Integrating healthcare and research genetic data empowers the discovery of 28 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[2,7], Alison Yeung[8,9], Helger G. Yntema[5], DDD Study, Lisenka E. L. M. Vissers[5], Jane Juusola[4], Caroline F. Wright[10], Han G. Brunner[5,7,11], Helen V. Firth[1,12], David R. FitzPatrick[13], 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] Department of Clinical Genetics, Maastricht University Medical Centre, Maastricht, 6202 AZ, the Netherlands
Victorian Clinical Genetics Services, Melbourne, Australia
[9] Murdoch Children's Research Institute, Melbourne, Australia
[10] 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
[11] GROW school for oncology and developmental biology, and MHENS school for mental health and neuroscience, Maastricht University Medical Centre, Maastricht, 6202 AZ, the Netherlands
[12] Department of Clinical Genetics, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK
[13] 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 285 significantly DD-associated genes, including 28 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 1,000 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.

54 **Introduction**

55 It has previously been estimated that ~42-48% of patients with a severe developmental
56 disorder (DD) have a pathogenic *de novo* mutation (DNM) in a protein coding gene[1,2]. However,
57 over half of these patients remain undiagnosed despite the identification of hundreds of
58 dominant and X-linked DD-associated genes. This implies that there are more DD relevant
59 genes left to find. Existing methods to detect gene-specific enrichments of damaging DNMs
60 typically ignore much prior information about which variants and genes are more likely to be
61 disease-associated. However, missense variants and protein-truncating variants (PTVs) vary in
62 their impact on protein function[3–6]. Known dominant DD-associated genes are strongly enriched
63 in the minority of genes that exhibit patterns of strong selective constraint on heterozygous
64 PTVs in the general population[7]. To identify the remaining DD genes, we need to increase our
65 power to detect gene-specific enrichments for damaging DNMs by both increasing sample sizes
66 and improving our statistical methods. In previous studies of pathogenic Copy Number Variation
67 (CNV), utilising healthcare-generated data has been key to achieve much larger sample sizes
68 than would be possible in a research setting alone[8,9].

69

70 **Improved statistical enrichment test identifies 285 significant DD-associated genes**

71 Following clear consent practices and only using aggregate, de-identified data, we
72 pooled DNMs in patients with severe developmental disorders from three centres: GeneDx (a
73 US-based diagnostic testing company), the Deciphering Developmental Disorders study, and
74 Radboud University Medical Center. We performed stringent quality control on variants and
75 samples to obtain 45,221 coding and splicing DNMs in 31,058 individuals (**Supplementary Fig.**
76 **1**; **Supplementary Table 1**), which includes data on over 24,000 trios not previously published.
77 These DNMs included 40,992 single nucleotide variants (SNVs) and 4,229 indels. The three
78 cohorts have similar clinical characteristics, male/female ratios, enrichments of DNMs by
79 mutational class, and prevalences of known disorders (**Supplementary Fig. 2**).

80 To detect gene-specific enrichments of damaging DNMs, we developed a method
81 named DeNovoWEST (*De Novo* Weighted Enrichment Simulation Test,
82 https://github.com/queenjobo/DeNovoWEST). DeNovoWEST scores all classes of sequence
83 variants on a unified severity scale based on the empirically-estimated positive predictive value
84 of being pathogenic (**Supplementary Fig. 3-4**). We perform two tests per gene: the first is an
85 enrichment test on all nonsynonymous DNMs and the second is a test designed to detect genes
86 likely acting via an altered-function mechanism. This second test combines an enrichment test
87 on missense DNMs with a test of linear clustering of missense DNMs within the gene. We then
88 applied a Bonferroni multiple testing correction accounting for 18,762 x 2 tests, which takes into
89 account the number of genes and two tests per gene.

90 We first applied DeNovoWEST to all individuals in our cohort and identified 281
91 significant genes, 18 more than when using our previous method[1] (**Supplementary Fig. 5**; **Fig.**
92 **1a**). The majority (196/281; 70%) of these significant genes already had sufficient evidence of
93 DD-association to be considered of diagnostic utility (as of late 2019) by all three centres, and
94 we refer to them as "consensus" genes. 54/281 of these significant genes were previously
95 considered diagnostic by one or two centres ("discordant" genes). Applying DeNovoWEST to
96 synonymous DNMs, as a negative control analysis, identified no significantly enriched genes
97 (**Supplementary Fig. 6**).
98 To discover novel DD-associated genes with greater power, we then applied
99 DeNovoWEST only to DNMs in patients without damaging DNMs in consensus genes (we refer
100 to this subset as 'undiagnosed' patients) and identified 94 significant genes (**Fig. 1b;**
101 **Supplementary Fig. 7; Supplementary Table 2**). While 61 of these genes were discordant
102 genes, we identified 33 putative 'novel' DD-associated genes. To further ensure robustness to
103 potential mutation rate variation between genes, we determined whether any of the putative
104 novel DD-associated genes had significantly more synonymous variants in the Genome
105 Aggregation Database[5] (gnomAD) of population variation than expected under our null mutation
106 model (Supplementary Note). We identified 11/33 genes with a significant excess of
107 synonymous variants. For these 11 genes we then repeated the DeNovoWEST test, increasing
108 the null mutation rate by the ratio of observed to expected synonymous variants in gnomAD.
109 Five of these genes then fell below our exome-wide significance threshold and were removed,
110 leaving 28 novel genes, with a median of 10 nonsynonymous DNMs in our dataset (**Fig. 1c**;
111 **Supplementary Table 3**). There were 314 patients with nonsynonymous DNMs in these 28
112 genes (1.0% of our cohort); all DNMs in these genes were inspected in IGV[10] and, of 198 for
113 which experimental validation was attempted, all were confirmed as DNMs in the proband. The
114 DNMs in these novel genes were distributed approximately randomly across the three datasets
115 (no genes with $p < 0.001$, heterogeneity test). Six of the 28 novel DD-associated genes are
116 further corroborated by OMIM entries or publications, including *TFE3*[11,12] for which patients were
117 described in two recent publications.
118 We also investigated whether some synonymous DNMs might be pathogenic by
119 disrupting splicing. We annotated all synonymous DNMs with a splicing pathogenicity score,
120 SpliceAI[20], and identified a significant enrichment of synonymous DNMs with high SpliceAI
121 scores (≥ 0.8, 1.56-fold enriched, $p = 0.0037$, Poisson test; **Supplementary Table 4**). This
122 enrichment corresponds to an excess of ~15 splice-disrupting synonymous mutations in our
123 cohort, of which six are accounted for by a single recurrent synonymous mutation in *KAT6B*
124 known to disrupt splicing[21].
125

