De novo variants in the non-coding spliceosomal snRNA gene *RNU4-2* are a frequent cause of syndromic neurodevelopmental disorders

Yuyang Chen[1,2], Ruebena Dawes[1,2*], Hyung Chul Kim[1,2*], Sarah L Stenton[3,4*], Susan Walker[5*], Alicia Ljungdahl[6,7], Jenny Lord[8], Vijay S Ganesh[3,4,9], Jialan Ma[3], Alexandra C Martin-Geary[1,2], Gabrielle Lemire[3,4], Elston N D'Souza[1,2], Shan Dong[6,7], Jamie M Ellingford[5,10,11], David R Adams[12], Kirsten Allan[13], Madhura Bakshi[14], Erin E Baldwin[15], Seth I Berger[16,17], Jonathan A Bernstein[18,19,20], Natasha J Brown[13,21], Lindsay C Burrage[22], Kimberly Chapman[17], Alison G Compton[13,21,23], Chloe A Cunningham[13,21], Precilla D'Souza[12], Emmanuèle C Délot[16], Kerith-Rae Dias[24,25], Ellen R Elias[26,27], Care-Anne Evans[24,28], Lisa Ewans[29,30,31], Kimberly Ezell[32], Jamie L Fraser[16,17], Lyndon Gallacher[13,21], Casie A Genetti[4,33], Christina L Grant[17], Tobias Haack[34,35], Alma Kuechler[36], Seema R Lalani[22], Elsa Leitão[36], Anna Le Fevre[13], Richard J Leventer[21,23,37], Jan E Liebelt[38,39], Paul J Lockhart[21,40], Alan S Ma[41,42], Ellen F Macnamara[12], Taylor M Maurer[19,20,43], Hector R Mendez[19,20,44], Stephen B Montgomery[19,20,45], Marie-Cécile Nassogne[46,47], Serena Neumann[33], Melanie O'Leary[3], Elizabeth E Palmer[30,31], John Phillips[33], Georgia Pitsava[48], Ryan Pysar[30,31,49], Heidi L Rehm[3,50], Chloe M Reuter[19,20,44], Nicole Revencu[51], Angelika Riess[35], Rocio Rius[21,52,53], Lance Rodan[4], Tony Roscioli[25,26,29], Jill A Rosenfeld[22], Rani Sachdev[30,31], Cas Simons[52,53], Sanjay M Sisodiya[54,55], Penny Snell[40], Laura St Clair[41], Zornitza Stark[13,21], Tiong Yang Tan[13,21], Natalie B Tan[13], Suzanna EL Temple[14,56], David R Thorburn[13,21,23], Cynthia J Tifft[12], Eloise Uebergang[23], Grace E VanNoy[3], Eric Vilain[57], David H Viskochil[15], Laura Wedd[52,53], Matthew T Wheeler[19,20,44], Susan M White[13,21], Monica Wojcik[4,34,58], Lynne A Wolfe[12], Zoe Wolfenson[12], Changrui Xiao[59], David Zocche[60], John L Rubenstein[7], Eirene Markenscoff-Papadimitriou[61], Sebastian M Fica[62], Diana Baralle[63,64], Christel Depienne[24], Daniel G MacArthur[52,53], Joanna MM Howson[65], Stephan J Sanders[6,7], Anne O'Donnell-Luria[3,4,50], and Nicola Whiffin[1,2,3]

1. Big Data Institute, University of Oxford, Oxford, UK
2. Centre for Human Genetics, University of Oxford, Oxford, UK
3. Broad Center for Mendelian Genomics, Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA
4. Division of Genetics and Genomics, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA
5. Genomics England, London, UK
6. Institute of Developmental and Regenerative Medicine, Department of Paediatrics, University of Oxford, Oxford, UK
7. Department of Psychiatry and Behavioral Sciences, UCSF Weill Institute for Neurosciences, University of California, San Francisco, USA
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. Undiagnosed Disesases Program, National Human Genome Research Institute, Bethesda, MD, USA
13. Victorian Clinical Genetics Services, Murdoch Children's Research Institute, Melbourne, VIC, Australia
14. Department of Clinical Genetics, Liverpool Hospital, Sydney, NSW, Australia
15. Division of Medical Genetics, Department of Pediatrics, University of Utah School of Medicine, Salt Lake City, UT, USA

16. Center for Genetic Medicine Research, Children's National Research Institute, Washington, DC, USA
17. Division of Genetics and Metabolism, Children's National Hospital, Washington, DC, USA
18. Department of Pediatrics, Stanford University School of Medicine, Stanford, CA, USA
19. GREGoR Stanford Site, Stanford University School of Medicine, Stanford, CA, USA
20. Center for Undiagnosed Diseases, Stanford University School of Medicine, Stanford, CA, USA
21. Department of Paediatrics, University of Melbourne, Melbourne, VIC, Australia
22. Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA
23. Murdoch Children's Research Institute, Melbourne, VIC, Australia
24. Neuroscience Research Australia, Sydney, NSW, Australia
25. Prince of Wales Clinical School, Faculty of Medicine, University of New South Wales, Sydney, NSW, Australia
26. Department of Pediatrics, Children's Hospital Colorado, Aurora, CO, USA
27. University of Colorado School of Medicine, University of Colorado, Aurora, CO, USA
28. New South Wales Health Pathology Randwick Genomics, Prince of Wales Hospital, Sydney, NSW, Australia
29. Discipline of Paediatrics and Child Health, Faculty of Medicine and Health, University of New South Wales, Sydney, NSW, Australia
30. Centre for Clinical Genetics, Sydney Children's Hospitals Network, Randwick, NSW, Australia
31. Genomics and Inherited Disease Program, Garvan Institute of Medical Research, Darlinghurst, NSW, Australia
32. Division of Medical Genetics & Genomic Medicine, Vanderbilt University Medical Center, Nashville, TN, USA
33. Manton Center for Orphan Disease Research, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA
34. Institute of Medical Genetics and Applied Genomics, University of Tübingen, Tübingen, Germany
35. Center for Rare Diseases Tübingen, University of Tübingen, Tübingen, Germany
36. Institute of Human Genetics, University Hospital Essen, University Duisburg-Essen, Essen, Germany
37. Royal Children’s Hospital, Melbourne, VIC, Australia
38. Paediatric and Reproductive Genetics Unit, South Australian Clinical Genetics Service, Women's and Children's Hospital, North Adelaide, SA, Australia
39. Repromed, Dulwich, SA, Australia
40. Bruce Lefroy Centre, Murdoch Children's Research Institute, Melbourne, VIC, Australia
41. Department of Clinical Genetics, Sydney Children's Hospitals Network Westmead, Sydney, NSW, Australia
42. Specialty of Genomic Medicine, University of Sydney, Sydney, NSW, Australia
43. Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA
44. Department of Medicine - Cardiovascular Medicine, Stanford University School of Medicine, Stanford, CA, USA
45. Department of Pathology, Department of Genetics, Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA, USA
46. Service de Neurologie Pédiatrique, Cliniques Universitaires Saint-Luc, UCLouvain, B-1200, Brussels, Belgium
47. Institut des Maladies Rares, Cliniques Universitaires Saint-Luc, UCLouvain, B-1200, Brussels, Belgium
48. Institute for Clinical and Translational Research, University of California, Irvine, CA, USA
49. Department of Clinical Genetics, The Children's Hospital at Westmead, Westmead, NSW, Australia
50. Center for Genomic Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA
51. Center for Human Genetics, Cliniques universitaires Saint-Luc, Université catholique de Louvain, Brussels, Belgium
52. Centre for Population Genomics, Garvan Institute of Medical Research and UNSW Sydney, Sydney, New South Wales, Australia
53. Centre for Population Genomics, Murdoch Children’s Research Institute, Melbourne, Victoria, Australia
54. Department of Clinical and Experimental Epilepsy, UCL Queen Square Institute of Neurology, London, UK
55. UK and Chalfont Centre for Epilepsy, Bucks, UK
56. School of Women's and Childrens's Health, University of New South Wales, Sydney, NSW, Australia
57. Institute for Clinical and Translational Science, University of California, Irvine, CA, USA
58. Division of Newborn Medicine, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA
59. Department of Neurology, University of California, Irvine, CA, USA
60. North West Thames Regional Genetics Service, Northwick Park & St Mark’s Hospitals, London, UK
61. Department of Psychiatry, Langley Porter Psychiatric Institute, UCSF Weill Institute for Neurosciences, University of California, San Francisco, USA
62. Department of Biochemistry, University of Oxford, Oxford, UK

63. School of Human Development and Health, Faculty of Medicine, University of Southampton, Southampton, UK
64. National Institute for Health Research (NIHR) Southampton Biomedical Research Centre, University Hospital Southampton National Health Service (NHS) Foundation Trust, Southampton, UK
65. 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]. Increasingly, 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 novel 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]. 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 119 individuals with NDD. The vast majority of individuals (77.3%) have the same highly recurrent single base-pair insertion (n.64_65insT). We estimate that variants in this region explain 0.41% of individuals with NDD. We demonstrate that *RNU4-2* is highly expressed in the developing human brain, in contrast to its contiguous counterpart *RNU4-1* and other U4 homologs, supporting *RNU4-2*'s role as the primary U4 transcript in the brain. Overall, this work underscores the importance of non-coding genes in rare disorders. It will provide a diagnosis to thousands of individuals with NDD worldwide and pave the way for the development of effective treatments for these individuals.

