LETTER

doi:10.1038/nature13394

Genome sequencing identifies major causes of severe intellectual disability

Christian Gilissen[1]*, Jayne Y. Hehir-Kwa[1]*, Djie Tjwan Thung[1], Maartje van de Vorst[1], Bregje W. M. van Bon[1], Marjolein H. Willemsen[1], Michael Kwint[1], Irene M. Janssen[1], Alexander Hoischen[1], Annette Schenck[1], Richard Leach[2], Robert Klein[2], Rick Tearle[2], Tan Bo[1,3], Rolph Pfundt[1], Helger G. Yntema[1], Bert B. A. de Vries[1], Tjitske Kleefstra[1], Han G. Brunner[1,4]*, Lisenka E. L. M. Vissers[1]* & Joris A. Veltman[1,4]*

Severe intellectual disability (ID) occurs in 0.5% of newborns and is thought to be largely genetic in origin[1,2]. The extensive genetic heterogeneity of this disorder requires a genome-wide detection of all types of genetic variation. Microarray studies and, more recently, exome sequencing have demonstrated the importance of *de novo* copy number variations (CNVs) and single-nucleotide variations (SNVs) in ID, but the majority of cases remain undiagnosed[3–6]. Here we applied whole-genome sequencing to 50 patients with severe ID and their unaffected parents. All patients included had not received a molecular diagnosis after extensive genetic prescreening, including microarray-based CNV studies and exome sequencing. Notwithstanding this prescreening, 84 *de novo* SNVs affecting the coding region were identified, which showed a statistically significant enrichment of loss-of-function mutations as well as an enrichment for genes previously implicated in ID-related disorders. In addition, we identified eight *de novo* CNVs, including single-exon and intra-exonic deletions, as well as interchromosomal duplications. These CNVs affected known ID genes more frequently than expected. On the basis of diagnostic interpretation of all *de novo* variants, a conclusive genetic diagnosis was reached in 20 patients. Together with one compound heterozygous CNV causing disease in a recessive mode, this results in a diagnostic yield of 42% in this extensively studied cohort, and 62% as a cumulative estimate in an unselected cohort. These results suggest that *de novo* SNVs and CNVs affecting the coding region are a major cause of severe ID. Genome sequencing can be applied as a single genetic test to reliably identify and characterize the comprehensive spectrum of genetic variation, providing a genetic diagnosis in the majority of patients with severe ID.

Whole-genome sequencing (WGS) is considered to be the most comprehensive genetic test so far[7], but widespread application to patient diagnostics has been hampered by challenges in data analysis, the unknown diagnostic potential of the test, and relatively high costs. In this study, the genomes of 50 patients with severe ID and their unaffected parents were sequenced to an average genome-wide coverage of 80 fold (Supplementary Table 1)[8]. Before inclusion in the study, patients underwent an extensive clinical and genetic work-up, including targeted gene analysis, genomic microarray analysis and whole-exome sequencing (WES)[6], but no molecular diagnosis could be established (Fig. 1).

On average, 98% of the genome was called for both alleles, giving rise to 4.4 million SNVs and 276 CNVs per genome (Supplementary Table 2). WGS identified an average of 22,186 coding SNVs per individual, encompassing more than 97% of variants identified previously by WES (Supplementary Tables 2, 3). We focused our analysis first on *de novo* SNVs and CNVs because of their importance in ID[4]. On average, 82 high-confidence potential *de novo* SNVs were called per genome (Supplementary Methods and Supplementary Table 4), which is in concordance with previous studies[9–11]. Systematic validation by Sanger sequencing of putative *de novo* variants in the protein-coding regions resulted in a total of 84 coding *de novo* mutations in 50 patients, giving rise to a protein-coding *de novo* substitution rate of 1.58 (Supplementary Methods and Supplementary Tables 5, 6, 7, 8). This rate exceeds all previously published substitution rates[11–15] obtained using WES (Supplementary Table 9), as well as inferred substitution rates ($P = 3.58 \times 10^{-5}$) (ref. 14). In addition, this set of *de novo* mutations is significantly enriched for loss-of-function mutations ($P = 1.594 \times 10^{-5}$; Supplementary Methods).

Next, we investigated whether *de novo* mutations occurred in genes that have previously been identified in other patients with ID and/or overlapping phenotypes such as autism, schizophrenia or epilepsy[12–19]. To this end, we compiled two sets of genes, one set containing 528 genes harbouring mutations in at least five patients with ID (referred to as 'known ID genes') and one list containing 628 genes harbouring mutations in at least one, but less than five patients (referred to as 'candidate ID genes') (Supplementary Methods). It has recently been shown that Mendelian disease genes are less tolerant to functional genetic variation than genes that do not cause any known disease[20]. In line with this, both the set of known ID genes and the set of candidate ID genes indeed showed significantly less tolerance for functional variation ($P < 1.0 \times 10^{-6}$ for both sets; Extended Data Fig. 1 and Supplementary Methods). Subsequent analysis of our 84 *de novo* mutations at the gene level revealed significantly more mutations in known ID genes than expected (nine genes, $P = 0.04$; Supplementary Table 10). Mutations in these known ID genes included four insertion/deletion events, two nonsense mutations and three highly conserved missense mutations, thereby

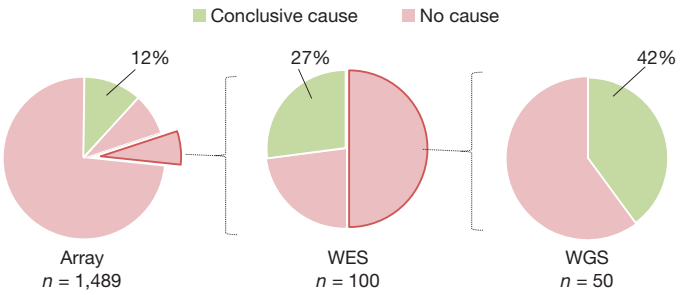


Figure 1 | Study design and diagnostic yield in patients with severe ID per technology. Diagnostic yield for patients with severe ID (IQ < 50), specified by technology: genomic microarrays, WES and WGS. Percentages indicate the number of patients in whom a conclusive cause was identified using the specified technique. Brackets indicate the group of patients in whom no genetic cause was identified and whose DNA was subsequently analysed using the next technology. WES data are updated with permission from ref. 6 (see Supplementary Methods).

[1]Department of Human Genetics, Radboud Institute for Molecular Life Sciences and Donders Centre for Neuroscience, Radboud University Medical Center, Geert Grooteplein 10, 6525 GA Nijmegen, the Netherlands. [2]Complete Genomics Inc. 2071 Stierlin Court, Mountain View, California 94043, USA. [3]State Key Laboratory of Medical Genetics, Central South University. 110 Xiangya Road, Changsha, Hunan 410078, China. [4]Department of Clinical Genetics, Maastricht University Medical Centre. Universiteitssingel 50, 6229 ER Maastricht, the Netherlands.
*These authors contributed equally to this work.

