6-23-2006

MULTIVARIATE ANALYSIS AND VISUALIZATION OF SPLICING CORRELATIONS IN SINGLE-GENE TRANSCRIPTOMES

Mark C. Emerick
Department of Physiology, Johns Hopkins School of Medicine, memeric1@jhem.jhmi.edu

Giovanni Parmigiani
Department of Oncology and Biostatistics, Johns Hopkins School of Medicine

William S. Agnew
Department of Neuroscience, Johns Hopkins School of Medicine

Multivariate Analysis and Visualization of Splicing Correlations in Single-Gene Transcriptomes

Mark C. Emerick[a*], Giovanni Parmigiani[b], William S. Agnew[c]

Departments of Physiology[a,c], Oncology[b], and Neuroscience[c], Johns Hopkins School of Medicine, and Biostatistics[b], Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205 USA

ABSTRACT

Through 'combinatorial splicing', RNA metabolism may create enormous structural diversity in the proteome. Functional interactions among multiple alternative domains can have a disproportionate impact on the phenotype, requiring integrated RNA-level regulation of molecular composition. Splicing correlations within molecules expressed from a single gene, where these effects would be greatest, provide valuable clues to functional relationships and targets for splicing regulation. We present tools to visualize complex splicing patterns in full-length cDNA libraries. Developmental changes in pair-wise correlations are presented vectorially in '*clock plots*' and linkage grids. Higher-order correlations are assessed via a loglinear model and Monte Carlo analysis with an empirical Bayes estimate of unobserved probabilities. log-linear coefficients are visualized in a '*spliceprint*,' a signature of splice correlations in the transcriptome. We present two novel metrics: the *developmental linkage index*, which measures the directional change in pair-wise correlation with tissue differentiation, and the *accuracy index*, a very simple goodness-of-fit metric that is more sensitive than the integrated squared error, applied to sparsely populated tables, and does not diverge at low variance, unlike chi-square. Considerable attention is given to sparse contingency tables, which are characteristic of single-gene libraries, but the methods apply to transcriptome analysis in general.

INTRODUCTION

Through alternative splicing at multiple sites, a single transcriptional unit may give rise to a complex array of isoforms—a 'mini transcriptome,' or single-gene transcriptome (SGT). Considerable effort is being invested in assembling a genome-wide compendium of sites of transcript and peptide variations (1, 2, 3, 4, 5, 6), the guiding principle being that combinatorial splicing may profoundly expand the proteome, and consequently the phenotypic repertoire, without increasing the number of structural genes (7).

Such phenotypic elaboration may arise through simple combinations of individual 'modular' domains, or through cooperative effects from multiple variable domains that interact functionally (8). The latter complication imposes significant regulatory demands—for the efficient selection of combinations of viable combinations—which may in part underly the expansion of the non-coding genome (9, 10). To understand gene function more fully, we must determine how all the possible alternative domains are actually combined by the RNA splicing process into working molecules. This requires structural analysis of large numbers of full-length transcripts expressed from *each* gene, approaching numbers currently available in full-length cDNA libraries derived from whole-genomes (11, 12, 13, 14, 15, 16) to ensure adequate representation of the less abundant variants.

EST, microarray, and proteomics methodologies for large-scale alternative splicing surveys share the limitation that by sampling fragments of macromolecules they cannot capture most intramolecular linkage information. That is, they primarily yield marginal splicing frequencies but not splicing correlations. The latter are invaluable in discerning the tissue- and cell-specific functional differentiation of splice variants (4, 17). Full-length cDNAs, by contrast, provide diaries of intramolecular splicing choices. We have developed robust methods for production of full-length, nonrecombinant, statistically representative single-gene libraries (SGLs) (18). An SGL is a vertical sample of the transcriptome, representing its basic building block, the SGT.

Initial analyses of moderate-scale SGLs (18, 19, 8) reveal fascinating developmental changes in both the extent and pattern of splicing linkage. These represent developmental changes in the regulatory programs that establish the selection rules for combining variable domains into functional ensembles, thus establishing the molecular phenotype. Recent developments in production of large-scale full-length cDNA libraries and high-throughput sequencing techniques (20, 21) promise to provide this information at high resolution and on a meaningful scale.

As large-scale SGT data become available, reliable statistical methods will be required to discern potentially complex splicing interactions. Two salient features of such data are (1) the particular relevance of higher order correlations, as they may relate to functional interactions within single molecules, and (2) sparse representation of the complete configuration space. Due to both functional constraints on domain combinations and the large number of potential configurations, a real SGL generally will typically comprise only a small portion of the total possible inventory of splice forms. Here we present tools for statistical analysis of high order splicing correlations in full-length SGLs, with careful attention to handling of sparse contingency tables. We provide simple, novel graphical visualizations that accentuate changes within groups of variable sites.

MATERIALS AND METHODS

Definitions

An alternative splice *site* is a categorical variable representing a region of the primary transcript that is subject to alternative splicing (we may omit the qualifier 'alternative' when the meaning is clear). We consider alternative

*Corresponding author

splicing a stochastic process, so that every alternative sequence *configuration* C_S, of any set of sites S, is a random variable with *splicing probability* $p_S(C_S)$. We say that a site j has a *strong splicing bias* if any $p_j(C_j)$ approaches 1. For a single site j, the configuration C_j may be represented by an integer between 0 and $g_j - 1$, where g_j is the number of sequence alternatives, or *multiplicity*, of j. For a cassette exon, spliced in or out as a unit, it is often convenient to assign 1 to the insertion and 0 to the deletion, although the reverse may occasionally be more convenient—if the insertion is rare, for example. What defines a site may also be flexible, depending on the purpose. An isolated cassette exon unambiguously defines a single binary site, with multiplicity 2. Two adjacent alternatively spliced cassette exons, however, may be considered two sites with $g = 2$ or a single site with $g = 4$. If the two exons are mutually exclusive then the best representation may be as a single site with $g = 3$.

If splicing at separate sites is independent (17, 22, 19), the expected frequency of each splice variant v is equal to its independent stochastic expectation

$$\phi_v = \prod_{j=1}^{k} f_j(v) \,, \tag{1}$$

where $f_j(v)$ is the marginal frequency of $C_j(v)$, the configuration of site j in splice variant v. The number of possible full-length transcript variants is

$$N_T = \prod_{j=1}^{k} g_j \,. \tag{2}$$

Computations and illustrations were made in the *R* programming language and environment (23).

Mutual information methods

We quantify splicing linkage between a pair of sites i and j with the mutual information $I(i,j)$, which measures the reduction in uncertainty about the configuration of one site when that of the other is specified (24): $I(i,j) = H(i) + H(j) - H(i,j)$, where $H(i) + H(j)$ is the expected entropy of C_{ij} given independent splicing at the observed marginal frequencies, and $H(i,j) = p(C_{ij}) \cdot log\ p(C_{ij})$ is the observed entropy.

While mutual information is non-negative, it is useful to define a directed, or 'configuration-specific' mutual information, which may be negative. The sign gives the direction of correlation between a specific pair of configurations, called the *reference configurations*. For example, if we define the reference configuration at a pair of binary sites as the insertion at both sites, then a negative value means that insertion at one site correlates with deletion at the other. The choice of reference configuration is arbitrary, and reversing the reference configuration for one site simply reverses the sign of the configuration-specific mutual information.

The *dependency*, $D(i|j) = I(i,j)/H(i)$, is the mutual information normalized to its maximum possible value, the total entropy of the 'dependent' site, i. It measures the degree to which the independent variable is a predictor of the dependent variable: the site with lower marginal entropy has a higher dependency on the other.

The developmental linkage index

To quantify developmental changes in linkage we introduce the developmental linkage index, S_D (Figure 1A-C). For a given pair of sites, we define the developmental linkage vector $D = (x, y)$, with components x — the splicing linkage in the fetal population, and y — the adult linkage. The difference $y - x$ is a simple measure of the developmental change in linkage: it is zero when linkage is the same ($x = y$) at both stages and maximal (positive or negative) for a complete reversal of linkage ($x = -y$). Scaling to the length of D gives

$$\begin{aligned} S_D &= (y - x)/|D| \\ &= sin\ \theta - cos\ \theta \\ &= \sqrt{2} sin\ \phi \end{aligned}$$

where θ is the angle between D and the x axis (Figure 1A) and $\phi = \theta - \pi/4$ is the angle between D and the transformed axis, x', representing unaltered linkage. In polar coordinates, D ranges in magnitude between about 1 and 1.4 times the larger of x and y, and has phase ϕ, the relative developmental change in linkage. Note that $sin\ \phi = y'/|D|$, so S_D is proportional to y', the perpendicular displacement of D from the x' axis.

S_D is a more straightforward index of linkage change than either ϕ or $sin\ \phi$ (*c.f.* Figure 1B): S_D is positive when splicing correlation becomes more positive or less negative with development ($0 < \phi < \pi$, magenta lines in Figure 1B and C), and if the correlation actually reverses direction from negative to positive, then S_D exceeds 1 ($\pi/4 < \phi < 3\pi/4$, quadrant II). Likewise, $S_D < 0$ reflects increasing negative (red lines), or decreasing positive correlation, and $S_D < -1$ means the linkage reverses, from positive to negative (quadrant IV). Figure 1C plots S_D as a function of the phase ϕ, annotated corresponding to Figure 1B.

Assessing higher-order linkages with log-linear models

Given a table of frequencies for N_T splice variants (Table S1), it is natural to arrange the data in a k-dimensional contingency table with k variables ($j = 1, \ldots, k$) of g_j categories ($C_j = 0, \ldots, g_j - 1$) each. This simplifies calculation of marginal frequencies. We fit the complete contingency table to a log-linear model (25), giving the log-frequency of each splice variant as a sum of coefficients $u_S(C_S)$, which measure the extent of mutual correlation among a set of sites S with configuration C_S. The most complete, or *saturated*, log-linear model is:

$$\begin{aligned} log\ p(C_{123\ldots k}) \quad &= \qquad\qquad\qquad\qquad\qquad\qquad (3) \\ \text{grand mean} \quad &\qquad u \\ \text{independence} \quad &+ \quad u_1(C_1) + u_2(C_2) + \cdots + u_k(C_k) \\ \text{main effects} \quad &+ \quad u_{12}(C_{12}) + \cdots + u_{k-1,k}(C_{k-1,k}) \\ \cdots \quad &+ \quad \cdots \\ \text{order k-1} \quad &+ \quad u_{123\ldots k}(C_{123\ldots k}) \end{aligned}$$

where subscripts refer to alternative splice sites. Each full-length splice variant $C_{123\ldots k}$ has a unique equation (3), giving N_T equations in all. The saturated model has a term for every possible subset of sites plus an intercept, u, with $\sum_{j=0}^{k} \binom{k}{j}$ terms in all. With all binary sites, the saturated model has 2^k equations of 2^k terms each. An unsaturated model is *hierarchical* if the presence of a u-term for any group of sites S implies a u-term for every subset of S.

For any site j in a set of sites S, the sum of $u_S(C_S)$ over all configurations of j is constrained to zero. Thus, for any C_S and C'_S differing only in the configuration of a binary site j, $u_S(C_S) = -u_S(C'_S)$. If all sites in S are binary, then all terms $u_S(C_S)$ have the same magnitude, $|u_S|$. Working with all binary sites thus simplifies the analysis, but does not otherwise alter the capabilities of the method. We use normalized frequencies, rather than total counts, to allow direct comparison of populations of different sizes without rescaling

Accuracy index

The 'accuracy,' $\mathcal{A}$, measures the extent to which a point, x_0, is centered within a distribution, $p(x)$. This has the advantage of extreme simplicity: it is the ratio of areas **a** and **b** in Figure 1D, where **a** is always the smaller of the two areas $\sum \{p(x) : x \le x_0\}$ and $\sum \{p(x) : x \ge x_0\}$. For a continuous pdf the two areas are $\int_{-\infty}^{x_0} p(x)dx$ and $\int_{x_0}^{\infty} p(x)dx$. Note that both **a** and **b** include $p(x_0)$. This is by design, as it yields a meaningful result for any x_0 and distribution $p(x)$. In the extreme low-variance limit, for example, if the $p(x)$ is an impulse $\delta(c)$, then $\mathcal{A}(x) = 0$ for all x except $x = c$, where $\mathcal{A}(x) = 1$. The accuracy is thus always defined, and ranges from 0, when x_0 lies completely outside the distribution, to 1, when x_0 is the median of the distribution.

