DIFFERENTIAL ANALYSIS OF GENE REGULATION BY HOXA1 AT ISOFORM RESOLUTION WITH RNA-SEQ (SUPPLEMENTAL MATERIAL)

C. TRAPNELL, D. G. HENDRICKSON, M. SAUVAGEAU, L. GOFF, J. L. RINN, AND L. PACHTER

1. Limitations of raw counts in differential analysis

Several groups have proposed methods for detecting differentially expressed genes by directly comparing fragment counts for each gene in two conditions, rather than assigning these "raw" counts to transcripts and then comparing the inferred gene expression levels. We show here that for genes with multiple isoforms, changes in raw counts between conditions do not necessarily imply a change in expression and vice versa.

Under the simplest multinomial model of RNA-Seq, the expression level[1] of transcript t in condition a is estimated by:

$$\hat{\rho}_t^a \quad = \quad \frac{X_t^a}{\eta_a} \cdot \frac{1}{l_t - m + 1} \cdot \left(\frac{1}{\sum_{s \in T} \frac{X_s^a}{\eta_a (l_t + m - 1)}} \right) \quad \propto \quad \frac{X_t^a}{\eta_a} \cdot \frac{1}{l_t - m + 1} \tag{1.1}$$

where X_t^a is the number of fragments that *unambiguously* map to transcript t, η_a is a normalization term for the size of the library for condition a (for example, the total number of fragments mapped in condition a), m is the length of the sequenced fragments, and l_t is the length of the transcript. We abbreviate $\tilde{l}_t = l_t + m - 1$ (also known as the "effective length" of t). Because the length of a transcript is a static feature that does not change between conditions, the following proposition is trivial:

Proposition 1. *The fold change in estimated expression of transcript t between conditions a and b is equal to the fold change in the number of fragments originating from that transcript:*

$$\Delta\rho_t \quad = \quad \frac{\hat{\rho}_t^a}{\hat{\rho}_t^b} \quad = \quad \frac{X_t^a}{X_t^b}. \tag{1.2}$$

We often wish to calculate the fold change in expression not just for a single transcript, but for a group of $k \geq 2$ transcripts $G = \{t_1, ..., t_k\}$. The expression level of this group in a condition is the sum of the expression levels for the transcripts in the group:

$$\rho_G \quad = \quad \sum_{t \in G} \rho_t \tag{1.3}$$

so that

$$\hat{\rho}_G \quad \propto \quad \sum_{t \in G} \frac{X_t^a}{\eta_a \tilde{l}_t}. \tag{1.4}$$

Thus the fold change for the group is:

$$\Delta\rho_G \quad = \quad \frac{\sum_{t \in G} \frac{X_t^a}{\eta_a \tilde{l}_t}}{\sum_{t \in G} \frac{X_t^b}{\eta_b \tilde{l}_t}} \quad = \quad \frac{\eta_b}{\eta_a} \cdot \frac{\sum_{t \in G} \frac{X_t^a}{\tilde{l}_t}}{\sum_{t \in G} \frac{X_t^b}{\tilde{l}_t}}. \tag{1.5}$$

That is, in order to calculate the fold change in G, we must know how many fragments came from each transcript. Typically, these counts are computed by simple inspection of fragment alignments against a genome or transcriptome: fragments that uniquely align to a single transcript can be unambiguously attributed to that transcript. However, in the human genome and many others, most genes are alternatively spliced, and thus a large fraction (potentially the vast majority) of reads will align to more than one transcript [6]. While transcript-level counts may not be directly computable, gene-level counts often are. That is, a fragment that aligns to a constitutive exon may have come from one of several alternative isoforms of a gene,

[1] Technically by "expression level" we mean relative abundance, and although the two may be interpreted differently the terms have been used interchangeably in previous RNA-Seq literature and we adopt that slight abuse of terminology as well.

but there is no question that the fragment came from that gene. A number of methods thus estimate a gene's expression as:

$$\rho_G \approx \frac{\sum_{t \in G} X_t^a}{\eta_a l_\epsilon} \tag{1.6}$$

where l_ϵ is the length of a "representative" model for the gene. Popular constructions for a representative model include taking the union of all exonic bases in the gene, or restricting to the set of all constitutive. We term the former the "exon-union" model and the latter the "exon-intersection" model (main text Figure 1). We show that a fold change in fragment counts based on either of these approaches is not equivalent to fold change in the expression of G.

Let the number of fragments originating from any transcript in G in condition a be written X_G^a, and the number of fragments counted under the exon union model be $U(X_G^a)$. Similarly, $I(X_G^a)$ represents the number of fragments counted for G under the exon-intersection model. For simplicity, consider the case where the libraries for condition a and b contain an equal number of fragments. The fold change in expression of G is

$$\Delta\rho_G = \frac{\sum_{t_i \in G} \frac{X_{t_i}^a}{\tilde{l}_{t_i}}}{\sum_{t \in G} \frac{X_{t_i}^b}{\tilde{l}_{t_i}}}. \tag{1.7}$$

Under the union counting scheme, the estimated fold change is accurate when

$$\frac{U(X_G^a)}{U(X_G^b)} = \frac{\sum_{t_i \in G} \frac{X_{t_i}^a}{\tilde{l}_{t_i}}}{\sum_{t \in G} \frac{X_{t_i}^b}{\tilde{l}_{t_i}}}. \tag{1.8}$$

Since the union scheme excludes no exons, and thus no fragments from the count, the above is equivalent to

$$\frac{X_G^a}{X_G^b} = \frac{\sum_{t_i \in G} \frac{X_{t_i}^a}{\tilde{l}_{t_i}}}{\sum_{t \in G} \frac{X_{t_i}^b}{\tilde{l}_{t_i}}}. \tag{1.9}$$

or

$$\sum_{t_i \in G} \left[X_G^a \frac{X_{t_i}^b}{\tilde{l}_{t_i}} - X_G^b \frac{X_{t_i}^a}{\tilde{l}_{t_i}} \right] = 0. \tag{1.10}$$

The above equation is satisfied for all values of $X_{t_i}^a$ and $X_{t_i}^b$ when the effective length of all transcripts in G is one, which is the case when the fragments are always the same length as the transcripts being sequenced. While this is sometimes the case for small RNA sequencing experiments, current read lengths ($\leq$ 150bp) are much shorter than most mRNAs. The equation may also be satisfied for certain assignments of $X_{t_i}^a$ and $X_{t_i}^b$ regardless of transcript length, but for most comparisons, a change in gene-level counts under the union model does not equal the underlying change in expression.

The derivation that changes in total fragment counts under the intersection counting scheme (written $I(X_g^a)$ and $I(X_g^b)$) are not equivalent to changes in gene expression proceeds similarly to the argument for the union scheme. The changes in total counts are equivalent to changes in gene expression only when

$$I(X_G^a) \sum_{t_i \in G} \frac{X_{t_i}^b}{\tilde{l}_{t_i}} - I(X_G^b) \sum_{t \in G} \frac{X_{t_i}^a}{\tilde{l}_{t_i}} = 0. \tag{1.11}$$

A change in counts under exon intersection, like the union model, may be drastically different than the change in true gene expression, and may even indicate fold change in the opposite direction. This problem is illustrated in Figure 1.

The discrepancy between change in count and change in expression clearly depends on the "geometry" of the gene in question. The length of each isoform and the specific fragments that fall on constitutive as opposed to unique exons determine the expression of the isoforms and the gene. The discrepancy between change in count and change in expression for a gene is driven by *isoform switching*, rather than differential

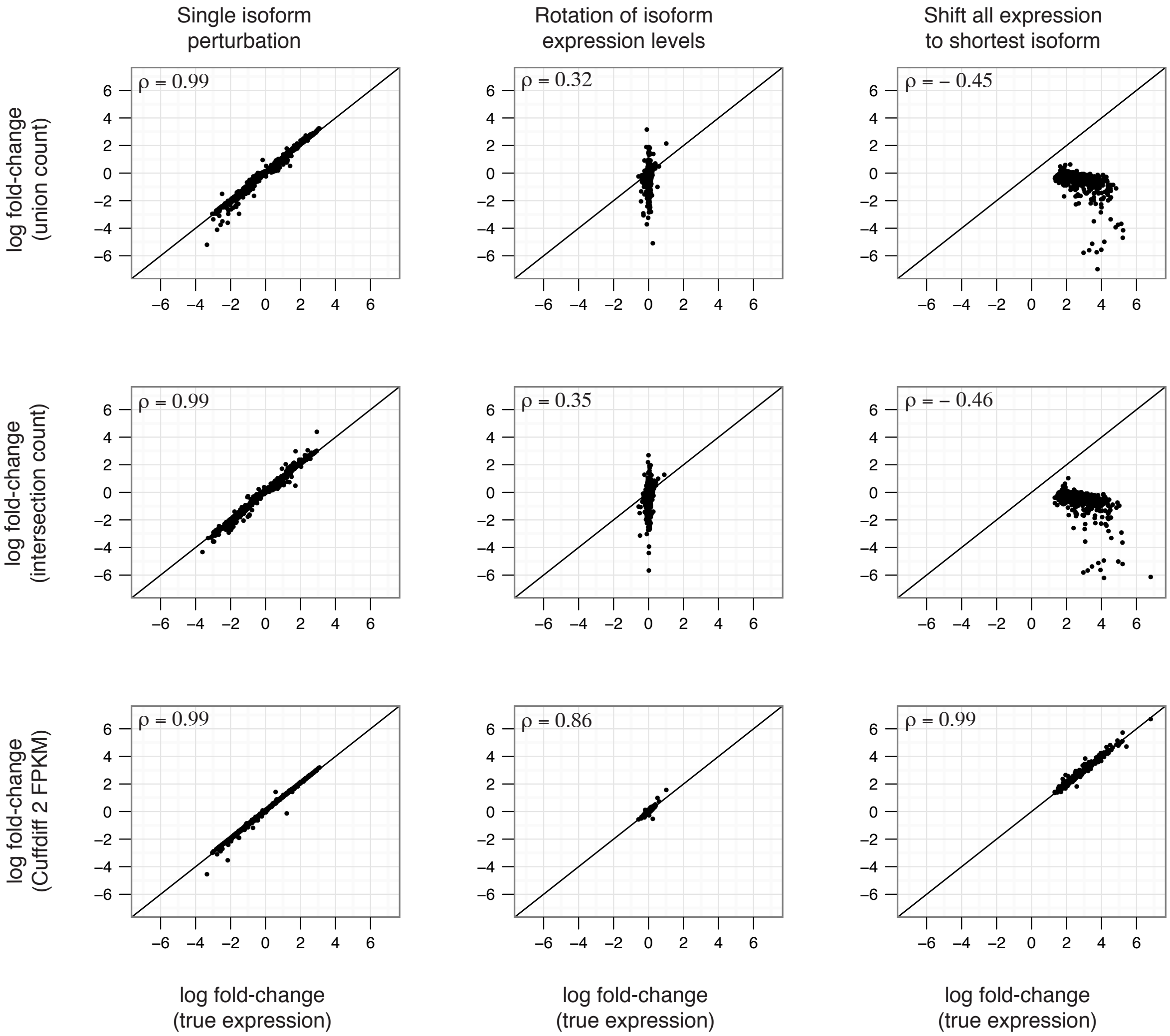

FIGURE 1. Simulation experiments illustrating the discrepancy between change in gene-level fragment count and change in gene expression. Each column of plots illustrates a different perturbation scenario (see Table 1). Left column: MAJOR-ISO, middle column: ROTATE, right column: SHORTEST-ISO. The top row of plots shows the fold change for each perturbed gene under the union count model as calculated by HT-Seq/DESeq. The bottom row shows the fold change in expression as estimated by Cuffdiff 2. Under MAJOR-ISO, relative abundance of each perturbed gene's isoforms with respect to one another is largely preserved, and so the union count scheme recovers the fold change in expression accurately. However, for the other two perturbation scenarios, there is substantial isoform switching, which drives changes in overall gene fragment count that do not reflect changes in expression.

expression. Due to the limited data on the prevalence of this phenomenon in real transcriptomes, it is difficult to assess the impact of the discrepancy on real experiments. We thus performed a simulation study using the human transcriptome and three different isoform switching scenarios, described in Table 1. Under the ROTATE scenario, the transcript abundances for a perturbed gene are simply permuted, leaving the total gene expression unchanged. The SHORTEST-ISO scenario transfers all of the fragments in a gene to the shortest isoform, resulting in a change in the expression of the gene and all of its expressed isoforms. MAJOR-ISO selects a random isoform of a gene and up- or down-regulates it, but leaves the others unchanged, producing a modest shift in gene expression and a mild isoform-switching effect. We then quantified the fold change in gene expression using Cuffdiff 2, and the fold change in gene-level counts using the HT-Seq and DESeq packages, and compared each to the true change in expression. The discrepancy between the change in count and the change in gene expression is shown in Figure 1. Under the MAJOR-ISO scenario, the discrepancy is very mild, but for the other two scenarios it is severe. Under the ROTATE scenario, the permuting of expression levels among transcripts of varying lengths induces a change in the total number of fragments generated from the gene, leading to the incorrect view that the genes are differentially expressed. Under the SHORTEST-ISO, the concentration of fragments on the shortest isoform results in an upregulation of gene expression without an increase in the number of fragments originating from the gene.

In contrast, the estimated expression levels produced by Cuffdiff 2 are highly concordant with the true changes in all three scenarios. The simulations results are corroborated by a comparison of naive counting and Cuffdiff 2 estimates on real expression data obtained by both sequencing and microarray technology (Figure 4). Taken together, these results highlight a fundamental weakness in the approach of estimating gene expression through raw fragment counts: (probabilistic) assignment of ambiguously mapped fragments and estimation of individual transcript abundance levels is crucial for estimating accurately the fold change even of *gene* expression levels.

2. Statistical methods

2.1. **Library size normalization.** In RNA-Seq the expression measurement for a transcript is a function of the number of fragments sequenced from that transcript. However, a sequencing instrument produces a variable number of sequencing fragments in each run. Thus, it is imperative to control for library size and sequencing depth when comparing the number of fragments sequenced from a transcript or a set of transcripts. Several methods have been proposed for this purpose. The first and most basic method is to simply divide each transcript's fragment count from each library by the total number of fragments sequenced from that library [11]. However, expression levels calculated by this method may suffer from bias in some circumstances because RNA-Seq read counts provide only a relative abundance measure. Consider two libraries with an identical number of fragments, and where all genes but one are equally expressed in both libraries. If the differentially expressed gene is also very highly expressed (and thus generates a large fraction of the fragments in each library), all of the other genes in the experiment will exhibit a change in their fragment count. Excluding the genes in the upper quartile of the fragment count distribution from the total library size reduces this effect [2]. Additional sources of bias stem from differences in the lower tail of the count distribution. These can be handled robustly by methods from Anders and Huber [1] or Robinson and Oshlack [13].

We implemented the geometric normalization method of [1], the quartile normalization method, and the traditional total count normalization as options selectable by the user, with the geometric normalization enabled by default in Cuffdiff 2. The geometric normalization method works as follows: in the notation of [1] the fragment count for each gene i in each library j is computed and stored in a matrix k. The size of library j is calculated as:

$$s_j = \text{median}_i \frac{k_{ij}}{\left(\prod_{\nu=1}^{m} k_{i\nu}\right)^{\frac{1}{m}}} \tag{2.1}$$

That is, for each gene, we compute the geometric mean of the number of fragments for that gene across the libraries and a library's size is taken as the median of the ratios of raw counts to the geometric mean. We thus transform the values k_{ij} to a count scale that is common to all libraries and controls for differences in library sizes and biases caused by differential expression at the tails of the count distributions. This transformation simply involves dividing each value k_{ij} by the appropriate size factor s_j.

It is important to note that the final normalized count values for a library depend on the other libraries used to produce the common scale. For reasons that will become clear below, we first apply the geometric normalization to each group of replicates for a given condition, producing an "internal" common count scale for each condition. We then compute the arithmetic mean of these scaled counts for each gene and use them as input for a second round of geometric mean normalization used when comparing conditions to one another. The internal scale is used whenever Cuffdiff 2 compares replicates of a condition to each other. The "external" scale is used when comparing conditions to each other. We denote the internal library size factor for library j as s_j and the external size factor as η_j.

2.2. **Replicate variability and assignment uncertainty.** An RNA-Seq experiment is inherently random at many levels. First, if biological samples are assayed in repeated experiments, there will be random fluctuations in transcript abundances even when controlling for experimental conditions. This is known as *biological variability.* Second, library construction involves random events, such as fragmentation, so that even two identical samples prepared for sequencing will differ in representative transcript abundances. Finally, with current sequencing technologies fragments from the cDNA library are sequenced at random. The latter two effects are typically termed *technical variability*, and together all the effects can be termed *replicate variability.* Thus, one can associate random variables $\{C_t\}_{t \in T}$ to all transcripts from a transcriptome T that represent the (random) number of fragments generated from each transcript in a given experiment. In what follows we assume that if in a sample the abundance of a transcript t is ρ_t and the total number of fragments generated is N, then the random variables C_t are functions of ρ_t and N.

The most natural and simplest model for RNA-Seq is Poisson. With such a model, the C_t are Poisson distributed so that

$$\mathbb{P}(C_t = k) = \frac{e^{\rho_t N}(\rho_t N)^k}{k!}. \tag{2.2}$$

Assuming that the number of fragments sequenced in each experiment is constant, Poisson distributions for each C_t arise naturally if one assumes that fragments are generated according to a multinomial model. Furthermore, if one assumes that the origin of fragments produced in an experiment can be uniquely determined, then the parameters for the Poisson distributions associated to transcripts can be easily estimated from the "fragment counts" which are sufficient statistics. Formally, if x_t is the number of observed fragments originating from transcript t in an experiment then the maximum likelihood estimate for the rate parameter for transcript t is $\hat{\rho}_t = \frac{x_t}{N}$. This formula can be easily adapted to the case when replicates have been performed, in which case the rate parameter is estimated as the average of the fragment counts for the transcript t. Differential analysis with the Poisson model is thus reduced to testing for the difference of two Poisson distributions [2].

Unfortunately, in practice, the C_t are not observed random variables. This is because of *assignment uncertainty*, namely the fact that many fragments map ambiguously to different transcripts. Formally, let F be the set of all fragments. The observed data are, for each $f \in F$, sets $S_f \subseteq 2^T$ where 2^T denotes all the possible subsets of transcripts. We let $\mathbf{S} = \{S_f\}_{f \in F}$. In a given experiment all that one can hope to recover are posterior distributions $Y_t = p(C_t|\mathbf{S})$ where $\mathbf{S}$ are the observed fragment mappings.

Thus, although we continue to assume that fragments are generated according to Poisson models, i.e., the C_t are Poisson random variables, for the purpose of differential analysis we model the observed counts by *mixtures* of Poisson random variables that incorporate our uncertainty in estimates of the underlying rates where the mixtures are given by

$$\sum_t \mathbb{P}(C_t = x_t|\mathbf{S}) \frac{e^{-x_t} x_t^k}{k!}. \tag{2.3}$$

The summation in Equation (2.3) does not necessarily result in a simply described distribution. However we note that it is a mixture of Poisson random variables, and it is therefore convenient to approximate the discrete posterior distributions for the Y_t with continuous Gamma distributions so that we obtain Gamma-Poisson mixtures. Specifically, if m_t is the first moment of the distribution Y_t and ψ_t the second central moment then we fit the posterior distribution for Y_t by a Gamma distribution $\Gamma(r, \theta)$ where

$$r_t \theta \quad = \quad m_t, \qquad\qquad r_t \theta^2 \quad = \quad \psi_t. \tag{2.4}$$

Fitting the moments we obtain

$$r_t \quad = \quad \frac{m_t^2}{\psi_t}, \qquad\qquad \theta_t \quad = \quad \frac{\psi_t}{m_t}. \tag{2.5}$$

The continuous mixture of Poisson random variables where the mixing rates are gamma distributed is a negative binomial distribution which we parametrize by $NB(r_t, p_t)$ where r_t is as above and

$$p_t = \frac{1}{1 + \theta_t} = \frac{m_t}{m_t + \psi_t} \tag{2.6}$$

The mean of $NB(r, p)$ is $\frac{r(1-p)}{p}$ and the variance is $\frac{r(1-p)}{p^2}$. Therefore, we have that the mean of the mixture is m_t, and the variance is $m_t + \psi_t$.

This shows that the negative binomial distribution, although useful for modeling *biological variability* [5], can serve a different purpose in RNA-Seq. As we have shown above it emerges naturally when accounting for the *uncertainty in fragment counts* due to ambiguously mapping fragments even when biological variability is not explicitly modeled. In the following section we describe how to extend the approach outlined above to the case where biological variability is also modeled. Specifically, in following [1], we review how the negative binomial distribution provides a better model for counts that captures biological variability (Section 2.2). For completeness, in Section 2.3 we describe in more detail how we estimate the posterior distributions of counts and their moments. We explain the connection to isoform abundance estimation algorithms such as [7, 8, 15, 12] and describe an approach to obtaining the moments of the posterior distributions of counts that avoids the need for computing the distributions explicitly. Then, in Section 2.4, we show how to incorporate assignment uncertainty in this more general setting. Instead of modeling mixtures of Poisson distributions with negative binomial distributions, we are lead to mixtures of negative binomial distributions that we model with beta negative binomial distributions.

2.3. **Modeling cross-replicate variability in transcript-level counts.** In this section we generalize the previous section to the case when we assume that the probability distributions C_t are negative binomial rather than Poisson. Since the Poisson distribution is a specific instance of a negative binomial distribution, the results of the previous section reduce to a special case of methods at we describe in this section. The use of the negative binomial distribution is motivated by the overdispersion in counts observed in experiments, and that can be attributed to biological variability.

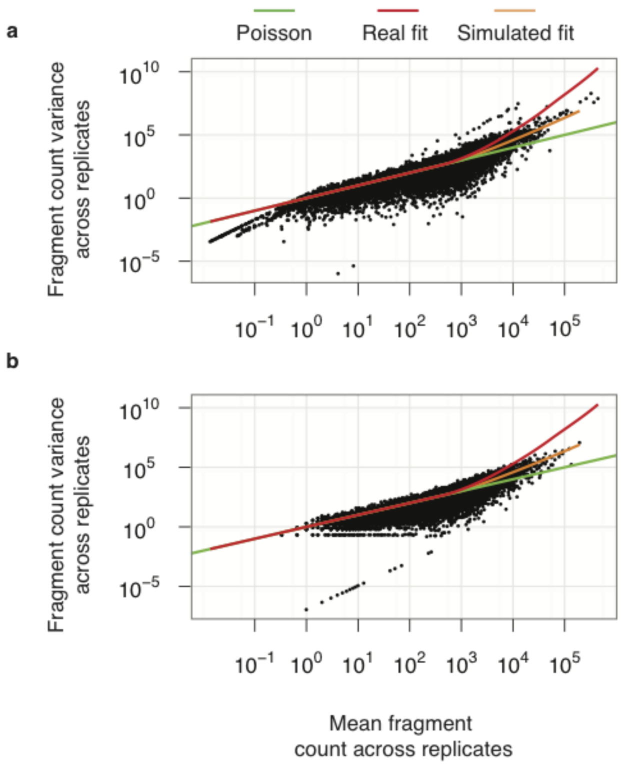

FIGURE 2. Simulated and real fragment count overdispersion, based on three replicates each.

Thus, we adopt the approach of [1] in assuming that the number of fragments x_t^j in replicate j from a transcript t can be modeled by a negative binomial distribution:

$$x_t^j \sim NB(\mu_t^j, {\sigma_t^j}^2). \tag{2.7}$$

To estimate the variances for these distributions, we mimic the procedure in [1] and fit a generalized linear model (GLM) of the gamma family through the cross-replicate (mean, variance) pairs for the gene-level scaled fragment counts computed during the inspection phase of the algorithm. This GLM is used to parametrize the negative binomial distribution that models each gene's cross-replicate fragment count.

The GLM fit is obtained using the local regression package LOCFIT [9], from which we obtain a function $V_E(N)$ for each condition E that yields the predicted variance across replicates in fragment counts given a value for the mean counts N across replicates. Figure 2 shows the extent of overdispersion in our own dataset (panel a) and our ability to simulate a similar pattern of overdispersion (panel b). The green lines illustrate the extent of dispersion expected with the Poisson model (where the variance equals the mean), and the red and orange curves are the functional fit we produce for the real and simulated data respectively using the procedure outlined above.

When a condition is represented by only a single replicate, the variability model is estimated from a different condition that is represented by more than one replicate. Each single-replicate condition uses the variance model borrowed from the condition with the most replicates. When all conditions have only a single replicate, we conservatively treat the conditions as replicates of each other [1]. In such a scenario, the implicit assumption is that most genes are not differentially expressed.

2.4. **Estimating fragment count distributions.** We use the likelihood function originally proposed by Trapnell *et al.*, [15] with the improvements introduced by Roberts et al. [14] in order to estimate fragment counts of individual isoforms. Transcript abundances are deconvolved for a condition by pooling the cDNA fragment alignments from all replicates into a single computation, with their contributions to the full likelihood scaled appropriately by the library depth factors introduced above. We also utilize the "rescue" strategy of Mortazavi *et al.* [11] to account for fragments that map to multiple to multiple loci.

Together, this corresponds to fragment count estimation based on maximum likelihood estimation from the complete likelihood function:

$$L(\rho|\mathbf{S}) \quad = \quad \left(\prod_{g\in G} \beta_g^{X_g}\right)\left(\prod_{g\in G}\prod_{f\in F_g}\sum_{t\in g}\gamma_t\cdot Q(f,t)\right) \tag{2.8}$$

where $\mathbf{S}$ are the fragment alignments as described above, and $\rho = \{\rho_t\}_{t\in T}$ are the transcript relative abundances. As in [15], the model partitions the fragments and transcripts into non-overlapping loci $G = g_1, ..., g_k$, and separates each transcript-level abundance parameter ρ_t into two parameters β_g and γ_t, such that $\rho_t = \beta_g\gamma_t$. The parameter β_g is simply the probability that a fragment drawn at random from the library falls in the locus g. The parameter γ_t is the probability that a fragment from locus g containing the set of transcripts T_g originated from transcript $t \in T_g$. In the above equation, $F_g \subseteq F$ is the set of fragments that map to locus g, and $Q(f,t)$ is a constant that incorporates a number of per-fragment normalizing effects modeled by the algorithm. It is defined as:

$$Q(f,t) = \frac{1}{s_f}\cdot m_f \cdot \frac{b(t, e_{5'}(t,f), e_{3'}(t,f))}{B(t, I_t(f))} \tag{2.9}$$

where s_f is internal size factor for the fragment's library and $(Qf,t) = 0$ if $t \notin S_f$.

The term $\frac{b(t,e_{5'}(t,f),e_{3'}(t,f))}{B(t,I_t(f))}$ captures the fragment bias parameters introduced by Roberts *et al.* [14]. In that work, Cufflinks was used to quantify transcript expression by iteratively refining bias parameters and transcript abundances to jointly maximize the likelihood function via coordinate ascent. However, because the marginal increase in likelihood dropped dramatically after just two iterations of the ascent, the algorithm terminates there for performance reasons. The scaling constant m_f allows Cuffdiff 2 to implement the "rescue" strategy of [11]: fragments that map to $n > 1$ loci are initially weighted $m_f = \frac{1}{n}$. As with the bias model, m_f is updated after each pass of maximization of the abundance parameters. This is equivalent to updating estimates for transcript abundances taking into account fragments mapping ambiguously to different loci with a single round of the EM algorithm. Finally, $X_g = \sum_{f\in F_g} m_f \cdot \frac{1}{\eta_f}$, where η_f is the external size factor for f. This deconvolution yields for each transcript an estimate of its relative abundance in the sample ρ_t.

As discussed in Section 2.1, incorporation of assignment uncertainty in differential analysis involves the determination of the posterior distributions of counts. To do this, in the notation of [15], we compute γ_t for each transcript by the EM algorithm. Next, we calculate the posterior expectation for the number of fragments originating from each transcript. We treat each fragment as a set of Bernoulli random variables, one for each transcript in S_t, each with success probability:

$$p_f^t \quad = \quad \frac{\gamma_t Q(f,t)}{\sum_{i \in S_f} \gamma_i Q(f,i)} \tag{2.10}$$

The posterior expectation on the fragments assigned to t in locus g can then be computed as $\sum_{f \in F_g} p_f^t$ by the law of iterated expectation. We track the variance and covariance on the posterior distribution for assigned fragments between transcripts i and j in a matrix ψ computed as:

$$\psi_{i,j} = \sum_{f \in F_g} \psi_{f_{i,j}} = \begin{cases} p_f^i(1-p_f^i) & \text{if } i = j \\ \\ -p_f^i p_f^j & \text{if } i \neq j \end{cases} \tag{2.11}$$

2.5. **Mixing negative binomial distributions.** Having captured both the cross-replicate variability and the uncertainty in transcript fragment counts for each replicate due to ambiguously mapped fragments, we are now ready to combine both. We propose that each transcript abundance ρ_t can be modeled by a variable X_t that is beta negative binomially distributed. The beta negative binomial can be interpreted as a mixture of negative binomials. The parameters of the beta negative binomial, r,α,and β, are determined by solving three equations. For notational simplicity, we let $p = \frac{\alpha-1}{\alpha+\beta-1}$ and we set

$$A = X_g \hat{\gamma}_t \tag{2.12}$$

$$B = V_E(X_g) \cdot \hat{\gamma}_t \tag{2.13}$$

$$C = \psi^g_{t,t} \tag{2.14}$$

The equations that must be solved are:

$$\frac{r(1-p)}{p} = A. \tag{2.15}$$

$$\frac{r(1-p)}{p^2} = B. \tag{2.16}$$

$$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{A^4}{B^4} \cdot \frac{C}{r^2}. \tag{2.17}$$

Equations 2.15 and 2.16 define the negative binomial distribution for the mean value of the parameter p that is beta distributed so that it matches the overdispersion estimated for that estimated count value. Equation 2.17 is based on the fact that we would like the mean value of the negative binomial distributions to vary according to the fragment assignment uncertainty captured by $\psi^g_{t,t}$. That is, we would like to have

$$\text{Var}\left[\frac{r(1-p)}{p}\right] = C. \tag{2.18}$$

By Taylor approximation,

$$\text{Var}\left[\frac{r(1-p)}{p}\right] \approx r^2 \left(\frac{B^4}{A^4}\right) Var[p] \tag{2.19}$$

from which we obtain Equation 2.17.

Equations 2.15 and 2.16 can be used together to derive that

$$r = \frac{A^2}{B-A} \tag{2.20}$$

and

$$p = \frac{A}{B}. \tag{2.21}$$

Note that r must be greater than zero, so we are implicitly assuming that $B > A$ which is equivalent to assuming that there is overdispersion. If $B = A$, the cross-replicate variance equals the mean, and we can simply revert to mixing Poisson distributions and use the negative binomial instead of the beta negative binomial. Similarly, if there is no uncertainty in read counts the beta negative binomial distribution for a transcript collapses reduces to a negative binomial distribution, making our method equivalent to previous count-based differential expression approaches when used for transcriptomes without alternative splicing or paralogous genes. We now solve for β as follows. Using the first equation, we have that

$$\alpha = 1 - \frac{A}{A-B} \cdot \beta. \tag{2.22}$$

Plugging this into the Equation 2.17, and using the formula for r, we obtain a cubic equation from which we can solve for β and thus obtain the remaining parameters of the beta negative binomial distribution:

$$-\frac{C}{B}\beta^3 + \left(A + \frac{4AC}{B^2} - \frac{4C}{B}\right)\beta^2 + \left(-A + B - \frac{5A^2C}{B^3} + \frac{10AC}{B^2} - \frac{5C}{B}\right)\beta + \frac{2A^3C}{B^4} - \frac{6A^2C}{B^3} + \frac{6AC}{B^2} - \frac{2C}{B} = 0.$$

The distributions for the fragment counts for each transcript allow us to estimate the significance of observed changes between two or more conditions.