©2014 Macmillan Publishers Limited. All rights reserved

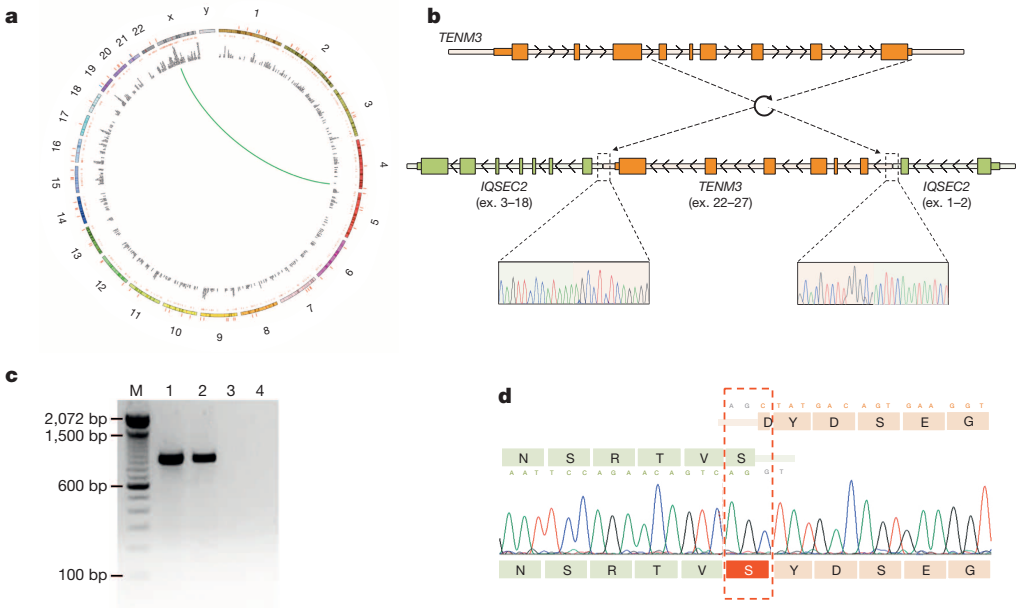


Figure 2 | Detected duplication of a chromosome 4 region into the X-chromosomal *IQSEC2* gene. **a**–**d**, Graphical representation of a *de novo* duplication–insertion event in patient 31. **a**, Circos plot with chromosome numbers and *de novo* mutations in the outer shell. Red bars represent genome-wide potential *de novo* SNVs, whereas blue lines represent potential *de novo* CNVs/structural variants. Inner shell represents the location of known ID genes (red marks) with the respective gene names. Green line illustrates a duplication event on chromosome 4, which is inserted into chromosome X. **b**, Details for inserted duplication event on chromosome X. The last six exons of *TENM3* are inserted in inverted orientation into intron 2 of *IQSEC2*, predicted to result in an in-frame *IQSEC2-TENM3* fusion gene. ex., exon. **c**, **d**, PCR (**c**) on and Sanger sequencing (**d**) of complementary DNA junction fragment in patient 31. Lanes in **c** represent the following: M, 100 bp marker; 1, cDNA of patient with cyclohexamide treatment; 2, cDNA of patient without cyclohexamide treatment; 3, control cDNA with cyclohexamide treatment; 4, control cDNA without cyclohexamide treatment. Our data verify the presence of a fusion gene in patient 31 that is suggested to escape nonsense-mediated decay.

also showing an enrichment for loss-of-function mutations ($P = 4.88 \times 10^{-6}$). Also, a significant enrichment for *de novo* mutations ($n = 10$) in candidate ID genes was identified ($P = 0.013$), including three loss-of-function mutations ($P = 0.02$) and three highly conserved missense mutations (Supplementary Table 11). These mutated known and candidate ID genes showed a diminished tolerance to functional variation ($P = 5.59 \times 10^{-6}$ and $P = 0.0042$, respectively), similar to what was observed for the entire set of known and candidate ID genes. These statistical analyses on SNVs together indicate that we not only identified significantly more *de novo* mutations in these 50 patients with severe ID, but also that they are more severe and occur more often in known or candidate ID genes.

In addition to the detection of *de novo* SNVs and small insertion/deletion events, a total of eight *de novo* structural variants, or CNVs, were identified and validated. These structural variants included five deletions, a tandem duplication, an interchromosomal duplication and one complex inversion/duplication/deletion event (Extended Data Table 1). All of these events had previously remained undetected by diagnostic microarray analysis. Three deletions were smaller than 10 kilobases (kb) in size, including two single-exon deletions and one intra-exonic deletion. Four of the *de novo* deletions encompassed a known ID gene and one a candidate ID gene, resulting in a significant enrichment for CNVs affecting known ID genes ($P = 0.015$). In addition, six *de novo* CNVs contained a gene in which exonic CNVs occur significantly more frequent in patients with ID ($n = 7{,}743$) compared to control individuals ($n = 4{,}056$) (Extended Data Table 1). Local realignment of sequence reads provided accurate single-nucleotide breakpoint information for six of the events, which was readily confirmed by breakpoint-spanning polymerase chain reactions (PCRs) (Extended Data Figs 2–5). Discordant reads not only provided the precise breakpoint sequences, but also positional information for duplicated sequences. In one case a partial duplication of *TENM3* on chromosome 4 was invertedly inserted into *IQSEC2* on the X chromosome. RNA studies confirmed the formation of a stable in-frame *IQSEC2-TENM3* gene fusion (Fig. 2), thereby suggesting that disruption of *IQSEC2*, a known ID gene, may well contribute to the patient's phenotype. The contribution of such fusion genes to disease is well known in tumorigenesis, but has only recently been systematically investigated in neurodevelopmental disorders[21].

Interestingly, three of ten *de novo* SNVs occurring in candidate ID genes seemed to be present in a mosaic state in the proband on the basis of the fraction of sequence reads containing the mutated allele. Sanger sequencing and amplicon-based deep sequencing confirmed the presence of mosaic mutations in these patients, at levels of 21% (*PIAS1*), 22% (*HIVEP2*) and 20% (*KANSL2*) (Extended Data Fig. 6), of which *KANSL2* is predicted to be deleterious owing to altered splicing. It is important that mosaic events like these can be detected by WGS as they are a known cause of genetic disease[22]. An additional advantage of genome sequencing over other approaches is that it may reveal pathogenic mutations in the non-coding part of the genome. In a systematic attempt to study the role of *de novo* non-coding mutations in ID, we selected all high-confidence candidate *de novo* mutations located either within the promoter regions, introns or untranslated regions of all known ID genes and validated 43 mutations (Supplementary Tables 12, 13). Annotation of these mutations using several ENCODE resources[23], including chromatin state segments of nine human cell types and transcription-factor-binding sites, did not reveal potential pathogenic non-coding mutations (Supplementary Methods). However, our understanding of non-coding variation is still limited and extensive functional follow-up will be required to determine its role in disease[23].

