
Article:
Martin, H.C., Jones, W.D., McIntyre, R. et al. (37 more authors) (2018) Quantifying the contribution of recessive coding variation to developmental disorders. Science, 362 (6419). pp. 1161-1164. ISSN 0036-8075

https://doi.org/10.1126/science.aar6731

Quantifying the contribution of recessive coding variation to developmental disorders

Short title: Recessive coding causes of developmental disorders

One Sentence Summary: Recessive coding variants explain a low fraction of undiagnosed developmental disorder patients.

Hilary C. Martin[1,*], Wendy D. Jones[1,2], James D. Stephenson[1,3], Juliet Handsaker[1], Giuseppe Gallone[1], Rebecca McIntyre[1], Michaela Bruntraeger[1], Jeremy F. McRae[1], Elena Prigmore[1], Patrick Short[1], Mari Niemi[1], Joanna Kaplanis[1], Elizabeth Radford[1,4], Nadia Akawi[5], Meena Balasubramanian[6], John Dean[7], Rachel Horton[8], Alice Hulbert[9], Diana S. Johnson[6], Katie Johnson[10], Dhavendra Kumar[11], Sally Ann Lynch[12], Sarju G. Mehta[13], Jenny Morton[14], Michael J. Parker[15], Miranda Splitt[16], Peter D Turnpenny[17], Pradeep C. Vasudevan[18], Michael Wright[16], Andrew Bassett[1], Caroline F. Wright[19], David R. FitzPatrick[20], Helen V. Firth[1,13], Matthew E. Hurles[1], Jeffrey C. Barrett[1,*] on behalf of the DDD Study

1. Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, U.K.
2. Great Ormond Street Hospital for Children, NHS Foundation Trust, Great Ormond Street Hospital, Great Ormond Street, London WC1N 3JH, UK.
3. European Molecular Biology Laboratory–European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK.
4. Department of Paediatrics, Cambridge University Hospitals NHS Foundation Trust, Cambridge, U.K.
5. Division of Cardiovascular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, U.K.
6. Sheffield Clinical Genetics Service, Sheffield Children's NHS Foundation Trust, OPD2, Northern General Hospital, Herries Rd, Sheffield, S5 7AU, U.K.
7. Department of Genetics, Aberdeen Royal Infirmary, Aberdeen, U.K.
8. Wessex Clinical Genetics Service, G Level, Princess Anne Hospital, Coxford Road, Southampton, SO16 5YA.
9. Cheshire and Merseyside Clinical Genetic Service, Liverpool Women's NHS Foundation Trust, Crown Street, Liverpool, L8 7SS, U.K.
10. Department of Clinical Genetics, City Hospital Campus, Hucknall Road, Nottingham, NG5 1PB, U.K.
11. Institute of Cancer and Genetics, University Hospital of Wales, Cardiff, U.K.
12. Temple Street Children's Hospital, Dublin, Ireland.
13. Department of Clinical Genetics, Cambridge University Hospitals NHS Foundation Trust, Cambridge, U.K.
14. Clinical Genetics Unit, Birmingham Women's Hospital, Edgbaston, Birmingham, B15 2TG, U.K.
15. Sheffield Clinical Genetics Service, Sheffield Children's Hospital, Western Bank, Sheffield, S10 2TH, U.K.
16. Northern Genetics Service, Newcastle upon Tyne Hospitals, NHS Foundation Trust
17. Clinical Genetics, Royal Devon & Exeter NHS Foundation Trust, Exeter, U.K.
18. Department of Clinical Genetics, University Hospitals of Leicester NHS Trust, Leicester Royal Infirmary, Leicester, LE1 5WW
19. University of Exeter Medical School, Institute of Biomedical and Clinical Science, RILD, Royal Devon & Exeter Hospital, Barrack Road, Exeter, EX2 5DW, U.K.
20. MRC Human Genetics Unit, MRC IGMM, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, U.K.

* Corresponding authors hcm@sanger.ac.uk and barrett@sanger.ac.uk

Large-scale sequencing can help uncover the genetic architecture of rare diseases. We estimated the genome-wide contribution of recessive coding variation from 6,040 exome-sequenced families from the Deciphering Developmental Disorders study. The proportion of cases attributable to recessive coding variants was 3.6% in patients of European ancestry, compared to 50% explained by *de novo* coding mutations. It was higher (31%) in patients with Pakistani ancestry, due to elevated autozygosity. Half of this recessive burden is attributable to known genes. We identified two genes not previously associated with recessive developmental disorders, *EIF3F* and *KDM5B*. Our results suggest that recessive coding variants account for a small fraction of currently undiagnosed individuals, and that the role of pathogenic noncoding variants, incomplete penetrance, and polygenic mechanisms need further exploration.

Genetic studies of rare diseases traditionally followed a phenotype-driven search for a shared genetic diagnosis in multiple individuals with a clinically similar presentation. Large-scale sequencing studies of more phenotypically heterogeneous patients have inverted this process, and have demonstrated the power of unbiased, genotype-first discovery of new disease genes (*1*–*3*). It is also possible to use the same datasets to characterise the overall genetic architecture of such disorders. For example, analyses of one such cohort, the Deciphering Developmental Disorders study (DDD) (*4*), discovered new dominant disease genes (*1*), estimated the fraction of patients with a causal *de novo* mutation in both known and as-yet undiscovered dominant genes (40-45%), and made predictions about the prevalence of such disorders (*5*).

Extending our analyses to other modes of inheritance may help design gene discovery studies and inform counselling about recurrence risk. It has been posited that there are thousands of recessive genes yet to be discovered (*6*, *7*), which could imply that recessive genes explain a large fraction of undiagnosed rare disease cases. However, attempts to estimate the prevalence of recessive disorders have been restricted to known disorders (*8*) or known pathogenic alleles (*9*). There has been no systematic attempt to quantify the overall recessive burden using large-scale sequencing data and a robust statistical and computational genetic framework.