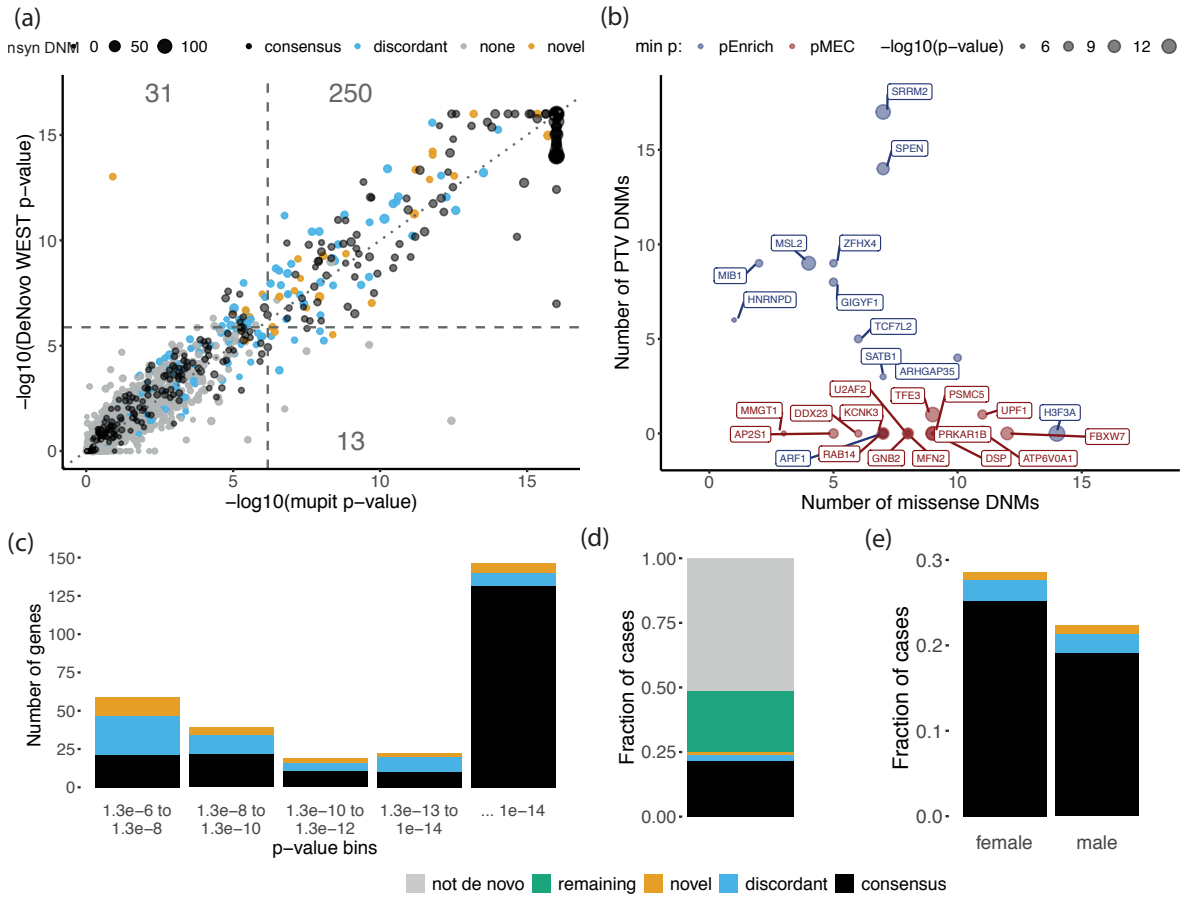

Figure 1: Results of DeNovoWEST analysis. (a) Comparison of p-values generated using the new method (DeNovoWEST) versus the previous method (mupit)[1]. These are results from DeNovoWEST run on the full cohort. The dashed lines indicate the threshold for genome-wide significance. The size of the points is proportional to the number of nonsynonymous DNMs in our cohort (nsyn). The numbers describe the number of genes that fall into each quadrant (b) The number of missense and PTV DNMs in our cohort in the novel genes. The size of the points are proportional to the $\log_{10}$(-p-value) from the analysis on the undiagnosed subset. The colour corresponds to which test p-value was the minimum (more significant) for these genes: non-synonymous enrichment test in blue (pEnrich), or missense enrichment and clustering test in red (pMEC). (c) 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 genes, blue represents the discordant 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*").

