promote the recognition of strong splice sites instead [Bentley, 2014]. Finally, the regulation of splicing decisions is not limited to the role of specific SFs, since fluctuations in the concentration of core components of the spliceosome are also known to influence the splicing outcome [Saltzman et al., 2011].

Overall, the above mentioned processes guarantee that splicing occurs in an accurate albeit flexible fashion. The accuracy of splicing is further increased by the many rearrangements that are required before the actual intron removal reaction can occur, and splicing errors are eliminated by the nonsense-mediated decay pathway. On the other hand, the accumulation of splice site mutations or the alteration in the function of spliceosomal components can lead to serious phenotypic consequences and, in fact, dysregulation of splicing has been linked to many diseases, including cancer [Ladomery, 2013; Padgett, 2012; Tazi et al., 2009].

1.3 Studying the transcriptome with RNA sequencing

In the past few years, RNA sequencing (RNA-seq) has become the method of choice for the study of transcriptome composition [Mortazavi et al., 2008; Wang et al., 2009]. Compared to microarrays, which constituted the first technology for the high throughput comparison of expression levels across conditions, RNA-seq offers a much bigger dynamic range to study gene expression patterns, and enables a much broader set of analyses without the need for intricate experimental designs [Malone and Oliver, 2011]. For example, besides standard differential gene expression analysis, popular applications of RNA-seq comprise the identification of novel transcribed regions, including fusion genes, the deconvolution of allele specific expression, and, as further explored in this thesis, the possibility to estimate transcript expression levels and to study differential splicing across conditions.

Since the introduction of the first sequencing machines in 2005, this technology has seen the rise and fall of many companies; however, following the acquisition of Solexa, Illumina's platforms have consolidated as the most commonly used. The reason behind such wide adoption of Illumina's systems is the large volume of information obtained from a typical sequencing run (*i.e.* sequencing depth), which, at a good ratio with the cost, compensates for the lower accuracy compared to other competitors [Mardis, 2013]. Thus, although microarrays can still be a

cheaper option to perform routine differential expression analysis at the gene level, the larger scope of applications and the decrease in the costs of sequencing (just announced to have reached the target of 1,000$ per human genome by Illumina while writing this thesis) explain the increasing popularity of RNA-seq.

In this section, I introduce the typical steps required to sequence a transcriptome with an Illumina platform, since this is the one that has been used to produce all the datasets analysed here. Moreover, I provide a detailed description of the most commonly used methods to study the transcriptome composition from RNA-seq data, with special emphasis on the analysis approaches used within the different chapters.

1.3.1 A typical sequencing workflow

The first step in transcriptome sequencing is library preparation, and consists of obtaining the starting material and converting it into a cDNA library that can be loaded into the sequencing machine (**Figure 1.5**) [van Dijk et al., 2014]. Following RNA extraction, the RNA species of interest are typically enriched through either polyA selection or ribodepletion. In both cases, the aim is to diminish the concentration of rRNAs, *i.e.* the most abundant species of RNA in the cell. With the first method, this is achieved through the use of oligo-dT beads, which enable the specific extraction of polyAdenylated RNAs, hence ensuring a good representation of mRNAs (**Figure 1.5** - step 1). Conversely, ribodepletion relies on the use of ribonucleases to specifically digest rRNAs, and has the advantage of not restricting the analyses to a specific type of RNA. Indeed, the term total RNA is typically used to refer to datasets produced with such protocol, while those obtained with the former method are commonly known as polyA-selected. Due to the simpler protocol and its lower price, polyA selection emerges as the most popular choice amongst the currently available RNA-seq datasets, with the notable exception of those studies aimed at characterising non-coding RNA species, which typically lack a polyA tail. The extracted RNA is then fragmented via hydrolysis with divalent cations and retro-transcribed into double stranded cDNA by using random hexamer primers, since the sequence of the obtained fragments is not known at this point (**Figure 1.5** - step 2). These steps are followed by the ligation of adapter sequences at both ends of each cDNA fragment (**Figure 1.5** - step 3). Such adapters satisfy two different purposes: on the one hand, they enable the hybridisation of those fragments into the flow cell,

where the sequencing takes place; on the other hand, they serve as primers for the sequencing reaction. Then, the resulting cDNA fragments are size-selected through gel electrophoresis to fit within the range required by the sequencing machine (typically 300-500 bp). Fragments outside this range will be missed; hence the existence of alternative protocols for the study of small RNAs [Zhuang et al., 2012]. Finally, the cDNA library is amplified by PCR.

