

1 Title

2 A systematic analysis of splicing variants identifies new diagnoses in the 100,000 Genomes Project.

3 Authors

4 Alexander J.M. Blakes^{1,2}, Htoo Wai¹, Ian Davies³, Hassan E. Moledian¹, April Ruiz⁴, Tessy Thomas⁴,
5 David Bunyan^{5,6}, N Simon Thomas^{5,6}, Christine P. Burren^{7,8}, Lynn Greenhalgh⁹, Melissa Lees¹⁰, Amanda
6 Pichini^{11,12}, Sarah F. Smithson¹¹, Ana Lisa Taylor Tavares^{12,13}, Peter O'Donovan¹², Andrew G.L.
7 Douglas^{14,1}, Genomics England Research Consortium, Splicing and Disease Working Group, Nicola
8 Whiffin¹⁵, Diana Baralle^{1,4,*}, Jenny Lord^{1,*,**}

9 * Joint senior author

10 ** Corresponding author: jenny.lord@soton.ac.uk

11 Affiliations

12 1. Human Development and Health, Faculty of Medicine, University of Southampton, Southampton,
13 UK; 2. National Heart and Lung Institute, Faculty of Medicine, Imperial College London, London, UK; 3.
14 Cancer Sciences, Faculty of Medicine, University of Southampton, Southampton, UK; 4. Wessex
15 Clinical Genetics Service, Princess Anne Hospital, Southampton, UK; 5. Wessex Regional Genetics
16 Laboratory, Salisbury District Hospital, Salisbury, UK; 6. School of Medicine, University of
17 Southampton, Southampton, UK; 7. Department of Paediatric Endocrinology and Diabetes, University
18 Hospitals Bristol and Weston NHS Foundation Trust, Bristol, UK; 8. Bristol Medical School, Dept of
19 Translational Health Sciences, University of Bristol, Bristol, UK; 9. Liverpool Centre for Genomic
20 Medicine, Crown Street, Liverpool, UK; 10. North East Thames Regional Genomics Service, Great
21 Ormond Street Hospital, London, UK; 11. Department of Clinical Genetics, University Hospitals Bristol
22 and Weston NHS Foundation Trust, Bristol, UK; 12. Genomics England, Dawson Hall, Chard House Square,

23 London, UK; 13. Cambridge University Hospitals NHS Foundation Trust, Cambridge Biomedical
24 Campus, Hills Road, Cambridge, UK; 14. Oxford Centre for Genomic Medicine, Oxford University
25 Hospitals NHS Foundation Trust, Oxford, UK; 15. Wellcome Centre for Human Genetics, University of
26 Oxford, Oxford, UK

27 Abstract

28 Genomic variants which disrupt splicing are a major cause of rare genetic disease. However, variants
29 which lie outside of the canonical splice sites are difficult to interpret clinically. Here, we examine the
30 landscape of splicing variants in whole-genome sequencing data from 38,688 individuals in the
31 100,000 Genomes Project, and assess the contribution of non-canonical splicing variants to rare
32 genetic diseases. We show that splicing branchpoints are highly constrained by purifying selection,
33 and harbour damaging non-coding variants which are amenable to systematic analysis in sequencing
34 data. From 258 *de novo* splicing variants in known rare disease genes, we identify 35 new likely
35 diagnoses in probands with an unsolved rare disease. We use phenotype matching and RNA studies
36 to confirm a new diagnosis for six individuals to date. In summary, we demonstrate the clinical value
37 of examining non-canonical splicing variants in participants with unsolved rare diseases.

38 Introduction

39 Improved diagnosis of rare genetic diseases remains a significant clinical and research challenge¹.
40 Diagnostic yields in individuals with rare diseases remain below 50%, despite extensive investigations
41 including whole-genome sequencing². The accurate interpretation of genomic variants in existing
42 sequencing data presents an important opportunity to narrow the diagnostic gap³.

43 Splicing is the process by which introns are removed from a pre-mRNA primary transcript. Almost all
44 human protein coding genes are spliced, and disruption of splicing is a major cause of rare genetic
45 diseases⁴. The improved interpretation of splicing variants is therefore a major opportunity to improve
46 clinical outcomes for individuals with undiagnosed rare disease⁵.

47 Already, “canonical splice site” (CSS) variants within 2bp of an exon-intron junction are widely
48 annotated as “loss of function” (LoF) variants, and are known to be strong diagnostic candidates in
49 “loss-of function” disorders⁶. The contribution of non-canonical splicing variants to rare disease is also
50 becoming increasingly recognised⁷. Up to 27% of pathogenic *de novo* splicing variants in exome-
51 sequencing data are found in non-canonical positions⁸. Several studies⁷⁻⁹ have developed the concept
52 of a “near-splice” region, usually tens of base pairs around an exon-intron junction, which contains
53 many conserved splicing motifs.

54 However, near-splice variants are under-reported in clinical databases⁸, and no standards exist for
55 their interpretation. Furthermore, variants distal to the near-splice region, including branchpoint
56 variants and deep intronic variants, can also disrupt splicing, and their overall contribution to rare
57 disease is unknown. Individual instances of pathogenic branchpoint variants have been previously
58 described^{10,11}, but they have not been systematically characterised in a large rare disease cohort.

59 Recently, large population genomic datasets have provided the statistical power necessary to measure
60 constraints on genetic variation within human populations. One powerful metric which uses this
61 approach is the mutability-adjusted proportion of singletons (MAPS)¹², which identifies classes of

62 variation which are subject to purifying selection, and are therefore likely to be deleterious. MAPS has
63 previously been calculated in many contexts, including for near-splice positions in the Exome
64 Aggregation Consortium (ExAC)⁸, and for upstream start-codon-creating variants in Genome
65 Aggregation Database (gnomAD)¹³.

66 Recent advances in computation and artificial intelligence have led to the development of numerous
67 *in silico* predictors for the prioritization of splicing variants¹⁴. For example, SpliceAI is a machine
68 learning tool which robustly predicts splice sites and splice-disrupting variants¹⁵, and out-performs
69 other algorithms in predicting splicing consequences from sequence data¹⁶. However, in clinical
70 variant interpretation, well-validated functional assays have greater weight than *in silico* predictions
71 of variant effect⁶, and functional validation of most splicing variants is still required to confirm a
72 molecular diagnosis.

