The contribution of non-canonical splicing mutations to severe dominant developmental disorders

AUTHORS

Jenny Lord[1], Giuseppe Gallone[1], Patrick J. Short[1], Jeremy F. McRae[1], Holly Ironfield[1], Elizabeth H. Wynn[1], Sebastian S. Gerety[1], Liu He[1], Bronwyn Kerr[2,3], Diana S. Johnson[4], Emma McCann[5], Esther Kinning[6], Frances Flinter[7], I. Karen Temple[8,9] , Jill Clayton-Smith[2,3], Meriel McEntagart[10], Sally Ann Lynch[11], Shelagh Joss[12], Sofia Douzgou[2,3], Tabib Dabir[13], Virginia Clowes[14], Vivienne P. M. McConnell[15], Wayne Lam[16], Caroline F. Wright[17], David R. FitzPatrick[1,16], Helen V. Firth[1,18], Jeffrey C. Barrett[1], Matthew E. Hurles[1], on behalf of the DDD study

AFFILIATIONS

[1] Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

[2]Manchester Centre for Genomic Medicine, St Mary's Hospital, Manchester University Hospitals NHS Foundation Trust Manchester Academic Health Sciences Centre

[3]Division of Evolution and Genomic Sciences School of Biological Sciences University of Manchester

[4]Sheffield Clinical Genetics Service, Sheffield Children's Hospital, OPD2, Northern General Hospital, Herries Road, Sheffield, S5 7AU

[5]Liverpool Women's Hospital Foundation Trust, Crown Street, Liverpool, L8 7SS

[6]West of Scotland Regional Genetics Service, NHS Greater Glasgow and Clyde, Institute of Medical Genetics, Yorkhill Hospital, Glasgow G3 8SJ, UK

[7]South East Thames Regional Genetics Centre, Guy's and St Thomas' NHS Foundation Trust, Guy's Hospital, Great Maze Pond, London SE1 9RT, UK

[8]Faculty of Medicine, University of Southampton, Institute of Developmental Sciences, Tremona Road, Southampton SO16 6YD

[9]Wessex Clinical Genetics Service, University Hospital Southampton, Princess Anne Hospital, Coxford Road, Southampton SO16 5YA, UK

[10]South West Thames Regional Genetics Centre, St George's Healthcare NHS Trust, St George's, University of London, Cranmer Terrace, London SW17 0RE, UK

[11]Temple Street Children's Hospital, Dublin 1, Ireland

[12]West of Scotland Regional Genetics Service, NHS Greater Glasgow & Clyde, Level 2, Laboratory Medicine Building, Queen Elizabeth University Hospital, Glasgow G51 4TF

[13]Northern Ireland Regional Genetics Centre, Belfast Health and Social Care Trust, Belfast City Hospital, Lisburn Road, Belfast BT9 7AB, UK

[14]North West Thames Regional Genetics Service, London North West University Healthcare NHS Trust, Northwick Park and St Mark's Hospitals, Watford Road, Harrow HA1 3UJ, UK

[15]Northern Ireland Regional Genetics Service, Belfast Health and Social Care Trust, Belfast City Hospital, Lisburn Road, Belfast BT9 7AB, Northern Ireland, UK

[16]MRC Human Genetics Unit, MRC IGMM, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, UK

[17]Institute of Biomedical and Clinical Science, University of Exeter Medical School, RILD Level 4, ED&E, Barrack Road, Exeter, EX2 5DW, UK

[18]East Anglian Medical Genetics Service, Box 134, Cambridge University Hospitals NHS foundation Trust, Cambridge Biomedical Campus, Cambridge, CB2 0QQ, UK

CONTACT DETAILS OF CORRESPONDING AUTHOR

Matthew E. Hurles, meh@sanger.ac.uk

Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

Abstract

Accurate and efficient pre-mRNA splicing is crucial for normal development and function, and mutations which perturb normal splicing patterns are significant contributors to human disease. We used exome sequencing data from 7,833 probands with developmental disorders (DD) and their unaffected parents to quantify the contribution of splicing mutations to DDs. Patterns of purifying selection, a deficit of variants in highly constrained genes in healthy subjects and excess *de novo* mutations in patients highlighted particular positions within and around the consensus splice site of greater disease relevance. Using mutational burden analyses in this large cohort of proband-parent trios, we could estimate in an unbiased manner the relative contributions of mutations at canonical dinucleotides (73%) and flanking non-canonical positions (27%), and calculated the positive predictive value of pathogenicity for different classes of mutations. We identified 18 likely diagnostic *de novo* mutations in dominant DD-associated genes at non-canonical positions in splice sites. We estimate 35-40% of pathogenic variants in non-canonical splice site positions are missing from public databases.

Introduction

Pre-mRNA splicing in humans is mediated by the major and minor spliceosomes, highly dynamic, metalloenzyme complexes comprised of five key small nuclear RNAs (snRNA), along with over 100 protein components and accessory molecules[1-3]. Accurate recruitment and function of the complex is reliant on a plethora of cis-acting regulatory elements encoded within the pre-mRNA itself, including the canonical acceptor and donor (or 3' and 5' splice) dinucleotides, the branchpoint, which offers a "tether" for lariat formation, the polypyrimidine tract (PolyPy), and a catalogue of exonic and intronic splice enhancers and silencers (ESE/ISE/ESS/ISS). Whilst our understanding of the underlying mechanistic processes regulating splicing has greatly increased in recent years, our ability to predict whether or not a mutation will affect splicing remains limited. However, with estimates that up to 50% of monogenic disease-causing variants may affect splicing[4; 5], a better understanding and more coherent approach to interpretation of variants affecting splicing is badly needed[6; 7]. With a plethora of *in silico* splicing pathogenicity predictors available, there is little consensus on what a "gold standard" for splicing pathogenicity prediction would be[8-10]. Whilst many of these methods perform well within the canonical splice site dinucleotides (CSS, the two highly-conserved bases flanking the acceptor and donor sites), their utility for other splice relevant regions is less clear[8]. In the clinical setting, often multiple algorithms and expert judgment are used to predict pathogenicity, while for large scale gene discovery research projects, classification of variants is often binary, with CSS mutations typically classified as likely splice affecting, whilst mutations in other splicing regulatory components are typically overlooked[11-13]. Both clinical and research interpretation of potential splice-disrupting variants has lacked a robust quantitative foundation.

We sought to assess the relative contribution of pathogenic, splice-altering mutations between the CSS and other, near-splice positions using both population-based and disease-focussed analyses utilising large scale exome sequencing data from the Deciphering Developmental Disorders (DDD) project[12] and ExAC[14].

We recruited 7,833 probands with undiagnosed developmental disorders, along with their parents, from clinical genetics centres across the UK and Ireland to the DDD study. Exome sequencing was performed to find likely genetic diagnoses underlying their conditions. Likely diagnostic *de novo* mutations (DNMs) in known developmental disorder (DD) genes have been found for ~25% of probands, while a smaller contribution of recessive disorders has also been observed[12; 13; 15], meaning over half the cohort currently lacks a molecular diagnosis.

Our analyses focus on near-splice site positions across a set of 148,244 stringently defined exons well covered (median coverage >15X at both CSS) across the DDD cohort (see Methods), including 25bp intronic and 11bp exonic sequence at the splice acceptor site, and 10bp intronic and 11bp exonic sequence at the splice donor site. For exonic positions, non-synonymous variants were removed to minimise the chances of observing effects not due to splicing regulation. To investigate selective constraint within the splicing region, we utilise exome sequencing data from 13,750 unaffected parents within the DDD study, as well as >60,000 aggregated exomes from ExAC[14].To investigate near-splice positions in a disease-centric way, we analysed >16,700 high confidence DNMs identified within coding and near-coding regions well covered by exome sequencing in the 7,833 probands.

Materials and methods

Cohort and sequencing

For full description of cohort and analytical methodology, see previous DDD publications[13; 16]. Briefly, 7,833 patients with severe, undiagnosed developmental disorders were recruited to the DDD study from 24 clinical genetics centres from across the UK and Ireland. Whole exome sequencing was conducted on the proband and both parents, with exome capture using SureSelect RNA baits (Agilent Human All-Exon V3 Plus with custom ELID C0338371 and Agilent Human All-Exon V5 Plus with custom ELID C0338371) and sequencing using 75 base paired-end reads using Illumina's HiSeq. Mapping was conducted to GRCH37 using the Burrows-Wheeler aligner (BWA, v0.59[17]) and variant identification was conducted using the Genome Analysis Toolkit (GATK, v3.5.0[18]). Variant annotation was conducted with Ensembl's Variant Effect Predictor (VEP), using Ensembl gene build 76[19]. DNMs were identified using DeNovoGear (v0.54)[20], and filtered using an in house pipeline - denovoFilter - developed by Jeremy F. McRae[13] (see web resources).

Defining exons of interest

We took exons from gencode v19 which met the following criteria: annotation_type = "exon", gene_type = "protein_coding", gene_status = "KNOWN", transcript_type = "protein_coding", transcript_status = "KNOWN", annotation != "level 3" (automated annotation), and tag = "CCDS", "appris_principal", "appris_candidate_longest", "appris_candidate", or "exp_conf" (n = 255,812 exons)[21]. We removed a small subset of exons which no longer met these criteria in the more recent, GRCH38 based gencode v22 release (leaving 253,275 exons). We removed any exons where the median coverage at the canonical acceptor or donor positions was <15X in two sets of DDD data

which used different exon capture methods (Agilent Human All-Exon V3 Plus with custom ELID C0338371 and Agilent Human All-Exon V5 Plus with custom ELID C0338371). 148,244 exons passed these criteria.

We annotated individual genomic positions relative to the acceptor and donor sites, removing any exons <14bp, and any positions which had multiple potential annotations. At the acceptor end, we considered 25bp of intronic sequence (acc-25 to acc-1) and 11bp exonic sequence (acc to acc+10). At the donor end, we considered 10bp of intronic sequence (don+1 to don+10) and 11bp exonic sequence (don to don-10). This yielded ~6.9 million near-splice positions of interest.

We define the polypyrimidine tract (PolyPy) region as acc-3, and acc-5 to acc-17, based on pyrimidine content > 70% in our exons of interest. We assess changes from a pyrimidine to a purine (PyPu) adjusting for the strand the exon is on.

Mutability adjusted proportion of singletons (MAPS)

In 13,750 unaffected parents enrolled as part of the DDD study, as well as >60,000 aggregated exomes from ExAC v0.3.1, we calculated the MAPS metric[14] using code developed in house by Patrick J. Short (see web resources). This was done for all splice positions, the last base of the exon split by reference nucleotide, and the PolyPy, split by PyPu vs all other changes, as well as VEP[19] ascertained synonymous, missense and nonsense sites across autosomal regions. To establish whether the MAPS metric was significantly different between PolyPy PyPu vs all other changes, a bootstrap resampling method was run with 1000 iterations.

