Large-scale discovery of novel genetic causes of developmental disorders

The Deciphering Developmental Disorders Study

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

Corresponding author:

Matthew Hurles
meh@sanger.ac.uk
Wellcome Trust Sanger Institute,
Wellcome Trust Genome Campus,
Hinxton, Cambridge, CB10 1SA, UK

Summary

Pathogenic variants in over 1,000 genes are already known to cause diverse developmental disorders (DDs) as a result of perturbed embryonic and fetal development, and yet the high prevalence of undiagnosed patients suggests that many more genetic causes remain undiscovered. We studied 1,133 children with severe, undiagnosed DDs, and their parents, using a combination of genome-wide assays to detect all major classes of genetic variation in the protein-coding portion of the genome. In addition to the 28% of children with pathogenic variants, predominantly *de novo* mutations, in 148 genes already robustly associated with DDs, we present compelling evidence for 12 recurrently mutated novel genes causing DDs in 35 individuals. Clustering of missense mutations in six of these new genes suggest an activating or dominant negative mechanism. Simulation studies and modelling of 32 candidate novel genes in zebrafish strongly suggest that many more novel genetic causes of DDs await discovery and that a genotype-driven strategy to finding them is likely to prove highly productive.

Introduction

Despite three decades of successful, predominantly ***phenotype-driven***, discovery of the genetic causes of monogenic disorders [1], up to half of children with severe developmental disorders of likely genetic origin remain without a genetic diagnosis. Especially challenging are those disorders rare enough to have eluded recognition as a discrete clinical entity, those whose clinical manifestations are highly variable, and those that are difficult to distinguish from other, very similar, disorders.

An alternative method of discovering novel genes underlying developmental disorders is suggested by the success in the past decade of discovering pathogenic chromosomal deletions and duplications across a broad range of developmental abnormalities, including both neurodevelopmental disorders and congenital malformations of different organ systems. This ***genotype-driven*** approach applies genome-wide discovery of genetic variation in large numbers of patients with diverse developmental abnormalities to identify small subsets of patients with similar pathogenic variants. This strategy has identified new disorders that could not be defined on a phenotypic basis alone [2] and expanded the phenotypic range associated with variants that had been initially discovered by focusing on patients with highly similar clinical features (e.g. [3,4]). The analysis of parent-offspring trios, which greatly facilitates the identification of *de novo* mutations, has catalyzed these discoveries.

More recently, exome sequencing of tens to hundreds of parent-offspring trios has highlighted the role of *de novo* mutations in protein-coding exons in intellectual disability, autism, schizophrenia, epilepsy and congenital heart defects [5-13]. Some of these papers have focused on diagnostic analyses of pathogenic *de novo* mutations in known disease-associated genes, while others have highlighted particular pathways enriched for pathogenic mutations, but none have achieved robust genome-wide discovery of novel genes underlying these disorders. Only with subsequent resequencing of plausible candidate genes in many additional patients has conclusive discovery of novel genes in these disorders been achieved [6,14] . We hypothesised that increasing the sample size of exome sequencing to over 1,000 trios with severe, undiagnosed developmental disorders should: (i) allow the conclusive genome-wide discovery of novel disease genes, (ii) broaden the phenotype-genotype correlation associated with known disease genes and (iii) enable the relative contribution of different classes of genetic variants to be quantified.

We have established a UK-wide network to recruit patients with diverse, severe undiagnosed developmental disorders, dominated by, but not limited to, neurodevelopmental disorders, through the 24 regional genetics services of the UK National Health Service (NHS) and Republic of Ireland. Here we describe the detailed analysis of 1,133 parent-offspring trios using a combination of exome sequencing, genome-wide SNP genotyping and exon-resolution detection of deletions and duplications using array comparative genomic hybridization (exome-aCGH), which has led to the conclusive identification of 12 novel disease genes and suggestive

evidence for a further 6 plausible disease genes for which more evidence is required to support their pathogenic role. We also present the results of zebrafish knockdown models for 32 genes, which as a set showed a striking enrichment for developmental defects compared to randomly knocked-out genes in zebrafish.

Summary of the phenotypic and genotypic data

The median age at last clinical consultation of these 1,133 children was 5.5 (age distribution shown in Extended Data Fig. 1). All children were phenotyped by clinical geneticists using the Human Phenotype Ontology (HPO). The median number of HPO terms used to describe the child's clinical features was 6, and ranged from 1 to 27, reflecting the mix of both generic and highly specific clinical presentations (Extended Data Fig. 2, Supplementary Table 1). Among the most common phenotypes were intellectual disability or developmental delay (87% of children), abnormalities revealed by cranial MRI (30%), and seizures (24%). As expected, the most common non-brain malformation was congenital heart defects (11%). Relevant family history, pertinent pregnancy and neonatal parameters, developmental milestones, and measurements of height, weight and head circumference were also recorded.

Most (849/1,101) families were sporadic (the child is the only affected family member), but 111 children had one or more parents with a developmental disorder that might share the same underlying genetic cause, and 124 had a similarly affected sibling (Extended Data Table 1).

We anticipated that prior clinical genetic testing would have diagnosed a high proportion of children with large, pathogenic CNVs, as well as those with canonical presentations of well-recognised genetic syndromes, thus enriching this research cohort of undiagnosed children for genetic causes of less well-recognised known syndromes as well as novel genetic disorders.

We exome sequenced 1,133 affected children and their parents, from 1,101 families, representing 1,071 unrelated children and 30 sibships. We also performed exome-aCGH on the children and genome-wide SNP genotyping on the children and their parents to enable analyses of deletions, duplications, uniparental disomy (UPD) and mosaic large chromosome rearrangements. After quality control, we obtained high quality exome-aCGH data on 1,009 children, and SNP genotyping data on 1,006 families.

We called single nucleotide variants (SNVs), insertion deletion events (indels), copy number variants (CNVs) and UPD from the exome data (Methods). We identified candidate *de novo* SNVs and indels for independent validation (Methods). We also detected CNVs from the exome-aCGH data and we inferred CNV inheritance and discovered UPD and mosaic large chromosomal rearrangements from SNP genotyping data. The SNVs, indels and CNVs were analysed jointly in the following analyses, allowing, for example, the identification of compound heterozygous CNVs and SNVs affecting the same gene.

The genetic ancestry of these children, is representative of the clinical population at need of diagnosis in the UK, being predominantly (~90%) of Northwest European ancestry (Extended Data Fig. 3). 51 children (4.5%) were from self-declared consanguineous unions, although a genetic analysis of parental kinship provided a more accurate picture of consanguinity, with 47 pairs of parents (4.1%) exhibiting kinship equivalent to, or in excess of second cousins, of which 33 were also self-declared consanguineous (Extended Data Table 2, Extended Data Fig. 4). The increased prevalence of (presumably recessive) developmental disorders as a result of parental relatedness is well attested [15],

In addition to the genetic data generated on these families with developmental disorders, to empower our analysis of exon-resolution deletions and duplications, we also performed exome-aCGH on 1,013 UK controls from the Scottish Family Health Study [16,17] and the UK Blood Service collection [18] (Methods).

From our exome sequencing and exome-aCGH data, we detected an average of 19,811 coding or splicing SNVs, 491 coding or splicing indels and 148 CNVs per child. From analyses of the SNP genotyping data [19] we identified 6 children with UPD and 5 children with mosaic large chromosomal rearrangements.

We identified 1,618 *de novo* variants (1,417 SNVs, 114 indels and 87 CNVs) in coding and non-coding regions, Supplementary Tables 2 and 3), of which 1,596 (98.6%) were validated using a second, independent assay, and the remainder were validated clinically. This represents an average of 1.12 *de novo* SNVs and 0.09 *de novo* indels in coding or splicing regions per child, which is within the range of previous, smaller, studies of children with developmental disorders [5-13], but slightly higher than most. The distribution of *de novo* SNVs and indels per child ranged from 0-7, and very closely approximated the Poisson distribution expected for random mutational events with little variation in mutation rate across children (Extended Data Fig. 5). These data rule out germline mutator phenotypes (greater than ten-fold increased mutation rate) in our patients, and we conclude that such mutator phenotypes, if indeed they exist, can only very rarely be a cause of developmental disorders.

Analysis of known causes of developmental disorders

To identify novel genetic causes of developmental disorders in these 1,133 children we must first identify children with pathogenic variants in genes that have been robustly associated with DD. We identified 28% (N=317) of these children with likely pathogenic variants (details in Supplementary Table 4 and ref [20]) in 1,142 known (published before Nov 2013) developmental disorder genes (24.3%), or with large deletions or duplications that encompass many genes (3.7%). The majority of these diagnoses involved *de novo* SNVs, indels or CNVs (Table 1). The single-gene diagnoses were distributed among 148 different genes, most of which (95/148) were only observed once in the 1,133 trios (Extended Data Figure 6). A handful of known, more mutable, haploinsufficient DD genes (e.g. *ARID1B*, *SATB2*, *SYNGAP1*, *ANKRD11*, *SCN1A*, *DYRK1A*, *STXBP1, MED13L*), each account for 0.5-1% of children in our

cohort. For 17 of these children (5% of diagnosed children) we identified two different genes with pathogenic variants, suggesting that some children with DD remain undiagnosed because they have two or more genetic disorders that result in a composite clinical phenotype that does not closely resemble a known condition. One example of this is a child with maternal uniparental disomy of chromosome 14 that likely accounts for her truncal obesity, short stature and intellectual disability, and a compound heterozygous pair of damaging variants in *TECTA*, on chromosome 11, which likely accounts for her hearing loss.

The diagnostic yield was higher in females (30.4%) than males (25.9%), which was largely due to an increased yield of diagnostic *de novo* mutations in females (Table 1). Although some of this effect was accounted for by *de novo* mutations in X-linked dominant genes due to the higher mutation rate of the paternally-inherited X chromosome, females also had a significantly higher diagnostic yield of autosomal *de novo* mutations (p=0.01, Fisher exact test). Larger sample sizes will be required to investigate more deeply whether there are substantive differences in genetic architecture between male and female children, as has been suggested for autism [21].

In addition to enriching for novel genetic causes of DD, this analysis of known genetic causes of DD in these children highlighted the high genetic heterogeneity in this research cohort, and suggested that the vast majority of novel genetic causes of DD observed recurrently among these 1,101 families would be *de novo* mutations in autosomal dominant or X-linked genes.

Burden analyses of classes of potentially pathogenic variants

Analyses that quantify and assess the significance of the enrichment in cases relative to controls (or null expectation) of a particular class of variation, so-called 'burden analyses', are highly informative for: (i) highlighting a particular class of variant as deserving of more detailed analysis, and (ii) estimating what proportion of a particular class of variant is likely to be pathogenic. For example, the observation of an enrichment of *de novo* deletions and duplications in children with autism [22] catalysed intensive investigation of *de novo* deletions across a broad range of neurodevelopmental disorders. However, not all of these *de novo* deletions are pathogenic, as evidenced by the presence of such events in 1-2% of children without apparent developmental disorders [23].

We observed a burden of *de novo* CNVs in the 1,133 trios, despite the fact that 77% of them had previously been screened for pathogenic CNVs using a lower resolution clinical microarray (we observed 87 in 1,133 proband trios versus 12 in 416 control trios from the Scottish Family Health Study, p-value: 0.0004). As expected, we observed a considerably lower burden of large CNVs in patients who had previously had clinical microarray testing (Extended Data Figure 7).

To evaluate the role of *de novo* SNVs and indels in causing DD, we first established a null expectation for the number of mutations of different functional consequences expected in every gene in the genome, by scaling gene-specific mutation rates that

account for gene length and sequence context [24] by the number of trios analysed (Methods). We then compared the observed number of protein-altering *de novo* SNVs and indels in known DD genes of different types to this null expectation. We observed no significant excess of any functional class of *de novo* SNVs or indels in autosomal recessive DD genes (Figure 1A). This suggests that only a small minority, if any, of the *de novo* mutations we observed in these genes are likely to be in *trans* to another damaging mutation, and thus causing the disorder observed in the child due to biallelic disruption of a gene. By contrast, we observed a highly significant excess of all 'functional' (coding and splice site variants excepting synonymous changes) classes of *de novo* SNVs and indels in dominant and X-linked DD genes (Figure 1B), within which *de novo* mutations can be sufficient to cause disease. Not all protein-altering mutations in known dominant and X-linked DD genes will be pathogenic, and these burden analyses can guide estimates of positive predictive values for different classes of mutations in known DD genes. The remaining, non-DD, genes in the genome also exhibit a more modest, but still significant, excess of functional, but not silent, *de novo* SNVs and indels (Figure 1C).

We compared the observed number of genes recurrently mutated with functional SNVs and indels in unrelated individuals with simulations derived from the mutation rate of every gene (Methods). We observed 96 such recurrently mutated genes, which represents a highly significant excess of 41 genes more than the median number (56) expected by chance from simulations (Figure 2A). This enrichment is even more pronounced (observed:29, expected:3) for genes with recurrent LoF mutations (Figure 2B). If we focus only on the *de novo* mutations in undiagnosed individuals we observe an excess of 22 (observed: 45, expected: 23) recurrently mutated genes with functional mutations (Figure 2A), and an excess of 8 (observed:9, expected:1) recurrently mutated genes with LoF mutations (Figure 2B). This significant excess of recurrently mutated genes in undiagnosed individuals strongly suggests that an appreciable fraction, but by no means all, of these recurrently mutated genes are novel DD genes.

