving recurrent germline mutation. I considered three potential contributory factors: (i) clinical ascertainment enriching for pathogenic mutations, (ii) greater mutability at specific sites, and (iii) positive selection conferring a proliferative advantage in the male germline, thus increasing the prevalence of sperm containing the mutation [67]. I observed strong evidence that all three factors contribute, but not necessarily mutually exclusively. Clinical ascertainment drives the observation that 65% of recurrent DNMs were in consensus genes, a 5.4-fold enrichment compared to DNMs only observed once ($p < 10^{-50}$, proportion test). Hypermutability underpins the observation that 64% of recurrent *de novo* SNVs occurred at hypermutable CpG dinucleotides[45], a 2.0-fold enrichment over DNMs only observed once ($p = 3.3 \times 10^{-68}$, χ^2 test). I also observed a striking enrichment of recurrent mutations at the haploinsufficient DD-associated gene *MECP2*, in which we observed 11 recurrently mutated

Symbol	Chr	Position	Ref	Alt	Consequence	Number recur	Likely mechanism	CpG	Somatic Driver Gene	Germline Selection Gene	DD status
PACS1	11	65978677	C	T	missense	36	activating	Yes	-	-	consensus
PPP2R5D	6	42975003	G	A	missense	22	dominant negative	-	-	-	consensus
SMAD4	18	48604676	A	G	missense	21	activating	-	Yes	-	consensus
PACS2	14	105834449	G	A	missense	13	dominant negative	Yes	-	-	discordant
MAP2K1	15	66729181	A	G	missense	11	activating	-	Yes	Yes	consensus
PPP1CB	2	28999810	C	G	missense	11	all missense/in frame	-	-	-	consensus
NAA10	X	153197863	G	A	missense	11	all missense/in frame	Yes	-	-	consensus
MECP2	X	153296777	G	A	stop gain	11	loss of function	Yes	-	-	consensus
CSNK2A1	20	472926	T	C	missense	10	activating	-	-	-	consensus
CDK13	7	40085606	A	G	missense	10	all missense/in frame	-	-	-	consensus
SHOC2	10	112724120	A	G	missense	9	activating	-	-	-	consensus
PTPN11	12	112915523	A	G	missense	9	activating	-	Yes	Yes	consensus
SMAD4	18	48604664	C	T	missense	9	activating	Yes	Yes	-	consensus
SRCAP	16	30748664	C	T	stop gain	9	dominant negative	Yes	-	-	consensus
FOXP1	3	71021817	C	T	missense	9	loss of function	Yes	-	-	consensus
CTBP1	4	1206816	G	A	missense	9	dominant negative	Yes	-	-	discordant

Table 4.2 Recurrent Mutations. *De novo* single nucleotide variants with more than 9 recurrences in the cohort annotated with relevant information, such as CpG status, whether the impacted gene is a known somatic driver or germline selection gene, and diagnostic gene group (e.g. consensus known). "Recur" refers to the number of recurrences. "Likely mechanism" refers to mechanisms attributed to this gene in the published literature

SNVs within a 500 bp window, nine of which were G to A mutations at a CpG dinucleotide. *MECP2* exhibits a highly significant twofold excess of synonymous mutations within the Genome Aggregation Database (gnomAD) population variation resource[102], suggesting that locus-specific hypermutability might explain this observation.

To assess the contribution of germline selection to recurrent DNMs, I initially focused on the 12 known germline selection genes (*FGFR2, FGFR3, PTPN11, HRAS, KRAS, RET, BRAF, CBL, MAP2K1, MAP2K2, RAF1, SOS1*), which all operate through activation of the RAS-MAPK signalling pathway[136, 137]. I identified 39 recurrent DNMs in 11 of these genes, 38 of which are missense. To determine if the observed mutations in germline selection genes are known to be activating, I first confirmed that these were all genes known to be acting through gain-of-function mechanisms. All of these genes are known monoallelic DD-associated genes annotated as having activating mutation consequences according to DDG2P[236]. I then confirmed that all these recurrent mutations are listed as pathogenic or likely pathogenic variants in ClinVar[115]. As expected, given that hypermutability is not the driving factor for recurrent mutation in these germline selection genes, these 39 recurrent DNMs were depleted for CpGs relative to other recurrent mutations (6/39 vs 425/692, p = 3.4×10^{-8}, χ^2 test). Positive germline selection has been shown to be capable of increasing the apparent mutation rate more strongly than either clinical ascertainment (10-100$\times$ in this dataset) or hypermutability (~10$\times$ for CpGs)[67]. However, only a minority of the most

