Autosomal recessive coding variants explain only a small proportion of undiagnosed developmental disorders in the British Isles

Hilary C. Martin[1,*], Wendy D. Jones[1,2], James Stephenson[1,3], Juliet Handsaker[1], Giuseppe Gallone[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], 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 Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, U.K.
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

We analyzed 7,448 exome-sequenced families from the Deciphering Developmental Disorders study to search for recessive coding diagnoses. We estimated that the proportion of cases attributable to recessive coding variants is 3.6% for patients of European ancestry, and 30.9% for those of Pakistani ancestry due to elevated autozygosity. We tested every gene for an excess of damaging homozygous or compound heterozygous genotypes, and found that known recessive genes showed a significant tendency towards having lower p-values (Kolmogorov-Smirnov $p=3.3x10^{-16}$). Three genes passed stringent Bonferroni correction, including a new disease gene, *EIF3F*, and *KDM5B*, which has previously been reported as a dominant disease gene. *KDM5B* appears to follow a complex mode of inheritance, in which heterozygous loss-of-function variants (LoFs) show incomplete penetrance and biallelic LoFs are fully penetrant. Our results suggest that a large proportion of undiagnosed developmental disorders remain to be explained by other factors, such as noncoding variants and polygenic risk.

Hundreds of autosomal recessive disease genes have been identified by studying large families with multiple affected individuals, bottlenecked populations, or populations with high levels of consanguinity [1–4]. Outside these circumstances, it has been difficult to find new genes for recessive diseases, especially those with a clinically variable presentation, such as non-syndromic intellectual disability (ID) [5]. It is challenging to identify multiple families with the same underlying novel recessive disorder when affected individuals do not share distinctive phenotypes. Even if multiple families who share a candidate recessive genotype can be found, it is important to consider the probability of observing it in multiple families by chance. For the same reasons, the prevalence of severe disorders that are due to recessive inheritance has been difficult to estimate, particularly in large populations with low levels of endogamy.

Here we describe an analysis of autosomal recessive coding variants in 7,448 exome-sequenced families with a child with a severe undiagnosed developmental disorder (DD). These were

recruited as part of the Deciphering Developmental Disorders (DDD) study from clinical genetics services across the UK and Ireland [6]. The DDD participants have highly variable clinical presentation, and 76% have British European ancestry, so have neither been through a recent population bottleneck nor have high levels of consanguinity. We recently estimated that 40-45% of this cohort have pathogenic *de novo* coding mutations, leaving 55-60% unexplained [7]. Using probabilistic genotype and phenotype matching in a subset of this cohort, we previously identified four new recessive disorders [8]. The increased sample size described here gives us better power to ask questions about the overall burden of recessive causality in this cohort and to identify new recessive disease genes.

Results

Genome-wide recessive burden

We hypothesized there should be a burden of biallelic (i.e. homozygous or compound heterozygous) genotypes predicted either to cause loss-of-function (LoF) or likely damage to the protein. For each of three possible genotypes (LoF on both alleles, damaging missense on both, or one of each), 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 (see Methods). We introduced three refinements to the framework we used previously [8]. Firstly, because our method is sensitive to inaccuracy in population frequency estimates of very rare variants in broadly-defined ancestry groups like “Europeans” or “South Asians”, we focused our analysis on the largest two subsets of the cohort that had homogenous ancestry, corresponding in a 1000 Genomes Project principal components analysis (Supplementary Figure 1) to Great British individuals and Punjabis from Lahore, Pakistan. We refer to these subsets as having European Ancestry or Pakistani Ancestry from the British Isles (EABI, PABI). Secondly, rather than using ExAC [9] to estimate the population frequencies of variants, we used the unaffected DDD parents. This was to avoid differences due to quality control of the sequencing and variant calling, and to allow us to phase rare variants in the same gene. Finally, we modified our ascertainment of autozygous segments (i.e. both alleles

inherited identical-by-descent from a recent shared ancestor) in order to avoid overcalling of regions of homozygosity that were not due to recent consanguinity. After these calibrations, the number of observed biallelic synonymous variants (which we do not expect to be involved in disease) closely matched what we would expect by chance (ratio=0.997 for EABI and 1.003 for PABI; p=0.6 and 0.4) (Figure 1A).

We observed no significant burden of biallelic genotypes of any consequence class in 1,366 EABI probands with a likely diagnostic *de novo* mutation, inherited dominant variant or X-linked variant, consistent with those probands' phenotypes being fully explained by the variants already discovered. 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 diagnostic variant in a known dominant or X-linked DD gene [6], or had at least one affected sibling, or >2% autozygosity. As expected due to their higher autozygosity (Supplementary Figure 2), PABI individuals had substantially more rare biallelic genotypes than EABI individuals (Figure 1). Ninety-two percent of the likely damaging rare biallelic genotypes observed in PABI samples were homozygous, 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 group (~1.4-fold enrichment in each; $p=3.5\times10^{-5}$ 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 (p=0.03), as well as a significant enrichment of compound heterozygous LoF/damaging missense genotypes in the EABI group (1.4-fold enrichment; $p=6\times10^{-7}$). In the EABI group, the enrichments became stronger and more significant at lower MAF, but the absolute number of excess variants fell slightly in some cases (Supplementary Figure 3). Thus, plausibly pathogenic variants are concentrated at rarer MAF, but some do rise to higher frequencies.

We next tested whether particular subsets of genes showed a higher burden of damaging biallelic genotypes (Supplementary Table 1). A set of 903 curated DD-associated recessive

genes showed significantly higher enrichment of biallelic LoF genotypes than other genes (OR=4.8; $p=4\times10^{-7}$ for EABI and PABI combined). Indeed, 48% of the observed excess of damaging biallelic genotypes was in these known DD-associated recessive genes. We also found a nominally significantly higher biallelic burden in genes annotated by ExAC as having high probability of being intolerant of LoFs in the recessive state (pRec>0.9) [9], and in genes that were sub-viable when knocked out homozygously in mice [10]. By contrast, we did not observe any 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 refined the method we previously developed [7] for estimating the proportion of probands who have a diagnostic variant in a particular class (see Methods). Our new method accounts for the fact that some of the variants expected by chance are actually causal; thus, it gives higher estimates than we previously reported for *de novo* mutations. We estimated that 3.6% of EABI probands have a recessive coding diagnosis, compared to 49.9% with a *de novo* coding diagnosis. In the PABI subset, recessive coding genotypes likely explain 30.9% of individuals, compared to 29.8% for *de novo* coding mutations. The contribution from recessive variants was nearly four times as high in EABI probands with affected siblings than those without affected siblings (12.0% versus 3.2%), and highest in PABI probands with high autozygosity (47.1%) (Figure 2). Supplementary Table 2 shows the 95% confidence intervals of these diagnostic fractions for different consequence classes in different sample subgroups. These estimates rely on another parameter, the proportion of genotypes in a particular class that are pathogenic (Supplementary Figure 4), but in fact, they are not very sensitive to this (see Methods).

Discovery of new recessive disease genes

In order to discover new recessive genes, we next tested each gene in either EABI alone or EABI+PABI for an excess of biallelic genotypes. We tested four combinations of the consequence categories described above (Methods) because, in some genes, biallelic LoFs

might be embryonic lethal and LoF/damaging missense compound heterozygotes might cause DD, but in other genes, including rare damaging missense variants in the analysis might drown out signal from truly pathogenic LoFs.

Three genes passed stringent Bonferroni correction ($p<3.4\times10^{-7}$, accounting for 8 tests for each of 18,630 genes), of which one, *THOC6*, is an established recessive DD-associated gene [11–13]. Thirteen additional genes had p-value$<10^{-4}$ (Table 1), eleven of which are known recessive DD-associated genes, and the distribution of p-values for all known recessive DD-associated genes was shifted significantly lower than that of all other genes (Kolmogorov-Smirnov test; $p<1\times10^{-15}$; Supplementary Figure 5). Summary statistics for all genes are given in Supplementary Table 3. For six of the genes in Table 1, one or more families had affected siblings who shared the biallelic genotypes, supporting their pathogenicity. Patients with biallelic damaging genotypes in *THOC6*, *CNTNAP1*, *KIAA0586*, and *MMP21* were significantly more phenotypically similar to each other than expected by chance (phenotypic p-value given in Table 1). Taken together, these observations validate our gene discovery approach, and suggest that our genome-wide significance threshold is likely conservative.

We observed five probands with an identical homozygous missense variant in *EIF3F* ($p=1.2\times10^{-10}$) (ENSP00000310040.4:p.Phe232Val), which is predicted to be deleterious by SIFT, polyPhen and CADD. There were an additional four individuals in the DDD cohort who were also homozygous for this variant but who had been excluded from our discovery analysis: two were siblings of distinct index probands, one had a potentially diagnostic inherited X-linked variant in *HUWE1* (subsequently deemed to be benign since it did not segregate with disease in his family), and one had no parental genetic data available. 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.

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 [14]. The secondary structure, domain architecture and 3D fold of EIF3F is well conserved between species but sequence similarity is low (29% between yeast and humans) (Figure 3A). The highly conserved Phe232 side chain is buried (solvent accessibility 0.7%) and likely plays a stabilising role, perhaps in conjunction with two other conserved aromatic amino acids (Figure 3B). The loss of the aromatic side chain in the Phe232Val variant would likely disrupt protein stability. Further work will be needed to understand how the Phe232Val variant affects EIF3F function, and how this causes DD.

