Outlier detection for improved differential splicing quantification from RNA-Seq experiments with replicates

Scott Norton [1], Jorge Vaquero-Garcia [1,2] and Yoseph Barash [1,2,*]

January 29, 2017

[1]Department of Genetics, Perelman School of Medicine, and
[2]Department of Computer and Information Science, School of Engineering,
University of Pennsylvania, Philadelphia, PA, 19104, USA.
[*]Correspondence should be addressed to yosephb@upenn.edu

Abstract

A key component in many RNA-Seq based studies is the production of multiple replicates for varying experimental conditions. Such replicates allow to capture underlying biological variability and control for experimental ones. However, during data production researchers often lack clear definitions to what constitutes a "bad" replicate which should be discarded and if data from failed replicates is published downstream analysis by groups using this data can be hampered. Here we develop a probability model to weigh a given RNA-Seq experiment as a representative of an experimental condition when performing alternative splicing analysis. Using both synthetic and real life data we demonstrate that this model detects outlier samples which are consistently and significantly different compared to samples from the same condition. Using both synthetic and real life data we perform extensive evaluation of the algorithm in different scenarios involving perturbed samples, mislabeled samples, no-signal groups, and different levels of coverage, and show it compares favorably with current state of the art tools.

Availability: Program and code will be available at majiq.biociphers.org

1 Introduction

Alternative splicing, the process by which segments of pre-mRNA can be arranged in different ways to yield distinct mature transcripts, is a major contributor to transcriptome complexity. In humans, over 90% of multi-exon genes are alternatively spliced, and most of those exhibit splicing variations which are tissue- or condition-dependent [10]. This key role of alternative splicing (AS) in transcriptome complexity, combined with the fact that aberrant splicing is commonly associated with disease state [17], has led to great efforts to accurately map transcriptome complexity, identify splicing variations between different cellular conditions, across developmental stages, or between cohorts of patients and controls.

Detection of splicing variations and the mapping of transcriptome complexity has been greatly facilitated by the development of technologies to sequence transcripts, or RNA-Seq. Briefly, RNA from the cells of interest, typically poly-A selected or ribo-depleted, are sheared to a specific size range, amplified, and sequenced. In most technologies used today the resulting sequence reads are typically around 100bp with read number varying greatly, from around 20 to 200M reads. The shortness of the reads, their sparsity, and various experimental biases make inference about changes in RNA splicing a challenging computational problem [1]. Consequently, many studies include several replicates of the conditions they are studying. Replicates are a key component in helping researchers distinguish between the biological variability they are trying to detect and variability associated with experimental or technical factors. However, what constitutes a "good" replicate or an outlier experiment is not always clear. Intuitively, an outlier is a sample which exhibits dis-proportionately large deviations in exon inclusion levels compared to other biological replicates. An outlier could be the result of a failed experiment or of some previously unknown variability cause (*e.g.*, different tissue source). Remarkably, despite the obvious importance of the question of what constitutes an outlier, this question has been mostly ignored in the literature. Instead, researchers are left to define outliers based on some heuristics which may not be ideal or carry unconscious biases. Thus, an important contribution of this work is to suggest a model which researchers could use to assess whether a set of replicates are "well behaved" or might include outliers.

Obviously, the presence of outliers can have deleterious effects on algorithms that aim to detect differential splicing between groups of experiments. Broadly, algorithms that aim to quantify differential splicing from RNA-Seq can be divided into two classes. The first, which includes tools such as RSEM [7] and Cuffdiff [15], aims to quantify full gene isoforms, typically by assuming a known transcriptome and assigning the observed reads to the various gene isoforms in the given transcriptome database. The second class of algorithms, which includes rMATS [13] and DEXSeq [2], works at the exon level detecting differential inclusion of those. Some algorithms such as SUPPA [5] can be considered a hybrid as they collapse isoform abundance estimates from other algorithms (e.g. SailFish [12] or SALMON [11]) to compute relative exon inclusion levels. Previous works showed that for the task of differential splicing, quantification algorithms that work at the

exon level generally perform better since they solve a simpler task and are less sensitive to isoform definitions or RNA-Seq biases within samples or along full isoforms [9]. Thus, for the comparative analysis section of this paper we focus on the second class of algorithms, and specifically on those that support replicates.

Recently, we published MAJIQ, a method to detect, quantify and visualize differential splicing between groups of experiments. Besides the details of its statistical model, two key features distinguish MAJIQ from the algorithms mentioned above. First, MAJIQ does not quantify whole gene isoforms as the first class of algorithms described, or only previously defined AS "types" (*e.g.*, cassette exons), as the second class of algorithms. Instead, MAJIQ defines a more general concept of "local splicing variations", or LSVs. Briefly, LSVs are defined as splits in a gene splice graph where a reference exon is spliced together with other segments downstream (single source LSV) or upstream of it (single target LSV, see Figure 1a). Importantly, the formulation of LSVs enables MAJIQ to capture all previously defined types of AS (Figure 1b) but also many other variations which are more complex (Figure 1c). Specifically, previously defined AS event types are all binary, involving only 2 alternative junctions, while over 30% of human LSVs are complex, involving three or more alternative junctions. The second important distinguishing element of MAJIQ is that it allows users to supplement previous transcriptome annotation with reliably-detected *de-novo* junctions from RNA-Seq experiments (Figure 1d). We found that even when using a well-annotated species such as mouse, normal tissue data, and the full Ensembl transcriptome, MAJIQ detects 32% more differentially spliced LSVs which involve unannotated junctions. We validated many splicing events involving *de-novo* junctions and showed these are highly reproducible. However, MAJIQ was built to handle only "good" replicate data. Thus, the second contribution of this work is to suggest a generalization of MAJIQ which enables down weighting of suspected outliers. Finally, the third contribution of this work is in extensive comparative analysis of MAJIQ and other algorithms in terms of reproducibility of inferred differential splicing events, false positives when no biological signal is expected, and independent validation using RT-PCR at varying degrees of read coverage.

The rest of this paper is organized as follows: Section 2.1 formulates the outlier model and the resulting generalization of MAJIQ, termed MAJIQout, Section 2.2 then describes the methods used to evaluate algorithm performance and to generate synthetic data, Section 3 details the comparative analysis on synthetic and real data of several algorithms for detecting differential splicing using replicates, followed by a discussion and future directions.

2 Methods

2.1 Outlier weight model

Let T be the set of RNA-seq experiments for which alternative junction inclusion is to be measured, and let $t \in T$ be one such experiment. All experiments constitute observations of reads mapping to L LSVs. Let $i = 1, 2, \ldots, L$, and let J be the number of junctions in LSV i with indices $j = 1, \ldots, J$. Then $\Psi_{i,j}^{(t)}$ is the inclusion ratio of junction j of LSV i within experiment t, with

$$\sum_{j=1}^{J} \Psi_{i,j}^{(t)} = 1, \tag{1}$$

and $\Psi_{i,j}^{(T)}$ is the inclusion ratio of junction j for the whole set of experiments, with the equivalent of Equation 1. Under the MAJIQ model, the set of $\Psi_{i,j}$ for LSV i has a Jeffrey's Dirichlet prior:

$$\{\Psi_{i,j}\}_{j=1}^{J} \sim \text{Dirichlet}\left(\frac{1}{J}, \cdots, \frac{1}{J}\right). \tag{2}$$

To simplify computations, we consider the marginal distribution of Ψ for each junction:

$$\Psi_{i,j} \sim \text{Beta}\left(\frac{1}{J}, \frac{J-1}{J}\right). \tag{3}$$

Define $D_{i,j}^{(t)}$ to be the number of reads mapping to junction j of LSV i in experiment t. Rather than using $D_{i,j}^{(t)}$ directly, MAJIQ applies a combination of GC bias corrections, stack removal, and bootstrapping from a zero-truncated negative binomial dispersion model over junction positions to return a per-junction read rate, μ. Let $\mu_{i,j,m}^{(t)}$ denote the mth bootstrapped read rate for junction j, where $m = 1, \ldots, M$. Define $\mu_{i,m}^{(t)} = \sum_{j=1}^{J} \mu_{i,j,m}^{(t)}$ to be the total read rate for LSV i, and let $\mu_{i,J\setminus j,m}^{(t)} = \mu_{i,m}^{(t)} - \mu_{i,j,m}^{(t)}$. Then

$$\{\Psi_{i,j}\}_{j=1}^{J} \mid \{\mu_{i,j,m}^{(t)}\}_{j=1}^{J} \sim \text{Dirichlet}\left(\frac{1}{J} + \mu_{i,1,m}^{(t)}, \cdots, \frac{1}{J} + \mu_{i,J,m}^{(t)}\right), \tag{4}$$

with marginal distribution

$$\Psi_{i,j} \mid \{\mu_{i,k,m}^{(t)}\}_{k=1}^{J} \sim \text{Beta}\left(\frac{1}{J} + \mu_{i,j,m}^{(t)}, \frac{J-1}{J} + \mu_{i,J\setminus j,m}^{(t)}\right). \tag{5}$$

In other words, Ψ is informed by the ratio of junction read rates. Indeed, as $\sum_{j=1}^{J} \mu_{i,j,m}^{(t)} \to \infty, E[\Psi_{i,j} \mid \{\mu_{i,j,m}^{(t)}\}_{j=1}^{J}] \to \frac{\mu_{i,j,m}^{(t)}}{\sum_{k=1}^{J} \mu_{i,k,m}^{(t)}}$ for each j.

