Quantification of the evolutionary constraint on intronic branch point sequences in the bovine and human genome

Running title: Evolutionary constraint on branch point sequences

Naveen Kumar Kadri and Hubert Pausch

Animal Genomics, ETH Zürich, Eschikon 27, 8315 Lindau, Switzerland

Corresponding author: Naveen Kumar Kadri (naveen.kadri@usys.ethz.ch)

Abstract

We predict the branch point sequence in 180,892 bovine introns and investigate the variability of this cis-acting splicing element at nucleotide resolution. A degenerate heptamer "nnyTrAy" with conserved thymine and adenine residues at positions 4 and 6 constitutes the bovine branch point sequence. This motif is located between 18 and 37 bp upstream of the 3' splice acceptor site. More than 90% of the introns contain canonical branch point sequence motifs with thymine and adenine residues at positions 4 and 6. Using a catalogue of 29.4 million variants detected in 266 cattle, we show that the conserved thymine and adenine residues are strongly depleted for mutations. These two intronic residues harbor 39% and 41% less variants than coding sequences suggesting extreme purifying selection. The evolutionary constraint is higher at the thymine residue than the branch point adenine itself in heptamers with a canonical branch point sequence motif. We replicate these observations in human branch point sequences with a catalogue of more than 115 million SNPs detected in 3,942 genomes from the gnomAD database. Our study suggests that the evolutionary constraint is stronger on intronic branch point sequences than coding regions in the bovine and human genome.

Keywords: Branch point sequence, regulatory variants, cattle

Introduction

Non-coding sequences of the genome are an important source of genetic variation. Transcriptome analyses in large cohorts of humans, model organisms and livestock have revealed regulatory elements in non-coding sequences that modulate gene expression and alternative splicing (Aguet et al. 2019; Liu et al. 2020). Such regulatory elements, in particular expression and splicing QTL, contribute disproportionately to the heritability of complex traits and diseases (Xiang et al. 2019)

Pre-mRNA splicing by the spliceosome is evolutionarily conserved across eukaryotes (Ast 2004). The spliceosome complex consists of five small ribonucleic protein particles (snRNP), namely U1, U2, U4, U5 and U6 and a multitude of other proteins. These sub-units of the spliceosome are sequentially recruited and assembled anew (Hoskins and Moore 2012) at the intron-exon boundary to remove non-coding sequences (introns) and ligate coding sequences (exons) together. Sequence motifs necessary for recruitment and assembly of the spliceosome are contained in both intronic and exonic sequences (Will and Lührmann 2011). Cis-acting intronic elements include the highly conserved splice acceptor and donor sites at the exon-intron boundaries, the branch point sequence and a polypyrimidine tract between the branch point and the 3' splice acceptor site (Coolidge et al. 1997).

The branch point is contained within a degenerate intronic heptamer called the branch point sequence (Reed and Maniatis 1988). This motif is typically between 18 and 37 bases upstream of the 3' splice site (Green 1986). During pre-mRNA splicing, the branch point sequence undergoes base-pairing with a conserved 6-bp motif (GUAGUA) of the U2 snRNA. All residues of the heptamer except the branch point undergo base-pairing (Taggart et al.

2017). The unpaired branch point, usually adenine, bulges out of the U2 snRNA-pre-mRNA duplex. An attack of the unpaired branchpoint nucleotide to the 5' splice site, frees the 5' exon and forms a lariat intermediate. Subsequently, the free 5' exon attacks the 3' splice site resulting in exon ligation, and release and degradation of the intron lariat (Hoskins and Moore 2012).

Mutations affecting the branch point sequence may compromise the assembly of the spliceosome, leading to dramatic changes in splicing and gene expression with disease consequences (Kuivenhoven et al. 1996; Taggart et al. 2012; Cooper et al. 2009). Despite their functional importance, branch point sequences are not systematically studied because they are not annotated in gene transfer files. Thus, widely used tools to predict functional consequences of genetic variants such as Ensembl's Variant Effect Predictor (McLaren et al. 2016) are largely blind to branch point sequences.

Intronic branch point sequences have been identified experimentally in humans and model organisms (Gao et al. 2008; Awan et al. 2013). Computational methods leveraging the features of a high confidence set of few experimentally detected branch point sequences enable systematic assessment of this intronic motif at a genome-wide scale (Zhang et al. 2017; Signal et al. 2018; Paggi and Bejerano 2017). We herein, for the first time, predict and characterize intronic branch point sequences in the bovine genome. Utilizing a variant catalogue established from 266 whole-genome sequenced cattle, we quantify evolutionary constraints on the intronic branch point sequence and other well-annotated genomic features and compare them to constraints on the same features in the human genome.

Results

The splice sites are more constrained than the coding sequence

To quantify the evolutionary constraint on functionally important sequence elements in the bovine genome, we extracted the coordinates of nine coding and non-coding genomic features from the annotation (Ensembl version 101) of the bovine reference genome (Rosen et al. 2020). Our analysis included 35,848 transcripts from 20,785 genes with 204,123 exons and 181,080 introns, as well as 17,759 3' UTR, 25,490 5' UTR, 30,363 stop codons, 29,000 start codons, 172,269 3' splice sites, and 173,749 5' splice sites. Using a variant catalogue established from whole-genome sequencing of 266 cattle from eleven breeds (Fig 1A) sequenced to an average 16.28 (±6.56)-fold coverage, we assessed the number of polymorphic sites overlapping these features.

From the variant catalogue, we considered 29,227,937 biallelic and 190,200 multiallelic SNPs. On average, we found a variable site every 85 base pairs (assuming an autosomal genome length of 2,489,368,272), i.e., 1.18 variable sites per 100 base pairs. Genic features (introns, UTRs, splice-sites, exons, start codons and stop codons) were less and intergenic regions more variable than the genome-wide average. The splice sites were the least variable feature with 0.12 and 0.17 variants per 100 bp, respectively, at 3'and 5' ends (Fig 1B). In exons, we observed 0.81 variants per 100 bp. Features with low variation were generally enriched for rare variants (Fig 1C).

