Integrating healthcare and research genetic data empowers the discovery of 49 novel developmental disorders

Joanna Kaplanis[1]*, Kaitlin E. Samocha[1]*, Laurens Wiel[2,3]*, Zhancheng Zhang[4]*, Kevin J. Arvai[4], Ruth Y. Eberhardt[1], Giuseppe Gallone[1], Stefan H. Lelieveld[2], Hilary C. Martin[1], Jeremy F. McRae[1], Patrick J. Short[1], Rebecca I. Torene[4], Elke de Boer[5], Petr Danecek[1], Eugene J. Gardner[1], Ni Huang[1], Jenny Lord[1,6], Iñigo Martincorena[1], Rolph Pfundt[5], Margot R. F. Reijnders[2], Alison Yeung[7,8], Helger G. Yntema[2], DDD Study, Lisenka E. L. M. Vissers[5], Jane Juusola[4], Caroline F. Wright[9], Han G. Brunner[5,10], Helen V. Firth[1,11], David R. FitzPatrick[12], Jeffrey C. Barrett[1], Matthew E. Hurles[1#]†, Christian Gilissen[2#], Kyle Retterer[4#]

[1] Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK
[2] Department of Human Genetics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, 6525 GA, the Netherlands
[3] Centre for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, 6525 GA, the Netherlands
[4] GeneDx, Gaithersburg, Maryland, USA
[5] Department of Human Genetics, Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Center, Nijmegen, 6525 GA, the Netherlands
[6] Human Development and Health, Faculty of Medicine, University of Southampton, UK
[7] Victorian Clinical Genetics Services, Melbourne, Australia
[8] Murdoch Children's Research Institute, Melbourne, Australia
[9] Institute of Biomedical and Clinical Science, University of Exeter Medical School, Research, Innovation, Learning and Development building, Royal Devon & Exeter Hospital, Barrack Road, Exeter EX2 5DW, UK
[10] Department of Clinical Genetics, GROW school for oncology and developmental biology, and MHENS school for mental health and neuroscience, Maastricht University Medical Centre, Maastricht, 6202 AZ, the Netherlands
[11] Department of Clinical Genetics, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK
[12] MRC Human Genetics Unit, MRC IGMM, University of Edinburgh, Western General Hospital, Edinburgh, UK

* contributed equally
[#] contributed equally
† To whom correspondence should be addressed: meh@sanger.ac.uk

Summary

De novo mutations (DNMs) in protein-coding genes are a well-established cause of developmental disorders (DD). However, known DD-associated genes only account for a minority of the observed excess of such DNMs. To identify novel DD-associated genes, we integrated healthcare and research exome sequences on 31,058 DD parent-offspring trios, and developed a simulation-based statistical test to identify gene-specific enrichments of DNMs. We identified 299 significantly DD-associated genes, including 49 not previously robustly associated with DDs. Despite detecting more DD-associated genes than in any previous study, much of the excess of DNMs of protein-coding genes remains unaccounted for. Modelling suggests that over 500 novel DD-associated genes await discovery, 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 dominant DDs.

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[1,2]. 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 relevant genes left to find. 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. However, missense variants and protein-truncating variants (PTVs) vary in their impact on protein function[3–6]. Known dominant DD-associated genes are strongly enriched in the minority of genes that exhibit patterns of strong selective constraint on heterozygous PTVs in the general population[7]. 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[8,9].

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

Following clear consent practices and only using aggregate, de-identified data, we pooled DNMs in patients with severe developmental disorders 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**; **Supplementary Table 1**), which includes data on over 24,000 trios not previously published. These DNMs included 40,992 single nucleotide variants (SNVs) and 4,229 indels. The three cohorts have similar clinical characteristics, male/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). DeNovoWEST scores all classes of sequence variants on a unified severity scale based on the empirically-estimated positive predictive value of being pathogenic (**Supplementary Fig. 3-4**). We then applied a Bonferroni multiple testing correction with independent hypothesis weighting[10] to incorporate a gene-based weighting using the selective constraint against heterozygous PTVs in the general population (s_{het}[11]), which is strongly correlated with the likelihood of being a dominant disease gene[7,11].

