Bioinformatic Approaches to Aligning Non-Canonical Splice-sites Using RNA-Seq Data

Annie Ehler
BIO 334 - Smith College

12-20-2020

Introduction:

Non-canonical splicing is a process in which exons and their subsequent intron junctions are manipulated to produce an alternative transcript. Most commonly, this occurs outside the canonical - or normal - regulatory mechanisms. However, the canonical mechanism of splicing does not differ far from alternative mechanisms, as cis-regulatory elements of the exon-intron junction are exploited by RNA binding proteins (RBPs) and small-nuclear ribonucleoproteins (snRNPs) in the spliceosome to remove the intron from the pre-mRNA. These cis-regulatory elements are abundant in both the genome and transcriptome as they are the primary binding sites for regulatory elements, which are, most commonly, transcription factors or RBPs in the spliceosome. (Wittkopp & Kalay, 2012) In the case of splicing, these cis-regulatory elements are found centrally to the introns and are bound on either side, 5' or 3', by exon-intron splice-site motifs. These splice-site motifs exploit the numerous free nucleotide-phosphate complexes (i.e., ATP, GTP, or AMP, all employed equally) in the nucleoplasm to undergo a transesterification reaction onto a 3' uracil in the case of canonical splicing, but could be replaced by any pyrimidine and still result in a biochemically viable splice-site motif. (Nelson & Cox, 2017) The exact process by which the cis-regulatory elements, trans-acting factors, and 5' splice-site (5'SS) and 3' splice-site (3'SS) motifs are exploited during canonical splicing are classified into three primary splicing mechanisms; group 1 self-splicing, group 2 self-splicing, and group 3 spliceosome-dependent splicing. The most common splice-site motifs based on their genetic (not transcriptomic) sequence are GT-AG, GC-AG, and AT-AC, with relative abundances of 98%, 0.5%, and 0.1% in the human genome respectively. These are given as their conserved 5'SS and 3'SS motifs sequenced in the exons post-splicing, and are referred to as the primary canonical splice-site motifs. (Dimon et al., 2010; Hu et al., 2012) The majority of alternative transcripts are the result of non-canonical trans-acting factors, which target the pre-mRNA by means of the same cis-regulatory elements. (Sibley et al., 2016) In most cases, these trans-acting factors interact with the cis-regulatory elements in order to exploit canonical splice site motifs outside of canonical exon-intron junctions resulting in non-canonical splicing. The best documented trans-acting factor responsible for non-canonical splicing is the U12-type spliceosome which takes advantage of both canonical and non-canonical 5'SS and 3'SS splice junction motifs. (Sibley et al., 2016; Turunen et al., 2013) In some cases of non-canonical splicing, canonical splice-site motifs are found expressed in repetition resulting in a prime site for alternative splicing along canonical splicing mechanisms by the U2-type spliceosome. (Dimon et al., 2010) For example, genes that are subject to common alternative splicing, such as those highly expressed in human CD4 immune cells, have been observed with repetitious 3'SS sequences of YAGYAG, with Y being any pyrimidine nucleotide. (Dimon et al., 2010) This exact motif is arguably the most common motif observed along all repetitious alternative splice-sites, as the triplet sequence conserves the open reading frame (ORF) while also conserving the canonical 3'SS motif; YAG. (Dimon et al., 2010) Additionally, non-canonical splicing can take advantage of cryptic splice-sites hidden within canonical introns, near or far from canonical exon splice-sites, which can be regulated by the aforementioned trans-acting factors. (Sibley et al., 2016; Turunen et al., 2013) Lastly, non-canonical splice-sites can take advantage of splicing mechanisms outside of the *well-defined rules of splicing*, even outside the mechanisms of the U12 non-canonical spliceosome. (Sibley et al., 2016) This is best observed in multiple species in the response regulatory mechanisms of the unfolded protein response (UPR), where the transcription factor *Xbp1* is non-canonically spliced in the cytosol by *Ire1α*, an endoribonuclease RBP, which removes a short 26nt fragment from the mature mRNA, causing an uneven shift in the ORF and triggers a differentiated transcriptional regulatory response on the behalf of *Xbp1*. The cytosolic nature of this splicing, as well as the mechanism of repurposing an endoribonuclease as the primary RBP for the splicing rather than employing the spliceosome, results in this event being unclassifiable as either U2 or U12 dependent splicing and unarguably non-canonical. (Bai et al., 2014)

Types of Non-Canonical and Alternative Splicing:

