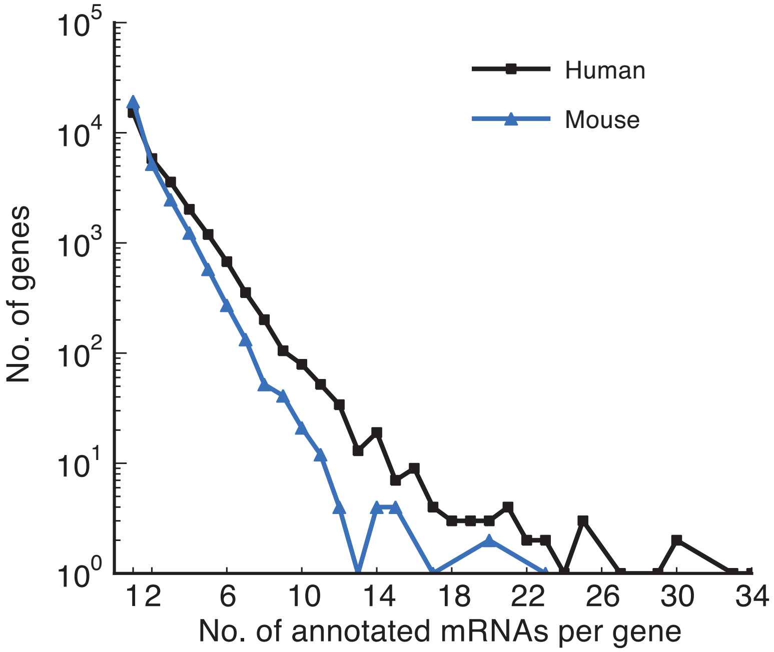
Nature Methods

Analysis and design of RNA sequencing experiments for identifying mRNA isoform regulation

Yarden Katz, Eric T Wang, Edoardo M Airoldi & Christopher B Burge

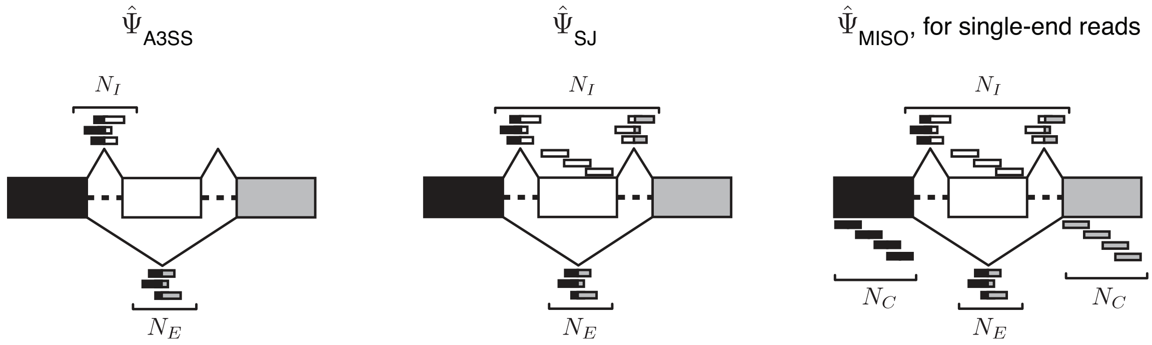
Supplementary Figures



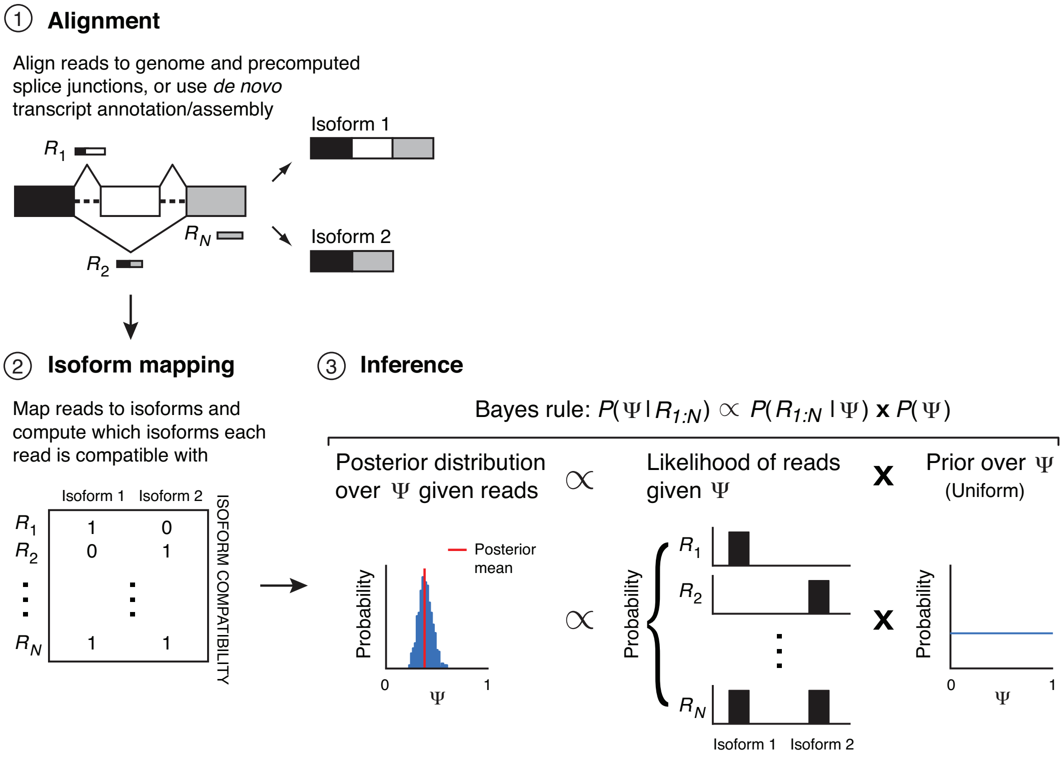
Supplementary Figure 1. Annotated isoforms in human and mouse genomes. Number of UCSC annotated mRNAs per gene, showing the large number of multi-isoform genes, even by conservative estimates that do not take into account RNA-Seq data.

Nature Methods: doi: 10.1038/nmeth.1528

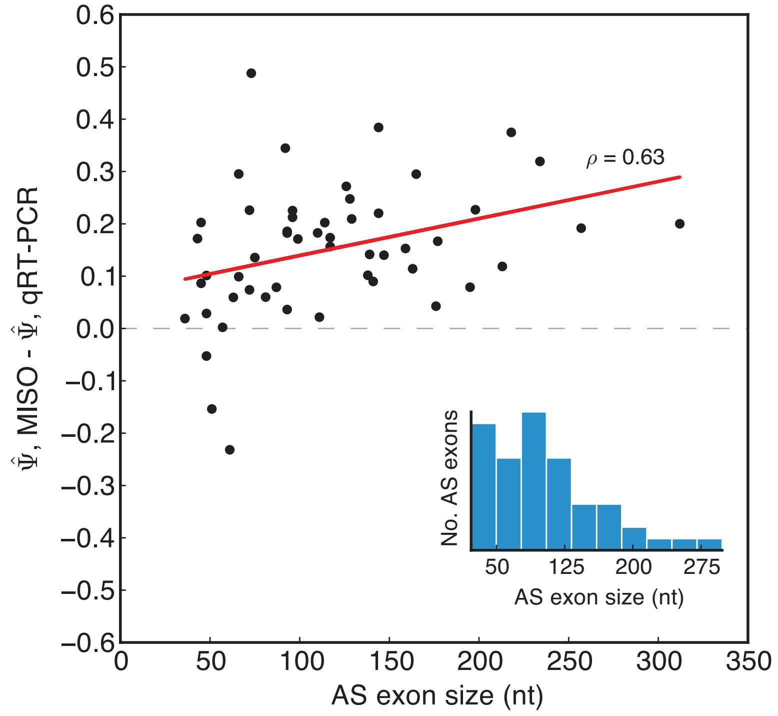
Read count	Supported isoform
N_I	Inclusive
N_E	Exclusive
N_C	Common


