Compound heterozygous mutations in the noncoding RNU4ATAC cause Roifman Syndrome by disrupting minor intron splicing

Daniele Merico1,*, Maian Roifman2,3,4,*, Ulrich Braunschweig5, Ryan K.C. Yuen1, Roumiana Alexandrova1, Andrea Bates6, Brenda Reid6, Thomas Nalpathamkalam1, Zhuzhi Wang1, Bhooma Thiruvahindrapuram1, Paul Gray7, Alyson Kakakios8, Jane Peake9,10, Stephanie Hogarth9,10, David Manson11, Raymond Buncic12, Sergio L. Pereira1, Jo-Anne Herbrick1, Benjamin J. Blencowe5,13, Chaim M. Roifman4,6 & Stephen W. Scherer1,13,14,15

Roifman Syndrome is a rare congenital disorder characterized by growth retardation, cognitive delay, spondyloepiphyseal dysplasia and antibody deficiency. Here we utilize whole-genome sequencing of Roifman Syndrome patients to reveal compound heterozygous rare variants that disrupt highly conserved positions of the RNU4ATAC small nuclear RNA gene, a minor splicesome component that is essential for minor intron splicing. Targeted sequencing confirms allele segregation in six cases from four unrelated families. RNU4ATAC rare variants have been recently reported to cause microcephalic osteodysplastic primordial dwarfism, type I (MOPD1), whose phenotype is distinct from Roifman Syndrome. Strikingly, all six of the Roifman Syndrome cases have one variant that overlaps MOPD1-implied structural elements, while the other variant overlaps a highly conserved structural element not previously implicated in disease. RNA-seq analysis confirms extensive and specific defects of minor intron splicing. Available allele frequency data suggest that recessive genetic disorders caused by RNU4ATAC rare variants may be more prevalent than previously reported.
Roifman Syndrome (OMIM 300258) was first described as a novel association of antibody deficiency, spondyloepiphyseal chondro-osseous dysplasia, retinal dystrophy, poor pre- and postnatal growth, cognitive delay and facial dysmorphism1–5. In spite of some variability, all subjects share remarkably identical dysmorphic, skeletal and immunological features1–4.

It was proposed that Roifman Syndrome might be a novel ciliopathy with immunodeficiency, because of retinal dystrophy and some early and transient bone changes2. X-linked inheritance was also suspected because most reported cases were males1,2,6,7. Candidate gene studies using targeted sequencing were unsuccessful in identifying causal variants.

Here we applied whole-genome sequencing in two affected siblings and exhaustive analysis of coding as well as noncoding variants. We identified rare compound heterozygous variants disrupting highly conserved elements of the small nuclear RNA (snRNA) gene RNU4ATAC (RefSeq NR_023343, OMIM 601428), which is essential for minor intron splicing8–11 and was reported to cause the recessive disorder microcephalic osteodysplastic primordial dwarfism, type I (MOPD1, OMIM 210710) (refs 10,11). Roifman Syndrome is phenotypically distinct from MOPD1 and presents a unique pattern of compound heterozygosity, which was confirmed in four unrelated families by targeted sequencing. About 800 genes have one (or less often more than one) minor intron and thus are dependent on the minor spliceosome for correct splicing9. Since they are involved in important cellular functions (DNA repair and replication, transcription, RNA processing, cell cycle, etc.) their incorrect splicing can alter cell functionality and viability. RNA-seq analysis confirmed specific alterations of minor intron splicing in Roifman Syndrome patients. In addition, we integrated RNA-seq results with other phenotypic evidence to prioritize genes whose splicing alteration is more likely implicated in Roifman Syndrome.

**Results**

Clinical features of Roifman Syndrome patients. We assembled almost all available Roifman Syndrome-affected subjects (six from four unrelated families) for this study (Fig. 1; see Supplementary Tables 1 and 2 and Supplementary Note 1 for a detailed subject description). All individuals shared the following facial features: a markedly long philtrum with a thin upper lip (Fig. 2), a narrow, tubular and upturned nose with hypoplastic alae nasi (Fig. 2), widely spaced eyes with long palpebral fissures and prominent lashes.

The six cases also presented highly characteristic skeleton and immune abnormalities. The proximal epiphyses of the femora were ‘bullet’ shaped or biconvex at an age one would expect them to be ‘squarer’. In addition, all six cases had brachydactyly, while four had transverse palmar creases and clinodactyly of the fifth digit.

While serum immunoglobulin levels were variable, all patients were unable to produce specific antibodies. Circulating B-cell number was on the lower end of the normal range, with mature B cell and memory B-cell numbers within normal ranges. T-cell number and function were completely normal (see Supplementary Table 2 for detailed immunological findings).

Finally, three of six patients had retinal dystrophy with extensive degeneration of the rod and cone systems.

**Whole-genome and targeted sequencing results.** We applied whole-genome sequencing of two affected individuals from kindred 1, to search for putative causal variants in an unbiased and hypothesis-free manner. Variants were prioritized based on sequencing quality, allele frequency in reference databases below 1%, gene product damage potential, zygosity and gene mode of inheritance.

None of the two siblings had any high-quality, rare, damaging homozygous variants.

No X-linked variant passing the prioritization criteria was shared. However, since the X-linked mode of inheritance had been proposed for Roifman Syndrome, we additionally investigated X-chromosome variants found in only one of the two siblings; they either did not have any known implication in human genetic disorders or mouse abnormal phenotypes (genes ARSH and HS6ST2-AS1), or they had modest protein-damaging potential and insufficient match to Roifman Syndrome phenotype (genes AFF2 and SH3KBP1, for more details see Supplementary Data 1, Supplementary Tables 3–5 and Supplementary Note 2).

While a dominant mode of inheritance was highly unlikely, we investigated three high-quality variants with very rare allele frequency (<0.1%) impacting genes with a dominant mode of inheritance and shared by the two siblings (GUCY2D, HTT and RPL11); unsurprisingly, on more detailed review, we found insufficient match to the Roifman Syndrome phenotype and only modest damaging potential (for more details see Supplementary Data 1, Supplementary Table 5 and Supplementary Note 2). Copy number and structural variant findings were also negative (see Supplementary Data 2 and 3 and Supplementary Note 2).