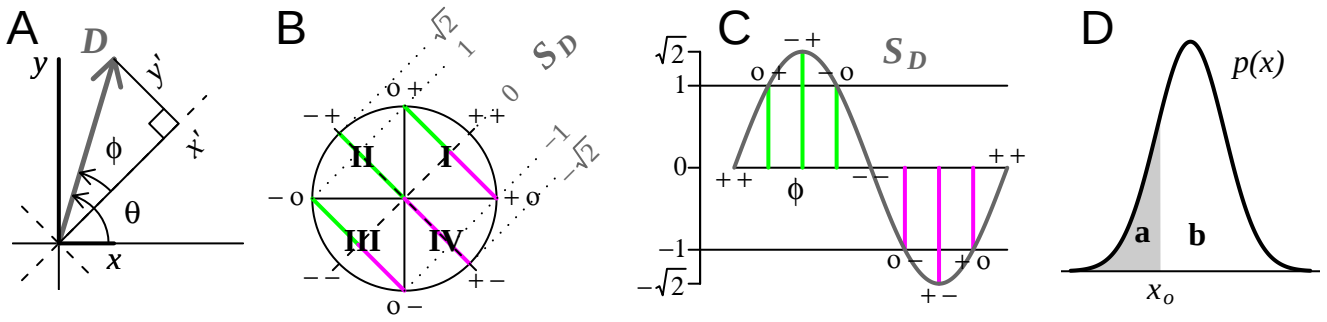

Fig. 1. Two novel metrics. A-C, Developmental Linkage index and B, Accuracy Index, $\mathcal{A} = \mathbf{a}/\mathbf{b}$. See text for details.

Supplementary data

Figures, tables and expressions labeled by numbers preceded by 'S' in the supplementary information.

RESULTS

Figure 2A plots splicing linkage between a pair of sites at one developmental stage *versus* another to display a developmental change in linkage. We call this a 'clock plot:' Each point is a vector whose magnitude measures the overall splicing linkage between the two sites and whose direction (displacement from the unit diagonal) indicates developmental regulation of linkage. Splicing may be developmentally regulated at both sites, but if they are regulated independently the plot point will fall on the origin, no matter how great the changes in splicing. If two sites are linked, but their *linkage* does not change with development, the point will lie away from the origin but on the diagonal. Thus, one pair of sites (1 and 2) shows a slight positive correlation between the reference configurations at the early stage, but this correlation increases greatly with development. The second pair also undergoes a developmental change in linkage, but in this case the sites become less correlated at the later stage. Linkage in this case is configuration-specific mutual information, which may be positive or negative. Plotting the configuration dependence allows us to see a reversal in the direction of correlation—manifest as a reflection about one axis—that may occur even in the absence of a change in mutual information.

Figure 2B is a clock plot displaying all 36 pairs of the nine alternative splice sites in the CACNA1G gene in fetal and adult human brain (data are in Supplementary Table S1). The points are dispersed primarily along the adult axis, indicating a general developmental increase in splicing linkage among most pairs of sites, an interesting exception being those that involve site **38B** (violet). Splicing at one pair of sites in particular, **25C** and **26**, is highly linked, with insertion at one site favoring deletion at the other in both stages of development, but much more strongly so in the adult than in the fetal brain. Several sites show considerable pair-wise splicing linkages with multiple other sites. We note that domains that correlate structurally in this way are good candidates for some kind of functional relationship, and multiple pair-wise splicing linkages to a single site, as seen here, may reflect either simple pair-wise functional interactions or a higher-order interrelationship. We explore the latter possibility in the next section.

Figure 2C, plots the (configuration-specific) *dependency* of **25C** on **26** (orange) and that of **26** on **25C** (blue). The dependency measures the extent to which splicing at one site predicts splicing at the other. Unlike mutual information, the dependency is an asymmetric function of the two sites, and may reveal relationships that are less apparent with mutual information: compare Figures 2D and E. In this case the strong developmental change in linkage is manifest more as a change in dependency of **26** on **25C** rather than the other way around. This reflects a greater discrepancy in entropy of those two sites in the fetal population than in the adult (Table S2).

The uncertainty, indicated by the dispersion around each data point in Figure 2C-E, for example, is obtained from 1000 simulated populations for each tissue, sampled by Monte Carlo from the Empirical Bayes estimate of the distribution of splice variants in each tissue, described below. For every pair of sites in each tissue, the mutual information (or dependency) values were binned into a histogram. Because splicing in the two tissues may safely be considered independent, the two-dimensional joint distribution for a pair of sites is the Cartesian cross product of the resulting bin-counts vectors from the two tissues. The error ring encloses the 95% most-probable values in this case.

Figure 3A displays linkage grids, showing the splicing dependency of all pairs of sites in a population that are statistically significant at the level of 99%. The rows give the dependent variables and columns the independent variables. In this way we may compare in adjacent grids the extent of splicing linkage in two populations for all pairs of sites exceeding a desired significance level. It is readily apparent in these plots that brain maturation entails the appearance, or strengthening of a considerable number of pair-wise correlations.

Figure 3B plots the dependency values for those sites showing a statistically significant *developmental change* in linkage. Thus, whereas in Figure 3A the significance was determined separately for each tissue, in Figure 3B it was determined for both jointly. Note that while sites **25C** and **26** show a high degree of negative dependency in both tissues (*c.f.* the red cells in the small white-bordered box within each grid of Figure A), the dependency of **25C** on **26** does not change significantly with development, whereas the reverse dependency increases somewhat. This reflects the different positions of the two points in Figure 2C, where the orange ring touches the diagonal.

We have introduced the *developmental linkage index*, S_D, to quantify these changes (*c.f.* methods). Figure 3C plots S_D and $|D|$

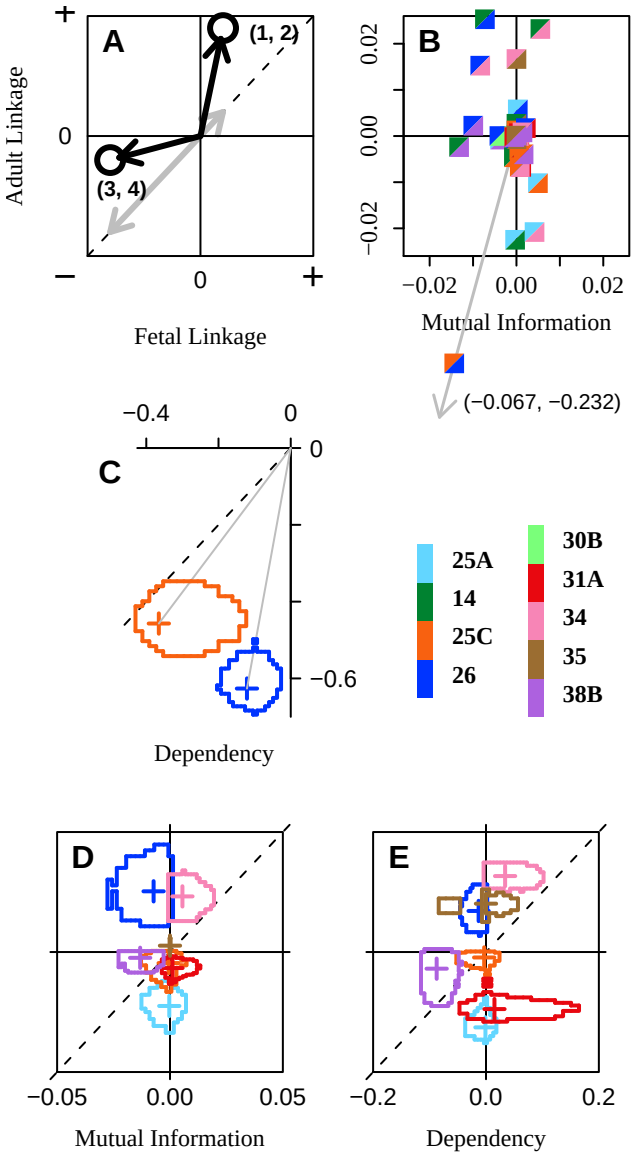

Fig. 2. Clock plots. **A**, Illustrative example. Each circle represents a pair of splice sites (e.g. sites '1' and '2'). The gray vectors depict splicing linkage that is present but does not change with development. **B**, Mutual information clock plot for all 36 pairs of the nine splice sites in the fetal and adult cDNA populations. Splice-site pairs are identified in the plot symbols by colors defined in the key. The reference configuration is the insertion for every site. **C**, Dependency clock plot for the single pair of sites $(\mathbf{25C}, \mathbf{26})$. orange: $D(\mathbf{25C}|\mathbf{26})$; blue: $D(\mathbf{26}|\mathbf{25C})$. **D**, Mutual information clock plot for the eight pairs involving site **14**: $I(x, \mathbf{14})$, $\forall x \neq \mathbf{14}$. **E**, Dependencies, $D(x|\mathbf{14})$, of the same pairs.

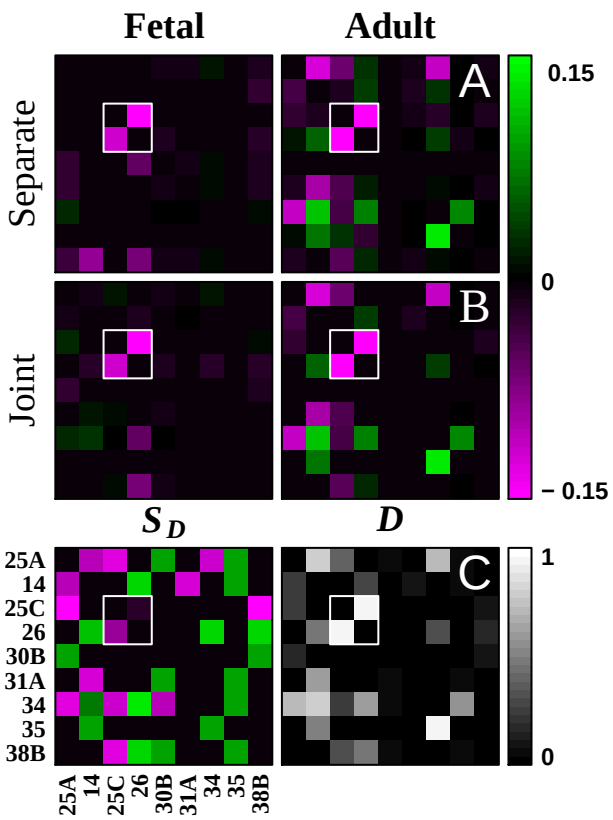

Fig. 3. Linkage grids. **A, B.** Dependency, $D(i|j)$, of splicing at one site i on a second site j is plotted for all pairs of sites in the fetal and adult cDNA populations. The abscissa lists the independent sites. The layout is the same for all grids. Only dependencies at or above 99% significance are displayed. Statistical significance was determined for the two tissue samples either separately (A) of jointly (B). For a given tissue, the same values are plotted in A and B when the linkage is significant in both. Note that a linkage may be significant in one tissue but not both *or vice versa*. **C.** The developmental linkage index (left) measures the extent of *change* in linkage at each pair of sites. Only changes significant at or above 99% are shown. The right panel plots the magnigtude of the developmental linkage vector, which gives an indication of the overall level of linkge at each pair of sites.

for those pairs of splice sites with nonzero S_D at 99% significance or greater. The left grid shows the dynamic range of directional changes, while the right shows the overall magnitudes of the linkages involved. An interesting point of comparison is the values within the white-bordered box representing sites **25C** and **26**. These cells are quite dim within the S_D grid, whereas they are bright in the other grids. This shows that, while there is strong splicing dependency between these two sites at both stages, and there is a statistically significant change in linkage with development, the extent of change is actually not very great in comparison to that at other pairs of sites. As we noted above, though, the dependency of **25C** on **26** undergoes a greater developmental change than the reverse dependency; this is more apparent in the plot of S_D than in the other plots. Other significant changes are much more obvious here as well.

Spliceprints

Up to this point we have considered splicing linkage between pairs of sites. It is possible for splicing to involve correlations of higher order. For example a segment may be deleted at site 1 only if segments are inserted at both of sites 2 and 3. This situation necessarily entails pair-wise correlations between sites 1 and 2 as well as 1 and 3, but a three-way linkage is more intricate than a collection of disjoint pair-wise linkages. We model

multi-site splicing with a log-linear model (25) to quantify higher-order linkages. Figure 4 displays the amplitudes of the log-linear coefficients from two developmental stages, for all terms of order 1 or higher in the saturated model. The contributing splice sites are not identified, but terms for subsets of the same cardinality are grouped together between vertical rules. In each plot, values above the zero line are coefficients derived from the Empirical Bayes estimate of the experimental population. For comparison, traces below the midline plot mean values from 1000 Monte Carlo populations. Because only the magnitudes are plotted, the ordinate values increase with distance from zero both upward and downward. Differential splicing regulation in the fetal and adult brain appears as different patterns in the upper and lower boxes. Notice that independent splicing gives coefficients that lie on the zero line, and that the more compressed fetal pattern indicates a lower level of splicing correlations, especially at higher orders. Although we focus only on the magnitudes of the log-linear coefficients in the present work, a wealth of information is present in their signs, which would admit a multidimensional extension of the clock plot analysis. Also, of course, the spliceprint is not limited to log-linear coefficients, and the coefficients may be presented in any desired order on the abscissa.

