nature neuroscience

Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability

Stefan H Lelieveld[1,5], Margot R F Reijnders[2,5], Rolph Pfundt[2], Helger G Yntema[2], Erik-Jan Kamsteeg[2], Petra de Vries[2], Bert B A de Vries[2], Marjolein H Willemsen[2], Tjitske Kleefstra[2], Katharina Löhner[3], Maaike Vreeburg[4], Servi J C Stevens[4], Ineke van der Burgt[2], Ernie M H F Bongers[2], Alexander P A Stegmann[4], Patrick Rump[3], Tuula Rinne[2], Marcel R Nelen[2], Joris A Veltman[2,4], Lisenka E L M Vissers[2,5], Han G Brunner[2,4,5] & Christian Gilissen[2,5]

To identify candidate genes for intellectual disability, we performed a meta-analysis on 2,637 *de novo* mutations, identified from the exomes of 2,104 patient–parent trios. Statistical analyses identified 10 new candidate ID genes: *DLG4*, *PPM1D*, *RAC1*, *SMAD6*, *SON*, *SOX5*, *SYNCRIP*, *TCF20*, *TLK2* and *TRIP12*. In addition, we show that these genes are intolerant to nonsynonymous variation and that mutations in these genes are associated with specific clinical ID phenotypes.

Intellectual disability (ID) and other neurodevelopmental disorders are in part due to *de novo* mutations affecting protein-coding genes[1–4]. Large scale exome sequencing studies of patient–parent trios have efficiently identified genes enriched for *de novo* mutations in cohorts of individuals with ID compared to controls[2] or on the basis of expected gene-specific mutation rates[5].

Here we sequenced the exomes of 820 patients with ID and their parents as part of routine genetic testing at the Radboud University Medical Center (RUMC) in the Netherlands. We identified 1,083 *de novo* mutations (DNMs) in the coding and canonical splice site regions affecting 915 genes (**Supplementary Tables 1** and **2** and **Supplementary Figs. 1–4**). In our cohort we detected an increased number of loss-of-function (LoF) mutations compared to controls (Fisher's exact test, $P = 9.38 \times 10^{-12}$; Online Methods) and enrichment for recurrent gene mutations (observed versus expected, $P < 1 \times 10^{-5}$; **Supplementary Fig. 5**).

Using an established framework of gene specific mutation rates[6], we calculated for each gene the probability of identifying the observed number of LoF or functional DNMs in our cohort (Online Methods). To validate this approach we first performed the analysis on the complete set of 820 ID patients. After Benjamini–Hochberg correction for multiple testing, 18 well-known ID-associated genes were significantly enriched for DNMs (**Supplementary Tables 3** and **4**). To optimize our analysis for the identification of new candidate genes in the RUMC cohort, we excluded all individuals with mutations in any of the known ID genes (Online Methods and **Supplementary Fig. 6**). Repeating the analysis for mutation enrichment, we identified four genes (*DLG4*, *PPM1D*, *SOX5* and *TCF20*) that were not, to our knowledge, previously associated with ID and that were significantly enriched for DNMs in our cohort (**Fig. 1**, **Table 1** and **Supplementary Table 5**). To achieve the best possible power for the identification of candidate ID genes, we next added data from four previously published family-based sequencing studies (**Supplementary Table 1**). The combined cohort included 2,104 patient–parent trios and 2,637 DNMs across 1,990 genes. After again excluding individuals with mutations in known ID genes, this cohort consisted of 1,471 individuals with 1,400 DNMs in 1,235 genes (Online Methods and **Supplementary Fig. 6**). Meta-analysis on this combined cohort identified ten candidate ID genes with more LoF DNMs or more functional DNMs than expected *a priori*. These ten genes included the four candidate ID genes previously identified in the RUMC cohort, as well as *RAC1*, *SMAD6*, *SON*, *TLK2*, *TRIP12* and *SYNCRIP* (**Fig. 1**, **Table 1** and **Supplementary Table 6**).

To further evaluate the identification of the ten candidate ID genes, we compared the phenotypes of the 18 RUMC individuals with DNMs in these genes. We observed strong phenotypic overlap for some of these genes (**Fig. 2**, **Supplementary Table 7** and **Supplementary Note**).

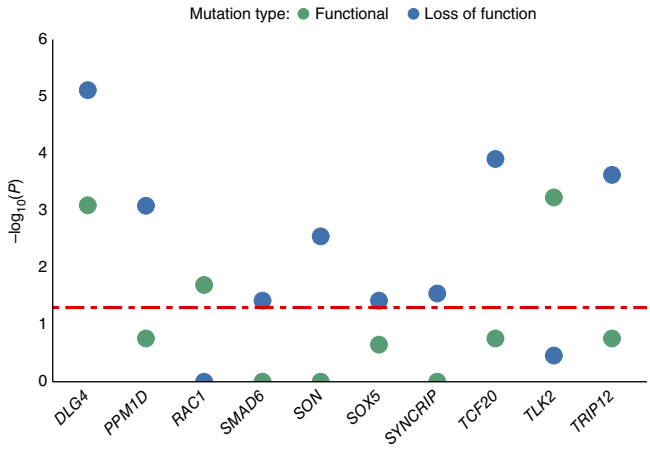

Figure 1 Genes enriched for LoF and functional DNMs in a cohort of 2,104 ID trios from multiple studies. The *y* axis shows the $-\log_{10}$-transformed, corrected *P*-value of the DNM enrichment as listed in **Table 1**. Corrected *P*-values based on LoF mutations are blue and corrected *P*-values based on functional mutations are green. Only genes with corrected *P*-values (LoF, functional, or both) less than the significance threshold (red dotted line, 0.05) are shown.

