Quantifying the contribution of recessive coding variation to developmental disorders

Hilary C. Martin*1, Wendy D. Jones*2, Rebecca McIntyre3, Gabriela Sanchez-Andrade4, Mark Sanderson4, James D. Stephenson1,2, Carla P. Jones1, Juliet Handsaker5, Giuseppe Gallone6, Michaela Bruntraeger7, Jeremy F. McRae4, Elena Prigmore7, Patrick Short5, Mari Niemi8, Joanna Kaplanis4, Elizabeth J. Radford4, Nadia Akawi9, Meena Balasubramanian4, John Dean2, Rachel Horton9, Alice Hulbert9, Diana S. Johnson9, Katie Johnson10, Dhavendra Kumar11, Sally Ann Lynnh2, Sarju G. Mehta7, Jenny Morton8, Michael J. Parker2, Miranda Splitt9, Peter D Turnpenny2, Pradeep C. Vasudevan10, Michael Wright7, Andrew Bassett1, Sebastian S. Gerety4, Caroline F. Wright9, David R. FitzPatrick9, Helen V. Firth1,2, Matthew E. Hurles7, Jeffrey C. Barrett8*, on behalf of the Deciphering Developmental Disorders Study

*Corresponding author. Email: hcm@sanger.ac.uk (H.C.M.); jeff.barrett@genomicsplc.com (J.C.B.)

We estimated the genome-wide contribution of recessive coding variation from 6,040 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, KDM5B and EIF3F, and functionally validated them with mouse and cellular models. Our results suggest that recessive coding variants account for a small fraction of currently undiagnosed non-consanguineous individuals, and that the role of noncoding variants, incomplete penetrance, and polygenic mechanisms need further exploration.

Large-scale sequencing studies of phenotypically heterogeneous rare disease patients can discover new disease genes (1–3) and characterize the genetic architecture of such disorders. In the Deciphering Developmental Disorders (DDD) study, we previously estimated the fraction of patients with a causal de novo coding mutation in both known and as-yet-undiagnosed disease genes to be 40–45% (4), and here we extend this approach to recessive variants. It has been posited that there are thousands of as-yet-undiagnosed recessive intellectual disability (ID) genes (5, 6), which could imply that recessive variants explain a large fraction of undiagnosed rare disease cases. However, attempts to estimate the prevalence of recessive disorders have been restricted to known disorders (7) or known pathogenic alleles (8). Here, we quantify the total autosomal recessive coding burden using a robust and unbiased statistical framework in 6,040 exome-sequenced DDD trios from the British Isles. Our approach provides a better-calibrated estimate of the exome-wide burden of recessive disease than previously published methods (3, 9).

We analyzed 5,684 European and 356 Pakistani probands (EABI, PABI - European or Pakistani Ancestry from the British Isles) (figs. S1 and S2) with developmental disorders (DDs). The clinical features are heterogeneous and representative of genetically undiagnosed DD patients from British and Irish clinical genetics services: 88% have an abnormality of the nervous system, and 88% have multiple affected organ systems (Fig. 1, fig. S3, and table S1). Clinical features are largely similar between EABI and PABI (Fig. 1 and table S1).

To assess the genome-wide recessive burden, we compared the number of rare (minor allele frequency, MAF, <1%)
biallelic genotypes observed in our cohort to the number expected by chance (10). We used the phased haplotypes from unaffected DDD parents to estimate the expected number of biallelic genotypes. Reassuringly, the number of observed biallelic synonymous genotypes matched the expectation (fig. S4). We observed no significant burden of biallelic genotypes of any consequence class in 1,389 probands with a likely diagnostic de novo, inherited dominant or X-linked variant. We therefore evaluated the recessive coding burden in the remaining 4,318 EABI and 333 PABI probands. This “undiagnosed” cohort were more likely to have a recessive cause because they did not have a likely dominant or X-linked diagnosis (11), had at least one affected sibling, or >2% autozygosity (Fig. 2A).

As expected due to their higher autozygosity (fig. S5), PABI individuals had more rare biallelic genotypes than EABI individuals (Fig. 2A); 92% of these were homozygous (rather than compound heterozygous), versus only 28% for the EABI samples. We observed a significant enrichment of biallelic loss-of-function (LoF) genotypes in both undiagnosed ancestry groups (Poisson P = 3.5 × 10^{-5} in EABI, P = 9.7 × 10^{-7} in PABI), and, in the EABI group, a nominally significant enrichment of biallelic damaging missense genotypes (P = 0.025) and a significant enrichment of compound heterozygous LoF/damaging missense genotypes (P = 6 × 10^{-5}) (Fig. 2A).