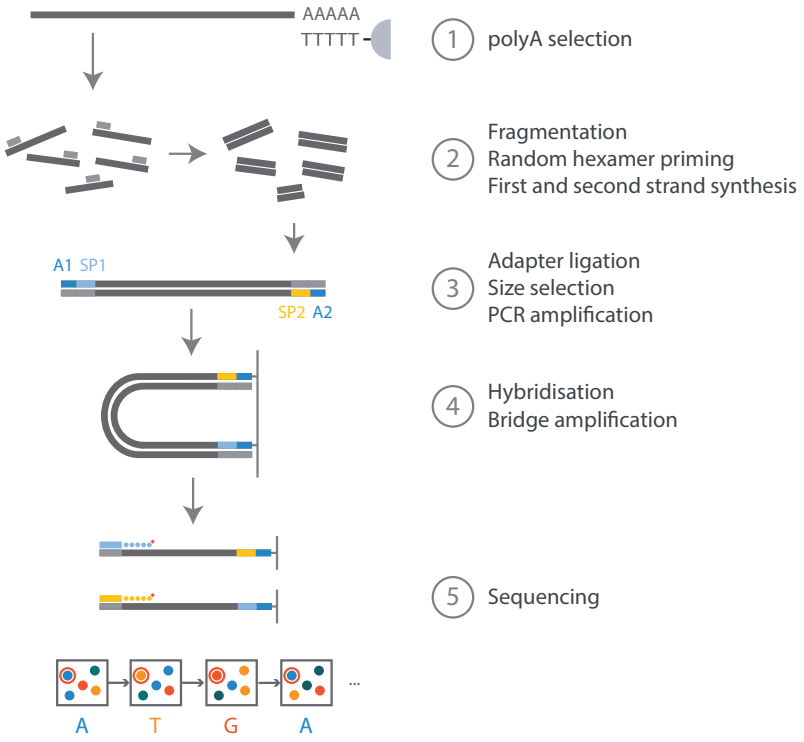


Figure 1.5| Overview of library preparation and sequencing steps in an Illumina platform. A typical paired-end workflow is illustrated here, which consists of ligating different adaptors at each end of the initial cDNA molecule. This enables sequencing each cDNA fragment from both ends, in two separate reactions, and has further advantages for the downstream bioinformatic analyses compared to single-end approaches. Adapted from Mardis [2013].

Once the library preparation procedure has finished, samples can be loaded into a flow cell for sequencing [Mardis, 2013]. Such flow cell is saturated with adapters that are complementary to the ones ligated at both ends of the cDNA fragments, consequently promoting the hybridisation of the denatured double

strand molecules. After this step, the starting material needs to be amplified once again in order to increase the signal for the sequencing reaction, this time through bridge amplification (**Figure 1.5** - step 4). Such process consists of the synthesis of fragments that are complementary to the hybridised cDNA molecules, which will in turn bend and hybridise with adjacent adapters, thus enabling subsequent rounds of synthesis. As a result, a large number of clusters with identical sequences will be formed, now ready to undergo sequencing. Illumina platforms rely on sequencing by synthesis to read the base pair composition of each cDNA cluster (**Figure 1.5** - step 5) [Bentley et al., 2008]. This reaction is based on the use of modified versions of the four bases, which differ from the standard nucleotides in the fact that they incorporate a reversible terminator, as well as a fluorescent dye. Hence, during each sequencing cycle, and following the addition of the necessary reagents, elongation will be blocked after a successful base incorporation, the identity of which can be recorded by measuring its fluorescent signal. Repetition of this process will lead to a set of images, which after interpretation with a base calling software, will be converted into a set of sequences or reads [Das and Vikalo, 2013]. Such reads represent the set of molecules expressed in the initial sample, and their length corresponds to the number of cycles performed during the sequencing reaction. Eventually, the obtained sequence information, together with the probability of a wrong base call at each given position of the read (*i.e.* Phred score), are stored in a plain text file in FASTQ format [Cock et al., 2010].

RNA-seq is not such an established technology as microarrays, and in spite of its many advantages, it still has some challenges. For example, regarding library preparation, it is known that the random hexamer priming step is not as random as initially proposed, since certain fragments have been observed to be preferentially converted to cDNA due to sequence composition [Hansen et al., 2010]. In the same category of sequence-dependent biases, the PCR amplification step has also been described to lead to differential amplification of fragments with higher or lower GC content [Benjamini and Speed, 2012], and it is known that failure to block the elongation reaction or to remove the fluorescent dye during the sequencing step can lead to wrong base calls [Metzker, 2010]. In most cases, the identification of such biases has been accompanied by the introduction of alternative protocols or analysis methods to overcome them. For example, several algorithms now try to take into account potential biases derived from the random

hexamer amplification step (*e.g.* Cufflinks [Trapnell et al., 2010], MMSEQ [Turro et al., 2011]). Alternative library preparation methods have also been proposed to account for PCR bias, whereby random barcodes (*i.e.* molecular identifiers) are used to quantify the absolute number of molecules [Shiroguchi et al., 2012]. Finally, some downstream analysis algorithms also incorporate information on the probability of a wrong base call at a given position of the read, as reported by the Phred score (*e.g.* Kim et al. [2013]).

On the other hand, alternative library preparation strategies can also add further information to the experiment. This is the case of strand-specific protocols, which are able to provide information on the strand from which each read originates [Levin et al., 2010]. Similarly, multiplexing emerges as a widely used approach to optimise the amount of data that can be obtained from each sequencing run, by enabling pooling of several samples into a single lane of the flow cell through the use of sequence identifiers (*i.e.* sample-specific barcodes) [Wong et al., 2013b]. Lastly, a very common strategy to overcome limitations on the read length and try to span larger regions consists of sequencing each cDNA fragment from both ends (*i.e.* paired-end sequencing, as opposed to the single-end strategy), which can be achieved through the use of modified adapters (**Figure 1.5** - step 3) [Mardis, 2013].

1.3.2 Read mapping strategies

The next step in a typical RNA-seq analysis pipeline consists of identifying, for each read, the genomic region from which it has originated. In RNA-seq, this task is equivalent to discovering the loci that are expressed in a given sample. In general, two different strategies exist to perform this task: on the one hand, reads can be aligned to the reference genome or transcriptome, provided that such information is available for the species of interest; on the other hand, they can be directly assembled into contigs (*i.e.* contiguously expressed regions) with the aim of reconstructing the set of expressed transcripts. The first strategy constitutes a much simpler approach, and it is typically the method of choice when working with model organisms.

Independently of the strategy used, read mapping is typically the most time consuming step of the analysis workflow, and the available tools make use of heuristic parameters such as the maximum number of allowed mismatches per read in order to speed up this task. Such processing can lead to information loss given a

decrease of quality at the 3′ end of the read, which emerges as a common profile when working with Illumina platforms due to the increased difficulty in interpreting the fluorescent signal as sequencing cycles accumulate [Minoche et al., 2011]. Thus, in order to avoid such reads being discarded, it is often useful to first perform a quality control and pre-filtering step, whereby read sequences can be shortened (*i.e.* trimmed) in terms of their quality (*e.g.* Andrews [2010]). Similarly, reads with overall low quality can be also removed, in order to speed up the subsequent mapping process.