Parental variants in high pLI genes

We looked at all variant positions overlapping with our splicing sites of interest within the unaffected parental data, and calculated the proportion of these sites which fell within genes with high probability of loss of function intolerance (pLI) scores[14] (> 0.9) for each position of the splice region, grouped sites, the last base of the exon split by reference allele, and the PolyPy split by PyPu vs all other changes, as well as VEP annotated synonymous, missense, and nonsense sites across autosomal regions.

***De novo* mutations**

DNMs were identified using DeNovoGear[20] as described in McRae *et al*, 2017[13], and a stringent confidence threshold (posterior probability > 0.8) was applied. We used triplet-based mutation rates[22] to calculate the expected number of DNMs across autosomal regions in the 7,833 probands, adjusting expected values for depth of sequencing coverage < 50 (exon depth <1, exp*0.119; exon

depth >1 and <50, exp*(0.119+0.204*log(depth))). We used the Poisson test to examine differences in the observed and expected values, and a 5% FDR correction to control for multiple testing (R v3.1.3). We stratified this analysis into known dominant, known recessive and non-DD associated genes using the DDG2P gene list, downloaded in June 2016. Genes with both recessive and dominant modes of inheritance were restricted to the dominant list.

We compared the relative distributions of CSS position variants with other protein truncating variants in the DDD and ExAC data.

Positive predictive values were calculated ((observed - expected) / observed) for CSS positions, combined and individually, don+5 sites, PolyPy PyPu, PolyPy other, other near splice exonic and intronic variants, as well as VEP defined missense and stop gained mutations.

We divided our exons into sextiles based on the pLI metric[14], and calculated the observed and expected number of DNMs in each group for don+5, PolyPy PyPu and synonymous variants (as above) to see if the enrichment of don+5 and PolyPy PyPu changes was concentrated in genes more likely to be intolerant of loss of function (LoF) mutations.

Potential diagnostic variants

DNMs overlapping with our near-splice positions of interest within dominant DDG2P genes were identified in DDD probands currently lacking a likely diagnosis (n = 5907). The HPO encoded[23] phenotypes of the probands were assessed by consultant clinical geneticist Helen V. Firth, along with the patient's recruiting clinician, and compared to the known clinical presentation of individuals with LoF mutations within those genes, classifying each variant as likely diagnostic, unlikely diagnostic, or unsure, depending on the strength of similarity between the proband and the disorder, and the specificity of the phenotype.

Validation of putative splicing variants

Eight variants were selected for validation via a minigene vector system. These comprised six likely diagnostic variants from the PolyPy, a PolyPy variant of uncertain clinical significance, and a likely diagnostic don+5 variant. Additionally, two untransmitted variants identified in unaffected parents within the same PolyPys as test variants were selected as negative controls. Details of all variants selected for validation are shown in Table S1.

Cloning splicing vectors

The minigene splice assay vector was adapted from that used in Singh *et al*, 2016[24], by replacing intron 1 with the first intron from the rat insulin 2 gene (Ins-2)(Rnor_6.0 Chr1:215857148-215857695). To generate individual assay vectors, either the 5' most 231bp (for the don+5 variant) or the 3' most 274bp (for PolyPy variants) of this vector was replaced with the appropriate endogenous intronic sequence encompassing the DNM of interest (Figure S3a and S3b), as described below.

First, proband genotypes (Table S1) were verified by capillary sequencing of genomic PCR products. Genomic regions containing the reference and alternate sequences were then either amplified by nested PCR, generated by site directed mutagenesis, or generated using gene synthesis (IDT). These fragments were sub-cloned, by Gibson Assembly (NEB), into our minigene vector. The regions assayed in our vectors are detailed by genomic coordinates in Table S1. Full sequence for all plasmids and primers available upon request.

In vitro splicing assay

HeLa cells were seeded into 12-well plates at a density of 160,000 cells per well, grown for 24 hours and transfected with 1 microgram of plasmid vector using Lipofectamine 3000 (Invitrogen). All transfections were carried out in duplicate and cultured for 48 hours. HeLa cells were cultured in DMEM (10% FCS + 1% pen/strep) at 37°C in a humidified incubator. Total RNA was extracted using a Micro RNeasy Qiagen kit and mRNA converted into cDNA using superscript IV (Invitrogen). RT-PCR was carried out using primers designed to span from exon 1 to exon 2, exon 2 to exon 3 and exon 1 to exon 3 and amplified on a thermocycler for either 25 or 35 cycles. Amplicons were capillary sequenced (GATC). For amplicons showing more than one splice variant (mixed capillary traces, for *CHD7*-Alt and *MBD5*-Alt), we cloned the PCR amplicons (Zero Blunt PCR cloning kit, Invitrogen) and sequenced individual colonies by capillary sequencing to identify the splice variants present. All PCR and sequencing primers available upon request.

Chromatograms were generated in R from .ab1 files using the sangerseqR[25] package (R v3.1.3), and likely consequences on the protein primary structure were generated using reference and alternative RNA sequences with the ExPASy Nucleotide Sequence Translation tool[26].

Splicing pathogenicity scores

Since our region of interest spanned >6 million individual positions, each with three potential single nucleotide changes, we were restricted in the choice of splicing pathogenicity prediction tools we could utilise, as many function primarily through a low throughput web interface model. We

identified three resources recently published which provide "genome wide" splicing pathogenicity scores. Two methods, dbscSNV's AdaBoost and RandomForest are based on ensemble learning combining predictions from multiple other splice prediction tools as well as conservation and CADD scores[27]. The targeted region at the acceptor end spans 14 bases (12 intronic, 2 exonic) and at the donor end spans 11 bases (8 intronic, 3 exonic). Spidex utilises deep learning methods trained on RNA sequencing data to estimate the consequence of variants on the "percent spliced in" of an exon, relative to the reference sequence[28]. Spidex scores positions up to 300bp from intron/exon boundaries, so provides greater coverage of our splicing region of interest. We also utilised the longer standing, and widely used MaxEntScan (MES)[29], for which perl scripts were available, allowing the tool to be run locally for all alternative alleles of all positions of interest. The metric used for MES was the percent difference between the scores for the reference and alternative alleles, with the greatest reduction in score classed as most pathogenic. All sites were also scored with CADD[30].

To allow cross-tool comparison, we ordered positions by increasing pathogenicity from each metric, and split positions into 20 brackets, such that the triplet based mutation rate for each bracket was equal, and the 20th bracket contained the positions with the most pathogenic scores. We calculated MAPS and the proportion of parental variants falling in high pLI genes for each bracket for all five metrics, as above, and looked at the number of DNMs in known dominant genes which fell in each bracket for the five metrics. Each of these analyses was conducted including and excluding CSS dinucleotide positions.

Splice region variants in the ClinVar database

We extracted all ClinVar[31] variants using the UCSC table browser[32] on 02.05.2017 and matched these against our splicing positions of interest, removing exonic sites with non-synonymous consequences. This resulted in 3603 positions with clinical significance recorded as pathogenic or likely pathogenic. We calculated the ratio of canonical to non-canonical splice positions within this data. Since each variant is present in this data only once, we used number of submissions as a proxy for allele count, and calculated the ratio of canonical to non-canonical variants adjusting for this. Differences between these observed values and our expectations, based on 27% of splice affecting mutations being in non-canonical positions, were assessed using Fisher's exact test (R v3.1.3).

Results

Signatures of purifying selection around the splice site

Since purifying selection acts to keep deleterious alleles rare, population variation data can be used to identify and assess the relative strengths of signals of purifying selection. To assess selective constraint acting on positions around the canonical splice site we used the MAPS metric[14] in 13,750 unaffected parents enrolled in the DDD study as well as >60,000 aggregated exomes from ExAC (Figure 1a). There is a high level of consistency between the data from the DDD and ExAC cohorts. The canonical splice acceptor and donor dinucleotides show a clear signal of purifying selection in both datasets. In the DDD data, we observed a difference between the strength of selection for the two positions within the canonical donor site, with the donor+1 site showing a signal of selection akin to stop-gained mutations, while the donor+2 site is lower, intermediate between the signal of selection observed at missense and nonsense sites. Concordant with this observation of apparently weaker purifying selection acting on variants at the CSS compared to nonsense variants, we identified that CSS variants represent 21.7% of likely protein truncating variants in highly constrained (pLI>0.9) genes in ExAC individuals without overt monogenic disorders, but only 13.9% of likely protein truncating DNMs in DDD probands. These proportions are significantly different ($p = 4.21 \times 10^{-6}$), and support the idea that, on average, CSS mutations have lower pathogenicity than other predicted truncating variants.

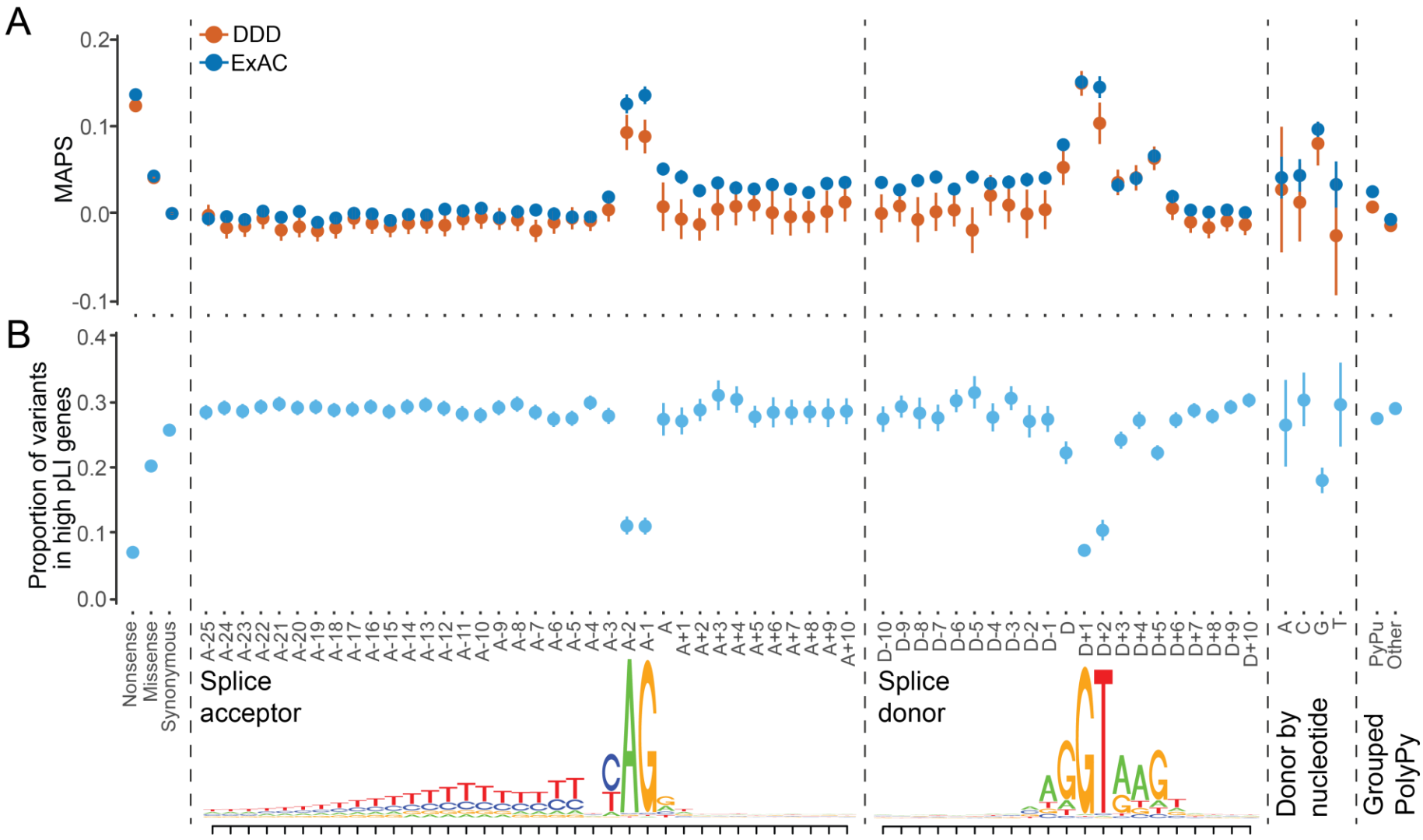
Outside of the CSS, other positions clearly show a signal of purifying selection beyond the background level, including the donor site (last base of the exon, which is particularly pronounced when the reference allele is G (Figure 1a)), and the intronic positions proximal to the canonical donor site, peaking at the don+5 position, which exhibits a signal of purifying selection intermediate between missense and nonsense variants. Although no sites within the PolyPy show a signal of purifying selection individually, when these sites are grouped together (Methods) and stratified by changes from a pyrimidine to a purine (PyPu), versus all other changes, there is a clear difference between the two types of variants, with PyPu changes exhibiting an increased signal of purifying selection when compared to non-PyPu changes ($p < 0.001$, Figure 1a, Figure S1).

