Supplementary Information:

Probabilistic Modeling of Alternative Splicing

In order to develop a probabilistic model of alternative splicing, which we approximate by the probability p_a of splicing to the second most abundant splicing state, we must first rigorously and mathematically describe p_a. To do so, we consider the following data: for each transcriptome, we examine the data set $\{D_i = \{(n_i, k_i)\}$, where n_i is the total number of reads of any given splice site S_i and k_i is the number of observations supporting the minor state. The p_a distribution is then a way of encoding this data such that $p_a(S_i) = n_i/k_i$. This distribution encodes the probability of alternative splicing events over each splice site in the data (Examples of such distributions are in Figure 3B and Supplementary Figure 5). Notice that for any splice site S_i and any given read of that particular splice site, there are two possibilities: S_i is either spliced to its major or minor splicing state (hereafter referred to as primary and secondary reads respectively).

Model M1: in which we model alternative splicing as a biased coin flip

The first obvious, naive model to implement is thus that of a biased coin flip: heads, for example, would signal a primary read, tails would signal a secondary read. This model M1 has a single free parameter *p*, the binomial probability, that is, the probability of seeing a secondary read for any given read of a splice site.

For an experiment with data $\{D_i = \{(n_i, k_i)\}$, the maximum likelihood estimate for *p*, $\hat{p} = \frac{\Sigma k_i}{\Sigma n_i}$.This is also trivially the posterior mode in a Bayesian framework, given a uniform prior distribution. Due to the large number of data points, choosing other non-informative priors, such as the Jeffreys Prior $Pr(p) = \frac{1}{\sqrt{p(1-p)}}$, does not alter the posterior mode. In Supplementary Figure 11A-C, we plot, for three representative experiments, the empirical probability distribution of p_a, along with the distribution predicted by model M1 using the maximum likelihood estimate. The model M1 clearly does not perform well (see Supplementary Figure 11A-C). In particular, it consistently underpredicts $P(p_{as})$ by orders of magnitude across all transcriptomes for $p_{as} > 0.1$.

Model M2: in which we allow the binomial probability to be drawn from a distribution

Above, we saw that, for any given experiment, having a single binomial probability *p* of alternative splicing occurrence was unable to capture the subtleties of the observed p_a distribution. For this reason, we next modeled alternative splicing by allowing p_a to be drawn from a distribution *P*. That is, for each splice site with n_i reads, we draw the probability of splicing p_i from *P* and then generate the number of minority reads k_i from the binomial distribution B(n_i , p_i) (note: if the generated k_i>n_i/2, we set k_i:= n_i-k_i to ensure that we retrieve the number of minority reads). As a first attempt, we assume that p_{as} is distributed uniformly on the interval [0,1]. This model has no free parameters. In Supplementary Figure 11D-F, we plot, for three representative experiments, the empirical probability distribution of p_a, along with the distribution predicted by model M2. We see that model M2 does not provide a good fit to the data as it generates, as expected, an approximately uniform distribution.

Note that model M1 can thought of as having a distributed p_a: it just so happens that in the case p_a is delta-distributed, that is, it has no variance. On the other hand, the variance of the uniform distribution assumed by model M2 has maximal variance. For this reason, we are led to consider utilizing distributions with an intermediate variance. As discussed in the main text, the data motivates us to consider the random variable p_a to be distributed according to a power law (also see Figure 4A).

Model M3: in which the binomial probability is distributed according to a power law

We define the model M3 as follows: assuming a given exponent $\alpha,$ for each (n_i, k_i) -pair we first generate a p_a from the power law distribution

$$P(p_a) = \frac{\alpha-1}{p_{min}^{-\alpha}-1} p_a^{-\alpha}, \text{ for } x_{min} < \alpha < 1 .$$

Note that the prefactor is a normalization term and guarantees that this is a bonafide probability distribution. It is also a subtlety of power-law distributions that they require a lower bound p_{min}in order for the integral to converge: in what follows, we used $p_{min} = 10^{-4}$ and decreasing p_{min} from this did not change the results. Next, we use the generated *p* to generate a number of minority reads k_i from the binomial distribution with binomial probability *p*, B(n,p). Thus, given the experimental data of (n_i, k_i)-pairs, for a given exponent, we are able to simulate the model M3 distribution.

It also follows that the probability of retrieving *k* minority reads, given *n* totals, can be described analytically:

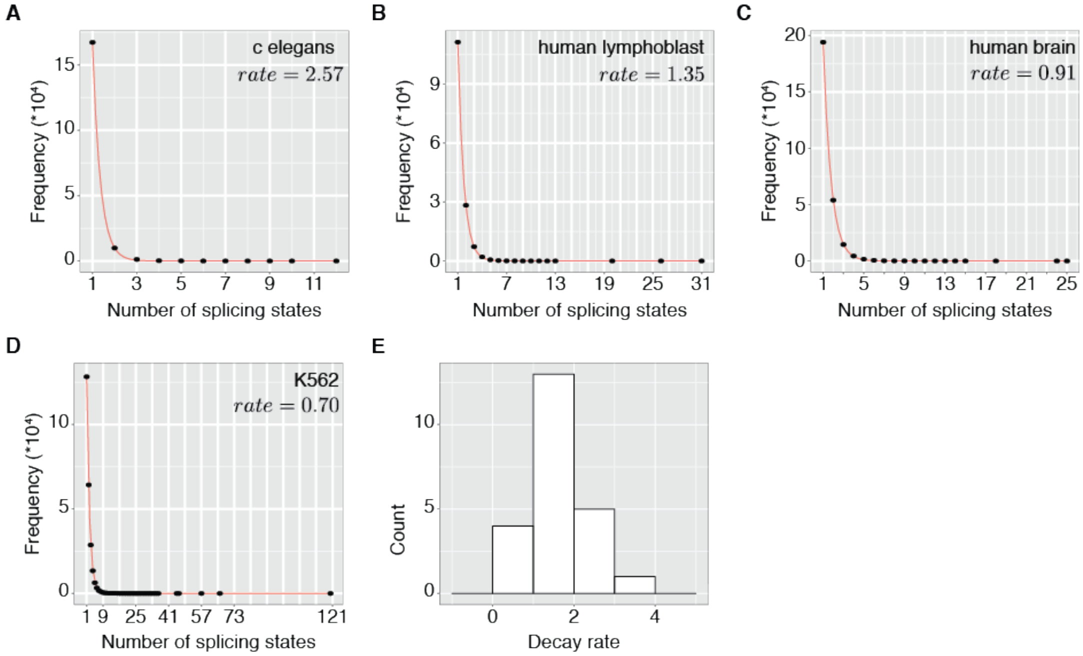
$$P(k|n) = \frac{n!\,(1-\alpha)}{k!\,(n-k)!\,(1-p_{min}^{1-\alpha})}\left(B(k-\alpha+1, n-k+1) - B(\alpha, k-\alpha+1, n-k+1\right),$$

where $B(k-\alpha+1, n-k+1)$ and $B(\alpha, k-\alpha+1, n-k+1)$ are the beta function and incomplete beta function respectively.

We need to estimate the parameter α for each transcriptome in light of the available data. Bayesian inference provides a theoretical framework in which to perform such parameter estimation. Indeed, we are interested in the mode $< \hat{\alpha} >$ of the posterior distribution $P(\alpha|D),$ this being the value of α that maximizes the probability of seeing the experimental data. Recall that the central equation of Bayesian inference is as follows: the *posterior* distribution $P(\alpha|D) = P(D|\alpha)P(\alpha)$, where $P(D|\alpha)$ is the *likelihood* (of seeing the data *D*, given α) and $P(\alpha)$ is the *prior* distribution on α. Due to the size of the data sets in question, the posterior is dominated by the likelihood function $P(D|\alpha)$and independent of the prior distribution (both the uniform prior $P_u(\alpha) = 1,$ for $0 < \alpha < 1$, and the Jeffreys prior $P_j(\alpha) = \frac{1}{1-\alpha}$, for $p_{min} < p < 1$, yield the exact same results). Thus the results reported are independent of the prior chosen and rely on only the data and no prior assumptions. Performing a parameter sweep by varying α allows us to calculate the posterior mode.

We eventually performed the parameter sweep on the interval [1.01,1.99] divided into 41 points (we did this as values > 2 consistently gave poor fits for representative data sets).

