

# Mutations in the U4 snRNA gene *RNU4-2* cause one of the most prevalent monogenic neurodevelopmental disorders

Received: 11 April 2024

Accepted: 23 May 2024

Published online: 31 May 2024

 Check for updates

Daniel Greene<sup>1,2</sup>, Chantal Thys<sup>3</sup>, Ian R. Berry<sup>4,5</sup>, Joanna Jarvis<sup>6</sup>, Els Ortibus<sup>7,8</sup>, Andrew D. Mumford<sup>5,9</sup>, Kathleen Freson<sup>10,3</sup> & Ernest Turro<sup>10,11</sup> ✉

Most people with intellectual disability (ID) do not receive a molecular diagnosis following genetic testing. To identify new etiologies of ID, we performed a genetic association analysis comparing the burden of rare variants in 41,132 noncoding genes between 5,529 unrelated cases and 46,401 unrelated controls. *RNU4-2*, which encodes U4 small nuclear RNA, a critical component of the spliceosome, was the most strongly associated gene. We implicated de novo variants among 47 cases in two regions of *RNU4-2* in the etiology of a syndrome characterized by ID, microcephaly, short stature, hypotonia, seizures and motor delay. We replicated this finding in three collections, bringing the number of unrelated cases to 73. Analysis of national genomic diagnostic data showed *RNU4-2* to be a more common etiological gene for neurodevelopmental abnormality than any previously reported autosomal gene. Our findings add to the growing evidence of spliceosome dysfunction in the etiologies of neurological disorders.

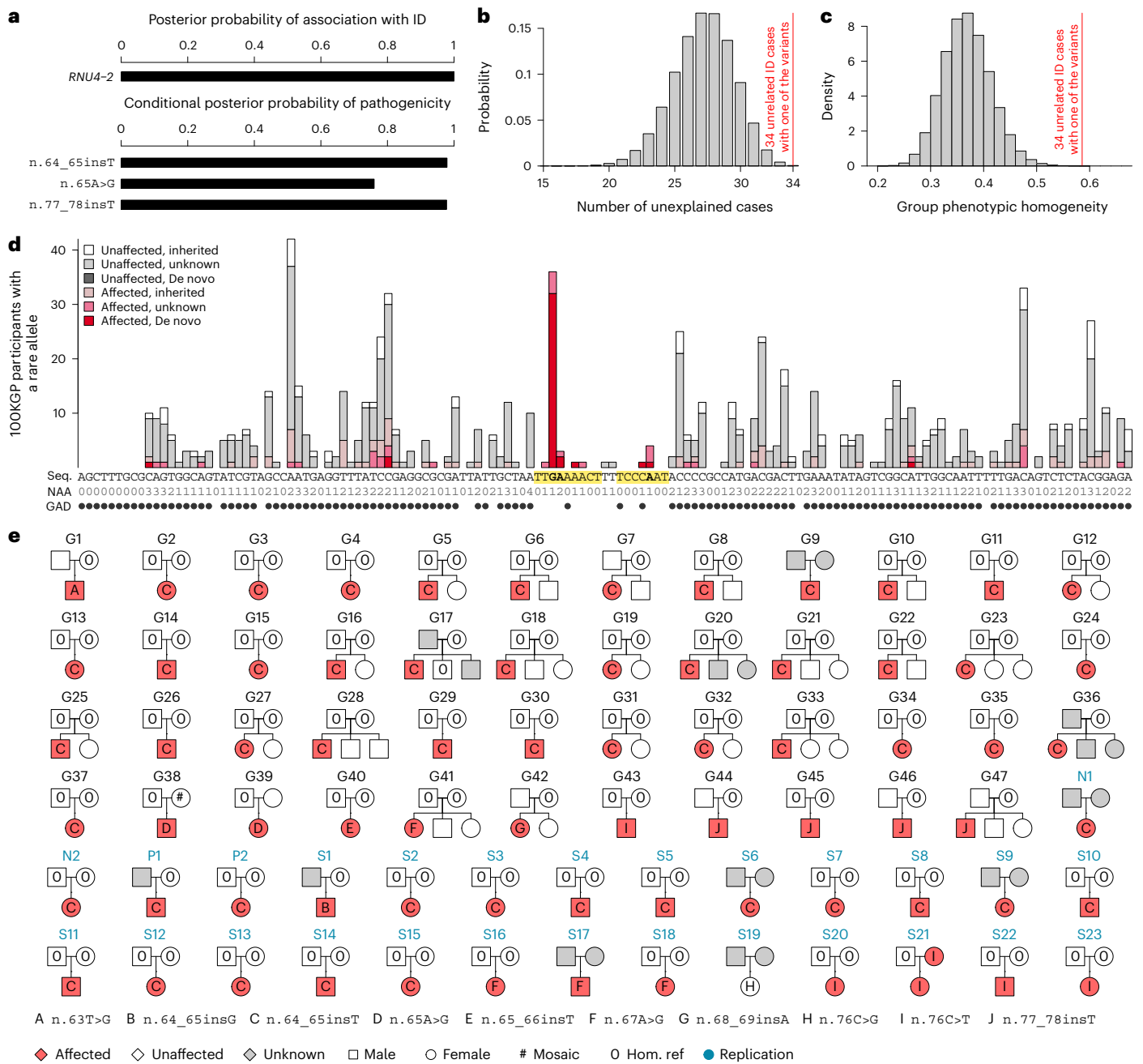
Although 1,427 genes have been identified confidently as etiological for intellectual disability (ID)<sup>1</sup>, most ID cases remain unexplained following genetic testing<sup>2</sup>. All but nine of the 1,427 known genes are protein coding<sup>3</sup>, in part because the largest genetic studies of ID have used whole-exome sequencing (WES)<sup>2</sup>, which typically omits noncoding genes. To identify noncoding etiologies of ID, we conducted a genetic association analysis using whole-genome sequencing (WGS) data on 77,539 participants enrolled in the 100,000 Genomes Project (100KGP). This study included 29,741 probands and 4,782 affected relatives assigned by expert clinicians to one or more of 220 ‘Specific Disease’ classes encompassing a wide range of pathologies<sup>4</sup>. The remaining 43,016 participants were unaffected relatives. We built a Rareservoir

database of genotypes and phenotypes<sup>4</sup> for all the study participants and applied the BeviMed genetic association method<sup>5</sup> to compare rare variant genotypes in 41,132 noncoding genes between 5,529 unrelated cases assigned to the Specific Disease class ID and 46,401 unrelated participants outside of that class.

We identified an extremely strong dominant association between rare variants in *RNU4-2* and the risk of ID (posterior probability of association (PPA)  $\approx 1$ , log Bayes factor = 55). *RNU4-2* is one of the genes encoding the U4 small nuclear RNA (snRNA) component of the small nuclear ribonucleoprotein (snRNP) U4, which in turn is one of the five snRNPs of the major spliceosome. The association was much stronger than that for any other noncoding gene (all other

<sup>1</sup>Department of Medicine, University of Cambridge, Cambridge, UK. <sup>2</sup>Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. <sup>3</sup>Department of Cardiovascular Sciences, Center for Molecular and Vascular Biology, KU Leuven, Leuven, Belgium.