In addition to the statistical analysis of our data, we also assessed the impact of our genome sequencing study in a clinical diagnostic setting, in which variant interpretation is combined with an evaluation of patients' phenotypes to make a diagnostic decision on a per patient basis. Therefore, all *de novo* coding mutations (SNVs and CNVs) were evaluated for pathogenicity on the basis of established diagnostic criteria (Supplementary Methods and Extended Data Table 1)[24–27]. For patients with *de novo* mutations in a known or candidate ID gene this clinical diagnostic assessment also included a comparison of the phenotype observed

©2014 Macmillan Publishers Limited. All rights reserved

Table 1 | Diagnostic yield by WGS for a pre-screened cohort of 50 ID trios

Genetic cause	Number of patients
Total positive diagnosis	21
Dominant *de novo*	20
Autosomal SNV	11
Autosomal CNV	5
X-linked SNV	2
X-linked CNV	2
Recessive	1
Homozygous	0
Compound heterozygous	1
X-linked	0
Candidate ID genes	8
No diagnosis	21

in our patient with those reported in the literature. Conclusive diagnoses were reached for fourteen patients with *de novo* mutations affecting a known ID gene (nine SNVs and five CNVs), as well as for six patients with *de novo* mutations affecting a candidate ID gene (four SNVs and two CNVs) (Extended Data Tables 1, 2 and Supplementary Tables 8, 14).

Although family history for ID was negative for all patients included in this study, we evaluated the presence of recessively inherited causes of disease due to mutations in known ID genes (Supplementary Table 10). We did not find X-linked maternally inherited variants in male patients consistent with the patient's phenotype, nor did we identify relevant homozygous or compound heterozygous SNVs on the autosomes. We did, however, identify a single proband carrying compound heterozygous deletions affecting the *VPS13B* gene, one of the known ID genes. Subsequent breakpoint sequencing confirmed that the 122 kb deletion, affecting exons 12–18, was paternally inherited whereas the 1.7 kb deletion of the last exon was maternally inherited (Extended Data Fig. 7). Notably, Cohen syndrome[28] was part of the differential diagnosis of this patient but no causative SNVs or CNVs were previously detected in this gene by direct Sanger sequencing or microarray analysis.

Taken together, a conclusive diagnosis was made in 21 of 50 patients with severe ID in this well-studied cohort (42%; Table 1 and Supplementary Table 15). The experimental set-up of our study allowed us to estimate the diagnostic yield of WGS in an unbiased cohort of such patients (Fig. 1). On the basis of established diagnostic rates for genomic microarrays (12%) and WES (27%) in patients from the same large cohort[6,26], the cumulative estimate for WGS to reach a conclusive genetic diagnosis is 62%, of which 60% by *de novo* events (39% SNVs, 21% CNVs) and 2% by recessive inheritance (Fig. 3 and Supplementary Methods). The role of *de novo* somatic mutations and *de novo* mutations outside the coding regions remains to be fully explored.

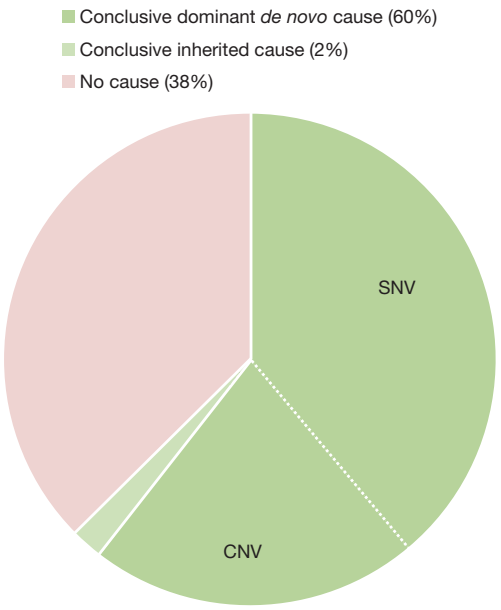


Figure 3 | Pie chart showing role of *de novo* mutations in severe ID. Contribution of genetic causes to severe ID on the basis of the cumulative estimates provided per technology. Our data indicate that *de novo* mutations are a major cause of severe ID. Note, small variants include SNVs and insertion/deletion events whereas large variants include structural variants and CNVs (>500 bp).

METHODS SUMMARY

Patients were selected to have severe ID (IQ < 50) and negative results on diagnostic genomic microarrays and exome sequencing[6] (Fig. 1). WGS was performed by Complete Genomics as previously described[8,29]. *De novo* SNVs were identified using Complete Genomics' cgatools 'calldiff' program. CNVs and structural variants were reported by Complete Genomics on the basis of read-depth deviations and discordant read pairs, respectively. *De novo* CNVs and structural variants were then identified by excluding variants with minimal evidence or overlapping with CNVs and structural variants identified in the parents or control data sets. All variants were annotated using an in-house analysis pipeline and subsequently prioritized for validation based on their confidence level (low/medium/high) and location in the genome (coding/non-coding). High-confidence candidate *de novo* mutations in non-coding variants in known ID genes were prioritized on the basis of evolutionary conservation and overlap with ENCODE chromatin state segments and transcription-factor-binding sites[23]. Statistical overrepresentation of mutations in known and candidate ID gene lists was calculated using Fisher's exact test based on RefSeq genes. Enrichments for loss-of-function CNV events were calculated using the exact Poisson test. To clinically interpret (*de novo*) mutations, each variant (both CNV and SNV) was assessed for mutation impact as well as functional relevance to ID according to diagnostic protocols for variant interpretation[6,24–27]. The diagnostic yield of WGS in an unbiased cohort was calculated based on cumulative estimates of diagnostic yield per technology (genomic microarray, WES and WGS).

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 7 March 2014; accepted 17 April 2014.
Published online 4 June 2014.

1. Ropers, H. H. Genetics of early onset cognitive impairment. *Annu. Rev. Genomics Hum. Genet.* **11,** 161–187 (2010).
2. Mefford, H. C., Batshaw, M. L. & Hoffman, E. P. Genomics, intellectual disability, and autism. *N. Engl. J. Med.* **366,** 733–743 (2012).
3. de Vries, B. B. *et al.* Diagnostic genome profiling in mental retardation. *Am. J. Hum. Genet.* **77,** 606–616 (2005).
4. Vissers, L. E. *et al.* A *de novo* paradigm for mental retardation. *Nature Genet.* **42,** 1109–1112 (2010).
5. Rauch, A. *et al.* Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* **380,** 1674–1682 (2012).
6. de Ligt, J. *et al.* Diagnostic exome sequencing in persons with severe intellectual disability. *N. Engl. J. Med.* **367,** 1921–1929 (2012).
7. Lupski, J. R. *et al.* Whole-genome sequencing in a patient with Charcot–Marie–Tooth neuropathy. *N. Engl. J. Med.* **362,** 1181–1191 (2010).
8. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327,** 78–81 (2010).
9. Michaelson, J. J. *et al.* Whole-genome sequencing in autism identifies hot spots for *de novo* germline mutation. *Cell* **151,** 1431–1442 (2012).
10. Kong, A. *et al.* Rate of *de novo* mutations and the importance of father's age to disease risk. *Nature* **488,** 471–475 (2012).
11. Jiang, Y. H. *et al.* Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing. *Am. J. Hum. Genet.* **93,** 249–263 (2013).
12. O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature* **485,** 246–250 (2012).
13. Iossifov, I. *et al. De novo* gene disruptions in children on the autistic spectrum. *Neuron* **74,** 285–299 (2012).
14. Neale, B. M. *et al.* Patterns and rates of exonic *de novo* mutations in autism spectrum disorders. *Nature* **485,** 242–245 (2012).
15. Sanders, S. J. *et al. De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485,** 237–241 (2012).
16. Epi4K Consortium & Epilepsy Phenome/Genome Project. *De novo* mutations in epileptic encephalopathies. *Nature* **501,** 217–221 (2013).
17. Gulsuner, S. *et al.* Spatial and temporal mapping of *de novo* mutations in schizophrenia to a fetal prefrontal cortical network. *Cell* **154,** 518–529 (2013).
18. Xu, B. *et al.* Exome sequencing supports a *de novo* mutational paradigm for schizophrenia. *Nature Genet.* **43,** 864–868 (2011).
19. Girard, S. L. *et al.* Increased exonic *de novo* mutation rate in individuals with schizophrenia. *Nature Genet.* **43,** 860–863 (2011).
20. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.* **9,** e1003709 (2013).