We describe an analysis of autosomal recessive coding variants in 6,040 exome-sequenced trios from the British Isles, recruited as part of the DDD study. We use a probabilistic method for identifying robust new recessive genes (*10*), in contrast to the heuristic filtering methods commonly applied (e.g. (*11–13*)). We extend this to estimate the overall burden of recessive causation in this cohort, and compare it between patients with different ancestries. Our approach overcomes drawbacks of previously published methods (*3*, *10*) that do not provide well-calibrated estimates of the exome-wide burden of recessive disease.

Results

Genome-wide recessive burden

We hypothesized there should be a burden of biallelic genotypes predicted either to cause loss-of-function (LoF) or functional impairment to a protein. For each of three possible genotype configurations (LoF on both alleles, damaging missense on both alleles, or one on each allele), we compared the number of observed rare (minor allele frequency, MAF, <1%) biallelic genotypes in our cohort to the number expected by chance given the population frequency of such variants and the gene-specific fraction of autozygous individuals (*14*). Because the expected number is sensitive to inaccuracy in population frequency estimates of very rare variants in broadly-defined ancestry groups, we focused our analysis on the largest two subsets of the cohort with homogenous ancestry (Fig. S1), corresponding in a principal components analysis using 1000 Genomes to Great British individuals and Punjabis from Lahore, Pakistan (Fig. S2). We refer to these subsets as having European Ancestry or Pakistani Ancestry from the British Isles (EABI, PABI).

We evaluated three methods for calculating the expected number of biallelic genotypes, using synonymous variants as a control. Firstly, we used the non-Finnish Europeans and South Asians from the Exome Aggregation Consortium (ExAC) (*15*) to estimate the allele frequencies for EABI and PABI respectively, as described previously (*10*). However, we found that the total observed number of biallelic synonymous genotypes was lower than the expected number (Fig.

S3). This is due to a combination of differences in sequence coverage, quality control, and ancestry between DDD and ExAC, and the lack of phased, individual-specific data in ExAC needed to avoid double-counting variants on the same haplotype within a gene. Secondly, we considered an approach (*3*) that uses per-gene mutation rates. While it was well-calibrated for individual genes, this method produced a significant underestimate of the total expected number of synonymous biallelic genotypes (Fig. S3). Finally, we used the phased haplotypes from unaffected DDD parents to estimate the expected number of biallelic genotypes (*14*). With this method, the number of observed biallelic synonymous genotypes closely matched what we would expect by chance (ratio=0.997 for EABI and 1.003 for PABI; Poisson p=0.6 and 0.4) (Fig. 1A), so we used it henceforth.

We observed no significant burden of biallelic genotypes of any consequence class in 1,389 probands with a likely diagnostic *de novo* mutation, inherited dominant variant or X-linked variant, consistent with those probands' phenotypes being fully explained by previously identified variants. We therefore evaluated the recessive coding burden in 4,318 EABI and 333 PABI probands whom we deemed more likely to have a recessive cause of their disorder because they did not have a likely dominant or X-linked diagnosis (*4*), or had at least one affected sibling, or >2% autozygosity. As expected due to their higher autozygosity (Fig. S4), PABI individuals had more rare biallelic genotypes than EABI individuals (Fig. 1). Ninety-two percent of the likely damaging rare biallelic genotypes observed in PABI samples were homozygous (rather than compound heterozygous), versus only 28% for the EABI samples. We observed a significant enrichment of biallelic LoF genotypes above chance expectation in both the EABI and PABI groups (~1.4-fold enrichment in each; Poisson $p=3.5\times10^{-5}$ for EABI, $p=9.7\times10^{-7}$ for PABI). We also observed a smaller enrichment of biallelic damaging missense genotypes which was nominally significant in the EABI group (Poisson p=0.03), as well as a significant enrichment of compound heterozygous LoF/damaging missense genotypes in the EABI group (1.4-fold enrichment; Poisson $p=6\times10^{-7}$). In the EABI group, the enrichments became stronger and more significant at lower MAF, but the absolute number of excess variants fell slightly (Fig. S5). Thus, plausibly pathogenic variants are concentrated at rarer MAF, but some do rise to higher frequencies.

We found that particular gene sets showed a higher burden of damaging biallelic genotypes amongst the 4,318 EABI and 333 PABI undiagnosed probands. A set of 903 clinically-curated DD-associated recessive genes showed very strong enrichment of damaging biallelic genotypes (1.7-fold; Poisson $p=6\times10^{-18}$ for EABI and PABI combined). Indeed, 48% of the observed excess of damaging biallelic genotypes lay in these known genes. We also found a significant enrichment of damaging biallelic genotypes in 3371 genes annotated as having high probability of being intolerant of LoFs in the recessive state (pRec>0.9) (*15*) (1.2-fold; Poisson $p=2\times10^{-8}$), even after excluding the known recessive genes used to train the model (1.1-fold; Poisson $p=4\times10^{-4}$), and in 189 genes that were sub-viable when knocked out homozygously in mice (*16*) (1.8-fold; Poisson $p=2\times10^{-3}$). By contrast, we did not observe any recessive burden in 243 DD-associated genes that act by a dominant LoF mechanism, nor in genes predicted to be intolerant of heterozygous LoFs (probability of LoF intolerance, pLI, >0.9) in ExAC. We did not see excess recessive burden in the diagnosed probands in any of these gene sets (Poisson p>0.05).