2.6. **Estimating fragment counts for genes and other groups of transcripts.** A posterior distribution on the relative abundance a transcript t will allow us to test for significance in observed changes in the assigned fragment count between two or more conditions. Because we are interested not just in differential analysis transcript-level expression but also in gene-level differential expression analysis, we must estimate the distribution on expression ρ_G for a group of transcripts $G = t_1, ..., t_m$ (e.g. the alternative isoforms of a gene). Given the posterior distributions $X_{t_1}, ..., X_{t_m}$, the total expression for G is

$$\mathrm{E}[\rho_G] = \sum_{t_i \in G} \mathrm{E}[\rho_{t_i}] \approx \sum_{t_i \in G} \frac{1}{M_j l_{t_i}} X_{t_i}. \tag{2.23}$$

This expression estimate for G has variance

$$\mathrm{Var}[\rho_G] \quad = \quad \sum_{t_i, t_j \in G} \frac{1}{M_j^2 l_{t_i} l_{t_j}} \mathrm{Cov}[X_{t_i}, X_{t_j}]. \tag{2.24}$$

Like the variance, covariance in assigned fragment counts between two transcripts t_i t_j is a result both of fragment mapping uncertainty and of true biological covariation across replicates. Our method does not analytically model the latter, which would likely require many replicates. To estimate the covariance in the number of fragments assigned to each transcript, we simulate the generation and subsequent assignment of sequencing fragments under our model. For each transcript, we draw a fragment count from a negative binomially distributed random variable, and then assign those fragments to all the transcripts according to the uncertainty matrix ψ. The negative binomial distribution for a transcript is parametrized such that it has a mean number of fragments equal to our maximum likelihood estimate and variance equal to the cross-replicate fit described in Section 3.3. To assign the fragments, we treat them as random variables drawn from a multinomial distribution parametrized by γ. The assigned count samples are then used to empirically estimate the variance-covariance matrix for fragment count assignment in each gene. By default, 1,000 rounds of fragment generation and assignment are performed for each gene (this is a parameter can be set as an option by the user).

2.7. **Testing for differential expression.** Estimating the variance and covariances for the number of fragments assigned to each transcript allows us to approximate the posterior distribution for the expression of G and thus perform gene-level differential expression analysis. We test for significance of observed changes exactly as in [15]. Briefly, the log-transformed ratio of expression constitutes a test statistic that is follows a standard normal distribution when divided by the variance of the transformed ratio. We perform a two-sided test for significance against a null hypothesis that the ratio is unity (no change). In the case when a gene or transcript has zero fragments in one condition, we perform a one-sided test using the estimated posterior distribution for that gene or transcript directly. This same method is used to test for differential expression at the transcript level.

We are also interested in changes in the *relative* abundances of transcripts from a given group. For example, changes in relative abundance among transcripts that share a common transcription start site (TSS) constitute differential splicing within that group.

We write the discrete distribution of relative transcript abundances as κ. Let $\hat{\kappa}^A, \hat{\kappa}^B$ denote the relative abundances for a set of transcripts in conditions A and B, respectively. Then from the delta method we have that the square root of the Jensen-Shannon divergence $d = \sqrt{JS(\hat{\kappa}^A, \hat{\kappa}^B)}$ has variance approximated by

$$\mathrm{Var}[d] \approx (\bigtriangledown d)^T \Sigma (\bigtriangledown d), \tag{2.25}$$

where Σ is a block-diagonal variance-covariance matrix formed by the individual $\hat{\kappa}$ variance-covariance matrices $\hat{\Sigma}^A$ and $\hat{\Sigma}^B$. When normalized by its variance, this test-statistic follows a Gaussian distribution that is truncated below zero, allowing a one-sided significance test on observed changes in relative abundance. This test was used in single-replicate comparisons in [15], but we found that with fewer than five replicates, Σ has a high error rate and the test consistently produced a higher than expected false discovery rate.

Cuffdiff 2 adopts a sampling-based approach to evaluate observed shifts in isoform abundance against the null hypothesis of no shift. Under the null hypothesis, $\hat{\kappa}^A$ and $\hat{\kappa}^B$ are assumed to be drawn from the same distribution, which we take to be a multivariate normal. First, we generate 100,000 random variates from the multivariate normal $\mathcal{N}(\hat{\kappa}^A, \hat{\Sigma}^A)$. We then compute the sampling distribution of the JS distance taken between these points by computing it on 100,000 random pairs. We can then estimate a $p-$value and reject the null hypothesis that the observed changes in relative abundance among a gene's isoforms actually arose by chance, having been drawn from this sampling distribution. Because $\mathcal{N}(\hat{\kappa}^B, \hat{\Sigma}^B)$ may better estimate the null distribution than $\mathcal{N}(\hat{\kappa}^A, \hat{\Sigma}^A)$, we repeat this procedure using $\hat{\kappa}^A$ and $\hat{\Sigma}^A$. We then take the mean of the two obtained p-values as the value for the observed change in the gene's relative isoform abundances.

3. Assessing Cuffdiff 2 accuracy

This section describes experiments demonstrating the overall accuracy of Cuffdiff 2, including its robustness in situations where count-based methods fail to recover changes in expression.

3.1. **Comparison with expression microarrays.** This section provides comparisons between Cuffdiff 2 fold change estimates across conditions in our HOXA1 knockdown experiment, and independently obtained fold change estimates using microarrays. Note that the same number of replicates (3) were used in each condition in both the arrays and the sequencing.

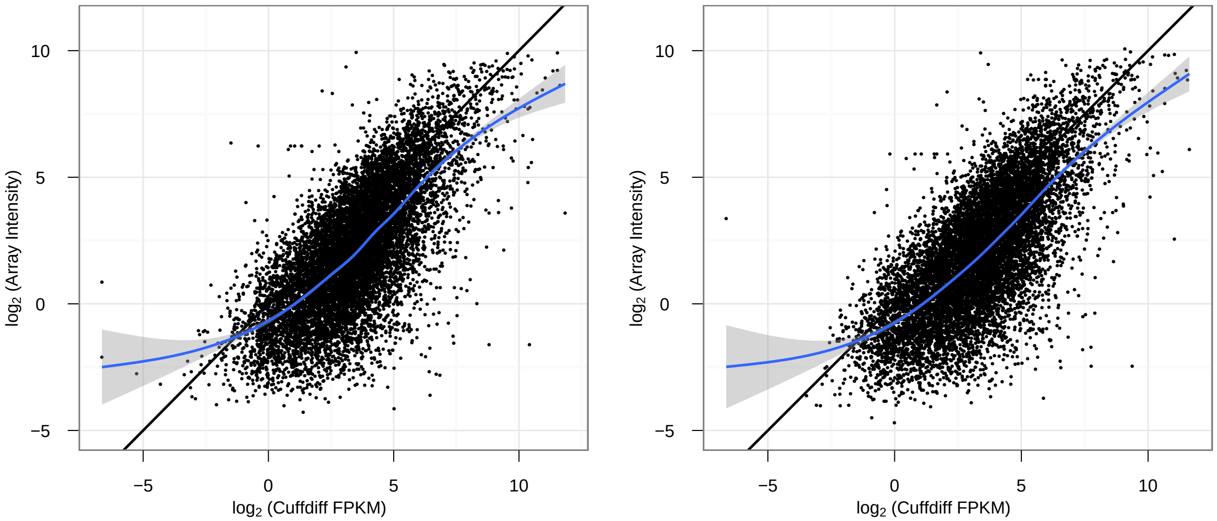

Figure 3. Intensity of Agilent Expression microarray probes compared against Cuffdiff 2-derived FPKM values in the scramble control (left) and the HOXA1 knockdown (right).

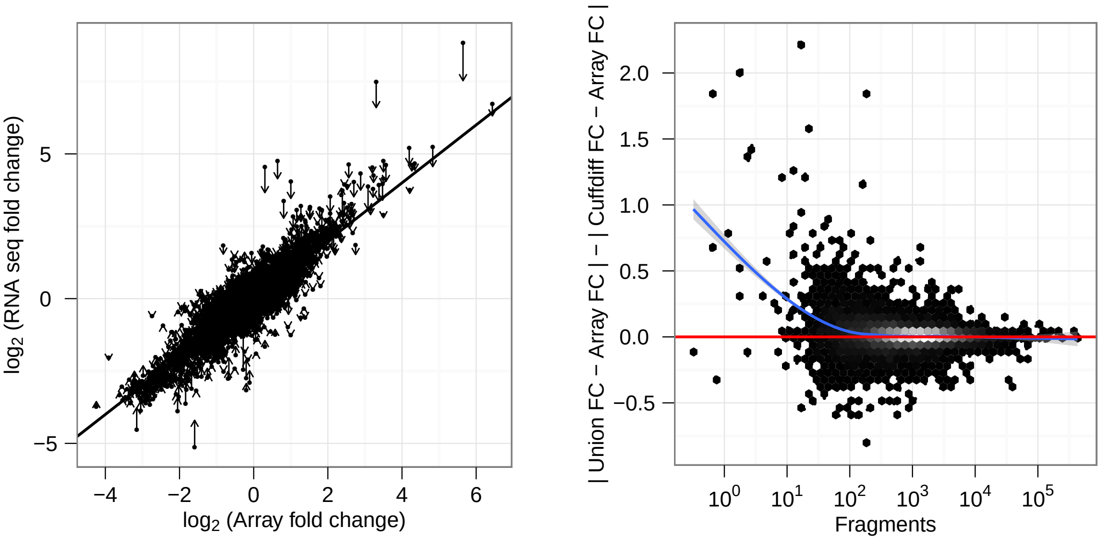

Figure 4. A comparison of Lung fibroblast gene expression change after knockdown of HOXA1 assessed by microarrays and RNA-Seq. (left) The vertical axis shows the change in each gene's expression (shown as an arrow) as calculated by HT-Seq/DESeq (which marks the tail of each arrow) and as calculated by Cuffdiff 2 (which marks the head of each arrow). The solid line indicates perfect concordance between the array measurement and the RNA-Seq measurement. (right) In the fibroblast data, the difference between the fold change estimates by Cuffdiff 2 and union count is worse at lower levels of expression.

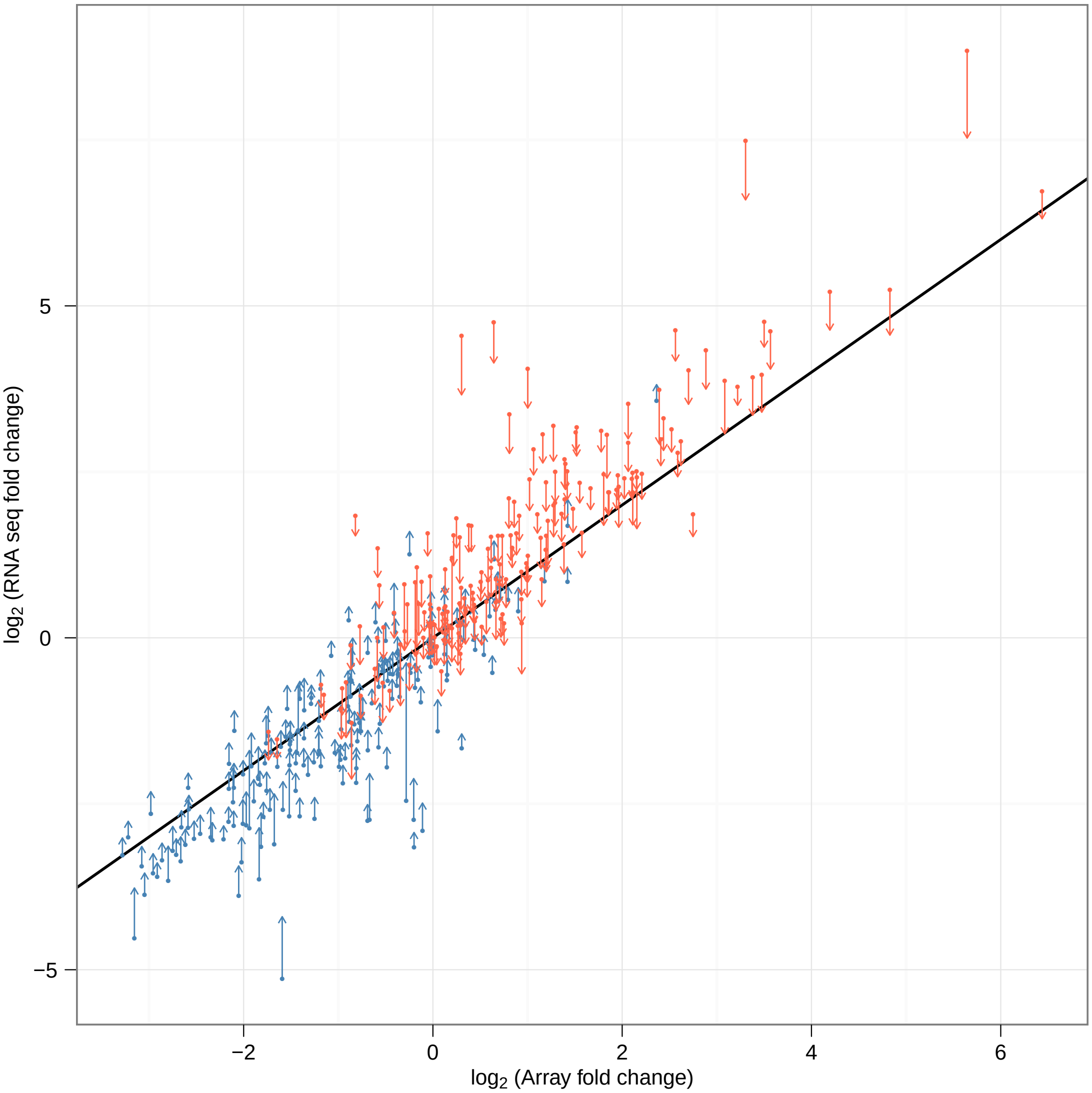

FIGURE 5. Enlarged version of Figure 3b. Gene expression by isoform deconvolution (tail of arrows) instead of by fold-change in gene-level fragment counts (head of arrows) improves agreement with microarrays. Genes shown are those where Cuffdiff and intersection-count fold changes are most discrepant (1% tails).

Nature Biotechnology: doi:10.1038/nbt.2450

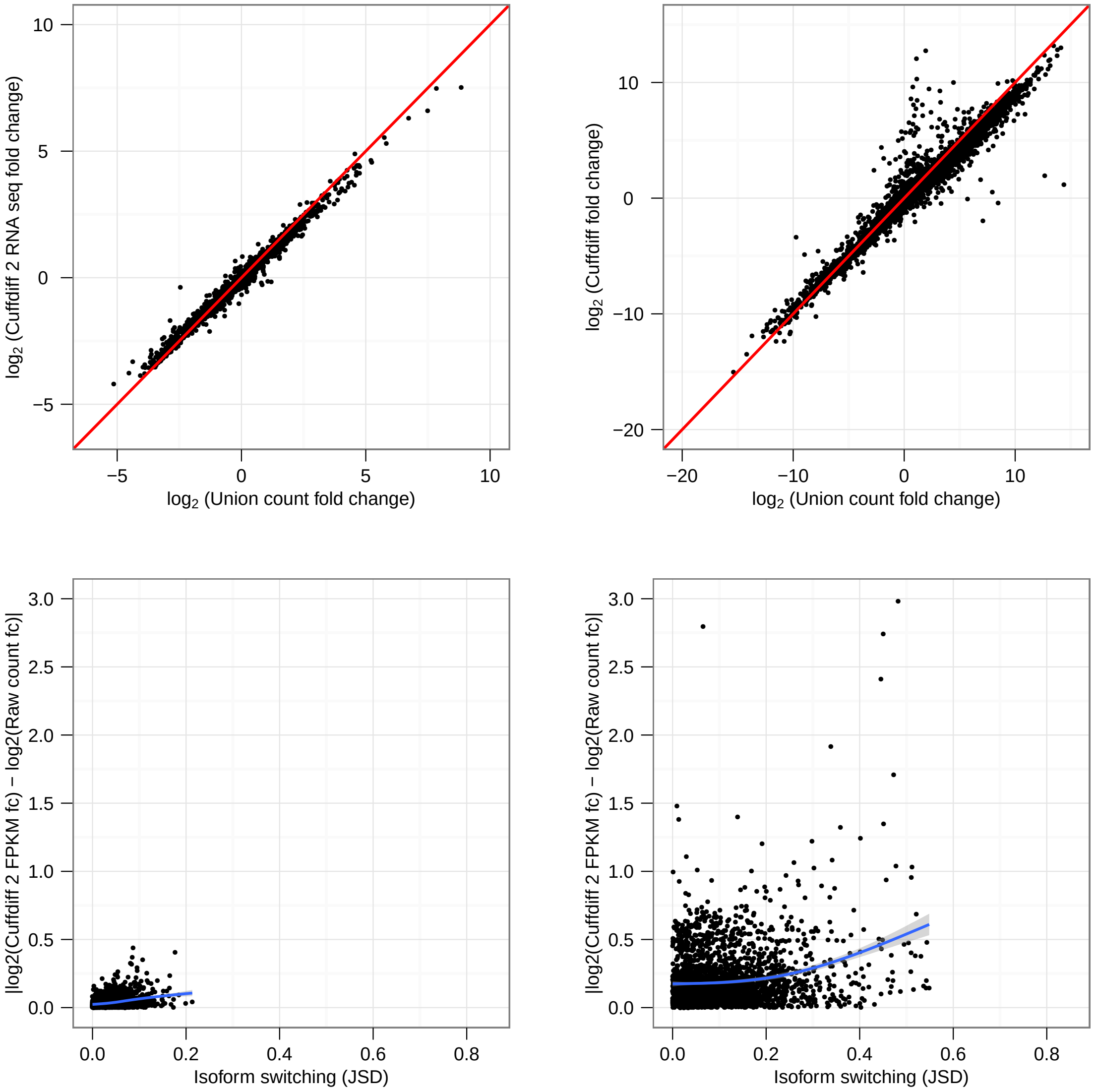

Figure 6. Cuffdiff 2 is highly concordant with other methods for differential analysis. (top left) Fold changes in gene expression in response to loss of HOXA1 in HLFs as measured by Cuffdiff 2 compared against those from the union count method. (top right) Fold changes in gene expression between HLFs and HESCs as derived by Cuffdiff 2 and union counts. However, the methods tend to disagree in genes with substantial changes in relative isoform abundance. Disagreement in fold change versus isoform switching as measured by the Jensen-Shannon distance between hLFs before and after HOXA1 knockdown (bottom left) and between hLFs and hESCs (bottom right). Bottom panels show genes with FPKM > 10 and normalized gene count > 100 to ensure reliable quantification of all isoforms.

3.2. **Comparison with MAQC qPCR data.** The Microarray Quality Control (MAQC) project [10] provided a large sample of gene expression measurements for two commercially available RNA samples, the Universal Human Reference (UHR) and brain (HBR). The MAQC project assayed gene expression for approximately 1,000 genes by both TaqMan qPCR and a variety of microarray platforms, and the MAQC samples were later assayed by RNA-Seq by several groups [2]. Here, we demonstrate that Cuffdiff 2 reports fold changes in expression between the samples that are highly concordant with the qPCR data. Furthermore, Cuffdiff 2 performs a differential analysis of these samples at competitive accuracy with the popular count-based tools DESeq and edgeR.

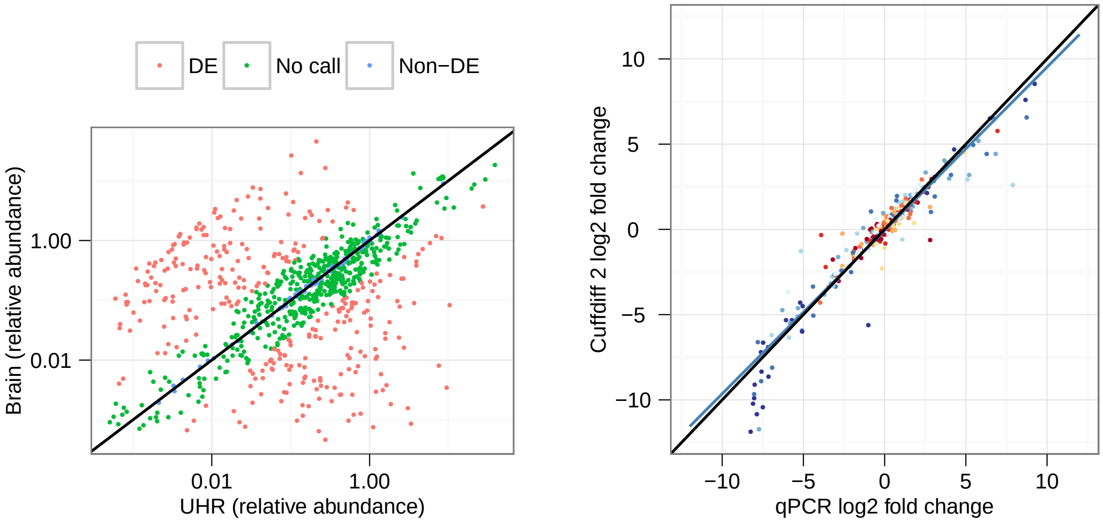

Figure 7. (left) The MAQC standard samples Brain and UHR as assayed by TaqMan qPCR. Following the methodology of [2], genes with a log2 fold change of greater than 2.0 were declared differentially expressed (DE) by qPCR, while genes with a log2 fold change of less than 0.2 were declared not DE. Genes with changes between these thresholds were said to have indeterminate DE status and not included in the true positive and true negative sets. (right) Cuffdiff 2 log2 fold changes in FPKM compared against qPCR fold changes. Plots show genes with two or more isoforms.

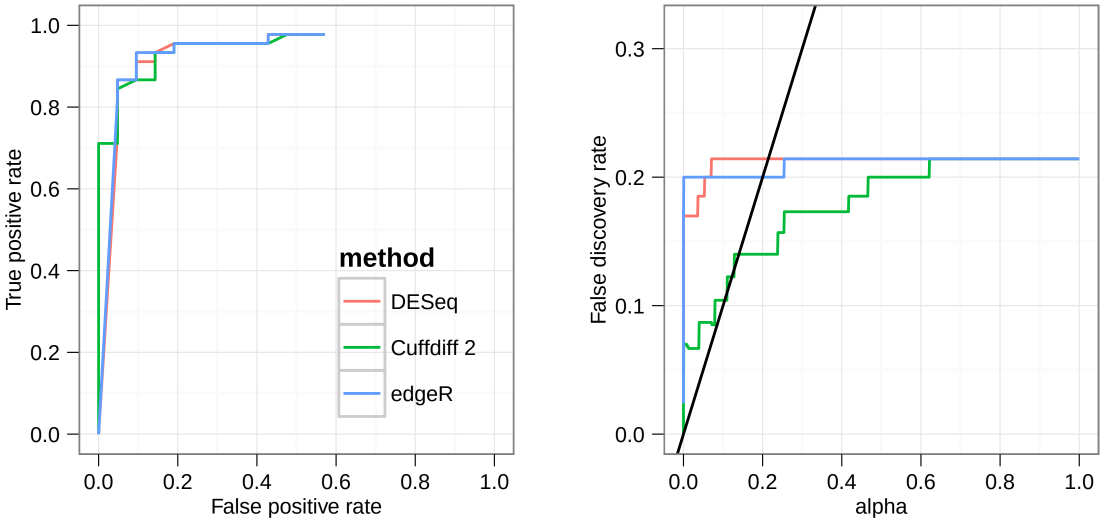

FIGURE 8. (left) ROC curves for Cuffdiff 2 and popular count-based tools for differential gene expression analysis for RNA-Seq. (right) False discovery rate for the same tools as a function of alpha, with the black line indicating the target FDR. Plots show genes with two or more isoforms.

3.3. **Comparison with Griffith qPCR data.** In their paper describing the RNA-Seq analysis platform ALEXA-Seq, Griffith *et al.* [4] analyzed MIP101, a human colorectal cell line sensitive to fluorouracil (5-FU), and compared to a resistant version (MIP/5FU). The authors reported a global disruption of splicing by comparing RNA-Seq read densities in alternative splicing features (e.g. cassette exons or alternative donor sites). The authors validated observed differential splicing events via an extensive qPCR study. Here, we demonstrate that Cuffdiff 2 recovers changes in isoform expression levels from the Griffith RNA-Seq data that are highly concordant with the qPCR data, similar to two independent tools for isoform-level RNA-Seq analysis. RSEM quantifies isoform expression levels using an EM-based approach similar to Cuffdiff 2, and ALEXA-Seq quantifies isoforms by comparing read densities in splicing features that distinguish them. RSEM[8] and Cuffdiff 2 were run on UCSC knownGenes from hg19. ALEXA-Seq values were taken directly from [4].

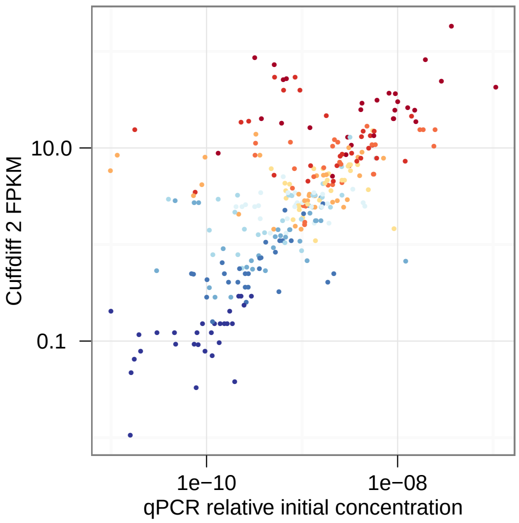

FIGURE 9. Isoform-level expression estimates produced by Cuffdiff 2 versus isoform-specific qPCR data for the MIP101 and MIP5FU datasets. Points are colored by the decile of estimated number of fragments mapped to each transcript.

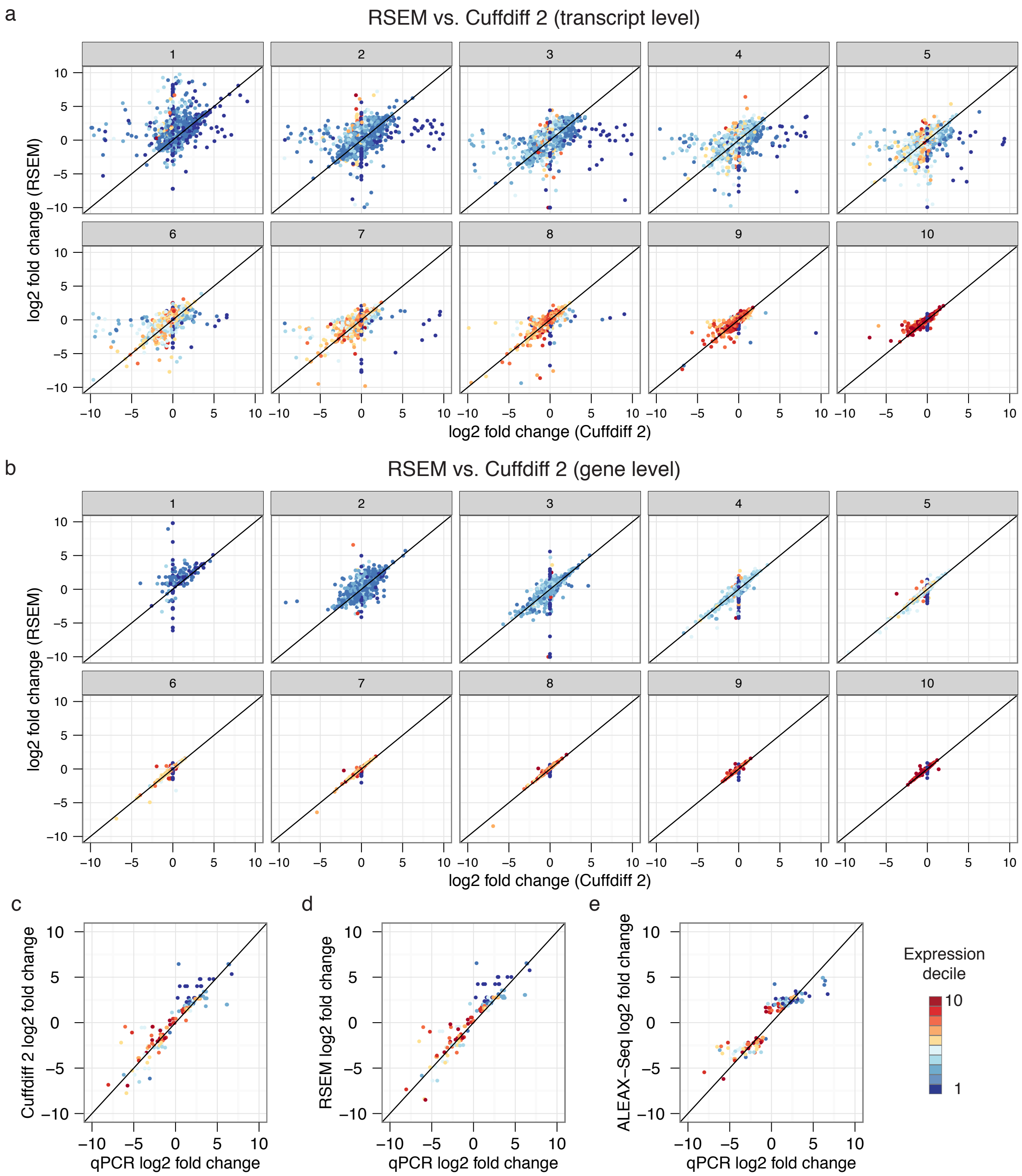

Figure 10. RNA-seq data described in Griffith *et al.* [4] analyzed by RSEM, Cuffdiff 2, and ALEXA-Seq. (a) Concordance of transcript-level log2 fold change in expression between the two Griffith *et al.* samples as measured by RSEM and Cuffdiff 2. The plot is faceted by decile of sequencing depth. (b) Concordance in fold changes at the gene level. Concordance of fold changes against the exon-specific qPCR probes from the Griffith *et al.* data as measured by (c) Cuffdiff 2, (d) RSEM, and (e) ALEXA-Seq. For RSEM and Cuffdiff 2, fold changes were determined by summing the expression of isoforms targeted by each probe and computing fold changes in the sum. These sums were compared against the qPCR-based fold changes for that amplicon. All points colored by expression decile (as determined by Cuffdiff 2 FPKM)

3.4. **Empirical estimation of false discovery rates.** Here, we provide an empirical estimate of the true false discovery rate of Cuffdiff 2 by running it on the MAQC datasets and our fibroblast data in ways that are both consistent and inconsistent with the design goals of these experiments. As expected, Cuffdiff 2 returns very few differentially expressed genes and transcripts, as well as few genes under going differential relative isoform output, in the “nonsense” contrasts. All values produced with a target FDR of 5%.

Group A	Group B	Genes	Genes	Genes	Transcripts	Differential	Promoter	CDS
						splicing	switching	switching
		(DESeq)	(edgeR)	(Cuffdiff 2)	(Cuffdiff 2)	(Cuffdiff 2)	(Cuffdiff 2)	(Cuffdiff 2)
UHR1-7	HBR1-7	14,391	14,905	9,559	7,033	116	228	107
HBR1,UHR2,HBR3, UHR4,HBR5,UHR6	HBR2,UHR3,HBR4, UHR5,HBR6,UHR7	0	0	0	0	0	0	0
HBR4,UHR4,HBR5, UHR6,HBR7,UHR2	HBR1,UHR3,HBR6, UHR1,HBR2,UHR5	0	0	0	0	0	0	0
HiSeq-A1KD1-3	HiSeq-SCR1-3	7,690	5,034	5,862	5,189	66	77	98
HiSeq-A1KD1 HiSeq-SCR2 HiSeq-A1KD3	HiSeq-SCR1 HiSeq-A1KD2 HiSeq-SCR3	1	70	0	0	1	0	3
HiSeq-SCR1 HiSeq-A1KD2 HiSeq-A1KD3	HiSeq-A1KD1 HiSeq-SCR2 HiSeq-SCR3	1	73	0	0	1	1	0
HiSeq-SCR1 HiSeq-SCR2 HiSeq-A1KD3	HiSeq-A1KD1 HiSeq-A1KD2 HiSeq-SCR3	0	0	0	0	3	1	2
MiSeq-A1KD1-3	MiSeq-SCR1-3	5,875	6,805	2,800	2,429	56	21	43
MiSeq-A1KD1 MiSeq-SCR2 MiSeq-A1KD3	MiSeq-SCR1 MiSeq-A1KD2 MiSeq-SCR3	0	0	0	0	0	0	0
MiSeq-SCR1 MiSeq-A1KD2 MiSeq-A1KD3	MiSeq-A1KD1 MiSeq-SCR2 MiSeq-SCR3	0	0	0	0	0	0	0
MiSeq-SCR1 MiSeq-SCR2 MiSeq-A1KD3	MiSeq-A1KD1 MiSeq-A1KD2 MiSeq-SCR3	0	0	77	22	0	0	0

Table 1. Empircal FDR assessment. Columns indicate differentially expressed genes and trancscripts, differentially spliced genes, and those undergoing promoter or CDS switching.