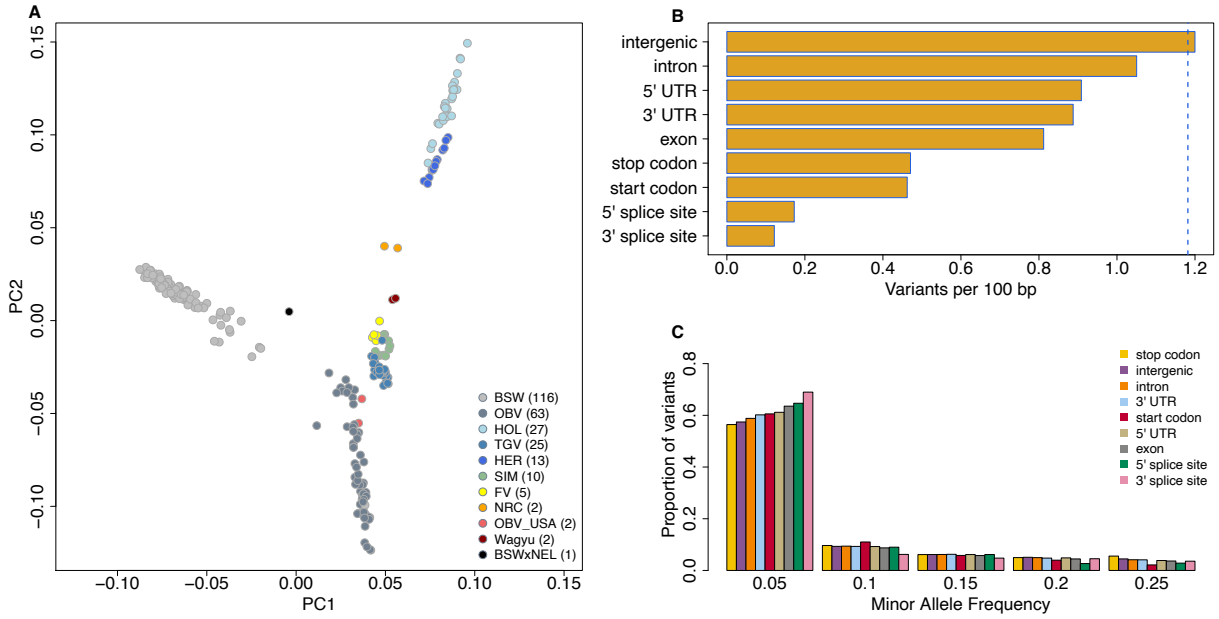

Figure 1. Constraints on well-annotated features uncovered from a large bovine variant catalogue. Plot of the first two principal components (PC) of a genomic relationship matrix among 266 cattle samples used to establish the variant catalogue (A). Individuals are colored according to breeds (BSW=Brown Swiss, OBV=Original Braunvieh, HOL=Holstein, TGV=Tyrolean Grey, HER=Hereford, SIM=Simmental, FV=Fleckvieh, NRC=Norwegian Red Cattle, NEL=Nellore). Values in parentheses indicate the number of samples per breed. Variation (B) and allele frequency spectrum (C) within nine annotated genomic features. Variation is expressed as the number of variable nucleotides per 100 bp of the feature. The blue dotted line indicates the average genome-wide variation.

Next, we, assessed constraints on nucleotides surrounding (±21 bp) the two splice sites. The number of polymorphic sites detected varied considerably at the exon-intron boundaries. The two splice sites with conserved GT and AG sequence motifs at 5' and 3' ends of introns, respectively, harbored 79% and 85% less variation than exonic nucleotides. Intronic nucleotides overlapping the splice donor site (last three exonic and first eight intronic nucleotides at the 5' splice site) and splice acceptor site (last twelve intronic and the first two exonic nucleotides at the 3' splice site) (Burge et al. 1993) were less variable than intronic positions on average. The first three intronic nucleotides next to the 5' splice site were strongly depleted for mutations. A consensus sequence motif "GTAAG" derived from

nucleotides overlapping 180,892 exon-intron boundaries indicated that these positions are conserved (Fig. 2B). At the splice acceptor site, the intronic nucleotide 2 bp upstream of the 3 'splice site was more variable than its surrounding nucleotides. The consensus sequence motif derived from 180,892 intron-exon boundaries also indicates that this position is not conserved (Fig. 2B). The relatively low variability at intronic positions upstream the 3' splice site is likely due to constraints on the polypyrimidine tract.

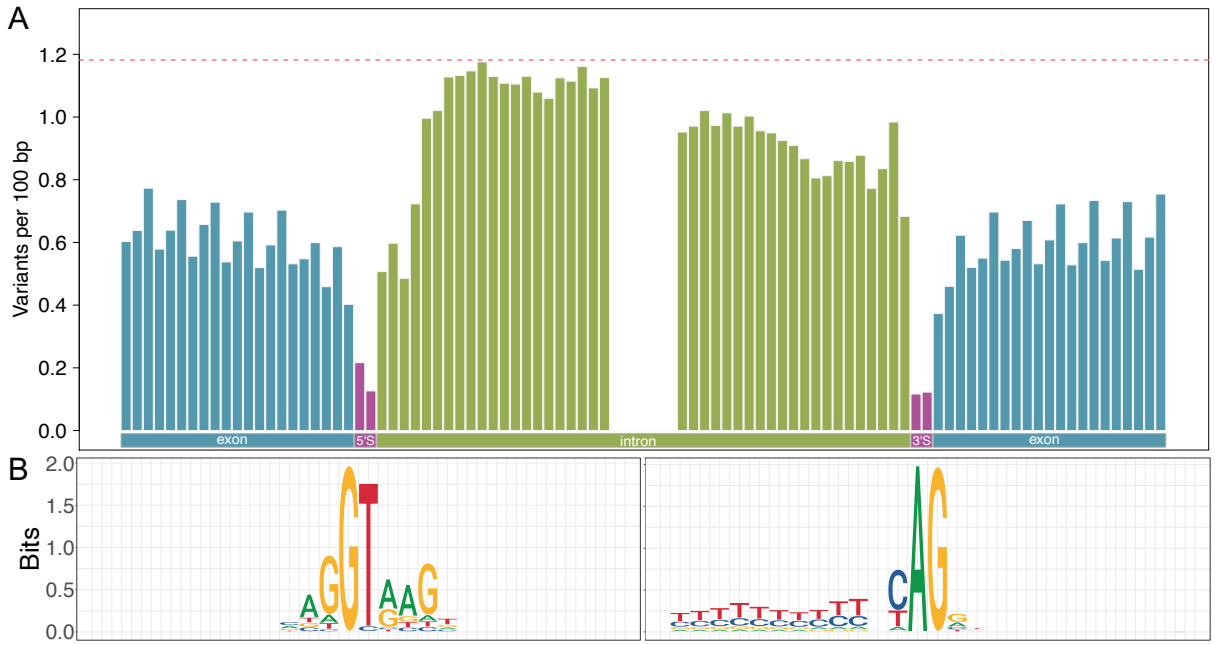

Figure 2. Selective constraint on splice sites.
Variation observed at 21 bases up- and downstream of the 3' and 5' splice sites (A). The red dotted line represents the average genome-wide variation. The conservation of nucleotides at 5' and 3' splice sites is shown in sequence logo plots (B).

In the exonic bases, the first and second positions of nucleotide triplets (group of 3 consecutive nucleotides) were more strongly depleted for variation than the third position reflecting wobble at the third position of the triplet. This periodic pattern is visible even though only a subset (52%) of exons starts with the first base of the coding triplet and ends with the third base of the coding triplet (Figure 2). We found 0.81, 0.68 and 0.97 variable

sites per 100 bp at nucleotides overlapping the first, second and third position, respectively, of coding triplets (Supplemental Figure 1). The last exonic position preceding the 5' splice site, however, was an exception to this periodic pattern, probably due to the contribution of a conserved guanine in canonical splicing (Fig. 2B).