We marginalize over the bootstrap samples by averaging their probability densities:

$$P\left(\Psi_{i,j} \mid \{\mu_{i,k,m}^{(t)}\}_{k=1}^{J}{}_{m=1}^{M}\right) = \frac{1}{M} \sum_{m=1}^{M} P(\left(\Psi_{i,j} \mid \{\mu_{i,k,m}^{(t)}\}_{k=1}^{J}\right) \tag{6}$$

To simplify notation, let $\Psi \mid \boldsymbol{\mu}_t = \Psi_{i,j} \mid \{\mu^{(t)}_{i,k,m}\}_{k \le J, m \le M}$

MAJIQ assumes that all the experiments in T are replicates of the same biological condition (tissue type, treatment, disease state, etc.). It follows that all experiments in T should share an underlying condition Ψ, denoted $\Psi^T_{i,j}$, $\forall i, j$. under this modeling assumption, Equation 5 generalizes to

$$\Psi_{i,j} \mid \boldsymbol{\mu}_{T,m} \sim \text{Beta}\left(\frac{1}{J} + \sum_{t \in T} \mu^{(t)}_{i,j,m}, \frac{J-1}{J} + \sum_{t \in T} \mu^{(t)}_{i,J\setminus j,m}\right), \tag{7}$$

where $\boldsymbol{\mu}_{T,m} = \{\boldsymbol{\mu}_{t,m}\}_{t \in T}$. A marginalization over $m = 1, \ldots, M$ exists and is a generalization of Equation 6. In this paper, we negate the replication assumption and consider the case where most but not all of the experiments in T represent the same experimental condition.

Definition 2.1. An *outlier* in T is an experiment in T which does not represent the same experimental or biological condition as the majority of the experiments in T. Specifically, s is an outlier in T if

$$\Psi^{(s)}_{i,j} \not\sim \Psi^{(T)}_{i,j} \tag{8}$$

for a sufficiently large proportion of LSVs.

Let $\rho_T(s)$ be the probability that replicate s is not an outlier in T, and define $\boldsymbol{\rho}_T = \{\rho_T(s)\}_{t \in T}$. We propose a generalized version of Equation 7 to estimate $\Psi_{i,j} \mid \boldsymbol{\mu}'_T$, where

$$\boldsymbol{\mu}'_T = \boldsymbol{\rho}_T \cdot \boldsymbol{\mu}_T. \tag{9}$$

In order to estimate $\rho_T(s)$ for suspected outlier s, we define a per-LSV metric of dissimilarity between Ψ distributions for each experiment relative to the group consensus.

Definition 2.2. Let X and Y be two continuous random variables with pdfs f_X and f_Y, respectively, such that at least one of their pdfs is nonzero on the interval I. The L_p *divergence* between X and Y is defined as

$$d_p(X, Y) = \left(\int_I |f_X(t) - f_Y(t)|^p dt\right)^{1/p}. \tag{10}$$

If X and Y are discrete random variables with pmfs f_X and f_Y, respectively, such that at least one of their pmfs is nonzero for $a \le k \le b$, then the L_p divergence between X and Y is defined as

$$d_p(X, Y) = \left(\sum_{k=a}^{b} |f_X(k) - f_Y(k)|^p\right)^{1/p}. \tag{11}$$

Setting $X = \Psi^{(t)}_{i,j}$ with pdf f_t, and $Y = \text{median}_{t \in T} \Psi^{(t)}_{i,j}$ with pdf f_{med}, in Equation 10, we have

$$d_p\left(\Psi^{(t)}_{i,j}, \text{median}_{t \in T} \Psi^{(t)}_{i,j}\right) = \left(\int_0^1 |f_t(\psi) - f_{\text{med}}(\psi)|^p\right)^{1/p} d\psi. \tag{12}$$

[1]

From this point, we define $d^{(t)}_{i,j} = d_p\left(\Psi^{(t)}_{i,j}, \text{median}_{t \in T} \Psi^{(t)}_{i,j}\right)$. We can summarize the L_p divergences of each replicate with respect to LSV i by taking the max divergence for each replicate over the junctions:

$$d^{(t)}_i = \max_{j \le J} d^{(t)}_{i,j}. \tag{13}$$

This leads into our primary postulate for outlier detection.

Postulate 1. *s is an outlier in T if $d^{(t)}_i$ is large for sufficiently many LSVs i.*

We say $d^{(t)}_i$ is large if it exceeds a predefined threshold $\tau > 0$. Intuitively, we can think of τ as a biologically informed definition for what constitutes a meaningful deviation. In the experiments that follow we found results were robust for a wide range of τ values (see below). Notably, for any reasonable τ definition we find that $d^{(t)}_i > \tau$ for multiple LSVs by chance alone. Let $K_t(\tau)$ be the set of LSVs i such that $d^{(t)}_i > \tau$, and let $K_T(\tau) = \bigcup_{t \in T} K_t(\tau)$. For fixed τ, we use the abbreviated notation K_t and K_T, respectively. Intuitively, K_t captures the total amount of significant variability in T, with more noisy data exhibiting large $|K_t|$ values. Importantly, if T has no outliers, we expect the high-divergence LSVs to be approximately evenly distributed across all replicates $t \in T$. That is,

$$E[|K_t|] = \frac{|K_T|}{|T|}. \tag{14}$$

[1] Because we bootstrap $D^{(}_{i,j}t)$ to capture more of the posterior variance in Ψ, we cannot explicitly define f_t in closed form. The median distribution, similarly, cannot be defined in closed form. To accommodate this, we discretize both distributions over fixed-width bins on the interval $[0, 1]$.

Thus it is natural to model $|K_t|$ as a Binomial(n, p) random variable with parameters $n = |K_T|$ and $p = |T|^{-1}$. In practice, however, the variance of the Binomial distribution (in this case, $|K_T|(|T|(1-|T|))^{-1}$) does not fit well variability of real data (data not shown). We account for this by letting $p \sim \text{Beta}(\alpha, \beta)$ with parameters α, β such that

$$\begin{aligned} \frac{\alpha}{\alpha+\beta} &= \frac{1}{|T|}, \\ \alpha+\beta &= \theta, \end{aligned}$$

where θ is a user-defined dispersion hyperparameter. In our experiments, setting $\theta = 0.10$ was sufficient to capture outlier samples in scenarios that included clear biological replicates. Under the full Beta-Binomial model, we finally define

$$\begin{aligned} \rho_T(t) &\propto P(|K_t| \geq |K_t|_{obs} \mid |K_T|, |T|, \tau, \theta) \\ &\sim \mathbf{BB}\left(|K_T|, \frac{\theta}{|T|}, \theta\left(1-\frac{1}{|T|}\right)\right). \end{aligned} \tag{15}$$

We further adjust these weights so that $P(|K_t| \geq E_\Theta[|K_t|] = 1$:

$$\rho_T(t) = \begin{cases} \frac{P_{BB}(|K_t||\Theta)}{P_{BB}(E_\Theta[|K_t|]|\Theta)}, & |K_t| > E_\Theta[|K_t|], \\ 1, & \text{else.} \end{cases} \tag{16}$$

2.2 Performance evaluation metrics

There is an inherent challenge in assessing the accuracy of methods for RNA-Seq analysis since the underlying true values are rarely known. Some works use synthetically-generated samples with specific transcripts spiked at different concentrations which may be very different from real life samples, while others resort to synthetic sequencing data generation under various simplifying assumptions. Instead, we focus here on using real life data with multiple replicates to assess *reproducibility* in different experimental setups as a mean of assessing the performance of Ψ and $\Delta\Psi$ quantification algorithms. Specifically, we use a reproducibility measure (RR) similar to the irreproducible discovery rate (IDR), which has been used extensively to evaluate ChIP-Seq peak calling methods [8] and, more recently, for methods detecting cancer driver mutations [14]. Conceptually, RR is a rank-based statistic, agnostic of an algorithm's model or scoring metric, which measures the proportion of high-ranked events (e.g. ChIP-Seq peaks or differentially-spliced events) that are also observed in a second, independent iteration of the same experiment. To compute the RR, an algorithm A is run on a "training" set, denoted $S1$, and outputs the number of differentially-spliced events (N_A), ranked by their relative significance or score. For any $n \leq N_A$, we then compute the size of the subset of events$(R_A(n) = n' \leq n)$ of those n events which are ranked in the n highest ranking events in a second "hidden" test set ($S2$). The reproducibility graph plots $R_A(n)$ as a function of n, with perfect reproducibility corresponding to a 45° line, and the reproducibility ratio RR statistic defined as the point $(R_A(N_A))$. We note that unlike the definition in [16], the RR graph is plotted as a function of n, n' and not $\frac{n}{N_A}, \frac{n'}{N_A}$ because the algorithms compared in this work varied greatly in terms of the overall number of events reported as significantly changing (N_A).