All nine individuals homozygous for the *EIF3F* variant had ID and six individuals had seizures (Supplementary Table 4). Affected individuals for whom photos were available did not have a distinctive facial appearance (Supplementary Figure 6). Features observed in three or more unrelated individuals were behavioural difficulties and sensorineural hearing loss. One of these individuals was previously published in a case report [15]. The phenotype in our patients appears distinct from the previously reported neurodegenerative phenotypes associated with variants in genes encoding eIF2B subunits [14]. Notably, one patient had skeletal muscle atrophy (Supplementary Figure 6), which is only reported in one other proband in the DDD study; in mice, *Eif3f* has been shown to play a role in regulating skeletal muscle size via interaction with the mTOR pathway [16]. None of the other individuals were either assessed to have or previously recorded to have muscle atrophy.

The second new recessive gene we identified was *KDM5B* ($p=1.1\times10^{-7}$) (Figure 4). *KDM5B* encodes a histone H3K4 demethylase. Other H3K4 methylases (*KMT2A, KMT2C, KMT2D, SETD1A*), demethylases (*KDM1A, KDM5A, KDM5C*), and two related reader proteins (*PHF21A, PHF8*) are known to cause neurodevelopmental disorders [17–19]. 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. Curiously, *KDM5B* is also enriched for *de novo* mutations in the DDD cohort [7] ($p=5.1\times10^{-7}$). Additionally, we saw nominally significant over-transmission of LoF variants from parents (p=0.002 including all families; p=0.02 when biallelic trios were excluded; transmission-disequilibrium test). This suggests that heterozygous LoFs in *KDM5B* confer an increased risk of DD but are not fully penetrant, which is consistent with the observation of 22 LoF variants in ExAC (pLI = 0), very unusual for dominant DD genes. 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 Methods and Supplementary Figure 7), nor of potentially modifying coding variants in likely interactor genes. There was also no evidence from the annotations in Ensembl (Figure 4B) or GTex data (https://gtexportal.org/home/) that the pattern could be explained by some LoFs being evaded by alternative splicing. We ran methylation arrays to search for an epimutation that might be acting as a modifier in the apparently monoallelic LoF carriers, but found none (Supplementary Figure 8). Together, these different lines of evidence suggest that heterozygous LoFs in *KDM5B* are pathogenic with incomplete penetrance, while homozygous LoFs are, as far as we can tell, fully penetrant.

The four individuals with biallelic *KDM5B* variants have ID and variable congenital abnormalities (Supplementary Table 5), in line with those seen in other disorders of the histone machinery [20]. 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 (e.g. Figure 4C). Structural abnormalities observed were agenesis of the corpus callosum and a cardiac defect each in one individual. However, in contrast to other disorders of the histone machinery where growth is often promoted or restricted, 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 latter group. Of

the 26 probands with inherited LoFs in *KDM5B*, five of them were reported to have a parent who had at least one clinical phenotype shared with the child (two mothers, three fathers). However, for only two families was this the parent who carried the LoF. There was no evidence for a parent-of-origin bias in which parent transmitted the LoF. Thus, the reason for the apparent incomplete penetrance of *KDM5B* LoF variants warrants further investigation.

Discussion

Despite the fact that there are more than twice as many known recessive than dominant DD genes, we found that recessive coding variants explain a much smaller fraction of patients in the DDD study than *de novo* dominant mutations. This is consistent with the fact that consanguinity is very rare in the UK, except in certain communities such as British Pakistanis [21,22]. There are few comparable quantitative estimates of the contribution of recessive coding causes to DD, but our estimate in the PABI subset (30.9%) is similar to the 31.5% reported by genetics clinics in Kuwait [23], which also has high levels of consanguinity. The proportion of all DD patients in the UK with a recessive coding cause is probably higher than our estimate because some recessive DDs are more easily diagnosed through current standard of care than dominant ones, and therefore are less likely to be recruited to a research study. For example, a consanguineous family history or the presence of multiple affected siblings prompt clinicians to consider recessive disorders, and recessive disorders of metabolism can often be diagnosed via biochemical testing.

There are also several reasons we might be underestimating the true burden of recessive coding causes within the DDD study. For example, it may be that the DDD parents are already enriched for damaging coding variants compared to the general population, and so use of these individuals as controls overestimates the population frequency of such variants. However, we made this more conservative choice because when we initially tried to use ExAC as controls, we

found that the number of observed rare biallelic synonymous genotypes in the DDD probands was significantly different from the expected number calculated using the ExAC frequencies [9]. We presume 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.

South Asian populations have been highlighted as particularly promising for discovering recessive genes, both because of high levels of autozygosity and increased frequency of pathogenic alleles due to bottlenecks in certain groups [24]. Despite this expectation, and the substantially higher burden of biallelic genotypes in the PABI subset versus EABI (Figure 2), they contributed little to our new gene discovery. While partially due to modest sample size, this was exacerbated by the consistent overestimation of rare variant frequencies in the small number of parents (700). Given the strong population structure in South Asia [25], it will be essential to have large, appropriately ancestry-matched control sets in future studies.

Studies in highly consanguineous populations would also allow investigation of the different ways that autozygosity may contribute to risk of rare genetic disorders. We previously showed that high autozygosity was significantly associated with lower risk of having a pathogenic *de novo* coding mutation in a known gene for DD [7]. This association was still significant once we controlled for the presence of at least one likely damaging biallelic genotype and other known factors (see Methods) ($p=0.003$). This suggests that autozygosity may increase the risk of DD via mechanisms other than a single homozygous coding variant, such as through the cumulative effect of multiple coding and/or noncoding variants. However, since overall autozygosity and the number of biallelic coding variants are correlated, it is difficult to disentangle these.

Neither of the new genome-wide significant genes we discovered in this analysis (*EIF3F* and *KDM5B*) would have been found by the traditional approach of collecting unrelated patients with the same highly recognisable disorder, because damaging biallelic genotypes in these

genes result in nonspecific and heterogeneous phenotypes. It is possible they could have been identified in large consanguineous families, although the *EIF3F* variant is much rarer in South Asians than non-Finnish Europeans in ExAC, so would be harder to find in the former population. In addition to its heterogeneous presentation, *KDM5B* is also unusual for a recessive gene because heterozygous LoFs appear to be be pathogenic with incomplete penetrance. Several *de novo* missense and LoF mutations in *KDM5B* had previously been reported in individuals with autism or ID [26–28], but LoFs had also been observed in unaffected individuals [27]. Disorders of the histone machinery normally follow autosomal dominant inheritance with *de novo* mutations playing a major role [20], so it is surprising that so many asymptomatic DDD parents carry LoFs in *KDM5B* (Figure 4). The other genes encoding H3K4 methylases and demethylases reported to cause dominant DD [19] all have a pLI score >0.99 and a very low pRec, in stark contrast to *KDM5B* (pLI=5×10^{-5}; pRec>0.999). LoFs in some other dominant ID genes appear to be incompletely penetrant [29], as do several microdeletions [30]. So far, the evidence suggests that biallelic LoFs in *KDM5B* are fully penetrant in humans, but interestingly, the homozygous knockout does show incomplete penetrance in mice, with only one strain presenting with neurological defects [31,32].

There are other examples of DD genes that show both biallelic and monoallellic inheritance, such as *NALCN* [33–35], *MAB21L2* [36], *ITPR1* [37,38] and *NRXN1* [39,40]. In *NALCN*, *MAB21L2* and *ITPR1*, heterozygous missense variants are thought to be activating or dominant-negative, so neither mirrors the situation in *KDM5B* in which we see biallelic LoFs, *de novo* LoFs, and *de novo* missense mutations that do not obviously cluster in the protein (p=0.437; method described in [41]). *NRXN1* is more similar: biallelic LoFs cause Pitt-Hopkins-like syndrome type 2 [40], which involves severe ID, whereas heterozygous deletions have been shown to predispose to a broad spectrum of neuropsychiatric disorders [39,40,42–46] with reduced penetrance and mild or no ID, but also to cause severe ID [44]. 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.

In summary, we have identified two new genome-wide significant recessive genes for DD (*EIF3F* and *KDM5B*). Additionally, we have shown that recessive coding variants make only a minor contribution to severe undiagnosed DD in EABI patients, but a much larger contribution in PABI patients. Our results suggest that identifying all the recessive DD genes would allow us to diagnose a total of 5.2% of the EABI+PABI subset of DDD, whereas identifying all the dominant DD genes would yield diagnoses for 48.6%. This has important implications for informing priors in clinical genetics. 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 noncoding recessive variants, as well as oligogenic and polygenic inheritance.

Online Methods

Family recruitment

Family recruitment has been described previously [6]. 7,832 trios from 7,448 families and 1,791 patients without parental samples were recruited at 24 clinical genetics centres within the United Kingdom National Health Service and the Republic of Ireland. Families gave informed consent to participate, and the study was approved by the UK Research Ethics Committee (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 patients were systematically phenotyped: detailed developmental phenotypes were recorded using Human Phenotype Ontology (HPO) terms [47], and growth measurements, family history, developmental milestones etc. were collected using a standard restricted-term questionnaire within DECIPHER [48]. DNA was collected from saliva samples obtained from the probands and their parents, and from blood obtained from the probands, then samples were processed as previously described [41].

Exome sequencing and variant quality control

Exome sequencing, alignment and calling of single-nucleotide variants and small insertions and deletions was carried out as previously described [7], as was the filtering of *de novo* mutations. For the analysis of biallelic genotypes, we chose thresholds for genotype and site filters to balance sensitivity (number of retained variants) and specificity (as assessed by Mendelian error rate and transition/transversion ratio). We removed sites with a strand bias test p-value < 0.001. We then set individual genotypes to missing if they had genotype quality < 20, depth < 7 or, for heterozygous calls, a p-value from a binomial test for allele balance < 0.001. Since the samples had undergone DNA capture with either the Agilent SureSelect Human All Exon V3 or V5 kit, we subsequently only retained sites that passed a missingness cutoff in both the V3 and the V5 samples. We found that, after setting a depth filter, the proportion of missing genotypes allowed had a more substantial effect on the number of Mendelian errors than genotype quality and allele balance cutoffs (Supplementary Figure 9). Thus, we ran the biallelic burden analysis on two different callsets, using a 10% (strict) or a 50% (lenient) missingness filter, and found that the results were very similar. We report results from the more lenient filter in this paper, since it allowed us to include more variants. Genotypes were set to missing for a trio if there was a Mendelian error, and variants were removed if more than one trio had a Mendelian error and if the ratio of trios with Mendelian errors to trios carrying the variant without a Mendelian error was greater than 0.1. If any of the individuals in a trio had a missing genotype at a variant, all three individuals were set to missing for that variant.

