nature genetics

A *de novo* paradigm for mental retardation

Lisenka E L M Vissers[1,2], Joep de Ligt[1,2], Christian Gilissen[1], Irene Janssen[1], Marloes Steehouwer[1], Petra de Vries[1], Bart van Lier[1], Peer Arts[1], Nienke Wieskamp[1], Marisol del Rosario[1], Bregje W M van Bon[1], Alexander Hoischen[1], Bert B A de Vries[1], Han G Brunner[1,3] & Joris A Veltman[1,3]

The per-generation mutation rate in humans is high. *De novo* mutations may compensate for allele loss due to severely reduced fecundity in common neurodevelopmental and psychiatric diseases, explaining a major paradox in evolutionary genetic theory. Here we used a family based exome sequencing approach to test this *de novo* mutation hypothesis in ten individuals with unexplained mental retardation. We identified and validated unique non-synonymous *de novo* mutations in nine genes. Six of these, identified in six different individuals, are likely to be pathogenic based on gene function, evolutionary conservation and mutation impact. Our findings provide strong experimental support for a *de novo* paradigm for mental retardation. Together with *de novo* copy number variation, *de novo* point mutations of large effect could explain the majority of all mental retardation cases in the population.

Recent studies[1,2] have indicated that humans have an exceptionally high per-generation mutation rate of between 7.6×10^{-9} and 2.2×10^{-8}. An average newborn is calculated to have acquired 50 to 100 new mutations in his or her genome, resulting in approximately 0.86 new amino-acid–altering mutations[2]. Spontaneous germline mutations can have serious phenotypic consequences when they affect functionally relevant bases in the genome. In fact, their occurrence may explain why diseases with a severely reduced fecundity remain frequent in the human population, especially when the mutational target is large and comprised of many genes. This would explain a major paradox in the evolutionary genetic theory of mental disorders[3,4]. In agreement with this hypothesis, *de novo* copy number variations (CNVs) are a known cause of schizophrenia, autism and mental retardation[5,6]. Much less is known about the frequency and impact of *de novo* point mutations in these common diseases. Whole genome or exome sequencing now permits the study of these mutations and their role in disease in a systematic genome-wide manner. This approach has recently been used to identify causative genes in several rare syndromes[1,7–10]. In addition, targeted resequencing of the coding exons of the X chromosome revealed nine genes associated with X-linked forms of mental retardation[11], showing the strength of these analyses in common diseases. In this study, we used a family based whole-exome–sequencing approach to test the *de novo* mutation hypothesis in an unselected cohort of individuals with mental retardation.

We sequenced the exomes of ten case-parent trios. All cases, eight males and two females, had moderate to severe mental retardation and a negative family history. Clinical evaluation did not lead to a syndromic or etiologic diagnosis (**Supplementary Note**). Prior cytogenetic analysis showed normal chromosomes, and array-based genomic profiling did not reveal *de novo* or other CNVs associated with mental retardation. In addition, fragile X syndrome was excluded by *FMR1* repeat expansion analysis. On average, we obtained 3.1 Gb of mappable sequence data per individual after exome enrichment (37 Mb of genomic sequence targeting ~18,000 genes) and sequencing on one quarter of a SOLiD sequencing slide (Online Methods and **Supplementary Table 1**). Color space reads were mapped to the reference genome. On average, 79.6% of the bases originated from the targeted exome, with 90% of the targeted exons covered at least ten times. The median exon coverage was 42-fold, indicating that the majority of variants present in each exome could be robustly detected using a custom bioinformatic analysis pipeline (**Supplementary Fig. 1**).