©2014 Macmillan Publishers Limited. All rights reserved

21. Rippey, C. *et al.* Formation of chimeric genes by copy-number variation as a mutational mechanism in schizophrenia. *Am. J. Hum. Genet.* **93,** 697–710 (2013).
22. Biesecker, L. G. & Spinner, N. B. A genomic view of mosaicism and human disease. *Nature Rev. Genet.* **14,** 307–320 (2013).
23. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
24. Bell, J. B. D., Sistermans, E. & Ramsden, S. C. *Practice guidelines for the Interpretation and Reporting of Unclassified Variants (UVs) in Clinical Molecular Genetics* (The UK Clinical Molecular Genetics Society and the Dutch Society of Clinical Genetic Laboratory Specialists, 2007).
25. Berg, J. S., Khoury, M. J. & Evans, J. P. Deploying whole genome sequencing in clinical practice and public health: meeting the challenge one bin at a time. *Genet. Med.* **13,** 499–504 (2011).
26. Vulto-van Silfhout, A. T. *et al.* Clinical significance of *de novo* and inherited copy-number variation. *Hum. Mutat.* **34,** 1679–1687 (2013).
27. Hehir-Kwa, J. Y., Pfundt, R., Veltman, J. A. & de Leeuw, N. Pathogenic or not? Assessing the clinical relevance of copy number variants. *Clin. Genet.* **84,** 415–421 (2013).
28. Kolehmainen, J. *et al.* Cohen syndrome is caused by mutations in a novel gene, *COH1*, encoding a transmembrane protein with a presumed role in vesicle-mediated sorting and intracellular protein transport. *Am. J. Hum. Genet.* **72,** 1359–1369 (2003).

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

Acknowledgements We thank R. Drmanac, K. Albers, J. Goeman, D. Lugtenberg and P. N. Robinson for useful discussions, and M. Steehouwer, P. de Vries and W. Nillesen for technical support. This work was in part financially supported by grants from the Netherlands Organization for Scientific Research (912-12-109 to J.A.V., A.S. and B.B.A.d.V., 916-14-043 to C.G., 916-12-095 to A.H., 907-00-365 to T.K. and SH-271-13 to C.G. and J.A.V.) and the European Research Council (ERC Starting grant DENOVO 281964 to J.A.V.).

Author Contributions Laboratory work: M.K., I.M.J., T.B., A.H., L.E.L.M.V. Clinical investigation: B.W.M.v.B., M.H.W., B.B.A.d.V., T.K., H.G.B. Data analysis: C.G., J.Y.H.-K., D.T.T., M.v.d.V., R.T. Generation of ID gene list: C.G., A.S., R.P., H.G.Y., T.K., L.E.L.M.V. Data interpretation: L.E.L.M.V., R.P., H.G.Y. Study design: J.A.V., H.G.B., R.L., R.K. Supervision of the study: H.G.B., L.E.L.M.V., J.A.V. Manuscript writing: C.G., J.Y.H.-K., H.G.B., L.E.L.M.V., J.A.V.

Author Information Data included in this manuscript have been deposited at the European Genome-phenome Archive (https://www.ebi.ac.uk/ega/home) under accession number EGAS00001000769. Reprints and permissions information is available at www.nature.com/reprints. Readers are welcome to comment on the online version of the paper. The authors declare competing financial interests: details are available in the online version of the paper. Correspondence and requests for materials should be addressed to J.A.V. (joris.veltman@radboudumc.nl).

©2014 Macmillan Publishers Limited. All rights reserved

METHODS

Patient selection. Patients were selected to have severe ID (IQ < 50) and negative results on diagnostic genomic microarrays and exome sequencing[6] (Fig. 1 and Supplementary Methods).

Whole genome sequencing. WGS was performed by Complete Genomics as previously described[8]. Sequence reads were mapped to the reference genome (GRCh37) and variants were called by local *de novo* assembly according to the methods previously described[29].

Identification of *de novo* small variants. *De novo* SNVs were identified using Complete Genomics' cgatools 'calldiff' program. On the basis of the rank order of the two confidence scores of a *de novo* mutation, we binned the variants in three groups: low confidence (at least one score < 0), medium confidence (both scores ≥ 0 but at least one < 5) and high confidence (both scores ≥ 5) (Supplementary Methods).

Identification of X-linked, recessive and compound heterozygous SNVs. Maternally inherited X-linked variants (in male patients), homozygous variants and compound heterozygous variant pairs were identified using the Complete Genomics' cgatools 'listvariants' and 'testvariants' programs to select variants according to their respective segregation. Compound heterozygous variants affecting the same gene were identified using RefSeq gene annotation (Supplementary Methods).

Identification of *de novo* CNVs and structural variants. CNVs were reported by Complete Genomics on the basis of read-depth deviations across 2 kb windows. Structural variants were reported by Complete Genomics based on discordant read pairs. *De novo* CNVs/structural variants were then identified by excluding variants with minimal evidence or overlapping with CNVs/structural variants identified in the parents or control data sets (Supplementary Methods).

Generation of lists for known and candidate ID genes. To prioritize and for subsequent interpretation of *de novo* variants for each patient individually, two gene lists were generated, one containing known ID genes (defined by five or more patients with ID having a mutation in the respective gene) and one containing candidate ID genes (defined by at least one but less than five patients with ID (or a related phenotype) showing a mutation in the respective gene) (Supplementary Methods).

Prioritization of clinically relevant SNVs and CNVs or structural variants. All SNVs were annotated using an in-house analysis pipeline. Variants were prioritized for validation in two distinct ways: (1) medium and high-confidence *de novo* SNVs and *de novo* CNVs/structural variants affecting coding regions and/or canonical splice sites; and (2) all potential *de novo* variants within known ID genes, irrespective of confidence level. Interpretation of coding *de novo* variants was performed as described previously[6]. High-confidence candidate *de novo* mutations in non-coding variants were prioritized on the basis of evolutionary conservation and overlap with ENCODE chromatin state segments and transcription-factor-binding sites (Supplementary Methods)[23].