<sup>4</sup>NHS South West Genomic Laboratory Hub, Southmead Hospital, Bristol, UK. <sup>5</sup>NHS South West Genomic Medicine Service Alliance, Bristol, UK. <sup>6</sup>Clinical Genetics Unit, Birmingham Women’s Hospital, Birmingham, UK. <sup>7</sup>Department of Development and Regeneration, KU Leuven, Leuven, Belgium. <sup>8</sup>Paediatric Neurology Department, University Hospitals of KU Leuven, Leuven, Belgium. <sup>9</sup>Bristol Medical School, University of Bristol, Bristol, UK. <sup>10</sup>Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. <sup>11</sup>Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. ✉e-mail: [ernest.turro@mssm.edu](mailto:ernest.turro@mssm.edu)

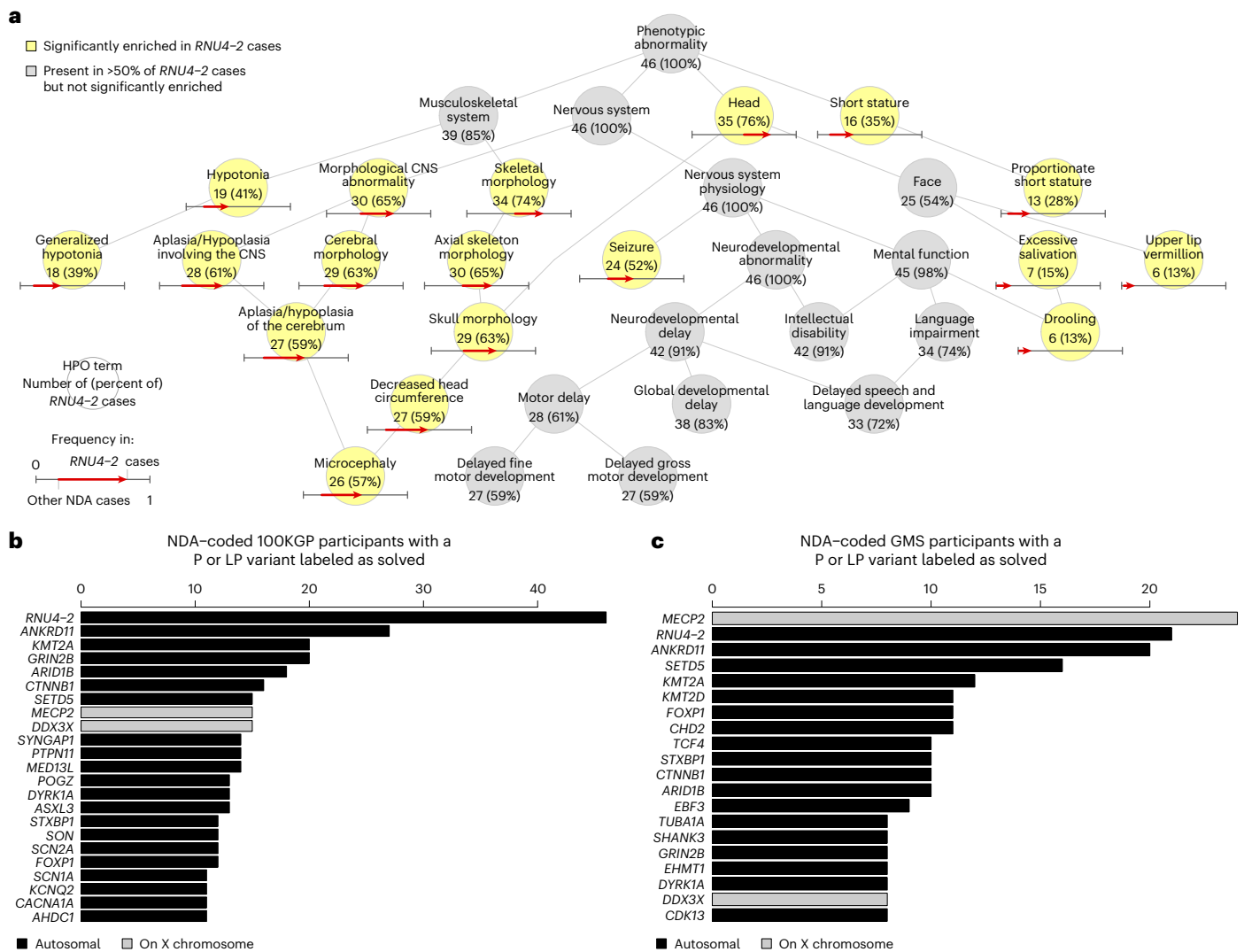


**Fig. 1 | Discovery and replication of *RNU4-2* as an etiological gene for a novel neurodevelopmental disorder.** **a**, BeviMed PPA between *RNU4-2* and the 100KGP Specific Disease ID. All other noncoding genes had a PPA < 0.5. Three *RNU4-2* variants had a conditional PPP > 0.5. **b**, Probability distribution for the number of unexplained cases among 34 randomly selected ID cases in the 100KGP. The actual number of unexplained ID cases (34) among the 34 ID cases with one of the three *RNU4-2* variants having a PPP > 0.5 is indicated with a red line. **c**, Distribution of phenotypic homogeneity scores (Methods) for randomly selected sets of 34 participants chosen from the 4,468 HPO-coded, unexplained, unrelated ID cases. The actual score for the 34 ID cases with one of the three *RNU4-2* variants having a PPP > 0.5 is indicated with a red line. **d**, For each of the 141 bases of the *RNU4-2* gene (Seq.), the number of participants with a rare allele at that position on the cDNA, stratified by affection status and inheritance information of the carried rare allele. The bases corresponding to the three

variants with a PPP > 0.5 are in bold. These three bases and adjacent bases for which no unaffected participants carry a rare allele are highlighted. At each base, the number of distinct rare alternate alleles (NAA) observed in the 100KGP is shown. Underneath, the presence of an alternate allele in gnomAD is indicated by a filled circle (GAD). **e**, Pedigrees for participants with a rare alternate allele at one of the highlighted bases. Pedigrees used for discovery have a 'G' prefix and are labeled in black. NBR, 100KGP Pilot and GMS pedigrees used for replication have an 'N', 'P' or 'S' prefix, respectively, and are labeled in blue. The affection status for pedigree members who were not study participants was obtained from pedigree tables, except where the proband was assigned to a Specific Disease that was unrelated to neurodevelopmental disorders despite having an NDA (this applies to the nine grayed out members of G9, G17, G20 and G36). One variant was called homozygous for the reference allele in a parent but found to be mosaic by inspection of aligned WGS reads (#).