73 Here, we perform a systematic analysis of splicing variants in whole-genome sequencing data from
74 38,688 individuals in the Rare Disease arm of the 100,000 Genomes Project (100KGP)¹⁷. We evaluate
75 the contribution of canonical, near-splice, and splicing branchpoint variants to rare genetic diseases
76 in this cohort. We show that splicing branchpoints harbour deleterious non-coding variants which are
77 amenable to systematic analysis in WGS data. We used a gene-agnostic approach to prioritise 258 *de*
78 *novo* splicing variants in families affected by a rare genetic disorder. Of these, at least 84 were already
79 considered to be diagnostic, and we identified an additional 35 variants which are likely to be
80 diagnostic given the available molecular, phenotypic, and *in silico* data. We confirmed a new molecular
81 diagnosis for six participants, including four out of five participants for whom RNA studies were
82 performed. Ultimately, we demonstrate the clinical and diagnostic value of examining both canonical
83 and non-canonical splicing variants in unsolved rare diseases.

84 Subjects and Methods

85 Cohort, sequencing, and tiering

86 This analysis was performed on whole-genome sequencing data from 38,688 participants in the Rare
87 Disease arm of the 100,000 Genomes Project¹⁸. These comprised 26,660 unaffected parents of rare
88 disease probands, and 12,028 participants (offspring) for whom trio WGS data was available. Only
89 participants for whom WGS data was aligned to GRCh38 were included in this study. Parents affected
90 by a rare genetic disease were excluded from the analysis of variant constraint (see below). Otherwise,
91 participants were not selected or stratified by any other criterion. The sequencing and bioinformatic
92 pipelines, as well as the “tiering” framework for variant prioritisation, have been previously
93 described¹⁷. Briefly, variants meeting filtering criteria and falling within applied virtual gene panels
94 were annotated as tier 1 (loss of function or *de novo* protein-altering variants), tier 2 (other variant
95 types eg missense, with correct mode of inheritance), or tier 3 (all other filtered variants). For example,
96 CSS variants in an appropriate gene panel are annotated as tier 1.

97 Defining CDS exons and near-splice positions

98 We identified coding sequences in high-confidence protein-coding transcripts from GENCODE v29¹⁹
99 (GRCh38) using the following filtering criteria: feature type = “CDS”, gene_type = “protein_coding”,
100 transcript_type = “protein_coding”, level != “level 3”, and tag = “CCDS”, “appris_principal_1”,
101 “appris_candidate_longest”, “appris_candidate”, or “exp_conf”. 401,314 CDS features (207,548
102 unique) met these criteria. Only autosomal CDS were included in the subsequent analyses. UTR and
103 other non-coding exons were not included in this analysis.

104 For each CDS feature, we annotated individual genomic positions with their positions relative to a
105 splice donor or acceptor site, excluding any sites with conflicting annotations. We defined the near-
106 splice region around the acceptor site as 25bp of intronic sequence (acceptor -25 to acceptor -1), and
107 11bp of exonic sequence (acceptor 0 to acceptor +10). Around the donor site, we included 11bp of

108 exonic sequence (donor -10 to donor 0) and 10 bp of intronic sequence (donor +1 to donor +10).
109 9,588,491 distinct near-splice positions were identified.

110 phyloP

111 We annotated each near-splice position with phyloP scores from multiple alignments of 99 vertebrate
112 species to the human genome (phyloP 100-way)²⁰ with pyBigWig, an open-source Python package²¹,
113 using BigWig files downloaded from the UCSC genome browser (hg38)^{22,23}.

114 SpliceAI

115 For every possible near-splice SNV in our positions of interest (i.e. all three possible single base
116 changes at each of the 9,588,491 positions), we annotated the predicted effect on splicing with
117 SpliceAI¹⁵. We annotated variants with pre-computed genome-wide SpliceAI v1.3 scores (available via
118 <https://github.com/Illumina/SpliceAI>) using BCFtools v1.9²⁴. A SpliceAI annotation was available for
119 28,265,193 variants (98.2% of 28,765,473 possible variants).

120 Aggregate SpliceAI scores for each near-splice position were calculated as the mean probability that
121 any variant at this position disrupts splicing. The probability that a given variant disrupts splicing was
122 calculated as the probability (P) of any one of the SpliceAI-predicted splicing events occurring, (i.e. 1 -
123 probability of *no* events occurring). The predicted splicing events are acceptor gain (AG), acceptor loss
124 (AL), donor gain (DG), or donor loss (DL), giving:

$$125 \quad P = 1 - ((1 - (AG)) * (1 - (AL)) * (1 - (DG)) * (1 - (DL)))$$

126 Mutability-adjusted proportion of singletons

127 In addition to the near-splice SNVs identified above, we also determined the set of all possible coding
128 SNVs in our exons of interest. These were annotated with the reference base for each position
129 (GRCh38, GenBank assembly accession GCA_000001405.15) and its immediate sequence context (1bp
130 either side) with bedtools version 2.27.1²⁵.

131 We annotated every possible coding SNV within our exons of interest with the Variant Effect Predictor
132 (VEP) version 99²⁶. For each variant, only the consequence in one transcript (typically the canonical
133 transcript, as determined by VEP's “–pick” flag), was used. Only synonymous, missense, and nonsense
134 variants were included in the subsequent analysis. Synonymous variants within a near-splice region
135 were classed as near-splice variants for the MAPS calculation and were excluded from the synonymous
136 variant set. Missense variants within a near-splice region were excluded from the analysis altogether.
137 Nonsense variants within a near-splice region were classed as nonsense variants and excluded from
138 the near-splice variant set.

139 We interrogated whole genome sequencing data from 26,660 unaffected parents in the Genomics
140 England (GEL) Rare Disease cohort for SNVs overlapping the near-splice and coding positions defined
141 above using BCFtools v1.9²⁴. Only variants passing all filters (see [https://research-
142 help.genomicsengland.co.uk/display/GERE/aggV2+Details](https://research-help.genomicsengland.co.uk/display/GERE/aggV2+Details)) within the GEL aggregated multi-sample
143 VCF were included. We identified 915,024 synonymous, 1,965,441 missense, 53,825 nonsense, and
144 672,528 near-splice variants, and calculated allele counts across the 26,660 unaffected parents for
145 each variant.

146 We calculated MAPS with custom Python scripts, adapting code written in R by Patrick J. Short²⁷
147 (<https://github.com/pjshort/dddMAPS>). We used the mutation rate of a given trinucleotide context
148 calculated by Samocha et al²⁸. The proportion of singletons for each position was adjusted for the
149 mutability of the immediate sequence context using a linear model trained on synonymous variants
150 within the same exons.

151 **Branchpoints**

152 Splicing branchpoint positions were identified by LaBranchoR, a machine-learning tool trained on
153 experimentally-validated branchpoints, which accurately identifies at least one branchpoint for the
154 majority of introns genome-wide²⁹. Pre-computed LaBranchoR scores are publicly available for every

155 position 1-70bp upstream of a splice acceptor (GENCODE v19, hg19)²⁹. For each intron, the highest
156 scoring position was annotated as the branchpoint (BP), totalling 195,863 putative branchpoints.