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]. To date, the overwhelming majority of known disease-causing variants are in the ~1.5% of the genome that directly encodes proteins[3]. 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].

Non-coding RNAs, which comprise 37.4% of processed exonic RNA sequence in humans[5], include important regulators of biological processes with diverse roles across cells and tissues[6]. Small nuclear RNAs (snRNAs) are a subcategory of non-coding RNAs that are key components of the spliceosome[7]. 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]. 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], while *RNU4ATAC* variants cause an autosomal recessive multisystem congenital disorder including microcephaly, growth retardation, and developmental delay (eponyms include Taybi Linder[10], Lowry-Wood[11] and Roifman syndromes[12]).

Here, we identify variants in *RNU4-2*, which encodes the U4 snRNA component of the major spliceosome, as a newly recognised autosomal dominant disorder. Using a cohort of 8,841 probands with genetically undiagnosed NDD in Genomics England (GEL)[4], 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 ~0.41% 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 explains 0.52% of undiagnosed NDD in Genomics England

We identified a highly recurrent single base insertion (GRCh38:chr12:120,291,839:T:TA; n.64_65insT) in *RNU4-2* in GEL[1]. This variant was initially identified as arising *de novo* in 38 probands recruited for genome sequencing with their unaffected parents[13]. 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 $P=2.75\times10^{-4}$), delayed gross motor development (n=26; OR=2.55; $P=1.64\times10^{-3}$), microcephaly (n=26; OR=6.62; $P=7.87\times10^{-10}$), delayed fine motor development (n=24; OR=2.61; $P=1.69\times10^{-3}$), hypotonia (n=18; OR=3.60; $P=7.09\times10^{-5}$), short stature (n=15; OR=3.54; $P=2.17\times10^{-4}$), drooling (n=7; OR=19.2; $P=2.83\times10^{-7}$), and absent speech (n=6; OR=6.23; $P=7.45\times10^{-4}$) compared to all other probands with NDD in GEL (n=12,203; diagnosed and undiagnosed) (**Figure 1A; Supplementary Table 1**).

The n.64_65insT variant is not found in 76,215 genome-sequenced individuals in gnomADv4.0[14], or in 245,400 individuals in the All of Us dataset[15]. It is seen in a single individual in the UK Biobank[16] (allele frequency=1.02×10^{-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 2)**.

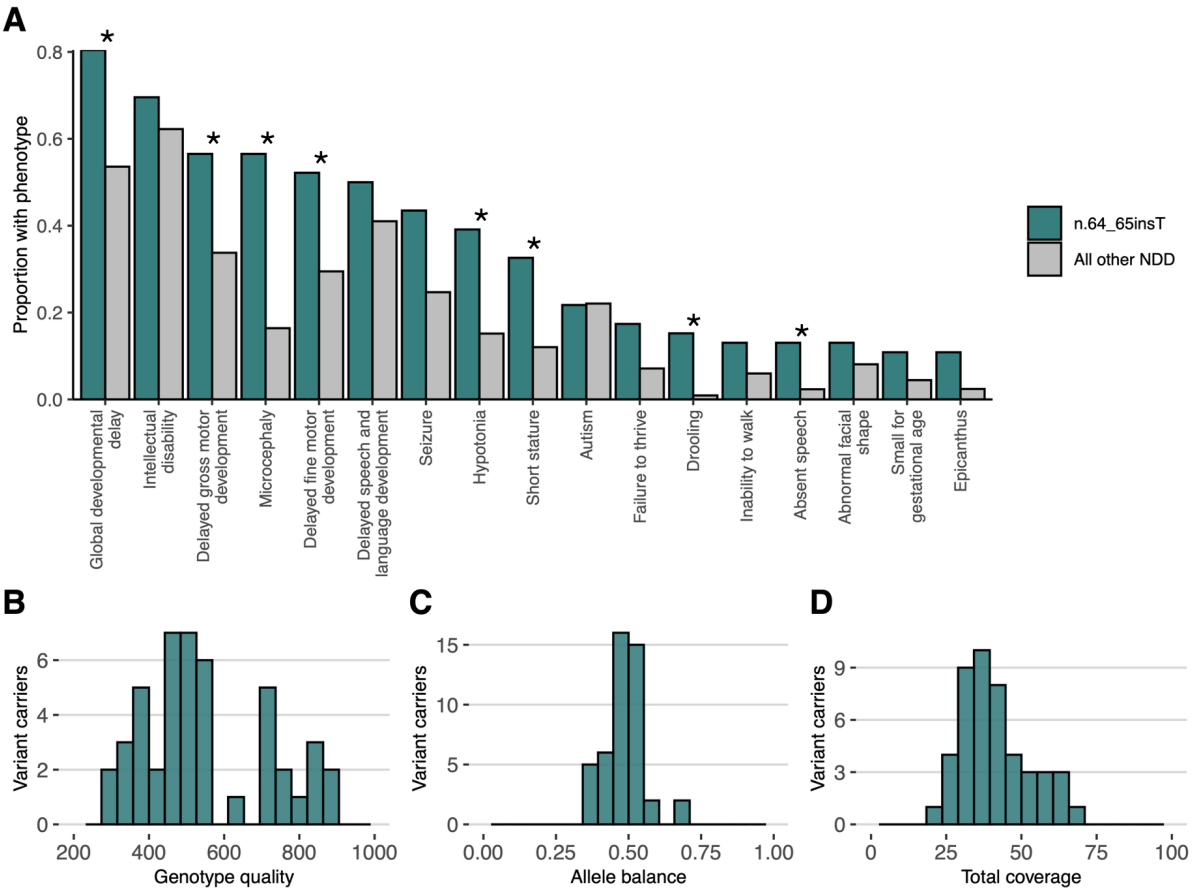


Figure 1: Characterisation of individuals with the n.64_65insT variant in GEL. (A) The proportion of individuals with human phenotype ontology (HPO) terms corresponding to phenotypes observed in ≥ 5 individuals with the n.64_65insT variant compared to all other individuals with NDD. Terms that are significantly enriched in individuals with the n.64_65insT variant are marked with a *. 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. A more detailed phenotypic characterisation of individuals with variants in *RNU4-2* is provided below. (B-D) Quality control metrics for the variant calls in all 46 individuals with the variant: (B) genotype quality scores, (C) allele balance, and (D) coverage.

Given the high recurrence rate of this 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 (**Figure 1B-D**) and manual inspection on IGV (**Supplementary Figure 1**). Finally, 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 (**Supplementary Figure 2**).

The n.64_65insT variant is within a highly constrained region with multiple NDD-causing variants

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 2A; Supplementary Figure 3**). Based on the population variant data, we defined a critical, highly constrained region as chr12:120,291,825-120,291,842.

