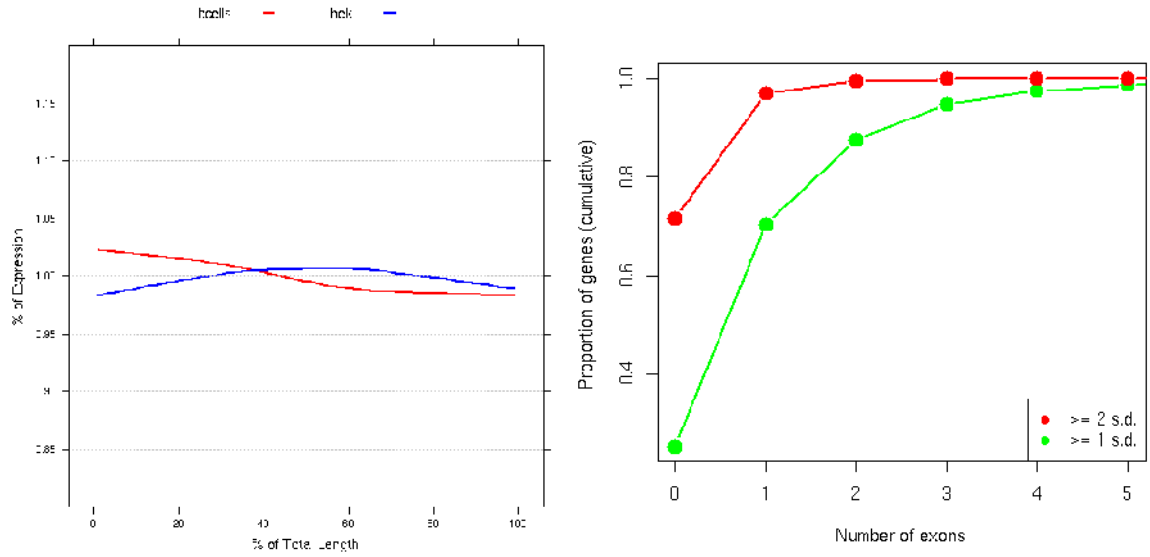
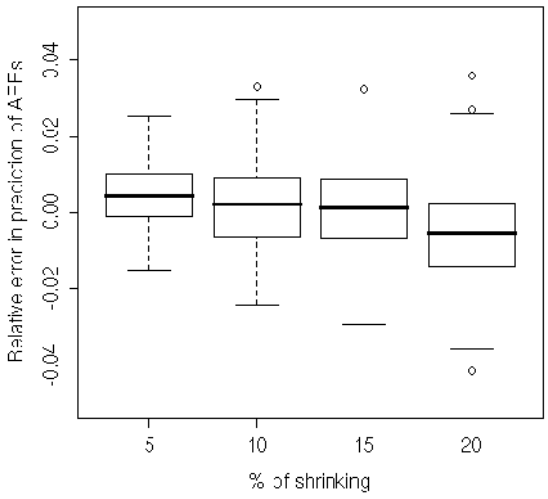
Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments

Hugues Richard, Marcel H. Schulz, Marc Sultan, Asja Nürnberger, Sabine Schrinner, Daniela Balzereit, Emilie Dagand, Axel Rasche, Hans Lehrach, Martin Vingron, Stefan A. Haas, Marie-Laure Yaspo.

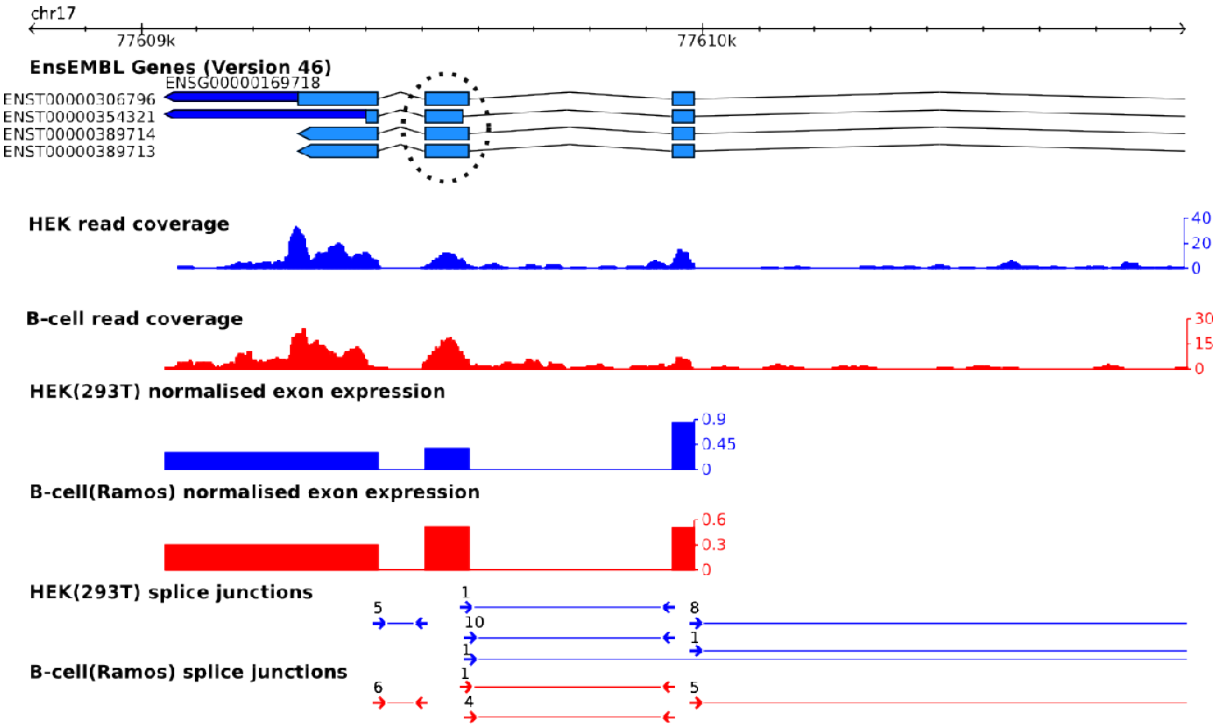
Supplemental Figures S1 to S12, Supplemental Methods and Supplemental Table Legends