3.5. **Sequencing library simulator.** Simulations were conducted through a set of custom scripts to match as closely as possible the properties of our real lung fibroblast RNA-Seq data. The simulator accepts three input files along with various options and parameters to be used during the run. The first input file is a set of annotated transcripts, which for all of our tests was the UCSC protein coding transcripts annotated in the human genome (hg19). The second input file is an assignment of relative abundances to transcripts in the annotation file. The third file is a table describing cross-replicate variability in fragment counts as a function of abundance. Key parameters for the simulator include the length of reads to generate, the mean cDNA fragment length and its variance, and the number of cDNA fragments to sequence. The simulation framework generates replicate libraries for the input condition and also creates libraries for a "perturbed" condition, where a random selection of transcripts are up- or down-regulated. The simulation framework reports the degree to which each has been modified so that Cuffdiff 2's accuracy can be assessed exactly.

The above simulator produces sequencing libraries as follows. First, the expected number of fragments from each transcript is calculated based on the number of total fragments sequenced in the library F and that transcript's abundance (given in FPKM and expressed here as ρ_t).

$$E[X_t] = \rho_t \left(\frac{\tilde{l}_t}{1000} \right) \cdot \left(\frac{T}{1000000} \right) \tag{3.1}$$

where $\tilde{l}_t$ is the effective length of the transcript. The expected number of fragments from transcript t is used in conjunction with the cross-replicate variability model to specify the actual number of fragments that will originate from t in the simulated library. The input file mentioned above is then used to parametrize a random number generator that returns variates from a negative binomial distribution $N(r, p)$ where

$$\begin{aligned} r &= \frac{E[X_t]^2}{E[X_t] - V_E(X_t)} \\ p &= \frac{E[X_t]}{V_E(X_t)} \end{aligned}$$

and $V_E(X_t)$ is the predicted cross-replicate variance in fragment counts for a transcript with mean fragment counts X_t in condition E. The simulator uses a variate returned by the generator as the actual number of fragments (appropriately scaled by total library size) generated for a transcript t in the simulated library.

In each library, cDNA fragments are generated fragment start position uniformly at random within the transcript's effective length, then selecting a fragment length from a normal distribution with mean of 180bp and variance of 50bp, similar to that in our real fibroblast data. For all test except the read length series described below, two 100bp reads for each fragment were then recorded into an alignment file readable by Cuffdiff 2. We thus base the simulation analysis that follows on a set of perfectly aligned reads, removing the confounding effects of errors in mapping fragments to the genome on differential expression accuracy.

The simulator generates a user-specified number of replicate libraries for the control condition libraries as specified above. To generate libraries for the perturbed condition, the underlying expression levels ρ_t for each transcript are modified before simulated sequencing in one of several ways. The user chooses the number of genes that are to be selected for perturbation. The user then chooses how the genes will be perturbed. For example, all isoforms of a gene may be multiplied by the same up- or down-regulation constant. Alternatively, one transcript of an alternatively spliced gene may be altered, while the others remain unperturbed. The simulation framework supports the perturbation modes listed in Table 1.

UNIFORM	Expression for all transcripts of a gene are multiplied by the same amount
ROTATE	Expression levels are permuted among transcripts
SHORTEST-ISO	Expression for all isoforms except the shortest are assigned to be zero. The shortest isoform's expression is set to the sum of the gene's isoforms control condition expression level.
MAJOR-ISO	A single isoform is selected at random from the gene and multiplied up or down by a constant $c \in [a, b]$, chosen at random. The direction of regulation is chosen at random. The user can specify a and b, and we use defaults of $a = 2, b = 10$.
MULTI-ISO	All isoforms of a gene are perturbed (independently) as described in MAJOR-ISO
SPECIALIZE	The fragments of a gene are all assigned to a randomly selected isoform, and all other isoforms are assigned zero expression. The selected isoform's expression is recalculated as the (normalized) gene fragment count divided by the selected transcript's length, resulting in change in gene expression but minimal change in gene-level fragment count.

Table 2. Simulator modes discussed in this study.

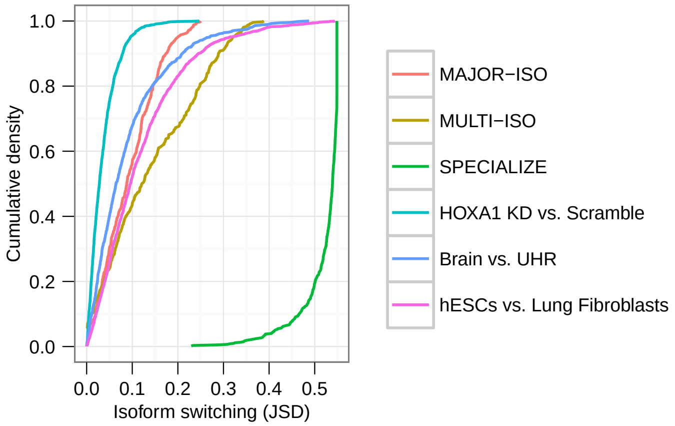

Figure 11. Cumulative densities of per-gene isoform switching observed in the contrasts performed in this paper. Isoform switching was measured as the Jensen-Shannon distance between the relative abundances of isoforms of each gene as estimated in the conditions in the contrast.

Once libraries were generated for each condition, the simulated alignment files were provided to Cuffdiff 2 to assess the software's accuracy in calling differentially expressed and regulated genes and transcripts.

3.6. **Assessment.** We use three metrics to assess performance:

Precision (P)	Fraction of genes Cuffdiff 2 calls significant that were truly perturbed in the simulation
Recall (R)	Fraction of genes that were perturbed in the simulation that Cuffdiff 2 calls significant

Table 3. Metrics for assessing Cuffdiff 2 accuracy *in silico*.

Each of these statistics is also calculated at the transcript level to assess Cuffdiff 2's performance in calling differentially expressed genes and transcripts independently. Since Cuffdiff 2 also calls differentially spliced TSS groups and identifies genes undergoing promoter switching, the simulation framework calculates precision, recall, and F-score to measure performance of these statistical tests as well.

3.7. **Tests.** The following figures describe the results of our simulation studies:

3.7.1. *Single isoform perturbation.* The MAJOR-ISO scenario perturbs a single isoform (not necessarily the major one) of each gene, resulting in minimal isoform switching.

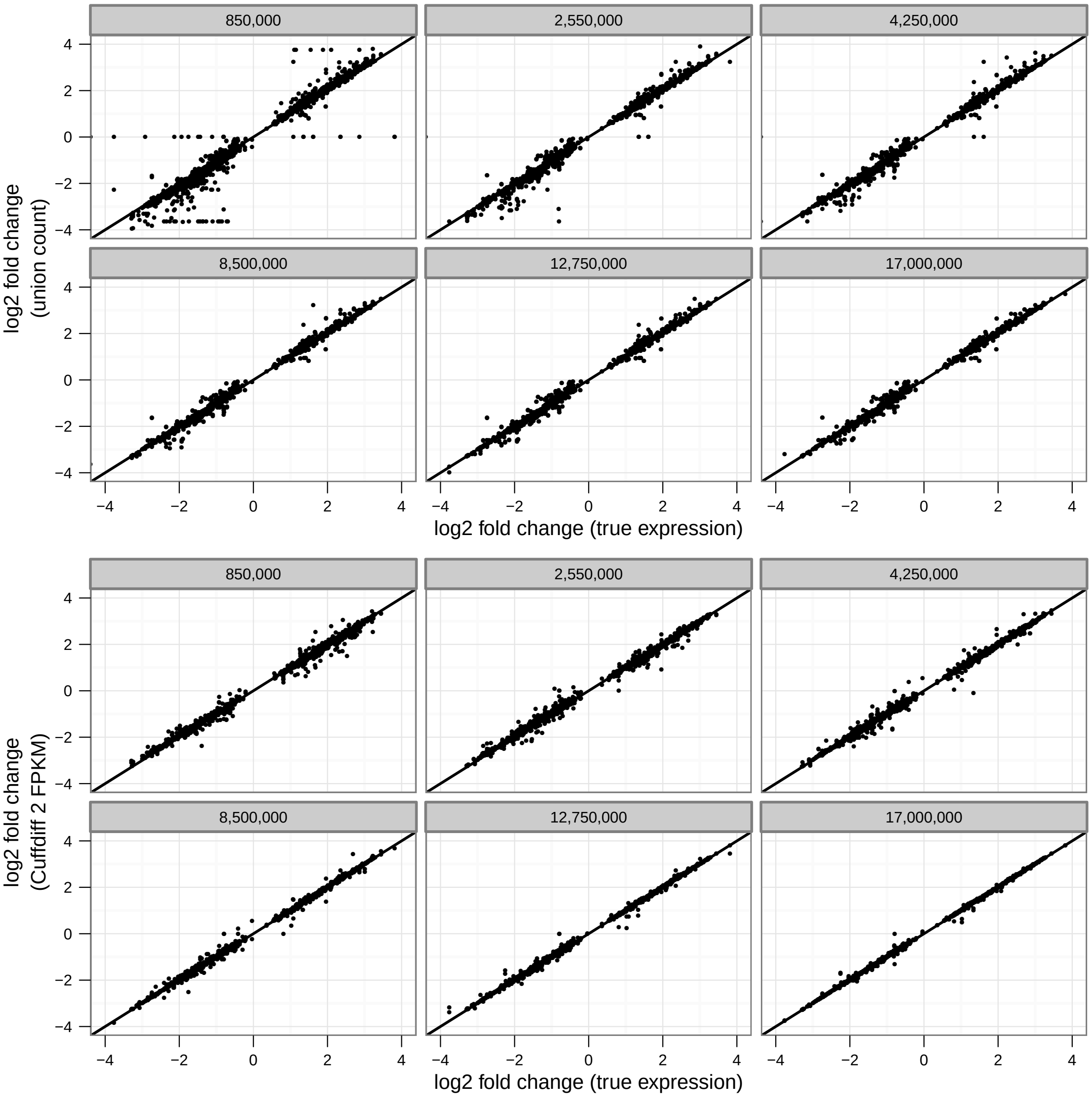

FIGURE 12. minimal isoform switching: Gene-level fold changes in perturbed genes as measured by union count (top) and Cuffdiff 2 FPKM (bottom). Plots are faceted by sequencing depth in the experiment.

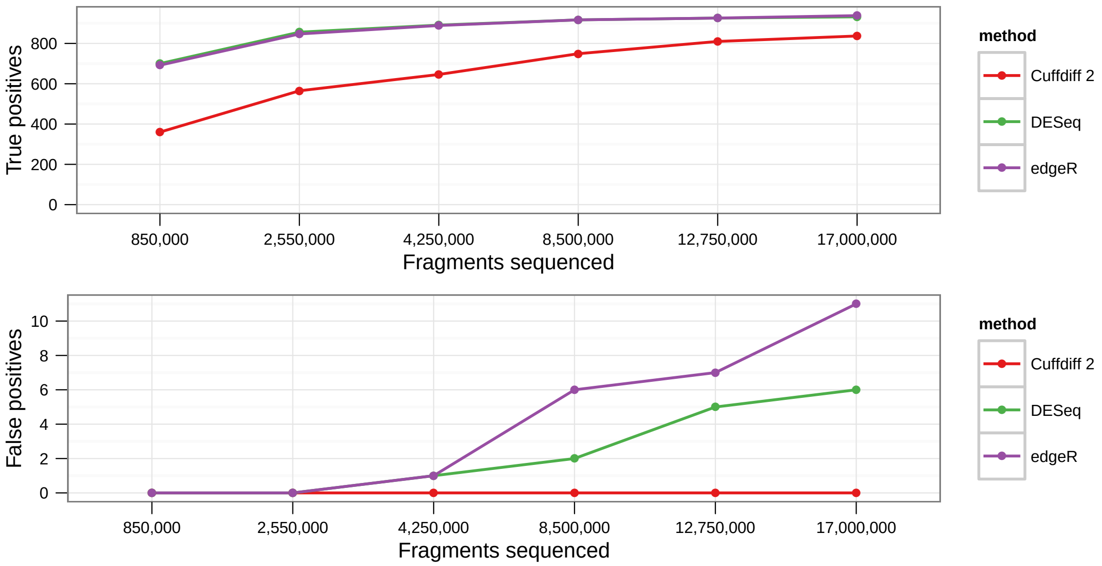

FIGURE 13. Minimal isoform switching: Gene-level true and false positives at different sequencing depths.

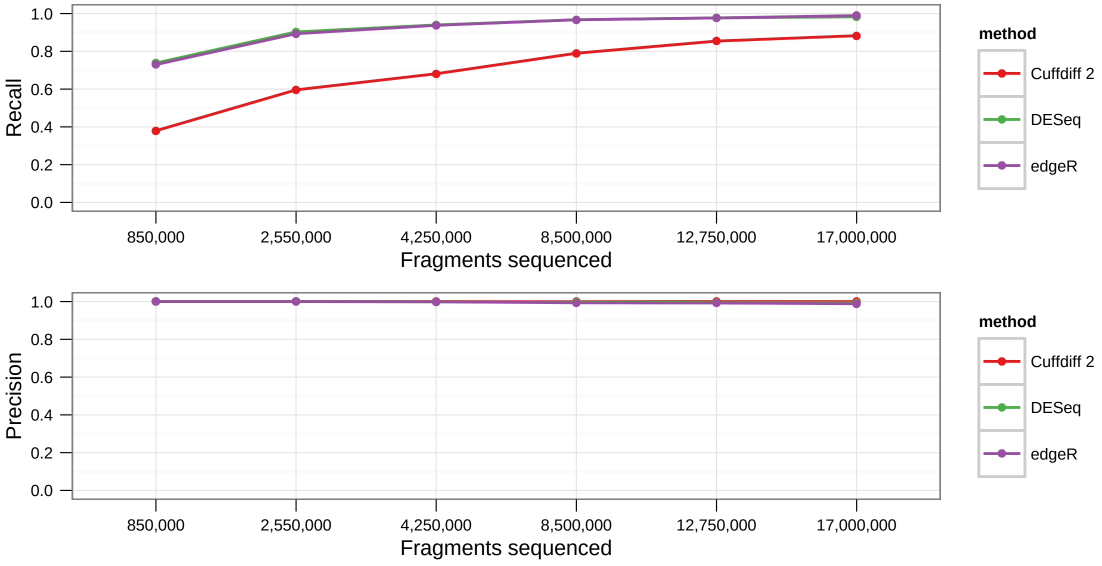

FIGURE 14. Minimal isoform switching: Gene-level precision and recall at different sequencing depths.

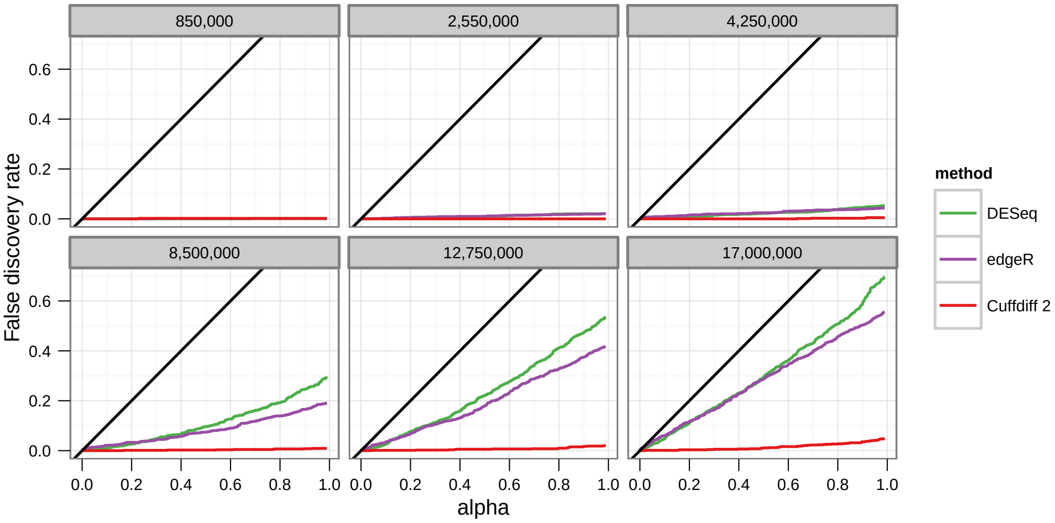

FIGURE 15. Minimal isoform switching: Gene-level false discovery rate as a function of alpha, with the black line indicating the target FDR. Plots are faceted by sequencing depth in the experiment.

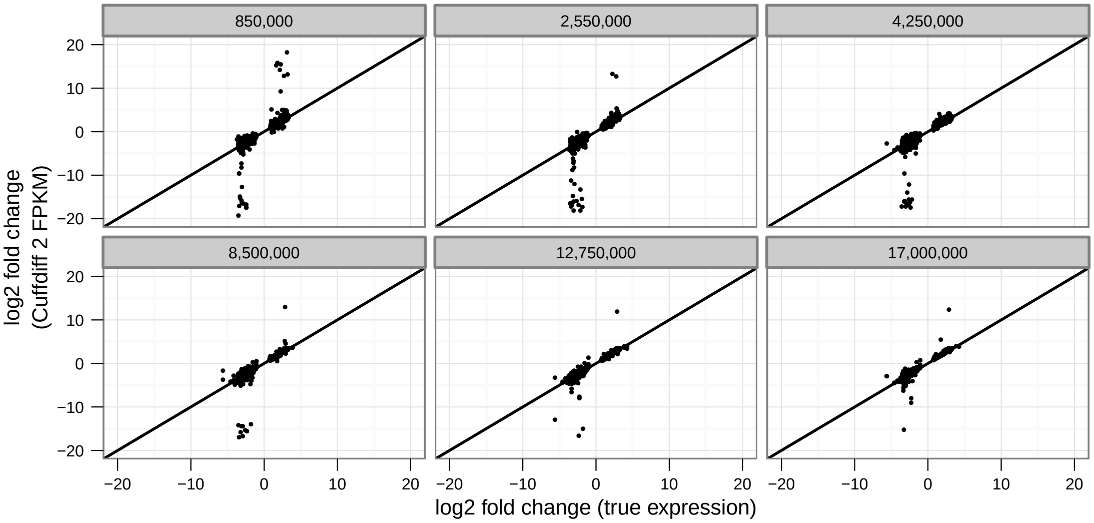

FIGURE 16. Minimal isoform switching: Transcript-level fold changes in perturbed genes as measured by union count (top) and Cuffdiff 2 FPKM (bottom). Plots are faceted by sequencing depth in the experiment.

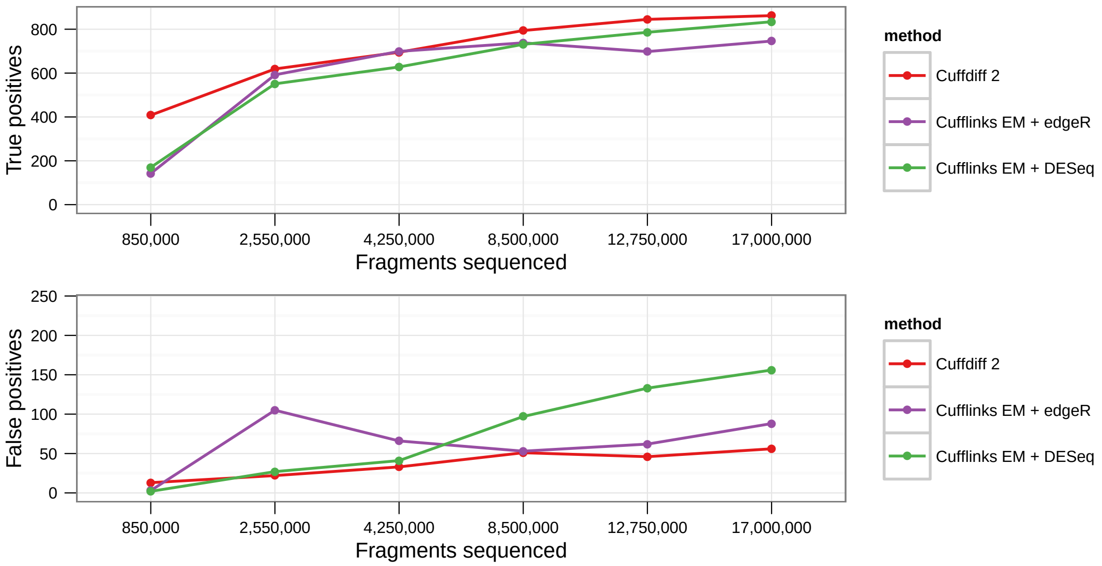

FIGURE 17. Minimal isoform switching: Transcript-level true and false positives at different sequencing depths.

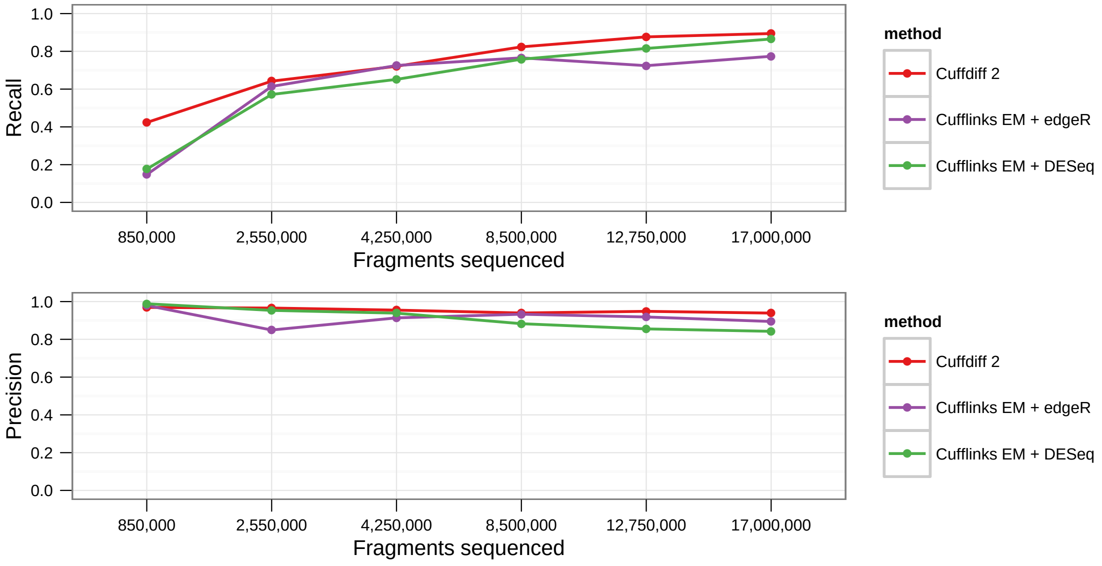

FIGURE 18. Minimal isoform switching: Transcript-level precision and recall positives at different sequencing depths.

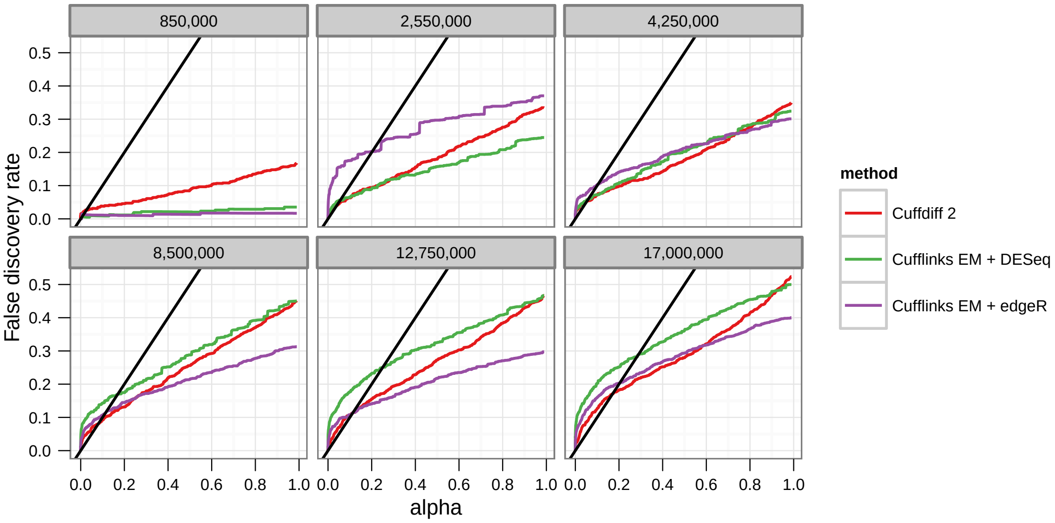

Figure 19. Minimal isoform switching: Transcript-level false discovery rate as a function of alpha, with the black line indicating the target FDR. Plots are faceted by sequencing depth in the experiment.

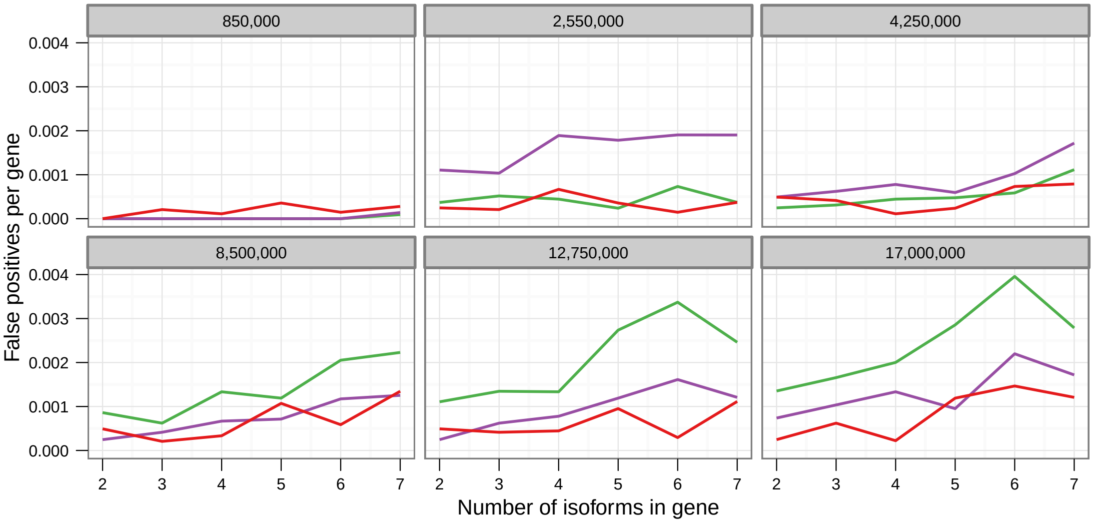

Figure 20. Minimal isoform switching:

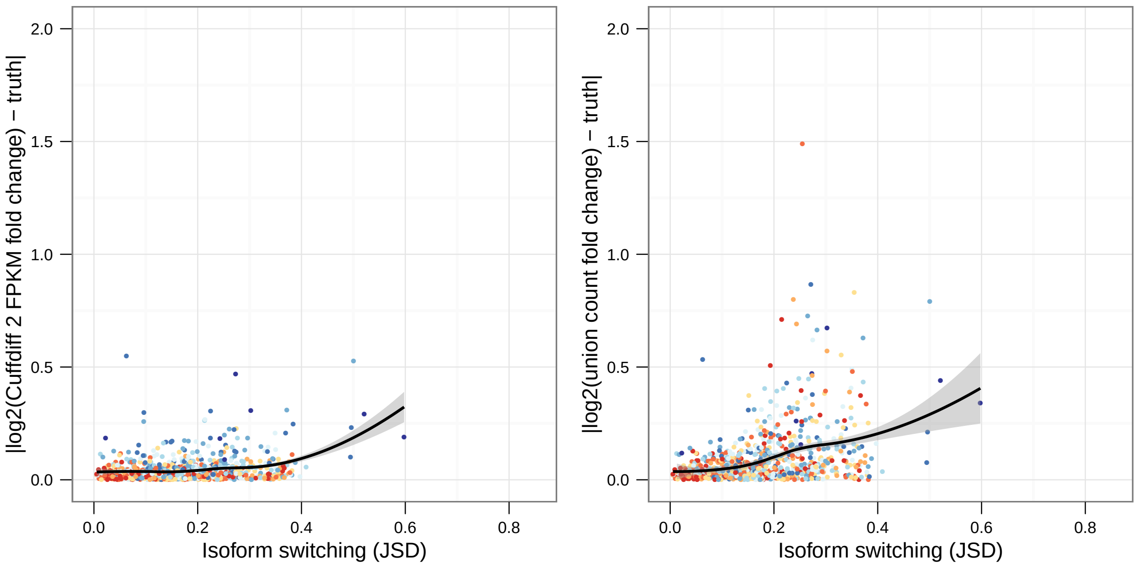

FIGURE 21. Minimal isoform switching: Error in estimation of change in gene expression as a function of isoform switching (measured as the Jensen-Shannon distance between relative isoform abundances) using Cuffdiff 2 (left) or union count (right)

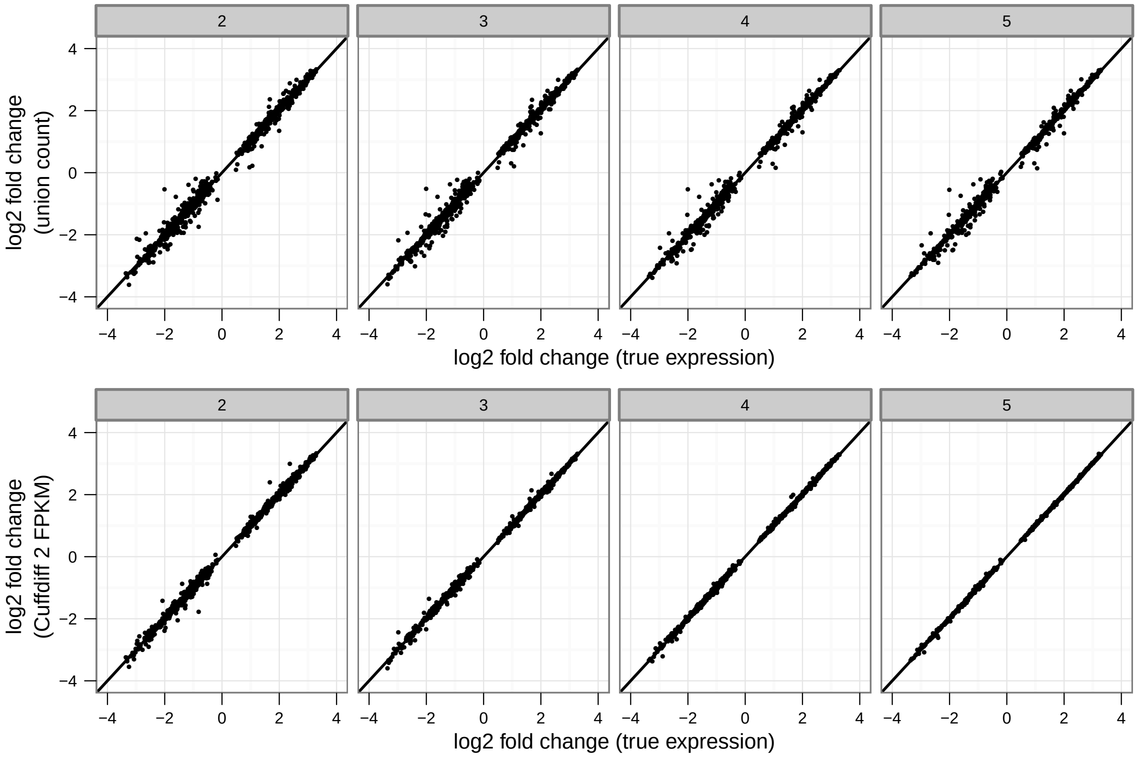

FIGURE 22. Minimal isoform switching: Gene-level fold changes in perturbed genes as measured by union count (top) and Cuffdiff 2 FPKM (bottom). Plots are faceted by sequencing depth in the experiment.

