Prevalence and architecture of de novo mutations in developmental disorders.

Nature 542, pp. 433-438. 10.1038/nature21062 file

Publishers page: http://dx.doi.org/10.1038/nature21062

Please note:
Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher’s version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See http://orca.cf.ac.uk/policies.html for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.
Prevalence and architecture of de novo mutations in developmental disorders

Deciphering Developmental Disorders Study

The genomes of individuals with severe, undiagnosed developmental disorders are enriched in damaging de novo mutations (DNMs) in developmentally important genes. Here we have sequenced the exomes of 4,293 families containing individuals with developmental disorders, and meta-analysed these data with data from another 3,287 individuals with similar disorders. We show that the most important factors influencing the diagnostic yield of DNMs are the sex of the affected individual, the relatedness of their parents, whether close relatives are affected and the parental ages. We identified 94 genes enriched in damaging DNMs, including 14 that previously lacked compelling evidence of involvement in developmental disorders. We have also characterized the phenotypic diversity among these disorders. We estimate that 42% of our cohort carry pathogenic DNMs in coding sequences; approximately half of these DNMs disrupt gene function and the remainder result in altered protein function. We estimate that developmental disorders caused by DNMs have an average prevalence of 1 in 213 to 1 in 448 births, depending on parental age. Given current global demographics, this equates to almost 400,000 children born per year.
Odds ratios are associated with increased risk of pathogenic DNMs.

Previously\textsuperscript{15}, we combined this analysis with 4,224 published DNMs in 3,287 affected individuals from thirteen exome- or genome-sequencing studies\textsuperscript{3–14} (Supplementary Table 2) that exhibited a similar excess of DNMs as our curated set of DD-associated genes (Extended Data Fig. 1). We found 93 genes with genome-wide significance ($P < 5 \times 10^{-7}$; Fig. 2), 80 of which had previous evidence of an association with DDs (Supplementary Table 3). We have developed visual summaries of the phenotypes associated with each gene to facilitate clinical use. In addition, we created anonymised, average face images from individuals with DNMs in genome-wide significant genes (Fig. 2) from ordinary (2D) clinical photos using previously validated software\textsuperscript{24}. These images highlight facial dysmorphologies specific to certain genes. After careful review by two experienced clinical geneticists, average face images for twelve genes were determined to be truly anonymised and of sufficient quality. To assess any increase in power to detect novel DD-associated genes, we excluded individuals with probably pathogenic variants in known DD-associated genes\textsuperscript{15}, leaving 3,158 probands from our cohort, along with 2,955 probands from the meta-analysis studies. In this subset, fourteen genes for which no statistically compelling previous evidence for DD causation was available achieved genome-wide significance: CDKN1B, CHD4, CNOT3, CSNK2A1, GNA11, KCNQ3, MSL3, PPM1D, PUF60, QRIC1H, SET, KMT5B (also known as SUV4202), TCF20 and ZBTB18 ($P < 5 \times 10^{-7}$; Table 1 and Extended Data Fig. 4). The clinical features associated with these newly confirmed disorders are summarized in Extended Data Figs 2, 3 and Supplementary Information. QRIC1H did not achieve genome-wide significance without including individuals with probably pathogenic variants in DD-associated genes. In addition to discovering novel DD-associated genes, we identified several new disorders linked to known DD-associated genes, but with different modes of inheritance or molecular mechanisms. We found that USP9X and ZCCH2 had a genome-wide significant excess of DNMs in female probands, indicating these genes have X-linked dominant modes of inheritance in addition to a previously reported X-linked recessive mode of inheritance in males\textsuperscript{25,26}. In addition, we found that truncating mutations in SMC1A were strongly associated with a novel seizure disorder ($P = 6.5 \times 10^{-19}$), whereas in-frame and/or missense mutations in SMC1A with dominant negative effects\textsuperscript{27} are a known cause of Cornelia de Lange Syndrome. Individuals with truncating mutations in SMC1A lacked the characteristic facial dysmorphism of Cornelia de Lange Syndrome.

We then explored two approaches for integrating phenotypic data into disease gene association: statistical assessment of HPO-term similarity between individuals sharing candidate DNMs in the same gene (as previously described\textsuperscript{28}) and phenotypic stratification based on specific clinical characteristics. Combining genetic evidence and HPO-term similarity increased the significance of some known DD-associated genes. However, significance decreased for a larger number of genes that caused severe DDs, but that are associated with non-discriminative HPO terms (Extended Data Fig. 5a). Although we did not incorporate categorical phenotypic similarity into the gene-discovery analyses described above, the systematic acquisition of phenotypic data from affected individuals within the DDD cohort enabled aggregate representations to be created for each gene that achieved genome-wide significance. We present these in the form of icon-based summaries of growth and developmental milestones (PhenIcons), heatmaps of the recurrently coded HPO terms and, where photos for at least ten children with mutations in the same gene were available, an anonymised average facial representation (Supplementary Information).

Twenty per cent of individuals had HPO terms that indicated seizures and/or epilepsy. We compared analysis within this phenotypically stratified group with gene-wise analyses of the entire cohort, to see whether it increased power to detect known seizure-associated genes (Extended Data Fig. 5b). Fifteen seizure-associated genes had genome-wide significance ($P < 5 \times 10^{-7}$) in both the seizure-only and the entire-cohort analyses. Nine seizure-associated genes had genome-wide significance ($P < 5 \times 10^{-7}$) in the entire cohort but not in the seizure subset. Of the 285 individuals with truncating or missense

---

**Table 1**

| Gene     | Odds ratio | 95% CI       |
|----------|------------|--------------|
| CDKN1B   | 1.6 × 10^4 | 0.392        |
| CHD4     | 0.190      | 0.00155      |
| CSNK2A1  | 0.346      | 0.278        |
| GNA11    | 0.0794     | 0.584        |
| KCNQ3    | 7.3 × 10^14| 5.7 × 10^14  |
| MSL3     | 8.0 × 10^11| 3.9 × 10^11  |
| PPM1D    | 3.9 × 10^11| 5.3 × 10^11  |
| PUF60    | 0.00115    | 0.00247      |
| QRIC1H   | 0.408      | 0.135        |
| SET      | 0.305      | 0.0564       |
| KMT5B    | 0.305      | 0.0564       |

**Figure 1** | Association of phenotypes with the presence of DNMs that are probably pathogenic. a, Odds ratios for binary phenotypes. Positive odds ratios are associated with increased risk of pathogenic DNMs when the phenotype is present. $P$ values are given (Fisher’s exact test). b, $\beta$ coefficients from logistic regression of quantitative phenotypes versus presence of a pathogenic DNM. All phenotypes aside from length of autosome regions were corrected for gender as a covariate. The developmental milestones (age to achieve first words, walk independently, sit independently and social smile) were log scaled before regression. The growth parameters (height, birthweight and occipitofrontal circumference (OFC)) were evaluated as absolute distance from the median. c. The relationship between length of autosome regions and chance of having a pathogenic DNM. The regression line and 95% CI are plotted as the dark grey line and grey shading, respectively. The autozygosity lengths expected under different degrees of consanguineous unions are shown as vertical dashed lines. n, number of individuals in each autozygosity group. Blue dots and blue lines, mean ± 95% CI. d, Relationship between age of fathers at child birth and the number of high confidence DNMs. Error bars, 95% CI. e, Relationship between age of mothers at child birth and number of high confidence DNMs. Error bars, 95% CI.
DNMs in known seizure-associated genes, 56% of individuals had no coded terms related to seizures and/or epilepsy. These findings suggest that the power of increased sample size far outweighs specific phenotypic expressivity owing to the shared genetic aetiology between individuals with and without epilepsy in our cohort. Despite this, nearly three times as many individuals with seizures had a DNAM in a seizure-associated gene compared to individuals without seizures (Extended Data Fig. 5c). With matched sample sizes, more genes exceeded genome-wide significance in seizure samples than in unstratified samples (Extended Data Fig. 5d). This highlights the cost–benefit effect of recruiting a phenotypically more homogenous cohort.

