Evidence for 28 genetic disorders discovered by combining healthcare and research data

Joanna Kaplanis1,40, Kaitlin E. Samocha1,40, Laurens Wiel1,4,40, Zhancheng Zhang1,4,40, Kevin J. Arvai6, Ruth Y. Eberhardt1, Giuseppe Gallone1, Stefan H. Lelieveld2, Nicolas C. Martin3, Jeremy F. McRae1, Patrick J. Short1, Rebecca I. Toone1, Elke de Boer3, Petr Danecek1, Eugene J. Gardner1, Ni Huang5, Jenny Lord15,9, Iñigo Martíncorena1, Ralph Pfundt1, Margot R. F. Reijnders5,7, Alison Yeung1,3, Helger G. Yntema1, Deciphering Developmental Disorders Study*, Lisinka E. L. M. Vissers5, Jane Juusola4, Caroline F. Wright10, Han G. Brunner5,7,11,12, Helen V. Firth13, David R. FitzPatrick14, Jeffrey C. Barrett1, Matthew E. Hurles1,41, Christian Gilissen2,41 & Kyle Retterer4,41

De novo mutations in protein-coding genes are a well-established cause of developmental disorders1. However, genes known to be associated with developmental disorders account for only a minority of the observed excess of such de novo mutations1,2. Here, we identify previously undescribed genes associated with developmental disorders, which integrates healthcare and research exome-sequence data from 31,058 parent–offspring trios of individuals with developmental disorders, and develop a simulation-based statistical test to identify gene-specific enrichment of de novo mutations. We identified 285 genes that were significantly associated with developmental disorders, including 28 that had not previously been robustly associated with developmental disorders. Although we detected more genes associated with developmental disorders, the majority of de novo mutations in protein-coding genes remains unaccounted for. Modelling suggests that more than 1,000 genes associated with developmental disorders have not yet been described, many of which are likely to be less penetrant than the currently known genes. Research access to clinical diagnostic datasets will be critical for completing the map of genes associated with developmental disorders.

Identification of 285 DD-associated genes

Following clear consent practices and only using aggregate, deidentified data, we pooled DNMs from patients with a DD from three centres: GeneDx (a US-based diagnostic testing company), the Deciphering Developmental Disorders study and Radboud University Medical Center. We performed stringent quality control on variants and samples to obtain 45,221 coding and splicing DNMs in 31,058 individuals (Supplementary Fig. 1 and Supplementary Table I), including data on 24,348 trios that have not previously been published. These DNMs included 40,992 single-nucleotide variants (SNVs) and 4,229 insertions or deletions (indels). The three cohorts have similar clinical characteristics, including male-to-female ratios, enrichments of DNMs by mutational class and prevalences of known disorders (Supplementary Fig. 2).

To detect gene-specific enrichments of damaging DNMs, we developed a method named DeNovoWEST (de novo weighted enrichment simulation test; https://github.com/queenjobo/DeNovoWEST).

It has previously been estimated that around 42–48% of patients with severe developmental disorder (DD) have a pathogenic de novo mutation (DNM) in a protein-coding gene. However, most of these patients remain undiagnosed despite the identification of hundreds of DD-associated genes. This indicates that there are more DD-relevant genes to find. Existing methods to detect the gene-specific enrichment of damaging DNMs do not incorporate all of the available information about which variants are more likely to be disease-associated; missense variants and protein-truncating variants (PTVs) vary in their impact on protein function3–6. Known dominant DD-associated genes are strongly enriched in the minority of genes that exhibit strong selective constraint on heterozygous PTVs7. To identify additional DD-associated genes, we need to increase our power to detect gene-specific enrichments of damaging DNMs by increasing sample sizes and improving our statistical methods. In previous studies of pathogenic copy number variations, the use of healthcare data has been key to achieving larger sample sizes than would be possible in a research setting alone8–10.

*These authors jointly supervised this work:
Matthew E. Hurles, Christian Gilissen, Kyle Retterer.
A list of authors and their affiliations appears in the online version of the paper. 1E-mail: meh@sanger.ac.uk

Published online: 14 October 2020
https://doi.org/10.1038/s41586-020-2832-5
Received: 8 October 2019
Accepted: 17 July 2020

DeNovoWEST scores all classes of sequence variants on a unified severity scale based on empirically estimated positive predictive values of being pathogenic (Supplementary Figs. 3, 4). We perform two tests per gene: an enrichment test on all nonsynonymous DNMs and a test designed to detect genes that probably act through an altered-function mechanism, which combines an missense enrichment test with a missense clustering test. We then applied a Bonferroni multiple-testing mechanism, which combines a missense enrichment test with a missense clustering test. We then applied a Bonferroni multiple-testing correction accounting for the number of genes (n = 18,762) and two tests per gene.

We first applied DeNovoWEST to all individuals in our cohort and identified 281 significantly enriched genes, 18 more than when using our previously published method (mupit1), run on the full cohort. Dashed lines indicate the threshold for genome-wide significance (one-sided, Bonferroni correction). Point size is proportional to the number of nonsynonymous DNMs in our cohort. The number of genes that fall into each quadrant are annotated. b, The number of missense and PTV DNMs in the novel cohort. Dashed lines indicate the threshold for genome-wide significance (one-sided, Bonferroni correction). Point size is proportional to the number of genes that fall into each quadrant. c, The distribution of significant P-values from analysis of the undiagnosed subset for discordant and novel genes; P-values for consensus genes come from the full cohort analysis. The number of genes in each P-value bin is coloured by diagnostic gene group (n = 285 significant genes; one-sided Bonferroni-corrected P-values). d, The fraction of patients (n = 31,058) with a nonsynonymous mutation in each diagnostic gene group. Green, the remaining fraction of patients (the offspring of the parent–offspring trios) expected to have a pathogenic de novo coding mutation; grey, the fraction of patients that are likely to be explained by other factors. e, The fraction of patients with a nonsynonymous mutation in each diagnostic gene group split by sex (n = 13,636 female patients; n = 17,422 male patients). In all panels, black, blue and orange represents consensus, discordant and novel genes, respectively.