highly recurrent mutations in this dataset are in genes that have been previously associated with germline selection. Nonetheless, several lines of evidence suggested that the majority of these most highly recurrent mutations are likely to confer a germline selective advantage. Based on the recurrent DNMs in known germline selection genes, DNMs under germline selection should be more likely to be activating missense mutations, and should be less enriched for CpG dinucleotides. Table 4.2 shows the 16 *de novo* SNVs observed nine or more times in our DNM dataset, only two of which are in known germline selection genes (*MAP2K1* and *PTPN11*). All but two of these 16 *de novo* SNVs cause missense changes, all but two of these genes cause disease by an altered-function mechanism, and these DNMs were depleted for CpGs relative to all recurrent mutations. Two of the genes with highly recurrent *de novo* SNVs, *SHOC2* and *PPP1CB*, encode interacting proteins that are known to play a role in regulating the RAS-MAPK pathway, and pathogenic variants in these genes are associated with a Noonan-like syndrome[244]. Moreover, two of these recurrent DNMs are in the same gene *SMAD4*, which encodes a key component of the TGF-beta signalling pathway, potentially expanding the pathophysiology of germline selection beyond the RAS-MAPK pathway. Confirming germline selection of these mutations will require deep sequencing of testes and/or sperm[137].

4.3.4 Evidence for incomplete penetrance and pre/perinatal death

Nonsynonymous DNMs in consensus or significant DD-associated genes accounted for half of the exome-wide nonsynonymous DNM burden associated with DD (Figure 4.5d). Despite the identification of 285 significantly DD-associated genes, there remains a substantial burden of both missense and protein-truncating DNMs in unassociated genes (those that are neither significant in my analysis nor on the consensus gene list). The remaining burden of protein-truncating DNMs is greatest in genes that are intolerant of PTVs in the general population (Figure 4.8a). Similarly, while I observed that unassociated genes were, overall, significantly less likely to be expressed in the fetal brain ($p = 4.42 \times 10^{-30}$, proportion test), I found that within genes intolerant of PTVs (pLI>0.9), unassociated genes were just as likely as significant genes to be expressed in the fetal brain ($p = 0.09$, proportion test; Figure 4.9). This suggests that more haploinsufficient (HI) disorders await discovery. The PTV enrichment dropped even further after removing nominally significant genes in my analysis which suggests that a larger sample size will increase the power to detect these (Figure 4.8b). I observed that PTV mutability (estimated from a null germline mutation model) was significantly lower in unassociated genes compared to DD-associated genes ($p = 4.5 \times 10^{-68}$, Wilcox rank-sum test 4.10a), which leads to reduced statistical power to detect