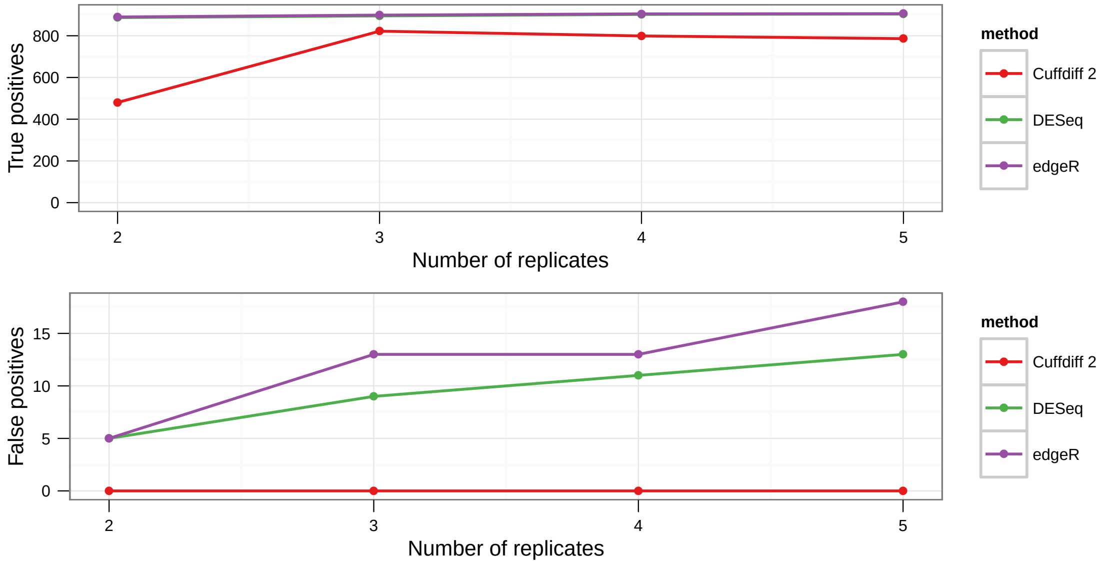

Figure 23. Minimal isoform switching: Gene-level true and false positives at different degrees of replication.

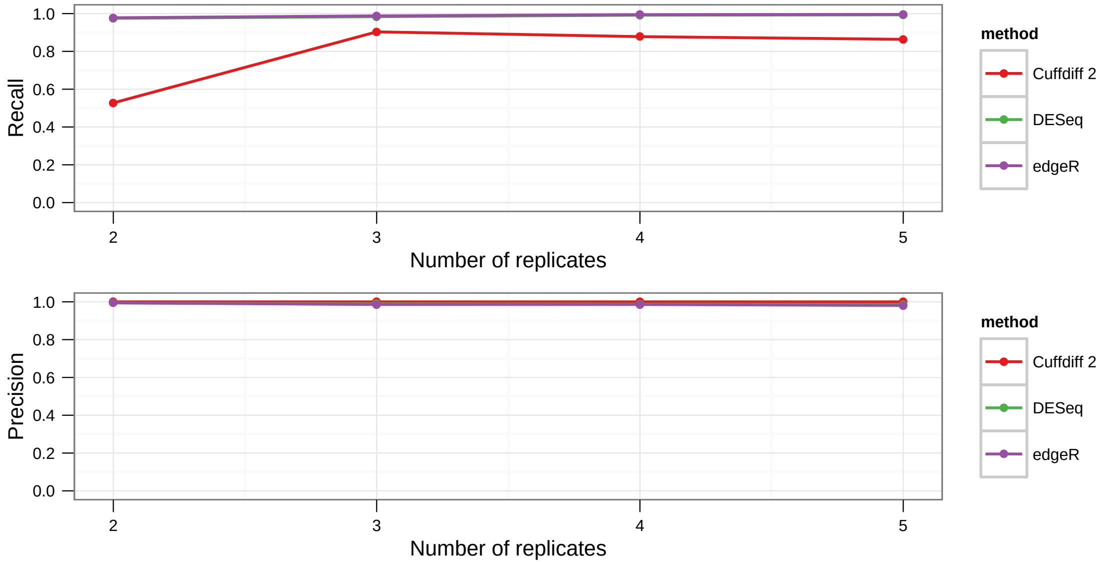

Figure 24. Minimal isoform switching: Gene-level precision and recall at different degrees of replication.

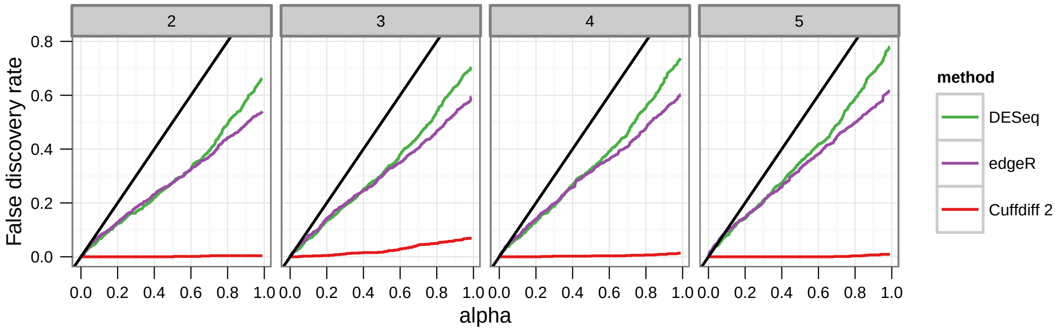

FIGURE 25. Minimal isoform switching: Gene-level false discovery rate as a function of alpha, with the black line indicating the target FDR. Plots are faceted by number of replicates in the experiment.

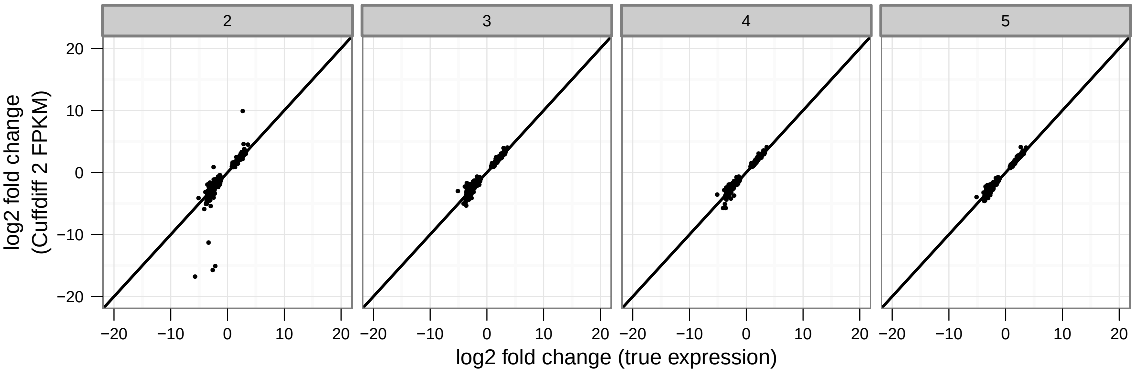

FIGURE 26. Minimal isoform switching: Transcript-level fold changes in perturbed genes as measured by union count (top) and Cuffdiff 2 FPKM (bottom). Plots are faceted by sequencing depth in the experiment.

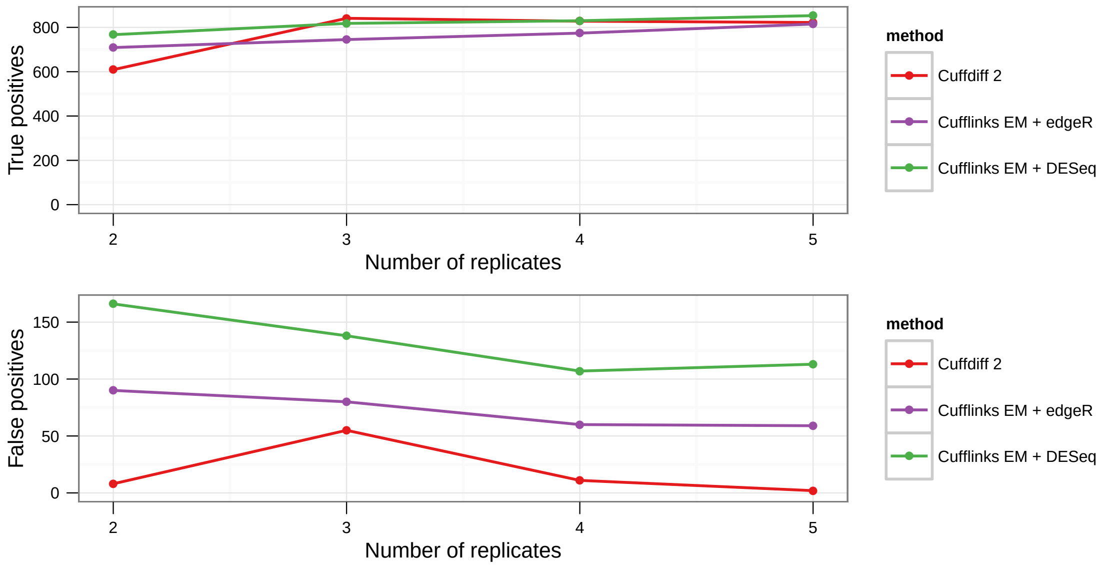

Figure 27. Minimal isoform switching: Transcript-level true and false positives at different degrees of replication.

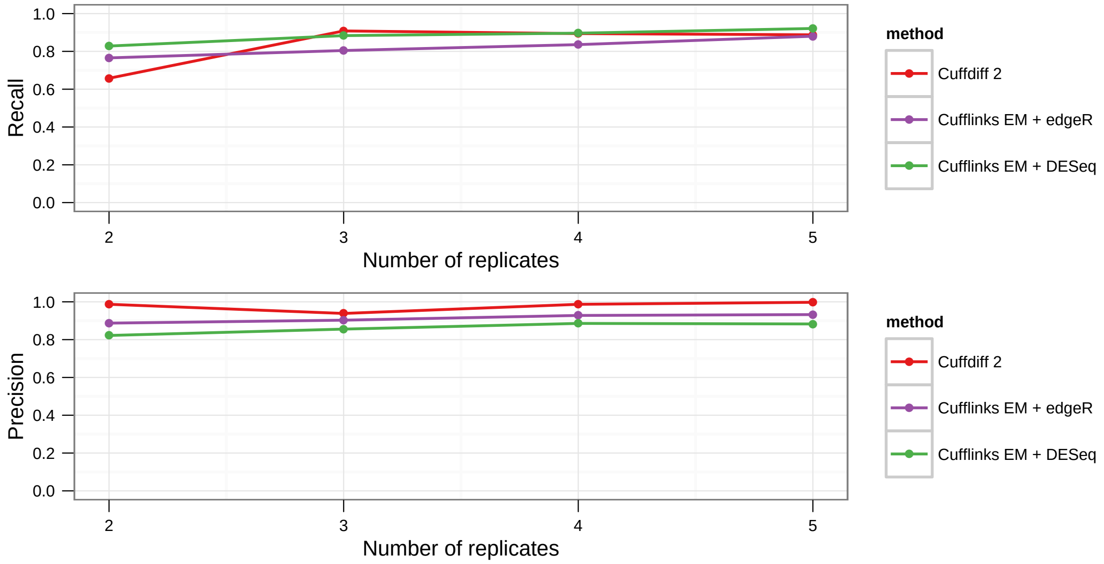

Figure 28. Minimal isoform switching: Transcript-level precision and recall at different degrees of replication.

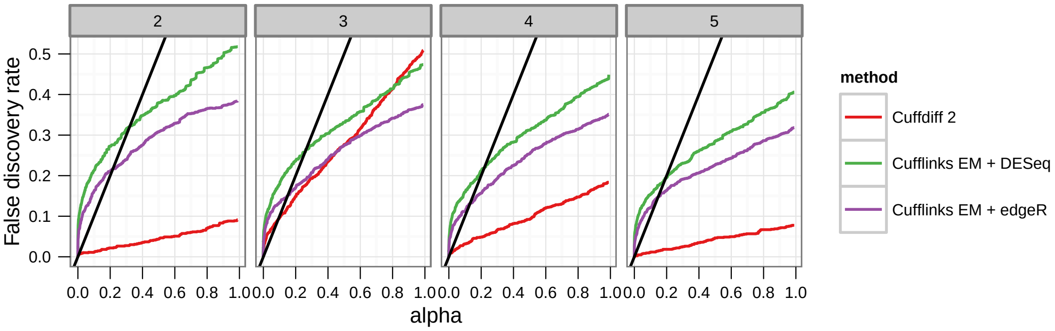

FIGURE 29. Minimal isoform switching: Transcript-level false discovery rate as a function of alpha, with the black line indicating the target FDR. Plots are faceted by number of replicates in the experiment.

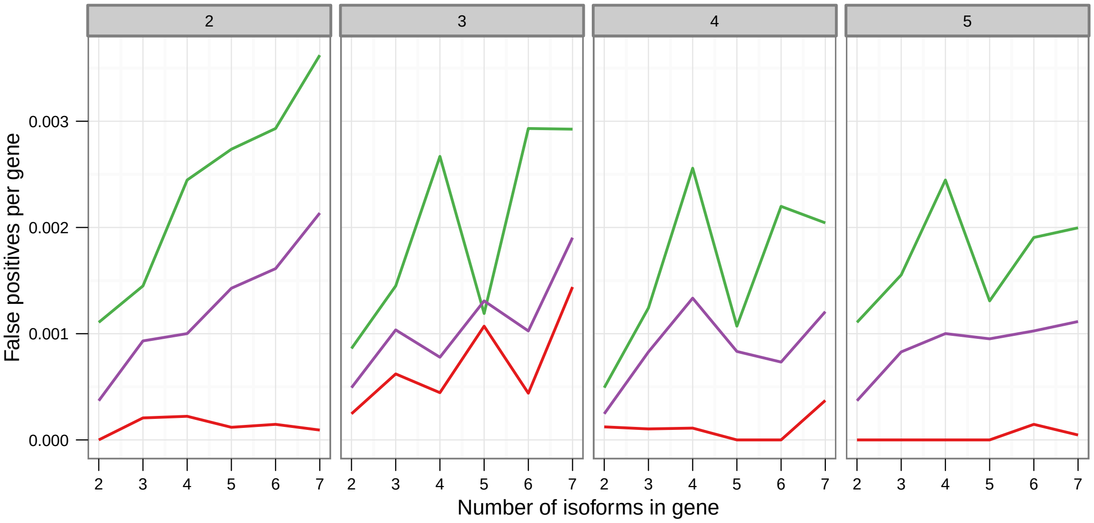

FIGURE 30. Minimal isoform switching: false positives transcripts per gene as a function of isoform diversity.

Nature Biotechnology: doi:10.1038/nbt.2450

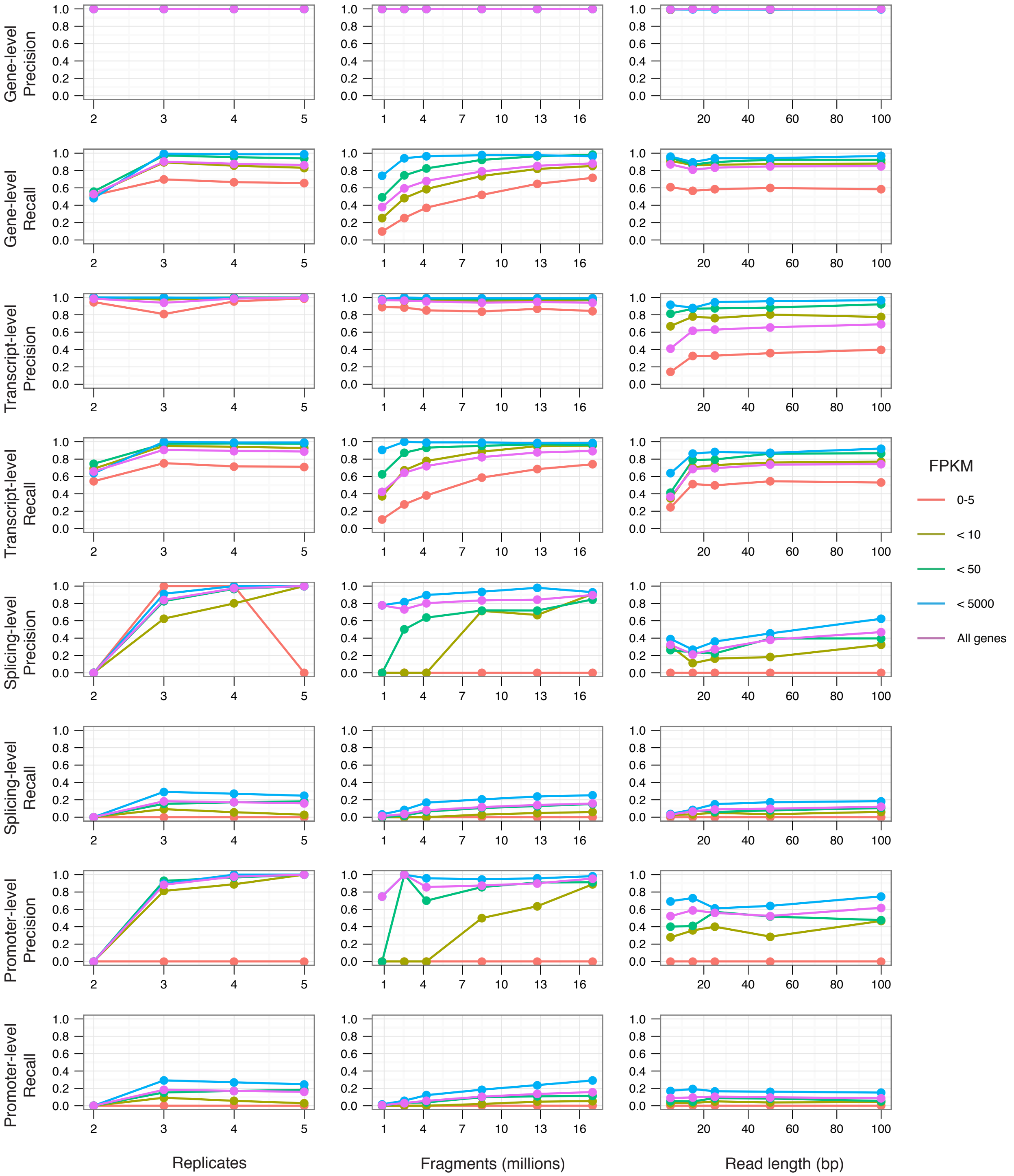

Figure 31. Accuracy of Cuffdiff 2 in the MAJOR-ISO simulation. Recall and precision of Cuffdiff 2's gene, transcript, splicing, and promoter switching tests under varying replication depths, sequencing depths, and read length.

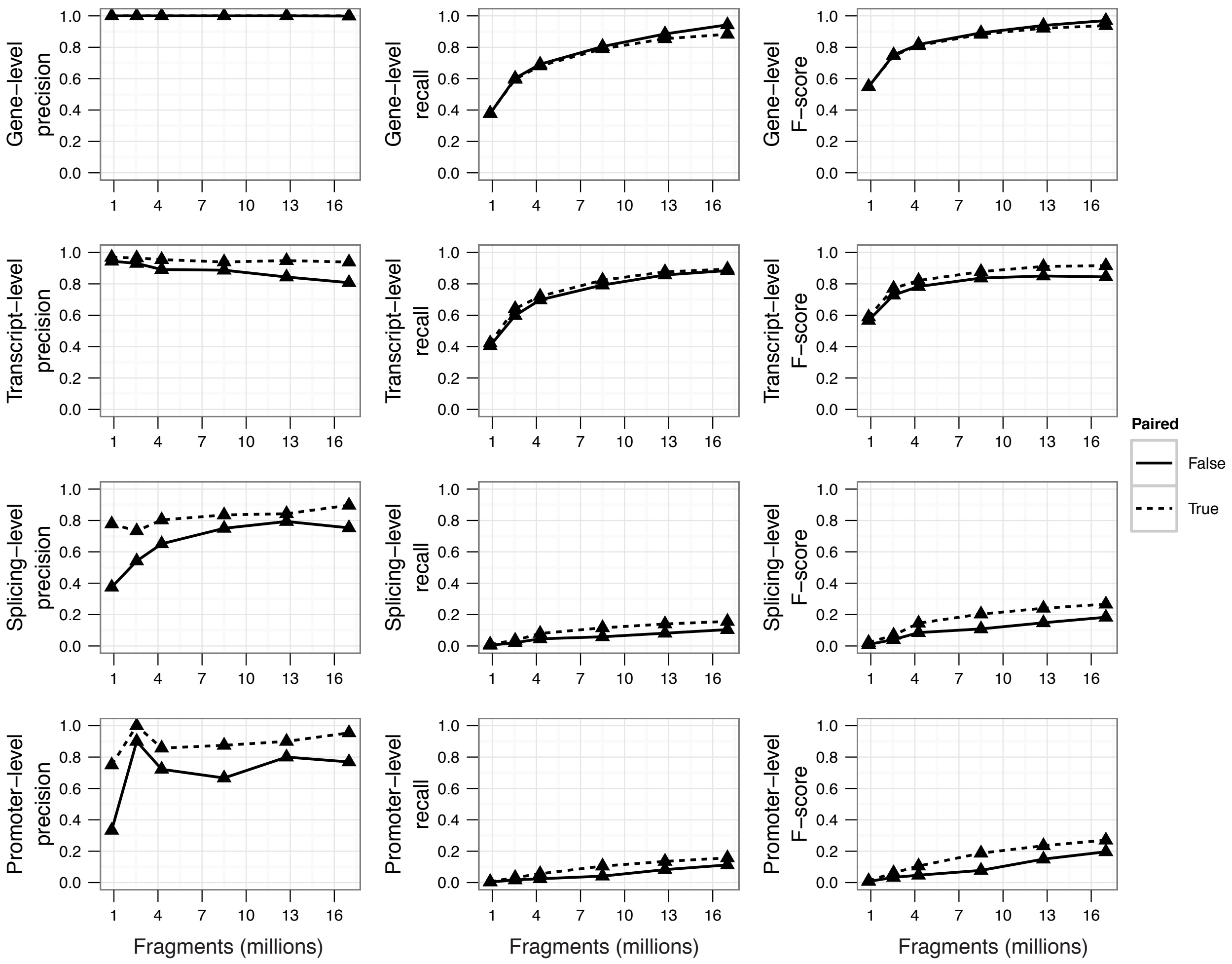

Figure 32. Accuracy of Cuffdiff 2 in the MAJOR-ISO simulation. Recall, precision, and F-score of Cuffdiff 2's gene, transcript, splicing, and promoter switching tests with and without paired-end RNA-Seq data.

3.7.2. *Multiple isoform perturbation.* The MULTI-ISO scenario perturbs all isoforms of each gene independently.

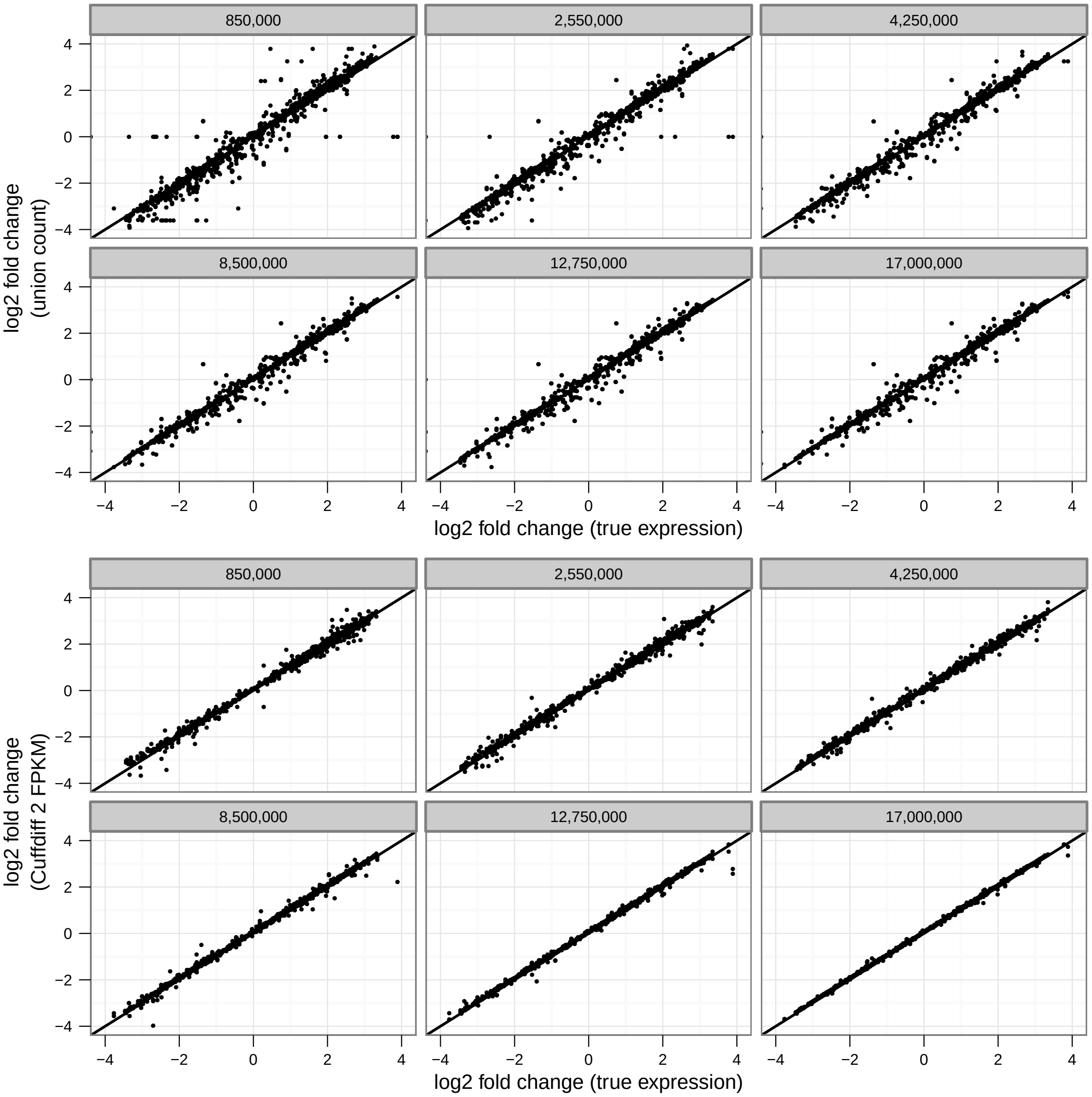

FIGURE 33. Moderate isoform switching: Gene-level fold changes in perturbed genes as measured by union count (top) and Cuffdiff 2 FPKM (bottom). Plots are faceted by sequencing depth in the experiment.

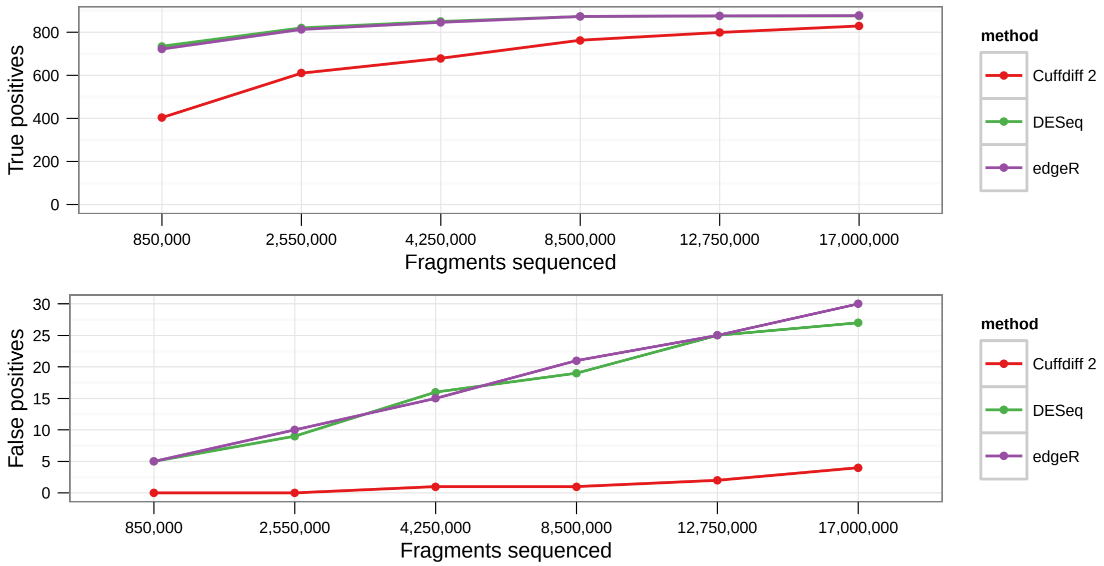

Figure 34. Moderate isoform switching: Gene-level true and false positives at different sequencing depths.

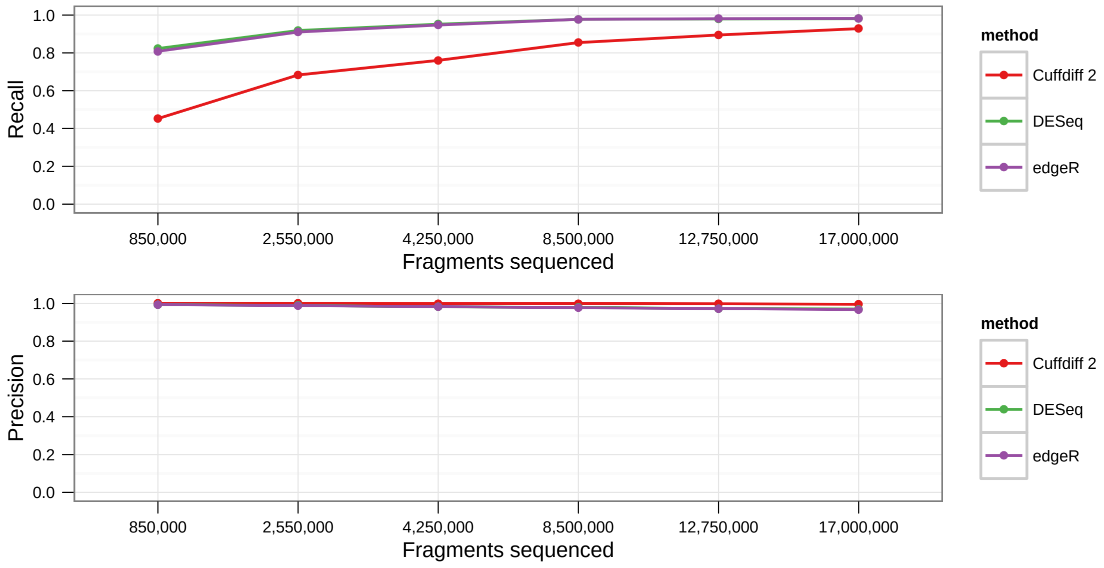

Figure 35. Moderate isoform switching: Gene-level precision and recall at different sequencing depths.

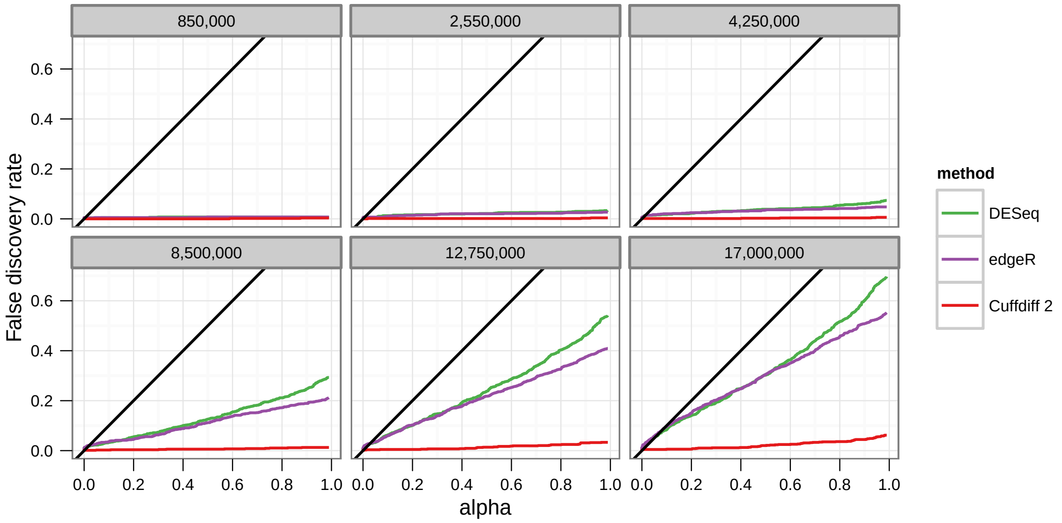

FIGURE 36. Moderate isoform switching: Gene-level false discovery rate as a function of alpha, with the black line indicating the target FDR. Plots are faceted by sequencing depth in the experiment.

