## De novo variants in the *RNU4-2* snRNA cause a frequent neurodevelopmental syndrome

Yuyang Chen1,2, Ruebena Dawes1,2\*, Hyung Chul Kim1,2\*, Alicia Ljungdahl3,4\*, Sarah L Stenton5,6\*, Susan Walker7\*, Jenny Lord8, Gabrielle Lemire5,6, Alexandra C Martin-Geary1,2, Vijay S Ganesh5,6,9, Jialan Ma5, Jamie M Ellingford7,10,11, Erwan Delage12, Elston N D'Souza1,2, Shan Dong3,4, David R Adams13, Kirsten Allan14, Madhura Bakshi15, Erin E Baldwin16, Seth I Berger17,18, Jonathan A Bernstein19,20,21, Ishita Bhatnagar22, Ed Blair22, Natasha J Brown14,23, Lindsay C Burrage24, Kimberly Chapman18, David J Coman25,26,27, Alison G Compton14,23,28, Chloe A Cunningham14,23, Precilla D'Souza13, Petr Danecek12, Emmanuèle C Délot17, Kerith-Rae Dias29,30, Ellen R Elias31,32, Frances Elmslie33, Care-Anne Evans29,34, Lisa Ewans35,36,37, Kimberly Ezell38, Jamie L Fraser17,18, Lyndon Gallacher14,23, Casie A Genetti6,39, Anne Goriely40,41, Christina L Grant18, Tobias Haack42,43, Jenny E Higgs44, Anjali G Hinch2, Matthew E Hurles12, Alma Kuechler45, Katherine L Lachlan46,47, Seema R Lalani24, François Lecoquierre48, Elsa Leitão45, Anna Le Fevre14, Richard J Leventer23,28,49, Jan E Liebelt50,51, Sarah Lindsay12, Paul J Lockhart23,52, Alan S Ma53,54, Ellen F Macnamara13, Sahar Mansour33, Taylor M Maurer20,21,55, Hector R Mendez20,21,56, Kay Metcalfe57, Stephen B Montgomery20,21,58, Mariya Moosajee59,60,61, Marie-Cécile Nassogne62,63, Serena Neumann38, Michael O'Donoghue64, Melanie O'Leary5, Elizabeth E Palmer35,36, Nikhil Pattani33, John Phillips38, Georgia Pitsava65, Ryan Pysar35,36,66, Heidi L Rehm5,67, Chloe M Reuter20,21,56, Nicole Revencu68, Angelika Riess42, Rocio Rius23,69,70, Lance Rodan6, Tony Roscioli29,30,34, Jill A Rosenfeld24, Rani Sachdev35,36, Charles J Shaw-Smith71, Cas Simons69,70, Sanjay M Sisodiya72,73, Penny Snell52, Laura St Clair53, Zornitza Stark14,23, Helen S Stewart22, Tiong Yang Tan14,23, Natalie B Tan14, Suzanna EL Temple15,74, David R Thorburn14,23,28, Cynthia J Tifft13, Eloise Uebergang28, Grace E VanNoy5, Pradeep Vasudevan75, Eric Vilain76, David H Viskochil16, Laura Wedd69,70, Matthew T Wheeler20,21,56, Susan M White14,23, Monica Wojcik6,39,77, Lynne A Wolfe13, Zoe Wolfenson13, Caroline F Wright78, Changrui Xiao79, David Zocche80, John L Rubenstein4, Eirene Markenscoff-Papadimitriou81, Sebastian M Fica82, Diana Baralle83,84, Christel Depienne45, Daniel G MacArthur69,70, Joanna MM Howson85, Stephan J Sanders3,4, Anne O'Donnell-Luria5,6,67, and Nicola Whiffin1,2,5

1. Big Data Institute, University of Oxford, Oxford, UK

2. Centre for Human Genetics, University of Oxford, Oxford, UK

3. Institute of Developmental and Regenerative Medicine, Department of Paediatrics, University of Oxford, Oxford, UK

4. Department of Psychiatry and Behavioral Sciences, UCSF Weill Institute for Neurosciences, University of California, San Francisco, CA, USA

5. Broad Center for Mendelian Genomics, Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA

6. Division of Genetics and Genomics, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA

7. Genomics England, London, UK

8. Sheffield Institute for Translational Neuroscience (SITraN), University of Sheffield, Sheffield, UK

9. Department of Neurology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

10. Manchester Centre for Genomic Medicine, Manchester University NHS Foundation Trust, Manchester, UK

11. Division of Evolution, Infection and Genomic Sciences, School of Biological Sciences, Faculty of Biology, Medicines and Health, University of Manchester, Manchester, UK

12. Human Genetics, Wellcome Sanger Institute, Hinxton, UK

13. Undiagnosed Disesases Program, National Human Genome Research Institute, Bethesda, MD, USA

14. Victorian Clinical Genetics Services, Murdoch Children's Research Institute, Melbourne, VIC, Australia

15. Department of Clinical Genetics, Liverpool Hospital, Sydney, NSW, Australia

16. Division of Medical Genetics, Department of Pediatrics, University of Utah School of Medicine, Salt Lake City, UT, USA

17. Center for Genetic Medicine Research, Children's National Research Institute, Washington, DC, USA

18. Division of Genetics and Metabolism, Children's National Hospital, Washington, DC, USA

19. Department of Pediatrics, Stanford University School of Medicine, Stanford, CA, USA

20. GREGoR Stanford Site, Stanford University School of Medicine, Stanford, CA, USA

21. Center for Undiagnosed Diseases, Stanford University School of Medicine, Stanford, CA, USA

22. Oxford Centre for Genomic Medicine, Oxford University Hospitals NHS Foundation Trust, Oxford, UK

23. Department of Paediatrics, University of Melbourne, Melbourne, VIC, Australia

24. Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA

25. Department of Metabolic Medicine, Queensland Children's Hospital, Brisbane, QLD, Australia

26. Faculty of Medicine, University of Queensland, Brisbane, QLD, Australia

27. School of Medicine, Griffith university, Gold Coast, QLD, Australia

28. Murdoch Children's Research Institute, Melbourne, VIC, Australia

29. Neuroscience Research Australia, Sydney, NSW, Australia

30. Prince of Wales Clinical School, Faculty of Medicine, University of New South Wales, Sydney, NSW, Australia

31. Department of Pediatrics, Children's Hospital Colorado, Aurora, CO, USA

32. University of Colorado School of Medicine, University of Colorado, Aurora, CO, USA

33. South West Thames Centre for Genomics, St George’s University Hospitals NHS Foundation Trust, London, UK

34. New South Wales Health Pathology Randwick Genomics, Prince of Wales Hospital, Sydney, NSW, Australia

35. Discipline of Paediatrics and Child Health, Faculty of Medicine and Health, University of New South Wales, Sydney, NSW, Australia

36. Centre for Clinical Genetics, Sydney Children's Hospitals Network, Randwick, NSW, Australia

37. Genomics and Inherited Disease Program, Garvan Institute of Medical Research, Darlinghurst, NSW, Australia

38. Division of Medical Genetics & Genomic Medicine, Vanderbilt University Medical Center, Nashville, TN, USA

39. Manton Center for Orphan Disease Research, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA

40. MRC Weatherall Institute of Molecular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, UK

41. NIHR Biomedical Research Centre, Oxford, UK

42. Institute of Medical Genetics and Applied Genomics, University of Tübingen, Tübingen, Germany

43. Center for Rare Diseases Tübingen, University of Tübingen, Tübingen, Germany

44. Liverpool Centre for Genomic Medicine, Liverpool Women’s Hospital, Crown Street, Liverpool, UK

45. Institute of Human Genetics, University Hospital Essen, University Duisburg-Essen, Essen, Germany

46. Wessex Clinical Genetics Service, University Hospital Southampton NHS Trust, Southampton, UK

47. Department of Human Genetics and Genomic Medicine, Southampton University, Southampton, UK

48. Univ Rouen Normandie, Inserm U1245 and CHU Rouen, Department of Genetics and reference center for developmental disorders, F-76000, Rouen, France

49. Royal Children’s Hospital, Melbourne, VIC, Australia

50. Paediatric and Reproductive Genetics Unit, South Australian Clinical Genetics Service, Women's and Children's Hospital, North Adelaide, SA, Australia

51. Repromed, Dulwich, SA, Australia

52. Bruce Lefroy Centre, Murdoch Children's Research Institute, Melbourne, VIC, Australia

53. Department of Clinical Genetics, Sydney Children's Hospitals Network Westmead, Sydney, NSW, Australia

54. Specialty of Genomic Medicine, University of Sydney, Sydney, NSW, Australia

55. Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA

56. Department of Medicine - Cardiovascular Medicine, Stanford University School of Medicine, Stanford, CA, USA

57. Manchester Centre for Genomic Medicine, St. Mary's Hospital, Manchester University NHS Foundation Trust, Health Innovation Manchester, Manchester, UK

58. Department of Pathology, Department of Genetics, Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA, USA

59. UCL Institute of Ophthalmology, London, UK

60. The Francis Crick Institute, London, UK

61. Moorfields Eye Hospital NHS Foundation Trust, London, UK

62. Service de Neurologie Pédiatrique, Cliniques Universitaires Saint-Luc, UCLouvain, B-1200, Brussels, Belgium

63. Institut des Maladies Rares, Cliniques Universitaires Saint-Luc, UCLouvain, B-1200, Brussels, Belgium

64. Nottingham University Hospitals NHS Trust, Nottingham, UK

65. Institute for Clinical and Translational Research, University of California, Irvine, CA, USA

66. Department of Clinical Genetics, The Children's Hospital at Westmead, Westmead, NSW, Australia

67. Center for Genomic Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

68. Center for Human Genetics, Cliniques universitaires Saint-Luc, Université catholique de Louvain, Brussels, Belgium

69. Centre for Population Genomics, Garvan Institute of Medical Research and UNSW Sydney, Sydney, NSW, Australia

70. Centre for Population Genomics, Murdoch Children’s Research Institute, Melbourne, VIC, Australia

71. Department of Clinical Genetics, Peninsula Regional Clinical Genetics Service, Royal Devon University Hospital, Exeter, UK

72. Department of Clinical and Experimental Epilepsy, UCL Queen Square Institute of Neurology, London, UK

73. UK and Chalfont Centre for Epilepsy, Chalfont St Peter, Buckinghamshire, UK

74. School of Women's and Children's Health, University of New South Wales, Sydney, NSW, Australia

75. Medical Genetics, University of Leicester, Leicester Royal Infirmary, Leicester, UK

76. Institute for Clinical and Translational Science, University of California, Irvine, CA, USA

77. Division of Newborn Medicine, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA

78. Department of Clinical and Biomedical Sciences, University of Exeter, Exeter, UK

79. Department of Neurology, University of California, Irvine, CA, USA

80. North West Thames Regional Genetics Service, Northwick Park & St Mark’s Hospitals, London, UK

81. Department of Psychiatry, Langley Porter Psychiatric Institute, UCSF Weill Institute for Neurosciences, University of California, San Francisco, CA, USA