To discover novel DD-associated genes with greater power, we applied DeNovoWEST to DNMs in patients without damaging DNMs in consensus genes (we refer to this subset as ‘undiagnosed’ patients) and identified 94 significant genes (Supplementary Fig. 7 and Supplementary Table 2), of which 33 were putative ‘novel’ DD-associated genes. To ensure robustness to potential mutation rate variation between genes, we determined whether any of the putative novel DD-associated genes had significantly more synonymous variants in the Genome Aggregation Database (gnomAD) of population variation than expected under our null mutation model (Supplementary Note). We identified 11 out of 33 genes with a significant excess of synonymous variants. For these 11 genes, we repeated the DeNovoWEST test, increasing the null mutation rate by the ratio of observed to expected synonymous variants in gnomAD. Five of these genes fell below our exome-wide significance threshold and were removed, leaving 28 novel genes, with a median of 10 nonsynonymous DNMs (Fig. 1b, c and Supplementary Table 3). There were 314 patients with nonsynonymous DNMs in these 28 genes (1.0% of our cohort); all of these DNMs were inspected in the Integrative Genomics Viewer (IGV) and, of the 198 patients for which experimental validation was attempted, all variants were confirmed to
The DNMs in these novel genes were distributed randomly across the three datasets (no genes with *P* < 0.001, heterogeneity test). In addition, 6 of the 28 novel DD-associated genes were corroborated by OMIM entries or publications, including TFE3, which was described in two recent publications. We also investigated whether some of the synonymous DNMs might be pathogenic by disrupting splicing. We identified a small but significant enrichment of synonymous DNMs with high values of the splicing pathogenicity score SpliceAI (≥0.8, 1.56-fold enriched, *P* = 0.0037, Poisson test; Supplementary Table 4). This enrichment corresponds to an excess of around 15 splice-disrupting synonymous DNMs in our cohort, of which 6 are accounted for by a recurrent synonymous DNM in *KAT6B* that is known to disrupt splicing.

Taken together, 25.0% of our cohort has a nonsynonymous DNM in one of the consensus or significant DD-associated genes (Fig. 1d). We noted significant sex differences in the autosomal burden of nonsynonymous DNMs (Supplementary Fig. 8). The rate of nonsynonymous DNMs in consensus autosomal genes was significantly higher in female individuals than male individuals (odds ratio = 1.16, *P* = 4.4 × 10⁻⁴, Fisher’s exact test; Fig. 1e), as noted previously. However, the exome-wide burden of autosomal nonsynonymous DNMs in all genes was not significantly different between undiagnosed male and female participants (odds ratio = 1.03, *P* = 0.29, Fisher’s exact test). This indicates that there are subtle sex differences in the genetic architecture of DDs, especially with regard to known and undescribed disorders. This could include sex-biased contributions of polygenic, oligogenic and/or environmental modifiers of phenotypic variation and thus clinical ascertainment.

**Characteristics of the novel DD-associated genes**

Based on 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 × 10⁻¹⁰, Wilcoxon rank-sum test; Fig. 2a and Supplementary Fig. 9). This suggests that these novel disorders less often result in distinctive and consistent clinical presentations, which may have made these disorders more difficult to discover using a phenotype-driven approach. Each of these novel disorders requires genotype–phenotype characterization, which is beyond the scope of this study.

Overall, novel DD-associated genes encode proteins that have very similar functional and evolutionary properties to consensus genes. The high-level functional similarity between known and novel DD-associated genes, nonsynonymous DNMs in the more recently described DD-associated genes are much more likely to be missense DNMs, and less likely to be PTVs. The DNMs in these novel genes were distributed randomly across the three datasets (no genes with *P* < 0.001, heterogeneity test). In addition, 6 of the 28 novel DD-associated genes were corroborated by OMIM entries or publications, including TFE3, which was described in two recent publications. We also investigated whether some of the synonymous DNMs might be pathogenic by disrupting splicing. We identified a small but significant enrichment of synonymous DNMs with high values of the splicing pathogenicity score SpliceAI (≥0.8, 1.56-fold enriched, *P* = 0.0037, Poisson test; Supplementary Table 4). This enrichment corresponds to an excess of around 15 splice-disrupting synonymous DNMs in our cohort, of which 6 are accounted for by a recurrent synonymous DNM in *KAT6B* that is known to disrupt splicing.

Taken together, 25.0% of our cohort has a nonsynonymous DNM in one of the consensus or significant DD-associated genes (Fig. 1d). We noted significant sex differences in the autosomal burden of nonsynonymous DNMs (Supplementary Fig. 8). The rate of nonsynonymous DNMs in consensus autosomal genes was significantly higher in female individuals than male individuals (odds ratio = 1.16, *P* = 4.4 × 10⁻⁴, Fisher’s exact test; Fig. 1e), as noted previously. However, the exome-wide burden of autosomal nonsynonymous DNMs in all genes was not significantly different between undiagnosed male and female participants (odds ratio = 1.03, *P* = 0.29, Fisher’s exact test). This indicates that there are subtle sex differences in the genetic architecture of DDs, especially with regard to known and undescribed disorders. This could include sex-biased contributions of polygenic, oligogenic and/or environmental modifiers of phenotypic variation and thus clinical ascertainment.

**Characteristics of the novel DD-associated genes**

Based on 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 × 10⁻¹⁰, Wilcoxon rank-sum test; Fig. 2a and Supplementary Fig. 9). This suggests that these novel disorders less often result in distinctive and consistent clinical presentations, which may have made these disorders more difficult to discover using a phenotype-driven approach. Each of these novel disorders requires genotype–phenotype characterization, which is beyond the scope of this study.