The large number of genes with genome-wide significance identified in the analyses above allows us to compare empirically different experimental strategies for novel gene discovery in a genetically heterogeneous cohort. We compared the power of exome and genome sequencing to detect genes with genome-wide significance, assuming that budget and not samples are limiting, under different scenarios of cost ratios and sensitivity ratios (Extended Data Fig. 6a). At current cost ratios (exome analysis costs 30–40% of genome analysis) and with a plausible sensitivity differential (genome analysis detects 5% more exonic variants than exome analysis\(^\text{26}\)), exome sequencing detects more than twice as many genome-wide significant genes. These empirical estimates were consistent with power simulations for identifying dominant loss-of-function genes (Extended Data Fig. 6b). In summary, although genome sequencing provides the greatest sensitivity to detect pathogenic variation in a single individual (or outside of the coding region), exome sequencing is more powerful for novel gene discovery for disease (and, analogously, probably currently delivers a lower cost per diagnosis).

Our previous simulations suggested that analysis of a cohort of 4,293 DDD families should be able to detect approximately half of all haploinsufficient DD-associated genes at genome-wide significance\(^\text{15}\). Empirically, we have identified 47% (50 out of 107) of haploinsufficient genes that have been previously robustly associated with neurodevelopmental disorders\(^\text{18}\). We hypothesized that genetic testing before recruitment into our study may have depleted the cohort of the most clinically recognizable disorders. Indeed, we observed that the genes associated with the most clinically recognizable disorders were associated with a significant, threefold lower enrichment of truncating DNMs than other DD-associated genes (\(P = 8.9 \times 10^{-20}\), approximately 40-fold enrichment for the most clinically recognizable disorders compared to around 120-fold enrichment for cryptic disorders; Fig. 3a). Removing these most recognizable disorders from the analysis, we identified 55% (42 out of 76) of the remaining haploinsufficient DD-associated genes. The known DD-associated haploinsufficient genes that did not reach genome-wide significance were clearly enriched in those with lower mutability, which we would expect to lower the power to detect for these analyses. We identified DD-associated genes (for example, NRXN2) with high mutability, low clinical recognizability and yet no signal of enrichment for DNMs in our cohort, as assessed by \(\Delta_{\text{AIC}}\) (the difference between the Akaike’s Information Criterion of model 1 and model 2) (Extended Data Fig. 7 and Supplementary Table 4). The current analyses call into question whether these genes really are associated with haploinsufficient neurodevelopmental disorders and highlight the potential for well-powered, gene-discovery analyses to refute previous credence in disease gene associations or previous inferences of an underlying haploinsufficient mechanism.

We estimated the prevalence of pathogenic missense and truncating DNMs within our cohort by increasing the stringency of called DNMs until the observed synonymous DNMs equated that expected under the null-mutation model (Extended Data Fig. 8a), and then quantifying the excess of observed missense and truncating DNMs across all genes (Fig. 3b). We observed an excess of 576 truncating and 1,220 missense mutations, suggesting that 41.8% (1,796 out of 4,293) of the cohort have a pathogenic DNAM. This estimate of the number of excess missense and truncating DNMs in our cohort is robust to varying the stringency of DNM calling (Extended Data Fig. 8b). The vast majority of synonymous DNMs are probably benign, as shown by the uniform distribution (Fig. 3d) among genes, irrespective of their tolerance for truncating variation in the general population (as quantified by the ‘probability of being loss-of-function intolerant’ \((P_{\text{LI}})\) metric\(^\text{25}\)). By contrast, missense and truncating DNMs are significantly enriched in genes with the highest probabilities of being intolerant to truncating variation (missense, \(P = 1.1 \times 10^{-47}\); truncating, \(P = 3.3 \times 10^{-45}\).

### Table 1 | Genes achieving genome-wide significant statistical evidence without previous compelling evidence for association with DDs

| Gene          | Missense | PTV | \(P\) value | Test   | Clustering |
|---------------|----------|-----|-------------|--------|------------|
| CDK13        | 10       | 1   | 3.2 \times 10^{-19} | DDD    | Yes        |
| GNAI1        | 7 (1)    | 1   | 2.1 \times 10^{-13} | DDD    | No         |
| CSNK2A1      | 7        | 0   | 1.4 \times 10^{-12} | DDD    | Yes        |
| PPM1D        | 0        | 5 (1) | 6.3 \times 10^{-12} | Meta   | No         |
| CNNOT3       | 5        | 2 (1) | 5.2 \times 10^{-11} | DDD    | Yes        |
| MSL3         | 0        | 4   | 2.2 \times 10^{-10} | DDD    | No         |
| KNCQ3        | 4 (3)    | 0   | 3.4 \times 10^{-10} | Meta   | Yes        |
| ZBTB18       | 1 (1)    | 4   | 1.4 \times 10^{-9}  | DDD    | No         |
| PUF60        | 4 (1)    | 3   | 2.6 \times 10^{-9}  | DDD    | No         |
| TCF20        | 1        | 5   | 2.7 \times 10^{-9}  | DDD    | No         |
| KMT5B        | 0 (2)    | 2 (3) | 2.9 \times 10^{-9}  | Meta   | No         |
| CHD4         | 8 (1)    | 1   | 7.6 \times 10^{-9}  | DDD    | No         |
| SET          | 0        | 3   | 1.2 \times 10^{-7}  | DDD    | No         |
| QRich1       | 0        | 3 (1) | 3.6 \times 10^{-7}  | Meta   | No         |

The numbers of unrelated individuals with independent DNMs are given for protein-truncating variants (PTV) and missense variants. Counts of individuals in other cohorts are given in brackets. The \(P\) value reported is the minimum \(P\) value from the testing of the DDD dataset or the meta-analysis dataset. The subset providing the \(P\) value is also listed. Mutations are considered clustered if the \(P\) value from proximity clustering of DNMs is less than 0.01.
In summary, we have shown that DNMs account for approximately half of the genetic architecture of severe DDs, and are split roughly equally between loss of function and altered function. Whereas most haploinsufficient DD-associated genes have already been identified, many DD-associated genes characterized by pathogenic missense DNMs remain to be discovered.

We estimated the birth prevalence of monoallelic DDs by using the germline-mutation model to calculate the expected cumulative germline-mutation rate of truncating DNMs in haploinsufficient DD-associated genes and scaling this upwards on the basis of the composition of excess DNMs in the DDD cohort described above (see Methods), correcting for disorders that were under represented in our cohort as a result of previous genetic testing (for example, clinically recognizable disorders and large pathogenic copy-number variations identified by previous chromosomal microarray analysis). This gives a mean prevalence estimate of 0.34% (95% CI = 0.31–0.37), or 1 in 295 births. By factoring in the paternal and maternal age effects on the mutation rate (Fig. 1), we modelled age-specific estimates of birth prevalence (Fig. 4) that range from 1 in 448 (both mother and father aged 20) to 1 in 213 (both mother and father aged 45). Assuming a yearly global birth rate of 18.6 live births per 1,000 individuals, and a mean age when giving birth of 26.6 years, nearly 400,000 of the 140 million annual births will have a DD caused by a DNM.

Figure 3 | Excess of DNMs. a, Enrichment ratios of observed to expected loss-of-function DNMs by clinical recognizability for dominant haploinsufficient neurodevelopmental genes as judged by two consultant clinical geneticists. Error bars, 95% CI. b, Enrichment of DNMs by consequence normalized relative to the number of synonymous DNMs. c, Proportion of excess DNMs with loss-of-function (LoF) or altered-function (PTV) mechanisms. Proportions are derived from numbers of excess DNMs by consequence, and numbers of excess protein-truncating (PTV) and missense DNMs in dominant haploinsufficient (HI) genes. d, Enrichment ratios of observed to expected DNMs by P11-constraint quantile for protein-truncating, missense and synonymous DNMs. Counts of DNMs in each lower and upper half of the quantiles are provided. Fig. 3d). The P11-based distributions were similar to distributions that used functional constraint31 (Extended Data Fig. 9). Only 51% (923 out of 1,796) of these excess missense and truncating DNMs are located in DD-associated dominant genes, with the remainder probably affecting genes not yet associated with DDs. A much higher proportion of the excess truncating DNMs (71%) than missense DNMs (42%) affected known DD-associated genes. This suggests that whereas most haploinsufficient DD-associated genes have already been identified, many DD-associated genes characterized by pathogenic missense DNMs remain to be discovered.

