SUPPA: a super-fast pipeline for alternative splicing analysis from RNA-Seq

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

[1]Computational Genomics, 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

ABSTRACT

High-throughput RNA sequencing allows genome-wide analyses of pre-mRNA splicing across multiple conditions. However, the increasing number of available datasets represents a major challenge in terms of time and storage required for analyses. Here we describe SUPPA, a computational pipeline to calculate relative inclusion values of alternative splicing events, exploiting fast transcript quantification of a known annotation. SUPPA provides a fast and accurate approach to calculate inclusion levels of alternative splicing events from a large number of samples, thereby facilitating systematic analyses in the context of large-scale projects using limited computational resources. SUPPA is available at https://bitbucket.org/regulatorygenomicsupf/suppa under the MIT license and is implemented in Python 2.7.

INTRODUCTION

Alternative splicing plays an important role in many cellular processes and has been linked to multiple diseases. High-throughput RNA sequencing has facilitated the study of alternative splicing at genome scale and under multiple conditions. However, as more datasets become available, a bottleneck appears in relation to the required time and storage space required for analysis, representing a major obstacle for large-scale projects and the analysis of publicly available data. We have developed SUPPA, an effective computational pipeline for the rapid calculation of the of relative inclusion values of alternative splicing events from a large number of samples.

The description of alternative splicing in terms of events facilitates their experimental validation and their characterization in terms of regulatory mechanisms; and is also motivated by the current limitations in transcript reconstruction from short sequencing reads. On the other hand, recent developments in the quantification of known transcripts have shown that considerable quality can be achieved at very high speed (Li et al. 2011, Patro et al. 2014, Zhang et al. 2014). These methods can thus provide in very short time an accurate census of mRNA molecules in a given condition for a deeply annotated genome as human (Harrow et al. 2012). The relative inclusion, or PSI, of a splicing event is generally defined as the fraction of mRNA isoforms that include an exon or a specific form of the event (Wang et al. 2008, Brosseau et al. 2010), which is often estimated from the reads falling specifically on either form of the event (Wang et al. 2008). SUPPA calculates PSI values directly from transcript isoform abundance values. We argue that this provides a fast and accurate approach to systematic splicing analysis.

METHODS

An alternative splicing event is a binary representation of a local splicing variation in a gene. In this context, the PSI value of an event can be defined as the ratio of the abundance of mRNAs that include one form of the event, F_1 over the abundance of mRNAs that contain either form of the event, $F_1 \cup F_2$. 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 calculates the PSI (Ψ_q) for an event as follows:

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

SUPPA performs this calculation in two steps. The first one, *generateEvents*, reads an annotation file in GTF format and produces information about alternative splicing events in the annotation. SUPPA calculates exon skipping events, alternative 5' and 3' splice-sites, mutually exclusive exons, intron retention, and alternative first and last exons (Supp. Fig. 1). The Ψ_q values for all events are then calculated in a second step with the operation *psiPerEvent*, which uses as input the output from *generateEvents* and the abundances for all transcripts in one or

* eduardo.eyras@upf.edu

more samples, which are previously obtained with a fast transcript quantification method (Figure 1). 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. SUPPA also includes a tool to combine multiple input files from the transcript quantification and to obtain PSI values for transcript isoforms. More details are given in the documentation of the software.

RESULTS

To calculate 107,506 alternative splicing events from the Ensembl 75 annotation (37.494 genes, 135.521 transcripts), *generateEvents* took 10 minutes on a 2.5 GHz Intel Xeon, and 2 mins and 43 secs in a 2.9 Ghz Intel Core i7 processor. On the other hand, *psiPerEvent* took less than a minute in both machines to obtain the Ψ_q values (output size 26 Mb) for these events using 8 RNA-Seq samples previously processed with Sailfish. Similarly, *psiPerEvent* took 4 mins and 50 secs on a 2.5 GHz Intel Xeon for 929 breast tumor samples from TCGA for 40411 events (Supp. Material). Considering that the events only need to be computed once, SUPPA speed is very competitive.

We performed a benchmarking analysis by comparing SUPPA Ψ_q values with those obtained from junction reads (Ψ_j), defined in (Supp. Fig 2) on a set of non-overlapping alternative splicing events (Supp. Fig. 3). First, simulated reads obtained with FluxSimulator (Griebel et al. 2012) were mapped to the genome with STAR (Dobin et al. 2013) and reads in junctions were counted with sjcount (Pervouchine et al. 2013) to calculate Ψ_j values. The same simulated reads were used to quantify transcript abundances with Sailfish (Patro et al. 2014). SUPPA Ψ_q shows a high correlation with Ψ_j (Pearson R=0.94) over 1041 non-overlapping events, including exons skipping, alternative 5'/3'splice-sites and mutually exclusive exon events (Figure 2 and Supp. Fig. 4). This correlation is also high (R=0.92) when comparing SUPPA results to the values obtained from the simulated number of molecules (Supp. Fig. 5).

We also used RNA sequencing from nuclear and cytosolic fractions from MCF7 and MCF10 cells. Correlations between biological replicates were high for SUPPA Ψ_q values (Pearson R= 0.91-0.93) and improved when only genes with TPM>1 (calculated as the sum of TPMs of the transcripts in each gene) were used (Pearson R=0.97-0.99) (Supp. Figs. 6-9). SUPPA correlations between replicates were in all cases superior to those the junction-based values (Supp. Fig. 10), calculated as above. Moreover, SUPPA systematically recovers more events than using junction reads at similar correlation value between replicates (Supp. Material). Comparison between Ψ_q and Ψ_j values, using events with more than 20 junction reads in genes with TPM>1 and, showed a good correspondence in all samples tested (Pearson R=0.73-0.78) (Supp. Figs. 11-14). Moreover, to account for 3' end sequencing biases (Supp. Fig. 15) the analysis was repeated quantifying protein-coding sequences only and comparing events occurring in CDS regions (see Supp. Material). This showed a considerable increase in the correlation between Ψ_q and Ψ_j values (R=0.89-0.93) (Figure 2 and Supp. Figs. 16-19).

As an additional benchmark, we calculated TPM values for transcript isoforms using the RNA-Seq data for lung adenocarcinoma from the TCGA project (https://tcga-data.nci.nih.gov/tcga/), and applied SUPPA to 55 tumor and paired normal samples. The Ψ values calculated with SUPPA for two splicing events in the genes NUMB and BIN1, known to be upregulated in lung tumors (Misquitta-Ali et al. 2011, Zong et al. 2014), show a good correlation with the Ψ values calculated with junction reads (Supp. Fig. 20). Further details are provided as supplementary material.

ACKNOWLEDGEMENTS

The authors acknowledge useful discussions with S. Mount, S. Janga, Y. Barash and M. Robinson. *Funding*: Spanish Government (BIO2011-23920 and Consolider RNAREG CSD2009-00080), Sandra Ibarra Foundation for Cancer, and Spanish National Institute of Bioinformatics (INB).

REFERENCES

Brosseau JP, et al. (2010) High-throughput quantification of splicing isoforms. *RNA* 16(2):442-9.

Dobin A, et al. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**(1):15-21.

Griebel T, et al. (2012) Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Res.* **40**(20):10073-83.

Harrow J, et al. (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**(9):1760-74.

Li B, et al. (2010) RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* **26**(4):493-500.

Li B, et al. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **4**;12:323.

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.

Patro R, et al (2014) Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat Biotechnol.* **32**(5):462-4.

Pervouchine et al. (2013) Intron-centric estimation of alternative splicing from RNA-seq data. *Bioinformatics* **29**(2):273-4

Wang ET, et al. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**(7221):470-6.

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

Zong FY, et al. (2014) The RNA-binding protein QKI suppresses cancer-associated aberrant splicing. *PLoS Genet.* 10(4):e1004289.

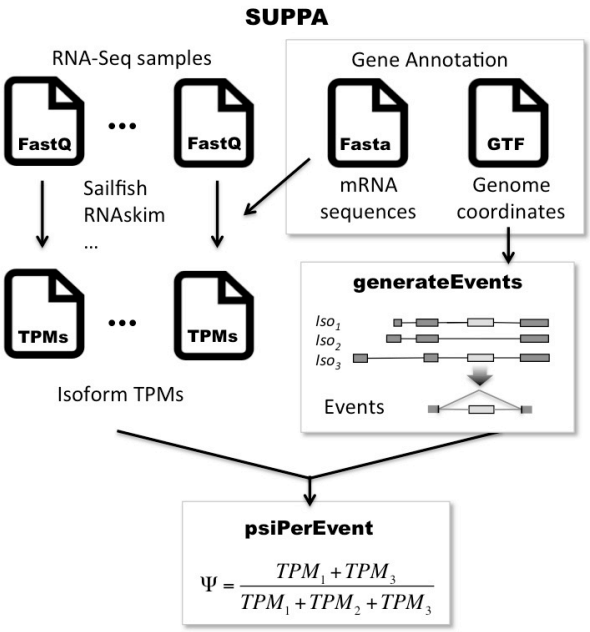


Fig. 1. Schematic description of how SUPPA works to define events from a gene annotation (*generateEvents*) and calculate event Ψ values from one or more transcript quantification files (*psiPerEvent*).

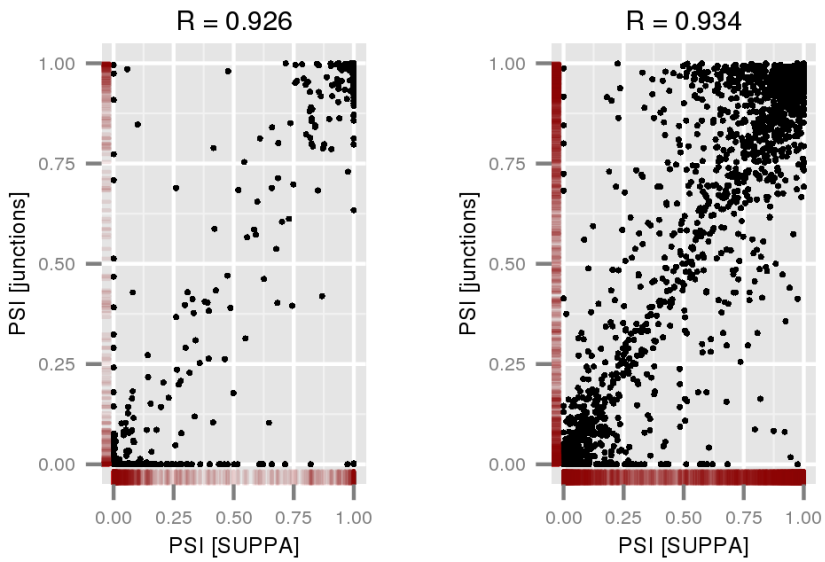


Fig. 2. Correlation (Pearson R) of the PSI values calculated with SUPPA (x-axis) and from junction reads (y-axis), respectively, for 1041 non-overlapping events in coding and non-coding regions using simulated reads (left) and for 2202 non-overlapping events in coding regions using sequencing reads from cytosolic mRNAs in MCF10 cells (right).