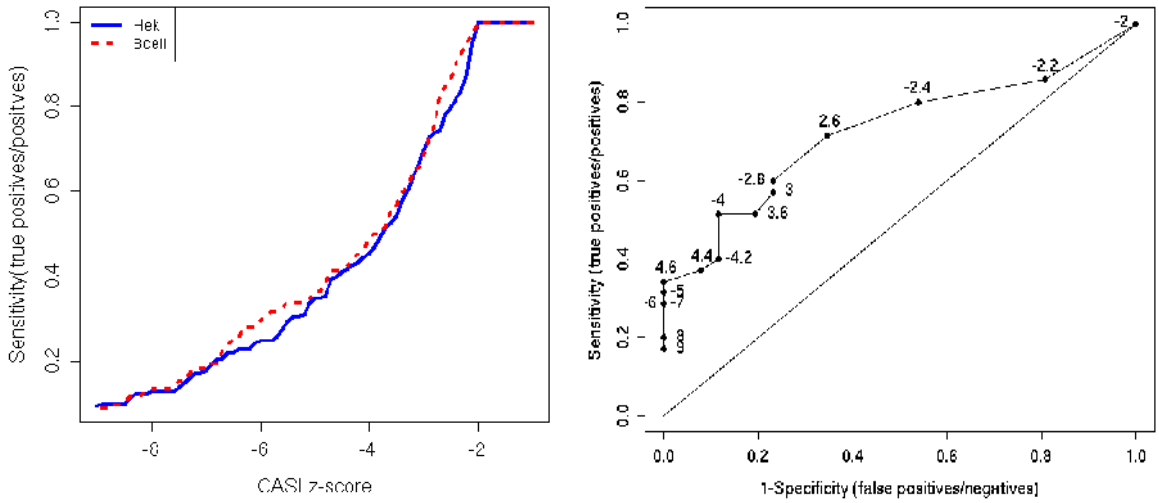
Supplemental Figure S 1: Read distribution across the gene and variability among exons. Left) Graph showing the percentage of reads mapping to each percentile of the length of Ensembl (v.46) genes. For each gene, all exons-except the first and last exon- were assembled to the longest possible transcript. Genes were also filtered for a minimum of 5 reads and a size below 15 kb (10,365 genes in HEK cells and 8,863 genes in B cells). The read distribution is homogeneous across the length of internal exons of the genes. Right) The cumulative proportion of multi-exon genes is reported as a function of the number of exons which expression level deviates more than one (green) or two (red) standard deviations (s.d.) from the average expression level. For instance, approximately 70% of all genes have at most one exon expressed within one s.d. of the average gene expression.



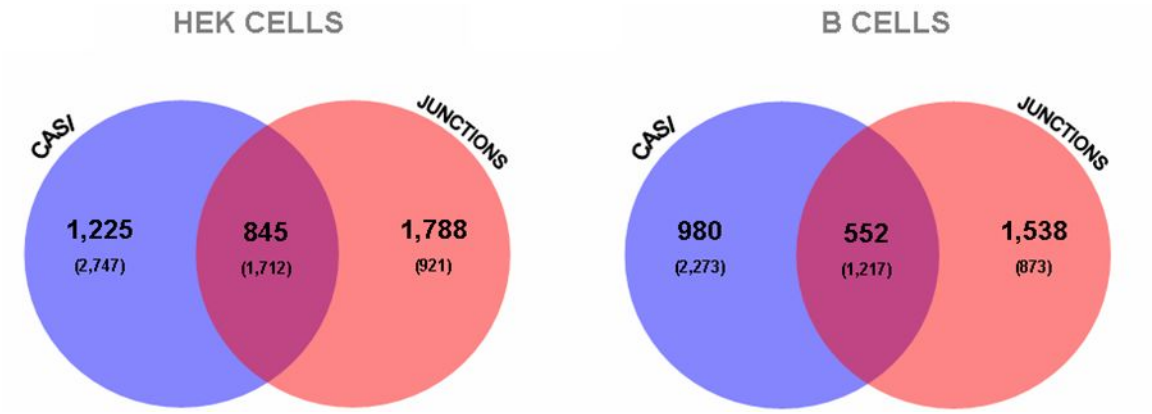
Supplemental Figure S 2: **Robustness estimation of CASI predictions for Hek data by bootstrapping**. The change in predicted AEEs is shown relative to the total number of predictions for the whole dataset (y-axis) and for 500 bootstrap samples. The x-axis shows the reduction in length that was introduced to an exon at random (p=0.25). Boxplots show the relative error for -4. Every box in a boxplot describes the 50% distribution; the line displays the median. The whiskers display the 1.5 IQR and remaining outliers are shown as dots.



