LETTER

doi:10.1038/nature14497

Identification of *cis*-suppression of human disease mutations by comparative genomics

Daniel M. Jordan[1]*, Stephan G. Frangakis[2]*, Christelle Golzio[2], Christopher A. Cassa[1], Joanne Kurtzberg[3], Task Force for Neonatal Genomics†, Erica E. Davis[2], Shamil R. Sunyaev[1]§ & Nicholas Katsanis[2]§

Patterns of amino acid conservation have served as a tool for understanding protein evolution[1]. The same principles have also found broad application in human genomics, driven by the need to interpret the pathogenic potential of variants in patients[2]. Here we performed a systematic comparative genomics analysis of human disease-causing missense variants. We found that an appreciable fraction of disease-causing alleles are fixed in the genomes of other species, suggesting a role for genomic context. We developed a model of genetic interactions that predicts most of these to be simple pairwise compensations. Functional testing of this model on two known human disease genes[3,4] revealed discrete *cis* amino acid residues that, although benign on their own, could rescue the human mutations *in vivo*. This approach was also applied to *ab initio* gene discovery to support the identification of a *de novo* disease driver in *BTG2* that is subject to protective *cis*-modification in more than 50 species. Finally, on the basis of our data and models, we developed a computational tool to predict candidate residues subject to compensation. Taken together, our data highlight the importance of *cis*-genomic context as a contributor to protein evolution; they provide an insight into the complexity of allele effect on phenotype; and they are likely to assist methods for predicting allele pathogenicity[5,6].

Understanding the nature and prevalence of genetic interactions has the potential to elucidate the evolutionary forces that act on protein residues, protein complexes and, more broadly, genomes. Some studies have reported that interactions are ubiquitous and contribute considerably to the evolutionary landscape[7,8], while others found that interactions are rare[9]. Even among those who agree that genetic interactions are important, the architecture of these interactions remains unclear: some studies find distinct interactions between two or three sites[10,11], while others propose a complex interaction network, effectively responding to aggregate properties of the entire protein or the entire genome[12,13].

One practical utility of comparative genomics has been highlighted by our appreciation of the large number of rare variants in humans and the difficulty in inferring their contribution to disease[2]. To prioritize variants of interest, frequency in control populations and evolutionary conservation have become two prominent filters. Conserved regions are considered more likely to be intolerant of variation[1]; programs such as PolyPhen[5] and SIFT[6] have employed this principle to predict functional effects of variants[2]. Although useful, these strategies are constrained, in part because they do not take into consideration the genomic context of the mutated allele. An allele can appear damaging in one sequence yet be neutral in an orthologous sequence of another species. This phenomenon, referred to as compensated pathogenic deviation (CPD), contributes an unknown, but potentially large, number of false negatives to the evaluation of functional sites[14,15].

To examine the prevalence of CPDs, and to identify such sites, we used comparative genomics. A typical, non-CPD allele should cause the same phenotype in any orthologous sequence, regardless of genetic background. By contrast, when a variant that causes human genetic disease is found in a wild-type orthologous sequence, it is likely that the genetic background of that species exerts a compensatory effect on such a variant: it suppresses the phenotype, and thus protects the variant from negative selection[11,14–16]. Previous studies have used this insight to quantify the fraction of CPD interspecies substitutions at ~10% (refs 14–16). Other studies have reported estimates of the inverse value, namely the fraction of pathogenic variants that are present as CPDs in other species, ranging from 2–18% (refs 17, 18). We set out to produce a new estimate of this value. We collected two data sets

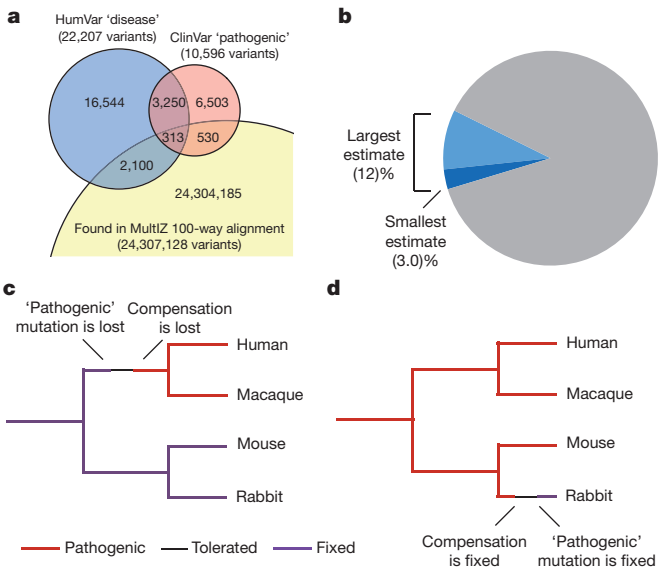


Figure 1 | Distribution of variants found in sequence alignments. **a**, Venn diagram showing sizes and overlap of the ClinVar and HumVar data sets, and how many are found in the multiple sequence alignment. **b**, Estimated number of human disease variants found in the alignment. The smallest estimate (3.0%, dark blue) comes from using the intersection of both variant data sets, requiring the variant to be absent from 6,503 human exomes, and filtering out alignments with low-quality scores. With any methodology, at least 88% of human disease variants (grey) are not found in the alignment. **c**, **d**, Potential mechanisms for the occurrence of CPDs in evolution. Branches where the variant is fixed are purple; branches where the variant is pathogenic are red. In **c**, the variant is present neutrally in an ancestor, but is lost in primates. Subsequent substitutions cause the ancestral allele to become pathogenic. In **d**, the variant is pathogenic in the ancestor, but mutations in a non-human branch cause it to become tolerated, and it arises later by mutation and becomes fixed.

[1]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA. [2]Center for Human Disease Modeling, Duke University, Durham, North Carolina 27701, USA. [3]Department of Pediatrics, Division of Pediatric Blood and Marrow Transplantation, Duke University, Durham, North Carolina 27710, USA.
*These authors contributed equally to this work.
§These authors jointly supervised this work.
†Lists of participants and their affiliations appear in the Supplementary Information.

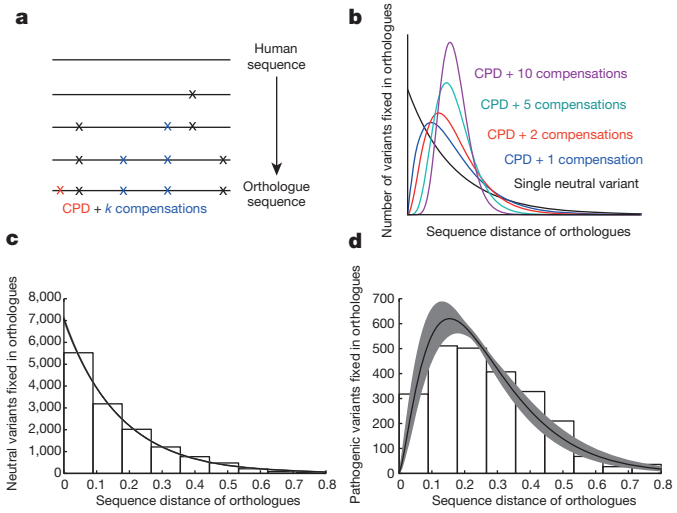


Figure 2 | Relationship between variants and evolutionary distance. **a**, Model for fixation of CPDs. Neutral changes (crosses) arise neutrally. Some of these (blue) compensate for alleles that would otherwise be pathogenic. The parameter *k* represents the number of compensatory changes required for each pathogenic allele. Once *k* compensatory changes have fixed, the CPD (red) fixes neutrally. **b**, The relationship between evolutionary distance and the number of variants in the alignment is expected to be different for individual benign variants (black) and pathogenic variants with different numbers of compensations (blue, one; red, two; cyan, five; magenta, ten). **c**, **d**, Observed distribution of missense variants annotated as neutral (**c**) or pathogenic (**d**) in HumVar and present in vertebrate orthologues (bars), with maximum likelihood fits (black lines) and 95% confidence bands (grey shading). Panel **d** corresponds to a fitted value for *k* of 1.44 ± 0.07.