Supplementary Material

SUPPA: a super-fast pipeline for alternative splicing analysis from RNA-Seq

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

[1]Computational Genomics, 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

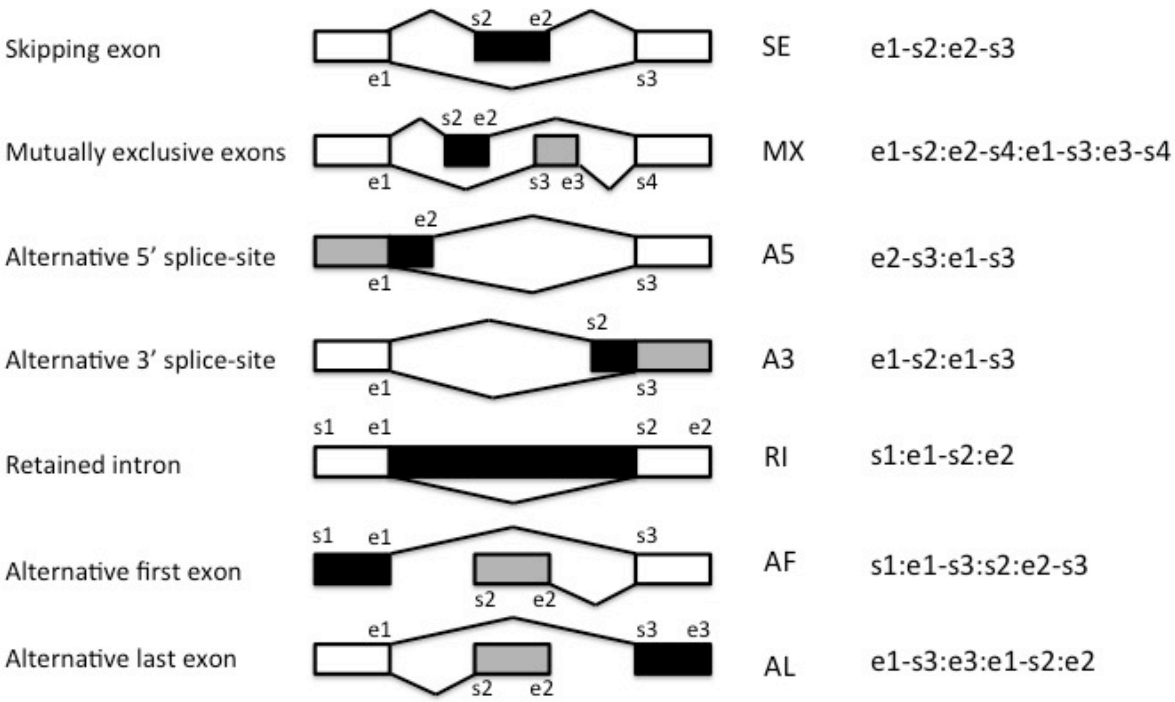
SUPPA Definition of events

Events are generated from the transcript annotation and are given an ID of the form

<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. The *seqname* is the field 1 from the GTF file (the chromosome typically). The *coordinates_of_the_event* is defined differently for each *even_type* (see Supplementary Figure 1). The PSI (Ψ) for an event is calculated with respect to one of the two forms of the event. Our convention is (See https://bitbucket.org/regulatorygenomicsupf/suppa for more details)

- **SE**: the PSI indicates the inclusion of the middle exon.
- **A5/A3**: the PSI indicates the inclusion of the form that minimizes the intron length
- **MX**: the PSI indicates the inclusion of the form that contains the alternative exon with the smallest start coordinate (the left-most exon), regardless of strand.
- **RI**: The PSI indicates the inclusion of the form that retains the intron.
- **AF/AL**: the PSI indicates the inclusion of the form that maximizes the intron length.



Supplementary Figure 1. Coordinate definition of the events by SUPPA. The type of event has an identifier (SE, MX, A5, A3, RI, AF, AL). Each event is also identified uniquely by a set of coordinates: The start (s) and end (e) coordinates for the different exonic regions involved in the event. The external coordinates of the event are only used for the RI, AF and AL events. 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. The gray area denotes the other alternative form of the event. For instance, for RI the inclusion level is given for the form that retains the intron.

Benchmarking

We downloaded the Ensembl [1] Release 75 annotation file in GTF format from the Ensembl FTP server, and filtered out all the annotations on chromosomes other than autosomes or sex chromosomes, obtaining 37,494 genes and 135,521 transcripts. We then used the SUPPA *generateEvents* operation:

```
/pathTo/python-2.7/python /pathTo/SUPPA/suppa.py generateEvents -i
/pathTo/ensembl_r75_filtered_annotation.gtf -o /pathTo/eventsOutput/ -e SE SS MX RI FL
```

to calculate all alternative splicing events present in the annotation. This produced a total of 107.506 events distributed as given in Supp. Table 1.

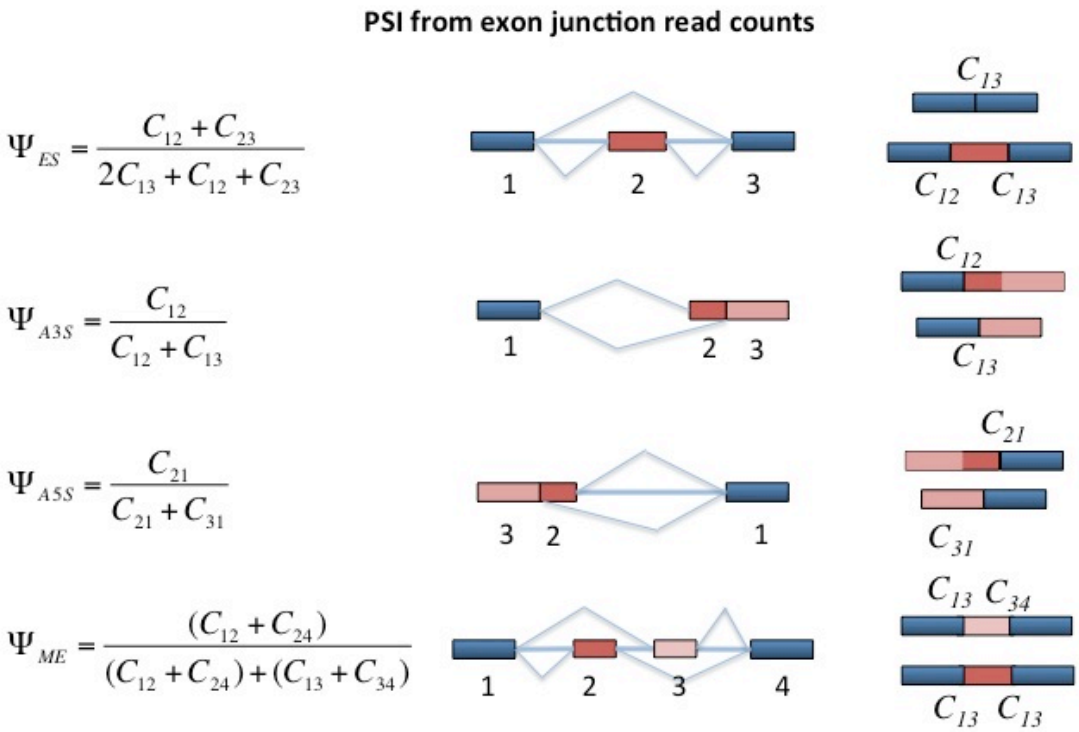
Supp. Table 1. Events calculated by SUPPA (*generateEvents*) from the Ensembl annotation.

Event type	Count
Skipping exon (SE)	21,930
Alternative 5'ss (A5)	8,417
Alternative 3'ss (A3)	9,787
Mutually exclusive (MX)	3,312
Intron retention (RI)	3,614
Alternative first exon (AF)	48,219
Alternative last exon (AL)	12,227
Total	107,506

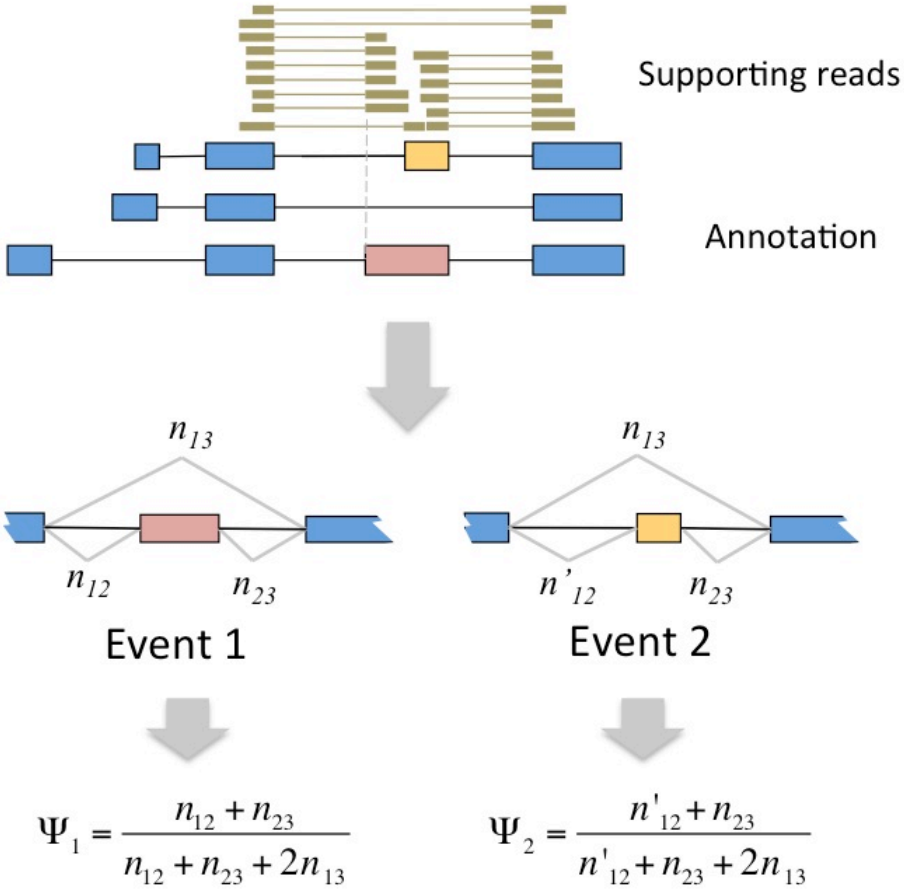
For the purpose of benchmarking, we compared the relative inclusion values obtained from transcript quantification (Ψ_q) produced by SUPPA, to those values obtained from junction reads (Ψ_j) (Supplementary Figure 2). We used events that do not overlap each other. The reason for this is that junctions are often shared between overlapping events, which introduce a confounding factor in the calculation of inclusion levels with junction-based methods (Supplementary Figure 3). Note that this confounding factor is circumvented by SUPPA, since the usage of a specific junction is given by the abundance of the transcripts that cover a form of the event, and not simply the junction (Supplementary Figure 3). Accordingly we selected for benchmarking the set of 8,636 alternative splicing events that are internal (exon skipping, alternative 5'/3' sites and mutually exclusive) and that do not overlap with any other event. Using coding transcripts only (see below), the number of non-overlapping events obtained was 9.337.

RNA Sequencing of MCF7 and MCF10 cells

MCF7 cells were cultured in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal bovine serum (FBS) and penicillin/streptomycin. MCF10A 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 under accession SRP045592 (http://www.ncbi.nlm.nih.gov/sra/?term=SRP045592)



Supplementary Figure 2. Calculation of relative inclusion levels from junction reads. Reads falling on the different junctions (C_{ij}) are counted. For each alternative splicing event, the relative inclusion level (Ψ) is defined in terms of the read counts on the junctions that describe the relevant splicing variation. Only internal events (exon-skipping, alternative 5'/3' splice-sites, mutually exclusive exons), described in the figure, were used for benchmarking.



Supplementary Figure 3. Junctions common to more than one alternative splicing event. The figure depicts an example of gene with two exon-skipping events that share a junction. The gene also contains an alternative 3' splice-site event that is not represented as an event in the figure. Methods based on junction-reads would allocate the reads falling on the common junction n_{23} to the calculation of Ψ for both skipping events. However, the non-common junctions (n_{12} and n'_{12}) indicate that the common junction should mostly contribute to Event 1 rather than to Event 2. As a result, the estimation of the Ψ value for Event 2 will be incorrect. This confounding factor is present in general for other types of events when they overlap each other. This effect would be circumvented by SUPPA, since the usage of a specific junction is given by the abundance of the transcripts that cover that junction, which also depends on the reads falling elsewhere along the transcript.

Simulated reads

Simulated human RNA-Seq data was generated with FluxSimulator [2] using the command:

/pathTo/flux-simulator-1.2.1/bin/flux-simulator -x -l -s -p /pathTo/parameters_file.par

with the parameters listed in Supp. Table 2. This resulted in a dataset of 45 million 2x76bp paired-end reads, which were split into two different FASTQ files, one containing the mate reads mapping in the sense strand of the transcripts and the other the mate reads in the antisense strand.

Supplementary Table 2 – Flux Simulator Parameters

Gene Expression Parameters	REF_FILE_NAME	/pathTo/ensembl_r75_filtered_annotation.gtf
	LOAD_CODING	True
	LOAD_NONCODING	True
	NB_MOLECULES	5.000.000
	EXPRESSION_K	-0.6
	EXPRESSION_X0	9.500
	EXPRESSION_X1	9.500^2
	TSS_MEAN	25
	POLYA_SCALE	NaN
	POLYA_SHAPE	NaN
Reverse transcription Parameters	RTRANSCRIPTION	True
	RT_PRIMER	RH
	RT_MIN	500
	RT_MAX	5.500
Fragmentation Parameters	FRAGMENTATION	true
	FRAG_METHOD	UR
	FRAG_SUBSTRATE	RNA
	FRAG_UR_D0	1
	FRAG_UR_DELTA	NaN
	FRAG_UR_ETA	NaN
Final Library Preparation Parameters	FILTERING	True
	PCR_PROBABILITY	0.05
	GC_MEAN	NaN
	GC_SD	NaN
Sequencing Parameters	READ_NUMBER	90.000.000
	READ_LENGTH	76
	PAIRED_END	True
	GEN_DIR	/pathTo/hg19_fasta/
	FASTA	True
	ERR_FILE	76
	UNIQUE_IDS	False

Read mapping and Ψ_j calculation

Simulated reads were mapped using STAR with parameters *--outSJfilterOverhangMin -1 -1 -1 -1* and *--sjdbScore 100* to use only annotated junctions. A genome index was previously generated with STAR over the filtered Ensembl Release 75 annotation using the --sjdbOverhang 75 parameter to adjust the splice junction database to the length of the RNA-Seq reads.