Finally, we reviewed genes with more than one heterozygous variant, thus potentially consisting of a compound heterozygous set. Only one set passed the prioritization criteria and was shared between the two siblings (Supplementary Table 5); corresponding to two heterozygous single-nucleotide substitutions in the autosomal RNU4ATAC gene encoding the highly conserved U4atac snRNA, an essential component of the minor spliceosome4–11. Compound heterozygous variants clustering in similar RNU4ATAC structural elements were subsequently identified by Sanger sequencing in four other Roifman

---

**Figure 1 | Pedigree of the six Roifman Syndrome cases (kindred 1–4).** The pedigrees show the RNU4ATAC compound heterozygous SNVs in the six genotyped cases of Roifman Syndrome; [ = ] indicates no variant detected.
syndrome-affected patients from three ethnically different families (Fig. 1 and Table 1). Analysis of more than 800 whole genome sequences with unrelated conditions revealed no homozygosity or potential compound heterozygosity at disease variant loci or other conserved positions (Supplementary Note 3).

### RNU4ATAC in MOPD1 and Roifman Syndrome

RNU4ATAC homozygous and compound heterozygous variants were recently reported to cause the recessive disorder MOPD1 (OMIM 210710) (refs 10,11). The lethality of MOPD1 revealed the importance of the minor spliceosome, which is required for correct splicing of about 800 human genes carrying minor introns, including many genes involved in essential functions9–11. MOPD1 and recessive isolated familial growth hormone deficiency (caused by compound heterozygous variants in the RNPC3 protein-coding gene)12 are the only known congenital disorders caused by the disruption of a minor spliceosome-specific component, that is, not present in the major spliceosome (for more details see Supplementary Note 4).

Typical MOPD1 is clearly distinguishable from Roifman Syndrome, as it is characterized by pre- and postnatal lethality, severe prenatal microcephaly and brain malformations, intractable epilepsy, short and bowed limbs, absent or sparse hair, neuroendocrine dysfunction and distinct facial features including proptotic eyes, large/prominent nose or downturned nasal tip and micrognathia13,14 (Table 2). Skin and retinal hypopigmentation have been described in a few cases15. Patients with a milder phenotype of MOPD1 have also been recently described15–18; they exhibit severe microcephaly, with poorly developed gyri and agenesis of corpus callosum, and typical dysmorphic features (striking micrognathia, absent eye brows, large prominent nose, dysplastic nails and, in some, agenesis of ear lobes). These features are not shared with Roifman Syndrome. Skeletal changes and eye pigmentation abnormalities in MOPD1 cases are also different from the epiphyseal dysplasia and severe retinal dysplasia typical of Roifman Syndrome. Finally, none of the MOPD1 cases, severe or mild, were reported to have evidence of primary immunodeficiency (Table 2).

## Variant impact on RNU4ATAC structural elements

The RNU4ATAC structural elements perturbed in MOPD1 and Roifman Syndrome suggest a molecular basis for the phenotypic differences between the two syndromes. The U4atac snRNP (snRNA ribonucleoprotein) is required for the formation of the catalytically active minor spliceosomal complex, by loading U6atac onto the U12-containing pre-spliceosomal complex in concert with the U5 snRNP19 (Supplementary Fig. 1). Regions of U4atac that base pair with

### Table 1 | Roifman Syndrome compound heterozygous variants detected in six affected individuals.

<table>
<thead>
<tr>
<th>Subject, kindred</th>
<th>Allele</th>
<th>U4atac snRNA pos</th>
<th>Sub</th>
<th>Freq 1000G</th>
<th>Freq cg1KG</th>
<th>Freq cgW597</th>
<th>PhyloP PMam</th>
<th>MOPD1/Novel</th>
<th>dbSNP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1–2 K1 Pat</td>
<td>13</td>
<td>C&gt;T</td>
<td>0</td>
<td>0</td>
<td>0.0008</td>
<td>2.57</td>
<td>Novel</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>1–2 K1 Mat</td>
<td>37</td>
<td>G&gt;A</td>
<td>0</td>
<td>0</td>
<td>2.63</td>
<td>Novel</td>
<td>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3, K2 Pat</td>
<td>13</td>
<td>C&gt;T</td>
<td>0</td>
<td>0.0008</td>
<td>2.57</td>
<td>Novel</td>
<td>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3, K2 Mat</td>
<td>48</td>
<td>G&gt;A</td>
<td>0</td>
<td>0.0008</td>
<td>2.63</td>
<td>Novel</td>
<td>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4–5, K3 Mat</td>
<td>16</td>
<td>G&gt;A</td>
<td>0</td>
<td>0.0008</td>
<td>2.63</td>
<td>Novel</td>
<td>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4–5, K3 Pat</td>
<td>51</td>
<td>G&gt;A</td>
<td>0.0014</td>
<td>0</td>
<td>1.37</td>
<td>MOPD1</td>
<td>rs188343279</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6, K4 Mat</td>
<td>8</td>
<td>C&gt;T</td>
<td>0</td>
<td>0.0011</td>
<td>2.57</td>
<td>Novel</td>
<td>rs370715569</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6, K4 Pat</td>
<td>118</td>
<td>T&gt;C</td>
<td>0</td>
<td>0</td>
<td>2.12</td>
<td>Novel</td>
<td>—</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Subject, kindred, subject and kindred index; Allele, maternal or paternal allele indication (all variants are compound heterozygous); U4atac snRNA pos, U4atac snRNA position (1 corresponds to the genomic coordinate 122,288,456 on chromosome 2, hg19 reference); Sub, substitution (reference > alternate); Freq 1000G, Freq cg1KG, Freq cgW597, alternate allele frequency in the 1000 Genomes project, and in the internal Complete Genomics control databases based on the 1000 Genomes subset and the Wellderly study (436 and 597 subjects, respectively); PhyloP PMam, UCSC placental mammal PhyloP score of genomic nucleotide conservation (score > 0 corresponds to negative selection); MOPD1/Novel, variant previously reported as causal for MOPD1, or reported for the first time as causing a genetic disorder; dbSNP, matching dbSNP138 record.