PPAs < 0.5). Conditional on the association, three variants, observed among 34 ID cases, had a BeviMed posterior probability of pathogenicity (PPP) > 0.5 (Fig. 1a). None of the 34 cases were previously

solved through an explanatory pathogenic (P) or likely pathogenic (LP) variant, representing a significant depletion compared to the 20.1% of ID cases overall explained through the 100KGP diagnostic



**Fig. 2 | Phenotypic characterization and prevalence of a novel neurological disorder.** **a**, Graph showing the is-a relationships among HPO terms present in at least half of the 46 NDA-coded *RNU4-2* cases identified or significantly enriched among these 46 cases relative to 9,112 other unrelated NDA-coded participants of the 100KGP. Terms are shortened to remove ‘Abnormality of (the)’ or ‘Abnormal’ for conciseness. The significantly overrepresented terms are highlighted. For each term, the number of cases with the term and the percentage that number represents out of 46 is shown. For each overrepresented term, the proportion of NDA-coded participants that are not *RNU4-2* cases with the term and the proportion of NDA-coded *RNU4-2* cases with the term are represented as the

base and the head of an arrow, respectively. **b**, Of the 9,112 NDA-coded cases in the 100KGP, the number solved through P or LP variants in a gene, provided at least 11 cases were diagnosed. In the case of *RNU4-2*, the number of NDA-coded cases with a rare variant in the highlighted region of Fig. 1d is shown instead of the number solved with P or LP variants. **c**, Of the 5,527 NDA-coded cases in the GMS, the number solved through P or LP variants in a gene, provided at least eight cases were diagnosed. In the case of *RNU4-2*, the number of NDA-coded cases with a rare variant in the highlighted region of Fig. 1d is shown instead of the number solved with P or LP variants.

pipeline ( $P = 4.86 \times 10^{-4}$ , one-sided binomial test; Fig. 1b). Based on Human Phenotype Ontology (HPO) terms assigned to 100KGP study participants, the 34 ID cases were phenotypically more homogeneous than expected by chance ( $P = 2.33 \times 10^{-5}$ , one-sided permutation test; Fig. 1c), further suggesting that the association was causal. In addition to the 34 ID cases, a further eight participants outside the Specific Disease class ID also carried one of the three variants. Seven of these had the HPO term ‘Neurodevelopmental abnormality’ (NDA) and one had related ICD10 (International Classification of Diseases, tenth revision) codes, such as ‘Unspecified intellectual disabilities’, providing internal replication among participants treated as controls in the association analysis. To account for variations in Specific Disease assignment and phenotype coding, we hereafter consider 100KGP participants with the NDA HPO term or an ‘Intellectual disabilities’ ICD10 code to be affected with a neurodevelopmental disorder.

Five additional variants within a few base pairs of the three variants with a PPP > 0.5 were also present exclusively in affected participants, forming two contiguous regions (n.62–70 and n.73–79) within which no unaffected participants had a rare variant. These regions were embedded within a genomic locus markedly depleted of variation in gnomAD<sup>6</sup> (Fig. 1d), consistent with the effect of purifying selection. Analysis of published secondary structure data of the U4 snRNA revealed that one region maps to a quasi-pseudoknot interaction between U4 and U6 while the other maps to an interaction between U4 and U6 called stem III’ (Extended Data Fig. 1). Of the 47 rare allele instances in the regions, which we observed in 47 cases, 37 were determined to be de novo on the basis of genotype calls in the respective parents, while 10 had unknown inheritance because of missing genotypes from one or both of the corresponding parents (Fig. 1d,e). The most common variant, n.64\_65insT, was observed as a de novo

mutation in 33 different families. Intriguingly, the G at position 64 of U4 snRNA is thought to contribute to the stability of the ACAGAGA loop of U6 snRNA, which binds 5' splice sites and induces splicing after U4–U6 unwinding<sup>8</sup>. Inspection of read alignments across the 77,539 100KGP participants to assess the quality of genotyping revealed that one of the parents was mosaic for n.65A>G, while the remainder had confident heterozygous or homozygous reference genotype calls (Extended Data Fig. 2).

We sought to replicate our findings in three additional collections: the National Institute for Health and Care Research (NIHR) BioResource for Rare Diseases (NBR)<sup>9</sup> (which includes 731 NDA-coded cases out of 7,388 participants enrolled for research), the 100KGP Pilot Project (which includes 291 NDA-coded cases out of 4,054 100KGP participants enrolled in a pilot phase before the 100KGP's main program was established) and the UK's Genomic Medicine Service (GMS) (which includes 5,527 NDA-coded cases out of 25,288 participants sequenced prospectively through the UK's National Health Service (NHS)). Across these three collections, we identified a further 27 probands with a rare variant within the two regions of interest, of whom 25 were NDA coded, one was coded with 'Abnormal brain morphology' terms such that NDA could be inferred, and one was unaffected (Fig. 1e). In total, 19 of the 26 affected probands acquired the variant de novo, six acquired the variant with inheritance that was unknown because of a lack of parental genotype data and one case inherited variant n.76C>T from an affected mother. The n.76C>T variant is the only one among the 74 pedigrees that is present in gnomAD, where it has an allele count of one in 152,108. While we observed this variant as a de novo mutation in four families, the inheritance in one family and the observation in gnomAD suggest that it may cause a less severe phenotype in some cases than the other variants we identified, particularly the predominant variant n.64\_65insT. Moreover, variant n.76C>G was observed in one case without NDA, further suggesting that certain mutations at nucleotide position 76 may have a limited or benign effect.

To further characterize this new syndrome, we analyzed the HPO terms of the 46 100KGP cases annotated with the NDA term (one of the 47 cases had consistent ICD10 codes but was erroneously not annotated with the NDA HPO term). Of these 46 cases, 91% had the ID term, 91% had 'Neurodevelopmental delay' and 61% had 'Motor delay', broadly in line with general term frequencies across ID cases (Fig. 2a). However, several terms were significantly overrepresented among the 46 *RNU4-2* cases, including 'Microcephaly' in 57% (versus 18% in other NDA-coded cases, Bonferroni-adjusted  $P = 3.23 \times 10^{-7}$ ), 'Drooling' in 13% (versus 1%,  $P = 6.93 \times 10^{-4}$ ), 'Proportionate short stature' in 28% (versus 7%,  $P = 7.60 \times 10^{-4}$ ), 'Generalized hypotonia' in 39% (versus 13%,  $P = 8.08 \times 10^{-4}$ ), 'Seizure' in 52% (versus 27%,  $P = 3.13 \times 10^{-2}$ ) and 'Abnormality of upper lip vermillion' in 13% (versus 2%,  $P = 4.00 \times 10^{-2}$ ) (Fig. 2a). No terms were significantly underrepresented in the *RNU4-2* cases.

To assess the prevalence of this new disorder, we compared the number of NDA-coded *RNU4-2* cases in the 100KGP (46 of 9,112 NDA-coded cases, 0.50%) with the number of solved cases with P or LP variants in other etiological genes for NDA. *RNU4-2* was the most prevalent etiology for NDA in the 100KGP (Fig. 2b). As the 100KGP was subject to genetic prescreening, we also assessed the prevalence among persons undergoing diagnostic WGS through the UK GMS, which prescreens only for aneuploidy, copy-number variants, tandem repeats and abnormal methylation. In the GMS, *RNU4-2* was the second most prevalent etiology for NDA (21 of 5,527 NDA-coded cases, 0.38%) after *MECP2* (24 cases), a long-established etiological X-linked gene for Rett syndrome<sup>10</sup> (Fig. 2c).