RNA-Seq reads from cytosolic and nuclear fractions in MCF7 and MCF10 cells were mapped using STAR [3] with parameters *--outSJfilterOverhangMin -1 -1 -1 -1* and *--sjdbScore 100* to use only annotated junctions. A genome index was previously generated with STAR over the filtered Ensembl Release 75 annotation using the *--sjdbOverhang 49* parameter to adjust the splice junction database to the length of the RNA-Seq reads.

Resulting alignment files in SAM format were converted to BAM format and then sorted using samtools [4], prior to counting the junction reads. Reads at junctions were counted with the sjcount software [5] from the BAM files for real and simulated data, using the *-read1 1* and *-read2 0* parameters. The relative inclusion values from junction reads were calculated using the formulae given in Supp. Fig. 2.

Supplementary Table 3. Commands used for the generation of junction reads, mapping of reads, and read counting on junctions.

	Analysis	**Command**
Operations to obtain Ψ_j	Generation of junction index with STAR for RNA-Seq data	/pathTo/STAR-2.3.1z10/STAR --runMode genomeGenerate --genomeDir /pathTo/genomeDir/ --genomeFastaFiles /pathTo/hg19_fasta/chr1.fa .. /pathTo/hg19_fasta/chrX.fa --sjdbGTFfile /pathTo/ensembl_r75_filtered_annotation.gtf --sjdbOverhang 49
	Generation of junction index with STAR for simulated data	/pathTo/STAR-2.3.1z10/STAR --runMode genomeGenerate --genomeDir /pathTo/genomeDir/ --genomeFastaFiles /pathTo/hg19_fasta/chr1.fa .. /pathTo/hg19_fasta/chrX.fa --sjdbGTFfile /pathTo/ensembl_r75_filtered_annotation.gtf --sjdbOverhang 75
	Mapping of reads with STAR	/pathTo/STAR-2.3.1z10/STAR --runMode alignReads --genomeDir /pathTo/genomeDir/ --readFilesIn /pathTo/reads_1.fastq /pathTo/reads_2.fastq --genomeLoad NoSharedMemory --outFileNamePrefix /pathTo/alignmentOutput/ --outSAMattributes All --sjdbScore 10 --outSJfilterOverhangMin -1 -1 -1 -1
	Conversion of reads from SAM to BAM, then to sorted BAM	**Step 1** /pathTo/samtools-0.1.18/samtools view -S -b /pathTo/alignmentOutput/alignment.sam > /pathTo/alignmentOutput/alignment.bam **Step 2** /pathTo/samtools-0.1.18/samtools sort /pathTo/alignmentOutput/alignment.bam /pathTo/alignmentOutput/alignment.sorted
	Quantification of junction reads with sjcount	/pathTo/sjcount/sjcount_v3 -bam /pathTo/alignment.sorted.bam -ssj /pathTo/alignment.ssj -read1 1 -read2 0

Transcript quantification and Ψ_q calculation

Sailfish [6] was used to generate the quantification of all transcripts in the filtered Ensembl Release 75 annotation file using the simulated and the real data. We first downloaded the FASTA sequences of the Ensembl transcripts corresponding to the same annotation as the GTF described earlier, and used it to generate the Sailfish index, selecting a k-mer size of 31 to minimize the number of reads mapping to different transcripts during the quantification step. We then proceeded with the quantification step by running Sailfish over the FASTQ files using the *-l "T=PE:O=><:S=AS"* flag. The process was analogous for coding transcripts. Only protein coding transcripts from Ensembl were quantified with Sailfish, and the GTF annotation (CDS lines) were used to generate events with SUPPA.

We used the *psiPerEvent* operation of SUPPA to calculate the Ψ_q values from the transcript quantifications obtained by Sailfish for all the 8,636 alternative splicing events generated above with the *generateEvents* module of SUPPA. The value Ψ_q is defined as the ratio of the set of mRNAs that include one form of the event, F_1 over the mRNAs that contain both forms of the event, F_1ÈF_2. Provided the expression levels of all transcripts isoforms calculated with Sailfish, here denoted as TPM_k, then we quantify the event using the equation (1) from the main text.

The value Ψ_q was calculated using the simulated and the real datasets. Additionally, we also calculated these values for the uncorrected and corrected TPMs given by Sailfish [6]. The commands used for these operations are given in Supp. Table 4.

Supplementary Table 4. Commands used for transcript quantification with Sailfish and for alternative splicing event quantification using SUPPA. The program *multipleFieldSelection.py* is included in the SUPPA distribution. Besides the *psiPerEvent* operation, SUPPA also allows to calculate the Y per isoform (i.e. normalized per gene) using *psiPerIsoform* (more details are provided at https://bitbucket.org/regulatorygenomicsupf/suppa)

	Analysis	**Command**
Operations to obtain Ψ_q	Transcript quantification with Sailfish	**Step 1** /pathTo/sailfish-0.6.3/bin/sailfish index -t /pathTo/ensembl_r75_filtered_annotation.fasta -k 31 -o /pathTo/indexDir/ **Step 2** /pathTo/sailfish-0.6.3/bin/sailfish quant -i /pathTo/indexDir/ -l "T=PE:O=><:S=AS" -1 /pathTo/reads_1.fastq -2 pathTo/reads_1.fastq -o /pathTo/quantificationOutput/
	Generation of alternative splicing events with SUPPA	/pathTo/python-2.7/python /pathTo/SUPPA/suppa.py generateEvents -i /pathTo/ensembl_r75_filtered_annotation.gtf -o /pathTo/eventsOutput/ -e SE SS MX RI FL
	Quantification of PSI for events from multiple from multiple samples	**Step 1** /pathTo/python-2.7/python /pathTo/SUPPA/multipleFieldSelection.py -i /pathTo/quantificationOutput/transcripts_tpm.sf -k 1-f 3 -o /pathTo/quantificationOutput/transcripts_tpm.tab **Step 2** /pathTo/python-2.7/python /pathTo/SUPPA/suppa.py psiPerEvent -i /pathTo/eventsOutput/exon_skipping_event_set.ioe -e /pathTo/quantificationOutput/transcripts tpm.tab -o /pathTo/psiFolder/events_psiq.tab

Correlation between Ψ_q and Ψ_j values

For the correlations between Ψ_q and Ψ_j values we first filtered out, in each sample separately, all the events that did not fulfil any of the following conditions:

a) Having a Ψ_q value greater or equal to 0. This is imposed simply because SUPPA gives a Ψ_q=-1 when the transcripts in the denominator of the Ψ_q formula have zero expression.

b) Having, at least, one read in the skipping junction. This condition is imposed to account for the fact that the junction-based Ψ_j can never give a value smaller than 1 if there are no skipping reads, whereas Ψ_q can give values < 1 without skipping reads, since it uses abundance values for all the transcripts.

c) Belonging to a gene with at least TPM > 1, where this TPM is the sum of the TPM values from all transcripts in that gene. We have observed that the estimation of Ψ_q is better when the average expression of the genes is sufficiently high. In fact, the $\Delta\Psi_q$ between replicates appears to be larger for genes with average expression TPM ≤ 1 (Supp. Figures 8 and 9).

d) Having at least 20 reads in total mapping at the junctions. Similarly to condition c), the estimation of PSI from junction reads is more robust when a minimum of reads is required to be at the junctions, as shown in (Supp Figure 10).

Additionally, as mentioned before, only non-overlapping were considered (Supp. Figure 3). Resulting number of events per sample are reported in Supp. Tables 5 and 6. The final number of events used for the correlation analysis are given in Supp. Table 7. Pearson correlation coefficient (R) was computed between Ψ_q and Ψ_j values, for both real and simulated data. Results are shown in Supp. Figures 4-8.

Supplementary Table 5. Number of events from the total of 8636 non-overlapping that pass the filter of either gene expression TPM>1, or having ≥20 reads in the exon junctions of the event. The number of events after imposing TPM>1 may differ between using the corrected or uncorrected transcript TPM values from Sailfish.

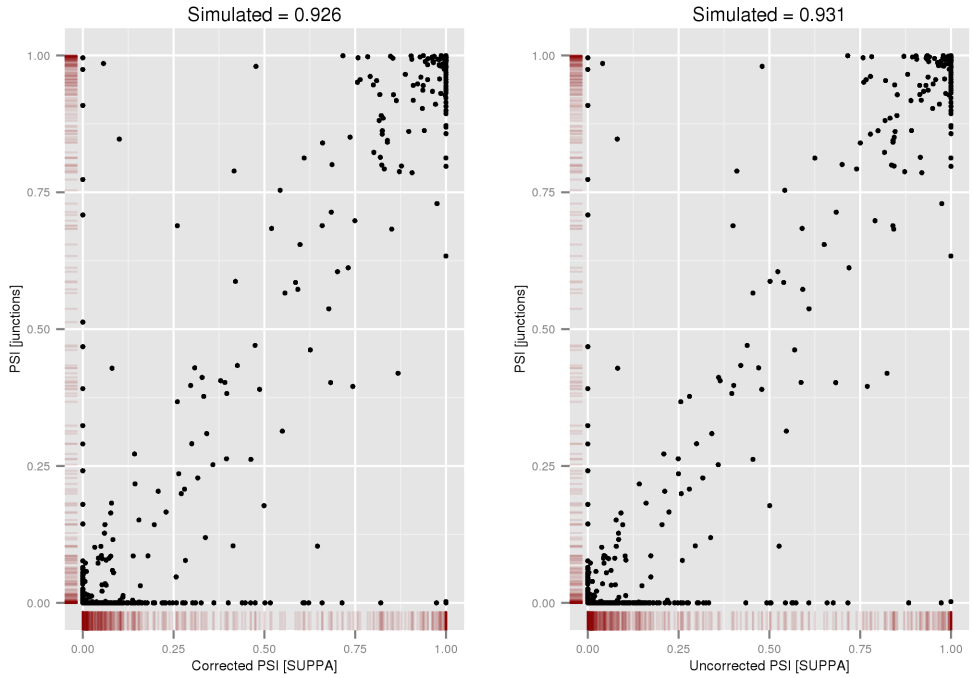
Sample	Events in genes with TPM > 1		Events with number of reads in junctions≥20
	Corrected	Uncorrected	
10C1	5085	5168	3170
10C2	5138	5185	3437
10N1	5697	5710	3852
10N2	5706	5703	3737
7C1	5155	5254	3129
7C2	5157	5239	3249
7N1	5735	5750	3446
7N2	5742	5769	3552

Supplementary Table 6. Number of events from the total of 9337 non-overlapping in coding regions that pass the filter of either gene expression TPM>1, or having ≥20 reads in the exon junctions of the event. The number of events after imposing TPM>1 may differ between using the corrected or uncorrected transcript TPM values from Sailfish.

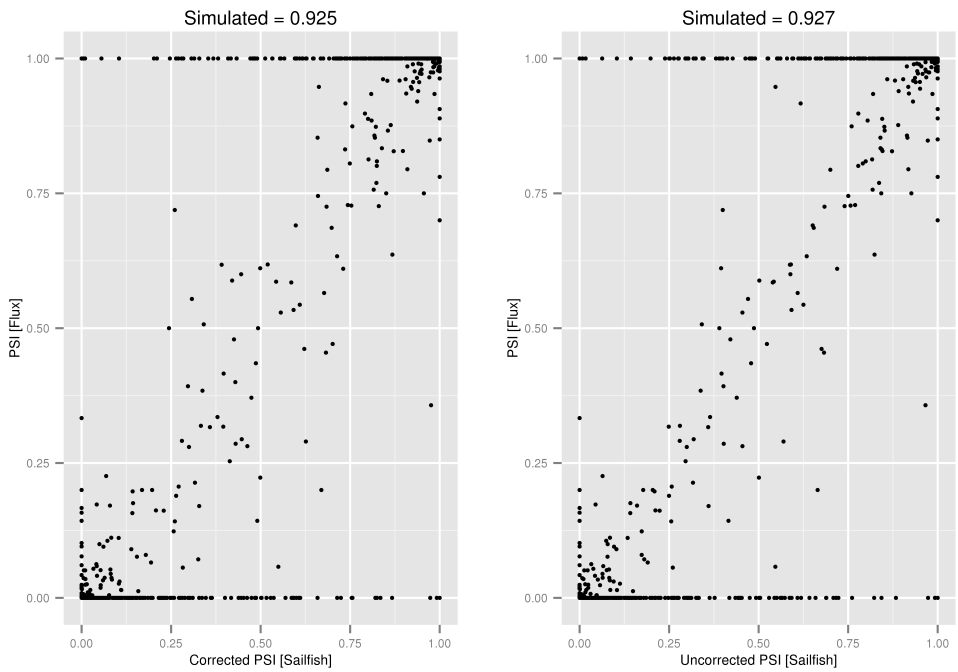
Sample	Events in genes with TPM > 1		Events with number of reads in junctions≥20
	Corrected	Uncorrected	
10C1	6395	6326	3233
10C2	6412	6341	3455
10N1	6724	6722	3772
10N2	6718	6717	3699
7C1	6420	6366	3175
7C2	6448	6375	3271
7N1	6771	6761	3439
7N2	6785	6773	3532

Supplementary Table 7. Number of events from the above lists that pass both filters and therefore were used for the correlation analysis between Ψ_q and Ψ_j values, for all transcripts and for coding transcripts. These are the events used for correlation calculation. The number of events after imposing TPM>1 may differ between the corrected and uncorrected Sailfish transcript TPM values.