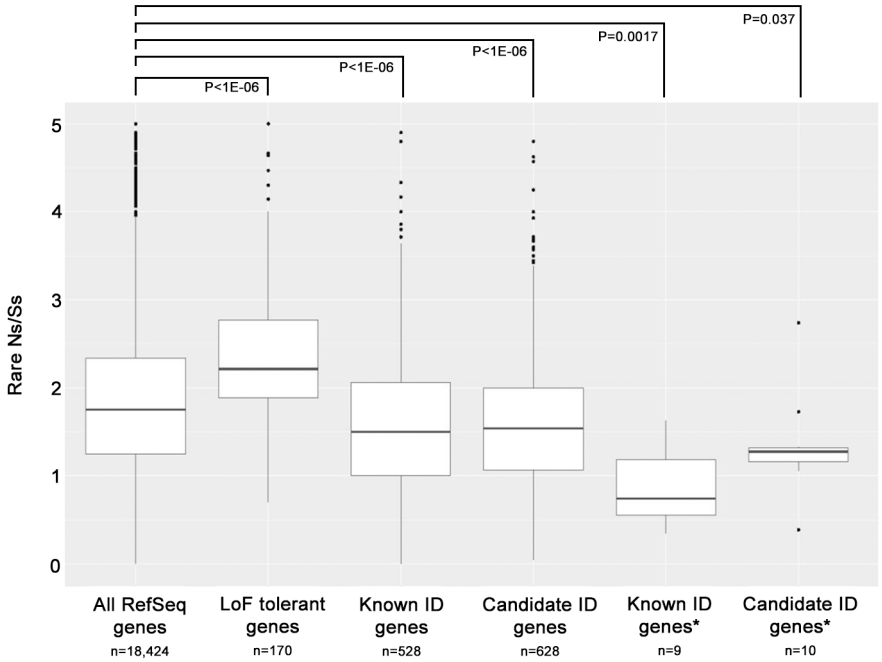
Clinical interpretation of mutations. To clinically interpret (*de novo*) mutations, each *de novo* mutation (both CNV and SNV) was assessed for mutation impact as well as functional relevance to ID according to diagnostic protocols for variant interpretation[24–27] that are used in our accredited diagnostic laboratory for genetic analysis (accredited to the 'CCKL Code of Practice', which is based on EN/ISO 15189 (2003), registration numbers R114/R115, accreditation numbers 095/103) (Supplementary Methods).

Statistical analysis. Overrepresentation of mutations in gene lists was calculated using Fisher's exact test based on the total coding size of all RefSeq genes and coding size of the genes from the respective gene list. Overrepresentation of loss-of-function mutations was calculated using Fisher's exact test based on published control cohorts. Enrichments for loss-of-function CNV events were calculated using the exact Poisson test. Enrichment for known ID genes was calculated using Fisher's exact test and odds ratios were calculated to compare the frequency of exonic CNVs in ID and control cohorts, respectively (Supplementary Methods).

Calculation of diagnostic yield. Our in-house phenotypic database contains 1,489 patients with severe ID who have all had a diagnostic genomic microarray in the time period 2003–2013. In 173 (11.6%) of these patients, a *de novo* CNV was identified as a cause of ID. Subsequently, 100 array-negative patients were subjected to WES, which resulted in a *de novo* cause for ID in 27% of patients. Of all WES-negative patients, 50 were selected for this WGS study, in which 42% obtained a conclusive genetic cause. Cumulative estimates were subsequently determined using the diagnostic yield per technology (Supplementary Methods).

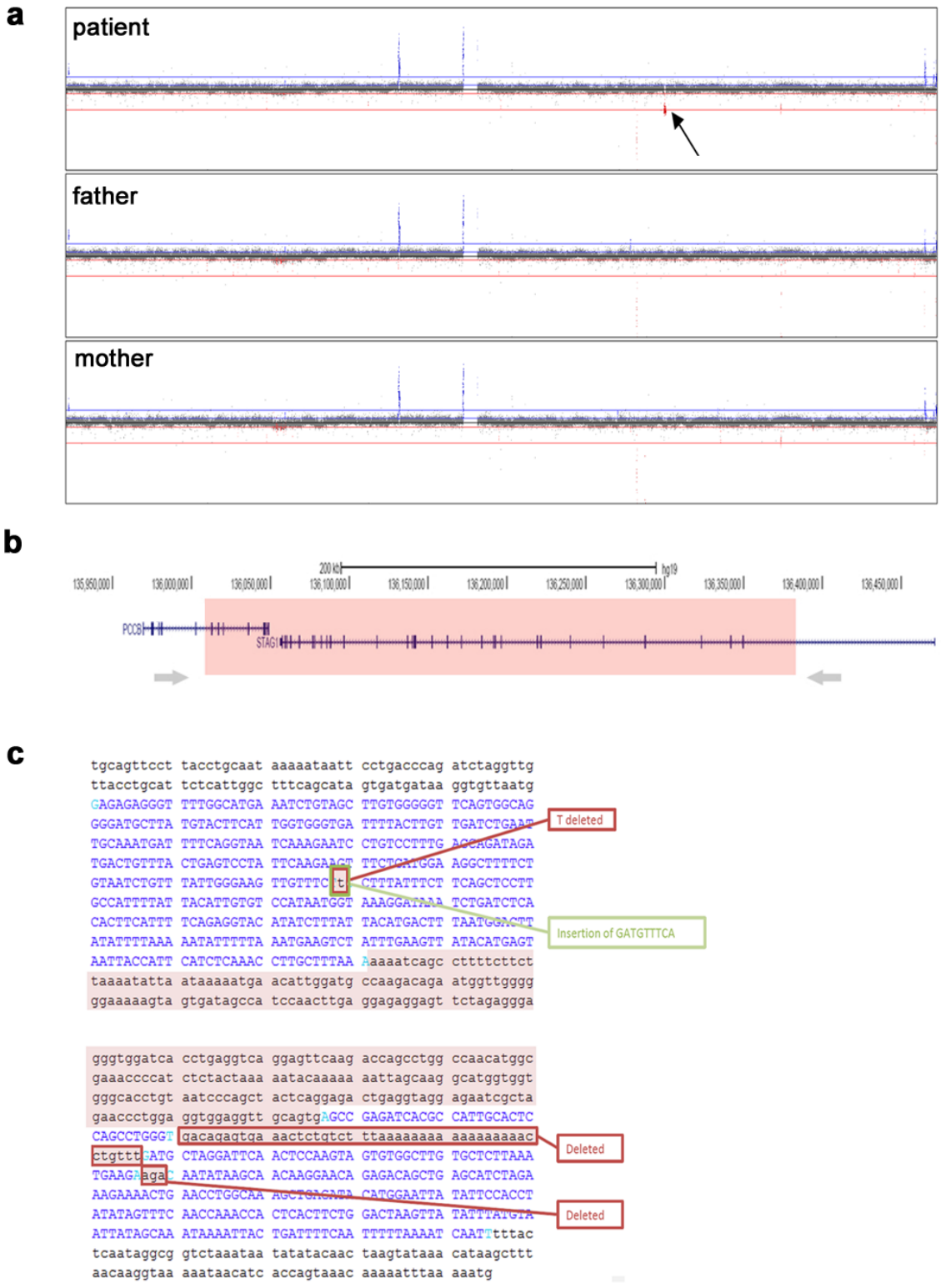
29. Carnevali, P. *et al.* Computational techniques for human genome resequencing using mated gapped reads. *J. Comput. Biol.* **19,** 279–292 (2012).
30. MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335,** 823–828 (2012).