There are many mRNA products diverging from the standard mature mRNA transcript that are due to non-canonical splicing. These isoforms have been classified into six main categories; cryptic exons, microexons, recursive splicing, circular RNAs (circRNAs), retained introns, and exonic introns or exitrons. (Sibley et al., 2016) Cryptic exons are similar to those discussed in the human CD4 cells, as they are the result of cryptic splice-sites, or novel splice-sites, resulting in exonization outside of the annotated gene models. Cryptic exons are often the result of transposable elements (TEs) causing a mutation during DNA replication which, in very few cases, leads to downstream exonization or, more commonly, cryptic exonization. However, the majority of cryptic exons - whether they're as a result of repetitious splice-site motifs or exonization - are caught by cellular quality control mechanisms, encode premature stop codons resulting in nonsense-mediated decay (NMD), or are near repressive sequences within the DNA, decreasing the resulting expression levels of the cryptic exons. However, some cryptic exons are selectively regulated by splicing factors as a result of complex auto- or cross-regulatory mechanisms which remain to be fully understood. (Sibley et al., 2016) More commonly found non-canonical exon isoforms are microexons, which are unusually short exons usually flanked by polypyrimidine intron tracts dividing the microexons in succession in the genome. This often results in an exploitable cis-regulatory element, making them a prime site for non-canonical splicing by the U12-type spliceosome. (Sibley et al., 2016) Hundreds of new microexons have been annotated in recent years - with over 60% of them being expressed in neuronal tissue - facilitating the differentiation and enrichment of interaction domains by maintaining the ORF of binding domain proteins – especially in the synapse and axon during neurogenesis. (Sibley et al., 2016) Retained introns are common in alternatively spliced transcripts as a byproduct of the mechanism in which the alternative splicing takes place. Most retained introns are caused by the downregulation of the spliceosome, short flanking introns that are not fully spliced, introns with a high GC content, or rather as a by-product of the depletion of exon junction complexes. (Sibley et al., 2016) Importantly, all causalities alter the interaction between the pre-mRNA and the spliceosome resulting in the retention of the introns in the nucleus. Therefore, most retained introns are caught by cellular quality control mechanisms. In some cases, RBPs selectively retain introns during splicing, which are the center of a wide range of research, as those introns usually contain the pattern of a canonical exon, such as maintaining the ORF. These differ from cryptic exons, as the full intron is retained, rather than just a portion as defined by a cryptic splice-site. (Sibley et al., 2016) Exitrons fall similarly to cryptic exons, as they are alternatively the product of the intronization of a portion of an exon, but similarly, exitrons fall along cryptic splice-sites. Unlike most alternative transcripts, exitrons often retain the ORF size constraints as a multiple of three but are still subject to NMD as they commonly cause frameshifts in the ORF. (Sibley et al., 2016) Exitrons, much like microexons, lead to protein variants, however, exitrons are more highly expressed in disordered regions of proteins rather than interaction sites. (Sibley et al., 2016) Furthermore, circRNAs and chimeric RNAs are caused by more advanced mechanisms of alternative splicing. CircRNAs are caused by head-to-tail splicing, where a cytosine-rich branch point motif is spliced to a splicing donor, resulting in the cyclization of the pre-mRNA. Alternatively, chimeric RNAs are the byproduct of the exons of different genes being spliced together. Proximally located genes are cis-spliced, whereas distally located – often very distant – genes are trans-spliced. CircRNAs are found at high relative expression levels in neuronal tissue where they have been experimentally observed sequestering miRNAs as apart of transcriptional regulatory networks. (Sibley et al., 2016) Finally, recursively spliced exons, also known as zero-length exons, are defined as *a sequence that combines the 3' and 5' splice-site consensus motifs. This allows an intron to be spliced in multiple consecutive steps. (Sibley et al., 2016)* First by using the 3' splice-site on the downstream intron in order to reconstitute the 5' splice-site which then splices the remaining intron. In line with other alternative splicing, recursive splicing has been documented in the brain, specifically on extremely long introns. Interestingly, steric blocking of multiple recursive splice-sites in two genes highly expressed in neuronal tissue failed to stop recursive splicing, alluding to an alternative mechanism other than U2-type, U12-type, or any other observed RBP mediated splicing mechanisms. (Sibley et al., 2016)

Non-Canonical Splicing in Clinical Studies:

As noted above, non-canonical splicing plays a large role in both neuronal gene differentiation and in complex transcription regulation. Non-canonical splicing has also been found to play a role in the regulation of intracellular stress responses, such as in the case of the non-canonical splicing of *Xbp1* as a response to the UPR. From this alone, it is clear that non-canonical splicing is integral to understanding the larger mechanisms of alternative transcription that lead to the many protein isoforms found both within cells and, more broadly, within tissues. Furthermore, non-canonical splicing is a leading cause of genetic disease, as over a third of all genetic pathologies are due to an irregularity either in the cis-regulatory elements of pre-mRNA or in the trans-acting factors that interact with the transcriptome. (Sibley et al., 2016) It was originally presumed that these non-canonical splicing events that led to disease pathogenesis were caused by alternative splice-sites proximal to canonical splice-sites resulting in cryptic splice-sites that could possibly be regulated with known genetic quality control mechanisms. However, it has been found that many diseases are caused by non-canonical splice-sites beyond the proximal bases surrounding known splice-site junctions. Notably, these distal non-canonical splice-sites lead to the pathogenesis of spinal muscular atrophy, colorectal cancer, and autism spectrum disorder. (Sibley et al., 2016) Additionally, for over half of diseases that are caused by cryptic splice-sites proximal to canonical splice-sites, the splice-site is the byproduct of non-canonical TEs in the transcriptome. These non-canonical TEs are created either through accruing single-nucleotide polymorphisms (SNPs) resulting, in most cases, in an *Alu* sequence, or, in rarer cases, through the insertion of an existing TE during either transcription or DNA replication. (Sibley et al., 2016) Alternatively, tumorigenesis during cancer pathogenesis is often caused by intron retention. The mechanism for such has been found to be vast. However, most commonly the intron retention is caused by competitive splicing of pre-mRNAs or due to repetitive somatic SNPs resulting in the mutation of one or both of the splice-sites (3'SS and 5'SS) surrounding an intron. (Sibley et al., 2016) Cancers have also been found to contain multiple chimeric transcripts, either through chromosomal rearrangement or deletion breakpoints. (Sibley et al., 2016) In the case of breast cancer, there are multiple instances of exitrons as a result of non-canonical splicing. (Sibley et al., 2016) From a therapeutic lens, non-canonical splicing mechanisms could be exploited to modify splicing events or modify splicing factors. Especially since splice factor modification has been seen as a potent tool against MYC-driven tumorigenesis. (Sibley et al., 2016) As for specific splicing events, therapies are currently being explored employing antisense oligonucleotides, uridine-containing small-nuclear RNAs, and CRISPR-Cas9 as methods for targeting cis-regulatory elements of the splice-site. (Sibley et al., 2016) CircRNAs are also being explored for their therapeutic uses, but their physiological function still largely remains a mystery. Preliminary studies have been done to employ circRNAs as *aptamers, trans-cleaving ribozymes, small interfering RNAs (siRNAs), or as sponges to sequester micro RNAs (miRNAs) or RBPs. (Sibley et al., 2016)*

Non-Canonical Splicing in Evolutionary Studies:

From a larger genetic evolutionary point of view, non-canonical splicing was first seen as nothing more than noise in the background, especially since the vast majority of the resulting alternative transcripts are caught by cellular quality control mechanisms. However, through the experimental discovery of the interactions between snRNPs and alternative transcripts as well as the role that transposable elements (TEs) play in both non-canonical splicing and in larger genetic mutations, non-canonical splicing has become a major mechanism for possible speciation. (Sibley et al., 2016) In the case of TEs in the transcriptome, it has been found that the repression of certain genes, specifically of genes encoding RBPs, gives rise to exonification through emerging *Alu* sequence TEs, which then degrade as TEs post-exonification by the upregulation of the formerly repressed RBPs. (Sibley et al., 2016) From this, multilevel selection theory supports the idea that both species- and clade-specific TEs may lead to speciation and, most notably, prevent extinction due to environmental stress. Following the role of TEs in speciation, it is hypothesized that cryptic splicing as a result of TEs may lead to species-specific splice-sites and novel gene regulatory networks. (Sibley et al., 2016) These theories lead to the exploration of neuronal evolution, especially as non-canonical splicing is upregulated in the brain, as the theories of TE guided evolution may have keen insights toward the evolution of the brain. (Sibley et al., 2016)

RNA-Seq Next-Generation Sequencing and Computation:

Such in-depth studies of non-canonical splice-sites would not have been possible without the capabilities of high-throughput sequencing of the transcriptome such as those available through RNA-Seq. Prior to RNA-Seq being widely available, computational biologists employed algorithms that compared the proteome to the genome, which then highlighted points of interest for novel transcriptomic regulation which then could be explored through short-sequence expression analysis methods such as RT-PCR. (Bai et al., 2014) Nowadays, RNA-Seq alongside modern computational approaches are plenty for identifying novel transcriptome variations, and most importantly, non-canonical transcripts and their subsequent splice-sites. (Sibley et al., 2016) In order to understand exactly how the computational approaches work, there must be an accurate intuitive context as to how RNA-Seq works. RNA-Seq works by sequencing individual bases within an experimentally derived transcriptome, which is usually extracted mRNA from a cell. This next-generation sequencing method works by sequencing smaller reads multiple times in order to form a relative snapshot of RNA expression levels as well as their sequence. (Wang et al., 2009) RNA-Seq next-generation sequencing protocols are relatively ubiquitous from one another but can be fine-tuned for specific experiments by the read length, the depth of coverage, and whether each read is a single-end read or a paired-end read. The read length defines the number of nucleobases sequenced at a time, usually somewhere around 50-100 base pairs or longer. The depth of coverage is how many times each genomic site is sequenced, which is multiplied by the read length to get the reads-per-sample number. Whether the reads are single-end or paired-end reads is dependent on the necessary accuracy of the sequencing. For single-end reads, each read length is sequenced from one side to the other. For paired-end reads, each read length is sequenced from one side to the other and then back again. (*Genome Sequencing*, 2019) However, RNA-Seq should not be thought of as a method for 'reading' the transcripts as one might a book, left to right, top to bottom, but rather a way of randomly grabbing a short read from a pool of nucleotides. This results in an output that accurately represents the expression levels of each read within the active transcriptome of the cell in which it was originally derived. (Wang et al., 2009) Given the inexact nature of sequencing - considering it is not evenly sequenced from one side to the other - poses a large computational question of how exactly to quickly and effectively identify the complete transcriptome from start to finish, let alone small perturbations in it. Existing algorithms to align and map RNA-Seq reads to a reference genome, giving an output of each annotated gene's expression level based on the frequency of its sequence in the RNA-Seq output. Common algorithms, such as those employed by Bowtie, TopHat, EXONERATE, and the Unix '*grep*' command have been tried and tested repeatedly, and consistently come through as trustworthy methods for aligning the reads to the genome and calculating the expression levels of the individual genes. (Bai et al., 2014) Bowtie works the most simplistically by directly aligning RNA-Seq reads to a reference genome through careful consideration of surrounding reads in-order to find the alignment with the highest fit to the reference genome. (Langmead et al., 2009) Bowtie then outputs the expression levels to each aligned gene. (Langmead et al., 2009) While TopHat utilizes Bowtie, its method of aligning reads involves an important step in its unique heuristic approach. TopHat utilizes Bowtie to get a primary alignment of the RNA-Seq reads to the reference genome, but then TopHat clusters areas of high alignment as consensus sequences, creating consensus "islands." TopHat then enumerates the sequences between the consensus islands in order to identify precise splice junctions with a bias towards canonical splice junctions. The output is then the individually aligned bases from the RNA-Seq data as well as the gene expression data, the same as what is offered in Bowtie. (Trapnell et al., 2009) EXONERATE works similarly to Bowtie by directly aligning the RNA-Seq reads to the reference genome using a selectable heuristic which then selects for the most statistically significant alignment without any additional steps. (Slater & Birney, 2005) Lastly, and most simply, is the Unix *grep* command, a terminal function built into any

Unix-based operating system. The command, *grep*, uses a regular expression algorithm to search a large body of text. In this case, the *grep* command searches the reference genome for each individual read, giving an output of aligned RNA-Seq reads. (Bai et al., 2014) Further steps must be taken to calculate expression levels for individual genes. However, the *grep* command offers precise mapping of reads with very low computational power.

On Alignment Algorithms:

All of these existing algorithms can identify small perturbations in the RNA-Seq transcriptome with a large enough reference genome and small enough seeding parameter, but this gives rise to two main issues. (Sibley et al., 2016) The first being that of the small seeding size, which is the length of bases that are primarily aligned in order to anchor an RNA-Seq read to the reference genome. Within these algorithms, a decrease in the seeding size greatly increases the number of false positives in their alignments. (Bai et al., 2014) The second problem being that many organisms do not have well-annotated transcriptomes, which can be mitigated by creating custom alignment files through, in most cases, combining genome reference files with computationally created splice-junction files. (Bai et al., 2014) However, neither of these problems nor the methods used for mending them can be used for quickly and accurately identifying the slight perturbations from the canonical transcriptome possible through alternative and non-canonical splicing. (Dimon et al., 2010) This is the reason for exploring more complex alignment algorithms. Although there are some methods used to identify some alternative transcripts, such as circRNAs, chimeric transcripts, retained introns, and exitrons. Alongside mapped sequences and gene expression data, the alignment algorithms can also be used to identify unalignable reads. These reads can be split in two and if the downstream portion of the read aligns to an upstream position on the reference genome (or vice versa), then that read is a possible candidate for being either circRNAs or a chimeric transcript. Additionally, the alignments can be scored for their splicing efficiency by calculating the ratio of intron retention as the ratio of exon-intron junction reads to the reads that span the junction or as the ratio of reads that cover the intron to the reads that cover flanking exons. (Sibley et al., 2016) This gives an overall picture of the possibility of alternative transcripts and can be compared to other cells of the same type to determine whether the ratio is abnormal. Whether the intron retention ratio is higher or lower than a baseline derived from other cells of the same type is an indicator of retained introns or an abundance of exitrons.