Prediction of bovine branch point sequences

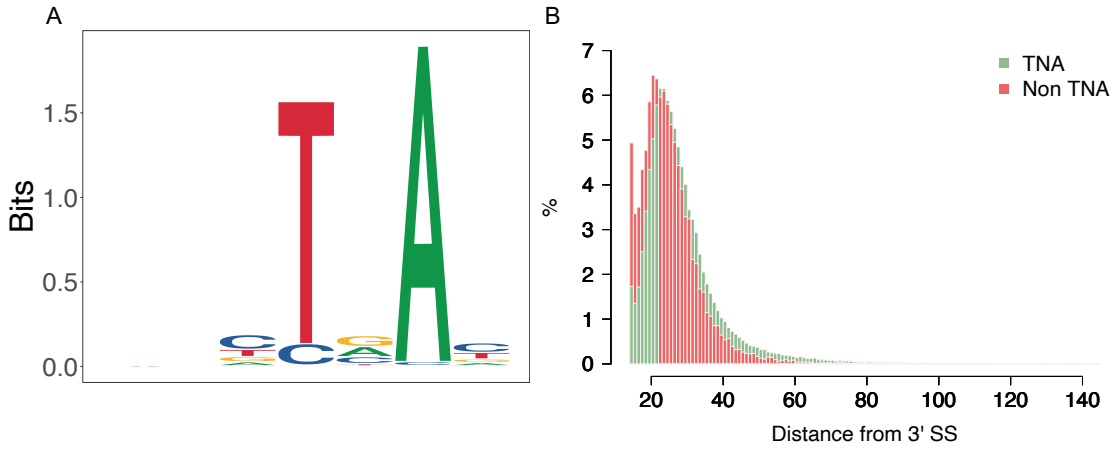

Figure 3: Characteristics of the branch point sequence in bovine introns.
Sequence logo plot generated from 180,892 branch point sequences identified in the bovine genome (A). Distribution of the distance between the predicted branch point and the 3' splice site for canonical (TNA) (green) and non-canonical (non-TNA) (red) branch point sequences(B).

The coordinates of splice sites are accessible from gene annotation files. However, other intronic features involved in the recognition and assembly of the spliceosome, such as the branch point sequence are not annotated. To make this feature amenable to our analysis, we extracted the sequences of 180,892 introns and predicted putative branch points with the BPP software package (Zhang et al. 2017).

Intron length ranged from 22 to 1,196,801 with a median of 1361 bp. A vast majority (n=173,166; 95.73%) of the introns had canonical GT and AG splice sites at 5' and 3' ends,

respectively. The second most frequent (n=6480; 3.59%) type of introns had GC and AG splice sites at 5' and 3' ends, respectively. The U12-type introns with AT and AC splice sites that are spliced by minor spliceosomes were exceedingly rare (n=125, 0.07%). A few introns (n=1121, 0.62%) had other splice sites.

The branch points resided between 14 and 145 bp upstream of the 3' splice site with a mean distance of 28.9±11.3 bp (Fig 3B). (Fig. 3A). The predicted branch point was predominantly adenine (n=178,440, 98.64%). Cytosine (n=2305, 1.27%) or thymine (n=147, 0.08%) residues were rarely predicted as branch points. A thymine residue was highly conserved (91.65%, n=165,788) at the 4th position of the heptamer. The heptamer preferentially contained purines (69.94%) at the 5th, and pyrimidines at the 3rd (70.30%) and 7th (70.16%) position.

The vast majority (90.56%, n=163822) of the heptamers contained a canonical "TNA" motif with thymine and adenine residues at positions 4 and 6. Heptamers with cytosine and adenine residues at position 4 and 6 ("CNA"), respectively, were the second most frequent (7.77%, n=14,049) class among the predicted branch point sequences.

Heptamers with a "TNA" motif were more distant to the 3' splice site than those with other motifs (29.19 bp vs 25.74 bp between the branch point residue and the 3' splice site; $p<2.2x10^{-16}$). Further, branch points were more distant from the 3' splice site in canonical (GT-AG type) than other introns (28.95 bp vs 27.00 bp, t-test: $p=2.0x10^{-51}$). Heptamers with a "TNA" motif were more frequent in GT-AG type than other introns (Supplemental Figure 2; 91.27% vs 74.62%: Fisher's exact test $p<2.5x10^{-323}$). Conversely, heptamers with the non-canonical "CNA" motif were more frequent (18.39% vs 7.29% $p=3.70x10^{-211}$) in non-canonical introns.

Strong selective constraint on the bovine branch point sequences

Next, we investigated the variation within the predicted branch point sequences and surrounding (±21) nucleotides. Overall, nucleotides within the heptamer were less variable than the (N=42) surrounding intronic nucleotides (Fig. 4). However, the constraint on the individual bases varied considerably within the heptamer. The reduced overall variability of the heptamer was primarily due to a striking depletion of variants at the predicted branch point (0.48 variants per 100 bp) and the residue at position 4 (0.50 variants per 100 bp) of the heptamer. Notably, these two bases were less variable than the exonic sequence (cf. Fig. 1B). The 5th position of the heptamer was also less variable (0.77 variants per 100 bp) than the intronic nucleotides surrounding the heptamer. However, the positions 1, 2, 3, and 7 of the heptamer (0.93 - 1.08 variants per 100 bp) were as polymorphic as adjacent intronic nucleotides.

The depletion of variation at positions 4 and 6 was more pronounced within heptamers that contained the canonical "TNA" motif. Within these heptamers, the thymine residue at position 4 (0.40 variants per 100 bp) was less variable than the branch point adenine (0.42 variants per 100 bp). The difference, however, was not statistically significant (Fisher's exact test; p=0.35).

We observed the same pattern at 179,093 branch point sequences predicted using LaBranchoR (Paggi and Bejerano 2018) (Supplemental Figure 3). The depletion of variation at positions 4 (0.31 variants per 100 bp) and 6 (0.32 variants per 100 bp) was even stronger in a subset of 109,199 branch point sequences that were placed at the same intronic position by both BPP and LaBranchoR (Supplemental Figure 4).