On average, we identified 21,755 genetic variants per individual with high confidence (**Table 1** and **Supplementary Fig. 2**). We developed an automated prioritization scheme to systematically identify all candidate dominant *de novo* mutations in each affected individual (**Fig. 1**). We first excluded all nongenic, intronic and synonymous variants other than those occurring at canonical splice sites. This first step reduced the number of candidates to an average of 5,640 non-synonymous and canonical splice site variants per affected individual. We further reduced this number to 143 by excluding all known, likely benign, variants by comparison with data from dbSNP database v130 and our in-house variant database. Next, we used the exome data from each case's parents to exclude all remaining inherited variants. This resulted in an average of five (with a range of two to seven) candidate *de novo* non-synonymous mutations per affected individual (**Table 1**).

For all 51 candidate mutations (**Supplementary Table 2**), we performed Sanger sequencing to (i) validate the mutations observed in the probands and (ii) validate the absence of the mutations in the parental DNA. Thirty-eight candidates could not be validated in the proband (covered by a median of five variant reads in the exome sequencing experiment), but 13 candidates could be validated

[1]Department of Human Genetics, Nijmegen Centre for Molecular Life Sciences and Institute for Genetic and Metabolic Disorders, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands. [2]These authors contributed equally to this work. [3]These authors jointly directed this work. Correspondence should be addressed to H.G.B. (h.brunner@antrg.umcn.nl) or J.A.V. (j.veltman@antrg.umcn.nl).

Received 11 August; accepted 18 October; published online 14 November 2010; doi:10.1038/ng.712

Table 1 Overview of all variants detected per proband and impact of the prioritization steps for selecting candidate non-synonymous *de novo* mutations

Trio	1	2	3	4	5	6	7	8	9	10	Average
High-confidence variant calls	20,810	21,658	21,338	22,647	17,694	22,333	21,369	22,658	24,085	22,962	21,755
After exclusion of nongenic, intronic and synonymous variants	5,556	5,665	5,691	5,991	4,607	5,567	5,716	5,628	5,985	5,994	5,640
After exclusion of known variants	165	159	157	155	120	136	120	149	96	171	143
After exclusion of inherited variants	4	7	3	7	7	2	2	6	6	7	5

(covered by a median of 17 variant reads). Parental analysis validated the *de novo* occurrence for 9 of these 13 mutations, detected in seven different individuals (**Table 2** and **Supplementary Figs. 3** and **4**). We did not identify these mutations in a total of 1,664 control chromosomes, nor did we see other likely pathogenic mutations identified in the affected genes in these control chromosomes, indicating that the population frequency of these types of *de novo* mutations in these genes will be lower than 0.22% (power = 0.95, $\alpha = 0.05$). Eight of the *de novo* mutations were present in a heterozygous state on the autosomes and one was present in a hemizygous state on the X chromosome. All *de novo* mutations occurred in different genes, including two genes recently implicated in mental retardation (**Table 2**). In addition to using a dominant disease model, we also analyzed the data for recessive forms of mental retardation. In the affected male of trio 10, we identified a maternally inherited non-synonymous variant in *JARID1C* (**Table 2**), which is a well-described X-linked mental retardation gene[12]. Subsequent analysis of this variant in DNA obtained from the affected individual's grandparents indicated that the mutation had occurred *de novo* in the mother of this proband. No conclusive evidence for autosomal recessive inheritance, either homozygous or compound heterozygous, was obtained for the other affected individuals.