Amongst the 4,651 EABI+PABI undiagnosed probands, a set of 903 clinically-curated DD-associated recessive genes showed a higher recessive burden (fig. S6) (1.7-fold; Poisson P = 6 × 10^{-19}) than average (1.1-fold for all genes). Indeed, 48% of the observed excess of biallelic genotypes lay in these known genes. By contrast, we did not observe any recessive burden in 243 DD-associated genes with a dominant LoF mechanism, nor in any gene sets tested in the 1,389 diagnosed probands (Poisson P > 0.05).

We developed a method to estimate the proportion of probands with a causal variant in a particular genotype class (10) in either known and as-yet-undiscovered genes. Unlike our previously published approach (4), this method accounts for the fact that some fraction of the variants expected by chance are actually causal (fig. S7). We estimated that 3.6% (~205) of the 5,684 EABI probands have a recessive coding diagnosis, compared to 49.9% (~2836) with a de novo coding diagnosis. Recessive coding genotypes explain 30.9% (~110) of the 356 PABI individuals, compared to 29.8% (~106) for de novos. The contribution from recessive variants was higher in EABI probands with affected siblings than those without (12.0% of 117 versus 3.2% of 5,098), and highest in PABI probands with high autozygosity (47.1% of 241) (Fig. 2B and table S2). In contrast, it did not differ between 115 PABI probands with low autozygosity and all 5,684 EABI probands.

We caution that the PABI results may be less reliable due to modest sample size (note the wide confidence intervals in table S2), exacerbated by consistent overestimation of rare variant frequencies in our limited sample of parents. Reassuringly, our estimated recessive contribution in PABI is close to the 31.5% reported in Kuwait (12), which has a similar level of consanguinity (13). Our results are consistent with previous reports of a low fraction of recessive diagnoses in European cohorts (3, 11, 14), but unlike those studies, our estimates further show that the recessive contribution in as-yet-undiscovered genes is also small. While it has been hypothesized that there are thousands of undiscovered recessive DD-associated genes (5, 6), our analyses suggest that the cumulative impact of these discoveries on diagnostic yield will be modest in non-consanguineous populations.

We next tested each gene for an excess of biallelic genotypes in the undiagnosed probands (table S3) (10). Three genes passed stringent Bonferroni correction (P < 3.4 × 10^{-7}) (10), THOC6 (previously reported (15)), EIF3F, and KDM5B. Thirteen additional genes had P < 10^{-4} (table S4), of which seven are known autosomal recessive DD-associated genes, and known genes were enriched for lower P values (fig. S8).

We observed five probands with an identical homozygous missense variant in EIF3F (binomial P = 1.2 × 10^{-10}) (ENSP00000310040.4:p.Phe232Val), plus four additional homozygous probands who had been excluded from our discovery analysis for various reasons (table S5). The variant (rs141976414) has a frequency of 0.12% in non-Finnish Europeans (one of the most common protein-altering variants in the gene), and no homozygotes were observed in gnomAD (http://gnomad.broadinstitute.org).

All nine individuals homozygous for Phe232Val had intellectual disability (ID) and a subset also had seizures (6/9), behavioral difficulties (3/9) and sensorineural hearing loss (3/9) (table S5). There was no obvious distinctive facial appearance (fig. S9). EIF3F encodes a subunit of the mammalian eIF3 (eukaryotic initiation factor) complex, which negatively regulates translation. The genes encoding eIF2B subunits have been implicated in severe autosomal recessive neurodegenerative disorders (16). We edited iPSC lines with CRISPR-Cas9 to be heterozygous or homozygous for the Phe232Val variant, and Western blots showed that EIF3F protein levels were ~27% lower in homozygous cells relative to heterozygous and wild-type cells (fig. S10), which may be due to reduced protein stability (fig. S11). The Phe232Val variant significantly reduced translation rate (Fig. 3A and fig. S12). Proliferation rates were also reduced in the homozygous but not heterozygous cells (Fig. 3B and fig. S13), although the viability of the cells was unchanged (fig. S14).