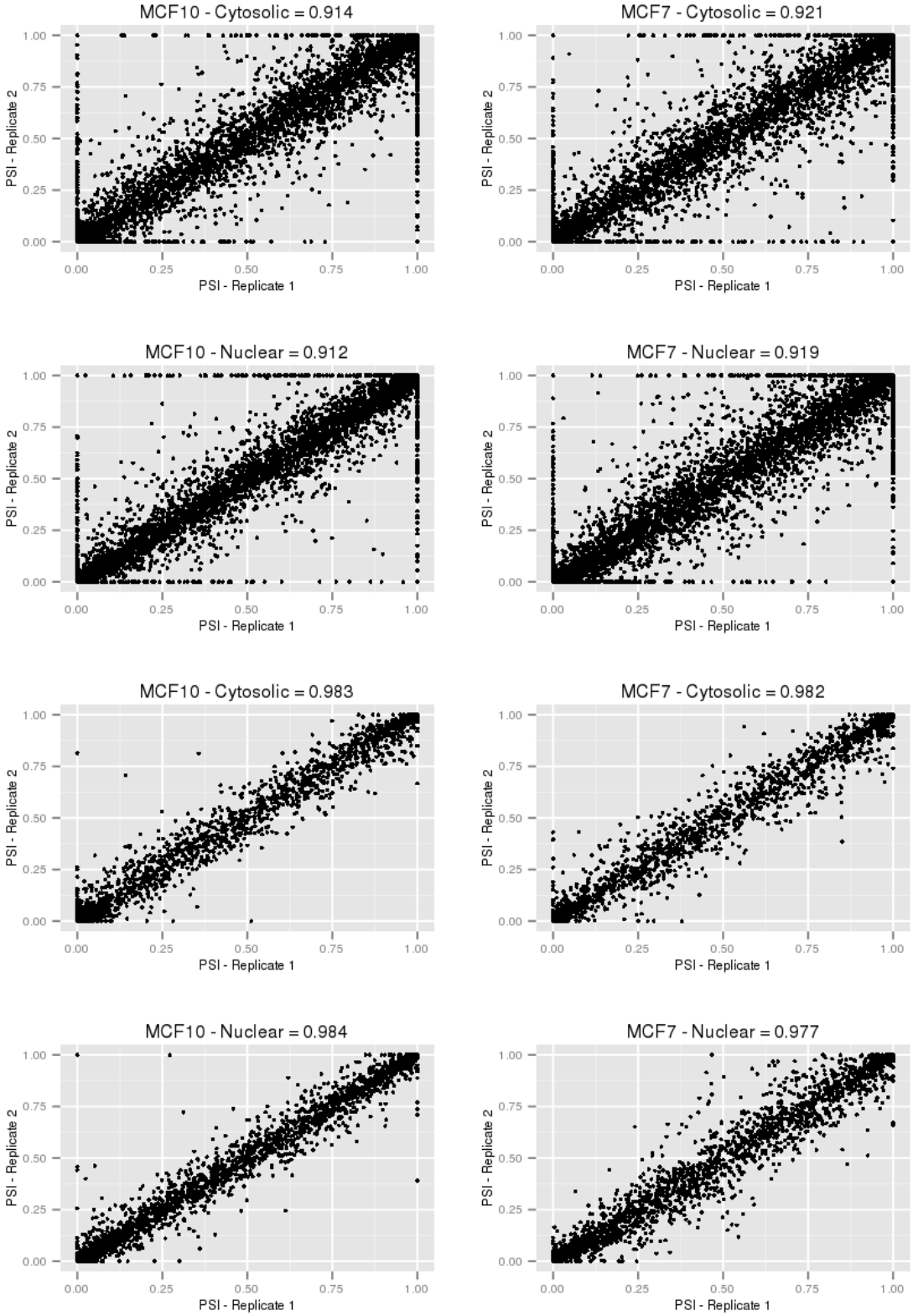
Sample	Final number of events used for correlation analysis: corrected (uncorrected)	Final number of events in coding transcripts used for correlation analysis: corrected (uncorrected)
MCF-10A – Cytosolic – Replicate 1	2324 (2324)	2203 (2202)
MCF-10A – Cytosolic – Replicate 2	2571 (2573)	2429 (2429
MCF-10A – Nuclear – Replicate 1	3000 (3000)	2763 (2763)
MCF-10A – Nuclear – Replicate 2	2888 (2889)	2687 (2687)
MCF-7 – Cytosolic – Replicate 1	2403 (2402)	2323 (2322)
MCF-7 – Cytosolic – Replicate 2	2543 (2543)	2437 (2437)
MCF-7 – Nuclear – Replicate 1	2835 (2836)	2698 (2698)
MCF-7 – Nuclear – Replicate 2	2896 (2897)	2744 (2744)
Simulated	1043 (1041)	-



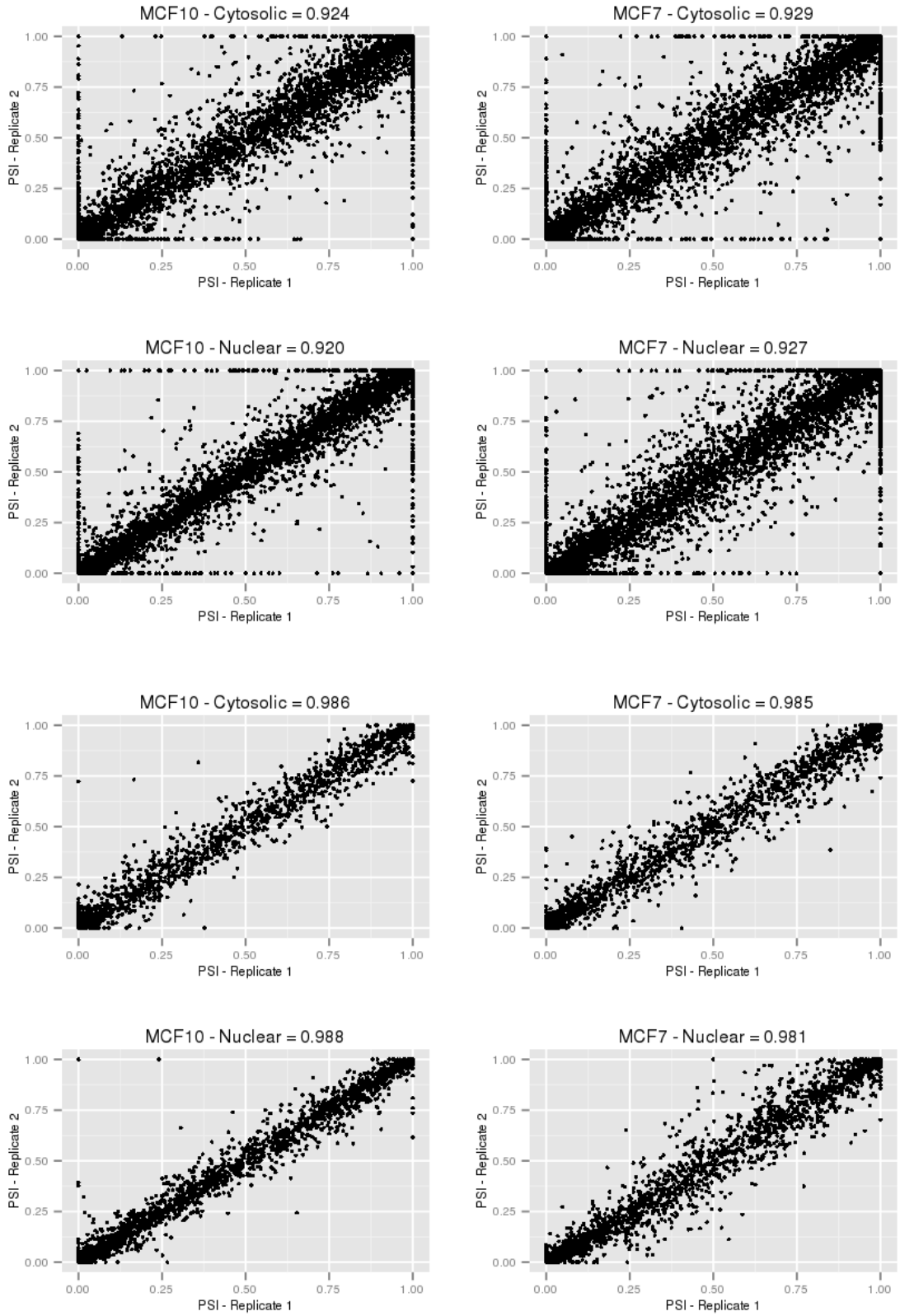
Supplementary Figure 4. Correlation between Ψ_q and Ψ_j for simulated data. The x-axis represents the value of Ψ_q from SUPPA (PSI [SUPPA] and the y-axis the value of Ψ_j (PSI [junction]) for the events in Supp. Table 4, using the simulated data and the corrected (left panel) and uncorrected (right panel) TPMs from Sailfish for Ψ_q. Ψ_j were obtained as described in the text. Correlation values (Pearson R) are given above each plot. The density of points along each axis is indicated in red.



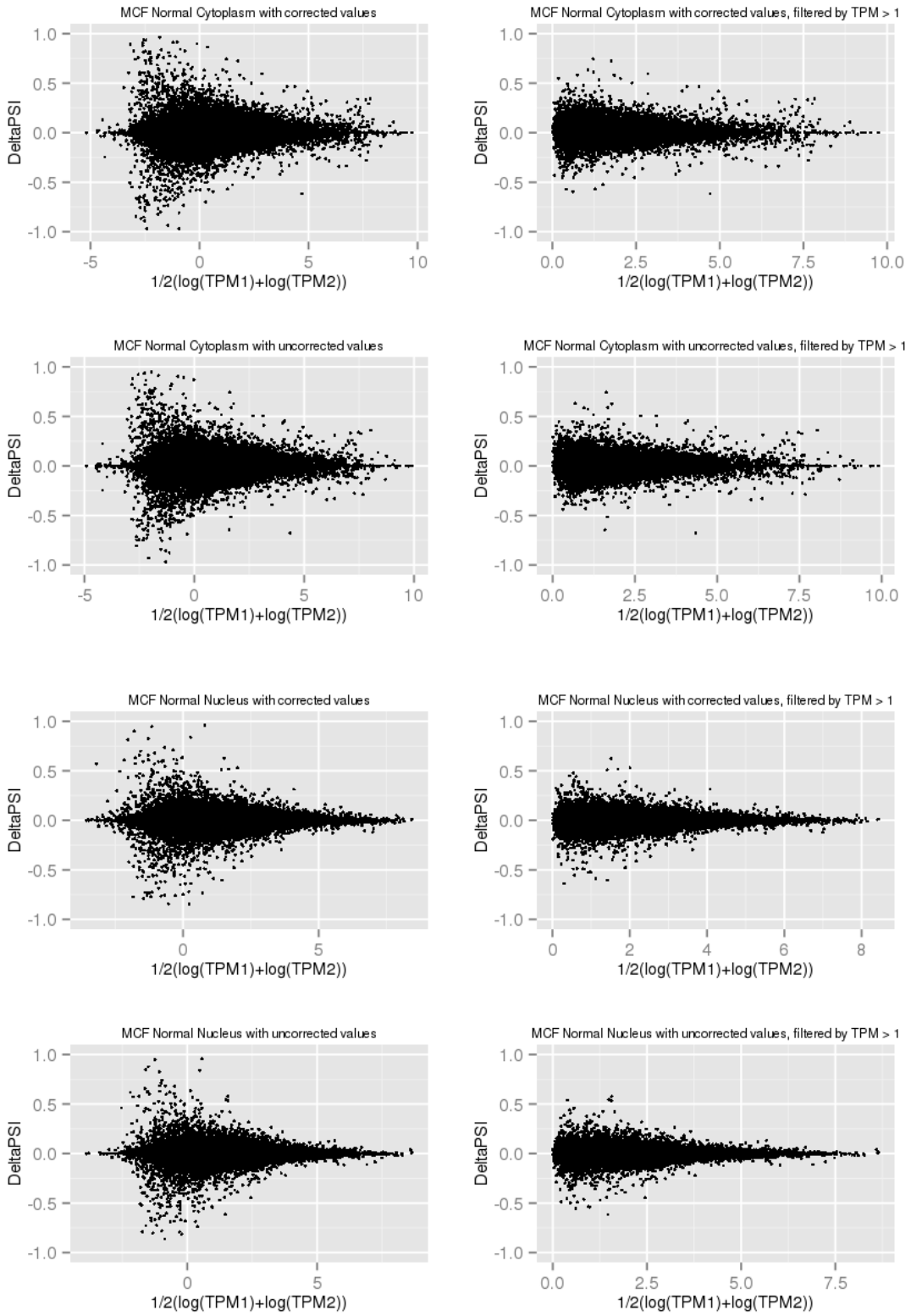
Supplementary Figure 5. Correlation between Ψ_q and the inclusion values calculated from the real count of RNA molecules from the simulation. As before, the x-axis represents Ψ_q (PSI [SUPPA]) quantified from the corrected (left) and uncorrected (right) TPMs calculated from Sailfish on the simulated data. The y-axis represents an inclusion value calculated with the same formula as SUPPA but using the real number of RNA molecules simulated by Flux Simulator. We only considered events belonging to a gene for which TPM > 1.



