Leveraging transcript quantification for fast computation of alternative splicing profiles

Gael P. Alamancos[1,*], Amadís Pagès[1,4,*], Juan L. Trincado[1], Nicolás Bellora[2], Eduardo Eyras[1,3,5]

Universitat Pompeu Fabra, E08003, Barcelona, Spain

[2]INIBIOMA, CONICET-UNComahue, Bariloche, Río Negro, Argentina

[3]Catalan Institution for Research and Advanced Studies, E08010 Barcelona, Spain

[4]Centre for Genomic Regulation, E08003, Barcelona, Spain

[5]corresponding author: eduardo.eyras@upf.edu

[*]These authors contributed equally

Abstract

Background

Alternative splicing plays an essential role in many cellular processes and bears major relevance in the understanding of multiple diseases, including cancer. High-throughput RNA sequencing allows genome-wide analyses of splicing across multiple conditions. However, the increasing number of available datasets represents a major challenge in terms of computation time and storage requirements.

Results

Here we describe SUPPA, a computational tool to calculate relative inclusion values (PSI or Ψ) of alternative splicing events, exploiting fast transcript quantification of a given annotation. SUPPA is more accurate than standard methods using simulated as well as real RNA sequencing data compared to experimentally validated events. We assess the variability in terms of the choice of annotation and provide evidence that using complete transcripts rather than more transcripts per gene provides better estimates. Moreover, SUPPA coupled with *de novo* transcript reconstruction methods does not achieve accuracies as high as using quantification of known transcripts, but remains better than existing methods. Finally, we also show that SUPPA is more than 1000 times faster than standard methods

Conclusions

SUPPA efficiently uses transcript quantification to obtain accurate estimates of local alternative splicing event Ψ values. Coupled with a fast transcript quantification method, SUPPA provides Ψ values at a much higher speed than existing methods without compromising accuracy. SUPPA facilitates the systematic splicing analysis of large datasets with limited computational resources. The software is implemented in Python 2.7 and is available under the MIT license at https://bitbucket.org/regulatorygenomicsupf/suppa.

Background

Alternative splicing plays an important role in many cellular processes and bears major relevance in the understanding of multiple diseases, including cancer (David & Manley 2010, Ward & Cooper 2010). Numerous genome wide surveys have facilitated the description of the alternative splicing patterns under multiple cellular conditions and disease states. These are mostly based on the measurement of local variations in the patterns of splicing, encoded as events, which have been studied using microarrays (Thorsen et al. 2008, Lapuk et al. 2010, Misquitta-Ali et al. 2011), RT-PCR platforms (Klinck et al. 2008), or RNA sequencing (Pan et al. 2008, Wang et al. 2008). The description of alternative splicing in terms of events facilitates their experimental validation using PCR methods and the characterization of regulatory mechanisms using sequence analysis and biochemical approaches (Bechara et al. 2013, Raj et al. 2014). Recent results also demonstrate that events are valuable in predictive and therapeutic strategies (Xiong et al. 2014, Hua et al. 2015). Events are generally defined as local variations of the exon-intron structure that can take two configurations, and are characterized by an inclusion level, also termed PSI or Ψ, which measures the fraction of mRNAs expressed from the gene that contain an specific form of the event (Venables et al. 2008, Wang et al. 2008). In terms of sequencing reads, Ψ is defined as the ratio of the density of inclusion reads to the sum of the densities of inclusion and exclusion reads (Wang et al. 2008, Shen et al. 2012). The simplest estimates of Ψ values are based on the junction reads or exon reads (Pan et al. 2008, Sultan et al. 2008, Wang et al. 2008, Cloonan et al. 2009), but other methods have been developed that take into account the uncertainty of quantification from single experiments (Katz et al. 2010), the comparison of two conditions (Katz et al. 2010, Griffith et al. 2010, Shen et al. 2012, Wu et al. 2011, Shi et al. 2013), as well as the handling of multiple replicates per condition (Shen et al. 2012, Brooks et al. 2011, Singh et al. 2011, Hu et al. 2013) and paired-replicates (Shen et al. 2014).

Current tools to process RNA sequencing data to study alternative splicing events can take more than a day to analyze a single sample and often require excessive storage, so they are not competitive to be applied systematically to large data sets, unless access to large computational resources is granted. In particular, methods for estimating Ψ values generally involve the mapping of reads to the genome or to a library of known exon-exon junctions, both of which require considerable time and storage capacity. Additionally, accuracy is often achieved at the cost of computing time. All this represents a major obstacle for the analysis of

large datasets, and in particular, for the re-analysis of public data and updates with new annotations or assembly versions. More importantly, these analyses remain unfeasible at small labs with limited computational resources. On the other hand, recent developments in the quantification of known transcripts have shown that considerable quality can be achieved at high speed (Li et al. 2011, Roberts et al. 2013, Patro et al. 2014, Zhang et al. 2014). These methods provide an estimate of the abundances of mRNA molecules in a given condition for an annotated genome. These developments raise the question of whether fast transcript quantification could be used to obtain accurate estimates of the Ψ values for local alternative splicing events.

In this article we describe SUPPA, a computational tool to leverage fast transcript quantification for rapid estimation of Ψ values directly from the abundances of the transcripts defining each event. Using simulated data we show that the Ψ values estimated by SUPPA, coupled to Sailfish or RSEM transcript quantification, are closer to the ground-truth than two standard methods, MATS and MISO. Additionally, using an experimentally validated set and matched RNA-Seq data we show that SUPPA achieves higher accuracy than MATS and MISO. We further assess the variability in terms of the choice of annotation and provide evidence that using complete transcripts rather than more transcripts per gene in the annotation provide better estimates. Moreover, we show that SUPPA coupled with *de novo* transcript reconstruction methods does not achieve accuracies as high as using the quantification of known transcripts, but remains better than existing methods. Finally, speed benchmarking provides evidence that SUPPA can obtain Ψ values at a much higher speed than existing methods without compromising accuracy. We argue that coupled to a fast transcript quantification method, SUPPA provides a fast and accurate approach to systematic splicing analysis. SUPPA facilitates the accurate splicing analysis of large datasets, making possible for labs with limited computational resources to exploit data from large genomics projects and contribute to the understanding of the role of alternative splicing in cell biology and disease.

Methods

SUPPA framework

SUPPA provides an effective and easy-to-use software to calculate the inclusion levels (Ψ) of alternative splicing events exploiting transcript quantification (Figure 2A). An alternative splicing event is a local representation of a splicing variation of a gene. It is the result of a summary of the contributions from all transcripts from a gene that covers a given region, and is generally represented as a binary choice, although more complex variations may happen. Accordingly, an event can be characterized in terms of the sets of transcripts that describe either form of the event, which can be denoted as F_1 and F_2. For instance, for an exon-skipping event, F_1 represents the transcripts that include the exon, whereas F_2 represents the transcripts that skip the exon. The PSI value of an event is defined as the ratio of the abundance of transcripts that include one form of the event, F_1, over the abundance of the transcripts that contain either form of the event, $F_1 \cup F_2$ (Venables et al. 2008, Wang et al. 2008, Katz et al. 2010, Shen et al. 2012). Given the abundances for all transcripts isoforms, assumed without loss of generality to be given in transcript per million units (TPM) (Li et al. 2010), which we denote as TPM_k, SUPPA then calculates the PSI (Ψ) for an event as follows:

$$\Psi = \sum_{k \in F_1} TPM_k \Big/ \sum_{j \in F_1 \cup F_2} TPM_j \qquad (1)$$