Variants were annotated with Ensembl Variant Effect Predictor [49] based on Ensembl gene build 83, using the LOFTEE plugin. The transcript with the most severe consequence was selected. We analyzed three categories of variant based on the predicted consequence: (1) synonymous variants; (2) loss-of-function variants (LoFs) classed as “high confidence” by LOFTEE (including the annotations splice donor, splice acceptor, stop gained, frameshift, initiator codon and conserved exon terminus variant); (3) damaging missense variants (i.e. those not classed as “benign” by PolyPhen or SIFT, with CADD>25). Variants were also annotated with MAF data from four different populations of the 1000 Genomes Project [50] (American, Asian, African and

European), two populations from the NHLBI GO Exome Sequencing Project (European Americans and African Americans) and six populations from the Exome Aggregation Consortium (ExAC) (African, East Asian, non-Finnish European, Finnish, South Asian, Latino), and an internal allele frequency generated using unaffected parents from the DDD.

Ancestry inference

We ran a principal components analysis in EIGENSOFT [51] on 5,853 common exonic SNPs defined by the ExAC project. We set genotypes with GL<20 to missing and excluded SNPs with >2% missingness, and then excluded samples with >5% missingness from this and all subsequent analyses. We calculated principal components in the 1000 Genomes Phase III samples and then projected the DDD samples onto them. We grouped samples into three broad ancestry groups (European, South Asian, and Other) as shown in Supplementary Figure 1 (right hand plots). By drawing ellipses around the densest clusters of DDD samples, we defined two narrower groups: European Ancestry from the British Isles (EABI) and Pakistani Ancestry from the British Isles (PABI).

For the burden and gene-based analysis, we primarily focused on these narrowly-defined EABI and PABI groups because it is difficult to accurately estimate population allele frequencies in more broadly defined groups. For example, in 4,942 European-ancestry probands, the number of observed biallelic synonymous variants was slightly higher than the number of expected (ratio = 1.06; $p=2.7\times10^{-4}$).

Calling autozygous regions

To call autozygous regions, we ran bcftools/roh[52] (bcftools version 1.5-4-gb0d640e) separately on the different broad ancestry groups. We LD pruned our data to avoid overcalling small runs of homozygosity as autozygous regions. Because rates of consanguinity differ dramatically between EABI and PABI, we chose r^2 cutoffs for each that brought the ratio of observed to expected biallelic synonymous variants with MAF<0.01 closest to 1 (see below for calculation of

the number expected): PLINK options --indep-pairwise 50 5 0.4 for EABI and --indep-pairwise 50 5 0.8 for PABI.

Defining sample subsets

We stratified probands by high autozygosity (>2% of the genome classed as autozygous), whether or not they had an affected sibling, and whether or not they already had a likely diagnostic dominant or X-linked exonic mutation (a likely damaging *de novo* mutation or inherited damaging variant in a known monoallelic DDG2P gene (http://www.ebi.ac.uk/gene2phenotype/) [if the parent was affected] or a damaging X-linked variant in a known X-linked DDG2P gene). The 4,458 patients who had no such diagnostic variants were included in the "undiagnosed" set, along with 193 patients who had biallelic genotypes in recessive DDG2P genes or potentially diagnostic variants in monoallelic or X-linked DDG2P genes but had high autozygosity or affected siblings. There were 1,366 EABI and 23 PABI probands in the diagnosed set, and 4,318 EABI and 333 PABI probands in the undiagnosed set. For the set of probands with affected siblings shown in Figures 1 and 2, we restricted to families from which more than one independent (i.e. non-MZ twin) child was included in DDD and in which the siblings' phenotypes were more similar than expected by chance given the distribution of HPO terms in the full cohort (HPO similarity p-value < 0.05 [8]).

For the burden analysis and gene-based tests, we removed 11 probands with uniparental disomy, and one individual from every pair of probands who were related (kinship > 0.044, estimated by PCRelate [53], equivalent to third-degree relatives). We also removed 924 parents reported to be affected, since one might expect these to be enriched for damaging variants compared to the general population, and 9 European parents with an abnormally high number of rare (MAF<1%) synonymous genotypes (>834, compared to the 99.9th percentile of 223), but we retained their offspring.

Burden analyses and gene-based tests

Variants were filtered on class (LoF, damaging missense or synonymous) and by different MAF cutoffs. Variants failing the MAF cutoff in any of the publicly available control populations, the full set of unaffected DDD parents, or the unaffected DDD parents in that population subset (PABI or EABI) were removed.

Following the approach we used previously [8], we calculated $B_{g,c}$, the expected number of rare biallelic genotypes of class *c* (LoF, damaging missense or synonymous) in each gene *g*, as follows:

$$E(B_{g,c}) = N_{prob}\lambda_{g,c}$$

where N_{prob} is the number of probands and $\lambda_{g,c}$ is the expected frequency of biallelic genotypes of class *c* in gene, calculated as follows:

$$\lambda_{g,c} = (1 - a_g){f_{c,g}}^2 + a_g f_{c,g}$$

where $f_{c,g}$ is the cumulative frequency of variants of class *c* in gene *g* with MAF less than the cutoff, and a_g is the fraction of individuals autozygous at gene *g*. An individual was defined as being autozygous if he/she had a region of homozygosity with any overlap of gene *g*; in practice, autozygous regions almost always overlapped genes completely rather than partially.

The rate of LoF/damaging missense compound heterozygous genotypes is:

$$\lambda_{g,LoF/miss} = (1 - a_g)[2f_{LoF,g}f_{miss,g}(1 - f_{LoF,g})]$$

To calculate the cumulative frequency of variants of class *c* in gene *g*, $f_{c,g}$, we first phased the variants in the parents based on the inheritance information. The cumulative frequency is then given by:

$$f_{c,g} = \frac{h_{c,g}}{N_{haps}}$$

where h_c is the number of parental haplotypes with at least one variant of class *c* in gene *g*, and N_{haps} is the total number of parental haplotypes.

For each gene, we calculated the binomial probability (given N_{prob} probands and rate $\lambda_{g,c}$) of the observed number of biallelic genotypes of class *c.* We did this for four consequence classes (biallelic LoF, biallelic LoF+LoF/damaging missense, biallelic damaging missense, and biallelic LoF+LoF/damaging missense+biallelic damaging missense) and for two sets of probands (EABI only, and EABI+PABI). We did not analyze PABI separately due to low power.

For the set of EABI only, we conducted a simple binomial test. For the combined EABI+PABI test, we took into account the different ways in which *n* or more probands with the relevant genotype could be distributed between the two groups and the probability of observing each combination using population-specific rates (e.g. two observed biallelic genotypes could be both seen in EABI, both in PABI, or one in each). We then summed these probabilities across all possible combinations to obtain an aggregate probability for sampling *n* or more probands by chance, as described in [8].

For some genes, $\lambda_{g,c}$ was estimated to be 0 in one or both populations because there were no variants in the parents that passed filtering. The vast majority of these also had $O(B_{g,c}) = 0$. We dropped these genes from the tests, but still included them in our Bonferroni correction. We also excluded 715 genes either because they were in the HLA region or because they were classed as having suspiciously many or suspiciously few synonymous or synonymous+missense variants in ExAC, leaving 18,630 genes. We thus set a significance threshold of 0.05/(8 tests × 18,630 genes) = $p<3.4\times10^{-7}$. For Supplementary Figure 5, we ordered the genes by their lowest p-value, randomized the order of genes with the same p-value, then tested for a difference in the distribution of ranks between recessive DDG2P genes and all other genes using a Kolmogorov-Smirnov test.

For the burden analysis, we summed up the observed and expected number of biallelic genotypes across all genes to give $O(B_c)$ and $E(B_c)$, then calculated their difference $O(B_c) -$

$E(B_c)$ (the excess) and their ratio $\frac{O(B_c)}{E(B_c)}$. Under the null hypothesis, we expect $O(B_c)$ to follow a Poisson distribution with rate $E(B_c)$.

We used a Fisher's exact test to compare the burden between different subsets of genes (Supplementary Table 1), applying it to a 2-by-2 table with the rows representing the number observed and expected.

Estimating the proportion of cases with diagnostic biallelic coding variants or *de novo* mutations

We are interested in estimating π_c, the proportion of probands with diagnostic variants of consequence class *c*. Under the null hypothesis in which none of the genotypes of class *c* are pathogenic, the number of such genotypes we expect to see in N_{pr}probands is:

$$E(b_{probands,c})_{null} = \lambda_{pr,c} N_{pr}$$

where $\lambda_{pr,c} = \sum_{g=1}^{G} \lambda_{g,c}$ is the total expected frequency of genotypes of class *c* across all genes. However, under the alternative hypothesis, suppose that some fraction $\varphi_{causal,c}$ of genotypes in class *c* cause DD, and some fraction $\varphi_{lethal,c}$are lethal. Assuming complete penetrance, we can thus split $E(b_{probands,c})$ into genotypes that are due to chance and those that are diagnostic:

$$E(b_{probands,c})_{alt} = (1 - \varphi_{causal,c} - \varphi_{lethal,c})\lambda_{pr,c} N_{pr} + \pi_c N_{pr} \frac{\varphi_{causal,c}\lambda_{pr,c}}{1 - e^{-\varphi_{causal,c}\lambda_{pr,c}}}$$

The component due to chance is $\left(1 - \varphi_{causal,c} - \varphi_{lethal,c}\right)\lambda_{pr,c} N_{pr}$ and $\frac{\varphi_{causal,c}\lambda_{pr,c}}{1-e^{-\varphi_{causal,c}\lambda_{pr,c}}}$is the average number of pathogenic genotypes per individual, given that the individual has at least one such genotype.