We acknowledge some key caveats regarding the usage of the reproducibility ratio RR and the number of significant events detected (N_A) to assess an algorithm's performance. First, both RR_A and N_A are not inherent characteristics of an algorithm A but rather a combination of an algorithm and a dataset. Furthermore, different algorithms may use different statistical criteria to call a splicing variation significantly changing. Consequently, their N_A may vary greatly. Second, reproducibility by itself is not a measure of accuracy as algorithms can be highly reproducible yet maintain a strong bias. In order to better assess accuracy of methods for differential splicing quantification, we perform two additional tests of performance. First, we assess a lower bound on the number of false positives (FP) by creating a balanced mix between experimental conditions. Consequently, the two groups being compared are expected to be identical mixes of biological conditions. The significantly changing events under this test (N^{ns}) are expected to be FPs. However, since we can not rule out inherent unknown bias even within the no-signal groups, we compute $R(N^{ns})$, expecting it to be close to 0. We then compute a conservative lower bound estimate on the False Discovery Rate (FDR) for a given algorithm A on dataset D as $FDR(A, D) \geq \frac{N_A^{ns} \cdot (1 - R(N_A^{ns}))}{N_A}$. Finally, as a second measure for an algorithm's accuracy, we used RT-PCR triplicate experiments from previous studies [16]. This measure is limited by the total number of events quantified, possible selection biases, and limitations of the experimental procedure. For example, for accurate quantification to be valid, careful reading of the gel bands (rather then qualitative calling of changes) need to be executed in triplicates. However, carefully executed RT-PCR provide valuable experimental validation and is considered the gold standard in the field.

2.3 Synthetic perturbation

To observe the impact of disagreement on Ψ in a controlled fashion, we use a real replicate and perturb it to create a synthetic new pseudo-replicate outlier using the following procedure:

1. Set $\theta \in [0, 1]$, $\delta \in [0, 1]$, and $\gamma > 0$.
2. Randomly sample $L \subset$ LSVs with $|L| = \theta|\text{LSVs}|$.
3. For $l \in L$ with per-junction read rates $\mu_{l,j}, j = 1, \ldots, J$:
   (a) Estimate $E[\Psi_{l,j}]$ for each junction.

(b) Sample $\varepsilon \sim U(0,1)$ and let

$$\sigma = \begin{cases} -1, & \varepsilon < E[\Psi_{l,1}], \\ 1, & \text{else.} \end{cases}$$

(c) Set $E[\Psi^*_{l,1}] = \min(\max(E[\Psi_{l,j}] + \sigma\delta, 0), 1)$.

(d) For $2 \leq j \leq J$, set

$$E[\Psi^*_{l,j}] = E[\Psi_{l,j}] + \frac{E[\Psi_{l,1}] - E[\Psi^*_{l,1}]}{J-1}.$$

(e) For $1 \leq j \leq J$, set

$$D^*_{l,j} = E[\Psi^*_{l,j}] \frac{\mu_{l,j}}{\sum_{k=1}^{J} \mu_{l,k}}.$$

4. For $l \in \text{LSVs} \setminus L$, set $\mu^*_{l,j} = \mu_{l,j}$.
5. For $l \in \text{LSVs}$, set $\mu^*_{l,j} = \gamma\mu^*_{l,j}$.

Observe that when $\gamma = 1$ and $0 \in \{\theta, \delta\}$, the synthetic perturbation does not alter μ for any LSV. We measure the effect of variations in θ, δ, and γ on ρ_T and RR by applying the above Ψ perturbation to one replicate in set $S1$.

2.4 Mislabeled sample

In an extreme case, we explore the effects of mislabeling a sample. We simulate this by swapping out one replicate in the set $S1$ with a sample from a different condition within the same dataset.

2.5 Source data

The results described here were derived using data from two different studies. Most of the analysis was done using RNA-seq data sourced from the Mouse Genome Project (MGP) transcriptome initiative [6]. The MGP dataset covers six tissues in *Mus musculus* with six biological replicates each, at 18-30 million reads per replicate. We supplement this data with a more recent study from [19] which includes twelve mouse tissues samples across eight time points. We use these data to test reproducibility across datasets, for behavior under no-signal conditions, and for comparison to biochemical quantifications of splicing.

3 Results

Figure 2 shows the effect of different synthetic perturbation of a replicate on the weight associated with that sample (ρ_T, left column), the number (N_A, middle column) of LSVs reported as differentially spliced with high confidence ($P(|\Delta\Psi| > 0.2) > 0.95\%$), and the reproducibility ratio ($RR_A = \frac{R_A}{N_A}$, right column). At $\delta = 0.6, \gamma = 1$ (top row) the outlier's weight ρ_t scales linearly in log scale to the fraction of LSVs perturbed and 10% is sufficient to drop ρ_t to 0.1. Consequently, MAJIQ detects up to approximately 400 false positives and reproducibility drops down to approximately 60% while both $N_{MAJIQout}$ and $RR_{MAJIQout}$ remain stable (Figure 2b,c). At $\theta = 0.3, \gamma = 1$ (middle row), increasing δ initially causes the weight on the outlier to decrease towards a positive infimum. For larger $\delta > 0.5$, the N_{MAJIQ} increases 4-fold with a corresponding 50% drop in RR_{MAJIQ}. At $\theta = 0.3, \delta = 0.6$, decreasing γ towards 0 causes the weight to shrink, suggesting that the algorithm is highly sensitive to low read counts. Indeed, without enough reads, the estimated Ψ distribution does not vary significantly from the prior. Unsurprisingly, increasing the read rates to 150% does not significantly affect the weight. It does, however, increase the unreliability of MAJIQ, tripling N_{MAJIQ} while nearly halving RR_{MAJIQ}. In all these cases, MAJIQout remains resistant to the perturbations.

Next, we evaluated the reproducibility with and without a sample swap for a large set of algorithms. In all cases, we used 2 heart samples vs. 2 liver samples for the validation set 2. Set 1 included either 2 liver samples compared with 2 heart samples (no swap) three livers compared with two hearts and one hippocampus sample (swap case). To produce the reproducibility curves we followed the following procedure. For MAJIQ and MAJIQout, N_A was defined over set 1 as the set for which $P(|\Delta\Psi| >= 0.2) > 0.95$ as in [16]. Similarly, for SUPPA and rMATS we define N_A as the set of significant events with $\Delta\Psi >= 0.2$; we use the provided $p - value <= 0.05$ in order to filter for significance. DEXSeq returns a $\log_2$ fold change value rather than a $\Delta\Psi$; in this case the rank is based on $\log_2$ fold change > 4, and we call an event significant if its adjusted $pvalue <= 0.05$. We also tried to rank rMATS hits by FDR rather than $\Delta\Psi$, but these rankings were far less reproducible than the $\Delta\Psi$-based rankings (data not shown).

Figure 3 summarizes the results for the evaluation procedure described above. One clear observation is the huge variation in the number of events reported as significantly changing by the different methods even when no outliers are present, ranging from 576 MAJIQ, through 1359 and 1686 (rMATS and SUPPA respectively) to 8096 (DEXSeq). When compared to the other algorithms without an outlier replica, both MAJIQ and MAJIQout exhibit significantly higher levels of reproducibility of the events ranking for significant changing events regardless of the N cutoff (light blue and light purple lines respectively). The higher reproducability is specifically notable for the several hundred top ranked events (inset figure). When an outlier is present, MAJIQ's N jumps to 886, and reproducibilty drops dramatically; MAJIQout, meanwhile, is not affected by the outlier (dark blue and dark purple lines respectively). The reproducibility ratio of rMATS, SUPPA and DEXSeq is generally lower, but these are not so sensitive to outliers. Noticeably, in some cases reproducibility even improves compared to the control, likely due to the introduction of the additional "good" liver sample in Set 1.