157 We converted each branchpoint to hg38 coordinates using the UCSC Liftover tool³⁰. We annotated
158 five positions upstream (-5 to -1) and five positions downstream (+1 to +5) of each branchpoint, as
159 well as every possible SNV at each of these positions using custom Python scripts. phyloP scores for
160 each position, and SpliceAI scores for each variant, were determined as above. We calculated the
161 MAPS statistic for these branchpoint positions in the same cohort of participants as described above.
162 Comparison of MAPS scores in branchpoint positions was made to the same set of coding variants as
163 described above.

164 **Statistics**

165 The null hypotheses that near-splice and branchpoint MAPS scores did not significantly differ from
166 synonymous variants were tested with two-sided Chi squared tests of the observed vs the expected
167 number of singletons in each variant class. In order that the synonymous MAPS did not equal zero, all
168 MAPS scores were first corrected by addition of the synonymous unadjusted proportion of singletons.
169 For each variant class the “observed” proportion of singletons was taken as the number of alleles
170 multiplied by the corrected MAPS score for that variant class. The “expected” number of singletons
171 was taken as the number of alleles multiplied by the corrected MAPS score for synonymous variants.
172 Multiple testing was accounted for by Bonferroni correction: 79 tests at alpha = 0.05 gave a
173 significance threshold of $<6.3 \times 10^{-4}$.

174 ***De novo* variants**

175 *De novo* mutations (DNMs) overlapping near-splice positions were identified from a set of 1,004,599
176 high confidence *de novo* calls in 13,949 trios from 12,609 rare disease families. The annotation pipeline
177 used to identify these variants is publicly available³¹. Briefly, a multisample VCF for each trio was
178 annotated for putative DNMs using custom scripts. Putative DNMs were then filtered by a series of

179 “Global”, “Base”, and “Stringent” filters (see reference³¹). Unless otherwise stated, our analyses were
180 performed on DNMs aligned to GRCh38 (870,559 DNMs in 12,028 trios).

181 At the outset of this project this dataset was not available. Preliminary work to identify candidate
182 diagnostic *de novo* variants was undertaken in a smaller set of 402,464 variants identified through a
183 custom filtering strategy by Patrick J. Short (Wellcome Sanger Institute, personal correspondence).
184 These variants were identified by applying post-processing filters to DNMs in 4,967 trios identified by
185 the GEL Platypus variant caller and aligned to GRCh38. They were filtered according to the following
186 criteria: genotype heterozygous in offspring and homozygous reference in both parents, no more than
187 one alternate allele read in either parent, variant allele frequency in the offspring between 0.3 and
188 0.7, greater than 20 sequencing reads in the offspring and both parents, fewer than 98 sequencing
189 reads in the offspring, no overlap with locus control regions, no overlap with hg38 “patch regions”, no
190 other DNM within 20bp in the same individual. Some candidate variants identified in this preliminary
191 dataset are not present in the larger *de novo* set, owing to differences in the filtering pipeline. Unless
192 explicitly stated, the data presented here are from the larger DNM set, above.

193 Candidate diagnostic variants

194 To identify candidate diagnostic near-splice and branchpoint variants, we annotated all GRCh38
195 autosomal *de novo* SNVs passing the “stringent” filters (above) with VEP (version 99). For each variant,
196 only the consequence in one transcript per gene (determined by VEP’s “--per_gene” flag) was
197 considered. We annotated these variants with SpliceAI as described above. We filtered for variants
198 overlapping our branchpoint or near-splice positions of interest, finding 3,672 such variants. Where a
199 variant had both a branchpoint and a near-splice annotation, only the near-splice annotation was kept.
200 We then filtered for variants overlapping any known monoallelic rare disease gene with a loss of
201 function mechanism using the G2P DD, G2P Eye, and G2P Skin gene lists³² (accessed 27/10/2021,
202 confirmed and probable genes only). In total we identified 258 candidate splicing DNMs (238 near-
203 splice, 20 branchpoint) in 255 participants, across 137 genes.

204 To identify new diagnoses in the cohort, we annotated these variants with tiering data, phenotype
205 data, and participant outcome data from the GEL bioinformatics pipeline³³. For each participant and
206 DNM, we manually reviewed the similarity between the HPO terms recorded at recruitment and the
207 phenotype expected for a loss-of-function variant in that gene. Excluding any participants whose case
208 was already solved through 100KGP, we identified 35 new likely diagnostic variants with at least a
209 plausible phenotype match. In each instance, we placed a clinical collaboration request with Genomics
210 England to recruit the participant to the Splicing and Disease Study for functional characterisation of
211 the variant.

212 Genomics England does not allow re-identification of participants outside of a secure research
213 environment. In order to protect participant identities, the HPO terms given here are “abstracted” by
214 moving up one level in the HPO hierarchy. For example “Tetralogy of Fallot” becomes “Conotruncal
215 defect”.

216 Functional validation

217 Samples from five participants underwent functional characterisation through the Splicing and
218 Disease study at The University of Southampton. Blood was collected in PAXgene Blood RNA tubes,
219 with the PAXgene Blood RNA kit (PreAnalytix, Switzerland) used to extract RNA. Random hexamer
220 primers were used to synthesise complementary DNA (cDNA) by reverse transcription using the High-
221 Capacity cDNA Reverse Transcription kit (ThermoFisher Scientific).

222 Reverse transcription polymerase chain reaction (RT-PCR) was used to test for splicing alterations.

223 Primers were designed for each variant to include at least two exons up- and downstream of the target
224 (primer sequences available upon request). Agarose gel electrophoresis was used to assess participant
225 vs control PCR products, and purified PCR products were analysed by Sanger sequencing.

226 Ethics

227 The 100,000 Genomes Project Protocol has ethical approval from the HRA Committee East of England
228 – Cambridge South (REC Ref 14/EE/1112). This study was registered with Genomics England under

229 Research Registry Projects 143, 165, and 166. The Splicing and Disease study has ethical approval from
230 the Health Research Authority (IRAS Project ID 49685, REC 11/SC/0269) and The University of
231 Southampton (ERGO ID 23056), with informed consent given for splicing studies in a research context.