Understanding the mechanism of action of a monogenic disorder is an important prerequisite for designing therapeutic strategies32. We tried to estimate the relative proportion of altered-function and loss-of-function mechanisms among the excess DNMs in our cohort, by assuming that the vast majority of truncating mutations operate by a loss-of-function mechanism and by using two independent approaches to estimate the relative contribution of the two mechanisms among the excess missense DNMs (see Methods). First, we used the observed ratio of truncating and missense DNMs within haploinsufficient DD-associated genes to estimate the proportion of the excess missense DNMs that probably act by loss of function (Fig. 3c). This approach estimated that 59% (95% CI = 55–64%) of the excess missense and truncating DNMs operate by loss of function, and 41% by altered function. Second, we took advantage of the different population genetic characteristics of known altered-function and loss-of-function DD-associated genes. Specifically, we observed that these two classes of DD-associated genes are differentially depleted of truncating variation in individuals without overt DDs (P11 metric30). We modelled the observed P11 distribution of excess missense DNMs as a mixture of the P11 distributions of known altered-function and loss-of-function DD-associated genes (Fig. 3e, f), and estimated that 63% (95% CI = 50–76%) of excess missense DNMs probably act by altered-function mechanisms. Incorporating the truncating DNMs that cause a loss-of-function mechanism, this approach estimated that 57% (95% CI = 48–66%) of excess missense and truncating DNMs operate by loss of function and 43% by altered function.
yet been described. This probably results from these disorders being individually rarer, being caused by a relatively small number of mis-
sense mutations within each gene. It would be valuable to estimate the
penetrance of DNMs in the genes we identified as exceeding genome-
wide significance, but we cannot formally assess penetrance with our
data. Future evaluations could integrate depletion of damaging variation
in large healthy populations with patterns of segregation in affected
families. Discovery of the remaining dominant DDs requires larger
studies and novel, more powerful, analytical strategies for disease-gene
association that leverage gene-specific patterns of population variation,
specifically the observed depletion of damaging variation. The integra-
tion of accurate and complete, quantitative and categorical phenotypic
data into the analysis will improve the power to identify ultrarare DDs
with distinctive clinical presentations. We have estimated the mean birth
prevalence of dominant monogenic DDs to be around 1 in 295,
which is greater than the combined impact of trisomies 13, 18 and
21 (ref. 33) and highlights the cumulative population morbidity and
mortality imposed by these individually rare disorders.

Note added in proof. Other recently published studies have also identi-
ﬁed DD-associations for several genes described here, namely CDK13
(ref. 34), CHD4 (ref. 34), CNKN2A1 (ref. 35), MSL3 (ref. 36), PPM1D
(ref. 37), TCF20 (ref. 37) and ZBTB18 (ref. 38).

Online Content Methods, along with any additional Extended Data display items and
Source Data, are available in the online version of the paper; references unique to
these sections appear only in the online paper.

Received 14 April; accepted 15 December 2016.
Published online 25 January 2017.

1. Sheridan, E. et al. Risk factors for congenital anomaly in a multiethnic birth
cohort: an analysis of the Born in Bradford study. Lancet 382, 1350–1359
(2013).
2. Ropers, H. H. Genetics of early onset cognitive impairment. Annu. Rev.
Genomics Hum. Genet. 11, 161–187 (2010).
3. de Ligt, J. et al. Diagnostic exome sequencing in persons with severe
intellectual disability. N. Engl. J. Med. 367, 1921–1929 (2012).
4. De Rubeis, S. et al. Synthetic, transcriptional and chromatin genes disrupted in
autism. Nature 515, 209–215 (2014).
5. Epi4K Consortium & Epilepsy Phenome/Genome Project. De novo mutations in
epileptic encephalopathies. Nature 501, 217–221 (2013).
6. EuroEPINOMICS-RES Consortium, Epilepsy Phenome/Genome Project &
Epi4K Consortium. De novo mutations in synaptic transmission pathways
including DNM1 cause epileptic encephalopathies. Am. J. Hum. Genet. 95,
360–370 (2014).
7. Froemer, M. et al. De novo mutations in schizophrenia implicate synaptic
networks. Nature 506, 179–184 (2014).
8. Gilissen, C. et al. Genome sequencing identiﬁes major causes of severe
intellectual disability. Nature 511, 344–347 (2014).
9. lossiof, I. et al. The contribution of de novo coding mutations to autism
spectrum disorder. Nature 515, 216–221 (2014).
10. lossiof, I. et al. De novo gene disruptions in children on the autistic spectrum.
Neuron 74, 285–299 (2012).
11. O’Roak, B. J. et al. Sporadic autism exomes reveal a highly interconnected
protein network of de novo mutations. Nature 485, 246–250 (2012).
12. Rauch, A. et al. Range of genetic mutations associated with severe non-
syndromic sporadic intellectual disability: an exome sequencing study. Lancet
380, 1674–1682 (2012).
13. Sanders, S. J. et al. De novo mutations revealed by whole-exome
sequencing are strongly associated with autism. Nature 485, 237–241
(2012).
14. Zaidi, S. et al. De novo mutations in histone-modifying genes in congenital
heart disease. Nature 498, 220–223 (2013).
15. Deciphering Developmental Disorders Study. Large-scale discovery of
novel genetic causes of developmental disorders. Nature 519, 223–228
(2015).
16. de Ligt, J., Veltman, J. A. & Vissers, L. E. M. Point mutations as a source of
de novo genetic disease. Curr. Opin. Genet. Dev. 23, 257–263 (2013).
17. Wilkie, A. O. The molecular basis of genetic dominance. J. Med. Genet. 31,
89–98 (1994).
18. Wright, C. F. et al. Genetic diagnosis of developmental disorders in the DDD
study: a scalable analysis of genome-wide research data. Lancet 385,
1305–1314 (2014).
19. Jacquemont, S. et al. A higher mutational burden in females supports a
“female protective model” in neurodevelopmental disorders. Am. J. Hum.
Genet. 94, 415–425 (2014).
20. Kong, A. et al. Rate of de novo mutations and the importance of father’s age to
disease risk. Nature 488, 471–475 (2012).
21. Rahbari, R. et al. Timing, rates and spectra of human germline mutation. Nat.
Genet. 48, 126–133 (2016).
22. Wong, W. S. et al. New observations on maternal age effect on germline de novo
mutations. Nat. Commun. 3, 1035 (2012).
23. Samocha, K. E. et al. A framework for the interpretation of de novo mutation in
human disease. Nat. Genet. 46, 944–950 (2014).
24. Ferry, Q. et al. Diagnostically relevant facial gestalt information from ordinary
photos. eLife 3, e02020 (2014).
25. Hirata, H. et al. ZC4H2 mutations are associated with arthrogryposis
multiplex congenita and intellectual disability through impairment of central
and peripheral synaptic plasticity. Am. J. Hum. Genet. 92, 681–695
(2013).
26. Homan, C. et al. Mutations in USP9X are associated with X-linked intellectual
disability and disrupt neuronal cell migration and growth. Am. J. Hum. Genet.
94, 470–478 (2014).
27. Liu, J. et al. SMC1A expression and mechanism of pathogenicity in probands
with X-linked Cornelia de Lange syndrome. Hum. Mutat. 30, 1535–1542
(2009).
28. Akawi, N. et al. Discovery of four recessive developmental disorders using
probabilistic genotype and phenotype matching among 4,125 families.
PLoS Genet. 9, e1003709 (2013).
29. Meynert, A. M., Ansari, M., FitzPatrick, D. R. & Taylor, M. S. Variant detection
sensitivity and biases in whole genome and exome sequencing. BMC
Bioinformatics 15, 247 (2014).
30. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans.
Nature 536, 285–291 (2016).
31. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic
tolerance to functional variation and the interpretation of personal genomes.
PLoS Genet. 9, e1003709 (2013).
32. Boycott, K. M., Vanstone, M. R., Bulman, D. E. & MacKenzie, A. E. Rare-disease
genetics in the era of next-generation sequencing: discovery to translation.
Nat. Rev. Genet. 14, 681–691 (2013).
33. Springett, A. et al. Congenital Anomaly Statistics 2011: England and Wales.
(2013).
34. Sifrim, A. et al. Distinct genetic mechanisms for syndromic and nonsyndromic
genital heart defects identiﬁed by exome sequencing. Nat. Genet. 48,
1060–1065 (2016).
35. Okur, V. et al. De novo mutations in CSNK2A1 are associated with
neurodevelopmental abnormalities and dysmorphic features. Hum. Genet.
135, 699–705 (2016).
36. El Chehadeh, S. et al. Dominant variants in the splicing factor PUF60 cause a
recognizable syndrome with intellectual disability, heart defects and short
stature. Eur. J. Hum. Genet. 25, 43–51 (2016).
37. Leieveld, S. H. et al. Meta-analysis of 2,104 trios provides support for 10 new
genomes for intellectual disability. Nat. Neurosci. 19, 1194–1196
(2016).
38. Cohen, J. S. et al. Further evidence that de novo missense and truncating
variants in ZBTB18 cause intellectual disability with variable features.
Clim. Genet. http://dx.doi.org/10.1007/s00037-016-1061-1 (2016).
Supplementary Information is available in the online version of the paper.