Deficit of splicing variants in highly constrained (pLI>0.9) genes in healthy individuals

We also examined the distribution of variants of different classes among genes that are known to be under different levels of selective constraint. Highly constrained genes should contain fewer deleterious variants than less constrained genes. We investigated the proportion of variants observed in the 13,750 unaffected parents which fell within highly constrained genes with pLI > 0.9

Figure 1 – **Signals of purifying selection around splice sites**

A. Selective constraint across splicing region in 13,750 unaffected parents of DDD probands and >60,000 aggregated exomes from ExAC. Mutability adjusted proportion of singletons (MAPS) shown for VEP annotated exonic sites, extended splice acceptor and splice donor regions, the last base of the exon, split by reference nucleotide, and grouped sites in the polypyrimidine tract region (PolyPy), split by changes from a pyrimidine to a purine (PyPu) vs all other changes. B. Proportion of variants in 13,750 unaffected parents of DDD probands which fall within genes with high pLI (>0.9) across VEP annotated exonic sites, extended splice acceptor and splice donor regions, the last base of the exon, split by reference nucleotide, and grouped sites in the polypyrimidine tract region, split by changes from a pyrimidine to a purine (PyPu) vs all other changes.


A
DDD
ExAC
MAPS
0.2
0.1
0.0
-0.1
B
Proportion of variants in high pLI genes
0.4
0.3
0.2
0.1
0.0
Nonsense
Missense
Synonymous
A-25 A-24 A-23 A-22 A-21 A-20 A-19 A-18 A-17 A-16 A-15 A-14 A-13 A-12 A-11 A-10 A-9 A-8 A-7 A-6 A-5 A-4 A-3 A-2 A-1 A A+1 A+2 A+3 A+4 A+5 A+6 A+7 A+8 A+9 A+10
Splice acceptor
D-10 D-9 D-8 D-7 D-6 D-5 D-4 D-3 D-2 D-1 D D+1 D+2 D+3 D+4 D+5 D+6 D+7 D+8 D+9 D+10
Splice donor
A
C
G
T
Donor by nucleotide
PyPu
Other
Grouped PolyPy

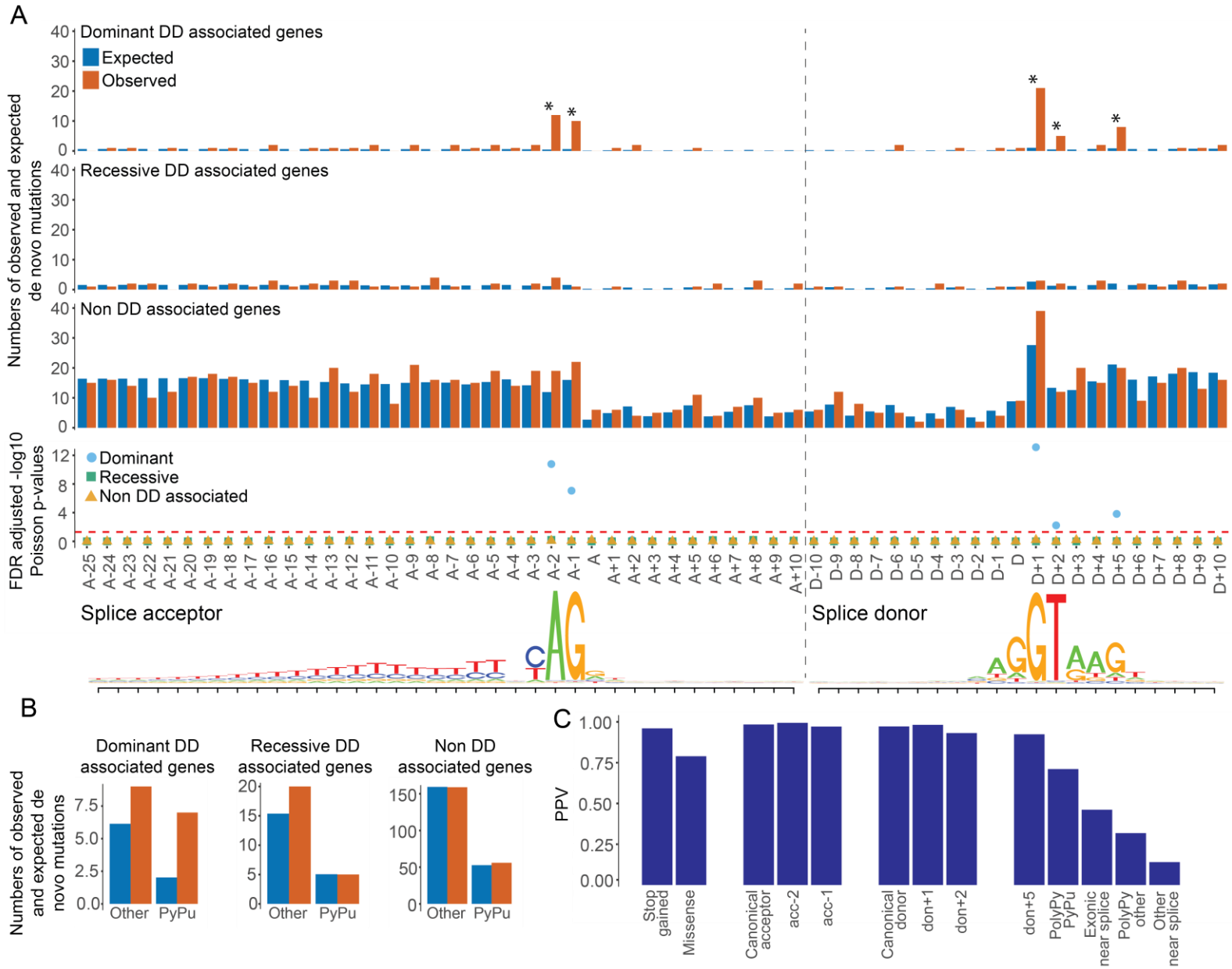
in our splicing regions of interest (Figure 1b). In the near splice positions at which the highest MAPS values were seen (CSS, donor, donor+5), we also observed a stronger depletion of variants in high pLI genes within the unaffected parents, again supporting the potential pathogenicity of variants at these positions. The proportion of parental variants in high pLI genes also recapitulates the signals of purifying selection seen in the MAPS analyses with regard to the donor position split by reference allele (Figure 1b) and the PolyPy region (Figure 1b, Figure S1), with the lowest proportions in high pLI genes observed for sites with the highest MAPS values.

Assessing the significance of mutational burden for different classes of splicing mutations

We identified 871 autosomal high confidence DNMs (non-synonymous consequences excluded) within canonical and near-splice regions of interest well covered by exome data in the 7,833 probands, allowing us to test for enrichment of DNMs relative to expectations based on a trinucleotide null model of mutation rate[22] across different sets of genes (DD-associated with dominant or recessive mechanisms, and non-DD associated, see Methods). Across recessive DD and non-DD associated genes, no enrichment of DNMs beyond the null expectation was observed (Figure 2a). In dominant DD genes, a significant cumulative excess of DNMs was noted across the full splicing region (Poisson p (FDR adjusted) = 2.60×10^{-14}, fold enrichment = 3.47), which remained significant upon exclusion of the canonical dinucleotide positions (Poisson p (FDR adjusted) = 0.0041, fold enrichment = 1.86). Individually, the four canonical splice site positions each showed a significant excess of DNMs (Poisson p (FDR adjusted), fold enrichment: acc-2 = 7.68×10^{-12}, 26.61; acc-1 = 6.02×10^{-8}, 16.57; don+1 = 2.60×10^{-14}, 20.06; don+2 = 0.0045, 9.99), as did the don+5 site (0.0001, 9.29). The greater fold enrichment in don+1 over don+2 sites suggests variants at don+2 may be less damaging, while the similar level of enrichment between don+5 and don+2 implies these positions harbour comparable levels of splice disrupting mutations. No individual positions within the PolyPy region showed an individual excess of DNMs, however, when the positions were considered cumulatively and split between PyPu and non-PyPu changes (Figure 2b), a significant excess of DNMs was observed in the PyPu group for dominant DD genes (Poisson p (raw) = 0.0048, fold enrichment = 3.46). These results are highly concordant with the signatures of purifying selection identified using the MAPS metric, and the deficit of parental variants in high pLI genes, providing multiple independent lines of evidence that the canonical positions do not contribute equally to splice site recognition, and that mutations in positions outside of the CSS can disrupt normal splicing.

Figure 2 – ***De novo*** **mutations around splice sites**

Enrichment of *de novo* mutations (DNMs) across the splicing region in 7,833 DDD probands A. Numbers of observed and expected DNMs across splicing region, in known dominant and recessive DD genes, as well as non-DD associated genes, with FDR corrected Poisson p-values. B. Aggregation of observed and expected numbers of DNMs in the polypyrimidine tract (PolyPy) region, with changes from a pyrimidine to a purine and all other changes shown separately for known dominant and recessive DD genes, as well as non-DD associated genes. C. Positive predictive values (PPVs) for *de novo* mutations in dominant DD-associated genes in positions across the splicing region, as well as VEP annotated stop gained and missense changes, calculated from observed and expected numbers of DNMs.