SUPPA reads the information about the events and the quantification of all transcripts to obtain the Ψ values using formula (1) (Figure 1A). SUPPA is agnostic of the actual methodology for quantifying transcripts and can read the quantification from multiple experiments in a single input file. Additionally, SUPPA also includes a function to obtain PSI values for transcript isoforms, defined as the abundance of a transcript normalized to the abundances of all transcripts from the same gene. For instance, for isoform k in gene G:

$$\Psi_{iso_k} = TPM_k \Big/ \sum_{j \in G} TPM_j \qquad (2)$$

SUPPA generates different alternative splicing events types from an input annotation file in GTF format: exon skipping (SE), alternative 5' and 3' splice-sites (A5/A3), mutually exclusive exons (MX), intron retention (RI), and alternative first and last exons (AF/AL)

(Figure 1B). Each event has a unique identifier that includes the gene symbol, the type of event, and the coordinates and strand that characterize the event:

<gene_id>;<event_type>:<seqname>:<coordinates_of_the_event>:strand

where *gene_id*, *seqname* and *strand* are obtained directly from the input annotation in GTF, *seqname* is the field 1 from the GTF file, generally the chromosome. The field *coordinates_of_the_event* is defined differently for each *even_type* (Figure 1B). The Ψ value for an event is calculated with respect to one of the two forms of the event (Figure 1B). Additionally, SUPPA provides in the output file the identifiers for the transcripts that describe either form of the event, from which will calculate the Ψ values. Moreover, SUPPA calculates events from a GTF file independently of the source. For instance, this can be a given annotation or the output from a method for *de novo* transcript reconstruction, or a combination of both.

Read mapping and PSI quantification

The Ensembl annotation (Release 75) (Flicek et al. 2014) and the RefSeq annotation (NM_ and NR_ transcripts) (Pruitt et al. 2014) (assembly hg19) were downloaded in GTF format from the Ensembl FTP server and the UCSC genome browser web server, respectively. All annotations on chromosomes other than autosomes or sex chromosomes were removed. In total, 37,494 genes and 135,521 transcripts were obtained for the Ensembl annotation, while 25,937 genes and 48,566 transcripts were obtained for the RefSeq annotation.

The mapping of reads for the analysis with MATS and MISO were performed with the MATS pipeline (Shen et al. 2012). This pipeline uses TopHat (Trapnell et al. 2009) and an input annotation to map the reads, assigning unmapped reads to de-novo splice junctions when possible, and filtering out those reads that map to more than one genomic position. The same annotation was used for read mapping to the genome or for transcript quantification in the comparisons. The mapping pipeline was run on simulated and real RNA-Seq reads. A total of 45 million 2x50bp paired-end reads were generated using FluxSimulator (Griebel et al. 2012) (parameter file described in the Supplementary Material). RNA-Seq reads from cytosolic fractions of MCF7 and MCF10 cells, as well as from the ESRP1-overexpression (ESRP1) and empty-vector (EV) experiments in MBA-MD-231 cells from (Shen et al. 2012) were also run through this mapping pipeline. Mapped reads for each of the datasets were used with MATs, which reported Ψ_{MATS} values for the different alternative splicing events (Supplementary

Table 1). Similarly, mapped reads in SAM format were converted to BAM format with samtools (Li et al. 2009) and analysed with MISO (Katz et al. 2010) to calculate the sets of Ψ_{MISO} values for each of the datasets (Supplementary Table 2).

SUPPA *generateEvents* operation was run to calculate all alternative splicing events present in each annotation. For instance, for the RefSeq annotation:

```
/pathTo/python-2.7/python  /pathTo/SUPPA/suppa.py  generateEvents  -i
/pathTo/refseq_annotation.gtf -o /pathTo/eventsOutput/ -e SE SS MX RI
```

This produced a total of 16714 and 66577 events from RefSeq and Ensembl, respectively (Supplementary Table 3). Sailfish (Patro et al. 2014) and RSEM (Li et al. 2011) were used to generate the quantification of all transcripts in the Ensembl and RefSeq annotations mentioned above using the same simulated and the real RNA-Seq datasets. The FASTA sequences of the transcripts corresponding to the same annotation as the GTF described earlier, were downloaded and used to generate the Sailfish index, selecting a *k*-mer size of 31 to minimize the number of reads assigned to multiple transcripts. Sailfish was then run using the FASTQ files for each read set using the flag *-l "T=PE:O=>>:S=U"*, and uncorrected and corrected (for sequence composition bias and transcript length) TPMs were calculated (Patro et al. 2014). On the other hand, for RSEM we used its own mapping pipeline (Li et al. 2011). The *psiPerEvent* operation of SUPPA was used to calculate the $\Psi_{Sailfish}$ and Ψ_{RSEM} values from the transcript quantifications obtained by Sailfish and RSEM, respectively, for the alternative splicing events generated above with the *generateEvents* module of SUPPA. The values $\Psi_{Sailfish}$ and Ψ_{RSEM} were calculated using the simulated and real datasets described earlier, MBA-MD-231 (ESRP1 overexpressed and empty-vector) and the MCF7/MCF10 datasets. This process was performed for the Ensembl and RefSeq annotations. The number of events for which SUPPA estimated a $\Psi_{Sailfish}$ or Ψ_{RSEM} values are given in Supplementary Tables 4 and 5. For the purpose of benchmarking, the PSI values obtained from transcript quantification ($\Psi_{Sailfish}$ and Ψ_{RSEM}) produced by SUPPA were compared to those values obtained from MATS (Ψ_{MATS}) and to those obtained from MISO (Ψ_{MISO}). Details of the commands used to run the different analyses are provided in Supplementary Tables 6-10.

Cufflinks analysis

The BAM files from 2 MBA-MD-231 datasets were used to run Cufflinks (Trapnell et al. 2010) in order to generate and quantify transcriptome annotations *de novo*. The mappings

were performed using TopHat and the Ensembl or RefSeq annotations, so that the starting data was exactly the same as for the other mapped-read based methods. A total of 47211 transcripts were predicted and quantified for the ESRP1 dataset, whereas 37.69 transcripts were predicted and quantified for the EV dataset. SUPPA *generateEvents* operation was then run on the GTF annotation generated by Cufflinks to calculate all the exon skipping events. This produced a total of 2566 and 2139 exon skipping events for the ESRP1 and EV datasets, respectively. Finally, SUPPA *psiPerEvent* operation was used to calculate the $\Psi_{Cufflinks}$ values from the transcript quantification obtained by Cufflinks. From these, the ones coinciding with the events from the RT-PCR experiments were used for the benchmarking.

Time benchmarking

All tools were run within the same node of an Oracle Grid Engine cluster. This node has 98Gb of RAM memory and consists of 24 AMD Opteron (1.4 GHz) processors. All tools were run in multi-threaded mode when possible, but time reported is the actual time the process used across all CPUs.

RNA sequencing

MCF7 cells were cultured in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal bovine serum (FBS) and penicillin/streptomycin. MCF10 cells were cultured in DMEM/F12 supplemented with 15 mM hepes buffer, 10% FBS, 10 mg/ml insulin, 25 ng/ml EGF, 100 ng/ml choleratoxin and 0.5 mg/ml hydrocortisone. Cell fractionation was performed with PARIS Kit from Ambion following their instructions. Two biological replicates from each sample were sequenced with Illumina HiSeq-2000 using a standard stranded protocol. Sequencing reads are available at SRA (SRP045592). We obtained an approximate amount of 42M and 46M paired-end reads for the two cytosolic MCF7 replicates, and 45M and 56M paired-end reads for the two cytosolic MCF10 replicates.