1.3.2.1 Alignment to the genome or transcriptome

A commonly used approach in the cases where a reference genome is available, is to align the reads directly to that sequence. Similarly, reads can be aligned to the transcriptome instead, provided that a good annotation exists. The main advantage of using this second strategy is that the alignment task is simplified due to the lack of intronic sequences; but this comes at the price of limiting the number of downstream analysis that can be performed (*e.g.* alignment to the transcriptome is not compatible with the identification of novel expressed regions nor the study of intronic expression levels). Thus, a good compromise is the use of hybrid approaches, as implemented in TopHat [Kim et al., 2013].

TopHat is a read mapping tool specially intended for RNA-seq data, since it enables alignment of the reads to the genome while taking into consideration the existence of splice junctions (**Figure 1.6**). It is based on Bowtie [Langmead and Salzberg, 2012], an independent algorithm for the alignment of short reads, and its main strength is the ability to detect exon-exon junctions without the need for any *a priori* knowledge on the annotation. However, the search can be simplified by providing such information, and in that case TopHat will first attempt to map the reads to the derived transcriptome. Those that fail to align will be subsequently queried against the genome (**Figure 1.6** - step 1). Alternatively, reads can also be mapped to the genome directly (**Figure 1.6** - step 2). In both cases, the goal is to assemble the initially aligned reads into exons, which might eventually become connected through spliced alignment (**Figure 1.6** - step 2). Reads that fail to align in this initial phase, as well as those that map with low alignment scores, are subsequently used to build a database of possible splice junctions, by splitting them into smaller segments and re-aligning those independently (**Figure 1.6** - step 3). In this context, a splice junction is reported whenever a read appears to span multi-

ple exons, *i.e.* in the cases in which an internal segment fails to align, or when two consecutive segments from the same read do not align contiguously on a given genomic locus. Next, the identified splice sites and their flanking sequences are concatenated into a novel transcriptome, which is then used to re-align the set of unmapped reads (**Figure 1.6** - step 4). In the case of paired-end data, each read is processed separately, and the alignments obtained are evaluated in a final phase by taking into account additional sources of information such as fragment length and orientation of the reads. Finally, all the information gathered during the mapping process is reported in SAM/BAM format [Li et al., 2009].

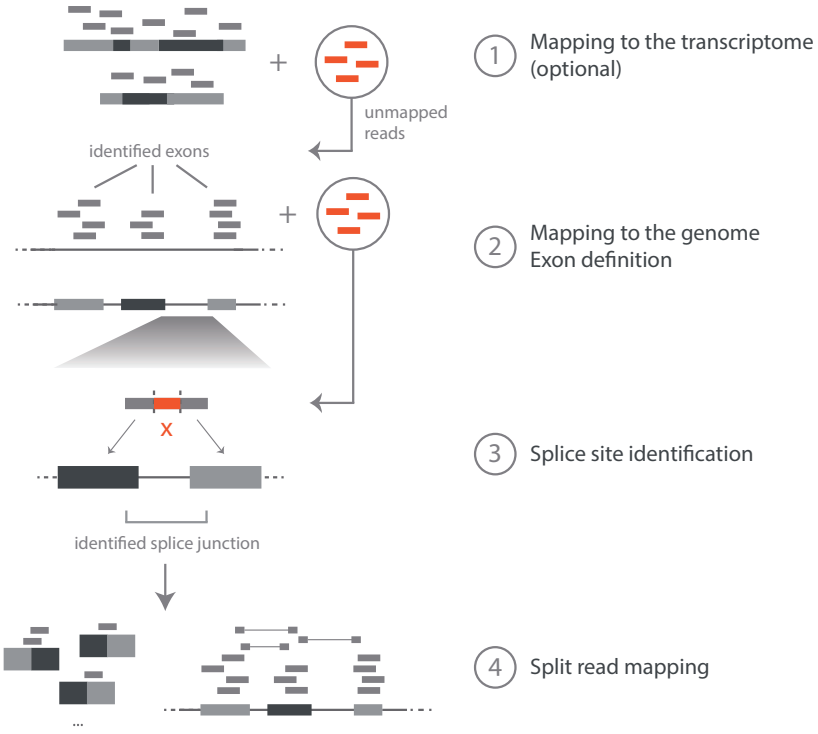


Figure 1.6 | Overview of the mapping algorithm implemented in TopHat. In the presence of an annotation file, TopHat uses a hybrid approach to uncover the genomic loci from which the detected reads could have originated. Alternatively, TopHat can directly align the reads to a reference genome. In both cases, the first step consists of identifying a set of expressed exons, and this is followed by the detection of splice junctions by using information from those reads that span multiple exons. Adapted from Kim et al. [2013].

1.3.2.2 *De novo* assembly

De novo assembly emerges as an advantageous strategy in the cases where the species of interest lacks a reference genome. Additionally, it can be used in situations where the genome composition of a given sample is expected to differ largely from that of the reference assembly (*e.g.* cancer samples). The goal here is to assemble the reads into sets of expressed regions (*i.e.* contigs), by relying on their overlap. Nonetheless, the short read length adds to the non-triviality of the task, and even though the use of paired-end data can simplify this process, lowly expressed regions are often difficult to solve. In terms of available software, Trinity [Grabherr et al., 2011] emerges as the most popular tool to perform this task; however, such methods are not used in this thesis and are covered elsewhere [Martin and Wang, 2011].