We also included 7,237 annotated regulatory sequences in our custom exome design, comprising validated enhancers [25], the most highly conserved non-coding elements in the genome [26], and likely enhancers identified from chromatin immunoprecipitation experiments [27] (Methods). We validated 74 *de novo* SNVs and indels in these regulatory sequences. This number of *de novo* mutations is consistent with the null expectation given the size of the mutational target that these regulatory sequences represent ($p>0.05$). Compared to other sites in these targeted regulatory sequences these *de novo* mutations were not enriched for highly conserved sites ($p>0.05$). Moreover, we did not observe a significant excess of recurrently mutated regulatory sequences.

We next evaluated a possible excess of potentially pathogenic inherited SNVs in our probands. In contrast to the *de novo* analyses described above, which use expected mutation rates to evaluate significance, we require a control group to compare against, so we constructed a set of "untransmitted diplotypes", corresponding to the pair of untransmitted haplotypes for each trio at every position in the genome

(Methods). This analytical strategy has the advantage of being robust to population structure. We first considered very rare (MAF < 0.0005%) inherited LoF variants, and observed a genome-wide trend towards over-transmission to probands (p=0.015). By contrast, very rare damaging (predicted damaging by both SIFT and PolyPhen) missense variants showed no such excess. We next evaluated biallelic (homozygous and compound heterozygous) LoF variants with minor allele frequency below 5% (Supplementary Table 5), and observed a 0.56-fold depletion of such variants (p=0.04) in probands with a likely dominant cause of their disorder (either a diagnostic *de novo* mutation or an affected parent) compared to other probands. Again we saw no enrichment in biallelic damaging missense variants (Extended Data Table 3), consistent with a similar observation in children with autism [28]. We looked specifically for enrichment in the list of 1,142 known DD genes, and saw stronger enrichment (Supplementary Table 5) than genome-wide, but we still observed 1 biallelic and 34 monoallelic rare LoF SNVs in the untransmitted diplotypes. These data suggest it is incorrect to assume that any damaging variants in known DD genes discovered in DD patients are certain to be pathogenic. These observations also imply that although inherited LoF variants (both monoallelic and biallelic) are likely contributing to DD in our patients, much larger sample sizes will be required to pinpoint specific DD genes in this way.

Novel developmental disorder genes

To identify genes that were enriched in damaging *de novo* mutations, we applied two statistical tests for every gene (Methods): first we tested for an overabundance of *de novo* LoF mutations in each gene, and second we tested for an overabundance of functional *de novo* mutations that are clustered within the coding sequence of each gene (as dominant negative or activating missense variants that are often clustered in this manner). We applied these two tests to all *de novo* mutations observed in 1,130 DDD children (removing one twin from each of 3 identical twin-pairs). Moreover, to increase power to detect DD genes, we also meta-analysed our data with *de novo* mutations observed in 2,347 published trios with developmental disorders which show etiological overlap with the patients studied here (we term this the meta-DD dataset). These include neurodevelopmental disorders such as intellectual disability [6,11], epileptic encephalopathy [5], autism [8-10,12] and schizophrenia [7], as well as congenital heart defects [13]. Figure 3 shows a comparison of the statistical evidence for an enrichment of LoF and functional *de novo* mutations in the DDD and meta-DD datasets. These analyses successfully identify 20 known DD genes at genome-wide significance ($p < 1.31 \times 10^{-6}$, a Bonferroni p value of 0.05 corrected for 38,504 tests [Methods]). Despite the broad phenotypic ascertainment in our data and the meta-analysed datasets, we can detect developmental disorder genes on statistical grounds alone, without incorporating considerations of phenotypic similarity or functional plausibility.

The most significantly mutated gene is the DDD dataset is *ARID1B*, with 11 independent LoF mutations. Also of note is *PACS1* with four identical missense mutations (which is the same mutation described previously in two similar patients [29]). The high statistical significance of *PACS1*, despite it having fewer

mutations than several other genes with similar mutation rates, is strengthened by the clustering of mutations within the gene.

For some known DD genes, our data increase significantly the number of patients with mutations in a specific gene, and thus allow a fuller characterization of the phenotype associated with mutations in that gene. For example, the six patients with *MED13L* mutations represent double the number of patients previously described with mutations in this gene and are the first reported single base mutations, with the previously described patients all having large structural rearrangements [30]. *MED13L* lies within a gene desert that contains many highly conserved non-coding sequences and is flanked by *TBX5*, a gene known to be critical in cardiac development [31]. In contrast to the previously reported patients, none of the six patients identified here are known to have congenital heart defects, potentially suggesting that larger variants might be perturbing the regulatory landscape in addition to truncating the gene.

We repeated the analysis of gene-specific enrichment for mutations described above, but excluding the 317 individuals with a known cause of their developmental disorder (as described above) in order to increase our power to detect novel DD genes. In this analysis the genetic data were integrated with phenotypic similarity of patients with mutations in the same gene, available data on model organisms and functional plausibility. We identified 12 novel disease genes with compelling evidence for pathogenicity (Table 2). The statistical test for mutation enrichment exceeds the genome-wide significance threshold of 1.36×10^{-6} in 9/12 of these (Methods), with the remaining three genes (*PCGF2*, *DNM1* and *TRIO*) lying just below this significance threshold. The two children with identical Pro65Leu mutations in *PCGF2*, which encodes a component of a Polycomb transcriptional repressor complex, share a strikingly similar facial appearance representing a novel and distinct dysmorphic syndrome. *DNM1* was previously identified as a candidate gene for epileptic encephalopathy (EE) [5]. Two of the three children we identified with *DNM1* mutations also had seizures, and a heterozygous mouse mutant has seizures [32]. Cumulatively the evidence strongly points to *DNM1* being a novel gene for EE. In the case of *TRIO*, in addition to two *de novo* missense SNVs, we identified an intragenic *de novo* 82kb deletion of 16 exons of *TRIO*. *De novo* intragenic deletions are rarer than smaller *de novo* LoF variants (SNVs and indels) in our cohort, and thus this observation adds considerable additional genetic evidence to the pathogenicity of mutations in *TRIO*.

One striking observation among the novel disease genes is that for four genes (*PCGF2*, *COL4A3BP*, *PPP2R1A* and *PPP2R5D*), like *PACS1*, we observed identical missense mutations in unrelated, phenotypically similar, patients (Figure 4). We hypothesise that the mutations in these four genes are operating by either dominant negative or activating mechanisms. For a fifth gene, *BCL11A*, we identified highly significant clustering of non-identical missense mutations.

The three individuals with *de novo* mutations in *COL4A3BP* have identical Ser132Leu mutations in the encoded protein, an intracellular transporter of ceramide.

Phosphorylation of this specific serine residue has previously been shown in mutagenesis studies to down-regulate transporter activity from the ER to the golgi [33] and this mutation is predicted to abrogate this regulation, presumably resulting in intra-cellular imbalances in ceramide and its downstream metabolic pathways.

Of the three individuals with *de novo* mutations in *PPP2R1A*, two have identical Arg182Trp mutations and one has a nearby Pro179Leu mutation. *PPP2R1A* encodes the constant scaffolding A subunit of the Protein Phosphatase 2 complex, which also comprises a constant catalytic C subunit and a variable regulatory B subunit. Precisely these two amino acids have been previously identified as sites of driver mutations in endometrial and ovarian cancer, and mutagenesis studies have shown that mutating either of these two residues in one of the 15 HEAT domains of this protein results in impaired binding of B subunits [34] .

Three of the four individuals with *de novo* mutations in *PPP2R5D* have identical Glu198Lys mutations in the B56 domain of the encoded protein, and the other individual has a nearby Pro201Arg mutation. Intriguingly, *PPP2R5D* encodes one of the possible B subunits of the same Protein Phosphatase 2 complex described above. The tight clustering of mutations in *PPP5RD* suggest a similar mechanism of perturbing interactions between subunits of this complex, although further functional studies will be required to confirm this hypothesis.

Three individuals have non-identical but clustered mutations in *BCL11A* (Thr47Pro, Cys48Phe and His66Gln), which encodes a newly recognized member of SWI-SNF complex [35] . Many other genes (e.g. *SMARCA2*, *ARID1B*) that encode members of the same complex are known dominant DD genes. The clustering of mutations is suggestive of a gain-of-function mechanism. Some of the other known DD genes in this complex are haploinsufficient (e.g. *ARID1B*), but others operate by a gain-of-function mechanism (e.g. *SMARCA2*). This key chromatin modifying complex is a hotspot for dominant DD genes, and it is noteworthy that we also observed two *de novo* mutations (1 LoF, 1 missense) in *SMARCD1*, which encodes another member of this complex, although this does not yet represent sufficiently compelling evidence to declare it a novel DD gene.

For several of these novel DD genes, the meta-analysis integrating published data increased the significance of enrichment. For example, a total of five *de novo* LoF variants in *POGZ* were identified, two from our cohort, two from recent autism studies and one from a recent schizophrenia study.

Six genes had suggestive statistical evidence of being novel DD genes, defined as being a p value for mutation enrichment less than $1x10^{-4}$ and being plausible from a functional perspective. *NAA10* is already known to cause an X-linked recessive developmental disorder in males [36], but here we identified missense mutations in females, suggesting a different, X-linked dominant, disorder. We expect that the majority of these genes will eventually accrue sufficient evidence to meet the stringent criteria we defined above for declaring a novel DD gene (Table 3).

We did not attempt an analogous statistical analyses of genes enriched for candidate pathogenic variants under other genetic models (e.g. X-linked, autosomal recessive) as our initial burden analyses suggested that these analyses would be severely under-powered in an analysis of 1,130 patients.

Assessment of candidate genes in animal models

To help direct future, deeper, functional experiments on the non-redundant role during development of candidate genes from this study we used two approaches. First, morphant-induced phenotypes were recorded in the first 5 days of zebrafish development. Second we performed a systematic review of perturbed gene function in human, mouse, xenopus, zebrafish and drosophila. In both approaches the animal phenotypes were compared to those seen in individuals in our cohort

We undertook an antisense-based loss of function screen in zebrafish to assess 32 candidate DD genes with *de novo* LoF, *de novo* missense or biallelic LoF variants from exome sequencing (Methods and Supplementary Data Table 6). The 32 human candidate genes corresponded to 39 zebrafish orthologues. Knockdowns of these zebrafish genes were repeated at least twice and all morpholinos were co-injected with *tp53* morpholino to eliminate off-target toxicity. Successful knockdown of the targeted mRNA could be confirmed using RT-PCR for 82.4% of genes (28/34) and 9/11 (82%) of genes that were tested gave an equivalent phenotype when knocked down by a second, independent morpholino. Knock-down of at least one or a pair of zebrafish orthologues of 65.6% of candidate DD genes (21 out of 32) resulted in perturbed embryonic and larval development (Table 4, Figure 5, Extended Data Figure 8 and Supplementary Data Table 7). A recent large scale Zebrafish mutagenesis study of 1,216 randomly selected genes found that only 6% give homozygous mutant phenotypes during the same stages of development [37], while a morpholino based screen of 150 selected genes encoding co-translationally translocated (CTT) proteins gave a 12% frequency of developmental phenotypes [38], suggesting at least a five-fold enrichment of developmentally non-redundant genes among the 32 selected for modelling. We then compared the phenotypes of the zebrafish morphants to those of the DDD individuals with *de novo* mutations or biallelic LoF variants in the orthologous genes (Table 4). 11/21 (52.4%) of the genes were categorised as *strong* candidates based on phenotypic similarity (Figure 5A). 7/11 were potential microcephaly genes whose gene knockdown in zebrafish gives significant reductions in both head measurements, and neural tissue (Figure 5B, Methods). 6/21 (28.6%) genes resulted in severe morphant phenotypes which could not be meaningfully linked to patient phenotypes. As many of our candidate DD genes carried heterozygous LoF variants (*de novo* mutations), it is to be expected that the severity of LoF phenotypes in zebrafish may exceed that observed in our patient cohort. In some cases, antisense dosage adjustments helped to strengthen the phenotypic concordance between model and patient (e.g. *ETF1*, *PSD2*). The genes with proven non-redundant developmental roles can reasonably be assigned higher priority for downstream functional investigations and genetic analyses (e.g. replication studies).

Our systematic review of gene perturbation in multiple species sought both confirmatory and contradictory (e.g. homozygous knock-out mutant is healthy) evidence from other animal models for these 21 apparently developmentally important genes. We identified 16 genes with solely confirmatory data, often from multiple different organisms, none with solely contradictory data, two with both confirmatory and contradictory evidence and three with no evidence either way (Extended Data Table 4).

Discussion

Our patient cohort was selected for severe or extreme developmental phenotypes presenting early in life and presumed likely to be genetic in origin for which a diagnosis had not proved possible using routinely available clinical investigations. Despite the broad clinical ascertainment of our patient cohort, and their likely genetic heterogeneity, through analysis of 1,133 parent-offspring trios we discovered 12 novel DD genes (and re-discovered many known DD genes) simply by virtue of these genes being highly significantly enriched for damaging *de novo* mutations. These results validate our genotype-driven strategy as complementary to the traditional phenotypic-driven strategy of selecting patients with specific clinical features for detailed study, and offers a productive avenue for the discovery of novel developmental disorders that result in highly variable or indistinct clinical presentations.

Our meta-analysis with previously published studies of developmental disorders allowed us to increase power to detect novel DD genes. These discoveries highlight the common genetic etiologies that exist between diverse neurodevelopmental disorders such as intellectual disability, epilepsy, autism and schizophrenia. These observations bolster previous observations based on large deletions and duplications shared between different neurodevelopmental disorders [39].

