Supplemental text, figures, and tables for:

Integrative analysis with ChIP-seq advances the limits of transcript quantification from RNA-seq

Peng Liu, Rajendran Sanalkumar, Emery H. Bresnick, Sündüz Keleş and Colin N. Dewey

Table of Contents

I. Processing RNA-seq and ChIP-seq data

I.A. Processing RNA-seq data

RNA-seq data were obtained from two sources: (i) five human and mouse cell lines from ENCODE (The ENCODE Project Consortium 2012; Stamatoyannopoulos et al. 2012) (Supplemental Table S1); (ii) sixteen cell types from a mouse hematopoietic differentiation study (Lara-Astiaso et al. 2014) (Gene Expression Omnibus accession number GSE60101).

Supplemental Table S1. ENCODE RNA-seq data sets. All data are from whole-cell fractions.

Species	Cell line	Sex	Treatment	DCC ID[1]	RNA extract	Run type	Read length (nt)	Nrep[2]
Human	K562	F	No	ENCSR000CPH	Long polyA+	Paired-end	76	2
	GM12878	F	No	ENCSR000COQ			76	
Mouse	CH12	F	No	ENCSR000CWD	Ribo-Zero-Gold		101	
	MEL	M	No	ENCSR000CWE			101	
	MEL	M	2% DMSO for 5 days	ENCSR000CWF			101	

[1] ENCODE DCC metadata database accession ID (https://www.encodeproject.org)
[2] Number of biological replicates

Transcript annotations were taken from GENCODE human version 19 and mouse version 4 (Harrow et al. 2012). Quantifications were performed on all genes with 'gene_type' annotated as 'protein_coding' and all isoforms from those genes. UCSC genome assemblies hg19 and mm10 were used for human and mouse, respectively.

RNA-seq reads were aligned with STAR v2.4.0h (Dobin et al. 2012) and quantified by RSEM v1.2.15 (Li and Dewey 2011) with command-line options from ENCODE's STAR-RSEM pipeline (manuscript in preparation; see https://github.com/ENCODE-DCC/long-rna-seq-pipeline for source code; the pipeline was implemented in the pRSEM package). Two variants of RSEM estimates were used. By 'RSEM ML', we refer to the maximum likelihood estimates obtained from RSEM's Expectation-Maximization algorithm. By 'RSEM', we refer to the posterior mean estimates obtained from Gibbs sampling with the Bayesian version of RSEM's probabilistic model with an initial pseudo-count of one for every isoform. 'RSEM' is the most comparable variant to pRSEM.

Two variants of pRSEM were used. 'pRSEM' refers to a pRSEM run with the default partition model. 'pRSEM no partition' refers to a pRSEM run where a single prior parameter is learned from all the isoforms in a training set without any partition.

Three variants of eXpress version 1.5.1 (Roberts and Pachter 2013) were used. 'eXpress' denotes an eXpress run under its default settings. 'eXpress O1B10' and 'eXpress O1B100' denote an eXpress run with one round of online EM, followed by ten or one hundred rounds of batch EM, respectively. All eXpress runs were supplied with command-line option '--rf-stranded' to match the orientation of the fragments in the ENCODE RNA-seq data sets we used. The same transcript alignments computed by STAR and used by RSEM and pRSEM were given to eXpress as input after sorting the alignment BAM files by read name (as required by eXpress).

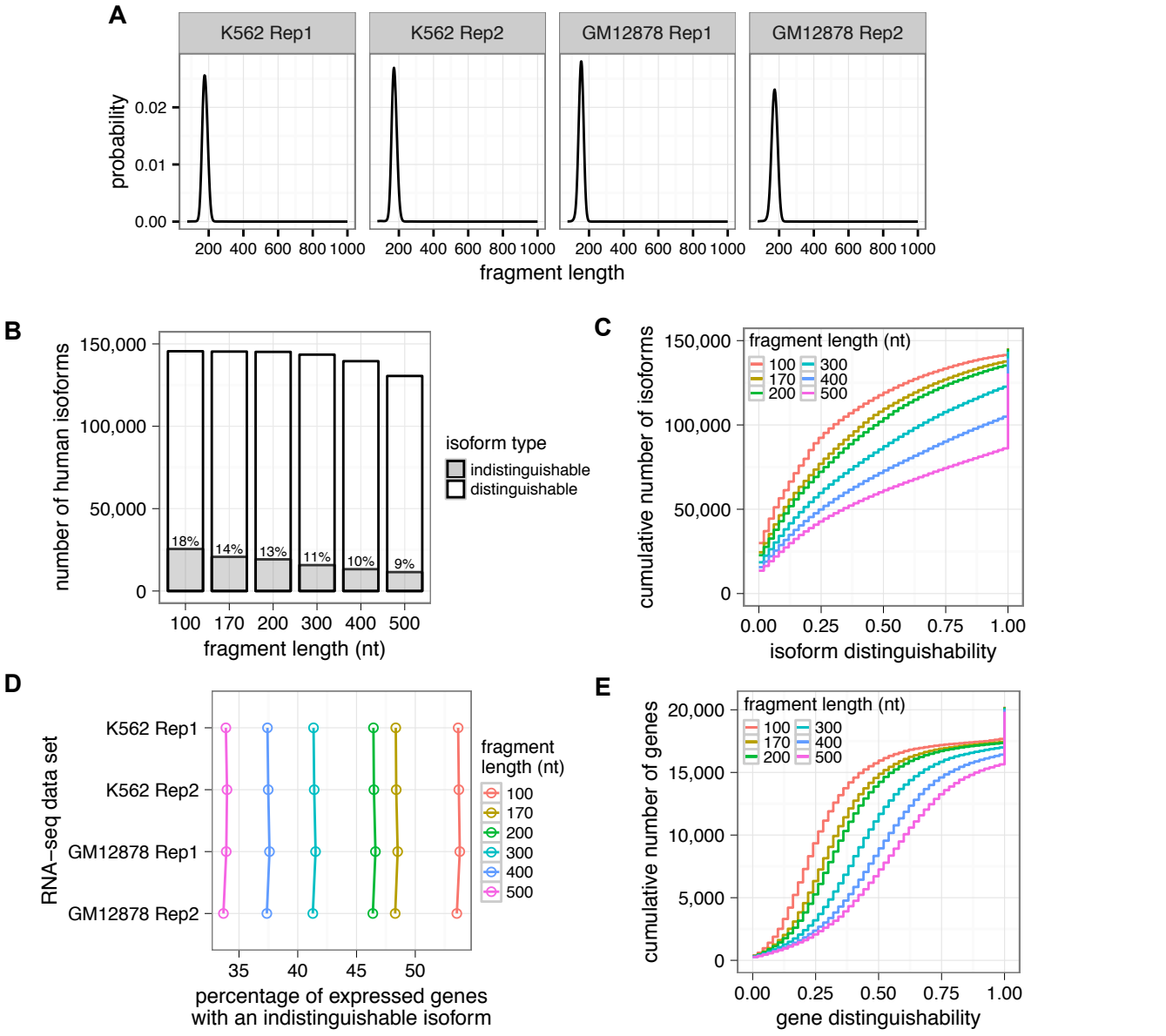
I.B. Distinguishability of isoforms and genes

The distinguishability of an isoform within an RNA-seq experiment depends on the uniqueness of the isoforms' exon(s) and junction(s) as well as the RNA-seq fragment length. Since all ENCODE human and mouse RNA-seq data sets used in this work were paired-end, we first obtained their fragment length distributions with RSEM and took the most probable length for each of them (Supplemental Figure S1A; Supplemental Figure S2A). We then determined a fragment length for each species by taking the average of the most probable length from each data set from that species (170 nt for human, 162 nt for mouse). Next, we enumerated all possible fragments of that length from each isoform and determined which of these could be uniquely mapped back to its parent isoform. Isoforms that were shorter than the fragment length (0.20% of human isoforms; 0.19% of mouse isoforms) were ignored because a numeric value characterizing the distinguishability of such an isoform could not be calculated. We defined an isoform's distinguishability as the ratio of the number of its uniquely mapped fragments over the total number of its fragments. Under this definition, an isoform with zero distinguishability, i.e. an indistinguishable isoform, must have all of its fragments identical to fragments from other isoforms. We defined a gene's distinguishability as the average distinguishability of all of its isoforms. Thus, all the isoforms from an indistinguishable gene are also indistinguishable.

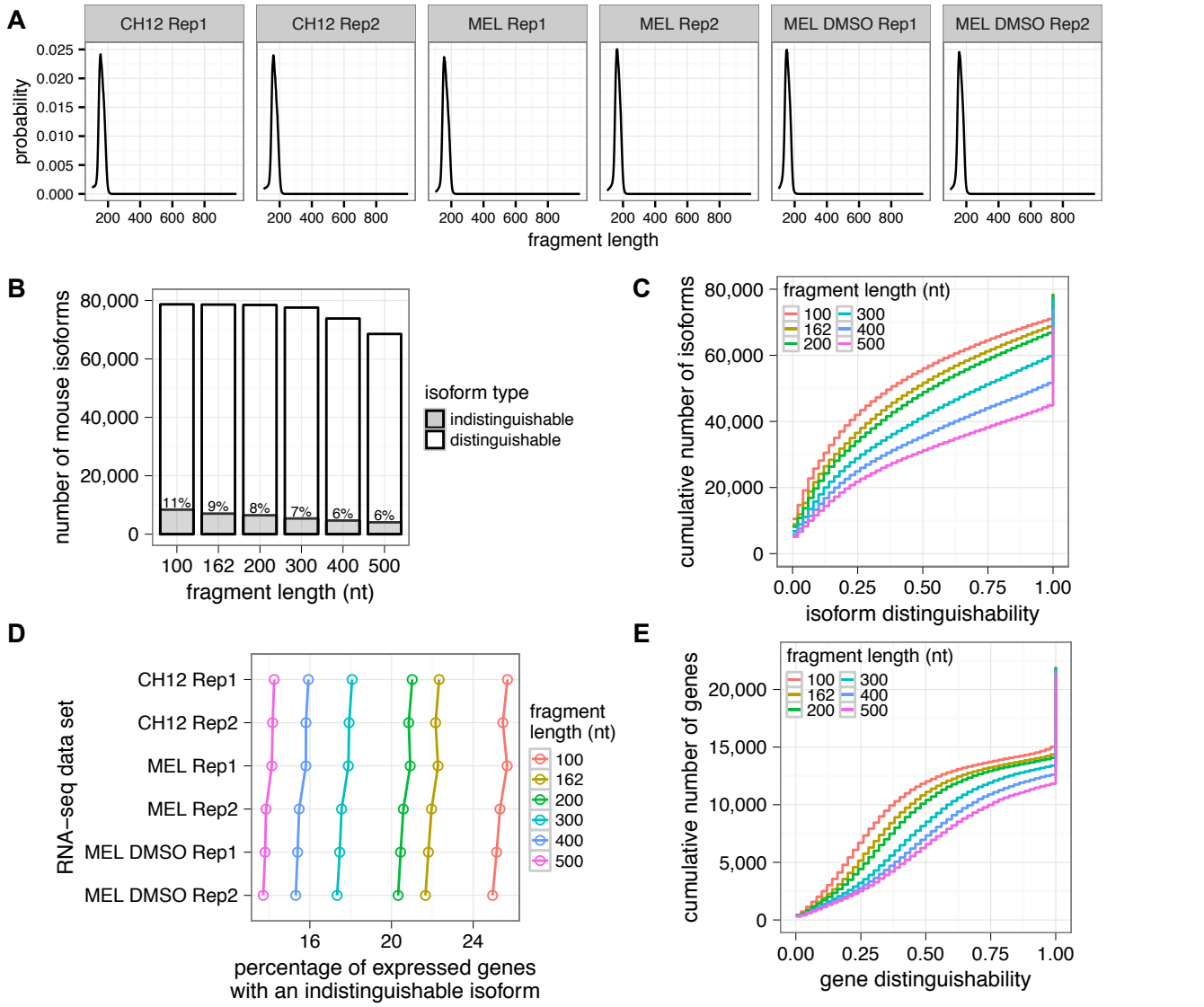
In human, 20,738 isoforms from protein-coding genes are indistinguishable, accounting for 14% of all isoforms (Supplemental Figure S1B). Moreover, there are more than 100,000 isoforms with distinguishability of 0.5 or less (Supplemental Figure S1C). At the gene level, more than 45% of expressed protein-coding genes contain at least one indistinguishable isoform (Supplemental Figure S1D) and almost 15,000 genes have distinguishability of 0.5 or less (Supplemental Figure S1E).

In addition to the average of the most probable fragment length (170 nt for human), we also used a number of other fragment lengths, ranging from 100 nt to 500 nt, to calculate isoform and gene distinguishability. As expected, a longer fragment length led to increased distinguishability at both the isoform- and the gene-level (Supplemental Figure S1, B to D). With a fragment length of 500 nt, the number of indistinguishable isoforms dropped to 11,478, roughly half the number when a fragment length of 170 nt was used (Supplemental Figure S1B). Similarly, the percentage of expressed genes with an indistinguishable isoform decreased to roughly 35% (Supplemental Figure S1D). Nevertheless, the number of indistinguishable isoforms and genes is sizable, even with large fragment lengths. Note that we took 100 nt as the lower bound because a read length of at least 100 nt is common for current RNA-seq experiments and the fragment length is typically longer than the read length. We did not use a fragment length longer than 500 nt, because there are already 15,100 transcripts (10% of all the transcripts) shorter than 500 nt and transcripts shorter than the fragment length were ineligible for the distinguishability calculations.