Splicing of eukaryotic pre-mRNA is catalyzed by large macromolecular complexes termed spliceosomes that comprise snRNPs and additional proteins<sup>11</sup>. The major spliceosome assembles on pre-mRNA through the sequential association of five snRNPs (U1, U2, U4, U5 and U6), each containing a unique snRNA, and a total of approximately

100 additional proteins<sup>12</sup>. Assembly of the major spliceosome is initiated by binding of U1 and U2 snRNPs to the pre-mRNA strand. Thereafter, the preassembled U4–U6–U5 tri-snRNP is recruited to form the inactive pre-B complex. The interaction between U4 and U6 snRNAs within the tri-snRNP (Extended Data Fig. 1) contributes to the structural stability of the pre-B complex, particularly of the critical ACAGAGA loop of U6 snRNA<sup>13</sup> that helps ensure the fidelity of the interaction with the intron. Activation of the major spliceosome is initiated by PRP28, which transfers the 5' splice site to U5 snRNP, thereby forming the major spliceosome B complex. This is followed by disassembly of the U4–U6 duplex by the helicase SNRNP200 (also known as BRR2), which is followed by pairing with U2 snRNP at the 3' splice site and complete removal of the intron<sup>8</sup>. Many components of the major spliceosome are ubiquitously expressed, including *RNU4-2* (Extended Data Fig. 3). However, some interacting protein components show tissue-specific expression, which might explain why pathogenic variants in genes encoding some of these components cause defects (spliceosomeopathies) that are restricted to subsets of cell lineages<sup>14</sup>. The same applies to genes encoding snRNAs of the minor spliceosome. Compound heterozygous and homozygous variants in *RNU4ATAC* and *RNU12* can cause recessive conditions restricted to certain organ systems in certain persons, albeit with apparently fully penetrant ID<sup>15–18</sup>. Before the present work, none of the five snRNAs of the major spliceosome were implicated in a human disorder. Very recently, biallelic loss-of-function variants in *WBP4*, which encodes a protein component of the major spliceosome, were found to cause a neurodevelopmental syndrome with remarkably similar phenotypes to those we describe here, including hypotonia, global developmental delay, severe ID, brain abnormalities and musculoskeletal defects<sup>19</sup>. WW domain-binding protein 4 (*WBP4*) delays the U4–U6 unwinding activity of SNRNP200 and thereby spliceosome activation<sup>20</sup>. Cryo-EM structures of the pre-B complex suggest that the quasi-pseudoknot promotes interaction between 5' splice sites and U6 snRNA's ACAGAGA loop to bring about SNRNP200 relocation in U4 snRNA, enabling SNRNP200 to unwind U4–U6 and trigger splicing activation<sup>8</sup>. Given that the predominant variant, n.64\_65insT, could disrupt the G64(U4)–A47(U6) interaction supporting the quasi-pseudoknot structure, a delay in U4–U6 unwinding activity similar to that observed in *WBP4* deficiency might be the underlying etiological mechanism of this newly described syndrome.

## Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41591-024-03085-5>.

## References

1. Martin, A. R. et al. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat. Genet.* **51**, 1560–1565 (2019).
2. Wright, C. F. et al. Genomic diagnosis of rare pediatric disease in the United Kingdom and Ireland. *N. Engl. J. Med.* **388**, 1559–1571 (2023).
3. Genomics England PanelApp. Available from <https://panelapp.genomicsengland.co.uk> (accessed on April 4, 2024), intellectual disability—microarray and sequencing (version 5.515).
4. Greene, D. et al. Genetic association analysis of 77,539 genomes reveals rare disease etiologies. *Nat. Med.* **29**, 679–688 (2023).
5. Greene, D., Richardson, S. & Turro, E. A fast association test for identifying pathogenic variants involved in rare diseases. *Am. J. Hum. Genet.* **101**, 104–114 (2017).



6. Chen, S. et al. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92–100 (2024).
7. Jakab, G. et al. *Chlamydomonas* U2, U4 and U6 snRNAs. An evolutionary conserved putative third interaction between U4 and U6 snRNAs which has a counterpart in the U4<sub>atac</sub>–U6<sub>atac</sub> snRNA duplex. *Biochimie* **79**, 387–395 (1997).
8. Charenton, C., Wilkinson, M. E. & Nagai, K. Mechanism of 5' splice site transfer for human spliceosome activation. *Science* **364**, 362–367 (2019).
9. Turro, E. et al. Whole-genome sequencing of patients with rare diseases in a national health system. *Nature* **583**, 96–102 (2020).
10. Amir, R. E. et al. Rett syndrome is caused by mutations in X-linked *MECP2*, encoding methyl-CpG-binding protein 2. *Nat. Genet.* **23**, 185–188 (1999).
11. Pan, Q. et al. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **40**, 1413–1415 (2008).
12. Wilkinson, M. E., Charenton, C. & Nagai, K. RNA splicing by the spliceosome. *Annu. Rev. Biochem.* **89**, 359–388 (2020).
13. Nguyen, T. H. et al. The architecture of the spliceosomal U4/U6-U5 tri-snRNP. *Nature* **523**, 47–52 (2015).
14. Griffin, C. & Saint-Jeannet, J. P. Spliceosomopathies: diseases and mechanisms. *Dev. Dyn.* **249**, 1038–1046 (2020).
15. He, H. et al. Mutations in U4<sub>atac</sub> snRNA, a component of the minor spliceosome, in the developmental disorder MOPD I. *Science* **332**, 238–240 (2011).
16. Merico, D. et al. Compound heterozygous mutations in the noncoding RNU4ATAC cause Roifman Syndrome by disrupting minor intron splicing. *Nat. Commun.* **6**, 8718 (2015).
17. Farach, L. S. et al. The expanding phenotype of RNU4ATAC pathogenic variants to Lowry Wood syndrome. *Am. J. Med. Genet. A* **176**, 465–469 (2018).
18. Elsaid, M. F. et al. Mutation in noncoding RNA RNU12 causes early onset cerebellar ataxia. *Ann. Neurol.* **81**, 68–78 (2017).
19. Engal, E. et al. Bi-allelic loss-of-function variants in *WBP4*, encoding a spliceosome protein, result in a variable neurodevelopmental syndrome. *Am. J. Hum. Genet.* **110**, 2112–2119 (2023).
20. Henning, L. M. et al. A new role for FBP21 as regulator of Brr2 helicase activity. *Nucleic Acids Res.* **45**, 7922–7937 (2017).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024

## Methods

### Enrollment criteria for the ID class in the 100KGP

The enrollment criteria for all specific diseases in the 100KGP are available from the 100KGP website<sup>21</sup>. The provided criteria for ID are reproduced here verbatim. “ID inclusion criteria. Moderate to severe/profound ID disproportionate to parental IQ unless the family history is consistent with an X-linked disorder; congenital onset; developmental delay;  $\pm$  clinical features suggestive of a specific syndrome; metabolic causes have been excluded. ID exclusion criteria. Antenatal history suggestive of non-genetic cause, proven congenital or neonatal infections; known genetic cause already identified; microarray analysis abnormal and clearly pathogenic. Prior genetic testing guidance. Results should have been reviewed for all genetic tests undertaken, including disease-relevant genes in WES data; the patient is not eligible if they have a molecular diagnosis for their condition; genetic testing should continue according to routine local practice for this phenotype regardless of recruitment to the project; results of these tests must be submitted through the ‘Genetic investigations’ section of the data capture tool to allow comparison of WGS with current standard testing. ID prior genetic testing genes. Testing of the following genes should be carried out prior to recruitment where this is in line with current local practice: for syndromes where the cause of disease is 1–2 genes these need to be excluded before recruitment, for example, for Kabuki syndrome, *KMT2D* and *KDM6A* should have been tested.”