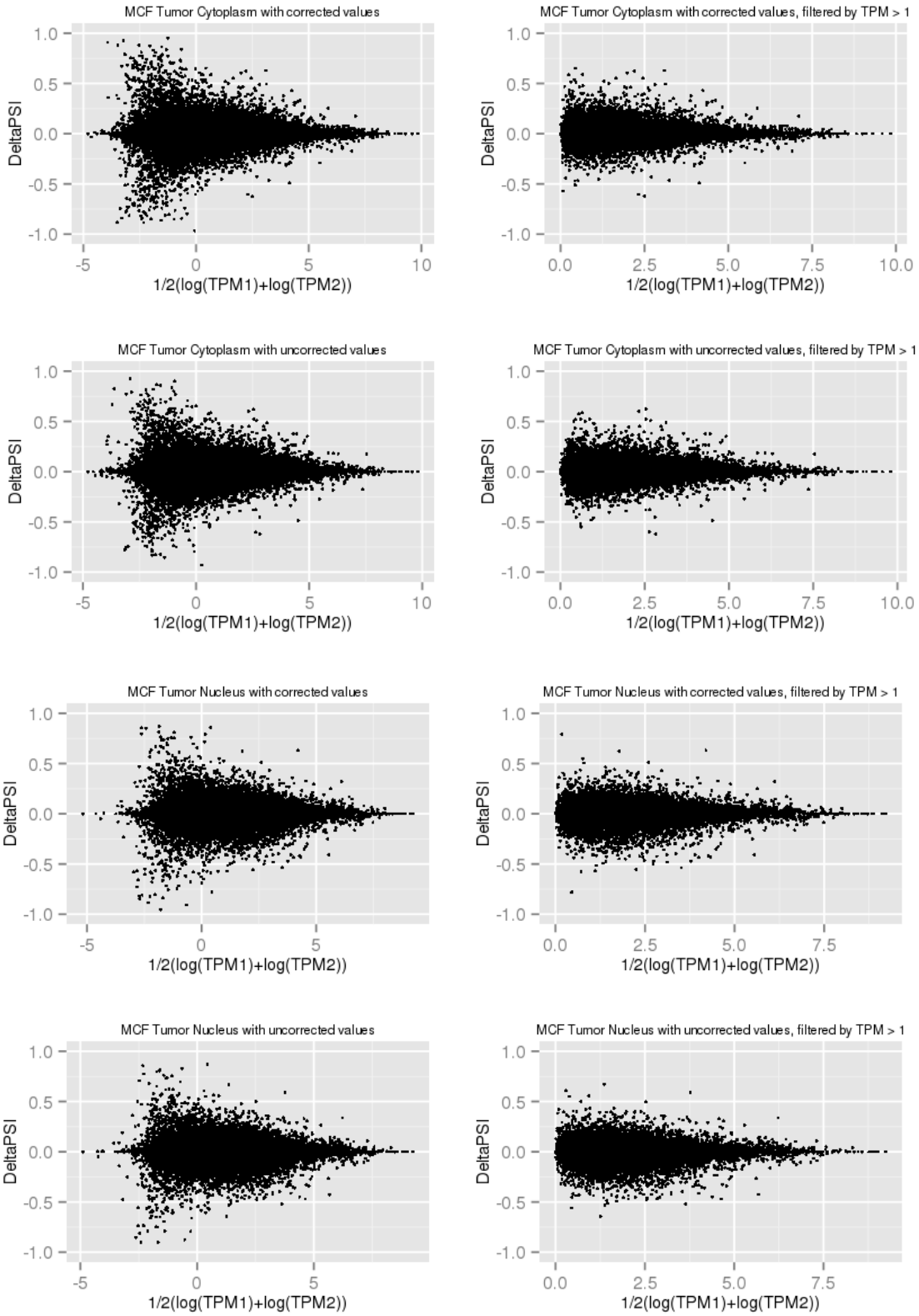
Supplementary Figure 6. Correlation between replicates for Ψ_q (corrected). Values were calculated for real RNA-Seq datasets using the transcript quantification from Sailfish in TPM units (corrected) for all events (4 upper panels) and for events in genes with TPM > 1 in both replicates (4 lower panels).



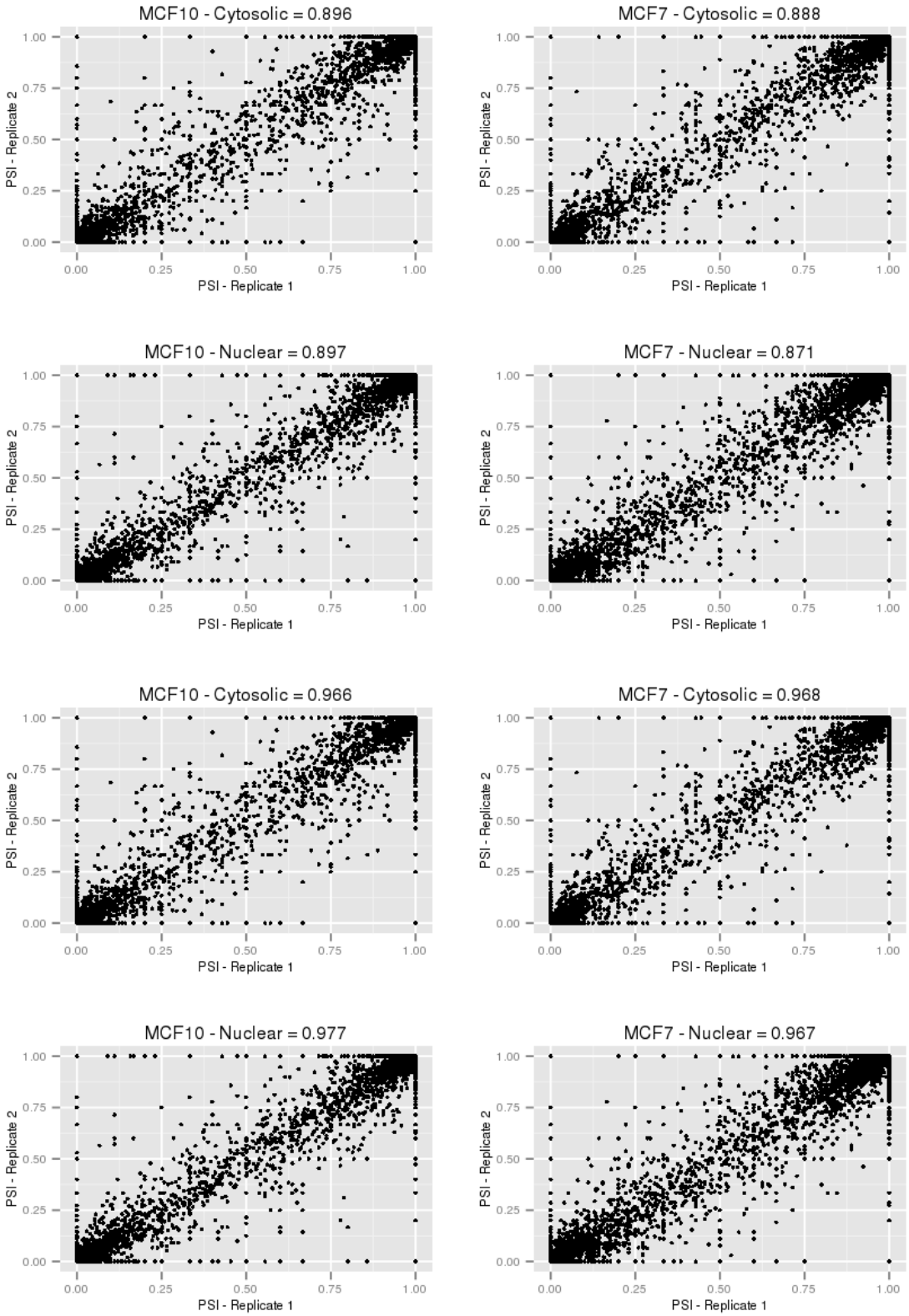
Supplementary Figure 7. Correlation between replicates for Ψ_q (uncorrected). Values were calculated for real RNA-Seq datasets using the transcript quantification from Sailfish in TPM units (uncorrected) for all events (4 upper panels) and for events in genes with TPM > 1 in both replicates (4 lower panels).



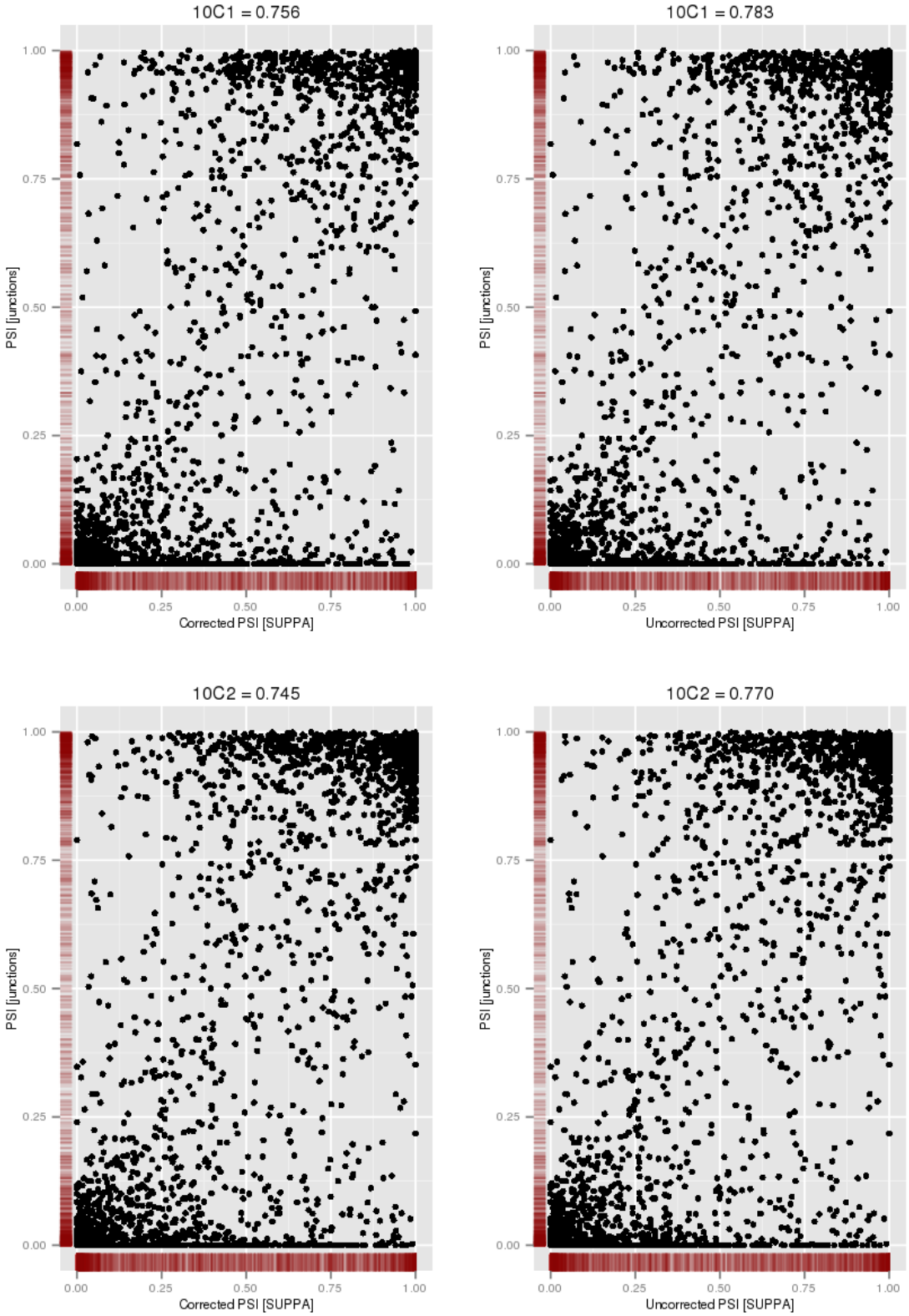
Supplementary Figure 8. MA-plots for SUPPA on the cytoplasm (four upper panels) and nuclear (four lower panels) for MCF10 RNA-Seq replicates. Plots of the $\Delta\Psi_q$ (DeltaPSI) (y-axis) as a function of the average gene expression in both replicates in $\log_2$(TPM) units (x-axis) for replicates 1 (TPM1) and 2 (TPM2). The left panels represent all the events, whereas the panels on the right represent those events for which the genes have TPM > 1 in both replicates. The plots were made using the corrected (top panels) and uncorrected (lower panels) TPM values estimated by Sailfish on the Ensembl transcripts.