82. Department of Biochemistry, University of Oxford, Oxford, UK

83. School of Human Development and Health, Faculty of Medicine, University of Southampton, Southampton, UK

84. National Institute for Health Research (NIHR) Southampton Biomedical Research Centre, University Hospital Southampton National Health Service (NHS) Foundation Trust, Southampton, UK

85. Human Genetics Centre of Excellence, Novo Nordisk Research Centre, Oxford, UK

\*contributed equally

Correspondence should be addressed to Nicola Whiffin (nwhiffin@well.ox.ac.uk)

**Abstract**

Around 60% of individuals with neurodevelopmental disorders (NDD) remain undiagnosed after comprehensive genetic testing, primarily of protein-coding genes[1](https://paperpile.com/c/8E4RdZ/8P6nU). Large genome-sequenced cohorts are improving our ability to discover new diagnoses in the non-coding genome. Here, we identify the non-coding RNA *RNU4-2* as a syndromic NDD gene. *RNU4-2* encodes the U4 small nuclear RNA (snRNA), which is a critical component of the U4/U6.U5 tri-snRNP complex of the major spliceosome[2](https://paperpile.com/c/8E4RdZ/jZgUk). We identify an 18 bp region of *RNU4-2* mapping to two structural elements in the U4/U6 snRNA duplex (the T-loop and Stem III) that is severely depleted of variation in the general population, but in which we identify heterozygous variants in 115 individuals with NDD. Most individuals (77.4%) have the same highly recurrent single base insertion (n.64\_65insT). In 54 individuals where it could be determined, the *de novo* variants were all on the maternal allele. We demonstrate that *RNU4-2* is highly expressed in the developing human brain, in contrast to *RNU4-1* and other U4 homologs. Using RNA-sequencing, we show how 5’ splice site usage is systematically disrupted in individuals with *RNU4-2* variants, consistent with the known role of this region during spliceosome activation. Finally, we estimate that variants in this 18 bp region explain 0.4% of individuals with NDD. This work underscores the importance of non-coding genes in rare disorders and will provide a diagnosis to thousands of individuals with NDD worldwide.

**Main**

Despite increasingly powerful genomic and analytic approaches for the diagnosis of rare developmental disorders, currently ~60% of individuals remain without an identified genetic diagnosis after genomic testing with current methods[1](https://paperpile.com/c/8E4RdZ/8P6nU). To date, the overwhelming majority of known disease-causing variants are in the ~1.5% of the genome that directly encodes proteins[3](https://paperpile.com/c/8E4RdZ/IPXd4). In contrast, the non-coding genome (that makes up the remaining 98.5%) has been relatively unexplored, especially regions far from protein-coding genes. Large-scale, systematic application of genome sequencing to clinical populations has increasingly enabled investigation of the contribution of variants in non-coding regions to genetic disorders[4](https://paperpile.com/c/8E4RdZ/AXc1Y).

Non-coding RNAs, which comprise 37.4% of processed exonic RNA sequence in humans[5](https://paperpile.com/c/8E4RdZ/HSATo), include important regulators of biological processes with diverse roles across cells and tissues[6](https://paperpile.com/c/8E4RdZ/gWPQM). Small nuclear RNAs (snRNAs) are a subcategory of non-coding RNAs that are key components of the spliceosome[7](https://paperpile.com/c/8E4RdZ/DwNr1). snRNAs complex with a multitude of proteins and other snRNA species in small nuclear ribonucleoprotein (snRNP) complexes to mediate the removal of introns from pre-mRNA transcripts[8](https://paperpile.com/c/8E4RdZ/GvKES). Many spliceosome components have a demonstrated role in human disorders, including two snRNA components of the minor spliceosome: *RNU12* variants cause autosomal recessive early-onset cerebellar ataxia[9](https://paperpile.com/c/8E4RdZ/Z8w7o), while *RNU4ATAC* variants cause an autosomal recessive multisystem congenital disorder including microcephaly, growth retardation, and developmental delay (eponyms include Taybi Linder[10](https://paperpile.com/c/8E4RdZ/wKJtA), Lowry-Wood[11](https://paperpile.com/c/8E4RdZ/pK8Sz) and Roifman syndromes[12](https://paperpile.com/c/8E4RdZ/aL1Zw)).

Here, we identify variants in *RNU4-2*, which encodes the U4 snRNA component of the major spliceosome, in an autosomal dominant disorder. Using a cohort of 8,841 probands with genetically undiagnosed NDD in the Genomics England 100,000 genomes project (GEL)[4](https://paperpile.com/c/8E4RdZ/AXc1Y), we identify variants in a critical 18 base-pair (bp) region in the centre of *RNU4-2* associated with a severe neurodevelopmental phenotype and estimate that variants in this region explain ~0.4% of individuals with neurodevelopmental disorders (NDD). We demonstrate that variants in this region are severely depleted from large population datasets. We show that NDD variants map to critical structural elements in the U4/U6 complex that are important to correctly position U6 ACAGAGA to receive the 5’ splice-site during initial spliceosome activation, and detail the expression of *RNU4-2* through brain development.

*A highly recurrent insertion in NDD*

We identified a highly recurrent single base insertion (GRCh38:chr12:120,291,839:T:TA; n.64\_65insT) in *RNU4-2* in GEL[1](https://paperpile.com/c/8E4RdZ/8P6nU). This variant was initially identified as arising *de novo* in 38 probands recruited for genome sequencing with their unaffected parents[13](https://paperpile.com/c/8E4RdZ/ks6Bf). Extending the search to include probands without data for both parents in the full GEL cohort, we identified an additional eight individuals with the n.64\_65insT variant; in all eight, the detectable inheritance is consistent with the variant having arisen *de novo* (i.e. where a single parent sample was available the variant was not detected in it). All of the 46 individuals with the variant have undiagnosed NDD (categorised as global developmental delay, intellectual disability, and/or autism spectrum disorder), corresponding to 0.52% of 8,841 probands with currently undiagnosed NDD in GEL. The n.64\_65insT variant is not found in any of 3,408 NDD probands with an existing genetic diagnosis, 21,817 probands with non-NDD phenotypes, or in 33,122 unaffected individuals. Individuals with the variant are significantly enriched for global developmental delay (n=37; OR=3.56; Fisher’s *P*=2.75x10-4), delayed gross motor development (n=26; OR=2.55; *P*=1.64x10-3), microcephaly (n=26; OR=6.62; *P*=7.87x10-10), delayed fine motor development (n=24; OR=2.61; *P*=1.69x10-3), hypotonia (n=18; OR=3.60; *P*=7.09x10-5), short stature (n=15; OR=3.54; *P*=2.17x10-4), drooling (n=7; OR=19.2; *P*=2.83x10-7), and absent speech (n=6; OR=6.23; *P*=7.45x10-4) compared to all other probands with NDD in GEL (n=12,203; diagnosed and undiagnosed) (**Extended Data** **Figure 1**).

The n.64\_65insT variant is not found in 76,215 genome-sequenced individuals in gnomADv4.0[14](https://paperpile.com/c/8E4RdZ/1g5QU), or in 245,400 individuals in the All of Us dataset[15](https://paperpile.com/c/8E4RdZ/a3huR). It is seen in a single individual in the UK Biobank[16](https://paperpile.com/c/8E4RdZ/lggip) (allele frequency=1.02x10-6) with a variant allele balance consistent with a true variant (23 reference and 18 [44%] alternate reads). This individual has an ICD-10 code for ‘personal history of disease of the nervous system and sense organs’ but no further phenotype data to assess a potential NDD diagnosis (**Supplementary Table 1**).

Given the high occurrence rate of this recurrent insertion, we wanted to rule out that it is a sequencing or mapping error, despite the overwhelming evidence of phenotype enrichment. Notably, the variant is a single A insertion after a run of four Ts, ruling out the most common cause of sequencing error for indels, polymerase slippage in homopolymer repeats. The variant calls were all high quality based on both analysis of quality metrics (**Supplementary Figure 1**) and manual inspection on IGV (**Supplementary Figure 2**). The genomic region surrounding the insertion and *RNU4-2* maps uniquely to a single region of the genome with short-read sequencing in GRCh38 and T2T CHM13v2.0/hs1. Finally, sequencing reads aligned to *RNU4-2* map with good quality (average 96 reads with mapping quality scores >20; **Supplementary Figure 3**).

*n.64\_65insT is in a highly constrained region*

The recurrent n.64\_65insT variant resides within the central region of *RNU4-2*, towards the 5’ end of an 18 bp region which is depleted of variants in population datasets compared with the rest of the gene (26% of all possible SNVs observed in UK Biobank compared to a median of 78% across the rest of the gene; **Figure 1a; Extended Data Figure 2a**). Based on the population variant data, we defined a critical, highly constrained region as chr12:120,291,825-120,291,842. We refer to this as the ‘critical region’ throughout the rest of the manuscript.

We searched for variants across this region in GEL, and also in additional cohorts containing undiagnosed individuals with NDD (see **methods**). In total, we identified 115 individuals with variants across this region, including 61 individuals in GEL (60 probands and one additional sibling) and 54 from additional cohorts (**Figure 1b**; **Extended Data Table 1**). For 86 of the 115 individuals, sequencing data for both parents were available to confirm the variants had arisen *de novo*. Where possible, we used nearby variants to determine the parental allele of origin of the variants. For 54 individuals where this could be confidently resolved (46 with n.64\_65insT; three with other insertion variants; five with SNVs), all 54 were present on the maternal allele. In one individual the n.65A>G variant appeared to be mosaic in the mother (54 reference and 8 alternate reads) and in another a SNV was maternally inherited (n.76C>T). This analysis also enabled us to determine likely *de novo* occurrence for five additional individuals where only one parent was sequenced. Sanger sequencing was used to confirm the presence of the variant in eight individuals with the n.64\_65insT variant. For seven of the eight, absence from both parents was also confirmed. In the eighth, the variant was confirmed as absent from the single available parent. In three families, the n.64\_65insT variant was identified as occurring *de novo* in both short and long read trio sequencing.

The vast majority of the 115 individuals have the initial n.64\_65insT variant (n=89; 77.4%). Five of the 11 additional variants are also single base insertions, including n.77\_78insT (GRCh38:chr12:120,291,826:T:TA), which is seen in six individuals, two of whom are affected siblings. Single base insertion variants in this region are strongly enriched in individuals with NDD 54/8,841 (0.61%) GEL undiagnosed NDD probands (55/10,388 individuals) have single base insertions compared to 2/490,132 individuals in the UK Biobank (OR=1,531; 95%CI:404,>16,384; Fisher’s *P*=3.3x10-92).