of missense single-nucleotide variants (SNVs), annotated as either benign or pathogenic, derived from two databases, one based on the literature ('HumVar')[5,19,20] and one based on clinical genetic laboratories and investigator submissions ('ClinVar')[20]. Although the two databases are not fully independent, the majority of pathogenic variants were listed in one or the other (Fig. 1a). Overall, these data sets comprised 69,905 human missense mutations across 13,040 genes. We compared this data set to orthologous proteins from 100 vertebrates. As expected, we found the mutant residue for a large number of likely neutral human variants to be fixed in orthologues. However, the number of pathogenic missense variants found in orthologues (CPDs) was surprisingly high: 5.6% ± 0.5% of ClinVar variants and 6.7% ± 0.4% of HumVar variants were found in the alignment of mammals. For all vertebrates, these numbers increase to 10.2% ± 0.7% and 12.0% ± 0.5%, respectively (Table 1).

Mindful of the possibility of false pathogenic annotations, we applied several filtering steps, including cross-referencing HumVar and ClinVar variant annotations with population frequency data, filtering on the basis of alignment quality[21], using alternative alignment methodologies, and requiring variants to be present in multiple species (Table 1 and Supplementary Note). Some filters did remove bona fide recessive alleles (since we did not allow carriers), as well as disease variants with incomplete penetrance, even though this class of alleles is, by definition, sensitive to genomic context and thus likely to be affected by compensation[22]. Nevertheless, all filtering steps retained a substantial number of variants (Supplementary Tables 1 and 2). Including only variants that pass all filtering steps and are detected in >1 vertebrate, we predict that the minimum estimate of CPDs in human patients is 3% (Fig. 1b). This is consistent with previous analyses, which have found that stringent filtering does not change the observed properties of CPDs to any notable extent[16,17]. As a final test, we extracted *post hoc* pathogenic alleles from three different sources, each of which used independent means for assessing pathogenicity in acute paediatric disorders; overall, CPD rates ranged once again between 3% and 9%; additional analyses of other possible sources of bias were likewise consistent with our initial observations and previous studies (see Supplementary Note and Supplementary Tables 3, 4, 5).

We next turned to the question of the structure of the genetic interactions underlying such sites. Broadly, there are two possibilities: suppression of the disease phenotype may be the result of a small number of discrete compensatory substitutions; or suppression may be caused by a global shift in the properties of the gene, or the whole genome, caused by numerous substitutions that, individually, have small effects. The difference between these two models should be visible in the distribution of CPDs among orthologous sequences. During evolution, variants arise stochastically through a Poisson process: the expected amount of evolutionary time required to produce a given substitution is distributed exponentially[23]. For a CPD, the distribution should be different; the presence of a CPD mandates the presence of all compensatory substitutions necessary for the CPD to be rendered neutral. As such, the expected evolutionary time required to produce a CPD is the sum of the times required to produce each compensatory substitution, followed by the time required to produce the CPD.

Previous studies have proposed different processes by which CPDs arise. The most plausible option is a neutral mechanism, in which the compensatory substitutions are neutral and arise/fix neutrally before the pathogenic substitution appears (Fig. 1c, d and Fig. 2a). In this case, the time required for each substitution to arise is given by an exponential distribution, and the time for all compensatory sites to arise is approximated by the convolution of multiple exponential distributions (a gamma distribution, in the case where all exponential distributions are identical). The number of exponential distributions included in the convolution corresponds to one plus the number of compensatory substitutions required, and it can be inferred from the shape of the distribution (Fig. 2b).

Although the evolutionary time separating two sequences is not observable directly, we can approximate it using sequence distance (one minus sequence identity)[24]. We plotted the number of missense variants observed as a function of sequence distance for neutral variants and for CPDs. Qualitatively, the shapes of both distributions match theoretical expectations. The two distributions are distinct from each other ($P = 1.6 \times 10^{-68}$, Kolmogorov–Smirnov two-sample test; Supplementary Tables 6, 7). Additionally, the observed distribution of CPDs is weighted towards shorter evolutionary distances, as expected if most CPDs require a small number of individual compensatory substitutions, as opposed to the normal distribution expected if CPDs require many individual compensatory substitutions (Fig. 2b, d). To obtain a more precise estimate of the number of compensatory substitutions, we used maximum likelihood to fit several versions of the convolution-of-exponentials model with different combinations of variant data sets and alignment strategies (Fig. 2c, d; see Methods

Table 1 | Range of estimates for prevalence of CPDs in human disease

	Unfiltered MultiZ alignment	High-quality MultiZ alignments	Mammalian subset of MultiZ alignment	Present in >1 species in MultiZ alignment	EPO alignment
HumVar	12.0% ± 0.5%	11.5% ± 0.5%	6.7% ± 0.4%	6.1% ± 0.3%	7.5% ± 0.4%
ClinVar	10.2% ± 0.7%	9.9% ± 0.7%	5.6% ± 0.5%	4.7% ± 0.5%	6.5% ± 0.6%
HumVar+ClinVar	9.3% ± 1.0%	8.5% ± 1.0%	5.3% ± 0.8%	3.9% ± 0.7%	5.5% ± 0.9%
HumVar+ClinVar+ESP	7.5% ± 1.0%	7.0% ± 1.0%	3.8% ± 0.7%	3.0% ± 0.6%	4.0% ± 0.8%

Fraction of likely pathogenic mutations in humans considered to be CPDs according to different filtering paradigms. Values represent the fraction of variants for which an alignment could be retrieved where the variant amino acid is present in an orthologue sequence; error ranges are Jeffreys 95% confidence intervals. ESP, NHLBI Exome Sequencing Project; EPO, Enredo–Pecan–Ortheus pipeline.