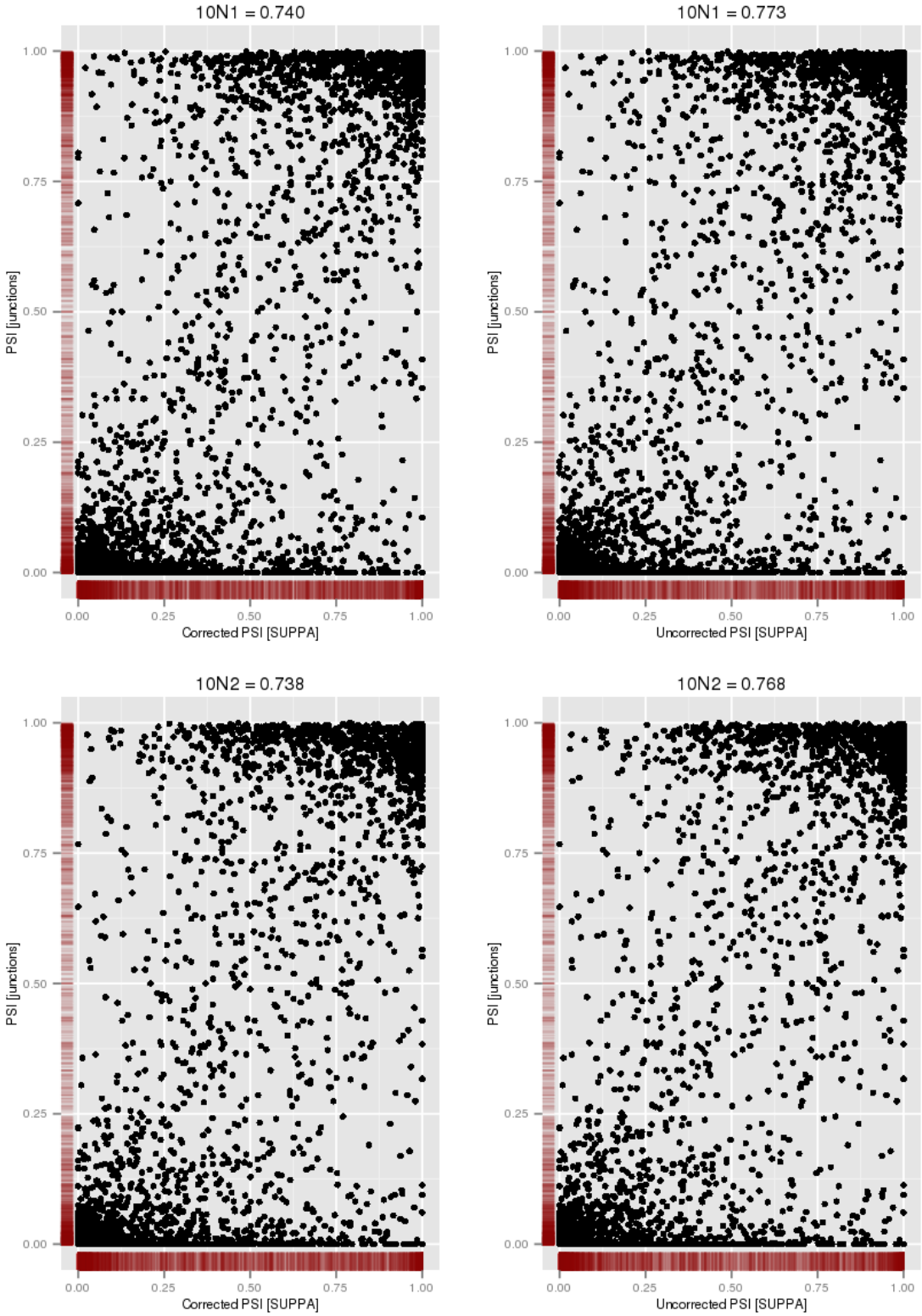
Supplementary Figure 9. MA-plots for SUPPA on the cytoplasmic (four upper panels) and nuclear (four lower panels) MCF7 RNA-Seq replicates. Plots of the $\Delta\Psi_q$ (DeltaPSI) (y-axis) as a function of the average gene expression in both replicates in $\log_2$(TPM) units (x-axis) for replicates 1 (TPM1) and 2 (TPM2). The left panels represent all the events, whereas the panels on the right represent those events for which the genes have TPM > 1 in both replicates. The plots were made using the corrected (top panels) and uncorrected (lower panels) TPM values estimated by Sailfish on the Ensembl transcripts.



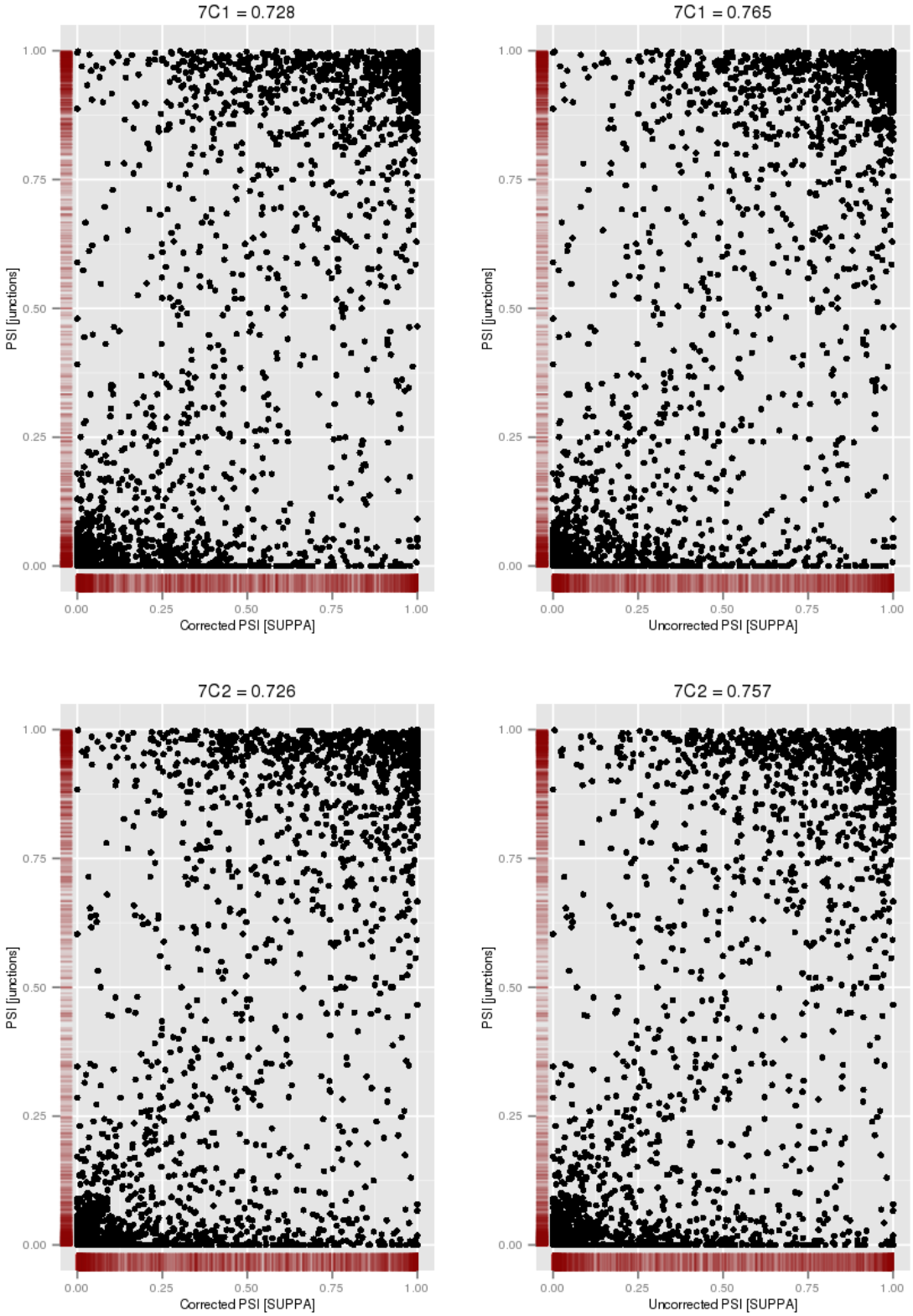
Supplementary Figure 10 – Correlation between replicates for Ψ_j values calculated from real RNA-Seq data. For all events (4 upper panels) and for events having at least 20 reads in total mapping to the junctions (four lower panels).



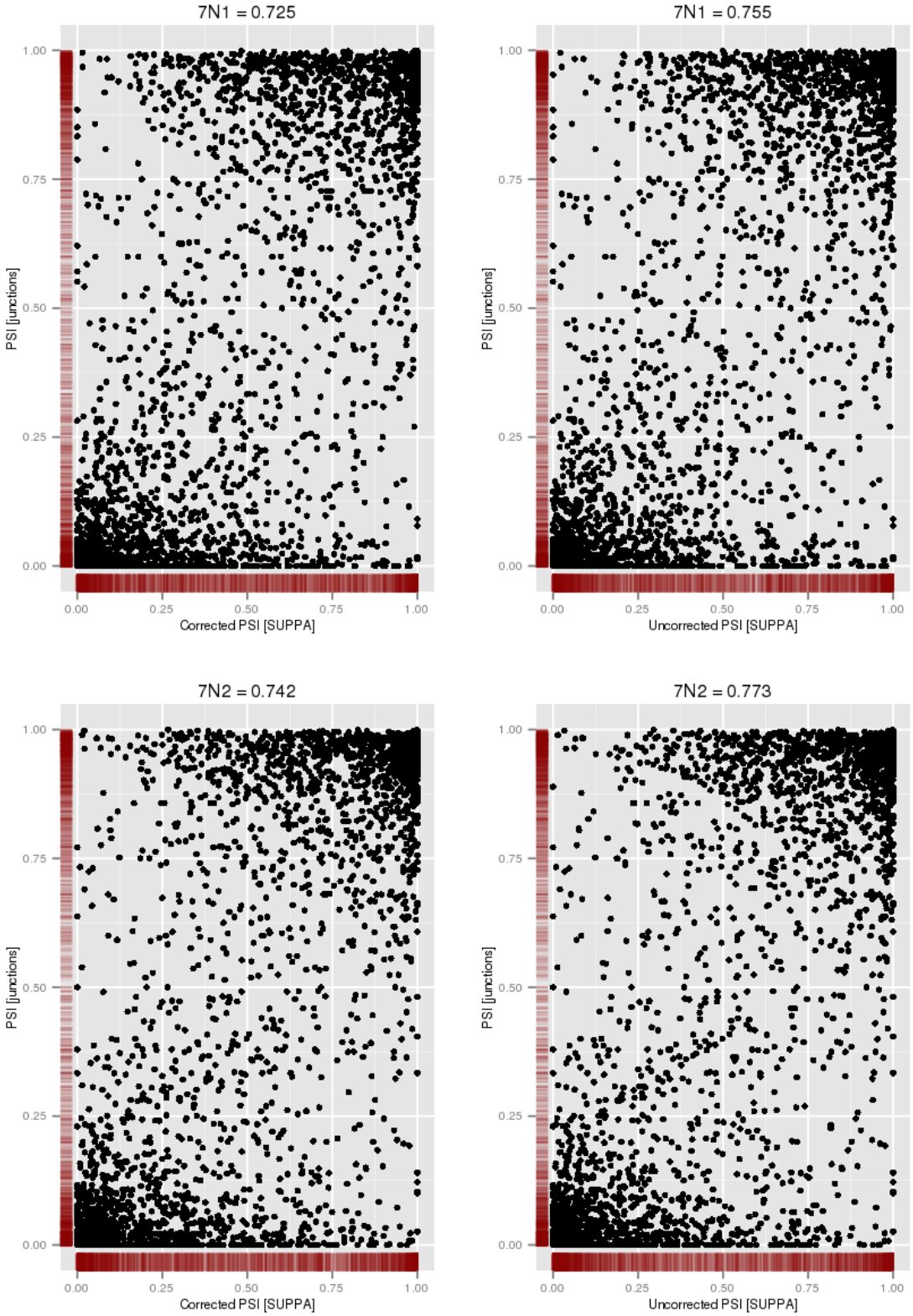
Supplementary Figure 11. Correlation between Ψ_q and Ψ_j for cytosolic RNA-Seq in MCF10 cells. The plots represent the relation between the Ψ_q (PSI [SUPPA]) (x-axis) and Ψ_j (PSI [junction]) (y-axis) for the MCF10 cytoplasmic replicates (10C1 and 10C2) and for corrected and uncorrected Sailfish data. All events are non-overlapping, genes have TPM>1 and events have more than 20 junction reads