In N_{pa} healthy parents, biallelic genotypes of class *c* are all due to chance from the portion of *c* that is not pathogenic:

$$E\left(b_{parents,c}\right) = (1 - \varphi_{causal,c} - \varphi_{lethal,c})\lambda_{pa,c} N_{pa}$$

where $\lambda_{pa,c}$ is the expected rate of biallelic genotypes of class *c* in the parents, given the cumulative frequencies estimated in the same set of people, and the autozygosity rates. We can thus obtain a maximum likelihood estimate for $\varphi_c = \varphi_{causal,c} + \varphi_{lethal,c}$ using $O_{pa,c}$, the observed number of biallelic genotypes of class *c* in the parents:

$$\widehat{\varphi_c} = 1 - \frac{O_{pa,c}}{\lambda_{pa,c} N_{pa}}$$

The 95% confidence interval for $\widehat{\varphi_c}$ is $(1 - \frac{O_{pa,c}+1.96\sqrt{O_{pa,c}}}{\lambda_{pa,c}N_{pa}}, \frac{O_{pa,c}-1.96\sqrt{O_{pa,c}}}{\lambda_{pa,c}N_{pa}})$. We show the estimates of $\widehat{\varphi_c}$ from the EABI and PABI parents in Supplementary Figure 4. To estimate π_c, we combined data from both populations for MAF<0.01 variants to estimate $\widehat{\varphi_c}$, and obtained the following maximum likelihood estimates and 95% confidence intervals: $\varphi_{LoF/LoF} =$ 0.141(0.046,0.238), $\varphi_{LoF/miss} =$ 0.083 (-0.009,0.175), and $\varphi_{miss/miss} =$ 0.007 (-0.028,0.042).

For biallelic genotypes, we can substitute $\widehat{\varphi_c}$ into the expression above, substitute $O_{pr,c}$for $E(b_{probands,c})$ and rearrange to obtain a maximum likelihood estimate for π_c:

$$\widehat{\pi_c} = (\frac{O_{pr,c}}{\lambda_{pr,c}N_{pr}} - (1 - \varphi_c))\frac{1-e^{-\varphi_{causal,c}\lambda_{pr,c}}}{\varphi_{causal,c}} \approx (\frac{O_{pr,c}}{\lambda_{pr,c}N_{pr}} - 1 + \varphi_c)\frac{1-e^{-\varphi_{causal,c}\lambda_{pr,c}}}{\varphi_{,c}}$$

We cannot disentangle $\varphi_{causal,c}$ and $\varphi_{lethal,c}$ with the available data, but we find that the ratio $\frac{\varphi_{causal,c}}{\varphi_{lethal,c}}$ makes very little difference to the estimate of $\widehat{\pi_c}$, so we make the assumption that $\varphi_{causal,c} = \varphi_c$. The 95% confidence interval is then $[(\frac{O_{pr,c}-1.96\sqrt{O_{pr,c}}}{\lambda_{pr,c}N_{pr}} - 1 + \varphi_c)\frac{1-e^{-\varphi_{,c}\lambda_{pr,c}}}{\varphi_{,c}}, (\frac{O_{pr,c}+1.96\sqrt{O_{pr,c}}}{\lambda_{pr,c}N_{pr}} - 1 + \varphi_c)\frac{1-e^{-\varphi_{,c}\lambda_{pr,c}}}{\varphi_{,c}}]$.

De novo mutations were called as previously described [7], selecting the threshold on pp_{DNM} (posterior probability of a *de novo* mutation) such that the observed number of synonymous *de novos* matched the number expected. Using Sanger validation data from an earlier dataset [41], we adjusted the observed number of mutations to account for specificity and sensitivity as follows:

$$O_{de\ novo,adjusted} = \frac{O_{de\ novo,raw}\alpha}{0.95\beta}$$

where α is the positive predicted value (the proportion of candidate mutations that are true positive at the chosen threshold) and β is the sensitivity to true positives at the same threshold. The adjustment by 0.95 is due to exome sequencing being only about 95% sensitive. The overall *de novo* mutation rate $\lambda_{de\ novo,c,pr}$ was calculated in different sets of probands using the model from [54], adjusting for sex, as described previously [7] .

Since we cannot estimate φ_c for *de novo* mutations using the parents, as we did for recessive variants, we instead set $\varphi_{DN\ LoF}$ to 0.099, the fraction of genes with pLI>0.99. This estimate is more speculative than the directly observed depletion of biallelic genotypes above, but we note that the estimate of $\pi_{DN\ LoF}$ for the full set of 7,832 trios only increases from ~0.129 to ~0.154 if we increase $\varphi_{DN\ LoF}$ from 0.01 to 0.3. To estimate $\varphi_{DN\ missense}$, we make use of this relationship:

$$\pi_c \approx \lambda_c \varphi_c N_{BI} Pr(DDD|c)$$

where N_{BI} is the population of the British Isles, and $Pr(DDD|c)$ is the probability that an individual is recruited to the DDD given he/she has a pathogenic mutation of class *c.* If we assume that this recruitment probability is the same for *de novo* missense mutations as for *de novo* LoFs, we can write:

$$\frac{\pi_{DN\ missense}}{\pi_{DN\ LoF}} = \frac{\lambda_{DN\ missense}\varphi_{DN\ missense}}{\lambda_{DN\ LoF}\varphi_{DN\ LoF}}$$

We know $\lambda_{DN\ missense}$ and $\lambda_{DN\ LoF}$, will assume $\varphi_{DN\ LoF} = 0.099$ so we can estimate $\pi_{DN\ LoF}$, and can thus write the number of *de novo* missense mutations we expect to see as:

$$E(m_{pr,DN\ missense}) = (1 - \varphi_{DN\ missense})\lambda_{prDN\ missense}N_{pr} +$$

$$N_{pr}\frac{(\varphi_{DN\ missense}\lambda_{prDN\ missense})^2\pi_{DN\ LoF}}{(\lambda_{DN\ LoF}\varphi_{DN\ LoF})(1-e^{-\varphi_{DN\ missense}\lambda_{pr,DN\ missense}})}$$

We calculated $E(m_{pr,DN\ missense})$ for a range of values of $\varphi_{DN\ missense}$ and found that $\varphi_{DN\ missense} = 0.036$ best matched the observed data, so we used this value for estimating $\pi_{DN\ missense}$.

Effect of autozygosity on risk of having a diagnostic *de novo*

We fitted a logistic regression on all EABI and PABI probands as follows:

$$y = \beta_0 + \beta_1 F + \beta_2 I_{LoF/LoF} + \beta_3 I_{LoF/miss} + \beta_4 I_{miss/miss} + \beta_5 I_{male} + \beta_6 age_m + \beta_7 age_d + \beta_8 I_{sib}$$

where y is an indicator for having a diagnostic *de novo* coding mutation (1) or not (0), *F* is the overall fraction of the genome in autozygous segments, $I_{LoF/LoF}$, $I_{LoF/miss}$ and $I_{miss/miss}$ are indicators for the presence of at least one biallelic genotype in the relevant class, age_m and age_d are the parental ages at birth for the mother and father respectively,and I_{male} and I_{sib} are indicators for being male and having an affected sibling respectively. In this joint model, the significant covariates were F(β_1 =-10.96; p=0.003), $I_{LoF/LoF}$ (β_2= -0.92; p=3x10^{-4}), I_{male}(β_5=-0.38; p=5x10^{-10}), age_d(β_7=0.017; p=0.005) and I_{sib}(β_8=-0.87; p=1x10^{-11}). The autozygosity effect is equivalent to a ~2-fold decreased chance of having a diagnostic *de novo* for a DDD patient who is offspring of first cousins (expected autozygosity= 6.25%).

Structural analysis of EIF3F

Human EIF3:f (pdb 3j8c:f) was submitted to the Protein structure comparison service PDBeFold at the European Bioinformatics Institute [55,56]. Of the close structural matches returned, the X-ray yeast structure pdb entry 4OCN was chosen to display the human variant position, as the structural resolution (2.25Å) was better than the human EIF3:f pdb 3j8c:f structure (11.6Å) and it was the most complete structure among the yeast models. In order to map the Phe232 variant onto the equivalent position on the yeast structure, the structural alignment from PDBeFold was used. Solvent accessibility was calculated using the Naccess software [57] using the standard parameters of a 1.4Å probe radius. Amino acid sequence conservation was calculated using the Scorecons server [58] and displayed using sequence logos [59].

Validation of *KDM5B* variants by targeted re-sequencing.

We re-sequenced all *KDM5B de novo* mutations and inherited LoF variants, with the exception of two large deletions. PCR primers were designed using Primer3 to amplify the site of interest, generating approximately a 230 bp product centred on the site. PCR amplification of the targeted regions was carried out using JumpStart™ AccuTaq™ LA DNA Polymerase (Sigma-Aldrich), using 40ng of input DNA from the proband and their parents. Unique identifying tag sequences were introduced into the PCR amplicons in a second round of PCR using KAPA HiFi HotStart ReadyMixPCR Kit (KapaBiosystems). PCR amplicons were pooled and 96 products were sequenced in one MiSeq lane using 250bp paired-end reads. Reference and alternate read counts extracted from the resulting bam files and were used determine the presence of the variant in question. In addition, read data were visualised using IGV.