©2014 Macmillan Publishers Limited. All rights reserved



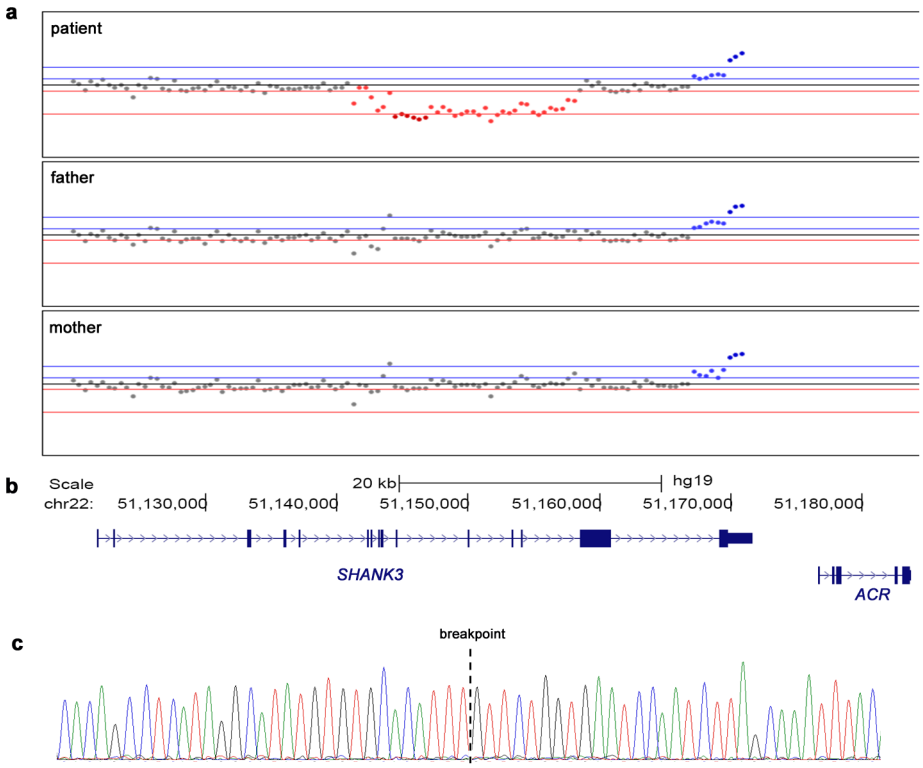
Extended Data Figure 1 | Boxplots of rare missense burden in different gene sets. Boxplots showing the difference in tolerance for rare missense variation in the general population. The vertical axis shows the distribution for each gene set of the number of rare (<1% in NHLBI Exome Sequencing Project) missense variants divided by the number of rare synonymous variants. From left to right the following gene sets are depicted: all 18,424 RefSeq genes, 170 loss-of-function tolerant genes from ref. 30, all 528 known ID genes (Supplementary Table 10), all 628 candidate ID genes (Supplementary Table 11), 9 known ID genes in which *de novo* mutations were identified in this study (Supplementary Table 8), and 10 candidate ID genes in which *de novo* mutations were identified in this study (Supplementary Table 8).

©2014 Macmillan Publishers Limited. All rights reserved



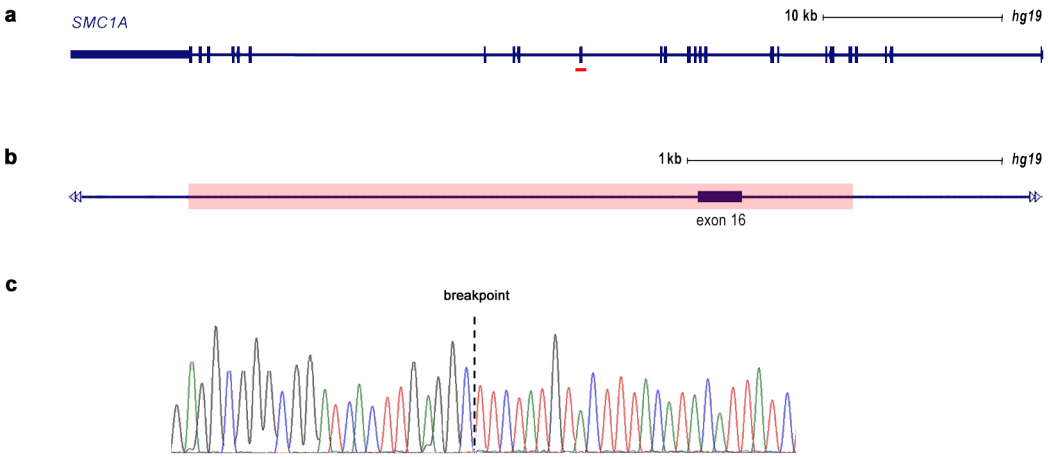
Extended Data Figure 2 | Structural variant involving *STAG1* (patient 40). **a**–**c**, CNV identified using WGS in patient 40, including the *STAG1* gene. **a**, Chromosome 3 profile (log_2 test over reference (T/R) ratios) based on read-depth information for patient, father and mother. Black arrow points towards the *de novo* event in patient 40. **b**, Genic contents of deletion. Grey arrows show primers used to amplify the junction fragment. **c**, Details on the proximal and distal breakpoints, showing the 'fragmented' sequence at both ends. Breakpoints are provided in Extended Data Table 1.

©2014 Macmillan Publishers Limited. All rights reserved



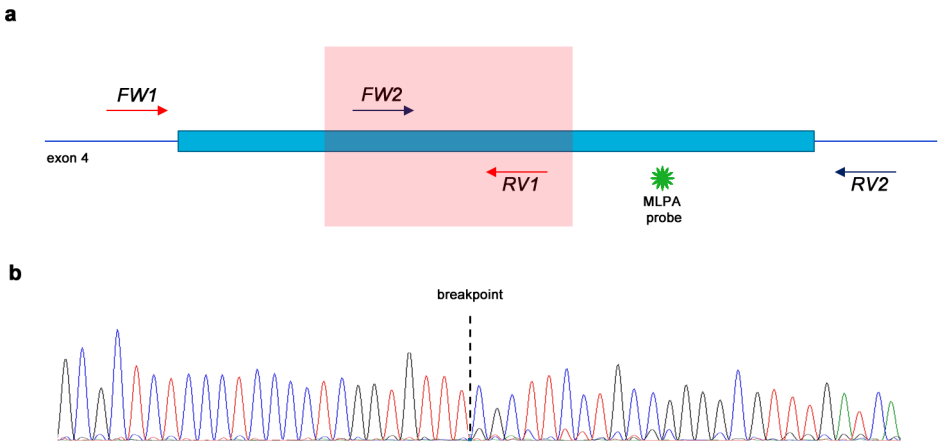
Extended Data Figure 3 | Structural variant involving *SHANK3* (patient 5). **a**–**c**, CNV identified using WGS in patient 5, including the *SHANK3* gene. **a**, Detail of chromosome 22 profile ($\log_2$ T/R ratios) based on read-depth information for patient, father and mother. Red dots in top panel show ratios indicating the *de novo* deletion in patient 5. **b**, Genic content of the deletion. **c**, Sanger validation for the junction fragment. Dotted vertical line indicates the breakpoint with sequence on the left side originating from sequence proximal to *SHANK3* and on the right side sequence that originates from sequence distal to *ACR*. Breakpoints are provided in Extended Data Table 1.