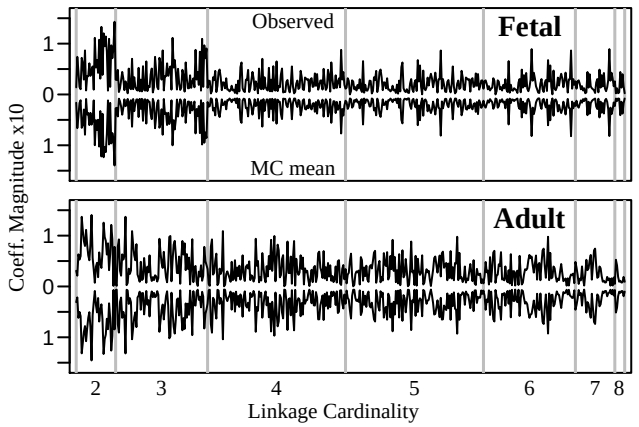

Fig. 4. Spliceprints. Log-linear coefficient magnitudes are plotted for all subsets of more than one site. The vertical scale is the same for all traces. Within a cardinality, k, the sequence of coefficients is determined by listing the 9 sites from left to right as in figure 3 (abscissa), and choosing groups of sites from left-most to right-most as follows: for sites A = **25A**, B = **14**, etc., cardinality-3 coefficients occur in the order ABC, ABD, ACD, BCD, ABE, ..., FHI, GHI.

Minimal-linkage models

We may wish to identify the smallest set of interactions that can account for the data within bounds of statistical significance. It may be surprising, for example, that some genes display nearly independent splicing at multiple sites, even in rather complicated tissues 17, 19, 8. In such cases one or two pair-wise interactions may account for any deviations from independence. We may identify those by first ranking the pairs in order of decreasing mutual information, then define a hierarchical model with only the most highly correlated first-order interaction terms. A least-squares fit to this model gives coefficients for a table of frequencies exhibiting only those interactions. We then ask whether this represents a possible parent distribution for the observed population.

A simple way to do this is to sample Monte Carlo populations from the fitted distribution, with the same number of transcripts, N, as the experimental population. For each MC population we then calculate its likelihood of arising from a reference distribution, ρ, for example:

$$\rho = N! \prod_{v=0}^{N_T-1} p_v^{n_v}/n_v! \qquad (4)$$

This is a multinomial distribution of N_T splice variant classes, v, each with expected probability p_v and abundance n_v in the sample: $\sum n_v = N$. The reference distribution is in fact arbitrary: we are not interested in the exact likelihood of our data; rather, we wish to find a model that generates populations of likelihood similar to the data with a given reference distribution. The fraction of MC likelihoods not exceeding the observed value measures the evidence against the model. This approach is an approximation to the method of posterior predictive assessment of model fit of Gelman *et al.* (26). Expression (4) constitutes their statistic T. We seek a model with sufficient departure from independence to be consistent with the observed data.

Figure 5 illustrates this method with the adult data. It shows log-likelihood histograms for six Monte Carlo ensembles of 1000 populations each. The reference distribution is the independent-splicing expectation [equation (1)] of the experimental population. The horizontal rule at -143 depicts the experimental log-likelihood. The first ensemble (left-most histogram) was sampled from the reference distribution, and shows that independent splicing is inconsistent with the experimental data. The reference distribution gives the 'cardinality-1' log-linear model, $log\ p(C_{123\ldots k}) = u + \sum_j u_j(C_j)$, with all independence terms and no interactions. The next histogram is obtained by adding a single interaction to this model: $u_{34}(C_{34})$, the pair-wise interaction between sites **25C** and **26**, identified by mutual information as the most highly correlated segments. Each subsequent ensemble is obtained from the previous by adding the next most-correlated pair. A minimum of five pair-wise linkages are thus required to account for the observed splicing correlations by a stochastic mechanism within bounds of 95% confidence.

Saturated models

Peptide domains that interact functionally are likely to exhibit statistical correlations reflecting enrichment of productive interactions or suppression of detrimental ones. Figures 2 and 3 show that many individual sites participate in pair-wise interactions with multiple other sites. Where these reflect functional interactions, we anticipate two important consequences, due to the fact that they occur within a close-packed, folded protein: (i) they are likely to be *transitive* in nature—*e.g.*, if sites A and B interact and sites B and C do, we expect that A and C will as well. Furthermore, we should expect that splicing at C will have effects that depend on A and B *together*, *i.e.* (ii) a set of sites may influence protein function as an integrated *ensemble*, rather than a collection of functionally separable modules. It then follows that their splicing regulation will exhibit mutual, higher-order interdependencies. Unsaturated models cannot capture these correlations accurately: A low-order model, as in Figure 5 for example, is from this perspective an

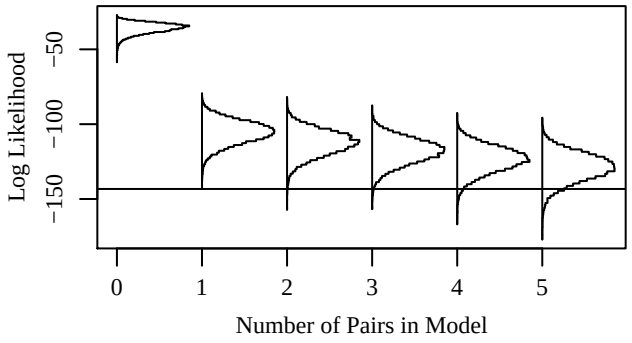

Fig. 5. Posterior predictive assessment of minimal model fit. A simple model may account for the observed distribution within admissible error limits, but it may cause one to overlook important effects in a network of interactions.

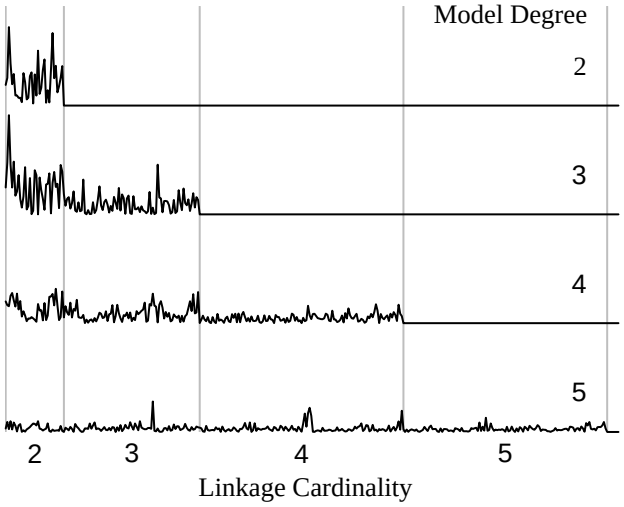

Fig. 6. Spliceprints of successively higher-order hierarchical models fit to the same data. Excluding high-order effects from the model misrepresents the lower-order interactions. The vertical scale is the same for all plots.

oversimplification, emphasizing a few low-order interactions at the expense of a wealth of information in the higher orders.

Figure 6 shows the distribution of amplitudes among the first 4 orders of interaction terms (cardinality $k = 2, \ldots, 5$) from fits of four hierarchical models to the same data (adult population). Each panel plots coefficients from a model including all terms of cardinality k and lower, but none higher. When excluded from the model, high-order interactions are 'absorbed' into lower-order terms. Notice, for example, the redistribution of *relative* amplitudes within the cardinality-3 coefficients as higher-order terms are included in the model. This occurs as weight from triplets in higher-order linkage groups is shifted to their higher-order coefficients when they are made available. Since we wish to compare coefficients estimated under identical models from parallel data sets, we use a saturated model to avoid confounding the low-order terms with higher-order effects. A nonzero coefficient for a set of sites then indicates a mutual splicing dependency among all sites in the set, in excess of any lower-order interactions that may be present among component subsets, and larger magnitudes reflect stronger correlations.

Empirical Bayes methodology

We are not interested in the precise values of the coefficients as much as the relative amplitudes of coefficients from two different populations, as compared side-to-side in Figure 4, for example. From this we may discern statistically significant developmental shifts in splicing linkages. This requires that we estimate the variance of the log-linear coefficients. These may be obtained from a saturated model fit to Monte Carlo populations sampled from an estimate of the parent distribution.

The simplest such estimate is the observed distribution itself (the bootstrap). While this makes no assumptions about the underlying mechanism, it assigns zero probability to the unobserved splice forms, which is obviously unreasonable. Increasing the experimental sample size, even by an order of magnitude, may not make the bootstrap applicable if the number of alternative splice sites is even moderately large (Figure S2): the probability space expands geometrically with the number of variables, so unless the sample is vastly larger than the number of classes the table of observed frequencies will contain a large number of zeros (empty cells). This is exaggerated when splicing at any site is strongly biased, as is common (*e.g.* Figure S5). Transcripts that combine rare splice configurations at multiple sites thus have a very low expectation, though we cannot assume that any empty cells would persist if we continued data collection indefinitely.

The empirical Bayes approach (27) enables an estimate of the parent distribution with plausible nonzero probabilities for the unobserved classes. This estimate (the *posterior* distribution) incorporates the observed distribution (the *likelihood*) with our current understanding of the underlying process (the *prior* distribution). We have found the 'pseudo-Bayes' estimator of Bishop *et al.* (25), a linear shrinkage estimator chosen for its simplicity, to be entirely adequate. Improvements may be made to the estimator—with nonlinear shrinkage, for example, but typically at the expense of added complexity. Our models do exhibit sensitivity to the choice of prior distribution, however, because of the sparse representation of splice forms in the experimental data. We present a thorough examination of different priors in the Supplementary Information. Because splicing at separate sites is approximately independent (Figure S4), equation (1) provides an excellent prior: the *tissue-specific independent marginals* prior. In this work we primarily use the *averaged-marginals* variant of this prior, obtained by averaging the fetal and adult expected frequencies.

The empirical Bayes methodology is open to the criticism that including experimental results in the prior may lead to duplicate use of evidence and subsequent underestimation of uncertainty. Purportedly 'uninformative' priors inadvertently introduce their own errors, however, mainly by forcing untenable splice correlations into the estimator (*c.f.* Supplementary Information; also Fig. 8). By making judicious use of the observed marginal frequencies in the prior we minimize this effect, and keep the focus of inference on the interactions.

Developmental changes in higher-order linkages

Figure 7 presents a statistical summary of Figure 4, obtained with the averaged-marginals prior. The adult population displays enhanced higher-order correlations compared to the fetal for groups of up to at least 6 sites. This agrees with the mutual information results (*e.g.* Figure 2B), with an interesting additional feature: the fetal and adult profiles are most divergent at cardinalities 4 and 5 with the gap closing toward cardinality 2. This shows that much of the difference in mutual information between the two tissues derives from extensive splicing correlations involving sets of considerably more than two sites, whereas the extent of isolated pair-wise interactions is more nearly comparable in the two tissues. Elsewhere we demonstrate that these higher-order splicing correlations correspond to nonlinear functional interactions involving multiple domains in the expressed ion channel protein (8).

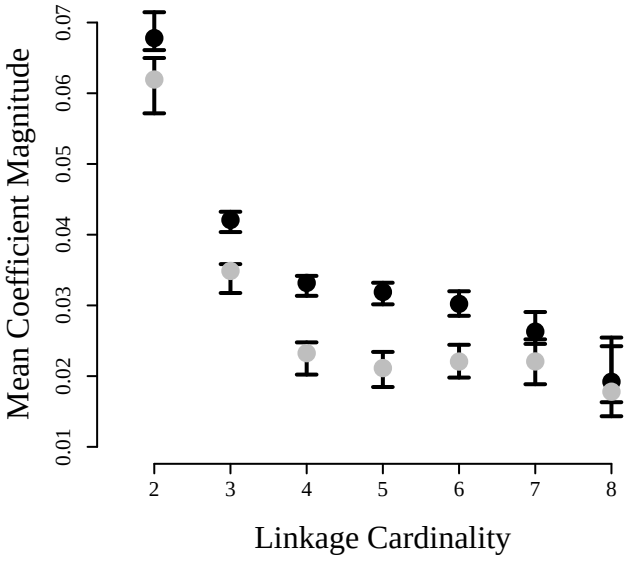

Fig. 7. Cardinality-averaged log-linear coefficients. All coefficients of the same degree are presented as a single average magnitude. 1000 Monte Carlo populations were sampled from the empirical Bayes posterior obtained with the 'averaged-marginals' prior and fetal (gray) or adult (black) likelihood. Error bars delimit the $2.5 - 97.5\%$ interquantile range for each distribution.

The SGT is a transcriptome

A realistic splicing model allows us to investigate the single-gene transcriptome with established methods of transcriptome analysis. The transcriptome is a highly complex assortment of gene products, but it exhibits a remarkably stable expression pattern. Only a few genes are expressed at a high level, while most genes are represented by only a few copies. It is not clear that this pattern should persist at the single-gene level. Different physiological inputs would affect the profile at different levels, so those aspects of gene-network topology that conspire to shape the aggregate gene expression profile, for example, may or may not be relevant to the selection of splice isoforms from a single gene in separate cells or tissues.