We first applied DeNovoWEST to all individuals in our cohort and identified 299 significant genes, 35 more than when using our previous method[1] (**Supplementary Fig. 5**; **Fig. 1a**). The majority (181/299; 61%) of these significant genes already had sufficient evidence of

DD-association to be considered of diagnostic utility (as of September 2018) by all three centres, and we refer to them as "consensus" genes. 69/299 of these significant genes were previously considered diagnostic by one or two centres ("discordant" genes). Applying DeNovoWEST to synonymous DNMs, as a negative control analysis, identified no significantly enriched genes (**Supplementary Fig. 6**). To discover novel DD-associated genes with greater power, we then applied DeNovoWEST only to DNMs in patients without damaging DNMs in consensus genes (we refer to this subset as 'undiagnosed' patients) and identified 118 significant genes (**Fig. 1b; Supplementary Fig. 7; Supplementary Table 2**). While 69 of these genes were discordant genes, we identified 49 'novel' DD-associated genes, which had a median of 10 nonsynonymous DNMs in our dataset (**Fig. 1c**; **Supplementary Table 3**). There were 500 patients with nonsynonymous DNMs in these 49 genes (1.6% of our cohort); all DNMs in these genes were inspected in IGV[12] 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). Fourteen of the 49 novel DD-associated genes have been further corroborated by recent OMIM entries or publications. In particular, seven of these 14 genes (*PPP2CA*[13], *ZMIZ1*[14], *CDK8*[15], *VAMP2*[16], *KMT2E*[17], *KDM6B*[18], and *TAOK1*[19]) have had genotype-phenotype studies recently published.

We also investigated whether some synonymous DNMs might be pathogenic by disrupting splicing. We annotated all synonymous DNMs with a splicing pathogenicity score, SpliceAI[20], and identified a significant enrichment of synonymous DNMs with high SpliceAI scores (≥ 0.8, 1.56-fold enriched, p = 0.0037, Poisson test; **Supplementary Table 4**). This enrichment corresponds to an excess of ~15 splice-disrupting synonymous mutations in our cohort, of which six are accounted for by a single recurrent synonymous mutation in *KAT6B* known to disrupt splicing[21].