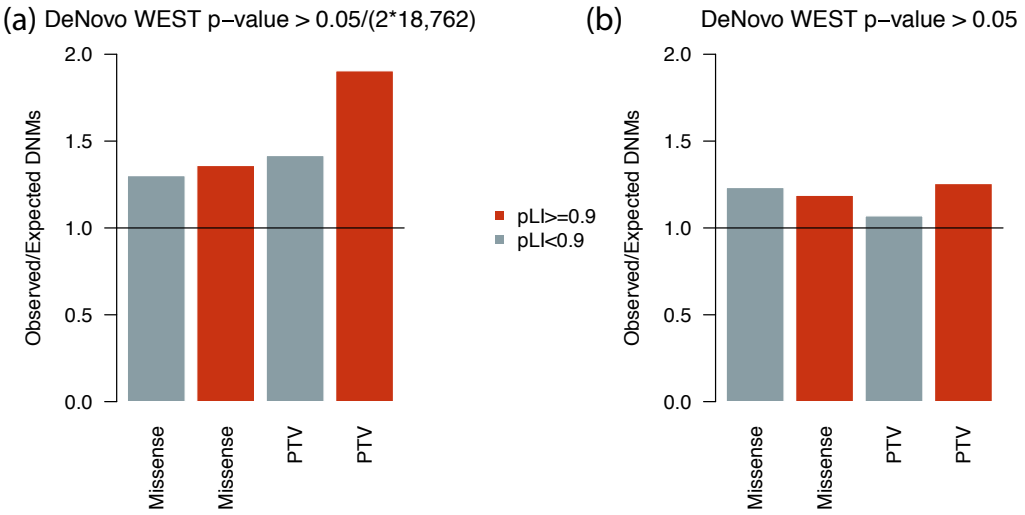

Fig. 4.8 DNM enrichment in non-significant genes (a) The enrichment of missense and PTVs separated by high pLI (red) and low pLI (blue). Unassociated genes are defined as those not on any diagnostic list (not consensus or discordant) and not significant in our analysis. (b) The same as (a) except here the genes have been subsetted further to those that are not nominally significant in our analysis (unadjusted p-value >0.05)

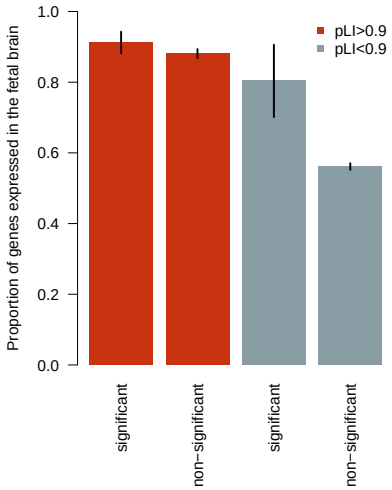

Fig. 4.9 Comparison of proportion of genes expressed in fetal brain. The proportion of genes between significant and non-significant genes in our analysis split by low (<0.9) and high (>0.9) pLI. Significant genes also includes all genes on the consensus diagnostic list.

DNM enrichment in unassociated genes. This is consistent with the hypothesis that many more HI disorders await discovery.