A
Dominant DD associated genes
Expected
Observed
Recessive DD associated genes
Non DD associated genes
Numbers of observed and expected de novo mutations
FDR adjusted -log10 Poisson p-values
Dominant
Recessive
Non DD associated
Splice acceptor
Splice donor
B
Dominant DD associated genes
Recessive DD associated genes
Non DD associated genes
Numbers of observed and expected de novo mutations
Other
PyPu
C
PPV
Stop gained
Missense
Canonical acceptor
acc-2
acc-1
Canonical donor
don+1
don+2
don+5
PolyPy PyPu
Exonic near splice
PolyPy other
Other near splice

1

Estimating positive predictive values for different classes of splice mutation

We used the fold-enrichment of the numbers of observed DNMs in dominant DD genes in the DDD cohort over the number expected under the null mutation model to calculate positive predictive values (PPV) for groupings of near splice site positions, independent of any prior clinical classification. We compared these with positive predictive values for other, more commonly disease-associated variant classes within the same exons of the same genes (Figure 2c). We observe minor differences in PPV for the individual positions of the canonical acceptor and donor sites, with don+2 showing the lowest PPV at 0.90, which is approximately the same as for the don+5 positions (PPV 0.89). Variants within the PolyPy region which change a pyrimidine for a purine have a PPV of 0.71, which is below the PPV for missense mutations (0.79), but still predicts a substantive number of pathogenic mutations arising from disruption of the PolyPy.

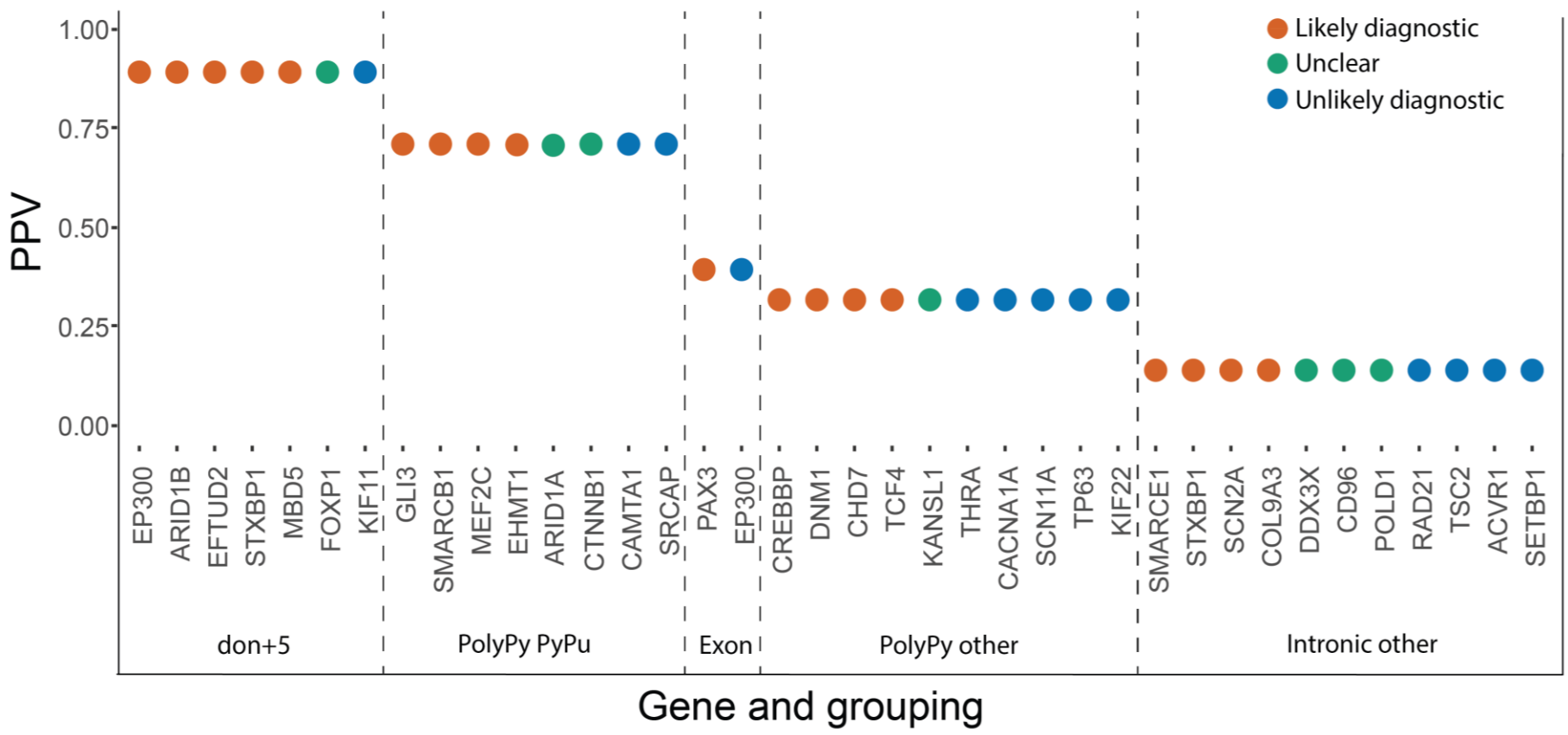
Despite the modest number of observed DNMs used to make these PPV estimates, we see striking concordance with the population based metrics described above (MAPS and deficit of splicing variants in high pLI genes in unaffected parents of DDD patients – Figure S2), suggesting these estimates are robust.

Identifying diagnostic non-canonical splice mutations

After exclusion of probands with likely diagnostic protein-coding or canonical splice site variants, 38 DNMs in our near splice site positions of interest in dominant DD genes were identified. The clinical phenotypes of patients carrying these mutations were reviewed by a consultant clinical geneticist, blinded to the precise mutation and PPVs estimated above, and the patient's recruiting clinician, to assess the phenotypic similarity between the proband and the disorder expected from a LoF mutation in that gene. The 38 variants were classified as likely diagnostic (Table 1), or unlikely diagnostic/unknown (Table 2), depending on the strength of phenotypic similarity. The clinical review resulted in 18 variants (47%) being classified as likely diagnostic, highly concordant with the number predicted from the overall PPV of non-canonical sites of 46%, moreover, a higher proportion of likely diagnostic variants were classified at sites with higher PPVs, calculated as described above (Figure 3). With 48 CSS DNMs observed within the same exons in our probands, we estimate that 73% of disease causing splice disrupting DNMs occur within the CSS, while 27% are in non-canonical, near-splice positions.

Figure 3 – Clinical classifications of non-canonical near splice *de novo* mutations

Relationship between clinical classifications of 38 splice region *de novo* mutations (DNMs) in undiagnosed DDD probands and positive predictive values (PPVs) calculated using observed and expected numbers of DNMs in 7,833 probands.

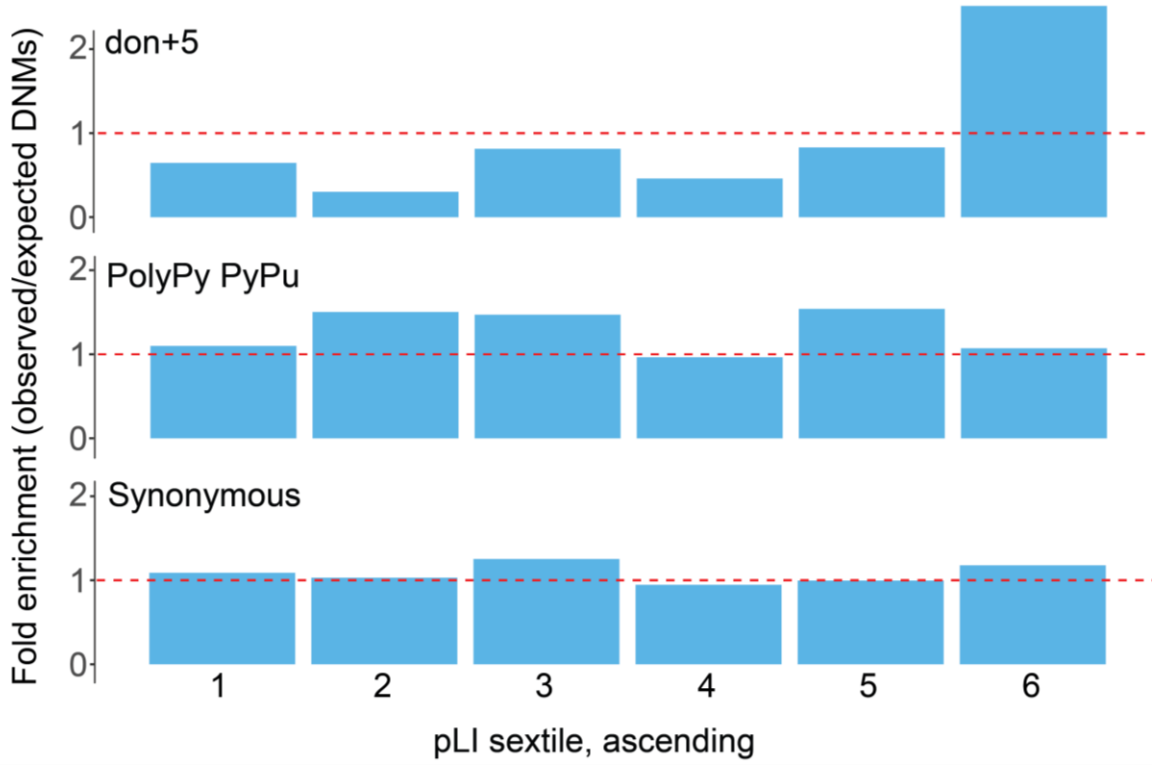


Eight DNMs were selected for functional validation via a minigene vector system, including six likely diagnostic PolyPy variants, a PolyPy variant of uncertain clinical significance, and a likely diagnostic don+5 variant, where both the phenotype of the patient and that associated with the gene (*MBD5*) are nonspecific, along with two negative controls (untransmitted variants identified in unaffected parents within the same PolyPys as test variants). For six of the variants selected for validation, differences in splicing between the reference and mutant constructs were observed (Figure S3a and S3b). The five PolyPy variants generating alterations in splice products all generated a cryptic splice site upstream of the canonical splice site, causing retention of part of the intron, in four instances leading to a frameshift effect, and in one leading to the inclusion of two additional amino acids in the protein sequence. The don+5 variant caused the utilisation of a second "GT" site within the reference sequence as a splice donor site, again causing retention of intronic sequence within the transcript, leading to a frameshift effect. For the *CHD7* variant, two splice products were observed, corresponding to the expected (wild type) splicing and the retention of 5bp intronic sequence. The *MBD5* variant gave multiple splice isoforms, with retention of 12bp intronic sequence being the most prevalent, but normally spliced, 19bp intronic retention, and complete intron retention also observed. Figure S3 shows the predominant isoform observed for these variants. One of the likely diagnostic PolyPy mutations, the PolyPy mutation of uncertain significance, and both negative controls showed no difference in splicing between the reference and mutant constructs (Figure S3c and S3d).