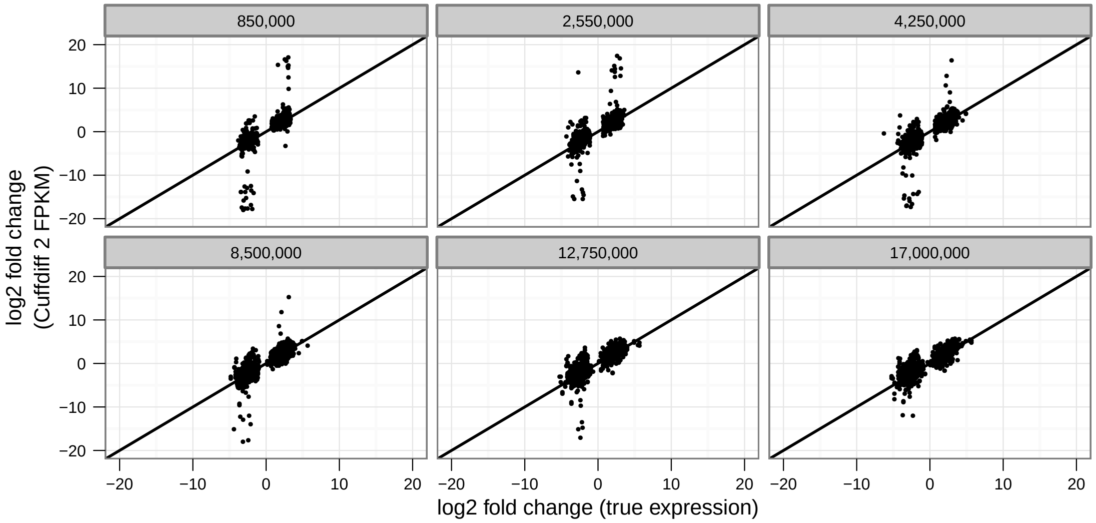

FIGURE 37. Moderate isoform switching: Transcript-level fold changes in perturbed genes as measured by union count (top) and Cuffdiff 2 FPKM (bottom). Plots are faceted by sequencing depth in the experiment.

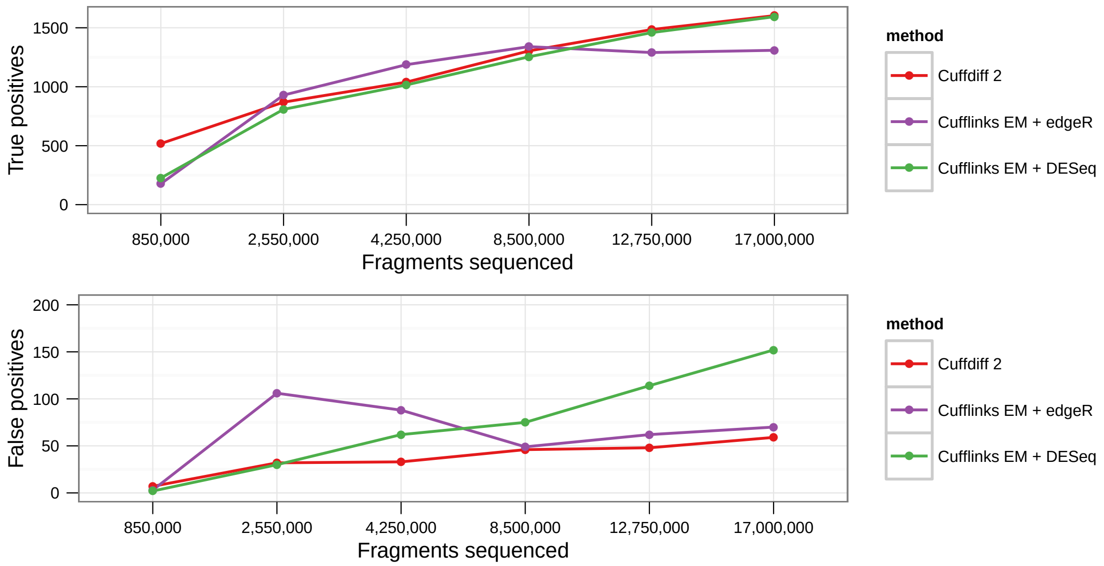

Figure 38. Moderate isoform switching: Transcript-level true and false positives at different sequencing depths.

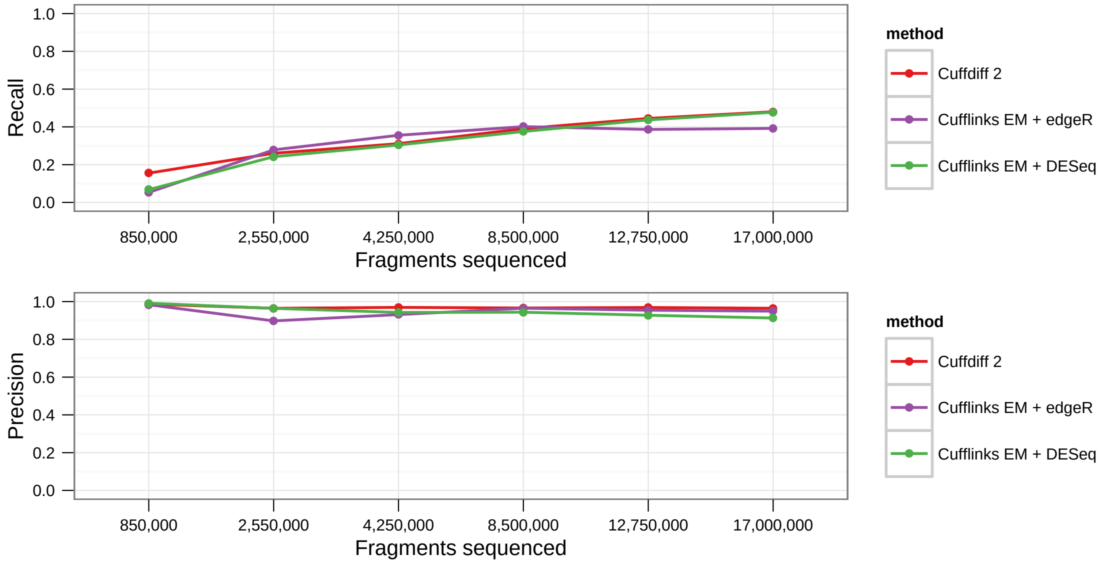

Figure 39. Moderate isoform switching: Transcript-level precision and recall positives at different sequencing depths.

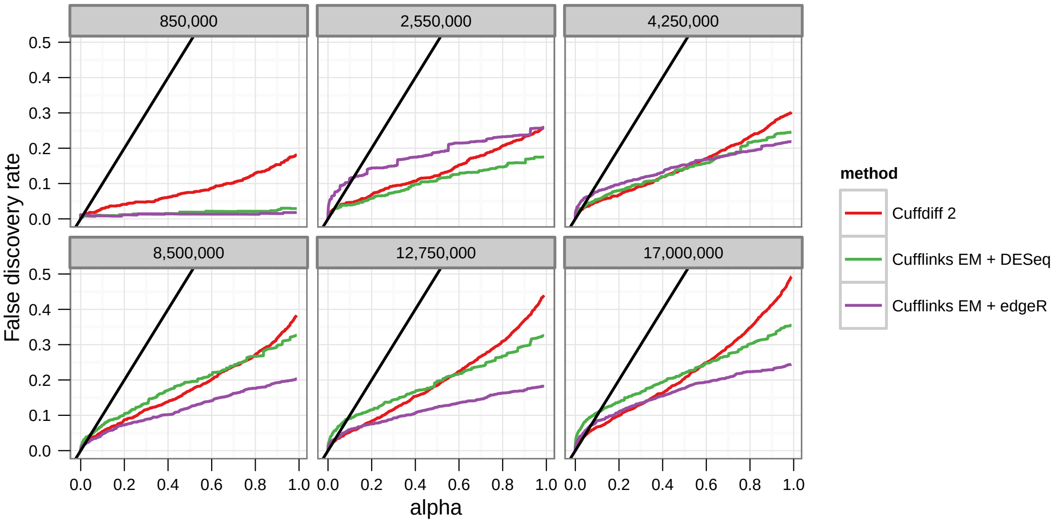

FIGURE 40. Moderate isoform switching: Transcript-level false discovery rate as a function of alpha, with the black line indicating the target FDR. Plots are faceted by sequencing depth in the experiment.

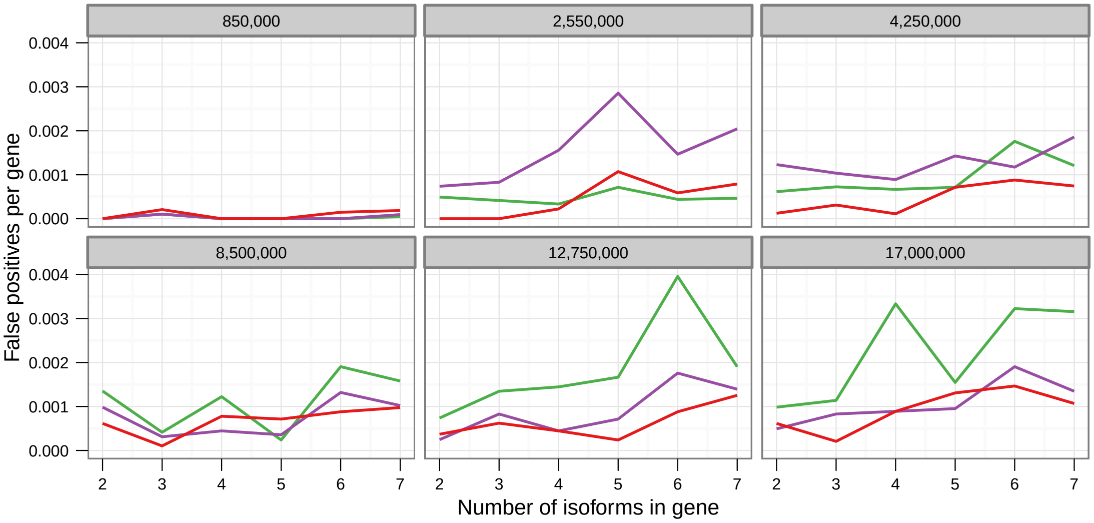

FIGURE 41. Minimal isoform switching:

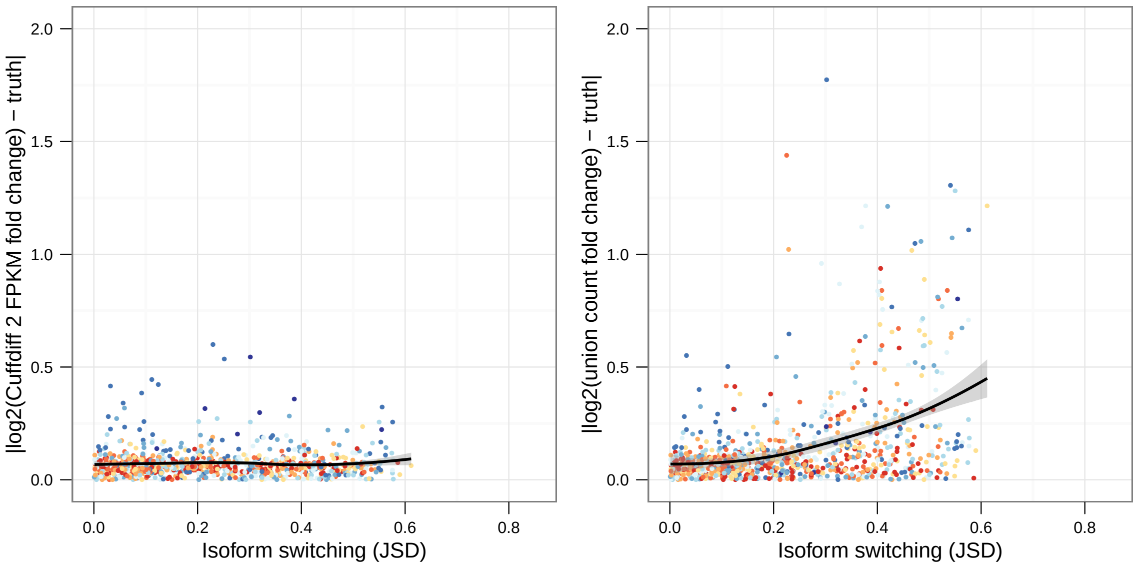

FIGURE 42. Minimal isoform switching: Error in estimation of change in gene expression as a function of isoform switching (measured as the Jensen-Shannon distance between relative isoform abundances) using Cuffdiff 2 (left) or union count (right)

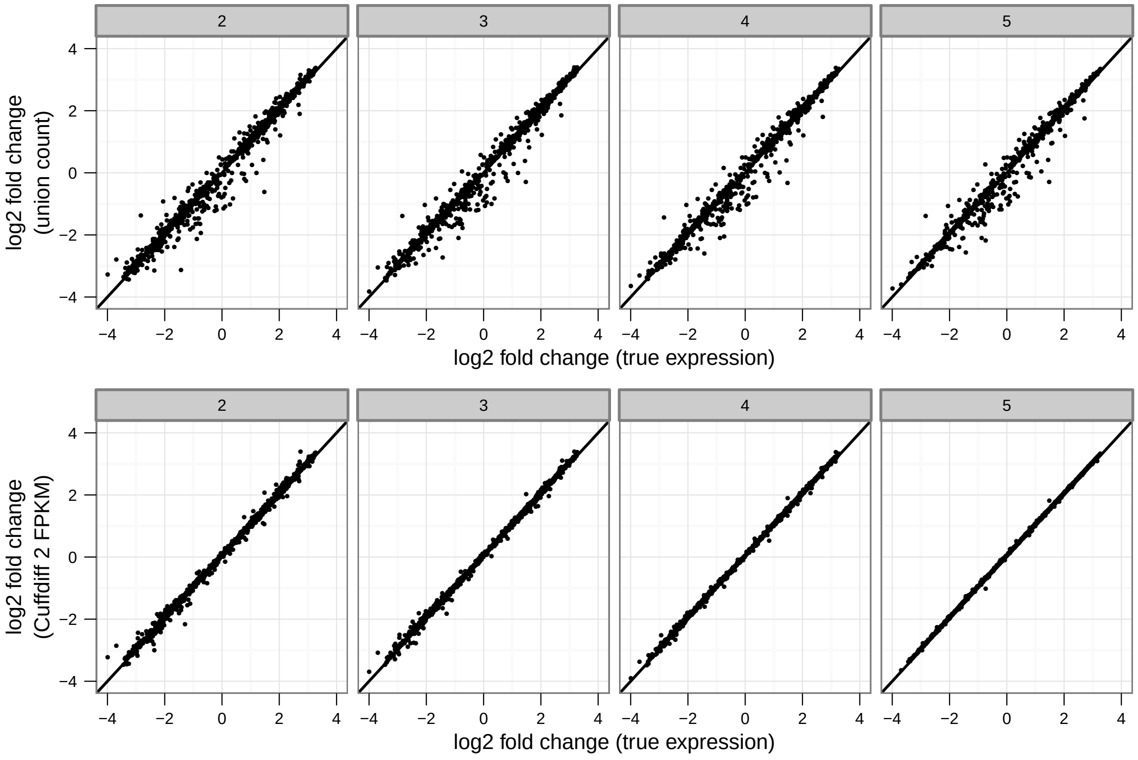

FIGURE 43. Moderate isoform switching: Gene-level fold changes in perturbed genes as measured by union count (top) and Cuffdiff 2 FPKM (bottom). Plots are faceted by sequencing depth in the experiment.

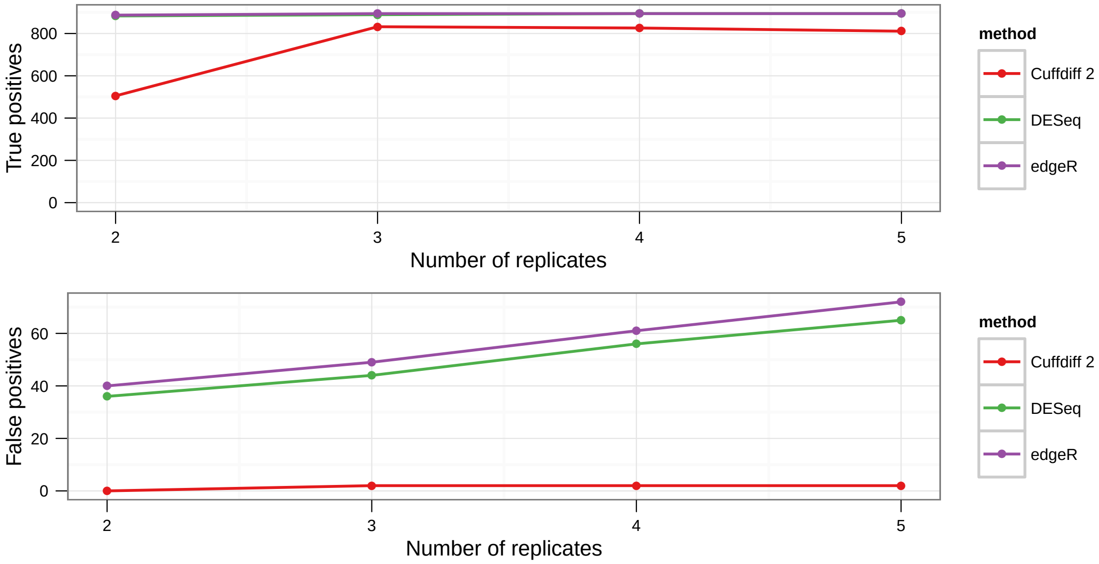

FIGURE 44. Moderate isoform switching: Gene-level true and false positives at different degrees of replication.

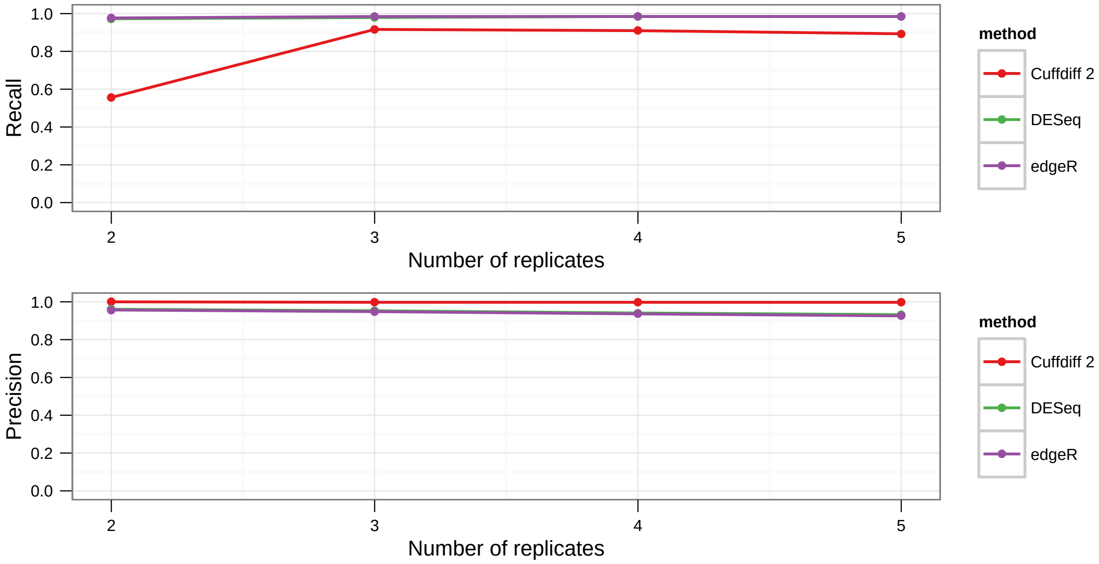

FIGURE 45. Moderate isoform switching: Gene-level precision and recall at different degrees of replication.

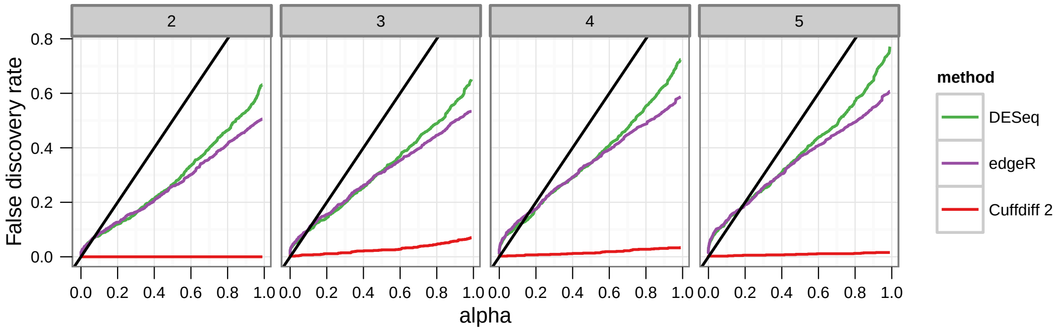

FIGURE 46. Moderate isoform switching: Gene-level false discovery rate as a function of alpha, with the black line indicating the target FDR. Plots are faceted by number of replicates in the experiment.

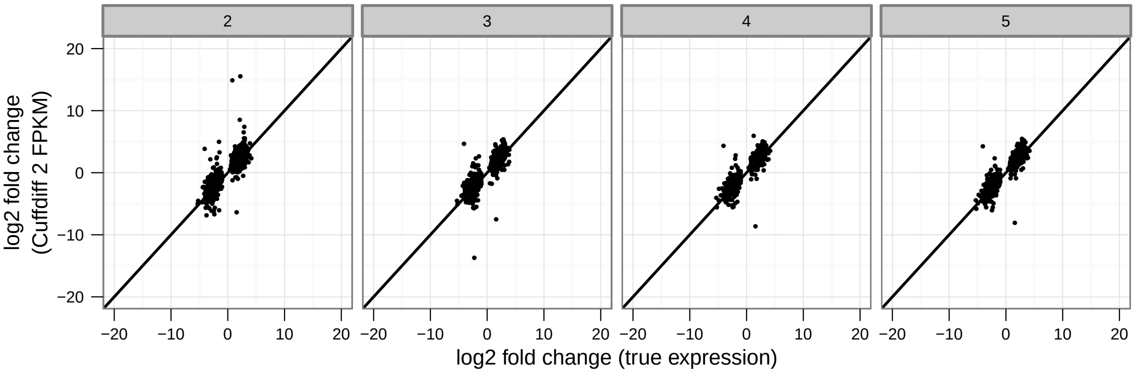

FIGURE 47. Moderate isoform switching: Transcript-level fold changes in perturbed genes as measured by union count (top) and Cuffdiff 2 FPKM (bottom). Plots are faceted by sequencing depth in the experiment.

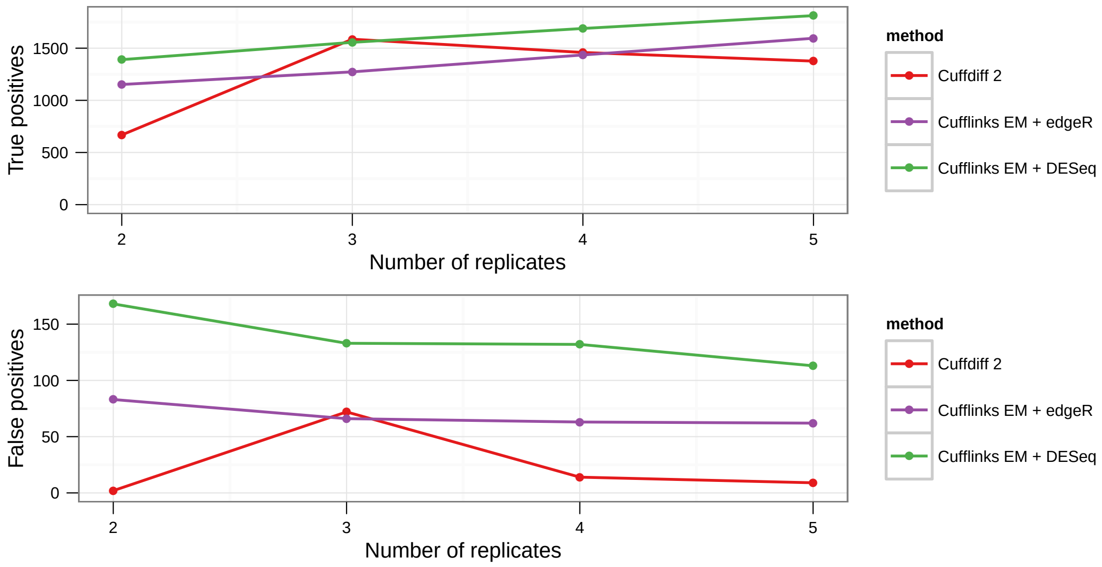

Figure 48. Moderate isoform switching: Transcript-level true and false positives at different degrees of replication.

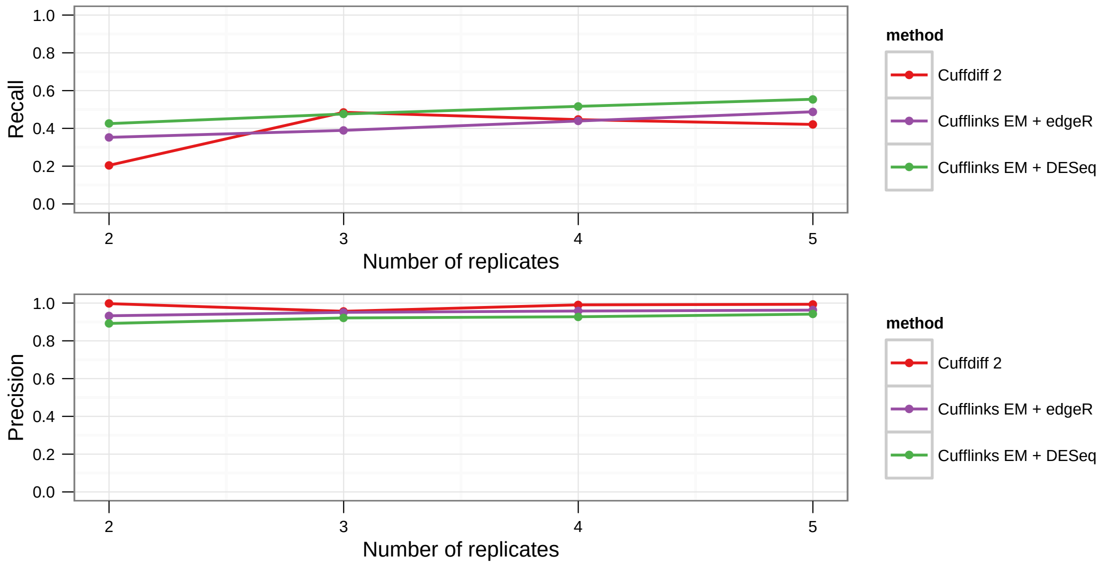

Figure 49. Moderate isoform switching: Transcript-level precision and recall at different degrees of replication.

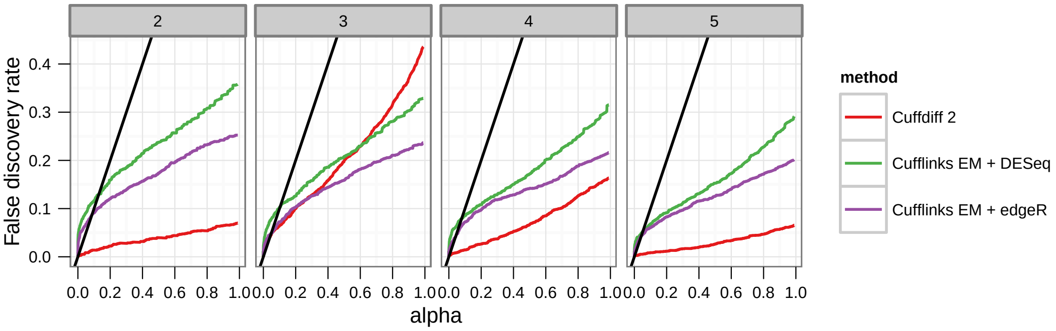

FIGURE 50. Moderate isoform switching: Transcript-level false discovery rate as a function of alpha, with the black line indicating the target FDR. Plots are faceted by number of replicates in the experiment.

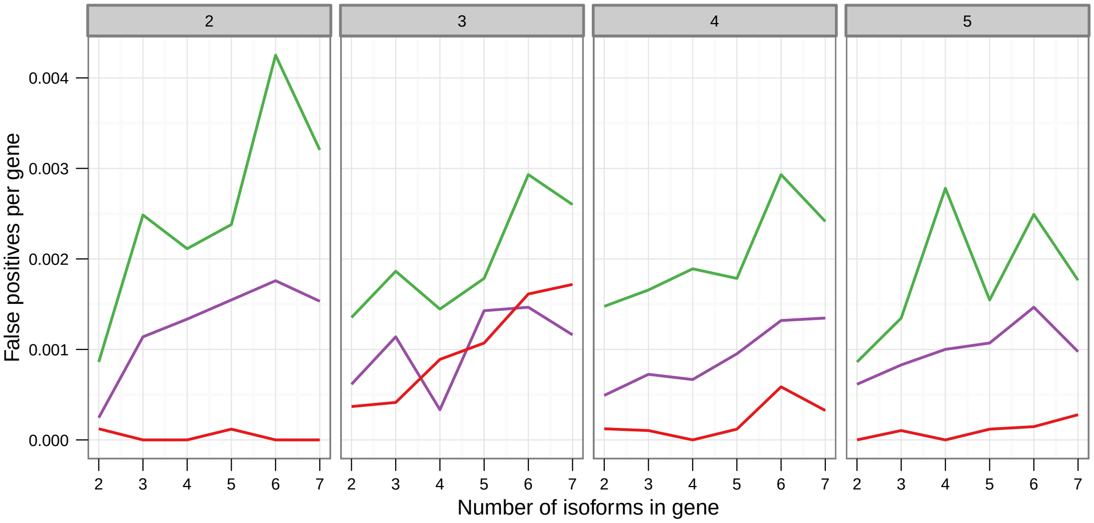

FIGURE 51. Moderate isoform switching: false positives transcripts per gene as a function of isoform diversity

Nature Biotechnology: doi:10.1038/nbt.2450

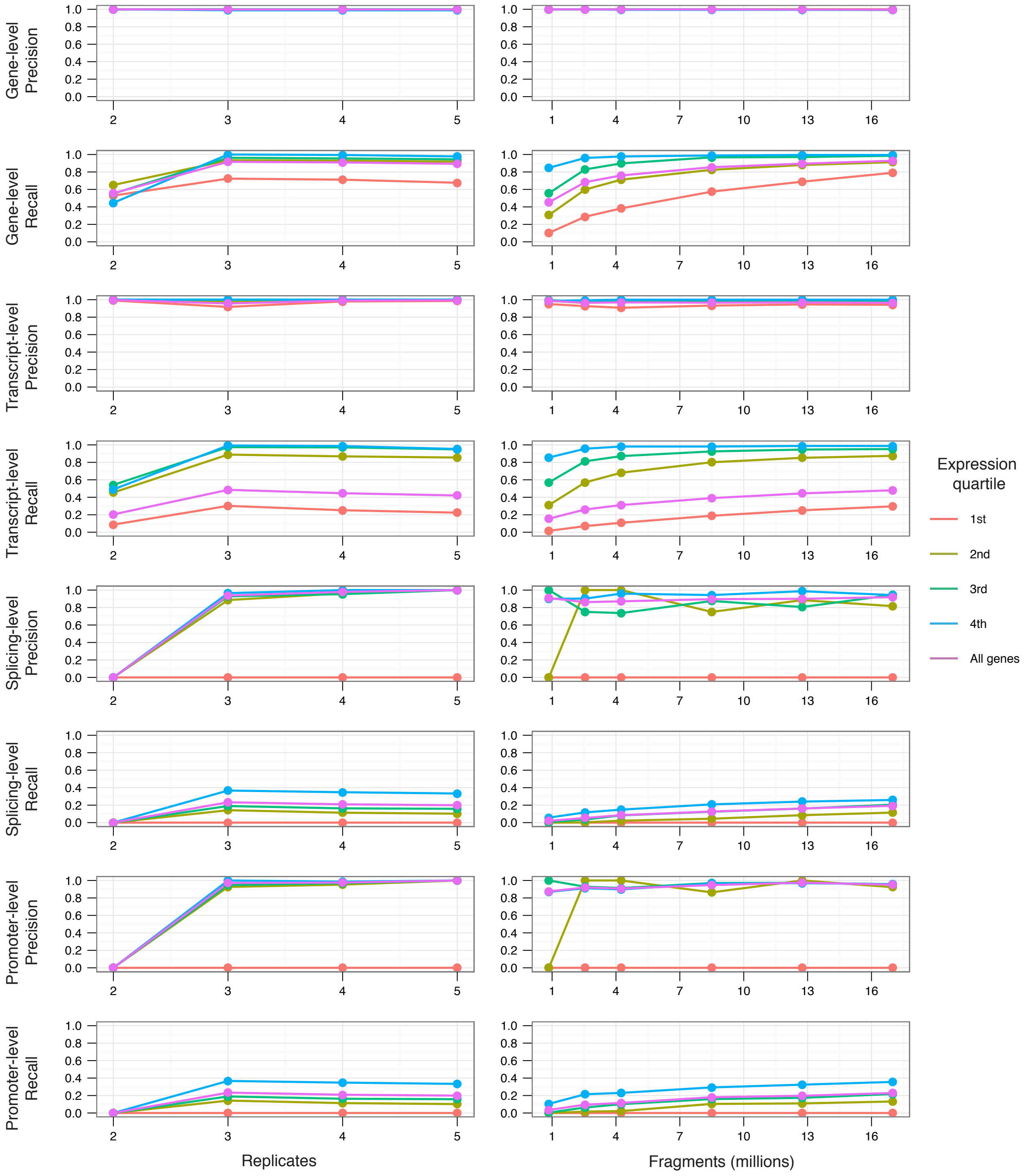

Figure 52. Accuracy of Cuffdiff 2 in the MULTI-ISO simulation. Recall and precision of Cuffdiff 2's gene, transcript, splicing, and promoter switching tests under varying replication depths, sequencing depths, and read length.