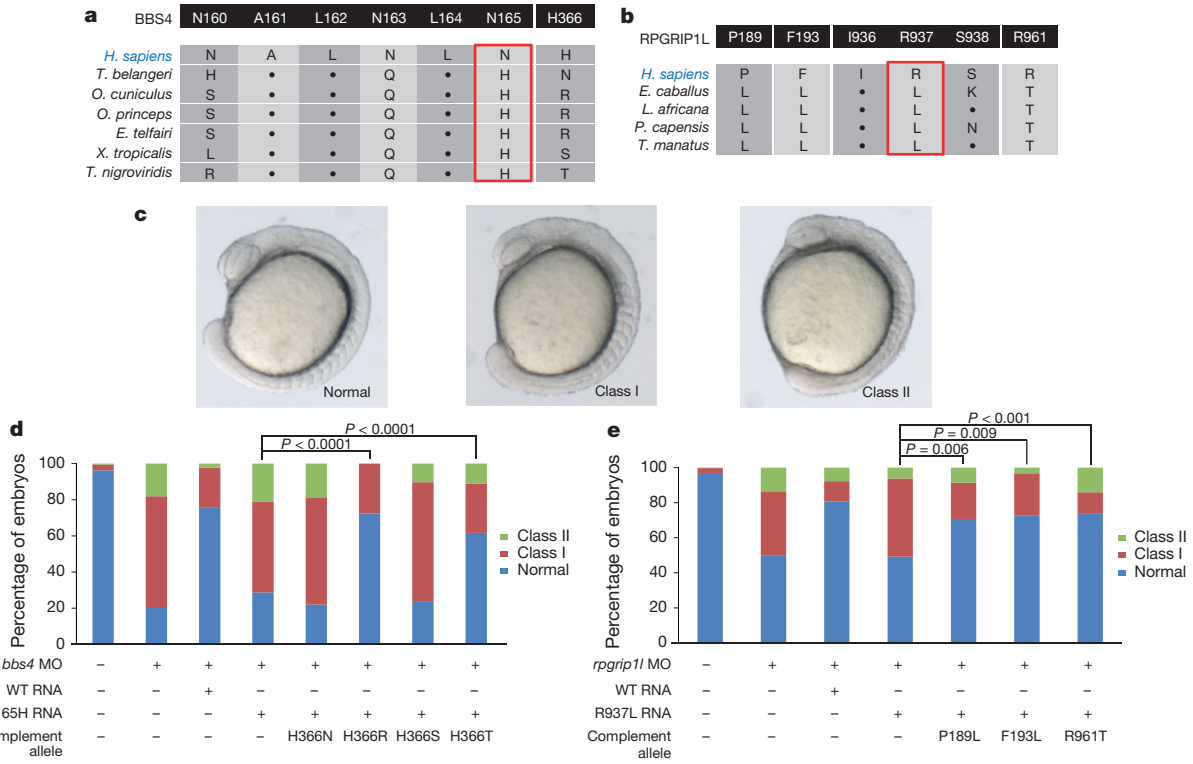


Figure 3 | Compensatory mutations rescue pathogenic alleles in *BBS4* and *RPGRIP1L*. **a**, The pathogenic BBS4 165H allele is fixed in six species. Secondary sites 160, 163 and 366 are possible CPDs. **b**, The pathogenic RPGRIP1L 937L allele is fixed in four species. The 189L, 193L and 961T alleles are present in all four species. **c**, Examples of zebrafish convergent extension phenotypic groups. **d**, Human RNA encoding the BBS4 165H mutation and either 366R or 366T can rescue the morphant phenotype; RNA encoding 165H mutation alone cannot. WT, wild type. **e**, Mutation of 189L, 193L or 961T, in the background of 937L *RPGRIP1L* mRNA, rescues the loss of function observed in 937L RNA. Significance was determined by χ^2 test. See Supplementary Table 9 for embryo counts.

and Supplementary Tables 6, 7). Most versions of the model fit best as the convolution of approximately two exponential distributions, supporting a mechanism in which most CPDs are compensated by simple pairwise interactions. Additionally, most models reported similar rates of evolution for neutral variants, CPDs and compensatory variants, suggesting that the target size for compensatory changes is small. We repeated these analyses with multiple different variant data sets and alignment strategies, finding similar results each time (Extended Data Fig. 1 and Supplementary Table 8).

These analyses predict that most CPDs could be rescued by one large-effect compensatory substitution. We tested this prediction experimentally. We posited that each vertebrate sequence that includes a CPD should also include its *cis*-compensatory allele. Therefore, every amino acid difference between the human sequence and the sequence of the orthologue(s) containing a CPD is a candidate compensatory substitution. Given the practical constraints of examining all possible compensatory substitutions in macromolecular complexes, we focused on substitutions within the same gene as the CPD.

Scanning our list of candidate CPDs, we noted two alleles in genes involved in ciliopathies: a protein-encoding p.N165H change in *BBS4* and a p.R937L variant in *RPGRIP1L*, which contribute pathogenic alleles to Bardet–Biedl syndrome and Meckel–Gruber syndrome, respectively[3,4]. These alleles were prioritized because: (1) Bardet–Biedl and Meckel–Gruber syndromes have a severe effect on reproductive fitness; (2) previous studies have established loss-of-function zebrafish phenotypes rescuable by human messenger RNA for both genes[4,25]; (3) *in vivo* complementation has indicated both alleles to be deleterious to human protein function[4,25]; and (4) we observed multiple species with the human mutant allele fixed: six species for BBS4 165H and four for RPGRIP1L 937L (Fig. 3a, b)—for this reason, both alleles were predicted to be benign (PolyPhen-2, SIFT, MutationAssessor).

Comparative genomic analysis identified 9 candidate sites in BBS4 and 32 candidate sites in RPGRIP1L (Supplementary Table 9). To test each site, we took advantage of the established convergent extension defects induced by morpholino (MO)-mediated suppression of *bbs4* or *rpgrip1l* in zebrafish[4,25]. Consistent with previous observations, suppression of *bbs4* or *rpgrip1l* induced convergent extension defects in 80% and 50% of embryos respectively ($n = 50$–100 embryos; Fig. 3c–e). Co-injection of MO with human wild-type mRNA rescued this phenotype, whereas injection with human mutant mRNA showed no improvement (Fig. 3d, e). We next tested the entire candidate complementing allelic series for each gene. For *BBS4*, the introduction of 2/9 candidate residues in *cis* with the 165H-encoding mRNA ameliorated the phenotype in a manner indistinguishable from wild-type mRNA. Strikingly, both complementing alleles affected the same amino acid and were specific to the compensatory changes: the 165H/366N and the 165H/366S behaved as null, whereas 165H/366R was indistinguishable from wild type; 165H/366T converted the functional null to a hypomorph (Fig. 3d and Extended Data Fig. 2a).

We observed a similar pattern for *RPGRIP1L*. Testing each of the 32 candidate sites identified three complementing events, two of which map to the same region: 937L/189L, 937L/193L and 937L/961T (Fig. 3e and Extended Data Fig. 2b). Testing each complementing allele individually showed them to be either extremely mild or benign (Supplementary Table 9). Finally, comparative genomic analysis showed that these data could explain the tolerance of the RPGRIP1L 937L change in all four species and of the BBS4 165H change in 4/6 species (Fig. 3a, b).

The above analysis is limited by its retrospective nature. We therefore tested the usefulness of our model in *ab initio* gene discovery. We have recently initiated a whole-exome sequencing (WES) and func-

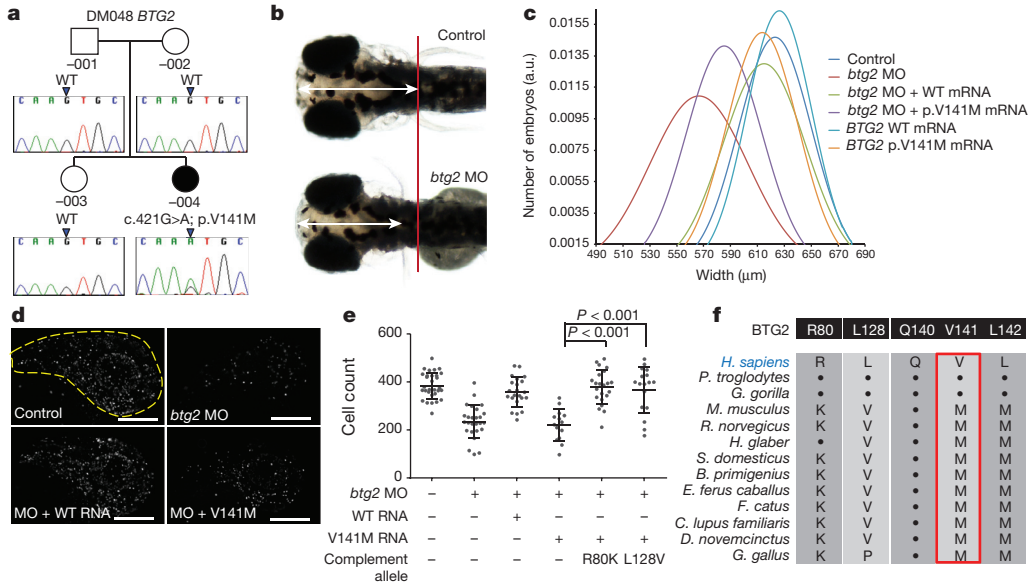


Figure 4 | A *de novo BTG2* p.V141M-encoding allele causes microcephaly. **a**, Pedigree DM048. Chromatograms show a *de novo* c.421G>A nucleotide change. WT, wild type. **b**, Suppression of *btg2* leads to head size defects. Dorsal view of uninjected control and *btg2* MO-injected zebrafish embryos at 4 dpf. White arrows show the distance measured from forebrain to hindbrain. Red line shows the protrusion of the pectoral fins in uninjected controls. **c**, Distribution of head size measurements at 4 dpf (Supplementary Table 10; white arrows in **b**), a.u., arbitrary units. **d**, 2 dpf zebrafish embryos stained for PH3. Human RNA containing the V141M mutation is unable to rescue the reduced proliferation of *btg2* morphants. Scale bars, 250 μm. **e**, Quantification of PH3-positive cells: human RNA with mutations V141M and either R80K or L128V can rescue knockdown of *btg2*. Error bars represent standard deviation. **f**, The 141M allele is fixed in 59/87 species besides primates, examples displayed here. See Supplementary Table 11 for PH3 quantification.