Adding the patients with pathogenic mutations in the 12 novel DD gene we discovered to those with pathogenic mutations in known DD genes increased the diagnostic yield from 28% to 31%. What, then, are the causes of the developmental disorders in the other 69% of patients? There are no obvious indications (e.g. fewer phenotype terms, older age of recruitment) that the undiagnosed patients are any less severely affected than the diagnosed patients. We anticipate that there are many more pathogenic, monogenic, coding mutations in these undiagnosed patients that we have detected, but for which compelling statistical evidence is lacking. Evidence supporting this comes from the significant enrichment in undiagnosed patients of functional mutations in genes predicted to exhibit haploinsufficiency (Extended Data Fig. 9), as well as the strong enrichment for developmental phenotypes in the zebrafish knock-down screen.

Our study of just over 1,000 trios has only 5-10% power to detect an averagely mutable haploinsufficient gene (Figure 6), whereas, studying 10,000 trios would

provide greater than 90% power to detect most haploinsufficient DD genes. In accordance with this modeling, the known DD genes that we have re-discovered in our analyses are greatly enriched for more mutable, longer, genes, for which we would expect to have most statistical power to detect a significant enrichment of damaging mutations. As the mutational target of haploinsufficient genes is significantly larger than that for the typically more localized mutations that act by dominant negative or activating mechanisms, we think it is reasonable to extrapolate that the discovery of such pathogenic mechanisms is even further from saturation than for haploinsufficient genes. Moreover, in contrast to our success with identifying novel dominant DD genes, we were unable to identify any novel recessive DD genes with compelling statistical evidence.

Taken together, and provided some necessary modeling assumptions (Supplementary Information), these considerations suggest that analysing 10,000 trios (from a largely outbred population such as the one studied here) should enable the discovery of most haploinsufficient genes causing DDs, but that studying well in excess of 10,000 trios will be required to detect most autosomal recessive causes of developmental disorders. These analyses motivate the global sharing of minimal genotypic and phenotypic data, such as through the DECIPHER web portal [40], to provide diagnoses for patients who would otherwise remain undiagnosed. Plausibly pathogenic *de novo* SNVs, indels and CNVs, and biallelic LoF variants in genes not yet associated with disease, observed in undiagnosed patients in our cohort are shared through DECIPHER (http://decipher.sanger.ac.uk).

We identified significant differences in the genetic architecture of developmental disorders between male and female probands, but not between major phenotype subgroups. The increased burden of monogenic disease among females with neurodevelopmental disorders has only recently started to become more apparent [21,41], and our observations strengthen this proposition. The predicted corollary is that male probands might be enriched for poly/oligogenic causation; however, testing this hypothesis will require further investigation in larger cohorts.

Given our limited power to detect pathogenic mutations that act through dominant negative or activating mechanisms, it was notable that in four of our novel genes (*COL4A3BP*, *PPP2R1A*, *PPP2R5D* and *PCGF2*) we observed identical *de novo* mutations in unrelated trios. Two hypotheses might explain this observation: first, that there is a vast number of different gain-of-function mutations in the human genome, of which we are just scratching the surface in this study, or second, that these particular variants are enriched in our cohort due to these mutations conferring a positive selective advantage in the germline [42]. Analysis of larger datasets will be required to distinguish between these hypotheses, although they are not necessarily mutually exclusive.

While we have adopted a predominantly statistical approach to discovering novel DD genes, for some genes it has also proven valuable to take into consideration the phenotypic similarity between patients sharing similar mutations relative to the broader set of patients (e.g. *PCGF2*), as well as functional data on individual variants

or genes (e.g. *COL4A3BP*). While it is difficult to quantify probabilistically these phenotypic and functional sources of evidence, used judiciously in tandem with strong statistical genetics support, these sources of evidence can add value to genotype-driven analyses.

Our study of developmental disorders shares both methodological and biological similarities with a recent meta-analysis of somatically mutated 'driver' genes in cancer [43]. The 262 cancer driver genes identified in that study exhibit a highly significant four-fold enrichment for known DD genes. In keeping with this observation 3/12 of the novel DD genes identified in this study are also among those 262 cancer driver genes. This sharing of variants, genes and pathways between cancer and embryonic development leads to both concerns and opportunities: concerns about the potential for an elevated risk of cancer among children with these specific developmental disorders, although the overlap between variants observed somatically and in the germline can be minimal [44], and opportunities to accelerate developing therapies for rare developmental disorders by leveraging advances and investments in cancer drug development.

Methods Summary

1,133 patients with severe, undiagnosed, developmental disorders and their parents were recruited and systematically phenotyped at 24 clinical genetics centres within the UK National Health Service and the Republic of Ireland. Patient and parental saliva or blood-extracted DNAs were assayed with a bespoke Illumina ArrayExpress genotyping array and customized Agilent SureSelect exome sequencing targeting additional non-coding regulatory sequences. Patients were also assayed with two Agilent one-million probe Comparative Genomic Hybridisation arrays (aCGH) that collectively targeted almost all coding exons as well as a genome-wide backbone of non-coding sites. Variants were called from exome sequencing using SAMtools, GATK, Dindel and an in-house CNV calling algorithm, CoNVeX. *De novo* SNV and indel mutations were detected using the DeNovoGear software [45]. Deletions and duplications were called from aCGH using an in-house algorithm, CNsolidate. Putative *de novo* SNVs and indels were validated using capillary sequencing and *de novo* CNVs validated using a variety of independent methods. Variants were annotated with their likely functional impact using the Ensembl Variant Effect Predictor (Supplementary Information). Diagnostic pathogenic variants were identified by clinical review of candidate variants flagged by applying Mendelian inheritance rules to a clinically-curated list of genes known to cause developmental disorders (DDG2P, available at http://decipher.sanger.ac.uk/ddd#ddgenes). Gene-specific rates for different functional classes of mutation that account for the length and sequence context of the gene [24] were used to estimate: (i) the expected number of mutations in 1,133 trios in each gene, (ii) the expected number of mutations in sets of genes and (iii) the expected number of recurrently mutated genes. Genes significantly enriched for mutations predicted to have a functional impact on the encoded protein were identified using Bonferroni correction for multiple testing. Mutations observed in the 1,133 patients were analysed in isolation and in combination with published mutations in 2,347 trios with overlapping developmental disorders. Thirty-two candidate novel DD genes were modeled using morpholino knockdown in Zebrafish and morphological phenotyping over the first five days of development (Supplementary Table 6).

References

1 OMIM. *Online Mendelian Inheritance in Man, OMIM*, <http://omim.org> (2014).

2 Cooper, G. M. *et al.* A copy number variation morbidity map of developmental delay. *Nat Genet* **43**, 838-846, doi:10.1038/ng.909 (2011).

3 Nik-Zainal, S. *et al.* High incidence of recurrent copy number variants in patients with isolated and syndromic Mullerian aplasia. *J Med Genet* **48**, 197-204, doi:10.1136/jmg.2010.082412 (2011).

4 Weiss, L. A. *et al.* Association between microdeletion and microduplication at 16p11.2 and autism. *N Engl J Med* **358**, 667-675, doi:10.1056/NEJMoa075974 (2008).

5 Allen, A. S. *et al.* De novo mutations in epileptic encephalopathies. *Nature* **501**, 217-221, doi:10.1038/nature12439 (2013).

6 de Ligt, J. *et al.* Diagnostic exome sequencing in persons with severe intellectual disability. *N Engl J Med* **367**, 1921-1929, doi:10.1056/NEJMoa1206524 (2012).

7 Fromer, M. *et al.* De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179-184, doi:10.1038/nature12929 (2014).

8 Iossifov, I. *et al.* De novo gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285-299, doi:10.1016/j.neuron.2012.04.009 (2012).

9 Neale, B. M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242-245, doi:10.1038/nature11011 (2012).

10 O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246-250, doi:10.1038/nature10989 (2012).

11 Rauch, A. *et al.* Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* **380**, 1674-1682, doi:10.1016/s0140-6736(12)61480-9 (2012).

12 Sanders, S. J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237-241, doi:10.1038/nature10945 (2012).

13 Zaidi, S. *et al.* De novo mutations in histone-modifying genes in congenital heart disease. *Nature* **498**, 220-223, doi:10.1038/nature12141 (2013).

14 O'Roak, B. J. *et al.* Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* **338**, 1619-1622, doi:10.1126/science.1227764 (2012).

15 Sheridan, E. *et al.* Risk factors for congenital anomaly in a multiethnic birth cohort: an analysis of the Born in Bradford study. *Lancet* **382**, 1350-1359, doi:10.1016/s0140-6736(13)61132-0 (2013).

16 Smith, B. H. *et al.* Cohort Profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness. *Int J Epidemiol* **42**, 689-700, doi:10.1093/ije/dys084 (2013).

17 Smith, B. H. *et al.* Generation Scotland: the Scottish Family Health Study; a new resource for researching genes and heritability. *BMC Med Genet* **7**, 74, doi:10.1186/1471-2350-7-74 (2006).

18 WTCCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-678, doi:10.1038/nature05911 (2007).
19 King, D. A. *et al.* A novel method for detecting uniparental disomy from trio genotypes identifies a significant excess in children with developmental disorders. *Genome Res* **24**, 673-687, doi:10.1101/gr.160465.113 (2014).
20 Wright, C. F. *et al.* Deciphering Developmental Disorders: Clinical Genome Sequencing Implemented in a Large-Scale Rare Disease Study. *Lancet* (in review).
21 Jacquemont, S. *et al.* A higher mutational burden in females supports a "female protective model" in neurodevelopmental disorders. *Am J Hum Genet* **94**, 415-425, doi:10.1016/j.ajhg.2014.02.001 (2014).
22 Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science* **316**, 445-449, doi:10.1126/science.1138659 (2007).
23 Itsara, A. *et al.* De novo rates and selection of large copy number variation. *Genome Res* **20**, 1469-1481, doi:10.1101/gr.107680.110 (2010).
24 Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat Genet*, doi:10.1038/ng.3050 (2014).
25 Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res* **35**, D88-92, doi:10.1093/nar/gkl822 (2007).
26 Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034-1050, doi:10.1101/gr.3715005 (2005).
27 May, D. *et al.* Large-scale discovery of enhancers from human heart tissue. *Nat Genet* **44**, 89-93, doi:10.1038/ng.1006 (2012).
28 Lim, E. T. *et al.* Rare complete knockouts in humans: population distribution and significant role in autism spectrum disorders. *Neuron* **77**, 235-242, doi:10.1016/j.neuron.2012.12.029 (2013).
29 Schuurs-Hoeijmakers, J. H. *et al.* Recurrent de novo mutations in PACS1 cause defective cranial-neural-crest migration and define a recognizable intellectual-disability syndrome. *Am J Hum Genet* **91**, 1122-1127, doi:10.1016/j.ajhg.2012.10.013 (2012).
30 Asadollahi, R. *et al.* Dosage changes of MED13L further delineate its role in congenital heart defects and intellectual disability. *Eur J Hum Genet* **21**, 1100-1104, doi:10.1038/ejhg.2013.17 (2013).
31 Li, Q. Y. *et al.* Holt-Oram syndrome is caused by mutations in TBX5, a member of the Brachyury (T) gene family. *Nat Genet* **15**, 21-29, doi:10.1038/ng0197-21 (1997).
32 Boumil, R. M. *et al.* A missense mutation in a highly conserved alternate exon of dynamin-1 causes epilepsy in fitful mice. *PLoS Genet* **6**, doi:10.1371/journal.pgen.1001046 (2010).
33 Kumagai, K., Kawano, M., Shinkai-Ouchi, F., Nishijima, M. & Hanada, K. Interorganelle trafficking of ceramide is regulated by phosphorylation-dependent cooperativity between the PH and START domains of CERT. *J Biol Chem* **282**, 17758-17766, doi:10.1074/jbc.M702291200 (2007).

34 Walter, G. & Ruediger, R. Mouse model for probing tumor suppressor activity of protein phosphatase 2A in diverse signaling pathways. *Cell Cycle* **11**, 451-459, doi:10.4161/cc.11.3.19057 (2012).

35 Kadoch, C. *et al.* Proteomic and bioinformatic analysis of mammalian SWI/SNF complexes identifies extensive roles in human malignancy. *Nat Genet* **45**, 592-601, doi:10.1038/ng.2628 (2013).

36 Rope, A. F. *et al.* Using VAAST to identify an X-linked disorder resulting in lethality in male infants due to N-terminal acetyltransferase deficiency. *Am J Hum Genet* **89**, 28-43, doi:10.1016/j.ajhg.2011.05.017 (2011).

37 Kettleborough, R. N. *et al.* A systematic genome-wide analysis of zebrafish protein-coding gene function. *Nature* **496**, 494-497, doi:10.1038/nature11992 (2013).

38 Pickart, M. A. *et al.* Genome-wide reverse genetics framework to identify novel functions of the vertebrate secretome. *PLoS One* **1**, e104, doi:10.1371/journal.pone.0000104 (2006).

39 Craddock, N. & Owen, M. J. The Kraepelinian dichotomy - going, going... but still not gone. *Br J Psychiatry* **196**, 92-95, doi:10.1192/bjp.bp.109.073429 (2010).

40 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, doi:10.1093/nar/gkt937 (2014).

41 Levy, D. *et al.* Rare de novo and transmitted copy-number variation in autistic spectrum disorders. *Neuron* **70**, 886-897, doi:10.1016/j.neuron.2011.05.015 (2011).

42 Goriely, A. & Wilkie, A. O. Paternal age effect mutations and selfish spermatogonial selection: causes and consequences for human disease. *Am J Hum Genet* **90**, 175-200, doi:10.1016/j.ajhg.2011.12.017 (2012).

