Nature. Author manuscript; available in PMC 2021 April 14.

Published in final edited form as:
Nature. 2020 October 01; 586(7831): 757–762. doi:10.1038/s41586-020-2832-5.

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

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 consortium authors †**, **Silvia Borras**[14], **Caroline Clark**[14], **John Dean**[14], **Zosia Miedzybrodzka**[14], **Alison Ross**[14], **Stephen Tennant**[14], **Tabib Dabir**[15], **Deirdre Donnelly**[15], **Mervyn Humphreys**[15], **Alex Magee**[15], **Vivienne McConnell**[15], **Shane McKee**[15], **Susan McNerlan**[15], **Patrick J. Morrison**[15], **Gillian Rea**[15], **Fiona Stewart**[15], **Trevor Cole**[16], **Nicola Cooper**[16], **Lisa Cooper-Charles**[16], **Helen Cox**[16], **Lily Islam**[16], **Joanna Jarvis**[16], **Rebecca Keelagher**[16], **Derek Lim**[16], **Dominic McMullan**[16], **Jenny Morton**[16], **Swati Naik**[16], **Mary O'Driscoll**[16], **Kai-Ren Ong**[16], **Deborah Osio**[16], **Nicola Ragge**[16], **Sarah Turton**[16], **Julie Vogt**[16], **Denise Williams**[16], **Simon Bodek**[17], **Alan Donaldson**[17], **Alison Hills**[17], **Karen Low**[17], **Ruth Newbury-Ecob**[17], **Andrew M. Norman**[17], **Eileen Roberts**[17], **Ingrid Scurr**[17], **Sarah Smithson**[17], **Madeleine Tooley**[17], **Steve Abbs**[12], **Ruth Armstrong**[12], **Carolyn Dunn**[12], **Simon Holden**[12], **Soo-Mi Park**[12], **Joan Paterson**[12], **Lucy Raymond**[12], **Evan Reid**[12], **Richard Sandford**[12], **Ingrid Simonic**[12], **Marc Tischkowitz**[12], **Geoff Woods**[12], **Lisa Bradley**[18], **Joanne Comerford**[18], **Andrew Green**[18], **Sally Lynch**[18], **Shirley McQuaid**[18], **Brendan Mullaney**[18], **Jonathan Berg**[19], **David Goudie**[19], **Eleni Mavrak**[19], **Joanne McLean**[19], **Catherine McWilliam**[19], **Eleanor Reavey**[19], **Tara Azam**[13], **Elaine Cleary**[13], **Andrew Jackson**[13], **Wayne Lam**[13], **Anne Lampe**[13], **David Moore**[13], **Mary Porteous**[13], **Emma Baple**[20], **Júlia Baptista**[20], **Carole Brewer**[20], **Bruce Castle**[20], **Emma Kivuva**[20], **Martina Owens**[20], **Julia Rankin**[20], **Charles Shaw-Smith**[20], **Claire Turner**[20], **Peter Turnpenny**[20], **Carolyn Tysoe**[20], **Therese Bradley**[21], **Rosemarie Davidson**[21], **Carol Gardiner**[21], **Shelagh Joss**[21], **Esther Kinning**[21], **Cheryl Longman**[21], **Ruth McGowan**[21], **Victoria Murday**[21], **Daniela Pilz**[21], **Edward Tobias**[21], **Margo Whiteford**[21], **Nicola Williams**[21], **Angela Barnicoat**[22], **Emma Clement**[22], **Francesca Faravelli**[22], **Jane Hurst**[22], **Lucy Jenkins**[22], **Wendy Jones**[22], **V. K. Ajith Kumar**[22], **Melissa Lees**[22], **Sam Loughlin**[22], **Alison Male**[22],

‡To whom correspondence should be addressed: meh@sanger.ac.uk.
#jointly supervised
†a list of authors and their affiliations appears at the end of the paper

Author contributions

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

Competing interests

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

Europe PMC Funders Author Manuscripts

Deborah Morrogh[22], Elisabeth Rosser[22], Richard Scott[22], Louise Wilson[22], Ana Beleza[23], Charu Deshpande[23], Frances Flinter[23], Muriel Holder[23], Melita Irving[23], Louise Izatt[23], Dragana Josifova[23], Shehla Mohammed[23], Aneta Molenda[23], Leema Robert[23], Wendy Roworth[23], Deborah Ruddy[23], Mina Ryten[23], Shu Yau[23], Christopher Bennett[24], Moira Blyth[24], Jennifer Campbell[24], Andrea Coates[24], Angus Dobbie[24], Sarah Hewitt[24], Emma Hobson[24], Eilidh Jackson[24], Rosalyn Jewell[24], Alison Kraus[24], Katrina Prescott[24], Eamonn Sheridan[24], Jenny Thomson[24], Kirsty Bradshaw[25], Abhijit Dixit[25], Jacqueline Eason[25], Rebecca Haines[25], Rachel Harrison[25], Stacey Mutch[25], Ajoy Sarkar[25], Claire Searle[25], Nora Shannon[25], Abid Sharif[25], Mohnish Suri[25], Pradeep Vasudevan[26], Natalie Canham[27], Ian Ellis[27], Lynn Greenhalgh[27], Emma Howard[27], Victoria Stinton[27], Andrew Swale[27], Astrid Weber[27], Siddharth Banka[28], Catherine Breen[28], Tracy Briggs[28], Emma Burkitt-Wright[28], Kate Chandler[28], Jill Clayton-Smith[28], Dian Donnai[28], Sofia Douzgou[28], Lorraine Gaunt[28], Elizabeth Jones[28], Bronwyn Kerr[28], Claire Langley[28], Kay Metcalfe[28], Audrey Smith[28], Ronnie Wright[28], David Bourn[29], John Burn[29], Richard Fisher[29], Steve Hellens[29], Alex Henderson[29], Tara Montgomery[29], Miranda Splitt[29], Volker Straub[29], Michael Wright[29], Simon Zwolinski[29], Zoe Allen[30], Birgitta Bernhard[30], Angela Brady[30], Claire Brooks[30], Louise Busby[30], Virginia Clowes[30], Neeti Ghali[30], Susan Holder[30], Rita Ibitoye[30], Emma Wakeling[30], Edward Blair[31], Jenny Carmichael[31], Deirdre Cilliers[31], Susan Clasper[31], Richard Gibbons[31], Usha Kini[31], Tracy Lester[31], Andrea Nemeth[31], Joanna Poulton[31], Sue Price[31], Debbie Shears[31], Helen Stewart[31], Andrew Wilkie[31], Shadi Albaba[32], Duncan Baker[32], Meena Balasubramanian[32], Diana Johnson[32], Michael Parker[32], Oliver Quarrell[32], Alison Stewart[32], Josh Willoughby[32], Charlene Crosby[33], Frances Elmslie[33], Tessa Homfray[33], Huilin Jin[33], Nayana Lahiri[33], Sahar Mansour[33], Karen Marks[33], Meriel McEntagart[33], Anand Saggar[33], Kate Tatton-Brown[33], Rachel Butler[34,35], Angus Clarke[34,35], Sian Corrin[34,35], Andrew Fry[34,35], Arveen Kamath[34,35], Emma McCann[35], Hood Mugalaasi[34,35], Caroline Pottinger[35], Annie Procter[34,35], Julian Sampson[34,35], Francis Sansbury[34,35], Vinod Varghese[34,35], Diana Baralle[36,37,38], Alison Callaway[36,37,38], Emma J. Cassidy[36,37,38], Stacey Daniels[36,37,38], Andrew Douglas[36,37,38], Nicola Foulds[36,37,38], David Hunt[36,37,38], Mira Kharbanda[36,37,38], Katherine Lachlan[36,37,38], Catherine Mercer[36,37,38], Lucy Side[36,37,38], I. Karen Temple[36,37,38], Diana Wellesley[36,37,38]

[14]North of Scotland Regional Genetics Service, NHS Grampian, Department of Medical Genetics Medical School, Aberdeen, UK [15]Northern Ireland Regional Genetics Centre, Belfast Health and Social Care Trust, Belfast City Hospital, Belfast, UK [16]West Midlands Regional Genetics Service, Birmingham Women's NHS Foundation Trust, Birmingham Women's Hospital, Birmingham, UK [17]Bristol Genetics Service (Avon, Somerset, Gloucs and West Wilts), University Hospitals Bristol NHS Foundation Trust, St Michael's Hospital, Bristol, UK [18]National Centre for Medical Genetics, Our Lady's Children's Hospital, Dublin, Ireland [19]East of Scotland Regional Genetics Service, Human Genetics Unit, Pathology Department, NHS Tayside, Ninewells Hospital, Dundee, UK [20]Peninsula Clinical Genetics Service, Royal Devon and Exeter NHS Foundation Trust, Clinical Genetics Department, Royal Devon & Exeter Hospital (Heavitree), Exeter, UK [21]West of Scotland Regional Genetics Service, NHS Greater Glasgow and Clyde, Institute of Medical Genetics, Yorkhill Hospital, Glasgow, UK [22]North East Thames Regional Genetics Service, Great Ormond Street Hospital for Children NHS Foundation Trust, Great Ormond Street Hospital, London,