Alternative Algorithms:

Read-Split-Walk:

Other algorithms set out to directly identify alternative transcripts, and ideally the canonical transcript they were identified from. This allows for the precise definition of the non-canonical splice-site. One such algorithm does that through a multistep process of aligning reads to the reference genome and then splitting the unmapped reads at multiple locations to identify non-canonical splice-sites. Accordingly, this algorithm has earned the name the Read Split Walk (RSW) algorithm. (Bai et al., 2014) This algorithm starts by aligning the reads to the reference genome using Bowtie and taking the unmapped reads as the output. The unmapped reads are then trimmed by two base pairs on each end and then split into two read-half pairs, creating a library of split read-half pairs. The read-half pairs are then aligned repeatedly at different splitting points using Bowtie, during which the algorithm tests different intron lengths between the read-half pairs. The aligned read-half pairs are then cross-compared and consolidated in order to define both the splice region and splice length. This is only in the case that the read-half pairs are within the intron length boundaries defined in the user parameters to the algorithm. If the read-half pairs are within the parameters, the splice region is then defined as the smallest region between two read-half pairs of the same read, and the splice length was defined based on the difference between the genomic coordinates of the ends of the read-half pairs. The algorithm then resolves read-split halves with multiple alignments to the maps of the now aligned reads, with a bias for smaller introns. Before the alignment is finalized, splice-site junctions are analyzed for their proximity to canonical splice-sites. All reads aligned within five base-pairs of a canonical splice-site are finalized, whereas the reads outside of canonical splice-site boundaries are further analyzed. Although, the number of base-pairs between the aligned splice-site versus the canonical splice-site is modifiable by the user. For a non-canonical splice-site to be finalized in the alignment, it must have at least two independent consensus reads by default to correct any false-positives. However, the number of consensus sequences necessary to finalize the alignment is able to be modified by the user. (Bai et al., 2014) Upon analysis of the default parameters, it was found that all aligned canonical splice junctions fell within 5 base pairs of the reference canonical splice junctions. Additionally, it was found that - dependent on the dataset - alignments to non-canonical splice-sites with only two to four independent consensus alignments were most likely false-positives. (Bai et al., 2014) When compared to Bowtie2, TopHat, EXONERATE, and the Unix *grep* command, the RSW algorithm either performed equally as well or better than the other algorithms at identifying non-canonical splice sites in mouse embryonic fibroblasts. This included a high number of consensus regions between the RSW algorithm's identified splice-sites and the annotated splice-sites in ENSEMBLE's proteomics reference database for human breast cancer cells. (Bai et al., 2014)

Omicsoft Sequence Aligner:

Alternatively, algorithms can utilize a reference transcriptome to get better reference coverage of their reads. This is exact methodology is employed by a commercially distributed alignment algorithm known as Omicsoft Sequence Aligner (OSA), aptly named after its commercial distributor, Omicsoft. Due to the fact that OSA utilizes a reference transcriptome as well as a reference genome, it was compared against other algorithms that do the same, namely RUM, RNASeqR, and TopHat as of late 2011. The first step in OSA's algorithm is to align the RNA-Seq reads to the reference transcriptome, which are then translated across to the reference genome. (Hu et al., 2012) This step increases speed and allows for better consideration of the intron alignment but introduces a bias for canonical intron bounds. However, the bias is slightly recovered by the fact that the alignment is run as a bitwise operation and that both alignment steps, to the transcriptome and genome respectively, take into consideration the 3'-end read quality scores from the RNA-Seq data. The second step of the algorithm considers the unaligned reads, which are then aligned directly to the reference genome based on the middle indel detection algorithm as the primary method for correcting the intron boundaries of unaligned reads. (Hu et al., 2012) The middle indel detection algorithm views introns as deletion mutations, as introns are 'deleted' from the transcriptome as pre-mRNA is processed from the genome into mature mRNA. By utilizing the middle indel detection algorithm, OSA is further able to recover some of the lost coverage in the double-backed alignment of the first step by considering any non-canonical transcripts as insertion or deletion mutations. This, however, introduces a bias for alternative transcripts that don't have multiple viable alignments, as they would have been processed in the first part of the algorithm, and a bias for reads that include motifs that can be interpreted as insertion or deletion mutations. Furthermore, OSA has a pre-programmed bias for canonical splice-site motifs, but does include some non-canonical, specifically U12-type, motifs; these motifs are GT-AG, GC-AG, and AT-AC. (Hu et al., 2012) This highlights a broader argument in the alignment of non-canonical transcripts, as the vast majority of non-canonical splice-site motifs do use one of the three canonical motifs, bringing into question what exactly is defined as a canonical versus non-canonical splice-site motif. However, not all non-canonical splice-site motifs abide by the commonality, most notably, the splice motif for the *Xpb1* transcription factor at its alternative splice-site is CA-AG. By simply changing the bias parameter from the three canonical motifs to a motif containing a pyrimidine nucleobase on one end and an arginine nucleobase followed by a guanine or cytosine nucleobase, OSA can include virtually all known splice motifs as that motif covers all biochemically possible group 1-type and group 2-type self-splicing motifs. (Nelson & Cox, 2017) The final step in the OSA algorithm is to simply condense steps one and two into a single alignment, which is done by considering paired-end read quality scores for each base as a factor in concatenating the separate alignments into one. (Hu et al., 2012) When compared to the other algorithms, OSA had a lower false-positive rate while being both faster and more effective in its alignment when compared to RUM, RNASeqR, and SOAP Splice on simulated datasets. Additionally, OSA was able to map more reads and identify more novel junctions when tested against TopHat (late 2011 version) on NGS datasets of real RNA-Seq data.

HMMSplicer:

Newer algorithms are beginning to take heed of machine learning in order to inform their alignments, and such is the case for HMMSplicer. HMMSplicer takes advantage of an unsupervised Hidden Markov Model (HMM) to map and align the RNA-Seq reads independently of any biological factors. While the algorithm still uses a reference genome, it is mostly used to seed the beginning of the reads in order to form a basis for its alignment. Although, the first step of the algorithm is to align any full-length reads to the reference genome using Bowtie. (Dimon et al., 2010) Notably, the algorithm only aligns full-length reads, decreasing the bias towards the reference sequence to negligible amounts in considering the RNA-Seq reads. The second step is similar to that of the RSW algorithm, in that the reads are split, but in the case of HMMSplicer, the reads are split evenly in half. The first read-half split is then aligned to the reference genome in a seeding step using Bowtie. (Dimon et al., 2010) Some of the transcripts in this step will align both read-split halves if their alignment is obvious, overshadowing the first step of aligning full-length reads with similar consideration of the bias. The third step involves training the HMM, which is done by creating an emission dataset of match/mismatch probabilities that the read will align to the seed sequence which is stored as a string of values. The HMM is then trained using an unsupervised Baum-Welch algorithm which defines expectation maximizations and optimizes the HMM parameters based on the maximizations. Each read is then run through the HMM twice, generating the most probably exon-intron junction, and classifying each now defined first exon into one of five bins (0-4) based on the probability of a correctly matched junction boundary. Defining the exon-intron junction is done by utilizing a two-state HMM. The initial state carefully aligns the read-half split to the genome considering the read quality for each base across all reads. The second state aligns the read-half split to the genome regardless of the quality scores. The junction boundary is then defined as the transition state between the two states. For reads with multiple probable junction boundaries, the shorter exon is chosen for later resolution. (Dimon et al., 2010) By using an unsupervised training algorithm, the HMM is able to define exon-intron junction boundary probabilities based on the RNA-Seq data alone while not basing the probability scores on the reference genome. This effectively eliminates bias towards the reference genome and is geared towards experiments concerning non-canonical splicing and alternative transcripts. Before starting the third step, the algorithm optionally, as so defined by the user, resolves what exons it can with multiple junction boundaries to whichever follows the most canonical motif. Other boundaries, especially those of low match probabilities, may also be resolved to canonical splice-site motifs. However, if this feature of correcting to canonical splice-site motifs is turned off, the reads are processed in the third step and resolved later. The third step utilizes a novel scoring algorithm, which first takes the sum of twice the bit score (akin to the BLAST bit score) of each nucleotide in each read-split half and multiplies it by the probability that the nucleotide call is correct given the positional scoring of paired-end RNA-Seq reads. (Fig. 1) The bit score is calculated as the normalized alignment score with respect to the alignment scoring algorithm itself for each read-half split. The sums for both of the read-split halves are then multiplied together and tested for the most optimal full-length alignment by subtracting half the maximum of the multiplied sum generated by moving both read-split halves relative to one another. (Fig. 2) This gives the product of the