tional testing paradigm to accelerate gene discovery in young children called Task Force for Neonatal Genomics (TFNG). Patients who display anatomical phenotypes amenable to functional modelling in zebrafish are evaluated by trio-based WES and have candidate alleles tested systematically *in vivo*[26].

We enrolled a 17-month-old female with an undiagnosed neuroanatomical condition hallmarked by microcephaly (Fig. 4a). We filtered WES data for non-synonymous and splice variants with a minor allele frequency of <1%, and we conducted a proband-centric trio analysis that yielded four candidates: *de novo* missense changes in *BTG2* and *NOS2*; and recessive missense variants in *TTN* and *LAMA1*. Testing of an unaffected sibling excluded *LAMA1*; *TTN*, a known dominant cardiomyopathy locus[27], is an unlikely driver.

To investigate the pathogenicity of the *BTG2* (p.V141M) and *NOS2* (p.P795A) protein-encoding changes, we studied *btg2* and *nos2* in zebrafish. Reciprocal use of Basic Local Alignment Search Tool (BLAST) between *Homo sapiens* and *Danio rerio* identified a single zebrafish *btg2* orthologue and two zebrafish *nos2* orthologues. We injected splice-blocking MO (sb-MO) or translational-blocking MO (tb-MO) (Extended Data Fig. 3) into zebrafish embryos (3 ng; $n = 80$ embryos per injection) and scored for head size defects at 4 days post-fertilization (dpf) by measuring the anterior–posterior distance between the forebrain and the hindbrain (Fig. 4b). For *nos2a/b* MO-injected embryos, we saw no differences at the highest dose injected (8 ng for *nos2a/b* sb-MOs; Supplementary Table 10). By contrast, we found a significant reduction of anterior structures in *btg2* morphants ($P < 0.0001$; Fig. 4b, c). Co-injection of wild-type human *BTG2* mRNA with tb-MO resulted in significant rescue ($P < 0.0001$; Fig. 4c). In contrast, injection of mRNA harbouring 141M was significantly worse at rescue than wild type ($P < 0.0001$; Fig. 4c).

BTG2 is a regulator of cell cycle checkpoint in neuronal cells[28] and is strikingly intolerant to variation in humans (Exome Variant Server (EVS)). To test the pathogenicity of 141M by a different assay, we performed antibody staining at 2 dpf (a time before the manifestation of microcephaly). We marked post-mitotic neurons in the forebrain with antibodies against neuronal HuC/HuD antigens, and we scored (blind, triplicate) on the basis of an established paradigm[29]. *btg2* morphants displayed a significant decrease in HuC/HuD staining ($P < 0.0001$; Extended Data Fig. 4). This defect was rescued with wild-type *BTG2* mRNA ($P < 0.05$), but could not be ameliorated by 141M-encoded mRNA co-injection (Extended Data Fig. 4). Importantly, co-injection of *btg2* tb-MO with two rare control EVS alleles (p.A126S and p.R145Q) resulted in rescue, providing evidence for assay specificity (Extended Data Fig. 4b). As a third test, we stained whole embryos with a phospho-histone H3 (PH3) antibody that marks proliferating cells. We counted the number of positive cells in a defined anterior region of embryos. We saw a significant reduction in cell proliferation in the heads of 2 dpf *btg2* morphants ($P < 0.0001$); this defect was likewise rescued by co-injection of wild-type mRNA, while 141M mutant rescue was indistinguishable from *btg2* tb-MO alone ($P = 0.38$; Fig. 4b, d). Combined, all three assays indicated that *BTG2* p.V141M is pathogenic and that haploinsufficiency of this gene probably contributes to the microcephaly of the proband.

Despite our functional and genetic data for p.V141M, this allele was predicted computationally to be benign. A likely reason is that, with the exception of primates, most BTG2 orthologues encode Met at the orthologous position (Fig. 4f). These data suggested that V141 might represent a CPD site in primates that branched from the ancestral methionine. To test this possibility, we identified nine BTG2 sites that co-evolved with 141M (Supplementary Table 11), which we mutagenized into the human construct encoding 141M. We then injected embryos with *btg2* MO; MO plus wild-type human *BTG2* mRNA; MO plus 141M-encoding mRNA; or MO plus 141M in *cis* with one of the nine candidate complementing alleles. Seven of the alleles had no effect (Supplementary Table 11). However, R80K- or L128V-encoded mRNA on the 141M backbone rescued the number of PH3-positive cells to wild-type levels (Fig. 4e and Extended Data Fig. 2c); both alleles were benign on their own (Supplementary Table 11). Taken together, our data indicated that 141M is deleterious in the human background, but the protection of this residue conferred by either Lys 80 or by Val 128 can explain >90% (54/59) of species encoding 141M (Fig. 4f).

To improve the scalability of detecting CPDs, we used our model of CPD evolution to develop a computational predictor for distinguish-

ing variants that are unlikely to be CPDs from those that might be CPDs, and to identify candidate compensations to aid experimental design (http://genetics.bwh.harvard.edu/cpd/). Initial testing of this tool intimated high negative predictive values but modest positive predictive values, probably due to the dearth of known CPDs (Supplementary Note).

Our results contrast with some previous studies that claim that epistasis is ubiquitous[7,10]; or that it is practically nonexistent[9]; or that it is commonly of higher order[12,13]. The most likely explanation for this discrepancy is that such studies have examined different kinds of variation and traits. For example, studies on the evolution of genetic incompatibilities rely on assumptions of high mutation rate and weak negative selection, assumptions that generally do not hold for the case of pathogenic missense variation[10,30]. The difference with the studies suggesting higher-order *cis*-interactions may be to do with the scale of evolutionary time our analyses probe: the span of hundreds of millions of years of evolution represented by the vertebrate alignment may not be long enough to reveal higher-order combinations of non-synonymous SNVs. Indeed, using neutral SNVs from the HumVar data set as a control, we estimate the vertebrate alignment has explored 12% of pairwise interactions between SNVs, compared to 0.6% of three-way interactions between SNVs. It is possible that higher-order interactions are common, but are not detectable without a deeper alignment.

Finally, considering the accelerated use of genome editing to model human pathogenic mutations in a variety of model organisms, our data highlight the critical need to not only pair computational predictions with functional studies, but also to evaluate the effect of human mutations in the context of the human sequence.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 26 October 2014; accepted 23 April 2015.

Published online 29 June 2015.