Results

Accuracy analysis with simulated data

Transcript abundances and corresponding paired-end reads were simulated using FluxSimulator (Griebel et al. 2012) with the RefSeq annotation as reference (Methods). The

reference set for accuracy analysis was built using events in genes that only have two transcripts in the RefSeq annotation and did not overlap other events. In these cases, the Ψ of the event is identical to the relative abundance of one of the two transcripts. The ground-truth Ψ values were then defined to be the relative abundances of the transcripts isoforms in these genes, where the transcript abundances were taken to be the simulated abundances. Simulated RNA-Seq reads were mapped (Methods) and used to calculate Ψ_{MISO} and Ψ_{MATS} values with MISO (Katz et al. 2010) and MATS (Shen et al. 2011), respectively. The same simulated reads were also used to quantify transcript abundances with Sailfish (Patro et al. 2014) and RSEM (Li et al. 2011), and $\Psi_{Sailfish}$ and Ψ_{RSEM} values were then calculated with SUPPA. Only genes with a total transcripts per million (TPM) abundance, calculated as the sum of the TPM of its transcripts, larger than 1 were considered. This resulted in a set of 144 events (Supplementary Data 1). Comparing the four estimated Ψ values with the ground-truth, the $\Psi_{Sailfish}$ and Ψ_{RSEM} values calculated with SUPPA show the highest correlations (Table 1) (Figure 2A). Calculating how different the estimated Ψ values are from the ground-truth, SUPPA Ψ values ($\Psi_{Sailfish}$ and Ψ_{RSEM}) show the closest behaviour, followed by MISO and MATS, which behave similarly (Figure 2B).

	Sailfish+SUPPA		**RSEM+SUPPA**		**MATS**		**MISO**	
	Pearson	*Spearman*	*Pearson*	*Spearman*	*Pearson*	*Spearman*	*Pearson*	*Spearman*
Synthetic	0,971	0,959	0,987	0,978	0,833	0.819	0,862	0,815

Table 1. Correlation values (Spearman and Pearson) between the estimated and ground-truth Ψ values using simulated data (Methods). The comparison involves 144 events, which can be found in Supplementary Data 1.

Accuracy analysis with experimentally validated events

To further validate the calculation of Ψ values with SUPPA, we used a set of 163 alternative splicing events validated by RT-PCR MDA-MB-231 cells under two conditions: with overexpression of the splicing factor ESRP1 and with an empty vector (EV) (Shen et al. 2012). Additionally, we used the RNA-Seq data obtained for the same samples (Shen et al. 2012). From both RNA-Seq samples we quantified the RefSeq transcripts using Sailfish and RSEM, and calculated the SUPPA $\Psi_{Sailfish}$ and Ψ_{RSEM} values. As before, the RNA-Seq reads were mapped to the genome to run MISO and MATS (Methods). From the 163 events, we finally compared those 60 that were present in the RefSeq annotation and for which we had Ψ

values for all methods (Supplementary Data 2). SUPPA and MISO show the best correlations with the experimental Ψ values, with similar performance for Sailfish+SUPPA and RSEM+SUPPA. The performances of each method were also similar performances for both the ESRP1 and EV datasets (Table 2) (Figure 3A). Calculating, as before, the absolute difference between the estimated and the experimental Ψ values for each event, we observe that SUPPA, either combined with Sailfish or RSEM, is more accurate than MISO and MATS (Figure 3B). Performing the same analysis using the Ensembl annotation we observe a general decrease of accuracy in all methods (Supplementary Figure 1) (Supplementary Table 11) and MISO performs better than the other methods.

	Sailfish+SUPPA		**RSEM+SUPPA**		**MATS**		**MISO**	
	Pearson	*Spearman*	*Pearson*	*Spearman*	*Pearson*	*Spearman*	*Pearson*	*Spearman*
ESRP1	0,778	0,769	0,795	0,779	0,556	0,539	0,701	0,715
EV	0,766	0,766	0,808	0,823	0,670	0,646	0,765	0,781

Table 2. Correlation values (Spearman and Pearson R) between the estimated Ψ values from ESRP1-overexpressed (ESRP) and empty-vector (EV) RNA-Seq datasets and the RT-PCR validation for the same samples (Shen et al. 2012). This comparison involves the 60 events that were in the RefSeq annotation and had a Ψ value from every method, which can be found in Supplementary Data 2.

Variability associated to replicates and annotation choice

The annotation choice seems to have in general an impact on the accuracy of Ψ estimates. To further study the dependency with the choice of annotation, we obtained RNA from two biological replicates of cytosolic fractions from MCF7 and MCF10 cells and performed sequencing using standard protocols (Methods). Correlation between replicates of the SUPPA Ψ values, using quantification with Sailfish on the RefSeq annotation, is high in all comparisons (Person R ~0.86-0.89) (Supplementary Figure 2). Restricting this analysis to genes with TPM>1, calculated as the sum of the TPMs from all transcripts in each gene, the correlation between replicates increases (Pearson R~0.95-0.97) (Supplementary Figure 2).

We then compared the results obtained using SUPPA for Sailfish quantifications on the RefSeq and Ensembl transcripts, using the set of events that are common to both annotation, and not overlapping with other events. SUPPA Ψ values were calculated using both replicas of the cytosolic MCF7 RNA-Seq data (similar results were observed for MCF10, data not

shown). The comparisons were performed for the 9301 (MCF7, replica 1) and 9287 (MCF7, replica 2) events that were found in both annotations. We observe a variability that does not depend on the difference between annotations in the number of transcripts used for the Ψ calculation (Figure 4A). Similarly, this variation is also independent of the difference in the number of transcripts annotated in the gene in which the event is contained (Supplementary Figure 3). Moreover, the variation in Ψ estimates is also independent of the mean expression estimate of gene in which the event is contained (Figure 4B). On the other hand, the dispersion of Ψ estimates comparing replicates decreases with mean expression of the gene (Figure 4C) and the dispersion for Ψ estimates as a result of differences in annotation is comparable to the dispersion found between replicates for lowly expressed genes.

Annotation-free estimation

The previous analyses suggest that incomplete annotations may lead to inaccurate transcript quantification, which will have in turn a negative impact on the Ψ estimates by SUPPA. The discovery of new transcripts and the completion of existing ones have been facilitated by the development of *de novo* transcript reconstruction and quantification methods, which recover the exon-intron structures from RNA-Seq reads mapped to the genome and estimate their abundances using different approaches (Trapnell et al. 2010, Li et al. 2011b, Li et al. 2011c, Li et al. 2012, Mezlini et al. 2012, Behr et al. 2013, Tomescu et al. 2013, Rossell et al. 2014, Maretty et al. 2014). As these methods produce an annotation of transcripts with their corresponding abundances, their output can be used with SUPPA to calculate alternative splicing events and their Ψ values. Thus, they provide an opportunity to assess whether a *de novo* prediction of transcripts structures and subsequent quantification from RNA-Seq data may lead to more accurate Ψ values than using a fixed annotation. To test this, we run Cufflinks with the *de novo* options with RNA-Seq data for the ESRP1 and EV samples (Methods). Using the resulting annotation, we calculated all possible alternative splicing events and their contributing transcripts with SUPPA. We then compared the Ψ values obtained from the *de novo* Cufflinks quantification with common events in the experimentally validated set (Shen et al. 2012): 83 for ESRP1 and 47 for EV (Supplementary Data 3). We observe that the correlation of $\Psi_{Cufflinks}$ values obtained with SUPPA (Figure 5) (Table 3) is slightly worse than the previously obtained $\Psi_{Sailfish}$ and Ψ_{RSEM} values (Figure 3) (Table 2), but better than the values obtained using the Ensembl annotation (Supplementary Figure 1)

(Supplementary Table 1).

	Cufflinks+SUPPA		MATS	
	Pearson	*Spearman*	*Pearson*	*Spearman*
ESRP1	0,613	0,627	0,467	0,457
EV	0,659	0,650	0,579	0,561

Table 3. Correlation values (Spearman and Pearson R) between the estimated Ψ values from ESRP1-overexpressed (ESRP) and empty-vector (EV) RNA-Seq datasets and the RT-PCR validation for the same samples (Shen et al. 2012). This comparison involves 83 events in the ESRP1 sample and 47 in the EV sample, which can be found in Supplementary Data 3.

