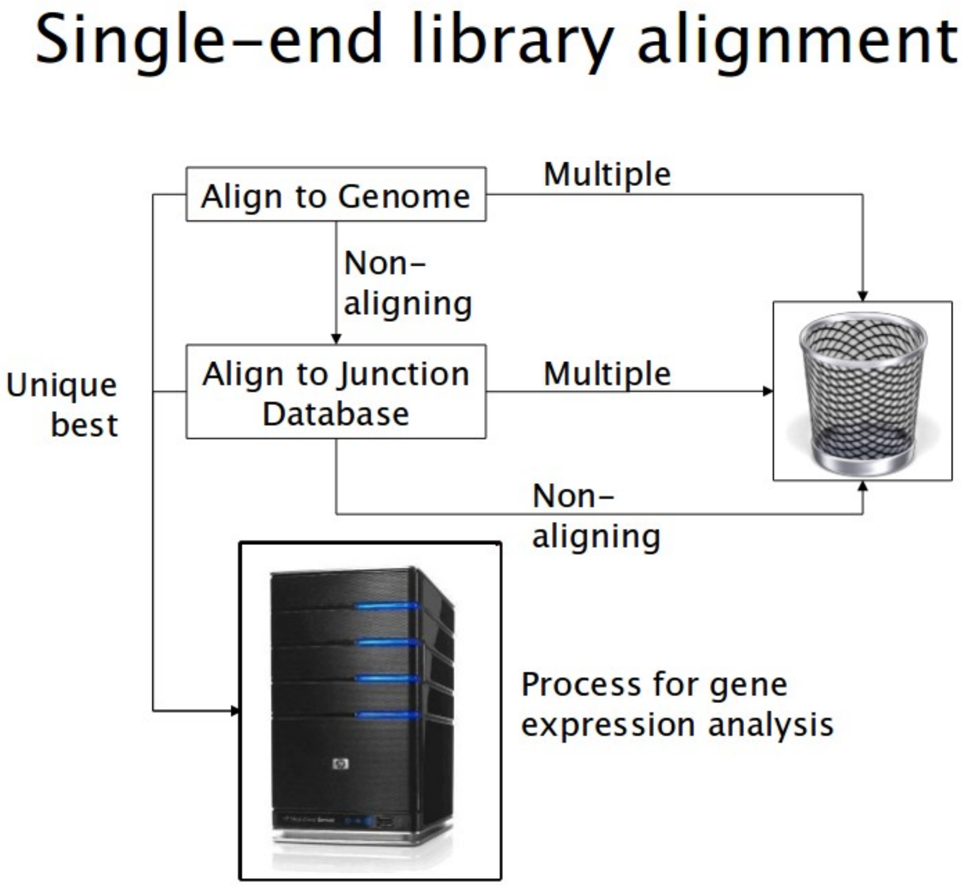
Supplementary Figures

Supplementary Figure 1. Single-end alignment flowchart.