Nonetheless, the basic characteristics of the transcriptome profile are also present in its building block, the SGT. Figure 8 shows 'frequency-of-frequencies' plots for simulated SGTs sampled from the tissue-specific independent-marginals distribution (filled dots). The reverse-J pattern, like those obtained in genome-wide expression profiles assayed by SAGE (28), reflects the complexity of both the transcript inventory and the tissue physiology in which this gene is expressed. The identical profile was obtained with the averaged-marginals estimator (squares), which places a lower reliance on the observed marginal splicing frequencies in either tissue. The uniform prior, however, gives an idiosyncratic L-shaped profile with an abrupt lower copy-number limit (open circles). This reflects the implicit exchangeability of the unobserved classes: all have the same low probability, but because the majority of forms are not observed, their cumulative probability in the estimator is large. This is one example of how the uniform prior (or any prior obtained by a small constant correction to the observed frequencies), although 'uninformative', is overly simplistic, and leads to artifacts.

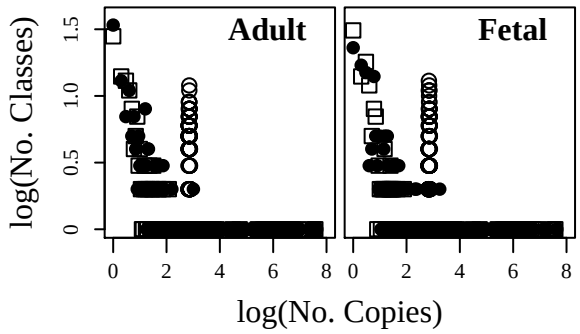

Fig. 8. Frequency-of-frequency plots. Monte Carlo populations of 10^8 transcripts each were sampled from the empirical Bayes estimate obtained with the fetal or adult likelihood and either the tissue-specific independent marginals (**dots**), averaged-marginals (**squares**), or uniform (**circles**) prior.

DISCUSSION

The human genome supports in the neighborhood of 23,000 protein coding genes (29), very similar to the number found in genomes of vastly simpler organisms, such as *C. elegans* and *Drosophila* (30, 31). To account for the increase in human phenotypic richness, therefore, the number of structural genes is not as important a factor as the way in which genes are used. Variations in gene expression levels, changes in the timing of expression, evolutionary adaptations that rearrange gene interactions as well as evolution of the coding sequence, and increased post-transcriptional modification of primary transcripts to diversify the products of single genes all play a role (32, 9).

Here we present tools to evaluate and visualize complex patterns of transcriptome variation, illustrated on populations of full-length cDNA splice variants from CACNA1G, the gene encoding the human $Ca_v3.1$ T-type calcium-channel α_1 subunit. In the course of brain maturation the transcriptome of this gene undergoes a transformation that would be largely invisible to a study of gene expression levels or a microarray- or EST-based splicing survey. The changes appear only in the complete structures of full-length transcripts, as alterations in splicing correlations at separate

sites within the same molecule. A standard analysis of pairwise correlations, while illuminating, is incomplete in an important way. Compared to the fetal, the adult transcriptome displays a marked increase in mutual information between many pairs of sites (Figure 3). The multivariate analysis, however, reveals two components of this increase: a modest elevation in disjoint pairwise linkages and a substantial increase in higher-order correlations that include linked pairs as a subset. Overall, splicing in the adult is far more restrictive than fetal splicing. This occurs at the same time as the range of cell types in which this gene is being expressed is diversifying, not constricting. This is consistent with the notion that splicing may need to be more stringently specified in the more intricate 'wiring' of the mature brain (8).

It is the grounding principle of this work, therefore, that splicing correlations will generally reflect functional interactions, and that these are likely to involve multiple domains. Splicing of physically linked domains should be co-regulated to inhibit detrimental interactions as well as to enhance beneficial ones. This relates directly to the complexity of the processes that regulate selection of alternative domains, the most important factor being whether the domains are modular or functionally interactive.

Modular domains may be shuttled in or out with predictable effects, independent of splicing at other sites. They may be used to conjoin functional activities—post-synaptic targeting with fast activation gating in an ion channel, for example. *Interactive domains*, in contrast, express a shared functional effect that exists only in the context of the ensemble. A specific effect cannot be independently defined for a single interactive domain: reconfiguring one such domain 'reinterprets' the functional influences of the others. That is, the molecular phenotype may be expressed as a linear combination of the effects of modular domains, but not so for interactive ones. As an example, deleting segment **38B** of the $Ca_v3.1$ calcium channel decreases the window current magnitude when **25C** is present, but increases it when **26** is present, and has no effect when both are absent; furthermore, it does not effect gating rates, except when **14**, **25C** and **26** were all absent, whereupon it speeds inactivation (8). Whether domains interact functionally depends on the domains, and modular and interactive qualities are not mutually exclusive.

The number of alternative molecular phenotypes is the same whether the sites are modular or interactive. In the former case, however, any given state is decomposable into identifiable subsets of phenotypes, whereas in the latter it is not. Functional interactions admit the possibility of introducing completely unpredictable, *qualitatively* novel behavior simply by reconfiguring an existing set of domains. In the course of evolution, the simple addition of a new variable domain may reinterpret the phenotypes of existing splicing patterns, enabling a rapid expansion of functional alternatives from the ancestral gene.

Nonlinear interactions may have various causes. Inserting one domain may simply block access to a binding site for a second domain, for example. Another possibility is an 'allosteric' type of interaction where electrical or conformational changes communicate through the protein interior. The consequences of such interactions may become even more complex when other genes are alternatively spliced in multiple ways. Current estimates of the number of alternatively spliced genes in humans range to $\sim 76\%$ of known genes (33), with an average of 3.9 splicing isoforms per gene (1). Furthermore, the frequency of alternative splicing is elevated in genes that mediate or modulate cell signaling and metabolic networks (33), increasing the likelihood of nonlinear, and largely unpredictable interactions *between genes* with alternatively spliced forms that communicate through such networks. Strong intergenic interactions are of course normal where proteins contact physically, as subunits of a multi-enzyme complex, or in a multi-subunit ion channel. There are 22-25 such genes for voltage-dependent calcium channels, all of which may be alternatively spliced. These may assemble in up to 840 stoichiometric complexes, excompassing as many as 20 variable sites each. Physiological channels may arise from as many as $\sim 9 \times 10^8$ transcript combinations. This is an enormous space of possibilities, just for calcium channels, that can be exploited in the refinement of neuronal networks.

We may expect splicing correlations to cross gene boundaries in such cases, though direct physical contact may not even be necessary in general (34). Splicing linkage analyses in high-throughput transcriptomics may provide a valuable compliment to direct peptide interaction studies, such as yeast two-hybrid, to reveal functional interactions that do not require strong physical contacts. It is interesting, in this light, that the notion of a reconfigurable 'interactome' (35) extends to variable domains within the protein interior.

The unpredictable consequences of functional interactions are amplified through ambiguity in the determinants of alternative splicing, which are not fully specified in the gene sequence. Complex mammalian genes support an intrinsic uncertainty in the structure of the expressed protein which is reduced through 'paragenetic' information residing outside the gene, within networks of *trans* regulatory factors, for example (36, 37, 38). Thus, very rare splice configurations may be produced under most conditions. Though any particular one may have a low probability, there is always a chance that a new form may arise, producing a protein that functions, albeit in an unusual way. A low level of such 'noise' may in fact be useful to a cell in a complex, unpredictable local environment. This certainly describes the mammalian brain, where humans have far outpaced other primates in the evolutionary divergence of phenotype. In keeping with this, the brain expresses a disproportionate diversity of alternative splicing, compared to other tissues (39).

In the context of expanding complexity in alternative splicing, interactions between variable domains therefore present a challenge to the regulatory processes that select them. A set of k modular domains may be configured through k sequential yes / no choices. Functional interactions, however, force a single, nondecomposable, selection from 2^k alternatives. The regulatory complexity thus increases exponentially with the number of sites if they interact, but only linearly if they do not. This is the cost of diversifying the proteome through combinatorial splicing. It returns a significant payoff, however, because the exponential expansion of the regulatory load is compensated by an expansion of phenotypic potential on the same scale. We have noted (8) that this regulatory burden could be escaped if some mechanism were available for somatic selection during development, based on feedback from the expressed transcriptome.

As full-length cDNA datasets become available the methods presented here will assist in defining the interaction landscape, revealing domain configurations that are selected in concert and providing insights into how domains within proteins interact functionally. Additionally, however, they should adapt well to

studies of clustering in the transcription of genes (and parts of genes). A compelling application is to the class of RNA transcripts that do not encode proteins (40). Though largely of unknown function, these ncRNAs comprise a large proportion of the transcriptome, representing roughly 50% of transcriptional units and covering 30 times more of the genome than the protein-coding mRNA, and they are elaborately processed (capped, polyadenylated and spliced—constitutively, alternatively and *trans*-spliced) (20). They likely represent an important component of the intricate web of RNA factors involved in the regulation of gene expression (41), including regulated alternative splicing.

ACKNOWLEDGMENTS

This work was supported by NIH award HL52307 and a Holden Targeted Research Award from the Epilepsy Foundation to W.S.A.

REFERENCES

1. Stamm, S., Riethoven, J. J., Le Texier, V., Gopalakrishnan, C., Kumanduri, V., Tang, Y., Barbosa-Morais, N. L. and Thanaraj, T. A. (2006) ASD: a bioinformatics resource on alternative splicing. *Nucleic Acids Res,* **34**(Database issue), 46–55.
2. Clark, T. A., Sugnet, C. W. and Ares, M. (2002) Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science,* **296**, 907–910.
3. Castle, J., Garrett-Engele, P., Armour, C. D., Duenwald, S. J., Loerch, P. M., Meyer, M. R., Schadt, E. E., Stoughton, R., Parrish, M. L., Shoemaker, D. D. and Johnson, J. M. (2003) Optimization of oligonucleotide arrays and RNA amplification protocols for analysis of transcript structure and alternative splicing. *Genome Biol,* **4**, R66.
4. Le, K., Mitsouras, K., Roy, M., Wang, Q., Xu, Q., Nelson, S. F. and Lee, C. (2004) Detecting tissue-specific regulation of alternative splicing as a qualitative change in microarray data. *Nucleic Acids Res,* **32**, e180.
5. Kapranov, P., Drenkow, J., Cheng, J., Long, J., Helt, G., Dike, S. and Gingeras, T. R. (2005) Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Res,* **15**, 987–997.
6. Aebersold, R. and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature,* **422**, 198–207.
7. Roberts, G. C. and Smith, C. W. J. (2002) Alternative splicing: combinatorial output from the genome. *Curr Opin Chem Biol,* **6**, 375–383.
8. Emerick, M. C., Stein, R., Kunze, R., McNulty, M. M., Regan, M. R., Hanck, D. A. and Agnew, W. S. (May 2, 2006) Profiling the array of $Ca_v3.1$ variants from the human T-type calcium channel gene CACNA1G: Alternative structures, developmental expression, and biophysical variations. *Proteins,* 10.1002/prot.20877.
9. Mattick, J. S. and Makunin, I. V. (2005) Small regulatory RNAs in mammals. *Hum Mol Genet,* **14** (Spec No 1), 121–132.
10. RIKEN Genome Exploration Research Group, Genome Science Group [Genome Network Project Core Group] and the FANTOM Consortium (2005) Antisense transcription in the mammalian transcriptome. *Science,* **309**, 1564–1566.
11. Yudate, H. T., Suwa, M., Irie, R., Matsui, H., Nishikawa, T., Nakamura, Y., Yamaguchi, D., Peng, Z. Z., Yamamoto, T., Nagai, K. *et al.* (2001) Variation in sequence and organization of splicing regulatory elements in vertebrate genes. *Nucleic Acids Res,* **29**, 185–188.
12. Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H. *et al.* (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature,* **420**, 563–573.
13. Stapleton, M., Carlson, J., Brokstein, P., Yu, C., Champe, M., George, R., Guarin, H., Kronmiller, B., Pacleb, J., Park, S. *et al.* (2002) A D*rosophila* full-length cDNA resource. *Genome Biol,* **3**, research0080.10080.8.
14. Strausberg, R. L., Feingold, E. A., Grouse, L. H., Derge, J. G., Klausner, R. D., Collins, F. S., Wagner, L., Shenmen, C. M., Schuler, G. D., Altschul, S. F. *et al.* (2002) Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc Natl Acad Sci U S A,* **99**, 16899–16903.
15. Kikuchi, S., Satoh, K., Nagata, T., Kawagashira, N., Doi, K., Kishimoto, N., Yazaki, J., Ishikawa, M., Yamada, H., Ooka, H., Hotta, I., Kojima, K. *et al.* (2003) Collection, mapping, and annotation of over 28,000 cDNA clones from *japonica* rice. *Science,* **301**, 376–379.
16. Sakurai, T., Satou, M., Akiyama, K., Iida, K., Seki, M., Kuromori, T., Ito, T., Konagaya, A., Toyoda, T. and Shinozaki, K. (2005) RARGE: a large-scale database of RIKEN Arabidopsis resources ranging from transcriptome to phenome. *Nucleic Acids Res,* **33**(Database issue), 647–650.
17. Neves, G., Zucker, J., Daly, M. and Chess, A. (2004) Stochastic yet biased expression of multiple dscam splice variants by individual cells. *Nat Genet,* **36**, 240–246.
18. Regan, M. R., Emerick, M. C. and Agnew, W. S. (2000) Full-length single-gene cDNA libraries: Applications in splice variant analysis. *Anal Biochem,* **286**, 265–276.
19. Regan, M. R., Lin, D. D., Emerick, M. C. and Agnew, W. S. (2005) The effect of higher order RNA processes on changing patterns of protein domain selection: a developmentally regulated transcriptome of type 1 inositol 1,4,5-trisphosphate receptors. *Proteins,* **59**, 312–331.
20. FANTOM Consortium, RIKEN Genome Exploration Research Group, and the Genome Science Group [Genome Network Project Core Group] (2005) The transcriptional landscape of the mammalian genome. *Science,* **309**, 1559–1563.
21. Zhu, J., Shendure, J., Mitra, R. D. and Church, G. M. (2003) Single molecule profiling of alternative pre-mRNA splicing. *Science,* **301**, 836–838.
22. Raymond, C. K., Castle, J., Garrett-Engele, P., Armour, C. D., Kan, Z., Tsinoremas, N. and Johnson, J. M. (2004) Expression of alternatively spliced sodium channel alpha-subunit genes. unique splicing patterns are observed in dorsal root ganglia. *J Biol Chem,* **279**, 46234–46241.
23. R Development Core Team (2005) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2005. ISBN 3-900051-07-0. http://www.R-project.org
24. Cover, T. M. and Thomas, J. A. (2005) Elements of information theory, Wiley-Interscience, Hoboken, N.J., 2^{nd} edition.
25. Bishop, Y. M. M., Fienberg, S. F. and Holland, P. W. (1977) Discrete multivariate analysis: Theory and practice, MIT Press, Cambridge, MA.
26. Gelman, A., Meng, X. L. and Stern, H. (1996) Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica,* **6**, 733-760.
27. Gelman, A. (2004) Bayesian Data Analysis, Chapman and Hall/CRC, Boca Raton, 2^{nd} edition.
28. Kuznetsov, V. A. (2001) Distribution associated with stochastic processes of gene expression in a single eukaryotic cell. *EURASIP J App Sig Proc,* **4**, 285–296.
29. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature,* **431**, 931–945.
30. *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode c. elegans: a platform for investigating biology. *Science,* **282**, 2012–2018.
31. Misra, S., Crosby, M. A., Mungall, C. J., Matthews, B. B., Campbell, K. S., Hradecky, P., Huang, Y., Kaminker, J. S., Millburn, G. H., Prochnik, S. E. *et al.* (2002) Annotation of the drosophila melanogaster euchromatic genome: a systematic review. *Genome Biol,* **3**, research0083.1-0083.22.
32. Claverie, J. M. (2001) Gene number. what if there are only 30,000 human genes?. *Science,* **291**, 1255–1257.
33. Johnson, J. M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P. M., Armour, C. D., Santos, R., Schadt, E. E., Stoughton, R. and Shoemaker, D. D. (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science,* **302**, 2141–2144.
34. Völker, J. and Breslauer, K. J. (2005) Communication between noncontacting macromolecules. *Annu Rev Biophys Biomol Struct,* **34**, 21–42.
35. Resch, A., Xing, Y., Modrek, B., Gorlick, M., Riley, R. and Lee, C. (2004) Assessing the impact of alternative splicing on domain interactions in the human proteome. *J Proteome Res,* **3**, 76–83.
36. Ladd, A. N. and Cooper, T. A. (2002) Finding signals that regulate alternative splicing in the post-genomic era. *Genome Biol,* **3**, 1–16.
37. Faustino, N. A. and Cooper, T. A. (2003) Pre-mRNA splicing and human disease. *Genes Dev,* **17**, 419–437.
38. Yeo, G., Hoon, S., Venkatesh, B. and Burge, C. B. (2004) Variation in sequence and organization of splicing regulatory elements in vertebrate genes. *Proc Natl Acad Sci U S A,* **101**, 15700–15705.
39. Grabowski, P. J. and Black, D. L. (2001) Alternative RNA splicing in the nervous system. *Prog Neurobiol,* **65**, 289–308.
40. Claverie, J. M. (2005) Fewer genes, more noncoding RNA. *Science,* **309**, 1529–1530.
41. Mattick, J. S. (2004) RNA regulation: a new genetics?. *Nat Rev Genet,* **5**, 316–323.

