Determining differential expression of splice variants in RNA-Seq

Ward L. Weistra

Master's thesis - September 2012

Plant-Microbe Interactions, Utrecht University

Under supervision of Marcel C. van Verk & Guido van den Ackerveken

Table of contents

Introduction

RNA sequencing is a hot topic. The vision used to be that determining genetic code (DNA) would explain organisms and diseases. However it's clear now that there is a large layer of regulation on top of the genetic code. Measuring which genes are actually transcribed to RNA gives insight to the path to proteins that are the next step down the line. Differences between the rate of transcription between multiple conditions, differential expression, give great insights in the origins and effects of these conditions.

To determine differential expression between different genes, different samples and different experiments quantified data should be normalized using a uniform method. Many protocols and algorithms have been developed over the last years to analyze RNA-seq data, although the field is still in its early days of standardization. A complicating factor in RNA-seq analysis is the existence of genes with multiple splice variants. These variants complicate the task of assigning reads to unique transcripts.

In this thesis we will give an overview of the tools and methods available in the literature for analyzing RNA-seq data including differential expression of splice variants for organisms that have a reference genome.

Measuring differential gene expression

With great advances in genome sequencing more and more full genomes are unraveled. However the genetic code is only part of the story, as the functioning of each cell is determined by the genes that are actually transcribed. To determine which genes are expressed within a sample and in order to quantify RNA levels within the cell can be measured. To achieve this on a genome wide scale many different methods have been developed, starting with the popular microarray.

Microarrays

Microarrays have short DNA or RNA sequences hybridized to wells on a plate. The mRNA that is isolated from a sample is reverse-transcribed to cDNA or cRNA and a fluorescent label is attached. By bringing

these fragments in contact with the complementary sequences spotted on the microarray they bind to their counterparts on the wells. After washing, and therefore removing any sequences that did not hybridize, the abundance of a binding sequence is measured by determining the level of fluorescence per spot. Brighter spots correspond to a higher abundance of a sequence and in this way represent a higher gene expression level.

Microarray experiments are high throughput and relatively inexpensive. Arrays with libraries of exon-spanning sequences even make it possible to determine relative abundance of known splice variants.
However, they have several limitations towards quantification of expression. It relies on previously known genome sequence to construct the array and only these sequences properly represented on the array are detected. Microarrays have a high background level of detection, due to sequences binding at multiple probes and background fluorescence that interferes with quantification of genes with a low expression level. And finally the method used to determine the quantity of highly expressed genes based upon the fluorescent signal makes saturation of the signal possible and makes it hard to compare relative levels between different experiments without complicated normalization (Wang et al., 2009).

Sequence-based approaches

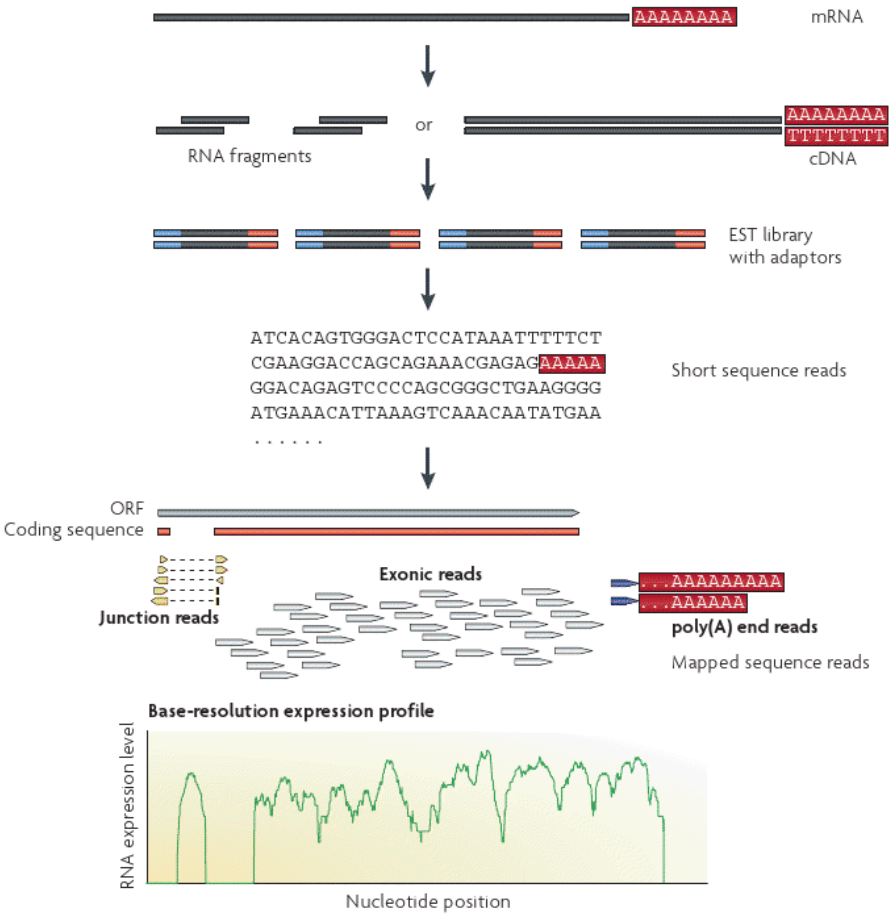
Sequence-based approaches such as serial analysis of gene expression (SAGE), cap analysis of gene expression (CAGE) and massive parallel signature sequencing (MPSS), also referred to as digital gene expression (DGE), isolate and quantify a specific tag within each sequence. In contrast to microarrays these methods directly determine part of the cDNA or cRNA sequence. By mapping these tags back on to the genome the origin can be determined. Quantitative analysis can be achieved by separating and quantifying the tags on an electrophoresis gel and determining the origin of a band by isolating it and multiplying it with Sanger sequencing or in case of massive parallel sequencing, the sequence of each tag is directly determined.
These methods are high throughput and provide precise, digital levels of expression with no saturation.
However, a significant fraction of the short tags cannot be uniquely mapped to the genome and most methods are based on the per base pair relatively expensive Sanger sequencing. Furthermore, since the analyzed tags are only part of the transcript, isoforms cannot be distinguished. (Wang et al., 2009)

RNA sequencing

As a relatively recent advancement RNA sequencing makes transcription analysis on a large scale possible with many advantages. RNA-seq refers to the experimental procedures that generate DNA sequence reads derived from the entire RNA molecule (Garber et al., 2011). It benefits from the introduction of next-generation and high throughput sequencing of DNA.

Before sequencing the mRNA is fragmented and reverse transcribed into cDNA with adaptor sequences ligated to one or both ends (Figure 1; Wang et al., 2009). After sequencing, the reads are mapped back to the reference genome or transcriptome. The reference genome can be the coding sequences or ORF's within a known genome (Genome guided protocol) or a reference genome may be constructed *de novo* (Genome independent protocol). The total amount of reads mapped to a gene is a direct measure of its transcription level.