Next, we repeated the reproducibility evaluation but with a different dataset and at different levels of coverage, varying from 100% through 50% to 25% of the original reads (Figure 4). Unlike the other algorithms, without parameter adjustments MAJIQ may be sensitive to low coverage since it relies on junction spanning reads. MAJIQ's default parameters, constructed for high-coverage data, require 10 reads from 3 different positions across a junction to define quantifiable events [16]. In order to maintain sufficient detection power at low coverage we adjusted these parameters to 3 reads from 2 positions, and also allowed the minimal number of samples including this event to drop to one. At the baseline of 100% coverage, this data included 4 replicates per tissue (cerebellum vs. liver) with an average of approximately 80M reads per sample. The larger number of replicates and higher coverage led to a much higher number of events identified as differentially spliced by all methods, and likely contributed to overall higher reproducibility as well. In terms of comparison between methods, the same trend remained, with MAJIQout comparing favorably with a stable reproducibility ratio of around 82% at different coverage levels. However, with the increased coverage and replicates compared to the data in Figure 2, MAJIQout denoted approximately 2000 events as deferentially spliced − similar to SUPPA, less than rMATS (∼2500), and significantly less than DEXSeq (¿8000). Finally, when testing the effect of lower coverage (x0.5, x0.25, dashed and dotted lines in Figure 4), we found some drop in reproducibility in most cases, with DEXSeq appearing to be the most sensitive to coverage levels.

In order to assess the fraction of false positives from the set of events reported by each method (FDR), we created no-signal groups from the datasets used in Figure 3 and Figure 4 by comparing two sets that involve an equal mix of replicates from the two tissues (see Section 2). This gave us a total of four no-signal groups for which we tested how many events were still determined as significantly changing. We found MAJIQ reports a lower number of events suspected to be false positive, with SUPPA and DEXSeq both suffering from high N^{ns} values and high variability between sets. This high variability may point to possible sensitivity to the dataset definition.

As expected, the set of events identified as deferentially spliced in the no signal groups also exhibited low reproducibility ratios ($R(N^{ns})$, see Figure S2). By combining N,N^{ns} and $R(N^{ns})$ as detailed in Section 2, we got a conservative lower bound on each methods FDR for each of the datasets. Figure 5b, shows MAJIQ had a significantly lower FDR estimate especially compared to SUPPA and DEXSeq.

Finally, we assessed the methods accuracy by RT-PCR as a function of the read coverage, either 100%, 50%, or 25% of the original reads (Figure S3). We downsampled cerebellum and liver timepoints CT28, CT40, and CT52 from the mouse circadian study [19] and correlated them with 50 RT-PCR $\Delta\Psi$ quantifications from the same tissue comparison. For the reduced-coverage experiments, we adjusted MAJIQ's execution parameters as described above. We found that on the original data (100%), MAJIQ recapitulates the results from [16] Figure S2.1B, and MAJIQout does not differ significantly ($R = 0.982$). By this metric, rMATS performs similarly to MAJIQout, while SUPPA slightly underperforms both algorithms. Decreasing the simulated read depth slightly decreases the number of LSVs which MAJIQ and MAJIQout are able to detect as quantifiable (47 at 50%, 43 at 25%), but correlation with the same RT-PCR quantifications remains high ($R = 0.962$ at 25%). rMATS maintains all events and performs similarly to MAJIQout on both downsampled fractions while SUPPA's correlation drops below 0.90 on the downsampled datasets.

4 Discussion

In this paper we developed a new model to automatically detect and down weight outliers in RNA-Seq datasets with replicates for splicing analysis. The problem of detecting outliers in batches of biological replicates has not received much attention in the literature as researchers are likely to simply discard samples before publication based on some heuristic. Such a heuristic may in turn reflect unconscious bias or cause good data to be lost. Next, by merging the outlier detection model into our previous algorithm, MAJIQ, we created a generalized version of the latter termed MAJIQout. We analyzed MAJIQ and MAJIQout using synthetic and real life data and showed MAJIQout maintains MAJIQ's favorable performance on data without outliers, and was also robust to outliers. When read coverage was low, MAJIQout was able to maintain relatively high detection power, high reproducibility, and high correlation to RT-PCR by adjusting its default execution parameters. However, since different datasets may suffer from different types of noise or biases, it is advisable for potential users to test algorithms using the kind of evaluation criteria introduced here, including reproducibility plots, no-signal groups, and RT-PCR. In addition, the methods tested here differ greatly in the set of features they offer. MAJIQ is the only one that offers the ability to detect complex splicing variations involving more than two alternative junctions, and couples these with interactive visualization and genome browser connectivity. It is also capable of supplementing a given transcriptome annotation with reliable *de-novo* junctions detected in the RNA-Seq data. While useful for even normal tissues [16], this feature is particularly relevant for disease studies, cases where uncharacteristic splicing is expected, and for species with poorly-annotated transcriptomes. Notably, the latest version of rMATS offers to include *de-novo* junctions but requires those to be at a predefined distance (a user-controlled parameter) so it can add those to annotated exons. In contrast, MAJIQ is able to detect completely novel junctions and exons. This ability of MAJIQ does come with a price of algorithm complexity and, consequently, execution time. While we did not perform detailed benchmarking, MAJIQ was much faster than rMATS and DEXSeq. However, SUPPA was much faster than all the other methods, as it assumes a known transcriptome and uses fast pseudoalignment algorithms such as SALMON to quantify each transcript's abundance. These assumptions may have deleterious effects on performance and might be at least partially responsible for the higher rate false positive we observed for SUPPA.

There are several important directions in which this work can be extended. First, MAJIQout can be further improved both in terms of memory consumption and running time. While we were able to process over 100 samples with the current implementation on machines with 64GB of memory, parsing several hundreds or thousands or samples is currently not feasible. Furthermore, all the algorithms compared here were designed for datasets with small sets of replicates. Large heterogeneous datasets, such as those created in cancer studies, are likely to benefit from different statistical models. Finally,

MAJIQ's improved quantifications can be used to subsequently derive new models for splicing codes and splicing predictions given genetic variations [3, 4, 18]. Such improvements form a promising path for future algorithm development.

Acknowledgements

We would like to thank Matthew R. Gazzara for helpful comments and suggestions.

Funding

This work has been supported by R01 AG046544 to YB.

References

[1] Gael P. Alamancos, Eneritz Agirre, and Eduardo Eyras. *Methods to Study Splicing from High-Throughput RNA Sequencing Data*, pp. 357–397. Humana Press, Totowa, NJ, 2014.

[2] Simon Anders, Alejandro Reyes, and Wolfgang Huber. Detecting differential usage of exons from rna-seq data. *Genome Research*, 22(10):2008–2017, 2012.

[3] Yoseph Barash, John A. Calarco, Weijun Gao, Qun Pan, Xinchen Wang, Ofer Shai, Benjamin J. Blencowe, and Brendan J. Frey. Deciphering the splicing code. *Nature*, 465(7294):53–59, May 2010.

[4] Yoseph Barash, Jorge Vaquero-Garcia, Juan González-Vallinas, Hui Yuan Xiong, Weijun Gao, Leo J. Lee, and Brendan J. Frey. AVISPA: a web tool for the prediction and analysis of alternative splicing. *Genome Biology*, 14(10):1–8, 2013.

[5] Juan C Entizne, Juan L Trincado, Gerald Hysenaj, Babita Singh, Miha Skalic, David J Elliott, and Eduardo Eyras. Fast and accurate differential splicing analysis across multiple conditions with replicates. *bioRxiv*, 2016.

[6] Thomas M. Keane, Leo Goodstadt, Petr Danecek, Michael A. White, Kim Wong, Binnaz Yalcin, Andreas Heger, Avigail Agam, Guy Slater, Martin Goodson, Nicholas A. Furlotte, Eleazar Eskin, Christoffer Nellaker, Helen Whitley, James Cleak, Deborah Janowitz, Polinka Hernandez-Pliego, Andrew Edwards, T. Grant Belgard, Peter L. Oliver, Rebecca E. McIntyre, Amarjit Bhomra, Jerome Nicod, Xiangchao Gan, Wei Yuan, Louise van der Weyden, Charles A. Steward, Sendu Bala, Jim Stalker, Richard Mott, Richard Durbin, Ian J. Jackson, Anne Czechanski, Jose Afonso Guerra-Assuncao, Leah Rae Donahue, Laura G. Reinholdt, Bret A. Payseur, Chris P. Ponting, Ewan Birney, Jonathan Flint, and David J. Adams. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, 477(7364):289–294, September 2011.

[7] Bo Li and Colin N. Dewey. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC Bioinformatics*, 12(1):323, 2011.

[8] Qunhua Li, James B. Brown, Haiyan Huang, and Peter J. Bickel. Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics*, 5(3):1752–1779, 2011.

[9] Ruolin Liu, Ann E. Loraine, and Julie A. Dickerson. Comparisons of computational methods for differential alternative splicing detection using rna-seq in plant systems. *BMC Bioinformatics*, 15(1):364, 2014.

[10] Q. Pan, O. Shai, L. J Lee, B. J Frey, and B. J Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*, 40(12):1413–1415, December 2008.