Another recessive gene we identified was KDM5B (binomial P = 1.1 × 10^{-7}) (Fig. 4), encoding a histone H3K4 demethylase. Three probands had biallelic LoFs passing our filters,
and a fourth was compound heterozygous for a splice site variant and a large gene-disrupting deletion. Several of these patients were recently reported with less compelling statistical evidence (17). Interestingly, KDM5B is also enriched for de novo mutations in our cohort (4) (binomial $P = 5.1 \times 10^{-7}$). We saw nominally significant over-transmission of LoFs from the mostly unaffected parents ($P = 0.002$, transmission-disequilibrium test) (table S6), but no parent-of-origin bias. Theoretically, all the KDM5B LoFs observed in probands might be acting recessively and heterozygous probands may have a second (missed) coding or regulatory hit or modifying epimutation. However, we found no evidence supporting this [see (10) (figs. S15 and S16)], nor of potentially modifying coding variants in likely interactor genes, nor that some LoFs avoid nonsense-mediated decay (Fig. 4B). Genome-wide levels of DNA methylation in whole blood did not differ between probands with different types of KDM5B mutations or between these and controls (fig. S17).

These lines of evidence, along with previous observations of KDM5B de novos in both autism patients and unaffected siblings (18), suggest that heterozygous LoFs in KDM5B are pathogenic with incomplete penetrance, while homozygous LoFs are likely fully penetrant. Several microdeletions (19) and LoFs in other dominant ID genes are incompletely penetrant (20). Other H3K4 methylases and demethylases also cause neurodevelopmental disorders (21). KDM5B is atypical; the others are mostly dominant (21), typically with $pLI > 0.99$ and very low $pRec$ scores, whereas KDM5B has $pLI = 5 \times 10^{-5}$ and $pRec > 0.999$ (22).

KDM5B is the only gene that showed significant enrichment for both biallelic variants and de novo mutations in our study. We saw significant enrichment of de novo missense (373 observed versus 305 expected; ratio = 1.25, upper-tailed Poisson $P = 1 \times 10^{-4}$) but not de novo LoF mutations across all known recessive DD genes (excluding those known to also show dominant inheritance). One hypothesis is that the de novo missense mutations are acting as a “second hit” on the opposite haplotype from an inherited variant in the same gene. However, we saw only two instances of this in the cohort, and besides, if they were driving the signal, we would expect to see a burden of de novo LoFs in recessive genes too, which we do not. A better explanation is that recessive DD genes are also enriched for dominant activating mutations. There are known examples of this; e.g., in NALCN (23, 24) and MAB21L2 (25), heterozygous missense variants are activating or dominant-negative, whereas the biallelic mechanism is loss-of-function. In contrast, the six de novo LoFs in KDM5B suggest it follows a different pattern. Of the twenty-one recessive genes with nominally significant de novo missense enrichment in our data, only one showed evidence of mutation clustering using our previously published method (1) (CTCI; $P = 0.03$), which could suggest an activating/dominant-negative mechanism. Larger sample sizes will be needed to establish which of these genes also act dominantly, and by which mechanism.

All four individuals with biallelic KDM5B variants have ID, variable congenital abnormalities (table S7) and a distinctive facial appearance (fig. S18). Other than ID, there were no consistent phenotypes or distinctive features shared between the biallelic and monoallelic individuals, or within the monoallelic group (table S7).

We created a mouse loss-of-function model for Kdm5b. Heterozygous knockout mice appear normal and fertile, while homozygous Kdm5b-null mice are subviable (44% of expected, from heterozygous in-crosses). This partially penetrant lethality, in addition to a fully penetrant vertebral patterning defect (fig. S19), is consistent with previously published work (26). We additionally identified numerous behavioral abnormalities in homozygous Kdm5b-null mice: increased anxiety, less sociability, and reduced long-term memory compared to wild-types (Fig. 4).

We have quantified the contribution of recessive coding variants in both known and as-yet-undiscovered genes to a large UK cohort of DD patients, and found that overall they explain a small fraction. Our methodology allowed us to carry out an unbiased burden analysis not possible with previous methods (fig. S4). We identified two new recessive DD genes that are less likely to be found by typical studies because they result in heterogeneous and nonspecific phenotypes, and presented strong functional evidence supporting their pathogenicity.