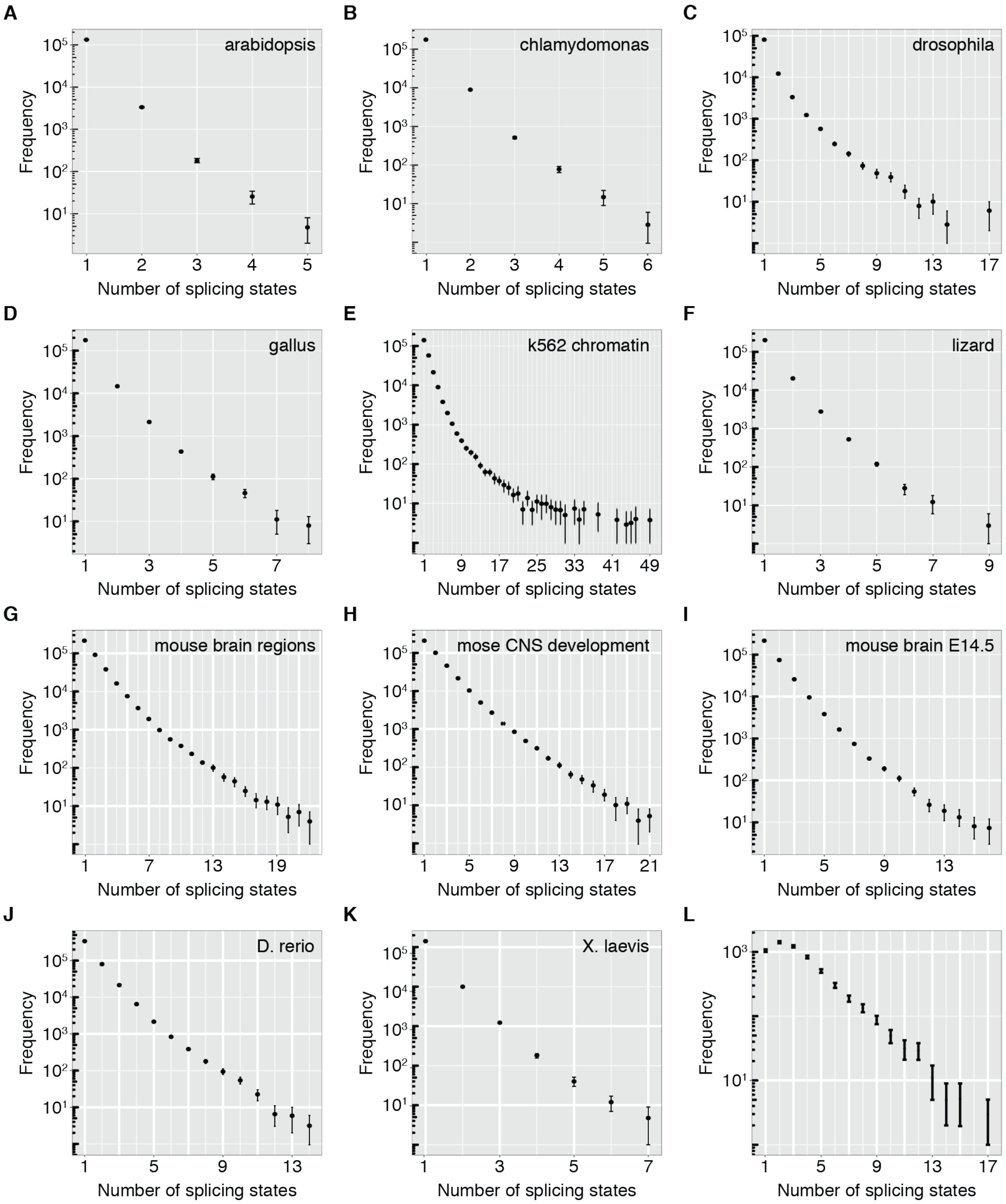
1. We took 10% of the data set (we will refer to this 10% as D_{10}) and performed what follows on it;
2. Given data set D_{10}, we simulate the model n = 5 times and bin the data into m = 20 bins to retrieve a 'simulated probability distribution'; we use this distribution to determine the likelihood function $P(D|\alpha)$(the results were verified using n = 10,15,20 and m = 5,10,15, 25);
3. We retrieved the posterior mode $\hat{\alpha}$;
4. We performed steps 1-3 k = 10 times to retrieve 10 values of $\hat{\alpha}$ and thus 10 values of $\hat{\beta} = 1/\hat{\alpha}$. We report $< \hat{\beta} >$ +/- SEM.



Supplementary Figure 1. Splicing state distribution decays exponentially.

(A-D) Splice site count (Frequency) versus number of splicing states for 4 transcriptomes are plotted in linear space. Red line indicates exponential fit, rate and transcriptome are given in inlets.

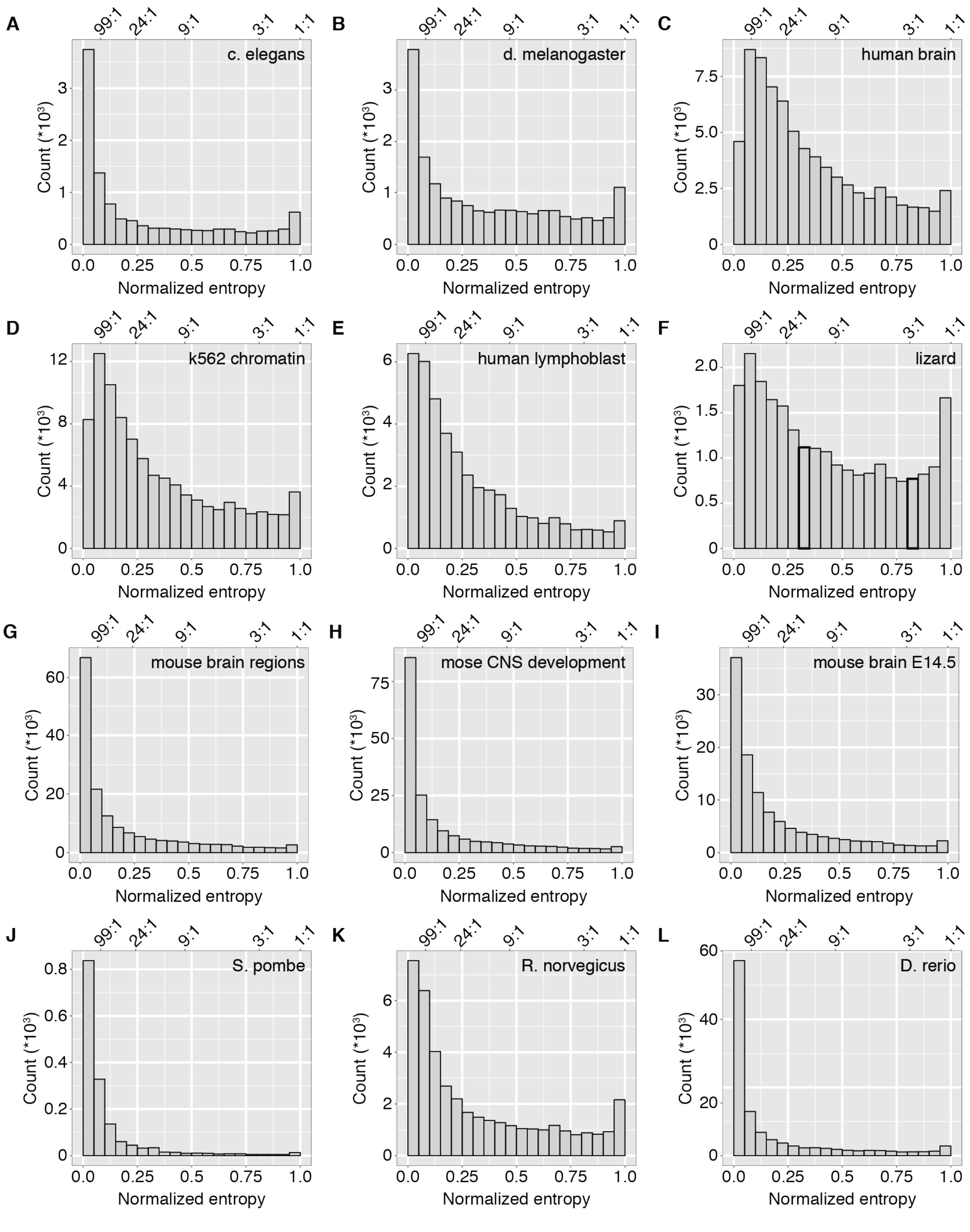
(E) Histogram of decay rates over all transcriptomes.