### Genetic association analysis

We constructed a Rareservoir database<sup>4</sup> in the Genomics England Research Environment containing rare variants extracted from 77,539 consented participants in each of the 41,132 noncoding canonical transcripts listed in Ensembl version 104. We annotated the participants with Specific Disease classes and HPO terms. We imputed the ID HPO term in cases in the ID class wherever the term had been omitted. We annotated and filtered the variants as previously described<sup>4</sup>. We applied the BeviMed<sup>5</sup> association test to each noncoding gene comparing unrelated participants in the ID class with unrelated participants outside of that class. The set of cases was constructed by selecting one case from each pedigree containing at least one person assigned to the ID class and the set of controls was formed by taking the intersection of the maximal set of unrelated participants provided by the 100KGP with participants not related to any of the cases.

Using BeviMed, we performed a Bayesian comparison of three models:

1. No association (prior probability 0.99)
2. Dominant association, taking variants with a probabilistic minor allele frequency (pMAF)<sup>9</sup> < 0.01% (prior probability 0.005)
3. Recessive association, taking variants with a pMAF < 0.1%, (prior probability 0.005).

Thus, the overall prior probability of association was 0.01. The hyperparameters were set to the default values given in the literature<sup>5</sup>. The PPA was the sum of the posterior probabilities of models 2 and 3. The analysis was carried out using R 3.6.2, making use of packages Matrix 1.2–18, dplyr 0.8.5, bit64 0.9–7, bit 1.1–14, DBI 1.1.0, RSQLite 2.1.4 and BeviMed 5.7.

### Phenotypic similarity analysis

To determine whether a given subset of HPO-annotated, unrelated participants was more phenotypically homogeneous than expected by chance, we applied the following approach: We computed the information content (IC) of each HPO term as  $-\log$  of its frequency. We chose Resnik’s<sup>22</sup> similarity function  $s$  to compute the similarity between two terms  $t_1$  and  $t_2$ :  $s(t_1, t_2) = \max_{t \in \text{anc}(t_1) \cap \text{anc}(t_2)} \text{IC}(t)$ , where  $\text{anc}(t)$  denotes the union of term  $t$  and its ancestral nodes in the HPO graph. We defined the similarity  $S$  between two sets of terms  $\phi_1$  and  $\phi_2$  (for example, those

attached to two study participants) using the best match average

function<sup>23</sup>:  $S(\phi_1, \phi_2) = \frac{1}{2|\phi_1|} \sum_{t_1 \in \phi_1} \max_{t_2 \in \phi_2} s(t_1, t_2) + \frac{1}{2|\phi_2|} \sum_{t_2 \in \phi_2} \max_{t_1 \in \phi_1} s(t_2, t_1)$ .

We defined the phenotypic homogeneity of a group of size  $k$  as the mean pairwise similarity:  $h(\phi_1, \phi_2, \dots, \phi_k) = \binom{k}{2}^{-1} \sum_{i=1}^{k-1} \sum_{j=i+1}^k S(\phi_i, \phi_j)$ .

To determine whether the homogeneity of a group of size  $k$  was significantly greater than expected by chance, we selected sets of  $k$  participants at random and obtained a Monte Carlo  $P$  value as the proportion of random sets that had a homogeneity greater than or equal to the homogeneity of the group.

### Phenotypic characterization

To identify enriched or depleted HPO terms among the 46 NDA-annotated cases with *RNU4-2* variants in the regions of interest, compared to unrelated NDA-coded participants without *RNU4-2* variants, we computed  $P$  values of association using Fisher’s two-sided exact test. We only tested enrichment for terms attached to at least five of the 46 cases and that belonged to the set of nonredundant terms at each level of frequency among the cases. To account for multiple comparisons, we adjusted the  $P$  values by multiplying them by the number of tests. An adjusted  $P$  value < 0.05 was deemed statistically significant. To visualize both common and distinctive HPO terms for *RNU4-2* cases, we selected terms that were either statistically significant or present in at least 50% of the cases, removed redundant terms at each level of frequency among the 46 cases and arranged the terms along with a nonredundant set of ancestral terms as a directed acyclic graph of is-a relations. These analyses were conducted using the ontologyX R package<sup>24</sup>.

### Ethics

Participants of the 100KGP, the 100KGP Pilot Project and the GMS were enrolled to the National Genomic Research Library under a protocol approved by the East of England–Cambridge Central Research Ethics Committee (20/EE/0035). NBR participants were enrolled under a protocol approved by the East of England–Cambridge South Research Ethics Committee (13/EE/0325). The Ethics Committee of University Hospitals Leuven approved genetic and experimental studies of a pedigree enrolled to the NBR in Belgium (ML3580/S50025 and S63666). Only participants who provided written informed consent for their data to be used for research were included in the analyses.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

### Data availability

Genetic and phenotypic data for the 100KGP study participants, the 100KGP Pilot study participants and the GMS participants are available through the Genomics England Research Environment through application at <https://www.genomicsengland.co.uk/join-a-gecip-domain>. Data pertaining to WGS data were obtained from a merged variant call format file (VCF) for 77,539 100KGP participants, a merged VCF for 4,054 100KGP Pilot participants, single-genome VCFs for 25,289 GMS participants (v3) and single sample gVCFs for 13,037 NBR participants. HPO phenotype data were obtained from the ‘rare\_diseases\_participant\_phenotype’ table (Main Program v13), ‘observation’ table (GMS v3) and ‘hpo’ table (Rare Diseases Pilot v3). Specific Disease class data were obtained from the ‘rare\_diseases\_participant\_disease’ table (Main Program v13). ICD10 codes were obtained from the ‘hes\_apc’ table (Main Program v13). Pedigree information was obtained from the ‘rare\_diseases\_pedigree\_member’ table (Main Program v13), ‘referral\_participant’ table (GMS v3) and ‘pedigree’ table (Rare Diseases Pilot v3). The explained or unexplained status of cases was obtained from the ‘gmc\_exit\_questionnaire’ tables (Main Program v18, GMS v3). Accession

codes for NBR data are given in the literature<sup>9</sup>. CADD version 1.5 (<https://cadd.gs.washington.edu/>), gnomAD version 3.0 (<https://gnomad.broadinstitute.org/>) and Ensembl version 104 (<http://may2021.archive.ensembl.org/index.html>) were used for variant annotation.