We performed the same calculations for the mouse genome. The fraction of isoforms or genes having low distinguishability was still substantial (Supplemental Figure S2). With the average of the most probable fragment length (162 nt), 9% of all the isoforms were indistinguishable (Supplemental Figure S2B) and more than 20% of expressed protein-coding genes had an indistinguishable isoform (Supplemental Figure S2D). As we observed in human, increasing the fragment length partially alleviates the low distinguishability issue (Supplemental Figure S2 , C and E).



Supplemental Figure S1. Distinguishability of human isoforms and genes defined by different fragment lengths. (A) Fragment length distributions of ENCODE human paired-end RNA-seq data sets estimated by RSEM; **(B)** Fractions of indistinguishable isoforms; **(C)** Cumulative distributions of distinguishability for all human isoforms; **(D)** Fractions of expressed genes with at least one indistinguishable isoform; **(E)** Cumulative distributions of distinguishability for all human genes.



Supplemental Figure S2. Distinguishability of mouse isoforms and genes defined by different fragment lengths. (A) Fragment length distributions of ENCODE mouse paired-end RNA-seq data sets estimated by RSEM; **(B)** Fractions of indistinguishable isoforms; **(C)** Cumulative distributions of distinguishability for all mouse isoforms; **(D)** Fractions of expressed genes with at least one indistinguishable isoform; **(E)** Cumulative distributions of distinguishability for all mouse genes.

I.C. Processing ChIP-seq data

ChIP-seq data sets were from the same sources as the RNA-seq data. For each ENCODE RNA-seq data set, we obtained ChIP-seq reads for both Pol II and its control (Supplemental Table S2). To test whether Pol II data from unmatched samples could provide an informative prior, we downloaded

Pol II ChIP-seq peak data for four other human cell lines (Supplemental Table S2). For the mouse hematopoietic differentiation data, for each cell type with an RNA-seq data set, we retrieved four types of histone modification ChIP-seq data sets: H3K4me1, H3K4me2, H3K4me3, and H3K27ac (Gene Expression Omnibus accession number GSE59636).

ChIP-seq reads were aligned with Bowtie v1.0.1 (Langmead et al. 2009) with command-line options '-q -v 2 -a --best --strata -m 1'. These options resulted in the reporting of only the best uniquely mapped reads. Since ENCODE Pol II ChIP-seq data sets all have controls, peaks were called by ENCODE's SPP and IDR pipeline (Landt et al. 2012) with an IDR threshold of 0.05. The ChIP-seq signal for a genomic interval was calculated in the same manner as in dPeak (Chung et al. 2013) and normalized by interval length. Due to the lack of ChIP-seq controls for the mouse hematopoietic differentiation samples, peaks were called by HOMER as described previously (Lara-Astiaso et al. 2014). For these samples, the ChIP-seq signal for a genomic interval was computed by counting the number of reads falling within that interval and then normalizing by interval length and read depth so that signals for different histone marks could be integrated. To avoid PCR artifacts resulted from the relatively low-input nature of the primary cell ChIP-seq data, the number of reads aligned to the same genomic interval was kept at a maximum of five per ChIP-seq replicate.

Supplemental Table S2. ENCODE ChIP-seq data sets for human and mouse.

ChIP-seq reads									
Species	**Cell line**	**Sex**	**Treatment**	**DCC ID[1]**	**Target**	**Nrep[2]**	**DCC ID[1]**	**Target**	**Nrep[2]**
Human	K562	F	No	ENCSR000BMR	POLR2A	2	ENCSR000BLJ	RevXlink-Chromatin	2
	GM12878	F	No	ENCSR000BIF	POLR2A-phosphoS5		ENCSR000BGH	RevXlink-Chromatin	5
Mouse	CH12	F	No	ENCSR000ERQ	POLR2A		ENCSR000ERT	IgG-mus	1
	MEL	M	No	ENCSR000EUC	POLR2A		ENCSR000EUF	IgG-mus	1
	MEL	M	2% DMSO for 5 days	ENCSR000ETG	POLR2A		ENCSR000ETD	IgG-rat	1
ChIP-seq peaks									
Human	A549	M	100 nM dexamethasone for 1 hour	ENCFF002CFW	POLR2A				
	H1-hESC	M	No	ENCFF002CJE	POLR2A				
	HeLa-S3	F	No	ENCFF002CJZ	POLR2A				
	HepG2	M	No	ENCFF002CKX	POLR2A				

[1] ENCODE DCC metadata database accession ID (https://www.encodeproject.org)
[2] Number of biological replicates

II. The pRSEM method

II.A. An overview of pRSEM

The framework of pRSEM is built on RSEM, which employs a generative model and an EM algorithm to estimate gene and isoform expression levels (Li and Dewey 2011; Li et al. 2010). Specifically, RSEM models the RNA-seq read sequencing process by taking into account transcript abundances, sequencing error, fragment length variation, read length variation, and read start position non-uniformity. RSEM's implementation also provides a Bayesian version of this model, in which transcript expression levels are modeled as latent variables from a Dirichlet distribution. By default, the prior parameters for the Dirichlet distribution are uniformly set to one (an uninformative prior) so that the maximum a posteriori estimates are equal to RSEM's maximum likelihood estimates. The framework of pRSEM takes advantage of this design and learns informative parameters for the Dirichlet prior using a training set of isoforms partitioned based on an external data set. In this way, pRSEM can leverage external information to supervise the allocation of multi-mapping reads and estimate transcript abundances. A single shared prior parameter is learned for each partition through maximization of the likelihood of a Dirichlet-multinomial model to fit the distributions of the read counts of the training set isoforms. In what follows, we use the derivation of a Pol II ChIP-seq prior as an example to describe these two steps. In addition to Pol II ChIP-seq, other data types, such as histone modification ChIP-seq, can also be used for deriving a prior. As described later in section II.F, pRSEM provides a testing procedure to determine if a given external data set can be used to derive an informative prior. This procedure also computes a score to rank multiple informative external data sets.

II.B. Building and partitioning a training set of isoforms

Given Pol II ChIP-seq data, a training set of isoforms is constructed by first selecting those isoforms that (i) are from single-isoform genes with a genomic span of at least 1,003 nucleotides, which ensures that the 'TSS region', 'body region', and 'TES region' of a gene do not overlap (see below for definitions of 'TSS region', 'body region', and 'TES region'); (ii) have TSSs that are more than 500 nucleotides from the TSS of any other isoform; and (iii) have genomic spans that do not overlap with the span of any other isoform on either strand. These criteria prevent any ambiguity in assigning ChIP-seq peaks and signals to isoforms. We further filter the training set by requiring isoforms to have an average mappability ≥ 0.8 for their TSS regions, body regions, and TES regions, where a 'TSS region' is defined as the 500 nucleotide flanking region of a TSS (5' end), a 'TES region' is defined as the 500 nucleotide flanking region of a transcription end site (TES, i.e. 3' end), and a 'body region' is defined as the genomic span of an isoform excluding its TSS and TES regions. With this filter, we can have high confidence in each segment's Pol II peak calls and signals by using uniquely mapped ChIP-seq reads. Mappability is defined as the alignability of 36-mers calculated by GEM (Derrien et al. 2012).

We have implemented six partition models in pRSEM. **Models I**-**V** are based on a single complementary data set and **Model VI** was developed to utilize information from multiple external data sets. Below we use Pol II ChIP-seq data as the complementary data set to describe **Models I** to **V** and use multiple histone modification ChIP-seq data sets to illustrate **Model VI**.

Model I, uses a binary partition established by the presence or absence of a Pol II peak overlapping with an isoform's TSS region, i.e., a Pol II TSS peak.

Model II. Isoforms are first partitioned as in Model I with the resulting 'no peak' set further partitioned into two subsets via a logistic regression model. This partition scheme was motivated by the bimodal distribution of the fragment counts of 'no peak' isoforms (red line in Figure 2B). The logistic regression model is specified by the predictive equation:

$$\ln\left(\frac{p_{has\_read}}{1-p_{has\_read}}\right) = \beta_0 + \beta_1 log_{10}(R_{GC}) + \beta_2 log_{10}(L_{eff}) + \beta_3 log_{10}(S_{TSS}) + \beta_4 I_{body} + \beta_5 I_{body} log_{10}(S_{body}) + \beta_6 (1 - I_{body}) log_{10}(S_{body}) + \beta_7 I_{TES} + \beta_8 I_{TES} log_{10}(S_{TES}) + \beta_9 (1 - I_{TES}) log_{10}(S_{TES})$$

where p_{has_read} is the probability of an isoform having a non-zero RNA-seq read count; R_{GC} is the ratio of an isoform's GC content over the mean GC content of all isoforms in the training set; L_{eff} is an isoform's effective length; S_{TSS}, S_{body}, and S_{TES} are the means of the Pol II ChIP-seq signal within an isoform's TSS region, body region, and TES region, respectively; I_{body} and I_{TES} are Boolean variables representing whether an isoform has a Pol II peak overlapping with its body region and TES region, respectively. The β_0 to β_9 are the intercept and coefficients obtained from fitting the logistic regression model to the training set. After fitting, isoforms in the 'no peak' set are divided based on whether or not their estimated p_{has_read} is less than 0.5.

Model III. This model is very similar to Model II with the difference being that instead of logistic regression, a linear regression model of the same form is used to divide the 'no peak' set. Instead of p_{has_read}, the model predicts an isoform's log read count. Isoforms in the 'no peak' set are binned based on their predicted read count with the upper bound of bin i given by:

$$b_i = \frac{i}{n} log_{10}\left(\frac{c^{max}}{c^{min}}\right) + log_{10}(c^{min})$$

where n is a user-defined number of bins, and c^{max} and c^{min} are the largest and smallest predicted read count, respectively. The interval for each bin is half-closed, $(b_{i-1}, b_i]$, except the interval for the first bin, which is $[\log_{10}(c^{min}), b_1]$.

Model IV. This model is the same as Model III, except that the 'with peak' set is further subdivided instead of the 'no peak' set.

Model V. Like Models III and IV, Model V uses a linear regression model to bin the isoforms, but unlike all other models, it does not initially partition by Pol II TSS peak. The predictive equation for this model is:

$$log_{10}(c) = \beta_0 + \beta_1 log_{10}(R_{GC}) + \beta_2 log_{10}(L_{eff}) + \beta_3 I_{TSS} + \beta_4 I_{TSS} log_{10}(S_{TSS}) + \beta_5 (1 - I_{TSS}) log_{10}(S_{TSS}) + \beta_6 I_{body} + \beta_7 I_{body} log_{10}(S_{body}) + \beta_8 (1 - I_{body}) log_{10}(S_{body}) + \beta_9 I_{TES} + \beta_{10} I_{TES} log_{10}(S_{TES}) + \beta_{11} (1 - I_{TES}) log_{10}(S_{TES})$$

where c is an isoform's read count; I_{TSS} is a Boolean variable representing whether or not an isoform has a Pol II TSS peak; the β_0 to β_{11} are the intercept and coefficients to be fitted; and all the other variables are the same as in the **Model II** equation.

Model VI. This model was developed to combine signals from multiple external data sets. Assuming we would like to utilize n types of histone modification ChIP-seq data, we use a logistic regression model specified by the predictive equation:

$$\ln\left(\frac{p_{expressed}}{1-p_{expressed}}\right) = \beta_0 + \sum_{i=1}^{n} \beta_i log_{10}\left(S_i^{TSS}\right)$$

where $p_{expressed}$ denotes the probability of an isoform being expressed, S_i^{TSS} is the ith type of histone modification signal for isoform's TSS region, and the β_0 to β_n are the intercept and coefficients obtained by fitting the model to the training set. After fitting, all isoforms are partitioned by whether $p_{expressed}$ is higher than 0.5 or not.

II.C. Learning prior parameters through a Dirichlet-multinomial model

Given a training set and partitioning of the isoforms, we use a Dirichlet-multinomial model and RSEM's posterior mean estimates to learn prior parameters for the partitions. Let T be the training set of isoforms, and $n_T = |T|$. Let c_i be the initial RNA-seq read count estimate for the ith isoform and $n_c = \sum_{i=1}^{n_T} c_i$ be the total number of reads initially assigned to isoforms in T. Let n_A denote the number of partitions and $f: [1, n_T] \rightarrow [1, n_A]$ denote the mapping from transcript indices to partition indices. The kth partition is associated with the Dirichlet parameter α_k, which is shared by all n_{Tk} isoforms in that partition. The parameters $\boldsymbol{\alpha}=\{\alpha_k \mid k \in [1, n_A]\}$, are what we learn from the training set.