43 Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495-501, doi:10.1038/nature12912 (2014).

44 Zenker, M. *et al.* Expansion of the genotypic and phenotypic spectrum in patients with KRAS germline mutations. *J Med Genet* **44**, 131-135, doi:10.1136/jmg.2006.046300 (2007).

45 Ramu, A. *et al.* DeNovoGear: de novo indel and point mutation discovery and phasing. *Nature methods* **10**, 985-987, doi:10.1038/nmeth.2611 (2013).

46 Huang, N., Lee, I., Marcotte, E. M. & Hurles, M. E. Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet* **6**, e1001154, doi:10.1371/journal.pgen.1001154 (2010).

47 Helsmoortel, C. *et al.* A SWI/SNF-related autism syndrome caused by de novo mutations in ADNP. *Nat Genet* **46**, 380-384, doi:10.1038/ng.2899 (2014).

Acknowledgements

We dedicate this paper to John Tolmie for his unwavering and enthusiastic support of the DDD project, and in memory of a deeply valued friend and colleague. We are indebted to the families for their participation and patience. We thank Mark Daly and Kaitlin Samocha for access to unpublished mutation rate estimates. We are grateful to Stephan Saunders, Damian Smedley, Don Conrad, Avinash Ramu and Ni Huang for access to data and algorithms. We thank the UK National Blood Service and the Generation Scotland: Scottish Family Health Study for access to DNA from controls. Generation Scotland has received core funding from the Chief Scientist Office of the Scottish Government Health Directorates CZD/16/6 and the Scottish Funding Council HR03006. The DDD study presents independent research commissioned by the Health Innovation Challenge Fund [grant number HICF-1009-003], a parallel funding partnership between the Wellcome Trust and the Department of Health, and the Wellcome Trust Sanger Institute [grant number WT098051]. The views expressed in this publication are those of the author(s) and not necessarily those of the Wellcome Trust or the Department of Health. The study has UK Research Ethics Committee approval (10/H0305/83, granted by the Cambridge South REC, and GEN/284/12 granted by the Republic of Ireland REC). The research team acknowledges the support of the National Institute for Health Research, through the Comprehensive Clinical Research Network.

Author information

Data can be accessed at the European Genome Phenome Archive under accession number EGAS00001000775.

Author list:

Fitzgerald TW [1*], Gerety SS [1*], Jones WD [1*], van Kogelenberg M [1*], King DA [1], McRae J [1], Morley KI [1], Parthiban V [1], Al-Turki S [1], Ambridge K [1], Barrett DM [1], Bayzetinova T [1], Clayton S [1], Coomber EL [1], Gribble S [1], Jones P [1], Krishnappa N [1], Mason LE [1], Middleton A [1], Miller R [1], Prigmore E [1], Rajan D [1], Sifrim A [1], Tivey AR [1], Ahmed M [2], Akawi N [1], Andrews R [1], Anjum U [3], Archer H [4], Armstrong R [5], Balasubramanian M [6], Banerjee R [1], Baralle D [2], Batstone P [7], Baty D [8], Bennett C [9], Berg J [8], Bernhard B [10], Bevan AP [1], Blair E [11], Blyth M [9], Bohanna D [12], Bourdon L [10], Bourn D [13], Brady A [10], Bragin E [1], Brewer C [14], Brueton L [12], Brunstrom K [15], Bumpstead SJ [1], Bunyan DJ [2], Burn J [13], Burton J [1], Canham N [10], Castle B [14], Chandler K [16], Clasper S [11], Clayton-Smith J [16], Cole T [12], Collins A [2], Collinson MN [2], Connell F [17], Cooper N [12], Cox H [12], Cresswell L [18], Cross G [19], Crow Y [16], D'Alessandro M [7], Dabir T [20], Davidson R [21], Davies S [4], Dean J [7], Deshpande C [17], Devlin G [14], Dixit A [19], Dominiczak A [22], Donnelly C [16], Donnelly D [20], Douglas A [23], Duncan A [21], Eason J [19], Edkins S [1], Ellard S [14], Ellis P [1], Elmslie F [3], Evans K [4], Everest S [14], Fendick T [17], Fisher R [13], Flinter F [17], Foulds N [2], Fryer A [23], Fu B [1], Gardiner C [21], Gaunt L [16], Ghali N [10], Gibbons R [11], Gomes Pereira SL [1], Goodship J [13], Goudie D [8], Gray E [1], Greene P [24], Greenhalgh L [23], Harrison L [2], Hawkins R [25], Hellens S [13], Henderson A [13], Hobson E [9], Holden S [5], Holder S [10], Hollingsworth G [15], Homfray T [3], Humphreys M [20], Hurst J [15], Ingram S [6], Irving M [17], Jarvis J [12], Jenkins L [15], Johnson D [6], Jones D [1], Jones E [16], Josifova D [17], Joss S [21], Kaemba B [18], Kazembe S [18], Kerr B [16], Kini U [11], Kinning E [21], Kirby G [12], Kirk C [20], Kivuva E [14], Kraus A [9], Kumar D [4], Lachlan K [2], Lam W [24], Lampe A [24], Langman C [17], Lees M [15], Lim D [12], Lowther G [21], Lynch SA [26], Magee A [20], Maher E [24], Mansour S [3], Marks K [3], Martin K [19], Maye U [23], McCann E [4], McConnell V [20], McEntagart M [3], McGowan R [7], McKay K [12], McKee S [20], McMullan DJ [12], McNerlan S [20], Mehta S [5], Metcalfe K [16], Miles E [16], Mohammed S [17], Montgomery T [13], Moore D [24], Morgan S [4], Morris A [22], Morton J [12], Mugalaasi H [4], Murday V [21], Nevitt L [6], Newbury-Ecob R [25], Norman A [12], O'Shea R [26], Ogilvie C [17], Park S [5], Parker MJ [6], Patel C [12], Paterson J [5], Payne S [10], Phipps J [11], Pilz DT [4], Porteous D [22], Pratt N [8], Prescott K [9], Price S [11], Pridham A [11], Procter A [4], Purnell H [11], Ragge N [12], Rankin J [14], Raymond L [5], Rice D [8], Robert L [17], Roberts E [25], Roberts G [23], Roberts J [5], Roberts P [9], Ross A [7], Rosser E [15], Saggar A [3], Samant S [7], Sandford R [5], Sarkar A [19], Schweiger S [8], Scott C [1], Scott R [15], Selby A [19], Seller A [11], Sequeira C [10], Shannon N [19], Sharif S [12], Shaw-Smith C [14], Shearing E [6], Shears D [11], Simonic I [5], Simpkin D [1], Singzon R [10], Skitt Z [16], Smith A [9], Smith B [22], Smith K [6], Smithson S [25], Sneddon L [13], Splitt M [13], Squires M [9], Stewart F [20], Stewart H [11], Suri M [19], Sutton V [23], Swaminathan GJ [1], Sweeney E [23], Tatton-Brown K [3], Taylor C [6], Taylor R [3], Tein M [12], Temple IK [2], Thomson J [9], Tolmie J [21], Torokwa A [2], Treacy B [5], Turner C [14], Turnpenny P [14], Tysoe C [14], Vandersteen A [10], Vasudevan P [18], Vogt J [12], Wakeling E [10], Walker D [1], Waters J [15], Weber A [23], Wellesley D [2], Whiteford M [21], Widaa S [1], Wilcox S [5], Williams D [12], Williams N [21], Woods G [5], Wragg C [25], Wright M [13], Yang F [1], Yau M [17], Carter NP [1], Parker M [27], Firth HV [1,5], FitzPatrick DR [24], Wright CF [1], Barrett JC [1], Hurles ME [1**] on behalf of the DDD study

[*] Joint first authors

[**] Corresponding author

Affiliations:

[1] Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

[2] Wessex Clinical Genetics Service, University Hospital Southampton, Princess Anne Hospital, Coxford Road, Southampton, SO16 5YA, UK and Wessex Regional Genetics Laboratory, Salisbury NHS Foundation Trust, Salisbury District Hospital, Odstock Road, Salisbury, Wiltshire, SP2 8BJ, UK and Faculty of Medicine, University of Southampton

[3] South West Thames Regional Genetics Centre, St George's Healthcare NHS Trust, St George's, University of London, Cranmer Terrace, London, SW17 0RE, UK

[4] Institute Of Medical Genetics, University Hospital Of Wales, Heath Park, Cardiff, CF14 4XW, UK and Department of Clinical Genetics, Block 12, Glan Clwyd Hospital, Rhyl, Denbighshire, LL18 5UJ, UK

[5] East Anglian Medical Genetics Service, Box 134, Cambridge University Hospitals NHS Foundation Trust, Cambridge Biomedical Campus, Cambridge, CB2 0QQ, UK

[6] Sheffield Regional Genetics Services, Sheffield Children's NHS Trust, Western Bank, Sheffield, S10 2TH, UK

[7] North of Scotland Regional Genetics Service, NHS Grampian, Department of Medical Genetics Medical School, Foresterhill, Aberdeen, AB25 2ZD, UK

[8] East of Scotland Regional Genetics Service, Human Genetics Unit, Pathology Department, NHS Tayside, Ninewells Hospital, Dundee, DD1 9SY, UK

[9] Yorkshire Regional Genetics Service, Leeds Teaching Hospitals NHS Trust, Department of Clinical Genetics, Chapel Allerton Hospital, Chapeltown Road, Leeds, LS7 4SA, UK

[10] North West Thames Regional Genetics Centre, North West London Hospitals NHS Trust, The Kennedy Galton Centre, Northwick Park And St Mark's NHS Trust Watford Road, Harrow, HA1 3UJ, UK

[11] Oxford Regional Genetics Service, Oxford Radcliffe Hospitals NHS Trust, The Churchill Old Road, Oxford, OX3 7LJ, UK

[12] West Midlands Regional Genetics Service, Birmingham Women's NHS Foundation Trust, Birmingham Women's Hospital, Edgbaston, Birmingham, B15 2TG, UK

[13] Northern Genetics Service, Newcastle upon Tyne Hospitals NHS Foundation Trust, Institute of Human Genetics, International Centre for Life, Central Parkway, Newcastle upon Tyne, NE1 3BZ, UK

[14] Peninsula Clinical Genetics Service, Royal Devon and Exeter NHS Foundation Trust, Clinical Genetics Department, Royal Devon & Exeter Hospital (Heavitree), Gladstone Road, Exeter, EX1 2ED, UK

[15] North East Thames Regional Genetics Service, Great Ormond Street Hospital for Children NHS Foundation Trust, Great Ormond Street Hospital, Great Ormond Street, London, WC1N 3JH, UK

[16] Manchester Centre for Genomic Medicine, St Mary’s Hospital, Central Manchester University Hospitals NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester M13 9WL

[17] South East Thames Regional Genetics Centre, Guy's and St Thomas' NHS Foundation Trust, Guy's Hospital, Great Maze Pond, London, SE1 9RT, UK
[18] Leicestershire Genetics Centre, University Hospitals of Leicester NHS Trust, Leicester Royal Infirmary (NHS Trust), Leicester, LE1 5WW, UK
[19] Nottingham Regional Genetics Service, City Hospital Campus, Nottingham University Hospitals NHS Trust, The Gables, Hucknall Road, Nottingham NG5 1PB, UK
[20] Northern Ireland Regional Genetics Centre, Belfast Health and Social Care Trust, Belfast City Hospital, Lisburn Road, Belfast, BT9 7AB, UK
[21] West of Scotland Regional Genetics Service, NHS Greater Glasgow and Clyde, Institute Of Medical Genetics, Yorkhill Hospital, Glasgow, G3 8SJ, UK
[22] University of Edinburgh, Institute of Genetics & Molecular Medicine, Western General Hospital, Crewe Road South, Edinburgh, EH4 2XU, UK
[23] Merseyside and Cheshire Genetics Service, Liverpool Women's NHS Foundation Trust, Department of Clinical Genetics, Royal Liverpool Children's Hospital Alder Hey, Eaton Road, Liverpool, L12 2AP, UK
[24] MRC Human Genetics Unit, MRC IGMM, University of Edinburgh, Western General Hospital, Edinburgh, EH4 2XU, UK
[25] Bristol Genetics Service (Avon, Somerset, Gloucs and West Wilts), University Hospitals Bristol NHS Foundation Trust, St Michael's Hospital, St Michael's Hill, Bristol, BS2 8DT, UK
[26] National Centre for Medical Genetics, Our Lady's Children's Hospital, Crumlin, Dublin 12, Ireland
[27] The Ethox Centre, Nuffield Department of Population Health, University of Oxford, Old Road Campus, Oxford, OX3 7LF, UK

Supplementary Information is linked to the online version of the paper at www.nature.com/nature

Tables

Table 1 Breakdown of diagnoses by mode and by sex

	Female (%)	Male (%)	Total (%)
Undiagnosed	**383 (69.6%)**	**433 (74.3%)**	**816 (72.0%)**
Diagnosed	**167 (30.4%)**	**150 (25.7%)**	**317 (28.0%)**
De novo mutation	124 (22.5%)	80 (13.7%)	204 (18.0%)
chrX	*24 (4.4%)*	*5 (0.9%)*	*28 (2.6%)*
autosomal	*100 (18.2%)*	*75 (12.9%)*	*176 (15.5%)*
Autosomal Dominant*	9 (1.6%)	11 (1.9%)	20 (1.8%)
Autosomal Recessive	20 (3.6%)	26 (4.5%)	46 (4.1%)
X-linked Inherited	1 (0.2%)	19 (3.3%)	20 (1.8%)
UPD/Mosaicism	4 (0.7%)	6 (1.0%)	10 (0.9%)
Composite	9 (1.6%)	8 (1.4%)	17 (1.5%)
Total	**550**	**583**	**1133**