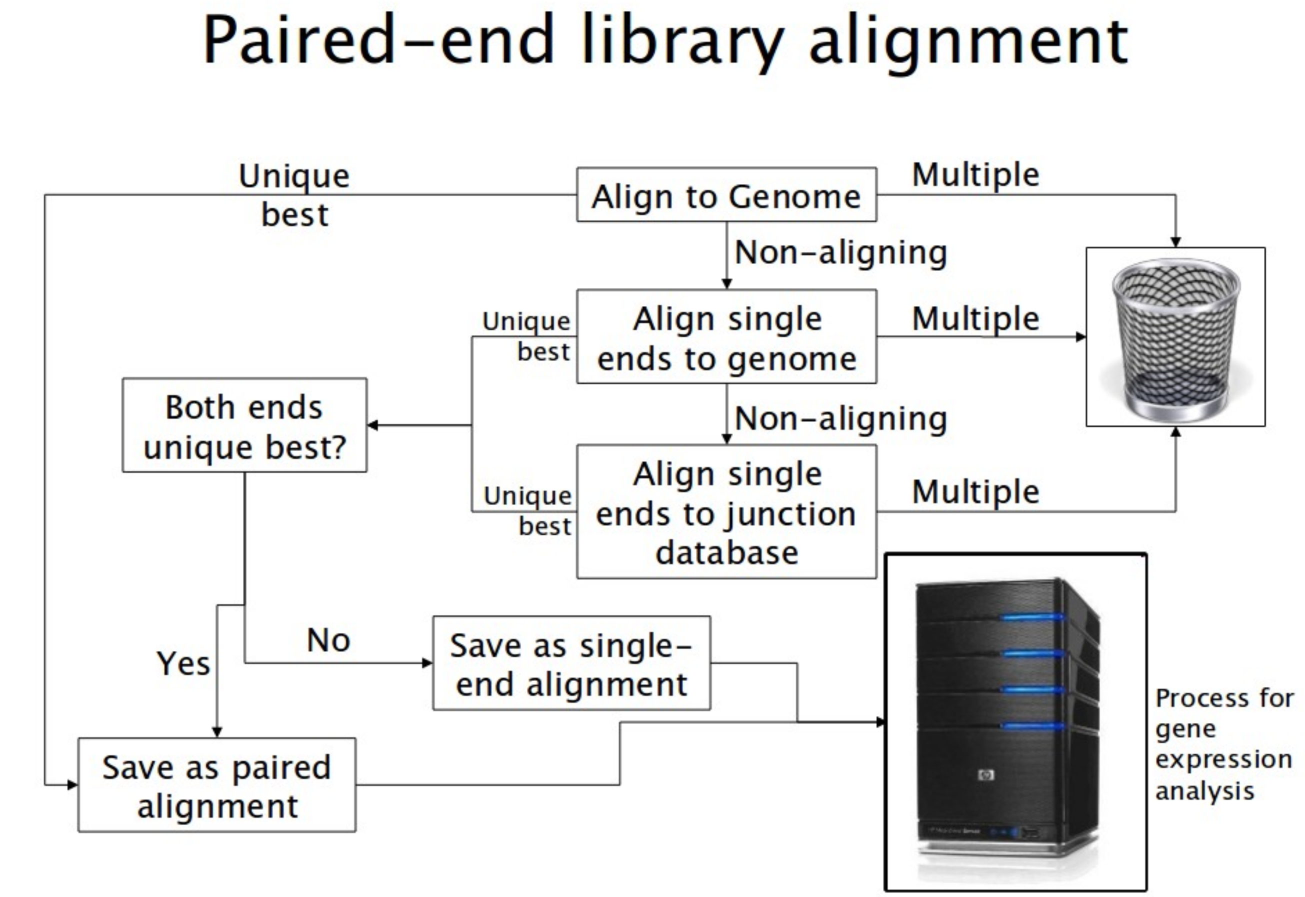
Single end reads are aligned to the genome with bowtie using the arguments “-k 2 -m 200 --best --strata”. Only best hits are considered. Multiple-best-aligning reads are discarded (right branch). Non-aligning reads are then aligned to the junction database using bowtie with the same options. Now multiple-best-aligning reads as well as non-aligning reads are discarded. Unique best genome and junction alignments are then further processed by the pipeline for gene expression analysis.


Single–end library alignment
Align to Genome
Multiple
Non–aligning
Align to Junction Database
Multiple
Unique best
Non–aligning
Process for gene expression analysis

Supplementary Figure 2. Paired-end alignment flowchart.