Aside from insertions, there is also a modest enrichment of SNVs in GEL NDD probands across the critical region (undiagnosed NDD: 6/8,841; UK Biobank 35/490,132; OR=9.51; 95%CI:3.27-22.8; Fisher’s *P*=8.16x10-5). We identified 15 individuals across cohorts with SNVs in this region (**Extended Data Table 1**; 10 confirmed *de novo*), all with phenotypes consistent with individuals with insertion variants. The identified SNVs cluster with the two regions harbouring insertion variants at the extreme ends of the 18 bp critical region (**Figure 1**). Conversely, SNVs in the central portion (particularly at nucleotides 71-74) are observed in both non-NDD individuals in GEL (n=2) and population controls, although all at low frequencies (**Extended Data Table 1**). Across the remainder of *RNU4-2* there is no significant enrichment of variants in undiagnosed NDD probands when compared with non-NDD probands (194/7,519 undiagnosed NDD; 521/19,428 non-NDD in GEL aggregated variant dataset[17](https://paperpile.com/c/8E4RdZ/8bPuw); OR=0.96; 95%CI:0.81-1.14; Fisher’s P=0.67).

In total, we identify variants in this 18 bp region in 115 individuals with NDD. This includes 60/8,841, or 0.68%, of all genetically undiagnosed NDD probands in GEL. In contrast, variants in this region are observed in 39/490,132 (0.008%) individuals in the UK Biobank (OR=85.8; 95%CI:56.4-131.6; Fisher’s P=1.84x10-78). As most individuals in GEL have had genetic testing prior to recruitment, we cannot estimate the overall prevalence of RNU4-2 variants in all cause NDD from this cohort. Instead, if we assume a diagnostic yield of 40% prior to defining our GEL undiagnosed NDD cohort, consistent with recent reports[1](https://paperpile.com/c/8E4RdZ/8P6nU), we estimate that variants in RNU4-2 could explain 0.4% of all NDD (calculated as 60 from an effective cohort size of 14,735 (8841 \* 1/0.6)).

U4 snRNA binds to U6 snRNA through extensive complementary base-pairing in the U4/U6.U5 tri-snRNP complex of the major spliceosome. Unwinding of U6 from U4 is essential to generate the catalytically active spliceosome[2](https://paperpile.com/c/8E4RdZ/jZgUk). The 18 bp critical region in *RNU4-2* maps to a region of U4 between the stem I region of complementary base-pairing to U6 and the 3’ stem-loop structures (nucleotides 62 to 79; **Figure 1c**). This region is known to be loaded into the active site of the *SNRNP200*-encoded BRR2 helicase, which mediates spliceosome activation by unwinding the U4/U6 duplex[2](https://paperpile.com/c/8E4RdZ/jZgUk). The highly recurrent n.64\_65insT variant is within a previously described ‘quasi pseudoknot’, or T-loop, structure[18](https://paperpile.com/c/8E4RdZ/OB72x) (**Figure 1d**). The region spanning nucleotides 76 to 78, where the recurrent n.77\_78insT variant resides, is involved in base-pairing with U6 in stem III[19](https://paperpile.com/c/8E4RdZ/RLNf5) (**Figure 1d**). Both of these regions are thought to stabilise U4/U6 pairing and accurately position the U6 ACAGAGA sequence to receive the 5’ splice site during initial spliceosome assembly, while U4 nucleotides in stem III are part of the loading site for BRR2. Nearby regions that are predicted to have important roles, such as the U4/U6 stem I binding region, are not enriched for variants in NDD probands.

*RNU4-2 variants disrupt 5’splice site use*

Given the importance of U4 snRNA in the spliceosome and previous observations of global disruption to splicing observed in other spliceosomopathies[20](https://paperpile.com/c/8E4RdZ/rf68c), we analysed RNA sequencing data from blood samples for five individuals from GEL. Three of these individuals have the highly recurrent n.64\_65insT variant, another has the other recurrent insertion, n.77\_78insT, and the final patient has an SNV (n.78A>C). The five individuals with *RNU4-2* variants had more abnormal splicing events than 378 controls with non-NDD phenotypes (mean 21.6 vs 4.5; Wilcoxon *P*=0.0126), but this was not significant after correcting for multiple testing. There was no difference in the number of genes that were significant outliers for expression (mean 1.8 vs 5.7; Wilcoxon *P*=0.94; **Extended Data Table 2**).

Consistent with the importance of the critical region in 5’ splice site recognition, the most pronounced difference was observed for abnormal splicing events corresponding to increased use of unannotated 5’ splice sites (mean 8.8 events in individuals with *RNU4-2* variants compared with 0.7 in both 378 unmatched controls and ten controls matched on genetic ancestry, sex and age at consent; Wilcoxon *P*=4.0x10-5 and *P*=5.7x10-3 respectively; **Figure 2a**; **Extended Data Table 2**). The individual with the SNV was not notably different from the four individuals with single base insertions (three significant events). Sequence motif analysis showed an increase in T at the +3 position and an increase in C at the +4 and +5 positions in the unannotated 5’ splice sites that were significantly increased in individuals with *RNU4-2* variants compared to decreased canonical sites (**Figure 2c**). These three positions of the 5’ splice site (+3, +4, and +5), which shift away from consensus in individuals with *RNU4-2* variants, pair directly with the U6 ACAGAGA region during spliceosome activation (**Figure 2d**).

Of all the detected abnormal splicing events, twelve of these were shared by two or more individuals with *RNU4-2* variants (**Extended Data Table 3**). Eleven of these twelve events (91.6%) corresponded to an increase in unannotated 5’ splice-site usage. None of these shared events were identified in any of the 378 controls. In contrast, when randomly sampling five control individuals across 10,000 permutations, the mean number of events shared by two or more individuals was 0.007, significantly less than the twelve in *RNU4-2* individuals (permutation *P*<1x10-4; **Figure 2b**). Five of the genes implicated in the twelve shared events are in the DDG2P database[23](https://paperpile.com/c/8E4RdZ/YKlCT) and/or were associated with NDD in a previous large-scale analysis[24](https://paperpile.com/c/8E4RdZ/YVmYD) (*NDUFV1*, *H2AC6*, *JMJD1C*, *MAP4K4*, and *SF1*). Ten of the twelve shared events impact the protein-coding sequence, with four predicted to cause a frameshift (**Extended Data Table 3**). Collectively, these results indicate a systematic shift in 5’ splice site usage in individuals with *RNU4-2* variants compared to controls. Future work should assess these patterns in a more disease-relevant tissue (e.g. brain) or in iPSC derived neuronal cells or organoid models. At present RNA from additional tissues from affected individuals is not available.

*Characterising the RNU4-2 NDD syndrome*

To characterise the phenotypic spectrum associated with variants in *RNU4-2*, we collected detailed phenotypic information for a subset of 49 individuals (42 with n.64\_65insT, three with other single base insertions, and four with SNVs; **Table 1**; **Supplementary Table 2**). Using these data, we find the *RNU4-2* syndromic NDD to be characterised by moderate to severe global developmental delay (four children with SNVs with moderate delay) and intellectual disability in all individuals. The majority (83%) achieved ambulation but at a delayed age (average 3.4 years, range 17 months to 7.5 years) with some noted to have a wide-based or ataxic gait. Only three individuals (two with an SNV) had fluent speech, some had a few words, and most were non-verbal. All but three were reported to have dysmorphic facial features. These facial features varied but consisted of a myopathic face with deep set eyes (some widely spaced and some narrowly spaced), epicanthus, wide nasal bridge, anteverted nares or underdeveloped ala nasi, large cupped ears (some posteriorly rotated), full cheeks, a distinctive mouth with full lips with downturned corners, high arched palate, and a large or protruding tongue (**Figure 3**). In comparison to the single base insertions, children with SNVs had fewer reports of severe global developmental delay (0/4 vs 34/40, Fisher’s *P*=0.0015).

Associated growth and neurodevelopmental phenotypes present in ≥75% of individuals include short stature, microcephaly (mostly congenital), speech abnormalities (mostly non-verbal), hypotonia, and seizures. Seizures had variable onset with a median of 3 years and ranging from the first year of life to 10 years of age (spanning infantile spasms, focal seizures and generalised tonic-clonic seizures, febrile seizures, and status epilepticus). Brain MRI showed a spectrum of abnormalities in the majority of individuals, most frequently reduced white matter volume, hypoplasia of the corpus callosum, ventriculomegaly, delayed myelination, and other non-specific abnormalities of the white matter. Involvement of multiple organ systems was reported for all individuals (with fewer systems reported as involved in individuals with SNVs), often including visual (optic nerve hypoplasia, cortical blindness, strabismus, nystagmus), gastrointestinal (constipation, reflux, feeding issues with need for a gastrostomy tube, poor growth), and bone/skeletal abnormalities (osteopenia, recurrent fractures, scoliosis, kyphosis, hip dysplasia), and in a lesser number of individuals, hearing, endocrine (hypothyroidism, growth hormone deficiency, pituitary abnormalities), limb, sleep, genitourinary, dental, cardiac, and cutaneous concerns (**Table 1; Supplementary Table 2**). No significant differences were noted in the presentation of male versus female individuals.

*Exome sequencing rarely finds n.64\_65insT*

The majority of individuals with NDD who undergo genetic testing currently have exome rather than genome sequencing. While *RNU4-2* is not directly captured by exome sequencing panels, there is a chance that off-target reads may map to the 18 bp critical region of *RNU4-2* and enable detection of variants in this region. To investigate this, we analysed individuals who are included in GEL and also have exome sequencing data in the Deciphering Developmental Disorders (DDD) cohort [1](https://paperpile.com/c/8E4RdZ/8P6nU). Across the DDD cohort, 3,408/13,450 individuals (25.3%) have at least one read mapping to the position of the n.64\_65insT variant (**Extended Data Figure 3**). The maximum number of mapping reads in any individual was five, which is below standard thresholds used to identify heterozygous variants. Of 1,755 individuals in both GEL and DDD, 22 have the n.64\_65insT variant (1.3%). Two of the 22 individuals (9.1%) each have a single read at the variant position in the exome sequencing data from DDD, but in each case it is identical to the reference sequence. The other 20 individuals have no reads mapping to *RNU4-2*. Nevertheless, others have reported successful identification and subsequent experimental validation of the n.64\_65insT variant identified initially only on one or two sequencing reads (public communication on X/Twitter with Steve Laurie and Konrad Platzer). These analyses suggest that while it is possible to identify individuals who may have variants in *RNU4-2* through exome sequencing data, the sensitivity of this approach is very low. Any variants identified through this approach will also need independent confirmation.

*Evaluating other snRNAs in NDD*

Given the identified importance of *RNU4-2* in NDD, we sought to determine whether other snRNA genes with no known association to NDD could also harbour novel diagnoses. We investigated 28 snRNA genes that are expressed in the brain, using multiple approaches (**Extended Data Table 4**). First, we tested for an overall enrichment of *de novo* variants in undiagnosed NDD probands compared to non-NDD probands across each snRNA with at least two identified *de novo* variants in probands with undiagnosed NDD (n=14) using the high-confidence *de novo* callset in GEL. Of the 13 genes other than *RNU4-2*, none showed a significant enrichment of *de novo* variants in undiagnosed NDD probands (all Fisher’s *P*>0.15).