Overall, novel DD-associated genes encode proteins that have very similar functional and evolutionary properties to consensus genes (Fig. 2b and Supplementary Table 5). Despite the high-level functional similarity between known and novel DD-associated genes, nonsynonymous DNMs in the more recently described DD-associated genes are much more likely to be missense DNMs, and less likely to be PTVs. The DNMs in these novel genes were distributed randomly across the three datasets (no genes with *P* < 0.001, heterogeneity test). In addition, 6 of the 28 novel DD-associated genes were corroborated by OMIM entries or publications, including TFE3, which was described in two recent publications. We also investigated whether some of the synonymous DNMs might be pathogenic by disrupting splicing. We identified a small but significant enrichment of synonymous DNMs with high values of the splicing pathogenicity score SpliceAI (≥0.8, 1.56-fold enriched, *P* = 0.0037, Poisson test; Supplementary Table 4). This enrichment corresponds to an excess of around 15 splice-disrupting synonymous DNMs in our cohort, of which 6 are accounted for by a recurrent synonymous DNM in *KAT6B* that is known to disrupt splicing.

Taken together, 25.0% of our cohort has a nonsynonymous DNM in one of the consensus or significant DD-associated genes (Fig. 1d). We noted significant sex differences in the autosomal burden of nonsynonymous DNMs (Supplementary Fig. 8). The rate of nonsynonymous DNMs in consensus autosomal genes was significantly higher in female individuals than male individuals (odds ratio = 1.16, *P* = 4.4 × 10⁻⁴, Fisher’s exact test; Fig. 1e), as noted previously. However, the exome-wide burden of autosomal nonsynonymous DNMs in all genes was not significantly different between undiagnosed male and female participants (odds ratio = 1.03, *P* = 0.29, Fisher’s exact test). This indicates that there are subtle sex differences in the genetic architecture of DDs, especially with regard to known and undescribed disorders. This could include sex-biased contributions of polygenic, oligogenic and/or environmental modifiers of phenotypic variation and thus clinical ascertainment.

**Characteristics of the novel DD-associated genes**

Based on 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 × 10⁻¹⁰, Wilcoxon rank-sum test; Fig. 2a and Supplementary Fig. 9). This suggests that these novel disorders less often result in distinctive and consistent clinical presentations, which may have made these disorders more difficult to discover using a phenotype-driven approach. Each of these novel disorders requires genotype–phenotype characterization, which is beyond the scope of this study.
Poisson test; Supplementary Fig. 11), whereas recurrent synonymous mutations in the TCGA are not significantly enriched (2.4-fold, \( P = 0.13 \), Poisson test). These results suggest that this observation is driven by the pleiotropic effects of these mutations in development and tumorigenesis, rather than because of hypermutability of these variants.

**Recurrent mutations**

We identified 773 recurrent DNMs (736 SNVs and 37 indels), observed in 2–36 individuals, which enabled us to systematically interrogate the factors that drive recurrent germline mutations. We considered three potential contributory factors: (1) clinical ascertainment that enriches for pathogenic mutations; (2) greater mutability at specific sites; and (3) positive selection that confers a proliferative advantage in the male germline\(^2\). We observed evidence that all three factors contributed to the occurrence of recurrent germline mutations; however, these factors are not mutually exclusive. Clinical ascertainment drives the observation that 65% of recurrent DNMs were in consensus genes, a 5.4-fold enrichment compared with DNMs that were observed only once \( (P < 10^{-16}, \) proportion test). Hypermutability underpins the observation that 64% of recurrent de novo SNVs occurred at hypermutable CpG dinucleotides\(^2\), a 2.0-fold enrichment over DNMs that were observed only once \( (P = 3.3 \times 10^{-16}, \chi^2 \) test).

To assess the contribution of germline selection to recurrent DNMs, we initially focused on the 12 known germline selection genes, which all operate through activation of the RAS–MAPK signalling pathway\(^25,26\). We identified 39 recurrent DNMs in 11 of these genes, 38 of which are missense and all of which are known to be activating in the germline (see Supplementary Information). As expected, given that hypermutability is not the driving factor for the recurrent mutations in these genes, these 39 recurrent DNMs were depleted for CpGs relative to other recurrent mutations (6 out of 39 compared with 425 out of 692 other recurrent mutations, \( P = 3.4 \times 10^{-8}, \chi^2 \) test).

Positive germline selection can increase the apparent mutation rate more strongly\(^23\) than either clinical ascertainment (10–100× in our dataset) or hypermutability (around 10× for CpGs). However, only a minority of the most highly recurrent mutations in our 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. On the basis of the observations above, DNMs under germline selection should be more likely to be activating missense mutations, and should be less enriched for CpG dinucleotides. Extended Data Table 1 shows the 16 de novo SNVs that were observed 9 or more times in our cohort, only 2 of which are in known germline selection genes. All but 2 of these 16 de novo SNVs cause missense changes, all but 2 of these genes cause disease by an altered-function mechanism, and these DNMs were depleted for CpGs relative to all recurrent mutations. Two of these genes with highly recurrent de novo SNVs, in **SHOC2** and **PPPC1B**, which encode interacting proteins that regulate the RAS–MAPK pathway; pathogenic variants in these genes are associated with a Noonan-like syndrome\(^27\). Moreover, two of these recurrent DNMs are in the same gene (**SMAD4**), which encodes a key component of the TGFβ signalling pathway, potentially expanding the pathophysiology of germline selection beyond the RAS–MAPK pathway. Confirming germline selection of these mutations will require deep sequencing analyses of the testes and/or sperm\(^28\).

**Incomplete penetrance and pre- or perinatal death**

Nonsynonymous DNMs in consensus or significant DD-associated genes accounted for half of the exome-wide nonsynonymous DNBM burden associated with DDs (Fig. 1b). Despite our 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 our analysis nor on the consensus gene list). This residual burden of protein-truncating DNMs is greatest in genes that are intolerant to PTVs in the general population (Supplementary Fig. 12), which suggests that many haploinsufficient disorders have not yet been described. We observed that PTV mutability (estimated from a null germline mutation model) was significantly lower in unassociated genes compared with DD-associated genes \( (P = 4.5 \times 10^{-48}, \) Wilcoxon rank-sum test; Fig. 3a), which leads to reduced statistical power to detect DNM enrichment in unassociated genes, consistent with our hypothesis that numerous haploinsufficient disorders have not yet been identified.