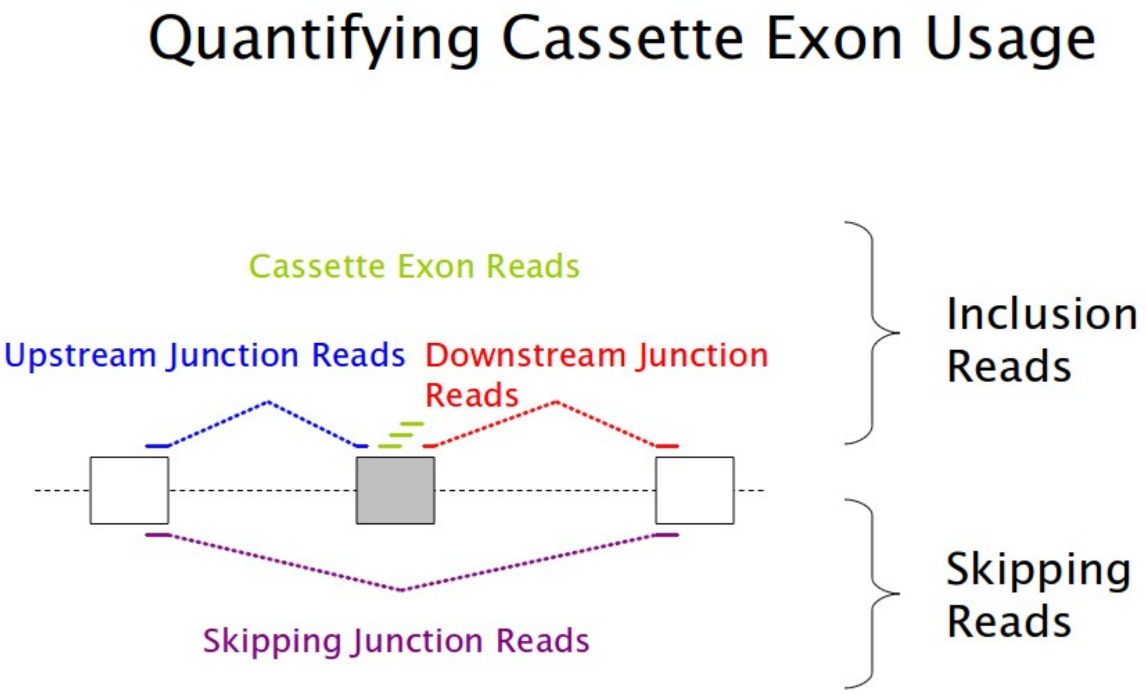
Paired end reads are aligned to the genome using bowtie with arguments “-X 20000 -k 2 -m 2”. Multiple-aligning reads are discarded. Single-ends of non-pair-aligning reads are then again aligned to the genome with bowtie (arguments “-k 2 -m 200 --best --strata”) and multiple-aligning reads are discarded. Non-aligning reads from this step are then aligned to the junction database using the same arguments. Finally multiple-aligning as well as non-aligning reads are discarded.

The unique best single-end alignments from the genome and junction steps are then re-paired. These pairs are combined with the genomic paired-end alignments from the first step, then, along with the single-end alignments that could not be re-paired, are sent downstream for further gene expression analysis processing.


Paired–end library alignment
Unique best
Align to Genome
Multiple
Non–aligning
Unique best
Align single ends to genome
Multiple
Both ends unique best?
Non–aligning
Unique best
Align single ends to junction database
Multiple
Yes
No
Save as single–end alignment
Save as paired alignment
Process for gene expression analysis