Transmission-disequilibrium test on *KDM5B* LoFs

We observed 15 trios in which one parent transmitted a LoF to the child, 5 trios in which one parent had a LoF that was not transmitted, 2 quartets in which one parent had a LoF that was transmitted to one out of two affected children, and 4 trios in which both parents transmitted a LoF to the child. We tested for significant over-transmission using the transmission-disequilibrium test as described by Knapp [60]. There were 7 LoFs (including one large deletion) observed in probands whose parents were not originally sequenced, which we excluded from the TDT. Of the six for which we attempted validation and segregation analysis, one was found to be *de novo* and five inherited.

Searching for coding, regulatory or epigenetic modifiers of *KDM5B*

We defined a set of genes that might modify *KDM5B* function as: interactors of *KDM5B* obtained from the STRING database of protein-protein interactions [61] (*HIST2H3A*, *MYC*, *TFAP2C*, *CDKN1A*, *TFAP2A*, *SETD1A*, *SETD1B*, *KDM1A*, *KDM2B*, *PAX9*) plus those mentioned by Klein *et al.* [62] (*RBBP4*, *HDAC1*, *HDAC4*, *MTA2*, *CHD4*, *FOXG1*, *FOXC9*), as well as all lysine demethylases, lysine methyltransferases, histone deacetylases, and SET domain-containing genes from

http://www.genecards.org/. The final list contained 95 genes. We looked for LoF or rare missense variants in these genes in the monoallelic *KDM5B* LoF carriers that might have a modifying effect, but found none that were shared by more than two of the *de novo* carriers.

We also looked for indirect evidence of a regulatory "second hit" near *KDM5B* by examining the haplotypes of common SNPs in the region (Supplementary Figure 7). DDD probands and a subset of their parents were genotyped on either the Illumina OmniExpress chip or the Illumina CoreExome chip. We performed variant and sample quality control for each dataset separately. Briefly, we removed variants and samples with high data missingness (>=0.03), samples with high or low heterozygosity, sample duplicates, individuals of African and East Asian ancestry, and SNPs with MAF<0.005. We then ran SHAPEIT2 [63] to phase the SNPs within 2Mb either side of *KDM5B*. To make Supplementary Figure 6, we used the `heatmap()` function in R to cluster the phased haplotypes using the default hierarchical clustering method (based on Euclidean distance).

We looked at methylation levels in the *KDM5B* LoF carriers to search for an "epimutation" (hypermethylation on or around the promoter) that might be acting as second hit. DNA from 64 DDD whole blood samples comprising 41 probands with a *KDM5B* variant and 23 negative controls was run on an Illumina EPIC 850K methylation array. Negative controls were selected from DDD probands with *de novo* mutations in genes not expressed in whole blood (*SCN2A*, *KCNQ2*, *SLC6A1*, and *FOXG1*), since we would not expect these to significantly impact the methylation phenotype in that tissue. Samples were randomised on the array to reduce batch effects, and were QCed using a combination of data from control probes and numbers of CpGs that failed to meet the standard detection p-value of 0.05. Based on these criteria, two samples failed and were excluded from further analysis (one of the negative controls and one of the inherited *KDM5B* LoF carriers). We analyzed a subset of CpGs in and around the *KDM5B* promoter region: the CpG island in the *KDM5B* promoter itself, and a CpG island in the promoter of *KDM5B-AS1*, a lnc-RNA not specifically associated with *KDM5B*, but also highly

expressed in the testis. We also extended analysis 5kb on either side of the start and stop sites of the *KDM5B* promoter. We examined the distribution of the beta values (the ratio of methylated to unmethylated alleles) at each of the CpGs in the 10kb region (Supplementary Figure 8).

Acknowledgements

We thank the families for their participation and patience. We are grateful to the Sanger Human Genome Informatics team, the Sample Management team, the Illumina High-Throughput team, the New Pipeline Group team, the DNA pipelines team and the Core Sequencing team for their support in generating and processing the data. We also thank Petr Danacek for help with calling the regions of homozygosity, and Kaitlin Samocha for useful discussions. 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.

Author Contributions

Exome sequence data analysis: H.C.M., J.F.M. Protein structure modelling: J.S., Methylation analysis: J.H., E.R. Clinical interpretation: W.D.J. Data processing: G.G., M.N., J.K., C.F.W. 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., J.H., J.C.B.

Competing financial interests

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

References

1. Al-Gazali, L. & Ali, B. R. Mutations of a country: a mutation review of single gene disorders in the United Arab Emirates (UAE). *Hum. Mutat.* **31,** 505–520 (2010).
2. Peltonen, L., Pekkarinen, P. & Aaltonen, J. Messages from an isolate: lessons from the Finnish gene pool. *Biol. Chem. Hoppe Seyler* **376,** 697–704 (1995).
3. Boycott, K. M. *et al.* Clinical genetics and the Hutterite population: a review of Mendelian disorders. *Am. J. Med. Genet. A* **146A,** 1088–1098 (2008).
4. Saftic, V., Rudan, D. & Zgaga, L. Mendelian diseases and conditions in Croatian island populations: historic records and new insights. *Croat. Med. J.* **47,** 543–552 (2006).
5. Ropers, H. H. Genetics of early onset cognitive impairment. *Annu. Rev. Genomics Hum. Genet.* **11,** 161–187 (2010).
6. Wright, C. F. *et al.* Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet* **385,** 1305–1314 (2015).
7. Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542,** 433–438 (2017).
8. Akawi, N. *et al.* Discovery of four recessive developmental disorders using probabilistic genotype and phenotype matching among 4,125 families. *Nat. Genet.* **47,** 1363–1369 (2015).
9. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).
10. Dickinson, M. E. *et al.* High-throughput discovery of novel developmental phenotypes. *Nature* **537,** 508–514 (2016).

11. Anazi, S. *et al.* Confirming the candidacy of THOC6 in the etiology of intellectual disability. *Am. J. Med. Genet. A* **170A,** 1367–1369 (2016).
12. Beaulieu, C. L. *et al.* Intellectual disability associated with a homozygous missense mutation in THOC6. *Orphanet J. Rare Dis.* **8,** 62 (2013).
13. Amos, J. S. *et al.* Autosomal recessive mutations in THOC6 cause intellectual disability: syndrome delineation requiring forward and reverse phenotyping. *Clin. Genet.* **91,** 92–99 (2017).
14. Fogli, A. & Boespflug-Tanguy, O. The large spectrum of eIF2B-related diseases. *Biochem. Soc. Trans.* **34,** 22–29 (2006).
15. Lynch, S. A. & Bushby, K. M. Congenital emphysema, cryptorchidism, penoscrotal web, deafness, and mental retardation--a new syndrome? *Clin. Dysmorphol.* **6,** 35–37 (1997).
16. Csibi, A. *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).
17. Shen, E., Shulha, H., Weng, Z. & Akbarian, S. Regulation of histone H3K4 methylation in brain development and disease. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **369,** (2014).
18. Singh, T. *et al.* Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders. *Nat. Neurosci.* **19,** 571–577 (2016).
19. Vallianatos, C. N. & Iwase, S. Disrupted intricacy of histone H3K4 methylation in neurodevelopmental disorders. *Epigenomics* **7,** 503–519 (2015).
20. Fahrner, J. A. & Bjornsson, H. T. Mendelian disorders of the epigenetic machinery: tipping the balance of chromatin states. *Annu. Rev. Genomics Hum. Genet.* **15,** 269–293 (2014).
21. Small, N., Bittles, A. H., Petherick, E. S. & Wright, J. Endogamy, Consanguinity and the Health Implications of Changing Marital Choices in the Uk Pakistani Community. *J. Biosoc. Sci.* **49,** 435–446 (2017).
22. Bittles, A. H. & Small, N. A. Consanguinity, Genetics and Definitions of Kinship in the Uk Pakistani Population. *J. Biosoc. Sci.* **48,** 844–854 (2016).
23. Teebi, A. S. Autosomal recessive disorders among Arabs: an overview from Kuwait. *J. Med. Genet.* **31,** 224–233 (1994).
24. Nakatsuka, N. *et al.* The promise of discovering population-specific disease-associated genes in South Asia. *Nat. Genet.* (2017). doi:10.1038/ng.3917
25. Reich, D., Thangaraj, K., Patterson, N., Price, A. L. & Singh, L. Reconstructing Indian population history. *Nature* **461,** 489–494 (2009).

26. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515,** 209–215 (2014).
27. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515,** 216–221 (2014).
28. Athanasakis, E. *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).
29. Ropers, H. H. & Wienker, T. Penetrance of pathogenic mutations in haploinsufficient genes for intellectual disability and related disorders. *Eur. J. Med. Genet.* **58,** 715–718 (2015).
30. Carvill, G. L. & Mefford, H. C. Microdeletion syndromes. *Curr. Opin. Genet. Dev.* **23,** 232–239 (2013).
31. Albert, M. *et al.* The histone demethylase Jarid1b ensures faithful mouse development by protecting developmental genes from aberrant H3K4me3. *PLoS Genet.* **9,** e1003461 (2013).
32. Zou, M. R. *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).
33. Köroğlu, Ç., Seven, M. & Tolun, A. Recessive truncating NALCN mutation in infantile neuroaxonal dystrophy with facial dysmorphism. *J. Med. Genet.* **50,** 515–520 (2013).
34. Al-Sayed, M. D. *et al.* Mutations in NALCN cause an autosomal-recessive syndrome with severe hypotonia, speech impairment, and cognitive delay. *Am. J. Hum. Genet.* **93,** 721–726 (2013).
35. Chong, J. X. *et al.* De novo mutations in NALCN cause a syndrome characterized by congenital contractures of the limbs and face, hypotonia, and developmental delay. *Am. J. Hum. Genet.* **96,** 462–473 (2015).
36. Rainger, J. *et al.* Monoallelic and biallelic mutations in MAB21L2 cause a spectrum of major eye malformations. *Am. J. Hum. Genet.* **94,** 915–923 (2014).
37. McEntagart, M. *et al.* A Restricted Repertoire of De Novo Mutations in ITPR1 Cause Gillespie Syndrome with Evidence for Dominant-Negative Effect. *Am. J. Hum. Genet.* **98,** 981–992 (2016).
38. Gerber, S. *et al.* Recessive and Dominant De Novo ITPR1 Mutations Cause Gillespie Syndrome. *Am. J. Hum. Genet.* **98,** 971–980 (2016).
39. Lowther, C. *et al.* Molecular characterization of NRXN1 deletions from 19,263 clinical microarray cases identifies exons important for neurodevelopmental disease expression. *Genet. Med.* **19,** 53–61 (2017).

40. Zweier, C. *et al.* CNTNAP2 and NRXN1 are mutated in autosomal-recessive Pitt-Hopkins-like mental retardation and determine the level of a common synaptic protein in Drosophila. *Am. J. Hum. Genet.* **85,** 655–666 (2009).
41. Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519,** 223–228 (2015).
42. Ching, M. S. L. *et al.* Deletions of NRXN1 (neurexin-1) predispose to a wide spectrum of developmental disorders. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **153B,** 937–947 (2010).
43. Huang, A. Y. *et al.* Rare Copy Number Variants in NRXN1 and CNTN6 Increase Risk for Tourette Syndrome. *Neuron* **94,** 1101–1111.e7 (2017).
44. Gregor, A. *et al.* Expanding the clinical spectrum associated with defects in CNTNAP2 and NRXN1. *BMC Med. Genet.* **12,** 106 (2011).
45. Marshall, C. R. *et al.* Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nat. Genet.* **49,** 27–35 (2017).
46. Lal, D. *et al.* Burden analysis of rare microdeletions suggests a strong impact of neurodevelopmental genes in genetic generalised epilepsies. *PLoS Genet.* **11,** e1005226 (2015).
47. Kohler, S. *et al.* Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am. J. Hum. Genet.* **85,** 457–464 (2009).
48. Bragin, E. *et al.* DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. *Nucleic Acids Res.* **42,** D993–D1000 (2014).
49. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17,** 122 (2016).
50. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
51. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38,** 904–909 (2006).
52. Narasimhan, V. *et al.* BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics* **32,** 1749–1751 (2016).
53. Conomos, M. P., Reiner, A. P., Weir, B. S. & Thornton, T. A. Model-free Estimation of Recent Genetic Relatedness. *Am. J. Hum. Genet.* **98,** 127–148 (2016).
54. Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46,** 944–950 (2014).
55. Krissinel, E. & Henrick, K. Multiple alignment of protein structures in three dimensions. *CompLife*

3695, 67–78 (2005).

56. Krissinel, E. & Henrick, K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr. D Biol. Crystallogr.* **60,** 2256–2268 (2004).
57. Hubbard, S. J. & Thornton, J. M. NACCESS-Computer Program. 1993. *Department of Biochemistry and Molecular Biology, University College London*
58. Valdar, W. S. J. Scoring residue conservation. *Proteins* **48,** 227–241 (2002).
59. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14,** 1188–1190 (2004).
60. Knapp, M. The transmission/disequilibrium test and parental-genotype reconstruction: the reconstruction-combined transmission/ disequilibrium test. *Am. J. Hum. Genet.* **64,** 861–870 (1999).
61. Szklarczyk, D. *et al.* The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.* **45,** D362–D368 (2017).
62. Klein, B. J. *et al.* The histone-H3K4-specific demethylase KDM5B binds to its substrate and product through distinct PHD fingers. *Cell Rep.* **6,** 325–335 (2014).
63. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10,** 5–6 (2013).
64. Yamamoto, G. L. *et al.* Rare variants in SOS2 and LZTR1 are associated with Noonan syndrome. *J. Med. Genet.* **52,** 413–421 (2015).
65. Akawi, N. A., Al-Jasmi, F., Al-Shamsi, A. M., Ali, B. R. & Al-Gazali, L. LINS, a modulator of the WNT signaling pathway, is involved in human cognition. *Orphanet J. Rare Dis.* **8,** 87 (2013).
66. Najmabadi, H. *et al.* Deep sequencing reveals 50 novel genes for recessive cognitive disorders. *Nature* **478,** 57–63 (2011).

Figure Legends

Figure 1: 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. Nominally significant p-values from a test of enrichment (assuming a Poisson distribution) are shown. The samples sizes are indicated in parentheses in the keys. Note that the set of probands with affected siblings includes only those whose siblings were also in DDD and appeared to share the same phenotype (see Methods).

Figure 2: Left: number of independent trio probands grouped by diagnostic category. The inherited dominant and X-linked diagnoses include only those in known genes, whereas the proportion of probands with *de novo* and recessive coding diagnoses was inferred as described in the Methods. Right: the proportion of probands in various EABI and PABI subsets inferred to have diagnostic *de novo* coding mutations or recessive coding variants.

Figure 3: 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 in fixed height characters represents the human EIF3F. Boxed characters are those 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. The blue backbone is from an X-ray structure of yeast 26S proteasome regulatory subunit RPN8 (PDB entry 4OCN), which is structurally virtually identical to human EIF3F (RMSD from PDB entry 3J8C:f <1Å). Amino acids conserved between yeast and human sequences as highlighted in panel a are shown in grey.

Figure 4: a) Summary of the damaging variants we found in *KDM5B* by mode of inheritance. b) Positions of likely damaging variants found in this and previous studies in the longest annotated transcript of *KDM5B*, ENST00000367264.2, with introns not to scale. Colours correspond to those shown in (a). There are no obvious differences in the spatial distribution of *de novo*

versus monoallelic or biallelic inherited LoFs within the gene, so it is does not seem that some are less likely to be truly LoF. The points with blue borders indicate the *de novo* mutations that had been previously reported in other studies. Two large deletions are not shown (one in a biallelic proband, another of unknown inheritance). All variants are listed in Supplementary Table 6. c) Anterior-posterior facial photographs of one of the individuals with biallelic *KDM5B* variants demonstrating narrow palpebral fissures, dark eyelashes, smooth philtrum and a thin upper vermillion border. Other affected individuals shared these features. Informed consent was obtained to publish these photographs.

Tables

Table 1: Genes enriched for damaging biallelic coding genotypes with $p<1\times10^{-4}$. The number of observed biallelic genotypes of different consequence classes is shown for the EABI and PABI probands. The lowest p-value out of the eight tests conducted is indicated, along with the details of the corresponding test (all combined: LoF + LoF/damaging missense + damaging missense) and the p-value for phenotypic similarity for the relevant probands. For all genes except *VPS13B*, the lowest p-value was achieved using EABI alone. 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 significant test	Note
	LoF	LoF/ damaging missense	damaging missense				
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	all combined	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 large 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	all combined	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[64]
LINS	2	0	0	8.2E-06	0.74	LoF	one proband has an affected sib who shares the variant; putative recessive gene[65,66]
POLR1C	0	1	2	1.4E-05	0.42	all combined	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

Supplementary Material

Supplementary Figure 1: Principal components analysis of the 1000 Genomes Phase 3 samples (left) with DDD samples projected on top of them (right). The ellipses used to define the EABI and PABI populations in DDD are shown on the PC2 versus PC3 plot.

Supplementary Figure 2: Histograms of levels of autozygosity across EABI and PABI probands.

Supplementary Figure 3: A) Burden (ratio of observed to expected) or B) excess (observed-expected) of biallelic genotypes in EABI undiagnosed probands for different MAF cutoffs. The dotted lines show 95% confidence intervals. The points for different consequence classes at the same MAF cutoff have been slightly scattered along the x-axis for ease of visualisation. Note that we do not show results for the PABI subset, because of inaccurate allele frequency estimates in this small sample.

Supplementary Figure 4: Estimates of φ, the proportion of biallelic genotypes that are lethal or cause DD. These were estimated from the parental data. See Methods for details. The points show maximum likelihood estimates and the lines show 95% confidence intervals. Points at the same MAF cutoff have been slightly scattered along the x-axis for ease of visualisation.

Supplementary Figure 5: Distribution of the ranks of minimum p-values per gene for known recessive genes versus all other genes. The order of genes with the same minimum p-value was randomised. A Kolmogorov-Smirnov (KS) test indicated that these distributions were significantly different.

Supplementary Figure 6: Anterior-posterior facial photographs of individuals with the homozygous Phe232Val variant in EIF3F. DECIPHER IDs are shown in the top right corner. Affected individuals did not have a distinctive facial appearance. Individual XXX

(leftmost) had muscle atrophy, as demonstrated in photographs of the anterior surface of the hands which show wasting of the thenar and hypothenar eminences.

Supplementary Figure 7: Plot showing haplotypes of common SNPs around *KDM5B* in individuals with *de novo* missense or LoF mutations or with monoallelic or biallelic LoFs. These is no evidence for a local haplotype shared by multiple probands with monoallelic LoFs that was not also present in an unaffected parent with a monoallelic LoF. The region shown lies between two recombination hotspots. The rows represent phased haplotypes, with orange and green rectangles corresponding to the different alleles at the SNPs at the positions indicated along the bottom. Hierarchical clustering has been applied to the haplotypes, as indicated by the dendrogram on the left, and the labels on the right indicate which individual carries the haplotype, and whether the individual was a proband carrying a *de novo* (purple), a biallelic LoF (dark green), or an inherited heterozygous LoF (yellow), or a parent carrying a heterozygous LoF (pink).