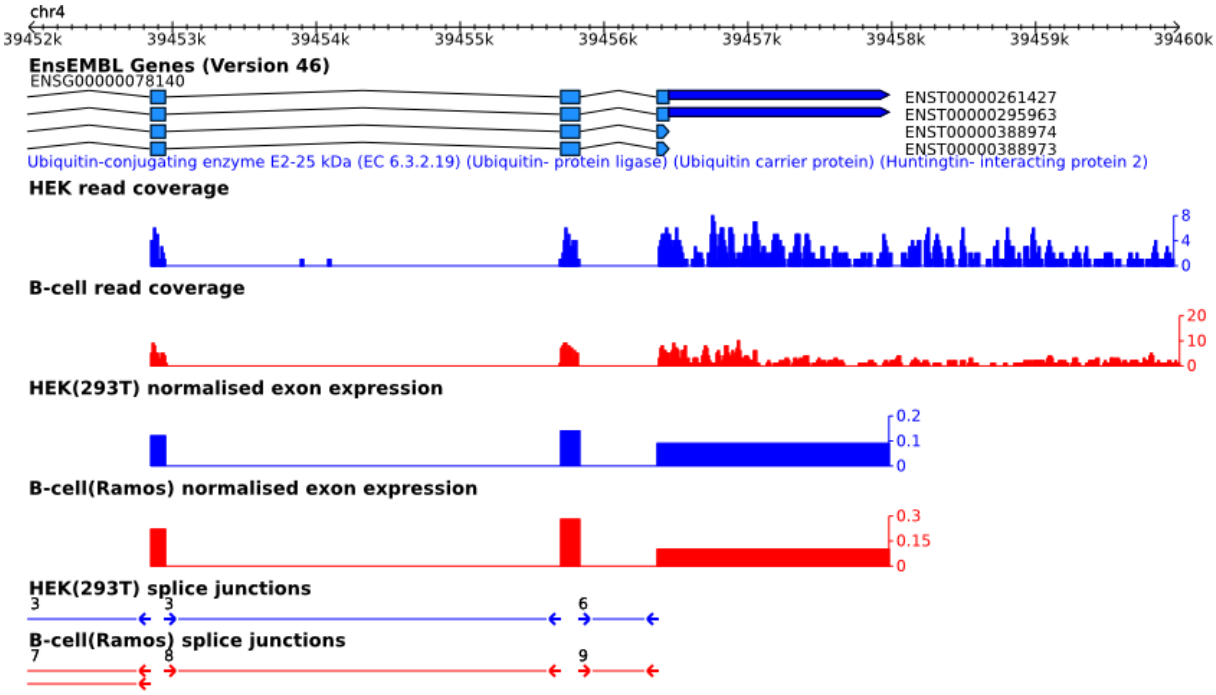
Supplemental Figure S 3: Alternative acceptor splice site in DUS1L. According to the annotation (ENSEMBL v.46), the circled exon in DUS1L shows an alternative acceptor splice site. This exon was predicted as AEE by the CASI method and had further one junction read identifying this event.



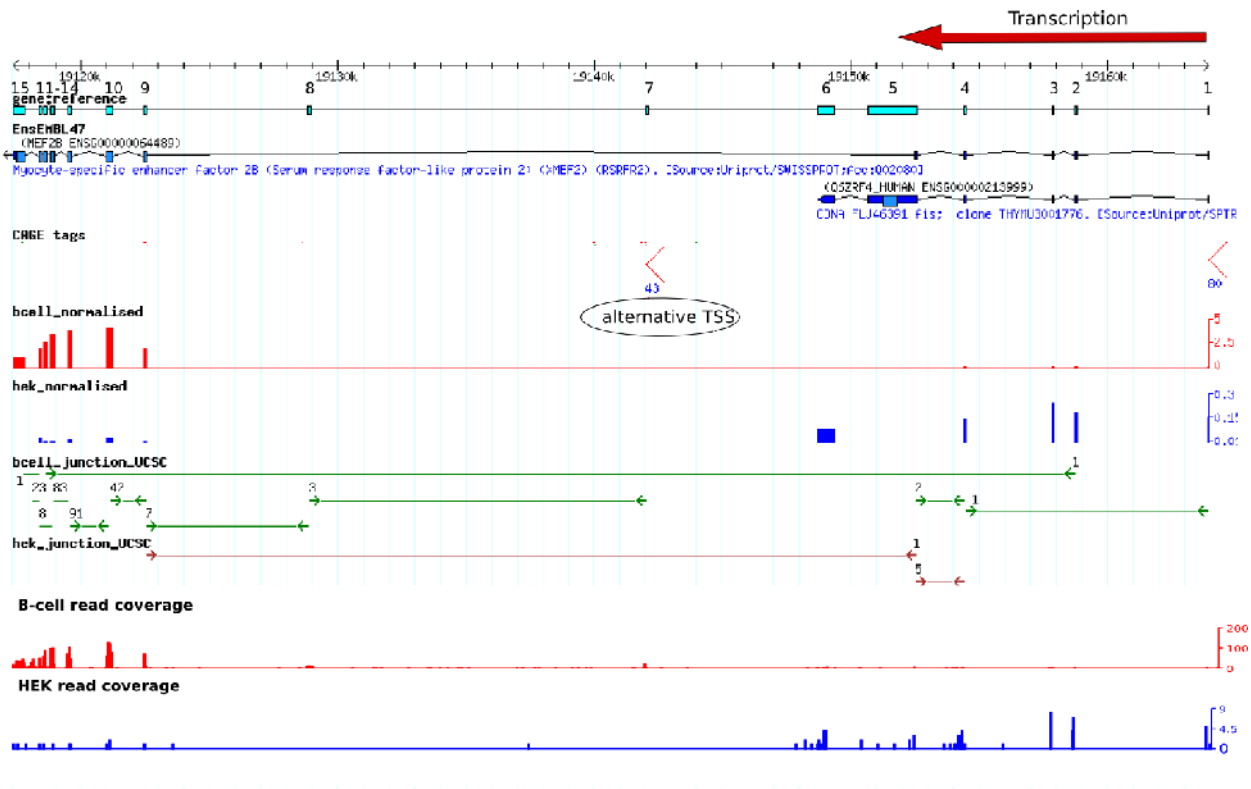
Supplemental Figure S 4: Validation of the CASI method with Splice junctions and RT-PCR. Left) Sensitivity of CASI predictions compared to splice junction reads. All AEEs detected by at least 3 splice junction reads are taken as the positive set. The y-axis shows the percentage of CASI AEEs that overlap with the positive set (Sensitivity) for different values of z on the x-axis. Right) ROC curve of RT-PCR results (positive/negative) testing 61 AEEs predicted by the CASI method. Each exon tested by RT-PCR was associated to its corresponding CASI value. CASI values are indicated for each data point. The best qualifier uses a CASI of -4 with a specificity of 89% and sensitivity of 51%. Note that the sensitivity for CASI predictions derived by comparison to splice junction reads (left figure) is highly similar with ~48% sensitivity for a z-score of -4 in both cell lines.