Given the concordant signals of purifying selection, enrichment of DNMs and number of likely diagnostic variants in the don+5 site and PolyPy (PyPu) region, we looked at the distribution of observed DNMs in genes with respect to their probability of being LoF intolerant (using the pLI metric[14], Figure 4). For synonymous variants, we observed no significant enrichment of DNMs in high pLI genes. For don+5 mutations, there is a clear excess of DNMs in genes most likely to be intolerant to LoF mutations in the DDD cohort, further supporting the likely pathogenicity of mutations in these positions. For the PolyPy PyPu mutations, although there is a nominally significant enrichment of DNMs in general, this does not show a significant skew towards high pLI genes in our cohort.

Figure 4 – Enrichment of *de novo* mutations by probability of loss of function intolerance

Enrichment (observed/expected) of *de* novo mutations (DNMs) by gene probability of loss of function intolerance (pLI), split in to sextiles for donor+5, pyrimidine to purine PolyPy, and synonymous sites. pLI scores encompassed by each sextile: 1 = 5.36E-91 - 0.000000605, 2 = 0.000000609 - 0.000558185, 3 = 0.000559475 - 0.027905143, 4 = 0.027908298 - 0.377456159, 5 = 0.377491926 - 0.919495985, 6 = 0.91955878 - 1.



Assessing splicing pathogenicity prediction tools

The population genetic metrics of purifying selection and mutation enrichment metric for pathogenicity that we have derived provide an orthogonal approach to assessing the accuracy of splicing pathogenicity prediction tools. We assessed four splicing pathogenicity prediction tools: two recently published genome-wide ensemble learning methods: AdaBoost and RandomForest, Spidex (based on deep learning trained on RNA sequencing data), and the longer standing, widely used MaxEntScan[27-29].

We divided the scores from each prediction tool, plus CADD[30], into 20 bins of equal mutation rate, to facilitate cross-method comparability. We calculated the MAPS for each bin of each of the scoring metrics for the splicing variants observed in the 13,750 DDD unaffected parents, and saw a strong positive correlation between pathogenicity metric and MAPS for all tools (Figure 5). AdaBoost had the highest absolute MAPS value for the top scoring bin, suggesting that it is best able to identify variants under the strongest purifying selection. The proportion of variants in the unaffected parents falling in genes with pLI > 0.9 broadly recapitulates this pattern, with fewer variants in high pLI genes in the highest scoring brackets for all metrics (Figure S4). We then looked at the distribution of scores for each tool for the 83 splicing DNMs observed in DDD probands in autosomal dominant DD-associated genes which were covered by all five scoring systems to compare performance of the metrics on mutations more likely to have a deleterious impact on splicing (Figure 6). Again, all metrics performed well, with the majority of DNMs being classified in the most deleterious score brackets. Here AdaBoost gave the highest area under the curve (AUC) value, suggesting it weighted these likely damaging variants as more deleterious than the other metrics comparatively. Interestingly, when CSS positions were removed from the analysis, AdaBoost remained the tool with the highest AUC. The largest reduction in the AUC metric was seen for Spidex and CADD, indicating these tools may be least informative for positions outside of the CSS. Upon removal of the CSS positions from the analyses of MAPS and deficit of parental variants in high pLI genes, similar results were revealed, with the highest AdaBoost scores retaining strong signals of purifying selection but a marked reduction in signal from the highest Spidex scores (Figure S5 and Figure S6).

Taken together, these data show a strong relationship between the considered splicing pathogenicity scoring systems and the general landscape of purifying selection on splicing control, but demonstrate that the utility of these systems in identifying likely diagnostic variants is limited outside of the CSS. However, it is worth noting that the scores reflect the probability of a variant affecting splicing, rather than overt phenotype. The high scoring variants may be affecting the splicing of transcripts, albeit not sufficiently to cause disease.

Figure 5 – Selective constraint and pathogenicity scores

Mutability adjusted proportion of singletons (MAPS) calculated for pathogenicity score brackets (least to most severe) in 13,750 unaffected parents from the DDD project, with Spearman correlation coefficient.

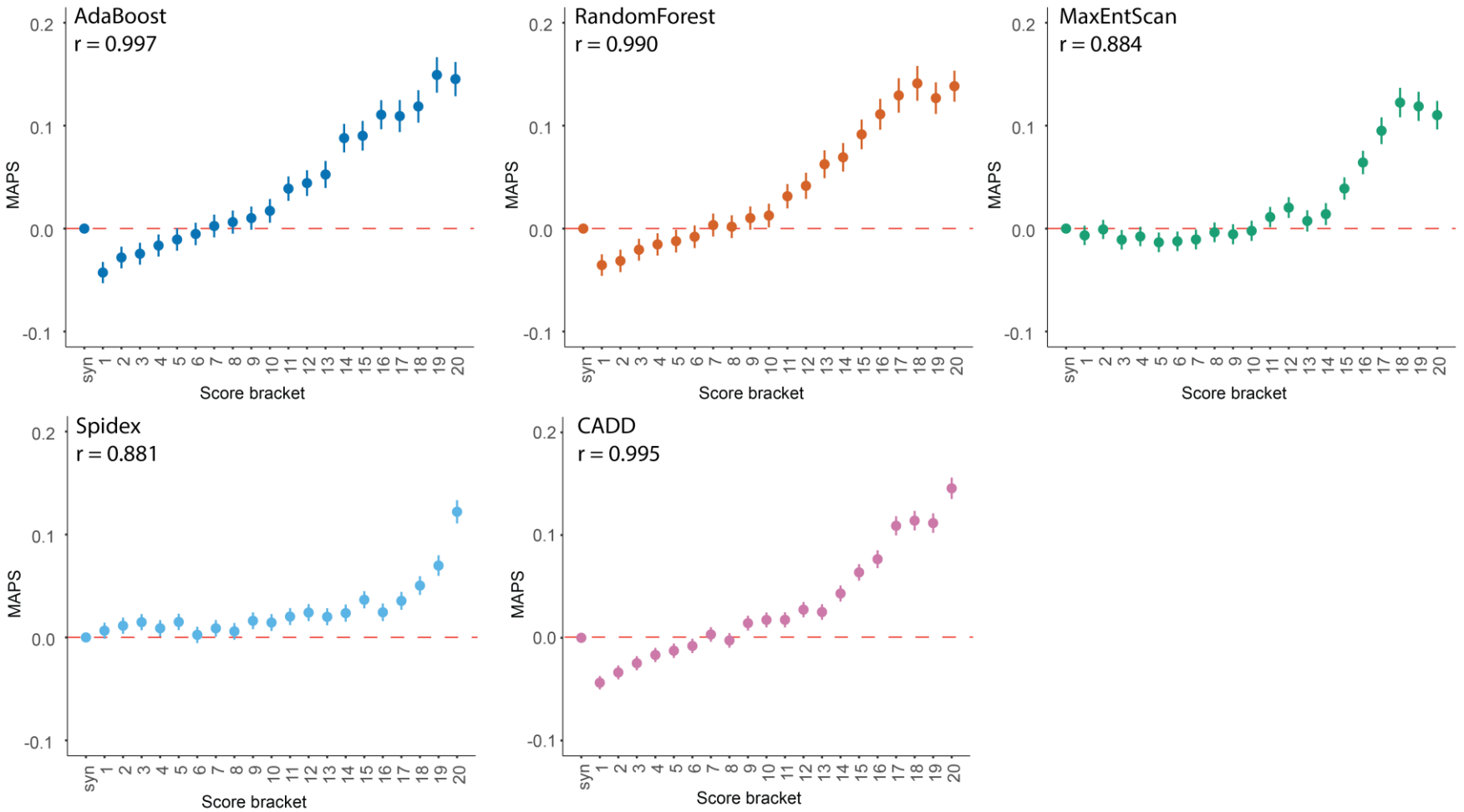
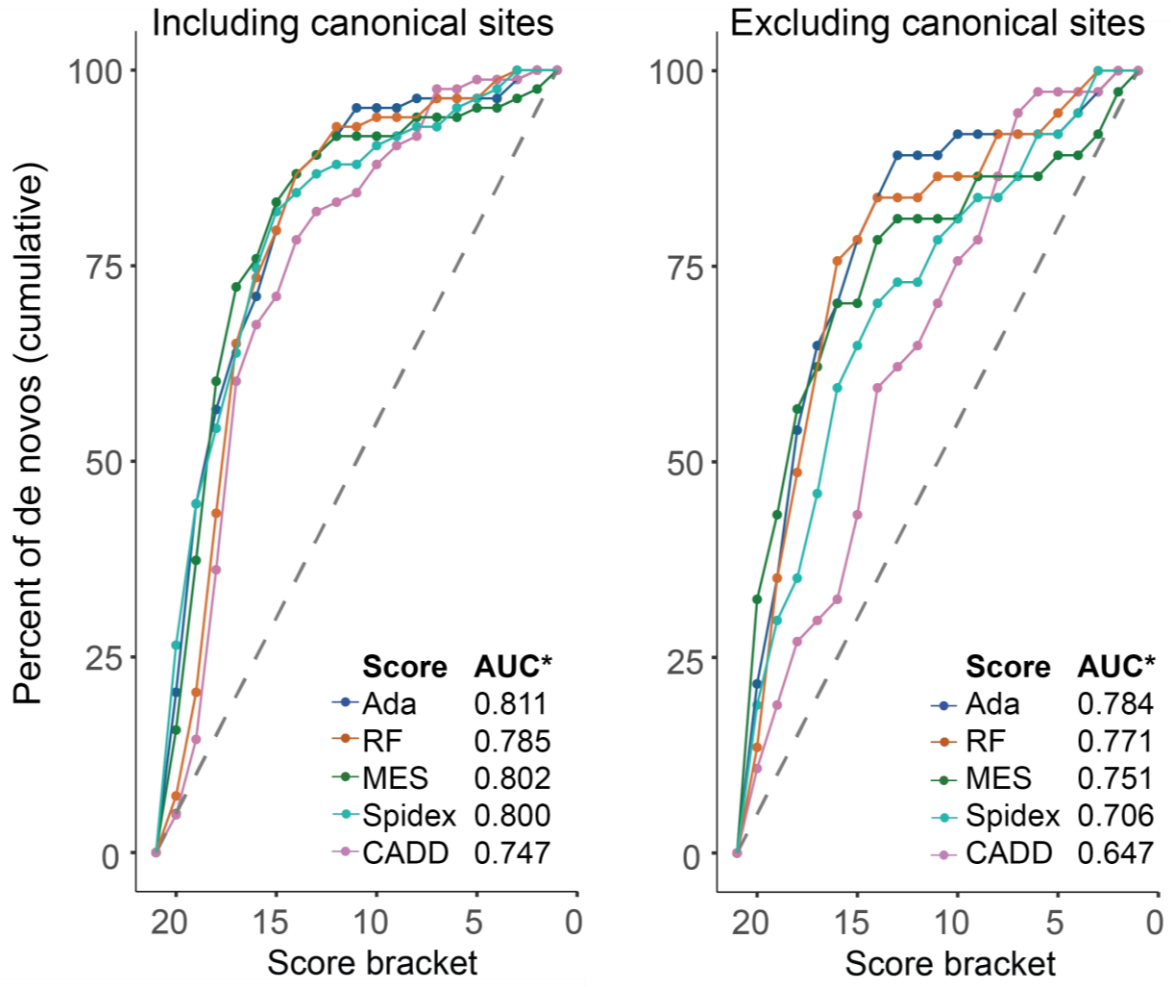


Figure 6 – Pathogenicity scores for observed near splice site *de novo* mutations

Cumulative percentage of *de novo* mutations (DNMs) in known dominant DD genes with decreasing pathogenicity score bracket, shown with canonical splice site positions included (left) and excluded (right). * AUC = area under curve



Discussion

Our analyses, taken together, suggest the pathogenic contribution of non-canonical splice sites has been under-appreciated. We estimate that around 27% of splice disrupting pathogenic mutations within the DDD cohort are in non-canonical positions. In sites with pathogenic or likely pathogenic clinical significance in ClinVar[31] overlapping with our splicing positions of interest (with non-synonymous consequences removed), we found 83.5% of variants fell within canonical positions, with just 16.5% in non-canonical positions. When adjusted for number of submissions as a proxy for allele count, this figure was 17.5%, perhaps indicative that recurrence strengthens evidence of pathogenicity. Both of these values are significantly below our estimate of 27% ($p = 1.22 \times 10^{-15}$ and $p = 2.2 \times 10^{-16}$ respectively), suggesting under-ascertainment of non-canonical splicing variants by around 35-40% in clinical databases.