Secondly, hypothesising that the burden of pathogenic variants in other snRNAs may be restricted to specific critical regions, as we see for *RNU4-2*, we used an 18 bp sliding window to identify snRNA regions that are depleted of variation in the UK Biobank compared to the overall variant burden across each gene. Notably, the regions with the highest depletion in *RNU4ATAC* correspond to two hotspots of pathogenic variants in ClinVar (chr2:121530923-121530946, chr2:121530984-121531007), however, the strength of the depletion in these regions is lower than in *RNU4-2 (minimum normalised proportion of observed -0.11 and -0.2 versus -0.5 for the depleted region in RNU4-2)*, consistent with lower selection acting on variants in *RNU4ATAC* that cause recessive disorders. We identified 14 regions in 13 unique snRNAs with a deviation from the median number of SNVs across the full gene of at least 20% (**Figure 4**; **Extended Data Table 5**). We repeated our *de novo* variant enrichment test in regions with at least two *de novo* variants in undiagnosed NDD probands (n=3). Only the conserved region in *RNU4-2* was significant (Fisher’s P=9.31x10-11; undiagnosed NDD probands n=37, non-NDD probands n=0; all other tests Fisher’s *P*>0.25).

Thirdly, we looked for recurrent *de novo* variants in undiagnosed GEL NDD probands that were absent from diagnosed NDD probands, non-NDD probands, and population controls. There are three *de novo* variants with an allele count ≥3 in the GEL undiagnosed NDD cohort, two in *RNU1-2* (chr1:16,895,992:C:T and chr1:16,896,002:A:G), and one in *RNVU1-7* (chr1:148,038,767:G:A). However, all three variants are observed at comparable frequencies in non-NDD probands and are also found at relatively high frequencies in population controls (all variants’ AF>0.5% in gnomAD 4.0).

Finally, given that variants in *RNU12* and *RNU4ATAC* are associated with recessive disease, we also tested for an enrichment of homozygous and compound heterozygous variants in undiagnosed NDD probands compared to non-NDD probands. We observed a nominal enrichment of variants in *RNU12* (11 probands with NDD vs 2 non-NDD probands; Fisher’s *P*=0.026), but this was not significant after correcting for multiple testing. We did not identify any significant associations across any other snRNA or when restricted to variants in our identified depleted regions (**Extended Data Table 4**; **Extended Data Table 5**).

*RNU4-2 is highly expressed in the brain*

Humans have multiple genes that encode the U4 snRNA, although only two of these, *RNU4-2* and *RNU4-1,* are highly expressed in the human brain (**Supplementary Table 3**). *RNU4-2* and *RNU4-1* are contiguous on chr12, both 141 bp long, and highly homologous, differing by four nucleotides (97.2% similarity). *RNU4-1* has a similar depletion of variants in population cohorts in the centre of the RNA (**Figure 4**), however, we do not observe an enrichment of variants in GEL in this central region (**Extended Data Figure 2b**). There is a variant equivalent to our highly recurrent variant in *RNU4-1* that is observed in six individuals in the UK Biobank dataset. There are no consistent phenotypes recorded in these six individuals (**Supplementary Table 1**).

To investigate the reason for variants in *RNU4-2*, but not *RNU4-1*, causing NDD, we analysed the expression of both *RNU4-1* and *RNU4-2* in the brain. First, we analysed the expression patterns of both genes across multiple developmental stages using bulk RNA-seq data from 176 human prefrontal cortex samples in BrainVar[25](https://paperpile.com/c/8E4RdZ/DXds5). The expression of *RNU4-1* and *RNU4-2* is tightly correlated (**Supplementary Figure 5**), however, *RNU4-2* is consistently expressed at a significantly higher level than *RNU4-1* (**Figure 5a**). Secondly, we assessed chromatin accessibility in the chromosome 12 locus containing both *RNU4-1* and *RNU4-2* using ATAC-seq data from two human prenatal prefrontal cortex samples. These data show a dramatic chromatin accessibility signal around *RNU4-2* and a much lower signal surrounding *RNU4-1*, again consistent with much higher expression of *RNU4-2* in the brain (**Figure 5b**). Overall, these data support the role of *RNU4-2* as the major U4 transcript in the brain.

**Discussion**

Here, we identified a highly constrained 18 bp region of *RNU4-2* in which variants cause a severe neurodevelopmental phenotype. We estimated that variants in this region could explain 0.4% of all NDD. As a comparison, the largest proportion of DD explained by a single gene in 13,449 individuals in the DDD cohort[1](https://paperpile.com/c/8E4RdZ/8P6nU) was 0.68% for *ANKRD11*, although we acknowledge that some genes and recognisable syndromes with longstanding associations (e.g., *MECP2*, *SCN1A*, *UBE3A*) will be depleted from this cohort. The proportion of NDD explained by variants in *RNU4-2* would be even higher if restricted to individuals with severe, syndromic NDD. This is consistent with the much lower rate of *RNU4-2* variants in cohorts recruited primarily for autism spectrum disorder (e.g., 3/7,149; 0.042% across SSC[26](https://paperpile.com/c/8E4RdZ/lhTFW), SPARK[27](https://paperpile.com/c/8E4RdZ/of9B4) and MSSNG[28](https://paperpile.com/c/8E4RdZ/gQh5i)).

Our findings underscore the value of large-scale genome sequencing datasets and the importance of considering variants outside of protein-coding regions. This region, despite being within a highly conserved non-coding exon, is not included in commercially available clinical exome sequencing capture, which primarily targets protein-coding exons[5](https://paperpile.com/c/8E4RdZ/HSATo). This discovery was empowered by the availability of increasingly large genome sequencing datasets from families affected by genetic disease around the world. Indeed, the scale and accessibility of the Genomics England dataset facilitated both the work reported here and a parallel discovery by an independent group[29](https://paperpile.com/c/8E4RdZ/OWuL). The detailed phenotypic characterisation included here will help prioritise individuals for targeted sequencing of *RNU4-2*.

For all individuals where we were able to confidently determine the parent of origin of the identified *RNU4-2* variants (n=54), the variants were observed to be on the maternal allele. This is in contrast to the well-established paternal bias observed for *de novo* small mutations[30](https://paperpile.com/c/8E4RdZ/i8xBu). The absence of any paternally derived variants in our cohort may be a consequence of negative selection in the male germline if *RNU4-2* plays an important role during spermatogenesis. It may also be a consequence of imprinting, for example if variants on a highly expressed paternal allele are embryonic lethal, while those on a weakly expressed maternal allele are survivable but result in NDD. Further work is needed to test these hypotheses.

The majority of individuals in our cohort have the highly recurrent n.64\_65insT variant. It is observed in 46 of 8,841 undiagnosed NDD probands in GEL. In contrast, the most recurrent protein-coding variant in a dataset of 31,058 individuals with developmental disorders[31](https://paperpile.com/c/8E4RdZ/dhk39) is observed in 36 individuals (0.12%; GRCh38:chr11:66211206:C:T; PACS1:p.Arg203Trp). The reasons for this high recurrence are unclear, but it could be driven by either a high endogenous mutation rate or positive selection in the germline. The latter has previously been described for so-called ‘selfish mutations’ associated with paternal age effects[32](https://paperpile.com/c/8E4RdZ/cot3U). One hypothesis is that germline selection is acting to increase the apparent frequency of the n.64\_65insT variant, for example through meiotic drive effects or by accelerating oocyte maturation[33](https://paperpile.com/c/8E4RdZ/di9LP). We do not see an association with maternal age for individuals with n.64\_65insT in GEL (mean 30.2 compared to 29.7 across other NDD probands; **Extended Data Figure 4**).

Alternatively, recurrence may be driven by a high mutation rate. This is consistent with the observed open chromatin state and very high expression of *RNU4-2* (**Figure 5**), as high levels of transcription are known to be correlated with increased mutation rate[34](https://paperpile.com/c/8E4RdZ/vhBKo). Hypermutability of short non-coding RNA genes, including snRNAs, has previously been documented[35,36](https://paperpile.com/c/8E4RdZ/cRlGc%2BQamvI). Consistent with this, a high variant density is observed across *RNU4-2* in the UK Biobank (**Extended Data Figure 5**). Despite the high number of variants in *RNU4-2* in UK Biobank, there are no individuals with homozygous variants and all observed variants are very rare (maximum allele frequency = 0.025%), consistent with strong negative selection acting on variants across *RNU4-2*. A high overall mutational burden does not, however, explain the high recurrence of this specific single base insertion. Local formation of secondary structure and base stacking is a known driver of biased small insertion mutations[37](https://paperpile.com/c/8E4RdZ/AX0VN). The high propensity of this region to form secondary structure when single-stranded may drive creation of this specific insertion. It is also possible that this variant is more compatible with live birth relative to other comparably recurrent mutations in the critical region.

The n.64\_65insT variant is one of six single base insertions that we observe in the 18 bp critical region in individuals with NDD, in a total of 100 individuals across cohorts. In contrast, single base insertions are very rare in population cohorts. While we do also observe some SNVs in this region in individuals with NDD, our initial data suggest these SNVs may result in a milder phenotype. However, given this observation is based on only four fully phenotyped individuals, it needs to be confirmed in larger cohorts. Saturation mutagenesis experiments that test the impact of different length insertions and deletions as well as SNVs across the length of *RNU4-2* will be important to understand the spectrum of deleterious mutations. The high proportion of single base insertion variants in individuals with NDD may indicate that steric conformational changes are needed to disrupt *RNU4-2* function. Specifically, insertion of a single base into the T-loop or stem III regions may destabilise the U4/U6 interaction and/or alter the positioning of the U6 ACAGAGA sequence and potentially disrupt the correct loading of the 5’ splice site into the fully assembled spliceosome. This hypothesised effect is supported by the observed systematic disruption to 5’ splice site usage observed in RNA-sequencing data from five individuals with *RNU4-2* variants. In particular, our observation that the +3, +4, and +5 positions of the 5’ splice site, which directly pair with the U6 ACAGAGA sequence, shift away from consensus at sites with increase usage in individuals with *RNU4-2* variants provides functional evidence that these variants disrupt accurate splice site recognition during spliceosome activation. Further, variants in U6 snRNA and protein components of the spliceosome situated in the proximity of our *RNU4-2* variants have recently been shown to alter 5’-splice site selection by changing the preference for nucleotides that pair with the U6 snRNA ACAGAGA, consistent with this region being involved in subtle regulation of alternative splicing[38,39](https://paperpile.com/c/8E4RdZ/18rVk%2B8OI5).

While two other snRNA genes, *RNU12* and *RNU4ATAC*, have been linked to different phenotypes, both are components of the minor spliceosome and are associated with recessive disorders. In contrast, here we implicate variants in a major spliceosome snRNA in a dominant disorder. We further explored whether other snRNA genes could explain undiagnosed cases. We did not find any other snRNAs, or constrained sub-regions of snRNAs, that were significantly enriched for either *de novo* variants or recessively inherited variants in NDD cases when compared with non-NDD probands. We note, however, that these tests have low power given the small size of the genes and regions (mean 139.5 bp and 28.1 bp, respectively). Variants in the regions we delineated should also be investigated in other disease cohorts.

In summary, we identify *RNU4-2* as a syndromic NDD gene, explaining ~0.4% of all individuals with NDD. Including *RNU4-2* in standard clinical workflows will end the diagnostic odyssey for thousands of NDD patients worldwide and knowledge of the gene responsible for disease will enable investigation of potential treatments for these individuals.