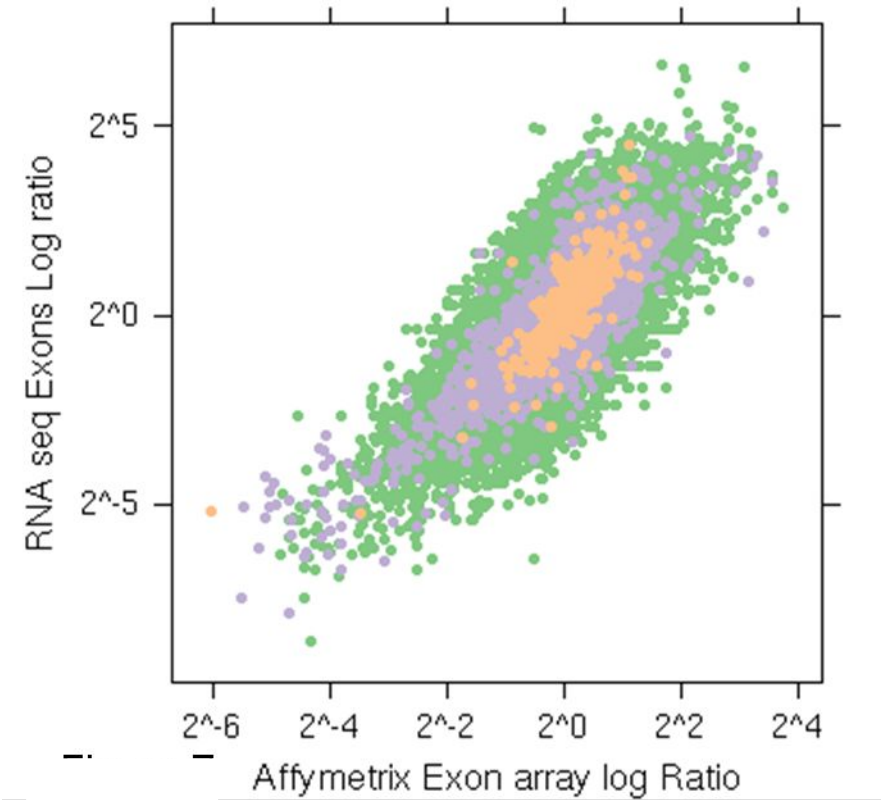
Supplemental Figure S 5: Comparison of alternatively spliced events identified with CASI (<=-4) and junctions reads. The venn diagrams show the number of genes, for each cell line, with at least one AEE according to CASI (blue) or splice junction reads (red). A gene is selected if any of its exons (3', 5' or internal exons) is flagged as an AEE. The comparison is based on the total set of genes analysed with CASI (12,140 in HEK and 10.417 for B cells). The numbers in brackets indicate the results for CASI<= -2.



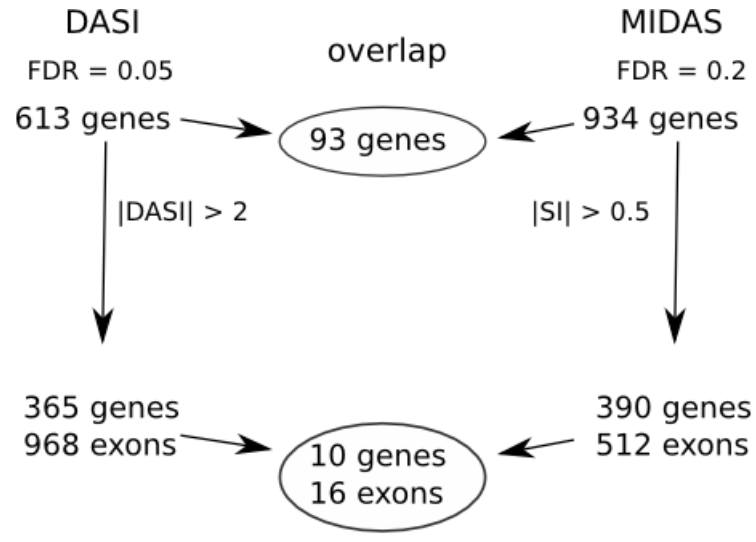
Supplemental Figure S 6: Evidence of an alternative polyadenylation site in HIP2. The figure shows the 3' end of HIP2 as annotated in ENSEMBL v.46 and the read coverage obtained by RNA-Seq for HEK (blue) and B (red) cells. In B cells, the drop in read coverage along the last exon suggests the presence of at least two alternatively polyadenylated forms specific to this cell line. HIP2 was previousely shown to be alternatively polyadenylated in proliferating T cells (Sandberg, Neilson, Sarma, Sharp and Burge 2008).