Estimates of the relative contribution of canonical and non-canonical splice site mutations are sparse in the literature. When comparing canonical and non-canonical mutations within HGMD, Krawczak *et al*.[33] stated canonical mutations accounted for 64% of mutations at donor sites and 77.4% of mutations at acceptor sites, giving an estimated non-canonical contribution of ~30% overall, while data taken from Caminsky *et al*.[34] put this estimate at around 43%. These values are much closer to our 27% estimate than to the ClinVar proportion of ~17%, despite our approach focussing on DNMs and dominant disorders, whereas the other two studies did not discriminate on mode of inheritance and included recessive disorders, which can also be caused by non-canonical splicing mutations[35; 36] and exonic variants.

Our analysis of non-canonical splice position mutations did not include exonic missense variants[4; 5; 33], nor did it explicitly include branchpoints[37-40], splicing enhancers and suppressors[41; 42] or deep intronic mutations[43; 44]. Detecting splice disrupting variants at these sites is even more challenging, as despite recent efforts[45-49], comprehensive catalogues of all branchpoints and ESE/ISE/ESS/ISS are currently unavailable, algorithms that predict the impact of mutations at such sites are not highly accurate, and some of these sites are not covered by exome sequencing. As such, our estimate of the contribution of non-canonical splicing position mutations is likely to be a lower bound. If the higher estimated non-canonical contributions from Krawczak and Caminsky are more accurate, our estimate of 35-40% under-ascertainment in clinical databases may be conservative, and the true extent of missed diagnoses may be even higher.

The nature of developmental disorders makes obtaining RNA samples from relevant tissues of patients (i.e. neural tissue) acutely problematic, so we investigated the effects on splicing of several

of the potentially diagnostic DNMs using a minigene vector system. We were able to demonstrate changes to splicing for five out of six likely diagnostic PolyPy variants as well as the likely diagnostic don+5 variant. We did not observe an effect on splicing for one likely diagnostic PolyPy variant, and one PolyPy variant of uncertain clinical significance. Although the accuracy of minigene assays when compared with patient RNA is generally high[50-52], known limitations of the system (e.g. lack of full endogenous genetic context[53; 54], and sensitivity to cell type utilised[55]) mean we cannot definitively state that the effects seen in the minigene assay would be the same in the full genetic, developmental and cellular context within the patient.

We envisage that greater appreciation of the importance of near splice site mutations will increase diagnostic yields, as well as providing increased power for the detection of new genetic associations, both within the field of rare disease and beyond. We highlight two challenges to improving detection of pathogenic non-canonical splice site mutations.

First, many commonly used tools for annotating the likely functional impact of variants do not discriminate between different non-canonical splice site positions with very different likelihoods of being pathogenic. Moreover, commonly used annotation tools differ in the ways in which variants are annotated, with splicing variants displaying the highest level of disagreement between tools[56]. This highlights the need for a more consistent and evidence based annotation of splicing variants. Of the positions shown in our analyses to be most damaging, don+5 sites are annotated by VEP[19] and SnpEff[57] as “splice_region_variant”, while most positions of the PolyPy are annotated as intronic, so are potentially easily overlooked. With Annovar’s[58] default settings, only the CSSs are flagged as splicing variants, although with both Annovar and SnpEff, the user can optionally extend the region to be annotated as splice variants. We note that Ensembl have recently implemented a VEP plugin which allows greater granularity in splice region annotation (see web resources), including annotating the don+5 and other near-donor positions, as well as the PolyPy region. This type of increased granularity of splicing annotation should facilitate consideration of these variants in future studies.

Second, current tools that predict the pathogenicity of non-canonical splice site mutations have limited accuracy, and it is not clear how to translate the scores that they output into a likelihood of pathogenicity. The quantitative framework that we introduced here of estimating PPVs for different classes of mutations by comparing the number of observed mutations to the number expected under a well-calibrated null model of germline mutation has much more direct relevance to clinical interpretation. We propose that the scores generated by such splicing prediction tools could be

calibrated by performing analogous analyses of mutation enrichment to estimate PPVs for different bins of scores. As the size of trio-based cohorts increases, the accuracy of calibration will improve.

In summary, our results demonstrate a significant contribution of non-canonical splicing mutations to the genetic landscape of DDs, a finding which is highly likely to be recapitulated across other monogenic disorders and contexts. We demonstrate disparities in the control of splice site recognition between the two positions of the canonical splice-donor site, and the importance of other, non-canonical positions (particularly the don+5 site and pyrimidine-removing mutations in the PolyPy region). These inferences are supported by both population genetic investigations of purifying selection, as well as a disease based approach, considering the burden of DNMs in ~8,000 children with severe DDs. Mutations at some non-canonical splicing positions convey a risk of disease similar to that of protein truncating and missense mutations, but are commonly under-represented in existing databases of disease-causing variants.

Description of supplemental data

Supplemental data contains six figures and one table.

Conflicts of interest

M.E.H. is a co-founder of, consultant to, and holds shares in, Congenica Ltd, a genetics diagnostic company.

Acknowledgments

We thank the families for their participation and patience. We are grateful to the Exome Aggregation Consortium for making their data and code available. We thank the Sanger Human Genome Informatics and DNA pipelines teams for their support in generating and processing the data. We are grateful to Adam Frankish for advice selecting an appropriate exon set, and to Sarah Hunt and Fiona Cunningham for help and advice regarding splice annotation, and the development of the VEP SpliceRegion.pm plugin. Thanks also go to Alex Henderson, Anand Saggar, Diana Baralle, Elizabeth Jones, Emma Wakeling, Fleur van Dijk, Joan Paterson, Joanna Jarvis, Kate Chandler, Katherine Lachlan, Miranda Splitt, Neeti Ghali, Rachel Harrison, Sahar Mansour, Shane Mckee, Susan Tomkins and Victoria McKay for providing phenotypic information and insight on the probands with variants near splice sites not deemed to be diagnostic. The DDD study presents independent research commissioned by the Health Innovation Challenge Fund (grant HICF110091003), a parallel funding partnership between the Wellcome Trust and the UK Department of Health, and the Wellcome Trust Sanger Institute (grant WT098051). The views expressed in this publication are those of the author(s) and not necessarily those of the Wellcome Trust or the UK Department of Health. The study has UK Research Ethics Committee approval (10/H0305/83, granted by the Cambridge South Research Ethics Committee and GEN/284/12, granted by the Republic of Ireland Research Ethics Committee).

Web resources

denovoFilter, https://github.com/jeremymcrae/denovoFilter

Gencode v19, https://www.gencodegenes.org/releases/19.html

MAPS, https://github.com/pjshort/dddMAPS

Ensembl's VEP, https://www.ensembl.org/info/docs/tools/vep/index.html

DDG2P, http://www.ebi.ac.uk/gene2phenotype

ExAC, http://exac.broadinstitute.org/

HPO, http://compbio.charite.de/hpoweb/showterm?id=HP:0000118

SangerSeqR, http://bioconductor.org/packages/release/bioc/html/sangerseqR.html

ExPASy, https://web.expasy.org/translate/

CADD, http://cadd.gs.washington.edu/

ClinVar, https://www.ncbi.nlm.nih.gov/clinvar/

SpliceRegion.pm, https://github.com/Ensembl/VEP_plugins/blob/release/88/SpliceRegion.pm

OMIM, https://www.omim.org/

References

1. Brody, E., and Abelson, J. (1985). The "spliceosome": yeast pre-messenger RNA associates with a 40S complex in a splicing-dependent reaction. Science 228, 963-967.
2. Hang, J., Wan, R., Yan, C., and Shi, Y. (2015). Structural basis of pre-mRNA splicing. Science 349, 1191-1198.
3. Scotti, M.M., and Swanson, M.S. (2016). RNA mis-splicing in disease. Nat Rev Genet 17, 19-32.
4. Ars, E., Serra, E., Garcia, J., Kruyer, H., Gaona, A., Lazaro, C., and Estivill, X. (2000). Mutations affecting mRNA splicing are the most common molecular defects in patients with neurofibromatosis type 1. Hum Mol Genet 9, 237-247.
5. Teraoka, S.N., Telatar, M., Becker-Catania, S., Liang, T., Onengut, S., Tolun, A., Chessa, L., Sanal, O., Bernatowska, E., Gatti, R.A., et al. (1999). Splicing defects in the ataxia-telangiectasia gene, ATM: underlying mutations and consequences. Am J Hum Genet 64, 1617-1631.
6. Cartegni, L., Chew, S.L., and Krainer, A.R. (2002). Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet 3, 285-298.
7. Baralle, D., and Buratti, E. (2017). RNA splicing in human disease and in the clinic. Clin Sci (Lond) 131, 355-368.
8. Tang, R., Prosser, D.O., and Love, D.R. (2016). Evaluation of Bioinformatic Programmes for the Analysis of Variants within Splice Site Consensus Regions. Adv Bioinformatics 2016, 5614058.
9. Jian, X., Boerwinkle, E., and Liu, X. (2014). In silico tools for splicing defect prediction: a survey from the viewpoint of end users. Genet Med 16, 497-503.
10. Houdayer, C., Caux-Moncoutier, V., Krieger, S., Barrois, M., Bonnet, F., Bourdon, V., Bronner, M., Buisson, M., Coulet, F., Gaildrat, P., et al. (2012). Guidelines for splicing analysis in molecular diagnosis derived from a set of 327 combined in silico/in vitro studies on BRCA1 and BRCA2 variants. Hum Mutat 33, 1228-1238.
11. Iossifov, I., O'Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D., Stessman, H.A., Witherspoon, K.T., Vives, L., Patterson, K.E., et al. (2014). The contribution of de novo coding mutations to autism spectrum disorder. Nature 515, 216-221.
12. Wright, C.F., Fitzgerald, T.W., Jones, W.D., Clayton, S., McRae, J.F., van Kogelenberg, M., King, D.A., Ambridge, K., Barrett, D.M., Bayzetinova, T., et al. (2015). Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. Lancet 385, 1305-1314.
13. Deciphering Developmental Disorders, S. (2017). Prevalence and architecture of de novo mutations in developmental disorders. Nature 542, 433-438.
14. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285-291.
15. Akawi, N., McRae, J., Ansari, M., Balasubramanian, M., Blyth, M., Brady, A.F., Clayton, S., Cole, T., Deshpande, C., Fitzgerald, T.W., et al. (2015). Discovery of four recessive developmental disorders using probabilistic genotype and phenotype matching among 4,125 families. Nat Genet 47, 1363-1369.
16. Deciphering Developmental Disorders, S. (2015). Large-scale discovery of novel genetic causes of developmental disorders. Nature 519, 223-228.
17. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754-1760.
18. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20, 1297-1303.
19. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol 17, 122.