Nature Biotechnology: doi:10.1038/nbt.2450

3.7.3. *Isoform switching.* The SPECIALIZE scenario assigns all the abundance of a gene to a single isoform (not necessarily the major one).

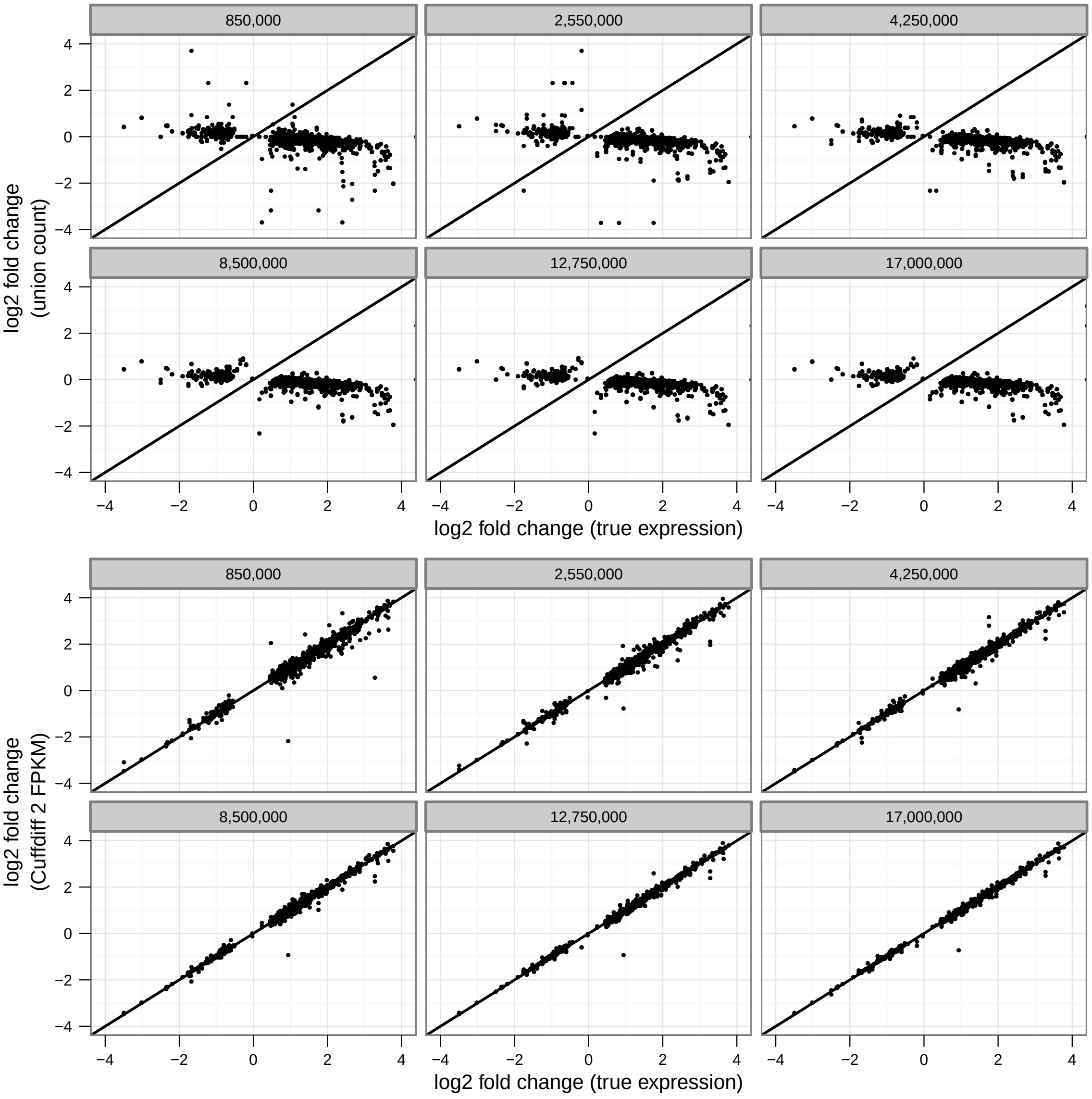

FIGURE 53. Heavy isoform switching: Gene-level fold changes in perturbed genes as measured by union count (top) and Cuffdiff 2 FPKM (bottom). Plots are faceted by sequencing depth in the experiment.

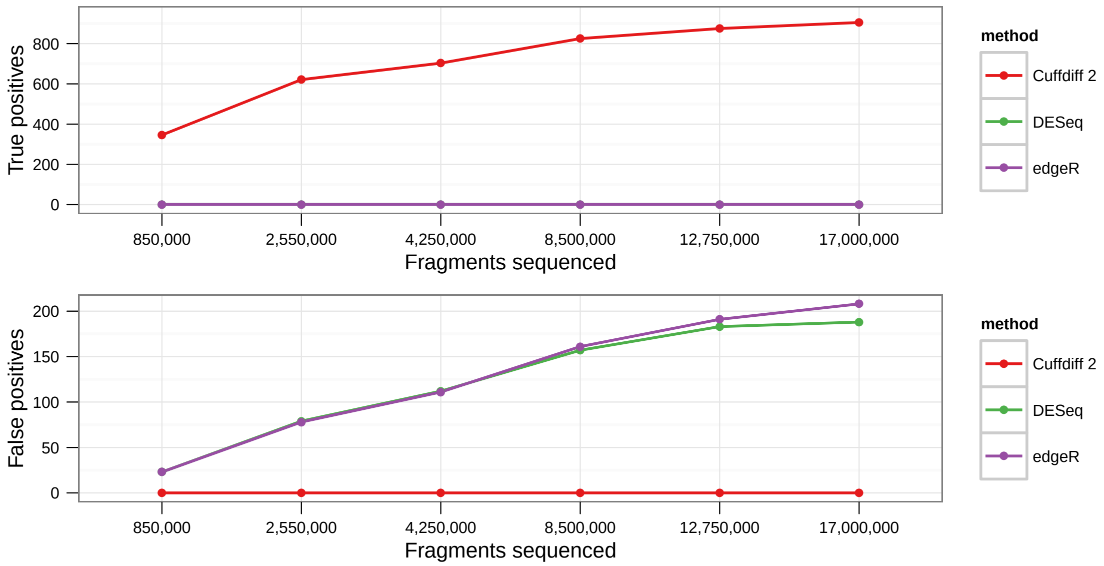

Figure 54. Heavy isoform switching: Gene-level true and false positives at different sequencing depths.

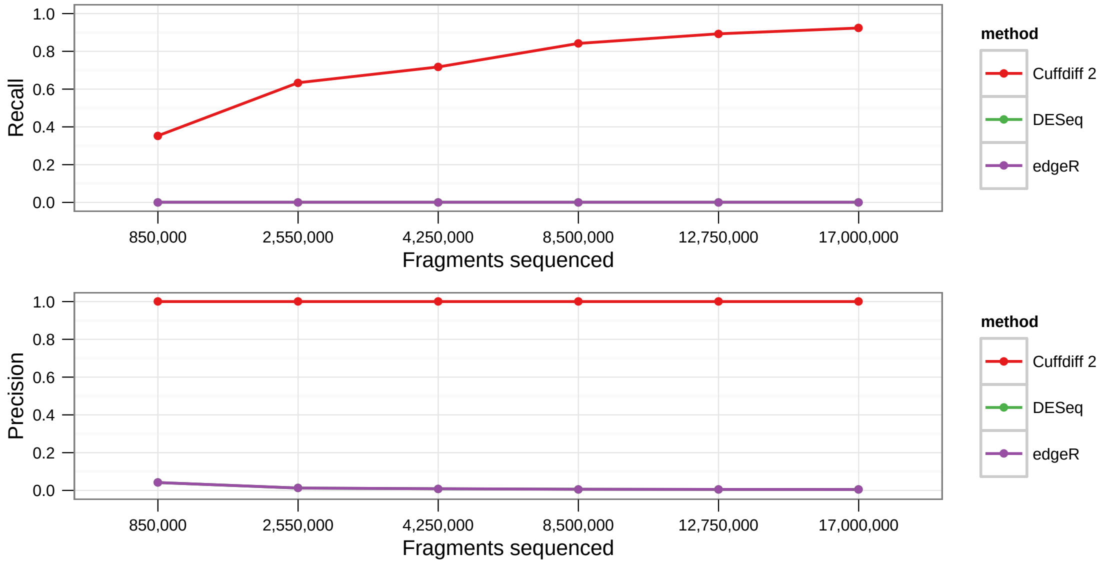

Figure 55. Heavy isoform switching: Gene-level precision and recall at different sequencing depths.

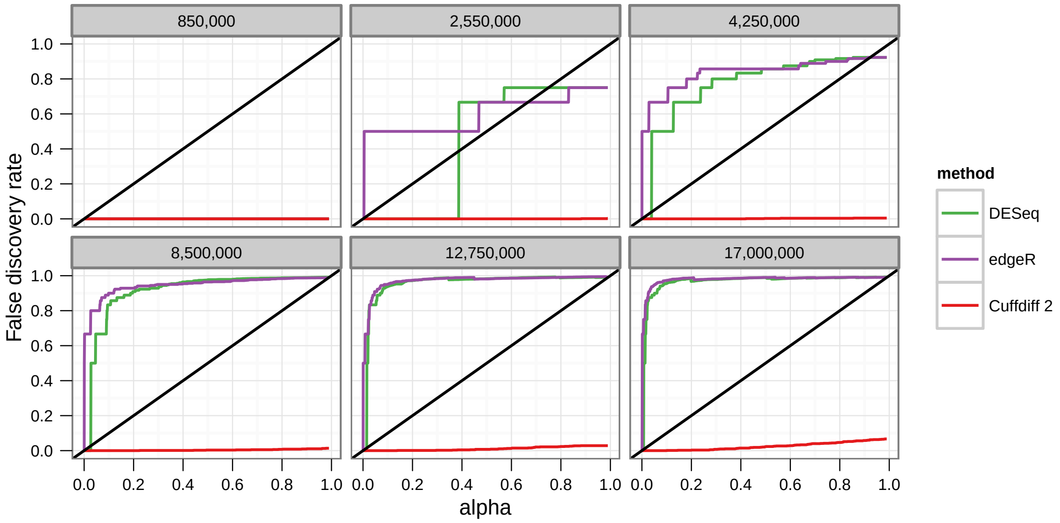

Figure 56. Heavy isoform switching: Gene-level false discovery rate as a function of alpha, with the black line indicating the target FDR. Plots are faceted by sequencing depth in the experiment.

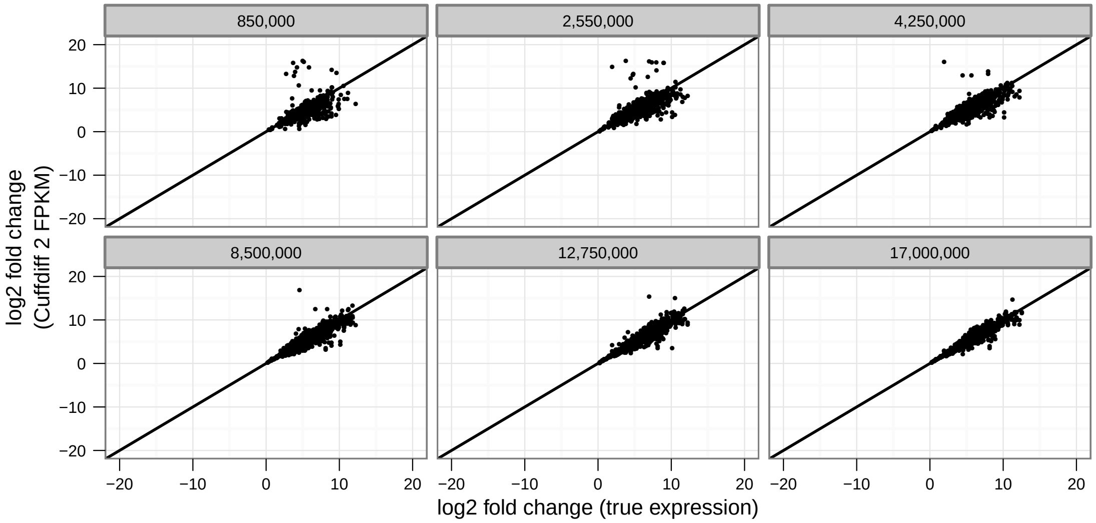

Figure 57. Heavy isoform switching: Transcript-level fold changes in perturbed genes as measured by union count (top) and Cuffdiff 2 FPKM (bottom). Plots are faceted by sequencing depth in the experiment.

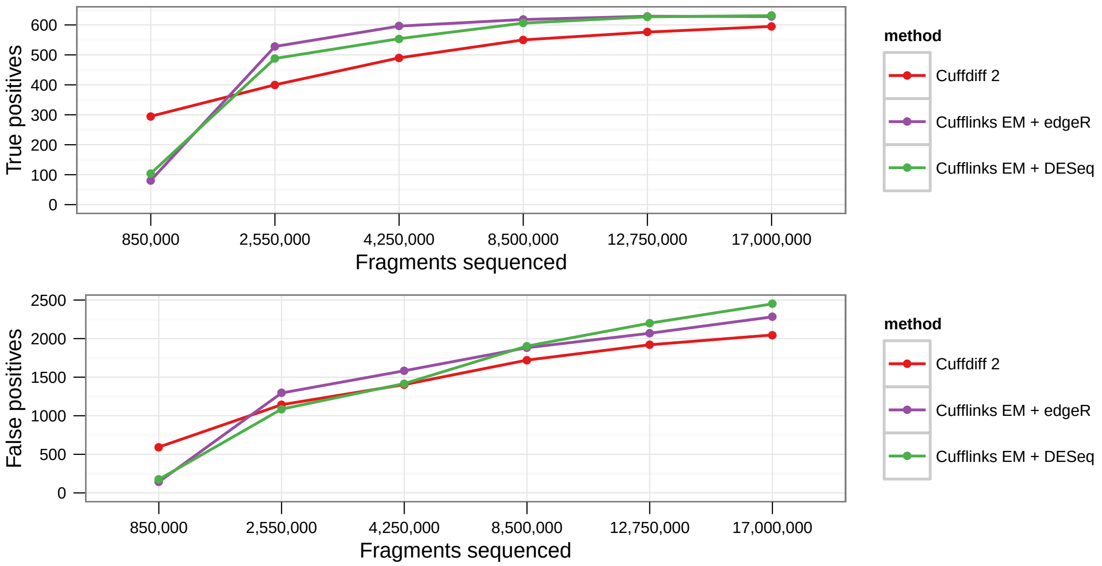

FIGURE 58. Heavy isoform switching: Transcript-level true and false positives at different sequencing depths.

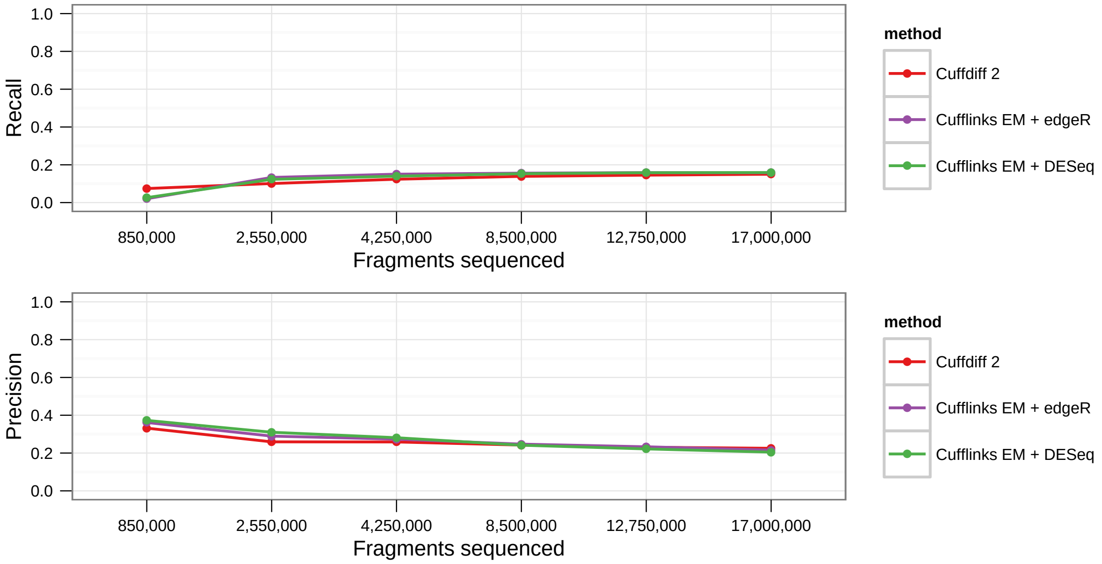

FIGURE 59. Heavy isoform switching: Transcript-level precision and recall positives at different sequencing depths.

‘

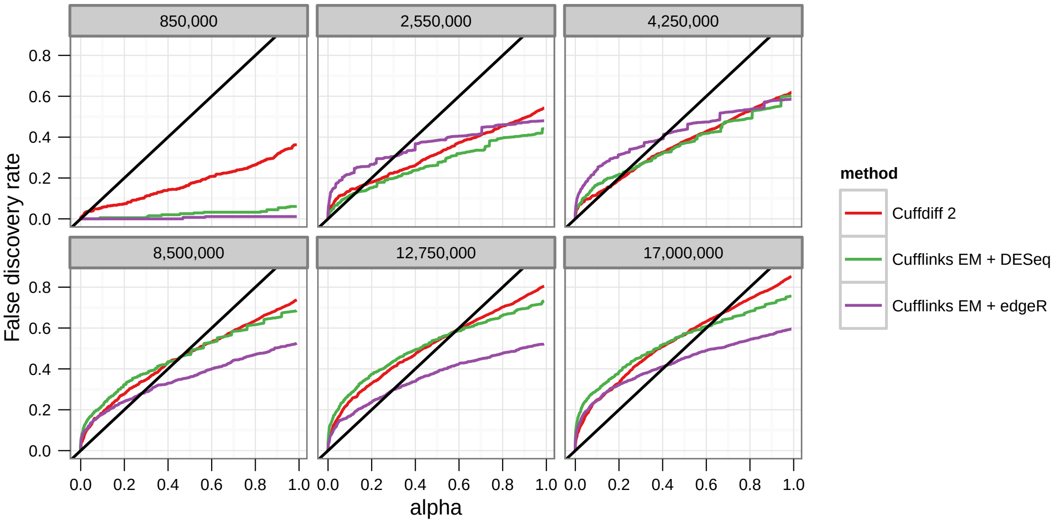

FIGURE 60. Heavy isoform switching: Transcript-level false discovery rate as a function of alpha, with the black line indicating the target FDR. Plots are faceted by sequencing depth in the experiment.

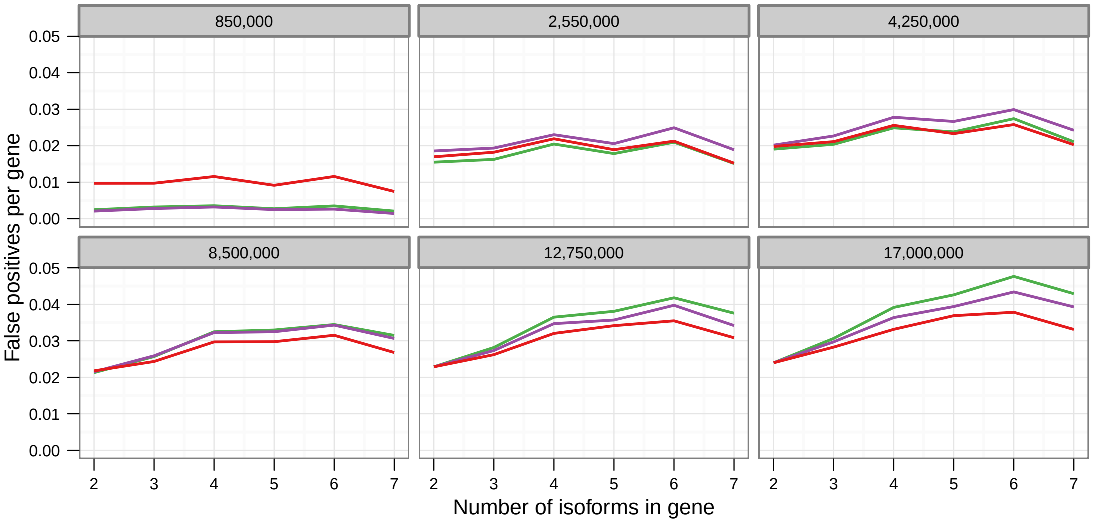

FIGURE 61. Heavy isoform switching:

Nature Biotechnology: doi:10.1038/nbt.2450

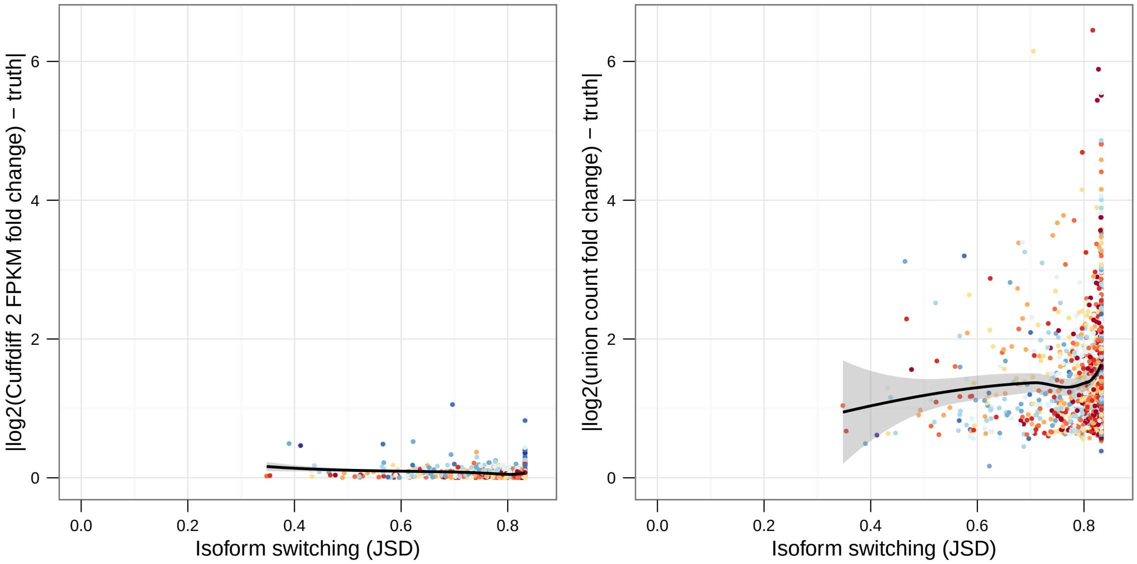

FIGURE 62. Heavy isoform switching: Error in estimation of change in gene expression as a function of isoform switching (measured as the Jensen-Shannon distance between relative isoform abundances) using Cuffdiff 2 (left) or union count (right).

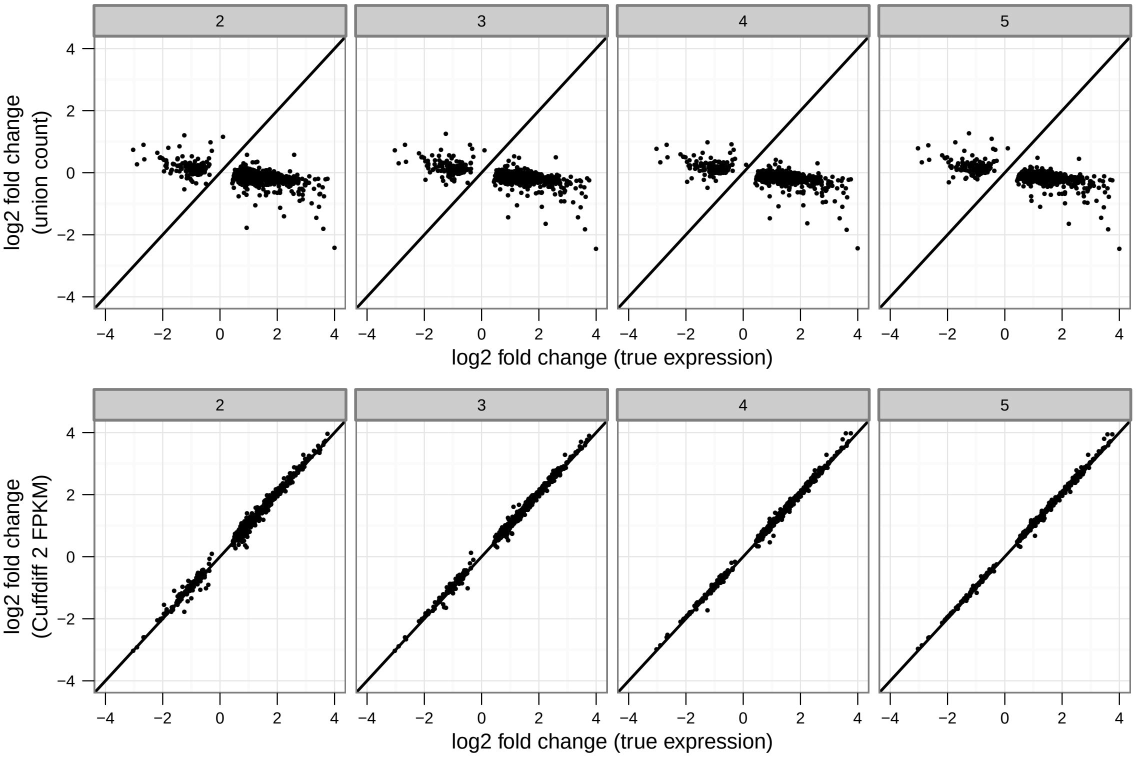

FIGURE 63. Heavy isoform switching: Gene-level fold changes in perturbed genes as measured by union count (top) and Cuffdiff 2 FPKM (bottom). Plots are faceted by sequencing depth in the experiment.

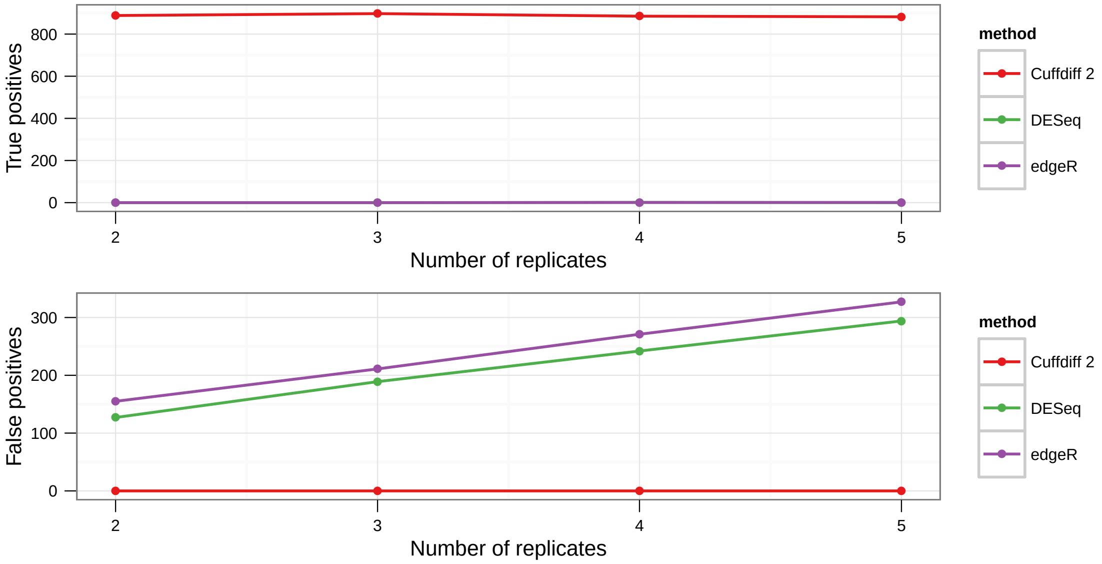

FIGURE 64. Heavy isoform switching: Gene-level true and false positives at different degrees of replication.

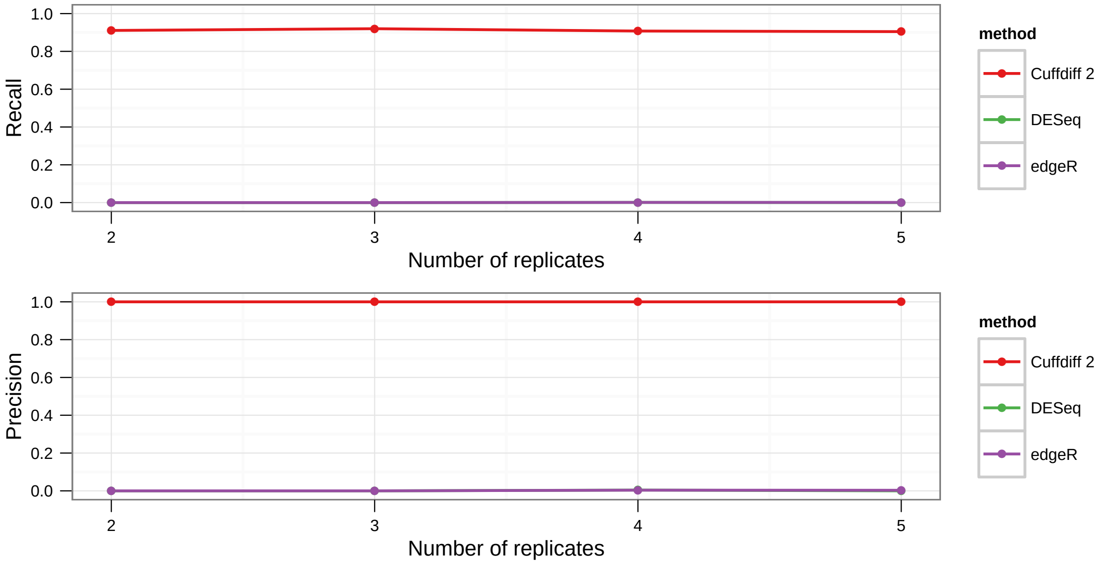

FIGURE 65. Heavy isoform switching: Gene-level precision and recall at different degrees of replication.