$\hat{\Psi}_{\text{A3SS}}$
$\hat{\Psi}_{\text{SJ}}$
$\hat{\Psi}_{\text{MISO}}$, for single-end reads
N_I
N_I
N_I
N_C
N_C
N_E
N_E
N_E

Supplementary Figure 2. Estimates of Ψ using single-end reads. The three Ψ estimates and the reads used in each estimate. N_I, N_E correspond to the number of reads supporting the inclusive isoform, respectively, while N_C corresponds to the number of reads supporting both isoforms (constitutive reads). The $\hat{\Psi}_{\text{SJ}}$ estimate shown corresponds to the estimate used in the majority of the analyses in[1], and $\hat{\Psi}_{\text{A3SS}}$ was also used in the same study for a subset of exons with an alternative splice site (see Supplementary Note for a proof of unbiasedness of these estimates.) The $\hat{\Psi}_{\text{MISO}}$ estimate shown corresponds to the analytic estimate from the MISO model (full derivation described in Supplementary Note), which is only obtained for single-end data. Estimates incorporate increasing amounts of information present in reads, with $\hat{\Psi}_{\text{A3SS}}$ using the least amount of information and $\hat{\Psi}_{\text{MISO}}$ using the most.

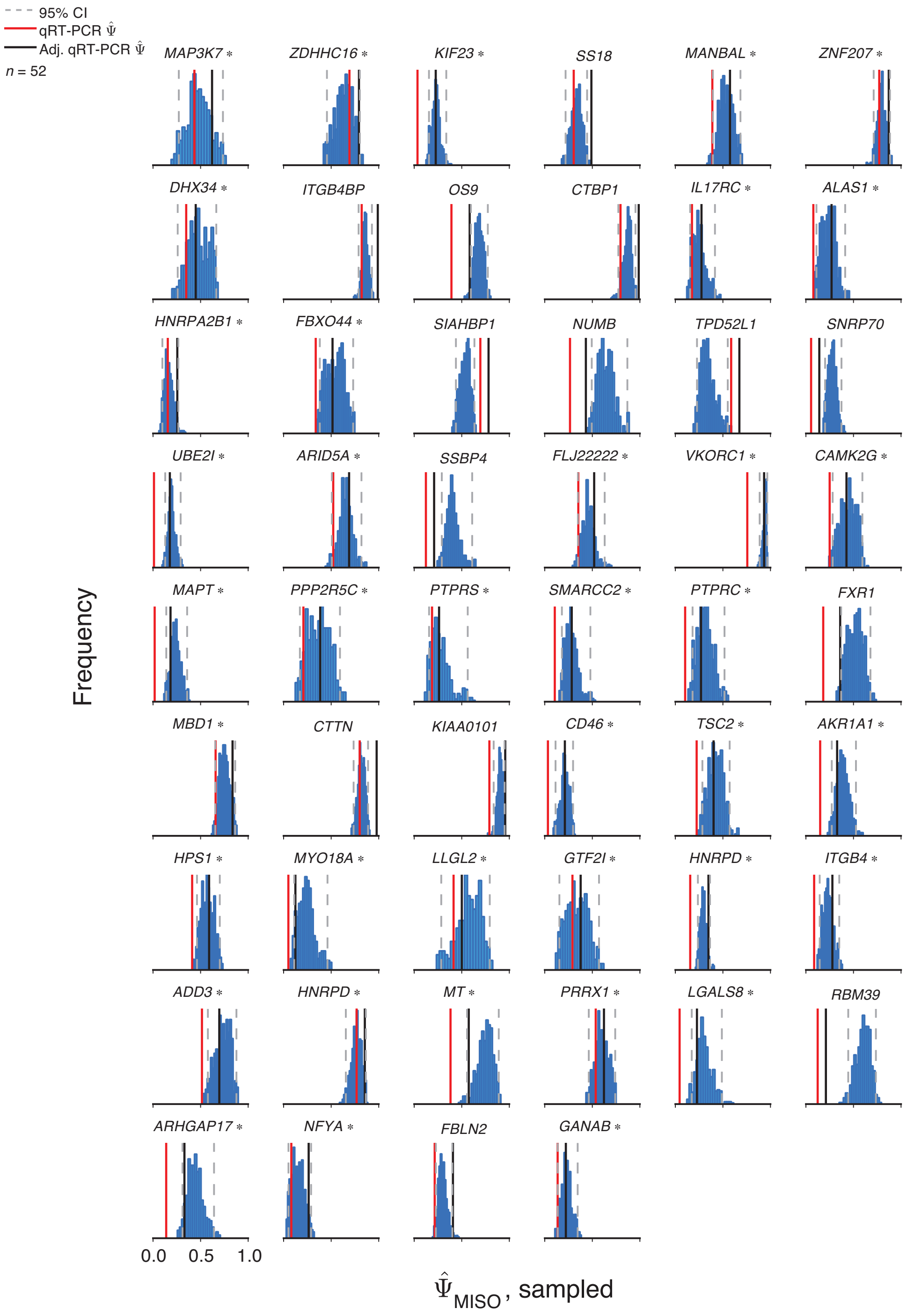

1 Alignment
Align reads to genome and precomputed splice junctions, or use *de novo* transcript annotation/assembly
R_1
R_2
R_N
Isoform 1
Isoform 2
2 Isoform mapping
Map reads to isoforms and compute which isoforms each read is compatible with
Isoform 1 Isoform 2
R_1 1 0
R_2 0 1
R_N 1 1
ISOFORM COMPATIBILITY
3 Inference
Bayes rule: $P(\Psi \mid R_{1:N}) \propto P(R_{1:N} \mid \Psi) \times P(\Psi)$
Posterior distribution over Ψ given reads
$\propto$
Likelihood of reads given Ψ
X
Prior over Ψ (Uniform)
Posterior mean
Probability
0
1
Ψ
R_1
R_2
R_N
Isoform 1
Isoform 2

Supplementary Figure 3. Steps of MISO statistical inference procedure from. Reads are aligned to the genome and a set of junctions that are either precomputed or discovered de novo using transcript annotation/discovery tools. Aligned reads are then mapped to isoforms, shown here for the case of a skipped exon, and represented as binary matrices that correspond to their compatibility with isoforms. Each each row i in the isoform compatibility matrix corresponds to a read, and each column j to an isoform, where the ijth entry is 1 if read i is consistent with isoform j and 0 otherwise. In this example, read R_1 is consistent only with the inclusive isoform (containing the white exon), R_2 consistent only with the exclusive isoform (excluding the white exon), while R_N consistent with both. Inference is performed by computing a probability distribution (the posterior) over Ψ given the reads. Bayes' rule states that this distribution is proportional to the product of our expectation about the value of Ψ (the prior, here taken to be uniformly distributed over $[0, 1]$) and the likelihood of observing the reads given Ψ (the likelihood). By summing over all possible assignments of reads to isoforms, weighting each assignment by its probability, the posterior distribution over Ψ is obtained. Inferences are then summarized by the mean of the posterior distribution, used as an estimate of Ψ, and confidence intervals that quantitate the confidence in the estimate (as described in Online Methods.)