20. Ramu, A., Noordam, M.J., Schwartz, R.S., Wuster, A., Hurles, M.E., Cartwright, R.A., and Conrad, D.F. (2013). DeNovoGear: de novo indel and point mutation discovery and phasing. Nat Methods 10, 985-987.
21. Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res 22, 1760-1774.
22. Samocha, K.E., Robinson, E.B., Sanders, S.J., Stevens, C., Sabo, A., McGrath, L.M., Kosmicki, J.A., Rehnstrom, K., Mallick, S., Kirby, A., et al. (2014). A framework for the interpretation of de novo mutation in human disease. Nat Genet 46, 944-950.
23. Kohler, S., Vasilevsky, N.A., Engelstad, M., Foster, E., McMurry, J., Ayme, S., Baynam, G., Bello, S.M., Boerkoel, C.F., Boycott, K.M., et al. (2017). The Human Phenotype Ontology in 2017. Nucleic Acids Res 45, D865-D876.
24. Singh, T., Kurki, M.I., Curtis, D., Purcell, S.M., Crooks, L., McRae, J., Suvisaari, J., Chheda, H., Blackwood, D., Breen, G., et al. (2016). Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders. Nat Neurosci 19, 571-577.
25. Hill, J.T., Demarest, B.L., Bisgrove, B.W., Su, Y.C., Smith, M., and Yost, H.J. (2014). Poly peak parser: Method and software for identification of unknown indels using sanger sequencing of polymerase chain reaction products. Dev Dyn 243, 1632-1636.
26. Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., de Castro, E., Duvaud, S., Flegel, V., Fortier, A., Gasteiger, E., et al. (2012). ExPASy: SIB bioinformatics resource portal. Nucleic Acids Res 40, W597-603.
27. Jian, X., Boerwinkle, E., and Liu, X. (2014). In silico prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Res 42, 13534-13544.
28. Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., Yuen, R.K., Hua, Y., Gueroussov, S., Najafabadi, H.S., Hughes, T.R., et al. (2015). RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. Science 347, 1254806.
29. Yeo, G., and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol 11, 377-394.
30. Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46, 310-315.
31. Landrum, M.J., Lee, J.M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., et al. (2016). ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res 44, D862-868.
32. Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J. (2004). The UCSC Table Browser data retrieval tool. Nucleic Acids Res 32, D493-496.
33. Krawczak, M., Thomas, N.S., Hundrieser, B., Mort, M., Wittig, M., Hampe, J., and Cooper, D.N. (2007). Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing. Hum Mutat 28, 150-158.
34. Caminsky, N., Mucaki, E.J., and Rogan, P.K. (2014). Interpretation of mRNA splicing mutations in genetic disease: review of the literature and guidelines for information-theoretical analysis. F1000Res 3, 282.
35. Brunham, L.R., Kang, M.H., Van Karnebeek, C., Sadananda, S.N., Collins, J.A., Zhang, L.H., Sayson, B., Miao, F., Stockler, S., Frohlich, J., et al. (2015). Clinical, Biochemical, and Molecular Characterization of Novel Mutations in ABCA1 in Families with Tangier Disease. JIMD Rep 18, 51-62.
36. Basel-Vanagaite, L., Hershkovitz, T., Heyman, E., Raspall-Chaure, M., Kakar, N., Smirin-Yosef, P., Vila-Pueyo, M., Kornreich, L., Thiele, H., Bode, H., et al. (2013). Biallelic SZT2 mutations cause infantile encephalopathy with epilepsy and dysmorphic corpus callosum. Am J Hum Genet 93, 524-529.

37. Di Leo, E., Panico, F., Tarugi, P., Battisti, C., Federico, A., and Calandra, S. (2004). A point mutation in the lariat branch point of intron 6 of NPC1 as the cause of abnormal pre-mRNA splicing in Niemann-Pick type C disease. Hum Mutat 24, 440.
38. Crotti, L., Lewandowska, M.A., Schwartz, P.J., Insolia, R., Pedrazzini, M., Bussani, E., Dagradi, F., George, A.L., Jr., and Pagani, F. (2009). A KCNH2 branch point mutation causing aberrant splicing contributes to an explanation of genotype-negative long QT syndrome. Heart Rhythm 6, 212-218.
39. Maslen, C., Babcock, D., Raghunath, M., and Steinmann, B. (1997). A rare branch-point mutation is associated with missplicing of fibrillin-2 in a large family with congenital contractural arachnodactyly. Am J Hum Genet 60, 1389-1398.
40. Aten, E., Sun, Y., Almomani, R., Santen, G.W., Messemaker, T., Maas, S.M., Breuning, M.H., and den Dunnen, J.T. (2013). Exome sequencing identifies a branch point variant in Aarskog-Scott syndrome. Hum Mutat 34, 430-434.
41. Liu, H.X., Cartegni, L., Zhang, M.Q., and Krainer, A.R. (2001). A mechanism for exon skipping caused by nonsense or missense mutations in BRCA1 and other genes. Nat Genet 27, 55-58.
42. Lorson, C.L., Hahnen, E., Androphy, E.J., and Wirth, B. (1999). A single nucleotide in the SMN gene regulates splicing and is responsible for spinal muscular atrophy. Proc Natl Acad Sci U S A 96, 6307-6311.
43. Cummings, B.B., Marshall, J.L., Tukiainen, T., Lek, M., Donkervoort, S., Foley, A.R., Bolduc, V., Waddell, L.B., Sandaradura, S.A., O'Grady, G.L., et al. (2017). Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci Transl Med 9.
44. Vaz-Drago, R., Custodio, N., and Carmo-Fonseca, M. (2017). Deep intronic mutations and human disease. Hum Genet.
45. Mercer, T.R., Clark, M.B., Andersen, S.B., Brunck, M.E., Haerty, W., Crawford, J., Taft, R.J., Nielsen, L.K., Dinger, M.E., and Mattick, J.S. (2015). Genome-wide discovery of human splicing branchpoints. Genome Res 25, 290-303.
46. Corvelo, A., Hallegger, M., Smith, C.W., and Eyras, E. (2010). Genome-wide association between branch point properties and alternative splicing. PLoS Comput Biol 6, e1001016.
47. Taggart, A.J., Lin, C.L., Shrestha, B., Heintzelman, C., Kim, S., and Fairbrother, W.G. (2017). Large-scale analysis of branchpoint usage across species and cell lines. Genome Res 27, 639-649.
48. Wang, Y., and Wang, Z. (2014). Systematical identification of splicing regulatory cis-elements and cognate trans-factors. Methods 65, 350-358.
49. Badr, E., ElHefnawi, M., and Heath, L.S. (2016). Computational Identification of Tissue-Specific Splicing Regulatory Elements in Human Genes from RNA-Seq Data. PLoS One 11, e0166978.
50. Bonnet, C., Krieger, S., Vezain, M., Rousselin, A., Tournier, I., Martins, A., Berthet, P., Chevrier, A., Dugast, C., Layet, V., et al. (2008). Screening BRCA1 and BRCA2 unclassified variants for splicing mutations using reverse transcription PCR on patient RNA and an ex vivo assay based on a splicing reporter minigene. J Med Genet 45, 438-446.
51. Thery, J.C., Krieger, S., Gaildrat, P., Revillion, F., Buisine, M.P., Killian, A., Duponchel, C., Rousselin, A., Vaur, D., Peyrat, J.P., et al. (2011). Contribution of bioinformatics predictions and functional splicing assays to the interpretation of unclassified variants of the BRCA genes. Eur J Hum Genet 19, 1052-1058.
52. van der Klift, H.M., Jansen, A.M., van der Steenstraten, N., Bik, E.C., Tops, C.M., Devilee, P., and Wijnen, J.T. (2015). Splicing analysis for exonic and intronic mismatch repair gene variants associated with Lynch syndrome confirms high concordance between minigene assays and patient RNA analyses. Mol Genet Genomic Med 3, 327-345.
53. Sangermano, R., Khan, M., Cornelis, S.S., Richelle, V., Albert, S., Elmelik, D., Garanto, A., Qamar, R., Lugtenberg, D., van den Born, L.I., et al. (2017). ABCA4 midigenes reveal the full splice spectrum of all reported non-canonical splice site variants in Stargardt disease. Genome Res.

54. Baralle, M., Skoko, N., Knezevich, A., De Conti, L., Motti, D., Bhuvanagiri, M., Baralle, D., Buratti, E., and Baralle, F.E. (2006). NF1 mRNA biogenesis: effect of the genomic milieu in splicing regulation of the NF1 exon 37 region. FEBS Lett 580, 4449-4456.
55. Lastella, P., Surdo, N.C., Resta, N., Guanti, G., and Stella, A. (2006). In silico and in vivo splicing analysis of MLH1 and MSH2 missense mutations shows exon- and tissue-specific effects. BMC Genomics 7, 243.
56. McCarthy, D.J., Humburg, P., Kanapin, A., Rivas, M.A., Gaulton, K., Cazier, J.B., and Donnelly, P. (2014). Choice of transcripts and software has a large effect on variant annotation. Genome Med 6, 26.
57. Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6, 80-92.
58. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38, e164.