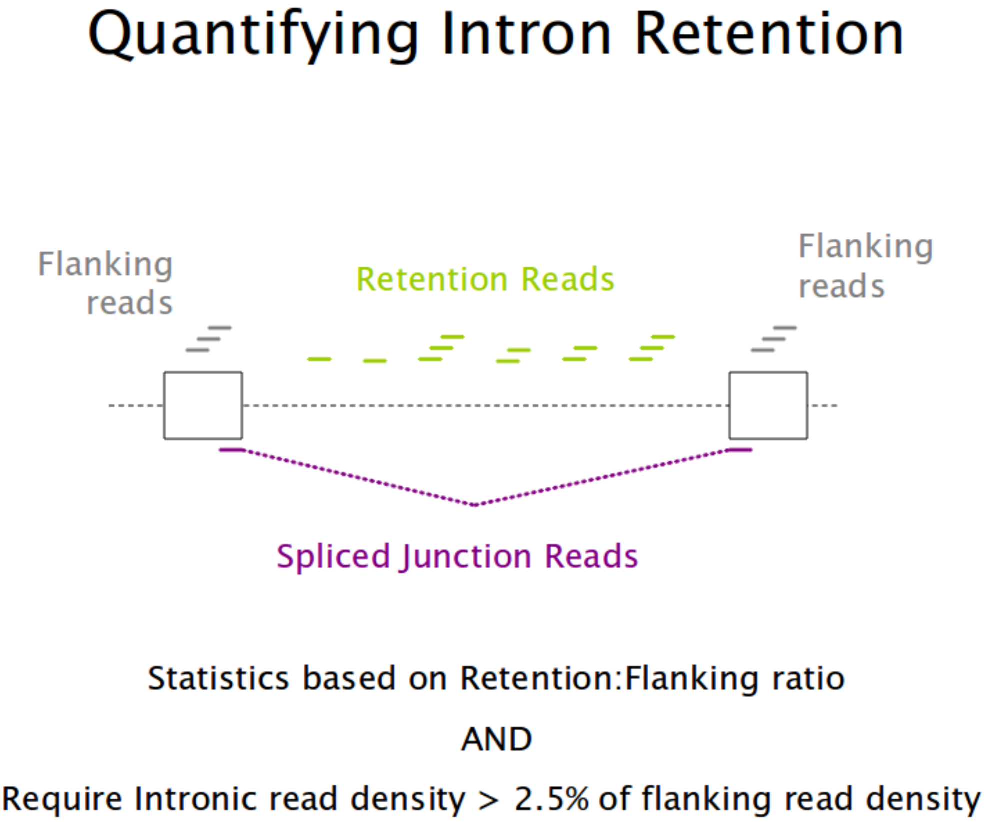
Supplementary Figure 3. Quantifying cassette exon usage.

Cassette exons are quantified by comparing inclusion reads to skipping reads. Inclusion reads include junction reads ending or beginning at either splice site of that exon as well as exon body reads. Skipping reads are junction reads that flank the candidate exon but are anchored in known splice sites of the host gene. The main reason for the anchor requirement is to prevent the “skipping reads” count for exons of intron-encoded genes from including junction reads of their host introns. For paired-end data reads are counted as inclusion or skipping if either mate is a skipping or inclusion alignment.



Supplementary Figure 4. Quantifying intron retention.

For intron retention, the number of reads aligning to the intron is compared to the number aligned to flanking introns (actually to the “locally constitutive” transcript---see Supplementary Figure 5). The number of splice junction reads is counted but not considered for statistical purposes. As a default, only introns with a read density (RPKM) at least 2.5% of the flanking regions are considered.