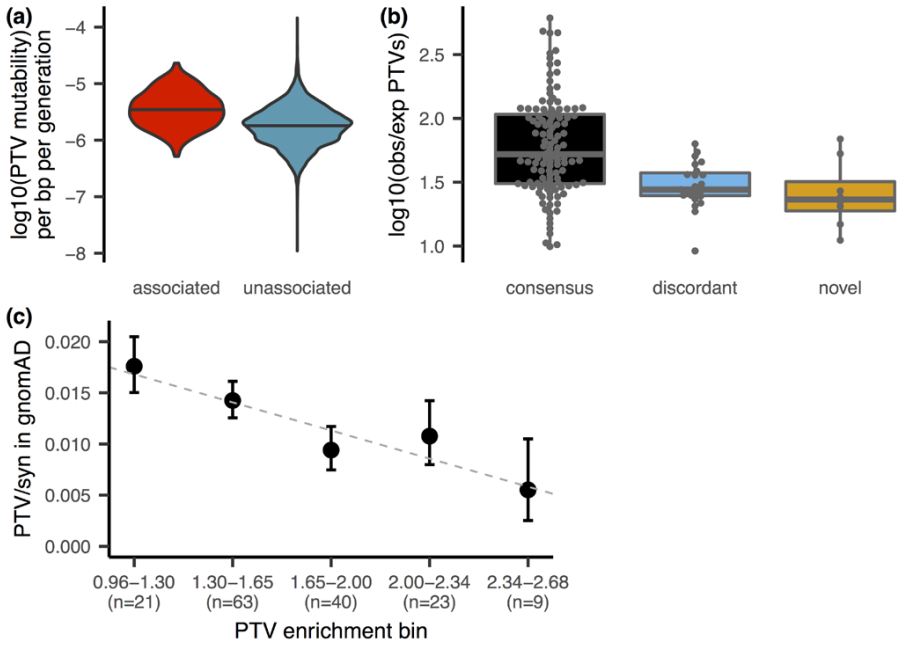

Fig. 4.10 Impact of penetrance on power. (a) PTV mutability is significantly lower in genes that are not significantly associated to DD in the analysis ("unassociated", coloured blue) than in DD-associated genes ("associated", coloured red; $p = 4.5 \times 10^{-68}$, Wilcox rank sum test). (b) Distribution of PTV enrichment in significant, likely haploinsufficient, genes by diagnostic group. (c) Comparison of the PTV enrichment in the cohort vs the PTV to synonymous ratio found in gnomAD, for genes that are significantly enriched for the number of PTV mutations in the cohort (without any variant weighting). PTV enrichment is shown as log10(enrichment). There is a significant negative relationship ($p = 0.031$, weighted regression). Figure (c) courtesy of Kaitlin Samocha

A key parameter in estimating statistical power to detect novel HI disorders is the fold-enrichment of *de novo* PTVs expected in as yet undiscovered HI disorders. I observed that novel DD-associated HI genes had significantly lower PTV enrichment compared to the consensus HI genes ($p = 0.005$, Wilcox rank-sum test; 4.10b). Two additional factors that could lower DNM enrichment, and thus power to detect a novel DD-association, are reduced penetrance and increased pre/perinatal death, which here covers spontaneous fetal loss, termination of pregnancy for fetal anomaly, stillbirth, and early neonatal death. To evaluate incomplete penetrance, Kaitlin investigated whether HI genes with a lower enrichment of protein-truncating DNMs in our cohort are associated with greater prevalences of PTVs in the general population. She observed a significant negative correlation ($p = 0.031$, weighted linear regression) between gene-specific PTV enrichment in our cohort and the gene-specific ratio of PTV to synonymous variants in the gnomAD dataset of population variation[102],

suggesting that incomplete penetrance does lower *de novo* PTV enrichment in individual genes in the cohort (Figure 4.10c).

A structural malformation (detected via ultrasound) during pregnancy is associated with pre/perinatal death, which here encompasses spontaneous fetal loss, stillbirth, early neonatal death, and termination of pregnancy for fetal anomaly. To understand the impact that pre/perinatal death may have had on our power to detect DD-associated genes a clinician (Allison Yeung) assigned each consensus DD gene known to be haploinsufficient (n=217) a low, medium or high likelihood of a patient with a pathogenic mutation in the gene presenting with a structural malformation on ultrasound. To verify that this classification was valid, I compared the proportion of individuals in the DDD study with nonsynonymous mutations in the three gene groups that were reported to have an abnormal scan during pregnancy (Figure 4.11a). I found that the proportion of patients with an abnormal ultrasound for those with a nonsynonymous DNM in a gene with a high or medium likelihood of abnormal ultrasound was significantly higher than for those with a DNM in a gene with a low likelihood classification (12.8% (low) vs 28.0% (medium/high), χ^2 p = 7.24×10^{13}). I also looked at DNMs called in 640 trios from the Prenatal Assessment of Genomes and Exomes (PAGE) study and found that there was a higher enrichment of non synonymous DNMs in genes with a medium/high likelihood of presenting with an ultrasound abnormality compared to the genes with a low likelihood but this was not significant (2.3 (low) vs 4.8 (medium/high) enrichment, p = 0.052, Poisson test; Figure 4.11d) [133].

Having verified this classification, I then calculated the ratio of the number of observed *de novo* PTVs in each group to the total number of expected *de novo* PTVs across each group. I observed that the fold-enrichment of protein-truncating DNMs in consensus HI DD-associated genes in the cohort was significantly lower for genes with a medium or high likelihood of presenting with a prenatal structural malformation (p = 4.6×10^{-5}, Poisson test, Figure 4.11c), suggesting that pre/perinatal death decreases power to detect some novel DD-associated disorders. For the DDD data, I also regressed the proportion of individuals with a nonsynonymous mutation in a consensus HI gene who had an abnormal ultrasound against the observed PTV enrichment in that gene but did not find a significant regression coefficient for proportion of abnormal ultrasounds (p = 0.33, quasipoisson GLM; Figure 4.11b).