Next, we evaluated the function of each mutated gene in relation to the disorder (**Table 2**). Three genes do not seem to play a role in biological pathways linked to mental retardation. *BPIL3* is involved in the innate immune response[13], whereas *PGA5* is involved in protease activity in the stomach[14]. The function of *ZNF599* is currently unknown. For the six other genes affected by *de novo* mutations, functional evidence suggests a role in mental retardation. Two mutations occurred in genes (*RAB39B* and *SYNGAP1*) that, when disrupted, are known to cause mental retardation (**Table 2**)[15,16]. For the remaining four mutated genes, evidence for a causal link with mental retardation is provided by model organisms and protein-protein interaction studies. *DYNC1H1* encodes a cytoplasmic dynein that acts as a motor for intracellular retrograde axonal transport. Heterozygous *Dync1h1*$^{+/-}$ mutant mice exhibit sensory neuropathy[17], and studies in zebrafish have shown the importance of *dync1h1* in correct nuclear positioning. Mislocalization of nuclei in the vertebrate central nervous system is likely to result in profound patterning defects and severely compromised function[18]. Notably, *DYNC1H1* interacts with *PAFAH1B1*, the gene associated with type I lissencephaly, which involves gross disorganization of the neurons within the cerebral cortex[19]. *YY1* encodes the ubiquitously expressed transcription factor yin-yang 1 and directs histone deacetylases and histone acetyltransferases, implicating chromatin remodeling as its main function. Complete ablation of *Yy1* in mice results in early embryonic lethality, whereas *Yy1* heterozygous mice display growth retardation, neurulation defects and brain abnormalities[20]. Recent studies show that YY1 interacts directly with MECP2; *MECP2* is mutated in Rett syndrome[21]. *DEAF1* encodes a transcription factor that regulates the 5-HT1A receptor in the human brain. Mutations in the *Drosophila DEAF1* ortholog result in early embryonic arrest, suggesting an essential role for the gene in early development[22]. Additional evidence is provided by *Deaf1*-deficient mice, which show neural tube defects including exencephaly[23]. Finally, *CIC* is a member of the HMG-box transcription factor superfamily, which is associated with neuronal and glial development of the nervous system. *CIC* is predominantly and transiently expressed in immature granule cells of the cerebellum, hippocampus and neocortex, suggesting a critical role in central nervous system development[24].

We next examined the evolutionary conservation of affected nucleotides (using the phyloP score), as well as the potential of the *de novo* mutations to affect the structure or function of the resulting proteins (using the Grantham score; **Table 2**). All *de novo* missense mutations and the inherited X-linked mutation were included in this analysis; no Grantham scores were available for the additional nonsense and frameshift mutations. Of note, *de novo* mutations in genes with a functional link to mental retardation showed a higher phyloP (mean, 4.7) and Grantham score (mean, 135) than mutations in genes without such a functional indication (mean phyloP score, −0.5 and mean Grantham score, 38). We also compared these scores to those for all non-synonymous variants in the dbSNP database as well as those in the Human Gene Mutation Database (HGMD). The distribution of phyloP scores and Grantham scores differed markedly between dbSNP and the HGMD (Online Methods and **Supplementary Fig. 5**). The four mutations in genes functionally linked to mental retardation all showed higher probability values for being observed in HGMD

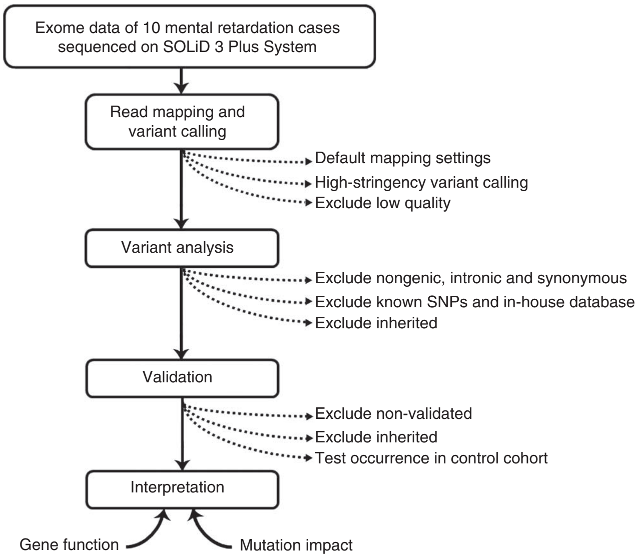
Exome data of 10 mental retardation cases sequenced on SOLiD 3 Plus System
Read mapping and variant calling
Default mapping settings
High-stringency variant calling
Exclude low quality
Variant analysis
Exclude nongenic, intronic and synonymous
Exclude known SNPs and in-house database
Exclude inherited
Validation
Exclude non-validated
Exclude inherited
Test occurrence in control cohort
Interpretation
Gene function
Mutation impact

Figure 1 Experimental work flow for detecting and prioritizing sequence variants. For all ten mental retardation trios, prioritization of variants observed in the probands was based on selection for non-synonymous changes of high quality only and exclusion of all variants previously observed in healthy individuals, together with those variants that were inherited from an unaffected parent. Interpretation of *de novo* variants was based on gene function and the impact of the mutation.