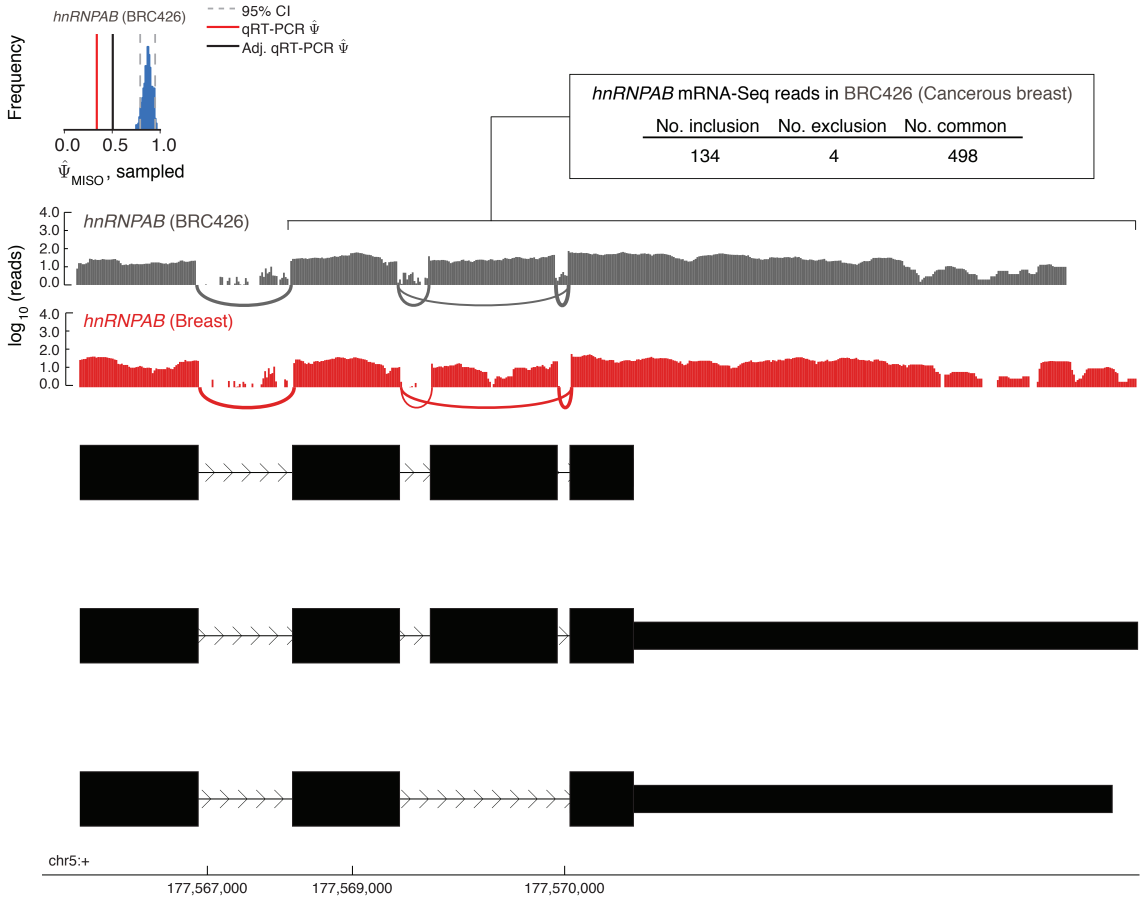


Supplementary Figure 4. Evidence for exon size-dependent qRT-PCR bias. Posterior marginals for each of the 52 alternative splicing events used in the breast cancer tissue sample[4].

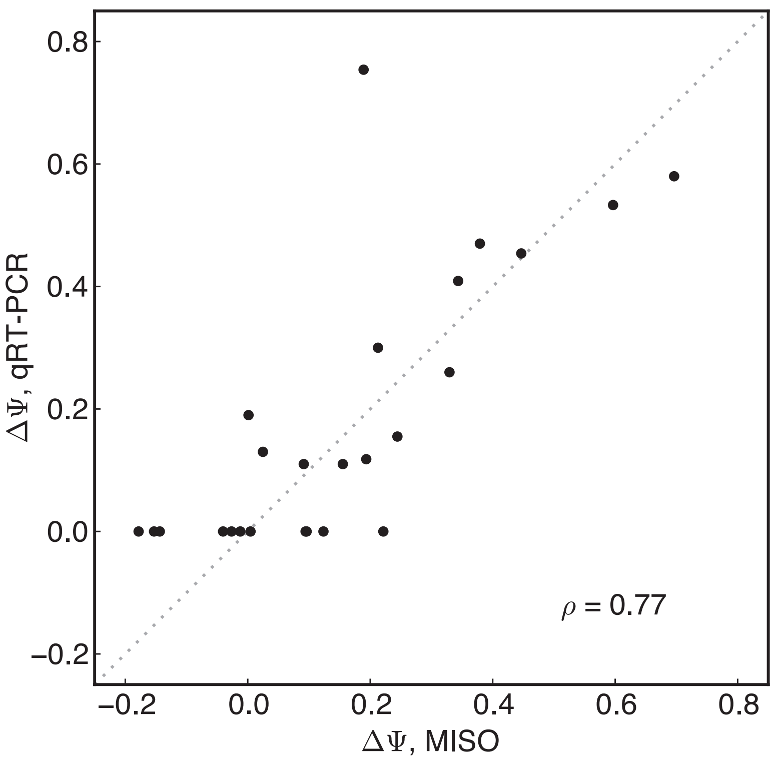
Nature Methods: doi: 10.1038/nmeth.1528



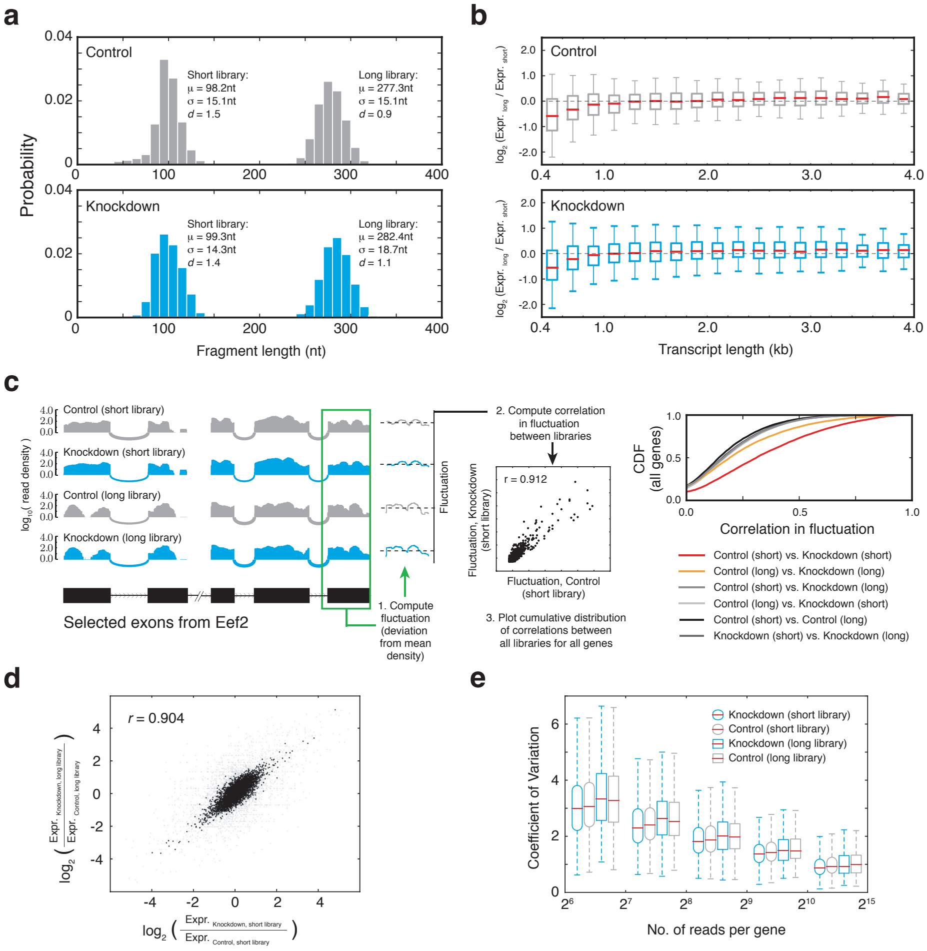
Supplementary Figure 5. MISO Ψ estimation for 52 alternative exons in breast cancer tissue. Posterior distributions for each of the 52 alternative splicing events used in the breast cancer tissue sample[4].