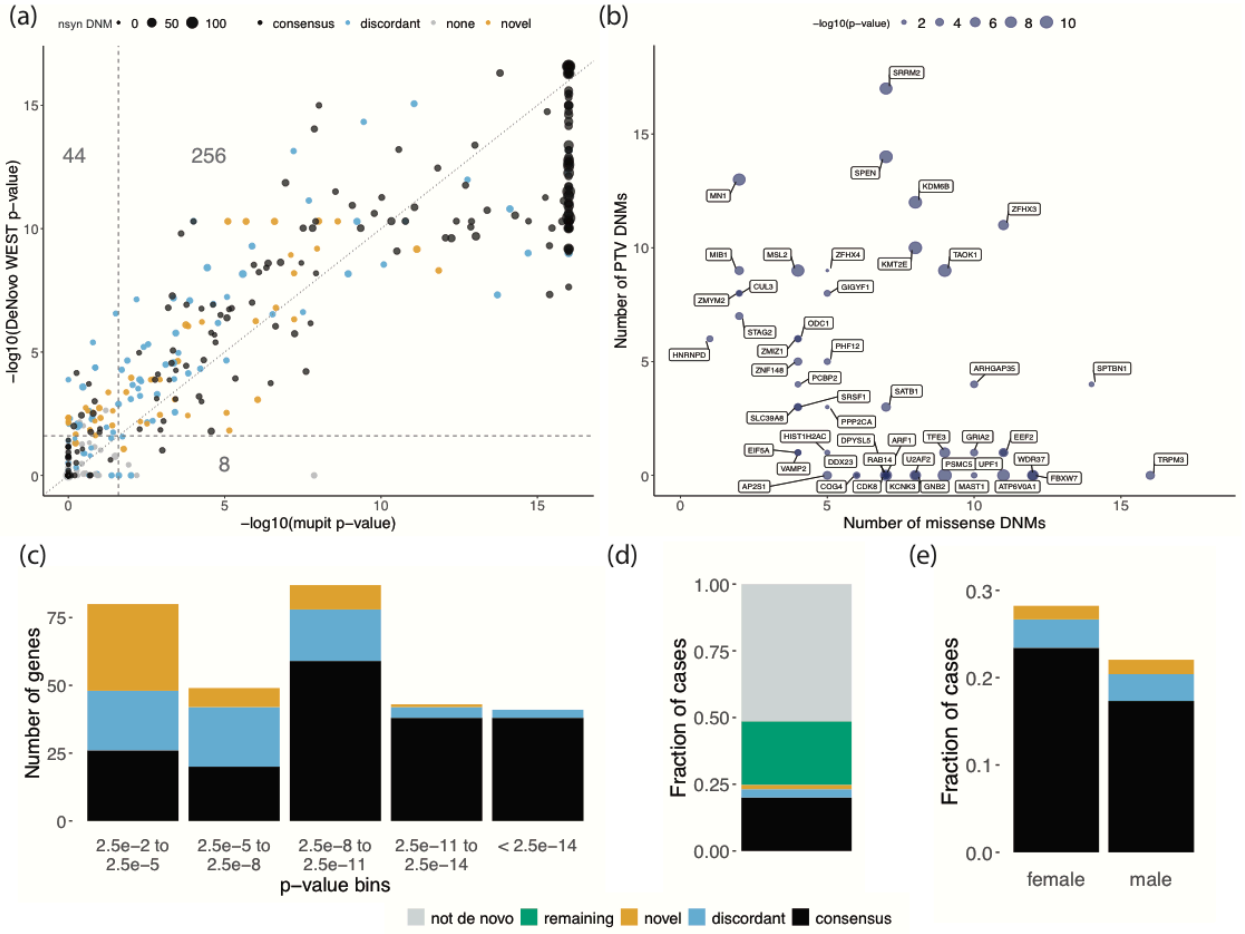

Figure 1: Results of DeNovoWEST analysis. (a) Comparison of p-values generated using the new method (DeNovoWEST) versus the previous method (mupit)[1]. These are results from DeNovoWEST run on the full cohort. The dashed lines indicate the threshold for genome-wide significance (the p-values have already been corrected for multiple testing). 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 (43 in the top left, 256 in the top right, and 8 in the bottom right). (b) The number of missense and PTV DNMs in our cohort in the 49 novel genes. The size of the points are proportional to the log10(-p-value) from the analysis on the undiagnosed subset. (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 p-value bin is coloured by diagnostic gene group. (d) The fraction of cases with a nonsynonymous mutation in each diagnostic gene group. (e) The fraction of cases with a nonsynonymous mutation in each diagnostic gene group split by sex. In all figures, black represents the consensus known genes, blue represents the discordant known genes, and orange represents the novel genes. In (c), 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").

Taken together, 24.8% of individuals in our combined cohort have 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 females than males (OR = 1.17, p = 1.1 x 10^{-7}, Fisher's exact test; **Fig. 1e**), as noted previously[1]. However, the exome-wide burden of autosomal nonsynonymous DNMs in all genes was not significantly different between undiagnosed males and females (OR = 1.03, p = 0.29, Fisher's exact test). This suggests the existence of subtle sex differences in the genetic architecture of DD, especially with regard to known and undiscovered disorders.

Characteristics of the novel DD-associated genes and disorders

Based on semantic similarity[22] 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 = 9.5 x 10^{-38}, Wilcoxon rank-sum test; **Fig. 2a**). 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, 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, e.g. developmental expression patterns, network properties and biological functions (**Fig. 2b; Supplementary Table 5**). 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 = 3.3 x 10^{-21}, chi-squared test). Sixteen of the 49 (33%) 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 = 6.5 x 10^{-10}) had one inframe deletion and nine missense DNMs, eight of which altered one of two amino acids that interact within the 3D protein structure: p.Pro320Arg and p.Arg325Trp (**Supplementary Fig. 9a-b**), and so is likely to operate via an altered-function mechanism. Additionally, we identified one novel DD-associated gene, *MN1*, with *de novo* PTVs significantly (p = 1.6 x 10^{-7}, Poisson test) clustered at the 3' end of its transcript (**Supplementary Fig. 9c**). This clustering of PTVs indicates the transcript likely escapes nonsense mediated decay and potentially acts via a gain-of-function or dominant negative mechanism[23], although this will require functional confirmation.