---

Figure 2 | Facial, retinal and skeletal features of Roifman Syndrome. Facial dysmorphic features (a–c) include a markedly long philtrum, thin upper lip and down-turned corners of the mouth. Retinal features are displayed for patient 3 at age 4 years: arteriolar attenuation (d, black arrow), wrinkling of the inner limiting membrane (d, short white arrow and e, black arrows) and pigmentary changes (d, long white arrow). Skeletal features are displayed for patients 1, 3 and 5, respectively: the radiographs show the proximal epiphyses of the femora with symmetric delayed ossification, as well as flattening and irregularity.
Table 2 | Roifman Syndrome and MOPD1 phenotypic manifestations.

<table>
<thead>
<tr>
<th>Phenotypic feature</th>
<th>MOPD1</th>
<th>Roifman Syndrome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre- and post-natal lethality</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Absent or sparse hair</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Dysplastic nails</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Severe micrognathia</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Agenesis of ear lobes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Philtrum</td>
<td>Variable</td>
<td>Long</td>
</tr>
<tr>
<td>Thin upper lips</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Nose</td>
<td>Large/prominent, downturned nasal tip</td>
<td>Tubular and upturned</td>
</tr>
<tr>
<td>Retinal changes</td>
<td>Hypopigmentation</td>
<td>Severe degradation of rods and cones</td>
</tr>
<tr>
<td>Skeletal abnormalities</td>
<td>Metaphyseal changes, flat acetalubar root</td>
<td>Spondyloepiphysseal dysplasia</td>
</tr>
<tr>
<td>Immunodeficiency</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Noncompaction of the heart</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Head</td>
<td>Severe prefrontal microcephaly</td>
<td>Mild microcephaly or normal head size</td>
</tr>
<tr>
<td>Structural brain abnormalities</td>
<td>Agenesis of corpus callosum, gyral anomalies, cortical atrophy, enlarged lateral ventricles, hypoplastic frontal lobes, hypoplastic pituitary gland, interhemispheric cysts, cerebellar vermis hypoplasia</td>
<td>None in all cases but in 1, who had partial agenesis of corpus callosum and hippocampal atrophy6</td>
</tr>
<tr>
<td>Intractable epilepsy</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Endocrine dysfunction</td>
<td>Growth hormone deficiency, low prolactin levels</td>
<td>None</td>
</tr>
</tbody>
</table>

U4atac, the stem I and stem II (at the 3’ and 5’ of U4atac, respectively), are separated by an intramolecular stem–loop (the 5’ stem–loop). Another stem–loop is present at the 3’ end of U4atac, followed by a sequence acting as binding site for Sm proteins, which are important for snRNP assembly and import into the nucleus. The stem II, the 5’ stem–loop, the stem I and the Sm-binding site are all enriched in highly conserved nucleotides and mutagenesis experiments have demonstrated their importance for correct minor splicesosomal activity, while part of the 3’ stem is fully dispensable, and other sequence elements have lower conservation20,21 (Fig. 3, see also Supplementary Figs 2 and 3).

Most MOPD1 causal variants cluster in the 5’ stem–loop (U4atac snRNA positions 30, 46 and 50–55), while only a few are found at other elements (U4atac snRNA position 66, connecting the single-strand region to stem I; position 111, part of the 3’ stem–loop; position 124, part of the Sm-binding site)10,11,15–18 (Fig. 3). In contrast, all Roifman Syndrome causal variants identified in the six cases are always compound heterozygous (Fig. 1 and Table 1), with one variant overlapping the MOPD1-implicated 5’ stem–loop critical region (U4atac snRNA positions 37, 48 and 51) or the Sm protein-binding site (U4atac snRNA position 118), while the other variant occurs at highly conserved positions in the stem II (U4atac snRNA positions 8, 13 and 16; Fig. 3). On the basis of the secondary structure and conservation profile of U4atac, we expect the Roifman Syndrome causal variants occurring in the 5’ stem–loop to have a destabilizing effect similar to MOPD1 variants. The alteration of the stem II is, however, unique to Roifman Syndrome (Fig. 3). While MOPD1 5’ stem–loop variants impair binding of the NHP2LI and PRPF3I proteins, stem II variants may affect the PRPF4/PRPF3/PP1H protein complex binding in this region22. Considering the greater severity of MOPD1, it is reasonable to hypothesize that these stem II variants have a weaker effect on minor splicesosomal function.

RNA-seq analysis. To confirm the presence of specific minor splicing alterations, RNA-seq was performed on mononuclear blood cells from kindred 1 (the affected son, that is, patient #2, and the unaffected carrier father and younger male sibling) and from kindred 2 (the affected son, that is, patient #3, and the unaffected carrier father). Sample clustering by gene expression showed perfect separation between the two affected and three unaffected samples (Supplementary Fig. 4).

We used a comprehensive set of curated splicing junctions and major/minor intron coordinates to evaluate splicing alterations, together with a recently developed RNA-seq analysis pipeline (vast-tools)23. We detected very consistent minor intron retention in affected subjects compared with unaffected subjects (median percentage of transcripts with intron retention, percentage intron retention (PIR) 25–40% for affected and PIR 2–4% for unaffected subjects, Wilcoxon two-tailed P value <10−15, Fig. 4a). These splicing changes were highly specific to minor introns, since major intron retention was small for all subjects (median PIR 0–1.5%, Fig. 4a). Moreover, other types of alternative splicing (that is, involving alternative cassette exon, microexons and 5’ or 3’ alternative splice sites) did not display significant difference between affected and unaffected subjects (Fig. 4b).

Interestingly, transcripts of minor intron genes displayed slightly increased (rather than reduced) expression levels in affected subjects, compared with other genes (edgeR24 median log2 expression ratio between affected and unaffected subjects: 0.054 for minor intron genes and –0.024 for other genes, Wilcoxon two-tailed P value = 1.419e-05, Fig. 4c). That is expected to successfully compensate the increased minor intron retention only for a minority of the genes (30%, Fig. 4d). This suggests a compensatory transcriptional upregulation of minor intron genes, which however is not fully successful at restoring optimal levels of correctly spliced transcripts.