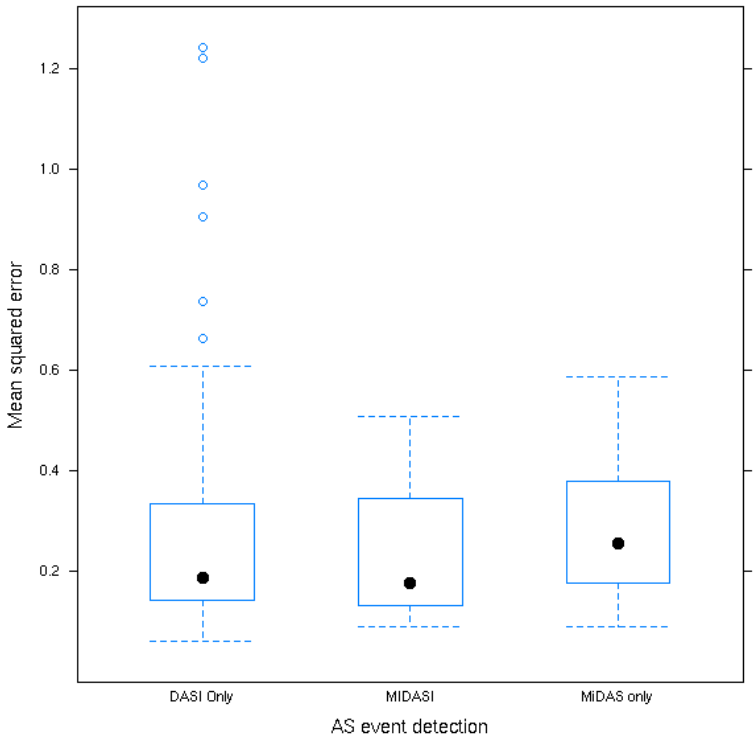
Supplemental Figure S 7: Snapshot of MEF2B in Gbrowse. This example shows that DASI enables the identification of alternative promoter usage events. The gene structure at the top is derived from EST data (Genenest cluster Hs78881) and includes exon 7 and 8, which were not annotated in the ENSEMBL database (v.46). The expression profile in both cell lines suggests an alternative transcript starting at exon 7, which is highly expressed in B cells and not in HEK cells. This is supported by CAGE tag evidences nearby exon 7, suggesting the presence of an alternative transcription start site.

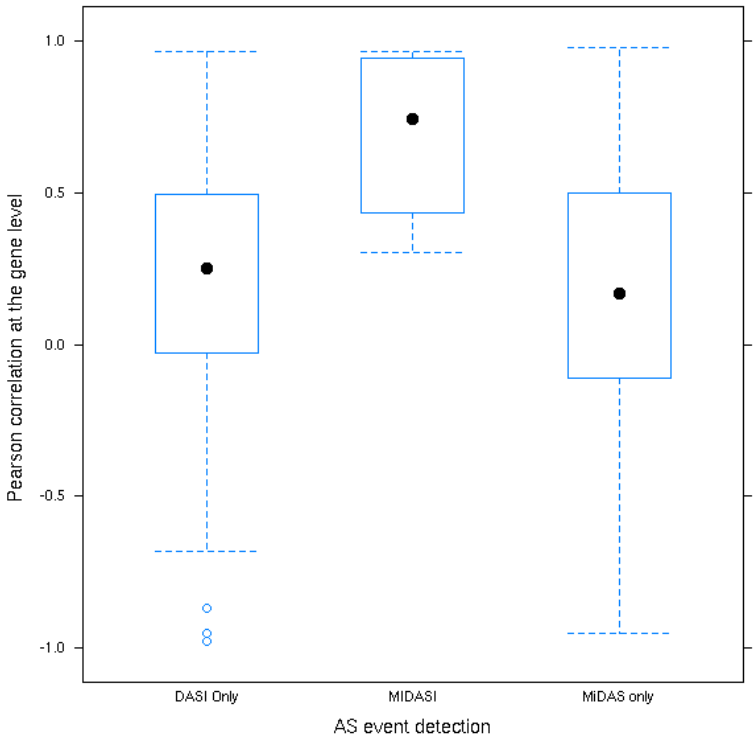


Supplemental Figure S 8: Comparison of RNA-Seq and Affymetrix exon arrays. The scatter plots shows the log ratio of the expression fold changes (HEK cells : B cells) from the affymetrix exon array (x-axis) versus the RNA-Seq platform (y-axis). Exons detected as present by both platforms are represented according to the average level of expression on the microarray: lower than 30 (green), between 30 and 80 (pink), or greater than 80 (orange). The overall correlation was good (PCC=0.73).