Let p_i denote the prior probability that a read originates from the ith isoform. We assume that the probabilities $\boldsymbol{p}=\{p_i \mid i \in [1, n_T]\}$ follow a Dirichlet distribution parameterized by $\boldsymbol{\alpha}$ and the partition function f. The read count $\boldsymbol{c}=\{c_i \mid i \in [1, n_T]\}$ then follows a multinomial distribution parameterized by $\boldsymbol{p}$ and n_c. Given read counts $\boldsymbol{c}$ for T, the log likelihood is:

$$\begin{aligned}
\ln\left(Pr(\boldsymbol{c}|\boldsymbol{\alpha})\right) &= \ln\left(\int Pr(\boldsymbol{c}|\boldsymbol{p}) Pr(\boldsymbol{p}|\boldsymbol{\alpha})\, d\boldsymbol{p}\right) \\
&= \ln\left(\int \frac{\boldsymbol{\Gamma}(n_c+1)}{\prod_{i=1}^{n_T}\boldsymbol{\Gamma}(c_i+1)} \prod_{i=1}^{n_T} p_i^{c_i} \frac{\boldsymbol{\Gamma}\left(\sum_{i=1}^{n_T}\alpha_{f(i)}\right)}{\prod_{i=1}^{n_T}\boldsymbol{\Gamma}\left(\alpha_{f(i)}\right)} \prod_{i=1}^{n_T} p_i^{\alpha_{f(i)}-1}\, d\boldsymbol{p}\right) \\
&= \ln\left(\frac{\boldsymbol{\Gamma}(n_c+1)\boldsymbol{\Gamma}\left(\sum_{i=1}^{n_T}\alpha_{f(i)}\right)}{\prod_{i=1}^{n_T}\left(\boldsymbol{\Gamma}(c_i+1)\boldsymbol{\Gamma}\left(\alpha_{f(i)}\right)\right)} \int \prod_{i=1}^{n_T} p_i^{c_i+\alpha_{f(i)}-1}\, d\boldsymbol{p}\right) \\
&= \ln\left(\frac{\boldsymbol{\Gamma}(n_c+1)\boldsymbol{\Gamma}\left(\sum_{i=1}^{n_T}\alpha_{f(i)}\right)}{\boldsymbol{\Gamma}\left(n_c+\sum_{i=1}^{n_T}\alpha_{f(i)}\right)} \prod_{i=1}^{n_T} \frac{\boldsymbol{\Gamma}\left(c_i+\alpha_{f(i)}\right)}{\boldsymbol{\Gamma}(c_i+1)\boldsymbol{\Gamma}\left(\alpha_{f(i)}\right)}\right) \\
&= \ln\left(\frac{\boldsymbol{\Gamma}(n_c+1)\boldsymbol{\Gamma}\left(\sum_{k=1}^{n_A} n_{Tk}\alpha_k\right)}{\boldsymbol{\Gamma}\left(n_c+\sum_{k=1}^{n_A} n_{Tk}\alpha_k\right)} \prod_{i=1}^{n_T} \frac{\boldsymbol{\Gamma}\left(c_i+\alpha_{f(i)}\right)}{\boldsymbol{\Gamma}(c_i+1)\boldsymbol{\Gamma}\left(\alpha_{f(i)}\right)}\right)
\end{aligned}$$

where $\boldsymbol{\Gamma}(x)$ is the gamma function and d$\boldsymbol{p}$ denotes integrating $\{p_i \mid i \in [1, n_T]\}$ over the simplex.

We learn $\boldsymbol{\alpha}$ via maximum likelihood estimation. For this optimization we employ a bound constrained BFGS algorithm to search over positive values of α. The algorithm requires the gradient of the log-likelihood, which is given by:

$$\frac{\partial}{\partial \alpha_j}\ln\left(Pr(\boldsymbol{c}|\boldsymbol{\alpha})\right) = \frac{\partial}{\partial \alpha_j}\ln\left(\frac{\boldsymbol{\Gamma}\left(\sum_{k=1}^{n_A} n_{Tk}\alpha_k\right)}{\boldsymbol{\Gamma}\left(n_c + \sum_{k=1}^{n_A} n_{Tk}\alpha_k\right)}\frac{\prod_{i:f(i)=j}\boldsymbol{\Gamma}\left(c_i + \alpha_j\right)}{\left(\boldsymbol{\Gamma}\left(\alpha_j\right)\right)^{\boldsymbol{n_{Tj}}}}\right)$$

$$= n_{Tj}\left(\boldsymbol{\Psi}\left(\sum_{k=1}^{n_A} n_{Tk}\alpha_k\right) - \boldsymbol{\Psi}\left(n_c + \sum_{k=1}^{n_A} n_{Tk}\alpha_k\right) - \boldsymbol{\Psi}\left(\alpha_j\right)\right) + \sum_{i:f(i)=j}\boldsymbol{\Psi}(c_i + \alpha_j)$$

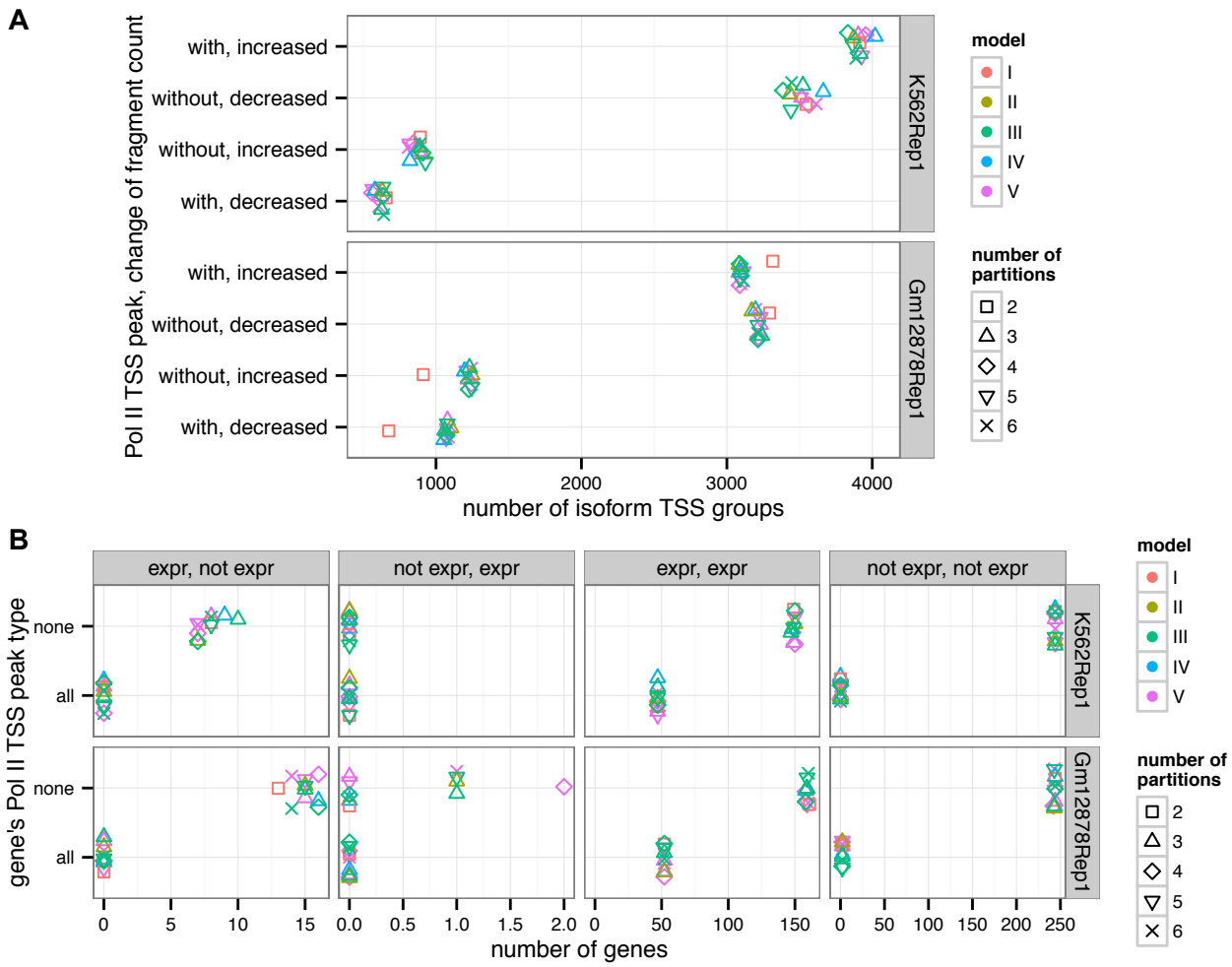
where $\boldsymbol{\Psi}(x)$ is the digamma function.

II.D. Comparison of partition models for a single complementary data set

We evaluated the first five models described in section II.B. on the K562 and GM12878 datasets. Models were compared in terms of: (i) the number of isoforms that had fragment count changes in agreement with their Pol II TSS peak status; and (ii) the number of genes that had expression status prediction changes in concordance with their Pol II TSS peak status. In the evaluation for the first metric, isoform TSS groups were selected in the same way as for Figure 3A. Based on their Pol II TSS peak status and change of counts, isoform TSS groups were classified into four categories: (i) with TSS peak and have fragment count increased ('with, increased'); (ii) without TSS peak and have fragment count decreased ('without, decreased'); (iii) without TSS peak and have fragment count increased ('without, increased'); (iv) with TSS peak and have fragment count decreased ('with, decreased'). After using a Pol II prior, we assumed that 'with peak' isoform TSS groups would have counts increased, whereas 'no peak' isoform TSS groups were more likely to have counts decreased. Thus, a good partition model would have a large number of isoform groups in (i) and (ii) and a small number in (iii) and (iv). No model was found to be overwhelmingly better than the others (Supplemental Figure S3A).

For the second metric, evaluation methods and gene selection were similar to those used for Figure 4B. We classified genes into four categories: (i) estimated as 'expressed' by RSEM, but not by pRSEM ('expr, not expr'); (ii) estimated as 'not expressed' by RSEM, but as 'expressed' by pRSEM ('not expr, expr'); (iii) estimated as 'expressed' by both RSEM and pRSEM ('expr, expr'); (iv) estimated as 'not expressed' by both RSEM and pRSEM ('not expr, not expr'). All comparisons were carried out on the first replicate of the K562 and GM12878 RNA-seq data sets. Model I and II have two and three partitions by definition, respectively. For Model III, four numbers of bins (2, 3, 4, 5) were applied to the 'no peak' set. For Model IV, we only divided the 'with peak' set into two bins, because larger numbers of bins resulted in one bin containing just a single isoform. For Model V, there were four numbers of bins (3, 4, 5, 6) applied on the whole training set. No model was found to be overwhelmingly better than the others with regard to this metric as well (Supplemental Figure S3B).

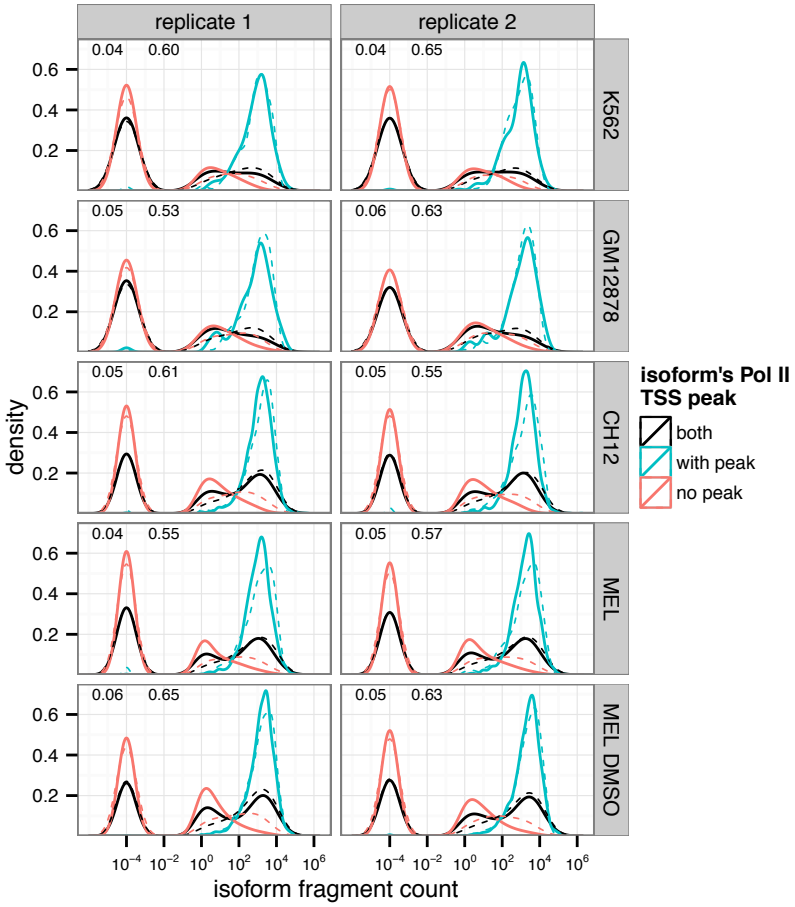
Therefore, we chose to use the simplest partition model (**Model I,** by Pol II TSS peak) as the default in pRSEM and for the remainder of the experiments in this work. Although the presented partition models were designed with Pol II ChIP-seq data in mind, they may be applicable to other types of external data that are informative regarding isoform abundances. For example, we expect that transcription factor ChIP-seq data, as well as other types of transcript sequencing data, e.g. RAMPAGE, can be used with these models to partition the training set of isoforms.



Supplemental Figure S3. No partition model was found to outperform the others. (A) Partition models were compared in terms of the number of isoform TSS groups that had a change of fragment count agree or disagree with their Pol II TSS peak status. **(B)** Partition models were compared by the number of genes that had their expression states changed after using pRSEM.

II.E. Learning priors from ENCODE human and mouse data

We applied pRSEM to five human and mouse cell lines from ENCODE. Every cell line had both RNA-seq and Pol II ChIP-seq data available (Supplemental Table S1 and Supplemental Table S2). We partitioned training set isoforms by whether they had a Pol II TSS peak or not (**Model I**). The pRSEM-learned prior fit the training set data well for all cell lines (Supplemental Figure S4).



Supplemental Figure S4. Pol II TSS peak data are informative for deriving pRSEM priors across cell lines and species. Empirical and fitted distributions of fragment counts for pRSEM training set isoforms, stratified by Pol II TSS peak status, for five human and mouse cell lines. Plots were generated in the same manner as Figure 2B. The estimated Dirichlet prior parameters (without peak, with peak) are shown in the top left of each panel.