We developed a method to estimate the proportion of probands who have a diagnostic variant in a particular genotype class (*14*). In contrast to our previously published approach (*5*), this method accounts for the fact that a fraction of the variants expected by chance are actually causal (Fig. S6); thus, it gives higher estimates than previously reported for the proportion of the cohort with causative *de novo* mutations. We estimated that 3.6% (~205) of the EABI probands have a recessive coding diagnosis, compared to 49.9% (~2836) with a *de novo* coding diagnosis. In the PABI subset, recessive coding genotypes likely explain 30.9% (~110) of the individuals, compared to 29.8% (~106) for *de novo* coding mutations. The contribution from recessive variants was nearly four times as high in EABI probands with similarly affected siblings than those without affected siblings (12.0% versus 3.2%), and highest in PABI probands with high autozygosity (47.1%) (Fig. 2; Table S2). In contrast, it did not differ significantly between PABI probands with low autozygosity and EABI probands.

Discovery of new recessive disease genes

To discover previously unidentified recessive genes, we next tested each gene for an excess of biallelic genotypes in undiagnosed probands from either EABI alone or EABI+PABI (Table S3) (*14*). Three genes passed stringent Bonferroni correction (binomial $p<3.4\times10^{-7}$, accounting for 8 tests for each of 18,630 genes): *EIF3F*, *KDM5B* and *THOC6*, of which the last is an established recessive DD-associated gene (*17*). Thirteen additional genes had $p<10^{-4}$ (Table 1). Eleven of these are known recessive DD-associated genes, and the distribution of p-values for all such known genes was shifted lower than that of all other genes (Kolmogorov-Smirnov test; $p<1\times10^{-15}$; Fig. S7). These observations validate our gene discovery approach, and suggest that our genome-wide significance threshold is conservative.

We observed five probands with an identical homozygous missense variant in *EIF3F* (binomial $p=1.2\times10^{-10}$) (ENSP00000310040.4:p.Phe232Val), predicted to be deleterious by SIFT, polyPhen and CADD. An additional four individuals in the DDD cohort were also homozygous for this variant but had been excluded from our discovery analysis for various reasons (Table S4). All probands had European ancestry and low overall autozygosity, and none of them (apart from the pairs of siblings) were related (kinship<0.02). In the gnomAD resource of population variation (http://gnomad.broadinstitute.org/), this variant (rs141976414) has a frequency of 0.12% in non-Finnish Europeans, and no homozygotes were observed. The observation that this single missense variant is driving all of the signal in *EIF3F* probably reflects the fact that it is one of the most common variants in the gene. *EIF3F* is essential in several cancer cell lines (*18*, *19*), so the biallelic knockout is likely to be lethal to humans; consistent with this, LoFs in this gene in gnomAD are very rare (MAF<0.01%).

All nine individuals homozygous for the *EIF3F* variant had ID and six had seizures (Table S4). Those for whom photos were available did not have a distinctive facial appearance (Fig. S8). Features observed in three or more unrelated individuals were behavioral difficulties and sensorineural hearing loss. One patient had skeletal muscle atrophy (Fig. S8), which is only reported in one other proband in the DDD study. This is notable because in mice, *Eif3f* has been shown to play a role in regulating skeletal muscle size via interaction with the mTOR pathway (*20*).

EIF3F encodes the F subunit of the mammalian eIF3 (eukaryotic initiation factor) complex, a negative regulator of translation. The genes encoding eIF2B subunits have been implicated in severe autosomal recessive neurodegenerative disorders (*21*). The secondary structure, domain architecture and 3D fold of EIF3F is conserved between species but sequence similarity is low (29% between yeast and humans) (Fig. 3A). The conserved Phe232 side chain mutated in our patients is buried (solvent accessibility 0.7%) and likely plays a stabilizing role (Fig. 3B). The loss of the aromatic side chain in the Phe232Val variant would likely disrupt protein stability. It is currently unknown how the Phe232Val variant affects EIF3F function and causes DD.

Another recessive gene we identified was *KDM5B* (binomial $p=1.1\times10^{-7}$) (Fig. 4), encoding a histone H3K4 demethylase. Other H3K4 methylases and demethylases are known to cause neurodevelopmental disorders (*22–24*). Three probands had biallelic LoFs passing our filters, and we subsequently identified a fourth who was compound heterozygous for a splice site variant and large gene-disrupting deletion. While our paper was under review, Faundes *et al. (25)* reported biallelic LoFs in *KDM5B* in three DD patients, including two of ours; our result provides robust statistical evidence that these are pathogenic. *KDM5B* is also enriched for *de novo* mutations in the DDD cohort (*5*) (binomial $p=5.1\times10^{-7}$). Additionally, we saw nominally significant over-transmission of LoF variants from the parents, who were almost all unaffected (p=0.002 including all families; transmission-disequilibrium test; Table S5). There was no evidence for a parent-of-origin bias in which parent transmitted the LoF.

We considered the possibility that all the *KDM5B* LoFs observed in probands might be, in fact, acting recessively and that the probands with apparently monoallelic LoFs had a second coding or regulatory hit on the other allele. However, we found no evidence supporting this hypothesis (see (*14*) and Fig. S9), nor of potentially modifying coding variants in likely interactor genes. There was also no evidence from the annotations in Ensembl or GTex data (https://gtexportal.org/home/) that the pattern could be explained by some LoFs avoiding nonsense-mediated decay (Fig. 4B). We searched for potential modifying epimutations, but found none (Fig. S10). These lines of evidence suggest that heterozygous LoFs in *KDM5B* are

pathogenic with incomplete penetrance, while homozygous LoFs are likely fully penetrant. Genome-wide analysis of DNA methylation levels in whole blood found no significant differences between probands with different types of *KDM5B* mutations or between these and controls (Fig. S11).

The four individuals with biallelic *KDM5B* variants have ID and variable congenital abnormalities (Table S6), in line with those seen in other disorders of the histone machinery (*26*). Affected individuals have a distinctive facial appearance with narrow palpebral fissures, arched or thick eyebrows, dark eyelashes, a low hanging columella, smooth philtrum and a thin upper vermillion border (Fig. 4C). However, in contrast to other disorders of the histone machinery, there was no consistent growth pattern. Other than ID, there were no consistent phenotypes or distinctive features shared between the biallelic and monoallelic individuals, or within the monoallelic group (Table S6).