RNA-seq is, unlike microarrays, not bound to detecting transcripts that correspond to existing genomic sequences. Also, RNA-seq has very low, to none, background signal because DNA sequences can be unambiguously mapped to their unique regions of the genome. It does not have an upper limit in detection which gives it a large dynamic range of expression levels over which transcripts can be detected. RNA-seq has also been shown to be highly accurate for quantifying expression levels and have very high levels of technical reproducibility. Furthermore, as single molecule sequencers like the Helicos technology and iontorrent semiconductor sequencing become available RNA-Seq analysis requires a lower input amount of RNA sample and biased library amplification before sequencing can be omitted (Wang et al., 2009).



▲ ***Figure 1. A typical RNA-experiment (Wang et al., 2009).*** *First the library of full length mRNAs is converted to cDNA fragments. Fragmentation can be done either at the RNA or DNA level. Sequencing adaptors are ligated to the fragments and a short sequence (read) is obtained from them, using high-throughput sequencing. The reads are then mapped back to the reference genome or transcriptome, either as a exonic read or as a junction reads spanning an intron. Based on this a base-resolution expression profile is assembled of which an example is shown at the bottom for a yeast ORF with one intron.*

In general a higher coverage, the amount of reads on average mapping on a single location, will return a higher resolution and will detect more low expressed transcripts. In RNA-seq paired-end sequencing, where also the complementary strand of a read is sequenced, both sides of a read are determined with high confidence. This advantage gives a higher resolution too and this makes it a good choice for complex genomes with a highly repetitive genomic sequence. To convert the sequenced reads into differential gene expression a number of steps need to be taken: e.a. read alignment, normalization and quantification of differences in transcript levels. These methods will be reviewed in the next chapter.

Analysis of RNA-seq data

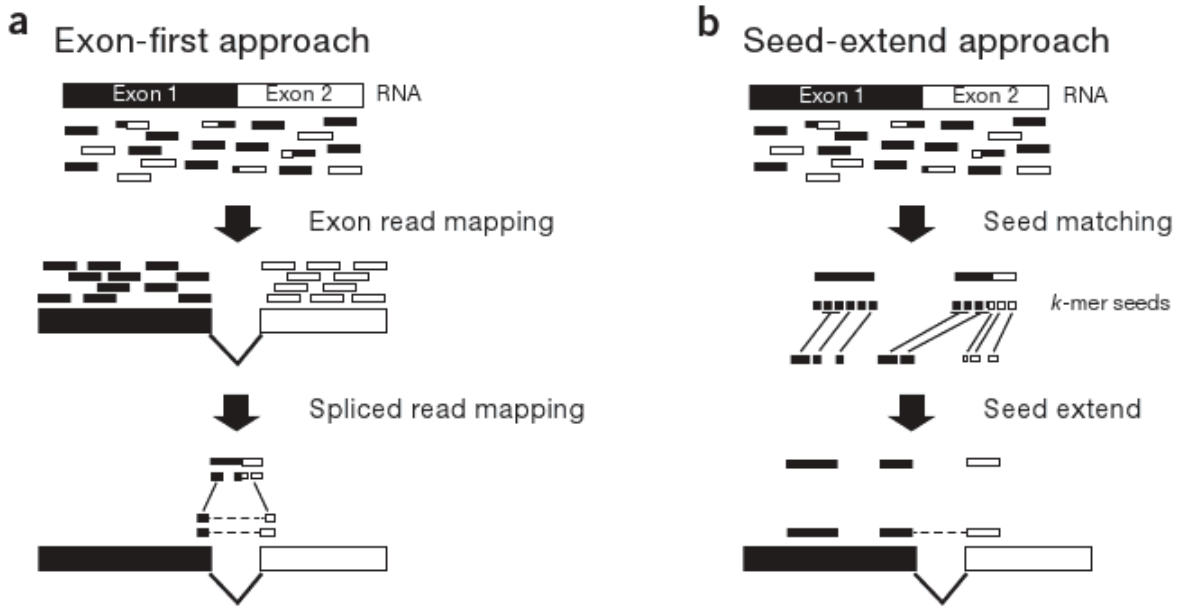
In this chapter the path from raw RNA-seq data, the nucleotide code of the many reads, to differential expression is reviewed. In this analysis four steps can be distinguished: the mapping of the reads to the reference genome, the reconstruction of the transcriptome, the quantification of the expression and the determination of differential expression. Many tools have been developed for each step, with each having its own advantages and limitations.

Read mapping

One of the most basic tasks after sequencing the RNA reads is mapping these reads back to the genome. The reads of RNA-seq pose an extra challenge to the typical bioinformatics problem of aligning sequence reads because of their limited size (~36–125 bases), considerable error rates, reads spanning exon-exon junctions and the large number of reads (Garber et al., 2011).

Two major classes of read alignment algorithms exist: 'unspliced aligners', which map reads directly to the transcriptome and don't allow for large gaps, and 'spliced aligners', which map their reads to the entire genome allowing for large gaps over intron-spanning regions.

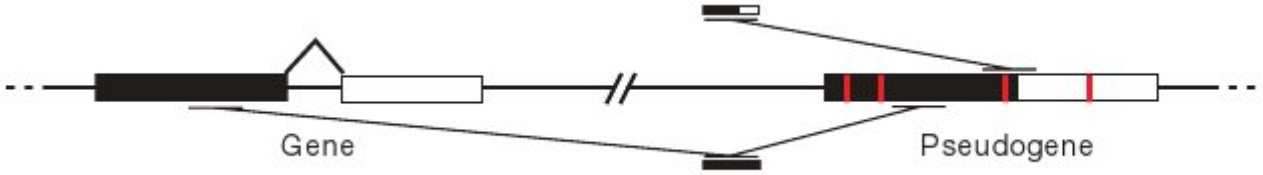
Unspliced aligners use two main strategies. Seed methods map short subsequences, the seeds, to the transcriptome. When a seed matches other more sensitive methods, such as Smith-Waterman alignment (Smith & Waterman, 1981), are used to extend to a full alignment. Examples of seed methods are Efficient Large-Scale Alignment of Nucleotide Databases (ELAND, part of the analysis pipeline bundled by Illumina with its sequencing instruments; Li & Homer, 2010), mapping and assembly with quality (MAQ; Li et al., 2008) and Stampy (Lunter & Goodson, 2011). Burrows-Wheeler transformation methods store the reference in an efficient data structure and search for perfect matches of the whole read on this. These include Burrows-Wheeler alignment (BWA), Bowtie and Short Oligonucleotide Analysis Package 2 (SOAP2, where the first SOAP was a seed method; Li & Homer, 2010). The speed of these methods decreases exponentially with the number of mismatches allowed. Since unspliced read aligners map to the known transcriptome they are limited to detecting known exons and junctions. Splicing events involving novel exons, for example after intron retention, are not identified.