1. Alföldi, J. & Lindblad-Toh, K. Comparative genomics as a tool to understand evolution and disease. *Genome Res.* **23,** 1063–1068 (2013).
2. Cooper, G. M. & Shendure, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nature Rev. Genet.* **12,** 628–640 (2011).
3. Katsanis, N. *et al. BBS4* is a minor contributor to Bardet-Biedl syndrome and may also participate in triallelic inheritance. *Am. J. Hum. Genet.* **71,** 22–29 (2002).
4. Khanna, H. *et al.* A common allele in *RPGRIP1L* is a modifier of retinal degeneration in ciliopathies. *Nature Genet.* **41,** 739–745 (2009).
5. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature Methods* **7,** 248–249 (2010).
6. Sim, N. L. *et al.* SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* **40,** W452–W457 (2012).
7. Breen, M. S., Kemena, C., Vlasov, P. K., Notredame, C. & Kondrashov, F. A. Epistasis as the primary factor in molecular evolution. *Nature* **490,** 535–538 (2012).
8. Weinreich, D. M., Delaney, N. F., DePristo, M. A. & Hartl, D. L. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312,** 111–114 (2006).
9. McCandlish, D. M., Rajon, E., Shah, P., Ding, Y. & Plotkin, J. B. The role of epistasis in protein evolution. *Nature* **497,** E1–2 (2013).
10. Corbett-Detig, R. B., Zhou, J., Clark, A. G., Hartl, D. L. & Ayroles, J. F. Genetic incompatibilities are widespread within species. *Nature* **504,** 135–137 (2013).
11. Gao, L. & Zhang, J. Why are some human disease-associated mutations fixed in mice? *Trends Genet.* **19,** 678–681 (2003).
12. Weinreich, D. M., Lan, Y., Wylie, C. S. & Heckendorn, R. B. Should evolutionary geneticists worry about higher-order epistasis? *Curr. Opin. Genet. Dev.* **23,** 700–707 (2013).
13. Chou, H. H., Chiu, H. C., Delaney, N. F., Segre, D. & Marx, C. J. Diminishing returns epistasis among beneficial mutations decelerates adaptation. *Science* **332,** 1190–1192 (2011).
14. Kondrashov, A. S., Sunyaev, S. & Kondrashov, F. A. Dobzhansky-Muller incompatibilities in protein evolution. *Proc. Natl Acad. Sci. USA* **99,** 14878–14883 (2002).
15. Kulathinal, R. J., Bettencourt, B. R. & Hartl, D. L. Compensated deleterious mutations in insect genomes. *Science* **306,** 1553–1554 (2004).
16. Soylemez, O. & Kondrashov, F. A. Estimating the rate of irreversibility in protein evolution. *Genome Biol. Evol.* **4,** 1213–1222 (2012).
17. Ferrer-Costa, C., Orozco, M. & de la Cruz, X. Characterization of compensated mutations in terms of structural and physico-chemical properties. *J. Mol. Biol.* **365,** 249–256 (2007).
18. Waterston, R. H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
19. Mottaz, A., David, F. P., Veuthey, A. L. & Yip, Y. L. Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar. *Bioinformatics* **26,** 851–852 (2010).
20. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42,** D980–D985 (2014).
21. Ahola, V., Aittokallio, T., Vihinen, M. & Uusipaikka, E. Model-based prediction of sequence alignment quality. *Bioinformatics* **24,** 2165–2171 (2008).
22. Giudicessi, J. R. & Ackerman, M. J. Determinants of incomplete penetrance and variable expressivity in heritable cardiac arrhythmia syndromes. *Transl. Res.* **161,** 1–14 (2013).
23. Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, 1984).
24. Povolotskaya, I. S. & Kondrashov, F. A. Sequence space and the ongoing expansion of the protein universe. *Nature* **465,** 922–926 (2010).
25. Zaghloul, N. A. *et al.* Functional analyses of variants reveal a significant role for dominant negative and common alleles in oligogenic Bardet-Biedl syndrome. *Proc. Natl Acad. Sci. USA* **107,** 10602–10607 (2010).
26. Katsanis, N., Cotten, M. & Angrist, M. Exome and genome sequencing of neonates with neurodevelopmental disorders. *Future Neurology* **7,** 655–658 (2012).
27. Herman, D. S. *et al.* Truncations of titin causing dilated cardiomyopathy. *N. Engl. J. Med.* **366,** 619–628 (2012).
28. Montagnoli, A., Guardavaccaro, D., Starace, G. & Tirone, F. Overexpression of the nerve growth factor-inducible *PC3* immediate early gene is associated with growth inhibition. *Cell Growth Differ.* **7,** 1327–1336 (1996).
29. Beunders, G. *et al.* Exonic deletions in *AUTS2* cause a syndromic form of intellectual disability and suggest a critical role for the C terminus. *Am. J. Hum. Genet.* **92,** 210–220 (2013).
30. Fraïsse, C., Elderfield, J. A. & Welch, J. J. The genetics of speciation: are complex incompatibilities easier to evolve? *J. Evol. Biol.* **27,** 688–699 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank Y. Liu and D. Balick for helpful discussions, M. Kousi for assistance with the NCL mutational list, and M. Talkowski, A. Kondrashov and G. Lyon for critical review of the manuscript. This work was supported by grants R01HD04260, R01DK072301 and R01DK075972 (N.K.); R01 GM078598, R01 MH101244, R01 DK095721 and U01 HG006500 (S.R.S.); R01EY021872 (E.E.D.); and a NARSAD Young Investigator Award (C.G.). N.K. is a Distinguished Brumley Professor.

Author Contributions D.M.J., S.G.F., S.R.S. and N.K. designed the overall study. D.M.J., C.A.C. and S.R.S. conceptualized the principle of CPDs and performed all computational analyses. S.G.F., E.E.D. and N.K. conceptualized the biological properties of CPDs and implemented *in vivo* testing with the assistance of C.G. J.K. referred the index patient and evaluated clinical data in the context of molecular discoveries. The Task Force for Neonatal Genomics constructed the platforms and methods for recruitment, ascertainment and evaluation of clinical and molecular data and return of results.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.R.S. (ssunyaev@rics.bwh.harvard.edu) or N.K. (nicholas.katsanis@dm.duke.edu).

METHODS

Data sets of known benign and pathogenic variants. Our primary training data set was HumVar, one of the training data sets for PolyPhen-2.2.3. We used the most recent public release at the time of this publication (December 2011), available for download at http://genetics.bwh.harvard.edu/pph2/dokuwiki/downloads. This data set is derived from SwissVar variant annotations[19]. It contains 22,207 variants annotated as pathogenic and 21,433 variants annotated as benign. We also used a data set of pathogenic variants derived from the June 2014 release of the ClinVar database[20]. This data set consists of all missense variants from ClinVar that are unambiguously (that is, classified the same by all submitters) and confidently (that is, not a 'Likely' annotation) annotated as 'Pathogenic'. It contains 10,596 variants annotated as pathogenic and 1,926 variants annotated as benign. The intersection of these two data sets contains 3,563 variants annotated as pathogenic and 454 variants annotated as benign. As an additional control, we required that the pathogenic variants be absent from 6,503 human exomes[31] (EVS; http://evs.gs.washington.edu/EVS/). This most stringent data set contains 3,062 variants annotated as pathogenic.

Comparative genomics screen for CPDs. We used the University of California, Santa Cruz (UCSC) MultiZ whole-genome alignments of 100 vertebrate sequences[32], downloaded from UCSC as translated exons. As an alternative alignment strategy, we used the EPO alignment of 37 eutherian mammal species[33], downloaded as nucleotide sequences and translated for all aligned species using the human open reading frame. In cases in which the alignment contained multiple sequences from the same species, only the sequence most similar to the human sequence was retained. Variants were classified as CPDs if the variant amino acid was found in the translated sequence of any vertebrate orthologue other than human or chimpanzee, with chimpanzee being excluded because presence in the chimpanzee sequence may be used as evidence for neutrality in variant annotation databases. The resulting data set of neutral and deleterious variants found in vertebrate orthologues is available (see Source Data for Fig. 1).