145 Taken together, 25.0% of individuals in our combined cohort have a nonsynonymous
146 DNM in one of the consensus or significant DD-associated genes (**Fig. 1d**). We noted
147 significant sex differences in the autosomal burden of nonsynonymous DNMs (**Supplementary**
148 **Fig. 8**). The rate of nonsynonymous DNMs in consensus autosomal genes was significantly
149 higher in females than males (OR = 1.16, $p = 4.4 \times 10^{-7}$, Fisher's exact test; **Fig. 1e**), as noted
150 previously[1]. However, the exome-wide burden of autosomal nonsynonymous DNMs in all genes
151 was not significantly different between undiagnosed males and females (OR = 1.03, p = 0.29,
152 Fisher's exact test). This suggests the existence of subtle sex differences in the genetic
153 architecture of DD, especially with regard to known and undiscovered disorders. This could, for
154 example, include sex-biased contribution of polygenic and/or environmental causes of DDs.
155
156 **Characteristics of the novel DD-associated genes and disorders**
157 Based on semantic similarity[22] between Human Phenotype Ontology terms, patients with
158 DNMs in the same novel DD-associated gene were less phenotypically similar to each other, on
159 average, than patients with DNMs in a consensus gene ($p = 2.3 \times 10^{-11}$, Wilcoxon rank-sum test;
160 **Fig. 2a; Supplementary Figure 9**). This suggests that these novel disorders less often result in
161 distinctive and consistent clinical presentations, which may have made these disorders harder
162 to discover via a phenotype-driven analysis or recognise by clinical presentation alone. Each of
163 these novel disorders requires a detailed genotype-phenotype characterisation, which is beyond
164 the scope of this study.
165 Overall, novel DD-associated genes encode proteins that have very similar functional
166 and evolutionary properties to consensus genes, e.g. developmental expression patterns,
167 network properties and biological functions (**Fig. 2b; Supplementary Table 5**). Despite the
168 high-level functional similarity between known and novel DD-associated genes, the
169 nonsynonymous DNMs in the more recently discovered DD-associated genes are much more
170 likely to be missense DNMs, and less likely to be PTVs (discordant and novel; $p = 1.2 \times 10^{-25}$,
171 chi-squared test). Fifteen of the 28 (54%) of the novel genes only had missense DNMs, and
172 only a minority had more PTVs than missense DNMs. Consequently, we expect that a greater
173 proportion of the novel genes will act via altered-function mechanisms (e.g. dominant negative
174 or gain-of-function). For example, the novel gene *PSMC5* (DeNovoWEST $p = 2.6 \times 10^{-15}$) had
175 one inframe deletion and nine missense DNMs, eight of which altered two structurally important
176 amino acids that are both in the AAA+ ATPase domain within the 3D protein structure:
177 p.Pro320Arg and p.Arg325Trp (**Supplementary Fig. 10a-b**), and so is likely to operate via an
178 altered-function mechanism. None of the novel genes exhibited significant clustering of *de novo*
179 PTVs.