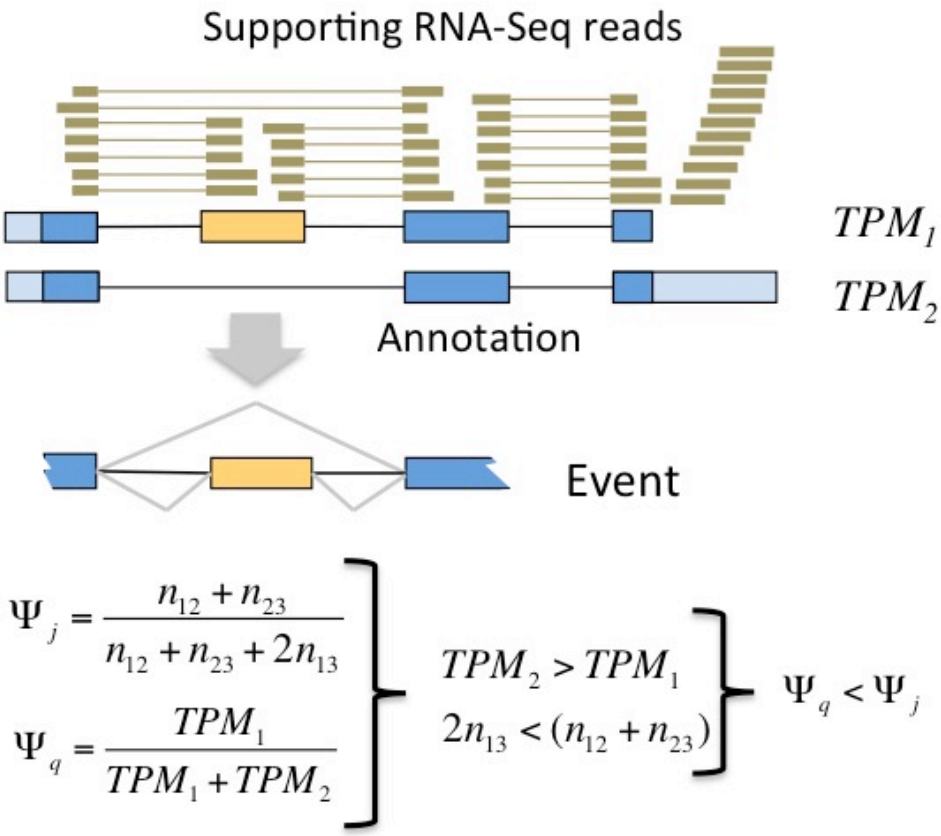
Supplementary Figure 12. Correlation between Ψ_q and Ψ_j for nuclear RNA-Seq from MCF10 cells. The plots represent the relation between the Ψ_q (PSI [SUPPA]) (x-axis) and Ψ_j (PSI [junction]) (y-axis) for the MCF10 nuclear replicates (10N1 and 10N2) and for corrected and uncorrected Sailfish data. All events are non-overlapping, genes have TPM>1 and events have more than 20 junction reads.



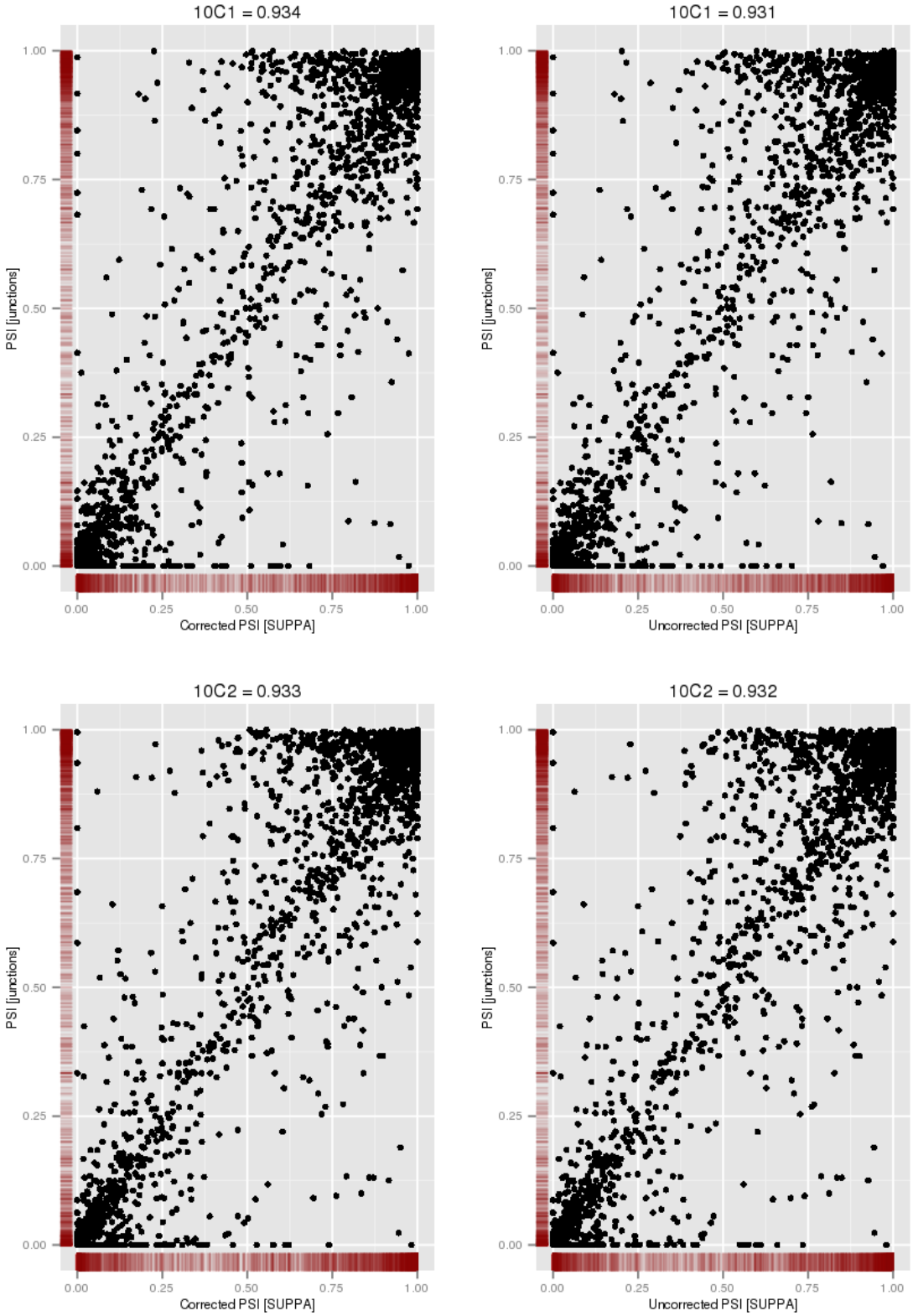
Supplementary Figure 13. Correlation between Ψ_q and Ψ_j for cytosolic RNA-Seq from MCF7 cells. The plots represent the relation between the Ψ_q (PSI [SUPPA]) (x-axis) and Ψ_j (PSI [junction]) (y-axis) for the MCF7 cytoplasmic replicates (7C1 and 7C2) and for corrected and uncorrected Sailfish data. All events are non-overlapping, genes have TPM>1 and events have more than 20 junction reads.



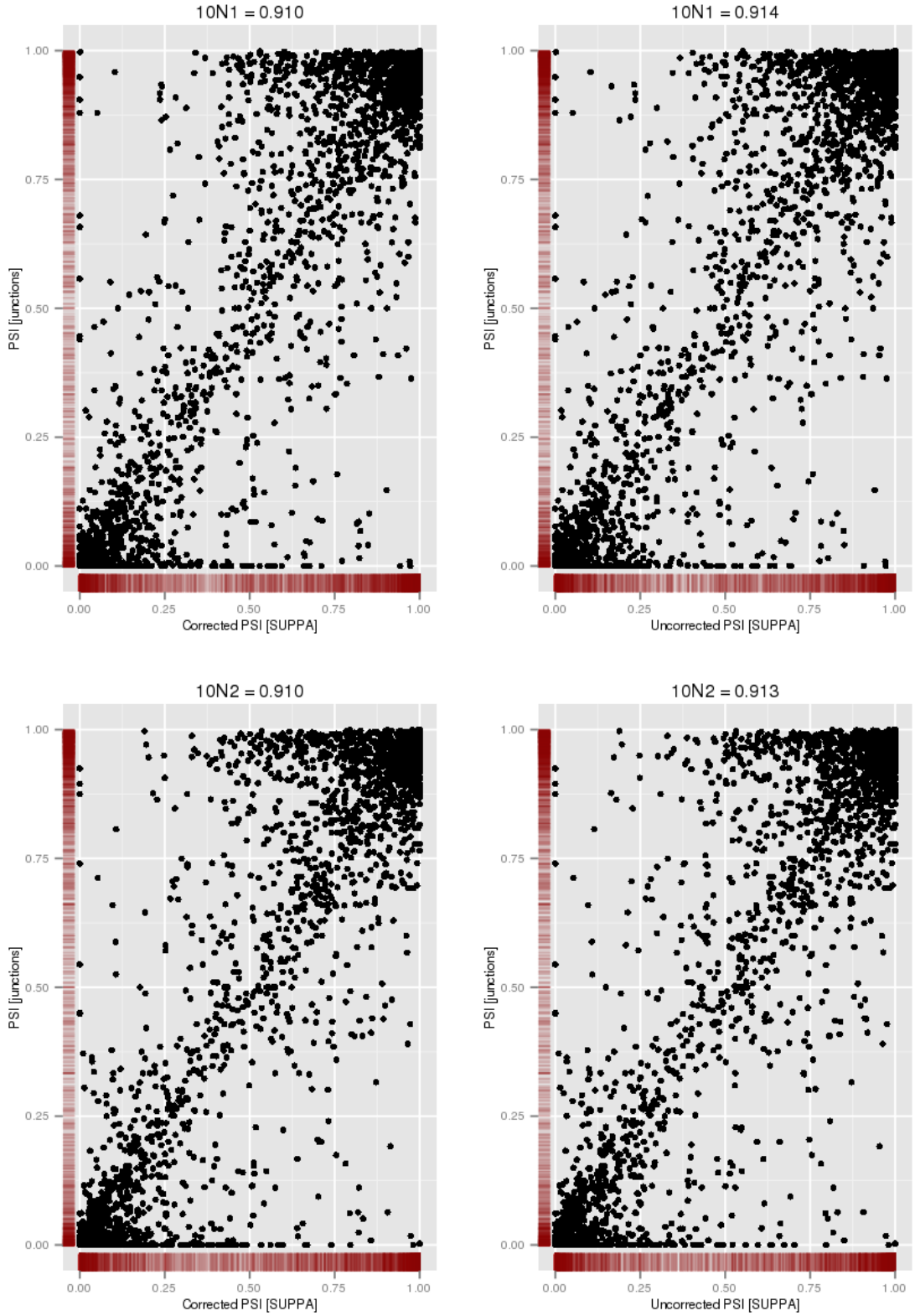
Supplementary Figure 14. Correlation between Ψ_q and Ψ_j for nuclear RNA-Seq from MCF7 cells. The plots represent the relation between the Ψ_q (PSI [SUPPA]) (x-axis) and Ψ_j (PSI [junction]) (y-axis) for the MCF7 nuclear replicates (7C1 and 7C2) and for corrected and uncorrected Sailfish data. All events are non-overlapping, genes have TPM>1 and events have more than 20 junction reads.