* Inherited from an affected parent

Table 2 Novel genes with compelling evidence for a role in DD

Evidence	Gene	*de novo*s DDD (Missense, LoF)	*de novo*s Meta (Missense, LoF)	P Value	Test	Mutation Clustering	Predicted Haploinsufficiency
De novo enrichment	*COL4A3BP*	3 (3,0)	5 (5,0)	4.10E-12	Meta	Yes	14.7%
	PPP2R5D	4 (4,0)	5 (5,0)	6.01E-12	DDD	Yes	19.7%
	ADNP	4 (0,4)	5 (0,5)	4.59E-11	Meta	No	9.8%
	POGZ	2 (0,2)	5 (0,5)	4.31E-10	Meta	No	30.0%
	PPP2R1A	3 (3,0)	3 (3,0)	2.03E-08	DDD	Yes	23.5%
	DDX3X	4 (3,1)	5 (3,2)	2.26E-07	DDD	No	12.7%
	CHAMP1	2 (0,2)	3 (0,3)	4.58E-07	Meta	No	52.9%
	BCL11A	3 (3,0)	4 (3,1)	1.03E-06	DDD	Yes	0.6%
	PURA	3 (1,2)	3 (1,2)	1.14E-06	DDD	No	9.4%
De novo enrichment + additional evidence	*DNM1*	3 (3,0)	5 (5,0)	1.43E-06	Meta	No	13.5%
	TRIO	2 (2,0)	7 (7,0)	5.16E-06	Meta	Yes	25.7%
	PCGF2	2 (2,0)	2 (2,0)	1.08E-05	DDD	Yes	37.7%

The table summarises the 12 genes with compelling evidence to be novel developmental disorder genes. The number of unrelated patients with independent functional or LoF mutations in the DDD cohort or the wider meta-analysis dataset including DDD patients is listed. The p value reported is the minimum p value from the testing of the DDD dataset and the meta-analysis dataset. The dataset that gave this minimal p value is also reported. Mutations are considered to be clustered if the p value of clustering of functional SNVs is less than 0.01. Predicted haploinsufficiency is reported as a percentile of all genes in the genome, with ~0% being highlight likely to be haploinsufficient and 100% very unlikely to be haploinsufficient, based on the prediction score described in Huang et al [46] updated to enable predictions for a higher fraction of genes in the genome. During submission, a paper was published online describing a novel developmental disorder caused by mutations in *ADNP* [47].

Table 3 Novel genes with suggestive evidence for a role in DD

Evidence	Gene	*de novo*s DDD (Missense, LoF)	*de novo*s Meta (Missense, LoF)	P Value	Test	Mutation Clustering	Predicted Haploinsufficiency
De novo enrichment + additional evidence	*NAA15*	1 (0,1)	3 (0,3)	1.64E-06	Meta	No	7.5%
	ZBTB20	3 (1,2)	3 (1,2)	4.84E-06	DDD	No	0.2%
	NAA10	2 (2,0)	3 (3,0)	8.28E-06	Meta	No	34.1%
	TRIP12	3 (1,2)	4 (2,2)	2.13E-05	Meta	No	3.8%
	USP9X	3 (1,2)	3 (1,2)	5.14E-05	DDD	No	3.8%
	KAT6A	2 (0,2)	2 (0,2)	7.91E-05	DDD	No	19.0%

The table summarises 6 genes with suggestive evidence to be novel developmental disorder genes. The number of unrelated patients with independent functional or LoF mutations in the DDD cohort or the wider meta-analysis dataset including DDD patients is listed. The p value reported is the minimum p value from the testing of the DDD dataset and the meta-analysis dataset. The dataset that gave this minimal p value is also reported. Mutations are considered to be clustered if the p value of clustering of functional SNVs is less than 0.01. Predicted haploinsufficiency is reported as a percentile of all genes in the genome, with ~0% being highly likely to be haploinsufficient and 100% very unlikely to be haploinsufficient, based on the prediction score described in Huang et al [46] updated to enable predictions for a higher fraction of genes in the genome.

Table 4 Zebrafish modeling identifies 21 developmentally important candidate genes

Gene	# patients	Variant	Patient phenotypes	Phenotypic concordance	Relevant knockdown phenotypes
BTBD9	2/1	Biallelic LoF/*De novo* Missense	Seizures, microcephaly, hypertonia	Strong	Reduced head size, brain volume
CHD3	1/2	*De novo* LoF/Missense	CNS and craniofacial defects	Strong	Abnormal head shape
DDX3X	1/3	*De novo* LoF/Missense	Moderately short stature, microcephaly, CNS defects	Strong	Reduced head size, brain volume
ETF1	1	*De novo* LoF	CNS and craniofacial defects, seizures, microcephaly, hypertelorism	Strong	Reduced head size, brain volume
FRYL	1	*De novo* LoF	Short stature, craniofacial and cardiac defects	Strong	Cardiac defects, reduced axis length
PKN2	1	*De novo* Missense	CNS, cardiac, ear, and craniofacial defects, growth retardation	Strong	Cardiac, craniofacial cartilage, and growth defects
PSMD3	1	*De novo* Missense	Microcephaly, muscular hypotonia, seizures, growth abnormality	Strong	Reduced head size and neural defects
SCGN	1	Biallelic LoF	Seizures, microcephaly, CNS defects	Strong	Reduced head size, brain volume
SETD5	1	*De novo* LoF	Seizures, CNS and cardiac defects, poor motor coordination	Strong	Reduced head size, cardiac defects, abnormal locomotion
THNSL2	2	Biallelic LoF	Microcephaly, CNS and ear defects	Strong	Reduced head size, brain volume, neural defects
ZRANB1	2	*De novo* Missense	Microcephaly, muscle defects, seizures	Strong	Reduced head size and neural defects
DPEP2	1	Biallelic LoF	CNS defects, growth retardation	Moderate	Growth reduction
PSD2	1	*De novo* LoF	CNS defects, hypertonia, seizures	Moderate	Abnormal musculature, CNS and locomotion
SAP130	1	*De novo* LoF	Short stature, hypotonia, hypotelorism	Moderate	Abnormal locomotion
CNOT1	1/1	*De novo* LoF/Missense	Short stature, cardiac, CNS, ear and craniofacial defects	Weak	Multisystem
DTWD2	1	*De novo* LoF	CNS defects, seizures	Weak	Multisystem
ILVBL	1	*De novo* LoF	CNS and craniofacial defects	Weak	Multisystem
NONO	1	*De novo* LoF	CNS and ear defects, hypotonia, growth retardation	Weak	Multisystem, with otic and growth defects
POGZ	2	*De novo* LoF	CNS and ear defects, hypotonia, seizures, coloboma	Weak	Multisystem
SMARCD1	1/1	*De novo* LoF/Missense	CNS defects, hypotonia	Weak	Multisystem
WWC1	1	*De novo* Missense	CNS defects, hypertelorism	None	None

This table summarises the 21 genes whose knockdown results in developmental phenotypes in zebrafish. "# patients" column indicates how many patients were identified as carrying variants in these genes. Split numbers indicate the breakdown of variant types (eg. for *BTBD9*, 2/1 is two biallelic LoF and one *de novo* missense carrying patients). A summary of the patient phenotypes is listed, as well as the relevant phenotypes observed in zebrafish knockdown experiments. Phenotypic concordance categories indicate the degree of overlap between the zebrafish phenotyping and the patient phenotypes. Weak concordance typically is the result of severe, multisystem phenotypes in zebrafish. See *Supplemental Materials* for more detailed phenotype information.

Figure legends

Figure 1 Expected and observed numbers of *de novo* mutations
The expected and observed numbers of mutations of different functional consequences in three mutually exclusive sets of genes are shown, along with the p value from an assessment of a statistical excess of observed mutations. The three classes of genes are described in the main text.

Figure 2 Excess of recurrently mutated genes
Each panel shows the observed number of recurrently mutated genes (diamond) and the distribution of the number of recurrently mutated genes in 10,000 simluations (boxplot) under a model of no gene-specific enrichment of mutations: **a.** all protein-altering mutations in all DDD children and undiagnosed DDD children, **b.** all LoF mutations in all DDD children and undiagnosed DDD children. Each diamond is annotated with the median excess of recurrently mutated genes, with 95% confidence intervals in brackets.

Figure 3 Gene-specific significance of enrichment for DNMs
The $-\log_{10}(p)$ value of testing for mutation enrichment is plotted only for each gene with at least one mutation in DDD children. On the X-axis is the p value of the most significant test in the DDD dataset, and on the Y-axis is the minimal p value from the significance testing in the meta-analysis dataset. Red indicates genes already known to be associated with developmental disorders (in DDG2P). Only genes with a p value of less than 0.05/18,272 (red lines) are labeled.

Figure 4 Five novel genes with clustered mutations
The domains (blue), post-translational modifications, and mutation locations (red stars) are shown for five proteins with highly clustered *de novo* mutations in unrelated children with severe, undiagnosed developmental disorders. For two proteins (COL4A3BP and PCGF2) where all observed mutations are identical, photos are shown to highlight the facial similarities of patients carrying the same mutation.

Figure 5 Candidate gene Loss of Function modeling in zebrafish reveals enrichment for developmentally important proteins.
a, Examples of developmental phenotypes: Knockdown of *pkn2a* results in reduced cartilaginous jaw structures (black arrows), knockdown of *fryl* results in cardiac and craniofacial defects (white arrowheads and arrows, respectively), while knockdown of *psmd3* results in smaller ear primordia (red arrows), and mis-patterned CNS neurons (compare red double arrows and brackets). **b**, Knockdown outcomes of 7 genes with variants present in microcephaly patients: Interocular measurements of brightfield images from control and LoF embryos reveal significant decreases in head size. A neuronal antibody stain (anti-HuC/D, green channel) labels the brains of control and morphant zebrafish. Measurements taken across the widest extent of the midbrain identify significant reductions in brain size, likely underlying the

concomitant head size reductions seen in brightfield. In **b**, tables show average percentage reduction in head and brain width, and p-values of a *t*-test.

Figure 6 Saturation analysis for detecting haploinsufficient genes
A boxplot showing the distribution of statistical power to detect a significant enrichment of LoF mutations across 18,272 genes in the genome, for different numbers of trios studied, from 1,000 trios to 12,000 trios.

Extended Data Tables (EDT)

EDT1. Family History
Self-declared family history (only first and second degree relatives recorded)

EDT2. Consanguinity
Self-declared versus consanguinity defined by identify by descent (IBD)

EDT3. Biallelic Loss of function and damaging functional variants
Rare (MAF < 5%) biallelic loss-of-function and damaging functional variants in uninherited diplotypes and probands. 'Likely dominant probands' refers to probands with a reported *de novo* mutation or affected parents, and 'other probands' to all remaining probands. 'DDG2P Biallelic' refers to confirmed and probable DDG2P genes with a biallelic mode of inheritance. See *Supplemental methods* for details of variant processing.

EDT4. Evidence of developmental role from animal models
Concordant ('C') and Contradictory ('D') data from different animal models as to the developmental role of 21 genes showing a developmental phenotype in zebrafish knockdown experiments.
[a] Damaging variant observed monoallelic ('mono') or biallelic ('bi) in patients.
[b] Concordance between phenotype in fish knockdown and patient
[c] Results of different morpholinos targeting the same gene
[d] Genome-wide significance (GWS) of mutation enrichment in patients
[e] Summary of evidence across all organisms

Extended Data Figures (EDF)

EDF1. Gestation Adjusted Decimal Age at Last Clinical Assessment
Histogram showing the distribution of the gestation adjusted decimal age at last clinical assessment across the 1133 probands. The dashed red line show the median age.

EDF2. Log10 of the Frequency of HPO Term Usage
Histogram showing the log10 of the number of times each HPO term was used within the 1133 proband patient records.

EDF3.Projection PCA plot of the 1133 probands
PCA plot of 1133 DDD probands projected onto a PCA analysis using 4 different HapMap population from the 1000 genomes project. Black: African, Red: European, Green: East Asian, Blue: South Asian and the 1133 DDD probands are represented by orange triangles.

EDF4. Self Declared and Genetically Defined Consanguinity
Overlaid histogram showing the distribution of kinship coefficients from KING

comparing parental samples for each trio. Green: Trios where consanguinity was not entered in the patient record on DECIPHER. Red: Trios consanguinity was declared in the patient record on DECIPHER.

EDF5. Number of Validated *de novo* SNVs and indels per Proband
Bar plot showing the distribution of the observed number of validated SNVs and indels per proband sample, and the expected distribution assuming a Poisson distribution with the same mean as the observed distribution.

EDF6. Number of Diagnoses per Gene
Histogram showing the number of diagnoses per gene for genes with at least two diagnoses from different proband samples.

EDF7. Burden of Large CNVs in 1133 DDD Proband Samples
Plot comparing the frequency of rare CNVs in three sample groups against CNV size. Y-axis is the on a log scale. Red: DDD probands who have not had previous microarray based genetic testing, Purple: DDD probands who have had negative previous microarray based genetic testing Green: DDD controls.