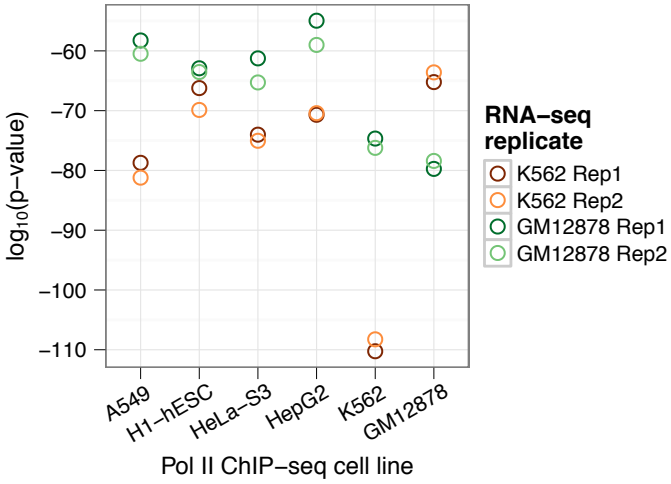
II.F. A testing procedure to select and compare complementary data sets

Investigators may not always have a Pol II ChIP-seq data set from the same condition as their RNA-seq data. In such cases, users may wish to know whether an unmatched external data set could be used for RNA-seq quantification with pRSEM. In pRSEM, we have implemented a testing procedure that provides users with two metrics regarding: (i) whether an external data set can provide an informative prior; and (ii) among multiple external data sets, which one is most informative. The first metric is a p-value indicating whether external information can significantly separate high read-count isoforms from low or zero read-count isoforms in the training set based on a Mann-Whitney test. The second metric is a log-likelihood calibrating how well the prior derived from a complementary data set fits the training set data.

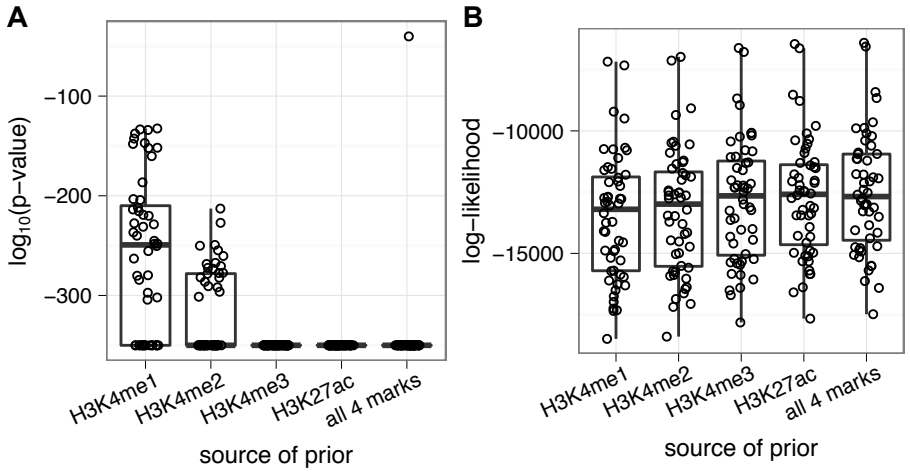
To demonstrate the use of these two metrics, we first considered ENCODE Pol II ChIP-seq data from six human cell lines (Supplemental Table S2) and applied them to the four RNA-seq data sets from the K562 and GM12878 cell lines (Supplemental Table S1). pRSEM's testing procedure resulted in p-values lower than 10^{-50} for all six Pol II data sets (Supplemental Figure S5) and indicated that they are all informative regardless of whether they are from the same condition or not. Comparison of log-likelihoods showed that Pol II data from the same cell line as the RNA-seq data always fit the training set the best (Figure 6A). For each of the two K562 RNA-seq replicates, the prior derived from K562 ChIP-seq data gave the highest log-likelihood compared to priors from the other five ChIP-seq data

sets, which were collected from cell lines other than K562 (the two panels on the left of Figure 6A). Similarly, priors derived from GM12878 Pol II ChIP-seq data have the highest log-likelihood for each of the two GM12878 RNA-seq replicates (the two panels on the right of Figure 6A). We noticed that there was a large difference in the log-likelihoods for the two GM12878 RNA-seq replicates. This is most likely because the second replicate had 25% more aligned RNA-seq fragments than the first replicate (78 million vs. 62 million) and the log-likelihood scales with the number of aligned RNA-seq fragments. In contrast, the two K562 RNA-seq replicates had similar read depth (67 million vs. 64 million). As a result, the log-likelihoods for the two K562 replicates are relatively close to each other. We would like to point out that pRSEM's testing procedure was developed for comparing different sources of prior on the same RNA-seq data set. For a given RNA-seq data set, the number of aligned RNA-seq fragment is a constant regardless of which source was used for the prior. Therefore, the fact that the log-likelihood scales with the number of aligned RNA-seq fragments is not an issue. In cases where RNA-seq data sets from multiple biological replicates are available and pRSEM users would like to select the best source of prior, our tests on the GM12878 and K562 RNA-seq data demonstrated that pRSEM's testing procedure would provide consistent results between RNA-seq replicates.

Supplemental Figure S5. Pol II ChIP-seq data from a different cell line can provide an informative prior for RNA-seq quantification. Comparison of p-values from the Mann-Whitney test on partitioned training set isoform fragment counts. Isoforms were divided into two groups based on their Pol II TSS peak status. Fragment counts were quantified on the two RNA-seq replicates from each of the K562 and GM12878 data sets.



Next, we applied pRSEM to 52 RNA-seq data sets from the sixteen mouse hematopoiesis cell types (Lara-Astiaso et al. 2014). Each cell type had four types of histone modification ChIP-seq data sets available, allowing us to evaluate if histone data could inform RNA-seq quantification and, of the



Supplemental Figure S6. Histone modification ChIP-seq data is informative for RNA-seq quantification. (A) Comparison of p-values from Mann-Whitney tests on partitioned training set isoform read counts. Isoforms were divided into two groups by histone TSS peak status or by a logistic model that utilizes all four types of histone modification ChIP-seq data; **(B)** Comparison of the log-likelihoods from the fit of pRSEM's Dirichlet-multinomial model on training set isoforms.

four marks, which was most informative. All four marks had p-values lower than 10^{-100} (Supplemental Figure S6A), indicating that they are all informative for RNA-seq multi-read allocation. The log-likelihoods obtained from using H3K4me3 and H3K27ac data were systematically better than those from the other histone marks (Supplemental Figure S6B), suggesting that these two marks are more informative. In addition, we developed a logistic model that utilizes information from all four types of histone modifications. A partition derived from the four marks combined resulted in a lower p-value and better log-likelihood (Supplemental Figure S6, A and B). For 48 out of 52 RNA-seq data sets, the combined model provided the highest log-likelihood (Figure 6B), suggesting that it is the most informative one. In summary, our testing procedure indicated that histone modification ChIP-seq data can be used to derive a prior for pRSEM and that integrating multiple histone mark data results in a prior that is generally better than one derived from a single mark.

II.G. Computational requirements

We performed experiments to measure the computational requirements for pRSEM, RSEM (PME), RSEM maximum likelihood, and three eXpress variants (Supplemental Table S3). The relative ordering of pRSEM and the RSEM variants in terms of running time was: RSEM ML < RSEM < pRSEM. RSEM is slower than RSEM ML because it additionally runs ten thousand rounds of Gibbs sampling. pRSEM is slower because it has to learn the prior parameters and run an additional set of Gibbs sampling rounds compared to RSEM. When using three CPUs on the human K562 data, RSEM's standard Gibbs sampling requires 9.3 hours of computing time (Supplemental Table S3, 19.5 vs. 10.2). Under the same settings, pRSEM's prior-learning and additional Gibbs sampling takes 9.4 hours to complete (Supplemental Table S3, 28.9 vs. 19.5), indicating that the prior-learning process is very fast and most of pRSEM's extra running time is spent on Gibbs sampling. Benefitting from RSEM's highly parallelized Gibbs sampling, this extra time can be reduced by half (4.5 hours) when running on eight CPUs (Supplemental Table S3, 14.1 vs. 9.6). Note that the running times for ChIP-seq peak calling were not included here because we assume that users will have peaks called once they obtained their ChIP-seq data. Also, we did not include the computational requirements for aligning RNA-seq reads since these will be the same for all methods shown here.

Compared to pRSEM and RSEM, eXpress with its default settings runs markedly faster (Supplemental Table S3). Running an additional ten rounds of batch EM with eXpress is still quicker than pRSEM if same number of CPUs were used. However, eXpress has limited parallelization abilities and can only use up to three CPUs in its current implementation. As a result, running an additional hundred rounds of batch EM on the human K562 data set required about a week. In contrast, running on eight CPUs—a very common configuration for workstations nowadays, pRSEM completes faster than eXpress with ten rounds of batch EM. Moreover, given that the three eXpress variants performed poorly in our qRT-PCR validations (Supplemental Figure S9 and Supplemental Figure S10 in section III.C), pRSEM compares favorably when considering both time-cost and accuracy.

Supplemental Table S3. Comparison of computational requirements for pRSEM, two variants of RSEM, and three variants of eXpress. All jobs ran on AMD 2.1GHz CPUs. Pol II TSS peak data was used to provide a prior for pRSEM.

Method	Number of CPUs	Mouse MEL[1]		Human K562*	
		Time (hours)	Memory (Gbytes)	Time (hours)	Memory (Gbytes)

pRSEM	3	13.1	2.9	28.9	9.0
RSEM	3	9.3	2.9	19.5	9.0
RSEM ML	3	5.3	2.9	10.2	8.5
eXpress	3	0.9	2.7	1.8	4.2
eXpress O1B10	3	7.3	2.7	17.4	4.2
eXpress O1B100	3	63.4	2.7	158.7	4.2
pRSEM	8	6.9	3.4	14.1	9.0
RSEM	8	5.2	3.4	9.6	9.0
RSEM ML	8	3.4	3.0	5.3	7.7

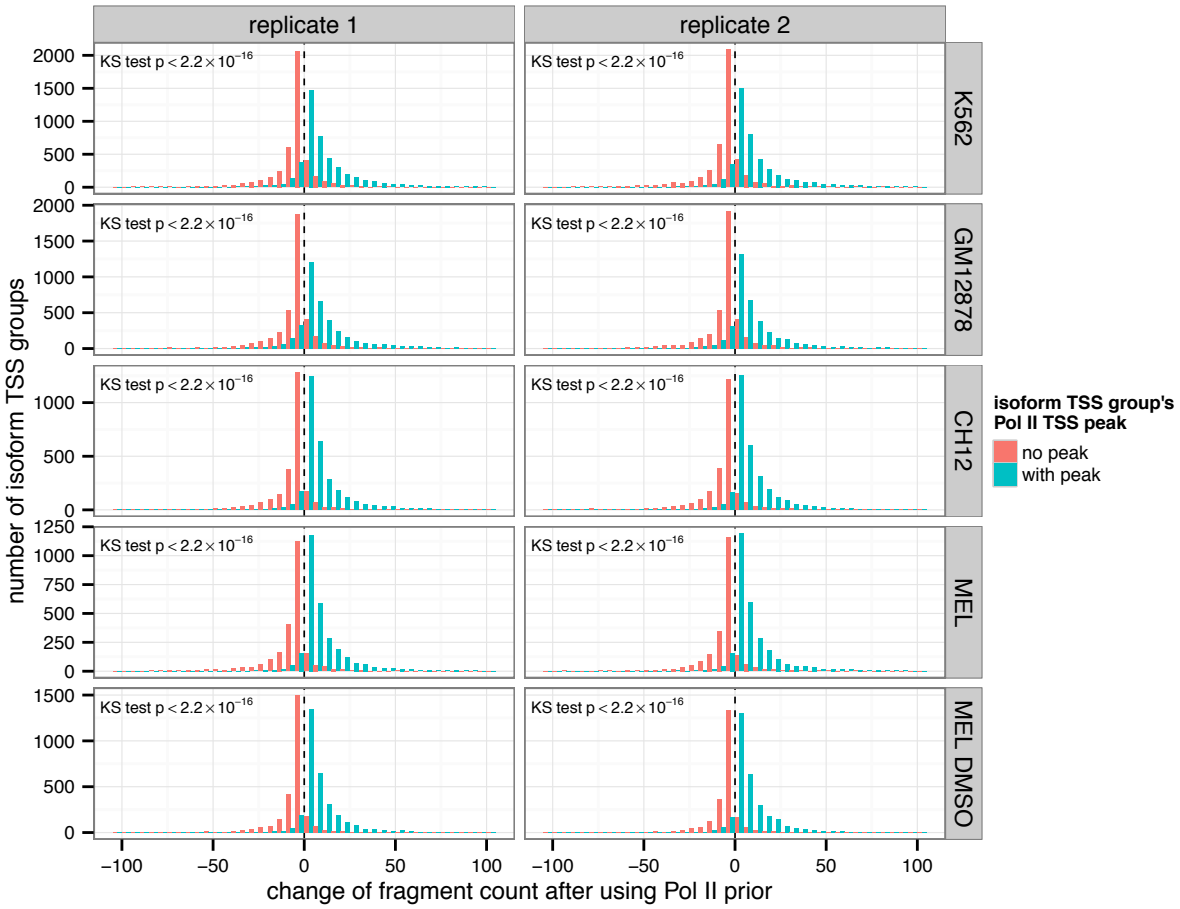
[1] RNA-seq data for mouse MEL cell line is 101 nt paired-end with 38.5 million reads aligned to transcripts.
* RNA-seq data for human K562 cell line is 76 nt paired-end with 67.4 million reads aligned to transcripts.

II.H. Software availability

The source code of pRSEM is available in the Supplemental Material. The latest version of pRSEM and a demo can be found at https://github.com/pliu55/RSEM/tree/pRSEM and https://github.com/pliu55/pRSEM_demo, respectively.