[11] Rob Patro, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. Salmon provides accurate, fast, and bias-aware transcript expression estimates using dual-phase inference. *bioRxiv*, 2016.

[12] Rob Patro, Stephen M Mount, and Carl Kingsford. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat Biotech*, 32(5):462–464, May 2014.

[13] Shihao Shen, Juw Won Park, Zhi-xiang Lu, Lan Lin, Michael D. Henry, Ying Nian Wu, Qing Zhou, and Yi Xing. rmats: Robust and flexible detection of differential alternative splicing from replicate rna-seq data. *Proceedings of the National Academy of Sciences*, 111(51):E5593–E5601, 2014.

[14] Collin J. Tokheim, Nickolas Papadopoulos, Kenneth W. Kinzler, Bert Vogelstein, and Rachel Karchin. Evaluating the evaluation of cancer driver genes. *Proceedings of the National Academy of Sciences*, 113(50):14330–14335, 2016.

[15] Cole Trapnell, David G Hendrickson, Martin Sauvageau, Loyal Goff, John L Rinn, and Lior Pachter. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotech*, 31(1):46–53, January 2013.

[16] Jorge Vaquero-Garcia, Alejandro Barrera, Matthew R Gazzara, Juan González-Vallinas, Nicholas F Lahens, John B Hogenesch, Kristen W Lynch, and Yoseph Barash. A new view of transcriptome complexity and regulation through the lens of local splicing variations. *eLife*, 5:e11752, February 2016.

[17] E. T Wang and A. T Cooper. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nature*, 8:749–761, 2007.

[18] Hui Y. Xiong, Babak Alipanahi, Leo J. Lee, Hannes Bretschneider, Daniele Merico, Ryan K. C. Yuen, Yimin Hua, Serge Gueroussov, Hamed S. Najafabadi, Timothy R. Hughes, Quaid Morris, Yoseph Barash, Adrian R. Krainer, Nebojsa Jojic, Stephen W. Scherer, Benjamin J. Blencowe, and Brendan J. Frey. The human splicing code reveals new insights into the genetic determinants of disease. *Science*, 347(6218), 2015.

[19] Ray Zhang, Nicholas F. Lahens, Heather I. Ballance, Michael E. Hughes, and John B. Hogenesch. A circadian gene expression atlas in mammals: Implications for biology and medicine. *Proceedings of the National Academy of Sciences*, 111(45):16219–16224, 2014.

338 # Main text figures

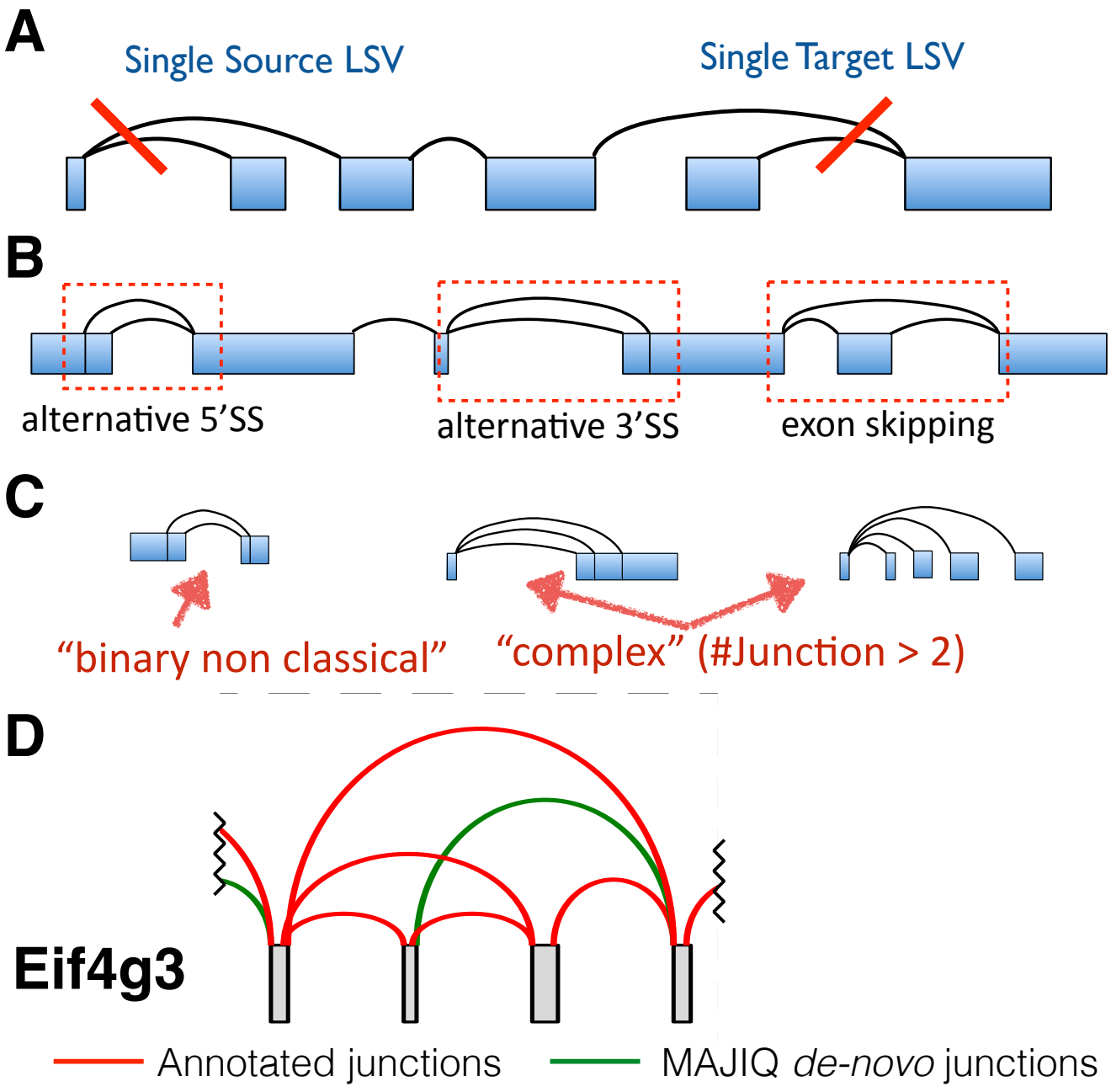

Figure 1: Illustration of the LSV definition. **a.** LSVs can be single-source (5' split) or single-target (3' join). **b.** The LSV definition is sufficient to explain the classical binary splicing event types. **c.** The LSV definition explains additional complexity observed in metazoan genomes, including non-classical binary events and complex (3 or more junctions) events. **d.** An example of a complex splicing event at the mouse Eif4g3 locus, augmented with *de-novo* junction detection by MAJIQ and validated by RT-PCR. Junctions drawn in red arise from the annotation and are supported by RNA-seq, whereas junctions drawn in green were detected from RNA-seq but not the annotation.