UK [23]South East Thames Regional Genetics Centre, Guy's and St Thomas' NHS Foundation Trust, Guy's Hospital, London, UK [24]Yorkshire Regional Genetics Service, Leeds Teaching Hospitals NHS Trust, Department of Clinical Genetics, Chapel Allerton Hospital, Leeds, UK [25]Nottingham Regional Genetics Service, City Hospital Campus, Nottingham University Hospitals NHS Trust, Nottingham, UK [26]Leicestershire Genetics Centre, University Hospitals of Leicester NHS Trust, Leicester Royal Infirmary (NHS Trust), Leicester, UK [27]Merseyside and Cheshire Genetics Service, Liverpool Women's NHS Foundation Trust, Department of Clinical Genetics, Royal Liverpool Children's Hospital Alder Hey, Liverpool, UK [28]Manchester Centre for Genomic Medicine, St Mary's Hospital, Central Manchester University Hospitals NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester, UK [30]Northern Genetics Service, Newcastle upon Tyne Hospitals NHS Foundation Trust, Institute of Human Genetics, International Centre for Life, Newcastle upon Tyne, UK [30]North West Thames Regional Genetics Centre, North West London Hospitals NHS Trust, The Kennedy Galton Centre and St Mark's NHS Trust Watford Road, Harrow, UK [31]Oxford Regional Genetics Service, Oxford Radcliffe Hospitals NHS Trust, Oxford, UK [32]Sheffield Regional Genetics Services, Sheffield Children's NHS Trust, Sheffield, UK [33]South West Thames Regional Genetics Centre, St George's Healthcare NHS Trust, St George's, University of London, London, UK [34]Institute of Medical Genetics, University Hospital of Wales, Cardiff, UK [35]Department of Clinical Genetics, Glan Clwyd Hospital, Rhyl, UK [36]Wessex Clinical Genetics Service, University Hospital Southampton, Princess Anne Hospital, Southampton, UK [37]Wessex Regional Genetics Laboratory, Salisbury NHS Foundation Trust, Salisbury District Hospital, Salisbury, UK [38]Faculty of Medicine, University of Southampton, Southampton, UK

, **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 [8]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]East Anglian Medical Genetics Service, Cambridge University Hospitals NHS

Foundation Trust, Cambridge, UK [13]MRC Human Genetics Unit, MRC IGMM, University of Edinburgh, Western General Hospital, Edinburgh, UK

[#] These authors contributed equally to this work.

Summary

De novo mutations (DNMs) in protein-coding genes are a well-established cause of developmental disorders (DD)[1]. However, known DD-associated genes only account for a minority of the observed excess of such DNMs[1,2]. 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, 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.

Introduction

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

Identification of 285 DD-associated genes

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

To detect gene-specific enrichments of damaging DNMs, we developed a method named DeNovoWEST (*De Novo* Weighted Enrichment Simulation Test, https://github.com/queenjobo/DeNovoWEST). DeNovoWEST scores all classes of sequence variants on a unified severity scale based on empirically-estimated positive predictive values of being pathogenic (Supplementary Fig. 3–4). We perform two tests per gene: an enrichment test on all nonsynonymous DNMs and a test designed to detect genes likely acting via an altered-function mechanism, which combines a missense enrichment test with a missense clustering test. We then applied a Bonferroni multiple testing correction accounting for the number of genes (n=18,762) and two tests per gene.

We first applied DeNovoWEST to all individuals in our cohort and identified 281 significantly enriched genes, 18 more than when using our previous method[1] (Supplementary Fig. 5; Fig. 1a). The majority (196/281; 70%) of these significant genes already had sufficient evidence of DD-association to be considered of diagnostic utility (as of late 2019) by all three centres, and we refer to them as “consensus” genes. 54/281 of these significant genes were previously considered diagnostic by one or two centres (“discordant” genes). Applying DeNovoWEST to synonymous DNMs, as a negative control analysis, identified no significantly enriched genes (Supplementary Fig. 6).

To discover novel DD-associated genes with greater power, we applied DeNovoWEST to DNMs in patients without damaging DNMs in consensus genes (we refer to this subset as ‘undiagnosed’ patients) and identified 94 significant genes (Supplementary Fig. 7; Supplementary Table 2), of which 33 were putative ‘novel’ DD-associated genes. To ensure robustness to potential mutation rate variation between genes, we determined whether any of the putative novel DD-associated genes had significantly more synonymous variants in the Genome Aggregation Database[6] (gnomAD) of population variation than expected under our null mutation model (Supplementary Note). We identified 11/33 genes with a significant excess of synonymous variants. For these 11 genes we repeated the DeNovoWEST test, increasing the null mutation rate by the ratio of observed to expected synonymous variants in gnomAD. Five of these genes fell below our exome-wide significance threshold and were removed, leaving 28 novel genes, with a median of 10 nonsynonymous DNMs (Fig. 1b–c; Supplementary Table 3). There were 314 patients with nonsynonymous DNMs in these 28 genes (1.0% of our cohort); all these DNMs were inspected in IGV[10] and, of 198 for which experimental validation was attempted, all were confirmed as DNMs. The DNMs in these novel genes were distributed randomly across the three datasets (no genes with $p < 0.001$, heterogeneity test). Six of the 28 novel DD-associated genes are corroborated by OMIM entries or publications, including *TFE3* [11,12] which was described in two recent publications.

We also investigated whether some synonymous DNMs might be pathogenic by disrupting splicing. We identified a small but significant enrichment of synonymous DNMs with high values of the splicing pathogenicity score SpliceAI[13] (≥ 0.8, 1.56-fold enriched, $p = 0.0037$, Poisson test; Supplementary Table 4). This enrichment corresponds to an excess of ~15 splice-disrupting synonymous DNMs in our cohort, of which six are accounted for by a recurrent synonymous DNM in *KAT6B* known to disrupt splicing[14].

Taken together, 25.0% of our cohort has a nonsynonymous DNM in one of the consensus or significant DD-associated genes (Fig. 1d). We noted significant sex differences in the autosomal burden of nonsynonymous DNMs (Supplementary Fig. 8). The rate of nonsynonymous DNMs in consensus autosomal genes was significantly higher in females than males (OR = 1.16, $p = 4.4 \times 10^{-7}$, Fisher's exact test; Fig. 1e), as noted previously[1]. However, the exome-wide burden of autosomal nonsynonymous DNMs in all genes was not significantly different between undiagnosed males and females (OR = 1.03, p = 0.29, Fisher's exact test).

This suggests the existence of subtle sex differences in the genetic architecture of DD, especially with regard to known and undiscovered disorders. This could include sex-biased contribution of polygenic, oligogenic and/or environmental modifiers of phenotypic variation and thus clinical ascertainment.

Characteristics novel DD-associated genes

Based on semantic similarity[15] between Human Phenotype Ontology terms, patients with DNMs in the same novel DD-associated gene were less phenotypically similar to each other, on average, than patients with DNMs in a consensus gene ($p = 2.3 \times 10^{-11}$, Wilcoxon rank-sum test; Fig. 2a; Supplementary Figure 9). This suggests that these novel disorders less often result in distinctive and consistent clinical presentations, which may have made these disorders harder to discover via a phenotype-driven approach. Each of these novel disorders requires genotype-phenotype characterisation, which is beyond the scope of this study.

Overall, novel DD-associated genes encode proteins that have very similar functional and evolutionary properties to consensus genes (Fig. 2b; Supplementary Table 5). Despite the high-level functional similarity between known and novel DD-associated genes, nonsynonymous DNMs in the more recently discovered DD-associated genes are much more likely to be missense DNMs, and less likely to be PTVs (discordant and novel; $p = 1.2 \times 10^{-25}$, chi-squared test). Fifteen (54%) of the 28 novel genes only had missense DNMs. Consequently, we expect that a greater proportion of the novel genes will act via altered-function mechanisms (e.g. dominant negative or gain-of-function). For example, the novel gene *PSMC5* (DeNovoWEST $p = 2.6 \times 10^{-15}$) had one inframe deletion and nine missense DNMs, eight of which altered two structurally important amino acids in the AAA+ ATPase domain, and so is likely to operate via an altered-function mechanism (Supplementary Fig. 10a–b). None of the novel genes exhibited significant clustering of *de novo* PTVs.

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 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)[16–18].

We observed a significant overlap between the 285 DNM-enriched DD-associated genes and a set of 369 previously described cancer driver genes[19] (overlap of 70 genes; $p = 1.7 \times 10^{-49}$, logistic regression correcting for s_{het}), as observed previously[20,21], as well as a significant enrichment of nonsynonymous DNMs in both overlapping and non-overlapping cancer genes (Supplementary Table 9). We observe 117 DNMs at 76 recurrent somatic mutations observed in at least three patients in The Cancer Genome Atlas (TCGA)[22]. By modelling the germline mutation rate at these somatic driver mutations, we found that recurrent nonsynonymous mutations in TCGA are enriched 21-fold in our cohort ($p < 10^{-50}$, Poisson test, Supplementary Fig. 11), whereas recurrent synonymous mutations in TCGA are not significantly enriched (2.4-fold, $p = 0.13$, Poisson test). This suggests that this observation is driven by the pleiotropic effects of these mutations in development and tumourigenesis, rather than hypermutability.