Speed benchmarking

The time needed by each methodology to obtain the Ψ values from a FastQ file depends on multiple different steps. To make a comparative assessment of computation times we therefore broke down the benchmarking into three different tasks, equivalent to the three necessary steps for the SUPPA analysis. The first step involves the calculation of alternative splicing events from an annotation file, which only needs to be carried once for a give annotation. To calculate 66577 alternative splicing events from the Ensembl 75 annotation (37494 genes, 135521 transcripts), SUPPA *generateEvents* took 20 minutes, whereas to calculate 16714 alternative splicing events from the RefSeq annotation (25937 genes, 48566 transcripts), it took 3 minutes.

The second step consists in the assignment of reads to transcripts and/or genomic positions. For the purpose of speed benchmarking of read-assignment to transcripts, we considered the transcript quantification by Sailfish or RSEM to be equivalent the read mapping for MISO and MATS. To perform the speed comparison we used the synthetic data (45 millions of paired-end reads) and both (ESRP1 and EV) RNA-Seq samples from the MDA-MB-231 cells pooled together (256 millions of single-end reads). Sailfish and STAR (Dobin et al. 2012) are the fastest to assign reads to their likely molecular sources, compared to TopHat and RSEM (Figure 6A). However, in this step, Sailfish also produces transcript abundance estimations in this step, whereas STAR output is still to be further processed.

The third and final step is the Ψ calculation from either transcript quantification (SUPPA) or from the mapped reads (MISO and MATS). SUPPA *psiPerEvent* operation took less than a minute to produce an output size of 1Mb for 16714 events and was >1000 times faster than

MISO and MATS on the same datasets (Figure 6B). In total, the total time from the raw reads in FastQ format to the Ψ values for Sailfish + SUPPA against the RefSeq annotation-derived events took 214 and 4022 seconds for the synthetic and the MDA-MB-231 samples, respectively. In summary, when used in conjunction with Sailfish, SUPPA is much faster than MISO and MATS, even if an ultra-fast aligner such as STAR (Dobin et al. 2012) is used.

Discussion

We have described SUPPA, a tool to calculate alternative splicing events from a given annotation and to estimate their Ψ values from the quantification of the transcripts that define the events. Using synthetic data and experimentally tested events, we have shown that SUPPA accuracy is generally higher than other frequently used methods. More importantly, SUPPA can obtain Ψ values at a much higher speed without compromising accuracy. Moreover, SUPPA needs very little configuration, a small number of command lines for preprocessing and running and has no dependencies on Python libraries.

RNA-Seq data presents a number of systematic biases that need correction for accurate transcript quantification (Hansen et al. 2010, Li et al. 2010, Roberts et al. 2011). In our analyses we did not observe much influence in the accuracy of SUPPA using corrected or uncorrected transcript quantification with Sailfish (data not shown). In fact, previous reports have already indicated that bias correction in RNA-Seq data does not influence much the estimation of Ψ values (Shen et al. 2012, Zhao et al. 2013). On the other hand, we did observe that there is variability in the estimation of Ψ values associated to the choice of annotation. In the comparison to experimental data, using Ensembl annotation provides slightly worse accuracy than using RefSeq annotation with SUPPA. Moreover, the observed variability between annotations does not depend on the difference in the number of transcripts per gene, on the number of transcripts used to describe the events, or the expression of the gene in which the event is contained. On the other hand, the variability is comparable to the expected variability for lowly expressed genes between biological replicates. Such variability is in fact also frequently observed in transcript quantification methods (Patro et al. 2014, Maretty et al. 2014). It should be noted that RefSeq annotation includes less transcripts per gene than Ensembl, but these transcripts are mostly full-length mRNAs. In particular, RefSeq transcripts generally have annotated untraslated regions, which generally hold a large contribution for the reads coming from a transcript, whereas a large proportion of Ensembl transcripts may be

incomplete. These facts, together with out results, suggest that the completeness of the transcript structures used, rather than the number of transcripts in genes, is of importance for an accurate estimate of transcript abundances, and consequently, for the accurate estimate of event Ψ values.

At the moment SUPPA generates 7 types of events: exon skipping, alternative 5' and 3' splice-sites, mutually exclusive exons, intron retention and alternative first and last exons (Wang et al. 2008). However, other more complex variations may take place. SUPPA model could be expanded from a binary to a ternary, or even further, description of events. However, these complex events are not always easy to test experimentally. On the other hand, the complexity may not always have to do with the number of possible conformations, but rather with a binary change that cannot be easily described in terms of one or two boundary changes, as observed recently for the gene QKI in lung adenocarcinomas (Sebestyen et al. 2015). We argue that a large proportion of the relevant splicing variation can be encapsulated with the binary events described by SUPPA and that more complex variations may be better described using transcript isoform changes (Sebestyen et al. 2015). Although SUPPA is limited to the splicing events available in the gene annotation, events can be expanded with novel transcript variants obtained by other means, like for instance, from a *de novo* transcript reconstruction and quantification method. We observed in this case accuracies similar to the test performed with the Ensembl annotations but lower than when performed with the RefSeq annotation.

As transcript quantification methods improve in accuracy and methods for RNA sequencing increase their efficiency and reliability, our knowledge of the census of RNA molecules in cells will keep on progressing. Although single molecule sequencing methods may eventually lead to the abandonment of transcript reconstruction methods, they are still costly and error prone, and quantification still relies on short read sequencing. Transcript quantification methods will therefore continue to be an essential component in the description of the RNA content in cells. However, fast reliable methods still depends on the annotation. Thus future efforts should focus on improving transcript annotations under multiple conditions, which ultimately will be essential for the correct description of the RNA content in cells and for the study of differential expression across conditions. In parallel to these advances, the local description of alternative splicing will remain a powerful representation of RNA variability in genes, as it facilitates the study of RNA regulation (Bechara et al. 2013, Raj et al. 2014) and has proven extremely valuable in predictive and therapeutic strategies (Xiong et al. 2014, Hua et al. 2015).

Finally, we have shown that when coupled to a fast transcript quantification method, SUPPA outperforms other methods in speed, without compromising the accuracy. This is of special relevance when analyzing large amount of samples. Accordingly, SUPPA facilitates the systematic analyses of alternative splicing in the context of large-scale projects using limited computational resources. We conclude that SUPPA provides a method to leverage fast transcript quantification for efficient and accurate alternative splicing analysis for a large number of samples.

Supplementary Materials

The three supplementary data files with the alternative splicing events tested can be found at https://bitbucket.org/regulatorygenomicsupf/suppa/downloads/Supplementary_Data.zip

Author contributions

EE and NB designed and supervised the study; GPA wrote the software with contributions from JLT and input from EE and NB; AP carried out the benchmarking analyses with contributions from JLT. EE and AP wrote the manuscript with input from the rest of the authors. All authors read and approved the final manuscript.

Acknowledgements

The authors acknowledge useful discussions with S. Mount, S. Janga, Y. Barash, M. Robinson and Y. Xing. This work was supported by grants from Plan Nacional I+D (BIO2011-23920) and Consolider (CSD2009-00080) grants from the Spanish Government, and by the Sandra Ibarra Foundation for Cancer (FSI 2013). GPA was partly funded by the Spanish National Institute of Bioinformatics (INB).

References

Bechara EG, Sebestyén E, Bernardis I, Eyras E, Valcárcel J. RBM5, 6, and 10 differentially regulate NUMB alternative splicing to control cancer cell proliferation. Mol Cell. 2013 Dec 12;52(5):720-33.

Behr J, Kahles A, Zhong Y, Sreedharan VT, Drewe P, Rätsch G (2013). MITIE: Simultaneous RNA-Seq- based transcript identification and quantification in multiple samples. *Bioinformatics* 2013, 29:2529–2538.

Brooks AN, Yang L, Duff MO, Hansen KD, Park JW, Dudoit S, Brenner SE, Graveley BR. (2011). Conservation of an RNA regulatory map between Drosophila and mammals. Genome Res. 21(2):193-202.

