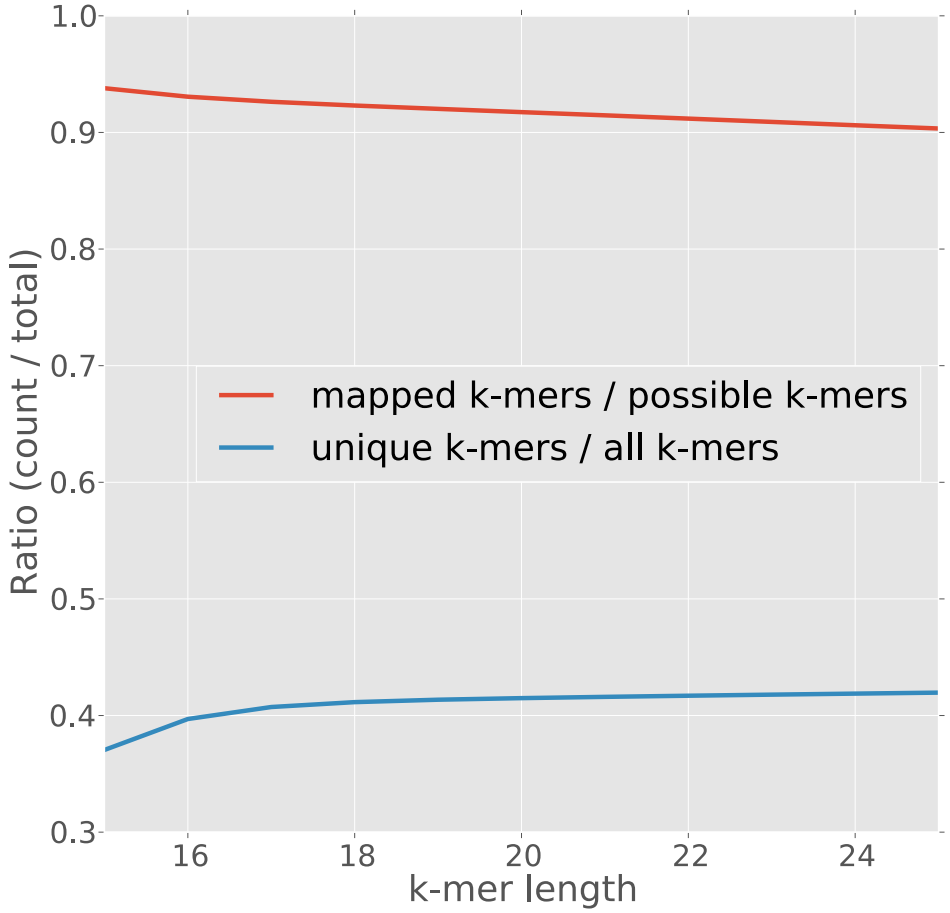
Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms

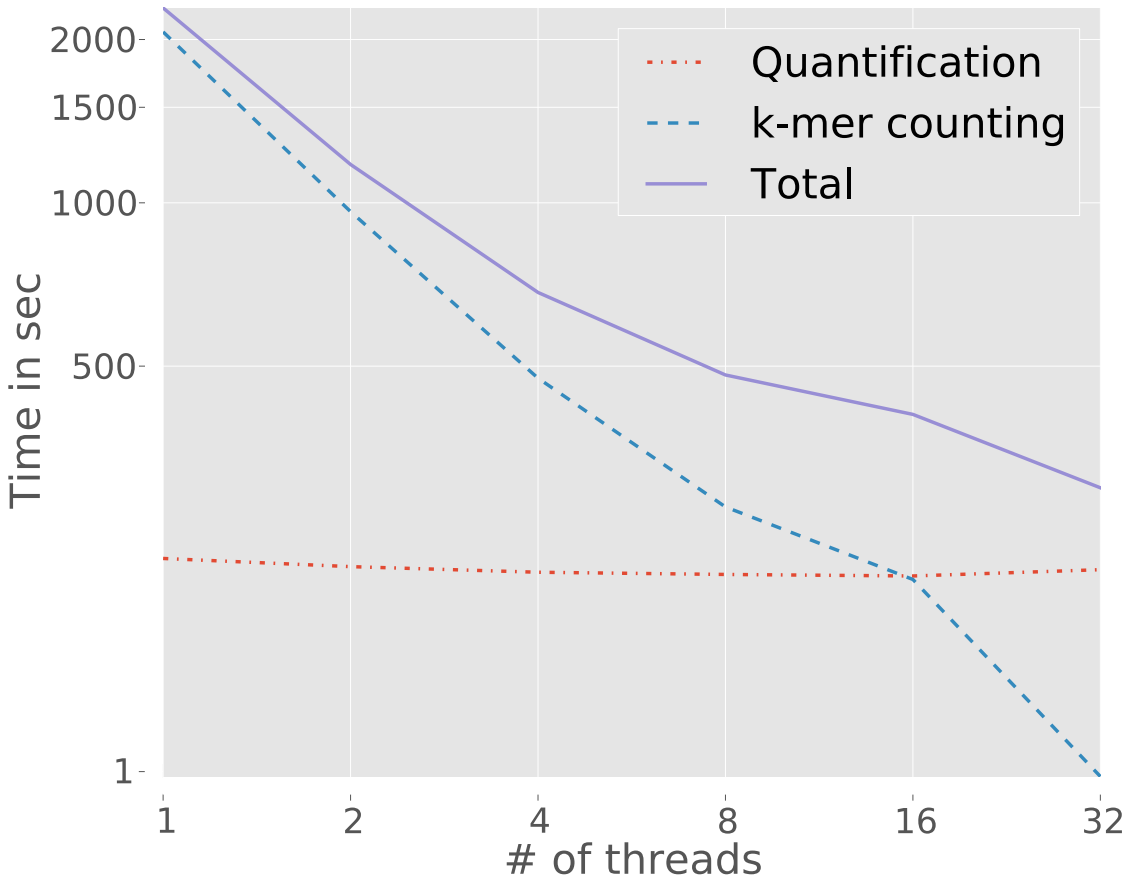
Rob Patro, Stephen M. Mount and Carl Kingsford

Supplementary Figure 1: Effect of k-mer length on retained data and k-mer ambiguity



Supplementary Figure 1: As the k-mer length is varied in the range $[15, 25]$ when processing the synthetic dataset, we observe that a longer k-mer length results in a slight decrease in data retention (denoted by the red line which shows the ratio of number of k-mers from the read set that were hashable to the total number of k-mers appearing in the set of reads). Simultaneously, we observe that the ratio of the number of unique k-mers (k-mers having a unique locus of origin) in the set of transcripts to the total number of k-mers in the set of transcripts (blue line) increases as we make k larger. It seems that, as expected, there is a trade-off in the choice of k, with a larger k resulting in less robustness to sequencing error but a higher fraction of unique k-mers and smaller k-mers providing more robustness to errors in the data but at the cost of increased ambiguity. However, since the differences are relatively small over a reasonably large range of k, we can expect the inference procedure to be fairly robust to this parameter. We use $k = 20$ for all experiments, and this is the default in Sailfish. However, we did not attempt to optimize this parameter when performing our experiments.

Supplementary Figure 2: Speed of counting indexed k-mers



Supplementary Figure 2: The time to count all of the k-mers and quantify transcript abundance in an 81M read dataset (`SRX016366`) as function of the number of concurrent hashing threads. Even with only a single thread, the counts for all k-mers in the dataset can be processed in 34 minutes and 26 seconds, while with 32 processing threads, all k-mers can be counted in only 1 minute and 28 seconds.