▲ ***Figure 2. The two types of spliced alignment methods (Garber et al., 2011).*** *a. In the exon-first approach reads are first mapped unspliced to the reference genome. Initially unmapped reads are fragmented and these seeds are mapped and extended to find spliced reads. b. The seed-extend approach splices all reads in to seeds, maps these and extends from there.*

Spliced aligners align their reads to the full genome. Seed-extend algorithms (Figure 2a) in this are equivalent to the seed methods above and include 'genomic short-read nucleotide alignment program' (GSNAP; Wu & Nacu, 2010) and 'computing accurate spliced alignments' (QPALMA; De Bona et al., 2008). Their first step is to break reads up in to small seeds and map these to the genome. Next these candidate regions are extended with more sensitive alignment methods. The Exon-first methods (Figure

2b) on the other hand first map full reads with unspliced aligners. For example, exon-first method TopHat invokes Bowtie for this. Next the unmapped reads are broken up into seeds and mapped independently to the full genome. The surroundings of mapped reads are then searched for possible connections between spliced reads. These exon-first methods further include 'RNA-Seq unified mapper' (RUM; Grant et al., 2011), MapSplice and SpliceMap (Garber et al., 2011).
Exon-first methods map reads first to the transcriptome and only splice the remaining reads to align these to the full genome. A risk in this procedure is that retrotransposed pseudogene copies of a gene might exclude an intron. Mapping reads spanning this intron using an exon-first method will give a biased mapping to the pseudogene because here the read will map already unspliced (Figure 3). In contrast seed-extend methods do not impose this bias to mapping unspliced versions first and they outperform exon-first approaches when mapping reads from polymorphic species (Garber et al., 2011).



▲ ***Figure 3 - Potential limitations of exon-first approach (Garber et al., 2011) -*** *Mapping reads using the exon-first approach has potential limitations in the case where a gene with an intron has a retrotransposed pseudogene without this intron. Reads from the gene spanning the splice junctions will map to the pseudogene.*

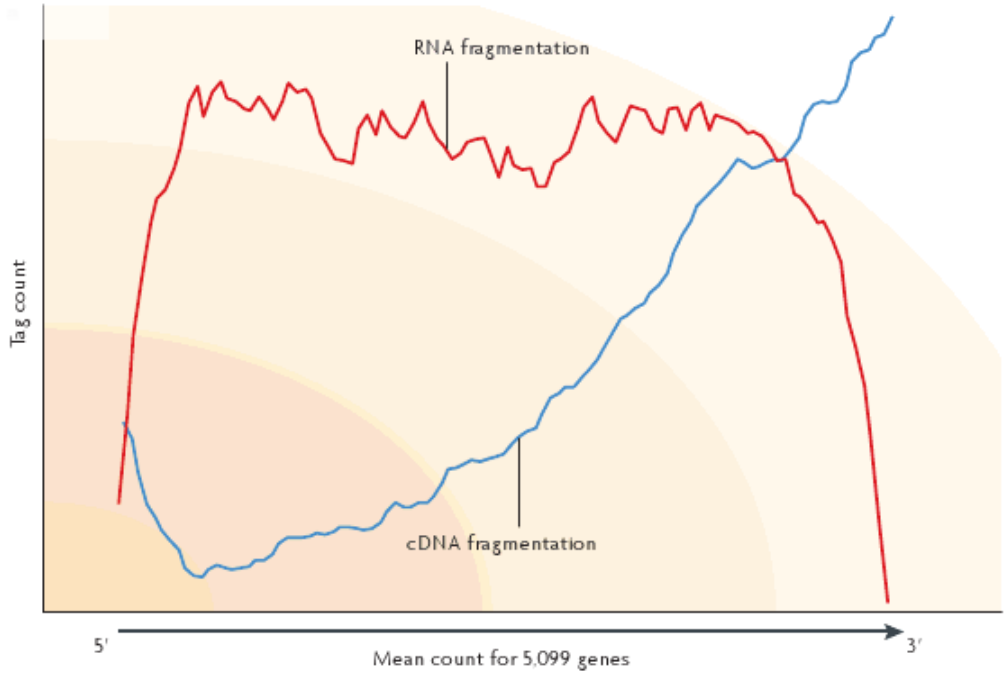
Transcriptome reconstruction

Once the reads are mapped to the reference it is important to determine which transcripts were present within the sample. To achieve this transformation of the read alignments to transcription units is needed, called transcriptome reconstruction (Garber et al., 2011).

If no reference genome or transcriptome is available for the organism at study, it is possible to construct the transcriptome *de novo* with a reconstruction algorithm such as Trans-ABySS (Robertson et al., 2010). For organisms whose genome or transcriptome is available the genome-guided route it the obvious choice.
These genome-guided methods include Cufflinks (Trapnell et al., 2010) and Scripture (Guttman et al., 2010), which both use TopHat as their read mapper. Both methods make a directed graph of bases or exons based on the mapped reads. Each fragment has one node in the graph and an edge is placed between each pair of compatible fragments. This graph is traversed to identify individual transcripts (Haas and Zody, 2010). Scripture reports the full set of all transcripts compatible with this graph, while Cufflinks reports only the minimal set needed to connect all nodes in the graph at least ones (Garber et al., 2011).

There are several factors making reconstruction of the transcriptome a challenging task. There can be several orders of magnitude difference between expression of transcripts, with some only represented by a few reads. Reads are short and genes can have many isoforms, making it hard to detect which isoforms were originally present within the sample (Garber et al., 2011). This issue will be covered in more depth in the next chapters.

The last problem to consider is the unequal distribution of reads along a transcript, called the fragmentation bias (Figure 4). Given the procedure of breaking the RNA molecules in to small reads the chance for a read to cover the edges of a gene is much smaller then further away. Using cDNA fragmentation this bias is strongly to the 3' end of the transcripts (Wang et al., 2009).



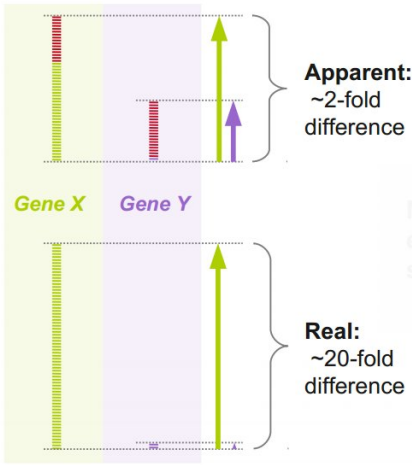
▲ ***Figure 4: The fragmentation bias (Wang et al., 2009).*** *mRNA fragmentation (red) leads to low coverage of the ends of a transcript, cDNA fragmentation leads to a bias of coverage to the 3' end.*

Expression quantification

Now the actual transcripts that were expressed are determined the task is to quantify the expression of each transcript. How do the number of mapped reads translate to the relative number of mRNA molecules present in the sample. Normalization steps should be taken to control for multireads, relative changes in abundance due to changes in overabundant transcripts, differences in the total amount of reads and differences in gene length.

The challenge in quantifying expression is the significant portion of sequence reads that match multiple locations within the genome. One solution is to split the number of reads over all matching positions, however as depicted in Figure 5, this can lead to incorrect conclusions about differential expression. Ultimately the best solution to the multi-mapping problem is using longer sequence reads or paired-end sequencing (Wang et al., 2009).