232 Code

233 All analyses were performed within a protected research environment which is available to registered
234 users (see [https://www.genomicsengland.co.uk/about-gecip/for-gecip-members/data-and-data-](https://www.genomicsengland.co.uk/about-gecip/for-gecip-members/data-and-data-access/)
235 [access/](https://www.genomicsengland.co.uk/about-gecip/for-gecip-members/data-and-data-access/)). Clinical data and tiering data were accessed through the LabKey application within this
236 research environment. The analyses were performed using Python 3.7.6 and Pandas 1.2.1. Figures
237 were generated in R 3.5.2 using RStudio 1.1.463. All code is available online at
238 https://github.com/alexblakes/100KGP_splicing.

239 Results

240 Signals of constraint at near-splice positions are replicated in a large healthy 241 cohort

242 To estimate the deleteriousness of variation in near-splice positions, we calculated aggregate
243 measures of evolutionary conservation, selective constraint, and pathogenicity prediction for
244 nucleotides within near-splice regions genome-wide.

245 Evolutionary conservation was measured by base-wise phyloP score. The CSSs are very highly
246 conserved (mean phyloP = 6.34) (Figure 1). Other intronic splicing positions with high phyloP scores
247 include the D+5 (3.44), D+4 (2.39), D+3 (1.97), A-3 (1.70), and D+6 (1.29) sites. Notably, the A-4
248 position is very weakly conserved (0.076). Coding positions are generally more highly conserved than
249 intronic sequences. The redundancy of third codon positions and bias for in-phase exons³⁴ is reflected
250 in lower phyloP scores at every third position, except for the donor 0 position (mean phyloP = 5.01),
251 which is more highly conserved than any other coding position.

252 To measure selective constraint at near-splice positions, we calculated the degree of purifying
253 selection acting at near-splice positions using MAPS¹². MAPS was calculated across near splice
254 positions genome-wide, using every observed synonymous, missense, nonsense, and near-splice SNV
255 in 207,548 distinct CDS exons for 26,660 unaffected parents in the 100KGP Rare Disease cohort. The
256 most significant signals of purifying selection are at the CSS, with MAPS scores of 0.089–0.146
257 ($p < 1.2 \times 10^{-43}$), approaching those of nonsense variants (0.16) (Figure 1). The non-canonical positions
258 with a MAPS score significantly above the synonymous baseline after Bonferroni correction include
259 the D-2, D0, D+3, D+4, D+5, D+6, A-3 and A+1 positions ($p < 6.3 \times 10^{-4}$). The MAPS scores at D0 and D+5
260 variants (MAPS = 0.057 and 0.067 respectively) are comparable to that at missense variants (0.052).
261 These results are highly concordant with previous near-splice MAPS calculations in the DDD and ExAC
262 datasets⁸.

263 A subset of splicing branchpoints are highly constrained

264 Having replicated earlier findings⁸ in our cohort, we expanded our analysis to examine splicing
265 branchpoints, which have not been previously characterised using MAPS.

266 We repeated our analysis of conservation, constraint, and SpliceAI predicted pathogenicity using a set
267 of 195,863 putative branchpoints predicted by LaBranchoR²⁹, a deep-learning tool trained on
268 experimentally-validated branchpoints.

269 Annotating each position with base-wise phyloP scores, we found modest conservation of BP0 (0.62)
270 and BP-2 (0.87), consistent with previous results²⁹ (Figure 1).

271 Next, we calculated the MAPS statistic for these branchpoint positions in the same cohort described
272 above. When all putative branchpoints were considered, only the BP-2 position has a significantly
273 higher MAPS score than the synonymous baseline (MAPS = 0.017, $p=1.3 \times 10^{-4}$) (Figure 1). However,
274 when only the most confident branchpoints are considered (LaBranchoR score >0.85 , $n=57,342$), the
275 BP0 (MAPS = 0.044, $p=7.6 \times 10^{-9}$) and BP-3 (0.024, $p = 3.7 \times 10^{-4}$) are also significantly constrained. These
276 data suggest that LaBranchoR-predicted branchpoints are functionally important, and that variants
277 near branchpoints may be a significant cause of rare disease.

278 Next, we calculated SpliceAI scores for every possible SNV around each branchpoint. Again, variants
279 at BP0 (0.15) and BP-2 (0.14) are nominally more likely to disrupt splicing than synonymous coding
280 variants (Figure 1). This trend is more pronounced when only the most confident branchpoints
281 (LaBranchoR score >0.85 , $n = 57,342$) are considered (mean SpliceAI BP0 = 0.22, BP-2 = 0.19).

282 New diagnostic candidates among near-splice *de novo* mutations

283 Having described three orthogonal metrics which independently suggest that certain near-splice and
284 branchpoint variants may be deleterious, we sought to identify new candidate diagnostic variants at
285 these positions.

286 We interrogated a set of 870,559 DNMs in 12,028 trios for potentially diagnostic splicing variants. We
287 identified 258 *de novo* SNVs overlapping near-splice or branchpoint regions of known monoallelic “loss
288 of function” rare disease genes in 255 individuals (Supplementary Table 1). Of these, 238 were in near-
289 splice positions, and 20 were within 5bp of a putative branchpoint. (12 variants had both a splice
290 acceptor and a branchpoint annotation; in these cases only the splice acceptor annotation (e.g. A- 25)
291 was kept).

292 Reviewing tiering data from the 100KGP bioinformatics pipeline, we found that of these 258 variants,
293 83 (32%) were “Tier 1”, 46 (18%) were “Tier 3”, and 129 (50%) were not tiered (Supplementary Figure
294 1). Of 59 CSS variants, 36 (61%) were “Tier 1”, nine (15%) were “Tier 3”, and 14 (24%) were not tiered.
295 Of ten donor +5 variants, four were “Tier 1”, two were “Tier 3” and four were not tiered
296 (Supplementary Figure 1). Annotation of these variants with SpliceAI generally highlighted variants at
297 positions with high MAPS scores (Supplementary Figure 2).

298 212 participants with a near-splice DNM had outcome data in the form of “exit questionnaires” from
299 their referring Genomic Medicine Centre. In 84/111 (76%) of solved cases, the diagnostic variant
300 matched our near-splice finding (Figure 2). This result gives us confidence in our approach to candidate
301 variant identification. Nevertheless, a significant proportion of participants with completed exit
302 questionnaires had unsolved cases (101/212, 48%). These included nine with a DNM in the CSS of a
303 known rare disease gene, one with a donor 0 variant, and four with donor +5 variants, which have
304 previously been estimated to have a 90% positive predictive value in rare disease diagnosis⁸ (Figure
305 2).

306 For each participant and DNM, we manually reviewed the similarity between the HPO terms recorded
307 at recruitment and the phenotype expected for a loss-of-function variant in that gene. Excluding any
308 participants whose case was already solved through 100KGP, we identified 35 new likely diagnostic
309 variants with at least a plausible phenotype match (Supplementary Table 2). In each instance, we

310 placed a clinical collaboration request with Genomics England to recruit the participant to the Splicing
311 and Disease Study for functional characterisation of the variant.