Table 2 Overview of all *de novo* variants identified by exome sequencing in ten individuals with unexplained mental retardation

Gene	Trio	Sex[a]	NM number	cDNA level change	Protein level change	PhyloP score	Grantham score	Probability of being observed in dbSNP[b]	Probability of being observed in HGMD[b]	Gene function
***De novo* mutations**										
DYNC1H1	1	M	NM_001376	c.11465A>C	p.His3822Pro	5.5	77	0.20	0.80	Retrograde axonal transporter; interacts with *PAFAH1B1* (mutation of which causes lissencephaly, a neurodevelopmental disorder)
ZNF599	1	M	NM_001007248	c.532C>T	p.Leu187Phe	–1.5	22	1.00	2.65×10^{-4}	Unknown
RAB39B	2	M	NM_171998	c.557G>A	p.Trp186X	4.8	–	–	–	Known X-linked mental retardation gene
YY1	3	M	NM_003403	c.1138G>T	p.Asp380Tyr	6.9	160	2.27×10^{-6}	1.00	Ubiquitously expressed transcription factor; mouse knockdown results in growth retardation, neurulation defects and brain abnormalities; interacts with *MECP2*, a known mental retardation gene
BPIL3	3	M	NM_174897	c.887G>A	p.Arg269His	0.5	29	0.97	0.03	Innate immune response
PGA5	4	F	NM_014224	c.1058T>C	p.Val353Ala	0.7	64	0.84	0.16	Precursor of pepsin
DEAF1	5	M	NM_021008	c.683T>G	p.Ile228Ser	4.9	142	0.01	0.99	Transcription factor; regulator of 5-HT1A receptor in the brain; mouse knockout causes neural tube defects
CIC	6	M	NM_015125	c.1474C>T	p.Arg492Trp	2.6	101	0.46	0.54	Granule cell development in central nervous system
SYNGAP1	8	F	NM_006772	c.998_999del	p.Val333AlafsX	3.3	–	–	–	Known autosomal dominant mental retardation gene
X-linked inherited mutations										
JARID1C	10	M	NM_001146702	c.1919G>A	p.Cys640Tyr	5.1	194	2.09×10^{-6}	1.00	Known X-linked mental retardation gene

[a]Sex of proband, with M for male and F for female. [b]Visual representation of probabilities are provided in **Supplementary Figure 5**. Grantham scores for nonsense (in *RAB39B*) and frameshift mutations (in *SYNGAP1*) could not be calculated.

(mean, 0.83) than for being observed in dbSNP (mean, 0.17). The three mutations in genes without a functional link to mental retardation showed an average probability of 0.94 for being observed in dbSNP and an average probability of 0.06 for being observed in HGMD (**Table 2**). Additionally, the inherited *JARID1C* mutation showed a probability of 1.00 for being in HGMD versus 2.09×10^{-6} for being in dbSNP.