III. Quantification of human and mouse RNA-seq data by pRSEM

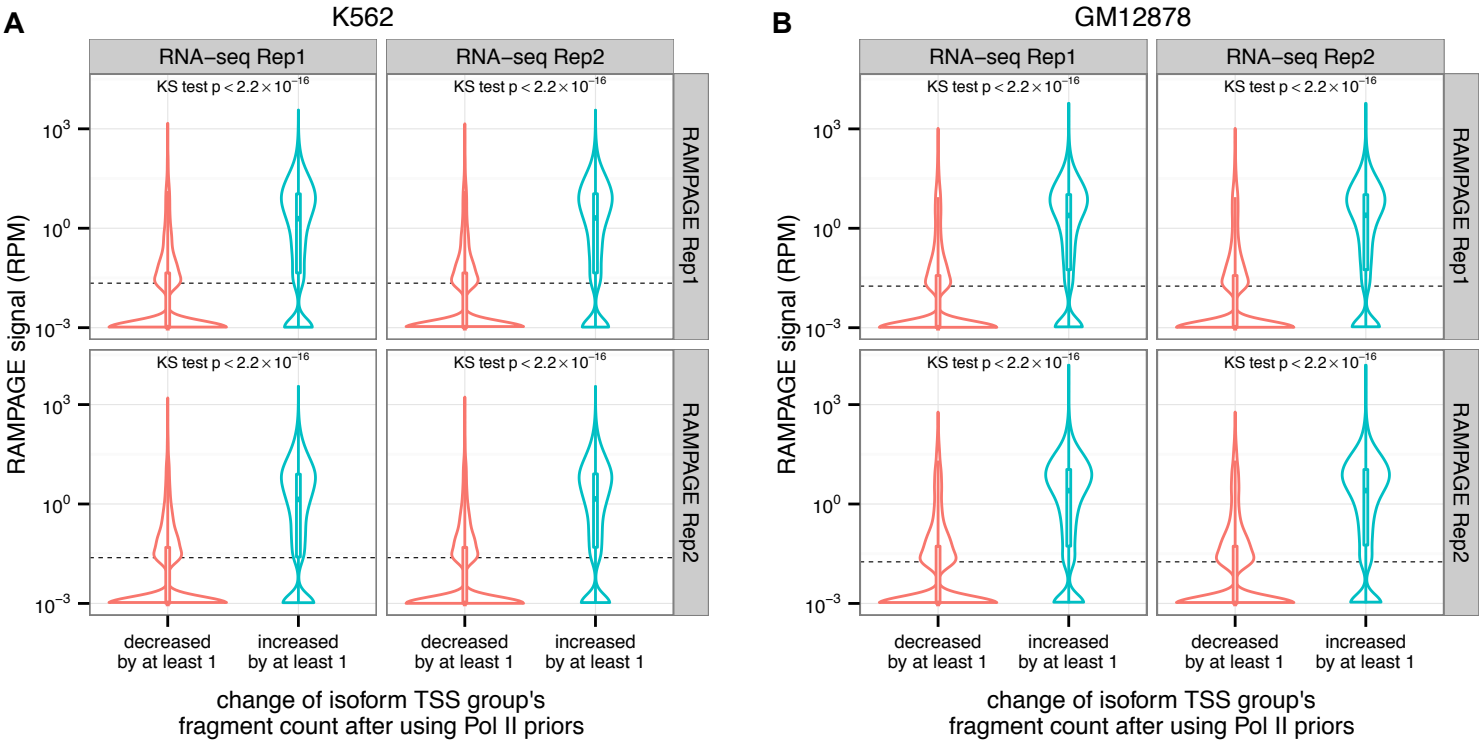
III.A. Allocating multi-mapping reads between isoform TSS groups



Supplemental Figure S7. pRSEM more accurately allocates multi-mapping reads between isoform TSS groups in both human and mouse data sets. Distributions of the change of fragment count between estimates from pRSEM and RSEM. Data shown are from the two replicates of each of five human and mouse cell line RNA-seq samples. The p-value from a Kolmogorov-Smirnov test is shown in the top left of each panel. Color code and data generation are the same as Figure 3A.

III.B. Validation of pRSEM estimates by RAMPAGE

An isoform's RAMPAGE signal was defined by first counting the number of reads that had their 5' ends map within the 100 nucleotide flanking region of the isoform's TSS, and then dividing that number by the total number (in millions) of RAMPAGE reads in that data set. An isoform group's RAMPAGE signal was defined similarly, except that the interval in which reads were counted was [TSS_{min} - 100, TSS_{max} + 100], where TSS_{min} and TSS_{max} were the lowest and highest coordinates of TSSs, respectively, for isoforms within that group.



Supplemental Figure S8. Allocation of multi-mapping reads by pRSEM is supported by RAMPAGE signals. Distribution of K562 **(A)** and GM12878's **(B)** RAMPAGE signals for isoform TSS groups that have fragment counts decreased by at least one and increased by at least one after using pRSEM instead of RSEM. Color code, line styles, and data generation are the same as in Figure 3C. Calculations were based on each cell line's two RNA-seq replicates and two RAMPAGE replicates.

III.C. Validation of pRSEM estimates by qRT-PCR

Isoform selection. Two sets of isoforms were selected for the validation of fold changes between two isoforms of the same gene in the MEL mouse cell line (Supplemental Data S1). First, we screened for genes that met the following criteria: (i) had a Pol II TSS peak status of 'mixed'; (ii) did not overlap or share RNA-seq reads with any other gene; (iii) had no more than five isoforms; (iv) had increases or decreases in isoform fragment counts (as compared with RSEM) that corresponded to the presence or absence of a Pol II TSS peak, respectively. For each selected gene, we looked for a pair of isoforms that fit the following criteria: (i) both isoforms had a unique exon region of at least 15 nucleotide for designing primers; (ii) one isoform had a Pol II TSS peak, had a read count increase of at least one with pRSEM, and had abundance ≥ 1 TPM as estimated by pRSEM; (iii) the other isoform did not have a Pol II TSS peak, had its read count decrease by at least one with pRSEM, had $log_2\left(\frac{TPM_{pRSEM}}{TPM_{RSEM}}\right) \leq -0.95$, where TPM_{pRSEM} and TPM_{RSEM} represent the abundances (in TPM) estimated by pRSEM and RSEM, respectively, and had $TPM_{RSEM} \geq 1$. Through these selection criteria, when comparing pRSEM against RSEM, the candidate genes always had reads transferred between their own isoforms and the reads were re-allocated from the 'no peak' isoform to the 'with peak' isoform.

Also, the differences between the pRSEM and RSEM estimates were large enough such that one would be definitively closer to the qRT-PCR measurements. For 'Set I.A' (Supplemental Data S1), we required that both isoforms had a GENCODE 'transcript_type' defined as 'protein_coding' or 'processed_transcript', that the selection criteria for pairs of isoforms were met in both of the MEL RNA-seq replicates, and that the 'no peak' isoform's TPM_{pRSEM} was at least one in both replicates. Under these strict criteria, expression of both isoforms could be detected by qRT-PCR with high confidence. For 'Set I.B' (Supplemental Data S1), we focused on isoforms that had a GENCODE 'transcript_type' defined as 'protein_coding'. Unlike Set I.A, the selection criteria for pairs of isoforms were only required to be met in at least one RNA-seq and we required that the 'no peak' isoform had $0.1 \leq TPM_{pRSEM} < 1$ in at least one MEL RNA-seq replicate. These relaxed criteria allowed us to obtain more candidates.

For validating estimated fold changes between two conditions, we selected isoforms from the mouse CH12, MEL, and MEL DMSO (MEL cell treated by 2% DMSO for five days) cell lines (Supplemental Data S3) with the following criteria: (i) the isoform had at least a 15 nucleotide unique exonic region for designing primers; (ii) its GENCODE 'transcript_type' was either 'protein_coding' or 'processed_transcript'; (iii) the ratio between the fold changes estimated by pRSEM and RSEM across the two conditions was either at least 2 or at most 0.5, where fold change was computed based on the average TPM from the two RNA-seq replicates in each condition; (iv) for one condition, the isoform had an abundance of at least one TPM and a non-zero fragment count as estimated by RSEM and pRSEM in both RNA-seq replicates. This criterion ensured that the expression of the isoform under this condition could be detected by qRT-PCR with high confidence; (v) for the other condition, the isoform had $TPM_{RSEM} \geq 1$ and $log_2\left(\frac{TPM_{RSEM}}{TPM_{pRSEM}}\right) \geq 1$ or had $TPM_{pRSEM} \geq 1$ and $log_2\left(\frac{TPM_{pRSEM}}{TPM_{RSEM}}\right) \geq 1$ in both RNA-seq replicates such that the fold changes of the isoform estimated by pRSEM and RSEM were different enough to be discriminated between with qRT-PCR measurements; (vi) the isoform had a Pol II TSS peak in only one condition.

Measuring isoform expression by qRT-PCR. Mouse erythroleukemia (MEL) and CH12 cells were maintained in 10% Fetal Bovine Serum containing RPMI 1640 with L-Glutamine, and 1% penicillin/streptomycin. CH12 cells were additionally supplemented with $1x10^{-5}$ M β-mercaptoethanol. Both cell lines were cultured under standard mammalian cell culture conditions, with 5% CO2 in a 37^0C incubator. MEL cells were treated with 2% DMSO for 2 days for MEL DMSO condition.

Total RNA was purified from MEL, DMSO treated MEL, and CH12 cells using TRIzol (Invitrogen). 2μg RNA was used to synthesize cDNA by Moloney murine leukemia virus reverse transcription (M-MLV RT) using a random hexamer-oligo dT primer cocktail. All cDNA synthesis reactions were preceded by a DNaseI treatment to remove any DNA contamination and a minus reverse transcriptase control was used to confirm the specificity. Real-time PCR was performed with SYBR green master mix (ABI). To compare the expression of different isoforms in MEL cells, the ddCt method was used. Primer pairs with similar amplification efficiencies were used for the analysis and all the values were normalized to 18S RNA expression. The *Prkci-001* sample was used as reference standard to generate the plots. Three independent experiments were carried out (Supplemental Data S2).

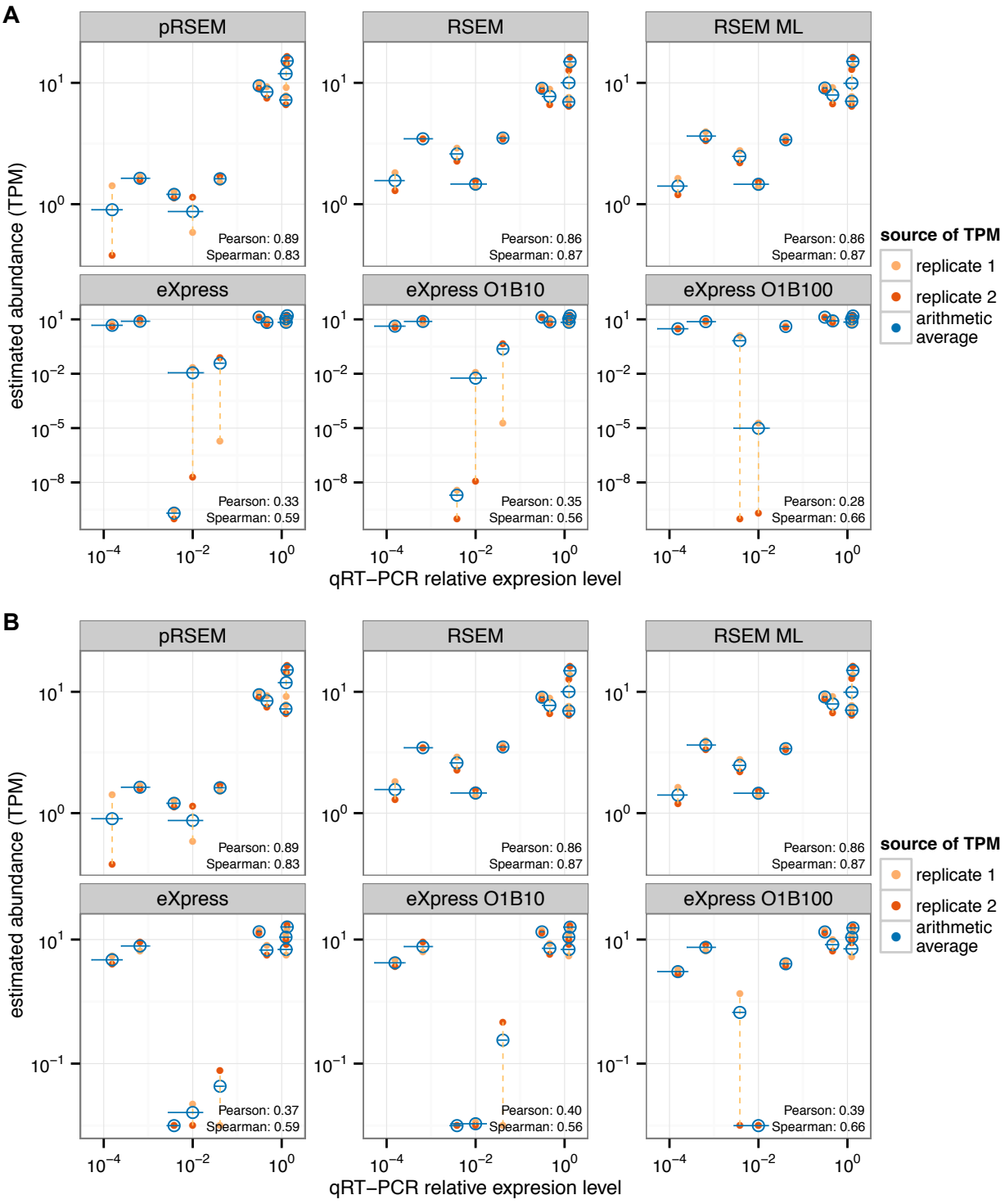
Relative expression of isoforms among different samples (MEL, MEL DMSO and CH12) was determined by a relative standard curve method. Serial dilutions (1:5) of cDNA sample from highest

expressed samples were used to generate the standard curve and relative expression of each isoform was determined from the curve using the StepOne plus analysis platform (ABI). All values were normalized to corresponding 18S values and fold changes were determined. Four independent experiments were performed (Supplementary Data S4). Isoforms for which primer pairs did not yield satisfactory amplification curves, had diverse primer efficiencies, or amplified intronic regions were excluded from the analysis. All primers are listed in Supplemental Table S4.