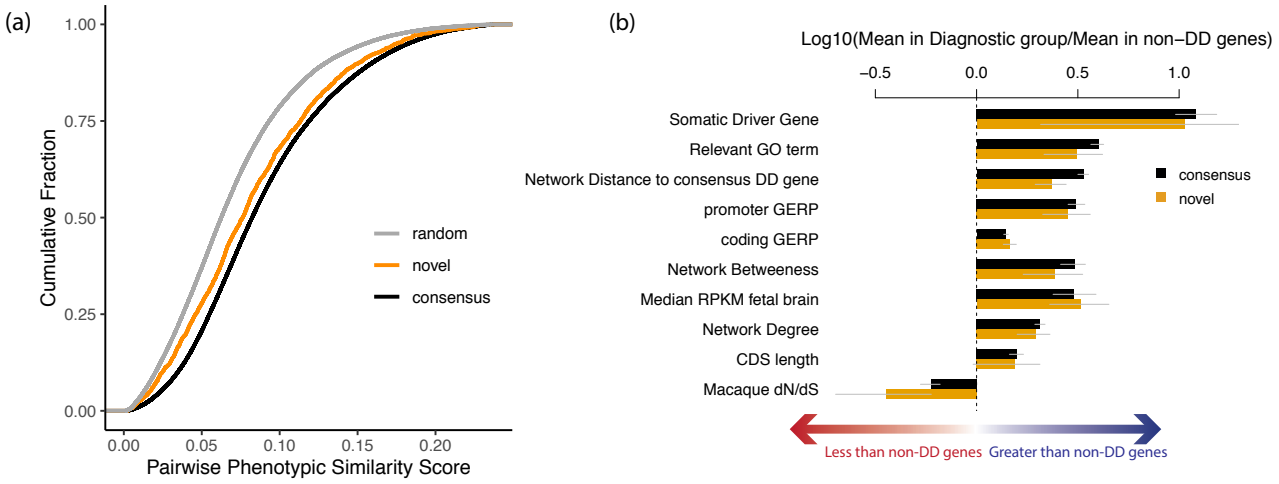

Figure 2: Functional properties and mechanisms of novel genes. (a) Comparing the phenotypic similarity of patients with DNMs in novel and consensus genes. Random phenotypic similarity was calculated from random pairs of patients. Patients with DNMs in the same novel DD-associated gene were less phenotypically similar than patients with DNMs in a known DD-associated gene ($p = 2.3 \times 10^{-11}$, Wilcoxon rank-sum test). (b) Comparison of functional properties of consensus 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 2.63-fold enrichment ($p = 2.2 \times 10^{-68}$, G-test) of missense DNMs residing in protein domains among consensus genes and a 1.80-fold enrichment ($p = 8.0 \times 10^{-5}$, G-test) in novel DD-associated genes, but no enrichment for synonymous DNMs (**Supplementary Table 6**). Four protein domain families in consensus genes were consistently enriched for missense DNMs (**Supplementary Table 7**): ion transport protein (PF00520, $p = 6.9 \times 10^{-4}$, G-test Bonferroni corrected), ligand-gated ion channel (PF00060, $p = 4.0 \times 10^{-6}$), protein kinase domain (PF00069, $p = 0.043$), and kinesin motor domain (PF00225, $p = 0.027$). Missense DNMs in all four enriched domain families have previously been associated with DD (**Supplementary Table 8**)[24–26].

We observed a significant overlap between the 285 DNM-enriched DD-associated genes and a set of 369 previously described cancer driver genes[27] (overlap of 70 genes; $p = 1.7 \times 10^{-49}$, logistic regression correcting for s_{het}), as observed previously[28,29], 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)[30]. 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

207 test, **Supplementary Fig. 11**), whereas recurrent synonymous mutations in TCGA are not
208 significantly enriched (2.4-fold, p = 0.13, Poisson test). This suggests that this observation is
209 driven by the pleiotropic effects of these mutations in development and tumourigenesis, rather
210 than hypermutability.
211
212 **Recurrent mutations and potential new germline selection genes**
213 We identified 773 recurrent DNMs (736 SNVs and 37 indels), ranging from 2-36
214 independent observations per DNM, which allowed us to interrogate systematically the factors
215 driving recurrent germline mutation. We considered three potential contributory factors: (i)
216 clinical ascertainment enriching for pathogenic mutations, (ii) greater mutability at specific sites,
217 and (iii) positive selection conferring a proliferative advantage in the male germline, thus
218 increasing the prevalence of sperm containing the mutation[31]. We observed strong evidence
219 that all three factors contribute, but not necessarily mutually exclusively. Clinical ascertainment
220 drives the observation that 65% of recurrent DNMs were in consensus genes, a 5.4-fold
221 enrichment compared to DNMs only observed once ($p < 10^{-50}$, proportion test). Hypermutability
222 underpins the observation that 64% of recurrent *de novo* SNVs occurred at hypermutable CpG
223 dinucleotides[32], a 2.0-fold enrichment over DNMs only observed once ($p = 3.3 \times 10^{-68}$, chi-
224 square test**)**. We also observed a striking enrichment of recurrent mutations at the
225 haploinsufficient DD-associated gene *MECP2*, in which we observed 11 recurrently mutated
226 SNVs within a 500bp window, nine of which were G to A mutations at a CpG dinucleotide.
227 *MECP2* exhibits a highly significant twofold excess of synonymous mutations within gnomAD[5],
228 suggesting that locus-specific hypermutability might explain this observation.
229 To assess the contribution of germline selection to recurrent DNMs, we initially focused
230 on the 12 known germline selection genes, which all operate through activation of the RAS-
231 MAPK signalling pathway[33,34]. We identified 39 recurrent DNMs in 11 of these genes, 38 of
232 which are missense and all of which are known to be activating in the germline (see
233 Supplement). As expected, given that hypermutability is not the driving factor for recurrent
234 mutation in these germline selection genes, these 39 recurrent DNMs were depleted for CpGs
235 relative to other recurrent mutations (6/39 vs 425/692, $p = 3.4 \times 10^{-8}$, chi-squared test).
236 Positive germline selection has been shown to be capable of increasing the apparent
237 mutation rate more strongly[31] than either clinical ascertainment (10-100X in our dataset) or
238 hypermutability (~10X for CpGs). However, only a minority of the most highly recurrent
239 mutations in our dataset are in genes that have been previously associated with germline
240 selection. Nonetheless, several lines of evidence suggested that the majority of these most
241 highly recurrent mutations are likely to confer a germline selective advantage. Based on the
242 recurrent DNMs in known germline selection genes, DNMs under germline selection should be
243 more likely to be activating missense mutations, and should be less enriched for CpG