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 119 individuals with variants across this region (**Table 1**), the vast majority of which have the initial n.64_65insT variant (n=92; 77.3%). For 92 of the 119 individuals, sequencing data for both parents was available to confirm the variants had arisen *de novo*. 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. The enrichment of single base insertion variants in this region in individuals with NDD is striking: 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}).

variant	nucleotide description	GEL NDD count (in Table 2)	Non-GEL NDD count** (in Table 2)	population cohort count
Single base insertions				
12:120291839:T:TA	n.64_65insT	46 (2)	46 (31)	1 (UK Biobank)
12:120291839:T:TC	n.64_65insG	0	2 (1)	0
12:120291826:T:TA	n.77_78insT	6*	0	0
12:120291827:T:TA	n.76_77insT	1	0	0
12:120291835:G:GT	n.68_69insA	1	0	0
12:120291838:T:TA	n.65_66insT	1	0	0
	Total	**55***	**48**	**1**
SNVs				
12:120291839:T:C	n.65A>G	2	0	0
12:120291826:T:G	n.78A>C	1	0	0
12:120291828:G:A	n.76C>T	1	6 (1)	1 (gnomAD v4)
12:120291835:G:A	n.69C>T	0	1 (1)	0
12:120291837:T:C	n.67A>G	1	3	0
12:120291841:A:C	n.63T>G	1	0	0
	Total	**6**	**10**	**1**

Table 1: Variants identified in individuals with NDD in the 18 bp critical region of *RNU4-2* (chr12:120,291,825-120,291,842). Numbers in brackets in NDD count columns correspond to individuals with detailed clinical information in **Table 2**. The count in population cohorts is shown only for variants observed in individuals with NDD. A full list of variants found across the region in population cohorts is in **Supplementary Table 3**. *count includes two siblings. **NHS GMS (n=21); MSSNG[17] (n=2); SSC[18] (n=1); GREGoR (n=10); Undiagnosed Diseases Network[19] (UDN; n=8); from personal communication/Matchmaker Exchange (n=17).

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 16 individuals across cohorts with SNVs in this region (**Table 1**; 11 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 2B**). 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 (**Supplementary Table 3**). Across the remainder of *RNU4-2* there is no significant enrichment of variants in undiagnosed NDD probands when compared to non-NDD probands (194/7,519 undiagnosed NDD; 521/19,428 non-NDD in GEL aggregated variant dataset[20]; 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 119 individuals with NDD. This includes 60/8,841, or 0.68%, of all genetically undiagnosed NDD probands in GEL (0.49% of all NDD probands). 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=1.84×10^{-78}).

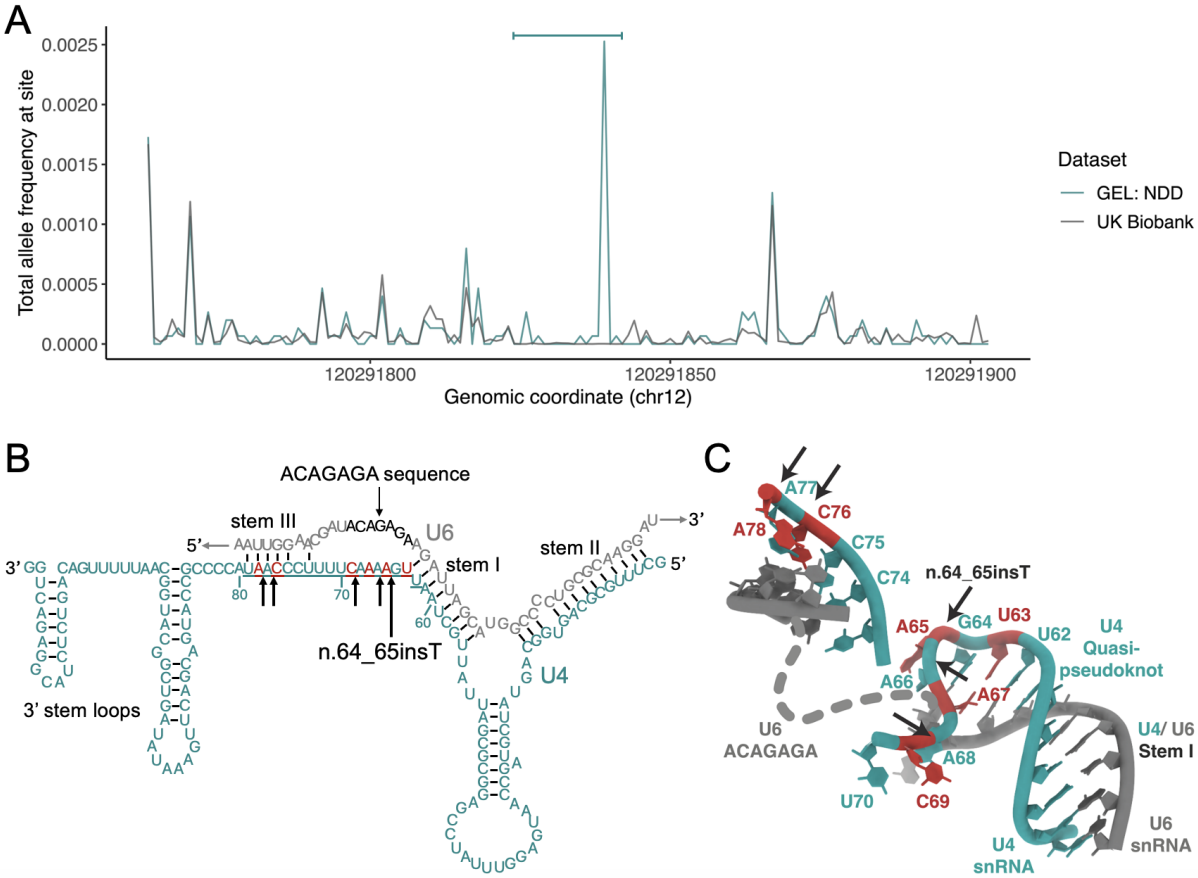


Figure 2: A highly structured 18 bp region of RNU4-2 that is critical for BRR2 helicase activity is enriched for variants in NDD and depleted in population cohorts. (A) Allele frequency of variants in 7,519 undiagnosed NDD probands GEL (teal) and the UK Biobank cohort (grey) across *RNU4-2*. The 18 bp critical region is marked by a horizontal bar at the top of the plot. (B) Schematic of U4 (teal) binding to U6 snRNA (grey). The 18 bp critical region is underlined. (C) The structure of U4 and U6 snRNAs resolved by cryoEM[21]. Created using RCSB Protein Data Bank[22] (structure 6QW6). In both (B) and (C) single base insertions identified in individuals with NDD are shown by black arrows and positions of SNVs by red nucleotides.

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 U4 and U6 is essential to generate the catalytically active spliceosome[2]. The 18 bp critical region in *RNU4-2* maps to a single-stranded 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 2B**). This region is known to be loaded into the active site of the *SNRNP200*-encoded BRR2 helicase, which mediates unwinding of the U4/U6 duplex[2]. The highly recurrent n.64_65insT variant is within a previously described 'quasi pseudoknot', or T-loop, structure[21] (**Figure 2C**). 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[23] (**Figure 2C**). Both of these regions are thought to stabilise the U4/U6 interaction and accurately position the U6 ACAGAGA sequence to receive the 5'splice site during spliceosome activation. Insertion of a single base into either of these structures 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. 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.

Variants in this crucial region cause a severe syndromic NDD phenotype

To characterise the phenotypic spectrum associated with variants in *RNU4-2*, we collected detailed phenotypic information for a subset of 36 individuals (33 with n.64_65insT, one with n.64_65insG, and two with SNVs; **Table 2**; **Supplementary Table 4**). Using these data, we find the *RNU4-2* syndromic NDD to be characterised by moderate to severe global developmental delay (two children with SNVs with moderate delay) and intellectual disability in all individuals. The majority (82%) achieved ambulation but at a delayed age (average 3.6 years, range 18 months to 7.5 years) with many noted to have a wide-based or ataxic gait. Only one individual (with an SNV) had fluent speech, some had a few words, and most were non-verbal. All but one 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 coll, 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.