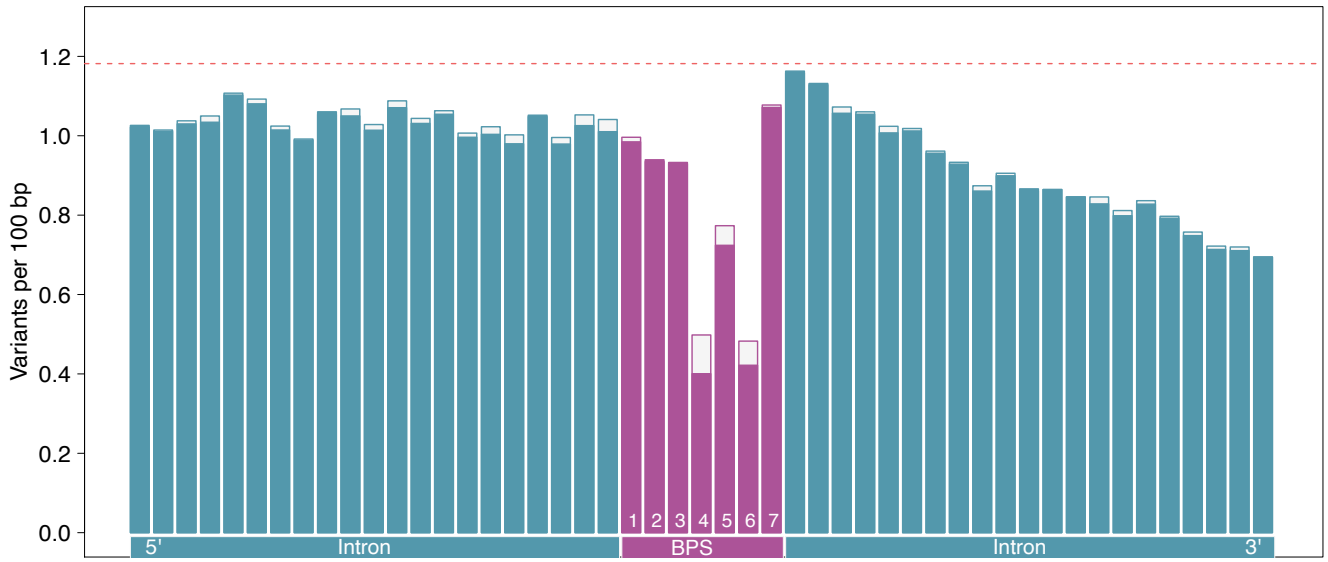

Figure 4. Local variation at the bovine branch point sequence (BPS).
The height of the bars indicates the variation (variants per 100 bp) surrounding 180,892 branch points while the solid bars indicate the variation surrounding a subset of 165,788 branch points encompassed by heptamers with a canonical "TNA" motif. The dotted horizontal red line indicates the average genome-wide variation.

The conserved thymine and adenine residues are enriched for transversions

Next, we investigated the mutational pattern at the canonical "TNA" motif. Specifically, we counted single nucleotide substitutions of thymine at position 4 (to adenine, cytosine, and guanine), and adenine at position 6 (to cytosine, guanine, and thymine) of the heptamer, and compared it to substitutions of adenine and thymine residues in other genomic features. Within all genomic features investigated, transitions were more frequent than transversions at both adenine and thymine residues. However, the transition to transversion (Ti/Tv) ratio varied in the features studied (Fig. 5). Our variant catalogue contained 656 and 691 SNPs, respectively, affecting thymine and adenine residues at positions 4 and 6 of canonical

heptamers, with a Ti/Tv ratio of 1.54 and 1.55, respectively. The Ti/Tv ratio was lower at these thymine and adenine residues than within the other features tested except the two splice sites. Importantly, their Ti/Tv ratio was significantly (Fisher's exact test; $p<1.0x10^{-04}$) lower compared to other intronic adenine (2.13) and thymine (2.12) nucleotides.

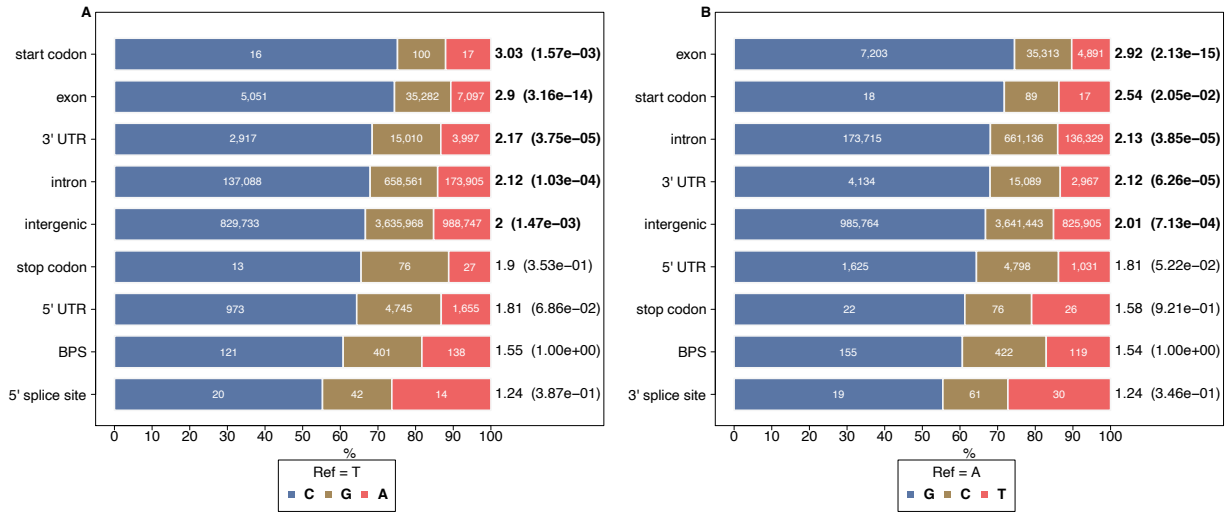

Figure 5. Ti/Tv ratio of different genomic features.

SNP mutation types of thymine at position 4 (a) and adenine at position 6 (b) within canonical ("TNA") branch point sequence motifs and other genomic features. The widths of differently colored bars indicate the proportion of substitution to the other three nucleotides. The number of substitutions observed in our variant catalogue is given in the middle of each differently colored bar. The Ti/Tv ratio for the features is indicated at the end of the bars. Values in the parenthesis are p values (<0.05 indicated in bold text) from pairwise Fisher's exact tests comparing the Ti/Tv ratio of nucleotides within the BPS and other genomic features. For splice sites, analysis was limited to a splice site where the canonical sequence motif (GT and AG at 5' and 3' splice sites respectively) contains the studied residues.

A conserved thymine residue is the least variable site in human branch point sequences

Following the approach introduced earlier, we considered 115,640,370 high quality SNPs segregating in 3,942 samples of diverse ancestry to estimate the variation in various functional features of the human genome. A larger number of sequenced human samples

provided a more comprehensive catalogue of SNPs including many that segregated at very low MAF. Thus, the average genome-wide variation was more than 3-fold higher (4.21 variable sites per 100 bp) in the human than cattle data. The variation within the genomic features relative to the genomic average, however, was largely consistent with the estimates obtained in the cattle data (Supplemental Text S1).