To assess whether mutations in novel genes are more likely to be associated with pre/perinatal death than consensus genes I compared the nonsynonymous *de novo* enrichment of these two groups in PAGE. I did not find a significant difference between the enrichment in novel genes (2.9) compared to consensus genes (3.4) (p = 0.44, Poisson test) however this analysis is not well powered. I also did not find that individuals in DDD with mutations

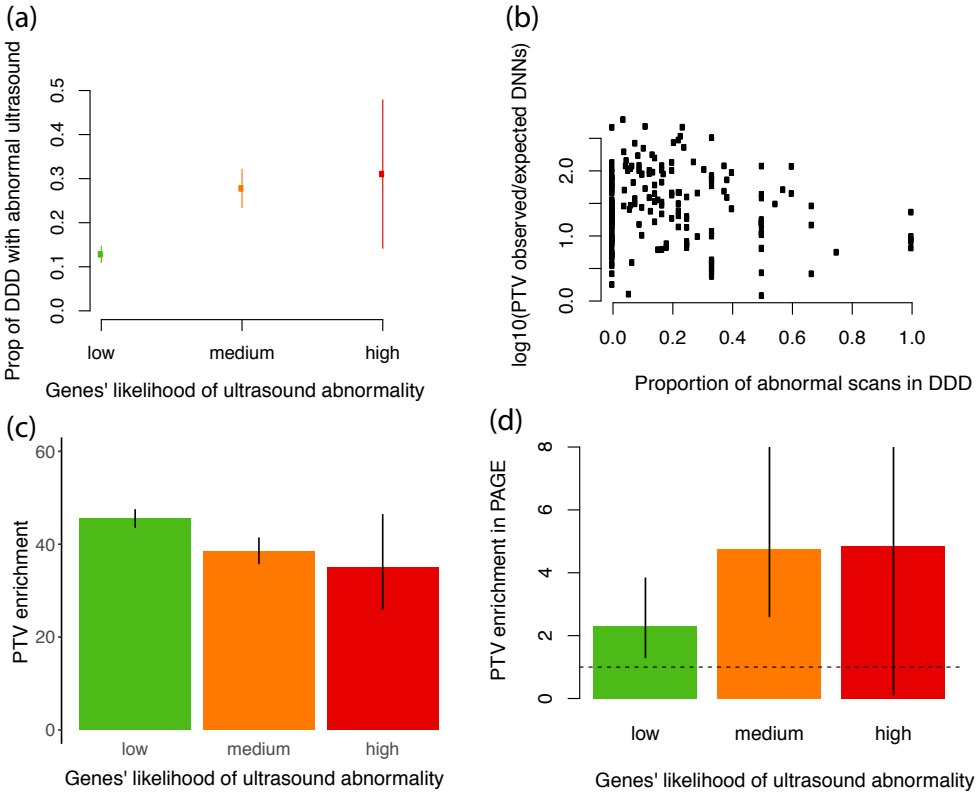

Fig. 4.11 Investigating impact of pre/perinatal death on power (a) The proportion of DDD probands that have been reported as presenting with an abnormal ultrasound furing pregnancy that have nonsynonymous DNMs in consensus genes. This has been split by whether the gene falls in the low, medium of high category of the clinician classified likelihood of presenting with a ultrasound abnormality. This has been coloured by red (high), orange (medium) and low (green). (b) This shows observed PTV enrichment in the cohort for each consensus known haploinsufficient gene against the proportion of individuals with a nonsynoymous DNM in those genes that have been reported in DDD as having an abnormal scan (c) Overall *de novo* PTV enrichment (observed / expected PTVs) across genes grouped by their clinician-assigned likelihood of presenting with a structural malformation on ultrasound during pregnancy in the cohort. PTV enrichment is significantly lower for genes with a medium or high likelihood compared to genes with a low likelihood ($p = 4.5 \times 10^{-5}$, Poisson test) (d) Overall *de novo* PTV enrichment (observed / expected PTVs) across genes grouped by their clinician-assigned likelihood of presenting with a structural malformation on ultrasound during pregnancy in the PAGE cohort.

in novel genes were significantly more likely to have an abnormal ultrasound compared to consensus genes (proportion abnormal for novel 0.24, proportion abnormal for consensus 0.17; χ^2 test $p = 0.08$) but again there is not sufficient power here.