Suppplement to Multivariate Analysis and Visualization of Splicing Correlations in Single-Gene Transcriptomes

Mark C. Emerick, Giovanni Parmigiani, William S. Agnew

June 9, 2006

Contents

Table S1

	Counts		Configuration								
Species	Fetal	Adult	**Δ25A**	**14**	**25C**	**26**	**30B**	**31A**	**34**	**35**	**38B**
17	1		○	○	○	○	●	○	○	○	●
25	9	36	○	○	○	○	●	●	○	○	●
33	1		○	○	○	●	○	○	○	○	●
49	1		○	○	○	●	●	○	○	○	●
57	64	11	○	○	○	●	●	●	○	○	●
88		2	○	○	●	○	●	●	○	○	○
89	4	250	○	○	●	○	●	●	○	○	●
93	1		○	○	●	○	●	●	●	○	●
137			○	●	○	○	○	●	○	○	●
145	1		○	●	○	○	●	○	○	○	●
153	33	27	○	●	○	○	●	●	○	○	●
157	5		○	●	○	○	●	●	●	○	●
169	3		○	●	○	●	○	●	○	○	●
177	1		○	●	○	●	●	○	○	○	●
184	10		○	●	○	●	●	●	○	○	○
185	111	66	○	●	○	●	●	●	○	○	●
189	5	15	○	●	○	●	●	●	●	○	●
209		10	○	●	●	○	●	○	○	○	●
216		11	○	●	●	○	●	●	○	○	○
217	8	289	○	●	●	○	●	●	○	○	●
221		4	○	●	●	○	●	●	●	○	●
223		4	○	●	●	○	●	●	●	●	●
249			○	●	●	●	●	●	○	○	●
313	8		●	○	○	●	●	●	○	○	●
409	8		●	●	○	○	●	●	○	○	●
441	13		●	●	○	●	●	●	○	○	●
473		22	●	●	●	○	●	●	○	○	●
477		11	●	●	●	○	●	●	●	○	●
Total:	287	758									

1 Datasets

We analyze a biological dataset described in Emerick *et al*. (2006). Human $Ca_v3.1$ structural variants in full-length single-gene libraries from adult and fetal whole brain. The species designation is the decimal equivalent of the bit-string representation of splice configurations on the transcript, in the order listed.

2 Illustration of modular and interactive domains

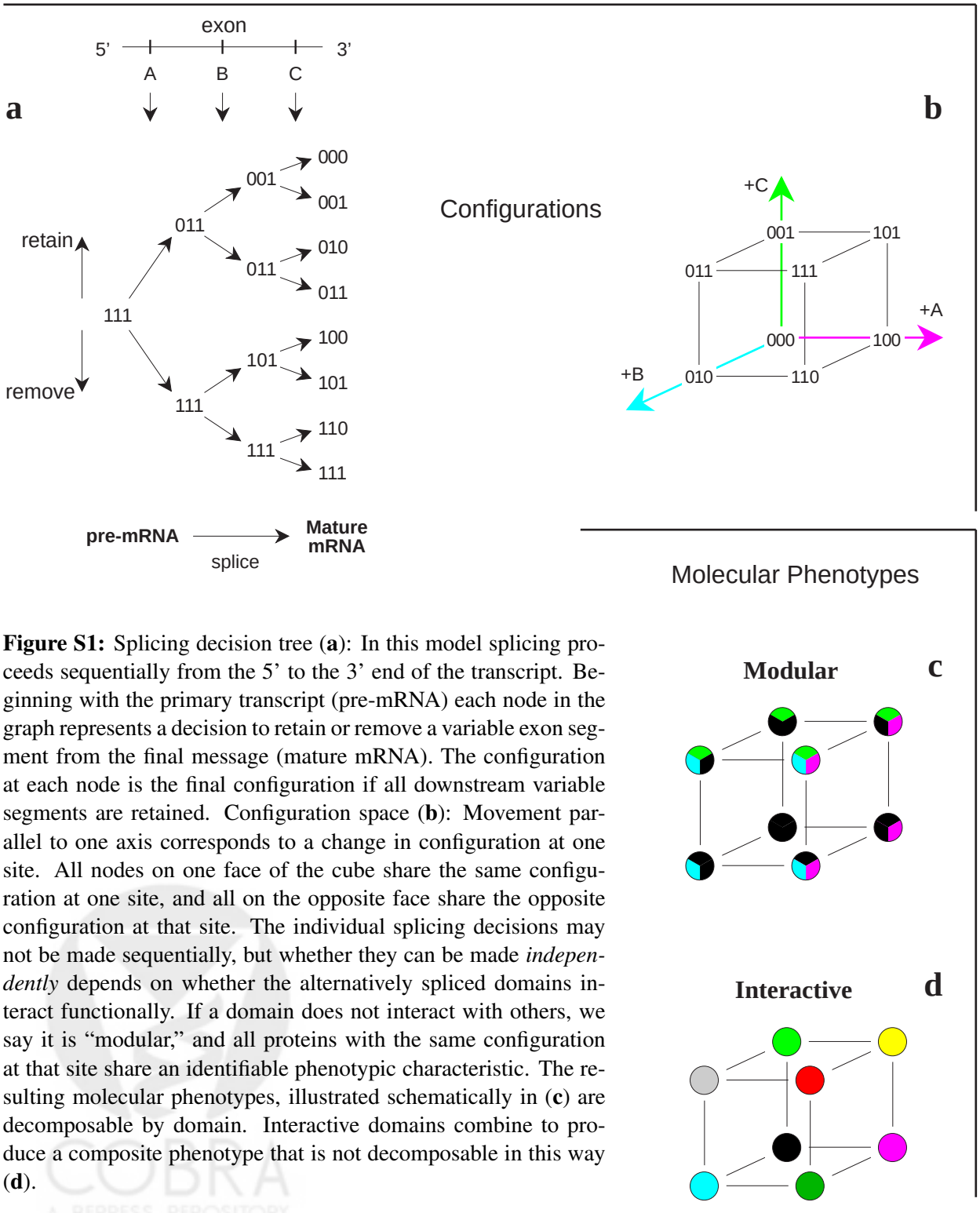

Figure S1: Splicing decision tree (**a**): In this model splicing proceeds sequentially from the 5' to the 3' end of the transcript. Beginning with the primary transcript (pre-mRNA) each node in the graph represents a decision to retain or remove a variable exon segment from the final message (mature mRNA). The configuration at each node is the final configuration if all downstream variable segments are retained. Configuration space (**b**): Movement parallel to one axis corresponds to a change in configuration at one site. All nodes on one face of the cube share the same configuration at one site, and all on the opposite face share the opposite configuration at that site. The individual splicing decisions may not be made sequentially, but whether they can be made *independently* depends on whether the alternatively spliced domains interact functionally. If a domain does not interact with others, we say it is "modular," and all proteins with the same configuration at that site share an identifiable phenotypic characteristic. The resulting molecular phenotypes, illustrated schematically in (**c**) are decomposable by domain. Interactive domains combine to produce a composite phenotype that is not decomposable in this way (**d**).

3 Empirical Bayes methodolgy

We estimate the frequency of splice variants in the parent distribution (the original tissue source) with the 'pseudo-Bayes' estimator, $\mathbf{p}^*$, of Bishop *et al*. (1975). This is a linear shrinkage estimator that takes a weighted average of the prior mean and the maximum likelihood estimate, shrinking the observed frequencies toward the prior, optimizing the weight to minimize the Euclidian distance between the estimate and the parent distribution (Carlin, 1996):

$$\mathbf{p}^* = (1-w)\mathbf{p} + w\boldsymbol{\lambda} \tag{S1a}$$

$$w = K/(N+K) \tag{S1b}$$

$$K = (N^2 - \sum_v x_v^2)/\sum_v (x_v - N\lambda_v)^2 \tag{S1c}$$

where N is the total number of transcripts in the cDNA library, x_v is the observed number of counts of each splice variant v, and λ_v is its prior probability. The likelihood, $\mathbf{p}$, is multinomial, $p(\mathbf{x}|\boldsymbol{\lambda}) \sim \prod \lambda_v^{x_v}$, [expression (4) in the paper], with integer parameters x_v. Our prior is Dirichlet: $p(\boldsymbol{\lambda}|K) \sim \prod \lambda_v^{\beta_v - 1}$, generally with non-integer parameters $\beta_v = K\lambda_v$. The posterior is therefore Dirichlet as well, with parameters $x_v + \beta_v$. We discuss extensively the choice of prior distribution.

3.1 Choice of prior distribution

Summary

If our goal were to model the observed population, the bootstrap would suffice. A realistic estimator should admit a small, nonzero probability for the unobserved classes, however. The two most common ways to do this are (i) a uniform prior, free of assumptions about the true distribution, and (ii) a prior obtained by adding a small constant to all observed frequencies and renormalizing. Both of these methods err in assigning equal probability to all unobserved classes. With a large number of such forms this becomes a large error, such that the observed classes are well modeled but the complete distribution is not. The unintended assumption, in effect, is that a high degree of splicing correlations conspire to lock out unobserved forms. A more likely scenario is that overall splicing is less restrictive and many of these forms would show up in larger samples.