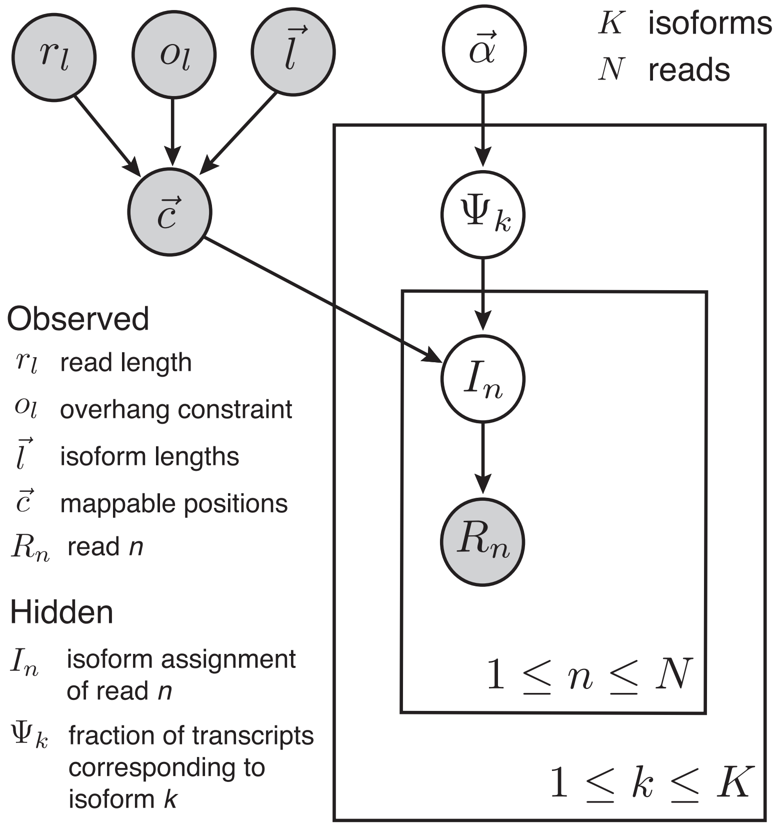
Supplementary Figure 6. mRNA-Seq data for hnRNPAB in breast cancer tissue. Data from breast cancer tissue (sample BRC426) from[4] and normal breast tissue (provided by Illumina, available on request).



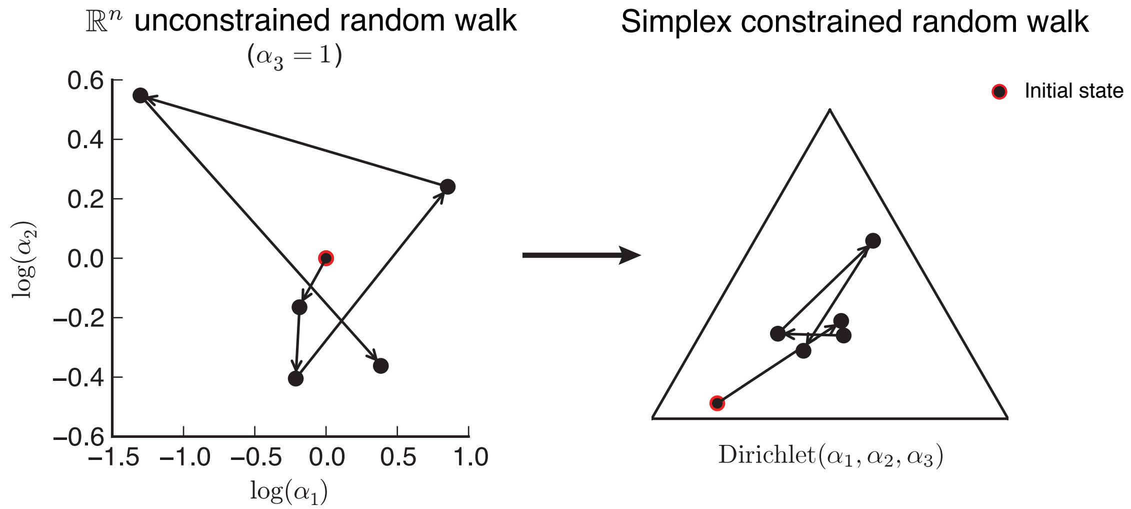
Supplementary Figure 7. Comparison of MISO $\Delta\Psi$ and qRT-PCR $\Delta\Psi$ values for hnRNP H dataset. Change in Ψ value based on MISO (x-axis) and qRT-PCR estimates (y-axis) for a set of 25 alternative exons in the hnRNP H control and knockdown data set are shown ($\rho = 0.77$).



Supplementary Figure 8. Comparison of read coverage fluctuations and gene expression for technical replicate libraries with short and long insert lengths. (**a**) The insert length distribution of pairs of libraries made from the same batch of RNA with mean fragment lengths of ∼98 nt and ∼277 nt. These are shown in both control (black) and CUGBP1 knockdown conditions. (**b**) Fold change between the short and long technical replicate libraries in control (top, black) and knockdown (bottom, blue) conditions. (**c**) Read coverage varies across exons, as illustrated on selected exons on Eef2. Deviation of sequence coverage from the mean can be computed for each gene. These values can be correlated between libraries, and the cumulative distribution function of correlation values can be plotted. (**d**) Changes in gene expression are preserved between libraries of differing insert lengths. (**e**) The coefficient of variation in read coverage across genes does not markedly differ between libraries of differing insert lengths.


r_l
o_l
$\vec{l}$
$\vec{\alpha}$
K isoforms
N reads
$\vec{c}$
Ψ_k
I_n
R_n
Observed
r_l read length
o_l overhang constraint
$\vec{l}$ isoform lengths
$\vec{c}$ mappable positions
R_n read *n*
Hidden
I_n isoform assignment of read *n*
Ψ_k fraction of transcripts corresponding to isoform *k*
$1 \leq n \leq N$
$1 \leq k \leq K$

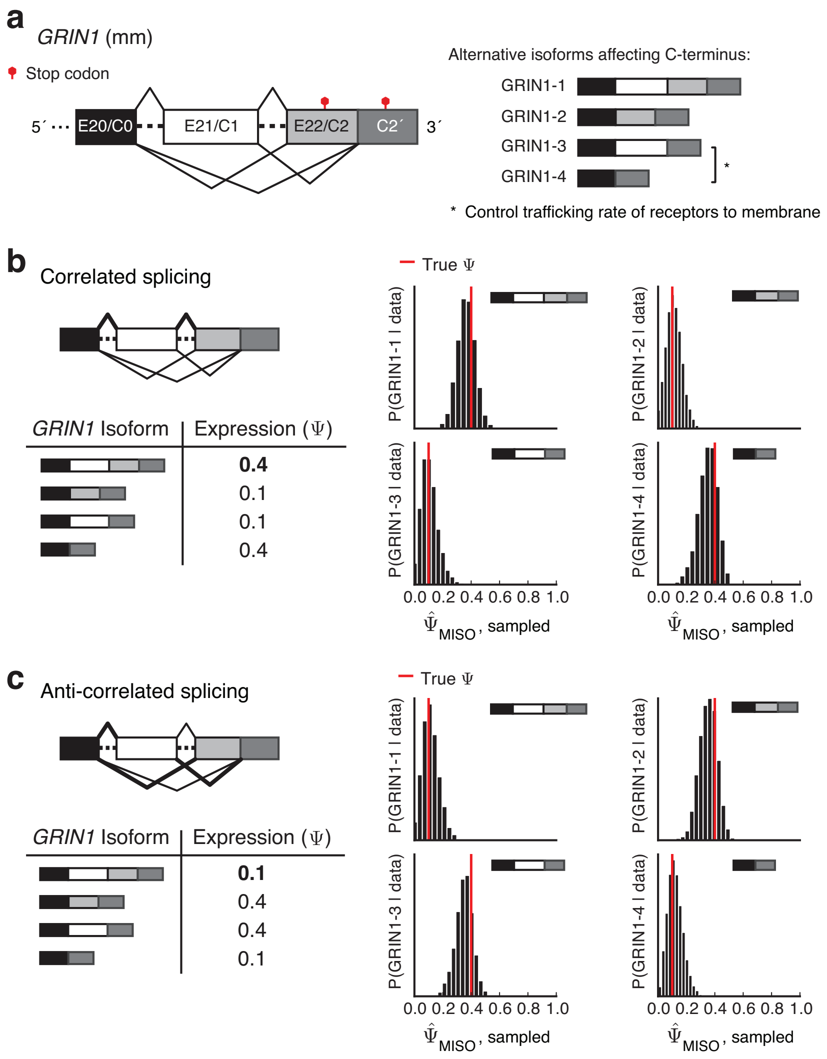
Supplementary Figure 9. Graphical model representation of MISO for single-end reads. A graphical representation of the probabilistic dependencies between variables in MISO, for a single gene with K isoforms. Shaded nodes represent observed variables, which include all the reads for the gene of interest, the parameters of the mRNA-Seq experiment and alignment procedure (the read length and the overhang length constraint) and features of the gene of interest (lengths of isoforms and the number of mappable positions in each isoform). The unshaded nodes represent random variables whose value are to be inferred from data, namely the Ψ value of each isoform k (Ψ_k) and the isoform from which each read was generated (I_n). The vector $\vec{\alpha}$ corresponds to the parameters of the Dirichlet prior distribution on isoform abundances, which is fixed to encode a uniform prior. MISO models the joint inference problem of finding the best set of Ψ values for the isoforms and the correct assignment of reads to the isoforms from which they were generated. For paired-end data, the probability of assigning a read to an isoform also depends on the parameters μ, σ of the insert length distribution, and are incorporated into the model as described in main text (for simplicity, these are not shown in the graphical model.)