Given:

$g_i = \textit{genome nucleotide at position } i$

$r_i = \textit{reported nucleotide at position } i$

$r'_i = \textit{reference nucleotide at position } i$

$q_i = \textit{quality score at position } i$

Score for one side is calculated as:

$$h = \sum_{i=0}^{L} \begin{matrix} g_i = r_i : P(r_i = r'_i | q_i) * 2 \textit{ bits} \\ g_i \neq r_i : 0 \end{matrix}$$

Figure 1: Dimon et al., 2010

$$s = h_a h_b - F * \max \begin{cases} h_{a'} h_b \\ h_a h_{b'} \end{cases}$$

Figure 2: Dimon et al., 2010

$$m_{even} = \frac{1200}{\left(\frac{l}{2} * 2\ bits\right)^2} \quad m_{odd} = \frac{1200}{\left(\frac{l-1}{2} * 2\ bits\right)\left(\frac{l+1}{2} * 2\ bits\right)}$$

Figure 3: Dimon et al., 2010

first equation and will remain positive as long as there is not a more probable full-length alignment. The second part of the scoring algorithm normalizes the scores to a range of 0-1200 for representation using the UCSC Genome Browser by calculating a theoretical maximum. This is done by dividing 1200 by the squared product of half the full read length multiplied by twice the bit score. For reads of uneven length, instead of squaring the product of the denominator, it is split into two separate products, one multiplying half of the length plus one to twice the bit score, and one multiplying half of the length minus one to twice the bit score, the products of which are multiplied together. (Fig. 3) The final equation in the scoring algorithm then takes the product of the first equation, giving the probability of a full-length alignment, and the second equation, normalizing the resultant to a theoretical maximum between 0 and 1200. (Fig. 4) (Dimon et al., 2010) For most alignment algorithms, the score is based on the alignment probability given the E-value. By utilizing a bit score for each individual nucleotide and read, HMMSplicer is able to utilize an information-based scoring algorithm which is normally only used with short reads for single proteins, such as the bit scoring algorithm utilized by BLAST. The fourth step in HMMSplicer's algorithm is to resolve any unaligned reads by comparing them to the now aligned reads. At this point, the possibility for shared junctions is considered, and microexons are scored and individually aligned. Any overlapping bases and shared junctions at this point are scored and their score is factored into the score of their previously aligned overlapping reads and bases. (Dimon et al., 2010) This all but completes the alignment, with the final step being to collapse the reads into a single alignment and generate the related expression values, as well as to filter the 5' and 3' end reads for possible non-canonical splice motifs. (Dimon et al., 2010) Of all the algorithms discussed, HMMSplicer performed the best, with three times as many reads mapped as TopHat on a simulated dataset with 1x read coverage and a read length of 70 base-pairs. HMMSplicer was also able to catch non-canonical splice-site motifs outside of the aforementioned canonical motifs, including the *Xbp1* motif, as it was listed in the top 0.5% splice-sites for possible non-canonical splicing. (Dimon et al., 2010)

$$final\ score = s * m$$

Figure 4: Dimon et al., 2010

Conclusions:

Of the three novel algorithms discussed, HMMSplicer showed the least amount of bias, even for the biological factors it was analyzing, the RSW algorithm mapped reads, even at non-canonical splice-sites with non-canonical motifs, with high accuracy and low latency, and the OSA algorithm gave the lowest false-positive rate on the baseline simulated datasets. (Bai et al., 2014; Dimon et al., 2010; Hu et al., 2012) While one thing is for sure, they all outperformed TopHat, to be able to truly cross-compare these algorithms, they must be run on the same dataset using the same reference dataset. Unfortunately, the field of computational genetics has yet to define ubiquitous parameters in which algorithms can be rigorously validated against, but there are still some biological questions that are shared evenly across the field which act as benchmark assessments. For instance, all three algorithms were tested on their recognition of the *Xbp1* splice-site. However, the parameters by which the algorithms were tested on *Xbp1* splice-site recognition was not consistent. There was no continuity in species, dataset, or reference sequence. Additionally, the algorithms mainly focused on identifying non-canonical splice-sites, and not on cryptic splice-sites with low transcriptional coverage. While HMMSplicer was tested on simulated data with low coverage, and on an NGS dataset with uneven coverage, there was no discussion as to whether or not the algorithm was able to identify cryptic splice-sites. While there is not much weight given to the importance of cryptic splice-sites, they are important to differentiate from other non-canonical splice-sites, as they lay outside of the canonical and non-canonical patterns of exon-intron splicing. (Sibley et al., 2016) Whether or not these discontinuities are resolved, these algorithms provide necessary computational functions in identifying novel alternative transcripts, which is paramount in furthering the understanding of how non-canonical and alternative splicing play a role in biological physiology.