**Acknowledgements**

We would like to thank all of the incredible patients and families who contributed to our study. We would like to thank Peter O'Donovan, Mitra Sato, Mimoza Hoti, Joanne Yang, Cassandra Smith, and Nour Elkhateeb from Genomics England for their help with clinician collaboration and Airlock requests. We would like to thank Jeffrey Barrett for his statistical advice.

YC is supported by a studentship from Novo Nordisk. NW is supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (220134/Z/20/Z). GL was supported by the Fonds de recherche en santé du Québec (FRQS), VSG by NIAMS K23AR083505, and SLS by a fellowship from Manton Center for Orphan Disease Research at Boston Children’s Hospital. The research was supported by grant funding from Novo Nordisk (NW), the Rosetrees Trust (PGL19-2/10025 to NW), the Simons Foundation Autism Research Initiative (SFARI #736613, SJS), the NIMH (R01 MH129751 to SJS), the HDR-UK Molecules to Health Records Driver Programme (SJS), the Australian National Health and Medical Research Council (1164479, 1155244, GNT2001513), the Australian NHMRC Centre for Research Excellence in Neurocognitive Disorders (NHMRC-RG172296), the Australian Medical Research Future Fund (MRF2007677, GHFM76747), NHGRI (U01HG011762, U01HG011745, U24HG011746, UM1HG008900, U01HG011755, R21HG012397, and R01HG009141), NINDS (U01NS134358, U01NS106845, U54NS115052, 1U24NS131172), the Chan Zuckerberg Initiative Donor-Advised Fund at the Silicon Valley Community Foundation (funder DOI 10.13039/100014989) grants 2019-199278, 2020-224274, (<https://doi.org/10.37921/236582yuakxy>), the US Department of Defense Congressionally Directed Medical Research Programs (PR170396), the National Institute of Neurological Disorders and Stroke of the National Institutes of Health (U01HG007709, U01HG007942, and U01HG010217), the NIHR Moorfields Biomedical Research Centre, and the Clinical Translational Core of the Baylor College of Medicine IDDRC (P50HD103555) from the Eunice Kennedy Shriver National Institute of Child Health and Human Development. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Eunice Kennedy Shriver National Institute of Child Health and Human Development or the National Institutes of Health. The Rare Disease Flagship acknowledges financial support from the Royal Children’s Hospital Foundation, the Murdoch Children’s Research Institute and the Harbig Foundation. Massimo’s Mission acknowledges funding support from the Australian Government Department of Health and Aged Care (EPCD000034). Sequencing and analysis were supported by the Deutsche Forschungsgemeinschaft (DFG) Research Infrastructure West German Genome Center (project 407493903) as part of the Next Generation Sequencing Competence Network (project 423957469). Short-read genome sequencing was carried out at the production site Cologne (Cologne Center for Genomics; CCG). CD received the DFG 458099954 as part of the DFG Sequencing call #3. We also thank Sabine Kaya for technical assistance, Christopher Schröder for bioinformatic analysis. SMS is supported by the Epilepsy Society.

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 National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK and the Medical Research Council have also funded research infrastructure.

**Author contributions**

Y.C., R.D., H.C.K., A.L., S.L.S., S.W., J.L., G.L., A.C.M-G., V.S.G., J.M., J.M.E., E.D., E.N.D’S., and S.D., analysed data and contributed to the figures and tables in the manuscript. J.L.R., E.M-P., S.M.F., D.B., C.D., D.G.M., J.M.M.H., S.J.S., A.O’D-L., and N.W. collected data, provided funding, and supervised the work. All other authors provided clinical and/or genomic data and are listed alphabetically. Y.C. and N.W. wrote the manuscript with input from all the authors.

**Competing interests**

NW receives research funding from Novo Nordisk and has consulted for ArgoBio studio.

SJS receives research funding from BioMarin Pharmaceutical. AODL is on the scientific advisory board for Congenica, was a paid consultant for Tome Biosciences, Ono Pharma USA Inc., and currently for Addition Therapeutics, and received reagents from PacBio to support rare disease research. HLR has received support from Illumina and Microsoft to support rare disease gene discovery and diagnosis. MHW has consulted for Illumina and Sanofi and received speaking honoraria from Illumina and GeneDx. SBM is an advisor for BioMarin, Myome and Tenaya Therapeutics. SMS has received honoraria for educational events or advisory boards from Angelini Pharma, Biocodex, Eisai, Zogenix/UCB and institutional contributions for advisory boards, educational events or consultancy work from Eisai, Jazz/GW Pharma, Stoke Therapeutics, Takeda, UCB and Zogenix. The Department of Molecular and Human Genetics at Baylor College of Medicine receives revenue from clinical genetic testing completed at Baylor Genetics Laboratories. JMMH is a full-time employee of Novo Nordisk and holds shares in Novo Nordisk A/S. DGM is a paid consultant for GlaxoSmithKline, Insitro, and Overtone Therapeutics and receives research support from Microsoft. All other authors declare no competing interests.

**Data availability**

Research on the de-identified patient data used in this publication from the Genomics England 100,000 Genomes Project and the NHS GMS dataset can be carried out in the Genomics England Research Environment subject to a collaborative agreement that adheres to patient led governance. All interested readers will be able to access the data in the same manner that the authors accessed the data. For more information about accessing the data, interested readers may contact research-network@genomicsengland.co.uk or access the relevant information on the Genomics England website: <https://www.genomicsengland.co.uk/research>. Genomic and phenotypic data from the GREGoR consortium (including the RGP cohort) and the UDN are available via dbGaP accession numbers phs003047 and phs001232.v5.p2, respectively, with at least annual data releases. Access is managed by a data access committee designated by dbGaP and is based on intended use of the requester and allowed use of the data submitter as defined by consent codes. The BrainVar data are available through the PsychENCODE Knowledge Portal: syn21557948 on [Synapse.org](http://synapse.org) (<https://www.synapse.org/#!Synapse:syn4921369> ). Raw ATAC-seq and ChIP-seq data are available on dbGAP: accession phs002033.v1.p1.

**Code availability**

Analysis of the 100,000 genomes project and NHS GMS data was performed inside the Genomics England Research Environment. We are happy to share the location of all code to registered users. Code used for analyses outside of Genomics England is available at Github: <https://github.com/Computational-Rare-Disease-Genomics-WHG/RNU4-2> and <https://github.com/francois-lecoquierre/genomics_shortcuts/blob/main/find_RNU4-2_recurrent_variant.py> .

**Supplementary information**

Supplementary Figures 1-5 and Tables 1-3

**References**

1. [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).](http://paperpile.com/b/8E4RdZ/8P6nU)

2. [Nguyen, T. H. D. *et al.* The architecture of the spliceosomal U4/U6.U5 tri-snRNP. *Nature* **523**, 47–52 (2015).](http://paperpile.com/b/8E4RdZ/jZgUk)

3. [Ellingford, J. M. *et al.* Recommendations for clinical interpretation of variants found in non-coding regions of the genome. *Genome Med.* **14**, 73 (2022).](http://paperpile.com/b/8E4RdZ/IPXd4)

4. [100,000 Genomes Project Pilot Investigators *et al.* 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. *N. Engl. J. Med.* **385**, 1868–1880 (2021).](http://paperpile.com/b/8E4RdZ/AXc1Y)

5. [Aspden, J. L., Wallace, E. W. J. & Whiffin, N. Not all exons are protein coding: Addressing a common misconception. *Cell Genom* **3**, 100296 (2023).](http://paperpile.com/b/8E4RdZ/HSATo)

6. [Nemeth, K., Bayraktar, R., Ferracin, M. & Calin, G. A. Non-coding RNAs in disease: from mechanisms to therapeutics. *Nat. Rev. Genet.* **25**, 211–232 (2024).](http://paperpile.com/b/8E4RdZ/gWPQM)

7. [Will, C. L. & Lührmann, R. Spliceosome structure and function. *Cold Spring Harb. Perspect. Biol.* **3**, (2011).](http://paperpile.com/b/8E4RdZ/DwNr1)

8. [Guthrie, C. & Patterson, B. Spliceosomal snRNAs. *Annu. Rev. Genet.* **22**, 387–419 (1988).](http://paperpile.com/b/8E4RdZ/GvKES)

9. [Elsaid, M. F. *et al.* Mutation in noncoding RNA RNU12 causes early onset cerebellar ataxia. *Ann. Neurol.* **81**, 68–78 (2017).](http://paperpile.com/b/8E4RdZ/Z8w7o)

10. [Edery, P. *et al.* Association of TALS developmental disorder with defect in minor splicing component U4atac snRNA. *Science* **332**, 240–243 (2011).](http://paperpile.com/b/8E4RdZ/wKJtA)

11. [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).](http://paperpile.com/b/8E4RdZ/pK8Sz)

12. [Merico, D. *et al.* Compound heterozygous mutations in the noncoding RNU4ATAC cause Roifman Syndrome by disrupting minor intron splicing. *Nat. Commun.* **6**, 8718 (2015).](http://paperpile.com/b/8E4RdZ/aL1Zw)

13. [De novo variant research dataset - Genomics England Research Environment User Guide.](http://paperpile.com/b/8E4RdZ/ks6Bf) <https://re-docs.genomicsengland.co.uk/de_novo_data/>[.](http://paperpile.com/b/8E4RdZ/ks6Bf)

14. [Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).](http://paperpile.com/b/8E4RdZ/1g5QU)

15. [All of Us Research Program Investigators *et al.* The ‘All of Us’ Research Program. *N. Engl. J. Med.* **381**, 668–676 (2019).](http://paperpile.com/b/8E4RdZ/a3huR)

16. [Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).](http://paperpile.com/b/8E4RdZ/lggip)

17. [Aggregated variant calls (AggV2) - genomics England research environment user guide.](http://paperpile.com/b/8E4RdZ/8bPuw) <https://re-docs.genomicsengland.co.uk/aggv2/>[.](http://paperpile.com/b/8E4RdZ/8bPuw)

18. [Charenton, C., Wilkinson, M. E. & Nagai, K. Mechanism of 5’ splice site transfer for human spliceosome activation. *Science* **364**, 362–367 (2019).](http://paperpile.com/b/8E4RdZ/OB72x)

19. [Wilkinson, M. E., Charenton, C. & Nagai, K. RNA Splicing by the Spliceosome. *Annu. Rev. Biochem.* **89**, 359–388 (2020).](http://paperpile.com/b/8E4RdZ/RLNf5)

20. [Griffin, C. & Saint-Jeannet, J.-P. Spliceosomopathies: Diseases and mechanisms. *Dev. Dyn.* **249**, 1038–1046 (2020).](http://paperpile.com/b/8E4RdZ/rf68c)

21. [Scheller, I. F., Lutz, K., Mertes, C., Yépez, V. A. & Gagneur, J. Improved detection of aberrant splicing with FRASER 2.0 and the intron Jaccard index. *Am. J. Hum. Genet.* **110**, 2056–2067 (2023).](http://paperpile.com/b/8E4RdZ/tkSKk)