312 New diagnoses among the cohort

313 Whole blood RNA samples were obtained for five participants with near splice DNMs. RT-PCR was
314 used to characterise the splicing impact of each variant (Supplementary Figure 3). Abnormal splicing
315 events (all exon skipping) were detected in four participants (participants 74 (*ARID1A*, A-3), 249 (*USP7*,
316 D+5), 259 (*TLK2*, D+5), 261 (*KAT6B*, D+5)). In the remaining participant (participant 32 (*PPP1R12A*, A-
317 21)), no disruption to splicing was observed (Table 1, Supplementary Figure 3). For two additional
318 participants where the candidate variant fell in a canonical splice site (participants 83 (*TAOK1*, A-2) &
319 94 (*PHIP*, A-2)), a new diagnosis was reached without the need for functional work based on ACMG
320 criteria with a PVS1 classification for these variants (Table 1, Supplementary Figure 3).

321 In summary, we demonstrate a functional splicing defect in four out of five participants recruited to
322 our study, and we have identified a new molecular diagnosis for six individuals to date.

323 Discussion

324 We examined WGS data from 38,688 individuals in the Rare Disease arm of the 100KGP to evaluate
325 the contribution of splicing variants to rare genetic diseases. Using a population-based metric of
326 constraint, MAPS, we showed that certain near-splice and (for the first time) branchpoint positions
327 are under strong purifying selection. We identified 258 *de novo* near-splice and branchpoint variants
328 in known disease genes in these families. We identified 35 likely diagnostic variants which had
329 previously been missed through the 100KGP, and we have confirmed a new molecular diagnosis for
330 six participants to date. Overall, we demonstrate the clinical value of examining both canonical and
331 non-canonical splicing variants in unsolved rare diseases.

332 Non-canonical splicing positions harbour deleterious splicing variants

333 We used three orthogonal approaches to estimate the deleteriousness of near-splice and branchpoint
334 variants: between-species conservation, within-species constraint, and splicing pathogenicity
335 prediction. The PhyloP, MAPS, and SpliceAI scores at splicing positions consistently highlight those
336 non-canonical splicing positions (especially D0 and D+5) which are likely to harbour damaging variants.
337 Indeed, three out of three D+5 variants in which we performed RNA studies caused exon skipping.
338 Importantly, although we use a cohort of unaffected parents as a proxy for a normal population, the
339 MAPS data we present is highly concordant with the strong signals of negative selection at which have
340 been previously described in the ExAC and DDD datasets⁸.

341 Extending this analysis to splicing branchpoints, we find strong signals of negative selection at a subset
342 of branchpoint positions. These results are consistent with other measures of constraint previously
343 described at bovine and human branchpoints³⁵. We also identified candidate diagnostic variants at
344 these positions, and we are awaiting RNA samples to functionally characterise these variants. The
345 disruption of splicing branchpoints may therefore make an important contribution to rare disease^{10,11},
346 and a systematic analysis of *de novo* variation at branchpoints is an exciting future research
347 opportunity.

348 The ACMG variant interpretation guidelines give special status to CSS variants as “very strong”
349 diagnostic candidates in disorders where LoF is a known disease mechanism⁶. This remains the case
350 in more detailed guidance for the interpretation of LoF variants which has recently been introduced³⁶.
351 However, the deleteriousness of splicing variants is not binary, but on a continuum, and can be
352 quantitatively compared to other variant classes. Previous estimates suggest that 46% of non-canonical
353 near-splice DNMs in dominant rare disease genes may be pathogenic, rising to 71% for pyrimidine to
354 purine transversions in the polypyrimidine tract, and 90% for D+5 variants⁸. The deleteriousness of
355 individual variants is contingent on many factors, such as local sequence context, the alternate
356 nucleotide, exon frame, exon length, and intron length⁹. For this reason, the systematic classification

357 of near-splice variants remains challenging, and clinical interpretation of these variants is still
358 dependent on expert phenotype matching and functional validation of candidate variants.

359 The functional characterisation of splicing variants can be challenging and requires adequate amounts
360 of good quality RNA. Our study is limited by the use of blood as a proxy for the most clinically relevant
361 tissue, although we affirm the utility of blood RNA analysis by identifying splicing defects in four out
362 of five samples tested. Whereas RT-PCR is a bespoke and low-throughput approach, going forward,
363 RNA-sequencing (RNA-seq) offers an unbiased and high-throughput alternative to simultaneously
364 detect and functionally characterise splicing variants. A whole-transcriptome RNA-seq pilot study has
365 recently been proposed for 100KGP, and use of RNA-seq in routine clinical practice could offer a much-
366 needed means to systematically and objectively interpret splicing variants³⁷.

367 **New rare disease diagnoses**

368 We identified 258 *de novo* SNVs overlapping near-splice or branchpoint regions of known monoallelic
369 “loss of function” rare disease genes in 255 individuals. Of these, at least 84 were already considered
370 to be diagnostic through 100KGP, and we identified an additional 35 variants which are likely to be
371 diagnostic given the available molecular, phenotypic, and *in silico* data. We confirmed a new molecular
372 diagnosis for six participants, including four participants for whom RNA studies were performed.

373 Surprisingly, several strong diagnostic candidates were apparently overlooked in the standard variant
374 interpretation pipeline, including at least nine CSS variants and four D+5 variants, all in known rare
375 disease genes. Of ten *de novo* D+5 variants, none were previously labelled as pathogenic, despite their
376 high prior probability of being diagnostic in this context⁸.

377 Clearly, many new diagnoses remain to be found. A recent analysis of 100KGP data in the context of
378 craniosynostosis found that expert-led review more than doubled diagnostic yields compared to the
379 standard pipeline³. An important factor is that the “virtual panels” applied to variant calls are outdated
380 and do not include recently discovered disease genes. Our phenotype-matching work suggests that

381 the clinical impact of near-splice variants has been under-ascertained in this cohort, and we are
382 continuing to recruit participants for functional assessment of these variants.

383 One obstacle to increasing the number of researcher-identified diagnoses in this context is the
384 difficulty of recontacting de-identified participants and clinicians through secure research
385 environments. The confidentiality of all participants in research is rightly a priority, and new pathways
386 must be developed to streamline the clinical-research interface in medical genomics.

387 Conclusion

388 In conclusion, the disruption of splicing is an important cause of rare disease among 100KGP
389 participants, but the contribution of non-canonical variants is still under-recognised. Splicing
390 branchpoints are another non-canonical and non-coding source of damaging splicing variants which
391 are amenable to systematic analysis in WGS data. The improved interpretation of splicing variants is
392 an area of great promise to genomic medicine and, above all, to individuals with rare diseases and
393 their families.