Overall, modern computational methods have allowed us to peer deep within the transcriptome, finding keys to unlocking the big picture of our genetics and their regulatory mechanisms. It is clear that understanding these mechanisms is integral to understanding speciation in evolution, disease pathogenesis, and genetic therapeutics. Due to the fast advancements made in RNA-Sequencing technology, it is now possible to not only calculate expression levels as a byproduct of genetic regulation but to zoom in on the precise points at which these regulatory mechanisms are acting. With this data, broad range genomic and transcriptomic experiments can be formed to identify the full range of non-canonical splicing events, whether they are caused by regulatory factors or genetic mutations leading to disease pathogenesis. Through these experiments, not only can the regulatory factors be identified, but their precise mechanisms can also be explored. By further understanding the regulatory mechanisms that act on non-canonical and cryptic splice-sites, the mechanisms by which the resulting alternative transcripts are able to make it past genomic quality control mechanisms can also be explored, and possibly exploited as another possible treatment route for genetic diseases and oncogenesis. Furthermore, the roles of the resulting alternative transcripts can be further explored, both in the context of broad species-specific physiology and in the much more focused tissue-specific physiology. Once there are enough annotated and understood non-canonical splice-sites, they can also be leveraged to answer questions concerning the evolution of species, the process of speciation, and the ability for species clades to physiologically adapt. Most notably, however, this data may shed some light on the evolutionary process of the canonical gene regulatory networks and mechanisms that drive the transcription and translation of the already discovered ordinary genes.

Bibliography

Bai, Y., Hassler, J., Ziyar, A., Li, P., Wright, Z., Menon, R., Omenn, G. S., Cavalcoli, J. D., Kaufman, R. J., & Sartor, M. A. (2014). Novel Bioinformatics Method for Identification of Genome-Wide Non-Canonical Spliced Regions Using RNA-Seq Data. *PLoS ONE*, *9*(7). https://doi.org/10.1371/journal.pone.0100864

Dimon, M. T., Sorber, K., & DeRisi, J. L. (2010). HMMSplicer: A Tool for Efficient and Sensitive Discovery of Known and Novel Splice Junctions in RNA-Seq Data. *PLOS ONE*, *5*(11), e13875. https://doi.org/10.1371/journal.pone.0013875

Genome Sequencing: Defining Your Experiment. (2019). Columbia University Department of Systems Biology: Irving Cancer Research Center. https://systemsbiology.columbia.edu/genome-sequencing-defining-your-experiment

Hu, J., Ge, H., Newman, M., & Liu, K. (2012). OSA: A fast and accurate alignment tool for RNA-Seq. *Bioinformatics*, *28*(14), 1933–1934. https://doi.org/10.1093/bioinformatics/bts294

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3), R25. https://doi.org/10.1186/gb-2009-10-3-r25

Nelson, D. L., & Cox, M. M. (2017). *Lehninger Principles of Biochemistry: Seventh Edition* (North American Edition). W. H. Freeman and Company.

Sibley, C. R., Blazquez, L., & Ule, J. (2016). Lessons from non-canonical splicing. *Nature Reviews. Genetics*, *17*(7), 407–421. https://doi.org/10.1038/nrg.2016.46

Slater, G. S. C., & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, *6*(1), 31. https://doi.org/10.1186/1471-2105-6-31

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics*, *25*(9), 1105–1111. https://doi.org/10.1093/bioinformatics/btp120

Turunen, J. J., Niemelä, E. H., Verma, B., & Frilander, M. J. (2013). The significant other: Splicing by the minor spliceosome. *Wiley Interdisciplinary Reviews. RNA*, *4*(1), 61–76. https://doi.org/10.1002/wrna.1141

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, *10*(1), 57–63. https://doi.org/10.1038/nrg2484

Wittkopp, P. J., & Kalay, G. (2012). Cis -regulatory elements: Molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics*, *13*(1), 59–69. https://doi.org/10.1038/nrg3095