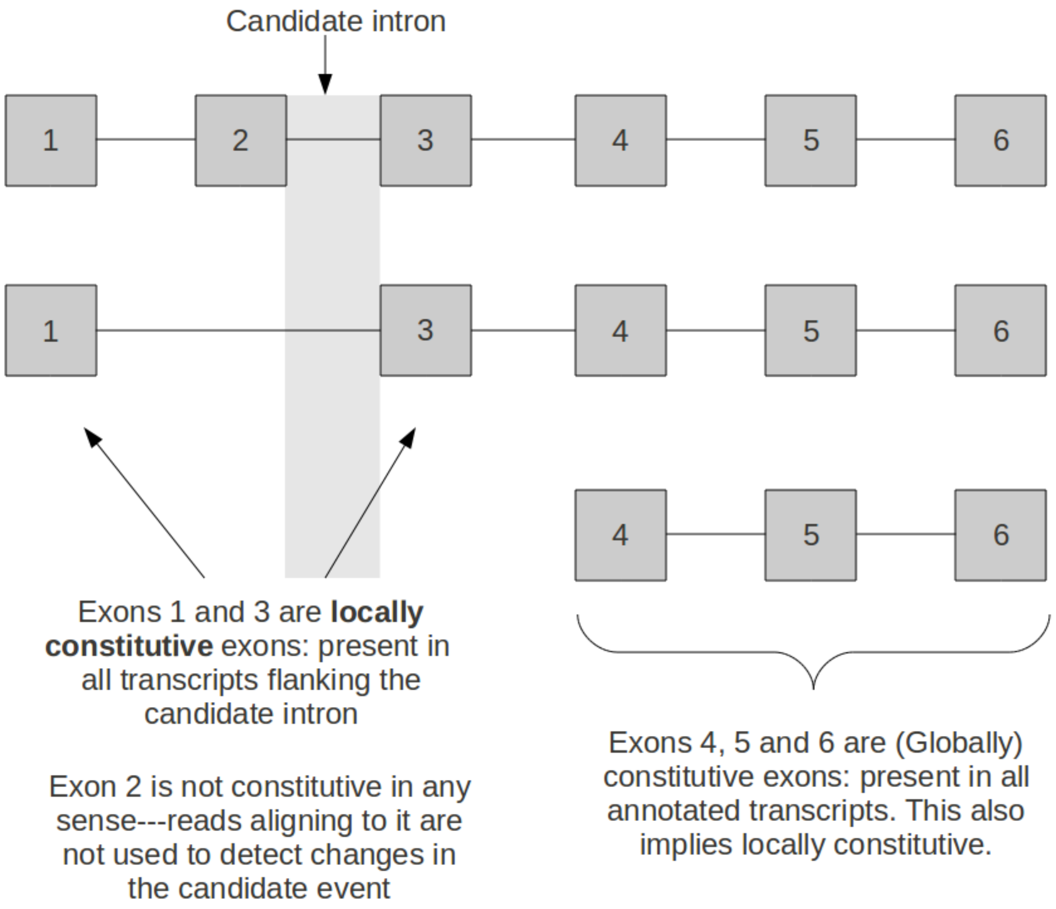
Quantifying Intron Retention

Statistics based on Retention:Flanking ratio

AND

Require Intronic read density > 2.5% of flanking read density

Supplementary Figure 5. Example of “locally constitutive” exons for a candidate intron.



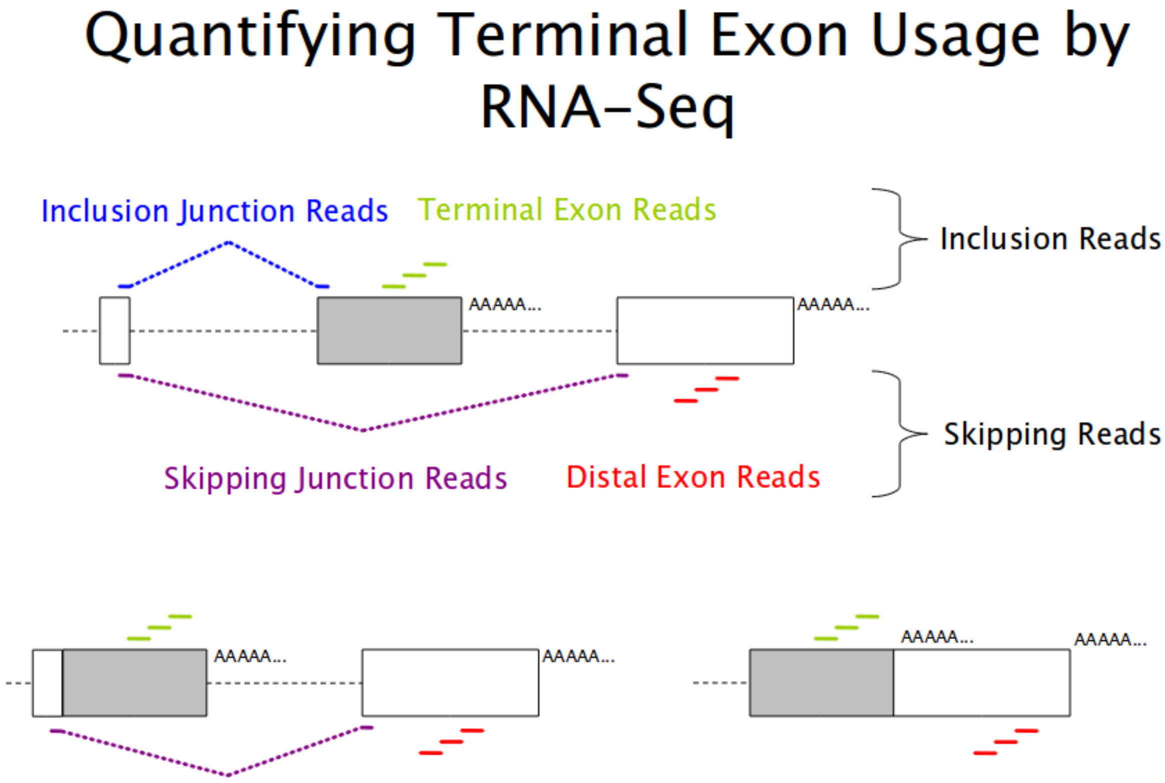
Exons 1 and 3 are **locally constitutive** exons: present in all transcripts flanking the candidate intron