Data presented in this paper were requested from the Genomics England Airlock on April 2, 2024 at 11:08 p.m. British Summer Time (BST). The manuscript was submitted to the Genomics England Publication Committee on April 8, 2024 at 5:45 a.m. BST and approved for submission on April 11, 2024 at 12:24 p.m. BST.

## Code availability

Software packages rsrv 1.0, bcftools 1.16, samtools 1.9 and perl 5 were used to build the noncoding 100KGP Rareservoir. The Rareservoir software is available from <https://github.com/turrogroupr/rsrv>. All R packages listed in the manuscript are available from the Comprehensive R Archive Network site (<https://cran.r-project.org>).

Expression data for *RNU4-2* were extracted from the file 'GTEx\_Analysis\_2017-06-05\_v8\_RNASeQCv1.1.9\_gene\_tpm.gct.gz' available from the GTEx Portal.

## References

21. Devereau, A. 100,000 Genomes Project Rare Disease Eligibility Criteria (Genomics England, 2018); <https://files.genomicsengland.co.uk/forms/Rare-Disease-Eligibility-Criteria.pdf>
22. Resnik, P. et al. Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res.* **11**, 95–130 (1999).
23. Köhler, S. et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am. J. Hum. Genet.* **85**, 457–464 (2009).
24. Greene, D., Richardson, S. & Turro, E. ontologyX: a suite of R packages for working with ontological data. *Bioinformatics* **33**, 1104–1106 (2017).
25. Agafonov, D. E. et al. Molecular architecture of the human U4/U6-U5 tri-snRNP. *Science* **351**, 1416–1420 (2016).
26. Boesler, C. et al. A spliceosome intermediate with loosely associated tri-snRNP accumulates in the absence of Prp28 ATPase activity. *Nat. Commun.* **7**, 11997 (2016).
27. Sarka, K., Katzman, S. & Zahler, A. M.A role for SNU66 in maintaining 5' splice site identity during spliceosome assembly. *RNA* **30**, 695–709 (2024).
28. Leung, A. K., Nagai, K. & Li, J. Structure of the spliceosomal U4 snRNP core domain and its implication for snRNP biogenesis. *Nature* **473**, 536–539 (2011).
29. Aguet, F. et al. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).

## Acknowledgements

This research was made possible through access to data in the National Genomic Research Library, which is managed by Genomics

England Limited (a wholly owned company of the Department of Health and Social Care). The National Genomic Research Library holds data provided by patients and collected by the NHS as part of their care and data collected as part of their participation in research. The National Genomic Research Library is funded by the NIHR and NHS England. The Wellcome Trust, Cancer Research UK and the Medical Research Council also funded research infrastructure. We thank NIHR BioResource volunteers for their participation and gratefully acknowledge NIHR BioResource centers, NHS Trusts and staff for their contribution. We thank the NIHR, NHS Blood and Transplant and Health Data Research UK as part of the Digital Innovation Hub Program. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. K.F. was supported by Katholieke Universiteit (KU) Leuven Special Research Fund (BOF) (C14/19/O96 and C14/23/121) and Research Foundation–Flanders (G072921N). D.G. and E.T. were supported by NIH awards R01HL161365 and R03HD111492 and E.T. was further supported by the Lowy Foundation USA.

## Author contributions

D.G. conducted analyses and cowrote the paper. C.T. performed experiments. I.R.B. and J.J. collected phenotype information. E.O. obtained consent and collected samples and phenotype information. A.M. provided clinical and biological interpretation and contributed to writing the paper. K.F. provided biological interpretation, designed and oversaw experiments and contributed to writing the paper. E.T. oversaw the study and cowrote the paper.

## Competing interests

The authors declare no competing interests.

## Additional information

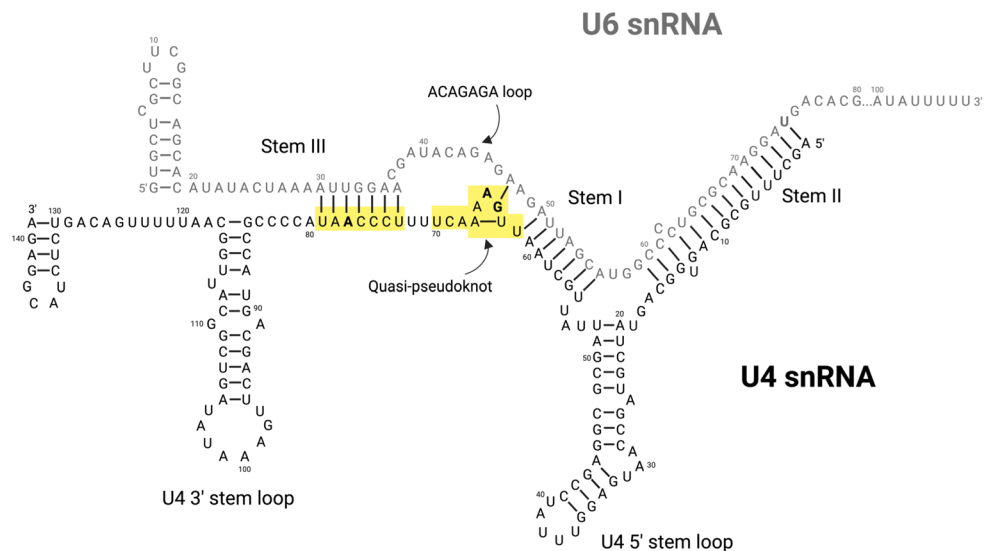
**Extended data** is available for this paper at <https://doi.org/10.1038/s41591-024-03085-5>.

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41591-024-03085-5>.

**Correspondence and requests for materials** should be addressed to Ernest Turro.

**Peer review information** *Nature Medicine* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Jerome Staal, in collaboration with the *Nature Medicine* team.