EDF8. Candidate gene Loss of Function modeling in zebrafish identifies developmentally important genes and concordance with patient phenotypes.
Each gene-specific panel includes **a**, patient information, including HPO terms, variant details (gene Ensembl id, inheritance, consequence, position and change in genome, transcript, and protein), **b,** zebrafish orthologue information including gene name, Ensembl gene id, morpholino sequence, knockdown confirmation assay, phenotypes relevant to those of patient, and concordance categories: *Strong* and *Moderate* concordance indicates specific phenotypes in animal models match those in the patient. *Weak* concordance indicates poor overlap, often due to severe multisystem phenotypes. **c** lists standardized phenotypic observations in LoF zebrafish embryos, including MO dose, stage, affected tissue or behaviour (*Entity)* and effect (*Quality*). Subsequent panels **d-m**, display pairs of size and stage matched images of control and LoF zebrafish embryos, highlighting specific phenotypes where relevant. (See *Supplemental methods* for further details)

BTBD9
btbd9 knockdown embryos show reduced body size and cardiac edema (**d** versus **e**), and microcephalic changes including reduced head size (**f,g** versus **i,j**) and smaller brain (green channel, **h** versus **k**). These head and CNS defects are concordant with the patient phenotypes.

CHD3
chd3 knockdown embryos display growth delays, including a smaller, abnormally shaped head and brain (**f** versus **g**), and curved body axis (d versus e). The CNS defects strongly suggest concordance with the patient phenotypes.

CNOT1

cnot1 knockdown embryos have numerous global developmental defects including a reduced body size at 24 hours (d versus e), malformed otic vescicle, and body axis curvature.

DDX3X
pl10 knockdown embryos show strong reductions in growth at higher MO doses (**d** versus **e**), and microcephalic changes in head size (**f,g** versus **i,j**) and a smaller brain (green channel, **h** versus **k**) at 2 days. The microcephaly and CNS changes in morphant embryos are consistent with patient phenotypes reported.

DPEP2
dpep2 knockdown embryos show severe growth delays resulting in reduced body size, axis curvature and brain size (**d** versus **e**). The reduced growth suggests possible concordance with the growth retardation reported for the patient.

DTWD2
dtwd2 knockdown embryos show lethal defects at high MO dose. At lower doses, *dtwd2* morphants have severe multisystem defects in head, brain, eye, somite and cardiac development resulting in dysmorphic embryos by day 2 (**d** versus **e**). The severity of these phenotypes precludes any assessment of concordance with the patient phenotypes(**a** versus **c**)

ETF1
At 6ng of *etf1a* MO, embryos show dramatic defects in early development including deformed notocord, somites, head, and CNS(**d** versus **e**). At lower doses of MO, embryos show a milder set of defects that include microcephalic changes in head and brain size (**h,i,j** versus **k,l,m**) consistent with patient phenotypes.

FRYL
fryl MO injected embryos show decreased body length (**b,** *relevant phenotypes***)**, malformed cardiac structures, and an abnormal head size at day 2 (**d,f** versus **e,g**). These are more pronounced at day 3, including craniofacial defects (**h** versus **i**, white arrows) and poor cardiac morphogenesis (**j** versus **k**, heart tissue in green channel). These phenotypes are highly concordant with the presence of reduced stature, cardiac, and craniofacial phenotypes reported for the patient.

ILVBL
Embryos injected with 6ng of *ilvbl* MO display severe multisystem phenotypes including absence of trunk structures, small head and eyes, reduced pigmentation, with many embryos dying by day 2 (**d** versus **e**). At lower doses, *ilvbl* morphants display brain edema, decreased body length, and a malformed head at 2 and 5 days (**f,h** versus **g,i**). The overlap of these severe zebrafish phenotypes with those of the patient is unclear, thus in weak concordance, despite an essential developmental role in the animal model.

NONO

Embryos injected with *nono* MOs show severe multisystem developmental defects (**d** versus **e**), including abnormal somite segmentation, reduced body length, decreased CNS and spinal cord volume, a short thick trunk, and cardiac malformations. Lower MO dosage reduces the severity of observed defects, while maintaining the complexity of the phenotype (**c**). This complex LoF outcome cannot be correlated to the observed patient phenotypes and is thus in weak concordance.

PKN2

Zebrafish embryos injected with *pkn2a* morpholinos show progressively more severe development and growth defects, including abnormal cardiac outflow tract (**f** versus **i**, black arrows), thinner CNS with edema (**g** versus **j**), and reductions in craniofacial cartilage affecting jaw structures (**h** versus **k**, black arrows) in a dose dependent manner. Reduced embryo length and head size reductions are apparent at days 2 and 3 (**d,g** versus **e,j**). This phenotypic spectrum is strongly concordant with the cardiac, craniofacial, and growth defects reported for the patient.

POGZ

pogza MO injected embryos display a number of defects affecting brain, eye, ear, trunk, heart, and pigment development as well as overall growth retardation (**d** versus **e**). This complex phenotype results in a weak concordance with the reported patient phenotypes.

PSD2

psd2a MO injected embryos have brain, trunk, heart and movements defects (**c**, and d versus e). Although complex, at lower MO doses specific phenotypes of defective movement and abnormal head and brain development suggest a moderate concordance with patient phenotypes.

PSMD3

psmd3 MO injected zebrafish embryos show defects in trunk, and ear development (**d,g** versus **e,j**), microcephalic changes in head and brain size (**f,h** versus **i,k**), as well as aberrant neuronal patterning (**h** versus **k**, red double arrows and brackets) at day 2. Overall, these phenotypes are strongly concordant with the described patient phenotypes.

SAP130

sap130 MO injected embryos display poor escape response to trunk touch stimuli at 2 days, and a failure to hatch from their chorions by day 5 (versus day 2 in control embryos) likely due to decreased locomotion. These movement defects are moderately concordant with the muscular phenotype in the patient's HPO terms.

SCGN

Embryos injected with *scgn* MOs show mildly reduced body size (**d** versus **e**), cardiac edema, and microcephalic changes including reduced head size (**f,g** versus **i,j**) and smaller brain (green channel, **h** versus **k**). The zebrafish head and

brain phenotypes at all doses identify a strong concordance with the patient phenotypes.

SETD5
setd5 MO injected embryos display cardiac and head size defects at day 2 (**f** versus **g**), as abnormal escape response locomotion (see *Supplemental video files*). These phenotypes are strongly concordant with the reported patient phenotypes (**c** versus **a**).

SMARCD1
High dose *smarcd1* MO injected embryos show severe phenotypes resulting in death by day 1. At lower doses, embryos show severe multisystem phenotypes including CNS, trunk, and heart defects (**d** versus **e**, and **c**).

THNSL2
thnsl2 knockdown embryos show microcephalic changes including reduced head size (**f,g** versus **i,j**) and smaller brain (green channel, **h** versus **k**). These phenotypes are consistent with the observed patients' microcephaly and CNS defects.

WWC1
wwc1 MO injected embryos have defects in ear primordia development (reduced otolith number, **d** versus **e**) and an abnormal escape response at day 2. These defects have no concordance with the patient phenotypes reported.

ZRANB1
Embryos co-injected with morpholinos for *zranb1a* and *zranb1b* display mild but significant reductions in head and brain size (**f,g,h** versus **i,j,k,** and **b**), consistent with the microcephaly and CNS defects present in both patients.

EDF9. Distribution of haplinsufficiency scores in selected sets of *de novo* mutations
Violin plot of haploinsufficiency scores in five sets of *de novo* mutations: Silent - all synonymous mutations, Diagnostic - mutations in known DD genes in diagnosed individuals, Undiagnosed_Func - all functional mutations in undiagnosed individuals, Undiagnosed_LoF - All LoF mutations in undiagnosed individuals, Undiagnosed_recur - mutations in genes with recurrent functional mutations in undiagnosed individuals. P values for a Mann-Whitney test comparing each of the latter four distributions to that observed for the silent (synonymous) variants are plotted at the top of each violin.

Supplementary Data Tables (S)

S1. HPO terms
Human Phenotype Ontology terms used within the 1133 probands, where 'shallow_count' is the number of time the specific term is used, and 'deep_count' is the number of times a term underneath that term in the ontology is used.

S2. *De novo* mutations
Validated *de novo* single nucleotide variants and indels within the 1133 probands. Patient IDs have been removed to preserve patient confidentiality, but a file with linking IDs can be provided upon request subject to a data access agreement. Key: Chr – Chromosome; Pos – chromosome coordinate on GRCh37; Gene – HGNC symbol; Transcript – Ensemble transcript ID; Ref/Alt – Reference and Alternate alleles observed; Type – SNV or Indel; Consequence – Most severe consequence predicted by VEP across all transcripts for the gene; AAchange – amino-acid change using single letter codes; Regulatory – presence of the variant within one of the three classes of targeted regulatory sequence described in Supplementary Information; Validation – experimentally validated using an independent technology ('DNM') or not ('Uncertain'); Diagnosis – variant is diagnostically pathogenic in a 'Known' or 'Novel' DD gene.

S3. *De novo* CNVs
De novo copy number variants within the 1133 probands detected using arrayCGH, exome sequencing or both. Patient IDs have been removed to preserve patient confidentiality, but a file with linking IDs can be provided upon request subject to a data access agreement. Key: Chr – Chromosome; Pos – approximate start chromosome coordinate on GRCh37;End – approximate end chromosome coordinate on GRCh37; Ref – reference base at start position; Alt – Deletion '<DEL>' or Duplication '<DUP>'; Call_source – called from exome data, aCGH data, or both; copy_number – estimated copy number: '1' – heterozygous deletion, '3' – heterozygous duplication, '.' uncertain copy number; w_score - confidence metric from aCGH analysis, larger is more confident; convex_score – confidence metric from exome analysis, larger is more confident; consequence - Most severe consequence predicted by VEP across all transcripts; Transcript – Ensembl transcript in which most severe consequence was observed; Gene – HGNC symbol in which most severe consequence was observed; Genes – all genes encompassed by the CNV.

S4. Diagnoses
All diagnostic variants, in both known and new genes, with links to the patient ID in DECIPHER. Key: DECIPHER_ID – ID in DECIPHER; Sex – Male or Female; Fam_Hist - Relatives with similar phenotypes; Chr - chromosome; Start - start chromosome coordinate on GRCh37; Stop - end chromosome coordinate on GRCh37; Gene – HGNC symbol; Gene_type – known or novel DD gene; Variant_type – class of variant; Genotype – heterozygous, homozygous, compound heterozygous, hemizygous, or type of UPD, or clonality of mosaicism; Ref/Alt – reference and alternate alleles observed; Consequence – predicted consequence on affected gene(s); Inheritance – de novo or inherited from mother, father, or both; Phenotypes – HP terms observed in proband.

S5. Biallelic loss-of-function variant counts
Biallelic, rare (MAF < 5%), loss-of-function variant counts for probands and uninherited diplotypes. See *supplemental methods* for details of variant processing. Patient IDs have been removed to preserve patient confidentiality,

but a file with linking IDs can be provided upon request subject to a data access agreement.

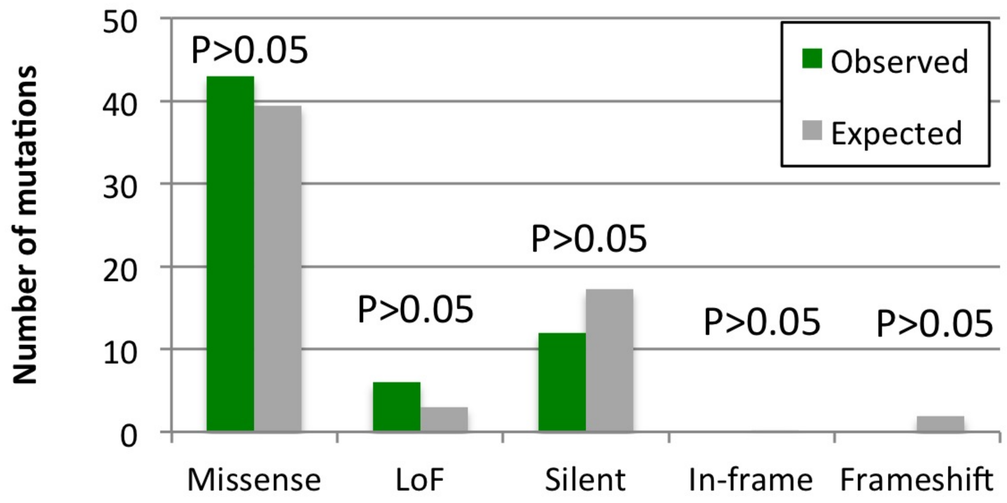
S6. Morpholinos and Primers

Both human and zebrafish gene names are listed, as well as Ensembl gene reference ids. For each gene, Decipher ids are listed for all patients carrying variants. The Variant column indicates type of variant, and number of probands with each type. Zebrafish phenotype indicates whether a developmental phenotype was detected upon gene knockdown. Double KD indicates that phenotypes were only seen with co-knockdown of both zebrafish orthologues. Morpholino sequence is listed as synthesised (orientation is antisense to RNA transcript). Where feasible, primers were selected to detect aberrant splicing in cDNA from injected embryos (Primer Sequences). Translation initiation/ATG blocking morpholinos do not affect mRNA splicing, therefore their activity cannot be detected by RT-PCRs (primers labeled N/A). ND indicates primers were not selected for these genes. For a subset of genes, a second replication morpholino was designed (Morpholino 2 Sequence). See Supplemental methods for additional details.