[1]Department of Human Genetics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, the Netherlands. [2]Department of Human Genetics, Donders Centre for Neuroscience, Radboud University Medical Center, Nijmegen, the Netherlands. [3]Department of Genetics, University Medical Center Groningen, Groningen, the Netherlands. [4]Department of Clinical Genetics, Maastricht University Medical Centre, Maastricht, the Netherlands. [5]These authors contributed equally to this work. Correspondence should be addressed to C.G. (christian.gilissen@radboudumc.nl).

Received 22 February; accepted 1 July; published online 1 August 2016; doi:10.1038/nn.4352

Table 1 Candidate ID genes

Gene		RUMC cohort		ID cohort		Gene description
		LoF	Functional	LoF	Functional	
DLG4	*q*	**1.13×10^{-4}**	0.086	**7.69×10^{-6}**	**8.02×10^{-4}**	Required for synaptic plasticity associated with NMDA receptor signaling. Depletion of DLG4 changes the ratio of excitatory to inhibitory synapses in hippocampal neurons.
NM_001365.4	*p*	6.56×10^{-9}	7.50×10^{-6}	2.24×10^{-10}	4.66×10^{-8}	
	c	*n* = 3	*n* = 3	*n* = 4	*n* = 5	
PPM1D	*q*	**0.047**	0.764	**8.22×10^{-4}**	0.174	Serine/threonine phosphatase that mediates a feedback regulation of p38–p53 signaling, thereby contributing to growth inhibition and suppression of stress-induced apoptosis.
NM_003620.3	*p*	5.45×10^{-6}	2.56×10^{-4}	9.57×10^{-8}	3.03×10^{-5}	
	c	*n* = 2	*n* = 2	*n* = 3	*n* = 3	
RAC1	*q*	n.d.	0.217	n.d.	**0.020**	Plasma-membrane-associated small GTPase involved in many cellular processes. In the synapses, it mediates the regulation of F-actin cluster formation by SHANK3.
NM_018890.3	*p*		3.80×10^{-5}		1.75×10^{-6}	
	c		*n* = 2		*n* = 3	
SMAD6	*q*	n.d.	n.d.	**0.037**	1	Mediates TGF-β activity and BMP–SMAD1 signaling. Functions as a transcriptional co-repressor.
NM_005585.4	*p*			8.29×10^{-6}	7.50×10^{-4}	
	c			*n* = 2	*n* = 2	
SON	*q*	1	1	**0.003**	1	Component of the spliceosome with pleiotropic roles during mitotic progression. Functions in efficient cotranscriptional RNA processing.
NM_138927.2	*p*	0.086	0.005	4.12×10^{-7}	1.67×10^{-3}	
	c	*n* = 1	*n* = 1	*n* = 3	*n* = 3	
SOX5	*q*	**0.016**	1	**0.038**	0.216	One of the transcription factors regulating embryonic development. Plays a critical role in neuronal progenitor development by regulating the timing of differentiation.
NM_006940.4	*p*	1.39×10^{-6}	3.98×10^{-4}	8.79×10^{-6}	5.83×10^{-5}	
	c	*n* = 2	*n* = 2	*n* = 2	*n* = 3	
SYNCRIP	*q*	1	1	**0.028**	1	Heterogeneous nuclear ribonucleoprotein (hnRNP) functioning in the CRD-mediated mRNA stabilization complex and SMN complex, and in the APOB RNA editing complex.
NM_006372.4	*p*	0.001	0.019	4.94×10^{-6}	1.24×10^{-3}	
	c	*n* = 1	*n* = 1	*n* = 2	*n* = 2	
TCF20	*q*	**6.22×10^{-6}**	**0.035**	**1.24×10^{-4}**	0.174	Transcriptional activator of matrix metalloproteinase 3 and (co)activator of various other transcriptional activators.
NM_005650.3	*p*	1.81×10^{-10}	1.00×10^{-6}	7.21×10^{-9}	3.71×10^{-5}	
	c	*n* = 4	*n* = 4	*n* = 4	*n* = 4	
TLK2	*q*	0.100	1	0.347	**5.86×10^{-4}**	Serine/threonine kinase regulating chromatin assembly. Involved in DNA replication, transcription and repair and in chromosome segregation.
NM_001284333.1	*p*	1.44×10^{-5}	4.20×10^{-4}	9.09×10^{-5}	1.70×10^{-8}	
	c	*n* = 2	*n* = 2	*n* = 2	*n* = 5	
TRIP12	*q*	0.273	1	**2.35×10^{-4}**	0.174	E3 ubiquitin ligase involved in the ubiquitin fusion degradation pathway. Guards against excessive spreading of ubiquitinated chromatin at damaged chromosomes in DNA repair.
NM_001284214.1	*p*	5.55×10^{-5}	0.003	2.05×10^{-8}	4.05×10^{-5}	
	c	*n* = 2	*n* = 2	*n* = 4	*n* = 4	