Exon 2 is not constitutive in any sense---reads aligning to it are not used to detect changes in the candidate event

Exons 4, 5 and 6 are (Globally) constitutive exons: present in all annotated transcripts. This also implies locally constitutive.

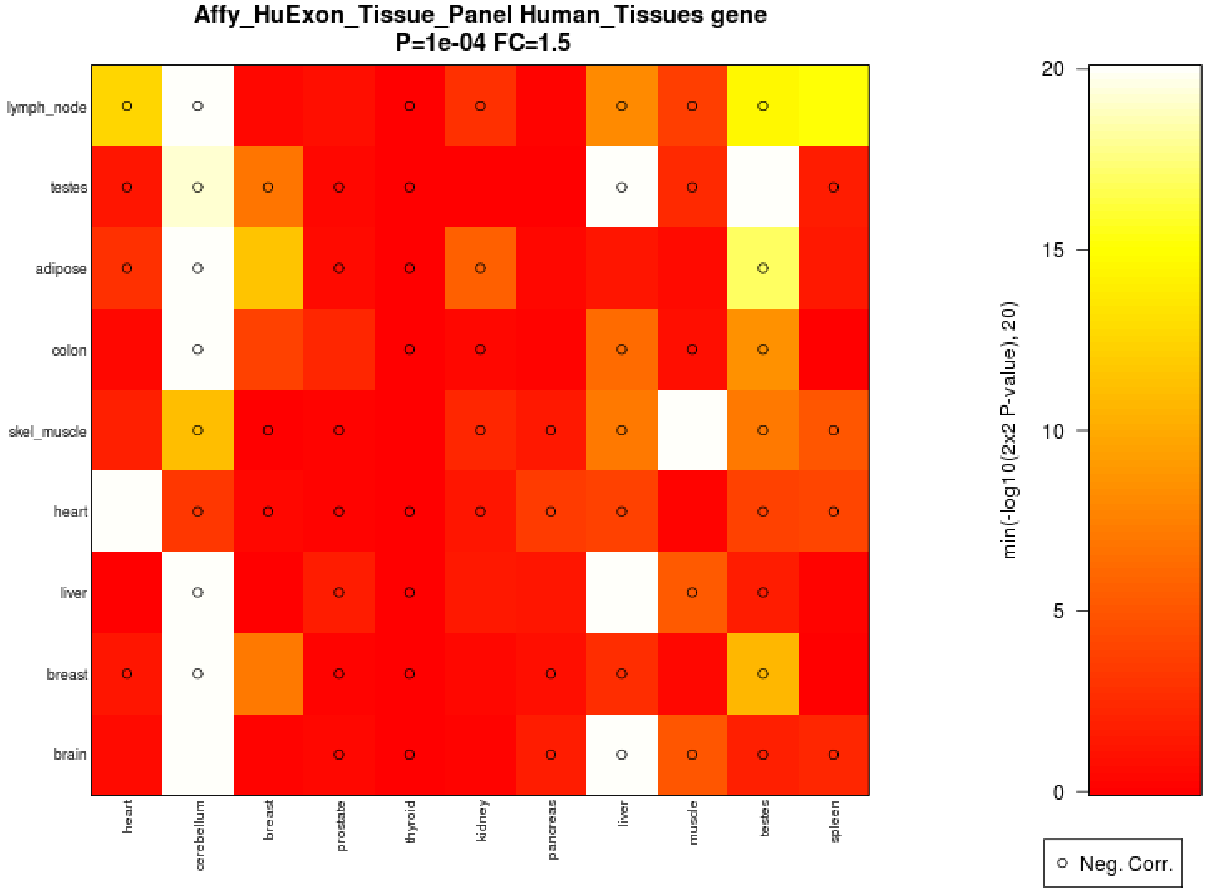
Supplementary Figure 6. Quantifying terminal exon usage.