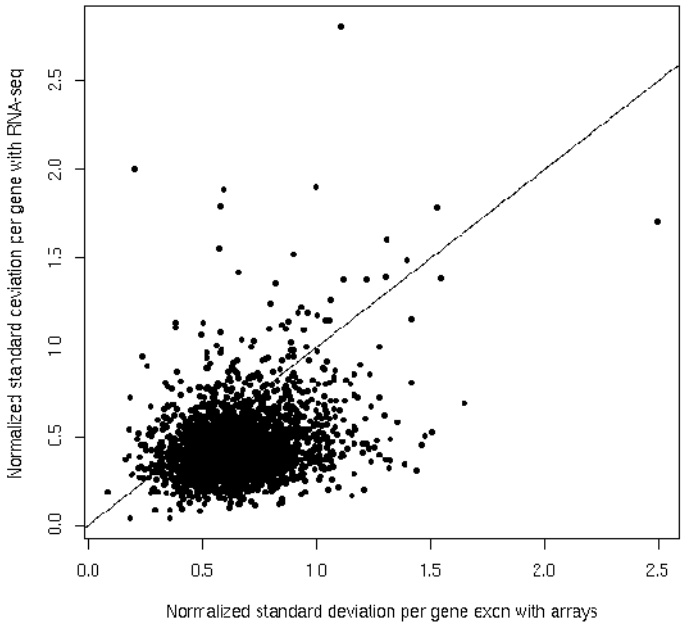


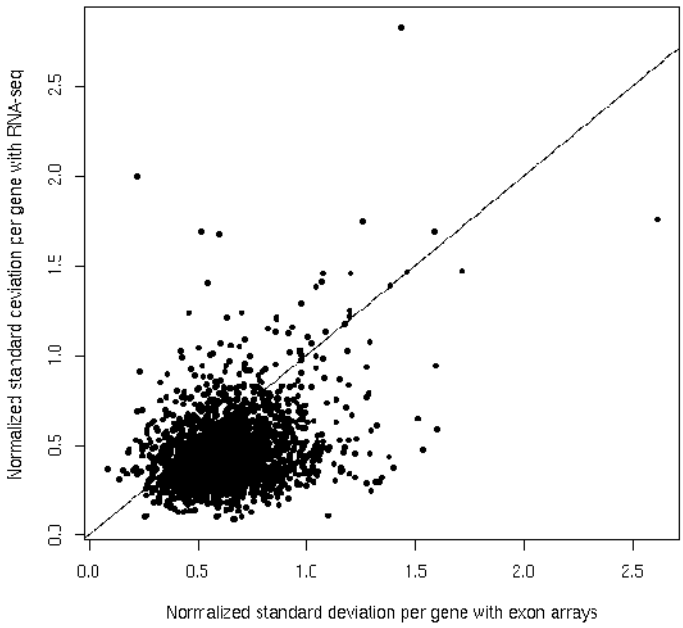
Supplemental Figure S 9: Comparison of RNA-Seq (DASI) and exon arrays (MIDAS) for differential splicing analysis between HEK and B cells. Both methods consist of 2 steps: 1) a corrected p-value identifies genes likely to be AS between HEK and B cell (FDR<=0.05 and FDR<=0.2 for DASI and MIDAS, respectively); 2) AS exons are scored according to a splicing index called DASI for RNA-Seq and SI for exon-arrays. The number of predicted AS genes that passes the threshold criteria at each steps are shown on the figure.