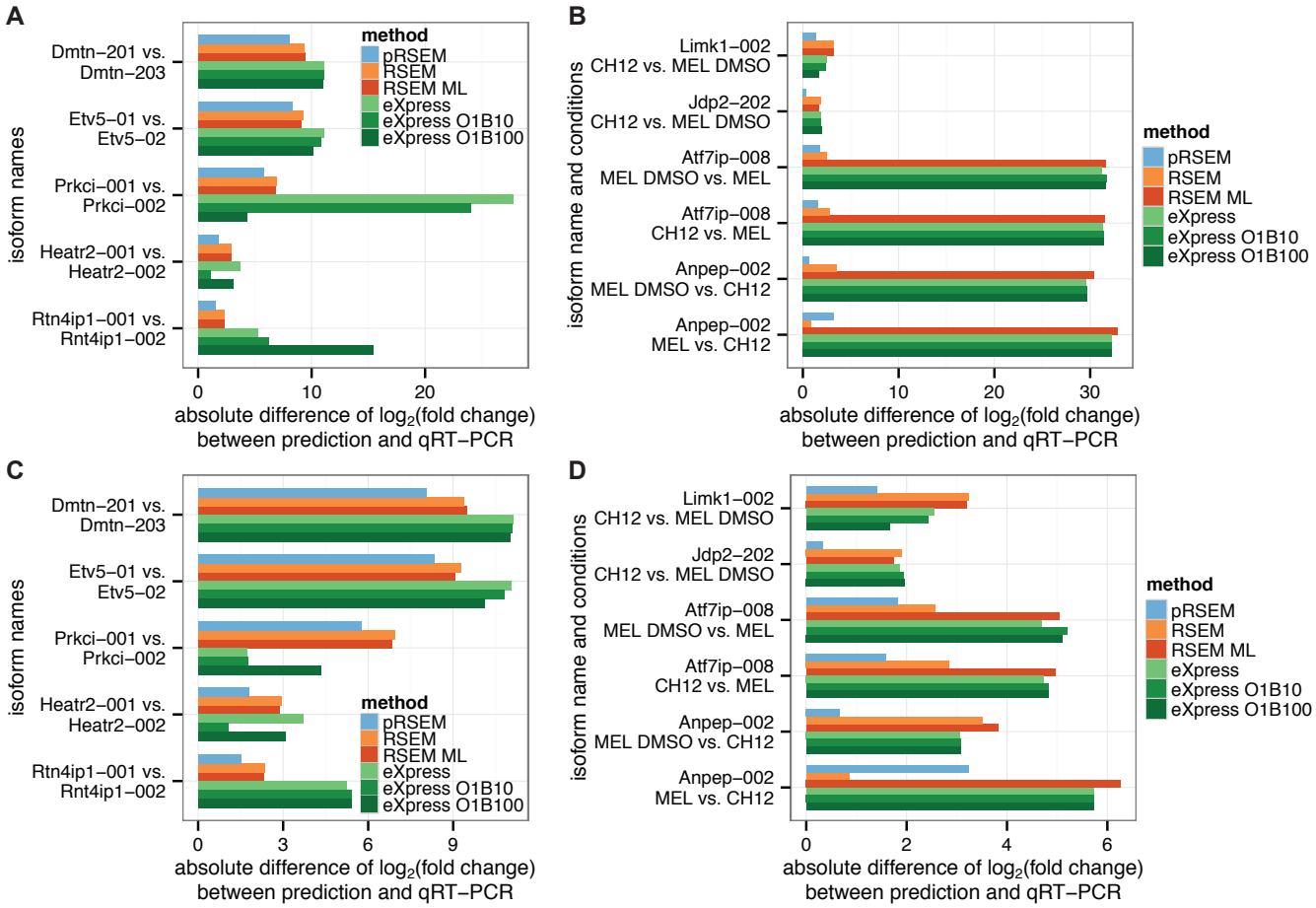
Supplemental Table S4. A list of primer pairs for qRT-PCR. Primer pairs that did not yield satisfactory amplification curve, had diverse primer efficiencies, or amplified intron regions are colored in grey.

Set	Isoform's Ensembl 78 ID	Primer (5' to 3')	
		Forward strand	Reverse strand
I.A	Prkci-001	ACCAGGTCCGGGTGAAA	ATATCTCGAACCTCACTGCAA
	Prkci-002	AGCAAAGGCTGTTGTTTTCC	CTGCAAAGTCCCTCAAAGGA
	Heatr2-001	CGGGTAGCCGTTATCGAA	CGACCACCGAGGTCACT
	Heatr2-002	TAACCTTTGTGGTTGGTTCCA	TCCTCCACCTCAGCCAGTGT
	Dmtn-201	CAAGACCCGAGAGCTTCCAA	AAGCCCCAGGAAGCAAAAGG
	Dmtn-203	GGAGCTGGCGAAGGA	GTTCAGGAGGGAGATCAGA
I.B	Rtn4ip1-001	TGATGTTACCTATCATACACTATCCAAATG	TGGCTCCATAACCACTTCTCATATT
	Rtn4ip1-002	ACCTGCAGAAGTGAATTGTTTGTC	ACATGAGCCCCCCATGCT
	Etv5-001	GAGTGGCCGCTCAGGAGTATC	TGCTTCCAAAGTCTCCGCTATC
	Etv5-002	GTTCCTGATGATGAGCAGTTTGTC	CACTGCAGTCCCGGCTCTAG
	Dcun1d3-001	GAACCTCTGGAGCAGCTGTTG	GAGTGGACCCCTCTGGATCA
	Dcun1d3-005	GGAATTAGAGTTGCCAGTCTGTGA	GCCACTGGCCCATGTACTTC
II	Atf7ip-008	GGGCTCCTTTGGGATTCAG	CGAGCCTTGAAGACTTTTTTCTG
	Anpep-002	GGGAGGAGGGCTTAGCTGTAA	CGGTAATCTACCTGGCACATGA
	Limk1-002	GATGGGGAAGCTTAGGCCAG	TACACTCGCAGCACTTAGCC
	Jdp2-202	CCGTCAGGCACATCAGGTT	TGCCCAGGCATCATAGCA
	Ehd1-002	GCACTAGCTCAGTAGCCTGAACTG	GGCCTGGAACTTGCTGAAGTC
	Pafah1b3-007	ACATCCTCCTCCTCACCTAGCA	CTACTTCGGGTTCCTTGTCTTTG
	Ubald2-002	CTGCGTCCCGTATTTTGTCC	ATCAGGGAAGTTGGGTGGTG

Comparison of qRT-PCR measurements with predictions. We compared qRT-PCR measurements with estimates from five quantification methods: pRSEM, RSEM, RSEM ML, eXpress, eXpress O1B10, and eXpress O1B100 (for explanations of each method, see section I.A). For the experiment measuring isoform abundances under the same condition, pRSEM and two RSEM variants had strong correlations with qRT-PCR, whereas the three eXpress variants had weak correlations (Supplemental Figure S9). When comparing the difference of fold changes to qRT-PCR measurements, pRSEM estimates had the smallest differences for three out of five pairs of isoforms and was always better than RSEM, RSEM ML, and eXpress with its default settings for all five pairs (Supplemental Figure S10, A and C; Supplemental Data S2). In qRT-PCR validation of one isoform under two conditions, pRSEM outperformed all other methods for five out of six cases (Supplemental Figure S10, B and D; Supplemental Data S4).



Supplemental Figure S9. pRSEM and RSEM estimates have stronger correlations with qRT-PCR measurements than those from eXpress variants. Comparison of expression levels between two isoforms from the same gene under the same condition. Minimum abundances from eXpress variants were set to 10^{-10} **(A)** or 10^{-2} **(B)**. At the lower right of each plot are Pearson and Spearman correlation coefficients calculated between log transformed arithmetic averages of estimated abundances (blue) and log transformed average relative expression levels. The arithmetic averages were derived from estimates on MEL RNA-seq's replicate 1 (light yellow) and replicate 2 (red). Error bars denote one standard deviation. RSEM ML: RSEM maximum likelihood; eXpress O1B10: eXpress run with one round of online EM followed by ten rounds of batch EM; eXpress O1B100: eXpress run with one round of online EM followed by 100 rounds of batch EM.



Supplemental Figure S10. pRSEM estimates are closer to qRT-PCR measurements than those from RSEM and eXpress variants. (A, C) Comparison of expression levels between two isoforms from the same gene under the same condition; **(B, D)** Comparison of an isoform's expression level in two different cell lines. Minimum average abundances from two RNA-seq replicates were set to 10^{-10} **(A, B)** or 10^{-2} **(C, D)**; Notations for quantification methods are the same as for **Supplemental Figure S9.**

III.D. Genome-wide biological implications of pRSEM abundance estimates

We carried out two genome-wide surveys to determine the biological implications of pRSEM's abundance estimates, given our results suggesting that they are more accurate than those of previous methods, such as RSEM. Our first survey examined genome-wide TSS activities and the second identified expressed isoforms for the sixteen cell types in the mouse hematopoietic differentiation study.

Genome-wide active TSSs. We compared transcriptome profiles from pRSEM and RSEM quantifications. Many TSSs were found to have different 'on' or 'off' calls between pRSEM and RSEM (Supplemental Table S5). In all five human and mouse cell lines, more than seven hundred TSSs were identified to be active by RSEM, but not by pRSEM. This finding is in line with pRSEM's strength of removing false positives, as shown in our qRT-PCR validations (section III.C) and data-driven simulations (section IV).

Supplemental Table S5. Number of active transcription start sites (TSSs) called by RSEM or pRSEM. A TSS is considered to be 'active' if it has abundance ≥ 1 TPM in all RNA-seq samples from the same cell line.

Cell line	Number of active TSS			
	RSEM	pRSEM	RSEM only	pRSEM only
K562	36,682	35,576	1,312	206
GM12878	36,193	35,006	1,425	238
CH12	24,242	23,226	1,111	95
MEL	22,675	21,588	1,176	89
MEL DMSO	26,621	25,970	705	54

Expressed genes and isoforms in hematopoietic cells. In our second survey, we applied pRSEM to sixteen types of primary cells from a mouse hematopoiesis differentiation study. We examined the extent to which the numbers of genes and isoforms called as expressed by pRSEM were different from those obtained from RSEM. Compared to the three ENCODE mouse cell lines used in the first survey, the primary cells used in this survey more closely resemble physiological states and data from them are more relevant to living systems. For every cell type, the numbers of expressed genes and isoforms called by pRSEM were always much smaller than those called by RSEM (Supplemental Table S6). This is similar to our observation in the first survey and is most likely the result of pRSEM's strength at removing false positives. The sets of expressed genes and isoforms determined by pRSEM are thus likely to contain less noise, which benefits downstream analyses that attempt to draw biological insights from such sets.

Supplemental Table S6. Number of expressed isoforms and genes called by RSEM and pRSEM for sixteen cell types from mouse hematopoietic differentiation. An isoform or a gene is defined as 'expressed' in a cell type if it has abundance ≥ 1 TPM in all of the RNA-seq samples for that cell type. Due to limited sequencing depth for these primary cells, an expressed isoform or gene is also required to have a non-zero RNA-seq read count.

Cell type	Number of isoforms					Number of genes				
	Total	Is expressed				Total	Is expressed			
		RSEM	pRSEM	RSEM only	pRSEM only		RSEM	pRSEM	RSEM only	pRSEM only
LT-HSC	78,754	23,408	20,127	3,396	115	22,019	11,290	10,409	883	2
ST-HSC		18,659	16,265	2,446	52		9,941	9,433	508	0
MPP		19,792	18,618	1,254	80		10,499	9,766	733	0
CMP		20,168	18,307	1,926	65		10,363	9,756	607	0
GMP		19,209	17,346	1,927	64		10,076	9,495	583	2
Mφ		14,625	12,284	2,406	65		8,299	7,958	343	2
Gn		11,391	9,408	2,015	32		6,142	5,982	160	0
Mo		19,744	16,721	3,099	76		9,336	9,041	295	0
CLP		15,786	14,425	1,407	46		9,089	8,446	644	1
B		21,883	18,345	3,610	72		9,596	9,175	422	1
CD4		18,349	15,219	3,187	57		8,714	8,406	309	1
CD8		17,896	14,896	3,071	71		8,930	8,483	447	0
NK		22,893	19,222	3,729	58		9,721	9,500	221	0
MEP		18,402	16,963	1,502	63		9,743	9,126	619	2

EryA		9,900	9,417	533	50		7,116	5,927	1,190	1
EryB		4,171	4,126	69	24		3,377	3,054	323	0