All genes listed reached statistical significance after Benjamini–Hochberg correction for enrichment of functional and/or LoF DNM in the RUMC or ID cohort. For each gene the Benjamini–Hochberg corrected *P*-value (*q*), uncorrected *P*-value (*p*) and raw counts (*c*) are shown. Significant Benjamini–Hochberg-corrected *P*-values are depicted in bold. n.d. (not defined) indicates genes without observed DNMs in the RUMC or ID cohort.

Additional genes that were close to statistical significance, such as *SETD2*, show phenotypic similarities suggestive of a shared genetic cause consistent with previous case reports[7,8] (**Supplementary Fig. 7** and **Supplementary Note**).

Studies have shown that genes involved in genetic disorders exhibit strongly reduced tolerance to nonsynonymous genetic variation compared to non-disease-associated genes. This is particularly evident for ID[3]. We found that a large set of well-known dominant ID genes (n = 444), along with the ten candidate ID genes, are highly intolerant of LoF variation[9] (median probability of being LoF-intolerant (pLI) of 0.95, $P < 1 \times 10^{-5}$ and median pLI of 0.99, $P < 1 \times 10^{-5}$, respectively; Online Methods, **Supplementary Fig. 8** and **Supplementary Table 8**). We noted that those ID genes that harbor only missense variants ('missense-only' genes) are among the most intolerant ID genes (median pLI of 0.99, $P < 1 \times 10^{-5}$; **Supplementary Fig. 8**). Additionally, we found that mutations in missense-only genes are more likely to cluster than mutations in genes for which we also identified LoF mutations (P = 0.01, Fisher's exact test; Online Methods and **Supplementary Table 9**).

There is considerable overlap of genes and molecular pathways involved in neurodevelopmental disorders (NDDs), such as autism spectrum disorder, schizophrenia, epileptic encephalopathy and ID[10]. Therefore, we performed a third analysis including 12 published family-based sequencing studies of various NDDs (**Supplementary Table 1** and **Supplementary Fig. 6**). Repeating our analysis in this NDD cohort, we identified seven genes significantly enriched for either LoF or functional DNMs (**Supplementary Fig. 9** and **Supplementary Table 10**). In line with our hypothesis, five of these identified genes were also identified in our previous analyses with individuals with ID only, whereas two genes (*SLC6A1* and *TCF7L2*) only reached significance in the NDD meta-analysis as a result of additional mutations in patients with phenotypes other than ID (**Supplementary Table 11**). Specifically, for two of the five candidate ID genes (*TLK2* and *TRIP12*) additional DNMs were identified in individuals with autism spectrum disorder and schizophrenia, suggesting that DNMs in these genes may lead to a broader phenotype than ID alone. For *TRIP12*, a similarly broad phenotype has been reported previously[4].

In summary, we identified ten candidate ID genes via a meta-analysis of whole exome sequencing data on 2,104 ID trios. The

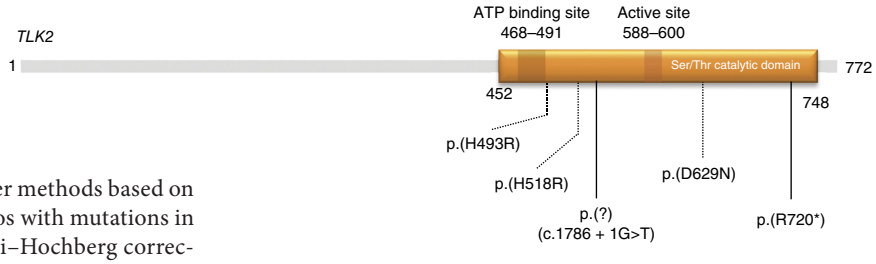

Figure 2 TLK2 protein (Q86UE8) with DNMs localized to the serine/threonine catalytic domain. Two individuals in the RUMC cohort were found to have a DNM in *TLK2*; they showed overlapping clinical features including facial dysmorphisms (**Supplementary Note**).

statistical framework used here differs from other methods based on gene-specific mutation rates by removing all trios with mutations in known disease genes and by applying Benjamini–Hochberg correction for multiple testing. Our study underscores the impact of DNMs on a continuum of neurodevelopmental phenotypes that impinge on a broad range of processes, including chromatin modifiers (*TRIP12* and *TLK2*), Fragile X Mental Retardation Protein (FMRP) target and synaptic plasticity genes (*DLG4*; **Supplementary Fig. 10**) and embryonically expressed genes (*PPM1D* and *RAC1*)[2]. Data from a similar systematic study of DNMs in neurodevelopmental disorders suggest that many, and possibly most, genes whose DNM causes severe developmental disorders are now known[11]. Yet only *TCF20* and *PPM1D* are shared between the 10 candidate genes in our study and the 14 genes identified by McRae *et al.*[11]. Thus, a large number of rare dominant developmental disorder genes may remain to be identified.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