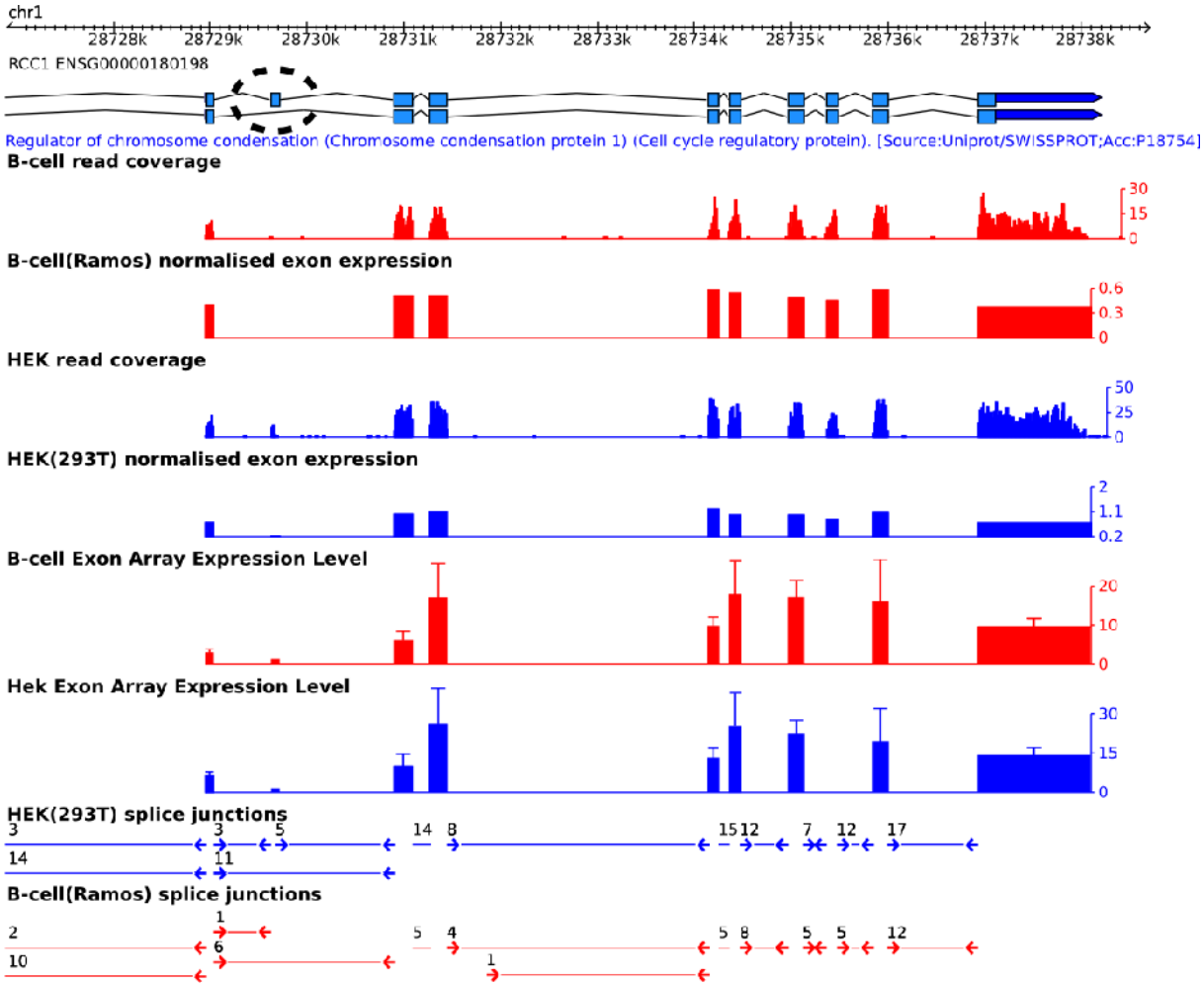


Supplemental Figure S 10: Top) Boxplot of the gene-wise mean squared error in exon-expression log-ratios measured with RNA-Seq and exon arrays. The error (y-axis) is shown for 1) genes with predicted AEE by DASI (DASI only), 2) genes with predicted exon with DASI and MIDAS (MIDASI), and 3) genes with predicted AS exon by MIDAS (MIDAS only). Bottom) Boxplot of gene-wise Pearson correlation of exon-expression log-ratios from RNA-Seq and exon arrays. The Pearson correlation is plotted according to the same three classes as above.





Supplemental Figure S 11: Scatterplots of the gene-wise normalized standard deviation (coefficient of variation) of exon expression values in B cells (top) and HEK cells (bottom) computed for exon arrays (x-axis) and RNA-Seq (y-axis). The coefficient of variation is consistently larger for exon array values in both cell lines (Wilcoxon, p-value<2.2 e-16).



Supplemental Figure S 12: Snapshot of the 3' end of RCC1 in Gbrowse. The circled exon in RCC1 is detected as alternatively spliced by the DASI method and verified by qPCR. However, it was not detected by the MIDAS method in exon arrays as the expression of this exon is below the background in both cell lines. The whiskers on the Exon Array Expression Level tracks denote the standard deviation as measured between the replicates.

Supplemental methods:

qPCR.The following experiment was performed to validate the use of the comparative Ct calculation method ($2^{-\Delta\Delta Ct}$). The amount of target, normalized to an endogenous reference and relative to a calibrator, is given by $2^{-\Delta\Delta Ct}$ ($\Delta\Delta$Ct is the difference in ΔCt for Ts65Dn and control sample and ΔCt is the difference in threshold cycles for target and reference).The $\Delta\Delta$Ct validation required approx. equal efficiencies of target and reference amplification. Therefore, standard curve assays were obtained for all target genes and references by measuring the transcripts levels obtained with specific primer sets on HEK and B cell cDNA sample diluted at two fold intervals. For each dilution, transcript levels were plotted against the Log value of the input cDNA concentration. qPCR efficiencies (*E*) were calculated from the slopes, where one cycle in the exponential phase yielded the efficiency: $E=10^{[-1/slope]}-1$. Transcripts analyzed here showed efficiency values close to 1 with a high linearity (Pearson correlation coefficient > 0.98).

EM update formulas. In our formalism, the hidden variables are the T_j. We derive an EM algorithm to estimate λ and q_j given $I_{i,j}$, y_i and p_i. Let's denote as $y_{i,j}$ the count of the form j in exon i, and m_j the total count of form j, the likelihood of the complete data is then written as:

$$\mathbf{P}(y_{1,1},\ldots,y_{n,k}) = \prod_{j=1}^{k}\left(\frac{\exp^{\lambda_j}\lambda_j^{m_j}}{m_j!}\cdot\binom{m_j}{y_{1,j}\ldots y_{n,j}}\cdot\prod_{i=1}^{n}\left(\frac{p_i}{\sum_l p_l I_{l,j}}\cdot I_{i,j}\right)^{y_{i,j}}\right.$$

$$\log\mathbf{P}(y_1,\ldots,y_n) = -\sum_{i-1}^{n}\mu_i+\sum_{i-1}^{n}y_i\log\mu_i-\sum_{i-1}^{n}\log y_i!$$

$$\text{with } \mu_i - \lambda\cdot p_i\sum_{j-1}^{k}q_j I_{i,j}$$

The E- and M-step of the algorithm are then computed as follow.