394 Supplemental data

395 Supplementary data include three figures and two tables.

396 Declaration of interests

397 The authors declare no competing interests.

398 Acknowledgements

399 The authors thank all participants and families involved in this research. We thank all clinicians and
400 contributors who helped to assess potential splicing diagnoses, including: Ellen Thomas, Jessica
401 Radley, Rebecca Igbokwe, Suresh Vijay, Deirdre Cilliers, Evan Reid, Mick Parker, David Hunt, Rachel
402 Keen, Ed Blair, Helen Firth, Peggy O'Driscoll, Chiara Marini Bettol, Monish Suri, John Barton, Angela

403 Barnicoat, Sahar Mansour, Melody Redman, Kate Barr, Debbie Fuller, Meena Balasubramanian, Julia
404 Rankin, Sian Ellard, Olga Tsoulaki, and Emma Kivuva.

405 The Baralle lab is supported by NIHR Research Professorship awarded to D.B. (RP-2016-07-011).
406 Functional work was additionally supported by a Wessex Medical Research Innovation Grant awarded
407 to J.L. NW is currently supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust
408 and the Royal Society (Grant Number 220134/Z/20/Z) and funding from the Rosetrees Trust. AB was
409 supported by funding from Health Education England. We acknowledge the NIHR Clinical Research
410 Network (CRN) in recruiting participants and the Musketeers Memorandum, as well as support from
411 the NIHR UK Rare Genetic Disease Consortium. We thank Patrick J. Short for providing a curated set
412 of de novo variants used in earlier iterations of these analyses.

413 This research was made possible through access to the data and findings generated by the 100,000
414 Genomes Project. The 100,000 Genomes Project is managed by Genomics England Limited (a wholly
415 owned company of the Department of Health and Social Care). The 100,000 Genomes Project is
416 funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer
417 Research UK and the Medical Research Council have also funded research infrastructure. The 100,000
418 Genomes Project uses data provided by participants and collected by the National Health Service as
419 part of their care and support.

420 Data and code availability

421 All code is available online at https://github.com/alexblakes/100KGP_splicing.

422 The data used in this study are available to registered users within a protected research environment
423 at <https://www.genomicsengland.co.uk/about-gecip/for-gecip-members/data-and-data-access/>.

424 References

- 425 1. International Rare Diseases Research Consortium. *Policies and guidelines*. (2013).
- 426 2. Wright, C. F., FitzPatrick, D. R. & Firth, H. V. Paediatric genomics: Diagnosing rare disease in
427 children. *Nat. Rev. Genet.* **19**, 253–268 (2018).
- 428 3. Hyder, Z. *et al.* Evaluating the performance of a clinical genome sequencing program for
429 diagnosis of rare genetic disease , seen through the lens of craniosynostosis. (2021).
430 doi:10.1038/s41436-021-01297-5
- 431 4. Sanders, S. J., Schwartz, G. B. & Farh, K. K. H. Clinical impact of splicing in neurodevelopmental
432 disorders. *Genome Med.* **12**, 1–5 (2020).
- 433 5. Wai, H., Douglas, A. G. L. & Baralle, D. RNA splicing analysis in genomic medicine. *Int. J.*
434 *Biochem. Cell Biol.* **108**, 61–71 (2019).
- 435 6. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint
436 consensus recommendation of the American College of Medical Genetics and Genomics and
437 the Association for Molecular Pathology. *Genet. Med.* **17**, 405–423 (2015).
- 438 7. Rivas, M. A. *et al.* Effect of predicted protein-truncating genetic variants on the human
439 transcriptome. *Science (80-.).* **348**, 666–669 (2015).
- 440 8. Lord, J. *et al.* Pathogenicity and selective constraint on variation near splice sites. *Genome Res.*
441 **29**, 159–170 (2019).
- 442 9. Zhang, S. *et al.* Base-specific mutational intolerance near splice sites clarifies the role of
443 nonessential splice nucleotides. *Genome Res.* **28**, 968–974 (2018).
- 444 10. Kapoor, R. R. *et al.* Persistent hyperinsulinemic hypoglycemia and maturity-onset diabetes of
445 the young due to heterozygous HNF4A mutations. *Diabetes* **57**, 1659–63 (2008).
- 446 11. Fadaie, Z. *et al.* BBS1 branchpoint variant is associated with non-syndromic retinitis

- 447 pigmentosa. *J. Med. Genet.* (2021). doi:10.1136/jmedgenet-2020-107626
- 448 12. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–
449 291 (2016).
- 450 13. Whiffin, N. *et al.* Characterising the loss-of-function impact of 5' untranslated region variants
451 in 15,708 individuals. *Nat. Commun.* **11**, 1–12 (2020).
- 452 14. Rowlands, C. F. & Baralle, D. Machine Learning Approaches for the Prioritization of Genomic
453 Variants Impacting Pre-mRNA Splicing. (2019).
- 454 15. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**,
455 535-548.e24 (2019).
- 456 16. Rowlands, C. *et al.* Comparison of in silico strategies to prioritize rare genomic variants
457 impacting RNA splicing for the diagnosis of genomic disorders. *Sci. Rep.* **11**, 20607 (2021).
- 458 17. Smedley, D. *et al.* 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care —
459 Preliminary Report. *N. Engl. J. Med.* **385**, 1868–1880 (2021).
- 460 18. Genomics England. *The National Genomics Research Library.* (2020).
461 doi:<https://doi.org/10.6084/m9.figshare.4530893.v7>
- 462 19. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic*
463 *Acids Res.* **47**, D766–D773 (2019).
- 464 20. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution
465 rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
- 466 21. Ryan, D. P. pyBigWig. (2015). Available at: <https://github.com/deeptools/pyBigWig>.
- 467 22. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
- 468 23. Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. BigWig and BigBed: Enabling
469 browsing of large distributed datasets. *Bioinformatics* **26**, 2204–2207 (2010).