22. [Brechtmann, F. *et al.* OUTRIDER: A Statistical Method for Detecting Aberrantly Expressed Genes in RNA Sequencing Data. *Am. J. Hum. Genet.* **103**, 907–917 (2018).](http://paperpile.com/b/8E4RdZ/XKDW2)

23. [Thormann, A. *et al.* Flexible and scalable diagnostic filtering of genomic variants using G2P with Ensembl VEP. *Nat. Commun.* **10**, 2373 (2019).](http://paperpile.com/b/8E4RdZ/YKlCT)

24. [Fu, J. M. *et al.* Rare coding variation provides insight into the genetic architecture and phenotypic context of autism. *Nat. Genet.* **54**, 1320–1331 (2022).](http://paperpile.com/b/8E4RdZ/YVmYD)

25. [Werling, D. M. *et al.* Whole-Genome and RNA Sequencing Reveal Variation and Transcriptomic Coordination in the Developing Human Prefrontal Cortex. *Cell Rep.* **31**, 107489 (2020).](http://paperpile.com/b/8E4RdZ/DXds5)

26. [Fischbach, G. D. & Lord, C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* **68**, 192–195 (2010).](http://paperpile.com/b/8E4RdZ/lhTFW)

27. [SPARK Consortium. Electronic address: pfeliciano@simonsfoundation.org & SPARK Consortium. SPARK: A US Cohort of 50,000 Families to Accelerate Autism Research. *Neuron* **97**, 488–493 (2018).](http://paperpile.com/b/8E4RdZ/of9B4)

28. [C Yuen, R. K. *et al.* Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nat. Neurosci.* **20**, 602–611 (2017).](http://paperpile.com/b/8E4RdZ/gQh5i)

29. [Greene, D. *et al.* Mutations in the U4 snRNA gene RNU4-2 cause one of the most prevalent monogenic neurodevelopmental disorders. *Nat. Med.* (2024) doi:](http://paperpile.com/b/8E4RdZ/OWuL)[10.1038/s41591-024-03085-5](http://dx.doi.org/10.1038/s41591-024-03085-5)[.](http://paperpile.com/b/8E4RdZ/OWuL)

30. [Jónsson, H. *et al.* Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**, 519–522 (2017).](http://paperpile.com/b/8E4RdZ/i8xBu)

31. [Kaplanis, J. *et al.* Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature* **586**, 757–762 (2020).](http://paperpile.com/b/8E4RdZ/dhk39)

32. [Wood, K. A. & Goriely, A. The impact of paternal age on new mutations and disease in the next generation. *Fertil. Steril.* **118**, 1001–1012 (2022).](http://paperpile.com/b/8E4RdZ/cot3U)

33. [Clark, F. E. & Akera, T. Unravelling the mystery of female meiotic drive: where we are. *Open Biol.* **11**, 210074 (2021).](http://paperpile.com/b/8E4RdZ/di9LP)

34. [Jinks-Robertson, S. & Bhagwat, A. S. Transcription-associated mutagenesis. *Annu. Rev. Genet.* **48**, 341–359 (2014).](http://paperpile.com/b/8E4RdZ/vhBKo)

35. [Seplyarskiy, V. *et al.* A mutation rate model at the basepair resolution identifies the mutagenic effect of polymerase III transcription. *Nat. Genet.* **55**, 2235–2242 (2023).](http://paperpile.com/b/8E4RdZ/cRlGc)

36. [Thornlow, B. P. *et al.* Transfer RNA genes experience exceptionally elevated mutation rates. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 8996–9001 (2018).](http://paperpile.com/b/8E4RdZ/QamvI)

37. [Pray, L. DNA replication and causes of mutation. *Nature education* **1**, 214 (2008).](http://paperpile.com/b/8E4RdZ/AX0VN)

38. [Zahler, A. M. *et al.* SNRP-27, the C. elegans homolog of the tri-snRNP 27K protein, has a role in 5’ splice site positioning in the spliceosome. *RNA* **24**, 1314–1325 (2018).](http://paperpile.com/b/8E4RdZ/18rVk)

39. [Parker, M. T. *et al.* m6A modification of U6 snRNA modulates usage of two major classes of pre-mRNA 5’ splice site. *Elife* **11**, (2022).](http://paperpile.com/b/8E4RdZ/8OI5)

40. Wersig, C. & Bindereif, A. Conserved domains of human U4 snRNA required for snRNP and spliceosome assembly. Nucleic Acids Res. 18, 6223–6229 (1990).

41. [Burley, S. K. *et al.* RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids Res.* **51**, D488–D508 (2023).](http://paperpile.com/b/8E4RdZ/IUWbY)

42. [Morales, J. *et al.* A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature* **604**, 310–315 (2022).](http://paperpile.com/b/8E4RdZ/rOqQB)

43. [Kang, H. J. *et al.* Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483–489 (2011).](http://paperpile.com/b/8E4RdZ/oQQ8j)

**Figure legends**

**Figure 1: A highly structured 18 bp region of RNU4-2 that is critical for BRR2 helicase activity is enriched for variants in NDD and depleted in population cohorts**. (a) Allele counts of *de novo* variants in 8,841 undiagnosed NDD probands in GEL (top; teal) and the UK Biobank cohort (bottom; grey) across *RNU4-2*. The 18 bp critical region, which is depleted of variants in the UK Biobank, is marked by a horizontal bar at the top of the plot. (b) Allele counts of additional variants identified in individuals with NDD in the critical 18 bp region. This includes 16 individuals with 7 variants without sequencing data for both parents in GEL and variants identified in individuals from the following additional cohorts (see **methods**): NHS GMS (n=19); MSSNG (n=2); SSC (n=1); GREGoR (n=10); Undiagnosed Diseases Network (UDN; n=6); from personal communication/Matchmaker Exchange (n=16). (c) Schematic of U4 (teal) binding to U6 snRNA (grey). The 18 bp critical region is underlined. Nucleotides 142 to 145 of U4 (in blue) are not within the GENCODE transcript of *RNU4-2* but are included in previous figures of the U4/U6 duplex in the literature on which this depiction is based[40](https://paperpile.com/c/8E4RdZ/FwcL) and are present in the RNA sequencing reads from human prefrontal cortex in BrainVar. (d) The structure of U4 and U6 snRNAs resolved by cryoEM[18](https://paperpile.com/c/8E4RdZ/OB72x). U4 residues in the critical region are labelled with the reference nucleotide and numbered according to the position along the RNA (e.g. U62 indicates a uracil residue in the reference sequence at position 62). Created using publicly accessible coordinated from the RCSB Protein Data Bank[41](https://paperpile.com/c/8E4RdZ/IUWbY) (structure 6QW6). In both (c) and (d) single base insertions identified in individuals with NDD are shown by black arrows and positions of SNVs by orange nucleotides.

**Figure 2: Individuals with *RNU4-2* variants have systematic changes in 5’ splice site usage.** (a) Boxplots of the number of abnormal splicing events (detected by FRASER2[21](https://paperpile.com/c/8E4RdZ/tkSKk)) at unannotated 5’ splice sites. The individuals with *RNU4-2* variants (n=5 individuals) have significantly more outlier events than both controls with non-NDD phenotypes (n=378 individuals) and controls matched on genetic ancestry, sex, and age at consent (n=10 individuals, 2 per case; Wilcoxon *P*=4.0x10-5 (W test statistic=1,766) and *P*=5.7x10-3 (W test statistic=45.5) respectively). Centre line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; maxima and minima represented as points. (b) The distribution of the number of abnormal splicing events at unannotated 5’ splice sites shared between two or more out of five randomly selected control individuals over 10,000 permutations (grey histogram). The number of shared events in individuals with *RNU4-2* variants is indicated as a dotted teal vertical line (n=11). (c) DNA sequence motifs around 5’ splice sites with increased and decreased usage in individuals with *RNU4-2* variants. Each plot shows the proportion of sites with each base at each position. 5’ splice sites with increased usage (top) have an increase in T at the +3 position (8/19 vs 0/36; Fisher’s *P*=6.2x10-5; OR=Inf; 95%CI:5.92-Inf) and an increase in C at the +4 (4/19 vs 0/36; Fisher’s *P*=0.011; OR=Inf; 95%CI:) and +5 (6/19 vs 1/36; Fisher’s *P*=0.0051; OR=15.3; 95%CI:2.09-Inf)) positions compared to decreased 5’ splice sites (bottom). The consensus sequence at 5’ splice sites in MANE transcripts[42](https://paperpile.com/c/8E4RdZ/rOqQB) is shown in **Supplementary Figure 4**. (d) The structure of the U6 snRNA paired with the 5’ splice site after 5’ splice site transfer. The three bases of the U6 ACAGAGA that directly pair with the 5’ splice site are shown in pink. The paired positions of the 5’ splice site (5’SS) are shown in green (A+3 and A+4) and yellow (G+5). Statistical tests in (a) and (c) are one-sided with unadjusted *P*-values.

**Figure 3. Clinical photographs showing facial features of affected individuals with variants in *RNU4-2*.** All individuals shown have the n.64\_65insT variant, except for Individual 44 in panel O (n.68\_69insA), Individual 45 in panel P (n.64\_65insG), and Individual 48 in panel Q (n.76C>T). (a) Individual 1 at 12 years old. (b) Individual 4 at 9 years old. (c) Individual 7 at 13 years old. (d) Individual 15 at 8 years old. (e) Individual 21 at 3.5 years old. (f) Individual 22 at 8 years old. (g) Individual 23 at 13 years old. (h) Individual 28 at 5 years old (left) and 9 years old (right). (i) individual 32 at 3 years old (left) and 12 years old (right). (j) Individual 36 at 11 months old (left) and 8 years old (right). (k) Individual 37 at 22 months old (left) and 16 years old (right). (l) Individual 38 at 2.5 years old (left) and 10 years old (right). (m) Individual 39 at 2 years old (left) and 12 years old (right). (n) Individual 43 at 8 years old (left) and 12 years old (right). (o) Individual 44 at 6 years old (left) and 19 years old (right). (p) Individual 45 at 6 years old (left and center) and 27 years old (right). (q) Individual 48 at 22 months old.

**Figure 4: Multiple snRNA genes have regions that are depleted of variation in the population.** The proportion of observed SNVs in 490,640 genome sequenced individuals in the UK Biobank, in sliding windows of 18 bp across each snRNA gene, normalised to the median value for each gene.