Longer genes will have a higher coverage of reads compared to smaller genes at the same expression level. Also, the difference in total amount of reads produced in each sequencing run will cause fluctuations in the number of fragments at each position. To normalize for this the reads per kilobase of transcript per million mapped reads (RPKM) can be calculated. However, as Bullard et al. (2010) demonstrates this metric is greatly influenced by changing levels of overabundant transcripts. Without proper normalization, a slightly higher expression of an overabundant gene will make less abundant transcripts seem to be strongly down regulated. Applying this normalization method, the sensitivity of RNA-seq for detecting differential expression is only as good or slightly better compared to microarray data. Instead they propose to scale the gene counts by a quantile (the upper-quartile) of the gene-count distribution, which greatly increases sensitivity without introducing noise and losing specificity.



▲ ***Figure 5: Complications when using multi-reads for determining differential expression.*** *When dividing reads with multiple matching locations equally over all matching positions a skewed representation of the actual differential expression can arise. Especially when one of the matching gene consists almost fully out of multi-reads the expression of this gene can appear to be only 2 fold lower, as in the upper picture, when actually all reads should have been mapped to the other gene and the real difference in expression was 20 fold. Figure taken from http://bioinformatics.ucdavis.edu.*

Differential expression

The last analysis step, after the transcript levels are quantified and, if not already included in the used differential expression method, normalized, is calculating the differential expression between different conditions. Many tools for evaluating differential expression in microarrays already existed, however, microarray data is continuous of nature (fluorescence of a spot) and RNA-seq data is in essence count data (number of mapped reads). Also, in RNA-seq the power to detect differential expression depends on the read count, and therefore on the sequencing coverage of the sample, the expression of the gene, and even the length of the gene (Garber et al., 2011; Bullard et al. 2010).

The Poisson distribution is the obvious choice for count data. It is shown that this distribution is indeed a good fit when technical replication is considered. However when biological replicates are included the Poisson distribution is a poor fit (Langmead et al., 2010). This leads to a high false positive rate in data sets with biological replicates due to underestimation of the sampling error (Oshlack et al., 2010). Ideally this error could be estimated using enough biological replicates, however few RNA-seq expression studies have enough replicates to achieve this (Garber et al., 2011).

To overcome this, methods try to model the biological variation in the count data and give a measure of significance. The count variation is modeled over replicates as a nonlinear function of the mean counts mostly using negative binomial distributions. These methods include EdgeR (Robinson & Oshlack, 2010), differential expression analysis of count data (DESeq; Anders & Huber, 2010) and Cuffdiff (Garber et al., 2011).

Alternative splicing

Functional, fully processed and spliced mRNA is preceded by precursor mRNA (pre-mRNA). In the steps toward its final form parts of the sequence, called introns, are removed from the pre-mRNA, leaving a sequence of only exons. Most exons are constitutive; they are always spliced or included in the final mRNA. However, many genes have multiple combinations of exons composing multiple different transcripts, a concept called alternative splicing. These multiple splice variants have different sequences and can result in different protein folding and function.

Abundance

Alternative processing of pre-mRNA is an important mode of genetic regulation in higher eukaryotes. Variation in splicing patterns is a major source of protein diversity from the genome. The estimated minimum level of human gene products that undergo alternative splicing is as high as 60%. Further, many transcripts have more than one alternative splice variant and some even up to thousands (Black, 2003).
In *Arabidopsis* 79% of genes include introns (Eckardt, 2002), and in 61% of these intron-containing genes AS was observed (Syed et al., 2012).

Function

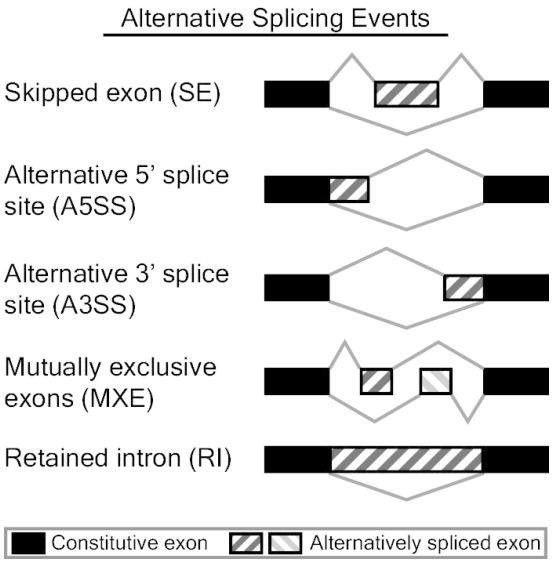
Alterations in splice sites can result in many different effects on the mRNA and protein products of a gene. Given that these products include or exclude parts of a sequence, the resulting isoforms can have different chemical and biological activity. Effects of small changes in peptide sequence include altered ligand binding, enzymatic activity, allosteric regulation, or protein localization.
AS variants can also behave as genetic switches, changing the effect of the product between multiple states based on the particular splice variant. These genetic switches are important in many cellular and developmental processes, including sex determination, apoptosis, axon guidance, cell excitation and contraction, and many others. Errors in the regulation of splicing have also been implicated as the basis for multiple diseases (Black, 2003).

Types

Classically five types of AS events are defined (Figure 6; Breitbart et al., 1987; Sammeth et al., 2008; Black, 2003):

- *Skipped exon (SE) / Cassette / Exon skipping (ES)* - An exon is included in some versions of the transcript, but not in another.
- *Alternative 5' splice site (A5SS) / Internal donor site / Alternative donor site (AD)* - An alternative splice junction is used at the 5' end (donor site), changing the 3' boundary of the upstream exon.
- *Alternative 3' splice site (A3SS) / Internal acceptor site / Alternative acceptor (AA)* - An alternative splice junction is used at the 3' end (acceptor site), changing the 5' boundary of the downstream exon.
- *Mutually exclusive exons (MXE/ME)* - An exon is included in one transcript, but another exon instead is included in another transcript.
- *Retained intron (RI) / Intron retention (IR)* - An intron is retained in some variants of a transcript. Distinguished from SE, because the retained intron is not flanked by introns itself.

While these five events conceptually cover most splicing events, many more complex events are found combining parts of above events (Sammeth et al. 2008). Alternative start and polyadenylation sites in mRNA transcription are no splice variants since the differencing step occurs already before splicing. However, these events give similar effects for mRNA and protein transcripts having parts of their sequences altered (Black, 2003). Alternative splicing poses difficulties for analyzing the RNA-seq data. However, RNA-seq also gives great opportunities to study this important regulatory mechanism in unprecedented detail.



▲ ***Figure 6 - Variants of alternative splicing.*** *In a skipped exon event one exon is, in some splice variants, not included. In alternative 5' splice sites the 5' end of a junction is different for splice variants, in alternative 3' splice sites this is true for the 3' end. With mutually exclusive exons some splice variants have exon A included, some variants exon B, but not both. A retained intron is an intron that would normally be spliced out, but is included in some variants. Image taken from http://rnaseq-mats.sourceforge.net/.*

Measuring differential expression of splice variants

Alternative splice variants pose a problem for measuring differential expression in all of the four above mentioned steps from mapping sequence reads to statistical analysis of differential expression. These steps will be reviewed again to see how alternative splicing complicates the analysis of RNA-seq data, how these hurdles can theoretically be overcome and how currently available tools implement these solutions.