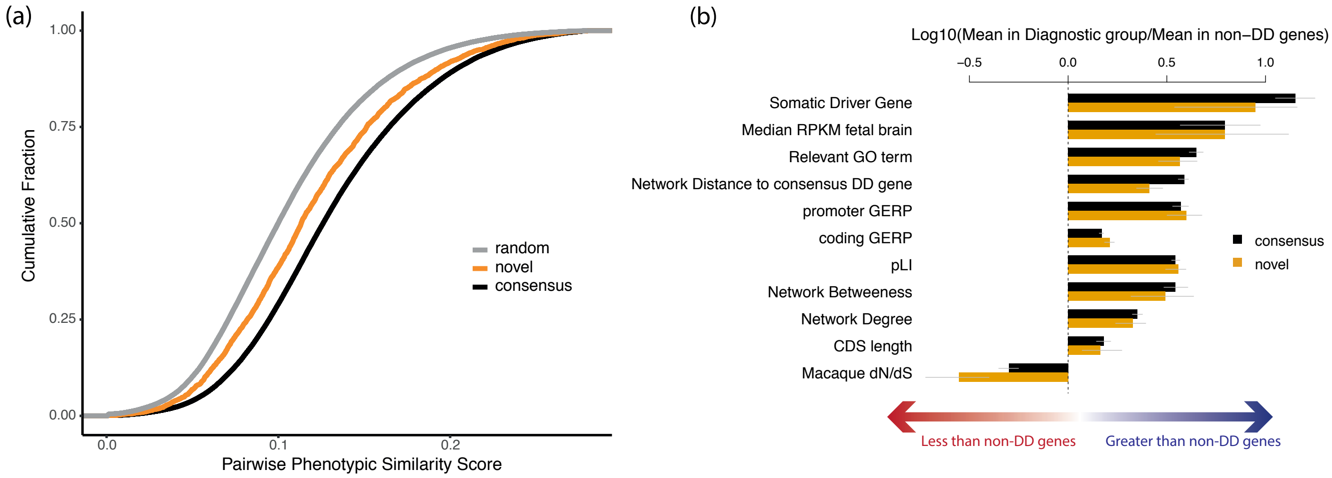

Figure 2: Functional properties and mechanisms of novel genes. (a) Comparing the phenotypic similarity of patients with DNMs in novel and consensus genes. Random phenotypic similarity was calculated from random pairs of patients. Patients with DNMs in the same novel DD-associated gene were less phenotypically similar than patients with DNMs in a known DD-associated gene ($p = 9.5 \times 10^{-38}$, Wilcoxon rank-sum test). (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.

We observed that missense DNMs were more likely to affect functional protein domains than other coding regions. We observed a 2.76-fold enrichment ($p = 1.6 \times 10^{-68}$, G-test) of missense DNMs residing in protein domains among consensus genes and a 1.87-fold enrichment ($p = 1.4 \times 10^{-4}$, G-test) in novel DD-associated genes, but no enrichment for synonymous DNMs (**Supplementary Table 6**). Three protein domain families in consensus genes were specifically enriched for missense DNMs (**Supplementary Table 7**): ion transport protein (PF00520, $p = 3.9 \times 10^{-7}$, G-test), ligand-gated ion channel (PF00060, $p = 6.7 \times 10^{-7}$, G-test), and protein kinase domain (PF00069, $p = 4.4 \times 10^{-2}$, G-test). Missense DNMs in all three enriched domain families, have previously been associated with DD (**Supplementary Table 8**)[24].

We observed a significant overlap between the 299 DNM-enriched DD-associated genes and a set of 369 previously described cancer driver genes[25] ($p = 1.7 \times 10^{-46}$, logistic regression correcting for s_{het}), as observed previously[26,27], as well as a significant enrichment of nonsynonymous DNMs in these genes (**Supplementary Table 9**). This overlap extends to somatic driver mutations: we observe 117 DNMs at 76 recurrent somatic mutations observed in at least three patients in The Cancer Genome Atlas (TCGA)[28]. By modelling the germline mutation rate at these somatic driver mutations, we found that recurrent nonsynonymous mutations in TCGA are enriched 21-fold in the DDD cohort ($p < 10^{-50}$, Poisson test, **Supplementary Fig. 9**), whereas recurrent synonymous mutations in TCGA are not significantly enriched (2.4-fold, p = 0.13, Poisson test). This suggests that this observation is driven by the pleiotropic effects of these mutations in development and tumourigenesis, rather than hypermutability.

Recurrent mutations and potential new germline selection genes

We identified 773 recurrent DNMs (736 SNVs and 37 indels), ranging from 2-36 independent observations per DNM, which allowed us to interrogate systematically the factors driving recurrent germline mutation. We 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[29]. We 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 68% of recurrent *de novo* SNVs occurred at hypermutable CpG dinucleotides[30], a 1.8-fold enrichment over DNMs only observed once ($p = 1.1 \times 10^{-59}$, proportion test**)**. We also observed a striking enrichment of recurrent mutations at the haploinsufficient DD-associated gene *MECP2*, in which we observed 11 recurrently mutated SNVs within a 500bp 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[5], suggesting that locus-specific hypermutability might explain this observation.