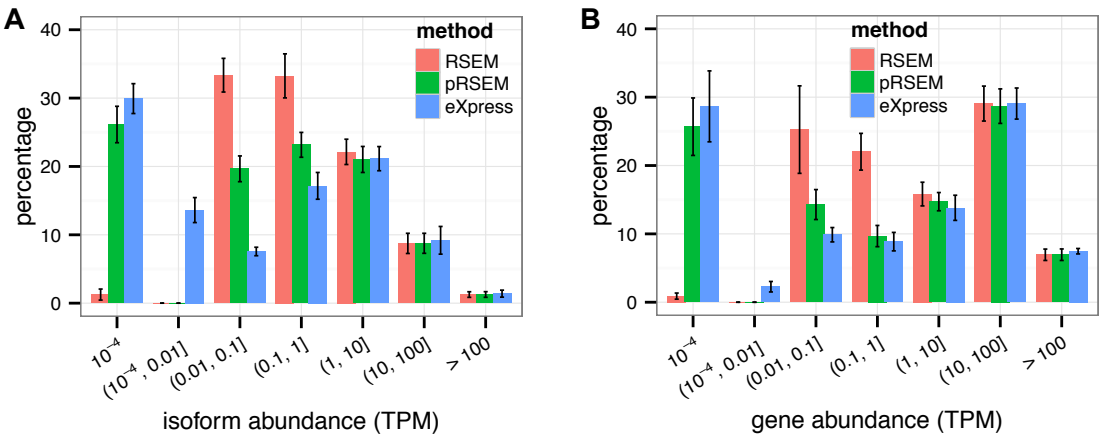
III.E. Differential expression

Given pRSEM's strength in reducing the number of false positive isoforms, we examined whether its estimates would lead to different differential expression (DE) results than those from RSEM estimates. We employed EBSeq (Leng et al. 2013) to make DE calls based on fragment counts from RSEM or pRSEM on the three mouse ENCODE cell lines: CH12, MEL, and MEL DMSO. Compared to alternative methods, EBSeq's advantage is that it will not only identify DE genes, but also DE isoforms. This feature allows us to make comparisons at the gene level as well as at the isoform level. We counted the number of DE genes and isoforms that were only called based on RSEM or pRSEM estimates. pRSEM and RSEM did not differ much in terms of identifying DE genes — the two methods disagreed on the DE call of 26 to 46 genes per comparison (Supplemental Table S7). In contrast, the two methods disagreed on the DE call for more than two thousand isoforms per comparison, with pRSEM estimates resulting in a larger number of DE isoform calls. (Supplemental Table S7). Such a large difference in the DE isoform calls would most likely lead to different functional characterizations between each pair of cell lines. Unfortunately, current Gene Ontology analysis is only available at the gene level and comprehensive functional annotations of isoforms are still lacking. Such limitations prevent us from performing further functional analysis on these large sets of DE isoforms.

Supplemental Table S7. Number of differentially expressed (DE) genes and isoforms only called based on RSEM or pRSEM estimates. DE genes or isoforms were controlled at a false discovery rate of 0.05.

Level	Cell line		Number of DE items	
	1	2	RSEM only	pRSEM only
gene	CH12	MEL	15	11
	CH12	MEL DMSO	16	11
	MEL	MEL DMSO	13	33
isoform	CH12	MEL	748	1,413
	CH12	MEL DMSO	986	1,476
	MEL	MEL DMSO	953	1,294

III.F. Isoform abundance profiles



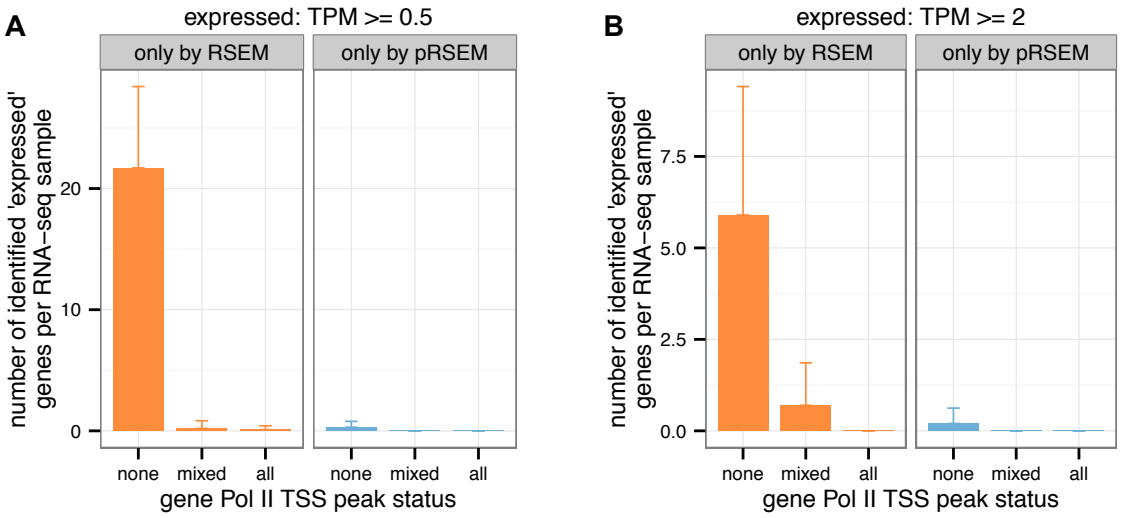
Supplemental Figure S11. Comparison of the distribution of transcript abundances estimated by RSEM, pRSEM, and eXpress. Percentages of isoforms **(A)** and genes **(B)** were calculated based on the two RNA-seq replicates from each of the K562, GM12878, CH12, MEL, and MEL DMSO cell lines. Error bars represent one standard deviation.

III.G. pRSEM identifies unexpressed genes misclassified by other methods

Supplemental Table S8. RSEM and pRSEM largely agree on the expression states for genes that do not overlap with any other gene and share RNA-seq reads. The comparisons were made at three different cutoffs: 0.5, 1.0, and 2.0 TPM, for defining 'expressed' genes.

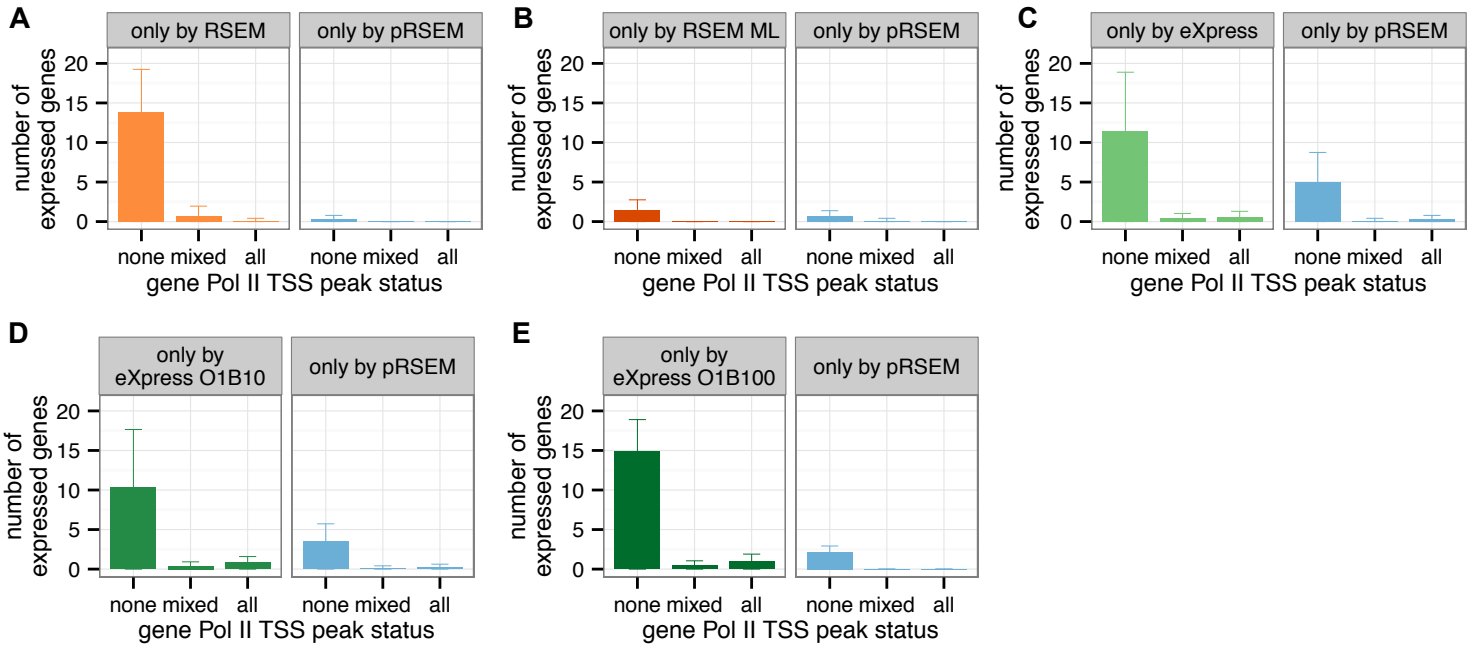
Cell line	Replicate index	RSEM and pRSEM agree		RSEM and pRSEM disagree	
		Number of genes	Percentage of genes	Number of genes	Percentage of genes
"expressed": TPM ≥ 0.5					
K562	1	554	96.3	21	3.7
	2	585	96.7	20	3.3
GM12878	1	560	96.9	18	3.1
	2	613	97.9	13	2.1
CH12	1	895	97.5	23	2.5
	2	921	98.0	19	2.0
MEL	1	835	95.6	38	4.4
	2	856	97.2	25	2.8
MEL DMSO	1	1064	97.5	27	2.5
	2	994	98.1	19	1.9
average		787.7	97.2	22.3	2.8
"expressed": TPM ≥ 1					
K562	1	567	98.6	8	1.4
	2	591	97.7	14	2.3
GM12878	1	565	97.8	13	2.2
	2	617	98.6	9	1.4
CH12	1	898	97.8	20	2.2
	2	918	97.7	22	2.3

MEL	1	852	97.6	21	2.4
	2	863	98.0	18	2.0
MEL DMSO	1	1078	98.8	13	1.2
	2	1001	98.8	12	1.2
average		795	98.1	15	1.9
“expressed”: TPM ≥ 2					
K562	1	566	98.4	9	1.6
	2	600	99.2	5	0.8
GM12878	1	575	99.5	3	0.5
	2	624	99.7	2	0.3
CH12	1	912	99.3	6	0.7
	2	932	99.1	8	0.9
MEL	1	858	98.3	15	1.7
	2	872	99.0	9	1.0
MEL DMSO	1	1085	99.5	6	0.5
	2	1008	99.5	5	0.5
average		803.2	99.1	6.8	0.9



Supplemental Figure S12. pRSEM eliminates far more false positive genes than false positive or false negative genes that it introduces under different ‘expressed’ cutoffs. Number of identified ‘expressed’ genes, in which RSEM and pRSEM disagreed, with an ‘expressed’ cutoff of TPM ≥ 0.5 **(A)** and TPM ≥ 2 **(B)**. Color code and data generation are the same as in Figure 4B.

At the gene level, we made pairwise comparisons of pRSEM with two variants of RSEM and three variants of eXpress. We counted the number of method-specific expressed genes stratified by their Pol II TSS peak status. For genes without peaks, the number of pRSEM-specific expressed genes was always lower than that of each other method (Supplemental Figure S13). Again, if we assume that a gene without a Pol II TSS peak should not be expressed, we can conclude that pRSEM identifies misclassified unexpressed genes by other methods.



Supplemental Figure S13. pRSEM identifies unexpressed genes misclassified by RSEM or eXpress. Expressed genes called by only one method from pairwise comparison of pRSEM with RSEM **(A)**; RSEM ML **(B)**; eXpress **(C)**; eXpress O1B10 **(D)**; and eXpress O1B100 **(E)**. The average and standard deviation of the number of genes were calculated from two RNA-seq replicates of cell lines: K562, GM12878, CH12, MEL, and MEL DMSO. Notations for quantification methods are the same as for **Supplemental Figure S9**.

IV. Evaluating pRSEM by data-driven simulations

IV.A. Sub-sampling experiments

We sub-sampled RNA-seq reads from ENCODE K562's RNA-seq replicate one (Supplemental Table S1), which had 113.6 million paired-end reads. We took random samples of 10%, 30%, and 50% of these reads as the input RNA-seq data sets for sub-sampling experiments. All of the remaining alignment and quantification processes were the same as those for the ENCODE RNA-seq data. There were 6.8 million, 20.3 million, and 33.8 million fragments aligned to transcripts for each of the sub-sampling data sets (Supplemental Table S9). We used RSEM ML estimates on K562 replicate one as the ground truth with which to determine false positives and false negatives. We note that this ground truth definition gave RSEM ML an advantage over the other methods in terms of quantification on the sub-sampled RNA-seq data.

IV.B. Simulation at full-sequencing depth

In the simulation at full sequencing depth, the total number of fragments was also based on ENCODE K562's RNA-seq replicate one (Supplemental Table S1). This data set had 113.6 million paired-end reads, 67.4 million of which aligned to transcripts with a noise parameter of 0.156 estimated by RSEM (Supplemental Table S9). We partitioned all isoforms by their TSS peak status according to K562 Pol II ChIP-seq data (Supplemental Table S2). For each partition, we drew each isoform's fragment-generating probability (θ in RSEM's probabilistic model) from the distribution learned from the training set isoforms. With the fragment-generating probabilities for all isoforms, we calculated their abundances (TPMs) and took it as the ground truth. Next, we employed RSEM's simulator (Li and Dewey 2011) to simulate paired-end reads based on the ground truth, total number of aligned reads, and the noise parameter.

Supplemental Table S9. Read depth for all RNA-seq data sets used in this work. Shown are the total numbers of sequenced fragments in RNA-seq data set and the numbers of fragments that aligned to isoforms. There is a substantial number of unaligned fragments, which can be attributed to fact that all ENCODE RNA-seq data were from the whole-cell fraction.