Recurrent mutations

We identified 773 recurrent DNMs (736 SNVs and 37 indels), observed in 2-36 individuals, which allowed us to interrogate systematically the factors driving recurrent germline mutation. We considered three potential contributory factors: (i) clinical ascertainment enriching for pathogenic mutations, (ii) greater mutability at specific sites, and (iii) positive selection conferring a proliferative advantage in the male germline[23]. We observed evidence that all three factors contribute, but not mutually exclusively. Clinical ascertainment drives the observation that 65% of recurrent DNMs were in consensus genes, a 5.4-fold enrichment compared to DNMs only observed once ($p < 10^{-50}$, proportion test). Hypermutability underpins the observation that 64% of recurrent *de novo* SNVs occurred at hypermutable CpG dinucleotides[24], a 2.0-fold enrichment over DNMs only observed once ($p = 3.3 \times 10^{-68}$, chi-square test).

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

Positive germline selection can increase the apparent mutation rate more strongly[23] than either clinical ascertainment (10-100X in our dataset) or hypermutability (~10X for CpGs). However, only a minority of the most highly recurrent mutations in our dataset are in genes that have been previously associated with germline selection. Nonetheless, several lines of evidence suggested that the majority of these most highly recurrent mutations are likely to confer a germline selective advantage. Based on the observations above, DNMs under germline selection should be more likely to be activating missense mutations, and should be less enriched for CpG dinucleotides. Extended Data Table 1 shows the 16 *de novo* SNVs observed nine or more times in our cohort, only two of which are in known germline selection genes. All but two of these 16 *de novo* SNVs cause missense changes, all but two of these genes cause disease by an altered-function mechanism, and these DNMs were

depleted for CpGs relative to all recurrent mutations. Two of these genes with highly recurrent *de novo* SNVs, *SHOC2* and *PPP1CB*, encode interacting proteins that regulate the RAS-MAPK pathway, and pathogenic variants in these genes are associated with a Noonan-like syndrome[27]. Moreover, two of these recurrent DNMs are in the same gene *SMAD4*, which encodes a key component of the TGF-beta signalling pathway, potentially expanding the pathophysiology of germline selection beyond the RAS-MAPK pathway. Confirming germline selection of these mutations will require deep sequencing of testes and/or sperm[26].

Incomplete penetrance and pre/perinatal death

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

A key parameter in estimating statistical power to detect novel HI disorders is the fold-enrichment of *de novo* PTVs expected in undiscovered HI disorders. We observed that novel DD-associated HI genes had significantly lower PTV enrichment compared to the consensus HI genes ($p = 0.005$, Wilcox rank-sum test; Fig. 3b). Two additional factors that could lower DNM enrichment, and thus power to detect a novel DD-association, are reduced penetrance and increased pre/perinatal death (due to spontaneous fetal loss, termination of pregnancy for fetal anomaly, stillbirth, or early neonatal death). To evaluate incomplete penetrance, we investigated whether HI genes with a lower enrichment of *de novo* PTVs in our cohort are associated with greater prevalences of PTVs in the general population. We observed a significant negative correlation ($p = 0.031$, weighted linear regression) between PTV enrichment in our cohort and the ratio of PTV to synonymous variants in gnomAD[6], suggesting that incomplete penetrance does lower *de novo* PTV enrichment in our cohort (Fig. 3c).

Additionally, we observed that the fold-enrichment of *de novo* PTVs in consensus HI DD-associated genes in our cohort was significantly higher for genes with a low likelihood of presenting with a prenatal structural malformation ($p = 4.6 \times 10^{-5}$, Poisson test, Fig. 3d), suggesting that pre/perinatal death decreases our power to detect some novel disorders (see supplement for details).

Hundreds of DD genes not yet discovered

Downsampling of our cohort and repeating enrichment analyses showed that the discovery of DD-associated genes has not plateaued (Extended Data Fig 1a). 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 *de novo* PTVs among genes as a function of varying numbers of undiscovered HI DD-associated genes and fold-enrichments of *de novo* PTVs in those genes. We found that the remaining PTV burden is most likely spread across ~1,000 genes with ~10-fold PTV enrichment (Extended Data Fig 1b). This fold enrichment is three times lower than in known HI DD-associated genes, suggesting that incomplete penetrance and/or pre/perinatal death is 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 1,000 genes with a substantially lower fold-enrichment than in currently known DD-associated genes (Supplementary Fig. 13).

We calculated that a sample size of ~350,000 parent-offspring trios would be needed to have 80% power to detect a 10-fold enrichment of *de novo* PTVs for an average gene. Using this inferred 10-fold enrichment among undiscovered HI genes, from our current data we can evaluate the likelihood that any gene i 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.

Discussion

In this study, we have presented evidence for 28 novel developmental disorders by developing an improved statistical test for mutation enrichment and applying it to a dataset of exome sequences from 31,058 parent-offspring trios. Most of the increased power to detect novel disorders comes from the increase in sample size, rather than the improved statistical test. These 28 novel genes account for 1.0% of our cohort, and their inclusion in diagnostic workflows will catalyse increased diagnosis of similar patients globally. The value of this study for improving diagnostic yield extends beyond these 28 novel genes; the total number of genes added to diagnostic workflows of the three participating centres (including newly validated discordant genes) ranged from 48-65 genes. We have shown that both incomplete penetrance and pre/perinatal death reduce our power to detect novel DDs postnatally, and hypothesise that one or both of these factors are operating more strongly among undiscovered DD-associated genes. In addition, we have identified a set of highly recurrent mutations that are strong candidates for novel germline selection mutations, which should result in a higher than expected disease incidence that increases dramatically with increased paternal age.

Our study is approximately three times larger than a recent meta-analysis of DNMs from a collection of individuals with autism spectrum disorder, intellectual disability, and/or a developmental disorder[28]. We identified ~2.3 times as many significantly DD-associated genes as this previous study when using Bonferroni-corrected exome-wide significance (285 vs 124). In contrast to meta-analyses of published DNMs, the harmonised filtering of

candidate DNMs across cohorts in this study should be more robust to cohort-specific differences in the sensitivity and specificity of detecting DNMs.

We inferred indirectly that developmental disorders with higher rates of detectable prenatal structural abnormalities had greater pre/perinatal death. The potential size of this effect can be quantified from the recently published PAGE study of genetic diagnoses in a cohort of fetal structural abnormalities[29]. In this latter study, genetic diagnoses were not returned to participants during the pregnancy, and so genetic diagnostic information could not influence pre/perinatal death. In the PAGE study data, 69% of fetal abnormalities with a genetically diagnosable cause died perinatally or neonatally. This emphasises the substantial impact that pre/perinatal death can have on reducing the ability to discover novel DDs from postnatal recruitment alone, and motivates the integration of genetic data from prenatal, neonatal and postnatal studies in future studies.

To empower our mutation enrichment testing, we estimated positive predictive values (PPV) of each DNM being pathogenic on the basis of their predicted protein consequence, CADD score[3], selective constraint against heterozygous PTVs across the gene (s_{het})[30], and, for missense variants, presence in a region under selective missense constraint[4]. These PPVs should also be informative for variant prioritisation in the diagnosis of dominant developmental disorders. Further work is needed to see whether these PPVs might be informative for recessive developmental disorders, and in other types of dominant disorders. More generally, we hypothesise that empirically-estimated PPVs based on variant enrichment in large datasets will be similarly informative in many other disease areas.

We adopted a conservative statistical approach to identifying DD-associated genes. In two previous studies using the same significance threshold, we identified 26 novel DD-associated genes[1,31]. All 26 are now regarded as being diagnostic, and have entered routine clinical diagnostic practice. Had we used a significance threshold of FDR < 10% as used in Satterstrom, Kosmicki, Wang et al[32], we would have identified 770 DD-associated genes. The FDR of individual genes depends on the significance of other genes being tested, so are not appropriate for assessing the significance of individual genes, but rather for defining gene-sets. There are 184 consensus genes that did not cross our significance threshold in this study. It is likely that many of these cause disorders that were under-represented in our study due to the ease of clinical diagnosis on the basis of distinctive clinical features or targeted diagnostic testing. These ascertainment biases will not impact the representation of novel DDs in our cohort.