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[31,32]. 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 supplement). 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 (9/39 vs 450/692, p = 0.0067, chi-squared test).

Positive germline selection has been shown to be capable of increasing the apparent mutation rate more strongly[29] than either clinical ascertainment (10-100X in our dataset) or hypermutability (~10X 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. 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 1** 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[33]. 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[32].

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	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 1: Recurrent Mutations. *De novo* single nucleotide variants with more than 9 recurrences in our 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 number of recurrence. “Likely mechanism” refers to mechanisms attributed to this gene in the published literature.

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 (**Fig. 1b**). Despite our identification of 299 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). The remaining burden of protein-truncating DNMs is greatest in genes that are intolerant of PTVs in the general population (**Supplementary Fig. 11**) suggesting that more haploinsufficient (HI) disorders await discovery. We estimated that our statistical power to detect the gene enrichment for protein-truncating DNMs expected for an HI disorder was lower in unassociated genes compared to the novel DD-associated genes ($p = 2.9 \times 10^{-6}$ Wilcox rank-sum test; **Fig. 3a**). However, the novel genes do not have significantly less power compared to the consensus genes ($p = 0.059$, Wilcox rank-sum test).

A key parameter in the above power analysis is the fold-enrichment of *de novo* PTVs expected in as yet undiscovered HI disorders, which we assumed above to be 37-fold, based on the average enrichment observed in known HI DD-associated genes. However, we observed that novel DD-associated HI genes had significantly lower PTV enrichment compared to the consensus HI genes ($p = 1.6 \times 10^{-5}$, Poisson test; **Fig. 3b**). 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, we 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. We observed a significant ($p = 0.031$, weighted linear regression) negative correlation between PTV enrichment in our cohort and the ratio of PTV to synonymous variants in the gnomAD dataset of population variation[5], suggesting that incomplete penetrance does lower *de novo* PTV enrichment in our cohort (**Fig. 3c**).

Additionally, we observed that the fold-enrichment of protein-truncating DNMs in consensus HI DD-associated genes in our cohort was significantly lower for genes with a medium or high likelihood of presenting with a prenatal structural malformation ($p = 0.0002$, Poisson test, **Fig. 3d**), suggesting that pre/perinatal death decreases our power to detect some novel DD-associated disorders (see supplement for details).