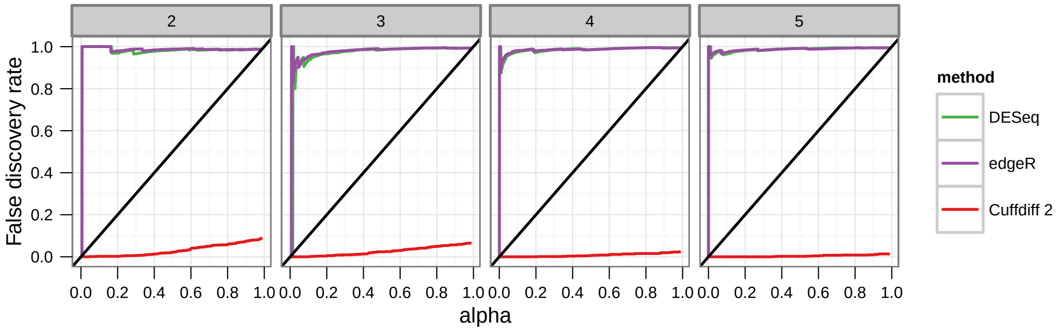

FIGURE 66. Heavy isoform switching: Gene-level false discovery rate as a function of alpha, with the black line indicating the target FDR. Plots are faceted by number of replicates in the experiment.

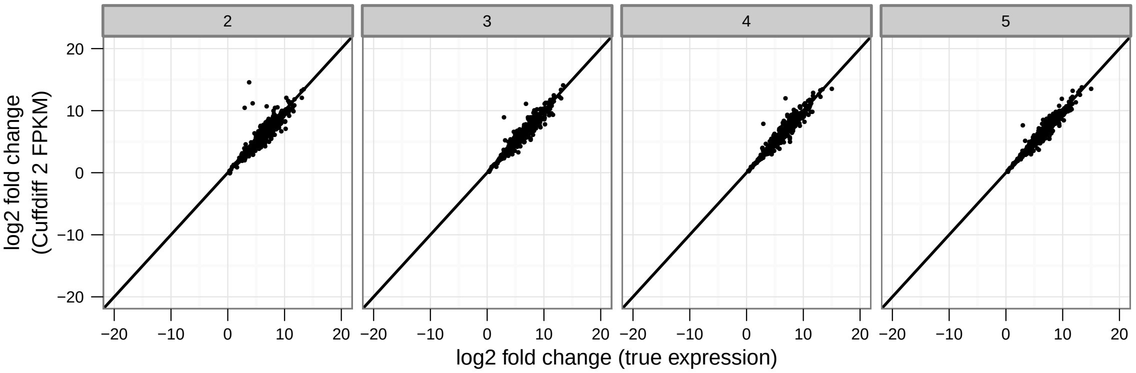

FIGURE 67. Heavy isoform switching: Transcript-level fold changes in perturbed genes as measured by union count (top) and Cuffdiff 2 FPKM (bottom). Plots are faceted by sequencing depth in the experiment.

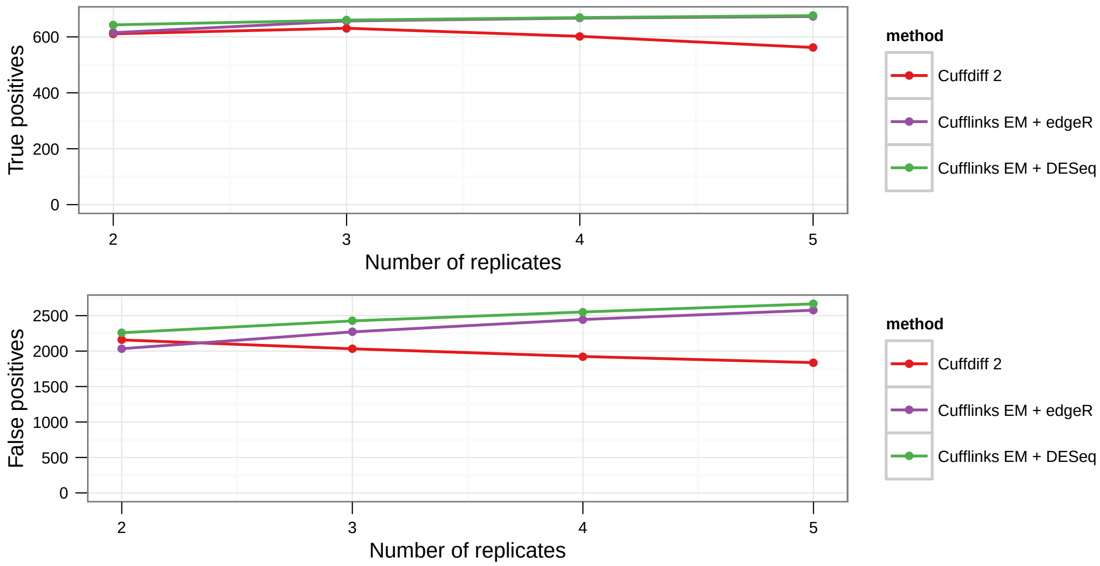

Figure 68. Heavy isoform switching: Transcript-level true and false positives at different degrees of replication.

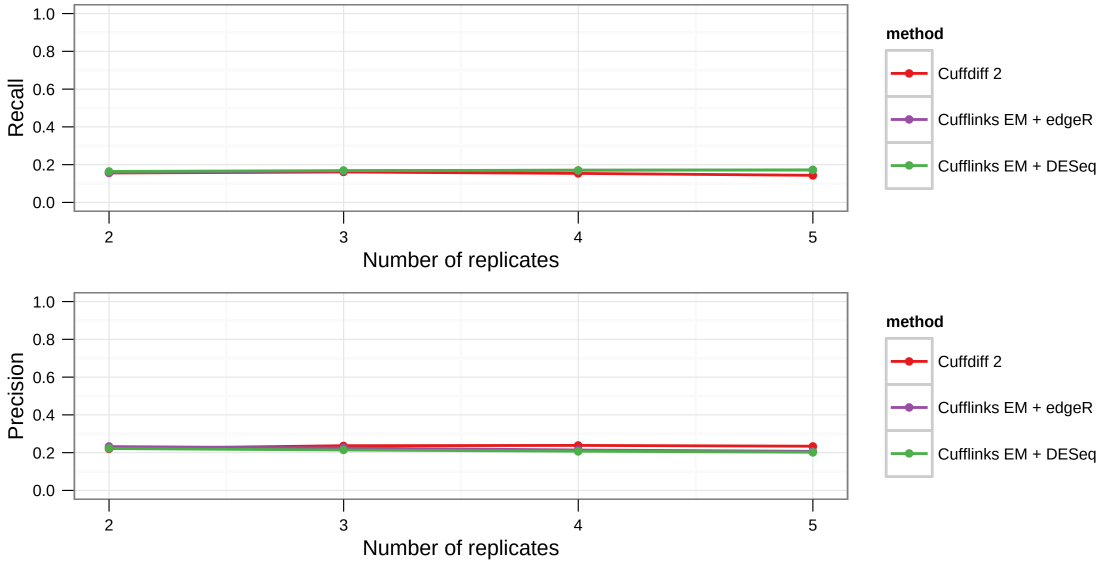

Figure 69. Heavy isoform switching: Transcript-level precision and recall at different degrees of replication.

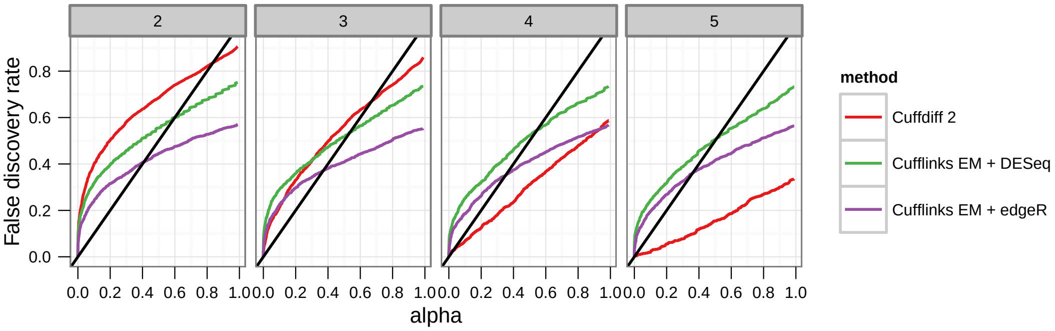

FIGURE 70. Heavy isoform switching: Transcript-level false discovery rate as a function of alpha, with the black line indicating the target FDR. Plots are faceted by number of replicates in the experiment.

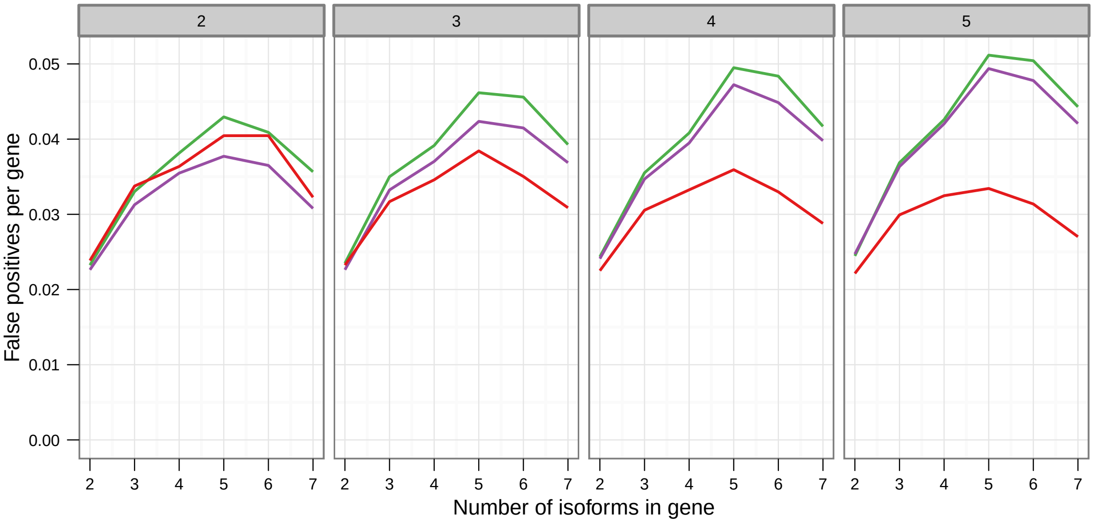

FIGURE 71. Heavy isoform switching: false positives transcripts per gene as a function of isoform diversity

Nature Biotechnology: doi:10.1038/nbt.2450

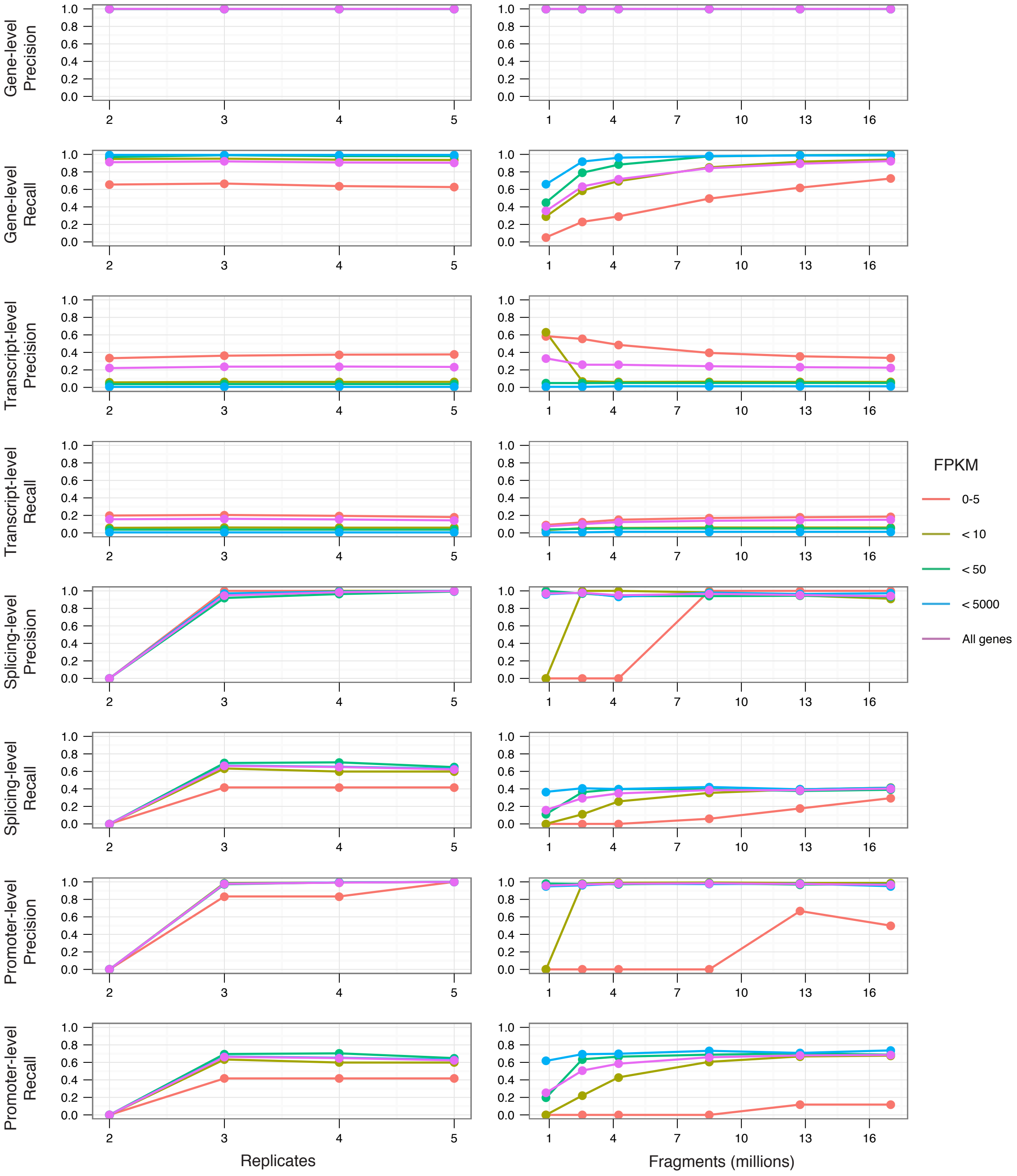

FIGURE 72. Accuracy of Cuffdiff 2 in the SPECIALIZE simulation. Recall and precision of Cuffdiff 2's gene, transcript, splicing, and promoter switching tests under varying replication depths, sequencing depths, and read length.

3.7.4. *Other profiling scenarios.* This section contains simulation experiments that examine the impact of heavy overdispersion among replicates. Also shown are performance results when simulated transcript abundances are derived from a mixed power/exponential rule as used in the FluxSimulator [3].

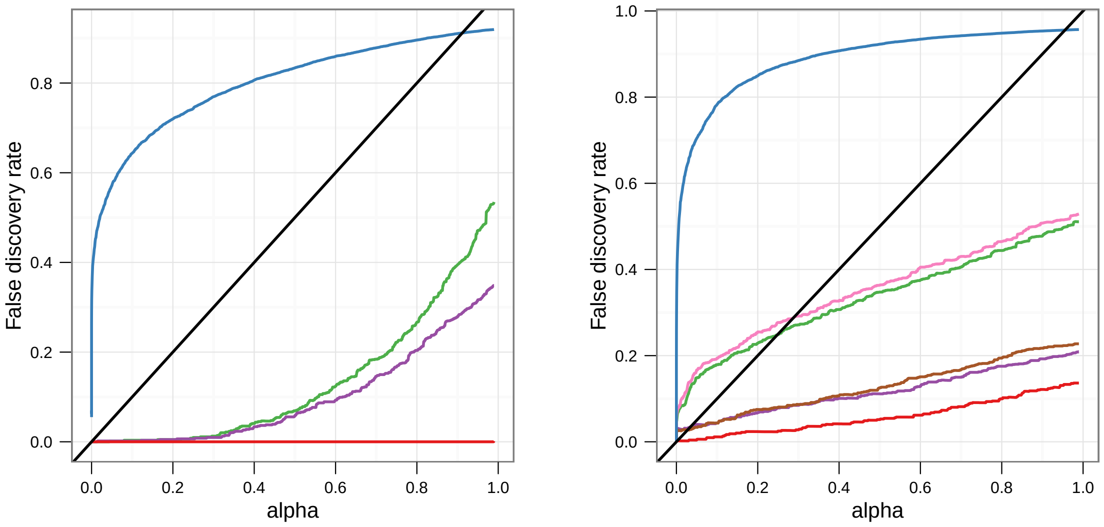

FIGURE 73. FDR for gene-level (left) and isoform-level (right) differential analysis with Cuffdiff 2 and other popular tools under high levels of cross-replicate variability. Overdispersion was set to 10-fold higher than the levels observed in our fibroblast experiment across all ranges of gene expression.

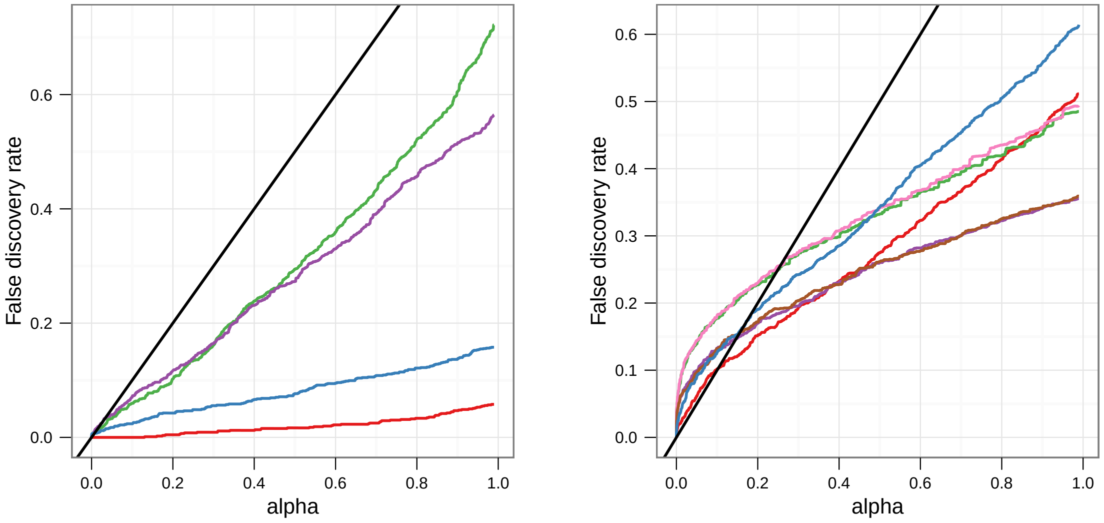

FIGURE 74. ROC and FDR for gene-level (left) and isoform-level (right) differential analysis with Cuffdiff 2 and other popular tools under levels of cross-replicate variability similar to that in our fibroblast data.

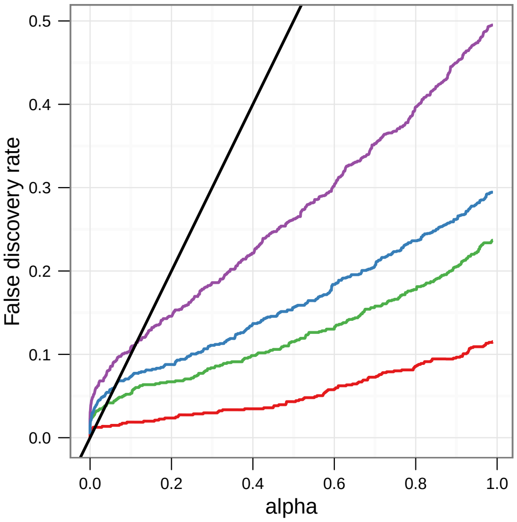

Figure 75. FDR of gene-level (left) differential analysis under simulations derived from abundance assignment policy used by FluxSimulator (right panels).

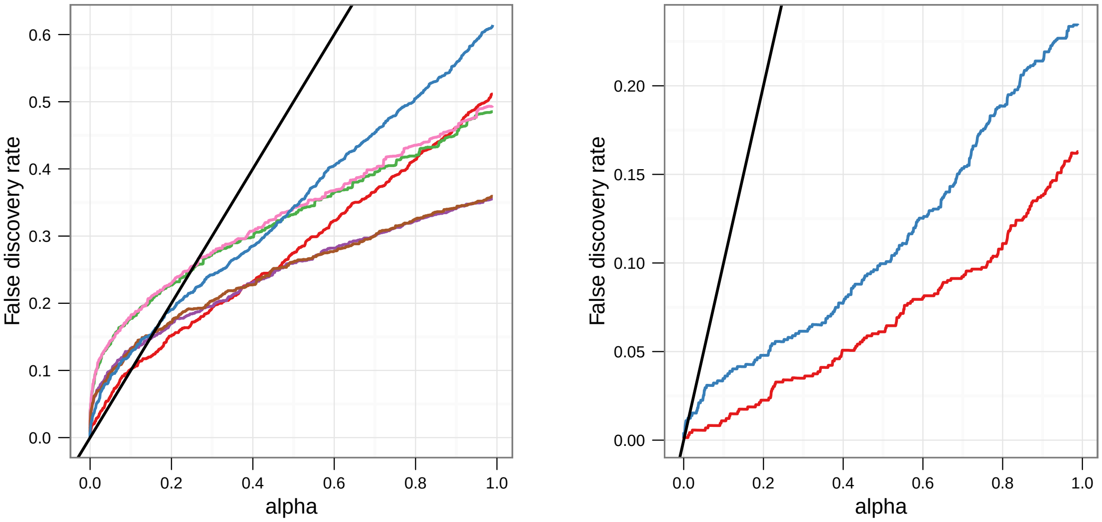

Figure 76. Performance of transcript-level differential analysis under simulated abundances derived from our fibroblast data (left) and derived from the default abundance assignment policy used by FluxSimulator (right).

4. Other supplemental figures

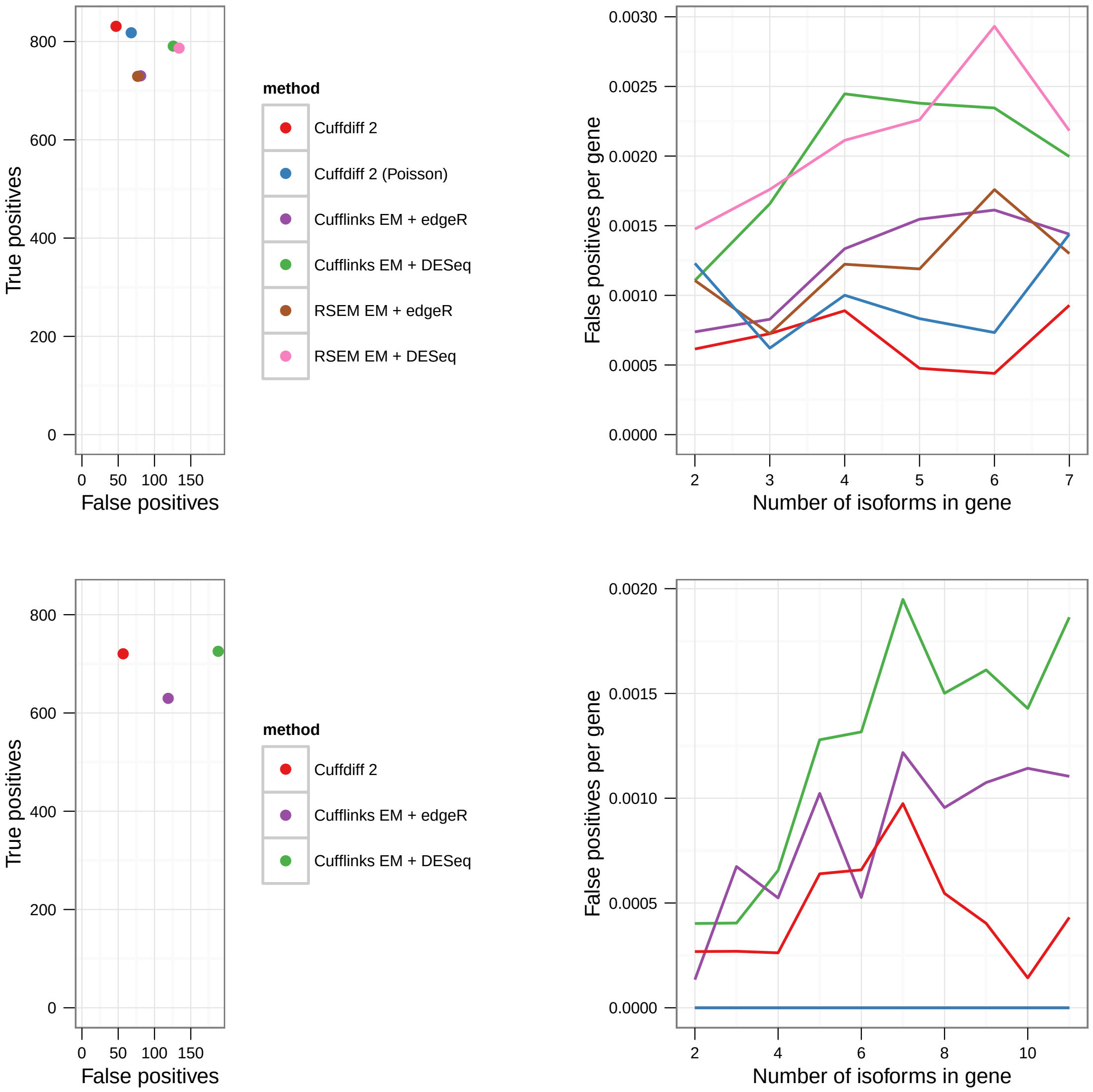

Figure 77. Performance of several approaches for isoform-level differential analysis. (Top left) True and false positives in the MAJOR-ISO simulation scenario (see below) using the UCSC hg19 transcriptome annotation. Cuffdiff 2 and several alternative pipelines that couple Cufflinks or RSEM transcript fragment count estimates with other count-based tools. (Top right) rate of false positive DE isoforms per gene as a function of gene splicing complexity. (Bottom left) True and false positives and (bottom right) false positive rates when using the GENCODE v12 annotation, which contains nearly twice as many isoforms for most genes as the UCSC hg19 annotation.

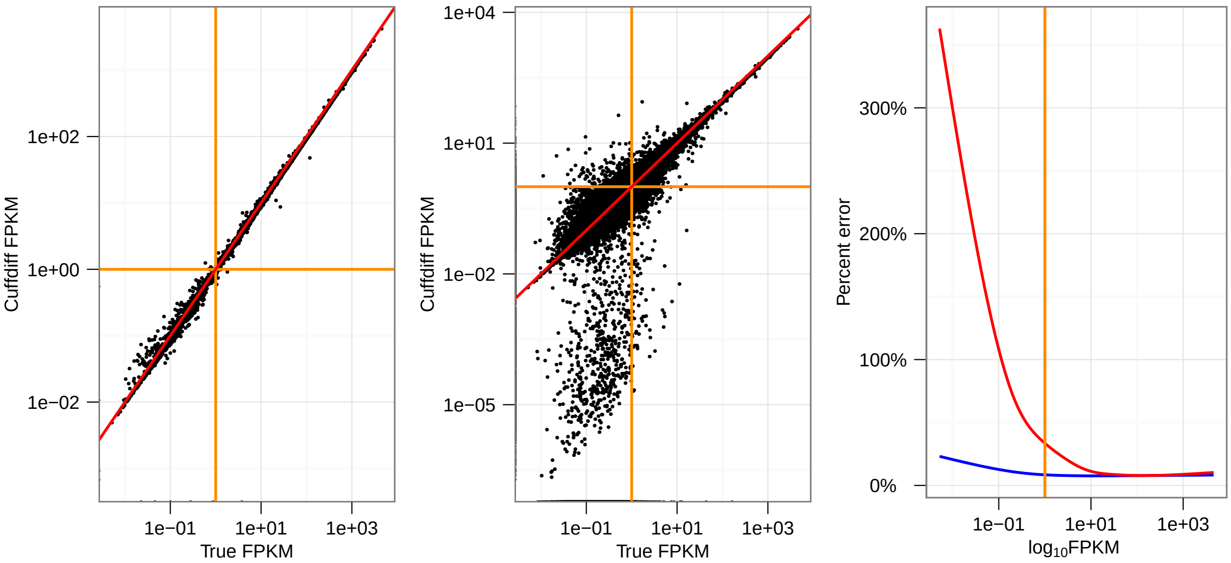

Figure 78. Accuracy of recovered FPKMs in simulation. Cuffdiff 2 FPKM estimates for genes (left) and transcripts (right) compared against the true values. Error decreases with expression level for genes (right, blue line), and transcripts (red line). Only multi-isoform genes are shown.

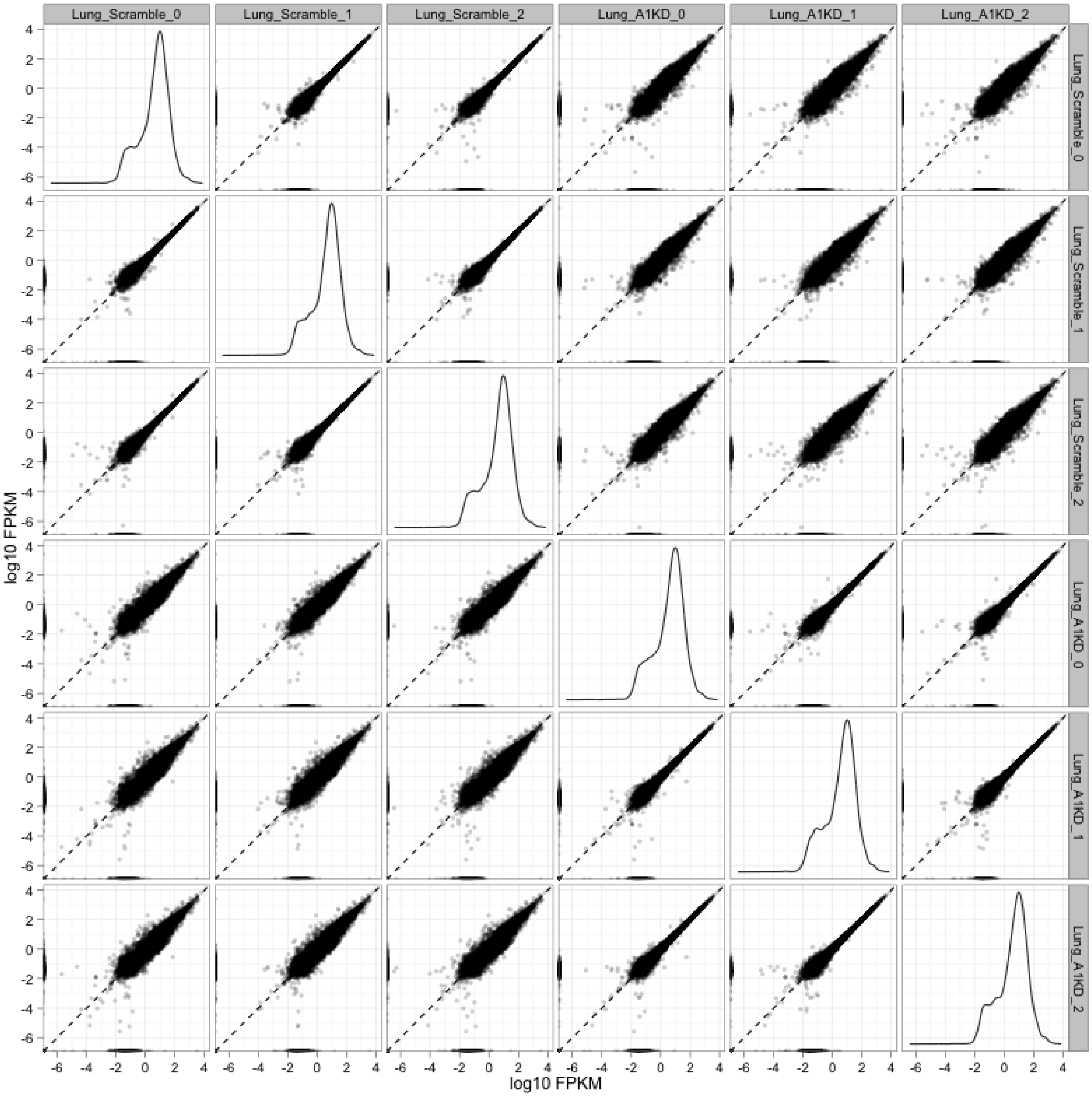

Figure 79. Scatterplot matrix across replicates of gene level FPKM values in the fibroblast experiment.