Supplementary Figure 2. Splicing state distributions.

(A-K) Splice site count (Frequency, log10) versus number of splicing states for 12 transcriptomes. For transcriptome information please refer to label in panel and Supplementary Table 1.

(L) Splice site count (Frequency, log10) versus number of splicing states for K562 splice sites with at least 2500 observations.

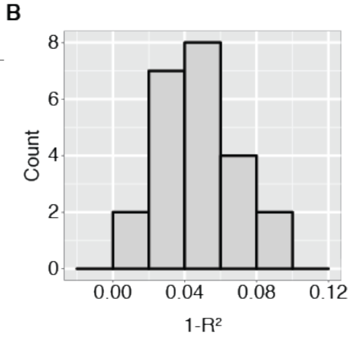


Supplementary Figure 3. Histogram of normalized entropies.

(A-L) Normalized entropy histogram over all splice sites of 12 transcriptomes. The ratios in a 2-state system (i.e. minor and major splicing state) yielding the corresponding normalized entropy are plotted (top of panel), indicating non-linear behavior. For transcriptome information please refer to label in panel and Supplementary Table 1.

A

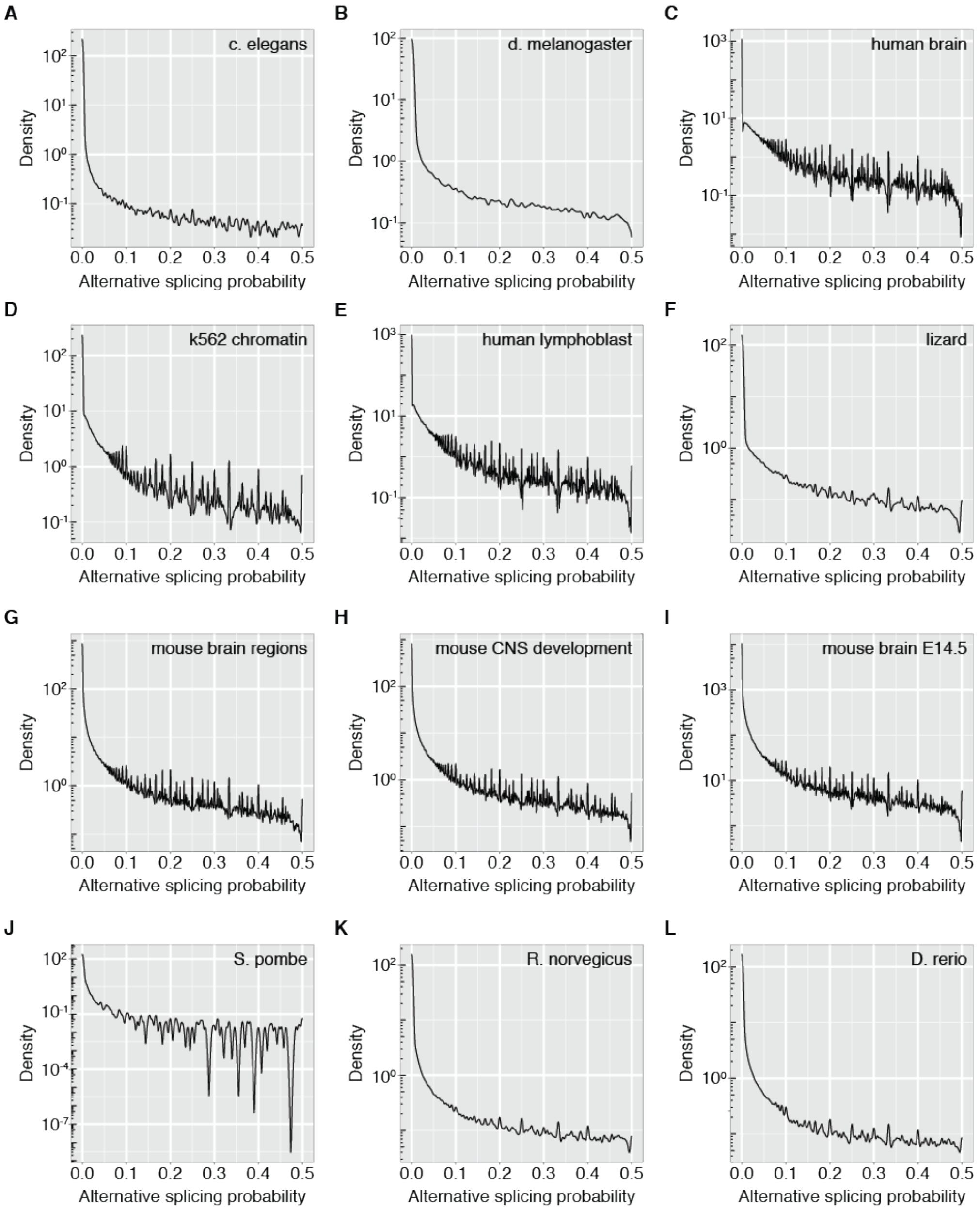
experiment	correlation coefficient
arabidopsis development	0.99
celegans development	0.99
chlamydomonas	0.99
drosophila tissues	0.96
frog brain	0.99
gallus brain	0.99
human brain	0.98
human K562	0.96
human K562 chromatin	0.95
human lymphoblast cambodian	0.98
human lymphoblast maya	0.98
human lymphoblast mbuti pygmy	0.98
human lymphoblast mozabite	0.98
human lymphoblast pathan	0.98
human lymphoblast san	0.98
human lymphoblast yakut	0.98
lizard brain	0.99
mouse brain regions	0.96
mouse CNS development	0.96
mouse whole brain	0.97
pombe	0.98
rat brain	0.98
zebrafish development	0.98



Supplementary Figure 4. Correlation of observed and approximated entropy.

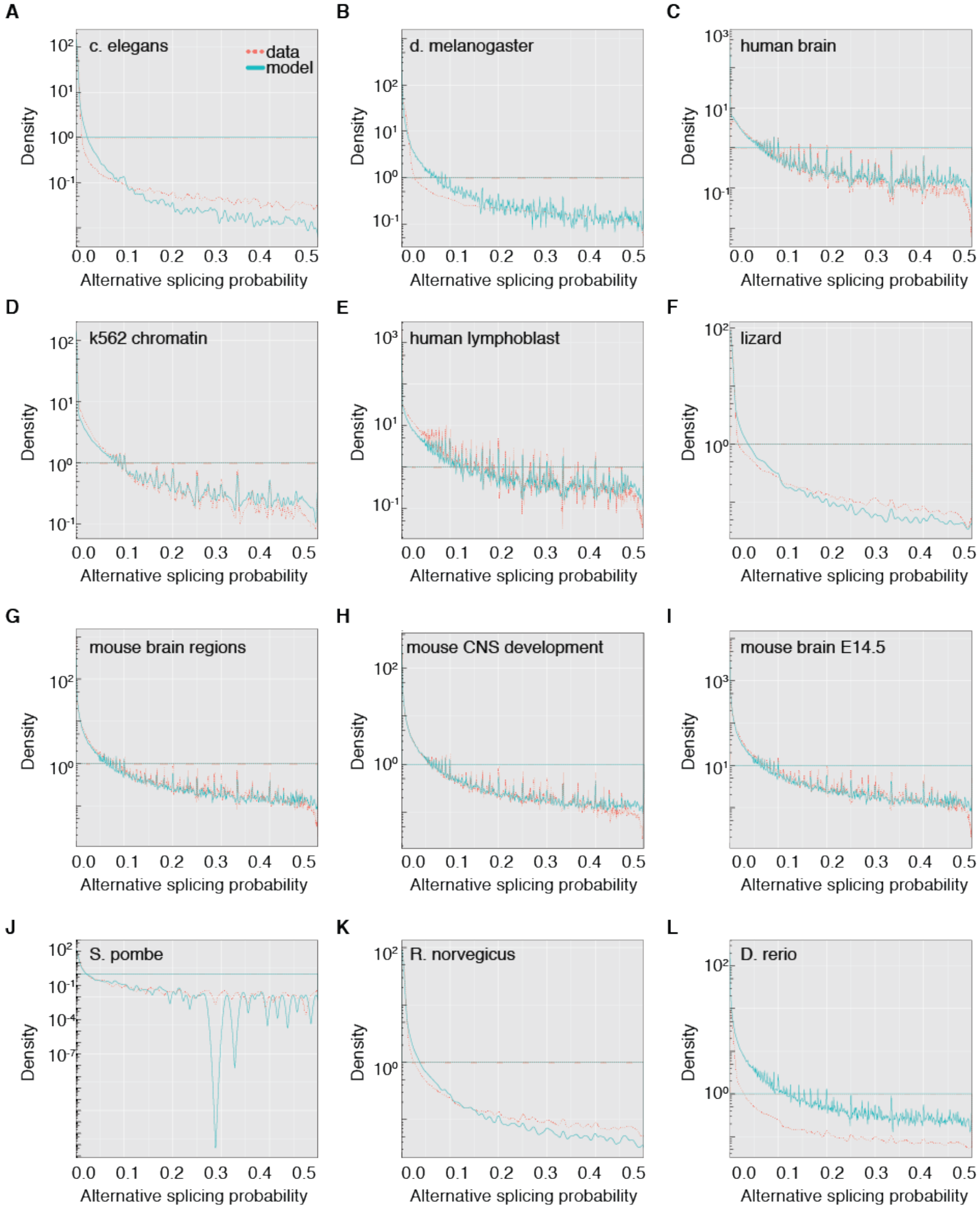
(A-B) Pearson correlation coefficient between observed and approximated normalized entropy for all transcriptomes.