Discussion

We found the contributions of *de novo* and recessive coding variants were approximately equal among DDD probands with Pakistani ancestry (both ~30%), but recessive causes contribute less than a tenth of the disease burden of dominant causes in the probands with European ancestry (3.6% vs 49.9% respectively). Our results are in line with previous reports of a low fraction of recessive diagnoses in European cohorts (*3*, *4*, *27*), but in contrast to those studies, our estimates include the recessive contribution in as-yet-undiscovered genes and are therefore unbiased. While it has been hypothesised that there are many more recessive DD-associated genes to be discovered (*6*, *28*), our analyses suggest that the cumulative impact of these discoveries on diagnostic yield will be modest in outbred populations.

Our use of DDD parental allele frequencies allowed us to carry out a properly calibrated burden analysis (Fig. S3), but does have caveats. Any enrichment in damaging coding variants in DDD parents compared to the general population will result in overestimates of the population frequency of such variants, as will the systematic difference between the true allele frequency of

a very rare variant and its estimate from a finite sample size (*15*). These effects could cause us to slightly underestimate the overall burden. Reassuringly, our estimate in the PABI subset (30.9%) is close to the 31.5% reported by genetics clinics in Kuwait (*29*), which has a similar level of consanguinity (*30*).

Our results have clinical value because they can be used to improve recurrence risk estimates for families with a particular ancestry and pattern of inheritance, but without a molecular diagnosis. Because our exact burden results derive from the patient population within the DDD study, extrapolating more widely requires some care. For example, the proportion of all DD patients in the British Isles with a recessive coding cause is probably higher than our estimate because some recessive DDs are relatively easily diagnosed through current screening and diagnostic practice (e.g. metabolic disorders) and therefore are less likely to be recruited to a research study. Furthermore, while the ascertainment of the DDD study is likely to be a reasonable proxy for undiagnosed patients from genetics clinics in the UK, country-specific properties of medical practice and levels of consanguinity may make the exact estimates less applicable elsewhere.

Recent papers have described the population structure and characteristics of South Asian populations and highlighted their potential for recessive disease gene discovery (*31*). Despite this expectation, and the substantially higher burden of recessive causation in the PABI subset (Fig. 2), they contributed little to our new gene discovery. This was partially due to modest sample size (which also explains the wide confidence intervals in Table S2) but was exacerbated by the consistent overestimation of rare variant frequencies.

Because damaging biallelic genotypes in *EIF3F* and *KDM5B* result in nonspecific and heterogeneous phenotypes, they are less likely to have been found by typical studies. *KDM5B* is unusual for a recessive gene because heterozygous LoFs appear to be pathogenic with incomplete penetrance. Several *de novo* missense and LoF mutations in *KDM5B* had previously been reported in individuals with autism or ID (*32*, *33*), but also in unaffected individuals (*32*). The other genes encoding H3K4 methylases and demethylases reported to cause DD are mostly dominant (*24*) and typically have a pLI score >0.99 and a very low pRec, in contrast to *KDM5B*

(pLI=5×10^{-5}; pRec>0.999). LoFs in some other dominant ID genes appear to be incompletely penetrant (*34*), as do several microdeletions (*35*). The evidence suggests that biallelic LoFs in *KDM5B* are fully penetrant in humans, but interestingly, the homozygous knockout shows incomplete penetrance in mice, with only one strain presenting with neurological defects (*36*, *37*). Until further studies clarify the true inheritance pattern of *KDM5B*-related disorders, caution should be exercised when counselling families about the clinical significance of heterozygous variants in this gene.

Our results suggest that identifying all recessive DD genes would allow us to diagnose 5.2% of the EABI+PABI subset of DDD, whereas identifying all the dominant DD genes would yield diagnoses for 48.6%. The high proportion of unexplained patients even amongst those with affected siblings or high consanguinity suggests that future studies should investigate a wide range of modes of inheritance including oligogenic and polygenic inheritance as well as noncoding recessive variants.

References and Notes

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

2. R. K. C. Yuen *et al.*, Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nat. Neurosci.* **20**, nn.4524 (2017).

3. S. C. Jin *et al.*, Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nat. Genet.* (2017), doi:10.1038/ng.3970.

4. C. F. Wright *et al.*, Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet*. **385**, 1305–1314 (2015).

5. Deciphering Developmental Disorders Study, Prevalence and architecture of de novo mutations in developmental disorders. *Nature*. **542**, 433–438 (2017).

6. H. H. Ropers, Genetics of early onset cognitive impairment. *Annu. Rev. Genomics Hum. Genet.* **11**, 161–187 (2010).

7. L. E. L. M. Vissers, C. Gilissen, J. A. Veltman, Genetic studies in intellectual disability and related disorders. *Nat. Rev. Genet.* **17**, 9–18 (2016).

8. P. A. Baird, T. W. Anderson, H. B. Newcombe, R. B. Lowry, Genetic disorders in children and young adults: a population study. *Am. J. Hum. Genet.* **42**, 677–693 (1988).

9. S. J. Schrodi *et al.*, Prevalence estimation for monogenic autosomal recessive diseases using population-based genetic data. *Hum. Genet.* **134**, 659–669 (2015).

10. N. Akawi *et al.*, Discovery of four recessive developmental disorders using probabilistic genotype and phenotype matching among 4,125 families. *Nat. Genet.* **47**, 1363–1369 (2015).

11. P. Biswas *et al.*, A mutation in IFT43 causes non-syndromic recessive retinal degeneration. *Hum. Mol. Genet.* (2017), doi:10.1093/hmg/ddx356.