The authors would like to thank the Exome Aggregation Consortium and the groups that provided exome variant data for comparison. A full list of contributing groups can be found at http://exac.broadinstitute.org/about. We thank all clinicians involved for referring individuals with ID for diagnostic exome sequencing. We thank J. Goeman for statistical advice and M. Hurles for discussions. We would also like to thank the participating individuals and their families. 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.; 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

C.G., L.E.L.M.V. and H.G.B. designed the study; S.H.L., M.R.F.R., C.G. and L.E.L.M.V. performed the analysis. R.P., H.G.Y., E.-J.K., T.R., S.J.C.S., A.P.A.S. and M.R.N. signed out initial diagnostic reports. P.d.V. performed Sanger validations. B.B.A.d.V., M.H.W., T.K., K.L., M.V., I.v.d.B., E.M.H.F.B., P.R. and M.R.F.R. collected patient phenotypes. S.H.L., M.R.F.R., J.A.V., H.G.B., L.E.L.M.V. and C.G. drafted the manuscript; all authors contributed to the final version of the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Gulsuner, S. *et al. Cell* **154**, 518–529 (2013).
2. Iossifov, I. *et al. Nature* **515**, 216–221 (2014).
3. Gilissen, C. *et al. Nature* **511**, 344–347 (2014).
4. O'Roak, B.J. *et al. Nat. Commun.* **5**, 5595 (2014).
5. Deciphering Developmental Disorders Study. *Nature* **519**, 223–228 (2015).
6. Samocha, K.E. *et al. Nat. Genet.* **46**, 944–950 (2014).
7. Luscan, A. *et al. J. Med. Genet.* **51**, 512–517 (2014).
8. Lumish, H.S., Wynn, J., Devinsky, O. & Chung, W.K. *J. Autism Dev. Disord.* **45**, 3764–3770 (2015).
9. Lek, M. *et al.* Preprint at *bioRxiv* http://dx.doi.org/10.1101/030338 (2016).
10. Krumm, N., O'Roak, B.J., Shendure, J. & Eichler, E.E. *Trends Neurosci.* **37**, 95–105 (2014).
11. McRae, J.F. *et al.* Preprint at *bioRxiv* http://dx.doi.org/10.1101/049056 (2016).

ONLINE METHODS

Recruitment of individuals with ID. The Department of Human Genetics from the Radboud University Medical Center (RUMC) is a tertiary referral center for clinical genetics. Approximately 350 individuals with unexplained intellectual disability (ID) are referred annually to our clinic for diagnostic evaluation. Since September 2011 whole exome sequencing (WES) has been part of the routine diagnostic work-up aimed at the identification of the genetic causes underlying disease[12]. For individuals with unexplained ID, a family-based WES approach is used which allows the identification of DNMs as well as variants segregating according to other types of inheritance, including recessive mutations and maternally inherited X-linked recessive mutations in males[13]. For this study, we selected all individuals with ID who had family-based WES using the Agilent SureSelect v4 enrichment kit combined with sequencing on the Illumina HiSeq platform in the time period 2013–2015. This selection yielded a set of 820 individuals, including 359 females and 461 males. The level of ID ranged between mild (IQ 50–70) and severe–profound (IQ <30).

Families gave informed consent both for the diagnostic procedure and for forthcoming research that could result in the identification of new genes underlying ID by meta-analysis, as presented here. Explicit consent for photo-publication was sought and granted by a subset of families.

Diagnostic whole exome sequencing. The exomes of 820 patient–parent trios were sequenced, using DNA isolated from blood, at the Beijing Genomics Institute (BGI) in Copenhagen. Exome capture was performed using Agilent SureSelect v4 and samples were sequenced on an Illumina HiSeq instrument with 101-bp paired-end reads to a median coverage of 75×. Sequence reads were aligned to the hg19 reference genome using BWA version 0.5.9-r16. Variants were subsequently called by the GATK unified genotyper (version 3.2-2) and annotated using a custom diagnostic annotation pipeline. Base-pair resolution coverage of the regions enriched by the SureSelect V4 kit were computed by BEDTools based on the regions as provided by the manufacturer. An average of 98.9% of Agilent SureSelect V4 enriched targets was covered by 10 or more reads for the RUMC cohort of 820 ID patients (**Supplementary Fig. 1**).

Identification of DNMs in 820 individuals with ID. The diagnostic WES process as outlined above only reports (*de novo*) variants that can be linked to the individuals' phenotypes. In this study, we systematically collected all DNMs located in the coding sequence (RefSeq) and/or affected canonical splice sites (canonical dinucleotides GT and AG for donor and acceptor sites; **Supplementary Fig. 4**), as identified in the 820 individuals with ID irrespective of their link to disease, to evaluate the potential relevance of genes for ID in an unbiased fashion using a statistical framework. DNMs were called as described previously[13]. Briefly, variants called within parental samples were removed from the variants called in the child. For the remaining variants, pileups were generated from the alignments of the child and both parents. Based on pileup results, variants were then classified into the following categories: 'maternal' (identified in the mother only), 'paternal' (identified in the father only), 'low coverage' (insufficient read depth in either parent), 'shared' (identified in both parents) and 'possibly *de novo*' (absent in the parents). Variants classified as possibly *de novo* were included in this study.