(C) Frequency histogram of residual variances over all transcriptomes.



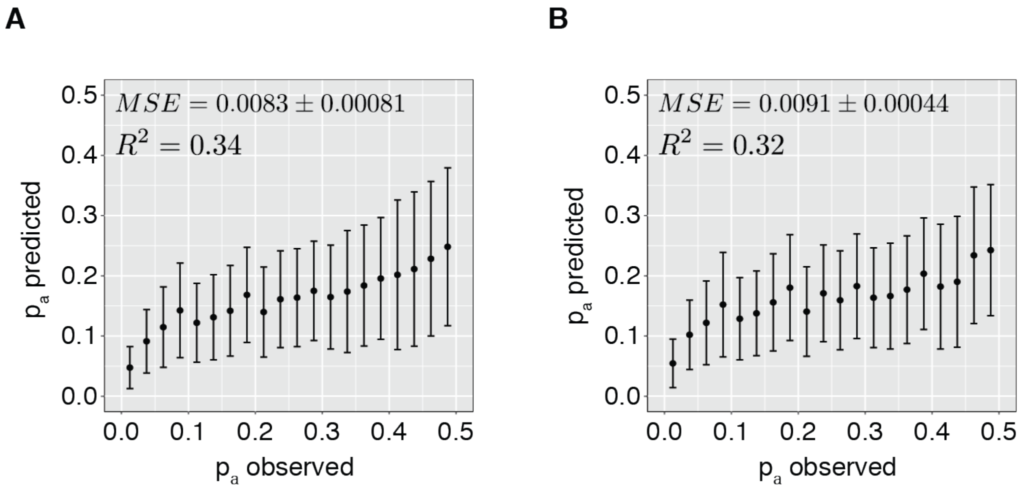
Supplementary Figure 5. Alternative splicing probability distributions.

(A-L) Log-density plot of alternative splicing probability for 12 transcriptomes. For transcriptome information please refer to label in panel and Supplementary Table 1.



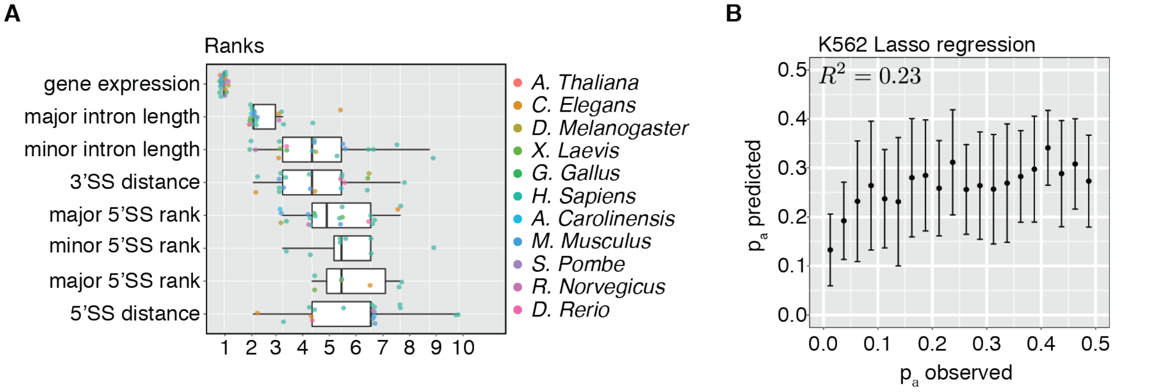
Supplementary Figure 6.