1.3.3 The estimation of expression levels

Once the reads have been assigned to a specific location in the genome or transcriptome, the next step in the RNA-seq analysis pipeline consists of estimating expression levels for the features of interest, typically genes and transcripts. Similarly to the scenarios encountered during the mapping step, the quantification of expression levels can be achieved by relying on existing information, but it can also be performed independently from any annotation, thus enabling *de novo* identification of transcribed regions (*i.e.* novel genes or unnanotated transcripts within known gene loci).

1.3.3.1 Gene expression levels

When working at the gene level, and provided a complete annotation exists, abundance estimation can be easily achieved by counting how many reads overlap each given locus (**Figure 1.7a**). Such count-based approach constitutes the starting point for many downstream analysis algorithms (*e.g.* DESeq2 [Love et al., 2014], DEXSeq [Anders et al., 2012]), and can be easily performed with the popular tool htseq-count [Anders et al., 2014]. However, despite this apparent simplicity, there are some challenges that need to be considered. First, in order not to over-estimate expression levels, reads that map to multiple locations in the genome, and which arise from repetitive or duplicated loci, need to be handled with care. In this situation, htseq-count adopts the most conservative approach and discards them, but other alternative strategies have been proposed in order to attempt to keep

the information from such multi-mapping reads. Generally, these consist of uniformly distributing them to all the mapped positions (*e.g.* Trapnell et al. [2010]), or probabilistically assigning them depending on the coverage at each mapping locus (*e.g.* Trapnell et al. [2010]; Turro et al. [2011]; first proposed by Mortazavi et al. [2008]). Second, special attention is required in the case of overlapping features. htseq-count offers several execution modes to deal with this scenario, even though in some cases reads remain ambiguously assigned (**Figure 1.7b**). Finally, despite not being intended for *de novo* quantification, htseq-count also gives the user some flexibility on how strictly the provided feature coordinates should be taken into account (**Figure 1.7b**).

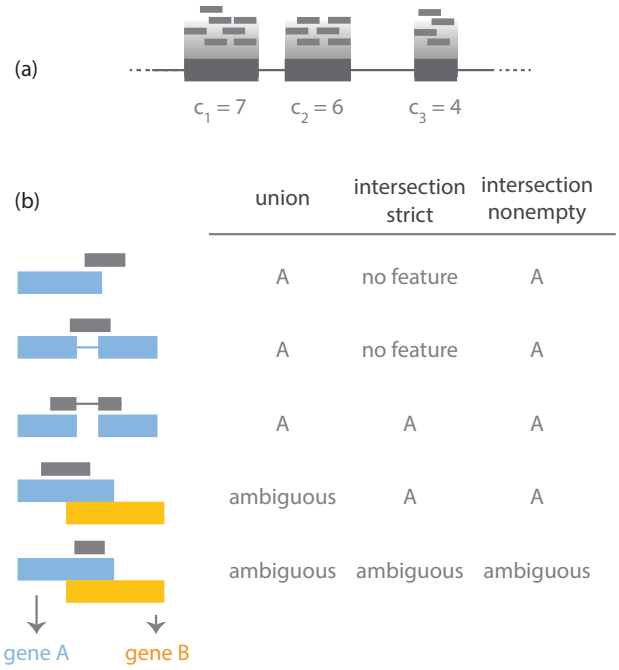


Figure 1.7| Overview of htseq-count.
(a) Illustration of the read counting concept. Expression estimation with htseq-count consists of counting the reads that overlap with the features of interest. In this example, any reads that fall outside the grey areas will not be considered.
(b) The three different execution modes available in htseq-count. htseq-count provides different counting modes to rescue reads that do not strictly overlap with the provided coordinates. These modes differ in how strictly the annotation is taken into account and in the behaviour adopted in the case of overlapping features. Adapted from Anders et al. [2014].

Alternatively, gene expression can be calculated after estimation of transcript expression levels, by aggregating the corresponding individual transcript abundances, as implemented for example in Cufflinks [Trapnell et al., 2010] and MMSEQ [Turro et al., 2011].

1.3.3.2 Transcript expression levels

On the other hand, the task of estimating expression levels becomes far more complicated when focusing on individual transcripts, since many reads will overlap with exons that are shared across multiple isoforms of the same gene. In this scenario, the question translates into attributing reads to specific transcripts, and further inference approaches are needed. The available algorithms typically rely on different sources of information in order to probabilistically estimate transcript expression levels, the most valuable one being those reads that map uniquely to one of the annotated transcripts within the loci. Moreover, reads that span two different exons (*i.e.* split reads) become especially informative. For example, splice junctions that involve cassette exons tend to provide unambiguous support for their inclusion or skipping. This is where paired-end information becomes most relevant: sequencing both ends of the initial cDNA fragment facilitates covering larger genomic regions, thus increasing the probability that a given read pair is mapped across different exons (*i.e.* spliced reads). Similarly, information on the fragment length distribution can also be used to deconvolute ambiguous assignments, by attributing a lower likelihood to those that would require extreme distances between the paired reads.

One of the most popular tools to estimate transcript expression levels is MISO [Katz et al., 2010], which formulates this task as a Bayesian inference problem [Beaumont and Rannala, 2004], whereby the goal is to find a probability distribution (the posterior) over transcript abundances (Ψ) given the observed RNA-seq data (**Figure 1.8**). Such distribution can be computed in terms of two quantities: the expectation about the value of Ψ before observation of the reads (the prior, set by MISO as a uniform distribution), and the probability of observing the data given a fixed value of Ψ (likelihood of the reads). Thus, following sampling across the space of Ψ values, all possible assignments of every read to each isoform are probabilistically evaluated, and this information is subsequently used to refine the search of the optimal set of Ψ values that best explain the observed data. Finally, an estimate of transcript abundances is obtained by calculating the mean over the

computed posterior distributions, and confidence intervals are also calculated as a measure of the certainty of the estimate.