We applied various quality measures to ensure that only the most reliable variant calls were included in the study: (i) all samples had fewer than 25 possibly *de novo* calls; (ii) each variant had at least 10× coverage in either parent (for example, high prior probability of being inherited); (iii) the location was not in dbSNP version137 (for example, a possible highly mutable genomic location) and (iv) each variant was called in a maximum of 5 samples in our in-house variant database (which eliminated variants that occur too frequently to be disease-causing given the incidence of ID in combination with the sample size of our in-house database); (v) each variant showed a variant read percentage >30%, or alternatively, >20% with >10 individual variant reads; and (vi) each variant had a GATK quality score of >400. For *de novo* variants called within a 5-bp window of each other within the same individual, variant calls were manually curated and merged into a single call (when occurring on the same allele). This set of criteria resulted in the identification of 1,083 potential DNMs in 820 individuals with ID.

Validation and categorization of DNMs. In a separate (unpublished) in-house study, we recently determined the predictive value for GATK quality scores in terms of the variant being validated by Sanger sequencing. A set of 840 variants called by the same version of GATK was retrospectively analyzed for the quality scores and validation statuses of each variant in the set. Based on this assessment, we determined that a GATK quality score ≥500 resulted in 100% of variants being validated by Sanger sequencing (data not shown). In addition to our in-house study, two other studies also found 100% Sanger validation rates for variants with GATK quality scores of ≥500 (ref. 14). Based on these results, we considered all variants with a GATK *Q*-score of ≥ 500 (n = 1,039) to be true DNMs. Nonetheless, a random set of 141 potential DNMs with GATK *Q*-scores of ≥500 were all confirmed by Sanger sequencing. All potential *de novo* variants with GATK *Q*-scores between 400 and 500 (n = 40) were subsequently validated by Sanger sequencing, and all were confirmed. All 20 DNMs of the reported candidate genes were confirmed by Sanger sequencing (**Supplementary Table 2** and **Supplementary Figs. 2–4**).

For further downstream statistical analysis (see below), DNMs were categorized by mutation type: (i) LoF DNM (n = 211), including nonsense (n = 77), frameshift (n = 97), canonical splice site (n = 27), start loss (n = 2), stop loss (n = 1) and premature stop codon resulting from an indel (n = 7); and (ii) functional DNM (n = 872), including all LoF mutations (n = 211), in-frame insertion/deletion events (n = 23) and all missense mutations (n = 638) (**Supplementary Fig. 4**). For variants within the same individual and within the same gene but more than 5 bp apart, the variant with the most severe functional effect was considered for the per-gene statistics (see below).

Evaluating the number of recurrently LoF and functional *de novo* mutated genes. We simulated the expected number of recurrently mutated genes by redistributing the observed number of mutations at random over all genes based on their specific LoF and functional mutation rates (see "Statistical enrichment of DNMs" below) as described by Samocha *et al*[6]. Based on results from 100,000 simulations, we calculated how many times the number of recurrently mutated genes was the same as or exceeded the observed number of recurrently mutated genes in the RUMC data set. We performed these simulations separately for LoF and functional DNMs (**Supplementary Fig. 5**). *P*-values were then calculated by taking the number of times the number of recurrently mutated genes exceeded the observed number of recurrently mutated genes and dividing by the number of simulations. In addition, *z*-values were computed by subtracting the mean value of the simulations from the observed value and dividing by the s.d. of the simulations.

Genes previously implicated in ID etiology. To evaluate whether the genes identified by our meta-analyses had been previously implicated in ID, we used two publicly available repositories of genes known to be involved in ID. First, we used our list of 707 genes, routinely used by our diagnostic setting to interpret WES results of individuals with ID[15]. Second, we downloaded a list of 1,424 genes associated with developmental disorders from the DDG2P database (http://www.ebi.ac.uk/gene2phenotype/gene2phenotype-webcode/cgi-bin/handler.cgi#); the list was compiled and curated by clinicians as part of the Deciphering Developmental Disorders (DDD) study to facilitate clinical feedback of likely causal variants[5]. In total the two lists compromised 1,537 unique genes. In this manuscript, the list of unique gene entries is referred to as "known ID genes" (**Supplementary Table 4**).