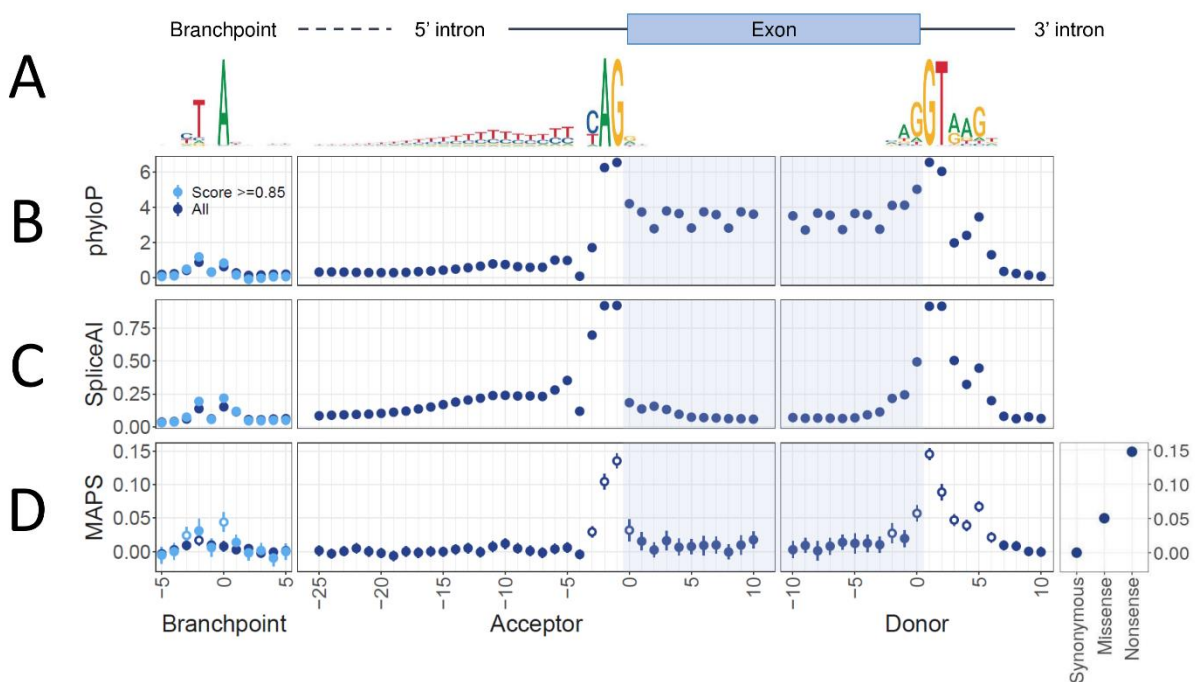
- 470 24. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, (2021).
- 471 25. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features.
472 *Bioinformatics* **26**, 841–842 (2010).
- 473 26. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 1–14 (2016).
- 474 27. Short, P. J. *et al.* De novo mutations in regulatory elements in neurodevelopmental disorders.
475 *Nature* **555**, 611–616 (2018).
- 476 28. Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in human disease.
477 *Nat. Genet.* **46**, 944–950 (2014).
- 478 29. Paggi, J. M. & Bejerano, G. A sequence-based, deep learning model accurately predicts RNA
479 splicing branchpoints. *Rna* **24**, 1647–1653 (2018).
- 480 30. UCSC. Lift Genome Annotations. Available at: <https://genome.ucsc.edu/cgi-bin/hgLiftOver>.
481 (Accessed: 25th August 2020)
- 482 31. Genomics England de novo variant research dataset. Available at: [https://research-](https://research-help.genomicsengland.co.uk/display/GERE/De+novo+variant+research+dataset)
483 [help.genomicsengland.co.uk/display/GERE/De+novo+variant+research+dataset](https://research-help.genomicsengland.co.uk/display/GERE/De+novo+variant+research+dataset). (Accessed:
484 23rd November 2021)
- 485 32. Thormann, A. *et al.* Flexible and scalable diagnostic filtering of genomic variants using G2P with
486 Ensembl VEP. *Nat. Commun.* **10**, 1–10 (2019).
- 487 33. Genomics England. *Rare Disease Results Guide*. (2020).
- 488 34. Tomita, M., Shimizu, N. & Brutlag, D. L. Introns and reading frames: Correlation between
489 splicing sites and their codon positions. *Mol. Biol. Evol.* **13**, 1219–1223 (1996).
- 490 35. Kadri, N. K., Mapel, X. M. & Pausch, H. The intronic branch point sequence is under strong
491 evolutionary constraint in the bovine and human genome. *Commun. Biol.* **4**, 1206 (2021).
- 492 36. Abou Tayoun, A. N. *et al.* Recommendations for interpreting the loss of function PVS1

- 493 ACMG/AMP variant criterion. *Hum. Mutat.* **39**, 1517–1524 (2018).
- 494 37. Cummings, B. B. *et al.* Improving genetic diagnosis in Mendelian disease with transcriptome
495 sequencing. *Sci. Transl. Med.* **9**, (2017).
- 496

497 Figures

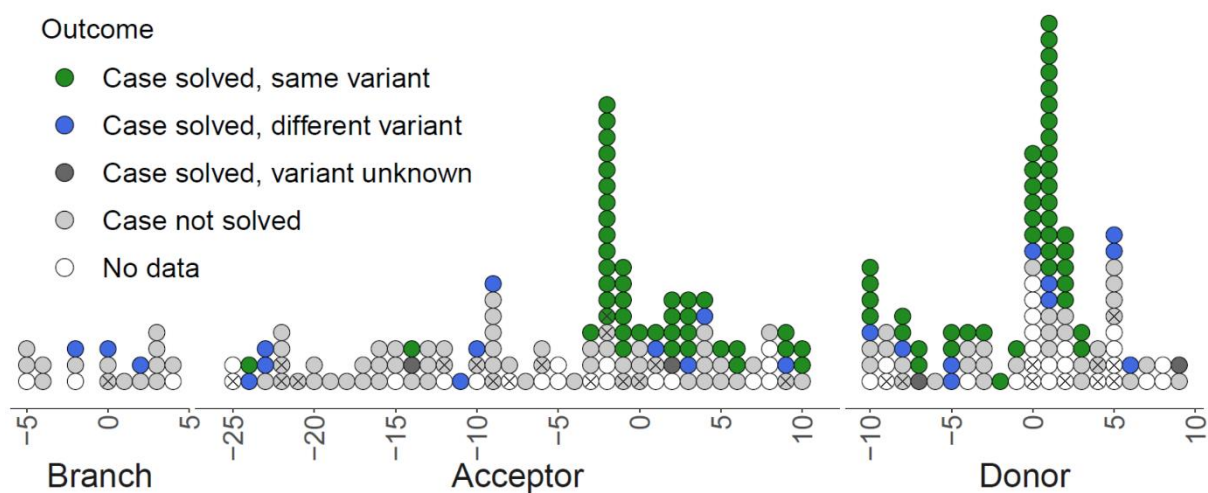
498 Figure 1

499 Conservation, predicted splice disruption, and constraint at near-splice and branchpoint positions
 500 across 207,548 CDS features in protein coding genes. **A:** Position-weight matrices and schematic
 501 indicating position of conserved splicing motifs relative to exon / intron boundaries. **B:** Mean phyloP
 502 100-way scores at splicing positions. Error bars indicate 95% confidence intervals. **C:** SpliceAI scores
 503 for all possible near splice SNVs. Scores represent the mean probability that any variant at this position
 504 disrupts splicing, as predicted by SpliceAI (see Methods). Error bars represent the 95% confidence
 505 interval. **D:** Mutability-adjusted proportion of singletons (MAPS) for both coding and near-splice SNVs.
 506 Error bars indicate 95% confidence intervals. Positions with a significantly higher MAPS than
 507 synonymous variants are indicated with open circles (see Methods). For branchpoint positions, dark
 508 blue points represent all putative branchpoints, whereas light blue points represent branchpoints with
 509 a LaBranchoR score >0.85