Supplementary Figure 8: Violin plots of the beta values (the ratio of methylated to unmethylated alleles) at each of the CpGs in the 10kb region around the *KDM5B* promoter. The CpGs within the *KDM5B* and *KDM5B-AS1* promoters are annotated below the plot, with coordinates relative to hg19. The bottom panel shows the negative controls (probands with likely causal *de novo* mutations in known DD genes not expressed in blood), and the other panels show probands with variants in *KDM5B* that are either biallelic (top panel), *de novo* (second panel) or monoallelic and inherited (third panel).

Supplementary Figure 9: Plots showing effect of variant filtering strategies on number of variants, Mendelian errors and Ti/Tv. We first set genotypes to missing based on genotype quality (GQ), depth (AD) and the p-value from a test of allele balance (p_{AB}), and then removed sites according to the proportion of missing genotypes.

Supplementary Table 1: Results from Fisher's Exact Tests for a difference in burden of damaging biallelic genotypes between different gene sets. The counts for EABI and PABI were combined, and we tested a 2-by-2 table in which the rows were the different gene sets and the columns were the observed and expected counts of biallelic genotypes.

Supplementary Table 2: Estimates of the π, the proportion of probands explained by diagnostic biallelic coding genotypes or *de novo* coding mutations, for different sample sets. Shown are the maximum likelihood estimates for π, and a 95% confidence interval. See Methods for how these were calculated.

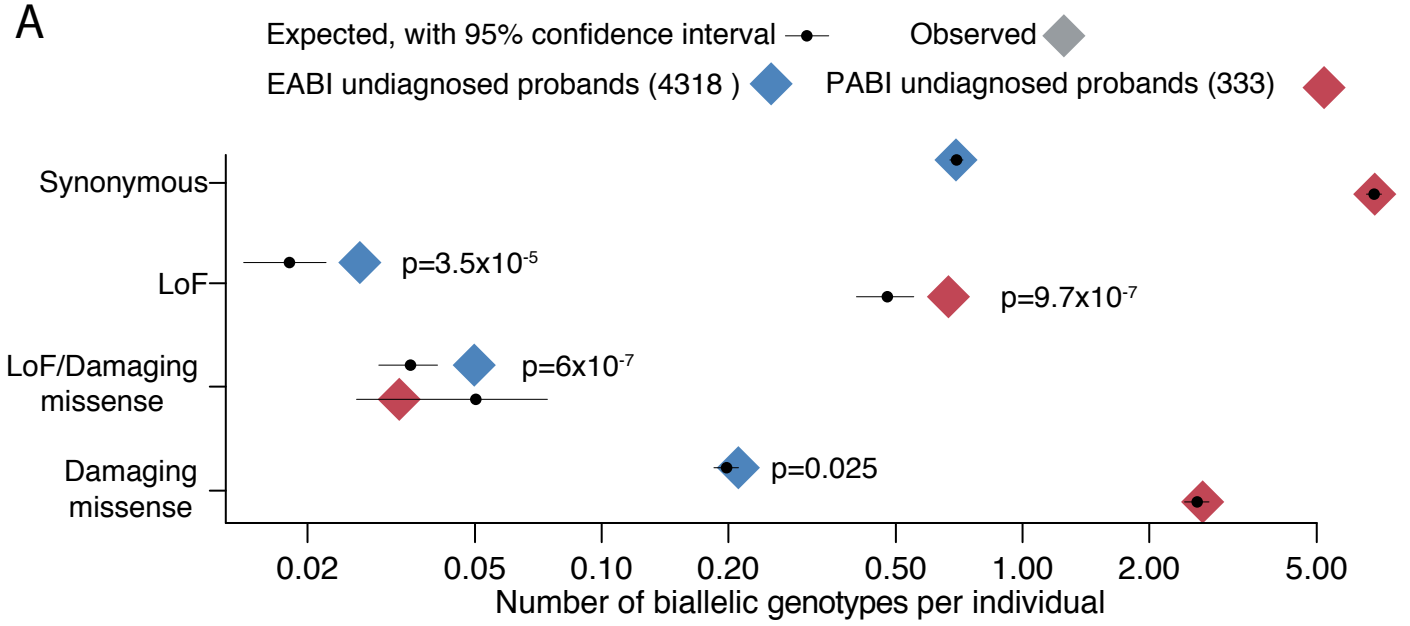
Supplementary Table 3: Results from tests of an excess of damaging biallelic genotypes for all genes. The lowest p-value out of the eight tests conducted for each gene is shown. We give results for the stringent ancestry filter (4318 EABI probands and 333 PABI probands), as shown in Table 1, as well as the lenient ancestry filter (4942 European ancestry probands and 498 South Asian ancestry probands).

Supplementary Table 4: Phenotypes of the nine patients homozygous for the *EIF3F* Phe232Val variant.

Supplementary Table 5: Phenotypes of the four probands with biallelic *KDM5B* variants.

Supplementary Table 6: *De novo* mutations in *KDM5B* from this and previous studies, and inherited LoFs in *KDM5B* from this study.

Figure 1


A
Expected, with 95% confidence interval
Observed
EABI undiagnosed probands (4318)
PABI undiagnosed probands (333)
Synonymous
LoF
LoF/Damaging missense
Damaging missense
p=3.5x10-5
p=9.7x10-7
p=6x10-7
p=0.025
0.02
0.05
0.10
0.20
0.50
1.00
2.00
5.00
Number of biallelic genotypes per individual

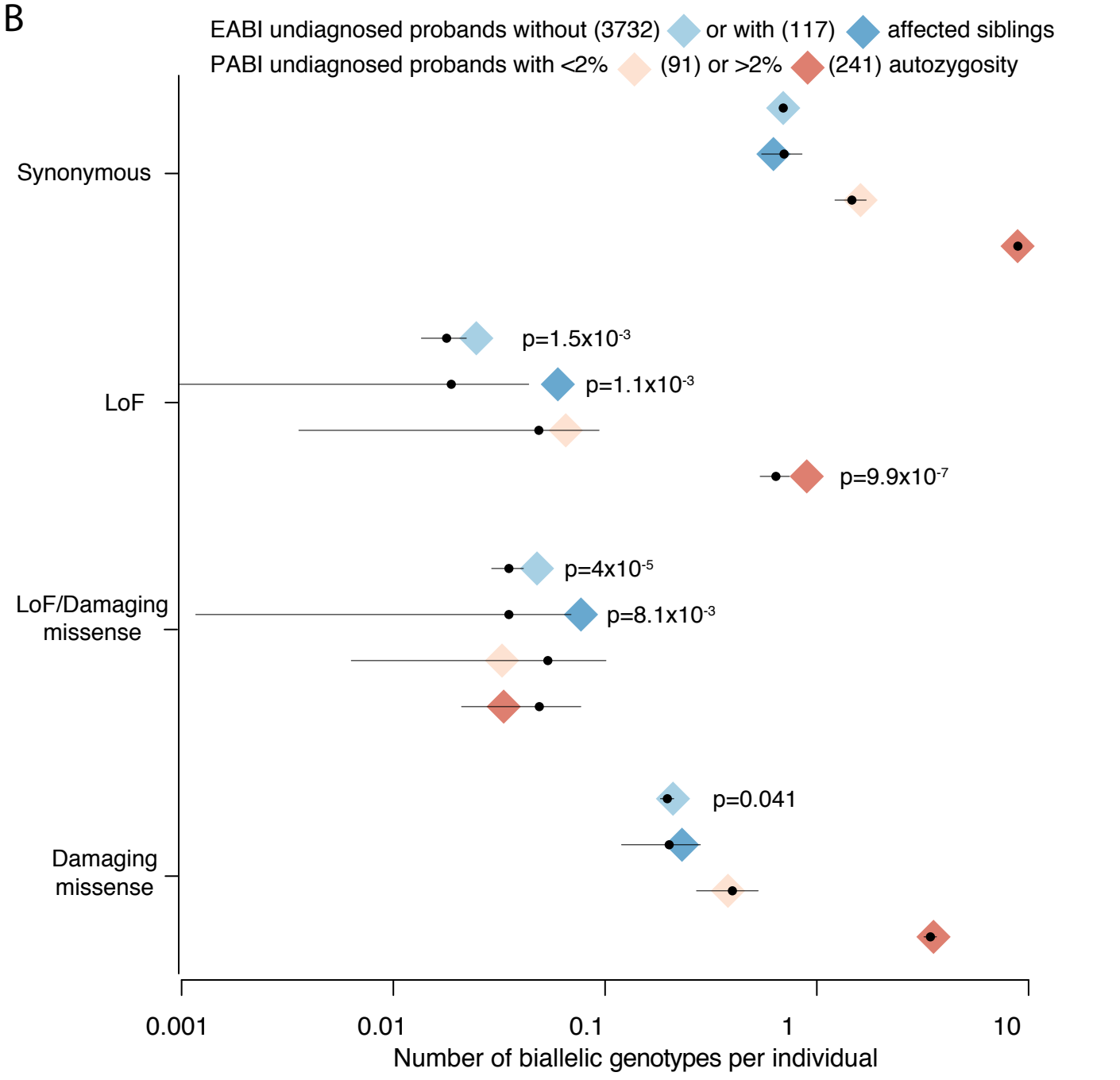

B
EABI undiagnosed probands without (3732)
or with (117)
affected siblings
PABI undiagnosed probands with <2%
(91) or >2%
(241) autozygosity
Synonymous
LoF
LoF/Damaging missense
Damaging missense
p=1.5x10-3
p=1.1x10-3
p=9.9x10-7
p=4x10-5
p=8.1x10-3
p=0.041
0.001
0.01
0.1
1
10
Number of biallelic genotypes per individual