Read mapping

For mapping RNA-seq reads a difference is made whether or not the complete reference transcriptome is available. If the transcriptome would be fully known unspliced alignment algorithms are faster in mapping the sequence reads. However, RNA-seq experiments frequently discover many new splice variants (e.g. Pan et al., 2008; Filichkin et al., 2010). The availability of the complete transcriptome regarding splice variants is thus for now a utopia. Unspliced aligners are limited to identifying known exons and junctions since they don't allow for large gaps in mapping reads to the genome or transcriptome. Analysis with an interest in splice variants is thus bound to spliced aligners. Furthermore, the problem with exon-first alignment algorithms described above gives a preference for seed-extend methods.

Read mapping algorithms with the ability to detect splice sites *de novo* have only been released for the last three years. Seven algorithms for this purpose are compared below: Seed-extend methods GSNAP and QPALMA, and the exon-first aligners RUM, MapSplice, SpliceMap, TopHat and SOAPsplice (Grant et al., 2011; Huang et al., 2011).

GSNAP (Wu & Nacu, 2010) is a seed-extend aligner basing its splice junction prediction on two factors. For evaluating possible splice events it uses both a probabilistic model, to score the chance for an acceptor and donor site, and a user submitted database of known splice exon-intron boundaries. The model is implemented as a maximum entropy model, which uses frequencies of nucleotides

neighboring a splice site to discriminate between true and false splice sites. GSNAP can handle not only short distance splicing events, but also long-distance intrachromosomal deletions or inversions, and interchromosomal translocations. It does suggest new combinations of known exons, but does not infer new splice sites.

QPALMA (De Bona et al., 2008) is also a seed-extend algorithm. One nice part is that it does take in to account the read's quality score given by the sequencer. However, it is limited to determining local exon–exon junctions and not unanticipated distant or interchromosomal gene fusion events (Wu & Nacu, 2010). The training set used for this is a set of previously known splice sites and thus the algorithm is biased to detecting only sites similar to these. This limits the detection of novel splice variants (Trapnell et al., 2009).

A well covered example of an exon-first algorithm able to detect splice sites *de novo* is TopHat (Trapnell et al., 2009; part of the Tuxedo suite; see also Figure 8a). TopHat first aligns full reads to the genome using Bowtie, with a main focus on matching the first 28 bp with high quality closest to the 5' end, because these are the most reliable in the Illumina sequencer. The mapped reads are clustered as putative exons.
Tophat can infer splice sites on the fly (Trapnell et al., 2012). The edges of these islands are searched for canonical introns (GT–AG exon-intron boundaries). The initially unmapped reads are tested to fit over these possible splice sites with a seed-extend method.
Recognition of introns and their splice sites by scanning for this canonical signatures is a wide spread practice. The dinucleotides splicing signal GT and AG for donor and acceptor sites appears in 98% known mammalian splice sites (Burset et al., 2000) and is also conserved in virtually all naturally occurring plant introns (Simpson & Filipowicz, 1996). TopHat is known to miss splicing events spanned by individual reads at a low level (Huang et al., 2011; Wu & Nacu, 2010).

SpliceMap (Au et al., 2010) does not use existing exon annotation and takes advantage of longer sequence reads. Reads are split in two halves and mapped to the genome with mappers like ELAND or SeqMap. Reads spanning splice junctions must have at least one half that maps to an exon. These seeds are then extended per base to find the canonical splice site (GT-AG) and matched to hits on the other site of the intron within range.

MapSplice (Wang et al., 2010) is not dependent on splice site features or intron length, consequently it can detect novel canonical as well as non-canonical splices. First reads are split in to smaller tags of typically 20-25 bp. These are mapped to the genome with any unspliced aligner, returning candidate alignments for each. If tags are not aligned, but the two tags upstream and downstream can be aligned these two are extended inwards to align the middle (unaligned) tag. If start or end tags are not aligned a short sequence of the respective start or end of the tag is searched in the direction where it should lie and the same extension starts from there. Wang et al. (2010) demonstrates that the algorithms performance is more sensitive and specific in a shorter amount of time on two synthetic data sets than TopHat and SpliceMap.

SOAPsplice (Huang et al., 2011; previously called SOAPals) first maps full reads to the genome. Initially unmapped reads are separated in two segments, so that the longest sequence at the 5' can be mapped to the reference. The boundary of the intron should be in the form of "GT-AG", "GC-AG" or "AT-AC", with preference for the first, canonical one. Paired-end and long reads can be used for additional filtering for false positives. Huang et al. demonstrates an advantage of SOAPsplice over TopHat, MapSplice and SpliceMap in a higher detection rate and a lower false positives rate at low coverage on two simulated data sets.

Finally, RUM (RNA-seq unified mapper; Grant et al., 2011) uses the speed of Bowtie to map reads to both the transcriptome and the full genome. These alignments are merged, with a preference for the transcriptome mapping. Next the initially unmapped sequences are aligned to the genome with BLAT (Blast Like Alignment Tool) and all results are merged into one final alignment. Junctions are found in aligned reads that have a large gap (by default 15 bp or more) and a known splice signal.

Grant et al. (2011) compared all above (except for QPALMA) on two artificially synthesized data sets of RNA-seq data, with the first set having low and the second set having moderate levels of polymorphisms and error rates. Consistently GSNAP and RUM performed with the highest base accuracy (the percentage of bases aligned to the right location) and lowest levels of false negative and false positive splice junctions in their mappings. MapSplice followed and TopHat, SOAPsplice and SpliceMap scored even lower.
The lower accuracy of SpliceMap, SOAPsplice and TopHat is especially visible in the second data set with moderate levels of polymorphisms and error rates, indicating a lower robustness against these variations.
Tests on mouse retina RNA-Seq analysis showed that most algorithms were able to accurately identify novel splice variants, except for a poor performance of TopHat. When comparing the runtimes the two most accurate algorithms where on the heavy side, with RUM outperforming GSNAP, especially in the data set with low levels of polymorphisms and errors.

Concluding, the seed-extend method GSNAP (no exon-first bias) or exon-first method RUM (higher speed and inferring new splice junctions) are the current read mapping algorithms with the best performance in *de novo* detection of splice junctions.

Transcriptome reconstruction

Alternative splicing makes the task of reconstructing a step harder. Mapped reads on the same genomic region might now be assigned to multiple different transcripts. To combine the correct reads to form a transcript a graph approach is taken by multiple algorithms. The only two detecting novel isoforms, not based on previous annotation are presented below.

Cufflinks (Figure 8b-c) starts by finding mapped reads that are mutually exclusive. Incompatible reads that display different splice junctions must have originated from different transcripts. Mutually exclusive reads are represented as points in a graph with edges to the left and right to compatible reads. The graph is traversed to make sure that every node is in at least one path, representing unique transcripts.