Supplementary Note 1: Additional benefits of the Sailfish approach

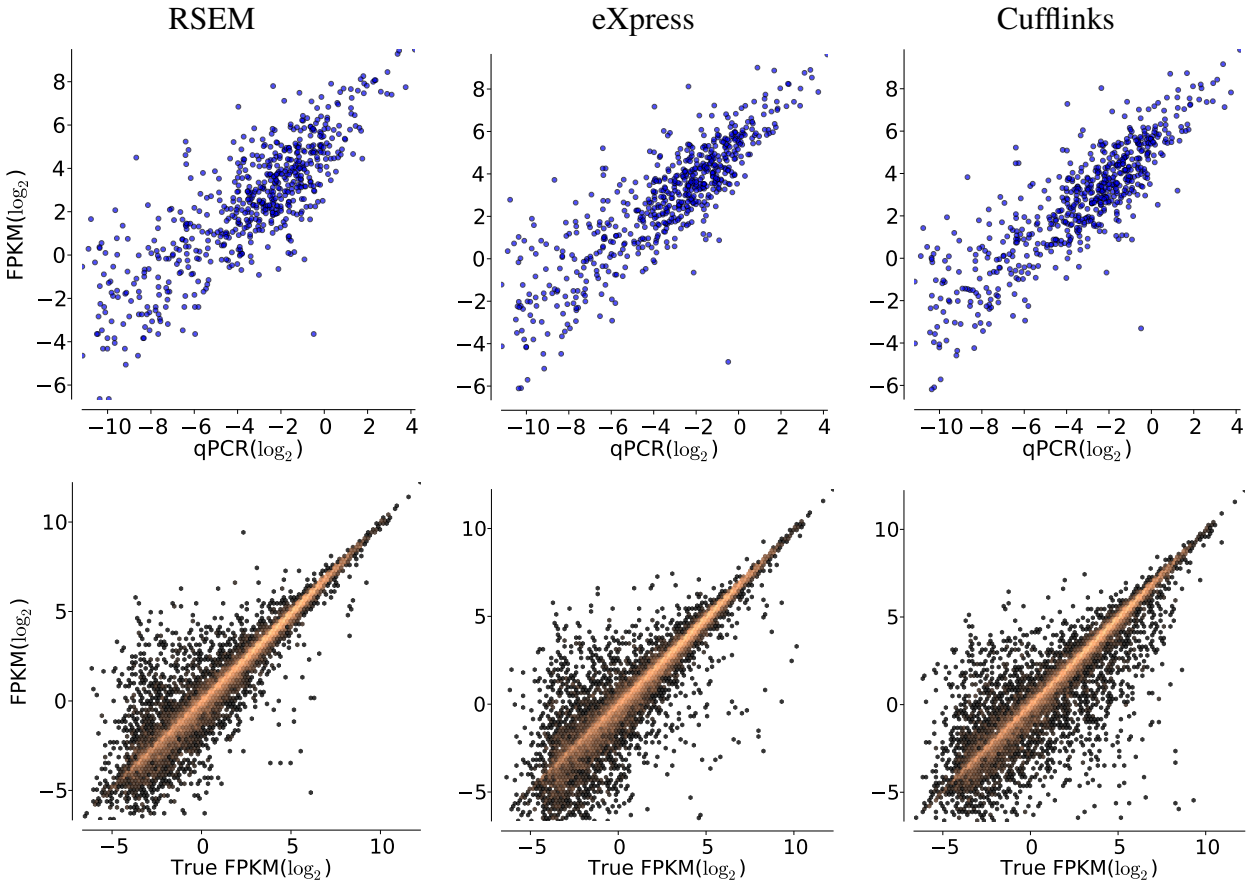
By not requiring read mapping, Sailfish avoids parameters specifying, for example, the number of mismatches to tolerate, total allowable quality of mismatched bases, gap open and extension penalties, whether and how much to trim reads, number and quality of alignments to report from the aligner and pass into the estimation procedure.

An additional benefit of our lightweight approach is that the size of the indexing and counting structures required by Sailfish are a small fraction of the size of the indexing and alignment files required by most other methods. For example, for the MAQC dataset described in Figure 2, the total size of the indexing and count files required by Sailfish for quantification was 3.1Gb, compared with much larger indexes and accompanying alignment files in BAM format used by other approaches (e.g., the 15.5Gb index and alignment file produced by Bowtie [1]). Unlike alignment files which grow with the number of reads, the Sailfish index files grow only with the number of unique k-mers and the complexity of the transcriptome's k-mer composition and are independent of the number of reads.

A third additional benefit is that, like eXpress, the memory usage of Sailfish is bounded by the size and complexity of the transcriptome and therefore independent of the number of reads processed.