Analysis of gene expression and minor intron retention using other RNA-seq analysis methods (cufflinks25, DESeq26 and iReckon27) led to similar findings (Supplementary Tables 6–14, Supplementary Fig. 5 and Supplementary Note 5).

Prioritization of genes altered in Roifman Syndrome. Minor intron genes are overall enriched in functions and phenotypes relevant to Roifman Syndrome (for example, brain and skeletal phenotypes, cell cycle regulation and signalling pathways, see Supplementary Fig. 6). To further prioritize genes whose splicing alteration leads to organ abnormalities observed for Roifman Syndrome, we integrated RNA-seq results based on different analytical methods (gene expression: edgeR, DESeq and cufflinks;
Figure 3 | U4atac snRNA secondary structure elements, positional conservation, MOPD1 and Roifman Syndrome causal variants. Elements of limited or null importance for splicing (dispensable region of 3′ stem-loop) have mutagenesis experiments showing no splicing alteration, are enriched in low conservation and do not present any known disease-causing variant. Elements of variable importance for splicing (3′ stem-loop, except dispensable region; single-strand region; 5′ stem-loop, except critical region) have mutagenesis experiments showing modest or no splicing alteration, or have not been probed by mutagenesis, but meet at least one of these criteria: (a) they present at least one MOPD1 causal variant (typically with reduced severity); (b) they are proximal to a MOPD1 causal variant cluster; (c) structural studies suggest they may have a functional role; these elements have mixed conservation, and only a few variants at more conserved positions may cause splicing alterations. Elements of major importance for splicing (stem II, critical region of the 5′ stem-loop, stem I, Sm protein-binding site) have mutagenesis experiments producing splicing alterations and/or overlap the MOPD1 variant cluster; in addition, they are expected to have a major functional role based on structural studies; finally, they are enriched in highly conserved positions, the majority of which are expected to cause splicing alterations in presence of variation. Positions are labelled as ‘high conservation’ if placental mammal or 100-vertebrate PhyloP ≥1.75, as ‘diverged’ if placental mammal and 100-vertebrate PhyloP are negative, and ‘moderate or low conservation’ otherwise. Parts of U6atac are displayed only in correspondence of U4atac−U6atac duplex structures. U4atac snRNA coordinate 1 corresponds to hg19 coordinate 122,288,456. Classification of MOPD1 causal variants as ‘full severity’ or ‘reduced severity’ is based on a thorough review of MOPD1 literature (for more details, see Supplementary Table 17) and biochemical assays of variant effect; the latter is particularly important for less severe MOPD1 forms presenting compound heterozygosity.

splicing alteration detection: vast-tools, cufflinks and iReckon). We identified 83 genes for which at least two methods detected increased minor intron retention and decreased correctly spliced transcripts (Supplementary Data 4). These genes were annotated for human and mouse phenotypes to further prioritize genes whose minor intron retention is expected to be more relevant to Roifman Syndrome pathogenesis (Supplementary Data 5). We thus identified 30 genes with phenotypes relevant to Roifman Syndrome (developmental, skeletal, immune, neurocognitive and retinal); of these, the most compelling candidates were ALG12, XRCC5 and SMC3. Their expression changes in Roifman Syndrome are summarized in Table 3, showing consensus detection of markedly increased minor intron retention and correctly spliced transcript isoform reduction; finally, for all three genes, minor intron retention is predicted to result into nonsense-mediated decay (Supplementary Data 4). ALG12 is an alpha-1,6-mannosyltransferase implicated in a recessive glycosylation disorder (OMIM 607143) with a phenotypic presentation similar to Roifman Syndrome. XRCC5 is a double-strand break repair gene also important for T-cell and B-cell receptor V(D)J recombination, whose mouse homozygous knockout causes growth retardation, severe combined immunodeficiency and retinal abnormalities. SMC3 is a component of the chromosome cohesion complex implicated in
the dominant Cornelia de Lange syndrome 3 (OMIM 610759), which is characterized by hand and feet abnormalities, and in some instances also mild intellectual disability. We additionally investigated genes whose correctly spliced transcripts are almost completely abrogated, using very stringent thresholds. We thus identified three genes: SLC9A9, WDFY1, and K1.

Figure 4 | Summary of RNA-seq findings. (a) PIR for major (grey) and minor (orange) introns, for kindred 1 and kindred 2 subjects, showing specific minor intron retention in affected subjects compared with unaffected subjects. (b) Affected-unaffected average difference in percentage alternative splicing (dPSI) for different alternative splicing categories (ALTA, alternative 5’ splice site; ALTD, alternative 3’ splice site; EX, cassette exon; INT, major intron retention; INTmin: minor intron retention; MIC, micro-exon); significant deviation from 0 is observed only for major introns (small yet significant reduction in affected) and minor introns (significant increase in affected). (c) Log2 of the mean expression ratio between affected and unaffected, for minor intron genes and all other genes, displaying a slight shift of minor intron genes towards higher expression in affected subjects. (d) Scatterplot showing that increased expression in affected relative to unaffected subjects (y axis, log2 ratio of the mean expression in the two groups) is insufficient to compensate for the increased minor intron retention (x axis, log2 ratio of the mean correctly spliced fraction in the two groups), with 30% of the genes over-compensated and 70% of the genes under-compensated (separated by the red dashed line); the linear correlation between the expression ratio and correctly spliced ratio is negative but modest (grey dashed line).