244 dinucleotides. **Table 1** shows the 16 *de novo* SNVs observed nine or more times in our DNM
245 dataset, only two of which are in known germline selection genes (*MAP2K1* and *PTPN11*). All
246 but two of these 16 *de novo* SNVs cause missense changes, all but two of these genes cause
247 disease by an altered-function mechanism, and these DNMs were depleted for CpGs relative to
248 all recurrent mutations. Two of the genes with highly recurrent *de novo* SNVs, *SHOC2* and
249 *PPP1CB,* encode interacting proteins that are known to play a role in regulating the RAS-MAPK
250 pathway, and pathogenic variants in these genes are associated with a Noonan-like
251 syndrome[35]. Moreover, two of these recurrent DNMs are in the same gene *SMAD4*, which
252 encodes a key component of the TGF-beta signalling pathway, potentially expanding the
253 pathophysiology of germline selection beyond the RAS-MAPK pathway. Confirming germline
254 selection of these mutations will require deep sequencing of testes and/or sperm[34].
255
256

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

257
258
259 **Table 1: Recurrent Mutations.** *De novo* single nucleotide variants with more than 9
260 recurrences in our cohort annotated with relevant information, such as CpG status, whether the
261 impacted gene is a known somatic driver or germline selection gene, and diagnostic gene group
262 (e.g. consensus). “Recur” refers to the number of recurrences. “Likely mechanism” refers to
263 mechanisms attributed to this gene in the published literature.
264
265

266 **Evidence for incomplete penetrance and pre/perinatal death**

267 Nonsynonymous DNMs in consensus or significant DD-associated genes accounted for
268 half of the exome-wide nonsynonymous DNM burden associated with DD (**Fig. 1b**). Despite our
269 identification of 285 significantly DD-associated genes, there remains a substantial burden of
270 both missense and protein-truncating DNMs in unassociated genes (those that are neither
271 significant in our analysis nor on the consensus gene list). The remaining burden of protein-
272 truncating DNMs is greatest in genes that are intolerant of PTVs in the general population
273 (**Supplementary Fig. 12**) suggesting that more haploinsufficient (HI) disorders await discovery.
274 We observed that PTV mutability (estimated from a null germline mutation model) was
275 significantly lower in unassociated genes compared to DD-associated genes ($p = 4.5 \times 10^{-68}$,
276 Wilcox rank-sum test **Fig. 3a**), which leads to reduced statistical power to detect DNM
277 enrichment in unassociated genes. This is consistent with our hypothesis that many more HI
278 disorders await discovery.

279 A key parameter in estimating statistical power to detect novel HI disorders is the fold-
280 enrichment of *de novo* PTVs expected in as yet undiscovered HI disorders. We observed that
281 novel DD-associated HI genes had significantly lower PTV enrichment compared to the
282 consensus HI genes ($p = 0.005$, Wilcox rank-sum test; **Fig. 3b**). Two additional factors that
283 could lower DNM enrichment, and thus power to detect a novel DD-association, are reduced
284 penetrance and increased pre/perinatal death, which here covers spontaneous fetal loss,
285 termination of pregnancy for fetal anomaly, stillbirth, and early neonatal death. To evaluate
286 incomplete penetrance, we investigated whether HI genes with a lower enrichment of protein-
287 truncating DNMs in our cohort are associated with greater prevalences of PTVs in the general
288 population. We observed a significant negative correlation ($p = 0.031$, weighted linear
289 regression) between gene-specific PTV enrichment in our cohort and the gene-specific ratio of
290 PTV to synonymous variants in gnomAD[5], suggesting that incomplete penetrance does lower *de*
291 *novo* PTV enrichment in individual genes in our cohort (**Fig. 3c**).

292 Additionally, we observed that the fold-enrichment of protein-truncating DNMs in
293 consensus HI DD-associated genes in our cohort was significantly lower for genes with a
294 medium or high likelihood of presenting with a prenatal structural malformation ($p = 4.6 \times 10^{-5}$,
295 Poisson test, **Fig. 3d**), suggesting that pre/perinatal death decreases our power to detect some
296 novel DD-associated disorders (see supplement for details).
297