4.3.5 Modelling reveals hundreds of DD genes remain to be discovered

To understand the likely trajectory of future DD discovery efforts, I downsampled the current cohort (to 5k, 10k, 15k, 20k and 25k individuals) and reran the enrichment analysis (Figure 4.12a). I observed that the number of significant genes has not yet plateaued. Increasing sample sizes should result in the discovery of many novel DD-associated genes. To estimate how many haploinsufficient genes might await discovery, I modelled the likelihood of the observed distribution of protein-truncating DNMs among genes as a function of varying numbers of undiscovered HI DD genes and fold-enrichments of protein-truncating DNMs in those genes. I found that the remaining HI burden is most likely spread across ~1000 genes with ~10-fold PTV enrichment (Figure 4.12b). This fold enrichment is three times lower than

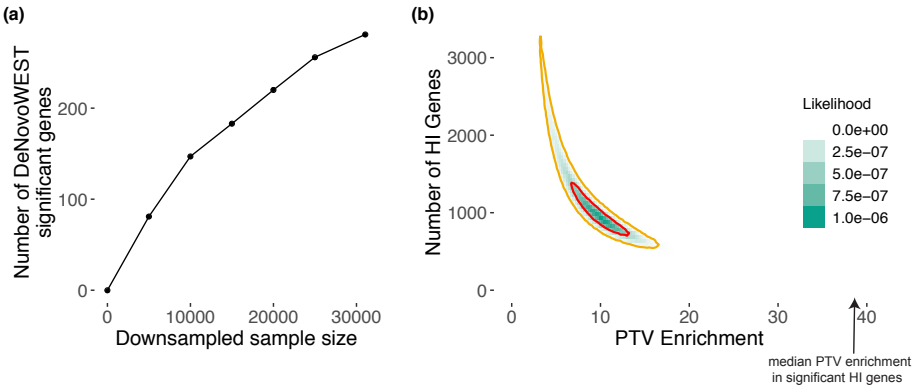

Fig. 4.12 Exploring the remaining number of DD genes. (a) Number of significant genes from downsampling full cohort and running DeNovoWEST's enrichment test. (b) Results from modelling the likelihood of the observed distribution of *de novo* PTV mutations. This model varies the numbers of remaining haploinsufficient (HI) DD genes and PTV enrichment in those remaining genes. The 50% credible interval is shown in red and the 90% credible interval is shown in orange. Note that the median PTV enrichment in significant HI genes (shown with an arrow) is 39.7

in known HI DD-associated genes, suggesting that incomplete penetrance and/or pre/perinatal death is much more prevalent among undiscovered HI genes. I modelled the missense DNM burden separately which allowed for different distribution of missense enrichments, as modelled by the gamma distribution, across DD genes. These distributions represent different scenarios for the proportion of genes acting via altered-function or loss-of-function mechanisms. I observed that the most likely architecture of undiscovered DD-associated genes is one that comprises over 1000 genes with a substantially lower fold-enrichment (~3 fold) than in currently known DD-associated genes (Figure 4.13 b). The most likely missense

enrichment distribution reflected the scenario where most of the missense mutations were acting via loss-of-function mechanisms (γ shape parameter of 20).