Cloonan, N., Forrest, A. R. R., Kolle, G., Gardiner, B. B. A., Faulkner, G. J., Brown, M. K., Taylor, D. F., Steptoe, A. L., Wani, S., Bethel, G., Robertson, A. J., Perkins, A. C., Bruce, S. J., Lee, C. C., Ranade, S. S., Peckham, H. E., Manning, J. M., Mckernan, K. J., and

David CJ, Manley JL.(2010) Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. Genes Dev. 24(21):2343-64.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan 1;29(1):15-21.

Flicek P, et al. Ensembl 2014. Nucleic Acids Res. 2014 Jan;42(Database issue):D749-55.

Griebel T, Zacher B, Ribeca P, Raineri E, Lacroix V, Guigó R, Sammeth M. Modelling and simulating generic RNA-Seq experiments with the flux simulator. Nucleic Acids Res. 2012 Nov 1;40(20):10073-83.

Griffith M, Griffith OL, Mwenifumbo J, Goya R, Morrissy AS, Morin RD, Corbett R, Tang MJ, Hou YC, Pugh TJ, Robertson G, Chittaranjan S, Ally A, Asano JK, Chan SY, Li HI, McDonald H, Teague K, Zhao Y, Zeng T, Delaney A, Hirst M, Morin GB, Jones SJ, Tai IT,

Marra MA. (2010). Alternative expression analysis by RNA sequencing. Nat Methods 7(10):843-7.

Hansen KD, Brenner SE, Dudoit S (2010) Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res 38(12):e131.

Hu Y, Huang Y, Du Y, Orellana CF, Singh D, Johnson AR, Monroy A, Kuan PF, Hammond SM, Makowski L, Randell SH, Chiang DY, Hayes DN, Jones C, Liu Y, Prins JF, Liu J. (2013). DiffSplice: the genome-wide detection of differential splicing events with RNA-seq. Nucleic Acids Res. 41(2):e39.

Hua Y, Liu YH, Sahashi K, Rigo F, Bennett CF, Krainer AR. Motor neuron cell-nonautonomous rescue of spinal muscular atrophy phenotypes in mild and severe transgenic mouse models. Genes Dev. 2015 Jan 12.

Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods. 2010 Dec;7(12):1009-15.

Klinck, R., Bramard, A., Inkel, L., Dufresne-Martin, G., Gervais-Bird, J., Madden, R., Paquet, E.R., Koh, C., Venables, J.P., Prinos, P., Jilaveanu-Pelmus, M., Wellinger, R., Rancourt, C., Chabot, B., Elela, S.A. (2008) Multiple alternative splicing markers for ovarian cancer. *Cancer Res.,* **68**, 657-63.

Lapuk, A., Marr, H., Jakkula, L., Pedro, H., Bhattacharya, S., Purdom, E., Hu, Z., Simpson, K., Pachter, L., Durinck, S., Wang, N., Parvin, B., Fontenay, G., Speed, T., Garbe, J., Stampfer, M., Bayandorian, H., Dorton, S., Clark, T.A., Schweitzer, A., Wyrobek, A., Feiler, H., Spellman, P., Conboy, J., Gray, J.W. (2010) Exon-level microarray analyses identify alternative splicing programs in breast cancer. *Mol. Cancer Res.,* **8**, 961-74.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9.

Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. RNA-Seq gene expression estimation

with read mapping uncertainty. Bioinformatics. 2010 Feb 15;26(4):493-500.

Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011 Aug 4;12:323.

Li W, Feng J, Jiang T (2011b). IsoLasso: a LASSO regression approach to RNA-Seq based transcrip- tome assembly. *J Comput Biol* 2011, 18:1693–1707.

Li JJ, Jiang C-R, Brown JB, Huang H, Bickel PJ (2011c): Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation. *Proc Nat Acad Sci USA* 2011, 108:19867–19872.

Li W, Jiang T: Transcriptome assembly and isoform expression level estimation from biased RNA-seq reads. *Bioinformatics* 2012, 28:2914–2921.

Maretty L, Sibbesen J, Krogh A. (2014). Bayesian transcriptome assembly. Genome Biol. 15(10):501.

Mezlini AM, Smith EJ, Fiume M, Buske O, Savich G, Shah S, Aparicion S, Chiang D, Goldenberg A, Brudno M: (2012). iReckon: Simultaneous isoform discovery and abundance estimation from RNA- seq data. *Genome Res* 2012, 23:519–529.

Misquitta-Ali CM, et al. (2011). Global profiling and molecular characterization of alternative splicing events misregulated in lung cancer. *Mol Cell Biol.* **31**(1):138-50.

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet. 2008 Dec;40(12):1413-5.

Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat Biotechnol. 2014 May;32(5):462-4.

Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, Farrell CM, Hart J, Landrum MJ, McGarvey KM, Murphy MR, O'Leary NA, Pujar S, Rajput B, Rangwala

SH, Riddick LD, Shkeda A, Sun H, Tamez P, Tully RE, Wallin C, Webb D, Weber J, Wu W, DiCuccio M, Kitts P, Maglott DR, Murphy TD, Ostell JM. RefSeq: an update on mammalian reference sequences. Nucleic Acids Res. 2014 Jan;42(Database issue):D756-63.

Raj B, Irimia M, Braunschweig U, Sterne-Weiler T, O'Hanlon D, Lin ZY, Chen GI, Easton LE, Ule J, Gingras AC, Eyras E, Blencowe BJ. A global regulatory mechanism for activating an exon network required for neurogenesis. Mol Cell. 2014 Oct 2;56(1):90-103.

Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol. 2011;12(3):R22.

Roberts A, Pachter L. Streaming fragment assignment for real-time analysis of sequencing experiments. Nat Methods. 2013 Jan;10(1):71-3.

Rossell, D., Attolini, C. S. O., Kroiss, M., & Stöcker, A. (2014). Quantifying alternative splicing from paired-end RNA-sequencing data. *The annals of applied statistics*, *8*(1), 309-330.

Sebestyén E, Zawisza M, Eyras E. Detection of recurrent alternative splicing switches in tumor samples reveals novel signatures of cancer. Nucleic Acids Res. 2015 Jan 10. doi: 10.1093/nar/gku1392

Shen S, Park JW, Huang J, Dittmar KA, Lu ZX, Zhou Q, Carstens RP, Xing Y. MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. Nucleic Acids Res. 2012 Apr;40(8):e61.

Shen S, Park JW, Lu ZX, Lin L, Henry MD, Wu YN, Zhou Q, Xing Y. rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc Natl Acad Sci U S A. 2014 Dec 5. pii: 201419161.

Shi Y, Jiang H. rSeqDiff: detecting differential isoform expression from RNA-Seq data using hierarchical likelihood ratio test. PLoS One. 2013 Nov 18;8(11):e79448.

Singh D, Orellana CF, Hu Y, Jones CD, Liu Y, Chiang DY, Liu J, Prins JF. (2011). FDM: a

graph-based statistical method to detect differential transcription using RNA-seq data. Bioinformatics 27(19):2633-40.

Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, Schmidt D, O'Keeffe S, Haas S, Vingron M, Lehrach H, Yaspo ML. (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science 321(5891):956-60.

Thorsen, K., Sørensen, K.D., Brems-Eskildsen, A.S., Modin, C., Gaustadnes, M., Hein, A.M., Kruhøffer, M., Laurberg, S., Borre, M., Wang, K., Brunak, S., Krainer, A.R., Tørring, N., Dyrskjøt, L., Andersen, C.L., ørntoft, T.F. (2008) Alternative splicing in colon, bladder, and prostate cancer identified by exon array analysis. *Mol. Cell. Proteomics*, **7**, 1214-24.

Tomescu AI, Kuosmanen A, Rizzi R, Mäkinen V, Veli M (2013). A novel min-cost flow method for estimating transcript expression with RNA-Seq. *BMC Bioinformatics* 14:S15.

Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009 May 1;25(9):1105-11.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 28(5):511-5.

Venables JP, Klinck R, Bramard A, Inkel L, Dufresne-Martin G, Koh C, Gervais-Bird J, Lapointe E, Froehlich U, Durand M, Gendron D, Brosseau JP, Thibault P, Lucier JF, Tremblay K, Prinos P, Wellinger RJ, Chabot B, Rancourt C, Elela SA. Identification of alternative splicing markers for breast cancer. Cancer Res. 2008 Nov 15;68(22):9525-31.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008 Nov 27;456(7221):470-6.

Ward, A. and Cooper,T. (2010) The pathobiology of splicing. J. Pathol., 220, 152–163.

Wu J, Akerman M, Sun S, McCombie WR, Krainer AR, Zhang MQ. SpliceTrap: a method to quantify alternative splicing under single cellular conditions. Bioinformatics. 2011 Nov 1;27(21):3010-6.

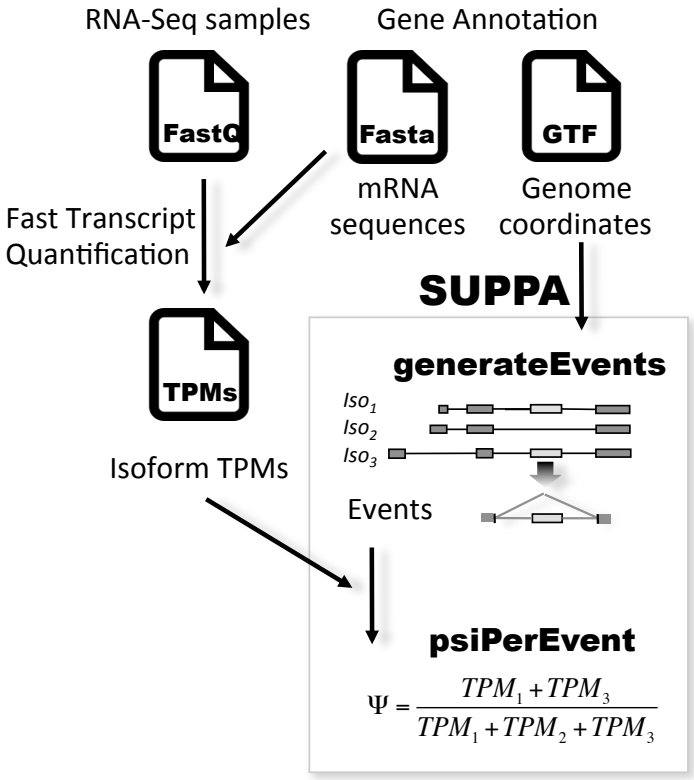
Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RK, Hua Y, Gueroussov S, Najafabadi HS, Hughes TR, Morris Q, Barash Y, Krainer AR, Jojic N, Scherer SW, Blencowe BJ, Frey BJ. RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. Science. 2015 Jan 9;347(6218):1254806.

Zhang Z & Wang W. (2014) RNA-Skim: a rapid method for RNA-Seq quantification at transcript level. *Bioinformatics* **30**(12):i283-i292.

Zhao K, Lu ZX, Park JW, Zhou Q, Xing Y (2013) GLiMMPS: Robust statistical model for regulatory variation of alternative splicing using RNA-Seq data. Genome Biol 14(7):R74.

Figure 1

(A)


RNA-Seq samples
Gene Annotation
FastQ
Fasta
GTF
mRNA sequences
Genome coordinates
Fast Transcript Quantification
SUPPA
TPMs
generateEvents
Iso_1
Iso_2
Iso_3
Events
Isoform TPMs
psiPerEvent
$\Psi = \frac{TPM_1 + TPM_3}{TPM_1 + TPM_2 + TPM_3}$

(B)

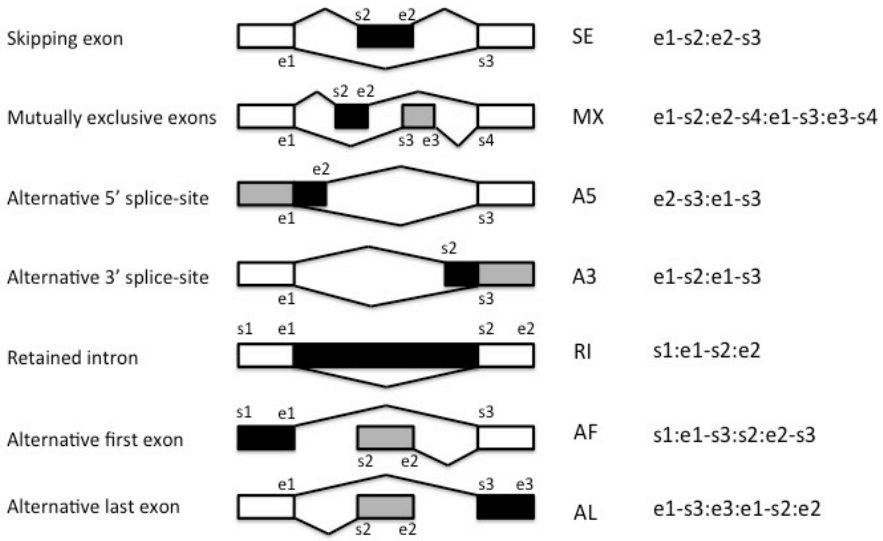

Skipping exon
SE
e1-s2:e2-s3
Mutually exclusive exons
MX
e1-s2:e2-s4:e1-s3:e3-s4
Alternative 5' splice-site
A5
e2-s3:e1-s3
Alternative 3' splice-site
A3
e1-s2:e1-s3
Retained intron
RI
s1:e1-s2:e2
Alternative first exon
AF
s1:e1-s3:s2:e2-s3
Alternative last exon
AL
e1-s3:e3:e1-s2:e2

Figure 1. SUPPA pipeline. **(A)** SUPPA calculates possible alternative splicing events from the annotation using the operation *generateEvents*. The annotation can be from a database or built from RNA-Seq data using a transcript reconstruction method. For each event, the transcripts contributing to either form of the event are store and the calculation of the Ψ value per sample for each event is performed using the transcript abundances per sample (TPMs) (Methods). From one or more transcript quantification files, SUPPA calculates for each event the Ψ value per sample with the operation *psiPerEvent*. SUPPA can use transcript quantification values obtained from any method. **(B)** Events generated from the annotation are given an identifier that includes a code for the event type (SE, MX, A5, A3, RI, AF, AL) and a set of start (s) and end (e) coordinates that define the event (shown in the figure) (Methods). In the figure, the form of the alternative splicing event that includes the region in black is the one for which the relative inclusion level (Ψ) is given: for SE, the PSI indicates the inclusion of the middle exon; for A5/A3, the form that minimizes the intron length; for MX, the form that contains the alternative exon with the smallest start coordinate (the left-most exon) regardless of strand; for RI, the form that retains the intron; and for AF/AL, the form that maximizes the intron length. The gray area denotes the alternative form of the event. For instance, for RI the inclusion level is given for the form that retains the intron.