E-step

Assuming current parameters are known, the a posteriori count for form j at step v can be written as:

$$\hat{m}_j = \mathbb{E}_{q_j^{(v)}}(Y_j \mid c_1,\ldots,c_n) = \sum_{i=1}^{n}\frac{p_i q_j^{(v)} I_{i,j}}{\sum_{l-1}^{k} p_l q_l^{(v)} I_{i,l}}\cdot c_i$$

M-step

Maximising the likelihood conditionnally on $\hat{m}_j$, we get:

$$\hat{\lambda} = \sum_j \hat{m}_j = c$$
$$\hat{q}_j = \sum_i p_i I_{i,j} \cdot \frac{\hat{m}_j}{c}$$

Convergence is assumed when the relative increase of the log-likelihood is lower than a fixed value ε. For all experiments, we fixed ε to 10^{-6}. Also to avoid convergence to local maximums, we always initialize the algorithm with multiple random seed (in our case 10), selecting the solution with the highest likelihood.

Quality score. The POEM algorithm produces the maximum likelihood solution, even in case of wrong annotation. In order to assess the quality of a set of estimated proportions $\left(\hat{q}_{1,K}, \hat{q}_k\right)$ for transcripts 1...k, we tested whether the observed and expected counts were significantly different (according to the counts observed on the exons):

$$\chi^2_G = \sum_{e=1}^{n} \frac{\left(y_e - Y_e^{\exp}\right)^2}{Y_e^{\exp}},$$

where $Y_e^{\exp}$ is computed according to (1) for $T_j = \hat{q}_j$.

χ^2_G follows a chi-square with $(n-1)$ degrees of freedom. The quality score is computed for each gene as the log10 of the p-value**.**

Supplemental Table Legends

Supplemental Table S1A: This table provides all genes with their AEEs for the cell-internal analysis in B cells. It lists the p-value, CASI value of the exons and additional gene information. For each exon that is supported by an alternative splice junction in B cells the unique junction identifier for junctions (see Supplement Sultan et al. 2008) is associated to it.

Supplemental Table S1B: This table provides all genes with their AEEs for the cell-internal analysis in Hek cells. It lists the p-value, CASI value of the exons and additional gene information. For each exon that is supported by an alternative splice junction in Hek cells the unique junction identifier for junctions (see Supplement Sultan et al. 2008) is associated to it.

Supplemental Table S2: A listing of the primer sequences and results for the 61 CASI exons that were tested for exon skipping with RT-PCR. The table describes which events are validated using the primers S1-R1 and S2-R1 and which have been tested additionally and were validated with the S3-R1 primer pair. All exon skipping events are listed with their additional support from a) EST data, b) ENSEMBl v46 and c) SJ Support, where column "SJ count" denotes the number of reads that fall on the alternative splice junction that supports the exon skipping event. The column additional evidence shows alternative explanations for alternative splicing, which was annotated by manual inspection, for the cases that could not been amplified with RT-PCR.

Supplemental Table S3: This table contains a set of 73,948 reference alternative exons annotated to ENSEMBL version 46 genes. The set of exons have either evidence from the ENSEMBL database, as indicated by listing the corresponding ENSEMBL transcript ids. Or the exons are predicted to be alternatively spliced based on EST data, as indicated by their Unigene Cluster ID. The column "position Est difference" can be used to find the part of the alternative exon by querying the GeneNest browser with the Unigene Cluster ID (http://genenest.molgen.mpg.de/).

Supplemental Table S4: This table lists the estimated splice form proportions for 1,487 (640 genes) in B cells and 1,920 transcripts (830 genes) in HEK cells. It lists all transcripts for which proportion estimations were calculated by the means of either the ENSEMBL transcript ID or the merged transcript ID (with the prefix ENST-M) that was created after merging one or more ENSEMBL transcripts. For each transcript, we listed their level of expression (relative abundance, estimated splice form proportion) computed with POEM. From this probability and the absolute normalized expression, measured with RNA-Seq, we derived the normalized expression for each transcript.

Supplemental Table S5: A listing of the primer sequences and results for the 24 events that were tested for exon skipping with quantitative RT-PCR for POEM validation. All experiments were conducted using the S1-R1 and S2-R1 primer pairs. Each primer pair's mean normalized expression is depicted as well as the standard deviation derived from the three replicate experiments. All exon skipping events are listed with their additional support from a) EST data, b) ENSEMBl v46 and c) SJSupport.

Supplemental Table S6: This table provides all genes with their AEEs for the differential analysis between Hek and B cells. It lists the p-value, DASI value of the exons and additional gene information. For each exon that is supported by an alternative splice junction in either B or Hek cells the unique junction identifier for junctions (see Supplement Sultan et al. 2008) is associated to it.

Supplemental Table S7: A listing of the primer sequences and results for the 16 DASI exons that were tested for exon skipping with quantitative RT-PCR. All experiments were conducted using the S1-R1 and S2-R1 primer pairs. Each primer pair's mean normalized expression is depicted as well as the standard deviation derived from the three replicate experiments. The column "validated DASI" depicts the cases which lie inside the 1.5 ratio-difference interval. All exon skipping events are listed with their additional support from a) EST data, b) ENSEMBl v46 and c) SJ Support, where column SJcount denotes the number of reads in both cell lines that fall on the alternative splice junction that supports the exon skipping event. The DASI value, the MIDAS detection value, and the Splicing Index is given for every exon.