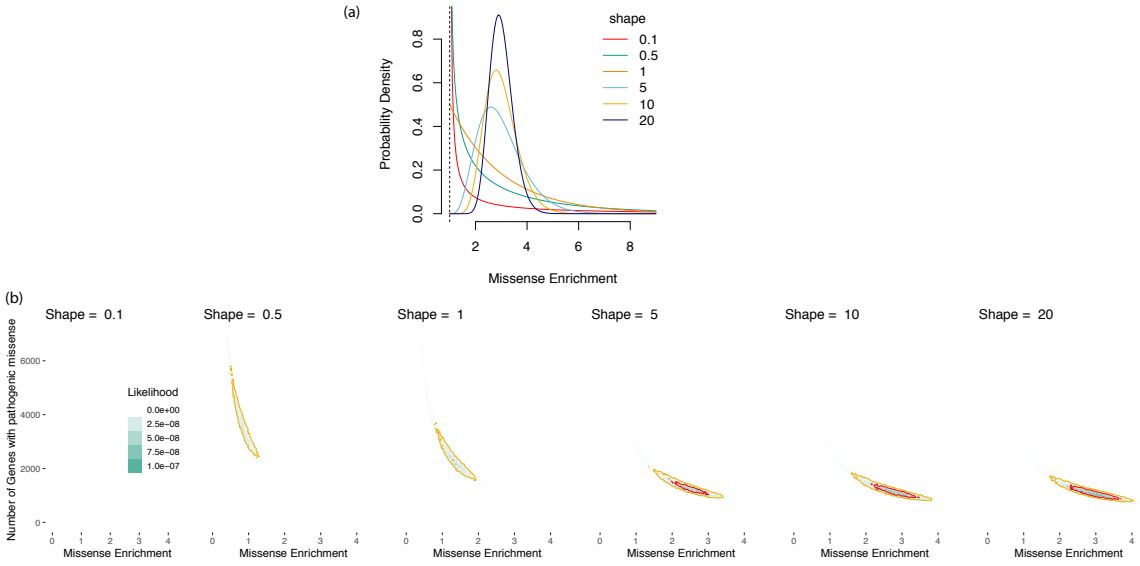

Fig. 4.13 Likelihood model for missense DNM enrichment(a) Depiction of γ distribution for 6 shape values used in simulations. Here the mean of each distribution is set at 2. (b) Likelihood of scenario under variety of shapes for γ distribution and under varying values of missense enrichment and number of genes with pathogenic missense variants. The 90% (yellow line) and 50% (red line) credible intervals are calculated across all shapes considered.

A sample size of ~350,000 parent-offspring trios would be needed to have 80% power to detect a 10-fold enrichment of protein-truncating DNMs for a gene with the median PTV mutation rate among currently unassociated genes (assuming Poisson test on PTV count enrichment). Using this inferred 10-fold enrichment among undiscovered HI genes, from our current data I can evaluate the likelihood that any gene in the genome is an undiscovered HI gene, by comparing the likelihood of the number of *de novo* PTVs observed in each gene to have arisen from the null mutation rate or from a 10-fold increased PTV rate. Among the ~19,000 non-DD-associated genes, ~1,200 were more than three times more likely to have arisen from a 10-fold increased PTV rate, whereas ~7,000 were three times more likely to have no *de novo* PTV enrichment.

4.4 Discussion

In this chapter, I identified 28 novel developmental disorders by developing an improved statistical test for mutation enrichment and applying it to a dataset of exome sequences

from 31,058 children with developmental disorders, and their parents. These 28 novel genes account for up to 1.0% of the cohort, and inclusion of these genes in diagnostic workflows will catalyse increased diagnosis of similar patients globally. The value of this study for improving diagnostic yield extends well beyond these 28 novel genes; once newly validated discordant genes are included, the total number of genes added to the diagnostic workflows of the three participating centres ranged from 48-65 genes. I have shown that both incomplete penetrance and pre/perinatal death reduce our power to detect novel DDs postnatally, and that one or both of these factors are likely operating considerably more strongly among undiscovered DD-associated genes. In addition, I have identified a set of highly recurrent mutations that are strong candidates for novel germline selection mutations, which would be expected to result in a higher than expected disease incidence that increases dramatically with increased paternal age. This study represents the largest collection of DNMs for any disease area, and is approximately three times larger than a recent meta-analysis of DNMs from a collection of individuals with autism spectrum disorder, intellectual disability, and/or a developmental disorder[31]. The analysis included DNMs from 24,348 previously unpublished trios, and I identified ~2.3 times as many significantly DD-associated genes as this previous study when using Bonferroni-corrected exome-wide significance (285 vs 124). In contrast to meta-analyses of published DNMs, the harmonised filtering of candidate DNMs across cohorts in this study should protect against results being confounded by substantial cohort-specific differences in the sensitivity and specificity of detecting DNMs.