Next, we predicted branch point sequences in 213,158 human introns using the BPP software (Zhang et al. 2017). The predicted branch point was predominantly adenine (98.96%), and it was placed 29.3±11.9 bp upstream of the 3' splice site. We observed a pattern of variability for nucleotides overlapping the heptamer that resembled our earlier findings in the bovine dataset. The branch point and the position 4 of the heptamer harbored 37% and 42% lower variation than the coding sequence, respectively. The position 4 of the heptamer was more strongly depleted for variation than the branch point (2.18 vs 2.39 variants per 100 bp; Fisher's exact test; $p=7x10\times 10^{-06}$, Figure 6). Both positions were depleted for variation more strongly in heptamers with "TNA" motifs (n=195,682; 91.80%). The lower variability of the thymine residue at position 4 compared to the branch point adenine was statistically more significant (1.86 vs 2.19 variants per 100 bp; Fisher's exact test; $p=3.9\times10^{-13}$) at the "TNA" than other motifs.

The Ti/Tv ratio was lower for the adenine (1.32) and thymine (1.46) residues within heptamers with a canonical motif than other genomic features tested except the splice sites, and it was significantly lower (Fisher's exact test; $p <2.0x10^{-17}$) compared to intronic adenine (1.98) and thymine (1.95) residues (Supplemental text S1)

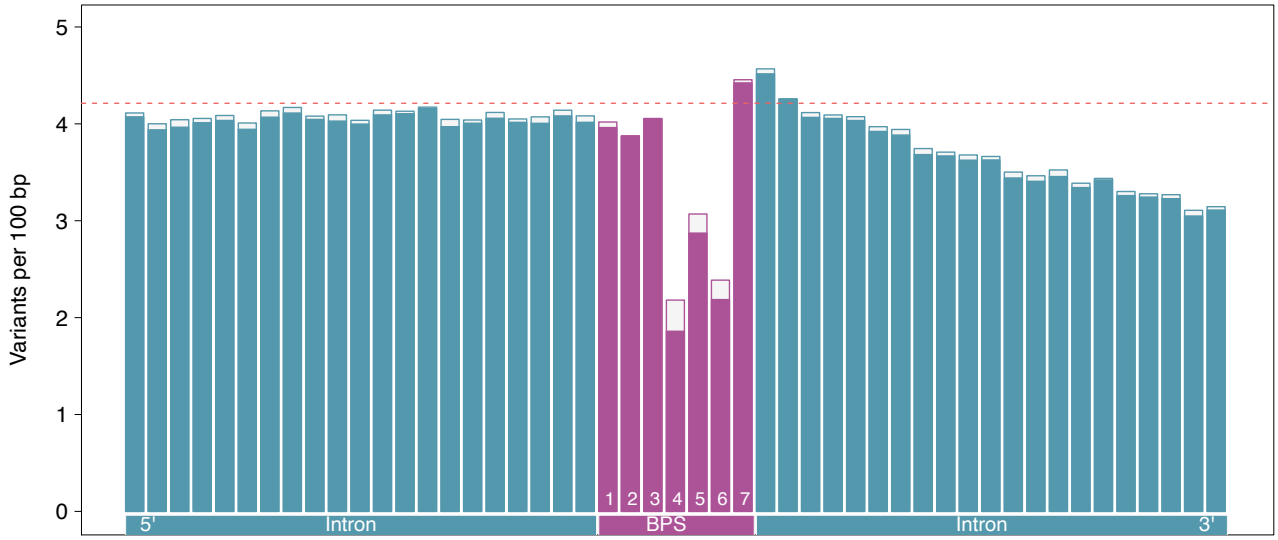

Figure 6. Local variation at the human branch point sequence (BPS).
The height of the bars indicates the variation (variants per 100 bp) surrounding 213,158 branch points while the solid bars indicate the variation surrounding a subset of 195,682 branch points encompassed by a heptamer with a canonical ("TNA") motif. The dotted horizontal red line indicates the average genome-wide variation.

Discussion

We used a catalogue of 29.4 million autosomal variants detected in 266 cattle to investigate evolutionary constraints on coding and non-coding features of the bovine genome. Our results show that the essential splice sites are less variable than any other feature investigated, suggesting extreme purifying selection. The strong constraint on splice regions of the bovine genome mirrors findings in human genomes (Telenti et al. 2016; Lord et al. 2019). While the splice acceptor and splice donor sites are accessible from gene annotation files, other motifs that are required for spliceosome assembly are barely annotated and thus poorly characterized. This motivated us to scrutinize the branch point sequence, a cis-acting intronic feature involved in pre-mRNA splicing. To the best of our knowledge, branch point sequences have not been predicted in bovine introns so far. Thus, our study is the first to

uncover a hitherto neglected feature of the bovine genome and study its evolutionary constraints at nucleotide resolution.

We considered up to 250 bp upstream of the 3' splice site to predict branch point sequences. A vast majority of the bovine branch points are between 18 and 37 bp upstream of the 3' splice-site, corroborating previously reported constraints on the placement of branch points relative to the 3' splice site (Green 1986; Nelson and Green 1989; Signal et al. 2018). While all four nucleotides may act as branch points in principle, a preference towards adenine and cytosine over uracil and guanine was observed in early mutational experiments (Hornig et al., 1986). We observed a very strong preference towards adenine at both bovine and human branch points. The preference towards adenine is stronger for the branch points detected in our study than for experimentally identified branch points in human introns (Gao et al. 2008; Mercer et al. 2015). This indicates that the computational prediction of branch point sequences as applied in our study prioritizes heptamers with canonical motifs.

The assembly of the spliceosome requires base pairing between a heptamer encompassing the branch point and the branch point recognition sequence "GUAGUA" of the U2 snRNA. Thus, the complementary sequence TACTA<u>A</u>C (the underlined branch point does not pair with the U2 snRNA, bulges out and acts as a nucleophile) was proposed as the most preferred branch site for mammalian pre-mRNA splicing (Zhuang et al. 1989). Our findings show that the TACTAAC motif rarely (n=384; 0.21%) constitutes the branch point sequence in bovine introns. This is different from yeast, where the TACTAAC motif is deeply conserved (Green 1986; Padgett et al. 1985). The consensus motif of the branch point sequence in bovine introns includes conserved thymine and adenine residues at positions 4 and 6. This motif

resembles branch point sequences detected in other mammalian species (Gao et al. 2008; Taggart et al. 2012; Corvelo et al. 2010; Green 1986; Bitton et al. 2014). Moreover, the purine and pyrimidine content of the last five residues is similar for predicted branch point sequences in cattle, other mammals, and yeast. Thus, our findings in 180,892 introns suggest that the branch point in bovine introns is contained within a degenerate heptamer with a consensus sequence of "nnyTrAy."