Table 3 | Prioritized genes whose splicing alteration contributes to Roifman Syndrome.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Expr. FC</th>
<th>VT Aff. MI PIR</th>
<th>VT Unaff. MI PIR</th>
<th>Cuff. MI FC</th>
<th>iReck. MI FC</th>
<th>VT CS FC</th>
<th>Cuff. CS FC</th>
<th>iReck. CS FC</th>
<th>Phen./Funct.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALG12</td>
<td>0.85–0.85</td>
<td>38.5%</td>
<td>8.0%</td>
<td>4.18</td>
<td>NA</td>
<td>0.57</td>
<td>0.62</td>
<td>0.56</td>
<td>Growth, neurodev., immune</td>
</tr>
<tr>
<td>XRCC5</td>
<td>1.07–1.18</td>
<td>28.2%</td>
<td>1.2%</td>
<td>Inf</td>
<td>7.95</td>
<td>0.79</td>
<td>0.99</td>
<td>0.57</td>
<td>Neurodev., immune, retinal</td>
</tr>
<tr>
<td>SMC3</td>
<td>0.92–1.17</td>
<td>31.9%</td>
<td>2.1%</td>
<td>15.25</td>
<td>15.57</td>
<td>0.64</td>
<td>0.85</td>
<td>0.48</td>
<td>Skeletal</td>
</tr>
<tr>
<td>SLC9A9</td>
<td>2.01–3.00</td>
<td>85.4%</td>
<td>2.7%</td>
<td>Inf</td>
<td>NA</td>
<td>0.31</td>
<td>0.00</td>
<td>0.00</td>
<td>Neurodev.</td>
</tr>
<tr>
<td>WDFY1</td>
<td>1.26–1.75</td>
<td>88.6%</td>
<td>12.8%</td>
<td>10.07</td>
<td>10.42</td>
<td>0.17</td>
<td>0.07</td>
<td>0.13</td>
<td>No phen., endosome</td>
</tr>
<tr>
<td>ZCCHC8</td>
<td>1.60–1.88</td>
<td>100%</td>
<td>27.1%</td>
<td>15.82</td>
<td>20.76</td>
<td>0.00</td>
<td>0.12</td>
<td>0.30</td>
<td>No phen., exosome NEXT</td>
</tr>
<tr>
<td>EXOSC1</td>
<td>0.87–0.93</td>
<td>NA</td>
<td>5.7%</td>
<td>Inf</td>
<td>NA</td>
<td>0.78</td>
<td>0.40</td>
<td>0.40</td>
<td>No phen., exosome core</td>
</tr>
<tr>
<td>EXOSC5</td>
<td>0.67–0.71</td>
<td>22.1%</td>
<td>9.1%</td>
<td>3.37</td>
<td>NA</td>
<td>0.57</td>
<td>0.59</td>
<td>0.41</td>
<td>No phen., exosome core</td>
</tr>
</tbody>
</table>

Cuff. CS FC, cufflinks correctly spliced isoform affected/unaffected fold change; iReck. CS FC, iReckon correctly spliced isoform affected/unaffected fold change; Cuff. MI FC, cufflinks minor intron retention isoform affected/unaffected fold change; Expr. FC, overall expression affected/unaffected fold change (interval based on edgeR, DESeq and cufflinks estimates); Gene, official gene symbol; iReck. MI FC, iReckon minor intron retention isoform affected/unaffected fold change; Phen./Funct., gene phenotype and function (note that function is reported only in absence of phenotype information, that is, ‘No phen.’); VT aff. MI PIR, vast-tools minor intron retention percentage in affected; VT unaff. MI PIR, vast-tools minor intron retention percentage in unaffected.
ZCCHC8. SLC9A9 was implicated by one study in autism and epilepsy risk\(^{35}\), while WDFY1 and ZCCHC8 are not implicated in any human genetic disorder or mouse abnormal phenotypes. However, ZCCHC8 is part of the NEXT (nuclear exosome targeting) complex, which is important for guiding the exosome degradation of malformed or byproduct transcripts such as PROMPTs (promoter upstream transcripts)\(^{34}\). Since reduced exosome function may contribute to increased persistence of minor intron retention transcripts, we investigated the splicing expression levels of the exosome core components and found two of them (EXOSC1 and EXOSC5) among the 83 prioritized genes, even though they have not been implicated in human genetic disorder or mouse phenotype yet. Detailed expression changes for these genes are also reported in Table 3.

**Discussion**

Using whole-genome sequencing of two affected siblings and targeted Sanger sequencing of four unrelated families, we have demonstrated that Roifman Syndrome is caused by compound heterozygous single-nucleotide variants (SNVs) in the minor splicingosomal snRNA gene RNU4ATAC, which was already implicated in a distinct and severe congenital disorder, MOPD1. Although both syndromes present growth and multi-system abnormalities, Roifman Syndrome has a different phenotype than MOPD1, even when considering the milder form of the latter. In addition, Roifman Syndrome causal variants always present a characteristic compound heterozygosity pattern: one variant affects RNU4ATAC elements already implicated in MOPD1, while the other variant affects a newly disease-implicated yet highly conserved element, the stem II.

RNA-seq analysis of two affected and three unaffected subjects revealed significantly higher minor intron retention in Roifman Syndrome patients compared with controls, which leads to reduced levels of correctly spliced transcripts for minor intron genes; we have also demonstrated that transcriptional alterations are highly specific of minor introns. Perhaps surprisingly, Roifman Syndrome patients did not exhibit overall reduced abundance of minor intron gene transcripts, as expected from a cell line model of short-term response to reduced minor spliceosome capacity\(^{35}\). On the contrary, a mild increase was detected, yet insufficient to fully compensate the splicing alterations. This finding is compatible with other reports showing that transcripts retaining minor introns can accumulate in the cell without being fully cleared by degradation machinery\(^{6,37}\); it may be further explained by partially reduced exosome functionality\(^{38}\), since two of its core components (EXOSC1 and EXOSC5) and one component of the nuclear exosome targeting complex (ZCCHC8) have reduced correctly spliced transcripts. In addition, it is also possible that a feedback response drives increased transcription for incorrectly spliced minor intron genes.

While Roifman Syndrome and MOPD1 are extremely rare, recurrent spontaneous abortions or congenital disorders with a broader phenotypic spectrum may be caused by homozygous or compound heterozygous variants altering any of the RNU4ATAC structural elements critical for splicing, with an estimated prevalence up to 1 in 30,000 pregnancies (Supplementary Note 3). Since RNU4ATAC is not targeted by many commercially available exome capture kits (Supplementary Figs 7–9 and Table 15), its contribution to Mendelian disorders may have been missed in other studies. This may also be the case for other noncoding minor spliceosome snRNA genes (RNU6ATAC, RNU11 and RNU12), which, unlike their major spliceosome snRNA counterparts, are present at single loci in the genome and can thus act as recessive Mendelian disease genes.