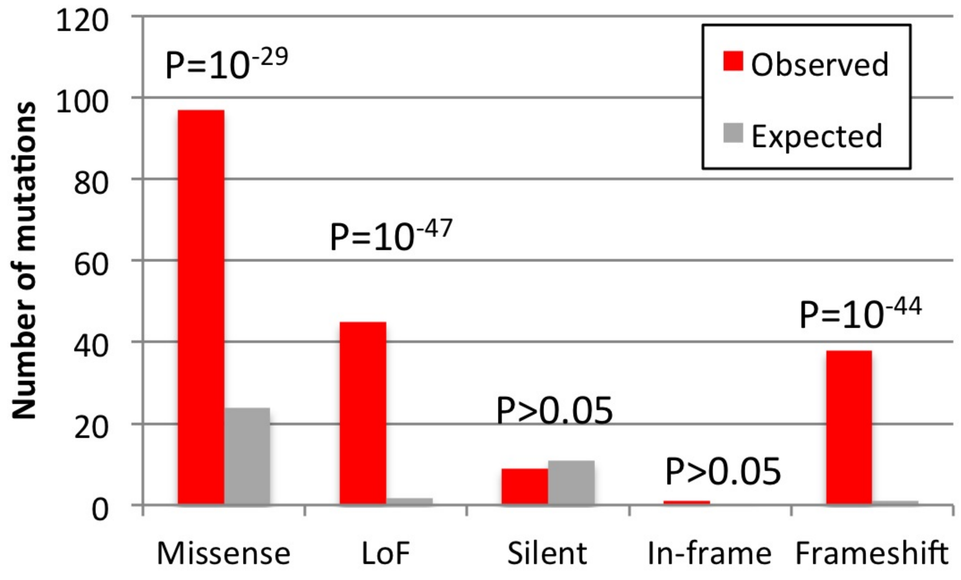
S7. Zebrafish genes and phenotypes table

Standardised phenotype ontological description for all genes showing developmental roles in zebrafish knockdown experiments. All knockdown animals were observed in comparison to control injected animals, and all morphological and locomotor defects were recorded in entity/quality format[6]. *Dose* indicates morpholino level relative to range injected in each experiment, *Amount* gives specific dose injected in nanograms per embryo. See *Supplemental methods* for additional details.

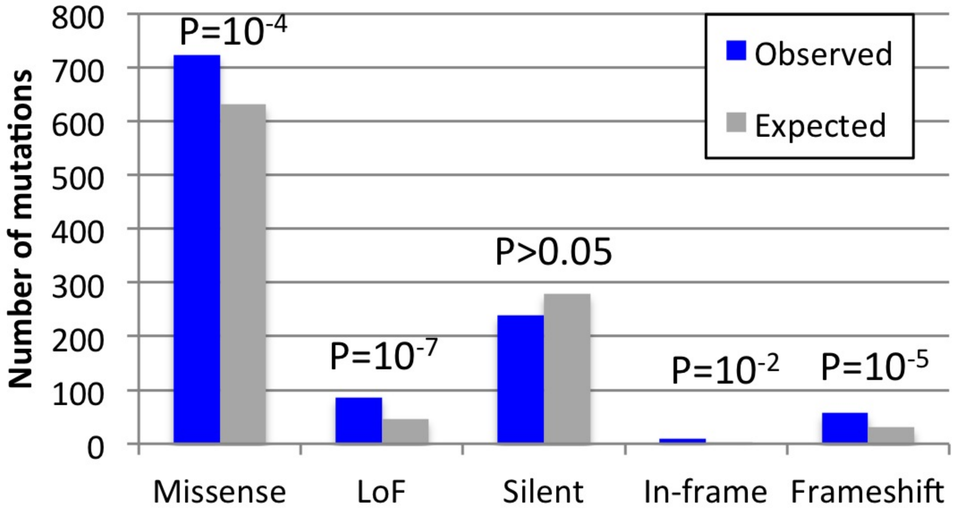
Recessive DD genes

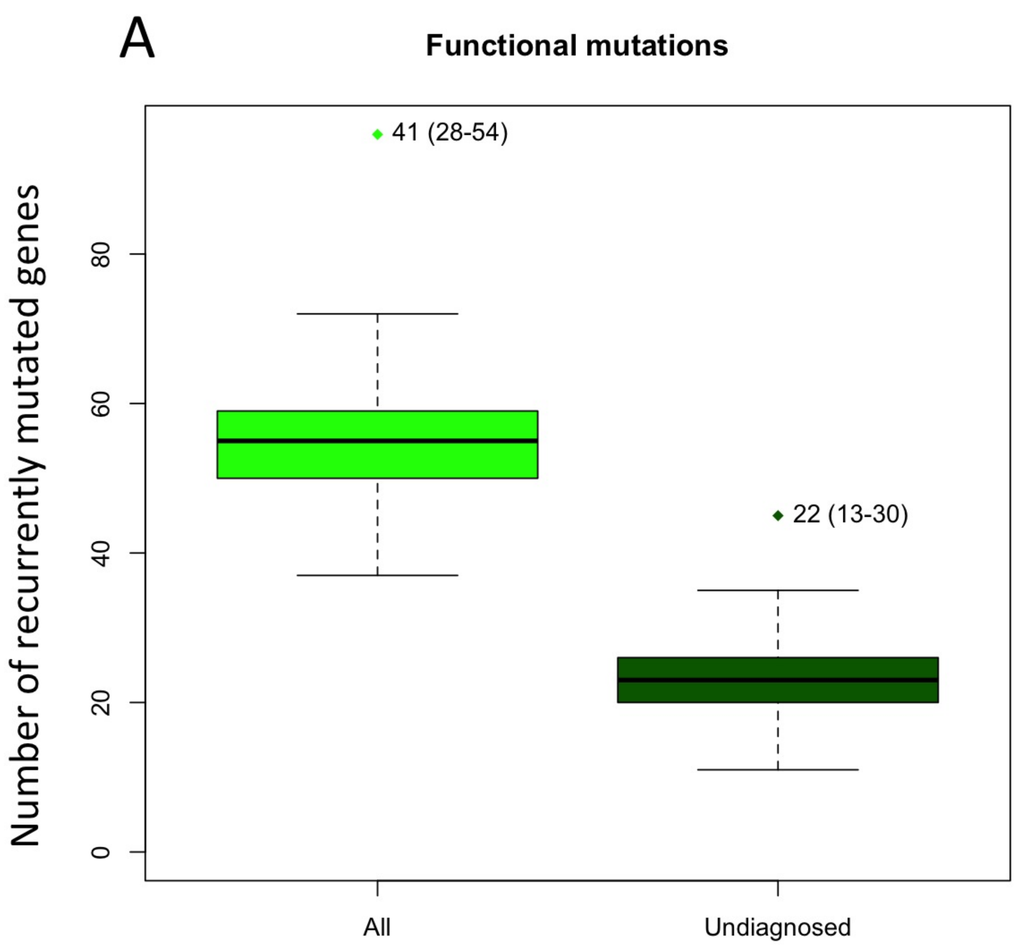


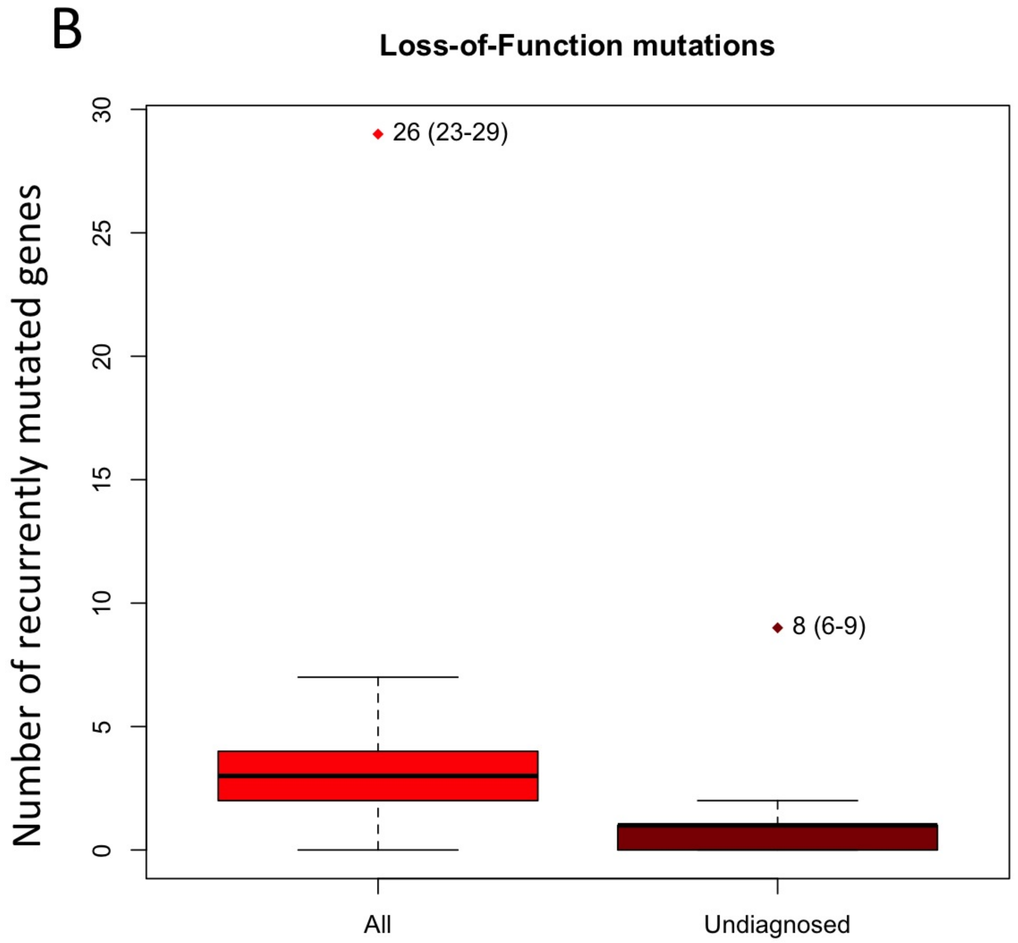
Dominant/XL DD genes

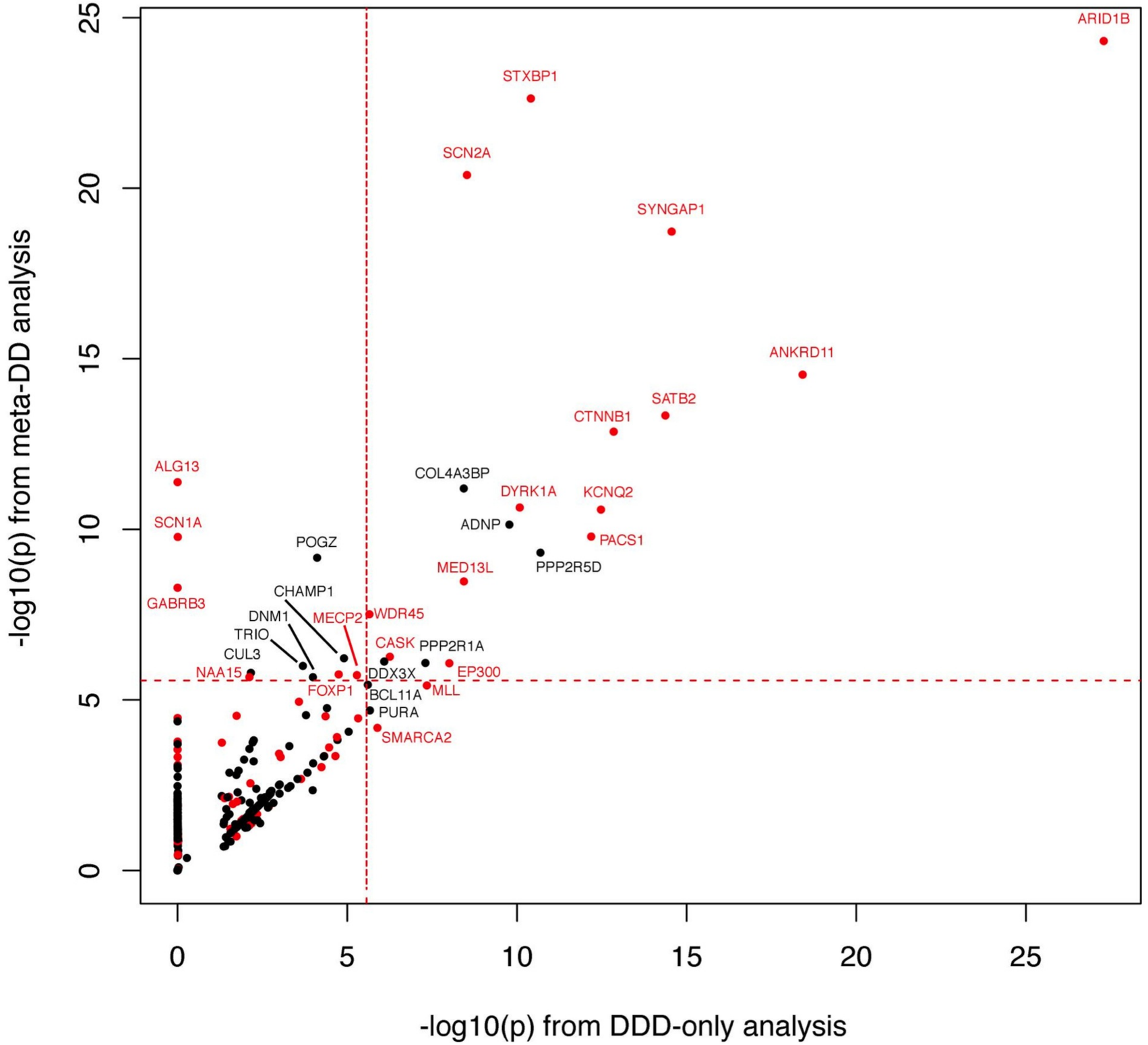


Non-DD genes




A
Functional mutations
Number of recurrently mutated genes
0
20
40
60
80
41 (28-54)
22 (13-30)
All
Undiagnosed