Our modelling suggested that likely over 1,000 DD-associated genes remain to be discovered, and that reduced penetrance and pre/perinatal death will reduce our power to identify these genes through DNM enrichment. Identifying these genes will require both improved analytical methods and greater sample sizes. As sample sizes increase, accurate modelling of gene-specific mutation rates becomes more important. In our analyses of 31,058 trios, we observed evidence that mutation rate heterogeneity among genes can lead to over-estimating the statistical significance of mutation enrichment based on an exome-wide mutation model. We advocate the development of more granular mutation rate models, based

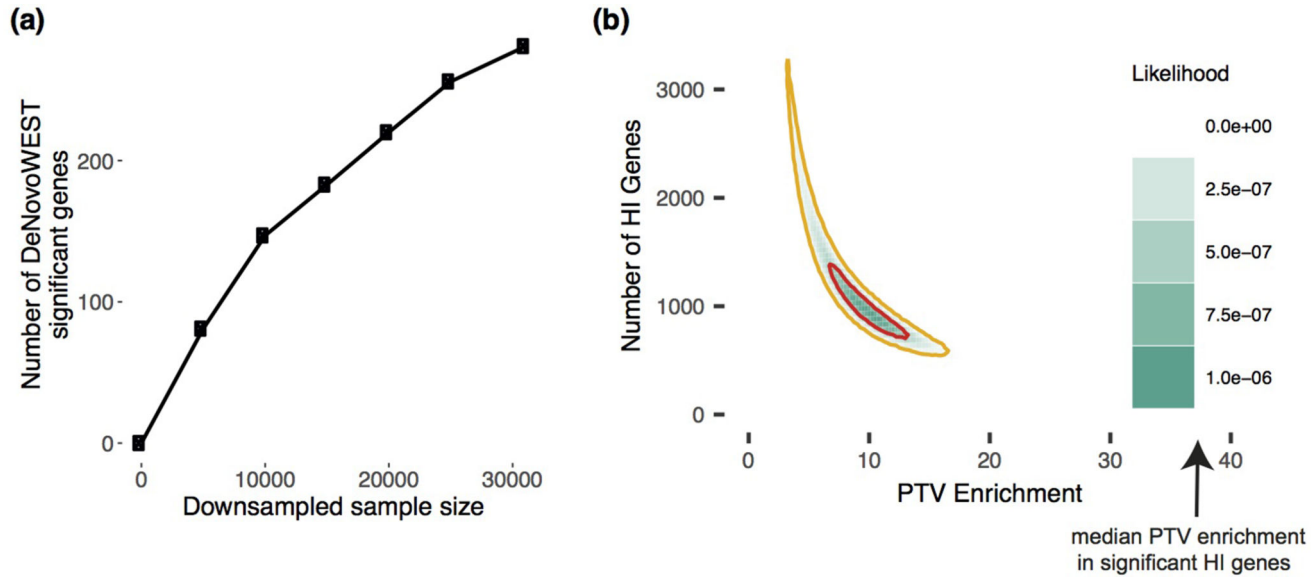
Europe PMC Funders Author Manuscripts

on large-scale population variation resources, that correct for all technical and biological complexities, to ensure that larger studies are robust to mutation rate heterogeneity.

We anticipate that the variant-level weights used by DeNovoWEST will improve over time. As reference population samples, such as gnomAD[6], increase in size, weights based on selective constraint metrics (e.g. s_{het}, regional missense constraint) will improve. Weights could also incorporate more functional information, such as expression in disease-relevant tissues. For example, we observe that DD-associated genes are significantly more likely to be expressed in fetal brain (Supplementary Fig. 14). Furthermore, novel metrics based on gene co-regulation networks can predict whether genes function within a disease-relevant pathway[33]. As a cautionary note, including more functional information may increase power to detect some novel disorders while decreasing power for disorders with pathophysiology different from known disorders. Our analyses also suggest that variant-level weights could be further improved by incorporating other variant prioritisation metrics, such as upweighting variants predicted to impact splicing, variants in particular protein domains, or variants that are somatic driver mutations during tumorigenesis. In developing DeNovoWEST, we explored applying both variant-level weights and gene-level weights in separate stages of the analysis, however, subtle but pervasive correlations between gene-level metrics (e.g. s_{het}) and variant-level metrics (e.g. regional missense constraint, CADD) presents statistical challenges to implementation. Finally, the discovery of less penetrant disorders can be empowered by analytical methodologies that integrate both DNMs and rare inherited variants, such as TADA[34]. Nonetheless, using current methods focused on DNMs alone, we estimated that ~350,000 parent-child trios would need to be analysed to have ~80% power to detect HI genes with a 10-fold PTV enrichment. Discovering non-HI disorders will need even larger sample sizes. Reaching this number of sequenced families will be impossible for an individual research study or clinical centre, therefore it is essential that genetic data generated as part of routine diagnostic practice is shared with the research community such that it can be aggregated to drive discovery of novel disorders and improve diagnostic practice.

Europe PMC Funders Author Manuscripts

Extended Data



Extended Data Figure 1. 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 genes that are significant and known to operate via a loss-of-function mechanism (shown with an arrow) is 39.7.

Extended Data Table 1
Recurrent Mutations.

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

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

Symbol	Chr	Position	Ref	Alt	Consequence	Recur	Likely mechanism	CpG	Somatic Driver Gene	Germline Selection Gene	DD status
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

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Authors

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 consortium authors †
Silvia Borras[14], Caroline Clark[14], John Dean[14], Zosia Miedzybrodzka[14], Alison Ross[14], Stephen Tennant[14], Tabib Dabir[15], Deirdre Donnelly[15], Mervyn Humphreys[15], Alex Magee[15], Vivienne McConnell[15], Shane McKee[15], Susan McNerlan[15], Patrick J. Morrison[15], Gillian Rea[15], Fiona Stewart[15], Trevor Cole[16], Nicola Cooper[16], Lisa Cooper-Charles[16], Helen Cox[16], Lily Islam[16], Joanna Jarvis[16], Rebecca Keelagher[16], Derek Lim[16], Dominic McMullan[16], Jenny Morton[16], Swati Naik[16], Mary O'Driscoll[16], Kai-Ren Ong[16], Deborah Osio[16], Nicola Ragge[16], Sarah Turton[16], Julie Vogt[16], Denise Williams[16], Simon Bodek[17], Alan Donaldson[17], Alison Hills[17], Karen Low[17], Ruth Newbury-Ecob[17], Andrew M. Norman[17], Eileen Roberts[17], Ingrid Scurr[17], Sarah Smithson[17], Madeleine Tooley[17], Steve Abbs[12], Ruth Armstrong[12], Carolyn Dunn[12], Simon Holden[12], Soo-Mi Park[12], Joan Paterson[12], Lucy

Raymond[12], Evan Reid[12], Richard Sandford[12], Ingrid Simonic[12], Marc Tischkowitz[12], Geoff Woods[12], Lisa Bradley[18], Joanne Comerford[18], Andrew Green[18], Sally Lynch[18], Shirley McQuaid[18], Brendan Mullaney[18], Jonathan Berg[19], David Goudie[19], Eleni Mavrak[19], Joanne McLean[19], Catherine McWilliam[19], Eleanor Reavey[19], Tara Azam[13], Elaine Cleary[13], Andrew Jackson[13], Wayne Lam[13], Anne Lampe[13], David Moore[13], Mary Porteous[13], Emma Baple[20], Júlia Baptista[20], Carole Brewer[20], Bruce Castle[20], Emma Kivuva[20], Martina Owens[20], Julia Rankin[20], Charles Shaw-Smith[20], Claire Turner[20], Peter Turnpenny[20], Carolyn Tysoe[20], Therese Bradley[21], Rosemarie Davidson[21], Carol Gardiner[21], Shelagh Joss[21], Esther Kinning[21], Cheryl Longman[21], Ruth McGowan[21], Victoria Murday[21], Daniela Pilz[21], Edward Tobias[21], Margo Whiteford[21], Nicola Williams[21], Angela Barnicoat[22], Emma Clement[22], Francesca Faravelli[22], Jane Hurst[22], Lucy Jenkins[22], Wendy Jones[22], V. K. Ajith Kumar[22], Melissa Lees[22], Sam Loughlin[22], Alison Male[22], Deborah Morrogh[22], Elisabeth Rosser[22], Richard Scott[22], Louise Wilson[22], Ana Beleza[23], Charu Deshpande[23], Frances Flinter[23], Muriel Holder[23], Melita Irving[23], Louise Izatt[23], Dragana Josifova[23], Shehla Mohammed[23], Aneta Molenda[23], Leema Robert[23], Wendy Roworth[23], Deborah Ruddy[23], Mina Ryten[23], Shu Yau[23], Christopher Bennett[24], Moira Blyth[24], Jennifer Campbell[24], Andrea Coates[24], Angus Dobbie[24], Sarah Hewitt[24], Emma Hobson[24], Eilidh Jackson[24], Rosalyn Jewell[24], Alison Kraus[24], Katrina Prescott[24], Eamonn Sheridan[24], Jenny Thomson[24], Kirsty Bradshaw[25], Abhijit Dixit[25], Jacqueline Eason[25], Rebecca Haines[25], Rachel Harrison[25], Stacey Mutch[25], Ajoy Sarkar[25], Claire Searle[25], Nora Shannon[25], Abid Sharif[25], Mohnish Suri[25], Pradeep Vasudevan[26], Natalie Canham[27], Ian Ellis[27], Lynn Greenhalgh[27], Emma Howard[27], Victoria Stinton[27], Andrew Swale[27], Astrid Weber[27], Siddharth Banka[28], Catherine Breen[28], Tracy Briggs[28], Emma Burkitt-Wright[28], Kate Chandler[28], Jill Clayton-Smith[28], Dian Donnai[28], Sofia Douzgou[28], Lorraine Gaunt[28], Elizabeth Jones[28], Bronwyn Kerr[28], Claire Langley[28], Kay Metcalfe[28], Audrey Smith[28], Ronnie Wright[28], David Bourn[29], John Burn[29], Richard Fisher[29], Steve Hellens[29], Alex Henderson[29], Tara Montgomery[29], Miranda Splitt[29], Volker Straub[29], Michael Wright[29], Simon Zwolinski[29], Zoe Allen[30], Birgitta Bernhard[30], Angela Brady[30], Claire Brooks[30], Louise Busby[30], Virginia Clowes[30], Neeti Ghali[30], Susan Holder[30], Rita Ibitoye[30], Emma Wakeling[30], Edward Blair[31], Jenny Carmichael[31], Deirdre Cilliers[31], Susan Clasper[31], Richard Gibbons[31], Usha Kini[31], Tracy Lester[31], Andrea Nemeth[31], Joanna Poulton[31], Sue Price[31], Debbie Shears[31], Helen Stewart[31], Andrew Wilkie[31], Shadi Albaba[32], Duncan Baker[32], Meena Balasubramanian[32], Diana Johnson[32], Michael Parker[32], Oliver Quarrell[32], Alison Stewart[32], Josh Willoughby[32], Charlene Crosby[33], Frances Elmslie[33], Tessa Homfray[33], Huilin Jin[33], Nayana Lahiri[33], Sahar Mansour[33], Karen Marks[33], Meriel McEntagart[33], Anand Saggar[33], Kate Tatton-Brown[33], Rachel Butler[34,35], Angus Clarke[34,35], Sian Corrin[34,35], Andrew Fry[34,35], Arveen Kamath[34,35], Emma McCann[35], Hood Mugalaasi[34,35], Caroline Pottinger[35], Annie Procter[34,35], Julian Sampson[34,35], Francis Sansbury[34,35], Vinod