This analysis of the mutated nucleotides and their impact on gene function strongly supports pathogenicity for six of the nine *de novo* mutations. Importantly, these six mutations occurred in genes with a functional link to mental retardation, two of which are known mental retardation genes. In contrast, three *de novo* variants in genes without a functional link did not appear to significantly affect protein function. Moreover, we identified a maternally inherited mutation in a known X-linked mental retardation gene that arose *de novo* in the proband's mother. Although we have not provided individual functional tests to prove causality, these data collectively provide strong evidence for a major role of *de novo* mutations in mental retardation. The identification of recurrent mutations in these genes in unrelated cases would provide additional proof for disease causality, but this may require the evaluation of thousands of affected individuals. The identification of subtle CNVs encompassing (part of) these genes may also provide additional proof for disease causality, as was shown recently for mutations in X-linked mental retardation genes[25]. As of yet, no such CNVs have been reported, nor have we found such CNVs in our diagnostic cohort of ~4,500 individuals with mental retardation (data not shown).

The discovery of nine *de novo* non-synonymous mutations in this cohort of ten affected individuals is concordant with the recently estimated background mutation rate of 0.86 amino-acid–altering mutations per newborn in controls[2], but it will be important to compare this result to similar data from healthy control trios when available. Notably, after applying the same systematic filtering approach and Sanger sequencing, we could only validate a single *de novo* synonymous mutation, which occurred in *GRIN1* (c.351C>T, seen in trio 10). This base pair is not conserved through evolution (phyloP score = –3.2) and does not seem to alter splicing, suggesting that this mutation is an unlikely candidate for causing mental retardation. Of note, the individual carrying this mutation also carries the *JARID1C* mutation. The observed ratio of non-synonymous to synonymous *de novo* mutations is far greater than would be expected for protein-coding genes under purifying selection and indicates that many of these mutations will result in a reproductive disadvantage. In contrast, the average non-synonymous to synonymous ratio reported in dbSNP for the six genes with predicted pathogenic mutations is significantly lower than that of the three genes with mutations reflecting the background mutation rate (Fisher's Exact test, $P = 0.0016$), which is to be expected for disease genes in the normal population.

In summary, our results suggest that *de novo* mutations are a major cause of unexplained mental retardation. These mutations can readily be identified using a family based exome sequencing approach and require only limited follow-up by Sanger sequencing. Our findings have implications for preventive and diagnostic strategies in mental retardation. Systematic genome-wide resequencing in parent-child trios may uncover further examples of this *de novo* paradigm for other human neurodevelopmental disorders.

URLs. 1000 Genomes Project, http://www.1000genomes.org; dbSNP, http://www.ncbi.nlm.nih.gov/projects/SNP/; HGMD, http://www.hgmd.cf.ac.uk/ac/index.php; R, http://www.r-project.org/.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Accession codes. The genomic reference sequence for *DYNC1H1* can be found under the GenBank accession number NM_001376; for *ZNF599* under NM_001007248; for *RAB39B* under NM_171998; for *YY1* under NM_003403; for *BPIL3* under NM_174897; for *PGA5* under NM_014224; for *DEAF1* under NM_021008; for *CIC* under NM_015125; for *SYNGAP1* under NM_006772; for *JARID1C* under NM_001146702; and for *GRIN1* under NM_021569.2.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank R. de Reuver and J. Hehir-Kwa for bioinformatics support in data analysis and personnel from the Sequencing Facility of our department for timely completion of Sanger sequencing of validation experiments. This work was funded in part by grants from The Netherlands Organization for Health Research and Development (ZonMW grants 916-86-016 to L.E.L.M.V., 917-66-36 and 911-08-025 to J.A.V. and 917-86-319 to B.B.A.d.V.), the EU-funded TECHGENE project (Health-F5-2009-223143 to J.d.L. and J.A.V.) and the AnEUploidy project (LSHG-CT-2006-37627 to A.H., B.W.M.v.B., H.G.B., B.B.A.d.V. and J.A.V.).