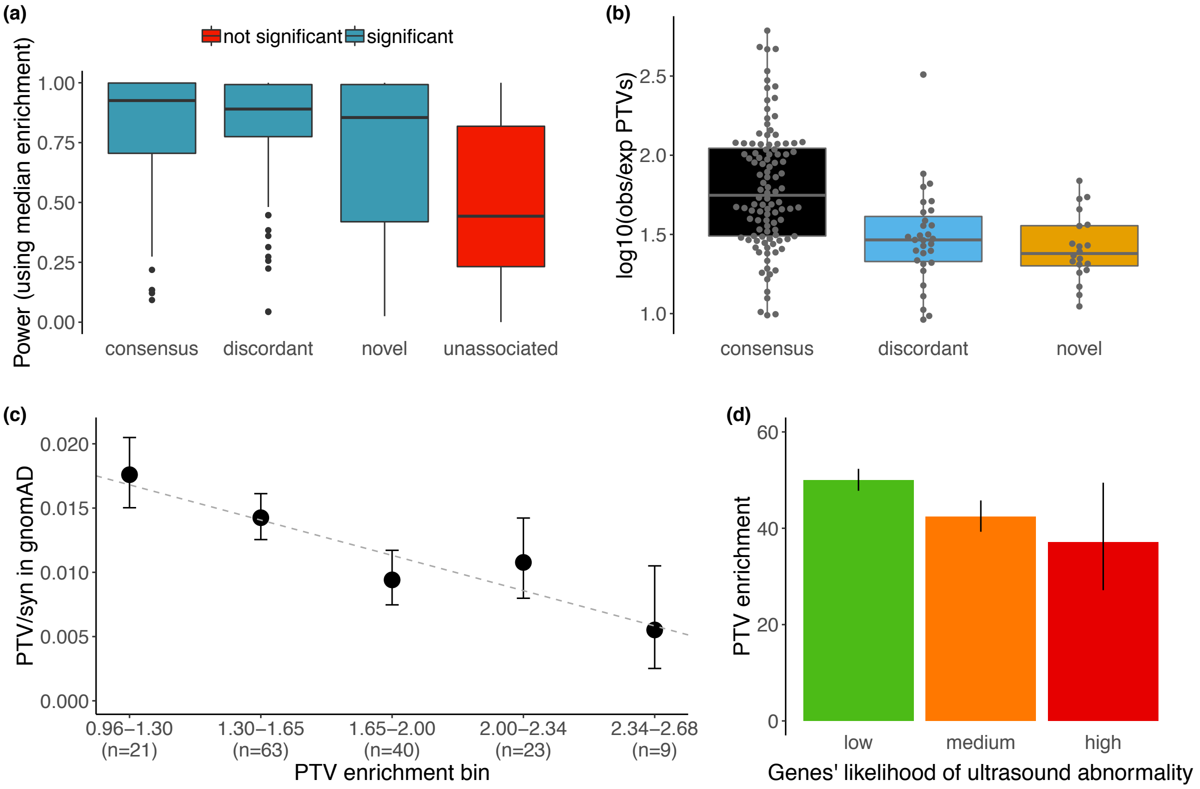

Figure 3: Impact of pre/perinatal death and penetrance on power. (a) Power to detect median PTV enrichment (36.6) in consensus known genes across diagnostic groups. Genes that were not significantly associated to DD in our analysis ("unassociated") have significantly lower power than the novel genes (p = 2.9 x 10^{-6}, Wilcox rank-sum test). (b) Distribution of PTV enrichment in significant, likely haploinsufficient, genes by diagnostic group. (c) Comparison of the PTV enrichment in our cohort vs the PTV to synonymous ratio found in gnomAD. PTV enrichment is shown as log10(enrichment). There is a significant negative relationship (p = 0.031, weighted regression). (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. PTV enrichment is significantly lower for genes with a medium or high likelihood compared to genes with a low likelihood (p = 0.002, Poisson test).

Modelling reveals hundreds of DD genes remain to be discovered

To understand the likely trajectory of future DD discovery efforts, we downsampled the current cohort and reran our enrichment analysis (**Fig. 4a**). We 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, we 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. We found that the remaining HI burden is most likely spread across ~500 genes with ~12-fold PTV enrichment (**Fig. 4b**). This fold enrichment is three times lower than in known HI DD-associated genes, suggesting that incomplete penetrance and/or pre/perinatal death is much more prevalent among undiscovered HI 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 over 500 genes with a substantially lower fold-enrichment than in currently known DD-associated genes (**Supplemental Fig. 12**).

We calculated that a sample size of ~200,000 parent-offspring trios would be needed to have 80% power to detect a 12-fold enrichment of protein-truncating DNMs for a gene with the median PTV mutation rate among currently unassociated genes. Using this inferred 12-fold enrichment among undiscovered HI genes, from our current data we 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 12-fold increased PTV rate. Among the ~19,000 non-DD-associated genes, ~1,100 were more than three times more likely to have arisen from a 12-fold increased PTV rate, whereas ~9,000 were three times more likely to have no *de novo* PTV enrichment.