Acknowledgements We thank the families for their participation and patience. We are grateful to the Exome Aggregation Consortium for making their data available. The DDD study presents independent research commissioned by the Health Innovation Challenge Fund (grant HICF-1009-003), a parallel funding partnership between the Wellcome Trust and the UK Department of Health, and the Wellcome Trust Sanger Institute (grant WT098051). The views expressed in this publication are those of the author(s) and not necessarily those of the Wellcome Trust or the UK Department of Health. The study has UK Research Ethics Committee approval (10/H0305/83, granted by the Cambridge South Research Ethics Committee and GEN/284/12, granted by the Republic of Ireland Research Ethics Committee). The research team acknowledges the support of the National Institutes for Health Research, through the Comprehensive Clinical Research Network. We thank the Sanger Human Genome Informatics team, the Sample Management team, the Illumina High-Throughput team, the New Pipeline Group team, the DNA pipelines team and the Core Sequencing team for their support in generating and processing the data. D.R.F. is funded through an MRC Human Genetics Unit program grant to the University of Edinburgh. Finally we acknowledge the contribution of two esteemed DDD clinical collaborators, J. Tolmie and L. Brueton, who died during the course of the study.

Author Contributions Patient recruitment and phenotyping: M.Ah., U.A., H.A., R.A., M.Ba., S.Ba., D.Bar., A.B., P.B., D.Bat., C.Be., J.Be., B.B., M.B.-G., E.B., M.Bl., D.Boh., S.Bou., L.Br., A.Br., C.Br., K.D., L.D., D.Da., J.D., C.De., G.D., A.D., A.Dob., A.Don., D.Donna., D.Donnie., C.Do., D.Dou., S.Dou., A.Due., J.E., S.EI., L.E., F.E., K.E., S.Ev., T.F., R.F., F.F., N.F., A.Frye., A.Frye., C.G., L.Ga., N.G., R.G., H.G., J.G., D.G., A.G., P.G., A.Gr., R.Har., L.Lha., V.H., R.Haw., S.Hel., A.H., S.Hew., E.H., S.Holden, M.Ho., S.Holder, G.H., T.H., M.Hu., J.H., S.I., M.I., L.I., A.J., J.J., L.J., D.Joh., E.J., D.Jos., S.J., B.Ka., S.K., B.Ka., H.K., U.K., E.Kin., G.K., C.K., E.Kiv., A.K., D.Ku., V.K., K., K., L.L., A.L., C.La., M.L., D.L., L., C.L., G.L., S.L., A.Mag., E.Ma., A.Mai., S.Ma., K.Mark., K.Mart., U.M., E.Mc., V.Mc., M.M., R.M., K.Mc., S.McK., D.M., S.McN., C.M., S.M., K.Me., Z.M., A.Mi., E.Mi., S.Moh., T.M., D.M., S.Mor., J.M., H.Mug., V.Mu., H.Mur., S.N., A.Ne., L.N., R.N.-E., A.No., R.O., C.O., K.-O., S.-O., M.J.P., C.Pa., J.Ph., D.TP., C.Po., J.Po., N.P., K.P., S.Pr., A.Pr., A.Pro., H.P., Q., L.P., J.R., J.Rank., L.Pa., D.R., L.R., E.Rob., J.Ro., P.R., G.R., A.R., E.Ros., A.Sag., S.Sa., J.S., R.Sa., A.Sar., S.Sc., P.Sc., I.Sc., A.Sc., A.ScI., A.ScII., A.ScIII., C.S., A.Sh., C.S.-S., E.Sheea., D.S., E.Sheer, I.Si., R.Si., Z.S., A.Sm., K.S., S.Sm., L.S., M.Sp., M.Sq., F.S., H.S., VSt., M.Su., YSu., E.Sw., K.T.-B., C.Ta., R.T., M.Tein, I.K.T., I.T., M.T., S.T., A.T., B.T., C.Tu., P.T., C.Ty., A.V., V.V., P.Va., J.V., E.V., S.Wa., S.Wa., J.W., A.W., D.We., M.Wh., S.Wil., D.Wi., N.W., L.W., G.W., C.W., M.Wr., L.Y., M.Y., H.V.F., and D.R.F. Sample and data processing: S.Clay., T.W.F., E.P., D.Ra., K.A., D.M.B., T.B., P., N.K., L.E.M., A.R.T., A.P.B., S.Br., E.C., I.C., E.G., S.G., L.Hii., B.H., R.K., D.P., M.Po., J.Rand., G.J.S., S.Wid. and E.W. Validation experiments: J.F.M., E.P., D.Ra., A.Si., N.K. and C.F.W. Study design: M.Pa., H.V.F., C.F.W., D.R., J.C.B. and M.E.H. Method development and data analysis: J.F.M., E.P., D.Ra., A.Si., N.K. and C.F.W. Study design: M.Pa., H.V.F., C.F.W., D.R., J.C.B. and M.E.H. Data interpretation: J.F.M., E.P., D.Ra., A.Si., N.K. and C.F.W. Study design: M.Pa., H.V.F., C.F.W., D.R., J.C.B. and M.E.H. Experimental and analytical supervision: M.Pa., H.V.F., C.F.W., D.R., J.C.B. and M.E.H. Project Supervision: M.E.H.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.E.H. (meh@sanger.ac.uk).

Reviewer Information Nature thanks D. Goldstein, B. Neale and the other anonymous reviewer(s) for their contribution to the peer review of this work.
Jeremy F. McRae 1, Stephen Clayton 1, Tomas W. Fitzgerald 2, Joanna Kaplanis 3, Elena Prigmore 1, Diana Rajan 4, Alejandro Sifirin 5, Stuart Atkin 6, Nadia Akiawi 7, Mohsan Ali 7, Kirsty Armdridge 1, Daniel M. Barrett 8, Tanya Bayzetova 9, Philip Jones 10, Wendy D. Jones 11, Daniel King 12, Netravati Krishnapa 13, Laura E. Mason 14, Tarjinder Singh 15, Adrian R. Tivey 16, Munazza Ahmed 17, Uruj Anjam 18, Hayley Archer 19, Ruth Armstrong 20, Jana Awada 21, Meena Balasubramaniam 22, Siddharth Banakar 23, Diana Baillie 24, Alison Barco 25, Natalia Cahir 26, Bruce Castle 27, Kate Chandler 28, Elena Chatzimichalaki 29, Deirdre Cilliers 30, Clarke 31, Susan Clasper 32, Jil Clayton-Smith 33, Virginia Clowes 34, Andrea Coates 35, Trevor Cole 36, Irena Colugi 37, Amanda Collins 38, Mornag Colman 39, Fiona Connolly 40, Nicola Cooper 41, Helen Cox 42, Lara出入境 43, Gareth Cross 44, Yanick Crow 45, Mariella D’Alessandro 46, Tabib Dabi 47, Rosemarie Davidson 48, Sally Davies 49, Dylan de Vries 50, John Dean 51, Charu Deshpande 52, Gemma Devlin 53, Abhijit Dixit 54, Angus Dobbie 55, Alan Donaldson 56, Donnanna 57, Deidre Donnelly 58, Carina Donnelly 59, Angela Douglas 60, Sofia Dousouz 61, Alexs Duncan 62, Jacobean 63, Eain Elbard 64, Ian Ellis 65, Frances Elsmie 66, Karenza Evans 67, Sarah Everett 68, Tina Fendick 69, Richard Fisher 70, Frances Flintoff 71, Siobhán Folks 72, Andrew Fryer 73, Carol Gardine 74, Lorraine Gaunt 75, Neeti Ghali 76, Richard Gibbons 77, Harinder Gill 78, Judith Goodship 79, David Goodie 80, Emma Gray 81, Andrew Green 82, Philip Greene 83, Lynne Greenhalgh 84, Susan Grubbie 85, Rachel Harrison 86, Lucy Harrison 87, Victoria Harrison 88, Rose Hawkins 89, Liu He 90, Ashley Hendry 91, Sarah Hewitt 92, Lucy hydrody 93, Emma Hobson 94, Simon Holden 95, Muriel Holder 96, Susan Holder 97, Georgina Hollingsworth 98, Tessa Hornfay 99, Mervyn Humphreys 100, Jane Hurst 101, Ben Hutton 102, Stuart Ingram 103, Melia Irving 104, Lily Islam 105, Andrew Jack 106, Ian Jason 107, Usha Jatin 108, Lucy Jebut 109, Chris Johnen 110, Jonathan Berg 111, Birgitta Bernhard 112, A. Paul Bevan 113, Maria Bitner-Glindzicz 114, Edward Blair 115, Moira Blyth 116, David Bohanna 117, Louise Bourdon 118, David Boul 119, Lisa Bradley 120, Angela Brady 121, Simon Brent 122, Carole Brewer 123, Kate Brunstrom 124, David J. Bunn 125, John Burno 126, Natalie Cahn 127, Liu He 128, Carol Gardiner 129, Norman Pratt 130, Alison Krasus 131, Lahore Kralwalt 132, Jonathon Waters 133, Aashr Weber 134, Diana Wellesley 135, Margo Whiteford 136, Sara Widaa 138, Sarah Wilcox 139, Emily Wilkinson 140, Denise Williams 141, Nicola Williams 142, Louise Wilson 143, Geoff Woods 144, Christopher Wrage 145, Michael Wright 146, Laura Yates 147, Matthew Yau 148, Chris Nellaker 149, Michael Parker 150, Helen V. Firth 151, Caroline F. Wright 152, David R. Fitzpatrick 153, Jeffrey C. Barrett 154, & Matthew E. Hurles 155.