A key parameter in estimating statistical power to detect novel haploinsufficiency is the fold enrichment of de novo PTVs expected in undescribed haploinsufficient disorders. We observed that novel DD-associated haploinsufficient genes had significantly lower PTV enrichment compared with the consensus haploinsufficient genes \( (P = 0.005, \) Wilcoxon rank-sum test; Fig. 3b). Two additional factors that could lower DNM enrichment, and thus the power to detect a novel DD association, are reduced penetrance and increased pre- or perinatal death (due to spontaneous fetal loss, termination of pregnancy because of a fetal anomaly, stillbirth or early neonatal death). To evaluate incomplete penetrance, we investigated whether haploinsufficient genes with a lower enrichment of de novo PTVs in our cohort are associated with a greater prevalence of PTVs in the general population. We observed a significant negative correlation \( (P = 0.031, \) weighted linear regression) between PTV enrichment in our cohort and the ratio of PTV to synonymous variants in gnomAD\(^6\), which suggests that incomplete penetrance does lower de novo PTV enrichment in our cohort (Fig. 3c).
Additionally, we observed that the fold enrichment of de novo PTVs in consensus haploinsufficient DD-associated genes in our cohort was significantly higher for genes with a low likelihood of presenting with a structural malformation of the fetus during prenatal screening \((P = 4.6 \times 10^{-8}, \text{Poisson test; Fig. 3d})\), which indicates that pre- or perinatal death decreases our power to detect some of the novel disorders (see Supplementary Information for details).

**Hundreds of DD genes have not yet been described**

Downsampling of our cohort and repeating enrichment analyses showed that the discovery of DD-associated genes has not plateaued (Extended Data Fig. 1a). Increasing the sample size should result in the discovery of many novel DD-associated genes. To estimate how many haploinsufficient genes have not yet been described, we modelled the likelihood of the observed distribution of de novo PTVs among genes as a function of varying numbers of undiscovered haploinsufficient DD-associated genes and fold enrichments of de novo PTVs in those genes. We found that the remaining PTV burden is most likely spread across around 1,000 genes with an approximately 10-fold PTV enrichment (Extended Data Fig. 1b). This fold enrichment is three times lower than in known haploinsufficient DD-associated genes, which suggests that incomplete penetrance and/or pre- or perinatal death is more prevalent among undiscovered haploinsufficient genes. We modelled the missense DNM burden separately and also observed that the most likely architecture of undiscovered DD-associated genes is one that comprises more than 1,000 genes with a substantially lower fold enrichment than in currently known DD-associated genes (Supplementary Fig. 13).

We calculated that a sample size of around 350,000 parent–offspring trios would be needed to have 80% power to detect a tenfold enrichment of de novo PTVs for an average gene. Using this inferred tenfold enrichment among undiscovered haploinsufficient genes, from our current data we can evaluate the likelihood that any gene is an undiscovered haploinsufficient 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 tenfold increased PTV rate. Among the approximately 19,000 non-DD-associated genes, around 1,200 were more than three times more likely to have arisen from a tenfold increased PTV rate, whereas approximately 7,000 were three times more likely to have no de novo PTV enrichment.

**Discussion**

Here we describe 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 parent–offspring trios. Most of the increased power to detect novel disorders comes from the increase in sample size, rather than the improved statistical test. These 28 novel genes account for 1.0% of our cohort, and their inclusion in diagnostic workflows will help to improve diagnosis of similar patients globally. The value of this study for improving diagnostic yield extends beyond these 28 novel genes; the total number of genes added to diagnostic workflows of the three participating centres (including newly validated discordant genes) ranged from 48 to 65 genes. We show that both incomplete penetrance and pre- or perinatal death reduced our power to detect novel DDs postnatally, and hypothesize that one or both of these factors are operating more strongly among undiscovered DD-associated genes. In addition, we identify a set of highly recurrent mutations that are strong candidates for novel germline selection mutations, which should result in a higher than expected disease incidence that increases markedly with increased paternal age.

Our study 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. We identified around 2.3 times as many significantly DD-associated genes as this previous study when using Bonferroni-corrected exome-wide significance (285 compared with 124). In contrast to meta-analyses of published DNMs, the harmonized filtering of candidate DNMs across cohorts in this study should be more robust to cohort-specific differences in the sensitivity and specificity of detecting DNMs.

We inferred indirectly that developmental disorders with higher rates of detectable prenatal structural abnormalities had a greater likelihood of pre- or 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. In the PAGE study, genetic diagnoses were not returned to participants during the pregnancy, and so genetic diagnostic information could not influence the incidence of pre- or perinatal death. In the PAGE study data, 69% of fetal abnormalities with a genetically diagnosable cause died perinatally or neonatally. This emphasizes the substantial effect that pre- or 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 studies.

To empower our mutation enrichment testing, we estimated positive predictive values that each DNM was pathogenic on the basis of their predicted protein consequence, CADD score, selective constraint against heterozygous PTVs across the gene \((s_{\text{het}})\), and, for missense variants, presence in a region under selective missense constraint. These positive predictive values should also be informative for variant prioritization in the diagnosis of dominant developmental disorders. Further work is needed to investigate whether these positive predictive values could be informative for recessive developmental disorders, and in other types of dominant disorders. More generally, we hypothesize that empirically estimated positive predictive values based on variant enrichment in large datasets will be similarly informative in many other disease areas.