Similar approaches are taken by other alternative methods, including Cufflinks [Trapnell et al., 2010] and MMSEQ [Turro et al., 2011], both of which are also used in this thesis. The main difference among the mentioned methods relies on the implementation of the inference approach, as well as the type of input required. Similarly to MISO, Cufflinks requires the reads to be mapped to a reference genome, but relies on a frequentist approach to find the expression levels that best explain the observed data, which does not allow quantification of the uncertainty around the obtained expression estimates. On the other hand, MMSEQ adopts a Bayesian model similar to the one of MISO, but requires mapping to the transcriptome, which limits the scope of the downstream analysis that can be performed. Furthermore, both MMSEQ and Cufflinks accommodate known sequence biases in their models, and are also able to retain information from reads that map to multiple genes.

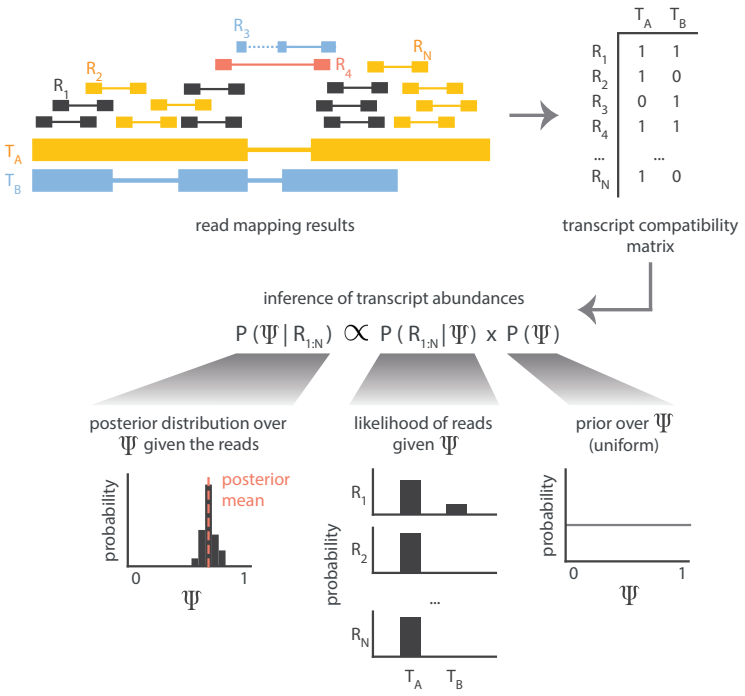


Figure 1.8 | Overview of the analysis workflow implemented in MISO for the estimation of transcript abundances. After alignment to the genome, MISO evaluates the compatibility of each read with all the transcripts annotated within a given gene. For example, in the scenario depicted here both read 2 (R_2) and read N (R_N) can only be detected if transcript A (T_A) is expressed, whilst read 3 (R_3) uniquely supports transcript B (T_B). Inference of expression levels is then done by calculating a probability distribution (the posterior) over such expression (Ψ) given the reads. Following Bayes' rule, this distribution can be obtained from the product of two terms: the likelihood of obtaining the observed set of reads given a fixed value of Ψ and the expectation on the value of Ψ before observation of the data (the prior). Hence, the inference problem translates into sampling from the space of Ψ values and evaluating all possible assignments of each read to each transcript. For example, given the larger number of reads that support the expression of transcript A in comparison to transcript B, higher expression of the latter will be probabilistically penalised, and this will contribute to the preferential assignment of ambiguous reads to the former. Fragment length information can also be used to deconvolute ambiguous assignments, as it is the case for read 4 (R_4): assigning such read to transcript B would imply an unusual distance between the paired reads, hence increasing the likelihood that it can be explained by transcript A. Finally, the overall probability of observing the reads given the evaluated Ψ value is obtained by combining the information from all reads, and this information is further used to calculate the posterior distribution. Adapted from Wang et al. [2010].

1.3.3.3 *De novo* transcript identification

One of the main advantages of RNA-seq over microarrays is the possibility to gather information on novel expressed loci in a more high throughput manner. In this context, Cufflinks [Trapnell et al., 2010] emerges as one of the most popular tools to achieve this task, thus complementing its aforementioned quantification capabilities (**Figure 1.9**). By relying on the output provided by TopHat [Kim et al., 2013], the Cufflinks assembler first identifies the expressed loci (*i.e.* genes) present in a given sample. Then, for each of them, it evaluates the observed data in search of a set of incompatible reads, *i.e.* reads which have necessarily originated from different transcripts. This step is followed by the construction of an overlap graph, whereby each read represents a node and each edge is used to connect compatible reads. Finally, Cufflinks tries to identify the minimum set of paths (*i.e.* transcripts) that explain such graph.

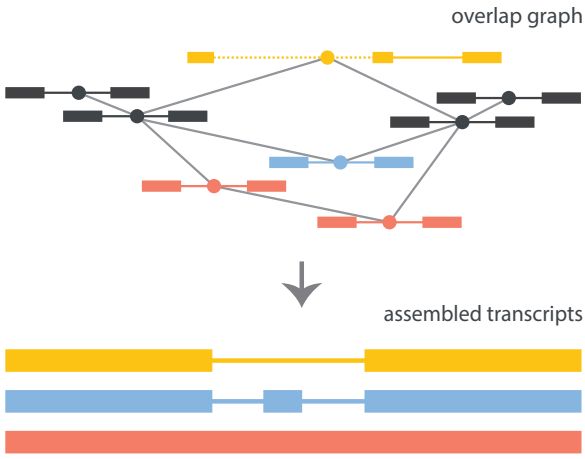


Figure 1.9| Overview of the *de novo* transcript identification algorithm implemented in Cufflinks. Three different incompatible sets of fragments exist in the example depicted here (*i.e.* yellow, blue, red). Black reads represent those that are compatible with any of the sets. Following the construction of an overlap graph that indicates the possible connections amongst the observed fragments, Cufflinks assembles the data into the minimum set of paths required to explain such graph. The identified transcripts can then be used for subsequent downstream analyses, including estimation of transcript expression levels. Adapted from Trapnell et al. [2010].