Supplementary Figure 10. Random walk sampling scheme for inference in MISO. A five-step random walk sampled from a Logistic-Normal proposal distribution in log space of the parameters of a Dirichlet distribution (left). Each step in this random walk parameterizes a Dirichlet distribution, from which corresponding points on the simplex can be drawn (right). The use of the Logistic-Normal proposal distribution allows efficient exploration of the space of $\vec{\Psi}$ values.

```
Input: Set of reads $R$, set of isoforms $G$ of a gene, number of iterations
          to run $M$
Output: Set $S$ of sampled $\vec{\Psi}$ values

Initialize $\vec{\Psi}_t = \vec{\Psi}_0$ randomly

Initialize assignments of reads to isoforms consistently

Set $S = \{\}$

foreach Iteration $m = 1,...,M$ do
    Propose $\vec{\Psi}_{new}$ from a distribution centered around $\vec{\Psi}_t$
    Compute the probability $\alpha$ of accepting $\vec{\Psi}_{new}$ (using Metropolis-Hastings ratio)
    With probability $\alpha$, set $\vec{\Psi}_{t+1} = \vec{\Psi}_{new}$, otherwise set $\vec{\Psi}_{t+1} = \vec{\Psi}_t$

    foreach Read $r \in R$ do

        foreach Isoform $g \in G$ do
            Compute probability $p_{r,g}$ of reassigning read $r$ to isoform $g$
        end

        Sample reassignment of read $r$ to an isoform $g \in G$ based on
        computed probabilities
    end

    Set $S = S \cup \{\Psi_{t+1}\}$
end

return $S$
```

Supplementary Figure 11. Sampling-based MISO inference algorithm. Algorithm for estimating the posterior distribution over $\vec{\Psi}$ by Markov chain Monte Carlo sampling is shown. The algorithm begins with a random initialization of isoform distributions and assignments of reads to isoforms, and then repeatedly proposes new isoform distributions. These proposals are probabilistically accepted or rejected. If rejected, the previous isoform distribution is used in the next step. Each read is then probabilistically reassigned to one of the gene's isoforms, based on the new isoform distribution. As the algorithm converges, it is expected that an isoform distribution and associated assignment of reads to isoforms will be sampled in proportion to their probability under the model.



Supplementary Figure 12. Isoform abundance estimation for four isoforms of *GRIN1* gene. (**a**) A combination of alternative 3′ splice sites and exon skipping produces four isoforms in the region encoding the C-terminus of *GRIN1*[5,6]. (**b**) MISO Ψ estimates for reads simulated from an underlying isoform distribution where splicing of exons 21/C1 and 22/C2 is correlated, causing the two exons to be frequently included together (simulated isoform abundances shown at left). The posterior marginal distribution, estimated using the Monte Carlo algorithm described in Online Methods, is shown for each isoform, with the correct value shown as a vertical red line. (**c**) Estimation of isoform abundance for the case where reads were simulated from an *GRIN1* isoform distribution where splicing of exons 21/C1 and 22/C2 is anti-correlated, so the exons are rarely included together. The Ψ values for 21/C1 and 22/C2 are equal (0.5) in both conditions, but in the correlated condition the conditional probability of including 21/C2 given that 22/C2 is included is 0.8, while in the anti-correlated condition it is 0.2.

Supplementary Tables

Supplementary Table 1. Short read datasets used in study

Source	No. mapped reads	Run type
Human heart (Illumina, HiSeq 2000)	156 M	PE, 2x50nt
Human heart (Illumina, GA2)	30 M	PE, 2x54nt
Human testes (Illumina, GA2)	17 M	PE, 2x54nt
Human breast cancer tissue[13]	11 M	PE, 2x36nt
HEK 293T cells, control[14]	16 M	SE, 36nt
HEK 293T cells, hnRNP H knockdown[14]	21 M	SE, 36nt
HEK 293T cells, hnRNP H CLIP-Seq (this work)	4 M	SE, 36nt

Supplementary Table 2. Events used in qRT-PCR validation of MISO on breast cancer data set

Gene	Chromosome	Strand	ASE coords	ASE size	PCR Psi	Adj. PCR Psi	MISO Psi	PsiSJ
MAP3K7	6	-	91311072-91310992	81	0.43	0.62	0.49	0.43
ZDHHC16	10	+	99203546-99203593	48	0.69	0.79	0.64	0.62
KIF23	15	+	67520161-67520472	312	0.03	0.23	0.23	0.18
SS18	18	-	21869885-21869793	93	0.3	0.49	0.34	0.4
MANBAL	20	+	35360580-35360696	117	0.39	0.57	0.55	0.54
ZNF207	17	+	27712600-27712647	48	0.77	0.87	0.8	0.78
DHX34	19	+	52571970-52572044	75	0.35	0.45	0.48	0.47
ITGB4BP	20	-	33332046-33331871	176	0.82	0.99	0.86	0.91
OS9	12	+	56400149-56400313	165	0.39	0.58	0.69	0.73
CTBP1	4	-	1225307-1225113	195	0.8	0.99	0.88	0.87
IL17RC	3	+	9937609-9937653	45	0.17	0.27	0.26	0.22
ALAS1	3	+	52207728-52207904	177	0.08	0.27	0.24	0.18
HNRPA2B1	7	-	26204011-26203976	36	0.16	0.26	0.17	0.48
FBXO44	1	+	11641177-11641272	96	0.34	0.51	0.55	0.49
SIAHBP1	8	-	144974874-144974824	51	0.7	0.78	0.54	0.58
NUMB	14	-	72815885-72815742	144	0.27	0.43	0.65	0.51