**Figure 5: *RNU4-2* is more highly expressed than *RNU4-1* in the prefrontal cortex.** (a) Levels of *RNU4-1* (grey) and *RNU4-2* (teal) expression in human dorsolateral prefrontal cortex at different developmental stages from BrainVar[25](https://paperpile.com/c/8E4RdZ/DXds5). Coloured lines correspond to the Loess smoothed average with the shaded regions representing 95% confidence intervals. Developmental stages are labelled with periods (1 to 12), spanning from embryonic development to late adulthood, that were defined previously[43](https://paperpile.com/c/8E4RdZ/oQQ8j). (b) ATAC-seq data from human prenatal prefrontal cortex shows substantially higher peaks of chromatin accessibility around *RNU4-2* than *RNU4-1*. Data for both 18 and 19 gestational weeks (GW) is shown to demonstrate replication.

|  |  |  |  |
| --- | --- | --- | --- |
| **Clinical feature** |  |  |  |
| Individuals (n) |  | 49 |  |
| Sex |  | 21 F, 28 M |  |
|  |  | **Median** | **Range** |
| Age at last evaluation (years) |  | 10  | 2 - 38 |
| Maternal age at birth (years)\* |  | 32 | 22 - 41 |
| Paternal age at birth (years)\*\* |  | 33 | 26 - 45 |
|  |  | **Count\*\*\*** | **Percentage** |
| Growth | IUGR | 8/45 | 18% |
| Short stature | 37/49 | 76% |
| Microcephaly | 37/48 | 77% |
| * congenital
 | 19/37 |  |
| * acquired
 | 9/37 |  |
| * not specified
 | 9/37 |  |
| Neurodevelopmental | GDD | 49/49 | 100% |
| * severe
 | 34/49 |  |
| * moderate
 | 10/49 |  |
| * not specified
 | 5/49 |  |
| Ambulatory (>5yo) | 30/36 | 83% |
| * abnormal gait
 | 7/30 |  |
| * not specified
 | 23/30 |  |
| Speech abnormality | 45/48 | 94% |
| * non-verbal
 | 35/45 |  |
| * few words
 | 10/45 |  |
| ID | 45/45 | 100% |
| Behavioural issues | 30/45 | 67% |
| ASD | 21/44 | 48% |
| Sleep issues | 15/32 | 47% |
| Hypotonia | 39/45 | 87% |
| Seizures | 37/48 | 77% |
| Abnormal brain MRI | 41/45 | 91% |
| Hearing | Hearing loss | 10/46\*\*\*\* | 22% |
| Vision | Vision issues | 38/48 | 79% |
| * Optic nerve hypoplasia
 | 8/37 | 22% |
| * Strabismus
 | 23/45 | 51% |
| * Nystagmus
 | 18/40 | 45% |
| Gastrointestinal | Constipation | 29/44 | 66% |
| GORD | 21/43 | 49% |
| Feeding difficulties | 32/42 | 76% |
| G-tube | 13/41 | 32% |
| Growth problems | 30/43 | 70% |
| Endocrine |  | 17/39 | 44% |
| Bone/skeletal |  | 27/42 | 64% |
| Limb |  | 23/42 | 55% |
| Genitourinary |  | 15/43 | 35% |
| Dental |  | 17/43 | 40% |
| Cardiac |  | 5/43 | 12% |
| Cutaneous |  | 25/44 | 57% |
| Dysmorphic facial features |  | 45/48 | 94% |

**Table 1: Clinical features of 49 individuals with *RNU4-2* variants.**

F, female; M, male; IUGR, intrauterine growth restriction; GDD, global developmental delay; ID, intellectual disability; ASD, autism spectrum disorder; MRI, magnetic resonance imaging; GORD, gastro-oesophageal reflux disease; G-tube, gastrostomy tube

\*maternal age available for 43/49 individuals

\*\*paternal age available for 41/49 individuals

\*\*\*denominator indicates the number of individuals for whom data were available

\*\*\*\*one individual has a dual diagnosis in *GJB2* which would account for the hearing loss

**Methods**

*Categorising participants in Genomics England*

We defined four groups of individuals in the Genomics England 100,000 genomes project v18 dataset. Individuals with NDD (n=13,812) were defined as those with human phenotype ontology (HPO)[44](https://paperpile.com/c/8E4RdZ/NsPCz) and/or International Classification of Diseases 10th Revision (ICD-10) codes[45](https://paperpile.com/c/8E4RdZ/Rj1JS) for global developmental delay (HP:0001263, HP:0012736, HP:0011344, HP:0011343, HP:0011342; ICD-10: R62, F80, F81, F82, F83, F88, F89), intellectual disability (HPO: HP:0001249, HP:0002187, HP:0010864, HP:0002342, HP:0001256, HP:0006887, HP:0006889; ICD-10: F70, F71, F72, F73, F78, F79), and/or autism (HPO: HP:0000717, HP:0000729, HP:0000753; ICD-10: F84), or who were recruited to GEL with a normalised specific disease of intellectual disability. NDD individuals were classified as diagnosed (n=3,424) if they were marked as solved or partially solved in the gmc\_exit\_questionnaire table or had an entry in the submitted\_diagnostic\_discovery table in GEL Labkey. The remaining 10,388 NDD individuals formed our undiagnosed NDD cohort. Of these, 8,841 are probands. We also identified 21,817 probands without NDD phenotypes (i.e. without the HPO and ICD10 codes detailed above) and 33,122 individuals reported to be unaffected. Our defined cohorts exclude anyone who has subsequently removed consent.

For the majority of our analyses, we used two previously defined datasets within GEL. First, a high-confidence set of *de novo* variants from 13,949 trios[13](https://paperpile.com/c/8E4RdZ/ks6Bf). As of 13 March 2024, this dataset includes 12,554 probands with consent: 5,426 probands with undiagnosed NDD, 2,352 with diagnosed NDD, and 4,776 non-NDD probands. *De novo* variants were filtered to those that pass the stringent\_filter. Second, an aggregated variant call set (aggV2)[17](https://paperpile.com/c/8E4RdZ/8bPuw) which contains 29,850 probands: 7,519 undiagnosed NDD, 2,903 diagnosed NDD, and 19,428 non-NDD.

*Identifying variants in population datasets*

We used data from gnomAD v4.0 (76,215 genome sequenced individuals)[14](https://paperpile.com/c/8E4RdZ/1g5QU), All of Us[15](https://paperpile.com/c/8E4RdZ/a3huR) (accessed via the publicly available data browser <https://databrowser.researchallofus.org/>; 245,400 genomes as of 28 March 2023) and the UK Biobank (490,640 genome sequenced individuals)[16](https://paperpile.com/c/8E4RdZ/lggip).

*Expanded NDD cohort and clinical data collection*

Clinical data were collected from research participants after obtaining written informed consent from the parents or legal guardians, with the study approved by the local regulatory authority. Samples were collected largely through personal communications (NW, AODL, DGM) as variants in this gene have not been prioritised in analysis. On entry into Matchmaker Exchange using the seqr[46](https://paperpile.com/c/8E4RdZ/60LWz) node, one match was made with GeneMatcher[47](https://paperpile.com/c/8E4RdZ/63CBT) (CD). NW reviewed the National Health Service Genome Medicine Service (NHS GMS; V3) dataset. Samples from NHS GMS were manually checked to remove duplicates with GEL. AODL and SS reviewed the Broad Center for Mendelian Genomics and the GREGoR consortium datasets and contacted the Undiagnosed Disease Network (UDN) through MTW. DGM contacted additional local collaborators. Clinical collaboration requests were submitted to GEL to contact recruiting clinicans and collect additional phenotypic information. Clinical data were collected and summarised for features seen across the cohort. Written consent was obtained to publish all photographs included in Figure 3.

We additionally searched 7,149 trios with autism spectrum disorder and 4,180 sibling control trios from three cohorts: Simons Simplex Collection (SSC; 2,383 cases; 1,938 controls)[26](https://paperpile.com/c/8E4RdZ/lhTFW), SPARK (3,144 cases; 2,190 controls)[27](https://paperpile.com/c/8E4RdZ/of9B4), and MSSNG (1,622 cases; 52 controls)[28](https://paperpile.com/c/8E4RdZ/gQh5i).

*Generating 1,000 random intergenic sequences*

Using the bedtools (v2.31.0) subtractBed function[48](https://paperpile.com/c/8E4RdZ/FUZK7) we retrieved regions on chromosome 12 that do not overlap with RefSeq transcripts aligned by NCBI. We further removed regions within 10 kbp of an annotated transcript and restricted the remaining regions to those at least 141 bp in length (n=611). We further removed regions overlapping the centromere. We then generated a set of 1,000 random sequences from each intergenic region and then randomly selected 1,000 non-overlapping regions from these.

*Identifying human snRNA genes*

We extracted genes with snRNA biotypes from Ensembl genome annotation v111. We filtered out known pseudogenes (i.e. with gene names marked with “P” or identified through manual curation). For each remaining gene, we used BrainVar[25](https://paperpile.com/c/8E4RdZ/DXds5) RNA-seq expression data to calculate the mean CPM value across the gene. We selected only genes with mean CPM value across all BrainVar samples >5, resulting in a dataset of 28 snRNA genes.

*Assessing variant depletion*

Given the high mutability of *RNU4-2* and other snRNA genes, coupled with strong selection pressures on variants, we did not think that conventional mutational models would be well calibrated to assess variant depletion. Instead, we devised a sliding window-based strategy to identify regions within snRNA genes that are relatively depleted of SNVs. We split genes into 18 bp sliding windows (chosen as it is the size of the region defined in *RNU4-2*) and tallied the number of SNVs observed in UK Biobank 500k genome sequencing data within that window, divided by the total number of possible SNVs (i.e. 18x3). The proportion of possible SNVs observed in each window was normalised to the median across all sliding windows in that gene (i.e. the per-gene median proportion observed was subtracted from each value). Depleted regions were defined as those spanning windows with a deviation from the per-gene median of at least 20%, i.e. normalised observed proportion of possible SNVs < -0.2. The same calculation was performed on 1,000 randomly selected 141 bp intergenic regions on chr12 (see above). A one-way approximative (Monte Carlo) Fisher-Pitman test was conducted to show the median observed proportion of possible SNVs was significantly higher for *RNU4-1* and *RNU4-2* compared to the distribution in the 1,000 random regions.

*RNA-sequencing of individuals with RNU4-2 variants*

Blood was collected from a subset of 100,000 Genomes Project probands in PaxGene tubes to preserve RNA at the time of recruitment. RNA was extracted, depleted of globin and ribosomal RNAs, and subjected to sequencing by Illumina using 100 bp paired-end reads, with a mean of 102M mapped reads per individual. Alignment was performed using Illumina’s DRAGEN pipeline (v3.8.4). FRASER2[21](https://paperpile.com/c/8E4RdZ/tkSKk) and OUTRIDER[22](https://paperpile.com/c/8E4RdZ/XKDW2) were used to detect abnormal splicing events and expression differences with samples run in batches of 500, both run via the DROP pipeline[49](https://paperpile.com/c/8E4RdZ/MFG2b)(v1.3.3). Significant outlier events were identified as those with false discovery rate (FDR) adjusted *P* < 0.05. The number of outlier events detected in five individuals with *RNU4-2* variants was compared to two different control sets: (1) ten individuals matched (two per *RNU4-2* individual) on genetic ancestry, sex and approximate age at consent and that didn’t have any NDD phenotypes; (2) 378 individuals with > 60 million mapped reads, age < 18, and with no NDD phenotypes. Sequence logo plots in Figure 2 and Supplementary Figure 6 were created in R (v4.0.2) using the ggseqlogo package.