AUTHOR CONTRIBUTIONS

J.A.V., L.E.L.M.V. and H.G.B. conceived the project and planned the experiments. B.B.A.d.V. and B.W.M.v.B. performed sample collection and reviewed phenotypes. L.E.L.M.V., A.H., I.J., M.S., P.d.V., B.v.L. and P.A. performed next-generation sequencing experiments using a custom pipeline set up by C.G. and A.H. J.d.L. and C.G. analyzed and interpreted the data with support from N.W. and M.d.R. L.E.L.M.V., P.d.V., I.J. and M.S. performed validation experiments. L.E.L.M.V., J.d.L. and J.A.V. prepared the first draft of the manuscript. All authors contributed to the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Roach, J.C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636–639 (2010).
2. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl. Acad. Sci. USA* **107**, 961–968 (2010).
3. Keller, M.C. & Miller, G. Resolving the paradox of common, harmful, heritable mental disorders: which evolutionary genetic models work best? *Behav. Brain Sci.* **29**, 385–404 (2006).
4. Uher, R. The role of genetic variation in the causation of mental illness: an evolution-informed framework. *Mol. Psychiatry* **14**, 1072–1082 (2009).
5. Cook, E.H. Jr. & Scherer, S.W. Copy-number variations associated with neuropsychiatric conditions. *Nature* **455**, 919–923 (2008).
6. de Vries, B.B. *et al.* Diagnostic genome profiling in mental retardation. *Am. J. Hum. Genet.* **77**, 606–616 (2005).
7. Ng, S.B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.* **42**, 30–35 (2010).
8. Lupski, J.R. *et al.* Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N. Engl. J. Med.* **362**, 1181–1191 (2010).
9. Hoischen, A. *et al. De novo* mutations of *SETBP1* cause Schinzel-Giedion syndrome. *Nat. Genet.* **42**, 483–485 (2010).
10. Sobreira, N.L. *et al.* Whole-genome sequencing of a single proband together with linkage analysis identifies a Mendelian disease gene. *PLoS Genet.* **6**, e1000991 (2010).
11. Tarpey, P.S. *et al.* A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation. *Nat. Genet.* **41**, 535–543 (2010).
12. Jensen, L.R. *et al.* Mutations in the *JARID1C* gene, which is involved in transcriptional regulation and chromatin remodeling, cause X-linked mental retardation. *Am. J. Hum. Genet.* **76**, 227–236 (2005).
13. Mulero, J.J. *et al.* Three new human members of the lipid transfer/lipopolysaccharide binding protein family (LT/LBP). *Immunogenetics* **54**, 293–300 (2002).
14. Taggart, R.T. *et al.* Relationships between the human pepsinogen DNA and protein polymorphisms. *Am. J. Hum. Genet.* **38**, 848–854 (1986).
15. Giannandrea, M. *et al.* Mutations in the small GTPase gene *RAB39B* are responsible for X-linked mental retardation associated with autism, epilepsy, and macrocephaly. *Am. J. Hum. Genet.* **86**, 185–195 (2010).
16. Hamdan, F.F. *et al.* Mutations in *SYNGAP1* in autosomal nonsyndromic mental retardation. *N. Engl. J. Med.* **360**, 599–605 (2009).
17. Chen, X.J. *et al.* Proprioceptive sensory neuropathy in mice with a mutation in the cytoplasmic Dynein heavy chain 1 gene. *J. Neurosci.* **27**, 14515–14524 (2007).
18. Tsujikawa, M., Omori, Y., Biyanwila, J. & Malicki, J. Mechanism of positioning the cell nucleus in vertebrate photoreceptors. *Proc. Natl. Acad. Sci. USA* **104**, 14819–14824 (2007).
19. Tai, C.Y., Dujardin, D.L., Faulkner, N.E. & Vallee, R.B. Role of dynein, dynactin, and CLIP-170 interactions in LIS1 kinetochore function. *J. Cell Biol.* **156**, 959–968 (2002).
20. He, Y. & Casaccia-Bonnefil, P. The Yin and Yang of YY1 in the nervous system. *J. Neurochem.* **106**, 1493–1502 (2008).
21. Forlani, G. *et al.* The MeCP2/YY1 interaction regulates ANT1 expression at 4q35: novel hints for Rett syndrome pathogenesis. *Hum. Mol. Genet.* **19**, 3114–3123 (2010).
22. Veraksa, A., Kennison, J. & McGinnis, W. DEAF-1 function is essential for the early embryonic development of *Drosophila. Genesis* **33**, 67–76 (2002).
23. Hahm, K. *et al.* Defective neural tube closure and anteroposterior patterning in mice lacking the LIM protein LMO4 or its interacting partner Deaf-1. *Mol. Cell. Biol.* **24**, 2074–2082 (2004).
24. Lee, C.J. *et al.* CIC, a member of a novel subfamily of the HMG-box superfamily, is transiently expressed in developing granule neurons. *Brain Res. Mol. Brain Res.* **106**, 151–156 (2002).
25. Whibley, A.C. *et al.* Fine-scale survey of X chromosome copy number variants and indels underlying intellectual disability. *Am. J. Hum. Genet.* **87**, 173–188 (2010).