Here I inferred indirectly that developmental disorders with higher rates of detectable prenatal structural abnormalities had greater pre/perinatal death. The potential size of this effect can be quantified from the recently published PAGE study of genetic diagnoses in a cohort of fetal structural abnormalities[133]. In this latter study, genetic diagnoses were not returned to participants during the pregnancy, and so the genetic diagnostic information itself could not influence pre/perinatal death. In the PAGE study data, 69% of fetuses with a genetically diagnosable cause for this anomaly died perinatally or neonatally, with termination of pregnancy, fetal demise and neonatal death all contributing. This emphasises the substantial impact that pre/perinatal death can have on reducing the ability to discover novel DDs from postnatal recruitment alone, and motivates the integration of genetic data from prenatal, neonatal and postnatal studies in future analyses.

To empower mutation enrichment testing, I estimated positive predictive values (PPV) of each DNM being pathogenic on the basis of their predicted protein consequence, CADD score, selective constraint against heterozygous PTVs in the gene (s_{het}), and, for missense variants, presence in a region under missense constraint in the general population [107, 24, 186]. These PPVs should also be highly informative for variant prioritisation in the diagnosis

of dominant developmental disorders. Further work is needed to see whether these PPVs might be informative for recessive developmental disorders, and in other types of dominant disorders. More generally, empirically-estimated PPVs based on variant enrichment in large datasets may be similarly informative in many other disease areas.

The approach taken here is statistically conservative in identifying DD-associated genes. In two previous published studies from the DDD, using the same significance threshold, 26 novel DD-associated genes were identified[215, 41]. All 26 are now regarded as being diagnostic, and have entered routine clinical diagnostic practice. There are 184 consensus genes that did not cross the significance threshold in this study. It is likely that many of these cause disorders that were under-represented in this study due to the ease of clinical diagnosis on the basis of distinctive clinical features or targeted diagnostic testing. These ascertainment biases are, however, not likely to impact the representation of novel DDs in this cohort. The modelling also suggested that likely over 1,000 DD-associated genes remain to be discovered, and that reduced penetrance and pre/perinatal death will reduce power to identify these genes through DNM enrichment. Identifying these genes will require both improved analytical methods and greater sample sizes. As sample sizes increase, accurate modelling of gene-specific mutation rates becomes more important. In this analyses of 31,058 trios, there was evidence that mutation rate heterogeneity among genes can lead to over-estimating the statistical significance of mutation enrichment based on an exome-wide mutation model. An important future direction would be to develop more granular mutation rate models, based on large-scale population variation resources, to ensure that larger studies are robust to mutation rate heterogeneity.

The variant-level weights used by DeNovoWEST could be improved over time. As reference population samples, such as gnomAD, increase in size, weights based on selective constraint metrics (e.g. s_{het}, regional missense constraint) will improve. Weights could also incorporate more functional information, such as expression in disease-relevant tissues [102]. For example, I observed that DD-associated genes are significantly more likely to be expressed in fetal brain (Figure 4.9). Furthermore, novel metrics based on gene co-regulation networks can predict whether genes function within a disease-relevant pathway[42]. As a cautionary note, including more functional information may increase power to detect some novel disorders while decreasing power for disorders with pathophysiology different from known disorders. Variant-level weights could be further improved by incorporating other variant prioritisation metrics, such as upweighting variants predicted to impact splicing, variants in particular protein domains, or variants that are somatic driver mutations during tumorigenesis. In developing DeNovoWEST, I initially explored applying both variant-level weights and gene-level hypothesis weights in separate stages of the analysis, however, subtle

but pervasive correlations between gene-level metrics (e.g. s_{het}) and variant-level metrics (e.g. regional missense constraint, CADD) presents statistical challenges to implementation. Finally, the discovery of less penetrant disorders can be empowered by analytical methodologies that integrate both DNMs and rare inherited variants, such as TADA [78]. Nonetheless, using current methods, I estimated that ~350,000 parent-child trios would need to be analysed to have ~80% power to detect HI genes with a 10-fold PTV enrichment. Discovering non-HI disorders will need even larger sample sizes. Reaching this number of sequenced families will be impossible for an individual research study or clinical centre, therefore it is essential that genetic data generated as part of routine diagnostic practice are shared with the research community such that it can be aggregated to drive discovery of novel disorders and improve diagnostic practice.