We adopted a conservative statistical approach to identifying DD-associated genes. In two previous studies using the same significance threshold, we identified 26 novel DD-associated genes. All 26 are now regarded as being diagnostic, and have entered routine clinical diagnostic practice. Had we used a significance threshold with a false-discovery rate of <10% as used previously, we would have identified 770 DD-associated genes. The false-discovery rate of individual genes depends on the significance of other genes being tested, which means that it is not appropriate for assessing the significance of individual genes, but can be useful for defining gene sets. There are 184 consensus genes that did not cross our significance threshold in this study. It is likely that many of these genes cause disorders that were underrepresented in our study due to the ease of clinical diagnosis on the basis of distinctive clinical features or targeted diagnostic testing. These ascertainment biases will not affect the representation of novel DDs in our cohort.

Our modelling suggests that there are probably more than 1,000 DD-associated genes that remain to be discovered, and that reduced penetrance and pre- or perinatal death will reduce our power to identify these genes using 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 our analyses of 31,058 trios, we observed evidence that mutation rate heterogeneity among genes can lead to overestimation of the statistical significance of mutation enrichment based on an exome-wide mutation model. We advocate the development of more granular mutation rate models, based on large-scale population variation resources, that correct for all technical and biological complexities, to ensure that larger studies are robust to mutation rate heterogeneity.

We anticipate that the variant-level weights used by DeNovoWEST will improve over time. As reference population samples, such as
gnomAD\textsuperscript{9}, increase in size, weights based on selective constraint metrics (for example, \(s_{\text{M}}\), or regional missense constraint) will improve. Weights could also incorporate more functional information, such as expression in disease-relevant tissues. For example, we observe that DD-associated genes are significantly more likely to be expressed in the fetal brain (Supplementary Fig. 14). Furthermore, new metrics based on gene co-regulation networks can predict whether genes function within a disease-relevant pathway\textsuperscript{9}. As a cautionary note, including more functional information may increase power to detect some new disorders while decreasing power for disorders with a pathophysiology that is different from known disorders. Our analyses also suggest that variant-level weights could be further improved by incorporating other variant prioritization metrics, such as upweighting variants predicted to affect splicing, variants in particular protein domains or variants that are somatic driver mutations during tumorigenesis. In developing DeNovoWEST, we explored the application of both variant-level weights and gene-level weights in separate stages of the analysis; however, subtle but pervasive correlations between gene-level metrics (for example, \(s_{\text{M}}\)) and variant-level metrics (for example, regional missense constraint or CADD) present 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\textsuperscript{10}. Nonetheless, using current methods focused on DNMs alone, we estimated that around 350,000 parent–child trios would need to be analysed to have around 80% power to detect haploinsufficient genes with a tenfold PTV enrichment. Discovering trios would need to be analysed to have around 80% power to detect haploinsufficient genes with a tenfold PTV enrichment. Discovering non-haploinsufficient disorders will need even larger sample sizes. Reaching this number of sequenced families will not be possible for an individual research study or clinical centre; it is therefore 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 previously undescribed disorders and improve diagnostic practice.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2832-5.

1. Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. Nature 542, 433–438 (2017).
2. Martin, H. G. et al. Quantifying the contribution of recessive coding variation to developmental disorders. Science 362, 1161–1164 (2018).
3. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. 46, 310–315 (2014).
4. Samocha, K. E. et al. Regional missense constraint improves variant deleteriousness prediction. Preprint at https://doi.org/10.1101/148353 (2017).
5. Kosmicki, J. A. et al. Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. Nat. Genet. 49, 504–510 (2017).
6. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581, 434–443 (2020).
7. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285–291 (2016).
8. Cooper, G. M. et al. A copy number variation morbidity map of developmental delay. Nat. Genet. 43, 838–846 (2011).
9. Coe, B. P. et al. Refining analyses of copy number variation identifies specific genes associated with developmental delay. Nat. Genet. 46, 1063–1071 (2014).
10. Robinson, J. T. et al. Integrative genomics viewer. Nat. Biotechnol. 29, 24–26 (2011).
11. Villegas, F. et al. Lysosomal signaling licenses embryonic stem cell differentiation via inactivation of Tc6. Cell Stem Cell 24, 257–270 (2019).
12. Diaz, J., Berger, S. & Leon, E. TFE3-associated neurodevelopmental disorder: a distinct recognizable syndrome. Am. J. Med. Genet. A 182, 584–590 (2020).
13. Jaganathan, K. et al. Predicting splicing from primary sequence with deep learning. Cell 176, 535–548 (2019).
14. Yilmaz, R. et al. A recurrent synonymous KAT6B mutation causes Say-Barber-Biesecker/Young-Simpson syndrome by inducing aberrant splicing. Am. J. Med. Genet. A 167, 3006–3010 (2015).
15. Wu, X., Pang, E., Lin, K. & Pei, Z.-M. Improving the measurement of semantic similarity between gene ontology terms and gene products: insights from an edge- and IC-based hybrid method. PLoS ONE 8, e66745 (2013).
16. Catterall, W. A., Dib-Hajj, S., Messieler, M. H. & Pietrobon, D. Inherited neuronal ion channelopathies: new windows on complex neurological diseases. J. Neurosci. 20, 11768–11777 (2008).
17. Lasser, M., Tiber, J. & Lowery, L. A. The role of the microtubule cytoskeleton in neurodevelopmental disorders. Front. Cell. Neurosci. 12, 165 (2018).
18. Hamilton, M. J. et al. Heterozygous mutations affecting the protein kinase domain of CDKL13 cause a syndromic form of developmental delay and intellectual disability. J. Med. Genet. 55, 28–38 (2018).
19. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. Cell 161, 1029–1041 (2017).
20. Qi, H., Dong, C., Chung, W. K., Wang, K. & Shen, Y. Deep genetic connection between cancer and developmental disorders. Hum. Mutat. 37, 1042–1050 (2016).
21. Ronan, J. L., Wu, W. & Crabtree, G. R. From neural development to cognition: unexpected roles for chromatin. Nat. Rev. Genet. 14, 347–359 (2013).
22. Cancer Genome Atlas Research et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. 45, 1113–1210 (2013).
23. Gorinly, A. & Wilkie, A. O. M. Paternal age effect mutations and selfish spermatogonial selection: causes and consequences for human disease. Am. J. Hum. Genet. 90, 175–200 (2012).
24. Duncan, B. K. & Miller, J. H. Mutagenic deamination of cytosine residues in DNA. Nature 287, 560–561 (1980).
25. Maher, G. J. et al. Visualizing the origins of selfish de novo mutations in individual seminiferous tubules of human testes. Proc. Natl Acad. Sci. USA 113, 2454–2459 (2016).
26. Maher, G. J. et al. Selfish mutations dysregulating RAS-MAPK signaling are pervasive in aged human testes. Genome Res. 28, 1779–1790 (2018).
27. Young, L. C. et al. SHOC2–MRAS–PPP complex positively regulates RAF activity and contributes to Noonan syndrome pathogenesis. Proc. Natl Acad. Sci. USA 115, E10576–E10585 (2018).
28. Coe, B. P. et al. Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. Nat. Genet. 51, 106–116 (2019).
29. Lord, J. et al. Prenatal exome sequencing analysis in fetal structural anomalies detected by ultrasonography (PAGE): a cohort study. Lancet 393, 747–757 (2019).
30. Cassa, C. A. et al. Estimating the selective effects of heterozygous protein-truncating variants from human exome data. Nat. Genet. 49, 806–810 (2017).
31. Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. Nature 519, 223–228 (2015).
32. Satterstrom, F. K. et al. Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. Cell 180, 568–584 (2020).
33. Deelen, P. et al. Improving the diagnostic yield of exome-sequencing by predicting gene–phenotype associations using large-scale gene expression analysis. Nat. Commun. 10, 2837 (2019).
34. He, X. et al. Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. PLoS Genet. 9, e1003671 (2013).