B
Loss-of-Function mutations
Number of recurrently mutated genes
0
5
10
15
20
25
30
26 (23-29)
8 (6-9)
All
Undiagnosed


-log10(p) from meta-DD analysis
-log10(p) from DDD-only analysis
0
5
10
15
20
25
ARID1B
STXBP1
SCN2A
SYNGAP1
ANKRD11
SATB2
CTNNB1
ALG13
COL4A3BP
DYRK1A
KCNQ2
SCN1A
ADNP
PACS1
POGZ
MED13L
PPP2R5D
CHAMP1
GABRB3
DNM1
MECP2
WDR45
TRIO
CASK
PPP2R1A
CUL3
NAA15
DDX3X
EP300
FOXP1
BCL11A
MLL
PURA
SMARCA2

COL4A3BP
(624aa)

Mutations: Ser132Leu, Ser132Leu, Ser132Leu

Post-translational modifications

Domains: PH, Coiled coil, FFAT, START

PCGF2
(344aa)

Mutations: Pro65Leu, Pro65Leu

Post-translational modifications

Domains: Zinc Finger, NLS, Pro/Ser rich

- Phosphoserine
- Phosphothreonine
- N-acetylalanine
- N6-acetyllysine

PPP2R5D
(602aa)

Mutations: Glu198Lys, Glu198Lys, Glu198Lys, Pro201Arg

Post-translational modifications

Domains: QP repeats, B56, NLS, SH3 binding

PPP2R1A
(589aa)

Mutations: Arg182Trp, Arg182Trp, Pro179Leu

Post-translational modifications

Domains: HEAT repeats PP2A subunit B binding, HEAT repeats PP2A subunit C binding

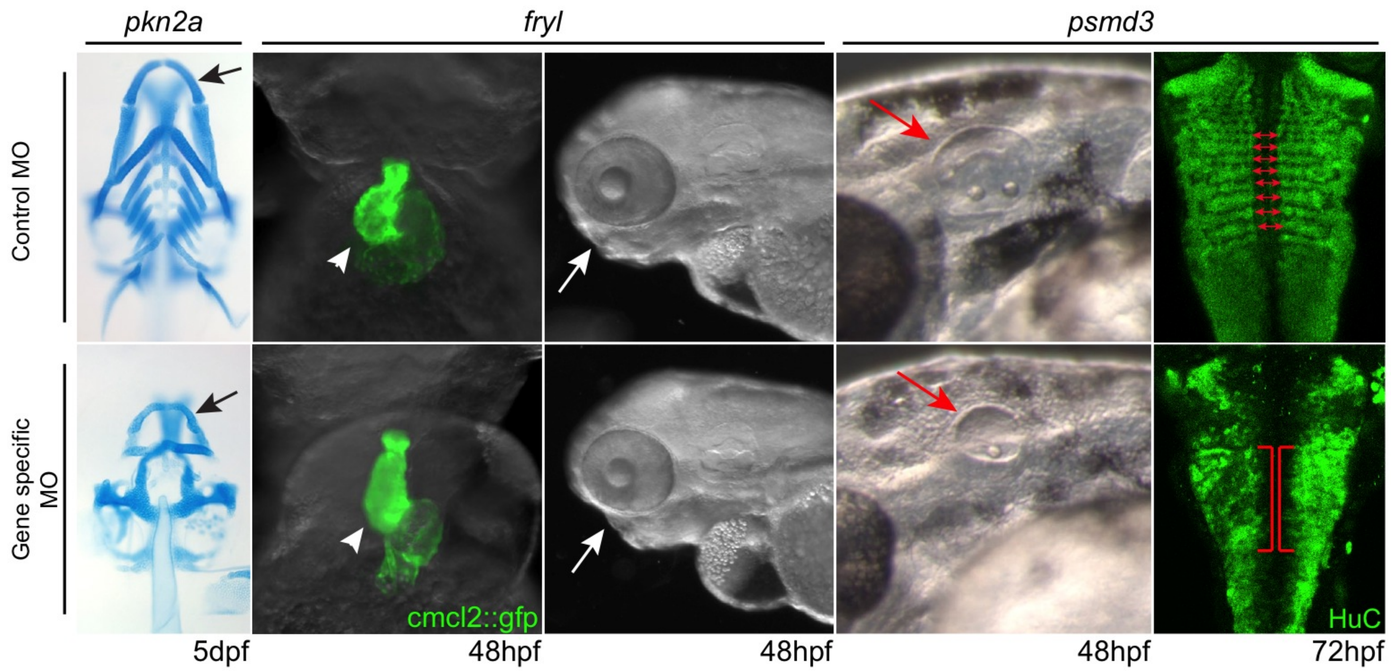
BCL11A
(835aa)

Mutations: Thr47Pro, Cys48Phe, His66Gln

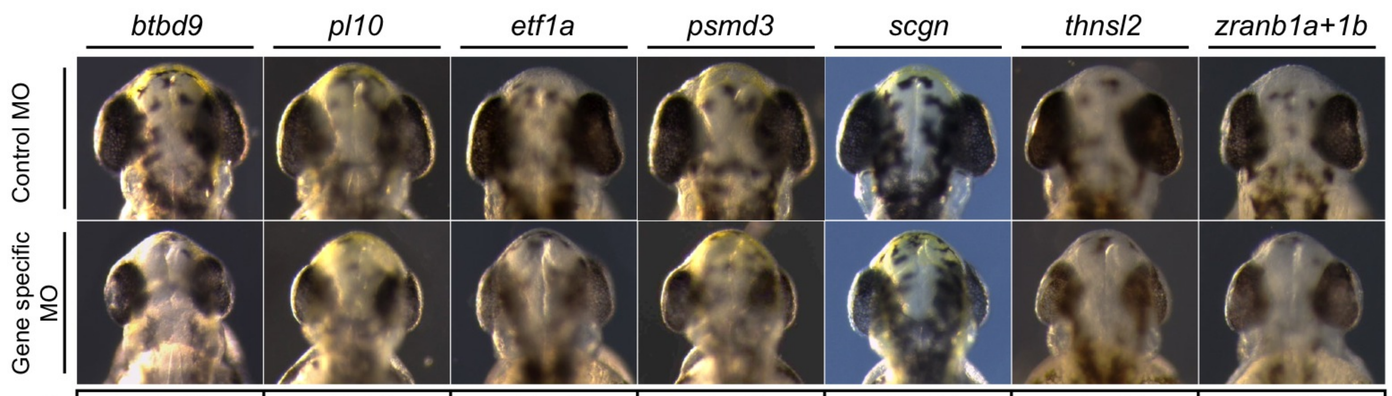
Post-translational modifications

Domains: Zinc Finger, Pro-rich, Zinc Fingers, Glu-rich, Zinc Fingers

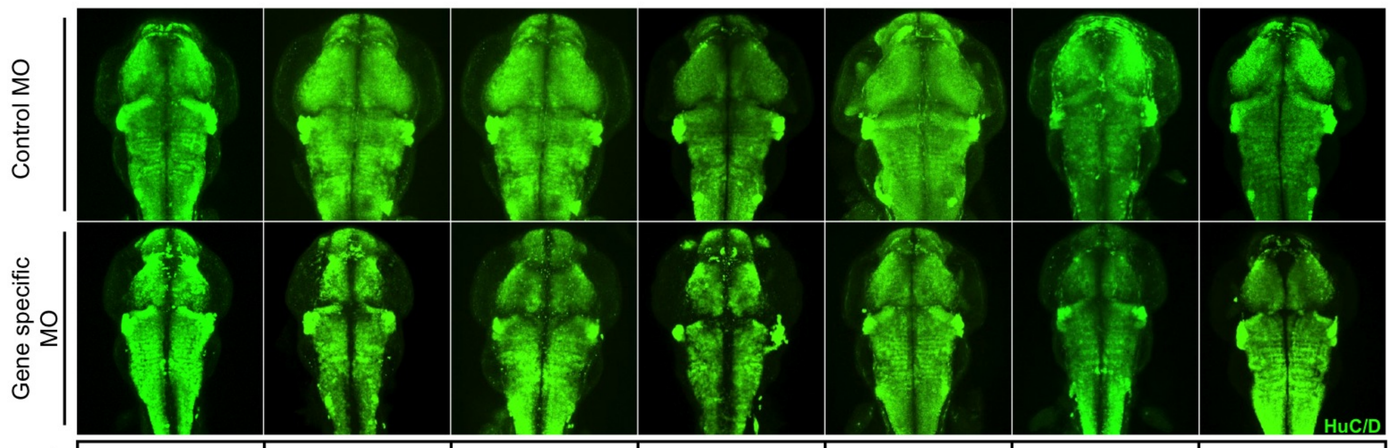
a



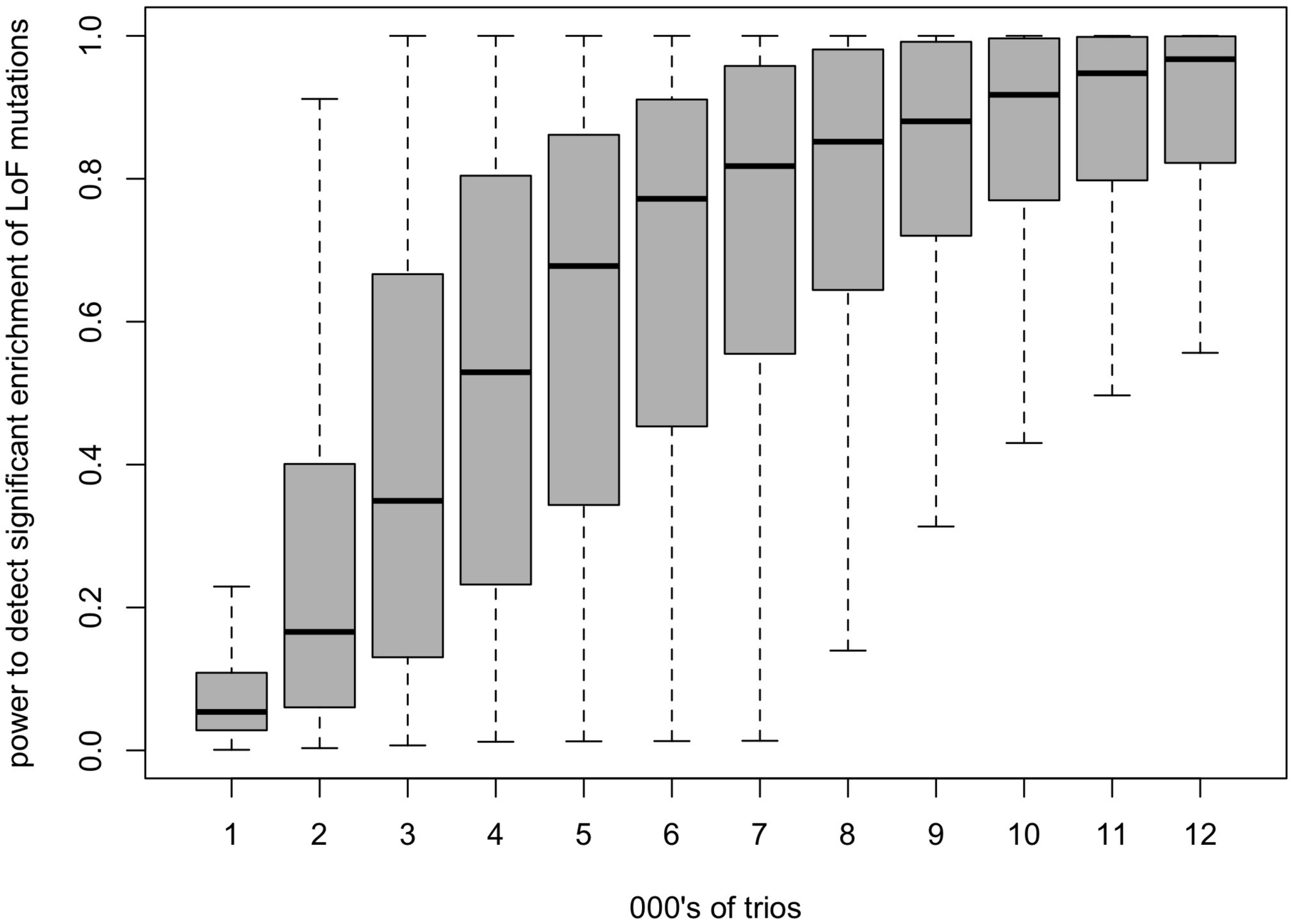
b



change in head size	-16.6% $p \le 4 \times 10^{-6}$	-23.1% $p \le 1.8 \times 10^{-5}$	-11.5% $p \le 2.3 \times 10^{-5}$	-8.9% $p \le 0.0005$	-12.3% $p \le 0.0003$	-15.1% $p \le 1.3 \times 10^{-12}$	-8.2% $p \le 5.4 \times 10^{-6}$



change in brain size	-15.3% $p \le 0.002$	-27.4% $p \le 7.8 \times 10^{-10}$	-13.4% $p \le 0.000014$	-16% $p \le 0.00027$	-17% $p \le 1.3 \times 10^{-15}$	-16.1% $p \le 1.6 \times 10^{-5}$	-9.2% $p \le 0.0015$


power to detect significant enrichment of LoF mutations
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
6
7
8
9
10
11
12
000's of trios