Alternative terminal exons (ATEs) are quantified by comparing proximal reads to distal reads. This figure shows as an example the 3' end, but the algorithm is identical for both 5' and 3' ends. For each gene, all transcript termini but the most distal are considered as alternative terminal exon events. (These events are really testing reads associated with the termini, and not with, for example, the splice sites of the terminal exons). Each terminus is categorized as “intronic” if it is within the intron of a more distally terminating transcript (top and bottom left) and “exonic” if in the exon of a more distally terminating transcript (bottom right). All termini are exonic or intronic, and it is also possible to be both (not shown). Distal reads include genomic reads that align to distal exon bodies or, in the case of intronic termini, junction reads that splice over the terminus but are anchored in known splice sites of the gene. Proximal reads for intronic termini are those that align to the terminal exon but are downstream of the last 5' splice site (for 3' termini; upstream of the first 5' splice site for 5' termini) proximal to the terminus. For purely exonic termini the entire exon through to the terminus is used.



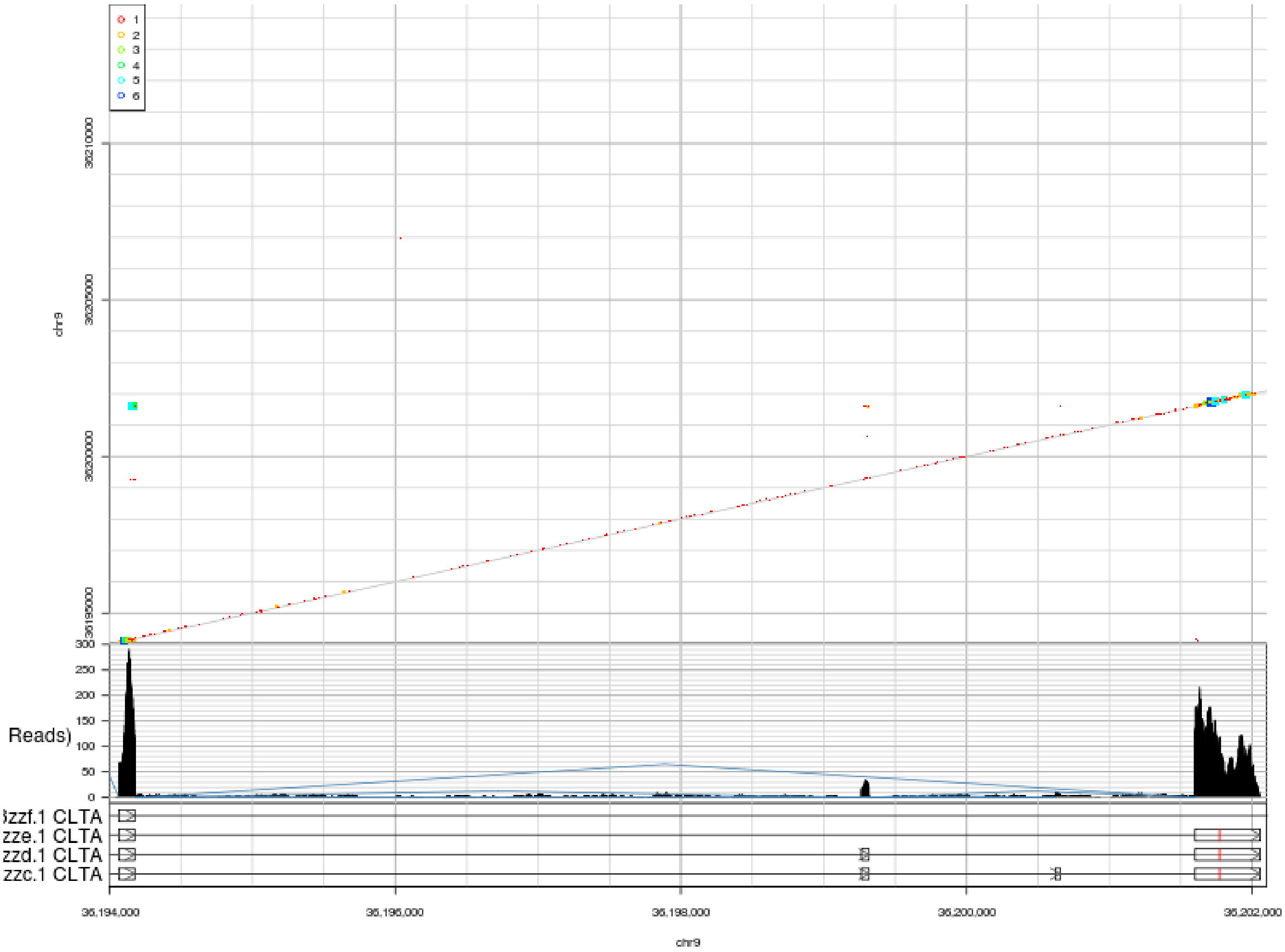
Supplementary Figure 7. Example of **heatmap** plot from *ExpressionPlot.*