(A-L) Comparison of Model M3 and data across 12 experiments. We plot the probability density function of the alternative splicing probability pas for both the the data and Model M3 (for all experiments) on log-linear axes after performing a kernel density estimation.



Supplementary Figure 7. Random forest model approximates observed alternative splicing probabilities.

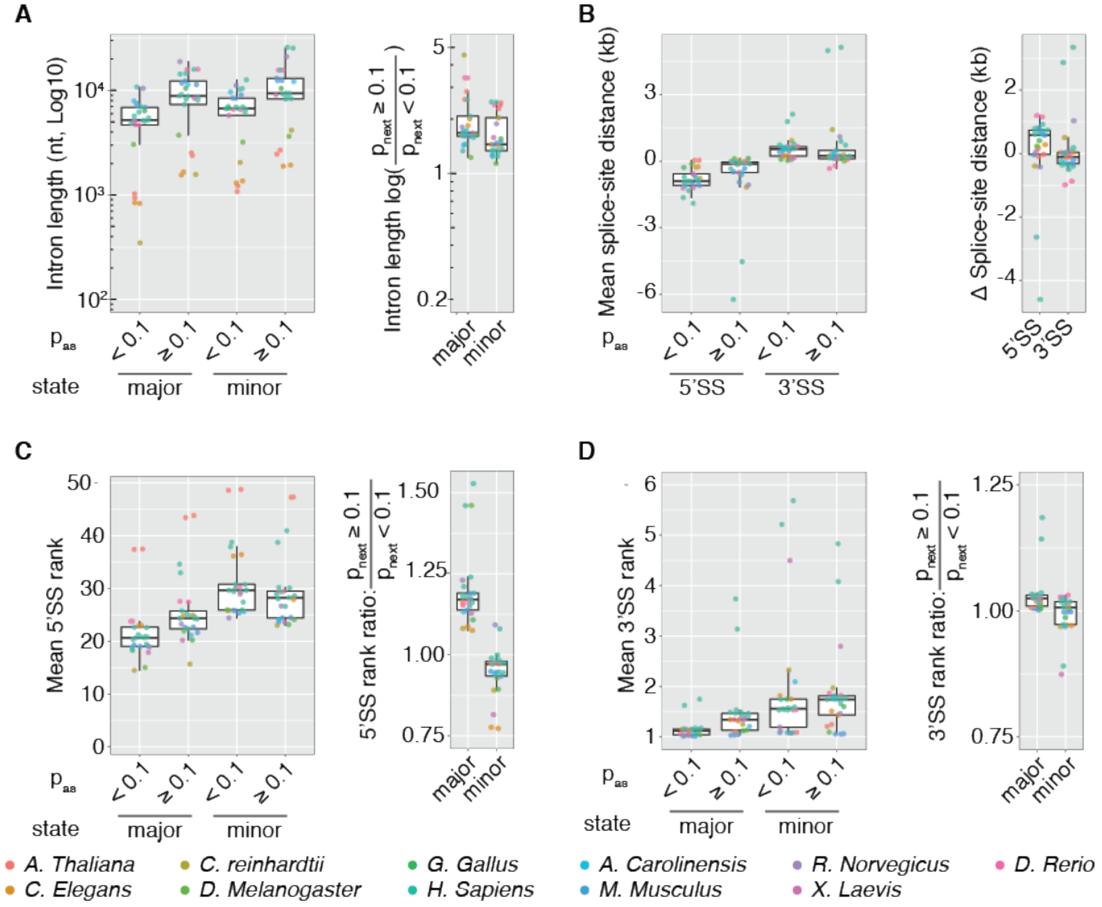
Alternative splicing probability (p_a predicted) predicted by random forest model versus observed (p_a observed) for full **(A)** and reduced **(B)** K562 feature set. Predicted values were grouped (0.025 bin) and mean and standard deviation plotted. Mean squared error (MSE) and variance explained (R^2) values are given.



Supplementary Figure 8. Species-specific random forest models identify a set of common features correlating with alternative splicing probabilities.

(A) Feature importance ranks were determined for each transcriptome (see Supplementary Table1). Feature ranks were grouped and plotted with increasing mean value from top to bottom. Feature name is indicated (left). Data points are color coded by species (legend on the right). Summary statistics are visualized with boxplots. Vertical bars indicate median value, boxes represent interquartile range.

(B) Alternative splicing probability (p_a predicted) predicted by Lasso regression model versus observed (p_a observed) for the reduced K562 feature set. Predicted values were grouped (0.025 bin) and mean and standard deviation plotted. Mean squared error (MSE) and variance explained (R^2) values are given.



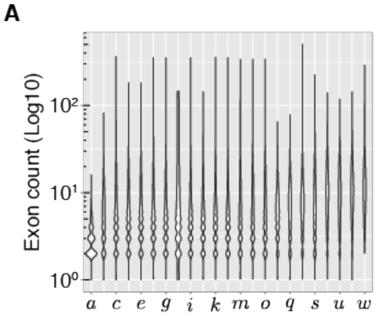
Supplementary Figure 9. Feature distributions for splice sites classified by alternative splicing probabilities.