Counting the number of variable sites within functional classes of annotations is a straightforward approach to study evolutionary constraints on genomic features (Telenti et al. 2016; Neininger et al. 2019; Castle 2011). Using a variant catalogue established from the whole genome sequencing of 266 cattle, we show that this method uncovers constraints at nucleotide resolution. For instance, we retrieved signals of purifying selection at the two essential splice sites as well as known patterns within coding triplets. The confirmation of well-known constraints suggests that this approach, despite its simplicity, reveals biologically valid signatures. The overall variability of the tested features is sensitive to the size of the variant catalogue, which in turn depends on the number of samples sequenced (Telenti et al. 2016). A larger variant catalogue in our human than cattle dataset yielded higher variation within all features studied. However, feature variability relative to the genomic average was similar in cattle and human, suggesting that constraints are comparable in both species. It is however worth noting that the large catalogue of variants in the human dataset enabled us to better distinguish variability in individual residues.

The variability at positions 4 and 6 of the predicted branch point sequences was strikingly lower than at surrounding intronic nucleotides and at coding sequences including the highly constrained second residue of coding triplets both in the bovine and human dataset. These

two intronic residues are strongly conserved in branch point sequences across vertebrates (Kol et al. 2005; Mercer et al. 2015). The fact that two evolutionarily conserved residues within the heptamer are strongly depleted for variants, suggests that they are under extreme purifying selection. It is likely that these intronic thymine and adenine residues harbor deleterious variants with fatal consequences on pre-mRNA splicing. Variants at these residues that perturb splicing have been associated with diseases in humans (Kuivenhoven et al. 1996; Li and Pritchard 2000; Královičová et al. 2004; Taggart et al. 2012; Cooper et al. 2009). However, to the best of our knowledge, phenotypic consequences of mutated branch point sequences have not been described yet in cattle; this is surprising given the strong constraint we uncovered on them. We suspect that this is partly because disease-causing variants are assumed to predominantly reside in coding sequences while intronic variants that are distant to the splice sites are rarely considered as candidate causal variants (Pausch et al. 2015; Duchesne et al. 2018; Derks et al. 2019; Bourneuf et al. 2017; Fritz et al. 2013; Störk et al. 2019). The genotyping of variants affecting branch point sequences using customized genotyping arrays in large cohorts might enable differentiating between tolerated and deleterious variants affecting this hitherto neglected intronic feature (Charlier et al. 2016).

The thymine residue at the fourth position of the heptamer is conserved across vertebrates more strongly than the branch point adenine itself (Mercer et al. 2015). Our study revealed that the thymine residue of canonical branch point sequences is less variable than the branch point adenine both in cattle and humans. Its intolerance to mutations is likely due to the fact that the conserved thymine residue is the only nucleotide in the heptamer that cannot undergo wobble base pairing with the recognition site of the U2 SnRNA (Mercer et al., 2015).

Moreover, it has a role (along with the branch point) in binding to the conserved pre-mRNA splicing factor SF1 in the early steps of spliceosome assembly (Berglund et al. 1997). In contrast, nucleotides other than adenine may serve as branch points (Hornig et al. 1986). We observed an excess of transversions among the SNPs affecting the thymine and adenine residues within the branch point sequence. Considering that cytosine is the second most preferred branch point residue, it seems plausible that some SNPs substituting the branch point adenine by cytosine do not compromise branch point activity. Disease-associated mutations affecting the two highly conserved residues at human branch point sequences are enriched for transitions (Královičová et al. 2004; Gao et al. 2008; Kralovicova et al. 2006). Thus, the excess transversions observed at the two nucleotides suggest that the variants detected in our cohort of healthy individuals are possibly enriched for tolerated variations.

The computationally predicted branch point sequences of our study likely contain falsely placed branch points. We suspect that true branch points are even stronger depleted for variation than suggested by our study. The fact that we observed less variation in a subset of branch point sequences that were predicted by two methods supports this hypothesis. On the other hand, our study considered only the most likely branch point per intron. Introns with multiple branch points are prevalent in the human genome (Mercer et al. 2015) and their branch points might be more tolerant to mutations due to the availability of alternative branch points. In any case, our findings indicate that systematic assessment of sequence variation affecting branch point sequences is warranted.

Methods

The bovine variant catalogue.

We considered single nucleotide polymorphisms (SNPs) identified in whole-genome sequence data of 266 cattle from 11 breeds to investigate variation in various bovine genomic features. All animals were sequenced with Illumina instruments using short paired-end reads at an average coverage of 10-fold or higher (mean=16.3). Accession numbers for the raw sequencing data of all animals of the present study are listed in supplemental_File S1.txt.

Following removal of adapter sequences using the fastp software package (Chen et al. 2018), high quality reads were aligned to the ARS-UCD1.2 assembly of the bovine genome (Rosen et al. 2020) using the mem-algorithm of the BWA software (Li 2013). We marked duplicate reads using Picard tools (http://broadinstitute.github.io/picard/) and sorted the alignments by coordinates using Sambamba (Tarasov et al. 2015). The base quality scores were recalibrated using the BaseRecalibrator module (Depristo et al. 2011) of the Genome Analysis Toolkit (GATK - version 4.1.4.1) using 110,270,189 unique positions from the bovine dbSNP version 150 as known variants.

Variant discovery and genotyping were carried out on the 29 autosomes following the GATK best practice guidelines (https://gatk.broadinstitute.org/hc/en-us/articles/360035535932-Germline-short-variant-discovery-SNPs-Indels-). We applied the HaplotypeCaller, GenomicsDBImport and GenotypeGVCFs modules from the GATK (version 4.1.4.1 (Poplin et al. 2017)) to discover and genotype SNPs and INDELs. Site-level hard filtering was applied using the VariantFiltration module of GATK and the raw genotypes

were refined, and sporadically missing genotypes were imputed using beagle4.1 (Browning and Browning 2007). The 29,835,670 high quality SNPs (29,124,151 bi-allelic and 711,519 multi-allelic) passing the hard filtering and segregating in 266 samples were subsequently used to study variation in bovine genomic features both at feature level and at nucleotide resolution. The genomic relationship matrix among the sequenced animals was constructed using 20,553,068 genomic variants segregating with a MAF ≥ 0.01 and principal component analysis was performed using GCTA (Yang et al. 2011)

Estimation of variation in functional features

We obtained the coordinates of genomic features from the Ensembl (release 101) annotation of the bovine reference genome (Rosen et al. 2020). We considered 35,848 transcripts from 20,785 protein-coding genes on the 29 autosomes with 204,123 exons and 181,080 introns, as well as 17,759 3' UTR, 25,490 5' UTR, 30,363 stop codons, 29,000 start codons, 172,269 3' splice sites, and 173,749 5' splice sites. Each nucleotide from the bovine reference genome (2,489,368,272 non-missing autosomal nucleotides) was assigned to one of the eight functional features according to the following priority order: stop codon, start codon, 5' splice site, 3' splice site, 3' UTR, 5' UTR, exons, introns. Bases that did not overlap any of the eight features were assigned to the intergenic category. Using the previously established variant catalogue (see above), we assessed the variability of each feature as the number of SNPs per 100 nucleotides assigned to the feature. The allele frequency distribution within each feature was assessed based on the minor allele frequency of all variants overlapping the features.