TPD52L1	6	+	125619943-125620003	61	0.59	0.67	0.35	0.33
SNRP70	19	+	54297183-54297254	72	0.05	0.14	0.28	0.46
UBE2I	16	+	1302349-1302605	257	0.01	0.18	0.2	0.31
ARID5A	2	+	96578785-96578923	139	0.52	0.69	0.66	0.6
SSBP4	19	+	18403163-18403228	66	0.12	0.21	0.42	0.44
FLJ22222	17	-	77945708-77945496	213	0.35	0.52	0.47	0.38
VKORC1	16	-	31012243-31012134	110	0.76	0.93	0.94	0.88
CAMK2G	10	-	75249409-75249296	114	0.25	0.43	0.45	0.47
MAPT	17	+	41423081-41423278	198	0.02	0.18	0.24	0.23
PPP2R5C	14	+	101453921-101454037	117	0.21	0.38	0.38	0.32
PTPRS	19	-	5167778-5167731	48	0.19	0.26	0.29	0.26
SMARCC2	12	-	54853080-54852988	93	0.11	0.28	0.29	0.32
PTPRC	1	+	196938139-196938282	144	0.1	0.27	0.32	0.28
FXR1	3	+	182175795-182175886	92	0.18	0.36	0.53	0.35
MBD1	18	-	46053839-46053702	138	0.66	0.84	0.76	0.81
HNRPAB	5	+	177569739-177569879	141	0.34	0.51	0.88	0.86
CTTN	11	+	69945224-69945334	111	0.8	0.98	0.82	0.87
KIAA0101	15	-	62456157-62455995	163	0.79	0.96	0.9	0.73
CD46	1	+	206030221-206030313	93	0.03	0.21	0.22	0.12
TSC2	16	+	2067600-2067728	129	0.22	0.4	0.43	0.57
AKR1A1	1	+	45790695-45790822	128	0.15	0.33	0.4	0.26

HPS1	10	-	100179636-100179538	99	0.41	0.59	0.58	0.62
MYO18A	17	-	24436792-24436748	45	0.05	0.12	0.25	0.15
LLGL2	17	+	71082129-71082171	43	0.41	0.5	0.59	0.55
GTF2I	7	+	73771134-73771196	63	0.29	0.38	0.35	0.22
HNRPD	4	-	83496860-83496714	147	0.16	0.35	0.3	0.24
ITGB4	17	+	71262731-71262889	159	0.09	0.28	0.24	0.28
ADD3	10	+	111882053-111882148	96	0.51	0.7	0.74	0.59
HNRPD	4	-	83511761-83511705	57	0.77	0.85	0.77	0.83
MT	22	-	41863248-41863031	218	0.38	0.57	0.76	0.53
PRRX1	1	+	168966042-168966113	72	0.54	0.62	0.61	0.51
LGALS8	1	+	234772838-234772963	126	0.04	0.23	0.32	0.22
RBM39	20	-	33791933-33791861	73	0.12	0.21	0.61	0.71
ARHGAP17	16	-	24858419-24858186	234	0.14	0.33	0.46	0.41
NFYA	6	+	41156528-41156614	87	0.08	0.26	0.16	0.09
FBLN2	3	+	13638276-13638416	141	0.22	0.41	0.31	0.4
GANAB	11	-	62158423-62158358	66	0.14	0.22	0.23	0.43

Supplementary Note: Details of statistical estimators and MISO inference algorithm

Estimates of Ψ. Multiple approaches for estimating Ψ values have been proposed. One simple estimate, $\hat{\Psi}_{\mathsf{A3SS}}$, considers the splicing event as a choice between the competing $3'$ splice sites ($3'$ss) of the alternative exon and the downstream constitutive exon, estimating the inclusion of the exon by the relative numbers of reads that join these $3'$ss to the upstream constitutive exon (Supplementary Fig. 2). This estimate was previously used for a subset of alternative splicing events.[1]

A more comprehensive estimate is $\hat{\Psi}_{\mathsf{SJ}}$, which estimates exon inclusion based on the combined read density in the body of the alternative exon and in the two junctions that involve the alternative exon, relative to the density of junction reads that join the upstream and downstream constitutive exons[1] (Figure 1c). The remaining reads that align to the bodies of the flanking constitutive exons could have derived from either isoform and are not used in $\hat{\Psi}_{\mathsf{SJ}}$.

More formally, $\hat{\Psi}_{\mathsf{SJ}} = \frac{D_I}{D_I + D_E}$, where D_I is the density of inclusion reads and D_E the density of exclusion reads. Let e_l be the length of the alternatively spliced exon, r_l be the length of mRNA-seq reads, and o_l the overhang constraint placed on splice junctions. Assuming all positions in the gene of interest are uniquely mappable, D_I and D_E are computed as follows:

$$D_I = \frac{N_I}{e_l - r_l + 1 + 2(r_l + 1 - 2o_l)}, D_E = \frac{N_E}{r_l + 1 - 2o_l}$$

where N_I and N_E are the number of reads supporting inclusion and exclusion reads, respectively. If non-uniquely mappable read starting positions exist, these are simply subtracted from the denominators of D_I and D_E.

Computing the maximum a posteriori estimate $\hat{\Psi}_{\mathsf{MISO}}$ for single-end reads. As explained in the main text, constitutive reads contain latent information about Ψ, and can be used to improve and stabilize Ψ estimates (Figure 1d). For exon-centric analyses, an analytic estimate can be computed, if only single-end reads are used and certain assumptions are made about the prior distribution $P(\Psi)$. This estimate is denoted $\hat{\Psi}_{\mathsf{MISO}}$ and can be derived by computing the *maximum a posteriori* (MAP) estimate of Ψ under the MISO model. (Note that in a subset of figures in the main text, $\hat{\Psi}_{\mathsf{MISO}}$ is used alternatively to denote the mean of the posterior distribution over Ψ obtained by MCMC-based inference, as indicated by the figure legends.)

Since the prior $P(\Psi)$ is 1 when the hyperparameters $\alpha = \beta = 1$, the MAP estimate and the MLE estimates are equal, and so we proceed by finding the MLE. Given $R_{1:N}$ reads and their isoform assignments $I_{1:N}$, the likelihood function $P(R_{1:N} \mid \Psi)$ is:

$$\begin{aligned} P(R_{1:N} \mid \Psi) &= \prod_{n=1}^{N} \sum_{I_n=1}^{2} P(R_n \mid I_n) P(I_n \mid \Psi) \\ &= \prod_{n=1}^{N} \left[\, P(R_n \mid I_n = 1)\Psi_f + P(R_n \mid I_n = 2)(1 - \Psi_f) \,\right] \end{aligned}$$

where $1 \leq n \leq N$. We'd like to find a value of $\hat{\Psi}$ that maximizes this likelihood, which is written as a function of Ψ_f. By the equivariance property of maximum likelihood, we can simply find the MLE $\hat{\Psi}_f$ of Ψ_f and transform it into $\hat{\Psi}$, since the two are one-to-one, using:

$$\hat{\Psi} = \frac{c_1 \hat{\Psi}_f}{c_1 - c_1 \hat{\Psi}_f + c_2 \hat{\Psi}_f} \tag{1}$$

To simplify the notation, let p_1 and p_2 stand for the probabilities of a read being generated from the first and second isoforms, respectively, assuming read lengths r_l:

$$p_1 = \frac{1}{m(rl, I_1)}, p_2 = \frac{1}{m(rl, I_2)}$$

Substituting our observation model into the likelihood gives:

$$\begin{aligned} P(R_{1:N} \mid \Psi_f) &= \prod_{n=1}^{N} (P(R_n \mid 1, \Theta)\Psi_f + P(R_n \mid 2, \Theta)(1 - \Psi_f)) \\ &= \prod_{n=1}^{N} \left[\, p_1 R_n^1 \Psi_f + p_2 R_n^2 (1 - \Psi_f) \,\right] \end{aligned}$$

Taking the log, we have:

$$\hat{\Psi}_f = \arg\max_{\Psi_f} \sum_{n=1}^{N} \log\left(p_1 R_n^1 \Psi_f + p_2 R_n^2 (1 - \Psi_f)\right)$$

Differentiating and setting the derivative to zero yields:

$$\begin{aligned} \frac{d}{d\Psi_f} \sum_{n=1}^{N} \log\left(p_1 R_n^1 \Psi_f + p_2 R_n^2 (1 - \Psi_f)\right) &= 0 \\ \sum_{n=1}^{N} \frac{p_1 R_n^1 - p_2 R_n^2}{p_1 R_n^1 \Psi_f + p_2 R_n^2 (1 - \Psi_f)} &= 0 \end{aligned}$$