We observe that the distribution of splice forms in both fetal and adult brain approximately follows the 'independent-marginals' expectation. Without justification for believing that unobserved forms would follow a different pattern, we may chose this distribution as our prior. This prior may be viewed as a smoothed modification of prior (ii), above, in which the probabilities of the unobserved classes taper off in a more realistic fashion, toward zero for forms that combine multiple low-frequency splice configurations. Because fetal splicing is more nearly independent than adult the fetal estimator gives more weight to its independence prior, and less weight to the actual data, than does the adult. A minor modification removes this effect: we reduce the amount of information about the tissue-specific marginal frequencies included in the prior by averaging the fetal and adult independence expectations into a single 'averaged-marginals' prior for both tissues. While omitting what we know about the developmental reversal at **25C** and **26**, this approach incorporates the information that splicing frequencies at the other 7 sites are preserved during development. The extent of shrinkage is approximately the same in the fetal and adult cases. The goodness of fit is similar for both and is not improved by arbitrarily reducing the shrinkage weight to favor the observed population.

Our criteria for assessing the prior distribution are that unobserved variants may have a small nonzero probability and that all observed frequencies should be among the more probable frequencies in the posterior distribution. A uniform prior is a simple option that allows modeling of the unobserved classes without imposing *a priori* differences between the unknown frequencies. This prior reproduces the observed behavior well (Figures S2A and S10, estimator 13) because the shrinkage gives little weight to the prior for either the fetal or adult tissue; w in equation (S1) is

Table S2: Loglinear model parameters for three priors.

	K		w	
	Fetal	Adult	Fetal	Adult
uniform	3.58	2.76	0.012	0.0036
tissue-specific marginals	146.76	32.80	0.338	0.0415
Averaged marginals	6.42	0.22	0.022	0.0003

0.004 for the adult and 0.012 for the fetal population (table S2). The posterior is thus the distribution of observed frequencies modified only slightly to give a nonzero probability to the unobserved classes. Shrinkage with this prior is preferable to the bootstrap in that it admits the possibility of sampling unobserved variants, but it does so in a rudimentary way, dismissing what we know about splicing at the separate sites. That is, some splice combinations are much more likely to occur than others, and it is just as unrealistic to suppose that all unobserved classes are equally likely (uniform prior) as it is to suppose that they are all impossible (bootstrap). In fact, simple independent assortment of multiple splice configurations, several with low probability, would produce many splice variants of extremely low probability while others would lie just beyond the detection limit at the current sample size.

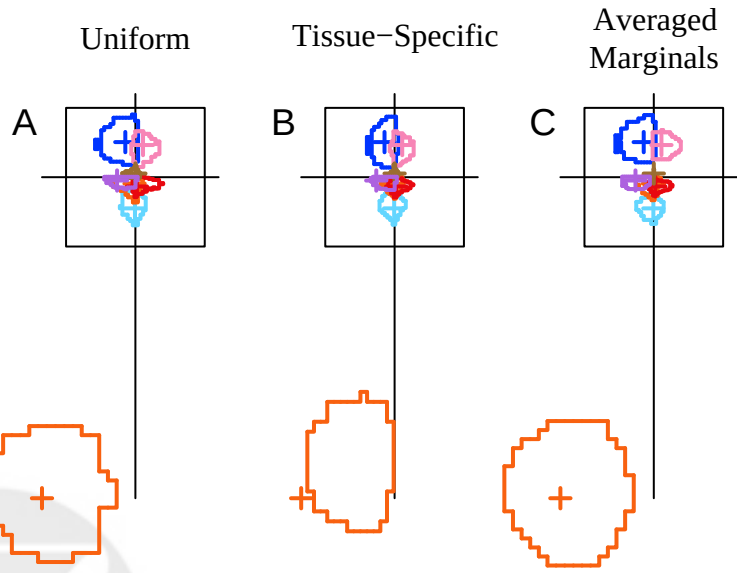

Figure S2: Mutual Information clock plots for three priors. Points within the box show pair-wise linkage of various sites to site **14**. The orange plots differ in depicting linkage between sites **25C** and **26**. Error rings give 95% confidence limits

Figure S3A and B illustrate the latter point: if splicing at the separate sites were completely independent, a library of 5,000 cDNAs from either adult or fetal brain would yield only about 14% of the 512 possible splice variants, with diminishing returns for larger samples. The yield is about the same in both tissues, reflecting a similar distribution of marginal frequencies in both. When splicing is not independent fewer forms are obtained, and while the fetal population appears close to independence, the adult is not. Figure S3 shows the dependence of species richness on sample size for several other estimators. The bootstrap (panels C and D) models only the observed variants.

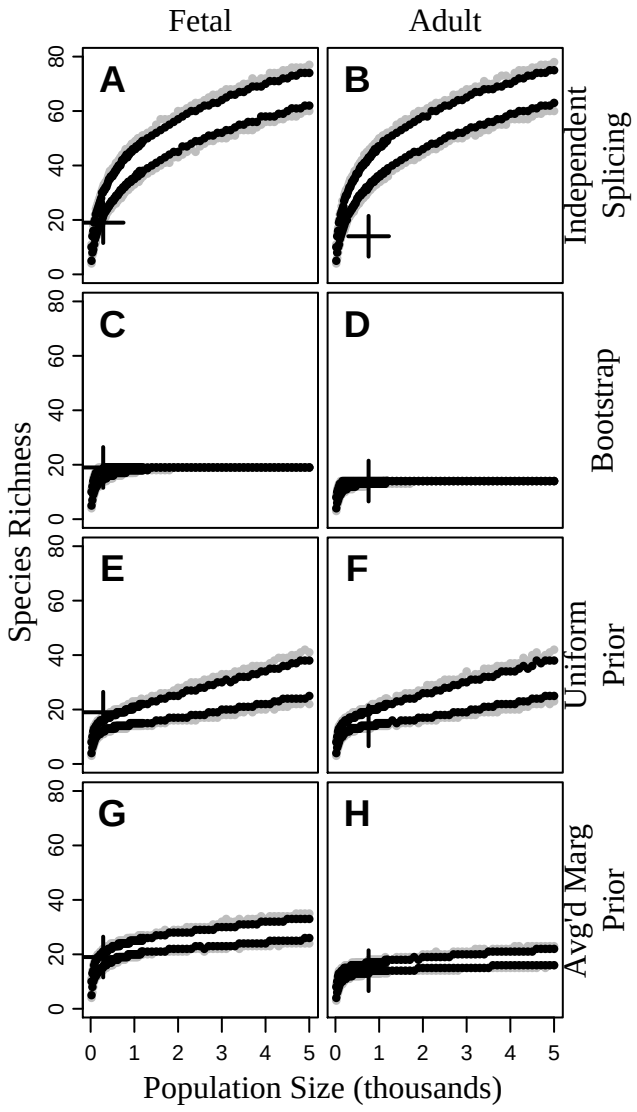

Figure S3: Species richness in observed and Monte Carlo-simulated cDNA populations. Monte Carlo populations of various sizes were sampled from four estimator distributions for each tissue. Estimators A-D were derived directly from the data: **A** and **B** were sampled from the expected distribution for independent splicing at the observed marginal frequencies of the individual splice sites (ϕ_v, equation 1 in the main paper); **C** and **D** were resampled from the observed frequencies of the splice variants. Estimators E-H were empirical-Bayes posteriors obtained with a uniform prior (**E**, **F**) or the "averaged-marginals" prior (**G**, **H**) discussed below. 1000 populations of each size were sampled. Black symbols mark the 5 and 95% quantiles, and gray mark the 1 and 99% quantiles. Crosses plot the observed values. This figure may be interpreted in two ways: (i) as a test of model adequacy; for example, independent splicing is insufficient to account for the observed adult population, and bootstrap models only the observed forms, and (ii) given an acceptable model it shows how many transcripts must be sampled to ensure that a desired number of forms represented. The bottom panels show that larger samples yield additional forms in both tissues, as generally expected, but progressively fewer in the adult than in the fetal.

Table S3: Marginal frequencies and entropies for splice configurations in table S1

C_j	**Δ25A**	**14**	**25C**	**26**	**30B**	**31A**	**34**	**35**	**38B**
$p_j(C_j)$									
Fetal	0.899	0.690	0.045	0.756	0.986	0.986	0.038	0	0.965
Adult	0.956	0.606	0.796	0.121	1.000	0.987	0.045	0.005	0.983
$H_j(C_j)$									
Fetal	0.327	0.619	0.184	0.556	0.074	0.074	0.162	0	0.152
Adult	0.180	0.671	0.506	0.369	0	0.069	0.184	0.031	0.086

The empirical Bayes estimate with uniform prior (E and F) captures the observed species richness only marginally in either population and gives an odd linear growth with sample size after an initial steep rising phase. The empirical Bayes estimate with averaged-marginals prior (G and H), brackets the observed values and shows a reasonable growth characteristic, with progressively diminishing yield, relatively flatter in the adult than the fetal, consistent with greater splicing linkage in the former. We sill discuss this estiamtor in greater detail. The mutual information analyses show that the overall observed splicing linkage is not high in either tissue, except for sites **25C** and **26** (Figure 1B in the main text). To a first approximation, the distribution of splice variants in both tissues follows the expected stochastic distribution for independent splicing of the separate sites (Figure S4). We may include this information in the prior distribution. Independent support for this comes also from Latour *et al*. (2004) and Monteil *et al*. (2000) who obtained a population of 68 long-range adult brain $Ca_v3.1$ cDNAs whose multinomial log-likelihood is consistent with our data, based on independent splicing at our observed marginal frequencies.

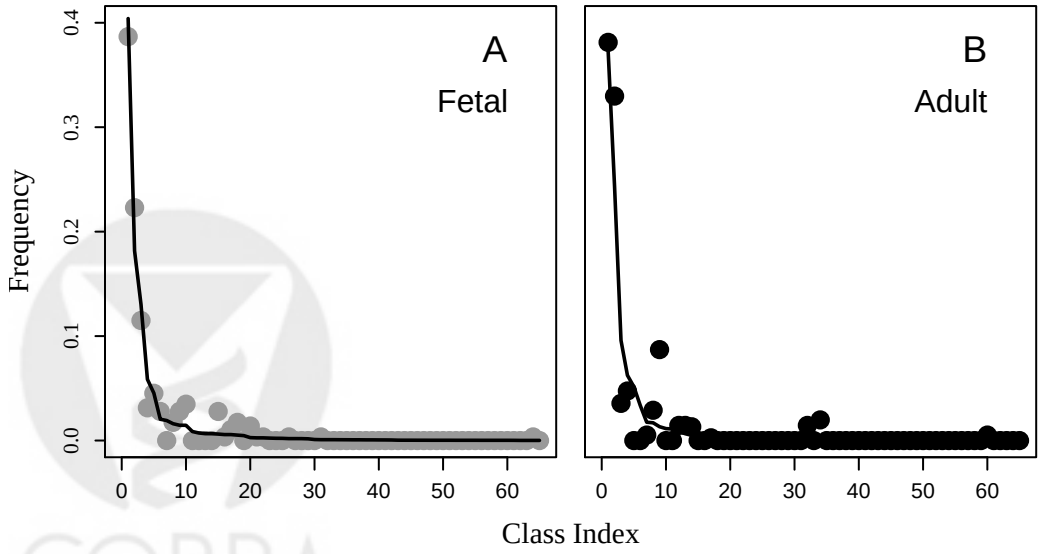

Figure S4: Splice variants ranked by expected frequency. The first 65 fetal splice variant classes are plotted in order of expected frequency given independent, stochastic splicing (**black curve**); a given class index does not generally refer to the same splice variant in the fetal and adult plots. Plot symbols indicate observed frequencies. All classes with more than one observed instance are plotted.

Several sites exhibit such a strong splicing bias (Figure S5 and table S3) that marginal frequencies cannot be estimated reliably with our sample size. Since these sites tend to yield the same configuration in both tissues, however, the tissue-averaged marginal frequencies were used for these sites (**30B**, **31A**, **34**, **35**, and **38B**), while tissue-specific values were used for the remaining sites (**14**, **25A**, **25B**, **26**). The stochastic expectation for independent splicing at these frequencies gives the tissue-specific 'independent-marginals' prior. As indicated in table S2, the fetal population conforms considerably to this prior, resulting in a high degree of shrinkage toward independence ($w = 0.34$). The adult frequencies are shrunk to a greater extent in this case ($w = 0.041$) than with the uniform prior, but markedly less than the fetal.