RNA-seq data set	Number of fragments (millions)	
	Total	Aligned
sub-sampling at 10.0% read depth of K562 Rep1	11.4	6.8
sub-sampling at 30.0% read depth of K562 Rep1	34.1	20.3
sub-sampling at 50.0% read depth of K562 Rep1	56.8	33.8
simulation at full sequencing depth as K562 Rep1	79.5	64.2
ENCODE human K562 Rep1	113.6	67.4
ENCODE human K562 Rep2	119.1	64.3
ENCODE human GM12878 Rep1	117.9	62.1
ENCODE human GM12878 Rep2	131.8	78.1
ENCODE mouse CH12 Rep1	140.2	47.3
ENCODE mouse CH12 Rep2	180.6	51.7
ENCODE mouse MEL Rep1	124.5	38.5
ENCODE mouse MEL Rep2	178.3	59.5

ENCODE mouse MEL DMSO Rep1	177.5	58.4
ENCODE mouse MEL DMSO Rep2	205.2	80.4

IV.C. Comparison of pRSEM with alternative quantification methods

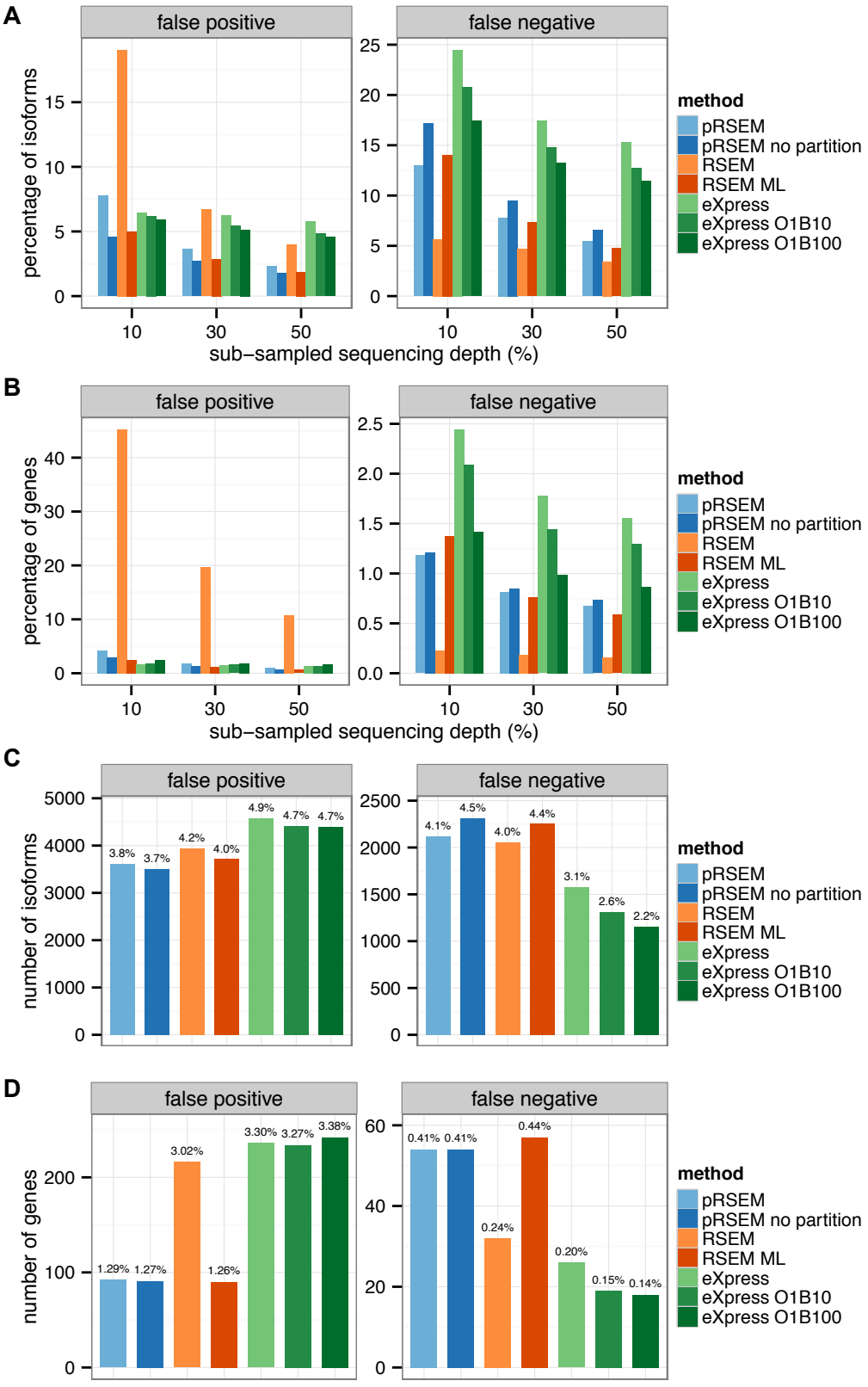
In the sub-sampling experiments, RSEM ML had a smaller false positive rate and a smaller false negative rate than pRSEM at both the isoform- and gene-level for most of the sub-sampling depths (Supplemental Figure S14, A and B). This is likely because of the fact that RSEM ML at 100% sampling depth was used as the ground truth. In simulations at full-sequencing depth, RSEM ML was not used as the truth and its advantage disappeared. pRSEM had smaller or comparable false positive rates and smaller false negative rates at both the isoform- and gene-levels (Supplemental Figure S14, C and D). Compared to all three eXpress variants, pRSEM generally had favorable false positive rates and false negative rates in the sub-sampling experiments (Supplemental Figure S14, A and B). In simulations at full sequencing depth, pRSEM had a much smaller number of false positives at the expense of false negatives at the isoform- and gene- levels (Supplemental Figure S14, C and D).

From our simulations, we found that running additional batch EM rounds with eXpress consistently increased eXpress's sensitivity and specificity in most of the experiments (Supplemental Figure S14). However, with respect to these metrics, the relative ranking of pRSEM compared to eXpress remained the same for all variants. Furthermore, additional batch EM rounds did not improve eXpress's accuracy in our qRT-PCR validations (Supplemental Figure S9; Supplemental Figure S10), and were time-consuming (Supplemental Table S3) due to eXpress's limited parallelization abilities in its current implementation.

IV.D. Comparison of pRSEM and RSEM on isoforms with uninformative priors

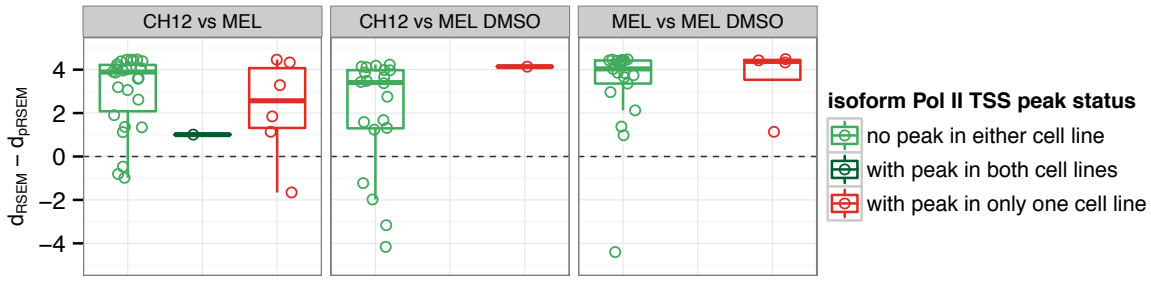
In our qRT-PCR validations, if one removes the criteria regarding the Pol II TSS peak status of the isoforms and the directionality of the difference between pRSEM and RSEM, 41 isoforms become candidates for validation. Of these, eleven have a Pol II TSS peak in one condition but not the other. The remaining isoform candidates do not have Pol II TSS peak in any of the two conditions. We decided against validating these isoforms because the Pol II information could not be explicitly connected to the difference between the pRSEM and RSEM estimates, leaving the differences difficult to explain as there are many factors at play including ChIP-seq multi-mapping read allocation, noise in peak calling, and IDR thresholds. Full validation for these isoforms would have required ChIP-qPCR experiments, which is beyond the scope of this work.

To circumvent the difficulties of validating isoforms for which the Pol II peak status was the same across conditions, we decided to perform data-driven simulations. We treated Pol II ChIP-seq peak data as the ground truth, and generated isoform expression levels as well as simulated RNA-seq reads in the same manner as the full-sequencing-depth simulations described above. We selected isoforms by the same criteria, but considered fold changes in both directions and did not place



Supplemental Figure S14. pRSEM has a lower false positive rate than alternative methods in data-driven simulations. Comparison of sub-sampling simulations at the isoform level **(A)** and gene level **(B)**; Comparison of simulations at full sequencing depth at the isoform level **(C)** and gene level **(D)**. 'pRSEM no partition' denotes a pRSEM run, where a uniform prior parameter was learned from a training set without any partition. Notations for the other quantification methods are the same as for **Supplemental Figure S9**.

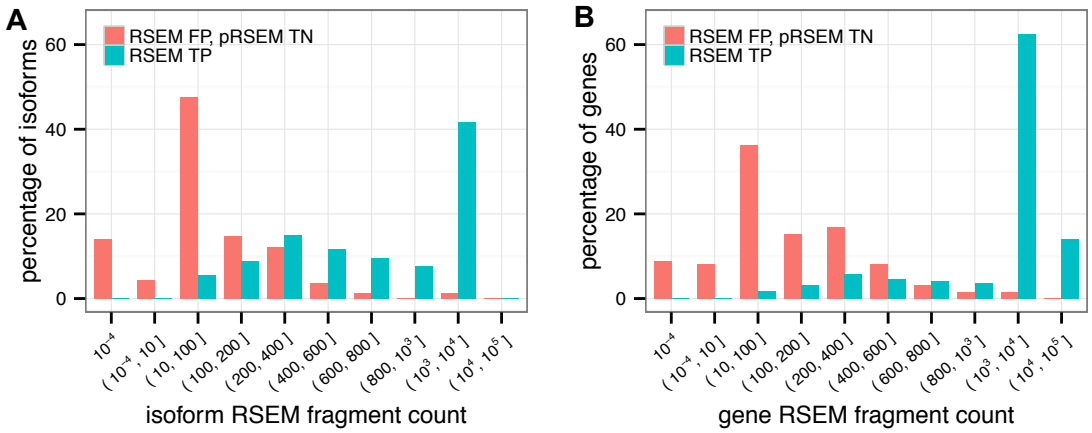
conditions on Pol II TSS peak status. We compared fold change differences between the simulated truth and RSEM or pRSEM (Supplemental Figure S15). For isoforms that have peaks in only one cell line, all pRSEM estimates, except one, had smaller difference to the truth than RSEM, which is in line with our qRT-PCR experiments. For isoforms with the same Pol II peak status across cell lines, pRSEM had better estimates than RSEM in the vast majority of cases. This simulation experiment suggests that, even for isoforms for which Pol II information is not explicitly informative, pRSEM still outperforms RSEM.



Supplemental Figure S15. pRSEM is more accurate than RSEM with respect to isoform expression fold change estimation between conditions for all isoform Pol II TSS peak status scenarios. d_{RSEM} and d_{pRSEM} represent the absolute difference of the log2 fold change between the truth and RSEM or pRSEM. The truth and simulated reads for each cell line were generated in the same manner as in the simulation at full sequencing depth in **Supplemental Figure S14**. Shown are isoforms that would be selected for qRT-PCR validation by our selection criteria when considering Pol II peak status (red), and those that would additionally fit our selection criteria if not considering Pol II TSS peak information (light green and dark green).

IV.E. Comparison of pRSEM with a naïve approach on eliminating false positives

We examined if a naive approach based on fragment counts could deliver results comparable to those of pRSEM. We compared the RSEM fragment count distributions for two categories of isoforms: (i) those called as false positives by RSEM and not called by pRSEM; and (ii) those called as true positives by RSEM. We found that the two distributions substantially overlap with each other. There are 32% isoforms of category (i) and 53% isoforms of category (ii) that have estimated fragment counts in the interval (100, 1000] (Supplemental Figure S16A). Under a naive approach, no matter where the cutoff is drawn, a substantial fraction of isoforms from either category would be misclassified. At the gene level, we observed a similar degree of overlap in the fragment count distributions. 45% and 21% genes from the two categories, respectively, have estimated fragment counts in the interval (100,1000] (Supplemental Figure S16B). Therefore, a naive approach that thresholds on estimated RSEM fragment counts cannot classify genes and isoforms as accurately as pRSEM.



Supplemental Figure S16. RSEM fragment count distributions of false positives called only by RSEM (cyan) overlaps substantially with the one of RSEM true positives (red). **(A)** Comparison at the isoform level; **(B)** Comparison at the gene level. FP: false positive; TN: true negative; TP: true positive.

V. Supplemental references

Chung D, Park D, Myers K, Grass J, Kiley P, Landick R, Keleş S. 2013. dPeak: High Resolution Identification of Transcription Factor Binding Sites from PET and SET ChIP-Seq Data. *PLoS Comput Biol* **9**: e1003246.

Derrien T, Estellé J, Marco Sola S, Knowles DG, Raineri E, Guigó R, Ribeca P. 2012. Fast computation and applications of genome mappability. *PLoS One* **7**: e30377.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2012. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**: 1760–1774.

Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, et al. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* **22**: 1813–1831.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.

Lara-Astiaso D, Weiner A, Lorenzo-Vivas E, Zaretsky I, Jaitin DA, David E, Keren-Shaul H, Mildner A, Winter D, Jung S, et al. 2014. Chromatin state dynamics during blood formation. *Science* **55**: 1–10.

Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BMG, Haag JD, Gould MN, Stewart RM, Kendziorski C. 2013. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics* **29**: 1035–1043.

Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**: 323.

Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. 2010. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* **26**: 493–500.

Roberts A, Pachter L. 2013. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat Methods* **10**: 71–73.

Stamatoyannopoulos JA, Snyder M, Hardison R, Ren B, Gingeras T, Gilbert DM, Groudine M, Bender M, Kaul R, Canfield T, et al. 2012. An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol* **13**: 418.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74.