Example plots from *ExpressionPlot* comparing tissue-enriched gene expression in human exon array tissue panel data (*x*-axis) and human RNA-Seq tissue panel data (*y*-axis). For each pair of tissues (t_e, t_r), where t_e is a tissue from the exon array panel and t_r is a tissue from the RNA-Seq panel, a 2x2 contingency table was constructed counting the number of genes significantly changed up and down in both tissues and data sets. Fisher's exact test or the chi-squared test was then applied and the *P*-value was then colored into the square (t_e, t_r). Dots indicate anti-correlation. Red colors are not significant, and significance increases through yellow and white. Notable positive correlations include (cerebellum, brain), (breast, breast), (liver, liver), (heart, heart), (muscle, skel_muscle), (breast, adipose), (testes, testes) and (spleen, lymph_node). None of the exon array tissues correlate well with colon. Adipose RNA-Seq correlates better than breast RNA-Seq with exon array breast by this measure but this may be due to deeper sequencing.



Supplementary Figure 8. Example of **pairplot** plot from *ExpressionPlot.*

This pairplot shows the vicinity of an alternatively spliced exon in Clta (clathrin light chain A) in a paired-end data set. Bottom track shows known transcripts in this region. Second track from bottom shows read pileup for this sample throughout the region. Top section shows the “pairplot”. At position (*x*, *y*), where *x* and *y* are coordinates within the chromosomal window, a point is plotted if there is a paired-end read whose plus-strand aligned read has its last base aligned to *x* and its minus-strand aligned read with its last base aligned to *y* (reads aligning to different chromosomes, same strand, or with *x* > *y - readlen* are not included in this plot, but counted and barplotted by the **pairdist** tool online). The color and weight of the point indicate the number of such reads. Multiple *y* values for a single *x* value indicate alternative splicing. For example, there are a few small red dots above the first exon corresponding to pairs spanning the first and second exon. The larger cyan dot above indicates reads spanning the first and fourth (last in this window) exon, and, concordant with the pileup plots, indicate that the skipping isoform is more abundant. Above the second exon one can also see a few reads going to the third as well as the fourth. Thus the double-inclusion isoform (1-2-3-4) is also present, but at an ever lower level than the single inclusion 1-2-4 isoform. No paired-end reads support a 1-3 splice, so the 1-3-4 isoform, if it exists at all, is scarce. This particular data set had an insert size close to twice the read length. This is why most of the reads lie along the gray diagonal *y*=*x.* (The underlying data set has been de-identified).


chr9
Reads)
3zzf.1 CLTA
zze.1 CLTA
zzd.1 CLTA
zzc.1 CLTA
chr9