Scripture also represents the mapped reads as a graph, starting with the spliced reads. Splice site information is used to detect the direction of transcripts. The connectivity graph is build drawing edges between any two bases connected by a spliced read gap. After this all paths are scored for their significance, comparing the coverage of it to the total coverage. Using the remaining significant paths all possible graphs are constructed. Finally, information from paired-end reads can be used to join graphs and remove unlikely ones.

One of the earlier algorithms taking this graph based approach is G-Mo.R-Se (Denoeud et al., 2008). This algorithm does detect many types of alternative splice variants, except for intron retentions, making it inferior to Cufflinks and Scripture.

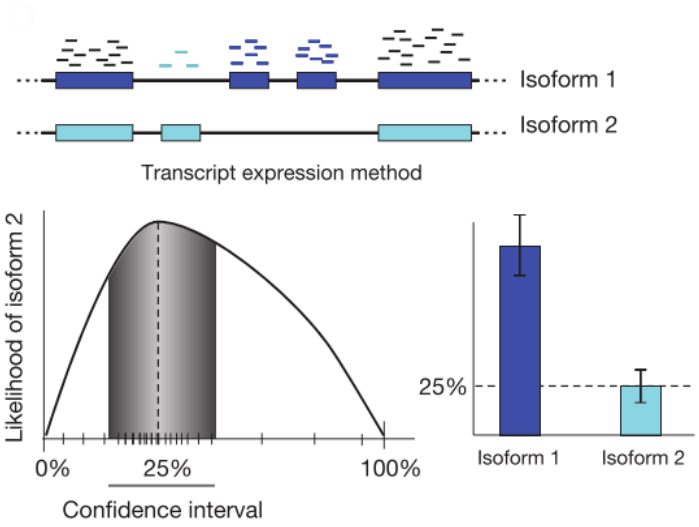
The largest difference between Scripture and Cufflinks is that the former returns all possible transcripts and the latter returns the minimal set needed to explain the reads (Garber et al., 2011). Choosing the right method depends on whether you want maximum sensitivity or maximum precision.

Expression quantification

Reads mapping to multiple locations in the genome, referred to as gene multireads, are a problem in regular expression quantification because the choice needs to be made to which location they should be assigned to. The same is true for isoforms with overlapping exons; the choice that needs to be made is to which transcript a read in these overlapping regions, called an isoform multiread, belongs.

One solution, which is used by alternative expression analysis by RNA sequencing (Alexa-seq; Griffith et al., 2010) is to only take in to account the expression of exons which are unique to an isoform. However, alternatively spliced genes with only non-unique exons can not be quantified this way.

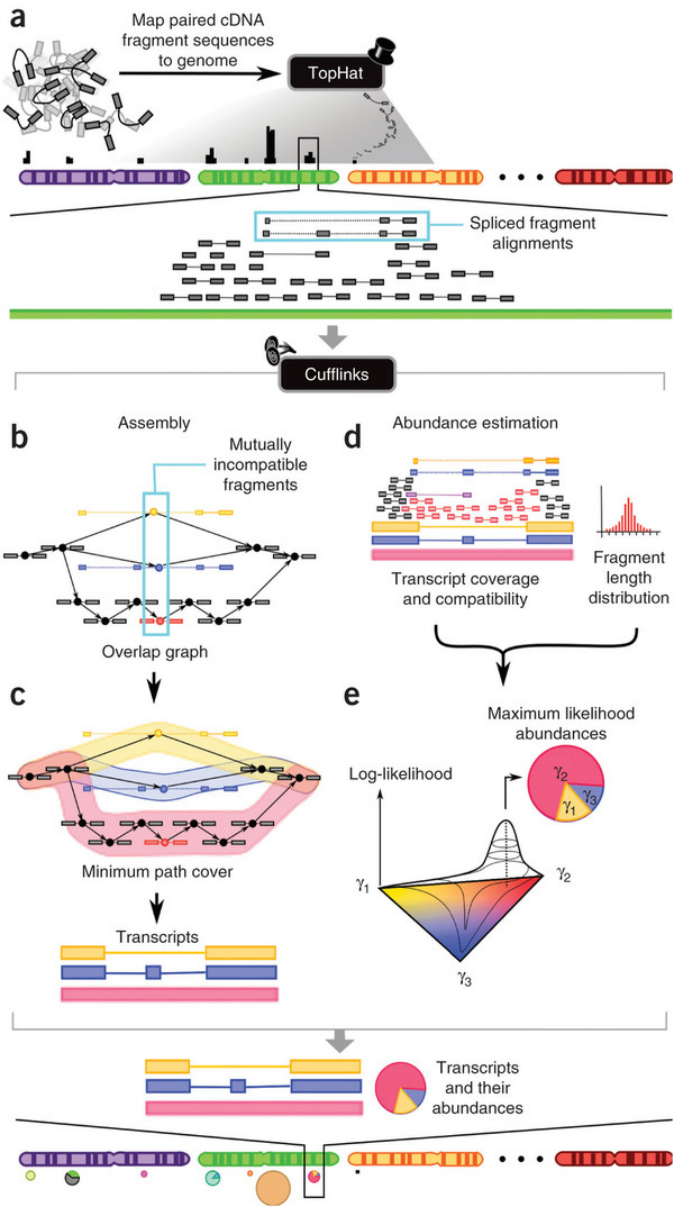
An alternative approach is to model the sequencing process and apply a maximum likelihood estimation (MLE) model to estimate the contribution of the different isoforms as shown in Figure 7. This statistical model, developed by Jiang and Wong (2009), describes how the counts mapped to the exons of a gene are related to the isoform-specific expression. By optimizing the MLE, the isoform abundance estimates are found that explain best the reads obtained. Since MLE is not an accurate expression estimate for genes expressed at low levels, Bayesian inference is used to sample alternative abundance estimates around the MLE and to calculate a confidence interval for it.



▲ ***Figure 7: Quantification of splice variants with maximum likelihood estimation (MLE) (Garber et al., 2011).*** *Above two isoforms are shown with their mapped reads. Reads for which the origin is clear are color coded, uncertain reads are depicted in black. In optimizing the MLE with the estimation of the abundance of both isoforms the uncertain reads are divided such that their distribution is as much in accordance with the reads with clear origin.*

Furthermore, the methods described in more detail below take advantage of paired-end reads in combination with the distribution of fragment size to advance the likelihood calculation. In the preparation of the RNA-seq library a size-selection step is used to control the mean length of inserted cDNA fragments. Using paired-end sequencing the total distribution of the selected fragments can be determined from read pairs that map to large, intron-less regions, like 3' UTRs.
This length distribution can then be used to enhance the inference of the quantity of different isoforms. If, for example, a read pair maps upstream and downstream of an alternatively spliced exon, the fragment distribution can be used to predict which read pairs belong to the inclusion and exclusion isoforms. This adds more reads of which can be determined from which isoform they originate, making for a better prediction of their abundance.