Figure 2

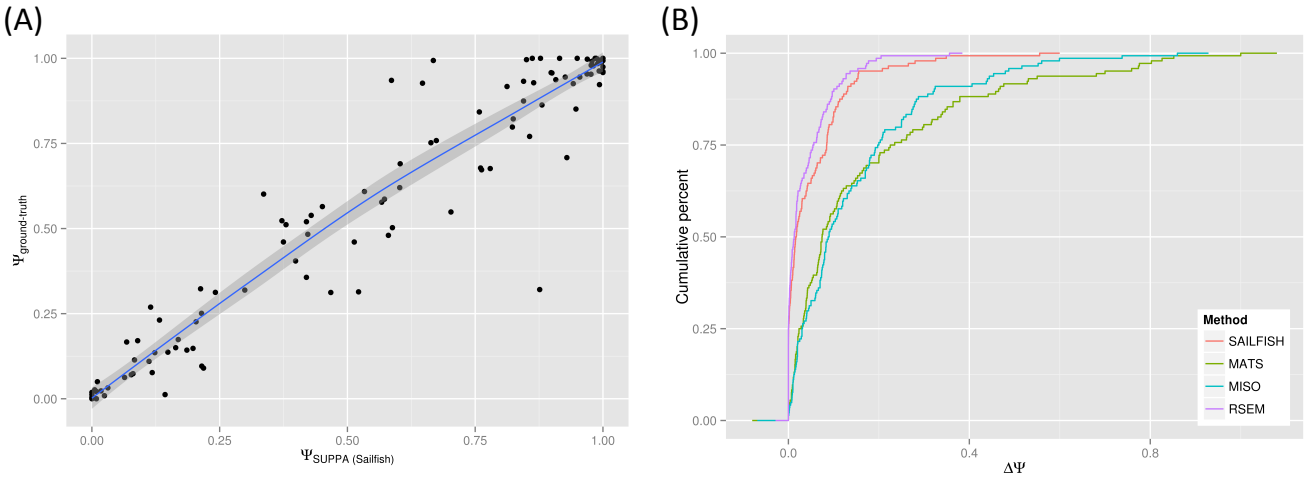


Figure 2. Benchmarking with simulated data. (A) Correlation of the ground-truth Ψ values (Methods) with those estimated with Sailfish+SUPPA using simulated data. The blue line and gray boundaries are the fitted curves with the LOESS regression method. **(B)** Cumulative distribution of the absolute difference between the ground-truth Ψ values and the ones estimated with Sailfish+SUPPA (SAILFISH), RSEM+SUPPA (RSEM), MISO and MATS. The lines describe the proportion of all events tested (Cumulative percent, y-axis) that are predicted at a given maximum absolute difference from the ground-truth value (**ΔΨ**, x-axis).

Figure 3

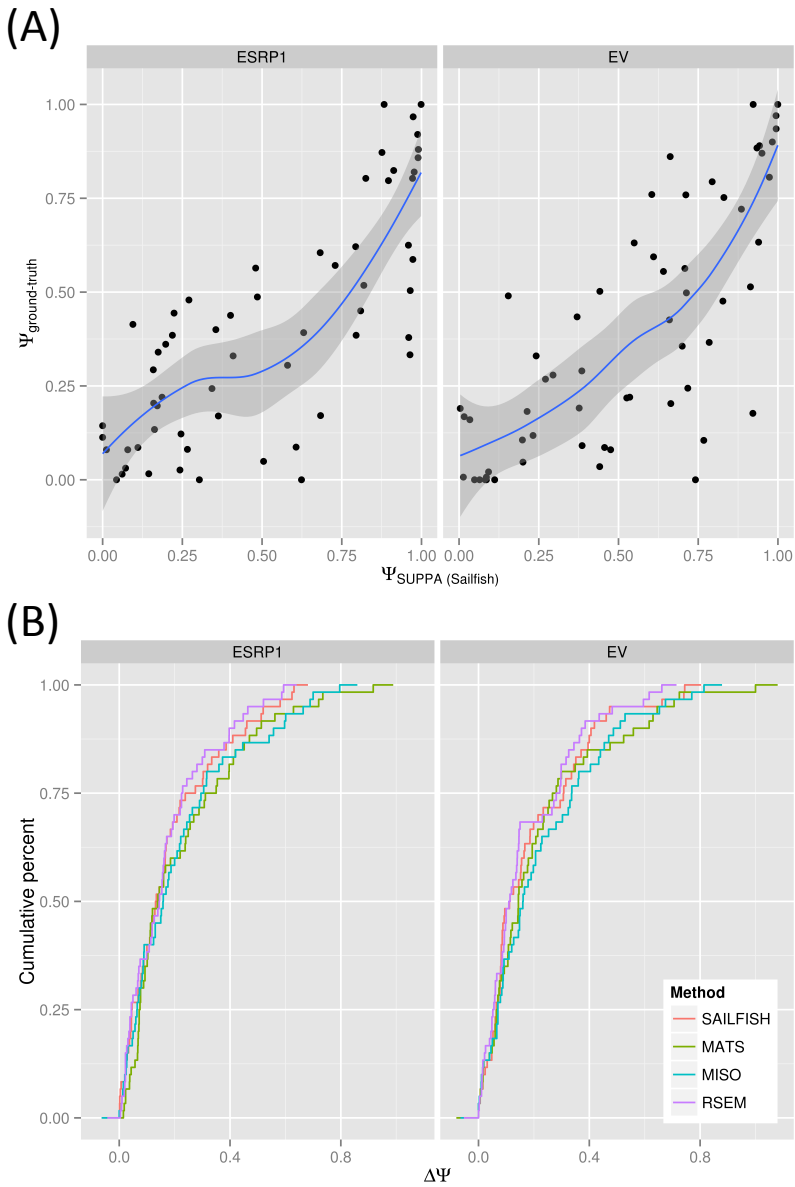


Figure 3. Benchmarking experimentally validated events. (A) Correlation of the experimental Ψ values with those estimated with Sailfish+SUPPA in MDA-MB-231 cells with (ESRP1, left panel) and without (EV, right panel) ESRP1 overexpression. Experimental Ψ values were obtained by RT-PCR (Shen et al. 2012) and estimated PSIs were obtained from RNA-Seq data from the same samples (Shen et al. 2012). The blue line and gray boundaries are the fitted curves with the LOESS regression method. **(B)** Cumulative distribution of the absolute difference between the same experimental Ψ values and the ones estimated with Sailfish+SUPPA (SAILFISH), RSEM+SUPPA (RSEM), MISO and MATS from RNA-Seq data from the same samples (Shen et al. 2012). The lines describe the proportion of all events (Cumulative percent, y-axis) that are calculated at a given maximum absolute difference from the RT-PCR value (**ΔΨ**, x-axis).