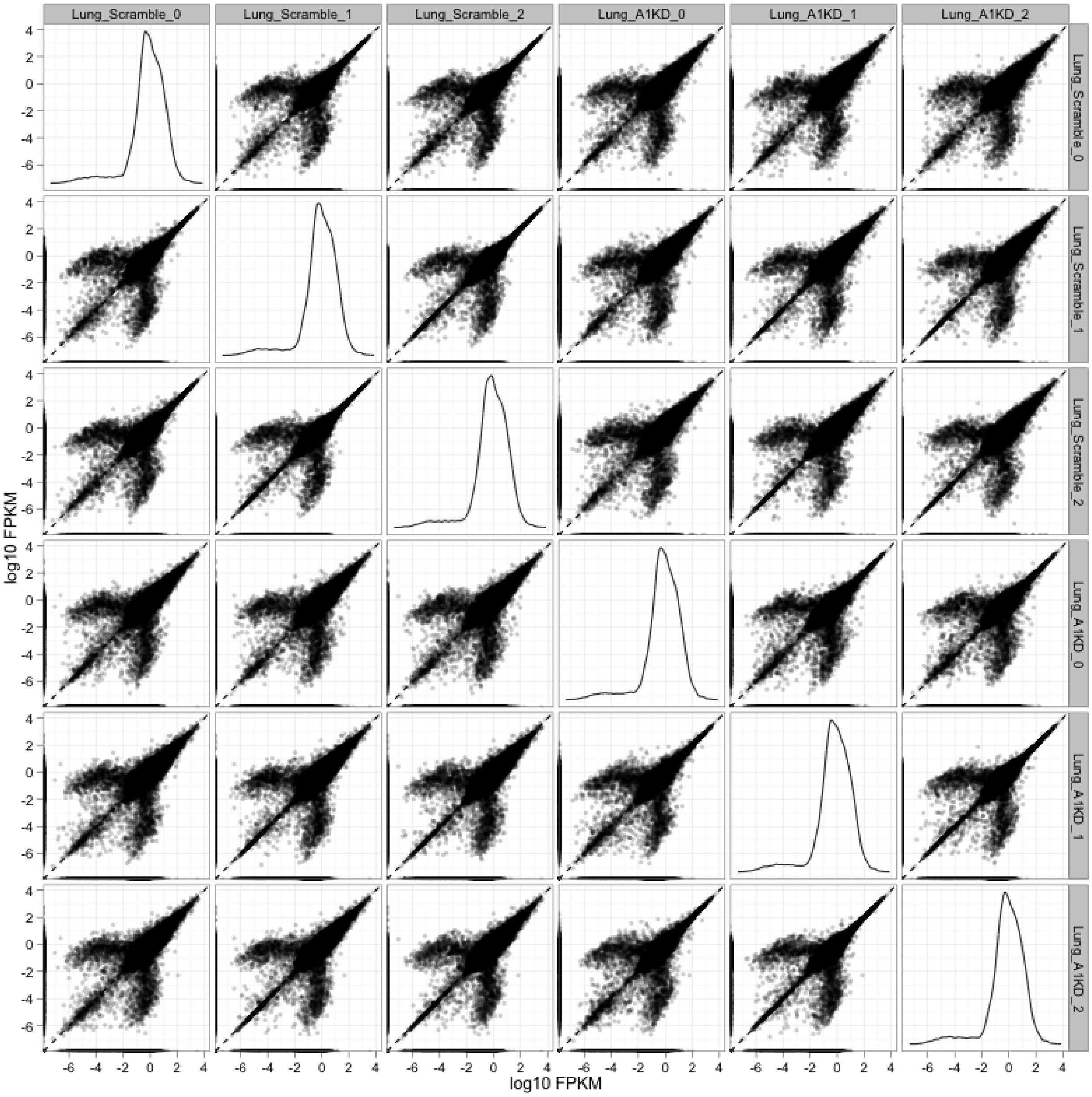

Figure 80. Scatterplot matrix across replicates of transcript level FPKM values in the fibroblast experiment.

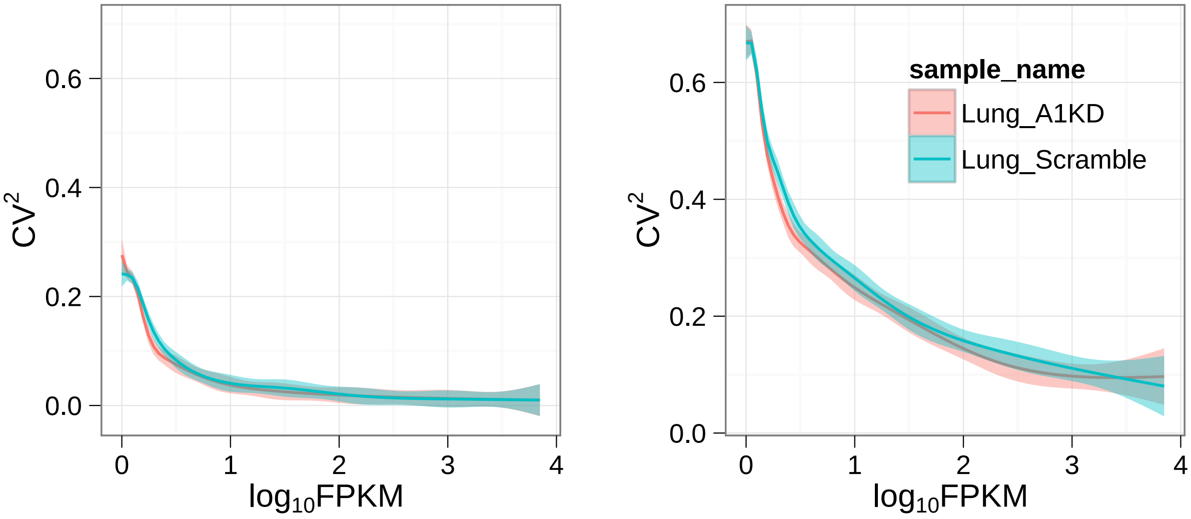

FIGURE 81. Squared coefficient of variation in FPKM values as a function of mean FPKM across replicates for genes (left) and transcripts (right).

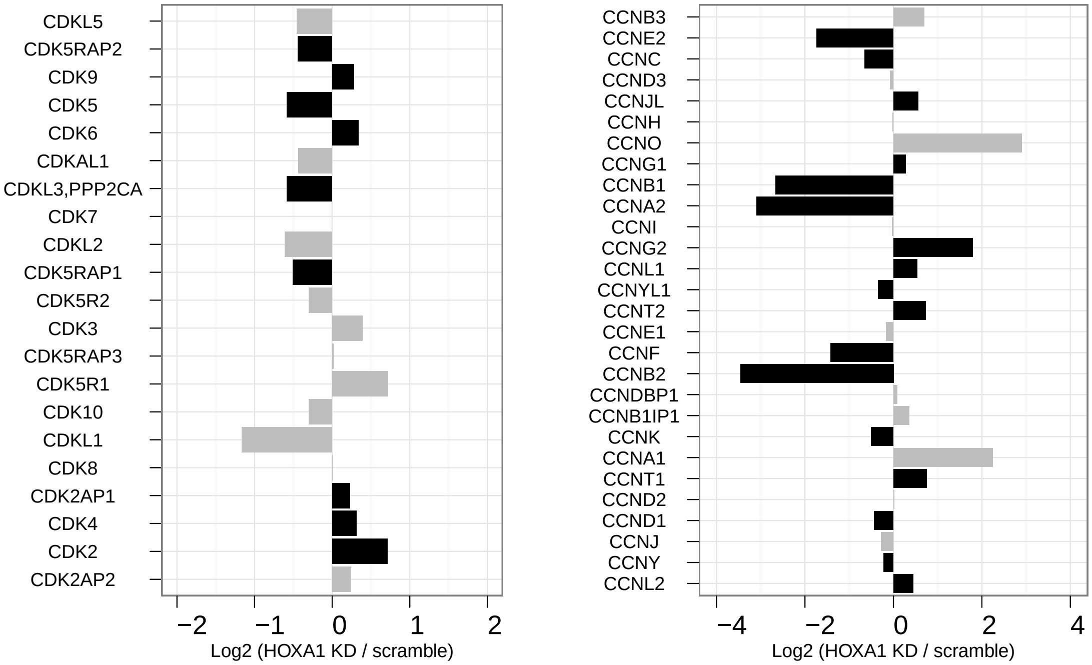

FIGURE 82. Numerous cyclins (right) and cyclin-dependent kinases (left) are significantly differentially expressed in response to HOXA1 knockdown. Black bars were declared significant, while gray bars were not.

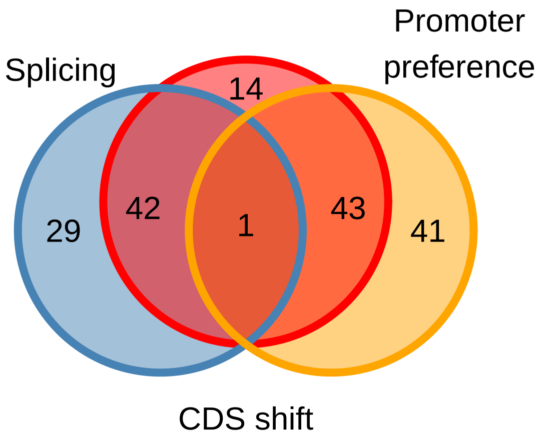

Figure 83. HOXA1 knockdown induces differential splicing and promoter preference shifts in hundreds of genes. These shifts often, but not always, involve isoforms of the same gene that have distinct coding sequences (CDS).

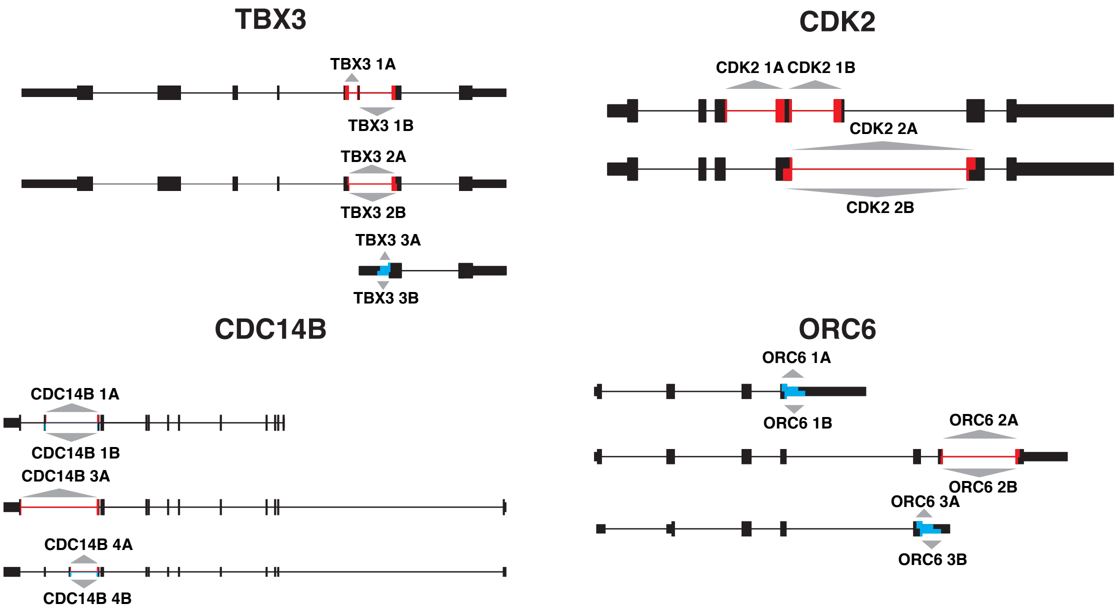

Figure 84. Isoform-specific qPCR probe designs for genes in Figure 5. Red probes indicate that at least one primer spans a distinguishing exon junction. Blue probes have primers that target distinguishing features, but do not span exon splice junctions, and may amplify primary transcript or genomic DNA.

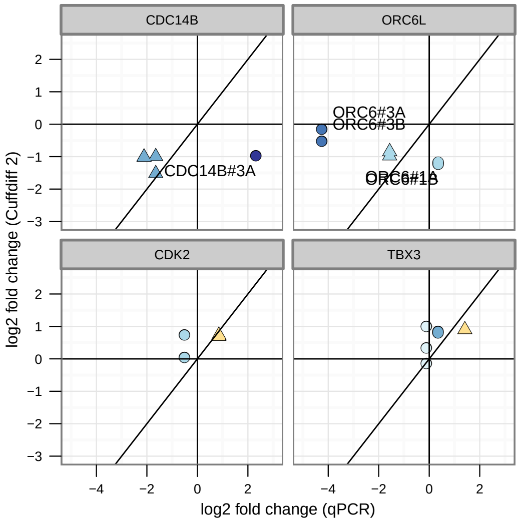

Figure 85. Cuffdiff 2 isoform-level log2 fold changes in expression compared to qPCR-based measurements. The plot is faceted by gene, and colored according to expression decile (with dark blue at the lowest expression decile, and dark red at the highest). Triangular points indicate that Cuffdiff 2 determined the isoform was significantly differentially expressed ($p \leq 0.01$). Circles indicate that Cuffdiff 2 did not report the isoform as significantly DE. Outliers are labeled, and we note that they correspond to probes that target primary transcript or isoforms at lower levels of expression (FPKM ≤ 2.0)

.

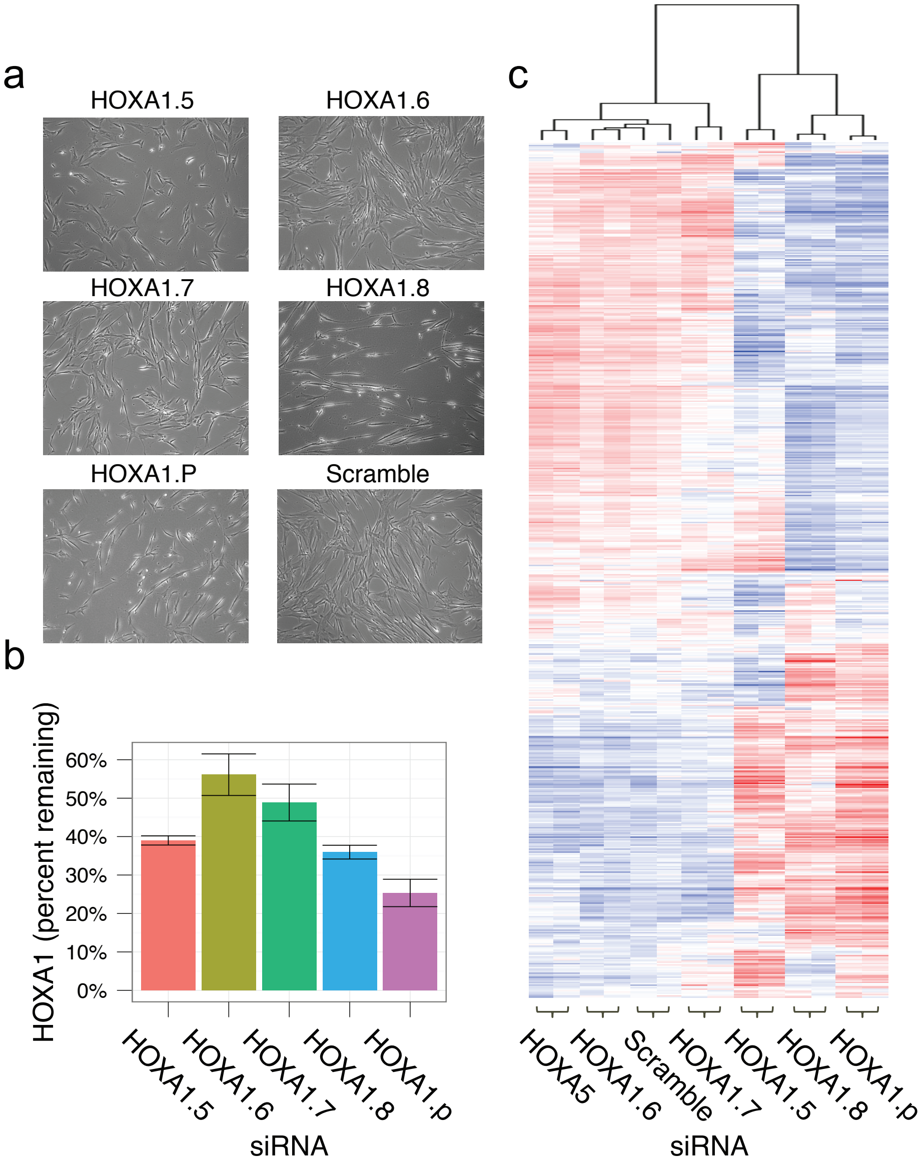

Figure 86. The individual HOXA1 siRNAs HOXA1.5 and HOXA1.8 recapitulate phenotype, knockdown efficiency, and gene expression signature of the combined siRNA pool HOXA1.P (a) Human lung fibroblasts transfected with individual HOXA1 siRNAs (HOXA1.5-8), a combined pool (HOXA1.P), and a scramble control. (b) HOXA1 knockdown efficacy of individual HOXA1 siRNAs and a combined pool assayed for with HOXA1 taqman QPCR. Percent remaining was calculated by normalizing relative initial concentrations to TBP and comparing to the TBP normalized scramble control. (c) Row-centered heatmap of log2 microarray intensity values for differentially expressed genes (10% FDR) post transfection with the HOXA1.P siRNA pool.

Nature Biotechnology: doi:10.1038/nbt.2450

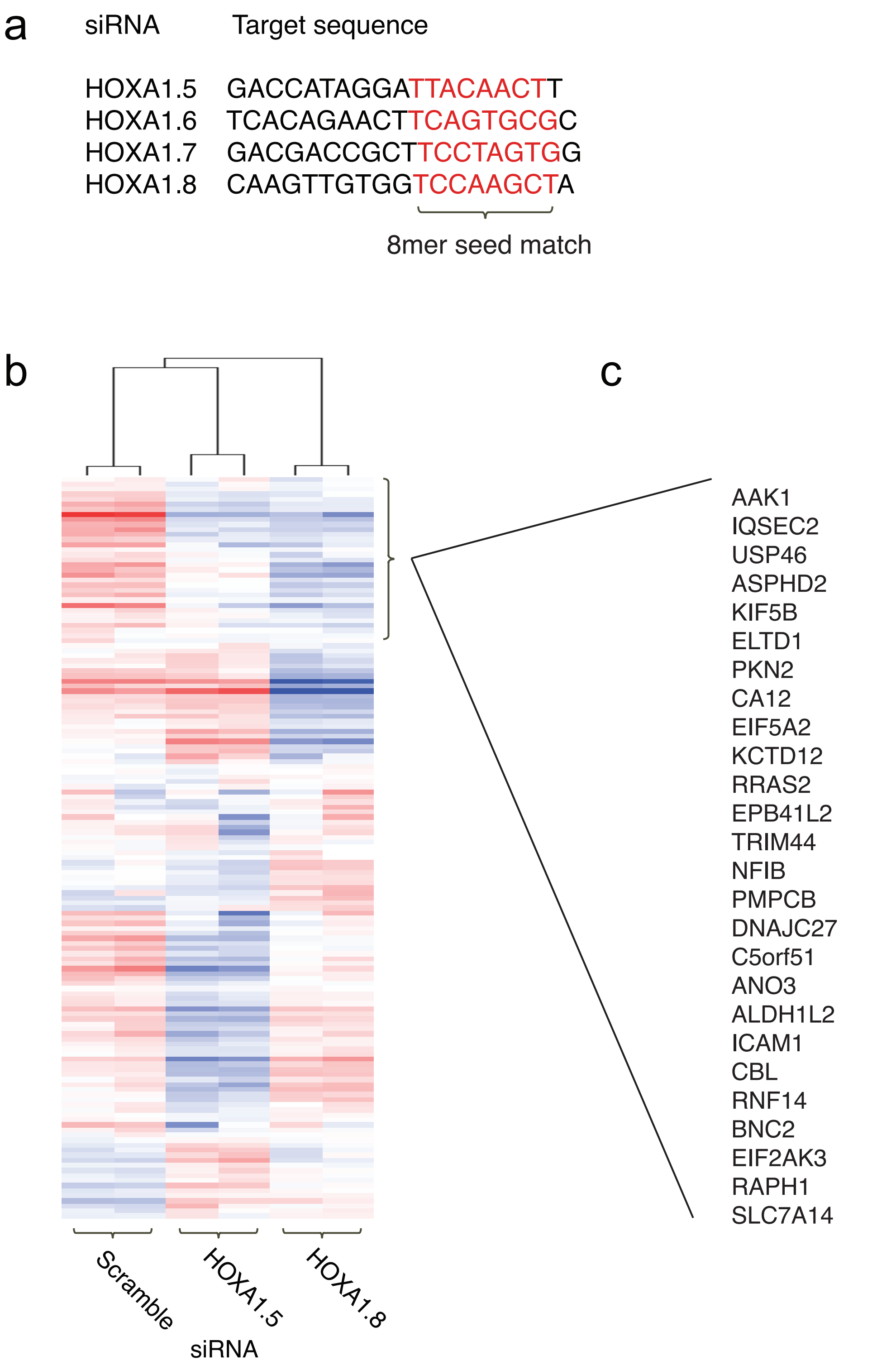

FIGURE 87. (a) Target sequences for individual HOXA1 siRNAs do not contain overlapping seed matches (red highlight). (b) Row-centered heatmap of log2 microarray intensity values for the 147 genes that contain seed matches to both the HOXA1.5 and HOXA1.8 siRNAs. (c) Subset of the 147 (above) genes that decrease significantly in response to transfection with either the HOXA1.5 and HOXA1.8 siRNAs.

Appendix A. Probability distributions

In this Appendix, for completeness, we review the families of probability distributions we work with.

Notation. The negative binomial distribution $NB(r, p)$ is defined by the distribution

$$f(k; n, p) \quad = \quad \binom{k+r-1}{r-1}(1-p)^k p^r \tag{A.1}$$

where k is a non-negative integer, $r > 0$ and $p \in (0, 1)$. The mean of the distribution is given by

$$\mu \quad = \quad \frac{r(1-p)}{p} \tag{A.2}$$

and the variance is

$$\sigma^2 \quad = \quad \frac{r(1-p)}{p^2}. \tag{A.3}$$

We will make use of two different (equivalent) interpretations of the negative binomial distribution. The first, is that if p is the probability of "success" of of a coin toss, then $f(k; n, p)$ is the probability of observing exactly k *failures* in the first $k + r - 1$ trials and then *success* in the $(k + r)$th trial.

The second interpretation is of the negative binomial in terms of mixtures of Poisson random variables where the Poisson parameter is Gamma distributed; we return to this following the definition of the Gamma distribution.

Notation. Note that the negative binomial distribution can be defined equivalently by

$$\tilde{f}(k; n, p) \quad = \quad \binom{k+r-1}{k}(1-p)^r p^k, \tag{A.4}$$

where the interpretation of $\tilde{f}$ is that it is the probability of observing exactly k *successes* in the first $k+r-1$ trials and then *failure* in the $k + r$th trial. In what follows, we will require the definition in Equation A.1.

Notation. The Gamma distribution $\Gamma(r, \theta)$ is defined by the distribution

$$f(x; r, \theta) \quad = \quad x^{r-1}\frac{e^{\frac{-x}{\theta}}}{\theta^r \Gamma(r)}, \tag{A.5}$$

where $x > 0$, the parameters $r, \theta > 0$ and $\Gamma(r)$ denotes the Gamma function given by

$$\Gamma(z) \quad = \quad \int_0^\infty t^{z-1}e^{-t}dt. \tag{A.6}$$

The mean of the distribution is given by

$$\mu \quad = \quad r\theta \tag{A.7}$$

and the variance is

$$\sigma^2 \quad = \quad r\theta^2. \tag{A.8}$$

Proposition 2. *The negative binomial distribution has an interpretation as a mixture of Poisson distributions. If X is a random variable such that, for a λ, $X|\lambda \sim Poisson(\lambda)$ where $\lambda \sim \Gamma(r, \theta)$ then $X \sim NB(r, \frac{1}{1+\theta})$. That is, if X is a mixture of Poisson distributions where the parameter λ in each Poisson is Gamma distributed, then the distribution of X is negative binomial.*

Notation. The beta distribution $Be(\alpha, \beta)$ is defined by the distribution

$$f(x; \alpha, \beta) \quad = \quad \frac{1}{B(\alpha, \beta)x^{\alpha-1}(1-x)^{\beta-1}} \tag{A.9}$$

where $B(\alpha, \beta)$ is the beta function, $\alpha, \beta > 0$ and $x \in (0, 1)$. The mean of the distribution is given by

$$\mu \quad = \quad \frac{\alpha}{\alpha+\beta} \tag{A.10}$$

and the variance is

$$\sigma^2 \quad = \quad \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{A.11}$$

Notation. The beta negative binomial distribution $BNB(r,\alpha,\beta)$ is defined by the distribution

$$f(k;r,p) \quad = \quad \frac{r^{(k)}\alpha^{(r)}\beta^{(k)}}{k!(\alpha+\beta)^{(r)}(r+\alpha+\beta)^{(k)}} \tag{A.12}$$

where $x^{(r)} = x(x+1)(x+2)\cdots(x+r-1)$, the parameters $\alpha,\beta > 0$ and r,k are non-negative integers. The mean of the distribution is given by

$$\mu \quad = \quad \frac{r\beta}{\alpha-1} \text{ if } \alpha > 1, \infty \text{ otherwise} \tag{A.13}$$

and the variance is

$$\sigma^2 \quad = \quad \frac{r(\alpha+r-1)\beta(\alpha+\beta-1)}{(\alpha-2)(\alpha-1)^2} \text{ if } \alpha > 2, \infty \text{ otherwise.} \tag{A.14}$$

Proposition 3. *The beta negative binomial distribution has an interpretation as a mixture of negative binomial distributions. If X is a random variable such that, for a parameter p, $X|p \sim NB(r,p)$ where $p \sim Be(\alpha,\beta)$ then $X \sim BNB(r,\alpha,\beta)$. That is, if X is a mixture of negative binomial distributions where the parameter p in each negative binomial is beta distributed, then the distribution of X is beta negative binomial.*

References

1. Simon Anders and Wolfgang Huber, *Differential expression analysis for sequence count data*, Genome biology **11** (2010), no. 10, R106.
2. James H Bullard, Elizabeth Purdom, Kasper D Hansen, and Sandrine Dudoit, *Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments.*, BMC bioinformatics **11** (2010), 94.
3. Thasso Griebel, Benedikt Zacher, Paolo Ribeca, Emanuele Raineri, Vincent Lacroix, Roderic Guigó, and Michael Sammeth, *Modelling and simulating generic RNA-Seq experiments with the flux simulator*, Nucleic Acids Research (2012), 1–11.
4. Malachi Griffith, Obi L Griffith, Jill Mwenifumbo, Rodrigo Goya, A Sorana Morrissy, Ryan D Morin, Richard Corbett, Michelle J Tang, Ying-Chen Hou, Trevor J Pugh, Gordon Robertson, Suganthi Chittaranjan, Adrian Ally, Jennifer K Asano, Susanna Y Chan, Haiyan I Li, Helen McDonald, Kevin Teague, YongJun Zhao, Thomas Zeng, Allen Delaney, Martin Hirst, Gregg B Morin, Steven J M Jones, Isabella T Tai, and Marco A Marra, *Alternative expression analysis by RNA sequencing.*, Nature methods **7** (2010), no. 10, 843–847.
5. Kasper D Hansen, Zhijin Wu, Rafael A Irizarry, and Jeffrey T Leek, *Sequencing technology does not eliminate biological variability.*, Nature Biotechnology **29** (2011), no. 7, 572–573.
6. D Hiller, H Jiang, W Xu, and W H Wong, *Identifiability of isoform deconvolution from junction arrays and RNA-Seq*, Bioinformatics **25** (2009), no. 23, 3056–3059.
7. Hui Jiang and Wing Hung Wong, *Statistical inferences for isoform expression in RNA-Seq.*, Bioinformatics **25** (2009), no. 8, 1026–1032.
8. B Li, V Ruotti, R M Stewart, J A Thomson, and C N Dewey, *RNA-Seq gene expression estimation with read mapping uncertainty*, Bioinformatics **26** (2010), no. 4, 493–500.
9. CR Loader, *Loader: LOCFIT for S-Plus (software and electronic... - Google Scholar*, Murray Hill, 1998.
10. MAQC Consortium, Leming Shi, Laura H Reid, Wendell D Jones, Richard Shippy, Janet A Warrington, Shawn C Baker, Patrick J Collins, Francoise de Longueville, Ernest S Kawasaki, Kathleen Y Lee, Yuling Luo, Yongming Andrew Sun, James C Willey, Robert A Setterquist, Gavin M Fischer, Weida Tong, Yvonne P Dragan, David J Dix, Felix W Frueh, Frederico M Goodsaid, Damir Herman, Roderick V Jensen, Charles D Johnson, Edward K Lobenhofer, Raj K Puri, Uwe Schrf, Jean Thierry-Mieg, Charles Wang, Mike Wilson, Paul K Wolber, Lu Zhang, Shashi Amur, Wenjun Bao, Catalin C Barbacioru, Anne Bergstrom Lucas, Vincent Bertholet, Cecilie Boysen, Bud Bromley, Donna Brown, Alan Brunner, Roger Canales, Xiaoxi Megan Cao, Thomas A Cebula, James J Chen, Jing Cheng, Tzu-Ming Chu, Eugene Chudin, John Corson, J Christopher Corton, Lisa J Croner, Christopher Davies, Timothy S Davison, Glenda Delenstarr, Xutao Deng, David Dorris, Aron C Eklund, Xiao-hui Fan, Hong Fang, Stephanie Fulmer-Smentek, James C Fuscoe, Kathryn Gallagher, Weigong Ge, Lei Guo, Xu Guo, Janet Hager, Paul K Haje, Jing Han, Tao Han, Heather C Harbottle, Stephen C Harris, Eli Hatchwell, Craig A Hauser, Susan Hester, Huixiao Hong, Patrick Hurban, Scott A Jackson, Hanlee Ji, Charles R Knight, Winston P Kuo, J Eugene LeClerc, Shawn Levy, Quan-Zhen Li, Chunmei Liu, Ying Liu, Michael J Lombardi, Yunqing Ma, Scott R Magnuson, Botoul Maqsodi, Tim McDaniel, Nan Mei, Ola Myklebost, Baitang Ning, Natalia Novoradovskaya, Michael S Orr, Terry W Osborn, Adam Papallo, Tucker A Patterson, Roger G Perkins, Elizabeth H Peters, Ron Peterson, Kenneth L Philips, P Scott Pine, Lajos Pusztai, Feng Qian, Hongzu Ren, Mitch Rosen, Barry A Rosenzweig, Raymond R Samaha, Mark Schena, Gary P Schroth, Svetlana Shchegrova, Dave D Smith, Frank Staedtler, Zhenqiang Su, Hongmei Sun, Zoltan Szallasi, Zivana Tezak, Danielle Thierry-Mieg, Karol L Thompson, Irina Tikhonova, Yaron Turpaz, Beena Vallanat, Christophe Van, Stephen J Walker, Sue Jane Wang, Yonghong Wang, Russ Wolfinger, Alex Wong, Jie Wu, Chunlin Xiao, Qian Xie, Jun Xu, Wen Yang, Liang Zhang, Sheng Zhong, Yaping Zong, and William Slikker, *The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements.*, Nature Biotechnology **24** (2006), no. 9, 1151–1161.

11. Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold, *Mapping and quantifying mammalian transcriptomes by RNA-Seq*, Nature methods **5** (2008), no. 7, 621–628.
12. Marius Nicolae, Serghei Mangul, Ion I Măndoiu, and Alex Zelikovsky, *Estimation of alternative splicing isoform frequencies from RNA-Seq data.*, Algorithms for molecular biology : AMB **6** (2011), no. 1, 9.
13. Alicia Oshlack, Mark D Robinson, and Matthew D Young, *From RNA-seq reads to differential expression results.*, Genome biology **11** (2010), no. 12, 220.
14. Adam Roberts, Cole Trapnell, Julie Donaghey, John L Rinn, and Lior Pachter, *Improving RNA-Seq expression estimates by correcting for fragment bias.*, Genome biology **12** (2011), no. 3, R22.
15. Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter, *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.*, Nature Biotechnology **28** (2010), no. 5, 511–515.