Statistical enrichment of DNMs. In our meta-analysis for ID and neurodevelopmental disorders we only included studies with minimum of 50 trios. For each gene, and each of the functional classes (LoF and functional), we used the corresponding gene-specific mutation rate (GSMR) as published by Samocha *et al.*[6] to calculate the probability of the number of identified DNMs in our cohort. For genes for which no GSMR was reported, we used the maximum GSMR of all reported genes (i.e., the GSMR of the gene *TTN*). We then calculated specific mutation rates for the two defined functional classes (LoF, functional). The GSMR for LoF DNMs was calculated by summing the individual GSMR for nonsense, splice site and frameshift variants; the GSMR for functional DNMs was calculated by summing the GSMR for the LoF variants with the missense mutation rate; and for genes for which variants from different functional classes were identified, we used the overall GSMR. For the stop-loss and start-loss mutations we used the LoF-rate and for in-frame indels, the functional rate. Null hypothesis testing was done using a one-sided exact Poisson test based on a sample size of

doi:10.1038/nn.4352

820 individuals with ID, representing 1,640 alleles for autosomal genes and 1,179 alleles for genes on the X chromosome (461 males).

For DNMs on chromosome X the correct mutation rate depends on the patient's gender as the mutation rate for fathers is higher than for mothers. Estimates show a 4:1 ratio of paternal to maternal DNMs[16]. Hence, male offspring, receiving their chromosome X exclusively from the mother, have therefore a lower mutation rate on chromosome X than estimated by the GSMR. This correction could, however, only be performed for the RUMC cohort, as gender information was not available for all studies included in the ID cohort. Notably, not correcting for this bias in male individuals for DNM in genes on the X chromosome will lead to less significant *P*-values for genes on the X chromosome, thereby potentially underestimating the significance of candidate ID genes located on the X chromosome. When a patient was found to have two DNMs in the same gene we ignored one of the two DNMs for the statistical enrichment analysis to avoid false positive results. In such cases the severity of the DNM protein effect was a factor in the choice of which DNM to ignore. For example, if a patient had one missense and one nonsense DNM in the same gene, the missense mutation was ignored in the statistical analysis.

Gene specific *P*-values were corrected for multiple testing based on the 18,730 genes present in the Agilent V4 exome enrichment kit times the number of tests (2), using the Benjamini–Hochberg procedure with an FDR of 0.05. In our cohort of 820 individuals with ID, conclusive diagnoses were already made based on DNMs in genes previously implicated in disease. The use of a multiple testing correction with a FDR of 0.05, in combination with a potential large number of DNMs in known ID genes, may artificially increase the significance of other genes because of an increasingly lenient correction for the least significant genes[17]. To verify that the identification of candidate ID genes was not inflated by this effect, we performed the analysis after removing all individuals with a DNM in one of the known genes (potential other DNMs in such individuals were also removed for further analysis). Incidentally, this also increased our statistical power. The mode of inheritance was not taken into account when removing individuals with a DNM in a known gene (for example, samples with a DNM in a recessive gene were excluded). This correction left 584/820 individuals with ID in the RUMC cohort, with 627 DNMs across 584 genes. Similarly, for the ID and neurodevelopmental cohort, we removed all individuals with a DNM in a known ID gene (and other DNMs in these individuals). For the ID cohort, 1,471 samples remained with 1,400 DNMs in 1,235 genes. For the neurodevelopmental cohort, 4,944 samples remained with 4,387 DNM across 3,402 genes (for a complete overview see **Supplementary Fig. 6**). We corrected for testing 34,386 genes (i.e., all 18,730 genes minus the 1,537 known ID genes multiplied by 2 for testing the LoF and functional categories).

Validation of the statistical approach by analysis of DNMs in a control cohort. To further confirm the validity of our statistical approach, we applied the same analyses to a set of DNMs identified in trios of healthy individuals and unaffected siblings. For this purpose, we downloaded and reannotated all DNMs identified in 1,911 unaffected siblings of individuals with ASD from Iossifov *et al.*[2] together with DNMs in controls (**Supplementary Table 12**). In total the control set contained 2,019 coding DNMs found in 2,299 trios. Notably, the protein coding DNM rate in the control cohort was markedly lower than that observed in the individuals with ID (0.91 DNMs versus 1.32 DNMs, respectively). Additionally, we observed no significant enrichment of recurrently mutated genes for LoF or functional mutations ($P = 0.60$ and $P = 0.12$, respectively; **Supplementary Fig. 11**).

For the control cohort we performed statistical analyses as described above and identified only one gene that was significantly enriched in functional DNMs. For *YIF1A* (FDR corrected $P = 0.01$) we identified a total of 3 missense DNMs and 1 frameshift DNM (**Supplementary Tables 13** and **14**). *YIF1A* may be involved in transport between the endoplasmic reticulum and the Golgi, and it has a pLI of 2.08×10^{-8} indicating this gene is a LoF-tolerant gene. We note that the control cohort consists mostly of healthy siblings from individuals with ASD, and, as such, may still have minor enrichments for mutations that lead to susceptibility to neurodevelopmental disorders.