Varghese[34,35], Diana Baralle[36,37,38], Alison Callaway[36,37,38], Emma J. Cassidy[36,37,38], Stacey Daniels[36,37,38], Andrew Douglas[36,37,38], Nicola Foulds[36,37,38], David Hunt[36,37,38], Mira Kharbanda[36,37,38], Katherine Lachlan[36,37,38], Catherine Mercer[36,37,38], Lucy Side[36,37,38], I. Karen Temple[36,37,38], Diana Wellesley[36,37,38]

[14]North of Scotland Regional Genetics Service, NHS Grampian, Department of Medical Genetics Medical School, Aberdeen, UK [15]Northern Ireland Regional Genetics Centre, Belfast Health and Social Care Trust, Belfast City Hospital, Belfast, UK [16]West Midlands Regional Genetics Service, Birmingham Women's NHS Foundation Trust, Birmingham Women's Hospital, Birmingham, UK [17]Bristol Genetics Service (Avon, Somerset, Gloucs and West Wilts), University Hospitals Bristol NHS Foundation Trust, St Michael's Hospital, Bristol, UK [18]National Centre for Medical Genetics, Our Lady's Children's Hospital, Dublin, Ireland [19]East of Scotland Regional Genetics Service, Human Genetics Unit, Pathology Department, NHS Tayside, Ninewells Hospital, Dundee, UK [20]Peninsula Clinical Genetics Service, Royal Devon and Exeter NHS Foundation Trust, Clinical Genetics Department, Royal Devon & Exeter Hospital (Heavitree), Exeter, UK [21]West of Scotland Regional Genetics Service, NHS Greater Glasgow and Clyde, Institute of Medical Genetics, Yorkhill Hospital, Glasgow, UK [22]North East Thames Regional Genetics Service, Great Ormond Street Hospital for Children NHS Foundation Trust, Great Ormond Street Hospital, London, UK [23]South East Thames Regional Genetics Centre, Guy's and St Thomas' NHS Foundation Trust, Guy's Hospital, London, UK [24]Yorkshire Regional Genetics Service, Leeds Teaching Hospitals NHS Trust, Department of Clinical Genetics, Chapel Allerton Hospital, Leeds, UK [25]Nottingham Regional Genetics Service, City Hospital Campus, Nottingham University Hospitals NHS Trust, Nottingham, UK [26]Leicestershire Genetics Centre, University Hospitals of Leicester NHS Trust, Leicester Royal Infirmary (NHS Trust), Leicester, UK [27]Merseyside and Cheshire Genetics Service, Liverpool Women's NHS Foundation Trust, Department of Clinical Genetics, Royal Liverpool Children's Hospital Alder Hey, Liverpool, UK [28]Manchester Centre for Genomic Medicine, St Mary's Hospital, Central Manchester University Hospitals NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester, UK [30]Northern Genetics Service, Newcastle upon Tyne Hospitals NHS Foundation Trust, Institute of Human Genetics, International Centre for Life, Newcastle upon Tyne, UK [30]North West Thames Regional Genetics Centre, North West London Hospitals NHS Trust, The Kennedy Galton Centre and St Mark's NHS Trust Watford Road, Harrow, UK [31]Oxford Regional Genetics Service, Oxford Radcliffe Hospitals NHS Trust, Oxford, UK [32]Sheffield Regional Genetics Services, Sheffield Children's NHS Trust, Sheffield, UK [33]South West Thames Regional Genetics Centre, St George's Healthcare NHS Trust, St George's, University of London, London, UK [34]Institute of Medical Genetics, University Hospital of Wales, Cardiff, UK [35]Department of Clinical Genetics, Glan Clwyd Hospital, Rhyl, UK [36]Wessex

Clinical Genetics Service, University Hospital Southampton, Princess Anne Hospital, Southampton, UK [37]Wessex Regional Genetics Laboratory, Salisbury NHS Foundation Trust, Salisbury District Hospital, Salisbury, UK [38]Faculty of Medicine, University of Southampton, Southampton, UK

, **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,#]

Silvia Borras[14], Caroline Clark[14], John Dean[14], Zosia Miedzybrodzka[14], Alison Ross[14], Stephen Tennant[14], Tabib Dabir[15], Deirdre Donnelly[15], Mervyn Humphreys[15], Alex Magee[15], Vivienne McConnell[15], Shane McKee[15], Susan McNerlan[15], Patrick J. Morrison[15], Gillian Rea[15], Fiona Stewart[15], Trevor Cole[16], Nicola Cooper[16], Lisa Cooper-Charles[16], Helen Cox[16], Lily Islam[16], Joanna Jarvis[16], Rebecca Keelagher[16], Derek Lim[16], Dominic McMullan[16], Jenny Morton[16], Swati Naik[16], Mary O'Driscoll[16], Kai-Ren Ong[16], Deborah Osio[16], Nicola Ragge[16], Sarah Turton[16], Julie Vogt[16], Denise Williams[16], Simon Bodek[17], Alan Donaldson[17], Alison Hills[17], Karen Low[17], Ruth Newbury-Ecob[17], Andrew M. Norman[17], Eileen Roberts[17], Ingrid Scurr[17], Sarah Smithson[17], Madeleine Tooley[17], Steve Abbs[12], Ruth Armstrong[12], Carolyn Dunn[12], Simon Holden[12], Soo-Mi Park[12], Joan Paterson[12], Lucy Raymond[12], Evan Reid[12], Richard Sandford[12], Ingrid Simonic[12], Marc Tischkowitz[12], Geoff Woods[12], Lisa Bradley[18], Joanne Comerford[18], Andrew Green[18], Sally Lynch[18], Shirley McQuaid[18], Brendan Mullaney[18], Jonathan Berg[19], David Goudie[19], Eleni Mavrak[19], Joanne McLean[19], Catherine McWilliam[19], Eleanor Reavey[19], Tara Azam[13], Elaine Cleary[13], Andrew Jackson[13], Wayne Lam[13], Anne Lampe[13], David Moore[13], Mary Porteous[13], Emma Baple[20], Júlia Baptista[20], Carole Brewer[20], Bruce Castle[20], Emma Kivuva[20], Martina Owens[20], Julia Rankin[20], Charles Shaw-Smith[20], Claire Turner[20], Peter Turnpenny[20], Carolyn Tysoe[20], Therese Bradley[21], Rosemarie Davidson[21], Carol Gardiner[21], Shelagh Joss[21], Esther Kinning[21], Cheryl Longman[21], Ruth McGowan[21], Victoria Murday[21], Daniela Pilz[21], Edward Tobias[21], Margo Whiteford[21], Nicola Williams[21], Angela Barnicoat[22], Emma Clement[22], Francesca Faravelli[22], Jane Hurst[22], Lucy Jenkins[22], Wendy Jones[22], V. K. Ajith Kumar[22], Melissa Lees[22], Sam Loughlin[22], Alison Male[22], Deborah Morrogh[22], Elisabeth Rosser[22], Richard Scott[22], Louise Wilson[22], Ana Beleza[23], Charu Deshpande[23], Frances Flinter[23], Muriel Holder[23], Melita Irving[23], Louise Izatt[23], Dragana Josifova[23], Shehla Mohammed[23], Aneta Molenda[23], Leema Robert[23], Wendy Roworth[23], Deborah Ruddy[23], Mina Ryten[23], Shu Yau[23], Christopher Bennett[24], Moira Blyth[24], Jennifer Campbell[24], Andrea Coates[24], Angus Dobbie[24], Sarah Hewitt[24], Emma Hobson[24], Eilidh Jackson[24], Rosalyn Jewell[24], Alison Kraus[24], Katrina Prescott[24], Eamonn Sheridan[24], Jenny Thomson[24], Kirsty Bradshaw[25], Abhijit Dixit[25], Jacqueline Eason[25], Rebecca Haines[25], Rachel Harrison[25], Stacey Mutch[25], Ajoy Sarkar[25], Claire Searle[25], Nora Shannon[25], Abid Sharif[25], Mohnish Suri[25], Pradeep Vasudevan[26], Natalie Canham[27], Ian Ellis[27], Lynn Greenhalgh[27], Emma Howard[27], Victoria Stinton[27], Andrew Swale[27], Astrid Weber[27], Siddharth Banka[28], Catherine Breen[28], Tracy