Alternatively, Cufflinks can also be used in conjunction with the existing annotation (*i.e.* annotation based transcript assembly). In this scenario, the annotated transcripts are used to generate artificial data points that are combined with the observed data during the assembly process, hence serving as a guide. Following transcript assembly, the novelty of the obtained transcripts is evaluated by comparing them to the annotation, and those that differ are reported as novel.

Overall, and similarly to *de novo* read assembly, the task of transcript assembly is not a trivial one, although it becomes simplified by the existence of mappings to the genome. Similarly to the situation encountered in the former scenario, lowly expressed regions are difficult to analyse, given that in those cases the algorithm is less likely to find a unique solution for the constructed graph. Finally, alternative start and end sites also become difficult to characterise, since all the paths are extended to the maximum.

1.3.4 Read count normalisation

Independently of the quantification approach followed, the result from such step is going to be an estimate on the number of reads that can be attributed to a certain feature, further referred to as counts. These counts will be proportional to the expression levels of the feature of interest; however, they will also depend on the length of the feature and the sequencing depth of the experiment (*i.e.* the total number of sequenced reads). In addition, further experimental biases have also been detected to have an impact on the counts detected in certain loci, as it is the case for the previously mentioned sequence-dependent biases. Altogether, these observations illustrate the need for normalisation in order to enable the comparison of read counts across different samples and features.

One of the measures commonly used to report expression levels derived from RNA-seq data is the Reads/Fragments per Kilobase per Million mapped reads (RPKMs or FPKMs, in the case of single-end or paired-end data, respectively) [Mortazavi et al., 2008]:

$$\hat{\mu}_{ij} = \frac{k_{ij}}{N_j l_i} \cdot 10^9$$

where:

$\hat{\mu}_{ij}$ = normalised expression for gene i in sample j
k_{ij} = observed counts for gene i in sample j
N_j = total number of reads in sample j (sequencing depth)
l_i = length for gene i

Given that this measure takes into account both the length of the feature of interest and the total number of mapped reads in the dataset (*i.e.* sequencing depth), it has become established as an intuitive measure of expression levels. However, this method is based on the assumption that the overall RNA levels are similar across samples, and hence it might fail to properly estimate the normalisation factors in cases where the compared libraries differ in their composition [Robinson and Oshlack, 2010]. For example, let us imagine two samples that express a common set of genes at similar levels, and let us consider an extra small set of highly expressed genes in one of them. Since the sequencing step can be understood as a sampling process, where it is more likely to detect reads from genes with high expression levels, the signal from commonly expressed genes will be lower in the latter sample, provided that both are sequenced at a similar depth. Hence, using the above mentioned normalisation method would lead to the interpretation that most genes undergo changes in expression across conditions; whilst the observed differences could be better explained by the isolated differential expression of the few non-overlapping genes (**Figure 1.10**).

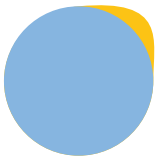
The above described scenario evidences the need for more robust normalisation methods, especially when the goal is to compare across libraries (*e.g.* in downstream analysis such as differential expression/splicing). An example of those methods is the one provided within the DESeq2 Bioconductor package [Love et al., 2014]. Such algorithm starts by calculating a geometric mean for each gene in order to capture the variability of the observed measurements across all the libraries (similar to obtaining a reference sample). Then, these values are used to normalise the initial counts, and finally, the library-specific normalisation factors are obtained from the median of the calculated ratios:

$$s_j = \underset{i:\, k_i^R \neq 0}{median} \frac{k_{ij}}{k_i^R}$$

where:

s_j = size factor for sample j
k_{ij} = observed counts for gene i in sample j
k_i^R = geometric mean for gene i across the m samples: $(\prod_{v=1}^{m} k_{iv})^{1/m}$

Other tools (*e.g.* Cufflinks [Trapnell et al., 2010] and MMSEQ [Turro et al., 2011]) enable also the correction of sequence-dependent biases, by attributing a weight to each position in the expressed loci based on its sequence context. The calculated weights are then used during the abundance inference step in order to model the non-uniform location of reads along the transcripts [Li et al., 2010; Roberts et al., 2011].



	sample A	sample B
gene 1	100	80
gene 2	100	80
...	...	...
gene 100	100	80
gene 101	0	2,000
total number of reads	10,000	10,000

Figure 1.10 | Limitations on the use of RPKMs for differential expression analysis. In this example, sample A and B express a common set of genes at similar levels; however, sample B also contains a highly expressed gene that is not present in the former. If both samples are sequenced at the same depth, the observed counts for the common set of genes will be lower in sample B, given the limited number of reads. In the context of differential expression analysis, the RPKM normalisation method would lead to the interpretation that all genes are differentially expressed, since it assumes homogeneity in library composition. In the scenario represented here such assumption does not hold, and the observed differences are better explained by the isolated differential expression of the gene unique to sample B. This highlights the need for more robust normalisation methods when comparison across libraries is attempted.

1.3.5 Differential expression

One of the most common uses of RNA-seq data is the assessment of differences in expression levels across conditions. Provided the corresponding counts have been obtained, such analysis can be performed both at the gene and transcript levels (differential gene/transcript expression), and one of the most popular tools to achieve that is the Bioconductor package DESeq2 [Love et al., 2014].

In general terms, DESeq2 relies on the use of Generalised Linear Models (GLMs) of the Negative Binomial (NB) family in order to address the significance of the detected changes in expression levels. The implemented analysis workflow first consists of normalising the observed counts in order to enable their comparison across libraries (**Figure 1.11** - step 1), as covered in the previous section. Next, for each gene, an estimate on the amount of variability that can be expected on the measurements from biological replicates is calculated (**Figure 1.11** - step 2), and finally, the differential expression test is performed (**Figure 1.11** - step 3).