ONLINE METHODS

Subjects. Ten individuals with unexplained moderate to severe mental retardation (with normal karyotypes and genomic profiles obtained using 250K SNP arrays) were selected for exome sequencing (**Supplementary Note**). Family history for mental retardation was negative for all cases. Nongenic causes for mental retardation, including pre-, peri- and post-natal infection and perinatal injury, were excluded. DNA was obtained from peripheral blood from the ten probands as well as from their unaffected parents. DNA isolation was performed using QIAamp DNA Mini Kit (QIAGEN), according to the instructions of the manufacturer. This study was approved by the Medical Ethics Committee of the Radboud University Nijmegen Medical Centre, and all participants signed written informed consent.

Library generation. Exome enrichment required 3 µg of genomic DNA, and an AB SOLiD Optimized SureSelect Human Exome Kit (Agilent) was used for enrichment, containing the exonic sequences of ~18,000 genes and covering a total of ~37 Mb of genomic sequence, as specified by the company. We followed the manufacturer's instructions (version 1.5) for enrichment with a minor modification, which was the reduction of the number of post-hybridization ligation-mediated PCR cycles from 12 cycles to 9 cycles.

SOLiD sequencing. The enriched exome libraries were subsequently used for emulsion PCRs, following the manufacturer's instructions (Life Technologies), based on a library concentration of 1 picomolar (pM) (version March 2010). For each sample, one-quarter of a sequencing slide (Life Technologies) was used on a SOLiD 3 Plus System.

Mapping of variants. Color space reads were mapped to the hg18 reference genome with the SOLiD bioscope software v1.2, which utilizes an iterative mapping approach. Single-nucleotide variants were subsequently called by the diBayes algorithm[26] using high stringency settings, requiring calls on each strand. Small insertions and deletions were detected using the SOLiD Small Indel Tool. We assumed a binomial distribution with a probability of 0.5 of sequencing the variant allele at a heterozygous position. Under this assumption, at least ten reads are required to obtain a 99% probability that at least two reads contain the variant allele. Variants and indels were selected using strict quality control settings, which included the presence of at least four unique variant reads (that is, having different start sites), as well as the variant being present in at least 15% of all reads. All called variants and indels were combined and annotated using a custom analysis pipeline (resulting in HCDiff files for each individual).

Custom bioinformatic analysis pipeline. All variants reported in the HCDiff files were filtered to ensure an optimal prioritization process. For this, we first excluded all nongenic, intronic (other than canonical splice sites) and synonymous variants, reducing the number of variants to an average of 5,640 per individual. Second, all known variants were excluded by comparison with data from dbSNP v130 as well as from our in-house variant database. At the time of this study, this in-house database contained variants from (i) 78 in-house performed 'exomes', contributing 515,480 variants, and (ii) the 1000 Genomes Project (see URLs) and published data from various other studies[27–29], contributing 3,059,835 variants, thereby bringing the number of variants in the in-house database to 3,525,278. Of note, if the variant observed in the proband occurred at a genomic position known in dbSNP v130, but the change present was different in the two (for example, A/C in dbSNP but A/T in the proband), the variant was not excluded from analysis. The filtering step using this data further reduced the average number of variants to 143 per proband.