**Methods**

**Patient information.** Patients or legal guardians provided informed consent in accordance with our Primary Immunodeficiency Registry & Tissue Bank protocol, Research Ethics Board Number 100000598. Patients 1, 2, 3, 4 and 5 were consented in our facility; patients 1, 2 and 3 have included consent to publish photos. Patient 6 signed our institutional consent for genetic analysis as well as signing the Nature informed consent.

**Whole-genome sequencing of kindred 1.** The whole genomes of the two affected siblings from kindred 1 were sequenced using the Complete Genomics platform\(^{39}\). The concentration of genomic DNA sample was measured by picogreen in triplicates and DNA quality was checked on a 2% agarose gel. About 11 μg of DNA was submitted to Complete Genomics for whole-genome sequencing. Complete Genomics employs high-density DNA nanobars that are concatamers of mate pairs with ~500 bp long; these identification is performed using proprietary non-sequenced, unchained read technology known as combinatorial probe-anchor ligation. Each mate pair includes 35 nucleotides of genomic DNA sequence as well as adaptor sequences required for combinatorial probe-anchor ligation sequencing; the average mate gap length is 300 bp. For both subjects, >97% of the genome was covered at depth ≥5 by uniquely aligned reads (Supplementary Table 3).

The following variants were called by the Complete Genomics pipeline (version 2.0.2, human genome reference hg19) (ref. 40): (i) SNVs and small insertions/deletions (indels); (ii) structural variants (based on abnormal junction and discordant mate pair clusters, with size typically 50–75,000 bp); (iii) copy number variations (based on normal junction coverage, with size typically ~2,000 bp). Whole-genome variant data are available on request (please contact the corresponding authors).

**SNV and indel annotation and prioritization.** Complete Genomics SNV and indels were annotated using a custom pipeline based on Annovar (August 2013 version)\(^{31}\). ReSeq gene models (downloaded from UCSC 2013 February 12), publicly available as well as internally maintained databases for allele frequency (http://www.Complete Genomics\(^{42}\), NHLBI-ESP\(^{43}\) and internal Complete Genomics control databases), genomic conservation (UCSC PhyloP and phastCons for placental mammals and 100 vertebrates\(^{44}\)) and variant impact predictors (SIFT\(^{45}\), PolyPhen2 (ref. 46), MutationAssessor\(^{47}\) and CADD\(^{48}\)). Please see the Supplementary Note 6 for a detailed description of the annotation fields and database versions.

Annotated variants mapping to coding or noncoding exonic sequence were further prioritized according to these criteria: (i) sequencing quality; (ii) allele frequency (restricting to rare variants); (iii) conservation and predicted impact (restricting to variants potentially damaging the gene product); (iv) variant pathogenic effect, as reported by disease variant databases; (v) zygosity and genomic mode of inheritance; and (vi) human disease and mouse abnormal phenotypes in which a gene is known to be implicated. Only high-quality, rare and potentially damaging variants were prioritized; while homozgyous and potential compound heterozygous variants were reported also for genes not implicated in human genetic disorders, heterozygous variants were reported only for genes implicated in a known dominant disorder according to Human Phenotype Ontology\(^{49}\) or Clinical Genomics Database\(^{50}\). A detailed description of the prioritization rules can be found in the Supplementary Note 7.

**Copy number variant (CNV) annotation.** Copy number gains and losses, reported by the Complete Genomics pipeline in the ‘cnvSegments’ DiploidBeta file, were separately annotated for frequency based on 50% reciprocal overlap with CNVs called in 54 unrelated control samples from the Complete Genomics diversity panel (pipeline version 2.0), and overlap with CNVs from the Database of Genomic Variants (November 2010 and March 2013 versions)\(^{51}\). CNVs were also annotated for overlapping gene transcripts and exons (ReSeq, downloaded March 2013). CNVs overlapping at least one genic exon and not found in the Complete Genomics diversity panel were inspected manually.

**Minor intron identification.** The U12db (ref. 52) was previously used as the authoritative source for minor introns\(^{10,12}\). However, it was last updated in January 2007 and it is based on the outdated genome build hg17/NCBI35. For this reason, we preferred to use U12db major and minor intron sequences to extract splicing consensus sequences, construct position-specific score matrices (PSSMs)\(^{23}\), and re-classify up-to-date hg19 ReSeq introns.

The intron 5 initial 15 bp (including the 5′ recognition splicing consensus sequence) and the intron 3′ terminal 39 bp (including the branching site splicing consensus sequence) were downloaded from U12 db for 487 GTAG minor introns, 208 ATAC minor introns and 82 major introns. The initial and terminal intronic discordant high-scoring exonic high-conserved GT/AT and AG/AC sequences were removed, as they are not highly discriminatory of major and minor introns. The resulting sequences were merged into 50-bp-long sequences, and MEME 4.9.1 (ref. 54) was used to identify over-represented sequences and construct corresponding position probability matrix. As expected, for minor introns, two over-represented sequences, a 5′ recognition consensus (ATCCCT, followed by less-conserved bases) and the branching site (TTTCT/T/A/G/AC, surrounded by less-conserved bases); for
major introns, we identified only the 5’ recognition consensus sequence (AAGTTT), while the branching site consensus was too degenerate and no over-represented sequence was found. These consensus sequences had good correspondence to curated ones.

We next created log-odds (LOD) PSSM by dividing each nucleotide probability by 0.25, while this is based on the simplistic assumption that the appearance of A, C, G or T is equally likely, we use the LODs to discriminate intronic sequences with high score either for the minor or for the major consensus sequences, thus the assumption is acceptable. In case the frequency of the base was 0, we assigned the LOD score of –100.