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
Deciphering Developmental Disorders Study

Silvia Borrás, Caroline Clark, John Dean, Zosia Miedzybrodzka, Alison Ross, Stephen Tennant, Tabib Dabi, Deirdre Donnelly, Mervyn Humphreys, Alex Magee, Vivienne McConnell, Shane McKee, Susan McNerlan, Patrick J. Morrison, Gillian Rea, Fiona Stewart, Trevor Cottle, Nicola Cooper, Lisa Cooper-Charles, Helen Cox, Lily Islam, Joanna Jarvis, Rebecca Keelagher, Derek Lim, Dominic McMullan, Jenny Morton, Swati Naik, Mary O’Driscoll, Kai-Ren Ong, Deborah Osio, Nicola Ragge, Sarah Turton, Julie Vogt, Denise Williams, Simon Bodek, Alan Donaldson, Alison Hills, Karen Low, Ruth Newbury-Ecob, Andrew M. Norman, Eileen Roberts, Ingrid Scarr, Sarah Smithson, Madeleine Tooley, Steve Abbis, Ruth Armstrong, Carolyn Dunn, Simon Holden, Soo-Mi Park, Joan Paterson, Lucy Raymond, Evan Reid, Richard Sandford, Ingrid Simonic, Marc Tischkowitz, Geoff Woods, Lisa Bradley, Joanna Comerford, Andrew Green, Sally Lynch, Shirley McQuaid, Brendan Mullany, Jonathan Berg, David Goudie, Eleni Navrak, Joanne McLean, Catherine McWilliam, Eleanor Reavey, Tara Azam, Elaine Cleary, Andrew Jackson, Wayne Lam, Anne Lampe, David Moore, Mary Porteous, Emma Bapte, Julia Baptista, Carole Brewer, Bruce Castle, Emma Kiviua, Martina Owens, Julia Rankin, Charles Shaw-Smith, Claire Turner, Peter Turnpenny, Carolyn Tysoe, Therese Bradley, Rosemarie Davidson, Carol Gardiner, Sholagh Joss, Esther Kinning, Cheryl Longman, Ruth McGowan, Victoria Murday, Danieli Pilz, Edwin Tobias, Margo Whiteford, Nicola Williams, Angela Barnicot, Emma Clement, Francesca Faravelli, Jane Hurst, Lucy Jenkins, Wendy Jones, V. K. Ajith Kumar, Melissa Lees, Sam Loughlin, Alison Male, Deborah Morrogh, Elisabeth Rosser, Richard Scott, Louise Wilson, Ana Beleta, Charu Deshpande, Frances Flintron, Muriel Holder, Melita Irving, Louise Izatt, Dragana Josifovska, Shelia Mohammed, Aneta Molenda, Leema Robert, Wendy Roworth, Deborah Roddy, Mia Ryten, Shu Yau, Christopher Bennett, Moira Blyth, Jennifer Campbell, Andrea Coates, Angius Dobbie, Sarah Hewitt, Emma Hobson, Eilidh Jackson, Rosalyn Jewett, Alkis Kraus, Katrina Prescott, Eamonn Sheridan, Jenny Thomson, Kirsty Bradshaw, Abhijit Dixit, Jacqueline Eason, Rebecca Haines, Rachel Harrison, Stacey Mitchum, Ajoy Sarkar, Claire Searle, Nora Shannon, Abd Safdah, Mohnish Suri, Pradeep Vasudevan, Natalie Canham, Ian Ellis, Lynn Greenhalgh, Emma Howard, Victoria Stinton, Andrew Swale, Astrid Weber, Siddharth Banka, Catherine Breen, Tracy Briggs, Emma Burkill-Wright, Kate Chandler, Jill Clayton-Smith, Dian Donnai, Sofia Douzgouz, Lorraine Gaunt, Elizabeth Jones, Bronwyn Kerr, Claire Langley, Kate Mcfall, Audrey Smith, Ronnie Wright, David Bourn, John Burn, Richard Fisher, Steve Hellins, Alex Henderson, Tara Montgomery, Miranda Split, Volker Straub, Michael Wright, Simon Zwindon, Zoe Allen, Birgitta Bernhard, Angela Brady, Claire Brooks, Louise Busby, Virginia Cloves, Neeti Ghali, Susan Holder, Rita Ilbıyely, Emma Wadling, Edward Blair, Jenny Carmichael, Deirdre Cilliers, Susan Clasper, Richard Gibbons, Usha Kiris, Tracy Lester, Andrea Nemeth, Joanna Poulton, Sue Price, Debbie Shears, Helen Stewart, Andrew Wilkie, Shadi Alibaba, Duncan Baker, Meena Balasubramanians, Diana Johnson, Michael Parkers, Oliver Quarell, Alison Stewart, Josh Willoughby, Charlene Crosby, Frances Elmslie, Tessa Homfray, Huilin Jin, Nynaya Lahiri, Sahar Mansour, Karen Marks, Meriel McEntagart, Anand Saggar, Kate Tatton-Brown, Rachel Butler, Angus Clarke, Sian Corriss, Andrew Fry, Arven Kamath, Emma McCann, Hood Mugalai, Caroline Pettinger, Annie Proctor, Julian Sampson, Francis Sanbury, Vinod Varghese, Diana Baralle, Alison Callaway, Emma J. Cassidy, Stacey Daniels, Andrew Douglas, Nicola Foulds, David Hunt, Mira Kharbanda, Katherine Lachlan, Catherine Mercer, Lucy Side, I. Karen Temple & Diana Wellesley.
Reporting summary
Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability
Sequence and variant-level data and phenotypic data for the DDD study data are available from the European Genome-phenome Archive (EGA; https://www.ebi.ac.uk/ega/) with study ID EGAS00001000775. The RadboudUMC sequence and variant-level data cannot be made available through the EGA owing to the nature of consent for clinical testing. To access the data, please contact C.G. (christian.gilissen@radboudumc.nl) with a request. Data sharing will be dependent on patient consent, diagnostic status of the patient, the type of request and the potential benefit to the patient. GeneDx data cannot be made available through the EGA owing to the nature of consent for clinical testing. GeneDx-referred patients are consented for aggregate, deidentified research and subject to US HIPAA privacy protection. As such, we are not able to share patient-level BAM or VCF data, which are potentially identifiable without a HIPAA Business Associate Agreement. Access to the deidentified aggregate data used in this analysis is available upon request to GeneDx. GeneDx has contributed deidentified data to this study to improve clinical interpretation of genomic data, in accordance with patient consent and in conformance with the ACMG position statement on genomic data sharing (details are provided in the Supplementary Note). Clinically interpreted variants and associated phenotypes from the DDD study are available through DECIPHER (https://decipher.sanger.ac.uk). Clinically interpreted variants from RUMC are available from the Dutch national initiative for sharing variant classifications (https://www.vkgl.nl/nl/diagnostiek/vkgl-datashare-database) as well as LOVD (https://databases.lovd.nl/shared/variants), where they are listed with 'VKGL-NL_Nijmegen' as the owner. Clinically interpreted variants from GeneDx are deposited in ClinVar (https://www.ncbi.nlm.nih.gov/clinvar) under accession number 26957 (https://www.ncbi.nlm.nih.gov/clinvar/submitters/26957/). Previously described datasets were from the Genome Aggregation Database (gnomAD v2.1.1; https://gnomad.broadinstitute.org/), The Cancer Genome Atlas (TCGA; https://portal.gdc.cancer.gov) and the Developmental Disorders Genotype-Phenotype Database (DDG2P; https://www.ebi.ac.uk/gene2phenotype/downloads).