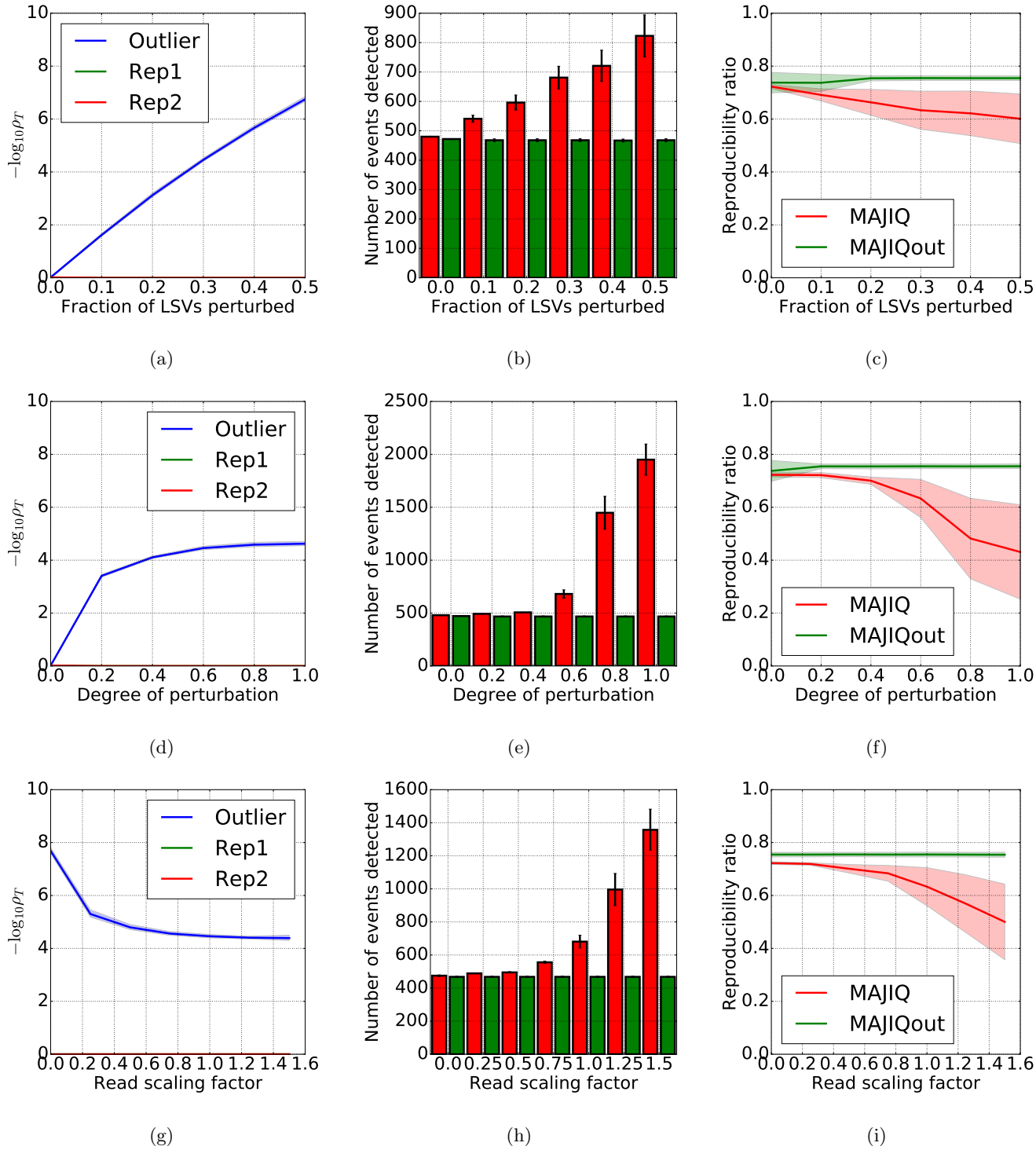

Figure 2: Effect of varying the synthetic perturbation hyperparameters on computed weights ρ_T (**a, d, g**), detection power (**b, e, h**), and reproducibility of detected events (**c, f, i**). **a-c.** Dependence on θ with $\delta = 0.60$ and $\gamma = 1.00$. **d-f.** Dependence on δ with $\theta = 0.30$ and $\gamma = 1.00$. **g-i.** Dependence on γ with $\theta = 0.30$ and $\delta = 0.60$.

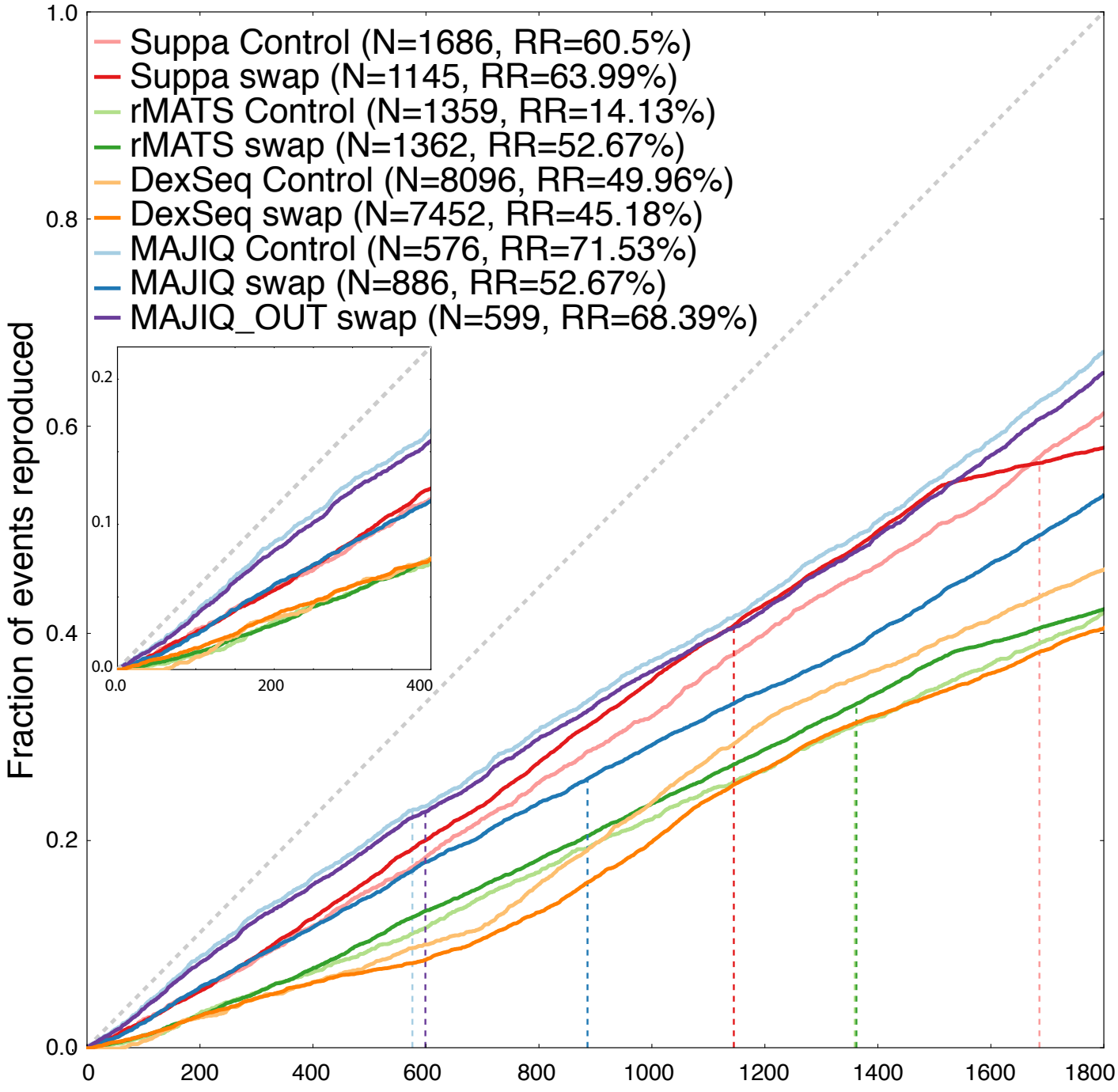

Figure 3: Reproducibility plots for differential splicing between tissues w/wo a mislabeled sample, as described in the text.