Associated growth and neurodevelopmental phenotypes present in ≥75% of individuals include short stature, microcephaly (mostly congenital), seizures (spanning infantile spasms, focal seizures and generalised tonic-clonic seizures, febrile seizures, and status epilepticus with variable onset from the first year of life, but most between 3-10 years of age), and hypotonia. Brain MRI showed a spectrum of abnormalities in the majority of individuals, most frequently reduced white matter volume, non-specific abnormalities of the white matter, hypoplasia of the corpus callosum, ventriculomegaly, and delayed myelination. Involvement of multiple organ systems was reported for all individuals, often including visual (optic nerve hypoplasia, cortical blindness, strabismus, nystagmus), gastrointestinal (constipation, reflux, feeding issues with need for a gastrostomy tube), and bone/skeletal abnormalities (osteopenia, recurrent fractures, scoliosis, kyphosis, hip dysplasia), and in a lesser number of individuals, hearing, endocrine (hypothyroidism, growth hormone deficiency), limb, sleep, genitourinary, dental, cardiac, and cutaneous concerns (**Table 2; Supplementary Table 4**).

Clinical feature			
Individuals (n)		36	
Sex		14 F, 22 M	
		Median	**Range**
Age at last evaluation (years)		8.25	0.5 - 29
Maternal age at birth (years)*		32	22 - 41
Paternal age at birth (years)*		33	26 - 41
		Count**	**Percentage**
Growth	IUGR	6/33	18%
	Short stature	30/35	86%
	Microcephaly	28/34	82%
	- congenital	16/28	
	- acquired	7/28	
	- not specified	5/28	
Neurodevelopmental	GDD	36/36	100%
	- severe	23/36	
	- moderate	6/36	
	- not specified	7/36	
	Ambulatory (>5yo)	18/22	82%
	- abnormal gait	7/18	
	- not specified	11/18	
	Speech abnormality	32/33	97%
	- non-verbal	25/32	
	- few words	7/32	
	ID	30/30	100%
	Behavioural issues	20/30	67%
	ASD	16/27	59%
	Hypotonia	32/33	97%
	Seizures	26/34	76%
	Abnormal brain MRI	30/33	91%
Hearing	Hearing loss	7/33***	21%
Vision	Vision issues	28/32	88%
	- Optic nerve hypoplasia	5/8	63%
	- Strabismus	17/26	65%
	- Nystagmus	14/18	78%
Gastrointestinal	Constipation	21/25	84%
	GORD	13/25	52%
	Feeding difficulties	27/32	84%
	G-tube	4/8	50%
	Growth problems	21/26	81%
Endocrine		14/24	58%
Bone/skeletal		20/23	87%
Limb		18/27	67%
Genitourinary		9/25	36%
Dental		11/27	41%
Cardiac		6/26	23%
Cutaneous		16/27	59%
Dysmorphic facial features		28/29	97%

Table 2: Clinical features of 36 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; GH, growth hormone; G-tube, gastrostomy tube
*parental age only available for 27/36 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

RNA-sequencing in blood does not show a global disruption to splicing

Given the importance of U4 snRNA in the spliceosome and previous observations of global disruption to splicing observed in other spliceosomopathies[24], 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). We did not see any significant difference in the number of gene expression outliers using OUTRIDER[25], or in the number of retained introns, or all outlier events using FRASER2[26] in the five individuals with *RNU4-2* variants compared with 5,409 controls (**Supplementary Table 5**). At present, RNA from additional tissues (e.g. brain samples) of affected individuals is not available. It is possible that the observed *RNU4-2* variants disrupt more subtle aspects of alternative splicing in a tissue-specific manner, as has been observed for other snRNA variants[27].

RNU4-2 is highly expressed across tissues and in the brain across development

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 6**). *RNU4-2* and *RNU4-1* are contiguous on chr12, both 141 bp long, and highly homologous, differing by four nucleotides (97.2% homology). *RNU4-1* has a similar depletion of variants in population cohorts in the centre of the RNA, however, we do not observe an enrichment of variants in GEL in this central region. Further, we do not observe an enrichment of variants in GEL in this central region (**Supplementary Figure 4**). 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 2**).

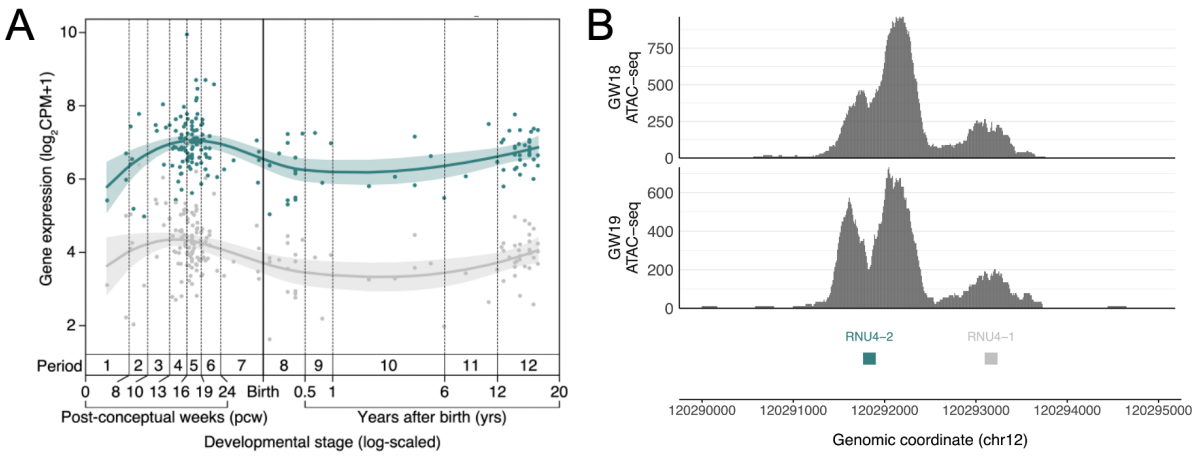


Figure 3: *RNU4-2* is more highly expressed than *RNU4-1* in the prefrontal cortex. (A) Levels of *RNU4-1* (grey) and *RNU4-2* (teal) expression at different developmental stages from BrainVar[28]. (B) ATAC-seq data from human prenatal prefrontal cortex (18 and 19 gestational weeks (GW)) with substantially higher peaks of chromatin accessibility around *RNU4-2* (teal) than *RNU4-1* (grey).

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[28]. 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 3A**). 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 3B**). Overall, these data support the role of *RNU4-2* as the major U4 transcript in the brain.

Multiple factors likely explain the high recurrence of the n.64_65insT variant

The n.64_65insT variant is highly recurrent. It is observed in 46/12,249 NDD probands in GEL (0.38%; or 0.52% of undiagnosed NDD probands). In contrast, the most recurrent protein-coding variant in a dataset of 31,058 individuals with developmental disorders[29] is observed in 36 individuals (0.12%; GRCh38:chr11:66211206:C:T; PACS1:p.Arg203Trp). The exact reasons for this high recurrence are unclear, however, we hypothesise three contributing factors. First, a high local mutation rate, which may be driven by the open chromatin state and very high levels of transcription (**Figure 3**). In UK Biobank, a median of 76% of all possible SNVs in *RNU4-2* are observed (calculated across 18 bp sliding windows). This is compared with 13% on average in 1,000 random intergenic sequences of the same length (141 bp; $P<0.001$, Monte-Carlo Fisher-Pitman test; **Supplementary Figure 6**). 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 high levels of selection acting on variants across *RNU4-2*.

Secondly, a high overall mutational burden does not 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[30]. The high propensity of this region to form secondary structure when single-stranded may drive creation of this specific insertion. Finally, it may be that germline selection is acting to increase the frequency of this specific variant, as has been shown for other highly recurrent sites[31]. While we see no association with paternal age (mean 33.1 in probands with *RNU4-2* variants and 33.4 across other NDD probands; **Supplementary Figure 7**), fully testing this hypothesis will require deep sequencing of testes or sperm samples.

No other spliceosomal snRNA genes are enriched for de novo variants in NDD

Given the newly 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 (**Supplementary Table 7**). 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=13) using the

high-confidence *de novo* callset in GEL. Of the 12 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**; **Supplementary Table 8**). 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 $P=1.34 \times 10^{-9}$; undiagnosed NDD probands n=33, non-NDD probands n=0; all other tests Fisher's *P*>0.25).