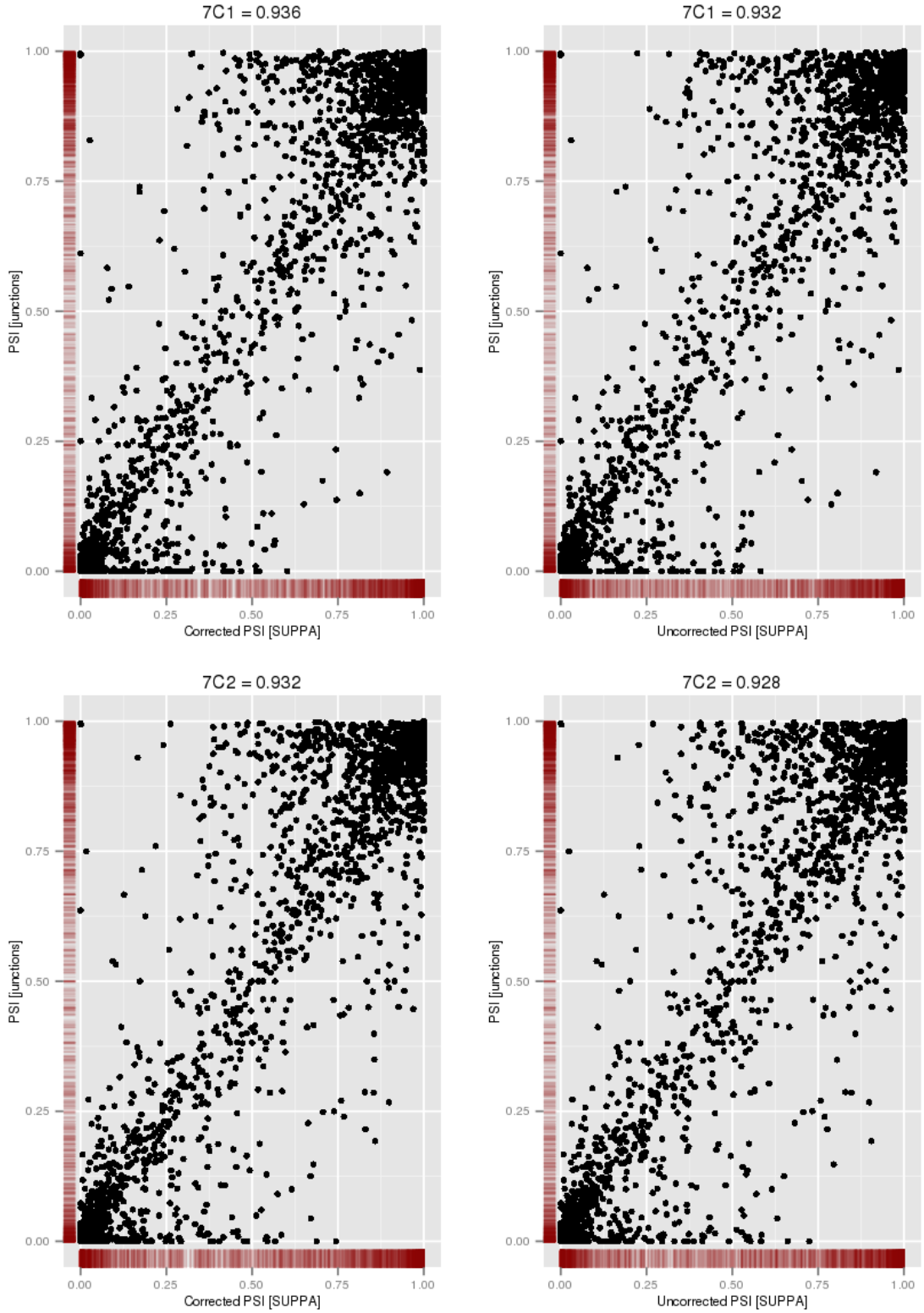
Supplementary Figure 15. Edge biases in transcript abundance quantification. In the figure we illustrate a gene with two transcript isoforms, where only one of them has an 3' UTR. Quantification of the transcripts may give a higher TPM to transcript 2 ($TPM_2 > TPM_1$) due to an excess of reads in the 3'UTR. On the other hand, junction reads may give the opposite result. In the figure, according to the junction reads, transcript 1 appears to have greater support. As a consequence, the PSI calculation according to junction reads will give a high inclusion to the alternative exon, i.e. $\Psi_j \sim 1$. On the other hand, since the transcript that skips the exon gets larger quantification, i.e. $TPM_2 > TPM_1$, PSI calculation based on transcript expression would give the opposite behaviour, i.e. $\Psi_j << 1$. To mitigate these effects, we provided also the benchmarking using the quantification of event in coding regions.



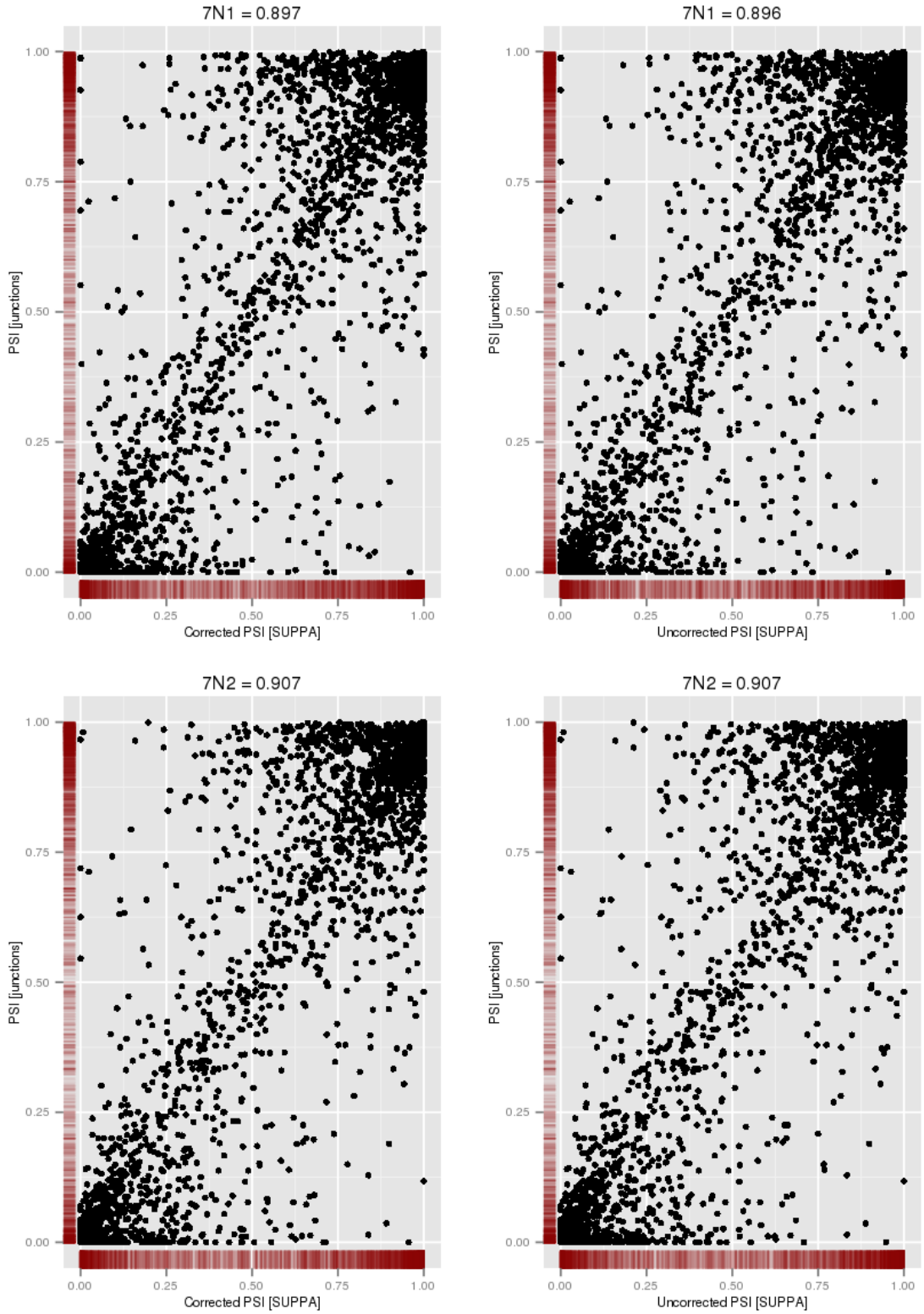
Supplementary Figure 16. Correlation between Ψ_q and Ψ_j for cytosolic RNA-Seq in MCF10 cells in coding regions. The plots represent the relation between the Ψ_q (PSI [SUPPA]) (x-axis) and Ψ_j (PSI [junction]) (y-axis) for the MCF10 cytoplasmic replicates (10C1 and 10C2) and for corrected and uncorrected Sailfish data using only events in coding regions. Quantification with Sailfish was performed only for the CDS annotations. As before, all events are non-overlapping, genes have TPM>1 and events have more than 20 junction reads.



Supplementary Figure 17. Correlation between Ψ_q and Ψ_j for nuclear RNA-Seq from MCF10 cells in coding regions. The plots represent the relation between the Ψ_q (PSI [SUPPA]) (x-axis) and Ψ_j (PSI [junction]) (y-axis) for the MCF10 nuclear replicates (10N1 and 10N2) and for corrected and uncorrected Sailfish data using only events in coding regions. Quantification with Sailfish was performed only for the CDS annotations. As before, all events are non-overlapping, genes have TPM>1 and events have more than 20 junction reads.



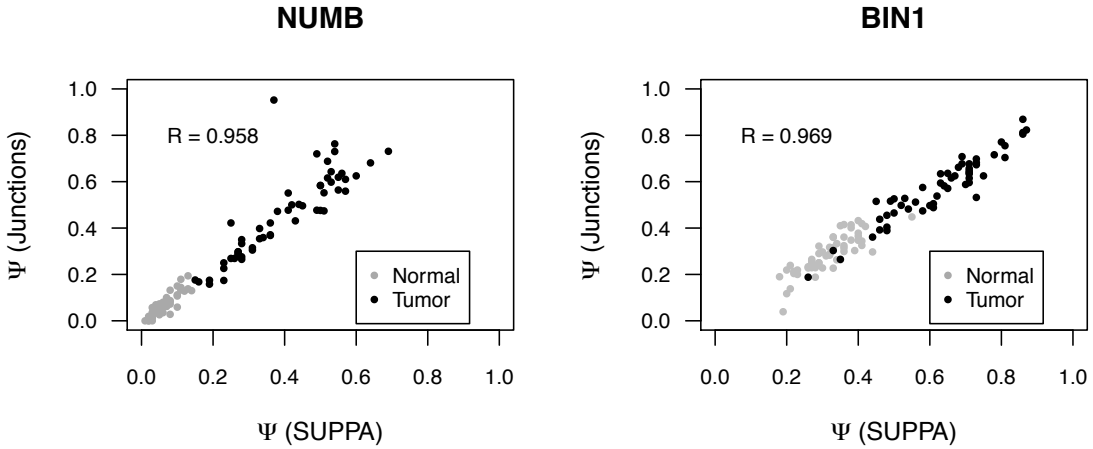
Supplementary Figure 18. Correlation between Ψ_q and Ψ_j for cytosolic RNA-Seq from MCF7 cells in coding regions. The plots represent the relation between the Ψ_q (PSI [SUPPA]) (x-axis) and Ψ_j (PSI [junction]) (y-axis) for the MCF7 cytoplasmic replicates (7C1 and 7C2) and for corrected and uncorrected Sailfish data using only events in coding regions. Quantification with Sailfish was performed only for the CDS annotations. As before, all events are non-overlapping, genes have TPM>1 and events have more than 20 junction reads.



Supplementary Figure 19. Correlation between Ψ_q and Ψ_j for nuclear RNA-Seq from MCF7 cells in coding regions. The plots represent the relation between the Ψ_q (PSI [SUPPA]) (x-axis) and Ψ_j (PSI [junction]) (y-axis) for the MCF7 nuclear replicates (7N1 and 7N2) and for corrected and uncorrected Sailfish data using only events in coding regions. Quantification with Sailfish was performed only for the CDS annotations. As before, all events are non-overlapping, genes have TPM>1 and events have more than 20 junction reads.

Benchmarking with TCGA data

The abundance of every transcript per sample was calculated in transcripts per million (TPM) from the transcript-estimated read counts provided by TCGA (https://tcga-data.nci.nih.gov/tcga/) for a total of 4114 tumor samples. No further normalization on the TPM values was performed. A total of 40411 alternative splicing events were calculated with SUPPA (*generateEvents*) using the UCSC gene annotation (June 2011) used by TCGA in GTF format. The data used is available at http://dx.doi.org/10.6084/m9.figshare.1061917



Supplementary Figure 20. Correlation (Pearson R) of the Ψ values calculated with SUPPA and with junction reads, using RNA-Seq from lung adenocarcinoma tumor and paired normal samples for two exon-skipping events.

REFERENCES

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

[2] 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.

[3] 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.

[4] 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.

[5] Pervouchine DD, Knowles DG, Guigó R. Intron-centric estimation of alternative splicing from RNA-seq data. Bioinformatics. 2013 Jan 15;29(2):273-4.

[6] 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.