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. 1MRC Human Genetics Unit, MRC IGMU, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, UK. 2Department of Engineering Science, University of Oxford, Parks Road, Oxford OX1 3PR, UK. 3Wessex Clinical Genetics Service, University Hospital Southampton, Princess Anne Hospital, Oxford Road, Southampton SO16 6YX, UK. 4Wessex Regional Genetics Laboratory, Salisbury NHS Foundation Trust, Salisbury District Hospital, Oadstock Road, Salisbury, Wiltshire SP2 8BJ, UK. 5Faculty of Medicine, University of Southampton, Building 85, Life Sciences Building, Highfield Campus, Southampton S017 1BJ, UK. 6South West Thames Regional Genetics Centre, St George’s Healthcare NHS Trust, St George’s, University of London, Cranmer Terrace, London SW17 ORE, UK. 7Institute of Medical Genetics, University Hospital of Wales, Heath Park, Cardiff CF14 4XW, UK. 8Department of Clinical Genetics, Block 12, Glan Clwyd Hospital, Rhyl, Denbighshire LL18 5UL, UK. 9East Anglian Medical Genetics Service, Box 143, Cambridge University Hospitals NHS Foundation Trust, Cambridge Biomedical Campus, Cambridge CB2 OQQ, UK. 10Shffield Regional Genetics Services, Sheffield Children’s NHS Trust, Western Bank, Sheffield S10 2TH, UK. 11Manchester Centre for Genomic Medicine, St Mary’s Hospital, Central Manchester University Hospitals NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester M13 9WL, UK. 12North East Thames Regional Genetics Service, Great Ormond Street Hospital for Children NHS Trust, Great Ormond Street Hospital, Great Ormond Street, London WC1N 3JH, UK. 13North of Scotland Regional Genetics Service, NHS Grampian, Department of Medical Genetics Medical School, Forestheri, Aberdeen AB25 2ZD, UK. 14East of Scotland Regional Genetics Service, Human Genetics Unit, Pathology Department, NHS Tayside, Ninewells Hospital, Dundee DD1 9SY, UK. 15Yorkshire Regional Genetics Service, Leeds Teaching Hospitals NHS Trust, Department of Clinical Genetics, Chapel Allerton Hospital, Chapeltown Road, Leeds LS7 4SA, UK. 16North West Thames Regional Genetics Centre, North West London Hospitals NHS Trust, The Kennedy Galton Centre, Northwick Park and St Mark’s NHS Trust Watford Road, Harrow HA1 3UL, UK. 17Oxford Regional Genetics Service, Oxford Radcliffe Hospitals NHS Trust, The Churchill Old Road, Oxford OX3 7LJ, UK. 18West Midlands Regional Genetics Service, Birmingham Women’s NHS Foundation Trust, Birmingham Women’s Hospital, Edgbaston, Birmingham B13 2TG, UK. 19Northern Genetics Service, Newcastle upon Tyne Hospitals NHS Foundation Trust, Institute of Human Genetics, International Centre for Life, Central Parkway, Newcastle upon Tyne NE1 3BZ, UK. 20Northern Ireland Regional Genetics Centre, Belfast Health and Social Care Trust, Belfast City Hospital, Lisburn Road, Belfast BT9 7AB, UK. 21Peninsula Clinical Genetics Service, Royal Devon and Exeter NHS Foundation Trust, Clinical Genetics Department, Royal Devon & Exeter Hospital (Heavitree), Gladesmore Road, Exeter EX1 2ED, UK. 22South East Thames Regional Genetics Centre, Guy’s and St Thomas’ NHS Foundation Trust, Guy’s Hospital, Great Maze Pond, London SE1 9RT, UK. 23Leicestershire Genetics Centre, University Hospitals of Leicester NHS Trust, Leicester Royal Inflammatory (NHS Trust), Leicester LE1 5SW, UK. 24Nottingham Regional Genetics Service, City Hospital Campus, Nottingham University Hospitals NHS Trust, The Gables, Hucknall Road, Nottingham NG5 1PB, UK. 25West of Scotland Regional Genetics Service, NHS Greater Glasgow and Clyde, Institute of Medical Genetics, Yorkhill Hospital, Glasgow G3 8SJ, UK. 26Bristol Genetics Service (Avon, Somerset, Gloucesters and West Wiltshire, University Hospitals Bristol NHS Foundation Trust, St Michael’s Hospital, Bristol BS2 8DT, UK. 27Wessex Clinical Genetics Service, Liverpool Women’s NHS Foundation Trust, Department of Clinical Genetics, Liverpool Children’s Hospital Alder Hey, Eaton Road, Liverpool L12 2AP, UK. 28National Centre for Medical Genetics, Our Lady’s Children’s Hospital, Crumlin, Dublin 12, Ireland. 29Department of Clinical Genetics, Block 12, Glan Clwyd Hospital, Rhyl, Denbighshire LL18 5UL, Wals, UK. 30Northfield Department of Obstetrics & Gynaecology, University of Oxford, Level 3, Women’s Centre, John Radcliffe Hospital, Oxford OX3 9DU, UK. 31Biostatistical Engineering, Department of Engineering Science, University of Oxford, Old Road Campus Research Building, Oxford OX3 7QD, UK. 32Big Data Institute, University of Oxford, Roosevelt drive, Oxford OX3 7UL, UK. 33The Etxe Centre, Nuffield Department of Population Health, University of Oxford, Old Road Campus, Oxford OX3 7LF, UK.

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.
METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Family recruitment. At 24 clinical genetics centres within the United Kingdom National Health Service and the Republic of Ireland, 4,293 patients with severe, undiagnosed DDs and their parents (4,125 families) were recruited and systematically phenotyped. The study has UK Research Ethics Committee approval (10/EE0385/83, granted by the Cambridge South Research Ethics Committee and GEN/284/12, granted by the Republic of Ireland Research Ethics Committee). Families gave informed consent for participation.

Clinical data (growth measurements, family history, developmental milestones, and so on) were collected using a standard restricted-term questionnaire within DECIPHER, and detailed developmental phenotypes for the individuals were entered using HPO terms. Saliva samples for the whole family and blood-extracted DNA samples for the probands were collected, processed and quality controlled as previously described.