Increased number of LoF mutations in RUMC cohort compared to controls. To reduce the impact of the enrichment kit used in the control set studies and RUMC cohort, we computed the intersection of all enrichment kits (Agilent SureSelect 37Mb ∩ Agilent SureSelect 50Mb ∩ Agilent SureSelect V4 ∩ Nimblegen SeqCap V2; **Supplementary Table 12**) using the "intersect" function of BEDTools. Only the LoF DNMs present in the 28,189,737-Mb intersection of the four enrichment kits were used in the analysis. The Fisher's exact test on the enrichment kits normalized LoF DNMs yielded a significant difference with $P = 9.38 \times 10^{-12}$ (RUMC: 157 LoF DNMs of a total of 805 DNMs; controls: 137 LoF DNMs of a total of 1,485 DNMs; OR = 2.38; CI: 1.85–3.07). The coverage and other relevant technical information of the control studies are listed in **Supplementary Table 12**. We note it is important to consider the coverage and false negative rates of all sequencing studies. So far, only a single study has attempted to provide a false negative rate (for example, mutations that are there but were not identified) for exome sequencing, and this was predicted to be <5% (ref. 2).

Attributing pLI for all protein coding genes. To determine the intolerance to LoF variation for each gene, we used the probability of LoF intolerance (pLI), which is based on data from the Exome Aggregation Consortium (ExAC) version 0.3.1, providing exome variants from 60,706 unrelated individuals[9]. The pLI is based on the expected versus observed variant counts to determine the probability that a gene is intolerant to LoF variants and is computed for a total of 18,226 genes. The closer a pLI is to 1, the more intolerant a gene is to LoF variants. The authors consider a pLI ≥ 0.9 as an extremely LoF-intolerant set of genes. The pLIs for the genes used in this study can be found in **Supplementary Table 8**. Intolerance to LoF variation was evaluated for the available pLIs of four gene sets: (1) 170 LoF-tolerant (LoFT) genes[18], (2) 404 housekeeping genes, involved in crucial roles in cell maintenance[19], (3) 1,359 genes with functional DNMs from the healthy control data set (**Supplementary Table 14**), and (4) 444 well-known dominant ID genes (**Supplementary Table 4**).

Gene set based evaluation of pLI. We evaluated the pLI by computing the expected median pLI for each gene set based on randomly drawing *n* pLI values from the complete set of 18,226 pLI annotated genes and calculating the median (where *n* is the number of genes in the gene set). By repeating this random sampling process 100,000 times, we computed the likelihood of the observed median pLI of a gene set compared to the expected median pLI by calculating the empirical *P*-value:

$$\text{empirical } P = \frac{\left(\sum_{i=1}^{N} m_i > m_{\text{observed}}\right) + 1}{N + 1}$$

where m is the median pLI of one simulation, m_{observed} is the observed median pLI and N is the total number of performed simulations ($N = 100{,}000$). In addition, the *z*-values were computed as described in the section "Evaluating the number of recurrently LoF and functional *de novo* mutated genes."

Based on the simulations we identified a significantly lower median pLI for the LoFT genes, which is in line with the LoFT nature of this gene set (observed 9.33×10^{-9} distribution simulations: $\mu = 0.04$, $\sigma = 0.03$; empirical $P < 1 \times 10^{-5}$; $z = 1.25$). For the healthy control set, the observed median pLI matched the simulated distribution of median pLI (observed 0.03; distribution simulations: $\mu = 0.03$, $\sigma = 0.01$; empirical $P = 0.31$; $z = 0.39$). For the housekeeping and dominant ID gene sets, the observed median pLI was significantly higher than the simulated distribution of median pLI (HK genes: observed: 0.87; simulated distribution: $\mu = 0.03$, $\sigma = 0.02$; empirical $P < 1 \times 10^{-5}$; $z = 54.05$ and dominant ID genes: observed: 0.95; simulated distribution: $\mu = 0.03$, $\sigma = 0.01$; empirical $P < 1 \times 10^{-5}$; $z = 61.54$). In addition, the median pLI of the housekeeping gene approximated (median pLI = 0.87) and the dominant ID gene set (median pLI = 0.95) surpassed the extremely LoF-intolerant threshold of 0.9, which is in line with the LoF-intolerant nature of housekeeping and dominant ID genes (**Supplementary Fig. 12**).

The set of ten novel candidate ID genes has a median pLI of 0.99 (observed 0.99; simulated distribution: $\mu = 0.14$, $\sigma = 0.20$; empirical $P < 1 \times 10^{-5}$; $z = 4.28$) which is, as observed for the dominant ID genes, above the extremely LoF-intolerant gene threshold of 0.9 (**Supplementary Fig. 8**). For the missense-only genes (with at least three missense mutations in the absence of LoF mutations, all of which were known dominant ID genes), we observe the highest median pLI of

0.9999 (observed: 0.9999; simulated distribution: $\mu = 0.09$, $\sigma = 0.14$; empirical $P < 1 \times 10^{-5}$; $z = 6.70$), illustrating that those known and candidate dominant ID genes that harbor only missense variants are among the most LoF-intolerant ID genes (**Supplementary Fig. 8**).

Attributing residual variation intolerance scores (RVIS) to all genes. In addition, the residual variation intolerance score (RVIS) was assessed in the same fashion as described for the pLI. The RVIS ranks genes based on whether they have more or less common functional genetic variations relative to the genome-wide expectation. The initial RVIS gene scores were computed based on the NHLBI-ESP6500 data set[20] and recently recomputed based on the ExAC v0.3 data set (http://genic-intolerance.org/). The genes from our study were annotated with RVIS scores based on ExAC (**Supplementary Table 8**).