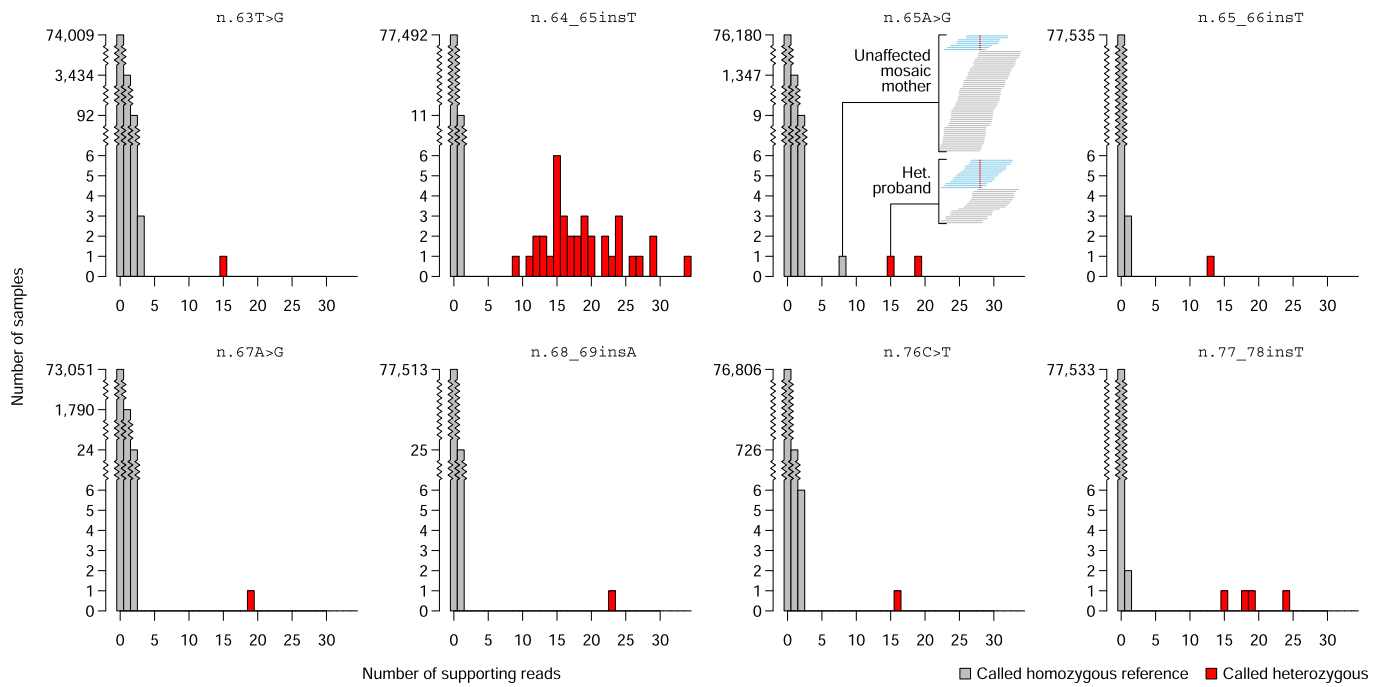
**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).



**Extended Data Fig. 1 | Organization of the U4/U6 snRNA duplex within the tri-snRNP U4/U6.U5.** Human U4 snRNA is stabilized by three intramolecular stem loop motifs (ribonucleotides 20–52, 85–117 and 127–144) and maintains an inhibitory interaction with U6 snRNA through extensive intermolecular pairings involving U4 ribonucleotides within stem I (56–61), stem II (1–16), stem III (73–79) and the U4 quasi-pseudoknot (62 and 64). This region is further stabilized by interactions between U4 ribonucleotides 68–70 and the RNA-binding protein 42 (RBM42)<sup>8,25,26</sup> and with more extensive contacts between this region and U4/U6.U5 tri-snRNP-associated protein 1 (SNUT1)<sup>27</sup> and U4/U6.U5 snRNP 27 kDa protein (SNRNP-27K)<sup>28</sup>. The U4 stem III, quasi-pseudoknot and associated RNA binding proteins determine the orientation of the U6 snRNA ACAGAGA loop, ensuring its accessibility for interaction with the 5' splice site of target introns

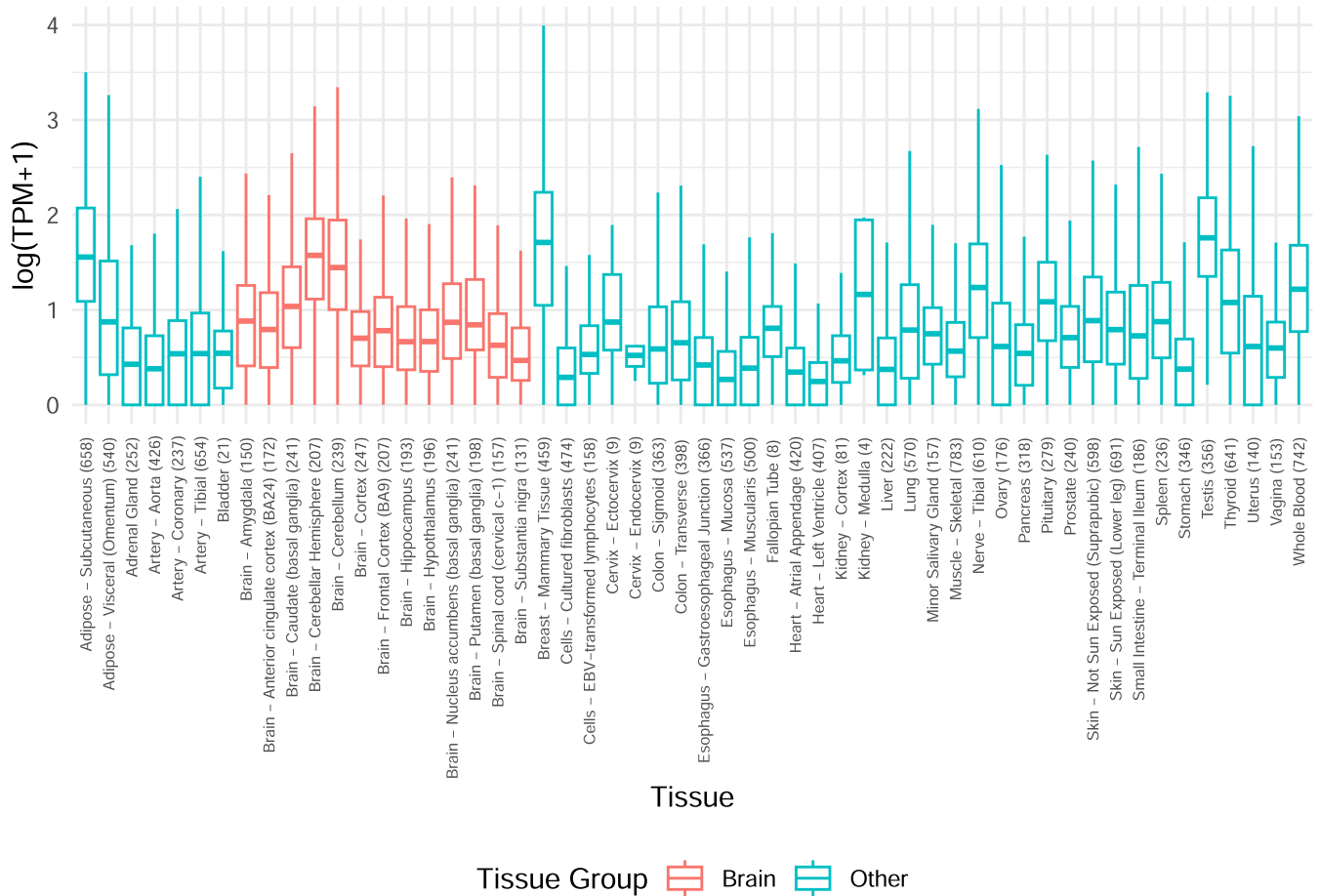
that is necessary for the catalytic activation of the spliceosome complex<sup>8</sup>. The variants associated with neurodevelopmental disorder lie in the central region of the U4/U6 snRNA duplex (yellow boxes; bolded ribonucleotides represent those harboring variants with a PPP > 0.5), are predicted to disrupt Watson-Crick interactions within U4 stem III (n.76 C > T and n.77\_78insT) or to disrupt configuration of the U4 quasi-pseudoknot (n.63 T > G; n.64\_65insG; n.64\_65insT; n.65A > G; n.65\_66insT; n.67 A > G and n.68\_69insA) thereby de-stabilizing U4 snRNA interactions with U6 snRNA and with RNA binding proteins that are necessary for correct spliceosome function. The image represents the predicted U4/U6 duplex secondary structure with ribonucleotide numbering according to the canonical *RNU4-2* and *RNU6-1* reference sequences ENST00000365668.1 and ENST00000383898.1 respectively.