Our results can be used to improve recurrence risk estimates for undiagnosed families with a particular ancestry and pattern of inheritance. Extrapolating our results more widely requires some care: our study is slightly depleted of recessive diagnoses since some recessive DDs (e.g., metabolic disorders) are relatively easily diagnosed through current clinical practice in the UK and less likely to have been recruited. Furthermore, country-specific diagnostic practices and levels of consanguinity may make the exact estimates less applicable outside the UK.

Overall, we estimated that identifying all recessive DD genes would allow us to diagnose 5.2% of the EABI+PABI subset of DDD, whereas identifying all 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


10. Materials and methods are available as supplementary materials.


ACKNOWLEDGMENTS


SUPPLEMENTARY MATERIALS

www.sciencemag.org/cgi/content/full/science.aar6/31/DC1
Materials and Methods
Figs. S1 to S20
Tables S1 to S7
References (27–54)

6 December 2017; resubmitted 10 August 2018
Accepted 29 October 2018
Published online 8 November 2018
10.1126/sciencea6731

First release: 8 November 2018
www.sciencemag.org (Page numbers not final at time of first release)
Fig. 1. Clinical features of DDD probands analyzed here. Proportion of probands in different groups with clinical features indicated, extracted from HPO terms. Asterisks indicate nominally significant differences between indicated groups (Fisher’s exact test).
Fig. 2. Contribution of recessive coding variants to genetic architecture in this study. (A) Number of observed and expected biallelic genotypes per individual across all genes. Nominally significant p-values from a Poisson test of enrichment are shown. (B) 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 (10), including those in as-yet undiscovered genes. Right: the proportion of probands in various patient subsets inferred to have diagnostic variants in the indicated classes.

Fig. 3. Functional consequences of the pathogenic EIF3F recessive missense variant. (A) The Phe232Val variant impairs translation. Plot shows median fluorescence intensity (MFI) in iPSC lines heterozygous or homozygous for or without the Phe232Val variant (correcting for replicate effects), measured using a Click-it protein synthesis assay (10). MFI correlates with methionine analog incorporation in nascent proteins. The p-value indicates a nonzero effect of genotype from a linear regression of MFI on genotype and replicate. Red lines: means. (B) The Phe232Val variant impairs iPSC proliferation in the homozygous but not heterozygous form. Results from a cell trace violet (CTV) proliferation assay, in which CTV concentration reduces on each division. The population of cells that have been through zero, one or multiple divisions is labeled.
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 KDM5B (ENST00000367264.2; introns not to scale), omitting two large deletions. Colors correspond to those shown in (A). There are no differences in the spatial distribution of LoFs by inheritance mode, nor in their likelihood of escaping nonsense-mediated decay by alternative splicing in GTEx (https://gtexportal.org/home/). (C to E) Behavioral defects of homozygous Kdm5b-null versus wild-type mice (n = 14-16). (C) Knockout mice displayed increased anxiety, spending significantly less time in the light compartment of the Light-Dark box. (D) Reduced sociability, in the three-chamber sociability test. Knockout mice spent less time investigating a novel mouse. (E) 24h memory impairment. While wild-type mice preferentially investigated an unfamiliar mouse over a familiar one, homozygous knockout mice showed no discrimination.
Quantifying the contribution of recessive coding variation to developmental disorders

Hilary C. Martin, Wendy D. Jones, Rebecca McIntyre, Gabriela Sanchez-Andrade, Mark Sanderson, James D. Stephenson, Caria P. Jones, Juliet Handsaker, Giuseppe Gallone, Michaela Bruntræger, Jeremy F. McRae, Elena Prigmore, Patrick Short, Mari Niemi, Joanna Kaplanis, Elizabeth J. Radford, Nadia Akawi, Meena Balasubramanian, John Dean, Rachel Horton, Alice Hulbert, Diana S. Johnson, Katie Johnson, Dhavendra Kumar, Sally Ann Lynch, Sarju G. Mehta, Jenny Morton, Michael J. Parker, Miranda Spillit, Peter D Turnpenny, Pradeep C. Vasudevan, Michael Wright, Andrew Bassett, Sebastian S. Gerety, Caroline F. Wright, David R. FitzPatrick, Helen V. Firth, Matthew E. Hurles, Jeffrey C. Barrett and on behalf of the Deciphering Developmental Disorders Study

published online November 8, 2018