©2014 Macmillan Publishers Limited. All rights reserved



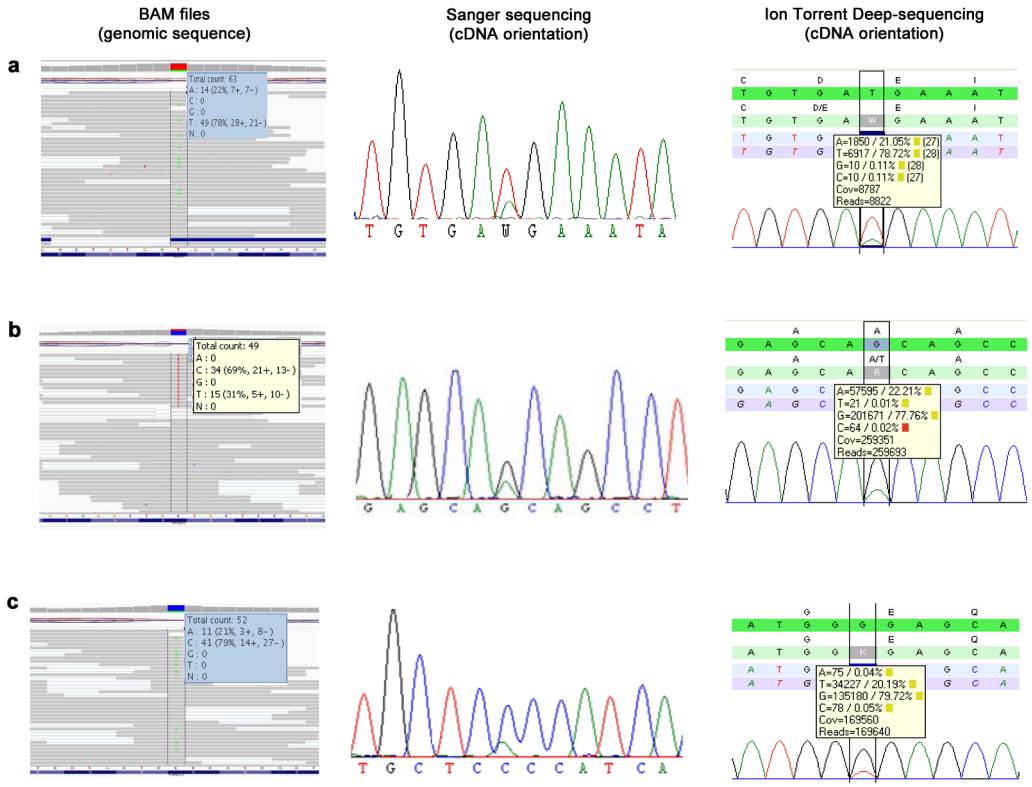
Extended Data Figure 4 | Single-exon deletion involving *SMC1A* (patient 48). **a**, Schematic depiction of the deletion identified in patient 48 involving a single exon of *SMC1A*. Pink horizontal bar highlights the exon that was deleted in the patient. **b**, Details at the genomic level of the deletion including exon 16, with Sanger sequence validation of the breakpoints. Junction is indicated by a black vertical dotted line. Breakpoints are provided in Extended Data Table 1.

©2014 Macmillan Publishers Limited. All rights reserved



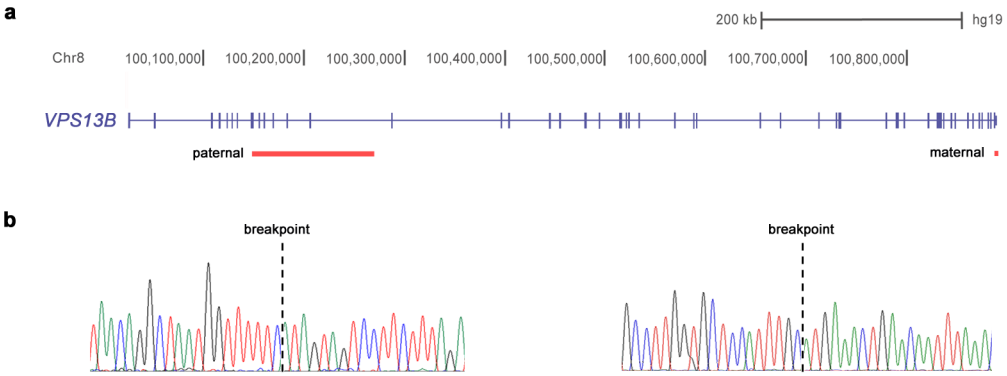
Extended Data Figure 5 | Intra-exonic deletion involving *MECP2* (patient 18). **a**, Schematic depiction of the deletion identified in patient 18, which is located within exon 4 of *MECP2*. Initial Sanger sequencing in a diagnostic setting could not validate the deletion as the primers used to amplify exon 4 removed the primer-binding sites (FW2 and RV1 respectively). Multiplex ligation probe amplification (MLPA) analysis for CNV detection showed normal results as the MLPA primer-binding sites were located just outside of the deleted region. **b**, Combining primers FW1 and RV2 amplified the junction fragment, clearly showing the deletion within exon 4. Of note, the background underneath the Sanger sequence is derived from the wild-type allele. Breakpoints are provided in Extended Data Table 1.

©2014 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 6 | Confirmation of mosaic mutations in *PIAS1*, *HIVEP2* and *KANSL2*. **a**–**c**, Approaches used to confirm the presence of mosaic mutations in *PIAS1* (**a**), *HIVEP2* (**b**) and *KANSL2* (**c**). Images and read-depth information showing the base counts in the BAM files (left) indicated that the variants/wild-type allele were not in a 50%/50% distribution. Sanger sequencing (middle) then confirmed the variant to be present in the patient, and absent in the parents (data from parents not shown), again indicating that the mutation allele is underrepresented. Guided by these two observations, amplicon-based deep sequencing using Ion Torrent subsequently confirmed the mosaic state of the mutations (right). On the basis of deep sequencing, percentages of mosaicism for *PIAS1*, *HIVEP2* and *KANSL2* were estimated at 21%, 22% and 20%, respectively.