Briggs[28], Emma Burkitt-Wright[28], Kate Chandler[28], Jill Clayton-Smith[28], Dian Donnai[28], Sofia Douzgou[28], Lorraine Gaunt[28], Elizabeth Jones[28], Bronwyn Kerr[28], Claire Langley[28], Kay Metcalfe[28], Audrey Smith[28], Ronnie Wright[28], David Bourn[29], John Burn[29], Richard Fisher[29], Steve Hellens[29], Alex Henderson[29], Tara Montgomery[29], Miranda Splitt[29], Volker Straub[29], Michael Wright[29], Simon Zwolinski[29], Zoe Allen[30], Birgitta Bernhard[30], Angela Brady[30], Claire Brooks[30], Louise Busby[30], Virginia Clowes[30], Neeti Ghali[30], Susan Holder[30], Rita Ibitoye[30], Emma Wakeling[30], Edward Blair[31], Jenny Carmichael[31], Deirdre Cilliers[31], Susan Clasper[31], Richard Gibbons[31], Usha Kini[31], Tracy Lester[31], Andrea Nemeth[31], Joanna Poulton[31], Sue Price[31], Debbie Shears[31], Helen Stewart[31], Andrew Wilkie[31], Shadi Albaba[32], Duncan Baker[32], Meena Balasubramanian[32], Diana Johnson[32], Michael Parker[32], Oliver Quarrell[32], Alison Stewart[32], Josh Willoughby[32], Charlene Crosby[33], Frances Elmslie[33], Tessa Homfray[33], Huilin Jin[33], Nayana Lahiri[33], Sahar Mansour[33], Karen Marks[33], Meriel McEntagart[33], Anand Saggar[33], Kate Tatton-Brown[33], Rachel Butler[34,35], Angus Clarke[34,35], Sian Corrin[34,35], Andrew Fry[34,35], Arveen Kamath[34,35], Emma McCann[35], Hood Mugalaasi[34,35], Caroline Pottinger[35], Annie Procter[34,35], Julian Sampson[34,35], Francis Sansbury[34,35], Vinod Varghese[34,35], Diana Baralle[36,37,38], Alison Callaway[36,37,38], Emma J. Cassidy[36,37,38], Stacey Daniels[36,37,38], Andrew Douglas[36,37,38], Nicola Foulds[36,37,38], David Hunt[36,37,38], Mira Kharbanda[36,37,38], Katherine Lachlan[36,37,38], Catherine Mercer[36,37,38], Lucy Side[36,37,38], I. Karen Temple[36,37,38], Diana Wellesley[36,37,38]

Affiliations

[14]North of Scotland Regional Genetics Service, NHS Grampian, Department of Medical Genetics Medical School, Aberdeen, UK [15]Northern Ireland Regional Genetics Centre, Belfast Health and Social Care Trust, Belfast City Hospital, Belfast, UK [16]West Midlands Regional Genetics Service, Birmingham Women's NHS Foundation Trust, Birmingham Women's Hospital, Birmingham, UK [17]Bristol Genetics Service (Avon, Somerset, Gloucs and West Wilts), University Hospitals Bristol NHS Foundation Trust, St Michael's Hospital, Bristol, UK [18]National Centre for Medical Genetics, Our Lady's Children's Hospital, Dublin, Ireland [19]East of Scotland Regional Genetics Service, Human Genetics Unit, Pathology Department, NHS Tayside, Ninewells Hospital, Dundee, UK [20]Peninsula Clinical Genetics Service, Royal Devon and Exeter NHS Foundation Trust, Clinical Genetics Department, Royal Devon & Exeter Hospital (Heavitree), Exeter, UK [21]West of Scotland Regional Genetics Service, NHS Greater Glasgow and Clyde, Institute of Medical Genetics, Yorkhill Hospital, Glasgow, UK [22]North East Thames Regional Genetics Service, Great Ormond Street Hospital for Children NHS Foundation Trust, Great Ormond Street Hospital, London, UK [23]South East Thames Regional Genetics Centre, Guy's and St Thomas' NHS Foundation Trust, Guy's Hospital, London, UK [24]Yorkshire Regional Genetics Service, Leeds Teaching Hospitals NHS Trust, Department of Clinical Genetics, Chapel Allerton Hospital, Leeds, UK [25]Nottingham Regional Genetics Service, City Hospital Campus, Nottingham University Hospitals NHS Trust, Nottingham, UK [26]Leicestershire Genetics Centre,

University Hospitals of Leicester NHS Trust, Leicester Royal Infirmary (NHS Trust), Leicester, UK [27]Merseyside and Cheshire Genetics Service, Liverpool Women's NHS Foundation Trust, Department of Clinical Genetics, Royal Liverpool Children's Hospital Alder Hey, Liverpool, UK [28]Manchester Centre for Genomic Medicine, St Mary's Hospital, Central Manchester University Hospitals NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester, UK [30]Northern Genetics Service, Newcastle upon Tyne Hospitals NHS Foundation Trust, Institute of Human Genetics, International Centre for Life, Newcastle upon Tyne, UK [30]North West Thames Regional Genetics Centre, North West London Hospitals NHS Trust, The Kennedy Galton Centre and St Mark's NHS Trust Watford Road, Harrow, UK [31]Oxford Regional Genetics Service, Oxford Radcliffe Hospitals NHS Trust, Oxford, UK [32]Sheffield Regional Genetics Services, Sheffield Children's NHS Trust, Sheffield, UK [33]South West Thames Regional Genetics Centre, St George's Healthcare NHS Trust, St George's, University of London, London, UK [34]Institute of Medical Genetics, University Hospital of Wales, Cardiff, UK [35]Department of Clinical Genetics, Glan Clwyd Hospital, Rhyl, UK [36]Wessex Clinical Genetics Service, University Hospital Southampton, Princess Anne Hospital, Southampton, UK [37]Wessex Regional Genetics Laboratory, Salisbury NHS Foundation Trust, Salisbury District Hospital, Salisbury, UK [38]Faculty of Medicine, University of Southampton, Southampton, UK

[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 [8]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]East Anglian Medical Genetics Service, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK [13]MRC Human Genetics Unit, MRC IGMM, University of Edinburgh, Western General Hospital, Edinburgh, UK

Acknowledgements

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

Data Availability

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

RadboudUMC sequence and variant level data cannot be made available through EGA due to the nature of consent for clinical testing. To access the data, please contact christian.gilissen@radboudumc.nl with a request. Data sharing will be dependent on patient consent, diagnostic status of the patient, type of request, and the potential benefit to the patient. GeneDx data cannot be made available through EGA due to the nature of consent for clinical testing. GeneDx-referred patients are consented for aggregate, de-identified research and subject to US HIPAA privacy protection. As such, we are not able to share patient-level BAM or VCF data, which is potentially identifiable without a HIPAA Business Associate Agreement. Access to the de-identified aggregate data used in this analysis is available upon request to GeneDx. GeneDx has contributed deidentified data to this study to improve clinical interpretation of genomic data, in accordance with patient consent and in conformance with the ACMG position statement on genomic data sharing (see Supplementary Note for details). Clinically interpreted variants and associated phenotypes from the DDD study are available through DECIPHER (https://decipher.sanger.ac.uk)

Clinically interpreted variants from RUMC are available from the Dutch national initiative for sharing variant classifications (https://www.vkgl.nl/nl/diagnostiek/vkgl-datashare-database) as well as LOVD (https://databases.lovd.nl/shared/variants), where they are listed with "VKGL-NL_Nijmegen" as the owner

Clinically interpreted variants from GeneDx are deposited in ClinVar (https://www.ncbi.nlm.nih.gov/clinvar). GeneDx's submitter ID is 26957 (https://www.ncbi.nlm.nih.gov/clinvar/submitters/26957/)

Genome Aggregation Database (gnomAD v2.1.1; https://gnomad.broadinstitute.org/)

The Cancer Genome Atlas (TCGA; https://portal.gdc.cancer.gov)

Developmental Disorders Genotype-Phenotype Database (DDG2P; https://www.ebi.ac.uk/gene2phenotype/downloads)

Code Availability

The DeNovoWEST method is available on GitHub (https://github.com/queenjobo/DeNovoWEST) along with code to recreate all figures in the manuscript. DOI: 0.5281/zenodo.3909398. Code to run the Phenopy method is also available on GitHub (https://github.com/GeneDx/phenopy).