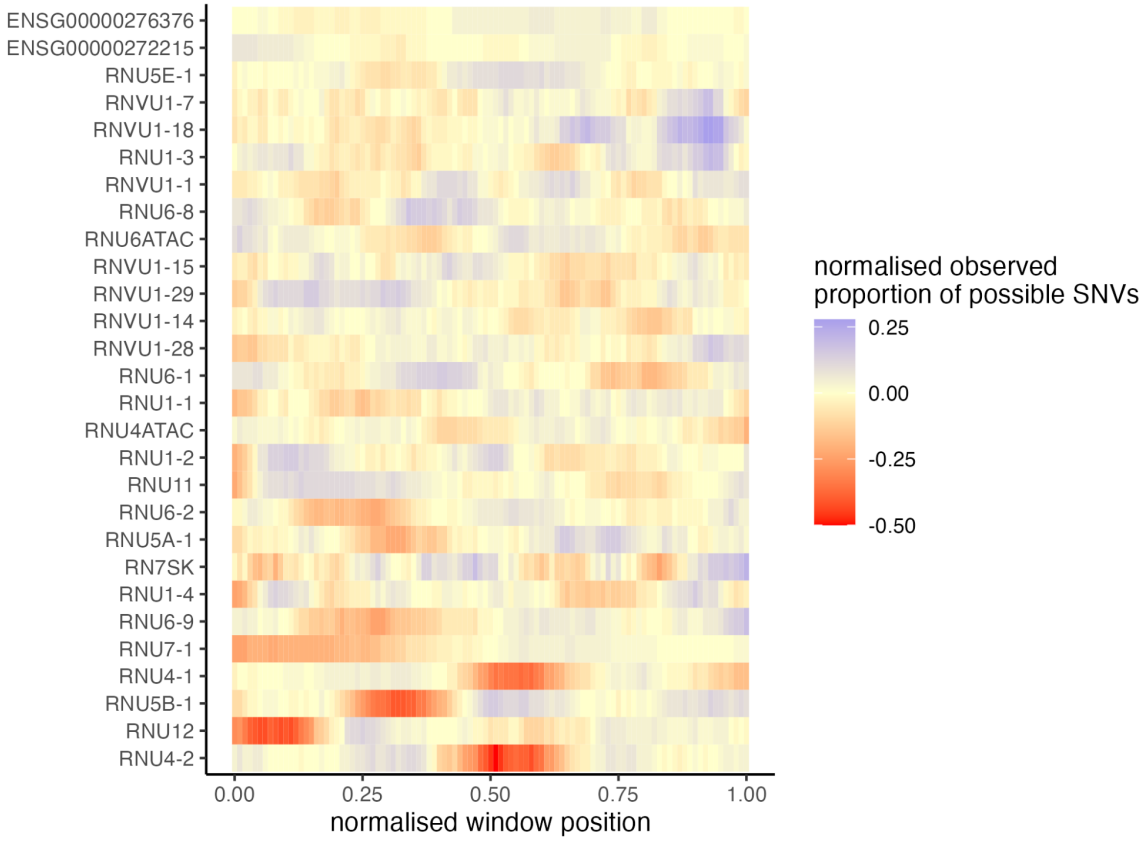


Figure 4: Multiple snRNA genes have regions that are depleted of variation in the population. The proportion of observed SNVs in 490,640 genome sequenced individuals in the UK Biobank, in sliding windows of 18 bp across each snRNA gene, normalised to the median value for each gene.

Finally, 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).

Discussion

Here, we identified a highly constrained 18 bp region of *RNU4-2* in which variants cause a severe neurodevelopmental phenotype. Variants in this region were identified in 0.68% of individuals with currently undiagnosed NDD in GEL. Assuming a diagnostic rate of 40% upstream of defining our undiagnosed NDD cohort, consistent with recent reports[29], we estimate that variants in *RNU4-2* could explain 0.41% of all NDD (60/(8841/6*10)). As a comparison, the largest proportion of DD explained by a single gene in a cohort of 31,058 individuals with DD[29] was 0.47% for *ARID1B*, 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[18], SPARK[37] and MSSNG[17]).

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 captured by commercially available clinical exome sequencing which primarily captures protein-coding exons[5]. The detailed phenotypic characterisation included here will help prioritise individuals for targeted sequencing of *RNU4-2*.

As *RNU4-2* is a snRNA component of the major spliceosome, we hypothesised that the identified variants would cause a global dysregulation of splicing. However, we did not see this in RNA-seq data derived from blood samples. There are several possible explanations for this result. Firstly, humans have multiple copies of U4-encoding genes, including *RNU4-1* and *RNU4-2*. While *RNU4-2* is the major U4 transcript in the brain, other U4 genes, such as *RNU4-1*, could be expressed at higher levels in blood and play a compensatory role. Future work should look for an effect on splicing in a more relevant cell type or tissue. Secondly, while retention of minor introns is observed in individuals with variants in minor spliceosome components[32], minor introns represent only a small fraction of all introns across the genome. Large-scale intron retention across major introns would likely be embryonic lethal. The identified variants in *RNU4-2* might therefore have a much more subtle and widespread effect on splicing which is harder to detect. Indeed, 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, consistent with this region being involved in subtle regulation of alternative splicing[33,34].

Finally, given the striking role of *RNU4-2* in NDD we explored whether other snRNA genes could explain undiagnosed cases. We did not find any other snRNAs, or constrained sub-regions of snRNAs, that were enriched for *de novo* variants in NDD cases. 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). Additionally, we did not explore whether variants in these snRNAs may cause recessive disorders. Variants in the regions we identified should also be investigated in other disease cohorts.

In summary, we identify *RNU4-2* as a novel syndromic NDD gene, explaining ~0.41% of all individuals with NDD. Including *RNU4-2* in standard clinical workflows will end the diagnostic odyssey for thousands of NDD patients worldwide and pave the way for development of effective treatments for these individuals.

Methods

Categorising participants in Genomics England

We defined four groups of individuals in GEL v18. Individuals with NDD (n=13,812) were defined as those with human phenotype ontology (HPO)[35] and/or International Classification of Diseases 10th Revision (ICD-10) codes[36] 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]. 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)[20] 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], All of Us[15] (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].

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 node, one match was made (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 reviewed the Broad Center for Mendelian Genomics and the GREGoR consortium datasets. DGM contacted additional local collaborators. Clinical data were collected and summarised for features seen across the cohort.

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)[18], SPARK (3,144 cases; 2,190 controls)[37], and MSSNG (1,622 cases; 52 controls)[17].

Generating 1,000 random intergenic sequences

Using the bedtools subtractBed function[38] 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[28] 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. FRASER2[26] and OUTRIDER[25] were used to detect abnormal splicing events and expression differences with samples run in batches of 500, both run via the DROP pipeline[39].

Analysing RNU4-2 and RNU4-1 expression

We used the BrainVar[28] 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[40], and gene-level read counts for GENCODE v31 human gene definitions were calculated with DEXSeq[41] and normalised to counts per million (CPM)[42].

Prenatal prefrontal cortex ATAC-seq data

Methods of generating ATAC-seq have been described previously[43], 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[44]. 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 *de novo* variants across each of 28 snRNA genes and 14 constrained sub-regions was assessed in undiagnosed NDD probands compared to non-NDD probands using the high-confidence *de novo* callset. Odds ratios and associated *P*-values were calculated using Fisher's exact test in R. A *P*-value threshold of 0.0031 was used to assess statistical significance as a Bonferroni correction accounting for 16 tests.

Acknowledgements

We would like to thank Peter O'Donovan, Mimoza Hoti and Joanne Yang from Genomics England for their help with clinician collaboration and Airlock requests.