Statistical models of variant density. We modelled the density of benign variants with an exponential distribution, with scale parameter β_{neut} representing the mean time to fixation of neutral alleles. We used three different models for the density of CPDs. (1) k compensatory changes fix at rate $1/\theta$, followed by the now-neutral CPD at the same rate. This is represented by a gamma distribution with shape parameter $k+1$ and scale parameter θ. (2) k compensatory changes fix at rate $1/\theta_1$, followed by the now-neutral CPD at an independent rate $1/\theta_2$. This is represented by the convolution of a gamma distribution with shape parameter k and scale parameter θ_1, and an exponential distribution with scale parameter θ_2. (3) k compensatory changes fix at rate $1/\beta_{\text{comp}}$, followed by the now-neutral CPD at the neutral rate $1/\beta_{\text{neut}}$.

We assume a reversible model of evolution, so that the same three models can apply both to the fixation of CPDs not present in an ancestral sequence and to the loss of CPDs that are present in an ancestral sequence. The random variable used in these models is the sequence distance to the closest sequence containing the variant, where sequence distance is defined as the fraction of aligned (that is, non-gapped) positions that are identical.

Fitting observed density to statistical models. We recorded for each variant found in a vertebrate orthologue the minimum number of amino acid differences between that variant and a vertebrate orthologue, not counting gapped sites, normalized by the length of the sequence. We then used maximum likelihood to fit the neutral model and each of the three pathogenic models described earlier, using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) optimization algorithm to maximize the likelihood functions. We repeated the fit using each of our filtered variant data sets and alignment methodologies, as well as discarding all variants that were only found in the alignment in a single sequence. All three models fit reasonably well, and all produced qualitatively similar results on all data sets and alignment methodologies. The exact fitted parameter values are found in Supplementary Table 8.

Prediction method. Our prediction method is implemented in Perl, and the source code is available (Supplementary Data 1-CPD Predictor Code). To calculate the probability that a variant is a CPD, we find the minimum distance to the variant in the multiple sequence alignment and apply Bayes' Law. We use the third likelihood model described earlier, where the CPD fixes at the neutral rate, using the maximum likelihood inferred parameter values. As our prior we use the well-established result that 10% of variants seen in another sequence are CPDs. Candidate compensation sites are identified by collecting all substitutions found in any sequence containing the candidate CPD, prioritizing sites that are substituted in many sequences over sites that are substituted in only a few sequences.

WES. Research study participants were enrolled upon informed consent according to protocols approved by the Duke University Internal Review Board. We conducted paired-end pre-capture library preparation by fragmenting genomic DNA through sonication, ligating it to the Illumina multiplexing PE adapters, and PCR amplification using indexing primers. For target enrichment/exome capture we enriched the pre-capture library by hybridizing to biotin-labelled VCRome 2.1 (ref. 34) in-solution Exome Probes at 47 °C for 64–72 h. For massively parallel sequencing, the post-capture library DNA was subjected to sequence analysis on an Illumina Hiseq platform for 100 bp paired-end reads (130× median coverage, >95% target coverage at 10×). Primary data were interpreted and analysed by Mercury 1.0; the output data from Illumina HiSeq were converted from bcl files to FastQ files by Illumina CASAVA 1.8 software, and mapped by the BWA program. We performed variant calls using Atlas-SNP and Atlas-indel[35].

Morpholino design. MOs targeting zebrafish *bbs4* and *rpgrip1l* were obtained from Gene Tools, and described previously[4,25]. MOs against zebrafish *btg2* and *nos2* were obtained from Gene Tools (Extended Data Fig. 3; sequences available upon request).

Site-directed mutagenesis. Mutant alleles were generated as described[25]. Sequences were validated via Sanger sequencing on Applied Biosystems 3730xl DNA Analyzer.

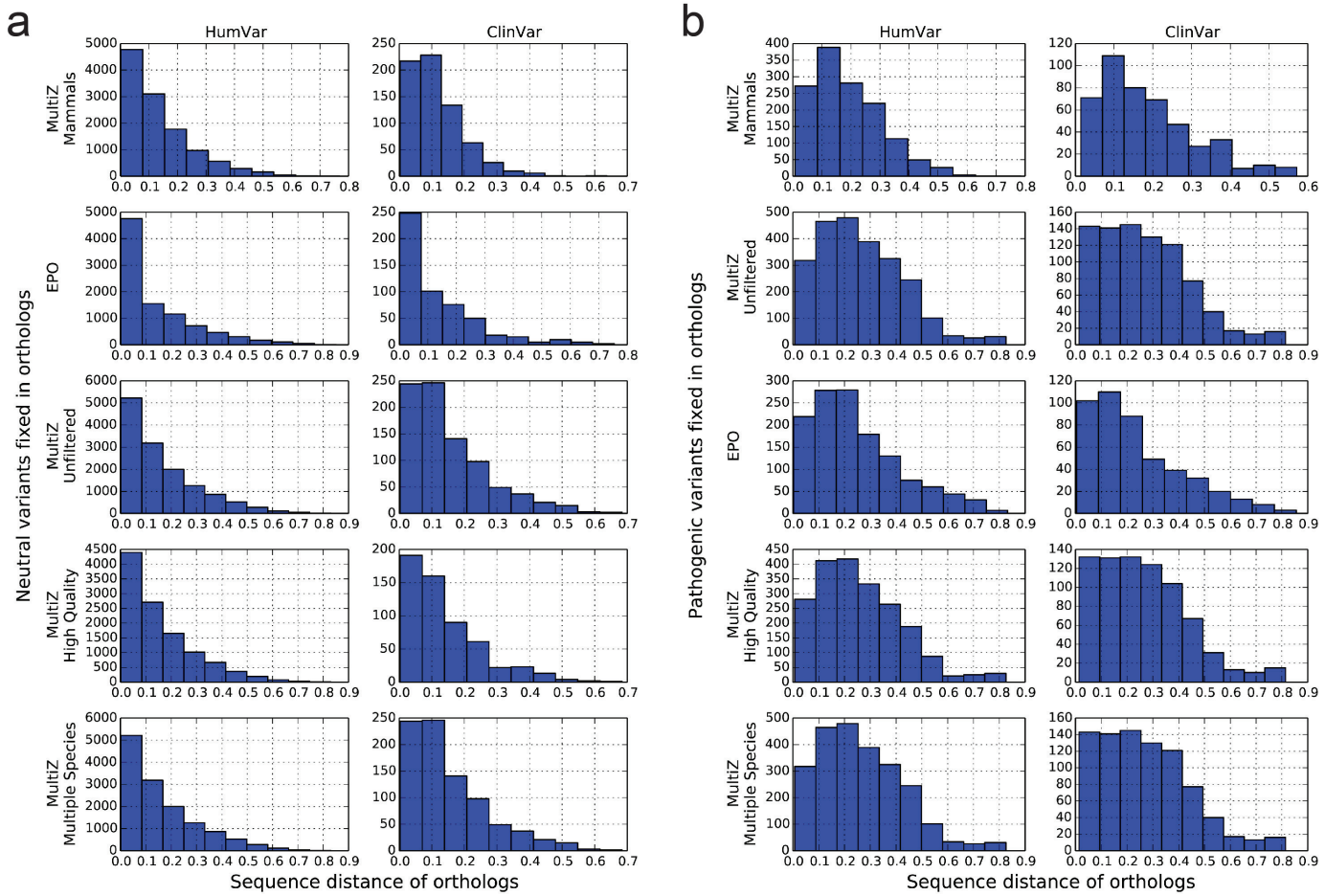
mRNA synthesis and zebrafish embryo injection. mRNA was transcribed *in vitro* as described[36] using SP6 Message Machine kit (Ambion). MO and mRNA concentrations were determined based on the combination by which wild-type mRNA efficiently rescued the morphant phenotype. The same concentrations were used for rescue with mutant mRNA or injection of mRNA alone. The MO and mRNA concentrations injected were as follows: 0.7 ng *bbs4* MO and 100 pg *BBS4* mRNA; 5 ng *rpgrip1l* MO and 100 pg *RPGRIP1L* mRNA; 3 ng *btg2* MO and 150 pg *BTG2* mRNA; 8 ng *nos2a*; 8 ng *nos2b*. All animal work was performed in accordance with the protocols and guidelines of the Duke Institutional Animal Care and Use Committee.