12. P. A. Long *et al.*, Recessive TAF1A mutations reveal ribosomopathy in siblings with end-stage pediatric dilated cardiomyopathy. *Hum. Mol. Genet.* **26**, 2874–2881 (2017).

13. H. Tawamie *et al.*, Hypomorphic Pathogenic Variants in TAF13 Are Associated with Autosomal-Recessive Intellectual Disability and Microcephaly. *Am. J. Hum. Genet.* **100**, 555–561 (2017).

14. Materials and methods are available as supplementary materials at the Science website.

15. M. Lek *et al.*, Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. **536**, 285–291 (2016).

16. M. E. Dickinson *et al.*, High-throughput discovery of novel developmental phenotypes. *Nature*. **537**, 508–514 (2016).

17. C. L. Beaulieu *et al.*, Intellectual disability associated with a homozygous missense mutation in THOC6. *Orphanet J. Rare Dis.* **8**, 62 (2013).

18. R. Marcotte *et al.*, Essential gene profiles in breast, pancreatic, and ovarian cancer cells. *Cancer Discov.* **2**, 172–189 (2012).

19. T. Wang *et al.*, Identification and characterization of essential genes in the human genome. *Science*. **350**, 1096–1101 (2015).

20. A. Csibi *et al.*, The translation regulatory subunit eIF3f controls the kinase-dependent mTOR signaling required for muscle differentiation and hypertrophy in mouse. *PLoS One*. **5**, e8994 (2010).

21. A. Fogli, O. Boespflug-Tanguy, The large spectrum of eIF2B-related diseases. *Biochem. Soc. Trans.* **34**, 22–29 (2006).

22. E. Shen, H. Shulha, Z. Weng, S. Akbarian, Regulation of histone H3K4 methylation in brain development and disease. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **369** (2014), doi:10.1098/rstb.2013.0514.

23. T. Singh *et al.*, Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders. *Nat. Neurosci.* **19**, 571–577 (2016).

24. C. N. Vallianatos, S. Iwase, Disrupted intricacy of histone H3K4 methylation in neurodevelopmental disorders. *Epigenomics*. **7**, 503–519 (2015).

25. V. Faundes *et al.*, Histone Lysine Methylases and Demethylases in the Landscape of Human Developmental Disorders. *Am. J. Hum. Genet.* **102**, 175–187 (2018).

26. J. A. Fahrner, H. T. Bjornsson, Mendelian disorders of the epigenetic machinery: tipping the balance of chromatin states. *Annu. Rev. Genomics Hum. Genet.* **15**, 269–293 (2014).

27. C. Gilissen *et al.*, Genome sequencing identifies major causes of severe intellectual disability. *Nature*. **511**, 344–347 (2014).

28. F. L. Raymond, P. Tarpey, The genetics of mental retardation. *Hum. Mol. Genet.* **15 Spec No 2**, R110–6 (2006).

29. A. S. Teebi, Autosomal recessive disorders among Arabs: an overview from Kuwait. *J. Med. Genet.* **31**, 224–233 (1994).

30. G. O. Tadmouri *et al.*, Consanguinity and reproductive health among Arabs. *Reprod. Health*. **6**, 17 (2009).

31. N. Nakatsuka *et al.*, The promise of discovering population-specific disease-associated genes in South Asia. *Nat. Genet.* (2017), doi:10.1038/ng.3917.

32. I. Iossifov *et al.*, The contribution of de novo coding mutations to autism spectrum disorder. *Nature*. **515**, 216–221 (2014).

33. E. Athanasakis *et al.*, Next generation sequencing in nonsyndromic intellectual disability: from a negative molecular karyotype to a possible causative mutation detection. *Am. J. Med. Genet. A*. **164A**, 170–176 (2014).

34. H. H. Ropers, T. Wienker, Penetrance of pathogenic mutations in haploinsufficient genes for intellectual disability and related disorders. *Eur. J. Med. Genet.* **58**, 715–718 (2015).

35. G. L. Carvill, H. C. Mefford, Microdeletion syndromes. *Curr. Opin. Genet. Dev.* **23**, 232–239 (2013).

36. M. Albert *et al.*, The histone demethylase Jarid1b ensures faithful mouse development by protecting developmental genes from aberrant H3K4me3. *PLoS Genet.* **9**, e1003461 (2013).

37. M. R. Zou *et al.*, Histone demethylase jumonji AT-rich interactive domain 1B (JARID1B) controls mammary gland development by regulating key developmental and lineage specification genes. *J. Biol. Chem.* **289**, 17620–17633 (2014).

38. G. L. Yamamoto *et al.*, Rare variants in SOS2 and LZTR1 are associated with Noonan syndrome. *J. Med. Genet.* **52**, 413–421 (2015).

39. N. A. Akawi, F. Al-Jasmi, A. M. Al-Shamsi, B. R. Ali, L. Al-Gazali, LINS, a modulator of the WNT signaling pathway, is involved in human cognition. *Orphanet J. Rare Dis.* **8**, 87 (2013).

40. H. Najmabadi *et al.*, Deep sequencing reveals 50 novel genes for recessive cognitive disorders. *Nature*. **478**, 57–63 (2011).

Acknowledgements

We thank the families for their participation and patience, the Sanger Human Genome Informatics team for their support in processing the data, P. Danacek for help with calling the regions of homozygosity, K. de Lange for help with making figures, K. Samocha for providing the mutability estimates, and A. Sakar for reviewing a patient.

Funding: 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. This study makes use of DECIPHER (http://decipher.sanger.ac.uk), which is funded by the Wellcome Trust.