As with any counting process, one would not expect the detected counts for a given gene to be exactly the same across all observations from a single condition. Hence, the underlying question in differential expression analysis is whether the counts observed across the two evaluated conditions are similar enough to be derived from the same distribution (null hypothesis), or whether they are better explained by two separate ones (alternative hypothesis). Given the nature of the data obtained from RNA-seq experiments, the Poisson distribution was first proposed to model noise intrinsic to the counting process [Marioni et al., 2008]. However, it was soon shown that while this approach works well for technical replicates, it underestimates the variability in measurements across biological replicates [Anders and Huber, 2010; Robinson et al., 2010]. As a result, the negative binomial distribution has been widely adopted to account for such over-dispersion:

$$k_{ij} = NB(\mu_{ij}, \sigma_{ij}^2)$$
$$\mu_{ij} = s_j q_{ij}$$

where:
- k_{ij} = observed counts for gene i in sample j
- μ_{ij} = distribution mean for gene i in sample j
- σ_{ij}^2 = dispersion for gene i in sample j
- s_j = size factor for sample j
- q_{ij} = quantity proportional to the concentration of cDNA fragments for gene i in sample j

The identification of the amount of variation across biological replicates is an essential step in the aforementioned workflow, since it enables for the evaluation of the significance of any changes detected. However, because of the low number of replicates typically available in RNA-seq experiments, such variation cannot be directly calculated, and needs to be estimated from the data instead. Following the assumption that genes with similar expression levels have similar sample-to-sample variance, DESeq2 obtains gene-specific variance estimates by taking into account not only the observed dispersion for each given gene, but also that of all other genes. This is achieved by fitting a regression curve to the data (*i.e.* average normalised counts *vs.* observed dispersion), which is subsequently used to modify the observed dispersion values.

Finally, by further decomposing the mean into a function of independent variables (*i.e.* the covariates), it is possible to take all known sources of variation into account:

$$log_2(\mu_{ij}) = \sum_r x_{jr} \beta_{ir}$$

where:
- μ_{ij} = mean for gene i in sample j
- x_{jr} = independent variable r for sample j
- β_{ir} = coefficient for gene i and variable r

Altogether, the algorithmic approach behind DESeq2 consists of fitting the model defined in the aforementioned equations for both the null and alternative hypotheses (reduced *vs.* full model, respectively), followed by the evaluation of the significance of the coefficient of interest (**Figure 1.11**).

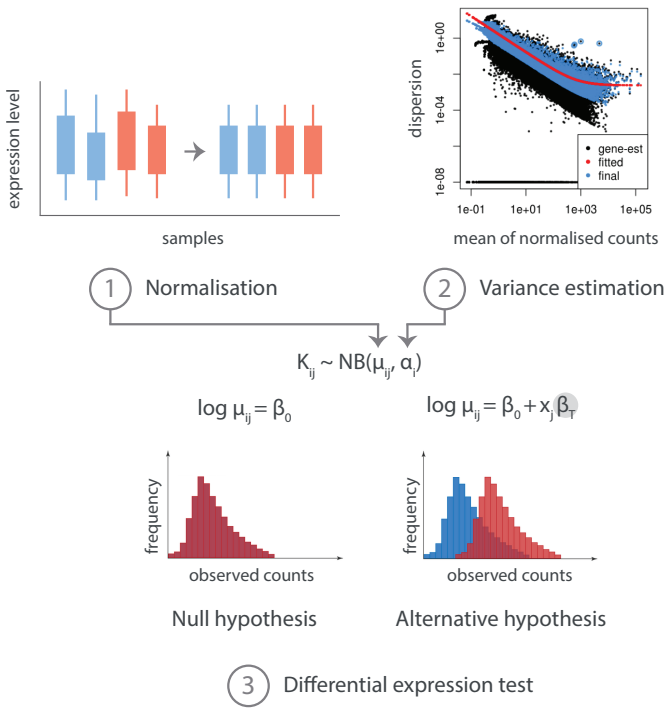


Figure 1.11 | Overview of the steps required for differential expression analysis using DESeq2. First, reads are normalised in order to enable comparison across libraries. Next, for each gene, an estimate of the amount of variability that can be expected across biological replicates is obtained. Given the typical low number of biological replicates in RNA-seq experiments, it is not possible to obtain such information directly from the data. Hence, DESeq2 relies on the observed dispersion from all genes instead (black dots), and by fitting a regression curve that explains the dependence of the dispersion on the mean (red line), further modifies the initial values through a process called shrinkage estimation of variability (blue dots). Finally, the obtained information is used to test the hypothesis that the observed counts originate from different distributions (alternative hypothesis). DESeq2 uses the negative binomial distribution to model both stochastic and biological noise, and it further relies on the use of GLMs to take all known sources of variation into account. Here, the independent variable x_j represents the experimental condition, and can be arbitrarily set to 0 in the case of controls and to 1 in the case of treated samples. Under the alternative hypothesis, this enables the existence of two different distributions that can explain potential differences in expression levels of the studied gene i. Conversely, under the null hypothesis, there is no need for such term, since in this case all counts arise from the same distribution. Hence, the general idea behind the differential expression test consists of deciding whether the inclusion of the variable x_j adds meaningful information to the model, and it translates into assessing the significance of the β_T coefficient (highlighted in grey).