We scanned the 5’ and 3’ sequences of known introns (hg19 RefSeq, downloaded from UCSC in April 2014) and calculated the following PSSM match scores (representing the LOD-transformed probability of observing the intronic sequence based on the LOD PSSM matrix): (a) for each intronic 5’ sequence we calculated the minor intron 5’ recognition match score and the major intron 5’ recognition match score; (b) for each intronic 3’ sequence we calculated the minor intron 3’ recognition match score and the major intron 3’ recognition match score. In particular, for the intron 5’ sequence of a new intron, we calculated the match score based on the 13 bases after the initial conserved dinucleotide, as in the MEME analysis we always observed the over-represented consensus sequence at that position; for the intron 3’, we slide a 19-base window over the last 40 bases (excluding the terminal conserved dinucleotide), calculated the match score for each window, and selected the maximum match score, as the MEME analysis showed a variable position of the corresponding over-represented sequence in the 3’ sequence. The match score represents the LOD-transformed probability of observing the intronic sequence based on the LOD PSSM matrix, and was defined as:

\[
S = \sum_{i=1}^{m} LODe_{i,j} \cdot seq_{pssm}(i,j) \cdot seq_{obs}(i,j)
\]

where \(i\) represents the i-th position in the scanned sequence, and \(LODe_{i,j} \cdot seq_{pssm}(i,j) \cdot seq_{obs}(i,j)\) represents the LOD score value at the i-th position of the PSSM for the nucleotide observed at position \(i\) in the scanned sequence; this notation is based on PSSM with columns (j index) corresponding to the four nucleotides \(seq_{pssm} = \{A, C, T, G\}\) and PSSM rows (i index) corresponding to specific positions.

Comparing the putative minor introns found in this analysis with the minor introns reported in U12db, we found that requiring a score < 2 for the 5’ match to the 5’ minor intron recognition PSSM and for the 3’ match to the minor intron-branching site PSSM was effective at discriminating minor introns from major introns. This lead to identification of 822 unique introns and 744 unique minor intron genes.

RNA-seq extraction and sequencing. RNA-seq was performed on mononuclear blood cells from three members of kindred 1 (affected son, corresponding to patient #2, unaffected father and sibling) and two members of kindred 2 (affected son, corresponding to patient #3 and unaffected father).

Total RNA sample quality control was performed using an Agilent Bioanalyzer 2100 RNA Nano chip following the Agilent Technologies’ recommendation. RNA library preparation was performed following the Illumina TruSeq RNA Sample Preparation V2 Guide (Rev. D, September 2012). Briefly, 1 μg of total RNA was used as the input material; poly(A) mRNA were enriched with oligo DT beads and the enriched RNA were fragmented for 6 min at 94 °C. Fragmented RNA was converted to double-stranded cDNA; end-repaired and adenylated at the 3’ end to create an overhang A to allow for ligation of TruSeq adapters with an overhang T; the fragments were cleaned up on a Bioanalyzer 2100 DNA High Sensitivity chip (Agilent Technologies). The RNA quality control was performed using the R package CummeRbund v2.6.1, based on the fragments per kilobase of transcript per million mapped fragments (FPKM) estimates generated by cufflinks.

RNA-seq bioinformatic alternative splicing analysis. Alternative splicing was analyzed using R package (version 0.2.19), a multi-module computational pipeline uses a broad range of evidence sources (RNA-seq, EST and cDNA, gene annotations and evolutionary conservation) to define splice junctions for the human transcriptome (hg19). That results in 258,603 potential alternative splicing events, comprising 74,233 cassette exon events, 478 microexons, 12,677 alternative 5’- and 3’-splice site events, 18,094 alternative 5’- and 3’-splice site events (including 66 minor introns). Of note, introns were considered even if they have never previously been detected as retained.

Read-1 and read-2 from RNA-seq paired-end reads were separately processed to produce two pairs of 50 bp fragment reads; this step is required for optimal mapping to junctions and counting. Whenever presence of adaptor sequences or low-quality stretches produce trimmed read lengths below 100 bp, the two 50-bp fragment pairs can overlap partially. Fragments mapping to multiple sites in the human genome, or to sites with overlapping transcribed sequence belonging to different genes, were discarded. Remaining fragments were aligned to libraries of exon-exon and exon-intron junction sequences, using bowtie with settings –m 1 –v 1; the 50-bp fragment pairs were mapped independently, while tracking from which original 100-bp read pair they were derived. We then counted how many original 100-bp read pairs were represented at each junction, and counts were finally normalized by the number of uniquely mappable 50-mer positions in each junction sequence. Per cent-spliced-in or PIR scores were calculated as previously described33,35–37.

Raw output from this pipeline was filtered using associated quality information. For cassette exons, we required a coverage score of ‘SOK’, ‘OK’ or ‘LOW’ (roughly corresponding to a minimum number of mapped reads per junction of 100, 20 or 10, respectively), and a junction score ≥ 1 (score of 0 for ‘OK’ or ‘LOW’ (corresponding to a ratio of numbers of reads mapping to the upstream and downstream junctions of less than twofold, or between two- and fivefolds, respectively; for details, see https://github.com/vastgroup/vast-tools). For alternative 5’- and 3’-splice sites and microexons, we required a coverage score of ‘SOK’, ‘OK’ or ‘LOW’. For intron retention events, we required counts of ≥ 15 total reads per event and a junction balance binomial test P value > 0.05. For all alternative splicing events, only events detected in at least two individuals were retained for analyses comparing affected and unaffected individuals. Differential PIR/PIR between affected and unaffected individuals were calculated as differences between averages in these groups.

Fold differences in correctly spliced fraction with respect to minor intron retention events (Fig. 4d) were calculated as log2((100 – PIRaffected)/ (100 – PIRunaffected)).

Predictions of introns whose retention triggers nonsense-mediated decay were derived from a previous publication53. The R package edgeR (ref. 24) was used to assess differential expression for the over/under-compensation assessment, please see next section for more details.

Note that, for different RNA-seq methods, the term ‘fold change’ means the ratio between the expression level (or other quantitative measure) between two conditions (affected/unaffected when not explicitly defined).

RNA-seq bioinformatics gene and isoform expression. Trimmed reads were aligned to the human reference sequence (hg19) using TopHat v2.0.10 (ref. 58) (emap option of the trimmed read pairs were concordantly aligned). For the edgeR differential gene expression analysis, read counts for gene model exons were extracted from TopHat alignments using HTseq v2.6.4 (ref. 59), with the ‘intersection strict’ setting (‘intersection strict’ ensures that reads only partially overlapping exons are not counted).