Exome sequencing. Genomic DNA (approximately 1μg) was fragmented to an average size of 150 base pairs (bp) and a DNA library was created using established Illumina paired-end protocols. Adaptor-ligated libraries were amplified and indexed using polymerase chain reaction (PCR). A portion of each library was used to create an equimolar pool comprising eight indexed libraries. Each pool was hybridized to SureSelect RNA baits (Agilent Human All-Exon V3 Plus with custom ELID C0338371 and Agilent Human All-Exon V5 Plus with custom ELID C0338371) and sequence targets were captured and amplified in accordance with the manufacturer’s recommendations. Enriched libraries were analysed by 75-base paired-end sequencing (Illumina HiSeq) following the manufacturer’s instructions.

Alignment and calling single-nucleotide variants, insertions and deletions. Mapping of short-read sequences for each sequencing lanelet was carried out using the Burrows-Wheeler aligner (BWA; version 0.5.9)1 backtrack algorithm with the GRCh37 1000 Genomes Project phase 2 reference (also known as h37d5). Sample-level BAM improvement was carried out using the Genome Analysis Toolkit (GATK; version 3.1.4)42 and SAMtools (version 0.1.19)43. This consisted of a realignment of reads around known and discovered indels (insertions and deletions) followed by base quality score recalibration (BQSR), with both steps performed using GATK. Lastly, SAMtools calmd was applied and indices were created.

Known indels for realignment were taken from the Mills Devine and 1000 Genomes Project Gold set and the 1000 Genomes Project phase low-coverage set, both part of the GATK resource bundle (version 2.2). Known variants for BQSR were taken from dbSNP 137, also part of the GATK resource bundle. Finally, single-nucleotide variants (SNVs) and indels were called using the GATK HaplotypeCaller (version 3.2.2); this was run in multisample calling mode using the complete dataset. GATK Variant Quality Score Recalibration (VQSR) was then computed on the whole dataset and applied to the individual-sample variant calling format (VCF) files. DeNovoGear (version 0.54)44 was used to detect SNV, insertion and deletion DNMs from child and parental exome data (BAM files).

Variant annotation. Variants in the VCF were annotated with minor allele frequency (MAF) data from a variety of different sources. The MAF annotations used included: (1) four different populations of the 1,000 Genomes Project (American, Asian, African and European), the UK10K cohort, the NHLBI GO Exome Sequencing Project (ESP), the Non-Finnish European (NF/E) subset of the Exome Aggregation Consortium (ExAC) and an internal allele frequency generated using unaffected parents from the cohort.

Variants in the VCF were annotated with Ensembl Variant Effect Predictor (VEP)46 based on Ensembl gene build 76. The transcript with the most severe consequence was selected and all associated VEP annotations were based on the predicted effect of the variant on that particular transcript; where multiple transcripts shared the same most severe consequence, the canonical or longest was selected. We included additional consequence for variants at the last base of an exon before an intron, where the final base is a guanine, since these variants appear to be as damaging as a splice-donor variant24. We categorized variants into three classes by VEP consequence: (1) protein-truncating variants (PTV): splice donor, splice acceptor, stop gained, frameshift, initiator codon and conserved exon terminus variant; (2) missense variants: missense, stop lost, frameshift, initiator codon, exon insertion, coding sequence and protein altering variant; (3) silent variants: synonymous.

DNM filtering. We filtered candidate DNM calls to reduce the false-positive rate but to maximize sensitivity, on the basis of previous results from experimental validation by capillary sequencing of candidate DNMs.41 Candidate DNMs were excluded if not called by GATK in the child, or called in either parent, or if they had a maximum MAF greater than 0.01. Candidate DNMs were excluded when the forward and reverse coverage differed between reference and alternative alleles, defined as $P < 10^{-3}$ using a Fisher’s exact test of coverage from orientation by allele summed across the child and parents.

Candidate DNMs were also excluded if they met two of the three following three criteria: (1) an excess of parental alternative alleles within the cohort at the DNMs position, defined as $P < 10^{-3}$ under a one-sided binomial test given an expected error rate of 0.002 and the cumulative parental depth; (2) an excess of alternative alleles within the cohort in DNMs in a gene, defined as $P < 10^{-3}$ under a one-sided binomial test given an expected error rate of 0.002 and the cumulative depth; or (3) both parents had one or more reads supporting the alternative allele.

If, after filtering, more than one variant was observed in a given gene for a particular trio, only the variant with the highest predicted functional impact was kept (protein truncating $>$ missense $>$ silent).

DNM validation. For candidate DNMs of interest, primers were designed to amplify 150–250 bp products centred around the site of interest. Default primer3 design settings were used with the following adjustments: GC clamp $= 1$, human mispriming library used. Site-specific primers were tagged with Illumina PCR primers along with unique barcodes enabling multiplexing of 96 samples. Barcodes were incorporated using Kapa HiFi mastermix (Kapa Biosystems). Samples were pooled and sequenced down one lane of the Illumina MiSeq, using 250 bp paired-end reads. An in-house analysis pipeline extracted the read count per site and classified inheritance status per variant using a maximum likelihood approach (see Supplementary Note).

Individuals with likely pathogenic variants. We previously screened 1,133 individuals for variants that contribute to their disorder.11–15. All candidate variants in the 1,133 individuals were reviewed by consultant clinical geneticists for relevance to the individuals’ phenotypes. Most diagnosable pathogenic variants occurred de novo in dominant genes but a small proportion also occurred in recessive genes or under other inheritance modes. DNMs within dominant DD-associated genes were very probable to be classified as the pathogenic variant for the individuals’ disorder. Owing to the time required to review individuals and their candidate variants, we did not perform a similar review in the remainder of the 4,293 individuals. Instead we defined probable pathogenic variants as candidate DNMs found in autosomal and X-linked dominant DD-associated genes, or candidate DNMs found in hemizygous DD-associated genes in males. 1,136 individuals in the 4,293 cohort had variants either previously classified as pathogenic11–15, or had a probably pathogenic DNM.

Gene-wise assessment of DNM significance. Gene-specific germline mutation rates for different functional classes were computed11,12 for the longest transcript in the union of transcripts overlapping the observed DNMs in that gene. We evaluated the gene-specific enrichment of PTV and missense DNMs by computing its statistical significance under a null hypothesis of the expected number of DNMs given the gene-specific mutation rate and the number of considered chromosomes.

We also assessed clustering of missense DNMs within genes, as expected for DNMs causing activating or dominant-negative mechanisms. We did this by calculating simulated dispersions of the observed number of DNMs within the gene. The probability of simulating a DNM at a specific codon was weighted by the number of its dominant genes but a small proportion also occurred in recessive genes or under other inheritance modes. DNMs within dominant DD-associated genes were very probable to be classified as the pathogenic variant for the individuals’ disorder. Owing to the time required to review individuals and their candidate variants, we did not perform a similar review in the remainder of the 4,293 individuals. Instead we defined probable pathogenic variants as candidate DNMs found in autosomal and X-linked dominant DD-associated genes, or candidate DNMs found in hemizygous DD-associated genes in males. 1,136 individuals in the 4,293 cohort had variants either previously classified as pathogenic11–15, or had a probably pathogenic DNM.

Gene-wise assessment of DNM significance. Gene-specific germline mutation rates for different functional classes were computed11,12 for the longest transcript in the union of transcripts overlapping the observed DNMs in that gene. We evaluated the gene-specific enrichment of PTV and missense DNMs by computing its statistical significance under a null hypothesis of the expected number of DNMs given the gene-specific mutation rate and the number of considered chromosomes.

We also assessed clustering of missense DNMs within genes, as expected for DNMs causing activating or dominant-negative mechanisms. We did this by calculating simulated dispersions of the observed number of DNMs within the gene. The probability of simulating a DNM at a specific codon was weighted by the number of its dominant genes but a small proportion also occurred in recessive genes or under other inheritance modes. DNMs within dominant DD-associated genes were very probable to be classified as the pathogenic variant for the individuals’ disorder. Owing to the time required to review individuals and their candidate variants, we did not perform a similar review in the remainder of the 4,293 individuals. Instead we defined probable pathogenic variants as candidate DNMs found in autosomal and X-linked dominant DD-associated genes, or candidate DNMs found in hemizygous DD-associated genes in males. 1,136 individuals in the 4,293 cohort had variants either previously classified as pathogenic11–15, or had a probably pathogenic DNM.

Gene-wise assessment of DNM significance. Gene-specific germline mutation rates for different functional classes were computed11,12 for the longest transcript in the union of transcripts overlapping the observed DNMs in that gene. We evaluated the gene-specific enrichment of PTV and missense DNMs by computing its statistical significance under a null hypothesis of the expected number of DNMs given the gene-specific mutation rate and the number of considered chromosomes.