Similarly to DESeq2, further available tools rely on the use of read counts in order to make assessments of differential gene expression (e.g. edgeR [Robinson et al., 2010] and baySeq [Hardcastle and Kelly, 2010]). These tools can also be used to study differential expression at the transcript level, but a common alternative approach consists of relying on the algorithms implemented in the framework of transcript abundance estimation, since those are able to take into account the uncertainty in the read assignment process. For example, this is the case with Cuffdiff2 [Trapnell et al., 2013] and MMDIFF [Turro et al., 2014], which can be executed following estimation of transcript expression levels with Cufflinks [Trapnell et al., 2010] and MMSEQ [Turro et al., 2011], respectively. Moreover, and consistent with the Bayesian model adopted, MMDIFF is also able to make use of the uncertainty in the expression estimates, thus adding further sensitivity to the differential transcript expression analysis.

1.3.6 Differential splicing

In the previous section, I have discussed briefly several approaches for the assessment of differential transcript expression. However, differences in absolute transcript abundance are not necessarily indicative of differential splicing (**Figure 1.12a**), and alternative analysis strategies are preferred when the focus lies on the latter. In general terms, changes in splicing patterns can be assessed through the identification of either differential exon usage (DEU) or differential transcript usage (DTU) events (**Figure 1.12b** and **c**, respectively), with advantages inherent to both approaches. On the one hand, exon-centric analysis strategies are completely independent from isoform reconstruction efforts, thus avoiding the uncertainty intrinsic to that task. Furthermore, while those rely on the existing annotation, such dependence is limited to the exonic coordinates, and this approach still enables the indirect identification of novel transcripts (*i.e.* novel alternatively spliced isoforms). On the other hand, the results from such exon-centric analysis are often difficult to interpret, and in this context transcript-centric analysis strategies emerge as an attractive alternative.

Interestingly, in terms of algorithm development, much of the effort has been focused on the identification of differential transcript expression events, with limited availability of tools for the study of changes in splicing. Amongst those, the Bioconductor package DEXSeq [Anders et al., 2012], was the first tool to account for

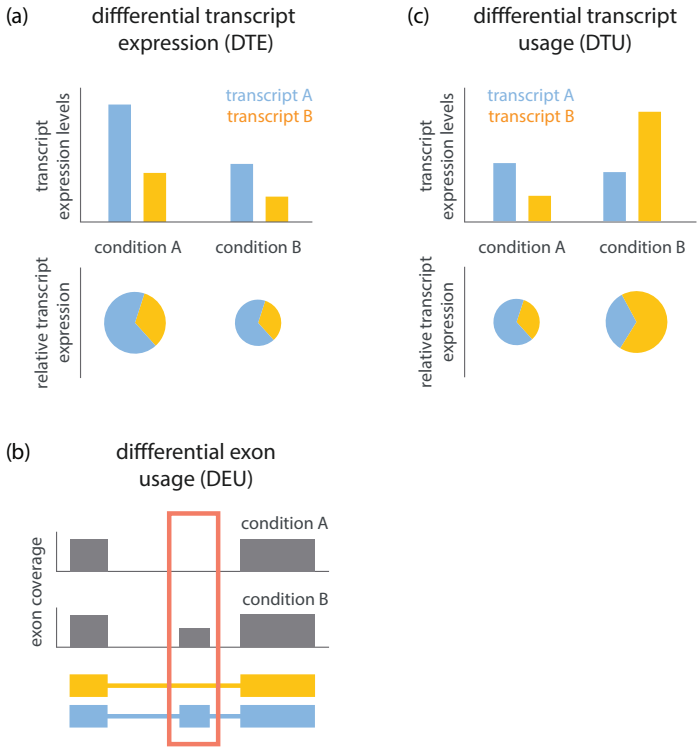


Figure 1.12 | Strategies for the study of changes in the abundance of alternative transcripts.
(a) Differential transcript expression concept. Differential transcript expression is analogous to differential gene expression, and does not necessarily imply differences in splicing.
(b) Differential exon usage concept. Differences in the read coverage for a given exon, relative to changes in the number of reads that overlap the other exons within the same gene, can be used as an indicator of differential splicing.
(c) Differential transcript usage concept. Differential transcript usage refers to those cases where there is a change in the transcript relative abundances, which is not necessarily linked to an overall change in expression levels. It constitutes the most direct strategy for the study of differential splicing.

biological variation in the analysis, a vital requirement for robust testing. Briefly, by relying on the same algorithmic principles as the aforementioned DESeq2, this method enables the identification of significant differences in the proportion of reads that overlap each exon, relative to the total number of reads that overlap the corresponding gene (DEU events; **Figure 1.12b**). On the other hand, the recently introduced tool MMDIFF [Turro et al., 2014] provides a method for the analysis

of DTU events (**Figure 1.12c**). MMDIFF is based on the use of Bayesian mixed models, whereby the uncertainty in transcript expression estimates can be incorporated into the regression models used for testing, thus improving the power to detect the events of interest. Altogether, the scarcity of tools to deconvolute differences in splicing from differences in expression, together with the lack of methods to infer the functional impact of the identified events, evidence that the computational pipelines for the analysis of RNA-seq data are still not completely established.

1.4 Aims of the thesis

The work presented in this thesis focuses on the use of RNA sequencing for the high throughput study of alternative transcript products in human samples. Overall, the goal is to improve the current understanding of splicing by addressing the following questions:

- What is the extent of transcriptome diversity? Are specific alternative transcripts preferentially produced within a given gene?
- How prevalent are changes in splicing patterns in cancer? How can we assess the potential functional impact of such changes?
- How do core spliceosomal factors participate in the regulation of splicing? What are the effects of disrupting such regulation on dynamic cellular processes such as cell division?
- Can the differential splicing events identified from transcriptomics data be recapitulated at the protein level?

The present chapter has provided an general introduction to the two central concepts behind this thesis, *i.e.* the splicing reaction and RNA sequencing. Further introductory remarks relevant to each of the aforementioned questions will be covered in the following chapters.