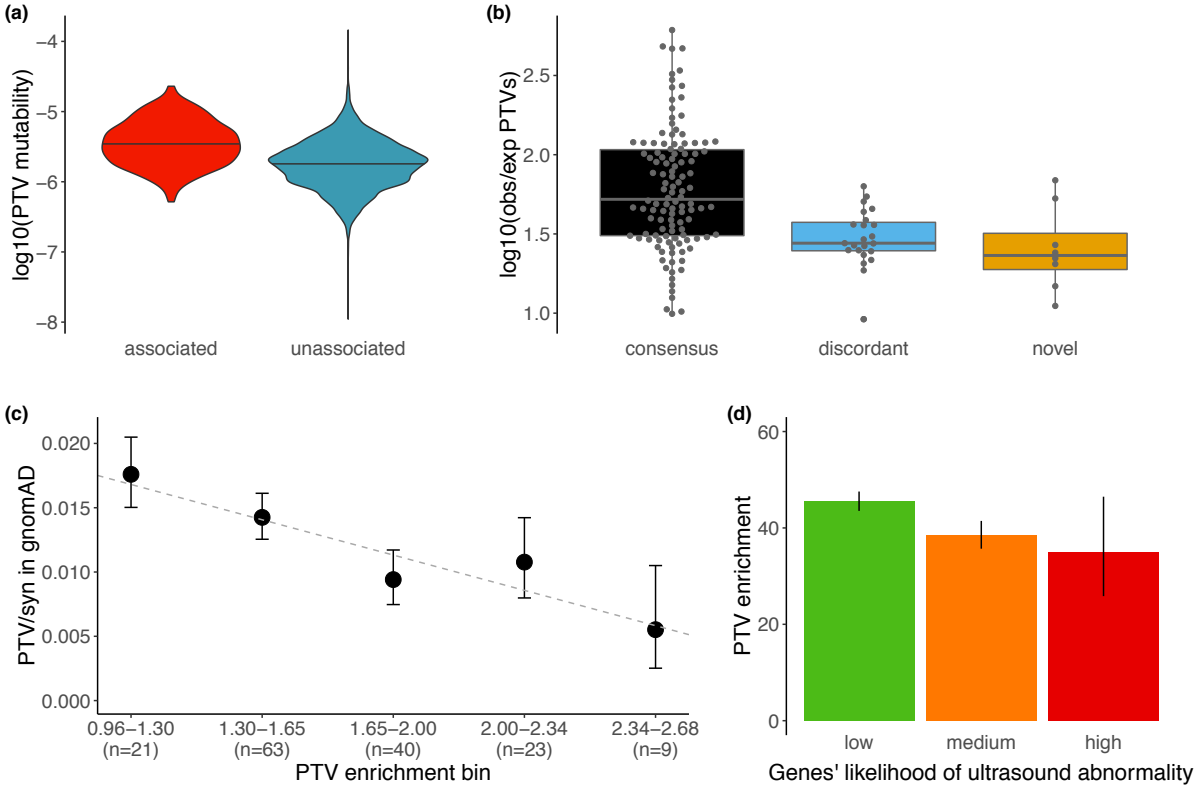

Figure 3: Impact of pre/perinatal death and penetrance on power. (a) PTV mutability is significantly lower in genes that are not significantly associated to DD in our analysis ("unassociated", coloured blue) than in DD-associated genes ("associated", coloured red; p = 4.6 x 10^{-68}, Wilcox rank sum test). (b) Distribution of PTV enrichment in significant, likely haploinsufficient, genes by diagnostic group. (c) Comparison of the PTV enrichment in our cohort vs the PTV to synonymous ratio found in gnomAD, for genes that are significantly enriched for the number of PTV mutations in our cohort (without any variant weighting). 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 = 4.6 x 10^{-5}, 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 ~1000 genes with ~10-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 1000 genes with a substantially lower fold-enrichment than in currently known DD-associated genes (**Supplemental Fig. 13**).

We calculated that a sample size of ~350,000 parent-offspring trios would be needed to have 80% power to detect a 10-fold enrichment of protein-truncating DNMs for a gene with the median PTV mutation rate among currently unassociated genes. Using this inferred 10-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 10-fold increased PTV rate. Among the ~19,000 non-DD-associated genes, ~1,200 were more than three times more likely to have arisen from a 10-fold increased PTV rate, whereas ~7,000 were three times more likely to have no *de novo* PTV enrichment.

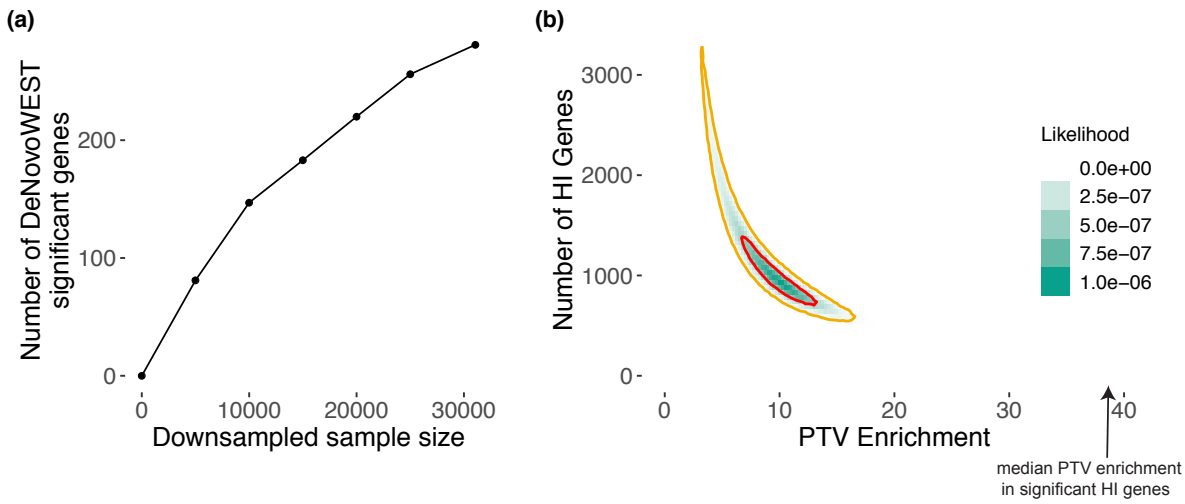

Figure 4: Exploring the remaining number of DD genes. (a) Number of significant genes from downsampling full cohort and running DeNovoWEST's enrichment test. (b) Results from modelling the likelihood of the observed distribution of *de novo* PTV mutations. This model varies the numbers of remaining haploinsufficient (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 39.7.