Figure 2

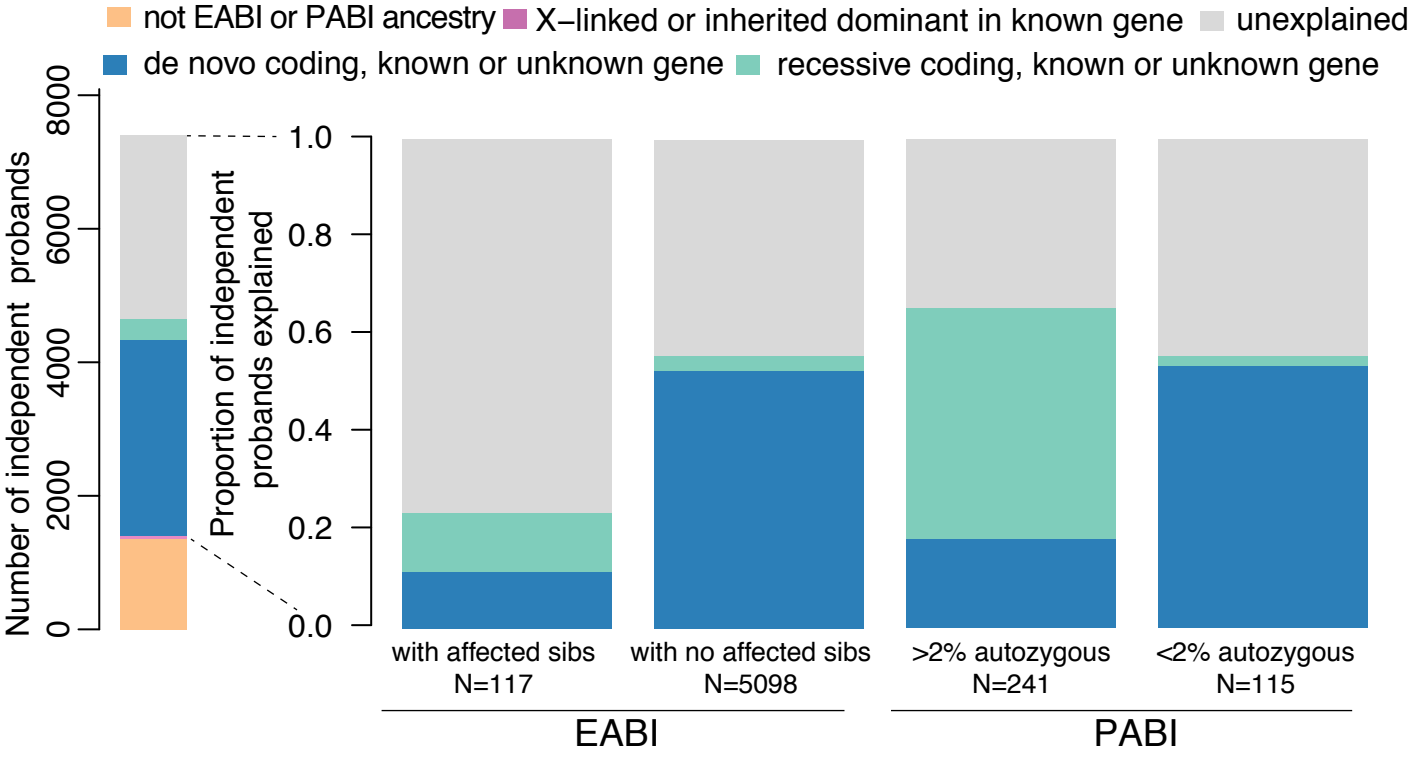
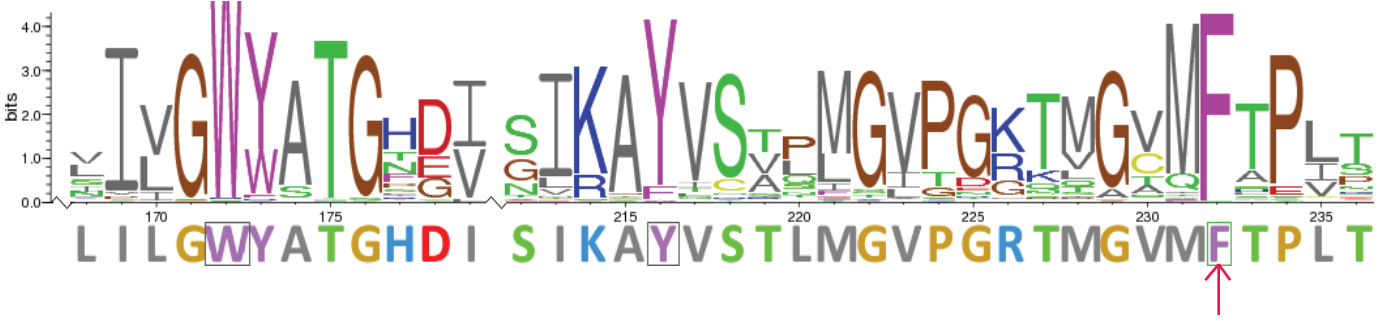

not EABI or PABI ancestry
X−linked or inherited dominant in known gene
unexplained
de novo coding, known or unknown gene
recessive coding, known or unknown gene
Number of independent probands
0
2000
4000
6000
8000
Proportion of independent probands explained
0.0
0.2
0.4
0.6
0.8
1.0
with affected sibs
N=117
with no affected sibs
N=5098
>2% autozygous
N=241
<2% autozygous
N=115
EABI
PABI

Figure 3

A


bits
4.0
3.0
2.0
1.0
0.0
170
175
215
220
225
230
235
L I L G W Y A T G H D I S I K A Y V S T L M G V P G R T M G V M F T P L T

B

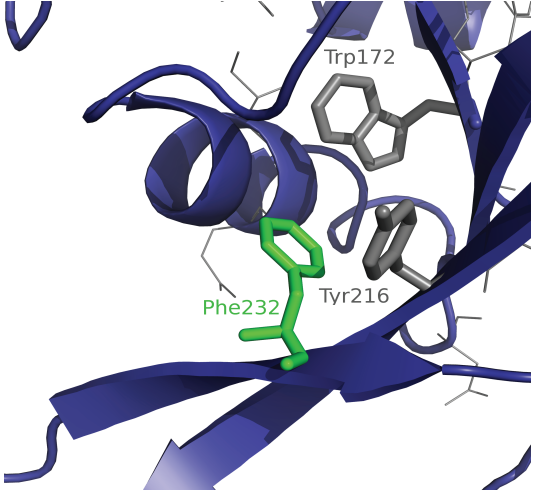
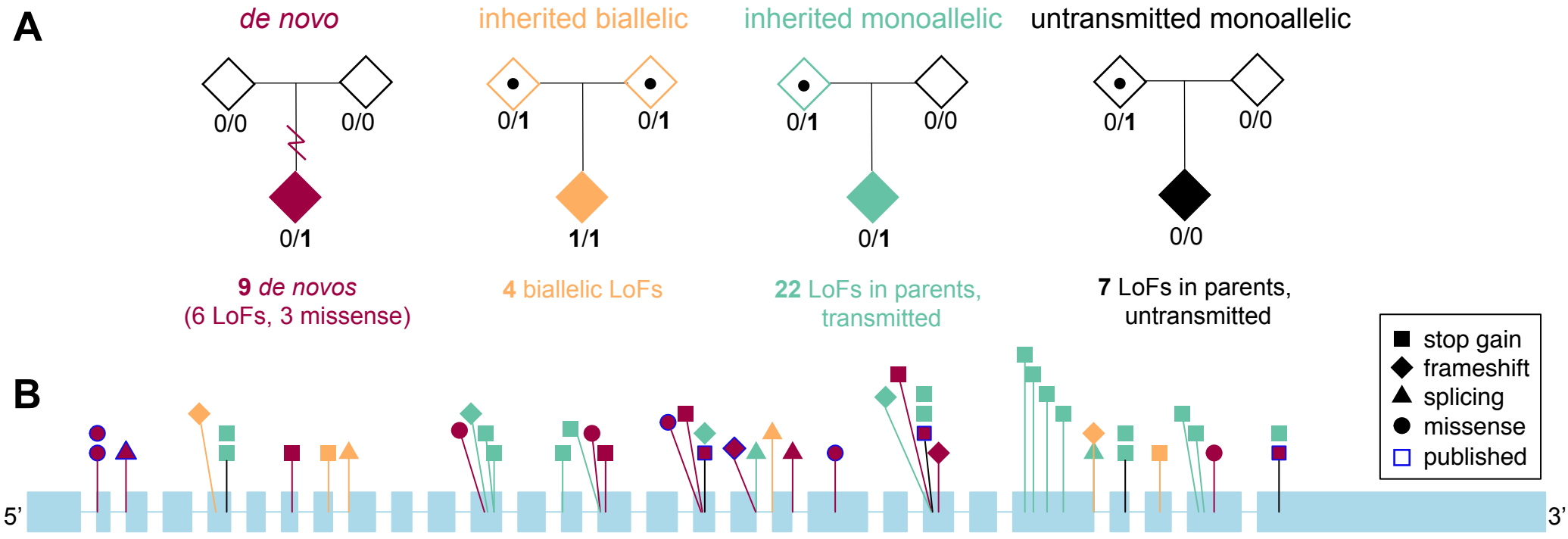

Trp172
Phe232
Tyr216

Figure 4



C

Photos will appear in published manuscript

Photos will appear in published manuscript