Table 1 – **Diagnostic *de novo* mutations in non-canonical dinucleotide near splice positions**

Variant and proband information for 18 *de novo* likely diagnostic splice region variants identified in previously undiagnosed DDD probands in known dominant DD-associated genes (hg19 coordinates)

chrom:pos_ref/alt	symbol	VEP annotation	Splice annotation	Associated disorder	HPO terms	HPO terms (translation)	Clinical classification
7:42063221_G/C	*GLI3*	intron_variant	acc-14	Greig Cephalopolysyndactyly Syndrome	HP:0001841, HP:0010709, HP:0011304	2-4 finger syndactyly, Broad thumb, Preaxial foot polydactyly	Likely pathogenic, full contribution
16:3819367_C/T	*CREBBP*	intron_variant	acc-13	Rubinstein-Taybi Syndrome Type 1	HP:0000028, HP:0000179, HP:0000248, HP:0000252, HP:0000347, HP:0000486, HP:0001263, HP:0001510, HP:0001831, HP:0002019, HP:0002205, HP:0004691, HP:0011304	2-3 toe syndactyly, Brachycephaly, Broad thumbs, Constipation, Cryptorchidism, Global developmental delay, Growth delay, Microcephaly, Micrognathia, Recurrent respiratory infections, Short toes, Strabismus, Thick lower lip vermilion	Likely pathogenic, full contribution
22:24143120_T/G	*SMARCB1*	intron_variant	acc-11	Rhabdoid Predisposition Syndrome 1 / Coffin-Siris Syndrome 3	HP:0000294, HP:0000680, HP:0000696, HP:0000750, HP:0001263, HP:0001763, HP:0001999, HP:0002205, HP:0002213, HP:0007021, HP:0007096, HP:0010830, HP:0010877, HP:0100543, HP:0003045, HP:0002926, HP:0000708, HP:0000545, HP:0002164, HP:0001212	Unilateral strabismus, Delayed eruption of permanent teeth, Delayed eruption of primary teeth, Delayed speech and language development, Global developmental delay, Cognitive impairment, Impaired tactile sensation, Pain insensitivity, Pes planus, Recurrent respiratory infections, Fine hair, Low anterior hairline, Abnormal facial shape, Hypoplasia of the optic tract, Abnormality of the patella,	Likely pathogenic, full contribution

						Abnormality of thyroid physiology, Behavioural abnormality, Myopia , Nail dysplasia, Prominent fingertip pads	
18:52895603_T/C	*TCF4*	intron_variant	acc-11	Pitt-Hopkins Syndrome	HP:0000122, HP:0000252, HP:0000545, HP:0000646, HP:0001263, HP:0001999	Abnormal facial shape, Amblyopia, Global developmental delay, Microcephaly, Myopia, Unilateral renal agenesis	Likely pathogenic, full contribution
5:88025173_A/C	*MEF2C*	splice_region_variant	acc-9	Mental Retardation-Stereotypic Movements-Epilepsy And/Or Cerebral Malformations	HP:0000179, HP:0001250, HP:0002500, HP:0006579, HP:0011344, HP:0100023	Abnormality of the cerebral white matter, Prolonged neonatal jaundice, Recurrent hand flapping, Seizures, Severe global developmental delay, Thick lower lip vermilion	Likely pathogenic, full contribution
9:130988306_G/A	*DNM1*	splice_region_variant	acc-8	Epileptic Encephalopathy	HP:0001250, HP:0001263, HP:0001319, HP:0009117, HP:0011228, HP:0011344, HP:0011947	Aplasia/Hypoplasia of the maxilla, Global developmental delay, Horizontal eyebrow, Neonatal hypotonia, Respiratory tract infection, Seizures, Severe global developmental delay	Likely pathogenic, full contribution
8:61763045_G/A	*CHD7*	splice_region_variant	acc-7	CHARGE / Kallmann Syndrome Type 5 / Idiopathic Hypogonadotropic Hypogonadism	HP:0000185, HP:0000202, HP:0000589, HP:0001263, HP:0002564, HP:0003508, HP:0011678	Cleft soft palate, Coloboma, Global developmental delay, Oral cleft, Proportionate short stature, Tetralogy of Fallot with pulmonary atresia and major aortopulmonary collateral arteries, obsolete Malformation of the heart and great vessels	Definitely pathogenic, full contribution
17:38801875_T/C	*SMARCE1*	splice_region_variant	acc-4	Coffin-Siris Syndrome 5	HP:0000750, HP:0004322, HP:0005484	Delayed speech and language development, Postnatal microcephaly, Short stature	Likely pathogenic, full contribution

1:27097607_C/A	*ARID1A*	splice_region_variant	acc-3	Coffin-Siris Syndrome 2	HP:0000179, HP:0000347, HP:0000364, HP:0000369, HP:0000377, HP:0000490, HP:0000973, HP:0001338, HP:0001511, HP:0008935	Abnormality of the pinna, Cutis laxa, Deeply set eye, Generalized neonatal hypotonia, Hearing abnormality, Intrauterine growth retardation, Low-set ears, Micrognathia, Partial agenesis of the corpus callosum, Thick lower lip vermilion	Likely pathogenic, full contribution
9:140728798_C/G	*EHMT1*	splice_region_variant	acc-3	9q Subtelomeric Deletion Syndrome / Kleefstra Syndrome 1	HP:0002558, HP:0002020, HP:0002021, HP:0012716, HP:0008915, HP:0000248, HP:0030812, HP:0001800, HP:0012433, HP:0100543, HP:0000750, HP:0040082, HP:0010864, HP:0005100	Supernumerary nipple, Gastroesophageal reflux, Pyloric stenosis, Moderate conductive hearing impairment, Truncal obesity, Brachycephaly, Enlarged tonsils, Hypoplastic toenails, Abnormal social behaviour, Cognitive impairment, Delayed speech and language development, Happy demeanor, Intellectual disability, severe, Premature birth following premature rupture of fetal membranes	Definitely pathogenic, full contribution
2:223160248_T/C	*PAX3*	splice_region_variant	don-1	Waardenburg Syndrome, Type 1 / Craniofacial-Deafness-Hand Syndrome	HP:0000218, HP:0000316, HP:0000460, HP:0000527, HP:0000581, HP:0000582, HP:0002829, HP:0007429, HP:0007603, HP:0009889, HP:0000426, HP:0008573, HP:0000579, HP:0000402	Arthralgia, Blepharophimosis, Few cafe-au-lait spots, Freckles in sun-exposed areas, High palate, Hypertelorism, Localized hirsutism, Long eyelashes, Narrow nose, Upslanted palpebral fissure, High nasal bridge, Low-frequency sensorineural hearing impairment, Lacrimal duct obstruction, Narrow ear canal	Likely pathogenic, partial contribution

2:166229861_A/G	*SCN2A*	splice_region_variant	don+4	Nonspecific Severe Id / Benign Familial Neonatal Infantile Seizures / Infantile Epileptic Encephalopathy	HP:0000717, HP:0001344, HP:0002342, HP:0001250	Absent speech, Autism, Intellectual disability, moderate, Seizures	Likely pathogenic, full contribution
9:130422391_A/G	*STXBP1*	splice_region_variant	don+4	Angelman/Pitt Hopkins Syndrome-Like Disorder / Epileptic Encephalopathy Early Infantile Type 4	HP:0001048, HP:0001252, HP:0002599, HP:0011344	Cavernous hemangioma, Head titubation, Muscular hypotonia, Severe global developmental delay	Likely pathogenic, full contribution
22:41556731_G/A	*EP300*	splice_region_variant	don+5	Rubinstein-Taybi Syndrome Type 2	HP:0000023, HP:0000213, HP:0000220, HP:0000322, HP:0000369, HP:0000414, HP:0000486, HP:0000490, HP:0000527, HP:0001263, HP:0001537, HP:0001771, HP:0005484, HP:0007993, HP:0008551, HP:0008850, HP:0100023, HP:0000717	Achilles tendon contracture, Bulbous nose, Deeply set eye, Global developmental delay, Inguinal hernia, Long eyelashes, Low-set ears, Malformed lacrimal ducts, Microtia, Postnatal microcephaly, Recurrent hand flapping, Severe postnatal growth retardation, Short philtrum, Strabismus, Thin vermilion border, Umbilical hernia, Velopharyngeal insufficiency, Autism	Likely pathogenic, full contribution

2:149221493_G/C	*MBD5*	splice_region_variant	don+5	Ehmt1-Like Intellectual Disability	HP:0000252, HP:0000664, HP:0001601, HP:0002020	Gastroesophageal reflux, Laryngomalacia, Microcephaly, Synophrys	Likely pathogenic, full contribution
9:130427615_G/C	*STXBP1*	splice_region_variant	don+5	Angelman/Pitt Hopkins Syndrome-Like Disorder / Epileptic Encephalopathy Early Infantile Type 4	HP:0000733, HP:0002066, HP:0002378, HP:0002943, HP:0003763, HP:0007359, HP:0010864	Bruxism, Focal seizures, Gait ataxia, Hand tremor, Intellectual disability, severe, Stereotypy, Thoracic scoliosis	Likely pathogenic, full contribution
17:42956919_C/T	*EFTUD2*	splice_region_variant	don+5	Mandibulofacial Dysostosis With Microcephaly	HP:0000253, HP:0000286, HP:0000384, HP:0000396, HP:0000412, HP:0011343	Epicanthus, Moderate global developmental delay, Overfolded helix, Preauricular skin tag, Progressive microcephaly, Protruding ear	Definitely pathogenic, full contribution
20:61452890_C/G	*COL9A3*	splice_region_variant	don+8	Multiple Epiphyseal Dysplasia Type 3	HP:0000729, HP:0000750, HP:0010529, HP:0000735, HP:0001382, HP:0008947, HP:0000736, HP:0000733	Autistic behavior, Delayed speech and language development, Echolalia, Impaired social interactions, Joint hypermobility, Muscular hypotonia, Short attention span, Stereotypy	Likely pathogenic, partial contribution

1

Table 2 – ***De novo*** **mutations in non-canonical near splice positions not thought to be diagnostic**

Genomic coordinates and annotations of 20 *de novo* splice region variants identified in undiagnosed DDD probands in known dominant DD-associated genes, deemed unlikely to be diagnostic based on lack of phenotypic match between proband and associated syndrome (hg19 coordinates)

Variant	Gene	VEP annotation	Splice Annotation	Clinical classification
18:42618432_G/T	*SETBP1*	intron_variant	acc-18	Likely benign
3:38988415_AC/A	*SCN11A*	intron_variant	acc-17	Likely benign
17:38240072_A/G	*THRA*	intron_variant	acc-16	Likely benign
19:13387958_G/A	*CACNA1A*	intron_variant	acc-16	Likely benign
3:189456422_A/G	*TP63*	intron_variant	acc-9	Likely benign
1:7309543_GTTT/GTT	*CAMTA1*	splice_region_variant	acc-8	Likely benign
16:30745810_C/G	*SRCAP*	splice_region_variant	acc-7	Likely benign
16:29816431_G/A	*KIF22*	splice_region_variant	acc-5	Likely benign
22:41543944_C/T	*EP300*	synonymous_variant	don-6	Likely benign
10:94381235_G/T	*KIF11*	splice_region_variant	don+5	Likely benign
16:2129206_G/A	*TSC2*	intron_variant	don+9	Likely benign
2:158594942_G/T	*ACVR1*	intron_variant	don+10	Likely benign
19:50912018_C/T	*POLD1*	synonymous_variant	acc-24	Uncertain
3:41266439_T/G	*CTNNB1*	splice_region_variant	acc-6	Uncertain
17:44159911_GC/G	*KANSL1*	splice_region_variant	acc-3	Uncertain
3:71021701_C/T	*FOXP1*	splice_region_variant	don+5	Uncertain
6:157431700_G/A	*ARID1B*	splice_region_variant	don+5	Uncertain
X:41196724_T/G	*DDX3X*	splice_region_variant	don+6	Uncertain
3:111366523_A/C	*CD96*	intron_variant	don+10	Uncertain
8:117869033_G/C	*RAD21*	intron_variant	acc-23	Uncertain