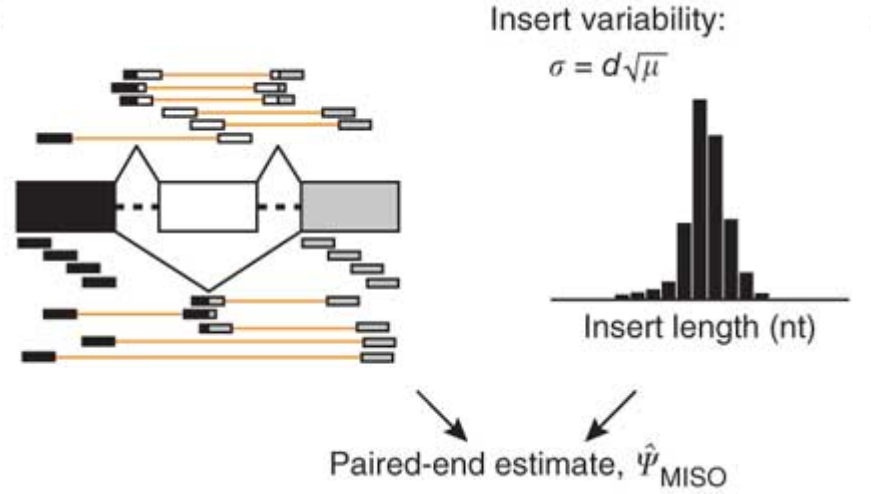
Cufflinks (Figure 8d-e), already mentioned above as a transcriptome reassembler, is one of the methods that use an extension of the model proposed by Jiang and Wong (2009) to quantify the abundance of isoforms. Expression of each transcript is determined by a statistical model using combinatorial optimization. Combinatorial optimization searches for an optimum object in a finite set of objects. Typically this set of object has a concise representation, like a graph, and the number of objects is so big that scanning all one by one is not an option. Solving the traveling salesman problem is an example of this (Schrijver, 2003).



▲ ***Figure 8: Overview of Cufflinks and the Tuxedo suite (Trapnell et al., 2010).*** *a. Cufflinks is part of the Tuxedo suite, just like TopHat which can be used to provide the input for Cufflinks. b. and c. are the transcriptome reconstruction part of Cufflinks. b. A graph is constructed where every read is a node and non-excluding reads are connected. c. This graph is traversed to construct the minimal number of transcripts needed to explain the reads. d. and e. are the expression quantification part of Cufflinks. d. The fragments are matched to the transcripts they could have originated from (represented here using colors). Cufflinks estimates transcript abundance using a statistical model in which the chance of observing a fragment is linear to the abundance of the transcript it originated from. Some reads (indicated by intermediate colors) can be assigned to multiple transcripts. Because in paired-end sequencing both ends of a fragment are sequenced their assumed length might differ depending on which transcript the read is assigned. Taking the distribution of fragment size this helps to assign a likelihood for the originating transcript. e. In the last step the program optimizes the distribution over the transcripts using combinatorial optimization. This results in the distribution that best explains the observed fragments, depicted here as a pie diagram.*

The program numerically maximizes a function that assigns a likelihood to all possible sets of relative abundances of all isoforms. Since the likelihood function is a non-negative linear model, it has a unique maximum, representing the abundances that best explain the observed fragments (Trapnell et al., 2010). To normalise for transcript length and total amount of sample the expression is given as fragments per kilobase of transcript per million fragments (FPKM), the fragment-equivalent of RPKM. Cufflinks at first assumed a uniform distribution of reads along a transcript. Because of the fragmentation process of RNA, read ends of transcripts have a lower coverage than middle parts. Taking this fragmentation bias in to account can explain up to 50 percent of the variation in coverage (Li et al., 2010). Cufflinks is later also updated to account for fragment bias (Roberts et al., 2011), just like the following algorithms.

The MISO (Mixture of Isoforms; Katz et al., 2010; figure 8) model is a stand alone algorithm using the same model to estimate isoform expression. However, here optimization is achieved with a technique based on Markov Chain Monte Carlo (MCMC) sampling. By starting with a Markov Chain with a default set of state transitions and stepwise optimizing these values a most likely distribution is determined (Diaconis, 2008).



▲ ***Figure 8: Quantification of isoforms with MISO (Katz et al., 2010).*** *Just like Cufflinks, MISO first uses reads supporting different isoforms to quantify their relative abundance. The example on the left shows reads supporting the inclusion of an exon (above) and the reads supporting the exclusion variant (below). It also uses length size distribution (right) of paired-end reads for determining to which transcript a read should be assigned. Using this evidence the maximum likelihood estimation model is optimized using a technique based on Markov Chain Monte Carlo sampling, resulting in Ψ_{MISO} , the estimation of relative isoform abundances.*

A third much used method for optimizing the prediction of isoforms is Expectation-Maximization, used by RSEM (Li et al., 2010; Li & Dewey, 2011), MMSEQ (Turro et al., 2011) and IsoEM (Nicolae et al., 2011). Expectation-Maximization consists of two iterative steps. At first a uniform distribution among transcripts is assumed. In the expectation step the log-likelihood of the current predicted distribution is calculated based on the data in the reads. In the maximization step the calculated log-likelihood is compared with the previous and if there is still a significant improvement the expectation step is started again with slightly adjusted distribution.

IsoEM is shown to outperform versions of RSEM and Cufflinks at the beginning of 2010 in a variety of quality metrics on synthetic and real RNA-Seq data sets (Nicolae et al., 2011). Also MMSEQ showed improved isoform estimates over the then latest RSEM for medium to low expression transcripts (Turro et al., 2011). The newer version of RSEM, in turn, is demonstrated to outperform IsoEM and Cufflinks in prediction quality on a simulated data set; however the best performance on a natural data set was achieved by the recent Cufflinks algorithm on a natural data set (Li & Dewey, 2011). There is a need for more and objective comparison of all algorithms on a well-controlled synthetic data set.

Important to note is that the incorrect or misassembled isoforms impact the certainty of the expression prediction greatly (Garber et al., 2011). Using a method for transcriptome reconstruction that generates the maximal isoform set, like Scripture, it is necessary to filter transcripts before quantification. Normalization as described for the general RNA-Seq analysis protocol, for e.a. transcript length, total read count and overabundant transcripts, is the same for isoforms as for other transcripts.

Differential expression

In essence the determination of differential expression of RNA-Seq data including alternative splice variants is no different from the procedure without. Now transcripts are obtained and read counts for each transcript are determined and normalized. Next, in comparing expression between genes and/or conditions, statistics should be used to determine whether two read counts are different and to what order of magnitude. Algorithms like Cuffdiff (Trapnell et al., 2012), DESeq (Anders & Huber, 2010) and edgeR (Robinson & Oshlack, 2010) can be used to achieve this.

However, edgeR and DESeq do only take raw gene count in to account and not counts over individual isoforms. Next to the fact that this does not yield any insight in the changes of relative isoform abundance, it also results in problems for determining the gene expression level in samples with alternative splicing. Both tools do not normalize for transcript length in their comparisons, which is no problem when comparing expression of two identical transcripts between samples. However, if a gene has multiple isoforms with different lengths, comparing the expression of this gene over two samples with different relative isoform abundance will indicate a change in gene expression even when the overall gene expression is the same. It is also possible, but most likely rare, that a change in relative isoform abundance cancels out a real change in gene expression resulting in a false negative.