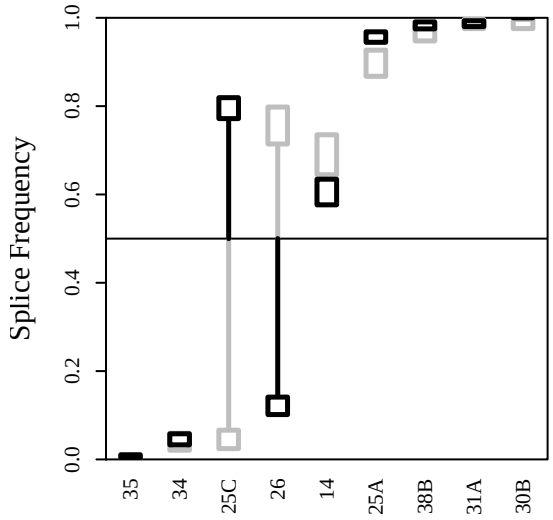

Figure S5: Marginal splicing probabilities. Boxes delimit the 95% confidence intervals for a binomial sample at the observed frequency in the fetal (gray) and adult (black) cDNA libraries. Vertical lines accentuate opposite-going developmental splicing shifts at sites **25C** and **26**.

We note that the high weight given to the fetal 'independent-marginals' prior, in contrast to the adult, reflects an important distinction between splicing of this gene in adult and fetal brain: fetal splicing linkage is relaxed compared to the adult, and this is of physiological importance. We emphasize that the resulting posterior distributions may very well reflect the true physiological parameters, but the fetal data are shrunk to such a degree that the observed frequencies no longer appear to be well represented in the posterior distribution (Figure S6A, blue symbols), and the pair-wise splicing linkage between sites **25C** and **26** is reduced in the posterior (Figure S2B). Unobserved variants are apparently overcompensated as well in both the fetal and adult populations (*c.f.* Figure S7, center column for each set of three priors).

To center our estimator better over the data, we should "back off" on the extent of shrinkage, and it is best to do so in a manner that is not preferential to one data set. If we simply average the observed fetal and adult marginal frequencies at each site, we obtain a single 'tissue-averaged independent-marginals' prior that is roughly equidistant from both populations, resulting in a similar, small extent of shrinkage for both tissues (table S2). This yields an accurate representation of the observed classes (Figure S6A and C, green crosses) and a faithful reproduction of the observed pair-wise splicing correlations (Figure S2C). Thus the posterior distribution is largely a reflection of the properties of the observed data in both tissues and is therefore a conservative estimate, in that we do not depend heavily on the closeness of either population to its independent-splicing

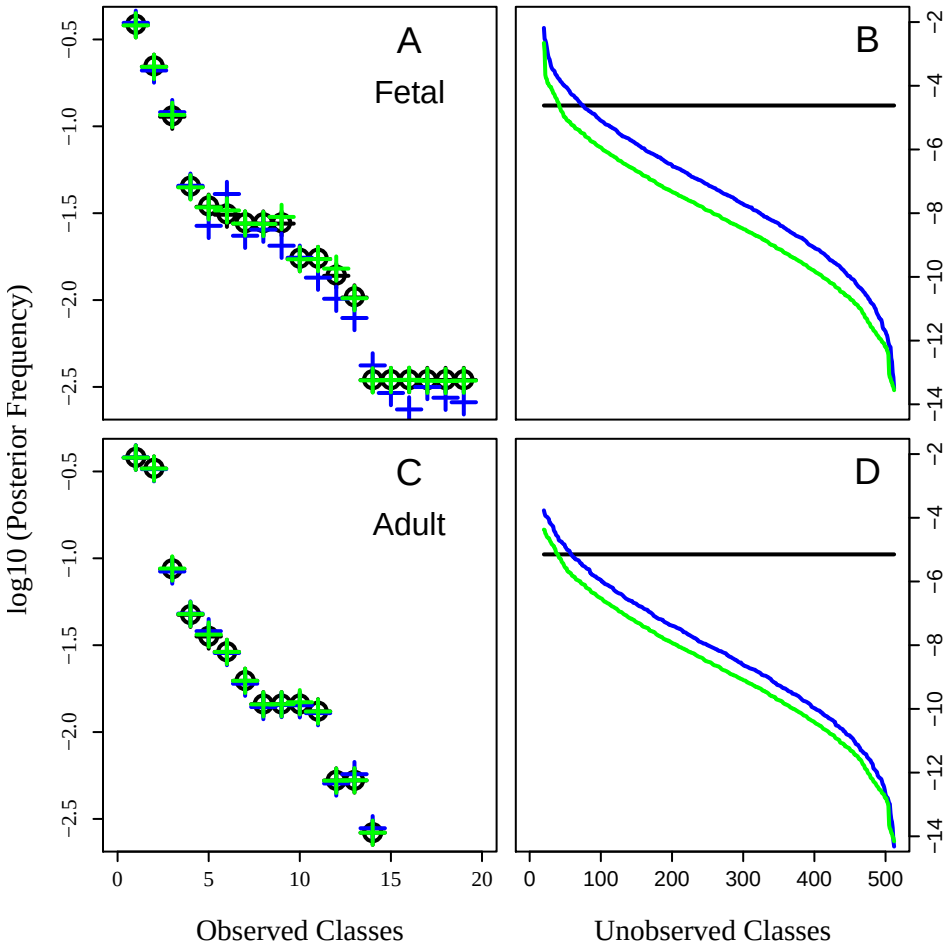

Figure S6: Empirical Bayes posterior distributions for three priors. Splice variant classes on the abscissa are listed in decreasing order of observed frequency (**A** and **C**) or expected frequency in the posterior distribution (**B** and **D**) for three priors: uniform (**black**), tissue-specific (**blue**), and averaged-marginals (**green**).

expectation. Unlike the uniform prior, however, both independent-marginals priors dampen contributions from highly unlikely splice configurations (compare either of the colored curves with the black curve in Figure S6B and D). The tissue-averaged prior improves on the tissue-specific prior by reducing the expected frequency of the first few unobserved classes (compare the two colored curves in Figure S6B and D).

This point is best illustrated in Figure S7. Each plot in this figure is a histogram of counts for a single splice variant in 1000 Monte Carlo populations sampled from the posterior estimate for one of three priors. The colored line demarks observed counts. The top row of plots in each set of three columns corresponds to the same observed splice variant (the one with the lowest counts in that tissue). The remaining plots in the column correspond to the first four unobserved classes, in order (downward) of decreasing posterior frequency. The tissue-specific prior over-estimates the first few unobserved classes, especially in the fetal samples. The averaged-marginals prior gives more expected behavior: the counts in the unobserved cells taper off more gradually in the fetal than the adult samples for this prior, consistent with less restrictive fetal splicing linkage. The uniform prior, by contrast, gives an unreasonably abrupt transition to very low counts in the unobserved

cells in both tissues, and all unobserved classes are in fact exchangeable.

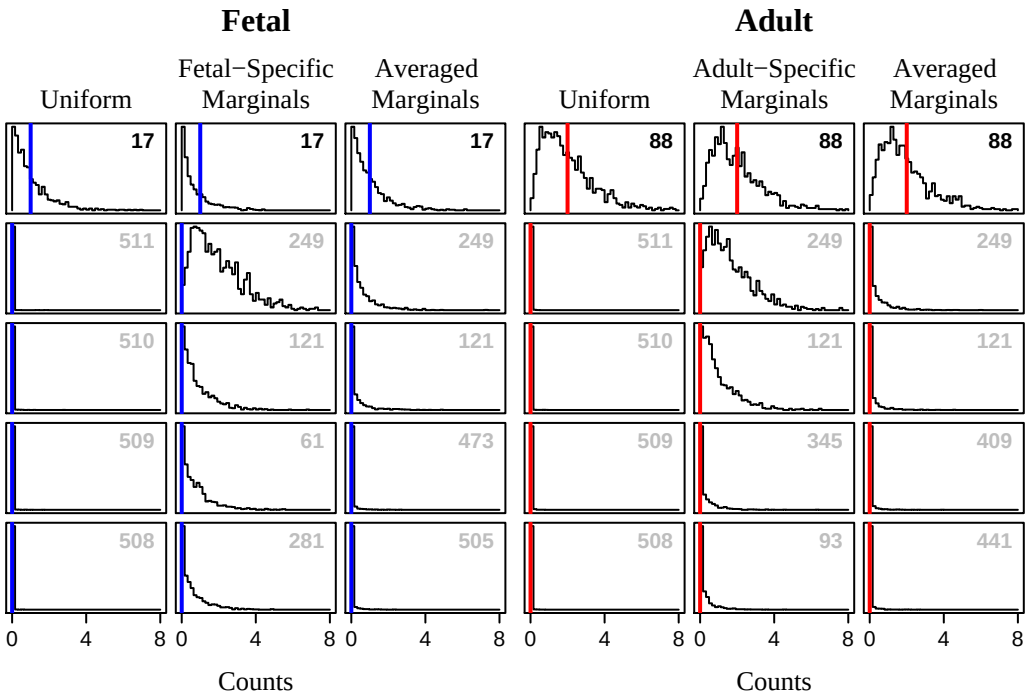

Figure S7: Low-frequency splice variants sampled from three estimators: details of the dependence on the prior. The top row depicts the least-frequent observed class. The remaining rows depict the four most frequent unobserved forms ranked in order (top to bottom) of decreasing frequency in the posterior. Fractional counts arise because the posterior distribution is continuous (Dirichlet).

It might be argued that by averaging out the marginal frequencies of **25C** and **26** we are neglecting the actual observation that these segments change inversely during development. This developmental switch has been noted not only by us, but by others as well (Monteil *et al.*, 2000; Latour *et al.*, 2004). We may account for this in the averaged-marginals prior in various ways. One possiblitiy is to assign a higher marginal frequency (0.75) to the 'major' configurations (**-25C and +26** in fetal and **+25C and -26**) and the complement one (0.25) to the minor configurations. This yields results that do not differ substantially from the tissue-specific independent-marginals prior. Alternatively, we may use the average observed 'major' and 'minor' frequencies in the same way. This produces results midway between that independent-marginals and averaged-marginals priors.

3.2 Assessing goodness of fit

An important criterion for selecting a data-based prior is that the experimental population should occur near the center of the posterior distribution. The distance of a sampled population $Y = y_1, y_2, \ldots, y_k$ from the experimental 'target' $X = x_1, x_2, \ldots, x_k$ is measured by the integrated squared error,

$$ISE = \sum_i (x_i - y_i)^2 \tag{S2}$$

Figure S8 plots the ISE for each of 1000 populations sampled from estimators derived from three priors. These priors are mostly indistinguishable by this criterion. A related measure is chi-square, in which each term of the sum (S2) is divided by the variance of y_i. This measure is inapplicable

for the 'independent-marginals' priors, because the denominator tends to zero for the extremely unlikely classes (Figure S9).

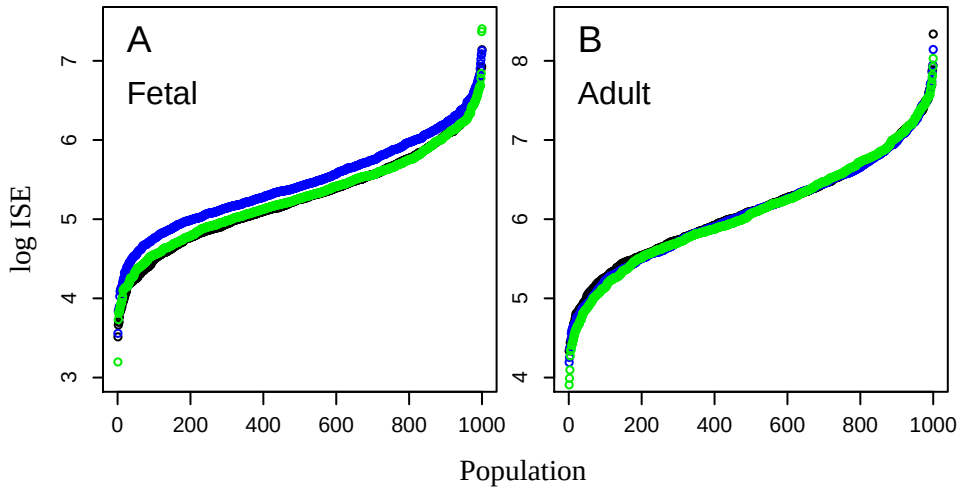

Figure S8: Overall goodness of fit measured by integrated squared error. Log(ISE) is plotted for each of 1000 populations sampled by Monte Carlo from the posterior distribution for each of three priors: uniform (**black**), tissue-specific (**blue**), and averaged-marginals (**green**).

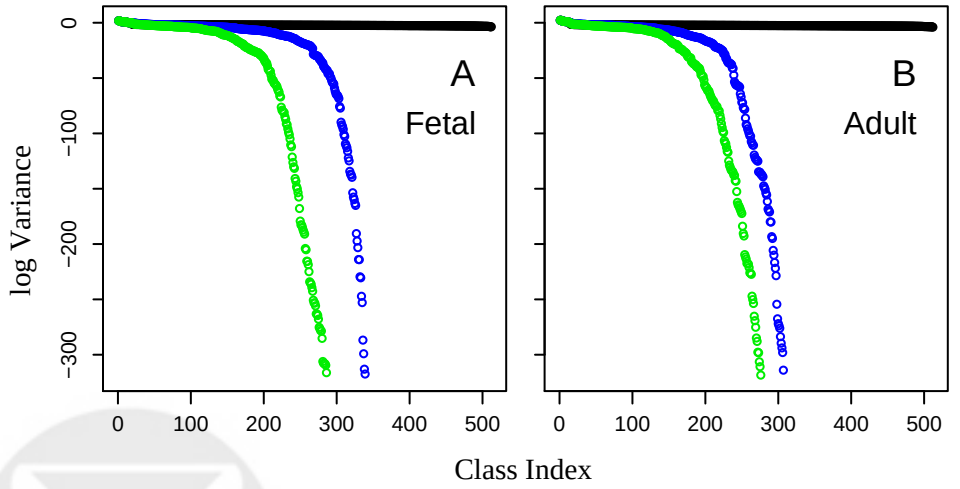

Figure S9: Dependence of the variance of different splice forms on the prior distribution. Log(Variance) is plotted in order of decreasing frequency in the posterior, for each of three priors: uniform (**black**), tissue-specific (**blue**), and averaged-marginals (**green**). Note that the variance of the uniform prior is not everywhere constant, but slightly elevated among the first few (observed) classes.