Splice sites were grouped into high (p_a> 0.1) and low (p_a ≤ 0.1) alternative splicing groups and compared with respect to correlated features, identified by machine learning algorithms. Mean feature values were determined for each transcriptome (see Supplementary Table 1) and data-points color coded by species (see legends). Summary statistics are visualized by boxplots. Horizontal lines indicate median values, boxes represent interquartile range.

(A) Mean intron length is increased for splice sites with high alternative splicing probabilities: Mean intron lengths (log10, in nucleotides) of major and minor splicing states for low and high p_a groups (left panel). Intron length ratios (log10, high/low) for major and minor splice sites (right panel).

(B) Splice site distances are derived between major and minor splicing states are derived in the direction of transcription. Mean distances are visualized (in kb) for 5'SS and 3'SS grouped by low and high p_a (left). Relative distances (in kb) between mean values in low and high groups are plotted (right). In agreement with co-transcriptional competition, splice sites with high alternative splicing probability have tightly spaced competing splicing partners.

(C-D) Mean splice site rank of 5'SS **(C)** and 3'SS **(D)** were determined for major and minor splicing state grouped by low and high pas (left panels). Rank ratios (high/low) were determined and visualized (right panels). Note that high splicing rank correlates with low splicing site strength.



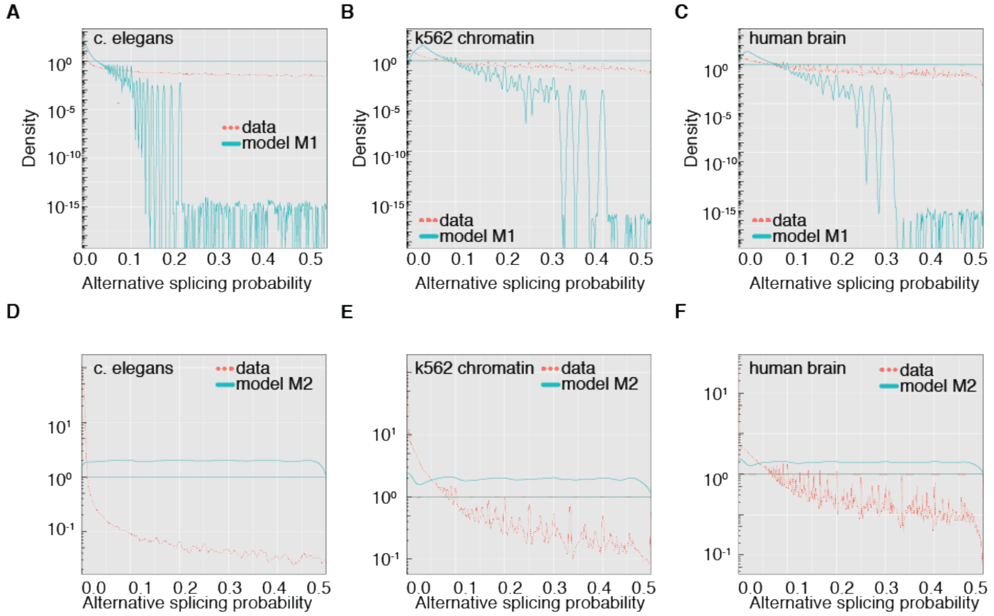
B

id	transcriptome	mean exon count
a	S. pombe	3.16
b	D. melanogaster	6.56
c	C. elegans	7.16
d	H. sapiens lymphoblast ma	7.26
e	H. sapiens lymphoblast c	7.26
f	H. sapiens lymphoblast p	7.26
g	H. sapiens lymphoblast mb	7.27
h	H. sapiens K562 mRNA	7.28
i	H. sapiens lymphoblast mo	7.28
j	H. sapiens lymphoblast y	7.30
k	H. sapiens lymphoblast s	7.41
l	H. sapiens K562 chromatin	7.47
m	H. sapiens brain	7.57
n	A. thaliana	8.04
o	M. musculus CNS	8.05
p	M. musculus brain regions	8.10
q	M. musculus brain	8.18
r	D. rerio	9.27
s	C. reinhardtii	9.58
t	R. norvegicus	11.89
u	A. carolinensis	12.15
v	G. gallus	12.71
w	X. laevis	12.99

Supplementary Figure 10. Number of exons per transcript.

(A) Exon count distribution (log10) over all splice sites are visualized by violin plots for all transcriptomes analyzed here. Transcriptomes are labeled by a single letter, referenced in **(B)**.

(B) Mean number of exons per transcript for all transcriptomes analyzed here. Note that only expressed genes are considered, leading to transcriptome specific differences in identical species.



Supplementary Figure 11.

(A-C) Comparison of Model M1 and data for three representative experiments. We plot the probability density function of the alternative splicing probability pas for both the the data and Model M1 on log-linear axes after performing a kernel density estimation.

(D-F) Comparison of Model M2 and data for three representative experiments. We plot the probability density function of the alternative splicing probability pas for both the the data and Model M2 on log-linear axes after performing a kernel density estimation.