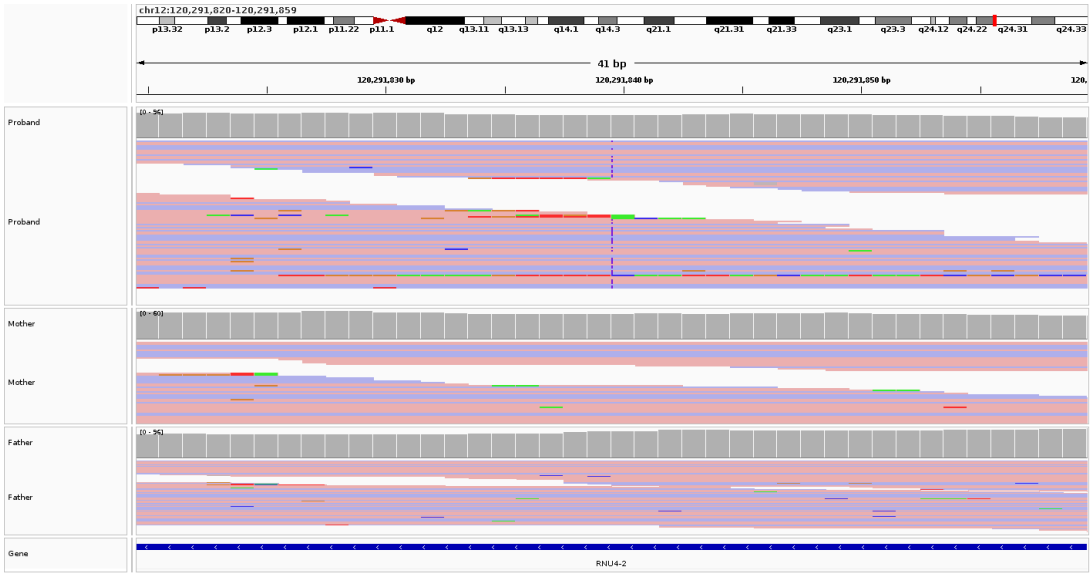
YC is supported by a studentship from Novo Nordisk. NW is supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (220134/Z/20/Z). GL was supported by the Fonds de recherche en santé du Québec (FRQS), VSG by NIAMS K23AR083505, and SLS by a fellowship from Manton Center for Orphan Disease Research at Boston Children's Hospital. The research was supported by grant funding from Novo Nordisk and 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), 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.

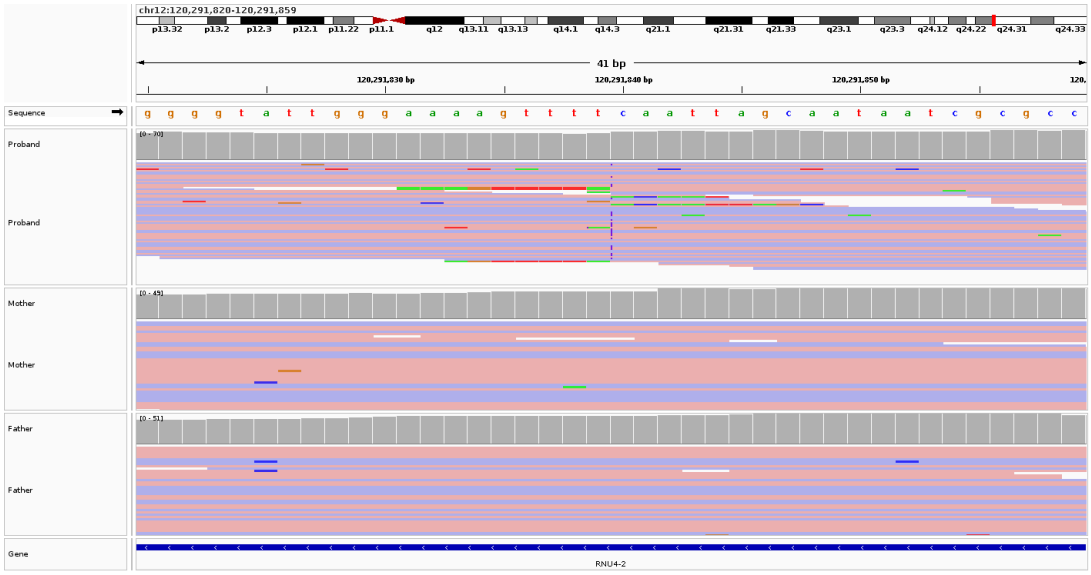
This research was made possible through access to data in the National Genomic Research Library, which is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The National Genomic Research Library holds data provided by patients and collected by the 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.

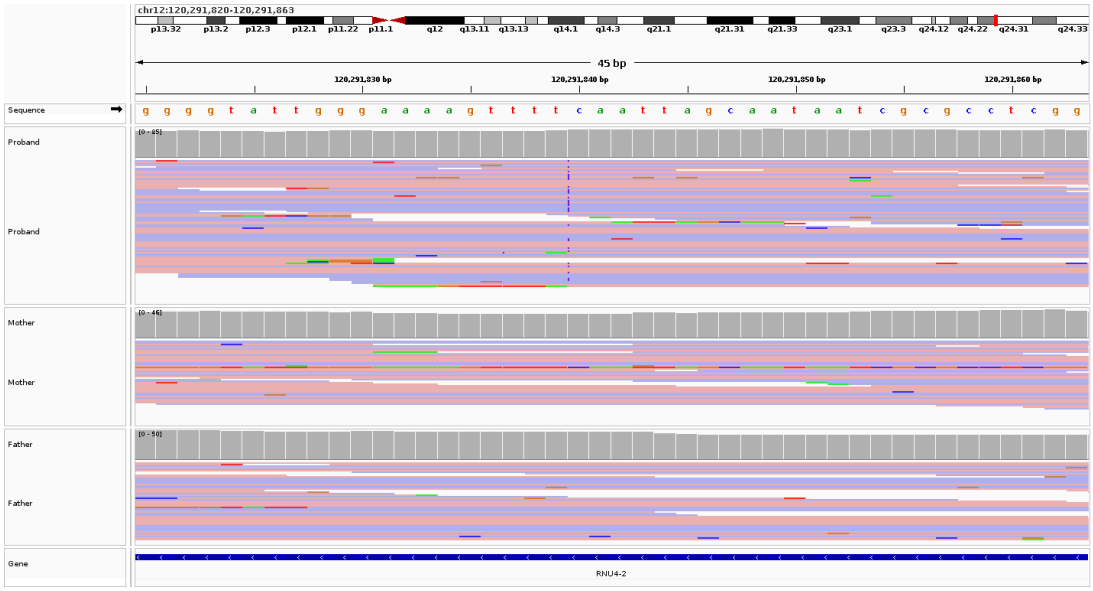
Competing interests

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 and Ono Pharma USA Inc., 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.