This sum can be rewritten in terms of three sufficient statistics: the number of reads supporting only isoform 1 (N_I), the number of reads supporting only isoform 2 (N_E), and the number of reads supporting both isoforms (N_C), to get:

$$\frac{N_I}{\Psi_f} - \frac{N_E}{1-\Psi_f} + \frac{N_C(p_1-p_2)}{p_1\Psi_f + p_2(1-\Psi_f)} = 0$$

This equation reduces to solving a quadratic equation, whose relevant solution is:

$$\hat{\Psi}_f = \frac{A - \sqrt{B+C}}{D}, \text{ where:}$$

$$\begin{aligned}
A &= N_I p_1 + N_C p_1 - 2N_I p_2 - N_E p_2 - N_C p_2 \\
B &= 4N_I p_2(N_I p_1 + N_E p_1 + N_C p_1 - N_I p_2 - N_E p_2 - N_C p_2) \\
C &= (-N_I p_1 - N_C p_1 + 2N_I p2 + N_E p_2 + N_C p_2)^2 \\
D &= 2(N_I p_1 + N_E p_1 + N_C p_1 - N_I p_2 - N_E p_2 - N_C p_2)
\end{aligned}$$

Now, $\hat{\Psi}_f$ can be plugged in to Equation 1 to obtain $\hat{\Psi}$, which is our MAP/MLE estimate. This resulting estimate is simply a function of the read counts N_I, N_E, N_C and the probabilities p_1 and p_2.

Proof that $\hat{\Psi}_{\mathsf{A3SS}}$ is unbiased. As an example of analytic estimates of Ψ, we show that the simplest estimate, $\hat{\Psi}_{\mathsf{A3SS}}$, is unbiased. Recall that $\hat{\Psi}_{\mathsf{A3SS}}$ uses only the reads from one inclusion junction and from the exclusion junction (Supplementary Fig. 2). Given a read length r_l and an overhang constraint of o_l, let J be the number of possible read starting positions in a junction:

$$J = r_l + 1 - 2o_l$$

Then $\hat{\Psi}_{\mathsf{A3SS}}$ can be defined as follows:

$$\begin{aligned}
\hat{\Psi}_{\mathsf{A3SS}} &= \frac{\frac{N_I}{J}}{\frac{N_I}{J} + \frac{N_E}{J}} \\
&= \frac{N_I}{N_I + N_E}
\end{aligned}$$

Proposition 1 (Unbiasedness of $\hat{\Psi}_{\mathsf{A3SS}}$) *The symmetric splice junction estimator $\hat{\Psi}_{\mathsf{A3SS}}$ is unbiased, i.e. $E(\hat{\Psi}_{\mathsf{A3SS}}) = \Psi$ for all Ψ.*

Proof Fix Ψ. Let l_1 be the length of the inclusive isoform, and l_2 be the length of the exclusive isoform. Then the number of reads possible from the two isoforms are c_1, c_2, respectively:

$$\begin{aligned} c_1 &= l_1 - r_l + 1 \\ c_2 &= l_2 - r_l + 1 \end{aligned}$$

These constants are used to compute Ψ_f, the probability of sequencing a read from the inclusive isoform, which is defined as follows:

$$\Psi_f = \frac{c_1 \Psi}{c_1 \Psi + c_2(1-\Psi)}$$

Recall that $J = r_l + 1 - 2o_l$. In general, a read generated in our model falls into one of four mutually categories. It could support: (1) the inclusive isoform, (2) the exclusive isoform, (3) both isoforms, or (4) be thrown out due to an overhang violation. Relative to this space of outcomes, the expected probabilities of inclusion and exclusion reads are as follows:

$$\begin{aligned} P(N_I) &= P(\text{inclusive isoform})P(\text{inclusion junction read} \mid \text{inclusive isoform}) \\ &= \Psi_f \frac{J}{c_1} \\ P(N_E) &= P(\text{exclusive isoform})P(\text{exclusion junction read} \mid \text{exclusive isoform}) \\ &= (1-\Psi_f)\frac{J}{c_2} \end{aligned}$$

To show unbiasedness, it suffices to show that $E(\frac{N_I}{N_I+N_E}) = \Psi$. Since $\hat{\Psi}_{\mathsf{A3SS}}$ uses only the N_I and N_E reads, we know that $N_I + N_E = n$, where n is the total number of reads used in the estimate. Therefore,

$$E\left(\frac{N_I}{N_I+N_E}\right) = E\left(\frac{N_I}{n}\right) = \frac{1}{n}E(N_I)$$

The expected number of inclusion reads in a sample of n reads, $E(N_I)$, is simply $n \times P(N_I)$. Since reads other than inclusion or exclusion reads are discarded in the $\hat{\Psi}_{\mathsf{A3SS}}$ estimate, the probability of an inclusion read must be normalized to account for the fact that these are the only two outcomes:

$$\begin{aligned} E(N_I) &= n \times \frac{P(N_I)}{P(N_I)+P(N_E)} \\ &= n \times \frac{\Psi_f \frac{J}{c_1}}{\Psi_f \frac{J}{c_1} + (1-\Psi_f)\frac{J}{c_2}} \end{aligned}$$

Substituting Ψ_f with its definition results in:

$$\begin{aligned}
E(N_I) &= n \times \frac{\dfrac{c_1\Psi}{c_1\Psi + c_2(1-\Psi)}\dfrac{J}{c_1}}{\dfrac{c_1\Psi}{c_1\Psi + c_2(1-\Psi)}\dfrac{J}{c_1} + \left(1 - \dfrac{c_1\Psi}{c_1\Psi + c_2(1-\Psi)}\right)\dfrac{J}{c_2}} \times \frac{c_1\Psi + c_2(1-\Psi)}{c_1\Psi + c_2(1-\Psi)} \\
&= n \times \frac{\Psi J}{\Psi J + (c_1\Psi + c_2(1-\Psi) - c_1\Psi)\frac{J}{c_2}} \\
&= n \times \frac{\Psi J}{\Psi J + (1-\Psi)J} \\
&= n \times \frac{\Psi}{\Psi + (1-\Psi)} \\
&= n \times \Psi
\end{aligned}$$

Thus, $\frac{1}{n}E(N_I) = \Psi$, which demonstrates that $\hat{\Psi}_{\mathsf{A3SS}}$ is unbiased. A similar argument holds for the $\hat{\Psi}_{\mathsf{SJ}}$ estimate used in[1].

Efficient estimation of isoform distributions for genes with many isoforms. We devised a Markov chain Monte Carlo (MCMC) inference scheme based on a novel proposal distribution. Considering the length information and length correction in our problem leads to violations of the mathematically convenient conjugacy properties of traditional Dirichlet-Multinomial mixture models. For this reason, the use of a standard Gibbs sampler is not possible. Instead, we use a hybrid MCMC sampler that combines the Metropolis-Hastings (MH) algorithm with a Gibbs sampler[2]. In MH, a *proposal distribution* Q is used to estimate the target distribution $P(\vec{x})$, where P can be evaluated up to proportionality on any set of states but cannot be easily sampled from. Transitions to different states of P are repeatedly proposed from Q, and these are stochastically accepted or rejected according to the *MH ratio*, α:

$$\alpha = \min\left(\frac{P(\vec{x}_{t+1})Q(\vec{x}_t;\vec{x}_{t+1})}{P(\vec{x}_t)Q(\vec{x}_{t+1};\vec{x}_t)}, 1\right) \qquad \text{(MH ratio)}$$

where α is the probability of transitioning to the proposed state $\vec{x}_{t+1}$ from the current state $\vec{x}_t$. The better the proposal distribution Q is at proposing probable values under P, the faster the sampling algorithm will converge to the correct distribution.

In our case, the target distribution is the posterior distribution on $\vec{\Psi}$ given a set of reads, $P(\vec{\Psi} \mid R_{1:N})$. In general, we expect any set of reads from a gene with many isoform to be well-explained by only a small set of closely related isoform distributions. In other words, we expect the model's probability mass to be peaked on a small set of $\vec{\Psi}$ values that explain the data, with little probability mass on other $\vec{\Psi}$ that encode a very different set of isoform abundances.