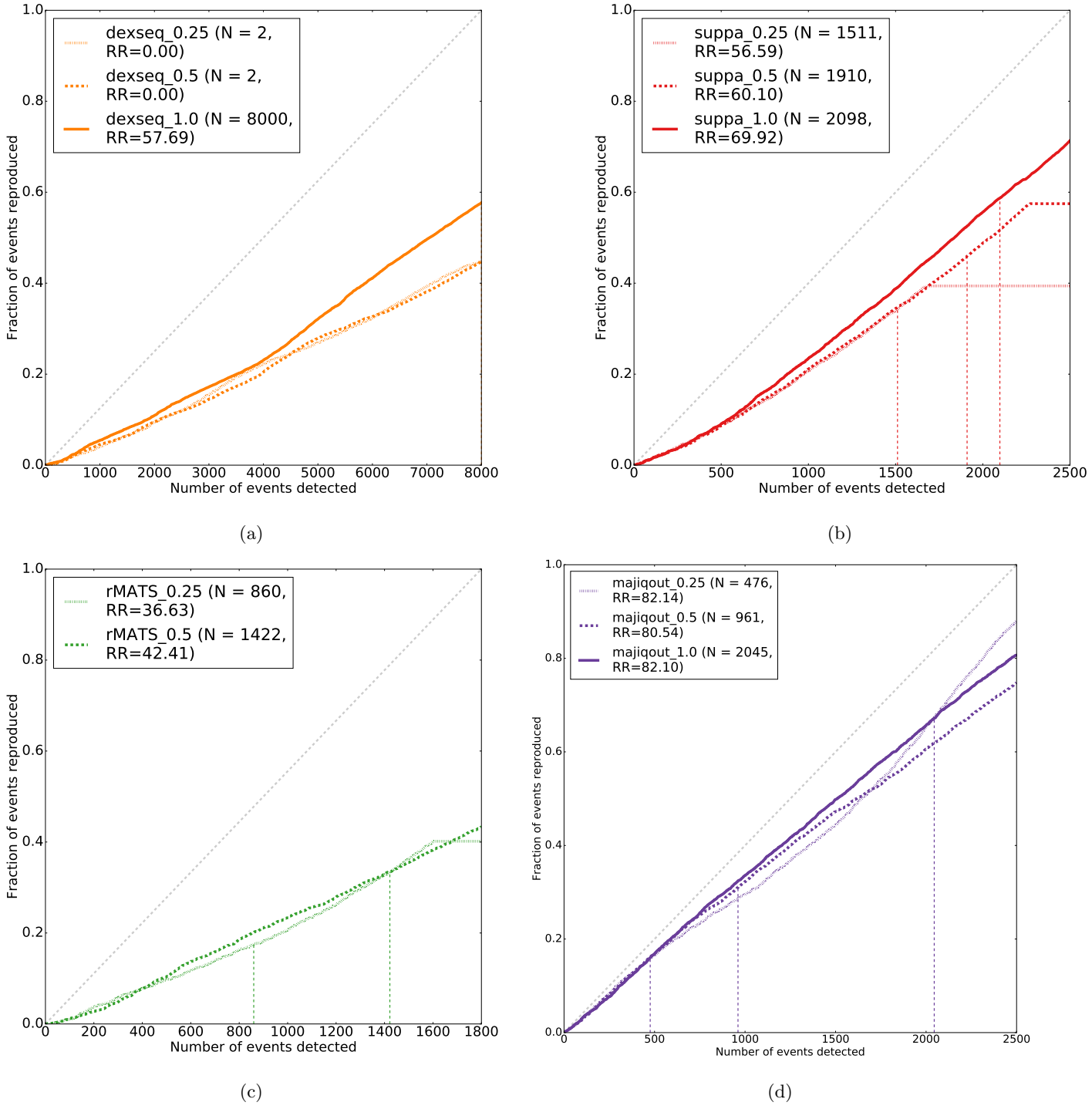

Figure 4: Reproducibility plots for differential splicing between tissues at different simulated coverage levels. **a.** DEXSeq; **b.** SUPPA; **c.** rMATS; **d.** MAJIQout.

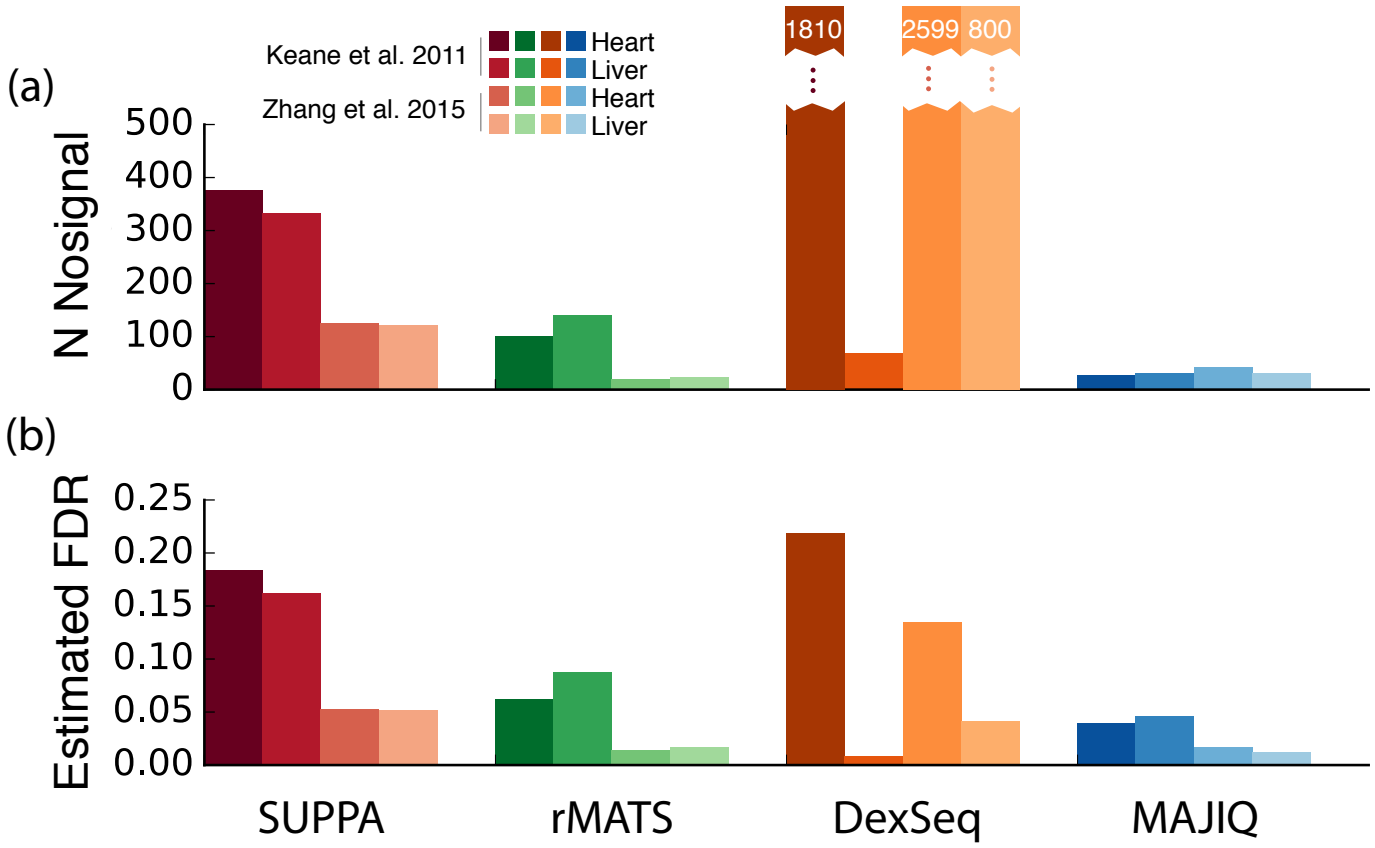

Figure 5: Evaluation of methods under no-signal conditions for SUPPA, rMATS, DEXSeq, and MAJIQ. **a.** Number of detected events. **b.** Lower-bound estimate of false discovery rate.

RT-PCR Correlations

	x1	x0.5	x0.25
SUPPA	0.906	0.899	0.883
rMATS	0.986	0.975	0.968
MAJIQ	0.982	0.970	0.962

(a)

 # RT-PCR Events Detected

	x1	x0.5	x0.25
SUPPA	37	37	37
rMATS	50	50	50
MAJIQ	50	47	43

(b)

Figure 6: **a.** Correlation coefficients of SUPPA, rMATS, and MAJIQ with RT-PCR from [16] as a function of read coverage. A total of 50 events were quantified. In order to enable comparison to other methods only simple cassette exons from the annotation database were tested and no complex variations were included. **b.** Size of the intersection between the events detected by each method and the subset of events for which RT-PCR $\Delta\Psi$ quantification was performed.