Figure 4

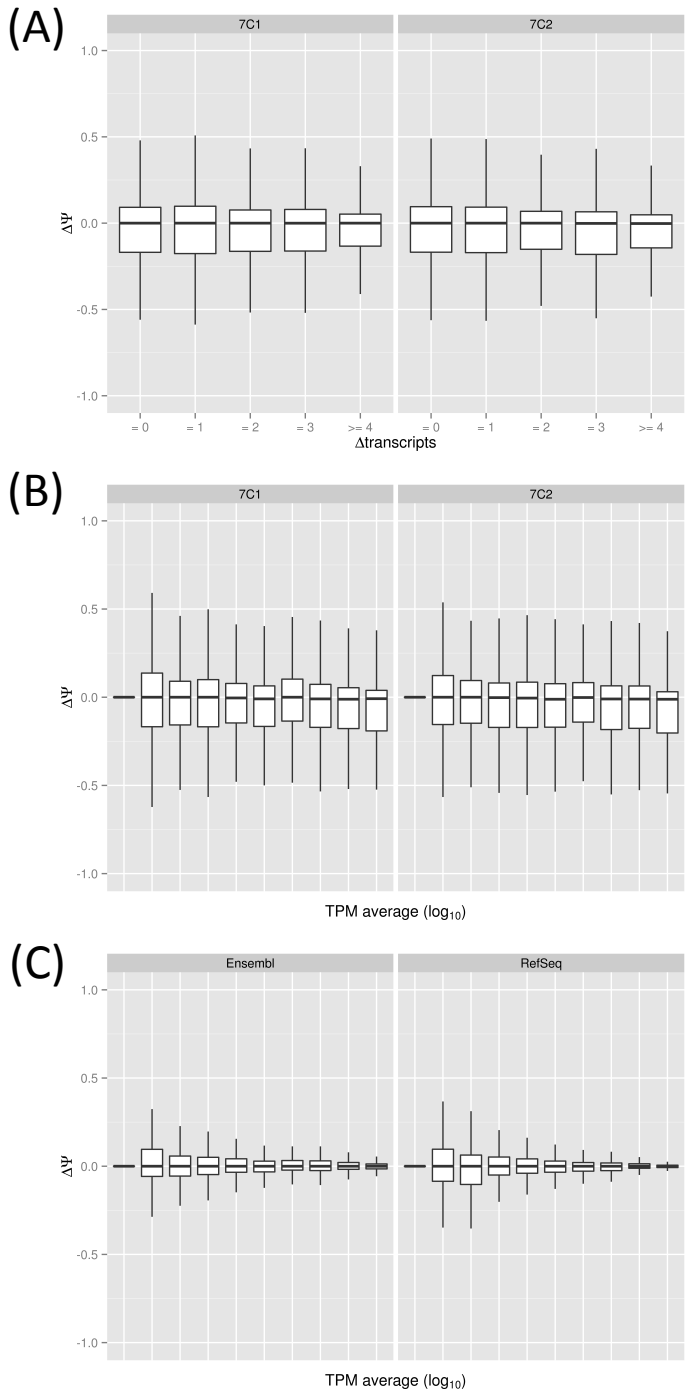


Figure 4. Annotation dependencies. Boxplots of the difference of Ψ values estimated by SUPPA for Ensembl and RefSeq annotations from Sailfish quantification (y axis) as a function of **(A)** the difference in the number of transcripts defining each event in Ensembl and RefSeq or as a function of **(B)** the mean expression of the gene in which the event is contained. X-axis in (B) is grouped into 10 quantiles according in the $\log_{10}$(TPM) scale. The variability is represented for both replicates (7C1 and 7C2) of the cytosolic RNA-Seq data from MCF7 cells. **(C)** Boxplots of the difference of Ψ values between replicates for the estimates from the Ensembl (left panel) and RefSeq (right panel) annotations as a function of the mean expression of the genes, grouped into 10 quantiles in the $\log_{10}$(TPM), for genes with TPM>0. Mean expression is calculated as the average of the $\log_{10}$(TPM) for the same gene in the two replicates for (C) or for the same gene in the two annotations in (B).

Figure 5

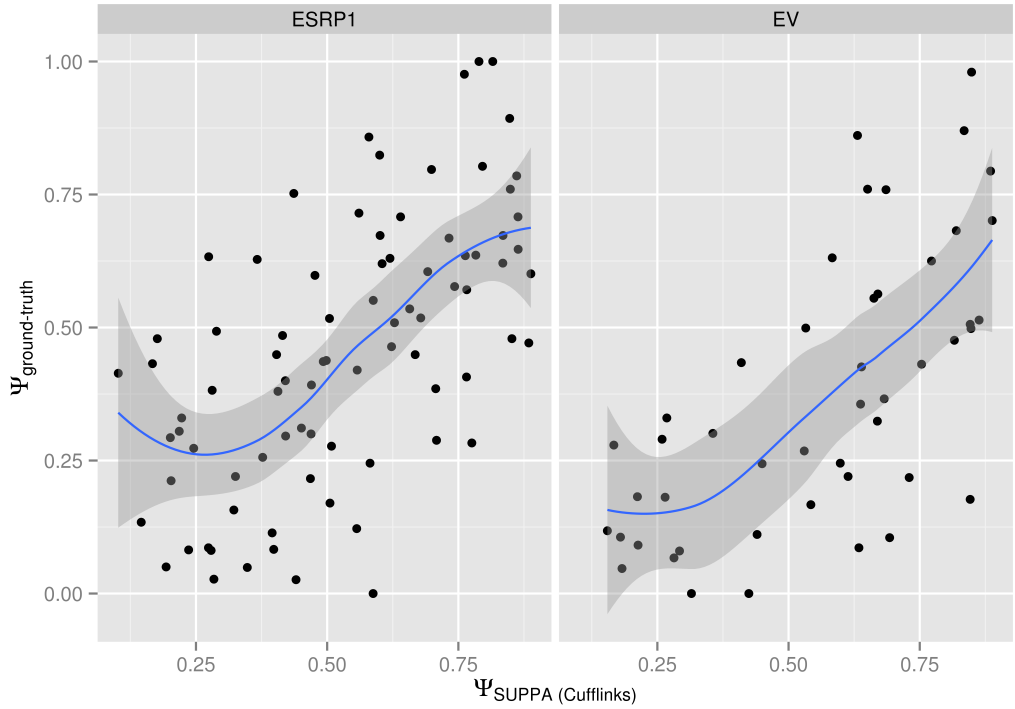


Figure 5. Annotation-free PSI estimation. Correlation of the experimental Ψ values with those estimated with Cufflinks *de novo* + SUPPA in MDA-MB-231 cells with (ESRP1, left panel) and without (EV, right panel) ESRP1 overexpression. Experimental Ψ values were obtained by RT-PCR (Shen et al. 2012) and estimated PSIs were obtained from RNA-Seq data in the same samples (Shen et al. 2012). The blue line and gray boundaries are the fitted curves with the LOESS regression method.

Figure 6

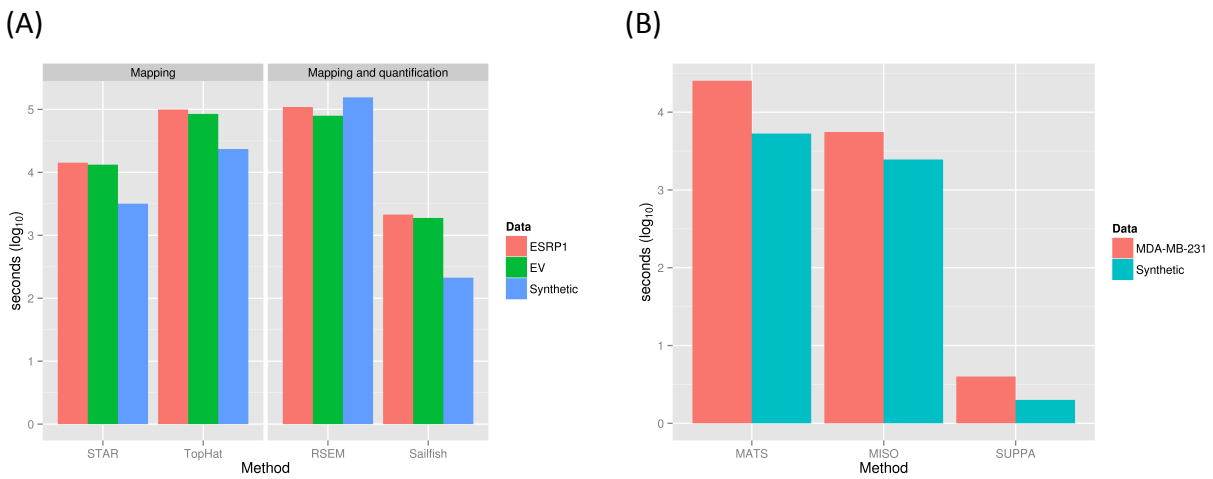


Figure 6. Speed benchmarking. (A) Time performance for read assignment/mapping to transcript/genome positions by RSEM, Sailfish, STAR and TopHat on the synthetic as well as the ESRP1 and EV RNA-Seq datasets separately (Methods). RSEM and Sailfish include the transcript quantification operation. **(B)** Time performance for the Ψ value calculation from the already mapped reads (MATS, MISO) or quantified transcripts (SUPPA). ESRP1 and EV samples were pooled for this benchmarking (MDA-MB-231). MATS time includes the calculation of the ΔΨ between samples, which we could not separate from the Ψ calculation (Shen et al. 2012). All tools were run in multi-threaded mode when possible. Time reported for all cases is the actual cumulative time the process used across all threads (Methods).