Classification and scoring of embryos. Embryos injected with *bbs4* or *rpgrip1l* MOs were classified into two graded phenotypes on the basis of the relative severity compared with age-matched uninjected controls from the same clutch, as described previously[25]. Comparisons between injections of MO alone, mRNA alone, mutant rescue, and wild-type rescue were performed by χ^2 test.

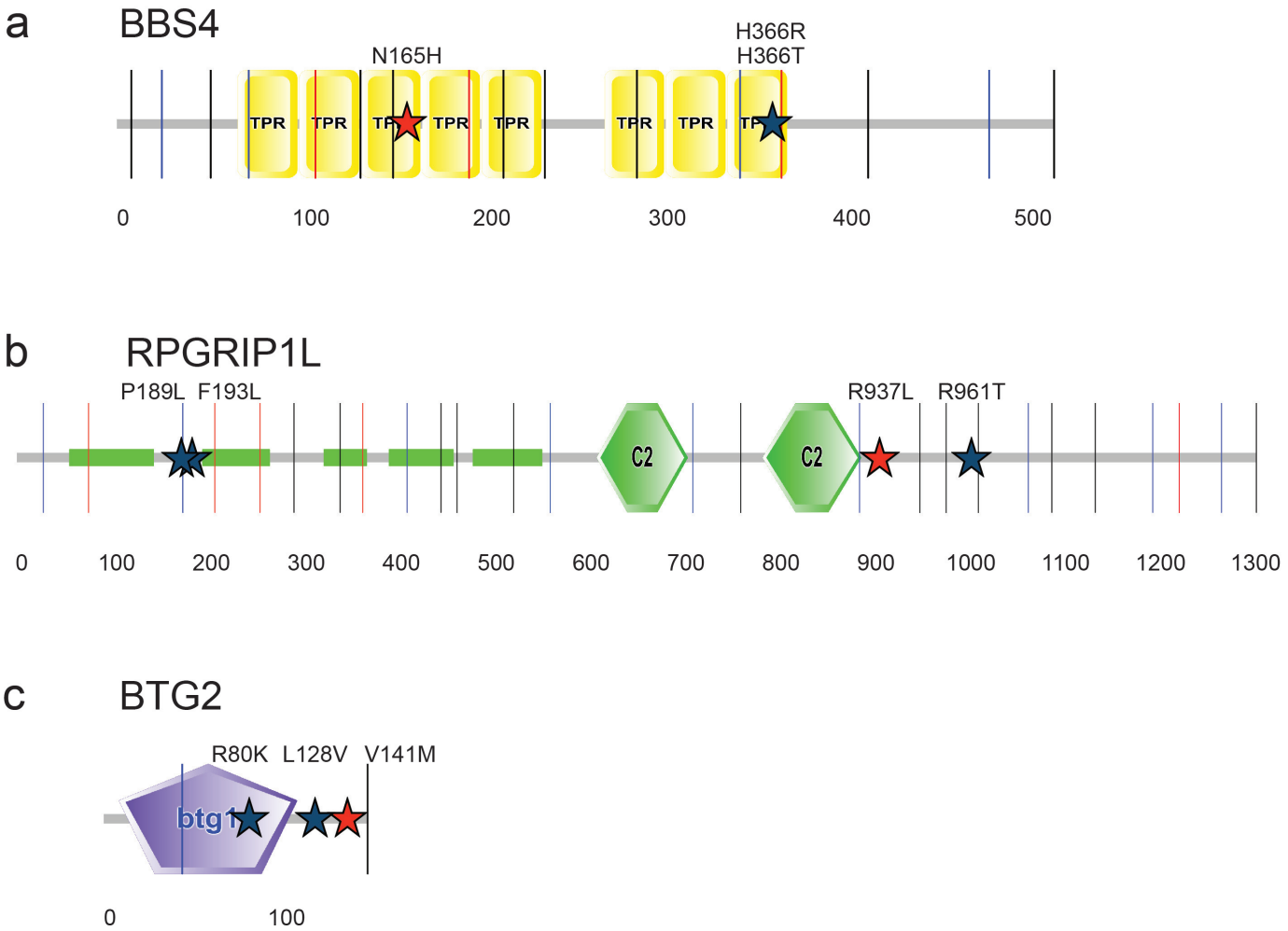
Embryos injected with *btg2* MO at were fixed in 4% paraformaldehyde at either 2 dpf or 4 dpf. Two days post-fertilization embryos were stained for HuC/HuD or PH3. HuC/HuD was scored and quantified as described[25]. Four days post-fertilization embryos were transferred to 1× PBS and bright-field dorsal images were captured; we assessed head size by measuring the distance from the anterior-most region of the forebrain to the hindbrain as defined by the attachment of the pectoral fins.

PH3-positive cell quantification was done using the Image-based Tool for Counting Nuclei (ITCN) plugin for the ImageJ software. Rolling background subtraction (25-pixel radius) and outlier removal (2.5-pixel radius, threshold = 5) were used to process images. Linear measurement of a typical cell was used to determine cell radius for ITCN analysis. Threshold level was set to 0.5. Statistical comparisons between groups were performed with a Student's *t*-test.

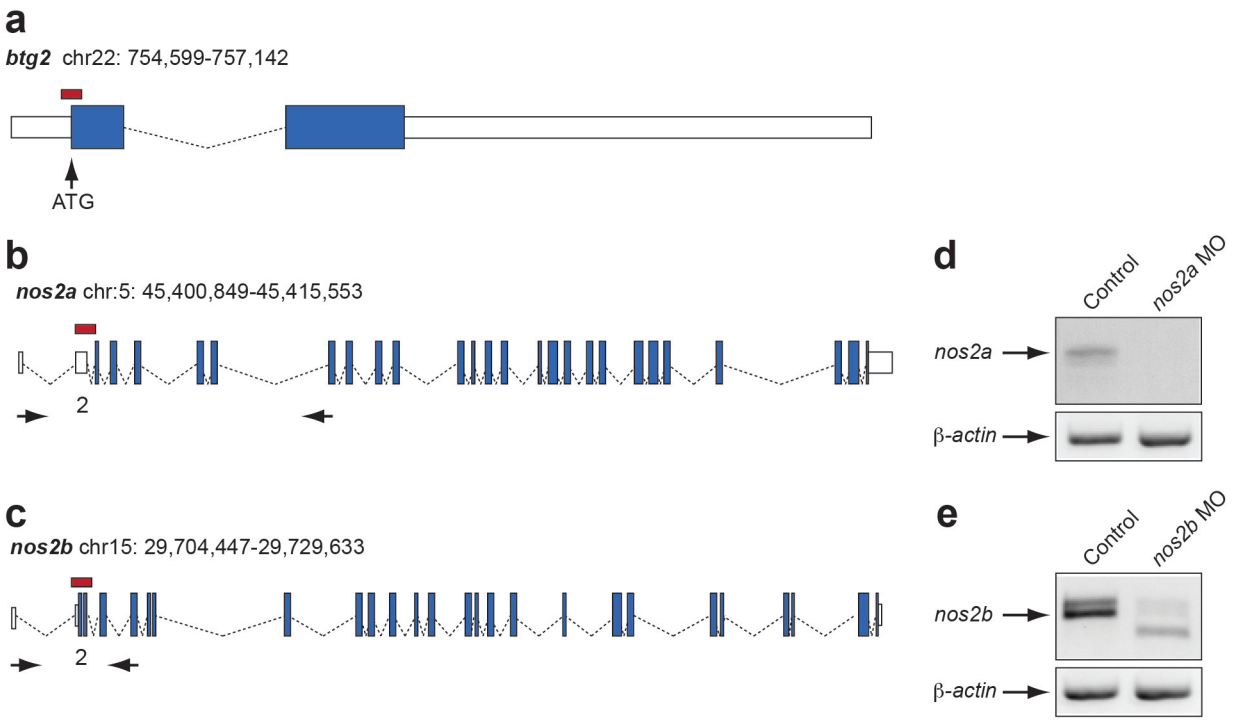
31. Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337,** 64–69 (2012).
32. Karolchik, D. *et al.* The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res.* **42,** D764–D770 (2014).
33. Flicek, P. *et al.* Ensembl 2014. *Nucleic Acids Res.* **42,** D749–D755 (2014).
34. Bainbridge, M. N. *et al.* Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities. *Genome Biol.* **12,** R68 (2011).
35. Challis, D. *et al.* An integrative variant analysis suite for whole exome next-generation sequencing data. *BMC Bioinformatics* **13,** 8 (2012).
36. Niederriter, A. R. *et al. In vivo* modeling of the morbid human genome using *Danio rerio*. *J. Vis. Exp.* **78,** e50338 (2012).