Code availability
The DeNovoWEST method is available on GitHub (https://github.com/queenjobo/DeNovoWEST) along with code to recreate all of the figures in the manuscript (https://doi.org/10.5281/zenodo.3909398). Code to run the Phenopy method is also available on GitHub (https://github.com/GeneDx/phenopy).

Acknowledgements
We thank the families and their clinicians for their participation and engagement, and our colleagues who assisted in the generation and processing of data. Inclusion of RadboudUMC data was in part supported by the Solve-RD project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 779257. This work was in part financially supported by grants from the Netherlands Organization for Scientific Research (897-17-353 to C.G.). The DDD study presents independent research commissioned by the Health Innovation Challenge Fund (grant number HICF-1009-003). This study makes use of DECIPHER, which is funded by the Wellcome Trust. The full acknowledgements can be found at www.ddduk.org/access.html. The DDD study authors acknowledges the work of R. Kelsell. Finally, we acknowledge the contribution of an esteemed DDD clinical collaborator, M. Bitner-Glindicz, who died during the course of the study.

Author contributions
J.K., K.E.S., L.W., K.J.A., M.E.H., C.G. and K.R. contributed to the generation of figures and writing of the manuscript. J.K., K.E.S., L.W., Z.Z., K.J.A., R.Y.E., G.G., S.H.L., H.C.M., J.F.M., E.d.B., R.P., M.R.F.R. and H.G.Y. contributed to the generation and quality control of data. J.K., K.E.S., L.W., Z.Z., K.J.A., R.I.T., J.F.M., P.D., E.J.G., N.H., J.L., I.M., A.Y. and K.R. developed methods, contributed data or performed analyses. H.C.M., L.E.L.M.V., J.J., C.F.W., H.G.B., H.V.F., D.R.F., J.C.B., M.E.H., C.G. and K.R. provided experimental and analytical supervision. M.E.H., C.G. and K.R. provided project supervision.

Competing interests
Z.Z., K.J.A., R.I.T., J.J. and K.R. are employees of GeneDx. J.J. and K.R. are shareholders of OPKO. M.E.H. is a co-founder of, consultant to and holds shares in Congenica, a genetics diagnostic company.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2832-5.

Correspondence and requests for materials should be addressed to M.E.H.