*Assessing the sensitivity to detect the n.64\_65insT variant in exome sequencing data*

We used a Python script that uses samtools mpileup to retrieve the coverage and base change at the n.64\_65 critical locus to identify putative carriers of the insertion (https://github.com/francois-lecoquierre/genomics\_shortcuts/blob/main/find\_RNU4-2\_recurrent\_variant.py). This was applied to exome sequencing data (32,681 CRAM files from probands and parents) from the DDD cohort.

*Analysing RNU4-2 and RNU4-1 expression*

We used the BrainVar[25](https://paperpile.com/c/8E4RdZ/DXds5) dataset to assess patterns of whole-gene expression of *RNU4-2* and *RNU4-1* in the human cortex across prenatal and postnatal development. This dataset includes bulk-tissue RNA-seq data from 176 de-identified postmortem samples of the dorsolateral prefrontal cortex (DLPFC, n=167 older than 10 post-conception weeks) or frontal cerebral wall (n=9 younger than 10 post-conception weeks), ranging from 6 post-conception weeks to 20 years of age. The 100 bp paired-read RNA-seq data from BrainVar were aligned to the GRCh38.p12 human genome using STAR aligner[50](https://paperpile.com/c/8E4RdZ/qo2zT)(v.2.4.2a), and gene-level read counts for GENCODE v31 human gene definitions were calculated with DEXSeq[51](https://paperpile.com/c/8E4RdZ/d8pLL) (v1.50.0) and normalised to counts per million (CPM)[52](https://paperpile.com/c/8E4RdZ/D6Gsg).

*Prenatal prefrontal cortex ATAC-seq data*

Methods of generating ATAC-seq have been described previously[53](https://paperpile.com/c/8E4RdZ/yqH7q), which is the source of the data shown here. Briefly, fresh prenatal (18 and 19 gestational weeks) brain samples were dissected within 2 hours of elective termination to extract the entire telencephalic wall, from the ventricular zone to the meninges. Intact nuclei were isolated by manually douncing the tissue on ice using a loose pestle douncer then lysed on ice for 10 minutes by adding a solution with 0.1% NP-40. Nuclei were spun down by centrifugation then resuspended and exposed to Tagmentation Enzyme for 30 minutes at 37C. The ATAC-seq library was generated using Illumina barcode oligos, amplified by high-fidelity PCR, and sequenced on the Illumina HiSeq 2500 using paired-end sequencing. Reads were aligned to GRCh38 using the ENCODE ATAC-seq pipeline with default parameters[54](https://paperpile.com/c/8E4RdZ/UpA0F). A UCSC Browser track of per nucleotide ATAC-seq counts was used to assess the region around *RNU4-2* and *RNU4-1*.

*Burden testing and statistical analysis*

The enrichment of both *de novo* variants and homozygous/compound heterozygous variants across each of 28 snRNA genes and 14 constrained sub-regions was assessed in undiagnosed NDD probands compared to non-NDD probands. *De novo* variants were identified from GEL’s high-confidence de novo callset. Homozygous and compound heterozygous variants were identified from genotyping data in individual participants’ VCF files. Homozygous variants were identified as variants that are heterozygous in both parents and homozygous in offspring. To identify compound heterozygous variants in a gene or region, we assessed whether there are ≥1 paternally inherited heterozygous variant and ≥1 maternally inherited heterozygous variant in the offspring. Multiallelic sites were excluded from this analysis. Homozygous variants and compound heterozygous variants were grouped together for burden testing. Odds ratios and associated *P*-values were calculated using a two-sided Fisher’s exact test in R. A *P*-value threshold of 0.0014 was used to assess statistical significance as a Bonferroni correction accounting for 35 tests.

**References**

44. [Gargano, M. A. *et al.* The Human Phenotype Ontology in 2024: phenotypes around the world. *Nucleic Acids Res.* **52**, D1333–D1346 (2024).](http://paperpile.com/b/8E4RdZ/NsPCz)

45. [ICD-10 version:2019.](http://paperpile.com/b/8E4RdZ/Rj1JS) <https://icd.who.int/browse10/2019/en>[.](http://paperpile.com/b/8E4RdZ/Rj1JS)

46. [Pais, L. S. *et al.* seqr: A web-based analysis and collaboration tool for rare disease genomics. *Hum. Mutat.* **43**, 698–707 (2022).](http://paperpile.com/b/8E4RdZ/60LWz)

47. [Sobreira, N., Schiettecatte, F., Valle, D. & Hamosh, A. GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. *Hum. Mutat.* **36**, 928–930 (2015).](http://paperpile.com/b/8E4RdZ/63CBT)

48. [Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).](http://paperpile.com/b/8E4RdZ/FUZK7)

49. [Yépez, V. A. *et al.* Detection of aberrant gene expression events in RNA sequencing data. *Nat. Protoc.* **16**, 1276–1296 (2021).](http://paperpile.com/b/8E4RdZ/MFG2b)

50. [Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).](http://paperpile.com/b/8E4RdZ/qo2zT)

51. [Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res.* **22**, 2008–2017 (2012).](http://paperpile.com/b/8E4RdZ/d8pLL)

52. [Bedre, R. *Reneshbedre/bioinfokit: Bioinformatics Data Analysis and Visualization Toolkit*. doi:](http://paperpile.com/b/8E4RdZ/D6Gsg)[10.5281/zenodo.3965241](http://dx.doi.org/10.5281/zenodo.3965241)[.](http://paperpile.com/b/8E4RdZ/D6Gsg)

53. [Markenscoff-Papadimitriou, E. *et al.* A Chromatin Accessibility Atlas of the Developing Human Telencephalon. *Cell* **182**, 754–769.e18 (2020).](http://paperpile.com/b/8E4RdZ/yqH7q)

54. [Lee, J. *et al*. *Kundajelab/atac\_dnase\_pipelines: 0.3.0*. doi:](http://paperpile.com/b/8E4RdZ/UpA0F)[10.5281/zenodo.156534](http://dx.doi.org/10.5281/zenodo.156534)[.](http://paperpile.com/b/8E4RdZ/UpA0F)

**Extended Data Figure legends**

**Extended Data Figure 1: HPO terms for individuals in GEL.** (a) The proportion of individuals with human phenotype ontology (HPO) terms corresponding to phenotypes observed in ≥ 5 individuals with the n.64\_65insT variant compared to all other individuals with NDD. Multiple HPO terms are significantly enriched in individuals with the n.64\_65insT variant after Bonferroni adjustment (marked with a \*) indicating that individuals with the n.64\_65insT variant have more phenotypic similarity than the GEL NDD cohort as a whole. Multiple terms relating to global developmental delay, intellectual disability, hypotonia, seizure, microcephaly, autism, and short stature have been collapsed into single phenotypes. Of note, this figure relates only to HPO terms entered for each individual into GEL, which may be incomplete. Error bars indicate ±1 standard error. (b) Data plotted in panel (a) including statistics from two-sided Fisher’s exact tests. A *P*-value threshold of 2.94x10-3 was used to assess statistical significance (Bonferroni adjusted for 17 tests).

**Extended Data Figure 2: Depletion of variants in the population in *RNU4-2* and *RNU4-1***. (a; top) Distance to the median proportion of all possible SNVs that are observed in the UK Biobank in 18 bp sliding windows across the length of RNU4-2. A clear region of depletion compared to the rest of the gene is observed in the centre. (bottom) Log transformation of the mean Roulette[35](https://paperpile.com/c/8E4RdZ/cRlGc) mutability across the 3 possible SNVs within a site. (b) Total allele frequency at each site of *RNU4-1* in undiagnosed NDD probands in GEL (teal) and the UK Biobank cohort (grey). In contrast to *RNU4-2* , variants in *RNU4-1* have higher allele frequencies. A similar region of depletion is seen in the centre of *RNU4-1* (quantified in Figure 4), but this is not enriched for variants in GEL NDD or non-NDD individuals.

**Extended Data Figure 3: Sequencing coverage in exome sequencing data**. The number of sequencing reads covering the position of the n.64\_65insT variant in 13,450 probands with exome sequencing in the DDD cohort. 3,408/13,450 probands (25.3%) have at least one read at the position.

**Extended Data Figure 4: Comparison of parental age**. Comparison of (a) paternal age for probands with fathers and (b) maternal age for probands with mothers recruited into GEL for individuals with variants in *RNU4-2* (teal) and all other NDD probands (grey). Centre line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range. Individual data points, including outliers, are not shown due to Genomics England restrictions. NS: not significant. Paternal age: mean 33.1 vs 33.4 in individuals with *RNU4-2* variants and other NDD probands, respectively (two-sided t-test *P*-value=0.771; t=-0.29 (-2.41 - 1.80)). Maternal age: mean 30.2 vs 29.7 in individuals with *RNU4-2* variants and other NDD probands, respectively (two-sided t-test *P*-value=0.505; t=-0.67 (-1.07 - 2.15)).

**Extended Data Figure 5: Assessing variant density in the UK Biobank**. Median proportion of possible SNVs observed in UK Biobank per 18 bp window across 1,000 intergenic regions on chromosome 12 (grey) and *RNU4-1*, *RNU4-2* (teal). A median of 76% of all possible SNVs in *RNU4-2* are observed compared with 13% on average in the intergenic sequences of the same length (141 bp; *P*<0.001, Monte-Carlo Fisher-Pitman test).

**Extended Data Table legends**

**Extended Data Table 1: Allele counts of variants in the critical 18 bp region of *RNU4-2* (chr12:120,291,825-120,291,842) in population controls and individuals with NDD**. Numbers in brackets in NDD count columns correspond to individuals with detailed clinical information in Table 1. \*variant found in an additional sibling. \*\*NHS GMS (n=19); MSSNG (n=2); SSC (n=1); GREGoR (n=10); Undiagnosed Diseases Network (UDN; n=6); from personal communication/Matchmaker Exchange (n=16).

**Extended Data Table 2: Outlier event counts from RNA-sequencing**. Outliers predicted by OUTRIDER and FRASER2 in RNA-seq data for five individuals with *RNU4-2* variants compared to ten matched controls and 378 unmatched controls. A *P*-value threshold of 0.005 was used to assess statistical significance (Bonferroni adjusted for 10 tests). All statistical tests are one-sided Wilcoxon rank-sum tests and the *P*-values are unadjusted.

**Extended Data Table 3: Shared splicing outlier events.** Splicing outlier events detected by FRASER2 that are shared between two or more individuals with *RNU4-2* variants. None of the events were observed in any of the 378 control individuals. \*Genes identified as associated with NDD in Fu *et al*. Nature Genetics 2023[24](https://paperpile.com/c/8E4RdZ/YVmYD) (PubMed ID 35982160). DDG2P: Developmental disorders gene 2 phenotype database.

**Extended Data Table 4: Burden testing across snRNAs.** Genomic coordinates of, and burden testing results for snRNA genes in 5,426 undiagnosed NDD probands against 4,776 non-NDD probands.

**Extended Data Table 5: Burden testing in sub-regions of snRNAs**. Sub-regions of snRNA genes identified as depleted of variation and burden testing results in these regions of variants in 5,426 undiagnosed NDD probands against 4,776 non-NDD probands.