Author contributions: Exome data analysis: H.C.M., J.F.M. Protein structure modelling: J.S., Methylation analysis: J.H. Clinical interpretation: W.D.J. Data processing: G.G., M.N., J.K., C.F.W., E.R. Experimental validation: E.P. Methods development: H.C.M., P.S., M.E.H., J.C.B. Data interpretation: H.C.M., J.S., J.H., N.A., M.E.H., J.C.B. Patient recruitment and phenotyping: M.B., J.D., R.H., A.H., D.S.J., K.J., D.K, S.A.L., S.G.M., J.M., M.J.P., M.S., P.D.T., P.C.V., and M.W. Experimental and analytical supervision: C.F.W., D.R.F., H.V.F., M.E.H., J.C.B. Project Supervision: J.C.B. Writing: H.C.M., W.D.J., J.S., M.B., A.B., J.H., M.E.H., J.C.B.

Competing interests: M.E.H. is a co-founder of, consultant to, and holds shares in, Congenica Ltd, a genetics diagnostic company.

Data and materials availability: Exome sequencing and phenotype data are accessible via the European Genome-phenome Archive (EGA) (https://www.ebi.ac.uk/ega/studies/EGAS00001000775) (Datafreeze 2016-10-03).

Supplementary Materials

Materials and Methods

Table S1 – S6

Fig S1 – S11

References (41-66)

Tables

Table 1: Genes enriched for damaging biallelic coding genotypes with $p<1\times10^{-4}$ (for eight tests; see (*14*)). Shown are the number of observed biallelic genotypes of different consequence classes, the lowest p-value out of the eight tests (achieved using EABI alone for all genes except for *VPS13B*), details of the corresponding test, and the p-value for phenotypic similarity for the relevant probands (*10*). Known recessive DD genes from the DDG2P list are indicated (http://www.ebi.ac.uk/gene2phenotype/).

gene	Biallelic genotypes counts for EABI (PABI if >0)	p-value - genotype	p-value - phenotype	Consequence class for most	Note

	LoF	LoF/ damaging missense	damaging missense			significant test	
EIF3F	0	0	5	1.2E-10	0.72	damaging missense	two probands had another affected sibling, both of whom were homozygous for the same variant
THOC6	0	1	3	4.4E-09	6.0E-05	LoF + LoF/damaging missense + damaging missense	known recessive gene
KDM5B	3	0	0	1.1E-07	0.53	LoF	previously reported as dominant gene; a fourth proband is compound heterozygous for a splice variant and CNV
CNTNAP1	2 (1)	1	0 (1)	1.8E-06	0.02	LoF+LoF/damaging missense	known recessive gene
KIAA0586	5	1	1	1.9E-06	0.05	LoF	known recessive gene; two probands have affected sibs, and both share the variants
NALCN	1	2	0	2.4E-06	0.37	LoF+LoF/damaging missense	known recessive gene; one proband has an affected sib who shares the variants
PIGN	0	3	1	2.5E-06	0.10	LoF + LoF/damaging missense + damaging missense	known recessive gene; one proband has an affected sib who shares the variants
ST3GAL5	0 (1)	2	0	2.7E-06	0.09	LoF+LoF/damaging missense	known recessive gene
ATAD2B	1	1	0	3.6E-06	0.88	LoF+LoF/damaging missense	one of our probands has an affected sib who shares the variants
LZTR1	0	3	0	5.6E-06	0.06	LoF+LoF/damaging missense	one of our probands has an affected sib who does not share both variants, so causality is dubious ;dominant missense mutations cause Noonan syndrome(*38*)
LINS	2	0	0	8.2E-06	0.74	LoF	one proband has affected sib who shares the variant; putative recessive gene(*39*, *40*)
POLR1C	0	1	2	1.4E-05	0.42	LoF + LoF/damaging missense + damaging missense	known recessive gene
MMP21	0	2	0	1.4E-05	2.0E-03	LoF+LoF/damaging missense	known recessive gene
MAN1B1	1	1	0	1.5E-05	0.62	LoF+LoF/damaging missense	known recessive gene
VPS13B	2 (1)	1	2	2.8E-05	0.05	LoF	known recessive gene
UBA5	0	2	1	3.7E-05	0.84	LoF+LoF/damaging missense	known recessive gene