345

346 **Discussion**

347 In this study, we have discovered 28 novel developmental disorders by developing an
348 improved statistical test for mutation enrichment and applying it to a dataset of exome
349 sequences from 31,058 children with developmental disorders, and their parents. These 28
350 novel genes account for up to 1.0% of our cohort, and inclusion of these genes in diagnostic
351 workflows will catalyse increased diagnosis of similar patients globally. We note that the value of
352 this study for improving diagnostic yield extends well beyond these 28 novel genes; once newly
353 validated discordant genes are included, the total number of genes added to the diagnostic
354 workflows of the three participating centres ranged from 48-65 genes. We have shown that both
355 incomplete penetrance and pre/perinatal death reduce our power to detect novel DDs
356 postnatally, and that one or both of these factors are likely operating considerably more strongly
357 among undiscovered DD-associated genes. In addition, we have identified a set of highly
358 recurrent mutations that are strong candidates for novel germline selection mutations, which
359 would be expected to result in a higher than expected disease incidence that increases
360 dramatically with increased paternal age.

361 Our study represents the largest collection of DNMs for any disease area, and is
362 approximately three times larger than a recent meta-analysis of DNMs from a collection of
363 individuals with autism spectrum disorder, intellectual disability, and/or a developmental
364 disorder[36]. Our analysis included DNMs from 24,348 previously unpublished trios, and we
365 identified ~2.3 times as many significantly DD-associated genes as this previous study when
366 using Bonferroni-corrected exome-wide significance (285 vs 124). In contrast to meta-analyses
367 of published DNMs, the harmonised filtering of candidate DNMs across cohorts in this study
368 should protect against results being confounded by substantial cohort-specific differences in the
369 sensitivity and specificity of detecting DNMs.

370 Here we inferred indirectly that developmental disorders with higher rates of detectable
371 prenatal structural abnormalities had greater pre/perinatal death. The potential size of this effect
372 can be quantified from the recently published PAGE study of genetic diagnoses in a cohort of
373 fetal structural abnormalities[37]. In this latter study, genetic diagnoses were not returned to
374 participants during the pregnancy, and so the genetic diagnostic information itself could not
375 influence pre/perinatal death. In the PAGE study data, 69% of fetal abnormalities with a
376 genetically diagnosable cause died perinatally or neonatally, with termination of pregnancy, fetal
377 demise and neonatal death all contributing. This emphasises the substantial impact that
378 pre/perinatal death can have on reducing the ability to discover novel DDs from postnatal
379 recruitment alone, and motivates the integration of genetic data from prenatal, neonatal and
380 postnatal studies in future analyses.

381 To empower our mutation enrichment testing, we estimated positive predictive values
382 (PPV) of each DNM being pathogenic on the basis of their predicted protein consequence,
383 CADD score[3], selective constraint against heterozygous PTVs across the gene (s_{het})[38], and, for
384 missense variants, presence in a region under selective missense constraint[4]. These PPVs
385 should also be highly informative for variant prioritisation in the diagnosis of dominant
386 developmental disorders. Further work is needed to see whether these PPVs might be
387 informative for recessive developmental disorders, and in other types of dominant disorders.
388 More generally, we hypothesise that empirically-estimated PPVs based on variant enrichment in
389 large datasets will be similarly informative in many other disease areas.
390 We adopted a conservative statistical approach to identifying DD-associated genes. In
391 two previous studies using the same significance threshold, we identified 26 novel DD-
392 associated genes[1,39]. All 26 are now regarded as being diagnostic, and have entered routine
393 clinical diagnostic practice. Had we used a significance threshold of FDR < 10% as used in
394 Satterstrom, Kosmicki, Wang et al[40], we would have identified 770 DD-associated genes.
395 However, as the FDR of individual genes depends on the significance of other genes being
396 tested, FDR thresholds are not appropriate for assessing the significance of individual genes,
397 but rather for defining gene-sets. There are 184 consensus genes that did not cross our
398 significance threshold in this study. It is likely that many of these cause disorders that were
399 under-represented in our study due to the ease of clinical diagnosis on the basis of distinctive
400 clinical features or targeted diagnostic testing. These ascertainment biases are, however, not
401 likely to impact the representation of novel DDs in our cohort.
402 Our modelling also suggested that likely over 1,000 DD-associated genes remain to be
403 discovered, and that reduced penetrance and pre/perinatal death will reduce our power to
404 identify these genes through DNM enrichment. Identifying these genes will require both
405 improved analytical methods and greater sample sizes. As sample sizes increase, accurate
406 modelling of gene-specific mutation rates becomes more important. In our analyses of 31,058
407 trios, we observed evidence that mutation rate heterogeneity among genes can lead to over-
408 estimating the statistical significance of mutation enrichment based on an exome-wide mutation
409 model. We advocate the development of more granular mutation rate models, based on large-
410 scale population variation resources, to ensure that larger studies are robust to mutation rate
411 heterogeneity.
412 We anticipate that the variant-level weights used by DeNovoWEST will improve over
413 time. As reference population samples, such as gnomAD[5], increase in size, weights based on
414 selective constraint metrics (e.g. s_{het}, regional missense constraint) will improve. Weights could
415 also incorporate more functional information, such as expression in disease-relevant tissues.
416 For example, we observe that DD-associated genes are significantly more likely to be
417 expressed in fetal brain (**Supplementary Fig. 14**). Furthermore, novel metrics based on gene