**Extended Data Fig. 2 | Mosaicism analysis.** For each of the eight rare variants carried by 100KGP participants in the highlighted region of Fig. 1c, truncated bar charts showing the distribution of the number of reads supporting the alternate allele. The participants called heterozygous for a given *RNU4-2* rare variant are represented in red, while other participants are represented in gray.

The embedded window in the panel for variant n.65 A > G shows the read pileups at this position in the heterozygous proband of pedigree G38 and his mother, who had a homozygous reference genotype call but is likely mosaic. The reads supporting the reference allele are in blue and those supporting the variant allele are in gray.



**Extended Data Fig. 3 | Expression of the gene encoding the U4 snRNA, *RNU4-2*, across tissues.** Boxplot of *RNU4-2* log expression across adult tissues subjected to RNA-seq by the GTEx Consortium<sup>29</sup>. The number of samples corresponding to each tissue type is shown in brackets. Brain tissue types are

highlighted. TPM: transcripts per million. The lower, center and upper edges of the boxes respectively indicate the lower quartile, median and upper quartiles. Whiskers are drawn up to the most extreme points that are less than 1.5 times the interquartile range away from the nearest quartile.

## Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection samtools 1.9; bcftools 1.16; perl 5.

Data analysis rsvr 1.0 (<https://github.com/turrogrouprsvr>); bcftools 1.16; samtools 1.9; perl 5; R 3.6.2; R packages Matrix 1.2-18, dplyr 0.8.5, bit64 0.9-7, bit 1.1-14, DBI 1.1.0, RSQLite 2.1.4, BeviMed 5.7, ontologyIndex 2.12, ontologySimilarity 2.7, ggplot2 3.5.0.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Genetic and phenotypic data for the 100KGP study participants, the 100KGP Pilot study participants and the GMS participants are available through the Genomics England Research Environment via the application at <https://www.genomicsengland.co.uk/join-a-gecip-domain>. Data pertaining to: WGS data were obtained from a merged variant call format file (VCF) for 77,539 100KGP participants, a merged VCF for 4,054 100KGP Pilot participants, single genome VCFs for 25,289 GMS

participants (v3) and single sample gVCFs for 13,037 NBR participants; HPO phenotype data from the 'rare\_diseases\_participant\_phenotype' table (Main Programme v13), 'observation' table (GMS v3) and 'hpo' table (Rare Diseases Pilot v3); Specific Disease class data from the 'rare\_diseases\_participant\_disease' table (Main Programme v13); ICD10 codes from the 'hes\_apc' table (Main Programme v13); pedigree information from the 'rare\_diseases\_pedigree\_member' table (Main Programme v13), 'referral\_participant' table (GMS v3), and 'pedigree' table (Rare Diseases Pilot v3); explained/unexplained status of cases from the 'gmc\_exit\_questionnaire' tables (Main Programme v18, GMS v3). Accession codes for NBR data are given in ref.[PMID:32581362]. CADD v.1.5 (<https://cadd.gs.washington.edu/>), gnomAD v.3.0 (<https://gnomad.broadinstitute.org/>) and Ensembl v.104 (<http://may2021.archive.ensembl.org/index.html>) were used for variant annotation.

## Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research](#).

|                             |   |
|-----------------------------|---|
| Reporting on sex and gender | Breakdown by genetically determined sex for the 100KGP discovery collection as provided in the Genomics England Research Environment: 40,332 female; 35,511 male; 1,696 not available.  |
| Population characteristics  | Collection of rare disease participants and relatives covering a wide range of pathologies. Breakdown by genetically determined most probable ancestry for the 100KGP discovery collection as provided in the Genomics England Research Environment: African: 2,762, Admixed American: 3,006; East Asian: 573; European: 63,493; South Asian: 7,705. Ages of 100KGP participants ranged between 0 and 110, with a lower quartile of 27, a median of 42 and an upper quartile of 58, with 18.4% under 18 overall.  |
| Recruitment                 | Participants were identified by clinicians as eligible for recruitment to the 100,000 Genomes Project or for clinical testing through the United Kingdom's National Health Service Genomic Medicine Centres. The eligibility criteria are available from the Genomics England web site ( <a href="https://www.genomicsengland.co.uk">https://www.genomicsengland.co.uk</a> ). Two cases were enrolled to the NIHR BioResource Rare Diseases (see <a href="https://doi.org/10.1038/s41586-020-2434-2">https://doi.org/10.1038/s41586-020-2434-2</a> for more information on recruitment to that study). The studies were presented to eligible patients or their guardians by their clinicians widely across the health system, minimising selection bias subject to the enrolment criteria. |
| Ethics oversight            | Participants of the 100KGP, the 100KGP Pilot Project and the GMS were enrolled to the National Genomic Research Library under a protocol approved by the East of England–Cambridge Central Research Ethics Committee (ref: 20/EE/0035). NBR participants were enrolled under a protocol approved by the East of England Cambridge South Research Ethics Committee (ref. 13/EE/0325). The Ethics Committee of University Hospitals Leuven approved genetic and experimental studies of a pedigree enrolled to the NBR in Belgium (ML3580/S50025 and S63666). Only participants who provided written informed consent for their data to be used for research were included in the analyses.   |

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences       Behavioural & social sciences       Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

|                 |   |
|-----------------|---|
| Sample size     | Statistical power to identify genetic associations with rare diseases depends on various factors including the sample sizes and genetic homogeneities of case groups. To our knowledge, a formal sample size calculation was not performed for the 100,000 Genomes Project. However, the study was informed by previous smaller studies showing sufficient power (see references in Turro et al. (2020), Nature). |
| Data exclusions | None.   |
| Replication     | Replication was achieved by analyzing genomes in the three collections other than the 100KGP discovery collection: the 100KGP Pilot project, the NIHR BioResource–Rare Diseases and the Genomic Medicine Service.   |
| Randomization   | Recruitment and genome sequencing were performed concurrently across rare disease categories, thus randomizing the order in which individuals were sequenced with respect to phenotype.   |
| Blinding        | This is an observational genetic study, not a clinical trial. As genome sequencing followed enrolment, participants and investigators were unaware of the participant genotypes generated by the 100KGP at enrolment.   |

## Reporting for specific materials, systems and methods



We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

- | n/a                                 | Included in the study                                  |
|-------------------------------------|--|
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Antibodies                    |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Eukaryotic cell lines         |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Palaeontology and archaeology |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Animals and other organisms   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Clinical data                 |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Dual use research of concern  |

### Methods

- | n/a                                 | Included in the study                           |
|-------------------------------------|---|
| <input checked="" type="checkbox"/> | <input type="checkbox"/> ChIP-seq               |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Flow cytometry         |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> MRI-based neuroimaging |