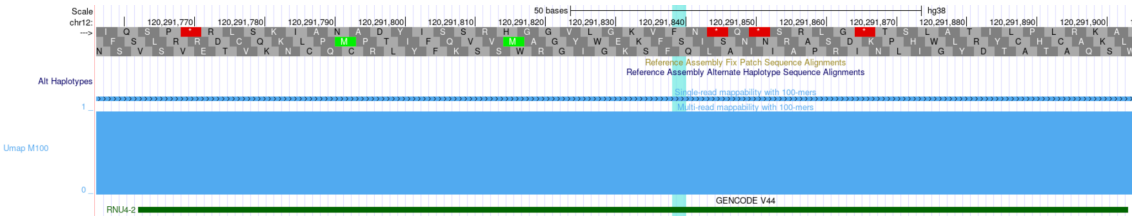
Supplementary Figures



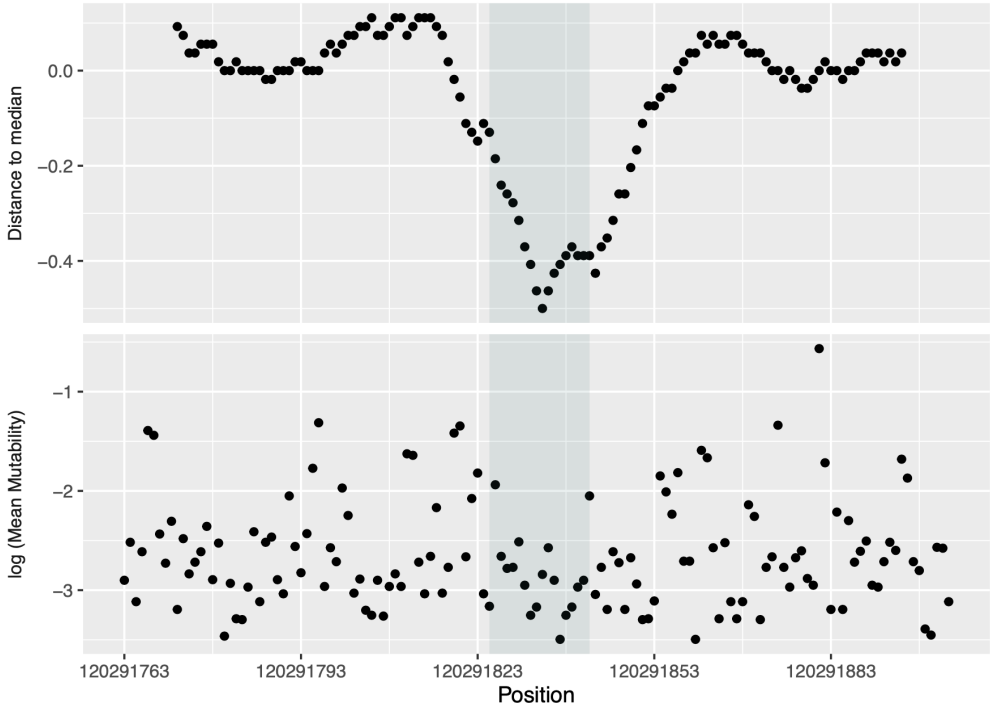




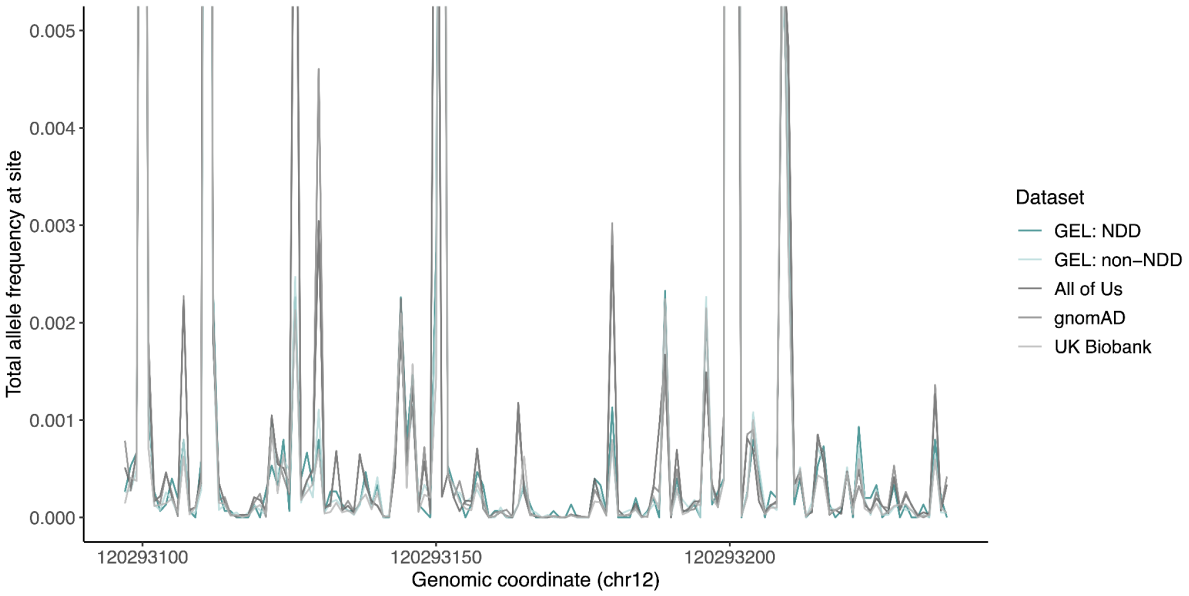
Supplementary Figure 1: Example IGV plots of the region surrounding the n.64_65insT variant in three trios.



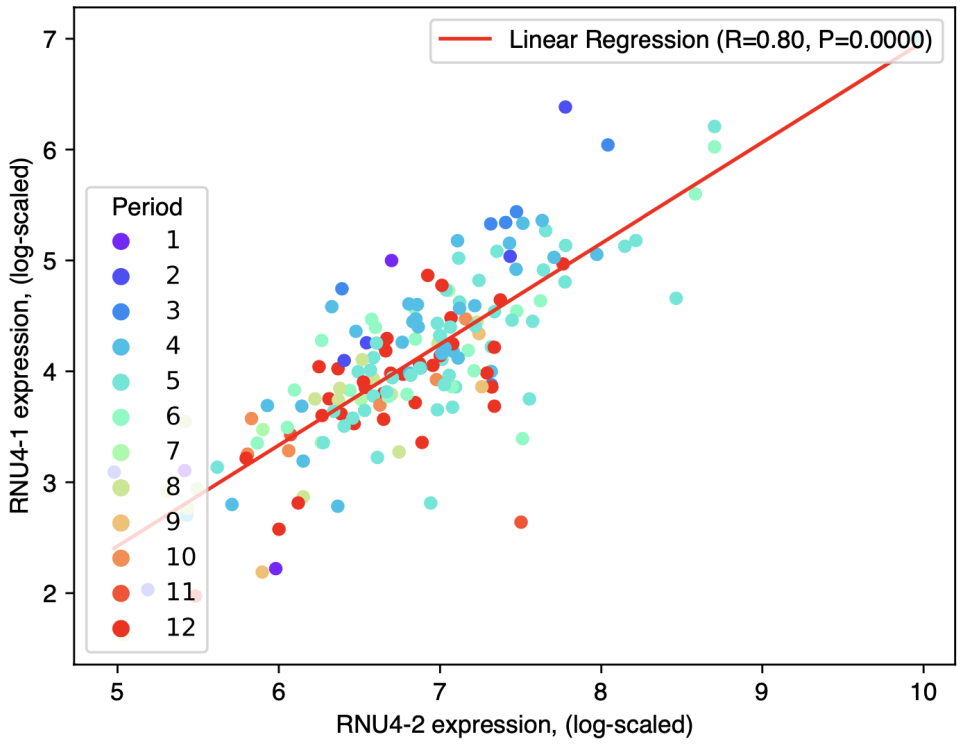
Supplementary Figure 2: Screenshot from the UCSC Genome Browser showing high mappability for 100-mers across the *RNU4-2* gene.



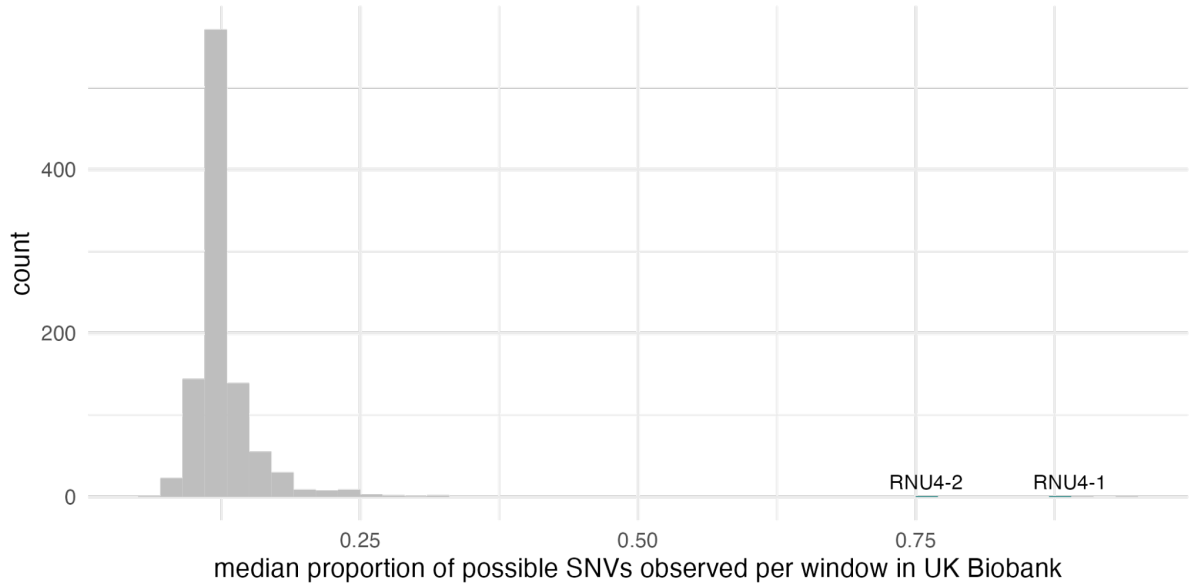
Supplementary Figure 3: (top) Distance to the median proportion of all possible SNVs that are observed in the UK Biobank in 18 bp sliding windows across the length of RNU4-2. A clear region of depletion compared to the rest of the gene is observed in the centre. (bottom) Log transformation of the mean Roulette[45] mutability across the 3 possible SNVs within a site.