418 co-regulation networks can predict whether genes function within a disease-relevant pathway[41].
419 As a cautionary note, including more functional information may increase power to detect some
420 novel disorders while decreasing power for disorders with pathophysiology different from known
421 disorders. Our analyses also suggest that variant-level weights could be further improved by
422 incorporating other variant prioritisation metrics, such as upweighting variants predicted to
423 impact splicing, variants in particular protein domains, or variants that are somatic driver
424 mutations during tumorigenesis. In developing DeNovoWEST, we initially explored applying
425 both variant-level weights and gene-level weights in separate stages of the analysis, however,
426 subtle but pervasive correlations between gene-level metrics (e.g. s_{het}) and variant-level metrics
427 (e.g. regional missense constraint, CADD) presents statistical challenges to implementation.
428 Finally, the discovery of less penetrant disorders can be empowered by analytical
429 methodologies that integrate both DNMs and rare inherited variants, such as TADA[42].
430 Nonetheless, using current methods focused on DNMs alone, we estimated that ~350,000
431 parent-child trios would need to be analysed to have ~80% power to detect HI genes with a 10-
432 fold PTV enrichment. Discovering non-HI disorders will need even larger sample sizes.
433 Reaching this number of sequenced families will be impossible for an individual research study
434 or clinical centre, therefore it is essential that genetic data generated as part of routine
435 diagnostic practice is shared with the research community such that it can be aggregated to
436 drive discovery of novel disorders and improve diagnostic practice.
437

438 **Acknowledgements**

439 We thank the families and their clinicians for their participation and engagement. We are very
440 grateful to our colleagues who assisted in the generation and processing of data. Inclusion of
441 RadboudUMC data was in part supported by the Solve-RD project that has received funding
442 from the European Union's Horizon 2020 research and innovation programme under grant
443 agreement No 779257. This work was in part financially supported by grants from the
444 Netherlands Organization for Scientific Research: 917-17-353 to CG. The DDD study presents
445 independent research commissioned by the Health Innovation Challenge Fund [grant number
446 HICF-1009-003]. This study makes use of DECIPHER which is funded by Wellcome. See
447 www.ddduk.org/access.html for full acknowledgement. The DDD study would like to
448 acknowledge the tireless work of Rosemary Kelsell. Finally we acknowledge the contribution of
449 an esteemed DDD clinical collaborator, M. Bitner-Glindicz, who died during the course of the
450 study.
451

452 **Data Access**

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

455 RadboudUMC sequence and variant level data cannot be made available through EGA due to
456 the nature of consent for clinical testing
457 GeneDx data cannot be made available through EGA due to the nature of consent for clinical
458 testing. GeneDx has contributed deidentified data to this study to improve clinical interpretation
459 of genomic data, in accordance with patient consent and in conformance with the ACMG
460 position statement on genomic data sharing (see Supplementary Note for details).
461 Clinically interpreted variants and associated phenotypes from the DDD study are available
462 through DECIPHER (https://decipher.sanger.ac.uk)
463 Clinically interpreted variants from RUMC are available from the Dutch national initiative for
464 sharing variant classifications (https://www.vkgl.nl/nl/diagnostiek/vkgl-datashare-database)
465 Clinically interpreted variants from GeneDx are deposited in ClinVar
466 (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. Robinson, J. T. *et al.* Integrative genomics viewer. *Nature Biotechnology* vol. 29 24–26 (2011).
11. Villegas, F. *et al.* Lysosomal Signaling Licenses Embryonic Stem Cell Differentiation via Inactivation of Tfe3. *Cell Stem Cell* **24**, 257–270.e8 (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. 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. Lasser, M., Tiber, J. & Lowery, L. A. The Role of the Microtubule Cytoskeleton in Neurodevelopmental Disorders. *Front. Cell. Neurosci.* **12**, 165 (2018).
26. Hamilton, M. J. *et al.* Heterozygous mutations affecting the protein kinase domain of cause a syndromic form of developmental delay and intellectual disability. *J. Med. Genet.* **55**, 28–38 (2018).
27. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **173**, 1823 (2018).
28. 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).
29. Ronan, J. L., Wu, W. & Crabtree, G. R. From neural development to cognition: unexpected roles for chromatin. *Nat. Rev. Genet.* **14**, 347–359 (2013).
30. Cancer Genome Atlas Research Network *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
31. 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).
32. Duncan, B. K. & Miller, J. H. Mutagenic deamination of cytosine residues in DNA. *Nature* vol. 287 560–561 (1980).
33. 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).
34. Maher, G. J. *et al.* Selfish mutations dysregulating RAS-MAPK signaling are pervasive in aged human testes. *Genome Res.* **28**, 1779–1790 (2018).
35. 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).

36. Coe, B. P. *et al.* Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. *Nat. Genet.* **51**, 106–116 (2019).
37. Lord, J. *et al.* Prenatal exome sequencing analysis in fetal structural anomalies detected by ultrasonography (PAGE): a cohort study. *Lancet* **393**, 747–757 (2019).
38. Cassa, C. A. *et al.* Estimating the selective effects of heterozygous protein-truncating variants from human exome data. *Nat. Genet.* **49**, 806–810 (2017).
39. Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**, 223–228 (2015).
40. 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.e23 (2020).
41. 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).
42. 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).