Next, for a dominant model of disease, we used the exome data from accompanying parents to exclude all inherited variants. This step further reduced the number of potential *de novo* variants to an average of 33 per proband. As not all variants identified in the exomes of the probands may have been sequenced at sufficient coverage in the parental samples, we checked all remaining variants in the exome data from the accompanying parents. In brief, even if only a single read showed the variant allele in one of the parental exome samples, the variant was excluded for validation in the proband. Simultaneously, we checked all remaining potential *de novo* indels for annotation differences in each child-parent trio and excluded those that were found to be identical variants in both parent and child. After this final check, an average of five potential *de novo* variants per proband remained for further validation.

To evaluate the presence of recessive mutations, variant filtering was essentially performed as described above, with the main difference being that uniquely inherited parental variants were not excluded here. The remaining variants were evaluated for the presence of compound heterozygous variants, as well as variants that were present in >80% of all reads. Subsequently, parental exome data were used for segregation analysis of the variants identified.

dbSNP and HGMD. To explore the pathogenicity of our *de novo* variants, the genomic evolutionary conservation score (phyloP) and the amino-acid change (Grantham) were compared to those scores present in dbSNP (build 130) and the HGMD (see URLs). All non-synonymous changes reported in dbSNP and HGMD were retrieved, and overlap between databases was removed from both datasets. In addition, non-synonymous variants in dbSNP with an OMIM disease entry, suggestive for a Mendelian phenotype, were omitted from the dbSNP dataset.

Next, quadratic discriminant analysis[30] was performed on these two datasets to determine the significance of the phyloP and Grantham scores as discriminating factors. Statistical tests were performed using the R statistics package (see URLs). The assumption of normality in the data required for the model was determined using Lilliefors (Kolmogorov-Smirnov) normality testing[31]:

PhyloP $D = 0.0626$, $P < 2.2 \times 10^{-16}$; Grantham $D = 0.0828$, $P < 2.2 \times 10^{-16}$; PhyloP × Grantham $D = 0.1395$, $P < 2.2 \times 10^{-16}$. D represents the maximum absolute difference between the empirical and hypothetical cumulative distribution function.

The combination of both scores together yielded the highest power to discriminate the two datasets, and as such, the combined value was used to calculate probabilities for our *de novo* variants to be observed in either database.

Validation experiments. Validation and *de novo* testing for candidate *de novo* mutations was performed using standard Sanger sequencing approaches. Primers were designed to surround the candidate mutation, and PCR reactions were performed using RedTaq Readymix PCR reaction mix (Sigma-Aldrich). Primer sequences and PCR conditions are available upon request. For all *de novo* mutations identified, an additional control cohort of 75 ethnically matched controls was tested for the presence of the same mutation by Sanger sequencing. Together with the results from 679 control individuals from the 1000 Genomes Project as well as the 78 'exomes' present in our in-house database, the control cohort for the *de novo* mutations encompassed 1,664 control chromosomes.

26. Marth, G.T. *et al.* A general approach to single-nucleotide polymorphism discovery. *Nat. Genet.* **23**, 452–456 (1999).
27. Ng, S.B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276 (2009).
28. Pushkarev, D., Neff, N.F. & Quake, S.R. Single-molecule sequencing of an individual human genome. *Nat. Biotechnol.* **27**, 847–852 (2009).
29. Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60–65 (2008).
30. Venables, W.N. & Ripley, B.D. *Modern Applied Statistics with S* (Springer, 4th edn., New York, New York, USA, 2002).
31. Lilliefors, H. On the Kolmogorov–Smirnov test for normality with mean and variance unknown. *J. Am. Stat. Assoc.* **62**, 399–402 (1967).