References

1. Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. Nature. 2017; 542:433–438. [PubMed: 28135719]
2. Martin HC, et al. Quantifying the contribution of recessive coding variation to developmental disorders. Science. 2018; 362:1161–1164. [PubMed: 30409806]
3. Kircher M, et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014; 46:310–315. [PubMed: 24487276]
4. Samocha KE, et al. Regional missense constraint improves variant deleteriousness prediction. bioRxiv. 2017; doi: 10.1101/148353
5. Kosmicki JA, et al. Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. Nat Genet. 2017; 49:504–510. [PubMed: 28191890]
6. Karczewski KJ, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020; 581:434–443. [PubMed: 32461654]
7. Lek M, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016; 536:285–291. [PubMed: 27535533]
8. Cooper GM, et al. A copy number variation morbidity map of developmental delay. Nat Genet. 2011; 43:838–846. [PubMed: 21841781]
9. Coe BP, et al. Refining analyses of copy number variation identifies specific genes associated with developmental delay. Nat Genet. 2014; 46:1063–1071. [PubMed: 25217958]
10. Robinson JT, et al. Integrative genomics viewer. Nature Biotechnology. 2011; 29:24–26.
11. Villegas F, et al. Lysosomal Signaling Licenses Embryonic Stem Cell Differentiation via Inactivation of Tfe3. Cell Stem Cell. 2019; 24:257–270.e8. [PubMed: 30595499]
12. Diaz J, Berger S, Leon E. TFE3-associated neurodevelopmental disorder: A distinct recognizable syndrome. Am J Med Genet A. 2020; 182:584–590. [PubMed: 31833172]
13. Jaganathan K, et al. Predicting Splicing from Primary Sequence with Deep Learning. Cell. 2019; 176:535–548.e24. [PubMed: 30661751]
14. Yilmaz R, et al. A recurrent synonymous KAT6B mutation causes Say-Barber- Biesecker/Young-Simpson syndrome by inducing aberrant splicing. Am J Med Genet A. 2015; 167A:3006–3010. [PubMed: 26334766]
15. Wu X, Pang E, Lin K, Pei Z-M. Improving the measurement of semantic similarity between gene ontology terms and gene products: insights from an edge- and IC-based hybrid method. PLoS One. 2013; 8
16. Catterall WA, Dib-Hajj S, Meisler MH, Pietrobon D. Inherited neuronal ion channelopathies: new windows on complex neurological diseases. J Neurosci. 2008; 28:11768–11777. [PubMed: 19005038]
17. Lasser M, Tiber J, Lowery LA. The Role of the Microtubule Cytoskeleton in Neurodevelopmental Disorders. Front Cell Neurosci. 2018; 12:165. [PubMed: 29962938]
18. Hamilton MJ, et al. Heterozygous mutations affecting the protein kinase domain of cause a syndromic form of developmental delay and intellectual disability. J Med Genet. 2018; 55:28–38. [PubMed: 29021403]
19. Martincorena I, et al. Universal Patterns of Selection in Cancer and Somatic Tissues. Cell. 2018; 173:1823.

20. Qi H, Dong C, Chung WK, Wang K, Shen Y. Deep Genetic Connection Between Cancer and Developmental Disorders. Hum Mutat. 2016; 37:1042–1050. [PubMed: 27363847]
21. Ronan JL, Wu W, Crabtree GR. From neural development to cognition: unexpected roles for chromatin. Nat Rev Genet. 2013; 14:347–359. [PubMed: 23568486]
22. Cancer Genome Atlas Research Network *et al*. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013; 45:1113–1120. [PubMed: 24071849]
23. Goriely A, Wilkie AOM. Paternal age effect mutations and selfish spermatogonial selection: causes and consequences for human disease. Am J Hum Genet. 2012; 90:175–200. [PubMed: 22325359]
24. Duncan BK, Miller JH. Mutagenic deamination of cytosine residues in DNA. Nature. 1980; 287:560–561. [PubMed: 6999365]
25. Maher GJ, et al. Visualizing the origins of selfish de novo mutations in individual seminiferous tubules of human testes. Proc Natl Acad Sci U S A. 2016; 113:2454–2459. [PubMed: 26858415]
26. Maher GJ, et al. Selfish mutations dysregulating RAS-MAPK signaling are pervasive in aged human testes. Genome Res. 2018; 28:1779–1790. [PubMed: 30355600]
27. Young LC, et al. SHOC2-MRAS-PP1 complex positively regulates RAF activity and contributes to Noonan syndrome pathogenesis. Proc Natl Acad Sci U S A. 2018; 115:E10576–E10585. [PubMed: 30348783]
28. Coe BP, et al. Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. Nat Genet. 2019; 51:106–116. [PubMed: 30559488]
29. Lord J, et al. Prenatal exome sequencing analysis in fetal structural anomalies detected by ultrasonography (PAGE): a cohort study. Lancet. 2019; 393:747–757. [PubMed: 30712880]
30. Cassa CA, et al. Estimating the selective effects of heterozygous protein-truncating variants from human exome data. Nat Genet. 2017; 49:806–810. [PubMed: 28369035]
31. Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. Nature. 2015; 519:223–228. [PubMed: 25533962]
32. Satterstrom FK, et al. Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. Cell. 2020; 180:568–584.e23. [PubMed: 31981491]
33. Deelen P, et al. Improving the diagnostic yield of exome- sequencing by predicting gene-phenotype associations using large-scale gene expression analysis. Nat Commun. 2019; 10:2837. [PubMed: 31253775]
34. He X, et al. Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. PLoS Genet. 2013; 9

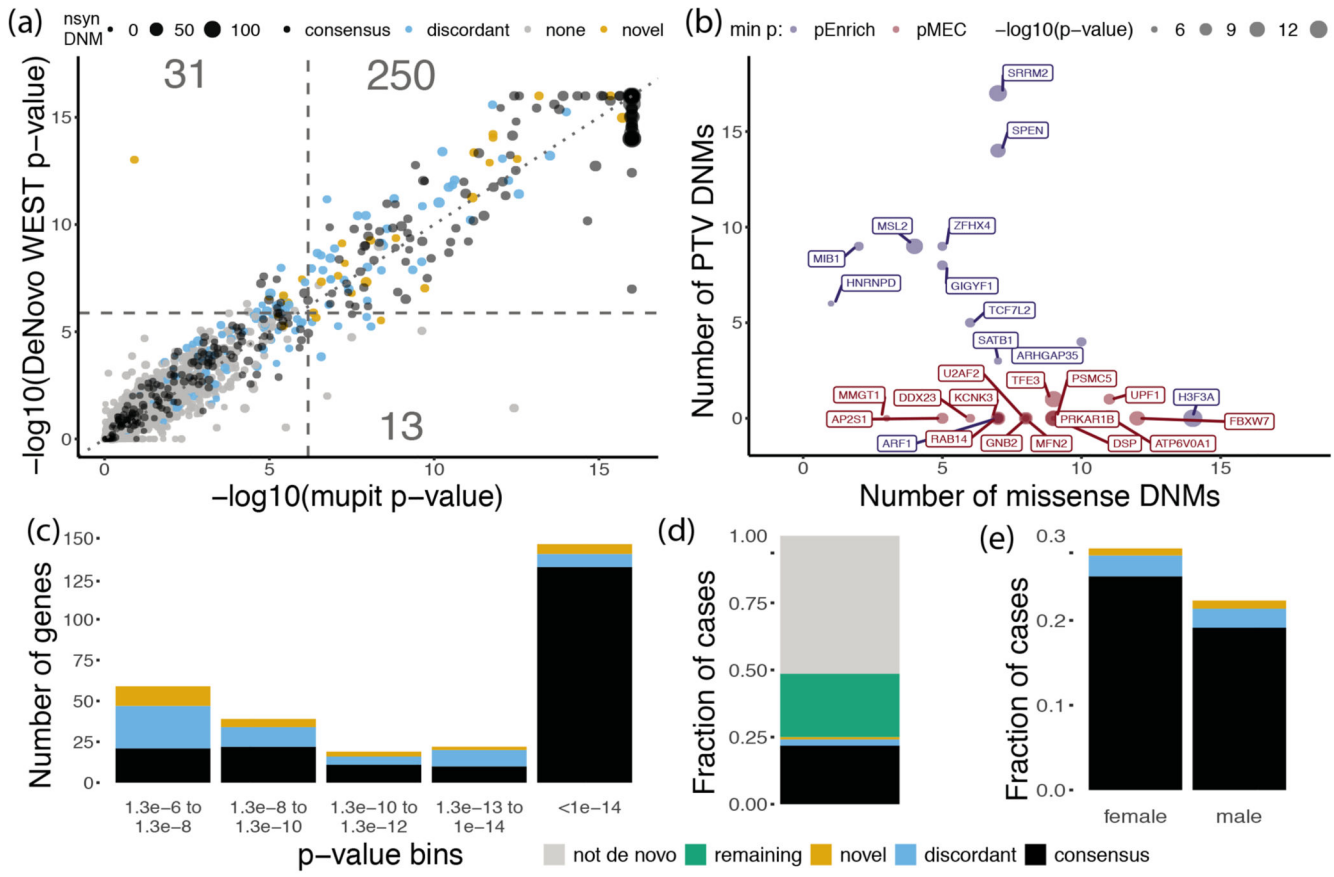


Figure 1. Results of DeNovoWEST analysis.