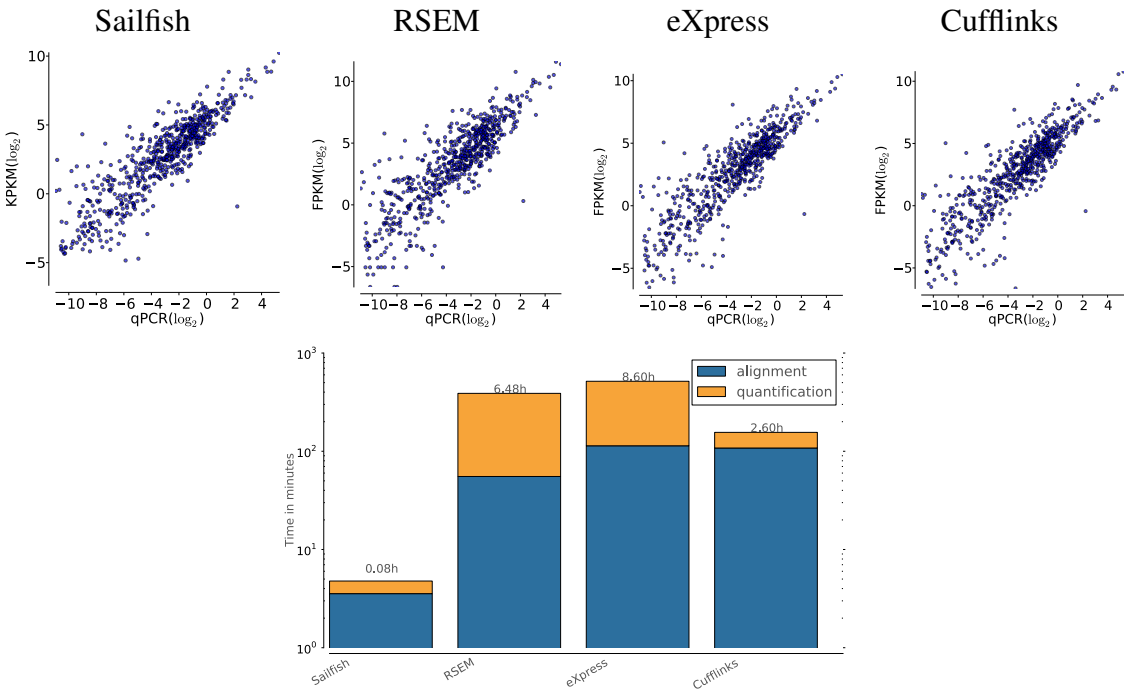
The k-mer-based approach allows for the creation of k-mer equivalence classes that result in a substantial reduction in the number of variables that must be considered during the EM procedure. For example, in the set of reference transcripts for which we estimate abundance using the Microarray Quality Control [2] data (Fig.2), there are 60,504,111 k-mers ($k = 20$), of which 39,393,132 appear at least once in the set of reads. However, there are only 151,385 distinct equivalence classes of k-mers with non-zero counts. Thus, our EM procedure needs to optimize the allocations of 151,385 k-mer equivalence classes instead of 39,393,132 individual k-mers, a reduction by a factor of 260.

Supplementary Figure 3: Correlation plots with qPCR on human brain tissue and synthetic data



Supplementary Figure 3: Correlation plots of RSEM, eXpress and Cufflinks for the data presented in Fig. 2. Each column is labeled with the method whose output was used to produce that column's plots. The top row of plots show the correlation between the computed FPKM and the qPCR-based expression estimates for the human brain tissue. The bottom row of plots shows the correlation between the computed FPKM and the true abundance of each transcript on the synthetic dataset. To generate the results shown here, eXpress was run using its default streaming expression estimation algorithm.

Supplementary Figure 4: Correlation with qPCR on universal human reference tissue



	Sailfish	RSEM	eXpress	Cufflinks
Pearson	0.88	0.85	0.88	0.87
Spearman	0.88	0.85	0.88	0.88

Supplementary Figure 4: The accuracy of four methods on a second dataset from the MACQ [2] study. The reads for this experiment were taken from SRA accession `SRX016367` (92,524,365 35bp single-end reads) and are from a mixture of different tissues (i.e. the Universal Human Reference or UHR). The same set of reference transcripts were used as in Fig. 2 of the main text. The relative accuracy and performance of the methods is similar to what we observed in the other MACQ dataset, with Sailfish, eXpress and Cufflinks all achieving comparable accuracy (all slightly more accurate than RSEM). Sailfish is ≈ 33 times faster than Cufflinks, the closest method in terms of speed.