Similarly, the variation at the nucleotide level was estimated by considering one nucleotide of a feature at a time. For instance, to estimate the variation at the first residue of the 5' splice site, we extracted the corresponding coordinates from n=173,749 5' splice sites. Thus, the variation at the first residue of the 5' splice site was expressed as $100 \times \frac{m}{n}$ (variants per 100 bp), where m was the number of SNPs overlapping these coordinates.

Prediction of branch point sequences

We used a mixture model implemented in the software program BPP (Zhang et al. 2017) to predict branch point sequences in the bovine introns. First, we obtained the weighted octanucleotides from the bovine introns following the approach proposed by Zhang and colleagues (Zhang et al. 2017). Briefly we extracted 14 bp long intronic sequences from the polypyrimidine tract region (3-16 bp upstream of the 3' splice site) and the background region (187-200 bp upstream of the 3' splice site) of 146,992 introns that were longer than 300 bp. The octanucleotide frequency was calculated in both regions separately and scored as proposed by Zhang and colleagues (Zhang et al. 2017).

Next, using these weighted octanucleotides and the position weight matrix (PWM) of predicted human branch point sequences (downloaded from https://github.com/zhqingit/BPP), we predicted the branch point sequence in bovine introns that were at least 20 bp long. We considered 180,292 intronic sequences (up to 250 bp upstream of 3' splice site) from protein-coding transcripts. Within each intron, we retained the branch point sequence with the highest score.

Branch point sequences were also predicted using the LaBranchoR software with default parameters (Paggi and Bejerano 2018). LaBranchoR predicts branch points within 70 bp

upstream of the 3' splice site. We considered 179,093 introns from protein-coding transcripts that were at least 70 bp long. As in the case of BPP only the branch point sequence the highest score was considered for each intron.

Validation with Human data

The coordinates for functional features were obtained from the gene transfer file of the Ensembl annotation (release 101) of the human genome (GRCh38). The human branch point sequences were predicted in 213,158 introns (>20 bp long) using the software program BPP (Zhang et al. 2017). The weighted octa-nucleotides and PWM were downloaded from https://github.com/zhqingit/BPP.

A variant catalogue was established from publicly available human data. Specifically, we downloaded vcf-files from the Human Genome Diversity Project (HGDP) and the 1000 genomes project (1KG) from the genome aggregate database (gnomAD) v3.0 (https://gnomad.broadinstitute.org/blog/2019-10-gnomad-v3-0/). The vcf files contained genotypes for 115,640,370 high quality SNPs (passing the VQSR) segregating in 3,942 samples of diverse ancestry.

Data Access

All raw sequencing data of cattle used in this study are available from the European Nucleotide Archive under accession numbers listed in supplemental_File_S1.txt. The human variant catalogue is available via gnomAD (https://gnomad.broadinstitute.org/blog/2019-10-gnomad-v3-0/).

Competing interest

The authors declare no competing interests.

Acknowledgements

This study was supported by grants from the Swiss National Science Foundation (310030 185229), the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 815668 (BovReg), and the Swiss Federal Office for Agriculture. The funding body was not involved in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript. We thank the Functional Genomics Center Zurich for generating DNA sequencing data.

References

Aguet F, Barbeira AN, Bonazzola R, Jo B, Kasela S, Liang Y, Parsana P, Aguet F, Battle A, Brown A, et al. 2019. The GTEx Consortium atlas of genetic regulatory effects across human tissues The Genotype Tissue Expression Consortium. *bioRxiv*.

Ast G. 2004. How did alternative splicing evolve? *Nat Rev Genet* **5**: 773–782.

Awan AR, Manfredo A, Pleiss JA. 2013. Lariat sequencing in a unicellular yeast identifies regulated alternative splicing of exons that are evolutionarily conserved with humans. *Proc Natl Acad Sci U S A* **110**: 12762–12767.

Berglund JA, Chua K, Abovich N, Reed R, Rosbash M. 1997. The splicing factor BBP interacts specifically with the pre-mRNA branchpoint sequence UACUAAC. *Cell* **89**: 781–787.

Bitton DA, Rallis C, Jeffares DC, Smith GC, Chen YYC, Codlin S, Marguerat S, Bähler J. 2014. LaSSO, a strategy for genome-wide mapping of intronic lariats and branch points using RNA-seq. *Genome Res* **24**: 1169–1179.

Bourneuf E, Otz P, Pausch H, Jagannathan V, Michot P, Grohs C, Piton G, Ammermüller S, Deloche MC, Fritz S, et al. 2017. Rapid Discovery of de Novo Deleterious Mutations in Cattle Enhances the Value of Livestock as Model Species. *Sci Rep* **7**: 1–19.

Browning SR, Browning BL. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* **81**: 1084–1097.

Burge chrisotpher B, TuschI T, Sharp PA. 1993. *Splicing of Precursors to mRNA by the Spliceosome*.

Castle JC. 2011. SNPs occur in regions with less genomic sequence conservation. *PLoS One* **6**.

Charlier C, Li W, Harland C, Littlejohn M, Coppieters W, Creagh F, Davis S, Druet T, Faux P, Guillaume F. 2016. NGS-based reverse genetic screen for common embryonic lethal mutations compromising fertility in livestock. *Genome Res* **26**: 1333–1341.

Chen S, Zhou Y, Chen Y, Gu J. 2018. Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**: i884–i890.

Coolidge CJ, Seely RJ, Patton JG. 1997. Functional analysis of the polypyrimidine tract in pre-mRNA splicing. *Nucleic Acids Res* **25**: 888–895.

Cooper TA, Wan L, Dreyfuss G. 2009. RNA and Disease. *Cell* **136**: 777–793.

Corvelo A, Hallegger M, Smith CWJ, Eyras E. 2010. Genome-wide association between branch point properties and alternative splicing. *PLoS Comput Biol* **6**: 12–15.

Depristo MA, Banks E, Poplin R, Garimella K V., Maguire JR, Hartl C, Philippakis AA, Del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**: 491–501.

Derks MFL, Gjuvsland AB, Bosse M, Lopes MS, Van Son M, Harlizius B, Tan BF, Hamland H, Grindflek E, Groenen MAM, et al. 2019. Loss of function mutations in essential genes cause embryonic lethality in pigs. *PLoS Genet* **15**: 1–22.

Duchesne A, Vaiman A, Frah M, Floriot S, Legoueix-Rodriguez S, Desmazières A, Fritz S, Beauvallet C, Albaric O, Venot E, et al. 2018. Progressive ataxia of Charolais cattle highlights a role of KIF1C in sustainable myelination. *PLoS Genet* **14**: 1–25.

Fritz S, Capitan A, Djari A, Rodriguez SC, Barbat A, Baur A, Grohs C, Weiss B, Boussaha M, Esquerré D, et al. 2013. Detection of Haplotypes Associated with Prenatal Death in Dairy Cattle and Identification of Deleterious Mutations in GART, SHBG and SLC37A2. *PLoS One* **8**: 2–9.