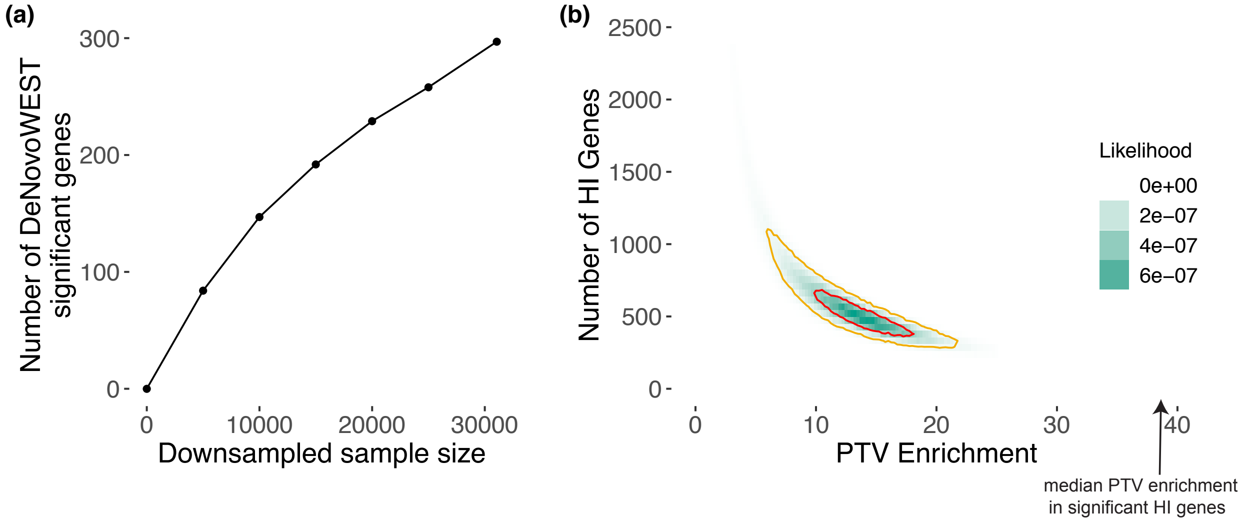

Figure 4: Exploring the remaining number of DD genes. (a) Number of significant genes from downsampling full cohort and running DeNovoWEST's enrichment test. (b) Results from modelling the likelihood of the observed distribution of *de novo* PTV mutations. This model varies the numbers of remaining haploinsufficient (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 38.8.

Discussion

In this study, we have discovered 49 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 49 novel genes account for up to 1.6% of our cohort, and inclusion of these genes in diagnostic workflows will catalyse increased diagnosis of similar patients globally. We 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, we 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.

Our 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[34]. Our analysis included DNMs from 24,348 previously unpublished trios, and we identified ~2.4 times as many significantly DD-associated genes as this previous study when using Bonferroni-corrected exome-wide significance (299 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 we 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[35]. 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 fetal abnormalities with a genetically diagnosable cause 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 our mutation enrichment testing, we estimated positive predictive values (PPV) of each DNM being pathogenic on the basis of their predicted protein consequence, CADD score[3] and presence in a region or gene under missense constraint in the general population[4]. 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, we hypothesise that empirically-estimated PPVs 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[1,36]. All 26 are now regarded as being diagnostic, and have entered routine clinical diagnostic practice. Had we used a significance threshold of FDR < 10% as used in Satterstrom, Kosmicki, Wang et al[37], we would have identified 737 DD-associated genes. However, as the FDR of individual genes depends on the significance of other genes being tested, FDR thresholds are not appropriate for assessing the significance of individual genes, but rather for defining gene-sets. There are 150 consensus genes that did not cross our significance threshold in this study. It is likely that many of these cause disorders that were under-represented in our 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 our cohort.

Our 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 our power to identify these genes through DNM enrichment. Identifying these genes will require both improved analytical methods and greater sample sizes. We anticipate that the variant-level and gene-level weights used by DeNovoWEST will improve over time. As reference population samples, such as gnomAD[5], increase in size, gene-level weights based on selective constraint metrics will improve. Gene-level weights could also incorporate more functional information, such as expression in disease-relevant tissues. For example, we observe that our DD-associated genes are significantly more likely to be expressed in fetal brain (**Supplementary Fig. 13**). Furthermore, novel metrics based on gene co-regulation networks can predict whether genes function within a disease relevant pathway[38]. As a cautionary note, including more functional information in the gene-level weights 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. Finally, the discovery of less penetrant disorders can be empowered by analytical methodologies that integrate both DNMs and rare inherited variants, such as TADA[39]. Nonetheless, using current methods, we estimated that ~200,000 parent-child trios would need to be analysed to have ~80% power to detect HI genes with a 12-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 is shared with the research community such that it can be aggregated to drive discovery of novel disorders and improve diagnostic practice.

Acknowledgements

We thank the families and their clinicians for their participation and engagement. We are very grateful to 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. 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 Wellcome. See www.ddduk.org/access.html for full acknowledgement. The DDD study would like to acknowledge the tireless work of Rosemary Kelsell. Finally we acknowledge the contribution of an esteemed DDD clinical collaborator, M. Bitner-Glindicz, who died during the course of the study.

Data Access

Sequence and variant level data and phenotypic data for the DDD study data are available through EGA study ID EGAS00001000775

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 (see Supplementary Note for details).