©2014 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 7 | Compound heterozygous structural variation affecting *VPS13B* (patient 12). **a**, **b**, CNVs of *VPS13B* identified using WGS in patient 12. **a**, Schematic representation of *VPS13B*, with vertical bars indicating coding exons. In patient 12 two deletions were identified, one ~122 kb in size which was inherited from his father, and another ~2 kb in size, which was inherited from his mother and consisted only of a single exon. **b**, Both CNV junction fragments were subsequently validated using Sanger sequencing. Left, junction fragment from the paternally inherited deletion. Right, junction fragment from the maternally inherited deletion. Breakpoints are provided in Extended Data Table 1.

©2014 Macmillan Publishers Limited. All rights reserved

Extended Data Table 1 | Large variants of potential clinical relevance identified using WGS and probability of exonic CNVs occurring in affected and control individuals for these loci

Trio	Type*	Genomic characterization	Size (kb)	CN	Origin	Genes affected	Affected (n=7,743)	Controls (n=4,056)	OR	CI	P-value#
5	CNV	chr22(GRCh37):g.51121756-51187704del	66	1	*De novo*	***SHANK3***; *ACR*	41	4	5.4	1.9-15.0	0.0013
18	SV	chrX(GRCh37):g.153295929-153296514del	0.6	1	*De novo*	***MECP2***§	41	0	22.6	1.4-367.6	0.04
31	CNV complex	chr4(GRCh37):g.183693432-183756173dup	62	3	*De novo*	*TENM3*	5	0	0.6	0.2-2.4	1.0
		Insertion point: chrX(GRCh37):g.53318362-53318363	-	-		***IQSEC2***	28	0	7.6	1.0-56.1	0.092
37	CNV complex	chr3(GRCh37):g.48532000-49156000dup†	624	3	*De novo*	n=22‖	22	0	23.6	1.4-388.7	0.033
		chr3(GRCh37):g.49298000-49848000dup†	550	3		n=20‖					
		chr3(GRCh37):g.49849505-49870969del	21	1		n=2‖					
		chr3(GRCh37):g.49872000-49958000dup†	86	3		n=4‖					
40	CNV	chr3(GRCh37):g.136003159delinsGATGTTTCA	-								
		chr3(GRCh37):g.136003363-136385607del	382	1	*De novo*	***STAG1***; *PCCB*	9¶	0	10.0	0.6-171.1	0.26
		chr3(GRCh37):g.136385640-136385685del	-								
		chr3(GRCh37):g.136385737-136385739del	-								
48	SV	chrX(GRCh37):g.53424894-53427008del	2.1	1	*De novo*	***SMC1A***§	33	0	17.3	1.1-282.9	0.075
49	SV	chr1(GRCh37):g.40247181-40256104dup	9	3	*De novo*	*BMP8B*§	2	0	2.6	0.1-54.6	1.0
50	CNV	chr16(GRCh37):g.29567295-30177916‡	611	1	*De novo*	n=29‖	31	4	4.1	1.4-11.5	0.0093
12	SV	chr8(GRCh37):g.100887349-100889133del	1.7	1	*Maternal*	***VPS13B***§☆	5	0	5.8	0.3-104.2	0.80
	CNV	chr8(GRCh37):g.100147792-100270123del	122	1	*Paternal*	***VPS13B***☆	0	0			

Genes highlighted in bold are either listed as known ID genes or candidate ID genes. Please note that all patients had 250K SNP microarrays. Re-evaluation of these data showed that for all but one CNV the number of probes within the region was insufficient, either because of the small genomic size of the CNV, or due to uneven genome-wide probe spacing leaving fewer probes than required for the hidden Markov model algorithms to be identified.

* Primary method used to identify the rearrangement (see also Supplementary Methods).

† Not assessed at base-pair level due to complexity of CNV event including an inversion, duplication and deletion.

‡ Not assessed at base-pair level as the CNV event, involving a known microdeletion syndrome region, is mediated by low-copy repeats.

§ Single exon.

‖ Number of genes affected rather than individual gene names are provided due to the large number of genes.

¶ Observed 13 times in Decipher.

#Corrected for multiple testing using Benjamini–Hochberg with a false discovery rate (FDR) of 0.1.

☆*VPS13B* recessive ID gene.

©2014 **Macmillan Publishers Limited. All rights reserved**

Extended Data Table 2 | *De novo* SNVs of potential clinical relevance identified using WGS

Trio	Gene	Protein effect	Mutation type	PhyloP‡	Gene Classification§
1	*NGFR*	p.(Cys122Arg)	Missense	4.97	-
2	*GFPT2*	p.(Thr680Ser)	Missense	6.02	-
6	*WWP2*	p.(Gly10Gly)†	Synonymous	-0.12	-
7	*TBR1*	p.(Gln373Arg)	Missense	3.51	Known
9	*WDR45*	p.(Cys344Alafs*67)	Frameshift		Known
13	*SMC1A*	p.(Asn788Lysfs*10)	Frameshift		Known
15	*SPTAN1*	p.(Glu91Lys)	Missense	5.69	Known
17	*ASUN*	p.(Gln99*)	Nonsense		-
21	*ALG13*	p.(Asn107Ser)	Missense	1.34	Known‖
21	*RAI1*	p.(Gln88*)	Nonsense		Known
22	*MED13L*	p.(Asp860Gly)	Missense	4.75	Candidate
24	*BRD3*	p.(Phe334Ser)	Missense	4.48	-
25	*SATB2*	p.(Gln310delinsHisCysLysAlaThr)	Insertion		Known
26	*PPP2R5D*	p.(Trp207Arg)	Missense	5.13	Candidate
27	*KCNA1*	p.(Thr371Ile)	Missense	5.69	Known
28	*SCN2A*	p.(Gln1521*)	Nonsense		Known
30	*MAST1*	p.(Pro1177Arg)	Missense	5.28	-
34	*APPL2*	p.(Ser329*)	Nonsense		-
41	*NACC1*	p.(Arg468Cys)	Missense	3.51	-
43	*POGZ*	p.(Arg1001*)	Nonsense		Candidate
46	*TBR1*	p.Thr532Argfs*144	Frameshift		Known
49	*KANSL2*	p.(Gly151Gly)†	Synonymous	1.58	Candidate

A dash indicates genes that have not yet been implicated in ID, but fulfil the criteria for diagnostic reporting of a pathogenic variant (that is, a possible cause for ID).
† Predicted effect on splicing.
‡ PhyloP score for nonsense and frameshift mutations is not provided as this are deleterious mutations regardless of their evolutionary conservation.
§ 'Known' refers to known ID gene whereas 'Candidate' refers to a gene that is listed on the candidate ID gene list.
‖ Since the inclusion of this patient in this study, the same *de novo* mutation in *ALG13* has been described elsewhere[16]. This may suggest that this mutation, despite its low conservation and the identification of a nonsense mutation in *RAI1*, may also contribute to the disease phenotype in this patient. See also Supplementary Table 8 legend.

©2014 Macmillan Publishers Limited. All rights reserved