(a) Comparison of p-values using the new method (DeNovoWEST) versus the previous method (mupit)[1], run on the full cohort. Dashed lines indicate the threshold for genome-wide significance (one sided, Bonferroni correction). Point size is proportional to the number of nonsynonymous DNMs in our cohort (nsyn). The number of genes that fall into each quadrant are annotated. (b) The number of missense and PTV DNMs in the novel genes. Point size is proportional to the $\log_{10}$(-p-value) from analysis of the undiagnosed subset. Point colour corresponds to which test p-value was more significant: non-synonymous enrichment test in blue (pEnrich), missense enrichment and clustering test in red (pMEC). (c) The distribution of significant p-values from analysis of the undiagnosed subset for discordant and novel genes; p-values for consensus genes come from the full cohort analysis. The number of genes in each p-value bin is coloured by diagnostic gene group (n = 285 significant genes; one-sided p-values, Bonferroni corrected). Green represents the remaining fraction of cases expected to have a pathogenic *de novo* coding mutation and grey is the fraction of cases that are likely to be explained by other factors. (d) The fraction of cases (n = 31,058) 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 (n = 13,636 female and 17,422 male). In all panels, black, blue and orange represents consensus, discordant and novel genes respectively.

Europe PMC Funders Author Manuscripts

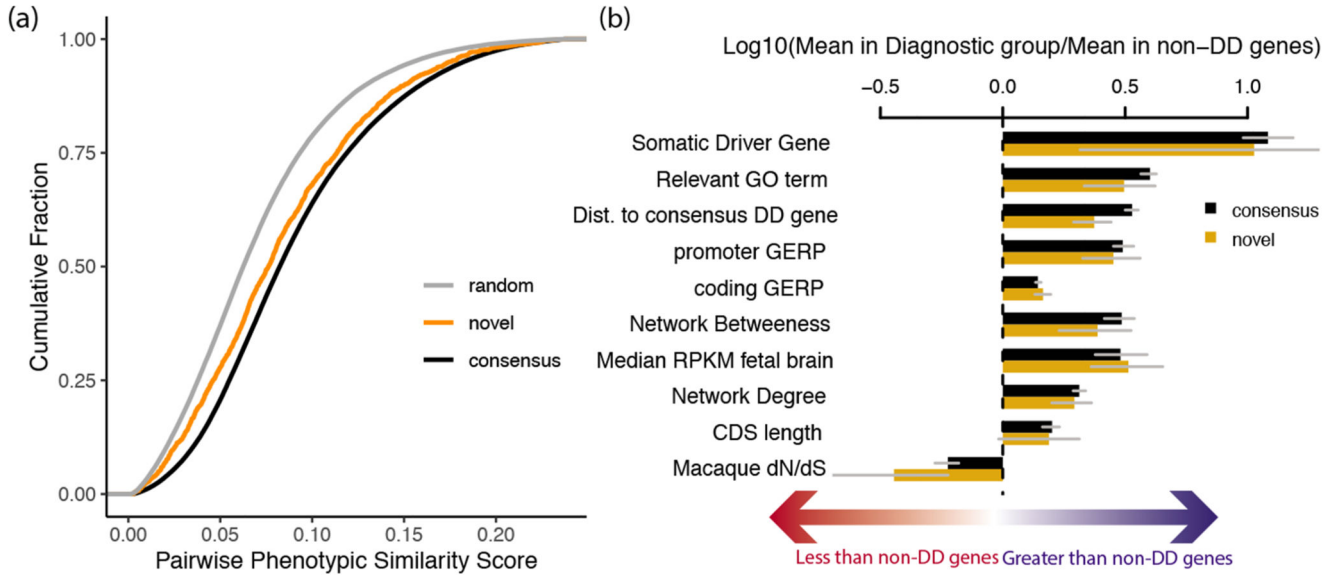


Figure 2. Properties of novel genes.

(a) The phenotypic similarity of patients with DNMs in novel and consensus genes. Random phenotypic similarity was calculated from random pairs of patients. Cases with DNMs in the same novel gene were less phenotypically similar than cases with DNMs in the same consensus gene ($p = 2.3 \times 10^{-11}$, two-sided Wilcoxon rank-sum test). (b) Comparison of properties of consensus (n = 380) and novel (n = 28) DD genes known to be differential between consensus and non-DD genes (95% bootstrapped confidence intervals shown).

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

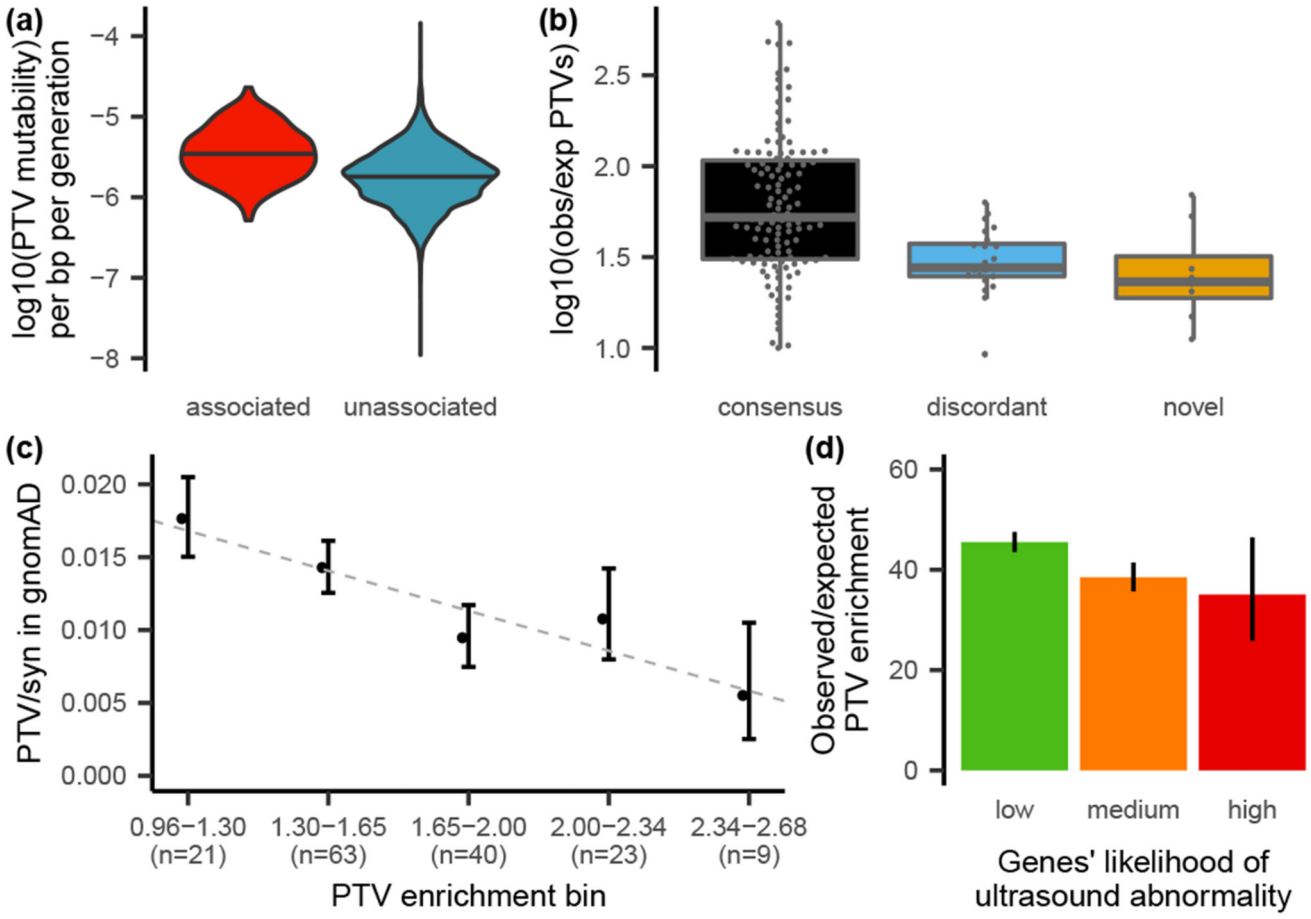


Figure 3. Factors influencing power.

(a) PTV mutability is significantly lower ($p = 4.6 \times 10^{-68}$, two-sided Wilcox rank sum test) in genes that are not significantly DD-associated (blue) than in DD-associated genes (red). Median depicted with a black horizontal line. (b) Distribution of PTV enrichment in significant, likely haploinsufficient, genes by category (118 consensus, 23 discordant, 8 novel genes). Lower and upper hinges correspond to first and third quantiles. Median depicted by a horizontal grey line. The upper and lower whiskers extend 1.5 times the inter-quartile range. (c) Comparison of PTV enrichment in our cohort vs the PTV to synonymous ratio in gnomAD, for genes that are significantly PTV-enriched in our cohort (without variant weighting; n = 156 genes). PTV enrichment bins labelled with log10(enrichment). Dashed line indicates regression. Confidence intervals are 95% of the rate ratio. (d) Overall PTV enrichment across genes grouped by likelihood of presenting with a structural malformation on prenatal ultrasound (145 low, 65 medium, 6 low genes). PTV enrichment is significantly higher for genes with a low likelihood compared to other genes ($p = 4.6 \times 10^{-5}$, two-sided Poisson test). Poisson 95% confidence intervals shown.