510 **Figure 2**

511 Participant outcomes for rare disease probands with *de novo* splicing variants in known monoallelic
512 loss-of-function rare disease genes. Each point represents a DNM in a rare disease proband. Points
513 are coloured by the clinical outcome for that individual. Crosses indicate variants which were
514 identified as likely new diagnoses in this study. Where a variant overlaps both a branchpoint and a
515 splice acceptor position, only the splice acceptor annotation is given.



516 Tables

517 Table 1

518 Diagnostic outcomes for seven individuals after clinical and functional characterisation of the splicing
519 variant. Five individuals underwent RNA studies, of which four received a new diagnosis. In two
520 additional individuals, a diagnosis was reached without the need for RNA studies. In total, a new
521 diagnosis was confirmed for six individuals. Note that the given HPO terms are “abstracted” (see
522 Methods) to protect confidentiality. * In participants 83 and 94, a new diagnosis was reached without
523 the need for functional evaluation. ** This exit questionnaire outcome was updated after the
524 participant was identified by this study.

ID	Chrom	Pos	Ref	Alt	Region	Site	Symbol	ENST	DS_any	DS_max	DS_max type	Tier	Max tier	Exit questionnaire	HPO terms (abstracted)	Splicing impact	Outcome
74	chr1	26767787	C	G	acceptor	-3	ARID1A	ENST00000324856	0.71	0.65	DS_AL	3	3	No data	Aplasia/Hypoplasia of the mandible, Advanced eruption of teeth, Abnormal pulmonary valve morphology, Abnormality of calvarial morphology, Abnormality of cardiovascular system morphology, Oral cleft	Exon skipping	New diagnosis
261	chr10	74989117	G	A	donor	5	KAT6B	ENST00000287239	0.98	0.98	DS_DL	3	3	Case not solved	Hypothyroidism	Exon skipping	New diagnosis
259	chr17	62596679	G	A	donor	5	TLK2	ENST00000326270	0.97	0.96	DS_DL	3	3	Case not solved	Abnormality of globe location, Abnormal facial shape, Facial asymmetry, Abnormal heart sound, Cutaneous syndactyly, Abnormal ear morphology, Neurodevelopmental delay, Short stature, Abnormal digit morphology, Decreased body weight, Intrauterine growth retardation, Abnormality of higher mental function, Motor delay, Language impairment, Gait disturbance, Abnormal location of ears	Exon skipping	New diagnosis
249	chr16	8905182	C	A	donor	5	USP7	ENST00000344836	0.932	0.8	DS_DL	3	3	Case not solved	Motor delay, Abnormal size of the palpebral fissure, Abnormal hair quantity, Abnormality of globe location, Facial hypertrichosis, Neurological speech impairment, Abnormality of higher mental function, Abnormal metatarsal morphology	Exon skipping	New diagnosis
94	chr6	79002126	T	G	acceptor	-2	PHIP	ENST00000275034	1	1	DS_AL	3	2	Case not solved	Abnormality of higher mental function, Motor delay, Neurodevelopmental delay, Macrotia, Language impairment, Finger clinodactyly, Abnormal muscle tone, Facial hypertrichosis	N/A	New diagnosis*
83	chr17	29491782	A	G	acceptor	-2	TAOK1	ENST00000261716	0.992	0.99	DS_AL	3	3	Case solved, same variant**	Renal agenesis, Hemangioma, Abnormality of joint mobility, Hypotonia, Increased head circumference, Neurodevelopmental delay	N/A	New diagnosis*
32	chr12	79808598	T	A	acceptor	-21	PPP1R12A	ENST00000450142	0	0	DS_AG	N/A	3	Case not solved	Increased head circumference, Abnormal thorax morphology, Bowing of the legs, Growth delay, Limb undergrowth, Abnormality of joint mobility, Short digit, Neurodevelopmental abnormality, Abnormality of movement	Normal splicing	Unsolved

526 Supplementary Figures

527 Supplementary Figure 1

528 Tiering data for splicing DNMs in known monoallelic loss-of-function rare disease genes. Each point
529 represents a DNM in a rare disease proband. Points are coloured by the “tier” of that variant in the
530 GEL annotation pipeline (see Methods).

531 Supplementary Figure 2

532 SpliceAI scores of splicing DNMs in known monoallelic loss-of-function rare disease genes. Each point
533 represents a DNM in a rare disease proband. Points are coloured by SpliceAI score; grey points indicate
534 that no SpliceAI annotation is available.

535 Supplementary Figure 3

536 Functional outcomes for participant samples which were characterised by RT-PCR. Each page
537 illustrates the following: a schematic of variant position relative to the exon/intron junction, a
538 schematic of the splicing consequence of each variant, gel electrophoresis of amplified RT-PCR
539 products for participant and control samples, Sanger sequencing trace for proband-specific bands.

540 Supplementary tables

541 Supplementary Table 1

542 Molecular details and clinical outcome data for the 258 prioritised DNMs. Rows in bold indicate those
543 variants identified as likely to be diagnostic, which are given in Supplementary Table 2 (“Main” DNM
544 set). * This exit questionnaire outcome was updated after the participant was identified by this study.

545 Supplementary Table 2

546 Molecular details and phenotypic data for 35 likely diagnostic variants in “unsolved” probands. Each
547 variant was judged to be at least a plausible diagnostic fit given the phenotype information available.
548 Note that the given HPO terms are “abstracted” (see Methods) to prevent participant re-identification.
549 Also note that of the 35 diagnostic candidates, 8 were identified in a preliminary set of DNMs used for
550 exploratory analysis at the inception of this project, and are not among the 258 prioritised DNMs
551 (Supplementary Table 1) described in more detail above (see Methods). Participants in bold are those
552 shown in Table 1. * This exit questionnaire outcome was updated after the participant was identified
553 by this study.