For the edgeR (ref. 24) analysis, library size normalization factors were calculated using the method ‘trimmed mean of M-values’ (TMM) and differential analysis was performed using the generalized linear models functions with default settings, with a design matrix specifying phenotype (affected and unaffected); at a false discovery rate (FDR) < 0.05, there were 500 differentially expressed genes. For the DESeq differential gene expression analysis, we used a common Gaussian distribution for testing differential expression.

Cufflinks/cuffmerger v2.0.2 (ref. 25) were used to assemble transcript isoforms from aligned reads, and cuffdiff to test for differential gene expression, differential transcript isoform expression and differential splicing output; the cufflinks pre-mRNA-fraction and min-isoform-fraction parameters were relaxed to 0.05 to maximize sensitivity. iReckon v1.0.8 (ref. 27) was also used to assemble transcripts from aligned reads and assess intron retention.

To confirm segregation of case and control samples, clustering was performed using the R package Gmmelbund v2.6.1, based on the per kilobase of exon per million mapped fragments (FPKM) estimates generated by cufflinks.

Prioritization of genes altered in Roifman Syndrome. For the genes with minor introns (744 genes), we integrated results from different tools into a final score,
indicating how many methods supported the presence of minor intron splicing retention and reduced correctly spliced transcripts.

For DESeq, edgeR and cufflinks differential gene expression analysis, we imported the log2 affected/unaffected fold change, the nominal \( P \) value and the FDR \( q \) value; for DESeq and cufflinks, we also imported the mean normalized counts (DESeq) and FPKM (cufflinks) for the two conditions (affected and unaffected). We found DESeq and edgeR to be highly correlated (Spearman rho of log fold change: 0.954 and 0.936 respectively; Spearman rho of nominal \( P \) value: 0.990), although edgeR found more genes significant for differential expression (at FDR \( < 0.2 \), edgeR: 54 minor intron genes, DESeq: 12); DESeq and edgeR were also correlated to cufflinks (Spearman rho of log fold change: 0.916 and 0.907, respectively; Spearman rho of nominal \( P \) value: 0.697 and 0.581, respectively).

Vast-tools splicing results were restricted to minor intron retention events, and whenever a gene had more than one minor intron (48/666 minor intron genes in the vast-tools splicing analysis), we imported only the one with the largest difference between affected and unaffected minor intron PIR. We used the minor intron PIR to calculate the log2 affected/unaffected fold change in correctly spliced transcript as: log2 (\( y \) / \( C0 \)). We defined the percentage of genes with over- or under-compensation in Fig. 4d by comparing the log2 affected/unaffected fold change in expression from edgeR (\( y \), following Fig. 4d) to the log2 affected/unaffected fold change in correctly spliced transcript from vast-tools (\( C0 \), following Fig. 4d): when \( y < 1 \), there is under-compensation (that is, deaths) correctly spliced transcript; percentages were reported with respect to minor intron genes without missing values for vast-tools and edgeR (222744 genes). For the combination of edgeR and vast-tools, we considered genes having a minor intron splicing alteration when log2 (vast-tools fold change correctly spliced) + log2 (edgeR expression fold change) < log2(1/1.2).

For cufflinks transcript isoform analysis, we defined 'correctly spliced' the isoform recognized by iReckon as 'known', and 'minor intron retention' the isoforms recognized by iReckon as 'intron retention' and including a minor intron; we then calculated total expression levels (reads per kilobase of exon per million isoforms recognized by iReckon as 'intron retention' and including a minor intron; FPKM of genes with over- or under-compensation in Fig. 4d by comparing the log2 affected/unaffected fold change in expression from edgeR (\( y \), following Fig. 4d) to the log2 affected/unaffected fold change in correctly spliced transcript from vast-tools (\( C0 \), following Fig. 4d): when \( y < 1 \), there is under-compensation (that is, deaths) correctly spliced transcript; percentages were reported with respect to minor intron genes without missing values for vast-tools and edgeR (222744 genes). For the combination of edgeR and vast-tools, we considered genes having a minor intron splicing alteration when log2 (vast-tools fold change correctly spliced) + log2 (edgeR expression fold change) < log2(1/1.2) and minor intron retention the isoforms recognized by iReckon as 'intron retention' and including a minor intron; we then calculated total expression levels (reads per kilobase of exon per million isoforms recognized by iReckon as 'intron retention' and including a minor intron; FPKM of genes with over- or under-compensation in Fig. 4d by comparing the log2 affected/unaffected fold change in expression from edgeR (\( y \), following Fig. 4d) to the log2 affected/unaffected fold change in correctly spliced transcript from vast-tools (\( C0 \), following Fig. 4d): when \( y < 1 \), there is under-compensation (that is, deaths) correctly spliced transcript; percentages were reported with respect to minor intron genes without missing values for vast-tools and edgeR (222744 genes). For the combination of edgeR and vast-tools, we considered genes having a minor intron splicing alteration when log2 (vast-tools fold change correctly spliced) + log2 (edgeR expression fold change) < log2(1/1.2).

**Acknowledgements**

D.Me., B.T., S.L.P. and J.-A.H. are supported by Science and Technology Innovation Centre funding from Genome Canada/Ontario Genomics Institute. U.B. is supported by a Human Frontier Science Program long-term fellowship. B.I.B. holds the Banbury Chair in Medical Research at the University of Toronto, and part of the work on the RNA-seq analysis methodology was supported by a CIHR grant held by B.I.B. C.M.R. is supported by the Canadian Center for Primary Immunodeficiency, the Jeffrey Modell Foundation and Immunodeficiency Canada. S.W.S. holds the GlaxoSmithKline-Canadian Institutes of Health Research Endowed Chair in Genome Sciences at the Hospital for Sick Children and University of Toronto. We thank The Centre for Applied Genomics (TCAG) at The Hospital for Sick Children for infrastructure support, funded by Genome Canada/ Ontario Genomics Institute and the Canada Foundation for Innovation (CFI).

**Author contributions**


**Additional information**

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/