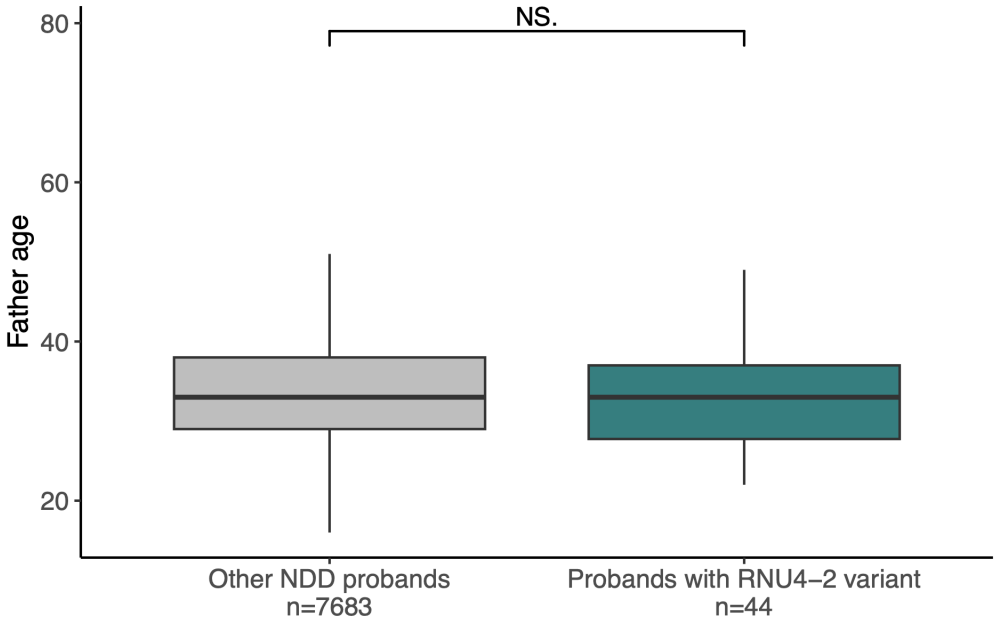
Supplementary Figure 4: Total allele frequency at each site of *RNU4-1* in five datasets. In contrast to *RNU4-2* (Figure 2a), variants in *RNU4-1* have higher allele frequencies. A similar region of depletion is seen in the centre of *RNU4-1* (quantified in Figure 4), but this is not enriched for variants in GEL NDD or non-NDD individuals.



Supplementary Figure 5: Correlation between *RNU4-1* and *RNU4-2* expression in RNA-seq data from human cortex across prenatal and postnatal development from BrainVar[28].



Supplementary Figure 6: 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).



Supplementary Figure 7: Comparison of paternal age for probands with fathers recruited into GEL.

Supplementary Tables

Supplementary Table 1: The number of probands with the n.64_65insT variant and all other individuals with NDD with HPO terms corresponding to phenotypes observed in ≥ 5 individuals compared to all other NDD probands. These data are plotted in Figure 1a. A P-value threshold of 2.94×10^{-3} was used to assess statistical significance (Bonferroni adjusted for 17 tests).

Supplementary Table 2: ICD10 and ICD9 codes for individuals with single base pair insertions between codons 64 and 65 of *RNU4-2* and *RNU4-1* in the UK Biobank.

Supplementary Table 3: Outliers predicted by OUTRIDER and FRASER2 in RNA-seq data for five individuals with *RNU4-2* variants compared to 5,409 controls. A P-value threshold of 0.017 was used to assess statistical significance (Bonferroni adjusted for 3 tests).

Supplementary Table 4: Detailed clinical information for 25 individuals with RNU4-2 variants. SNVs are highlighted in pink, and the individual with an alternate indel in blue. Blank spaces indicate that data were not provided.

Supplementary Table 5: Detailed phenotypic information for individuals with the n.64_65insT variant across cohorts.

Supplementary Table 6: Mean expression of U4 genes in prefrontal cortex across all samples in BrainVar.

Supplementary Table 7: Genomic coordinates of, and burden testing results for snRNA genes.

Supplementary Table 8: Sub-regions of snRNA genes identified as depleted of variation and burden testing results in these regions.

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).
2. Nguyen, T. H. D. *et al.* The architecture of the spliceosomal U4/U6.U5 tri-snRNP. *Nature* **523**, 47–52 (2015).
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).
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).
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).
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).
7. Will, C. L. & Lührmann, R. Spliceosome structure and function. *Cold Spring Harb. Perspect. Biol.* **3**, (2011).
8. Guthrie, C. & Patterson, B. Spliceosomal snRNAs. *Annu. Rev. Genet.* **22**, 387–419 (1988).
9. Elsaid, M. F. *et al.* Mutation in noncoding RNA RNU12 causes early onset cerebellar ataxia. *Ann. Neurol.* **81**, 68–78 (2017).
10. Edery, P. *et al.* Association of TALS developmental disorder with defect in minor splicing component U4atac snRNA. *Science* **332**, 240–243 (2011).
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).
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).
13. De novo variant research dataset - Genomics England Research Environment User Guide. https://re-docs.genomicsengland.co.uk/de_novo_data/.
14. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
15. All of Us Research Program Investigators *et al.* The 'All of Us' Research Program. *N. Engl. J. Med.* **381**, 668–676 (2019).
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).
17. 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).
18. Fischbach, G. D. & Lord, C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* **68**, 192–195 (2010).
19. Ramoni, R. B. *et al.* The Undiagnosed Diseases Network: Accelerating Discovery about Health and Disease. *Am. J. Hum. Genet.* **100**, 185–192 (2017).
20. Aggregated variant calls (AggV2) - genomics England research environment user guide. https://re-docs.genomicsengland.co.uk/aggv2/.
21. Charenton, C., Wilkinson, M. E. & Nagai, K. Mechanism of 5' splice site transfer for human spliceosome activation. *Science* **364**, 362–367 (2019).
22. 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).
23. Wilkinson, M. E., Charenton, C. & Nagai, K. RNA Splicing by the Spliceosome. *Annu. Rev. Biochem.* **89**, 359–388 (2020).
24. Griffin, C. & Saint-Jeannet, J.-P. Spliceosomopathies: Diseases and mechanisms. *Dev. Dyn.* **249**, 1038–1046 (2020).
25. 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).
26. 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).
27. Dvinge, H., Guenthoer, J., Porter, P. L. & Bradley, R. K. RNA components of the spliceosome regulate tissue- and cancer-specific alternative splicing. *Genome Res.* **29**, 1591–1604 (2019).
28. 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).
29. Kaplanis, J. *et al.* Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature* **586**, 757–762 (2020).
30. Pray, L. DNA replication and causes of mutation. *Nature education* **1**, 214 (2008).
31. Goriely, A. & Wilkie, A. O. M. Paternal age effect mutations and selfish spermatogonial selection: causes and consequences for human disease. *Am. J. Hum. Genet.* **90**, 175–200 (2012).
32. Verma, B., Akinyi, M. V., Norppa, A. J. & Frilander, M. J. Minor spliceosome and disease. *Semin. Cell Dev. Biol.* **79**, 103–112 (2018).
33. 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).
34. Shen, A. *et al.* U6 snRNA m6A modification is required for accurate and efficient cis- and trans-splicing of C. elegans mRNAs. *bioRxiv* (2023) doi:10.1101/2023.09.16.558044.
35. Gargano, M. A. *et al.* The Human Phenotype Ontology in 2024: phenotypes around the world. *Nucleic Acids Res.* **52**, D1333–D1346 (2024).
36. ICD-10 version:2019. https://icd.who.int/browse10/2019/en.
37. 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).
38. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
39. Yépez, V. A. *et al.* Detection of aberrant gene expression events in RNA sequencing data. *Nat. Protoc.* **16**, 1276–1296 (2021).
40. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

41. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res.* **22**, 2008–2017 (2012).
42. Bedre R. *Reneshbedre/bioinfokit: Bioinformatics Data Analysis and Visualization Toolkit*. doi:10.5281/zenodo.3965241.
43. Markenscoff-Papadimitriou, E. *et al.* A Chromatin Accessibility Atlas of the Developing Human Telencephalon. *Cell* **182**, 754–769.e18 (2020).
44. Lee, J. *et al*. *Kundajelab/atac_dnase_pipelines: 0.3.0*. doi:10.5281/zenodo.156534.
45. 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).