Peer review information
Nature thanks Ipsita Agarwal, James Lupski, Shamil Sunyaev and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints.
Extended Data Fig. 1 | Exploring the remaining number of DD genes.  
**a**, Number of significant genes after downsampling the full cohort and running the enrichment test of DeNovoWEST. **b**, The likelihood of the observed distribution of de novo PTV mutations was modelled. 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 genes that are significant and known to operate through a loss-of-function mechanism (as indicated by an arrow) is 39.7.
Extended Data Table 1 | Recurrent mutations

| Symbol | Chr | Position | Ref | Alt | Consequence | 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   | 40065606 | A   | G   | missense    | 10    | all missense/in frame | -   | -                  | -                      | consensus |
| SHOC2  | 10  | 112724120| A   | G   | missense    | 9     | activating       | -   | -                  | -                      | consensus |
| PTPN11 | 15  | 112915523| A   | G   | missense    | 9     | activating       | -   | 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 |

De novo SNVs with more than nine recurrences in our cohort annotated with relevant information, such as CpG status, whether the affected gene is a known somatic driver or germline-selection gene, and diagnostic gene group (for example, consensus). ‘Recur’ refers to the number of recurrences. ‘Likely mechanism’ refers to the mechanisms attributed to this gene in the published literature.
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
  Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. mean) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted. Give P values as exact values wherever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

Each center [e.g. GeneDx, DDD, Radboud UMC] generated its own data.

Data analysis

We put all code and data needed to recreate the main text figures in a Github repository:
https://github.com/queenjobo/DeNovoWEST
DOI: 0.5281/zenodo.3909398
Code to run the Phenopy method is also available on Github:
https://github.com/GeneDx/phenopy
We downloaded SpliceAI scores (whole_genome_filtered_spliceai_scores.vcf.gz) from the Illumina’s BaseSpace website on 30 November 2018.

For manuscripts utilizing custom algorithms or software that are not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. Github). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Sequence and variant level data and phenotypic data for the DDD study data are available through EGA study ID EGAS000001000775
RadboudUMC sequence and variant level data cannot be made available through EGA due to the nature of consent for clinical testing
GeneDx data cannot be made available through EGA due to the nature of consent for clinical testing. GeneDx has contributed deidentified data to this study to improve clinical interpretation of genomic data, in accordance with patient consent and in conformance with the ACMG position statement on genomic data sharing
Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- [x] Life sciences  
- [ ]Behavioural & social sciences  
- [ ] Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/hr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| Sample size | This is an observational study; the sample size is 31,058 exome sequenced parent-child trios. No power calculations were done prior to analyses; we used all available samples. |
|-------------|--------------------------------------------------------------------------------------------------|
| Data exclusions | As detailed in our supplement, we removed individuals who were duplicates across sequencing centers, which was a pre-established criteria of our project. For all analyses, we use the final 31,058 trios and 45,221 de novo mutations. |
| Replication | There is no replication within this study, but we highlight that any new genes should be followed up with a detailed genotype-phenotype study. |
| Randomization | NA. This was a case-only study, so there was no need for randomization. |
| Blinding | NA. This was a case-only study, so there was no need for blinding. |

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

| n/a | Involved in the study |
|-----|-----------------------|
| X   | Antibodies            |
| X   | Eukaryotic cell lines |
| X   | Palaeontology         |
| X   | Animals and other organisms |
| X   | Human research participants |
| X   | Clinical data         |

### Methods

| n/a | Involved in the study |
|-----|-----------------------|
| X   | ChiP-seq              |
| X   | Flow cytometry        |
| X   | MRI-based neuroimaging |
**Human research participants**

**Policy information about studies involving human research participants**

**Population characteristics**

Individuals with developmental disorders, as well as their parents, were recruited to one of three studies/centers: the Deciphering Developmental Disorders (DDD) Study, GeneDx, or Radboud University Medical Center (RUMC). After removing duplicate individuals, there were 31,058 unique sequenced trios, which included 17,422 male probands and 13,636 female probands. All probands have a severe neurodevelopmental disorder and/or congenital anomalies, abnormal growth parameters, dysmorphic features, and unusual behavioural phenotypes. For RUMC, all probands have intellectual disability (IQ < 70). No covariates (e.g., age) were used in this study.

**Recruitment**

Recruitment varied by center. More details are given in the supplement.

Deciphering Developmental Disorders (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. These analyses involve 9,858 trios from 3,307 families, a subset of whom have been analysed in previous publications. Patients typically had prior genetic testing (e.g., an array or a single gene test) before recruitment into the study.

GeneDx: Patients were referred to GeneDx for clinical whole-exome sequencing for diagnosis of suspected Mendelian disorders as described in Retterer et al 2016. Patients were selected for inclusion in this study based on having one or more HPO phenotypes overlapping the inclusion criteria for the DDD study.

Radboud University Medical Center: The Department of Human Genetics from the Radboud University Medical Center (RUMC) is a tertiary referral center for clinical genetics. For this study, we selected all individuals with intellectual disability who had family-based whole-exome sequencing using the Agilent SureSelect v4 and v5 enrichment kit combined with sequencing on the Illumina HiSeq platform in the time period 2013-2018.

Since patients in DDD and RUMC were typically screened for known causes of developmental disorders before recruitment into their respective studies, these two cohorts tend to be depleted of previously established and clinically recognizable causes of developmental disorder (e.g., Trisomy 21, Kabuki syndrome, etc.).

**Ethics oversight**

**DDD:** The study was approved by the UK Research Ethics Committee [10/H0305/83 granted by the Cambridge South Research Ethics Committee, and GEN/218/12 granted by the Republic of Ireland Research Ethics Committee].

GeneDx: The study was conducted in accordance with all guidelines set forth by the Western Institutional Review Board, Puyallup, WA [WIRB 20162523].

Radboud University Medical Center: This study was approved by the institutional review board ‘Commissie Mensgebonden Onderzoek Regio Arnhem-Nijmegen’ under number 2011/188.

Note that full information on the approval of the study protocol must also be provided in the manuscript.