Figures

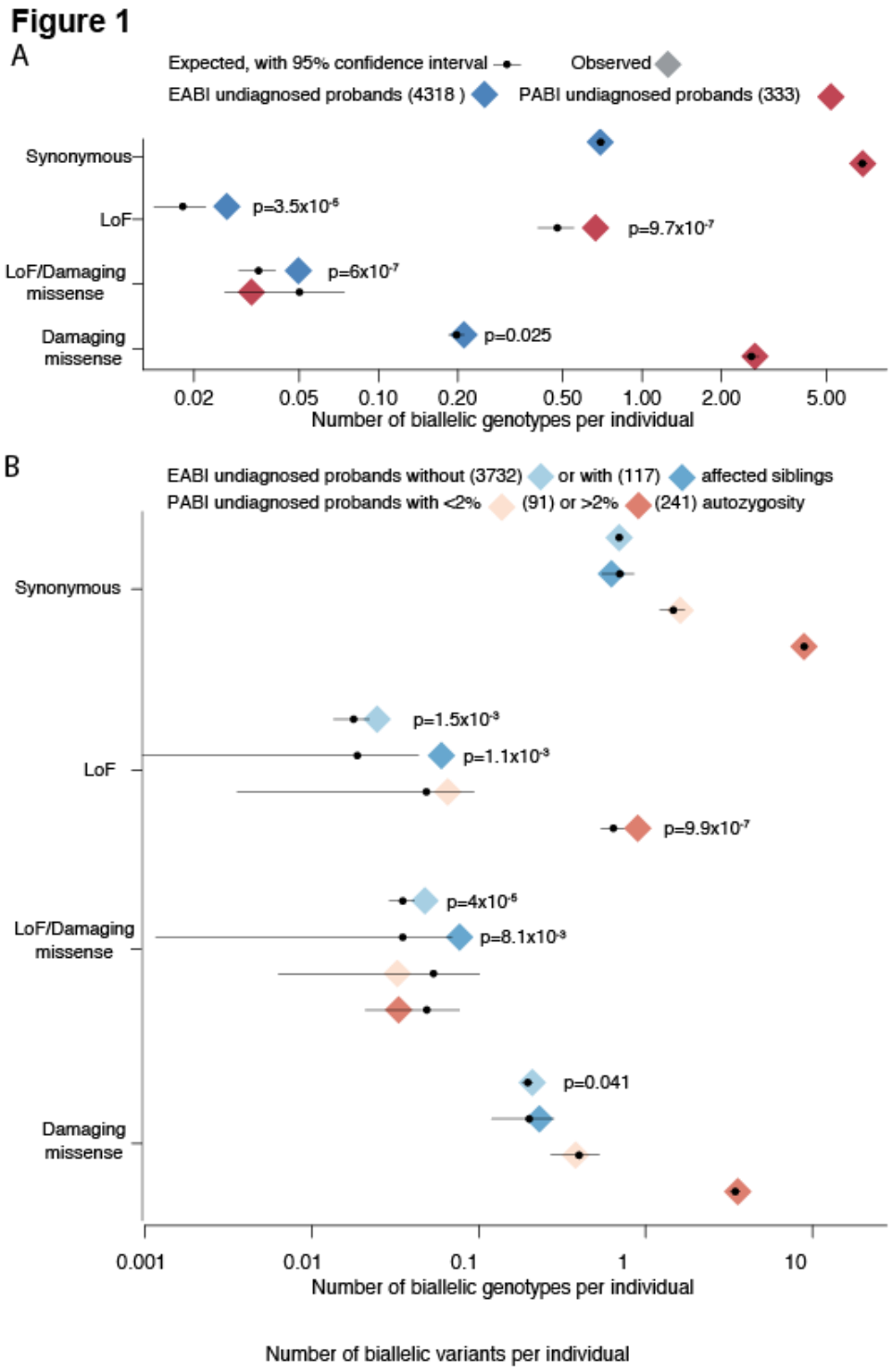

Fig. 1: Enrichment of damaging biallelic genotypes in DDD probands. Number of observed and expected biallelic genotypes per individual for all genes in A) undiagnosed EABI and PABI probands, and B) different subsets of undiagnosed probands (see (*14*)). Samples sizes are

indicated in parentheses in the keys. Nominally significant p-values from a Poisson test of enrichment are shown.

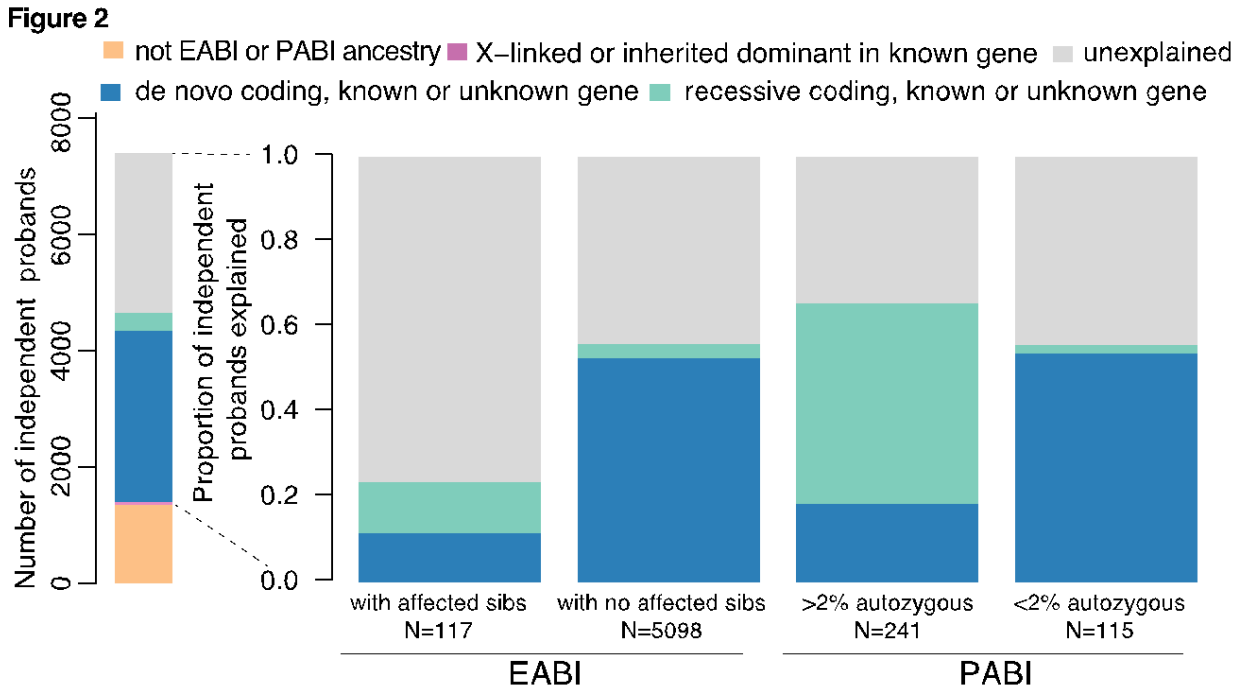

Fig. 2: Proportion of probands explained by recessive coding variants versus *de novo* coding mutations. Left: number of probands grouped by diagnostic category. The inherited dominant and X-linked diagnoses (narrow pink bar) include only those in known genes, whereas the proportion of probands with *de novo* and recessive coding diagnoses was inferred as described in (*14*). Right: the proportion of probands in various EABI and PABI subsets inferred to have diagnostic variants in the indicated classes.