DEXSeq (Anders et al., 2012) is an adjusted version of DESeq that indicates differential exon usage in genes, helping to detect differential isoform expression.

Cuffdiff tracks changes in the relative abundance of two types of isoforms: transcripts sharing a common transcription start site (TSS), and in the relative abundances of transcripts from the same gene but with different promoters. The former shows changes in splicing of the same pre-mRNA, and the latter shows changes in relative promoter use within a gene (Trapnell et al., 2012). The statistics used to achieve this is a little different, taking into account the variance originating from the uncertainty of mapping reads to multiple transcripts (Cufflinks; http://cufflinks.cbcb.umd.edu/howitworks.html). Expression of a gene is simply calculated by adding up the expression of the splice variants.

Conclusions

In recent years great advances have been made in analyzing the differential expression of alternatively spliced transcripts. For all four steps of the analysis path there are suitable tools available to address the problems alternative splicing causes.

- For mapping reads on the reference genome the important factor for alternative splicing detection is the ability to suggest splice junctions *de novo*. Two methods stand out with their results and approach: RUM, a fast exon-first mapping algorithm, and GSNAP, a seed-extend method.
- For reconstructing the transcriptome and inferring alternative splice variants Cufflinks and Scripture are the best options, using a graph based algorithm to return respectively the minimal and maximal set of isoforms needed to explain the mapped reads.
- Determining read counts for isoforms is done via optimizing a maximum likelihood estimation model to find a distribution over isoforms that most closely matches the data. Many algorithms are at hand that implement this MLE with different statistical bases. However, objective comparisons of their relative performance on a strictly controlled artificial data set are unavailable.
- Cufflinks seems to be the most extensive method to determine differential expression of isoforms and changes in relative isoform abundance between samples.

Most of the reviewed tools take compatible files as input and output, so theoretically tools at different steps of the analysis can be arbitrarily combined. However, the performance of algorithms may depend on which tool has provided the input, following from small differences in assumptions and procedure in their predecessors.

The short reads of RNA-seq pose many problems in recombining the data to find differential expression due to placement uncertainty. However, as Li & Homer (2010) note, as the reads sizes increase and their cost decreases, in a few years long reads will dominate the sequencing landscape. This will lead to less uncertainty for placing reads mapping to repeats, low complexity regions or alternative splice variants and will increase the quality of the outcome of our RNA-seq data analysis.

Literature

Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome biology* **11**, R106 (2010).

Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome research* (2012).

Au, K. F. F., Jiang, H., Lin, L., Xing, Y. & Wong, W. H. H. Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic acids research* **38**, 4570–4578 (2010).

Black, D. L. Mechanisms of alternative pre-messenger RNA RNA splicing. *Annual Review of Biochemistry* **72**, 291–336 (2003).

Breitbart, R. E., Andreadis, A. & Ginard, B. N. Alternative Splicing: A Ubiquitous Mechanism for the Generation of Multiple Protein Isoforms from Single Genes. *Annual Review of Biochemistry* **56**, 467–495 (1987).

Bullard, J., Purdom, E., Hansen, K. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11**, 94+ (2010).

Burset, M., Seledtsov, I. A. & Solovyev, V. V. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic acids research* **28**, 4364–4375 (2000).

De Bona, F., Ossowski, S., Schneeberger, K. & Rätsch, G. Optimal spliced alignments of short sequence reads. *Bioinformatics* **24**, i174–i180 (2008).

Denoeud, F. *et al.* Annotating genomes with massive-scale RNA sequencing. *Genome biology* **9**, R175 (2008).

Diaconis, P. The Markov chain Monte Carlo revolution. *Bulletin of the American Mathematical Society* **46**, 179–205 (2008).

Eckardt, N. A. Alternative Splicing and the Control of Flowering Time. *The Plant Cell Online* **14**, 743–747 (2002).

Filichkin, S. A. *et al.* Genome-wide mapping of alternative splicing in Arabidopsis thaliana. *Genome Research* **20**, 45–58 (2009).

Garber, M., Grabherr, M. G., Guttman, M. & Trapnell, C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature methods* **8**, 469–477 (2011).

Grant, G. R. *et al.* Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics* **27**, 2518–2528 (2011).

Griffith, M., Griffith, O. & Mwenifumbo, J. Alternative expression analysis by RNA sequencing. *Nature methods* **7**, (2010).

Guttman, M. *et al.* Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnology* **28**, 503–510 (2010).

Haas, B. J. & Zody, M. C. Advancing RNA-Seq analysis. *Nature Biotechnology* **28**, 421–423 (2010).

Huang, S. *et al.* SOAPsplice: Genome-Wide ab initio Detection of Splice Junctions from RNA-Seq Data. *Frontiers in genetics* **2**, (2011).

Jiang, H. & Wong, W. H. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics (Oxford, England)* **25**, 1026–32 (2009).

Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature methods* **7**, 1009–15 (2010).

Langmead, B., Hansen, K. D. & Leek, J. T. Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome biology* **11**, R83 (2010).
Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* **12**, 323 (2011).
Li, B., Ruotti, V., Stewart, R. M., Thomson, J. a & Dewey, C. N. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics (Oxford, England)* **26**, 493–500 (2010).
Li, H. & Homer, N. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics* **11**, 473–483 (2010).
Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research* **18**, 1851–8 (2008).
Lunter, G. & Goodson, M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome research* **21**, 936–9 (2011).
Nicolae, M., Mangul, S., Măndoiu, I. I. & Zelikovsky, A. Estimation of alternative splicing isoform frequencies from RNA-Seq data. *Algorithms for molecular biology* **6**, 9 (2011).
Oshlack, A., Robinson, M. & Young, M. From RNA-seq reads to differential expression results. *Genome Biology* **11**, 220+ (2010).
Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40**, 1413–1415 (2008).
Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L. & Pachter, L. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome biology* **12**, R22 (2011).
Robertson, G. *et al.* De novo assembly and analysis of RNA-seq data. *Nature methods* **7**, 909–12 (2010).
Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology* **11**, R25 (2010).
Sammeth, M., Foissac, S. & Guigó, R. A general definition and nomenclature for alternative splicing events. *PLoS computational biology* **4**, e1000147+ (2008).
Schrijver, A. *Combinatorial optimization*. (Springer-Verlag: Berlin, Heidelberg, New York, 2003).
Simpson, G. G. & Filipowicz, W. Splicing of precursors to mRNA in higher plants: mechanism, regulation and sub-nuclear organisation of the spliceosomal machinery. *Plant Molecular Biology* **32**, 1–41 (1996).
Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *Journal of molecular biology* **147**, 195–7 (1981).
Syed, N. H., Kalyna, M., Marquez, Y., Barta, A. & Brown, J. W. Alternative splicing in plants - coming of age. *Trends in plant science* (2012)
Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols* **7**, 562–578 (2012).
Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28**, 511–515 (2010).
Turro, E. *et al.* Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome biology* **12**, R13 (2011).
Wang, K. *et al.* MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Research* **38**, e178 (2010).
Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics* **10**, 57–63 (2009).
Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873–881 (2010).