RVIS scores for gene sets were compared in the same way as for the pLI (**Supplementary Fig. 13**). Again, we found the new candidate ID genes to be significantly more intolerant than any random set of genes found (observed median RVIS: 8.47, distribution simulations: $\mu = 50.05$, $\sigma = 15.08$; empirical $P = 4.60 \times 10^{-4}$; z: −2.76), like the known dominant ID genes (**Supplementary Fig. 13**). For the dominant missense-only genes we again observed the lowest median RVIS of 3.56 (distribution simulations: $\mu = 50.02$, $\sigma = 10.42$; empirical $P < 1 \times 10^{-5}$; z: −4.46; **Supplementary Fig. 13**).

Estimating clustering of DNMs. The spatial distributions of missense, frameshift and nonsense DNMs were analyzed for clustering within the respective gene they occurred in based on 100,000 simulations. The locations of observed DNMs were randomly sampled over the coding exons of the gene and the distances (in base pairs) between the mutations were normalized for the total coding size of the respective gene. The geometric mean (the nth root of the product of n numbers) of all mutation distances between the DNMs was taken as a measure of clustering. A pseudocount (adding 1 to all distances and 1 to the gene size) was applied to avoid a mean distance of 0 when there were identical mutations.

Based on the prior distance distribution of the 100,000 simulations, we computed a gene-based empirical probability of the observed distance for dominant ID genes with 3 or more DNMs ($n = 64$ genes) in the ID set of 2,104 trios. A total of 21 genes contained only missense mutations ("missense-only" group) and 43 genes contained frameshift, nonsense or a combination of frameshift, nonsense and missense DNMs ("LoF + functional" group). In 21 genes of the missense-only group, 5 genes had empirical probabilities below the significant threshold of 0.05/64, whereas only 1 of the 43 LoF + functional genes had an empirical probability below the significance threshold (**Supplementary Table 9**). Fisher's exact test was used to compute the statistical significance ($P = 0.012$; OR = 12.56; CI: 1.26–632.65).

Clinical evaluation of selected patients. All patients were referred by clinical geneticists for diagnostic evaluation and overall patient characteristics were comparable to those of a previously published cohort[13]. To confirm the identification of the candidate ID genes, we compared the phenotypes of individuals with a DNM in any one of the ten candidate genes and in two genes (*SLC6A1* and *TCF7L2*) significantly enriched in the neurodevelopmental cohort. Comparison of phenotypes was only possible for 8 of 12 genes in which at least 2 individuals with ID were in the RUMC cohort (7 of 10 candidate ID genes, and 1 of 2 candidate NDD genes). A table listing these clinical details is provided in **Supplementary Table 7**. Detailed clinical data for other published individuals is mostly not available. For *TLK2* and *SETD2* a more detailed phenotypic comparison was performed (see **Supplementary Note** for case studies).

Statistics. Statistical significance was calculated using R statistical computing software version 3.1.0. Two-tailed Fisher's exact tests (significance level α of 0.05) were used to analyze statistical significance between groups for the number of LoF DNMs and number of clustered DNMs. The gene-specific analysis of excess numbers of LoF and functional DNMs was performed using one-sided exact Poisson tests with gene-specific mutation rates taken from the Samocha *et al.* study[6]. The gene-specific P-values were corrected for multiple testing based on the 18,730 genes present in the Agilent V4 exome enrichment kit times the number of tests (2; LoF and functional), using the Benjamini–Hochberg procedure with an FDR of 0.05. Data distribution was assumed to be normal, but this was not formally tested.

For the statistical testing based on random sampling we used the "sample" and "sample.int" functions (without replacement) from R version 3.1.0 with a random sample size n of 100,000. By comparing the observed value to the distribution of the random samples, the empirical P-value was computed. In addition the z-value was computed by subtracting the mean value of the simulations from the observed value and dividing by the standard deviation of the simulations.

A **Supplementary Methods Checklist** is available.

Code availability. The R code used to perform the statistical analyses is available upon request.

Data availability. The data that support the findings of this study are available as **Supplementary Tables 1–14**.

12. Neveling, K. *et al. Hum. Mutat.* **34**, 1721–1726 (2013).
13. de Ligt, J. *et al. N. Engl. J. Med.* **367**, 1921–1929 (2012).
14. Strom, S.P. *et al. Genet. Med.* **16**, 510–515 (2014).
15. Genome Diagnostics Nijmegen. *Gene Panel: Intellectual Disability* https://www.radboudumc.nl/Informatievoorverwijzers/Genoomdiagnostiek/en/Pages/Intellectualdisability.aspx (2015).
16. Kong, A. *et al. Nature* **488**, 471–475 (2012).
17. Goeman, J.J. & Solari, A. *Stat. Med.* **33**, 1946–1978 (2014).
18. MacArthur, D.G. *et al. Science* **335**, 823–828 (2012).
19. Zhu, J., He, F., Song, S., Wang, J. & Yu, J. *BMC Genomics* **9**, 172 (2008).
20. Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S. & Goldstein, D.B. *PLoS Genet.* **9**, e1003709 (2013).