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)

Clinically interpreted variants from GeneDx are deposited in ClinVar (https://www.ncbi.nlm.nih.gov/clinvar)

References

1. Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433–438 (2017).
2. Martin, H. C. *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. *bioRxiv* 148353 (2017). doi:10.1101/148353
5. Karczewski, K. J. *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv* 531210 (2019). doi:10.1101/531210
6. 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).
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. Ignatiadis, N., Klaus, B., Zaugg, J. B. & Huber, W. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nat. Methods* **13**, 577–580 (2016).
11. Cassa, C. A. *et al.* Estimating the selective effects of heterozygous protein-truncating variants from human exome data. *Nat. Genet.* **49**, 806–810 (2017).
12. Robinson, J. T. *et al.* Integrative genomics viewer. *Nature Biotechnology* **29**, 24–26 (2011).

13. Reynhout, S. *et al.* De Novo Mutations Affecting the Catalytic Cα Subunit of PP2A, PPP2CA, Cause Syndromic Intellectual Disability Resembling Other PP2A-Related Neurodevelopmental Disorders. *Am. J. Hum. Genet.* **104**, 139–156 (2019).

14. Carapito, R. *et al.* ZMIZ1 Variants Cause a Syndromic Neurodevelopmental Disorder. *Am. J. Hum. Genet.* **104**, 319–330 (2019).

15. Calpena, E. *et al.* De Novo Missense Substitutions in the Gene Encoding CDK8, a Regulator of the Mediator Complex, Cause a Syndromic Developmental Disorder. *Am. J. Hum. Genet.* **104**, 709–720 (2019).

16. Salpietro, V. *et al.* Mutations in the Neuronal Vesicular SNARE VAMP2 Affect Synaptic Membrane Fusion and Impair Human Neurodevelopment. *Am. J. Hum. Genet.* **104**, 721–730 (2019).

17. O'Donnell-Luria, A. H. *et al.* Heterozygous Variants in KMT2E Cause a Spectrum of Neurodevelopmental Disorders and Epilepsy. *Am. J. Hum. Genet.* **104**, 1210–1222 (2019).

18. Stolerman, E. S. *et al.* Genetic variants in the KDM6B gene are associated with neurodevelopmental delays and dysmorphic features. *Am. J. Med. Genet. A* **179**, 1276–1286 (2019).

19. Dulovic-Mahlow, M. *et al.* De Novo Variants in TAOK1 Cause Neurodevelopmental Disorders. *Am. J. Hum. Genet.* (2019). doi:10.1016/j.ajhg.2019.05.005

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

21. 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* **167A**, 3006–3010 (2015).

22. 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).

23. Coban-Akdemir, Z. *et al.* Identifying Genes Whose Mutant Transcripts Cause Dominant Disease Traits by Potential Gain-of-Function Alleles. *Am. J. Hum. Genet.* **103**, 171–187

(2018).

24. Catterall, W. A., Dib-Hajj, S., Meisler, M. H. & Pietrobon, D. Inherited neuronal ion channelopathies: new windows on complex neurological diseases. *J. Neurosci.* **28**, 11768–11777 (2008).
25. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **173**, 1823 (2018).
26. 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).
27. Ronan, J. L., Wu, W. & Crabtree, G. R. From neural development to cognition: unexpected roles for chromatin. *Nat. Rev. Genet.* **14**, 347–359 (2013).
28. Cancer Genome Atlas Research Network *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
29. Goriely, 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).
30. Duncan, B. K. & Miller, J. H. Mutagenic deamination of cytosine residues in DNA. *Nature* **287**, 560–561 (1980).
31. Maher, G. J. *et al.* Visualizing the origins of selfish de novo mutations in individual seminiferous tubules of human testes. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 2454–2459 (2016).
32. Maher, G. J. *et al.* Selfish mutations dysregulating RAS-MAPK signaling are pervasive in aged human testes. *Genome Res.* **28**, 1779–1790 (2018).
33. Young, L. C. *et al.* SHOC2-MRAS-PP1 complex positively regulates RAF activity and contributes to Noonan syndrome pathogenesis. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E10576–E10585 (2018).
34. Coe, B. P. *et al.* Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. *Nat. Genet.* **51**, 106–116 (2019).
35. Lord, J. *et al.* Prenatal exome sequencing analysis in fetal structural anomalies detected by

ultrasonography (PAGE): a cohort study. *Lancet* **393**, 747–757 (2019).

36. Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**, 223–228 (2015).

37. Kyle Satterstrom, F. *et al.* Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *bioRxiv* 484113 (2019). doi:10.1101/484113

38. 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).

39. 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).