Figure 3

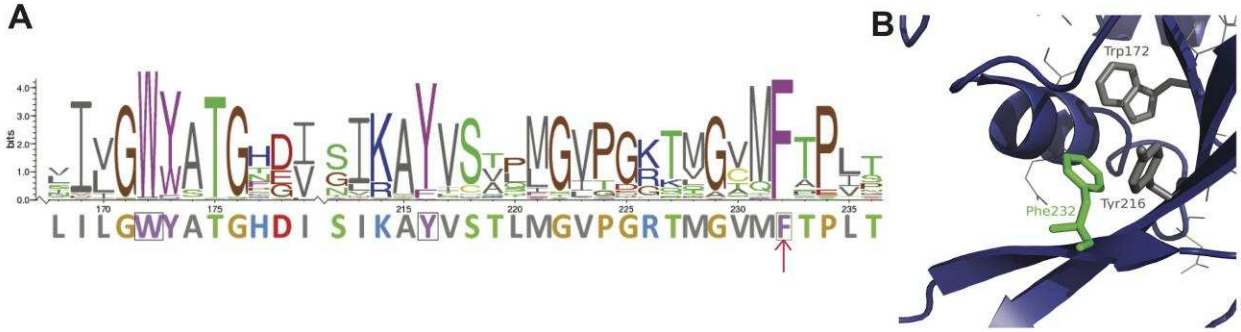

Fig. 3: Predicted effects of a pathogenic recessive missense variant in *EIF3F*. A) Section of the amino acid sequence logo for EIF3F where the strength of conservation across species is indicated by the size of the letters. The sequence below represents the human EIF3F. Boxed characters are the aromatic residues conserved between humans and yeast and proximal in space to Phe232. B) Structure of the section of EIF3F containing the Phe232Val variant, highlighted in green. Amino acids conserved between yeast and human sequences as highlighted in panel A are shown in grey. See (*14*) for details of structure prediction.

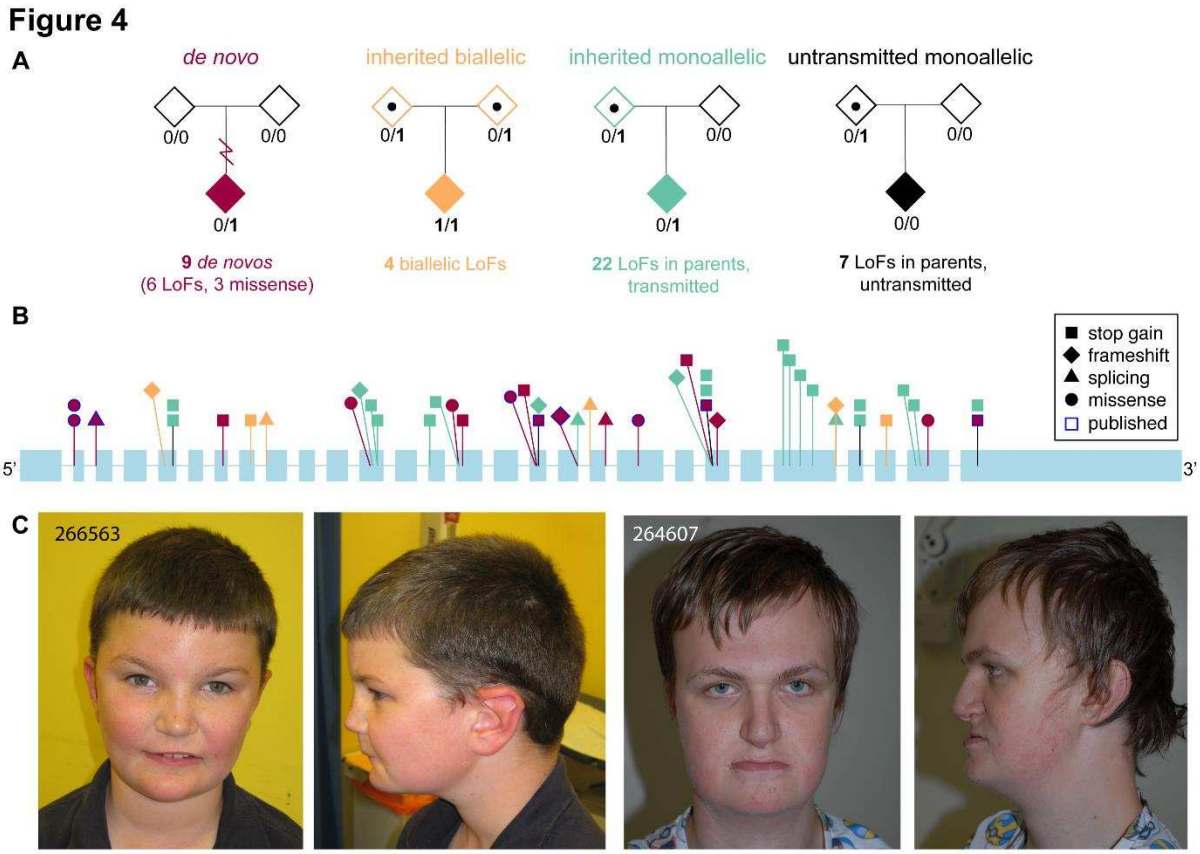

Fig. 4: *KDM5B* is a recessive DD gene in which heterozygous LoFs are incompletely penetrant. A) Summary of damaging variants found in *KDM5B*. B) Positions of likely damaging variants found in this and previous studies in the longest annotated transcript of *KDM5B*, ENST00000367264.2 (introns not to scale). Colors correspond to those shown in (A). Note the lack of obvious differences in the spatial distribution of *de novo* versus monoallelic or biallelic inherited LoFs within the gene. Two large deletions have been omitted. All variants are listed in Table S6. C) Anterior-posterior facial photographs of two of the individuals with biallelic *KDM5B* variants. Informed consent was obtained to publish these.