We are at present interested primarily in the statistical significance of differences in splicing-linkage structures between two populations, and not in making the most precise measurement of the properties of any particular population. Nonetheless, the bias-variance trade-off achieved by usual mean squared error measures of fit did not give satisfactory results in distinguishing priors that are clearly differentiable by direct inspection of clockplots, for example (Figure S2). The 'accuracy' index ($\mathcal{A}$), summed over all species in a population, is a very simple metric that works well for this purpose. Figure S10 plots the mean accuracies of all splice variants (squares) or just the observed

subset (circles) for populations sampled from 14 different posterior distributions, derived from six different priors by various degrees of shrinkage. The 'independent-marginals' priors are the most accurate, overall, while the uniform prior is satisfactory for the observed variants only. Posteriors 7-10 derive from four 'independent-marginals' prior variations by direct application of equations (S1). The accuracy drops off with posteriors 8-10, especially for the observed fetal variants. This is because a closer fit of the prior to the likelihood results in larger values of K in equation (S1c) (the 'Bishop' weight) and thus greater shrinkage toward independence, away from the largest observed values. The first six posteriors (except for #5) derive from independent-marginals priors, but with the prior weight reduced 'manually' to varying extents below the Bishop value. The observed accuracies plateau near the level set by the Bishop-weighted averaged-marginals prior (#7). The accuracies for the unobserved variants increase slightly because higher weight is given to the observed frequency of zero, resulting in a more compact distribution for these classes in the Monte Carlo populations. We emphasize that the apparent improvement in accuracy that results from decreasing the prior weight below the Bishop weight is mainly due to reducing the probabilities of unobserved variants. This is not justified in general, as equations (S1) optimize the weight for the choice of prior (Bishop, Fienberg et al. 1975). Furthermore, decreasing the prior weight in this way does not have a significant effect on the relative magnitudes of the log-linear coefficients for the interaction terms (Figure S12, discussed below).

The 'averaged-marginals' prior is a very simple, intuitive prior that, with straightforward Bishop shrinkage, (1) reproduces the observed pair-wise splicing correlations in the posterior distribution, (2) is well centered with respect to both the complete set of variants as well as just the observed subset, and (3) does not incorporate the plainly unreasonable assumption that all unobserved variants are equally likely, and gives a justifiable distribution of prior probabilities for those variants.

Figure S11 presents fetal/adult comparative spliceprints from Monte Carlo populations sampled from posteriors derived by Bishop shrinkage from several of the priors just discussed. The shape of the profiles is fairly stable across priors, except in two cases (panels E and F, the tissue-specific marginals prior and a close variant) where increased shrinkage toward independence produces a flattened fetal profile (compare Figure S2B).

Figure S12 is an assessment of the effects of shrinkage on the relative magnitudes of coefficents. It is a version of Figure 8 in the main text, in which mean coefficient magnitudes are *relative* to order-1 (cardinality-2) interactions. Grayscale histograms correspond to Bishop shrinkage according to equations (S1) (yielding $w = 0.022$ in fetal and 0.0003 in adult), while blue shading corresponds to linear shrinkage at a constant, lower extent ($w = 0.002$ in both fetal and adult). Relative magnitudes are stable to variation in the extent of shrinkage with the same prior, though reducing the shrinkage gives smaller variances.

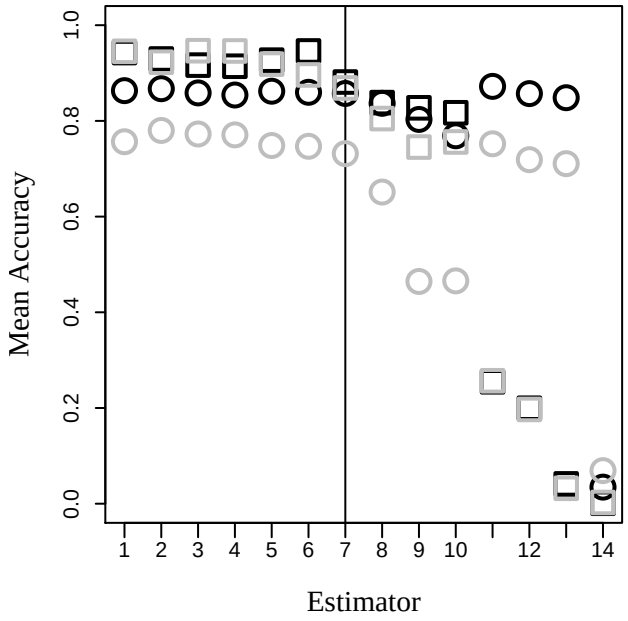

Figure S10: Estimator assessment. For each of 14 estimators, the mean accuracy ($\mathcal{A}$, *c.f. methods*) is plotted for all 512 classes (squares) or for just the observed subset (circles), using 1000 populations sampled by Monte Carlo from the posterior distribution, $\mathbf{p}^*$, derived by equation (S1a). Gray = fetal, Black = Adult. The three priors discussed most extensively are uniform (13), tissue-specific marginals (10), and averaged-marginals (7), all with weights determined by equation (S1c). Estimators 1 through 6 are obtained by 'manually' reducing the extent of shrinkage, over-weighting the data relative to to the Bishop weight, w in equation (S1c). This converges to the bootstrap at low weight, has little effect with a sound prior, and defeats the Bishop weighting rationale. The others are obtained as follows: Where indicated, sensitivity to shrinkage is tested by specifying K or w directly, rather than applying equation (S1c). Prior **1**: Tissue-specific prior with $w = 1/N$; **2-9**: Averaged-marginals prior with (2) $w = 1/758$, (3) $w = 1/500$, (4) $w = 1/287$, (5) a single "minor" frequency—equal to the average of the observed frequencies of **25C** in the fetal population and **26** in the adult— assigned to those two configurations, the corresonding average "major" frequency assigned to fetal **26** and adult **25C**, and K set to $1/25$ of the value determined by applying equation (S1c) (compare prior 9), (7) (Averaged marginals for all sites), (8) the minor and major frequencies (defined in prior 6) set to 0.75 and 0.25, respectively, (9) Same as prior 6, but with K determined by equation (S1c); **10**: tissue-specific marginals; **11**: Uniform, $K = 1$; **12**: Random—512 draws from a uniform distribution on $[0,1]$; **13**: Uniform; **14**: Unique random prior for each Monte Carlo population. Priors 6 and 8 incorporate the **25C/26** developmental switch into the averaged-marginals prior, progressively de-emphasizing the specific observed frequencies at those sites.

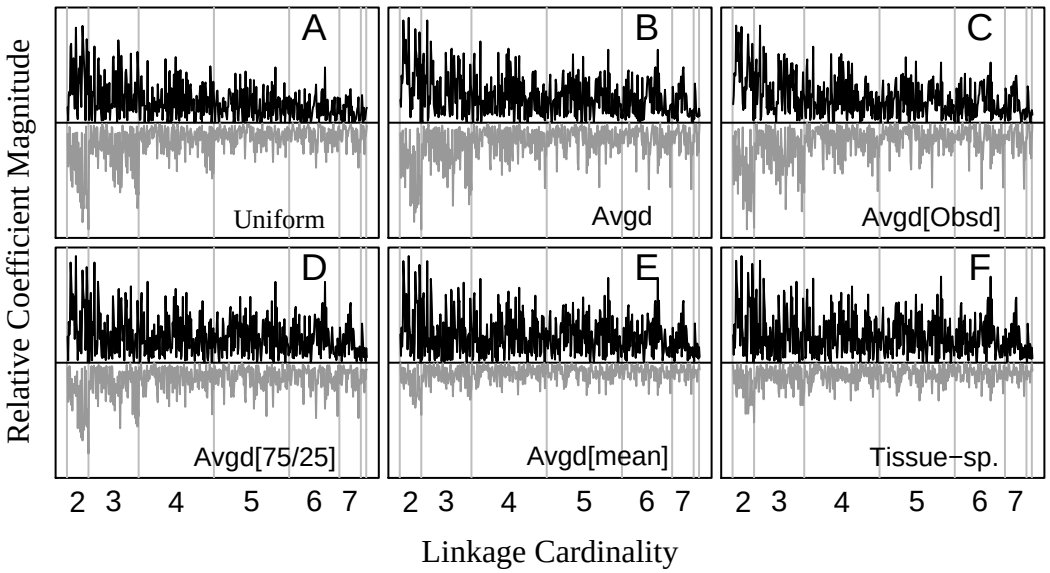

Figure S11: Comparative spliceprints for various fetal and adult estimators. Adult values (black) are plotted above and fetal values (gray) below the midline in each plot. Each trace plots the mean of the coefficients derived from loglinear model fits to Empirical Bayes estimates from 1000 MC populations. Priors are numbers 13 (**A**), 7 (**B**), 5 (**C**), 8 (**D**), 6 (**E**), and 10 (**F**) of Figure S10. Plots B-E use variants of the averaged-marginals prior. Notes in brackets indicate how minor/major frequencies at **25C** and **26** were obtained (see figure S10 legend for priors 5, 6, and 8).

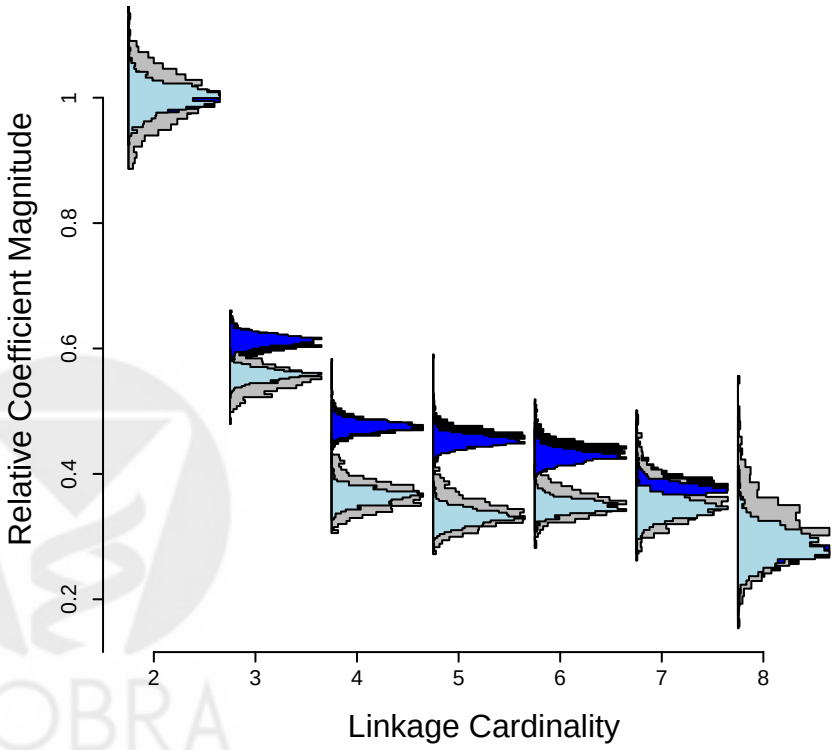

Figure S12: Distribution of cardinality-averaged loglinear coefficients. Relative coefficient magnitudes are stable to differing extents of shrinkage.

References

Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (1975) Discrete multivariate analysis: Theory and practice. Cambridge: MIT Press.

Carlin BP, Louis TA. (1996) Bayes and empirical Bayes methods for data analysis. 1st ed. London, New York: Chapman and Hall

Emerick, M.C., Stein, R., Kunze, R., McNulty, M., Regan, M.R., Hanck, D.L., and Agnew, W.S. (2005) Profiling the Array of $Ca_v3.1$ Variants from the Human T-type Calcium Channel Gene CACNA1G: alternative structures, developmental expression and biophysical variations. *Proteins: struct. funct. bioinform.*, bioinform., Published Online: 2 May 2006..

Monteil A, Chemin J, Bourinet E, Mennessier G, Lory P, Nargeot J (2000) Molecular and functional properties of the human alpha(1G) subunit that forms T-type calcium channels. J Biol Chem 275: 6090–6100.

Latour I, Louw DF, Beedle AM, Hamid J, Sutherland GR, and G. W. Zamponi GW (2004) Expression of T-type calcium channel splice variants in human glioma. Glia 48: 2, pp. 112–119.