Supplementary figures

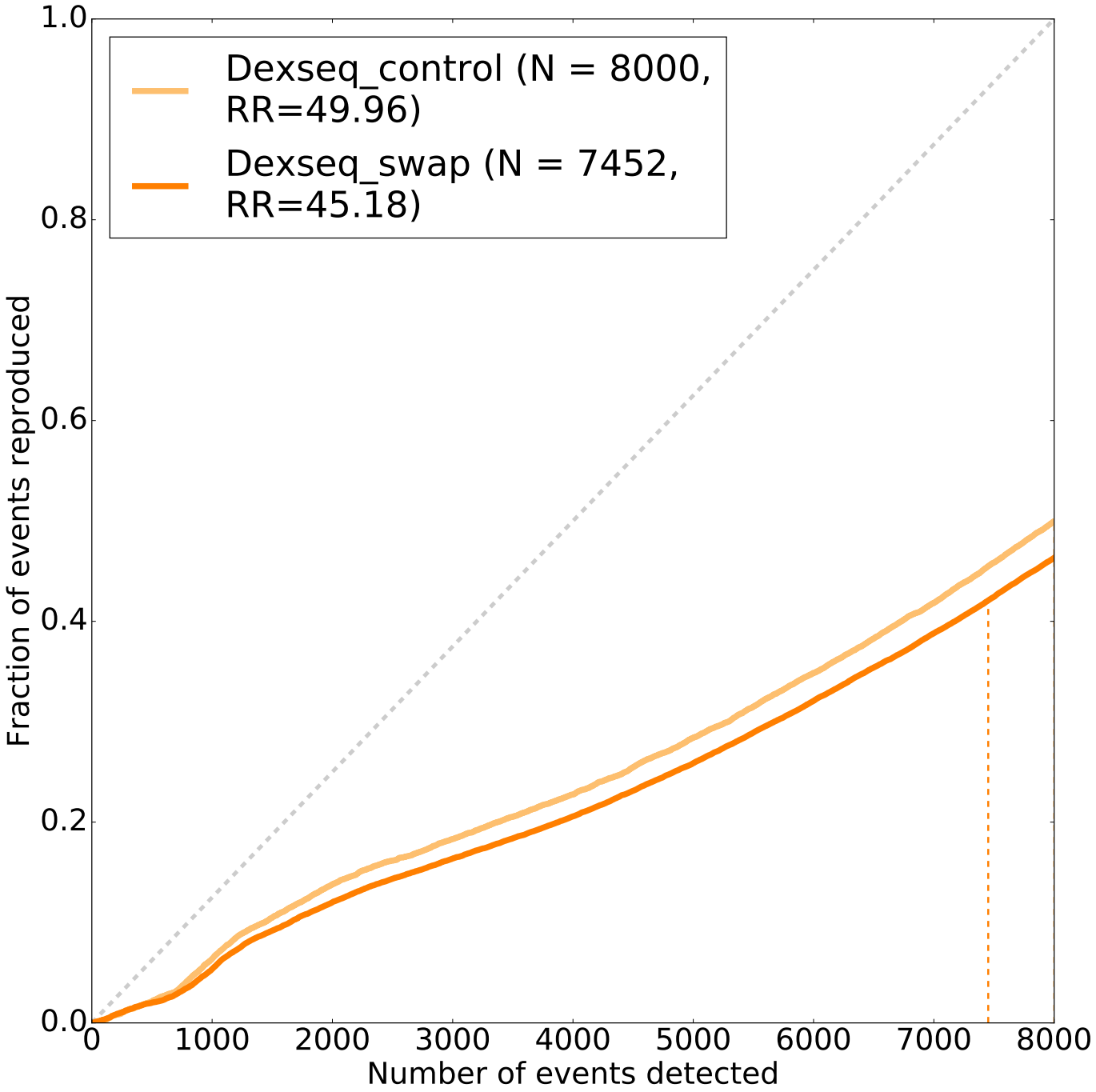

Figure S1: **Supplement to Figure 3.** Extended RR plot for DEXSeq showing the large number of splicing events flagged as significantly changing in the replicate swap experiment. Events beyond the 8000th call were ignored.

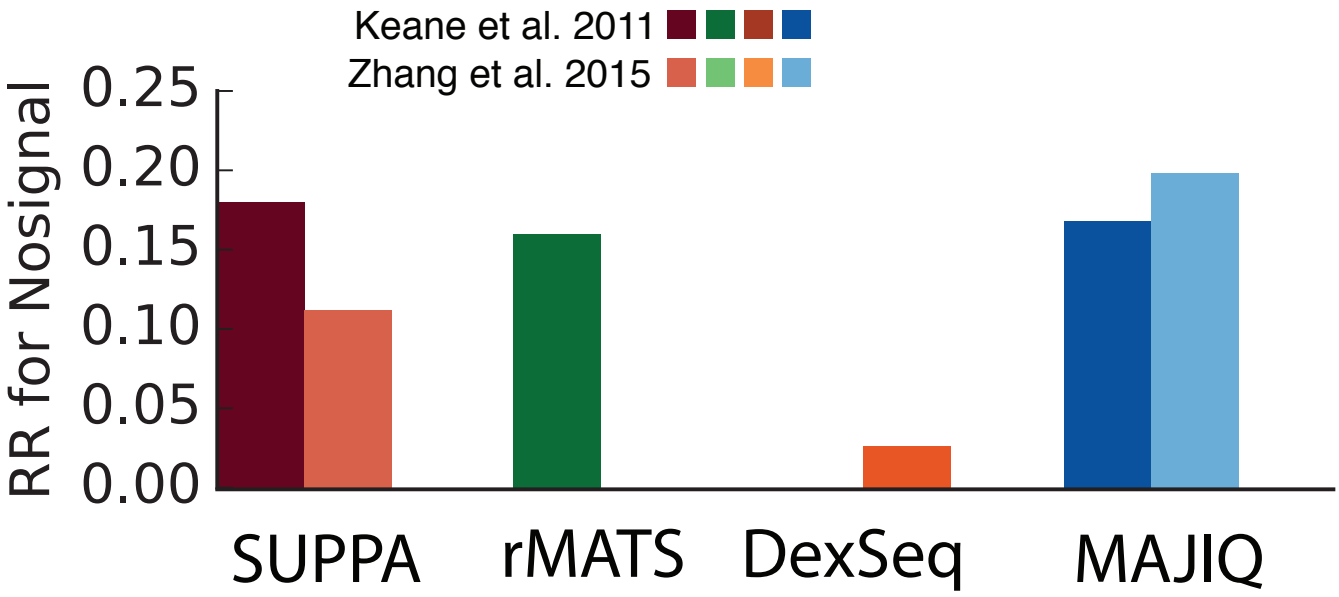

Figure S2: **Supplement to Figure 5.** Fractions of events detected in the no-signal runs which were reproduced in a second, disjoint no-signal experiment. These values were used to estimate the false discovery rates reported in Figure 5b.

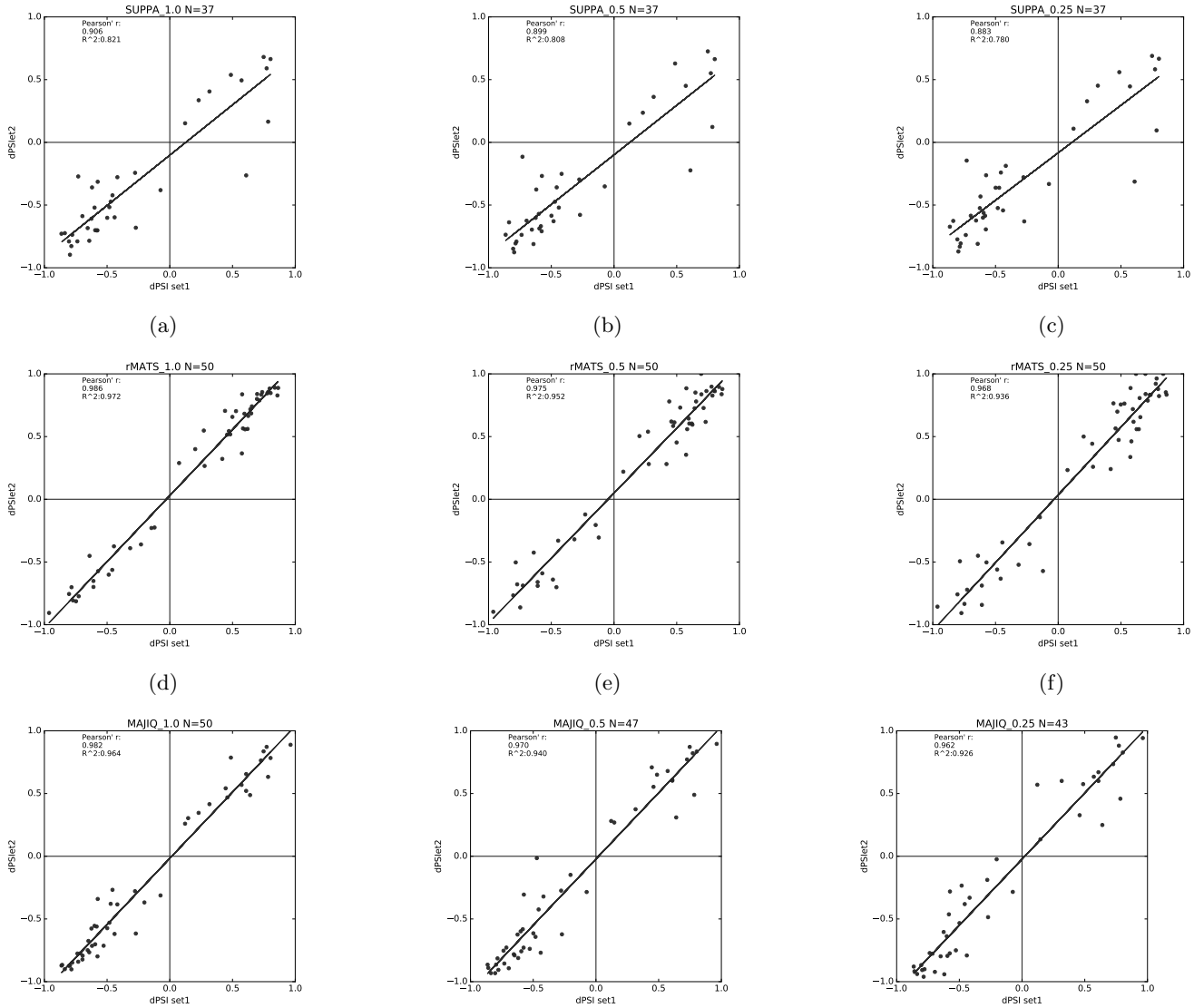

Figure S3: **Supplement to Figure 4.** RT-PCR correlations for 50 LSVs in the Hogenesch Cerebellum vs. Liver experiment, with downsampling. **a-c.** SUPPA; **d-f.** rMATS; **g-i.** MAJIQ. **a,d,g.** No downsampling; **b,e,h.** 50% downsampled; **c,f,i.** 25% downsampled.