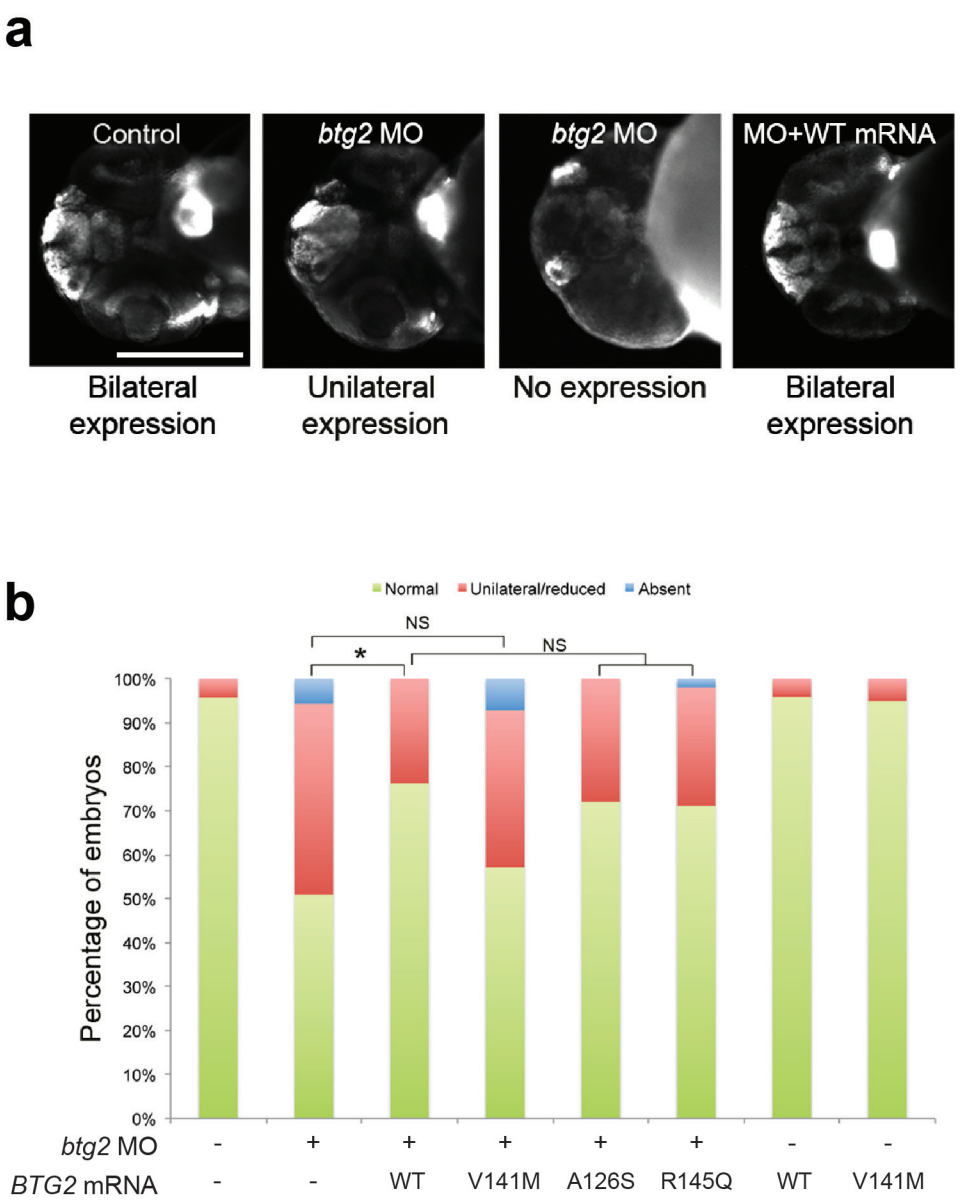
Extended Data Figure 1 | Different alignment methodologies with HumVar and ClinVar produce qualitatively similar alignments. **a**, **b**, Distributions of missense variants annotated as neutral (**a**) or pathogenic (**b**) in the HumVar and ClinVar data sets, with each of the five alignment strategies described in the text (MultiZ unfiltered, MultiZ mammals-only, EPO, MultiZ with alignment quality filter, MultiZ with >1 sequence filter). All distributions are quantitatively similar. Compare with Fig. 2c, d.



Extended Data Figure 2 | Protein domain structure of functionally tested human disease genes. **a**, Schematic of BBS4 (519 amino acids) is depicted with eight tetratricopeptide (TPR) domains (yellow); **b**, RPGRIP1L (1,315 amino acids) has multiple coiled-coil domains (green rectangles) and two protein kinase C conserved region 2 (C2) domains (green hexagons); and **c**, BTG2 (158 amino acids) has one BTG1 domain (purple pentagon). Disease-causing alleles are shown with red stars; complementing alleles are represented with blue stars; amino acid number scale in increments of 100 is shown below each schematic.



Extended Data Figure 3 | Evaluation of *btg2* and *nos2a/b* MOs. **a**–**c**, Schematic of the *D. rerio btg2*, *nos2a* and *nos2b* loci. Blue boxes, exons; dashed lines, introns; white boxes, untranslated regions; red boxes, MOs; ATG indicates the translational start site; arrows, polymerase chain reaction with reverse transcription (RT–PCR) primers; number indicates the targeted exon. **d**, **e**, Agarose gel images of *nos2a/b* RT–PCR products.



Extended Data Figure 4 | HuC/HuD staining and quantification of 2 dpf zebrafish embryos confirms pathogenicity of BTG2 V141M. **a**, Suppression of *btg2* leads to a decrease of HuC/HuD levels at 2 dpf. Representative ventral images of control, *btg2* morphants (images show unilateral or absent HuC/HuD expression), and a rescued embryo injected with a *btg2* MO plus human *BTG2* wild-type (WT) mRNA. Scale bar, 250 µm. **b**, Percentage of embryos with normal, bilateral HuC/HuD protein levels in the anterior forebrain or decreased/unilateral HuC/HuD protein levels in embryos injected with *btg2* MOs alone or MOs plus human *BTG2* wild-type or variant mRNAs (p.V141M, index case; p.A126S and p.R145Q, control alleles). $^*P < 0.05$ (two-tailed *t*-test comparisons between MO-injected and rescued embryos; $n = 38–78$ per injection batch).