We also assessed clustering of missense DNMs within genes, as expected for DNMs causing activating or dominant-negative mechanisms. We did this by calculating simulated dispersions of the observed number of DNMs within the gene. The probability of simulating a DNM at a specific codon was weighted by the number of its dominant genes but a small proportion also occurred in recessive genes or under other inheritance modes. DNMs within dominant DD-associated genes were very probable to be classified as the pathogenic variant for the individuals’ disorder. Owing to the time required to review individuals and their candidate variants, we did not perform a similar review in the remainder of the 4,293 individuals. Instead we defined probable pathogenic variants as candidate DNMs found in autosomal and X-linked dominant DD-associated genes, or candidate DNMs found in hemizygous DD-associated genes in males. 1,136 individuals in the 4,293 cohort had variants either previously classified as pathogenic11–15, or had a probably pathogenic DNM.
The dataset used for this work may contain multiple photos for one patient. To avoid biasing the average face mesh towards these individuals, we computed an average face for each patient and use these personal averages to compute the final average face. Finally, to avoid any image in the composite dominating owing to variance in illumination between images, we normalized the intensities of pixel values within the face to an average value across all faces in each average. The composite faces were assessed visually to confirm successful ablation of any individually identifiable features. Visual assessment of the composite photographs by two experienced clinical geneticists, alongside the individual patient photos, was performed for all 93 genome-wide significant DD-associated genes for which clinical photos were available for more than one patient, to remove potentially identifiable composite faces as well as quality control on the automated composite face generation process. Eighty-one composite faces were excluded leaving the twelve de-identified composite faces that are shown in Fig. 2 and Extended Data Fig. 3. Each of the twelve composite faces that passed de-identification and quality control was generated from photos of ten or more patients.

Assessing power of incorporating phenotypic information. We previously described a method to assess phenotypic similarity by HPO terms among groups of individuals sharing genetic defects in the same gene28. We examined whether incorporating this statistical test improved our ability to identify dominant genes at genome-wide significance. Per gene, we tested the phenotypic similarity of individuals with DNMs in the gene. We combined the phenotypic-similarity P value with the genotypic P value per gene (the minimum P value from the DDD-only and meta-analysis) using Fisher’s method. We examined the distribution of differences in P value between tests without the phenotypic-similarity P value and tests that incorporated the phenotypic-similarity P value.

Many individuals (854, 20%) of the DDD cohort experience seizures. We investigated whether testing within the subset of individuals with seizures improved our ability to identify gene clusters for seizure-specific genes. A list of 102 seizure-associated genes was curated from three sources: a gene panel for Ohtahara syndrome, a currently used clinical gene panel for epilepsy and a panel derived from DD-associated genes19. The P values from the seizure subset were compared to P values from the complete cohort.

Assessing power of exome versus genome sequencing. We compared the expected power of exome sequencing versus genome sequencing to identify disease genes. Within the DDD cohort, 55 dominant DD-associated genes achieve genome-wide significance when testing for enrichment of DNMs within genes. We did not incorporate missense DNM clustering owing to the large computational requirements for assessing clustering in many replicates.

We estimated the proportion of excess DNMs with a loss-of-function mechanism. We identified the optimal mixing proportion for the loss-of-function and altered-function mechanisms. We found that the excess of PTV DNMs within dominant haploinsufficient DD-associated genes had a greater skew towards genes with high intolerance for loss-of-function variants than the excess of missense DNMs in dominant non-haploinsufficient genes. We binned genes by the probability of being loss-of-function intolerant30 constraint decile and calculated the observed excess of missense DNMs in each bin. We modelled this binned distribution as a two-component mixture with the components representing DNMs with a loss-of-function or altered-function mechanism. We identified the optimal mixing proportion for the loss-of-function and altered-function DNMs from the lowest goodness of fit (from a spline fitted to the observed data, per decile) to missense/inframe indels in all genes across a range of mixtures.

The excess of DNMs with a loss-of-function mechanism was calculated as the excess of DNMs with a VEP loss-of-function consequence, plus the proportion of the excess of missense DNMs at the optimal mixing proportion.

We estimated the proportions for loss of function and altered function. We counted PTV and missense/inframe indel DNMs within dominant haploinsufficient genes to estimate the proportion of excess DNMs with a loss-of-function mechanism, but which were classified as missense/inframe indel. We estimated the proportion of excess DNMs with a loss-of-function mechanism as the PTV excess plus the PTV excess multiplied by the proportion of loss of function classified as missense.

Prevalence of DDs from dominant DNMs. We estimated the birth prevalence of monoallelic DDs by using the germline-mutation model. We calculated the expected cumulative germline-mutation rate of truncating DNMs in 238 haploinsufficient DD-associated genes. We scaled this upwards based on the composition of excess DNMs in the DDD cohort using the ratio of excess DNMs (n = 1,816) to DNMs within dominant haploinsufficient DD-associated genes (n = 412). Around 10% of DDs are caused by copy-number variations (CNVs) or other copy-number events, such as insertions and deletions.

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.
different combinations of parental ages, given our estimates of the extra DNMs per year from older mothers and fathers. We scaled the prevalence to different combinations of parental ages using the ratio of expected mutations at a given age combination to the number expected at the mean cohort parental ages.

To estimate the annual number of live births with DDs caused by DNMs, we obtained country population sizes, birth rates, age at first birth\textsuperscript{51}, and calculated global birth rate (18.58 live births per 1,000 individuals) and age at first birth (22.62 years), weighted by population size. We calculated the mean age when giving birth (26.57 years) given a total fertility rate of 2.45 children per mother\textsuperscript{52}, and a mean interpregnancy interval of 29 months\textsuperscript{53}. We calculated the number of live births given our estimate of DD prevalence caused by DNMs at this age (0.00288), the global population size (7.4 billion individuals) and the global birth rate.

**Code availability.** Source code for filtering candidate DNMs, testing DNM enrichment, DNM clustering and phenotypic similarity can be found here: https://github.com/jeremymcrae/denovoFilter, https://github.com/jeremymcrae/mupit, https://github.com/jeremymcrae/denovonear and https://github.com/jeremymcrae/hpo_similarity.

**Data availability.** Exome sequencing and phenotype data are accessible via the European Genome-phenome Archive (EGA) under accession number EGAS00001000775 (https://www.ebi.ac.uk/ega/studies/EGAS00001000775). Details of DD-associated genes are available at www.ebi.ac.uk/gene2phenotype. All other data are available from the corresponding author upon reasonable request.