In light of this unimodal probability landscape, a proposal distribution that uniformly proposes random isoform distributions is unlikely to find values that fit the data. A standard strategy for solving problems of this form using sampling is to use a proposal distribution that "drifts"—or forms a *random walk*—over the sampled variable's state space. In a random walk proposal, the proposed value is typically the previously sampled value plus some noise (e.g. the previous proposal, corrupted by normally distributed noise.)

A challenge in defining a random walk proposal in our case is that isoform distributions are constrained to sum to one—i.e., they must be probability distributions. Therefore, a random walk where a proposal is drawn from a normal distribution centered on the previously sampled isoform distribution will not work. To overcome this, we formulated a random walk using the *Logistic-Normal distribution*[3], a distribution on the simplex that generalizes the more commonly used Dirichlet distribution. With the Logistic-Normal it is possible to to formalize the idea that the newly proposed isoform distribution is drawn from a distribution whose mean is the previously sampled distribution, meaning that only small changes to the current isoform distribution are proposed, while still respecting the constraint that proposed values must sum to one. Intuitively, this allows the algorithm to 'hone in' on the region of highly probable isoform distributions for a given data set, and move around in that space, without spending too much time sampling lower probability regions.

The random walk is defined over the parameters of the distribution from which $\vec{\Psi}$ is drawn, in log space, allowing the sampled values to range unconstrained over the space of real numbers. Each draw of a set of parameters then parameterizes a Dirichlet distribution from which an isoform distribution is drawn. Supplementary Figure 10 shows proposals drawn according to this process, illustrating how a five-step unconstrained random walk on the parameters of the distribution in log space induces a random walk in the constrained space of the 2D simplex, where each point represents a probability vector. Our algorithm exploits the fact that a random walk over the parameters of a distribution—which can be conveniently 'drifted over' unconstrained—can be used to define a random walk over the values drawn from this distribution, which in this case are constrained to lie within the simplex.

Our sampling algorithm, shown in Supplementary Figure 11, proceeds by repeatedly proposing new values for $\vec{\Psi}$, which are stochastically accepted in proportion to their probability under the model. First, a new isoform distribution is proposed, which is probabilistically accepted or rejected based on the MH ratio, as described above. For each proposed isoform distribution, the algorithm then probabilistically reassigns each read to a new isoform, which completes one iteration of the

Nature Methods: doi: 10.1038/nmeth.1528

algorithm. As the number of iterations increases, the algorithm is guaranteed to eventually sample isoform distributions and assignments of reads to isoforms in proportion to their posterior probability in our model.

A distribution that can capture the desired random walk in simplex space is the *Logistic-Normal* distribution, parameterized by a mean $\vec{\mu}$ and covariance matrix Σ, and denoted $L_k(\vec{\mu}, \Sigma)$ for the k-dimensional case[3]. A k-dimensional vector $\vec{\theta}$ can be sampled from $L_{k-1}(\vec{\mu}, \Sigma)$ by first sampling a vector v from a multivariate normal distribution $\text{Normal}(\vec{\mu}, \Sigma)$ and then taking its inverse logistic transform, logit^{-1}:

$$\vec{\theta} = \text{logit}^{-1}(v) = \frac{e^v}{1 + \sum_{k=1}^{K} e^{v_k}}$$

The vector v can be obtained back via the logit transform $v = \log(\theta/\theta_{k+1})$, where $\theta_{k+1} = 1 - \sum_{k=1}^{K} \theta_k$. The probability density for $\vec{\theta}$ with parameters $\vec{\mu}, \Sigma$ is:

$$|2\pi\Sigma|^{-\frac{1}{2}} \left(\prod_{j=1}^{k+1} \theta_j \right)^{-1} \exp\left[-\frac{1}{2} \{\log(\theta/\theta_{k+1}) - \vec{\mu}\}^{\mathrm{T}} \Sigma^{-1} \{\log(\theta/\theta_{k+1})\} \right] \qquad (2)$$

Given a fixed covariance matrix Σ for the proposal distribution, the sampling scheme is as follows:

1. Initialize $\vec{\mu}_t, \vec{\Psi}_t = \vec{\mu}_0, \vec{\Psi}_0$, and initialize I to a consistent assignment
2. For $m = 1, \ldots, M$ iterations,
   (a) Propose $\vec{\mu}_{t+1}, \vec{\Psi}_{t+1}$:

   $$\begin{aligned} \vec{\mu}_{t+1} &\sim \text{Normal}(\vec{\mu}_t, \Sigma) \\ \vec{\Psi}_{t+1} &= \text{logit}^{-1}(\vec{\mu}_{t+1}) \end{aligned}$$

   (b) Let $\vec{\mu}_t, \vec{\Psi}_t = \vec{\mu}_{t+1}, \vec{\Psi}_{t+1}$ with probability α (otherwise, keep values from step t):

   $$\alpha = \min\left(\frac{P(\vec{\Psi}_{t+1}, I_{1:N}, R_{1:N}) Q(\vec{\Psi}_t; \vec{\Psi}_{t+1})}{P(\vec{\Psi}_t, I_{1:N}, R_{1:N}) Q(\vec{\Psi}_{t+1}; \vec{\Psi}_t)}, 1 \right)$$

   Here, Q is a Logistic-Normal (following the probability density given in Equation 2) with mean equal to the previous time step's $\vec{\Psi}$, excluding its last element:

   $$Q(\vec{\Psi}_{t+1}; \vec{\Psi}_t) \sim L_{k-1}(\log([\Psi_t^1, \ldots, \Psi_t^{k-1}]), \Sigma)$$

   And similarly for $Q(\vec{\Psi}_t; \vec{\Psi}_{t+1})$. The joint distributions in the MH ratio factor into a product of conditionals, as explained in Online Methods.

(c) Gibbs step: for $j = 1, \ldots, N$ reads,

i. Compute the probability $a_{j,k}$ of reassigning read R_j to the kth isoform, for every isoform $1 \leq k \leq K$:

$$a_{j,k} = P(I_j = k \mid R_{1:N}, I_{1:N} \backslash \{I_j\}, \vec{\Psi}_{t+1})$$

ii. Sample reassignment of $I_j \sim \text{Multinomial}(1, [a_{j,1} \ \cdots \ a_{j,K}])$

Note that Step (c) is the usual Gibbs sampling step for reassigning data points to components in mixture models.

Cross-validation adjustment of qRT-PCR values. Given the apparent length bias in the qRT-PCR estimates of Ψ in Figure 2, we computed an adjusted set of qRT-PCR Ψ values using cross-validation, in order to estimate the overlap between these values and the Bayesian confidence intervals of $\hat{\Psi}_{\mathsf{MISO}}$. Exons were binned by length ($n = 3$ bins) and the events used for validation were split in four sets. An adjusted qRT-PCR estimate of Ψ was then computed for each set by adding the average $\Delta\Psi$ of MISO and qRT-PCR estimates in the remaining three sets, bounding the resulting value in $[0, 1]$. The adjusted and raw Ψ estimates for MISO and qRT-PCR, along with the Bayesian confidence intervals, are shown in Supplementary Figure 5.

References

1. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456** (2008).

2. Liu, J. S. *Monte Carlo Strategies in Scientific Computing (Springer Series in Statistics)* (Springer, 2008).

3. Aitchison, J. & Shen, S. M. Logistic-normal distributions: Some properties and uses. *Biometrika* **67** (1980).

4. Venables, J. P. *et al.* Cancer-associated regulation of alternative splicing. *Nature Structural & Molecular biology* **16**, 670–676 (2009).

5. Zukin, R. & Bennett, M. Isoforms of the NMDAR1 receptor subunit. *Trends in Neurosciences* **18** (1995).

6. Lee, J.-A. *et al.* Depolarization and CaM Kinase IV Modulate NMDA Receptor Splicing through Two Essential RNA Elements. *PLoS Biology* **5** (2007).