Gao K, Masuda A, Matsuura T, Ohno K. 2008. Human branch point consensus sequence is yUnAy. *Nucleic Acids Res* **36**: 2257–2267.

Green MR. 1986. PRE-mRNA SPLICING. 671–708.

Hornig H, Aebi M, Weissmann C. 1986. Effect of mutations at the lariat branch acceptor site on β-globin pre-mRNA splicing in vitro. *Nature* **324**: 589–591.

Hoskins AA, Moore MJ. 2012. The spliceosome : a flexible , reversible macromolecular machine. *Trends Biochem Sci* **37**: 179–188.

Kol G, Lev-Maor G, Ast G. 2005. Human-mouse comparative analysis reveals that branch-site plasticity contributes to splicing regulation. *Hum Mol Genet* **14**: 1559–1568.

Královičová J, Houngninou-Molango S, Krämer A, Vořechovský I. 2004. Branch site haplotypes that control alternative splicing. *Hum Mol Genet* **13**: 3189–3202.

Kralovicova J, Lei H, Vorechovsky I. 2006. Phenotypic Consequences of Branch Point Substitutions. *Hum Mutat* **0**: 1–6.

Kuivenhoven JA, Weibusch H, Pritchard PH, Funke H, Benne R, Assmann G, Kastelein JJP. 1996. An intronic mutation in a lariat branchpoint sequence is a direct cause of an inherited human disorder (fish-eye disease). *J Clin Invest* **98**: 358–364.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arxive* **00**: 1–3.

Li M, Pritchard PH. 2000. Characterization of the effects of mutations in the putative branchpoint sequence of intron 4 on the splicing within the human lecithin:cholesterol acyltransferase gene. *J Biol Chem* **275**: 18079–18084.

Liu S, Gao Y, Canela-Xandri O, Wang S, Yu Y, Cai W, Li B, Pairo-Castineira E, D'Mellow K, Rawlik K, et al. 2020. A comprehensive catalogue of regulatory variants in the cattle

transcriptome. *bioRxiv*.

Lord J, Gallone G, Short PJ, McRae JF, Ironfield H, Wynn EH, Gerety SS, He L, Kerr B, Johnson DS, et al. 2019. Pathogenicity and selective constraint on variation near splice sites. *Genome Res* **29**: 159–170.

MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, et al. 2012. A systematic survey of loss-of-function variants in human protein-coding genes. *Science (80-)* **335**: 823–828.

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F. 2016. The Ensembl Variant Effect Predictor. *Genome Biol* **17**: 1–14.

Mercer TR, Clark MB, Andersen SB, Brunck ME, Haerty W, Crawford J, Taft RJ, Nielsen LK, Dinger ME, Mattick JS. 2015. Genome-wide discovery of human splicing branchpoints. *Genome Res* **25**: 290–303.

Neininger K, Marschall T, Helms V. 2019. SNP and indel frequencies at transcription start sites and at canonical and alternative translation initiation sites in the human genome. *PLoS One* **14**: 1–21.

Nelson KK, Green MR. 1989. specific RNA-binding activity. 1562–1571.

Padgett RA, Grabowski PJ, Konarska MM, Sharp PA. 1985. Splicing messenger RNA precursors: branch sites and lariat RNAs. *Trends Biochem Sci* **10**: 154–157.

Paggi JM, Bejerano G. 2017. A sequence-based, deep learning model accurately predicts RNA splicing branchpoints. *bioRxiv* 1647–1658.

Paggi JM, Bejerano G. 2018. A sequence-based, deep learning model accurately predicts RNA splicing branchpoints. *RNA* **24**: 1647–1658.

Pausch H, Schwarzenbacher H, Burgstaller J, Flisikowski K, Wurmser C, Jansen S, Jung S,

Schnieke A, Wittek T, Fries R. 2015. Homozygous haplotype deficiency reveals deleterious mutations compromising reproductive and rearing success in cattle. *BMC Genomics* **16**: 1–13.

Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, Kling DE, Gauthier LD, Levy-Moonshine A, Roazen D, et al. 2017. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* 1–22.

Reed R, Maniatis T. 1988. The role of the mammalian branchpoint sequence in pre-mRNA splicing. *Genes Dev* **2**: 1268–1276.

Rosen BD, Bickhart DM, Schnabel RD, Koren S, Elsik CG, Tseng E, Rowan TN, Low WY, Zimin A, Couldrey C, et al. 2020. De novo assembly of the cattle reference genome with single-molecule sequencing. *Gigascience* **9**: 1–9.

Signal B, Gloss BS, Dinger ME, Mercer TR. 2018. Machine learning annotation of human branchpoints. *Bioinformatics* **34**: 920–927.

Störk T, Nessler J, Anderegg L, Hünerfauth E, Schmutz I, Jagannathan V, Kyöstilä K, Lohi H, Baumgärtner W, Tipold A, et al. 2019. TSEN54 missense variant in Standard Schnauzers with leukodystrophy. *PLoS Genet* **15**: 1–21.

Taggart AJ, Desimone AM, Shih JS, Filloux ME, Fairbrother WG. 2012. Large-scale mapping of branchpoints in human pre-mRNA transcripts in vivo. *Nat Struct Mol Biol* **19**: 719–721.

Taggart AJ, Lin CL, Shrestha B, Heintzelman C, Kim S, Fairbrother WG. 2017. Large-scale analysis of branchpoint usage across species and cell lines. *Genome Res* **27**: 639–649.

Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. 2015. Sambamba: Fast processing of NGS alignment formats. *Bioinformatics* **31**: 2032–2034.

Telenti A, Pierce LCT, Biggs WH, Di Iulio J, Wong EHM, Fabani MM, Kirkness EF, Moustafa A, Shah N, Xie C, et al. 2016. Deep sequencing of 10,000 human genomes. *Proc Natl Acad Sci U S A* **113**: 11901–11906.

Will CL, Lührmann R. 2011. Spliceosome structure and function. *Cold Spring Harb Perspect Biol* **3**: 1–2.

Xiang R, Van Den Berg I, MacLeod IM, Hayes BJ, Prowse-Wilkins CP, Wang M, Bolormaa S, Liu Z, Rochfort SJ, Reich CM, et al. 2019. Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits. *Proc Natl Acad Sci U S A* **116**: 19398–19408.

Yang J, Lee SH, Goddard ME, Visscher PM. 2011. GCTA: A tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**: 76–82.

Zhang Q, Fan X, Wang Y, Sun MA, Shao J, Guo D. 2017. BPP: A sequence-based algorithm for branch point prediction. *Bioinformatics* **33**: 3166–3172.

Zhuang Y, Goldstein AM, Weiner AM. 1989. UACUAAC is the preferred branch site for mammalian mRNA splicing. *Proc Natl Acad Sci U S A* **86**: 2752–2756.