39. Bragin, E. et al. DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. *Nucleic Acids Res.* **42**, D993–D1000 (2014).
40. Köhler, S. et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am. J. Hum. Genet.* **85**, 457–464 (2009).
41. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
42. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
43. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
44. Ramu, A. et al. DeNovoGear: de novo indel and point mutation discovery and phasing. *Nat. Methods* **10**, 985–987 (2013).
45. Abecasis, G. R. et al. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
46. McLaren, W. et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069–2073 (2010).
47. Felzenszwalb, P. F., Girshick, R. B., McAllester, D. & Ramanan, D. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**, 1627–1645 (2010).
48. Xiong, X. & De la Torre, F. Supervised Descent method and its applications to face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 532–539 (Portland, 2013).
49. Cooper, G. M. et al. A copy number variation morbidity map of developmental delay. *Nat. Genet.* **43**, 838–846 (2011).
50. Sago, G. S. et al. Array CGH in patients with learning disability (mental retardation) and congenital anomalies: updated systematic review and meta-analysis of 19 studies and 13,926 subjects. *Genet. Med.* **11**, 139–146 (2009).
51. Central Intelligence Agency. *The World Factbook*. Vol. 2016 (2016).
52. The World Bank. Fertility rate, total (births per woman). in *World Development Indicators* (2016).
53. Copen, C. E., Thoma, M. E. & Kirmeyer, S. Interpregnancy Intervals in the United States: Data From the Birth Certificate and the National Survey of Family Growth. In *National Vital Statistics Reports* Vol. 64 (National Center for Health Statistics, 2015).
Extended Data Figure 1 | Proportion of individuals with a DNM that is probably pathogenic. Only individuals with protein-altering or protein-truncating DNMs in dominant or X-linked dominant DD-associated genes, or males with DNMs in hemizygous DD-associated genes were included. The proportions given are for those individuals with any DNMs rather than the total number of individuals in each subset. Cohorts included in the DNM meta-analyses are shaded blue.
Extended Data Figure 2 | Phenotypic summary of genes without previous compelling evidence. Phenotypes are grouped by type. The first group indicates numbers of individuals with DNMs per gene divided by sex (m, male; f, female), and by functional consequence (NSV, nonsynonymous variant; PTV, protein-truncating variant). The second group indicates mean values for growth parameters: birthweight (bw), height (ht), weight (wt) and occipitofrontal circumference (OFC). Values are given as standard deviations from the healthy population mean derived from ALSPAC (Avon longitudinal study of parents and children) data. The third group indicates the mean age for achieving developmental milestones: age of first social smile, age of first sitting unassisted, age of first walking unassisted and age of first speaking. Values are given in months. The final group summarizes HPO-coded phenotypes per gene, as number of HPO terms within different clinical categories.
Extended Data Figure 3 | Phenotypic summary of individuals with DNMs in genes achieving genome-wide significance. Phenotypes are grouped by type. The first group indicates numbers of individuals with DNMs per gene divided by sex (m, male; f, female), and by functional consequence (NSV, nonsynonymous variant; PTV, protein-truncating variant). The second group indicates mean values for growth parameters: birthweight (bw), height (ht), weight (wt) and occipitofrontal circumference (OFC). Values are given as standard deviations from the healthy population mean derived from ALSPAC data. The third group indicates the mean age for achieving developmental milestones: age of first social smile, age of first sitting unassisted, age of first walking unassisted and age of first speaking. Values are given in months. The final group summarizes HPO-coded phenotypes per gene, as number of HPO terms within different clinical categories.
Extended Data Figure 4 | Dispersion of DNMs and domains for each novel gene. a, CDK13. b, CHD4. c, CNOT3. d, CSNK2A1. e, GNAI1. f, KCNQ3. g, MSL3. h, PPM1D. i, PUF60. j, QRICH1. k, SET. l, KMT5B. m, TCF20. n, ZBTB18.
Extended Data Figure 5 | Effect of clustering by phenotype on the ability to identify genome-wide significant genes. a, Comparison of $P$ values derived from genotypic information alone versus $P$ values that incorporate genotypic information and phenotypic similarity. b, Comparison of $P$ values from tests in the complete DDD cohort versus tests in the subset with seizures. Genes that were previously linked to seizures are shaded blue. c, Proportion of cohort with a DNM in a seizure-associated gene, stratified by seizure-affected status. Error bars, 95% CI. d, Comparison of power to identify genome-wide significant genes in probands with seizures, versus the unstratified cohort, at matched sample sizes.
Extended Data Figure 6 | Power of genome versus exome sequencing to discover dominant genes associated with DDs. 

a, The number of genes exceeding genome-wide significance was estimated at three different fixed budgets ($USD1, 2 or 3 million) and a range of relative sensitivities for genomes versus exomes to detect DNMs. The number of genes identifiable by exome sequencing are shaded blue, whereas the number of genes identifiable by genome sequencing are shaded green. The regions where exome sequencing costs 30–40% of genome sequencing are shaded with a grey background, which corresponds to the price differential in 2016.

b, Simulated estimates of power to detect loss-of-function genes in the genome at different cohort sizes, given fixed budgets.
Extended Data Figure 7 | Gene-wise significance of neurodevelopmental genes versus the expected number of mutations per gene. Points are shaded by clinical recognizability classification (blue and brown points denote cryptic and distinctive disorders, respectively).
Extended Data Figure 8 | Stringency of DNM filtering. **a**, Sensitivity and specificity of DNM validations within sets filtered using varying thresholds of DNM quality (posterior probability of DNM). The analysed DNMs were restricted to sites identified within the earlier 1,133 trios\textsuperscript{15}, where all candidate DNMs underwent validation experiments. The labelled value is the quality threshold at which the number of candidate synonymous DNMs equals the number of expected synonymous mutations under a null germline mutation rate. **b**, Excess of missense and loss-of-function DNMs at varying DNM quality thresholds. The DNM excess is adjusted for the sensitivity and specificity at each threshold.
Extended Data Figure 9 | Enrichment of DNMs by consequence type, across functional constraint quantiles for residual variation intolerance scores. A comparison of enrichment for residual variation intolerance score (RVIS) values generated from ESP6500 data (ref. 31) versus ExAC data (obtained from http://genic-intolerance.org/) are provided.
Extended Data Table 1 | Phenotypes tested for association with having a pathogenic DNM

| Category          | Phenotype                      | Type          | Value          | 95% CI          | P-value   |
|-------------------|--------------------------------|---------------|----------------|----------------|-----------|
| Post-natal        | abnormal cranial MRI           | Odds ratio    | 1.365          | 1.125 – 1.656  | 0.002     |
|                   | feeding problems               | Odds ratio    | 1.176          | 1.01 – 1.369   | 0.039     |
|                   | neonatal intensive care        | Odds ratio    | 0.896          | 0.762 – 1.054  | 0.190     |
|                   | anticonvulsant drugs           | Odds ratio    | 0.582          | 0.246 – 1.377  | 0.270     |
| Pre-natal         | bleeding                       | Odds ratio    | 0.892          | 0.714 – 1.114  | 0.346     |
|                   | maternal illness               | Odds ratio    | 0.908          | 0.764 – 1.079  | 0.278     |
|                   | maternal diabetes              | Odds ratio    | 0.787          | 0.504 – 1.229  | 0.341     |
|                   | abnormal scan                  | Odds ratio    | 0.839          | 0.692 – 1.017  | 0.078     |
|                   | assisted reproduction          | Odds ratio    | 0.868          | 0.554 – 1.36   | 0.584     |
|                   | increased nuchal translucency  | Odds ratio    | 1.432          | 0.903 – 2.271  | 0.126     |
| Family history    | consanguinity                  | Odds ratio    | 0.234          | 0.138 – 0.397  | 8.0 x 10^-11 |
|                   | similar phenotype parents      | Odds ratio    | 0.295          | 0.184 – 0.474  | 5.7 x 10^-3  |
|                   | similar phenotype relatives    | Odds ratio    | 0.553          | 0.402 – 0.761  | 1.5 x 10^-3  |
|                   | similar phenotype siblings     | Odds ratio    | 0.311          | 0.23 – 0.421   | 7.3 x 10^-4  |
|                   | only patient affected          | Odds ratio    | 2.478          | 2.001 – 3.068  | 3.9 x 10^-3  |
|                   | X-linked inheritance           | Odds ratio    | 0.839          | 0.436 – 1.613  | 0.752     |
|                   | Multiple births                | Beta          | 0.043          | -0.058 – 0.144 | 0.403     |
|                   | History of pregnancy loss      | Beta          | -0.039         | -0.155 – 0.078 | 0.516     |
| Developmental milestones | first words         | Beta          | 0.205          | 0.081 – 0.328  | 0.001     |
|                   | walked independently           | Beta          | 0.125          | 0.016 – 0.235  | 0.025     |
|                   | sat independently              | Beta          | 0.050          | -0.069 – 0.17  | 0.408     |
|                   | social smile                   | Beta          | 0.072          | -0.066 – 0.211 | 0.305     |
| Growth            | height                         | Beta          | 0.008          | -0.111 – 0.126 | 0.897     |
|                   | birthweight                    | Beta          | -0.018         | -0.135 – 0.098 | 0.756     |
|                   | OFC                            | Beta          | -0.094         | -0.215 – 0.026 | 0.125     |
|                   | weight                         | Beta          | -0.331         | -1.278 – 0.615 | 0.493     |
| Age               | age at assessment              | Beta          | 0.116          | 0.015 – 0.217  | 0.025     |
|                   | gestation                      | Beta          | 0.079          | -0.033 – 0.19  | 0.167     |
|                   | father’s age                   | Beta          | 0.137          | 0.027 – 0.247  | 0.015     |
|                   | mother’s age                   | Beta          | 0.108          | -0.003 – 0.219 | 0.056     |
| Other             | phenotypic terms (n)           | Beta          | 0.104          | 0.004 – 0.203  | 0.041     |
|                   | autozygosity length            | Beta          | -0.185         | -0.254 – -0.115| 1.6 x 10^-7|
|                   | sex (male)                     | Odds ratio    | 0.750          | 0.646 – 0.87   | 1.6 x 10^-3|

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.