Supplementary Note 2: Additional details of accuracy analysis

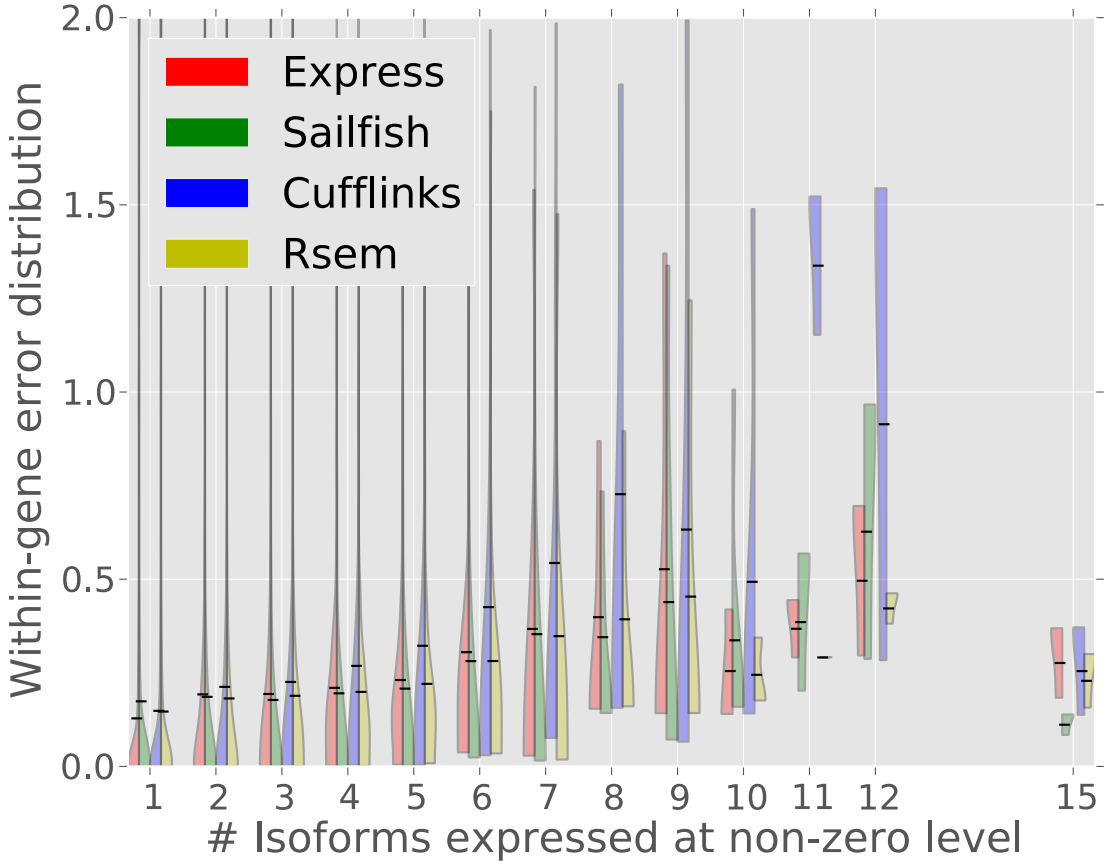
We compare predicted abundances using correlation coefficients (Pearson & Spearman), root-mean-square error (RMSE), and median percentage error (medPE). These metrics allow us to gauge the accuracy of methods from different perspectives. For example, the correlation coefficients measure how well trends in the true data are captured by the methods, but, because the Pearson correlation is taken in the log scale, it discounts transcripts with zero (or very low) abundance in either sample. Both eXpress and Cufflinks produced a number of outlier transcripts, with very low but non-zero estimated abundance, which would substantially degrade some of the metrics (particularly the RMSE). To eliminate the effect of such low-abundance outliers, we set to zero, for all methods, any estimated K/FPKM less than or equal to 0.01, a cutoff chosen because it removed the outliers while discarding only a small number of truly expressed transcripts. The Spearman correlations are not log transformed and therefore include 0 or near-0 abundance transcripts in the synthetic tests. For the qPCR-based tests, due to the relatively low number of transcripts that were experimentally measured, only transcripts with non-zero measured and estimated expression were included in the correlation.

Supplementary Note 3: Parameters for simulated data

The simulated RNA-seq data was generated by the FluxSimulator [3] v1.2.1 with the following parameters.

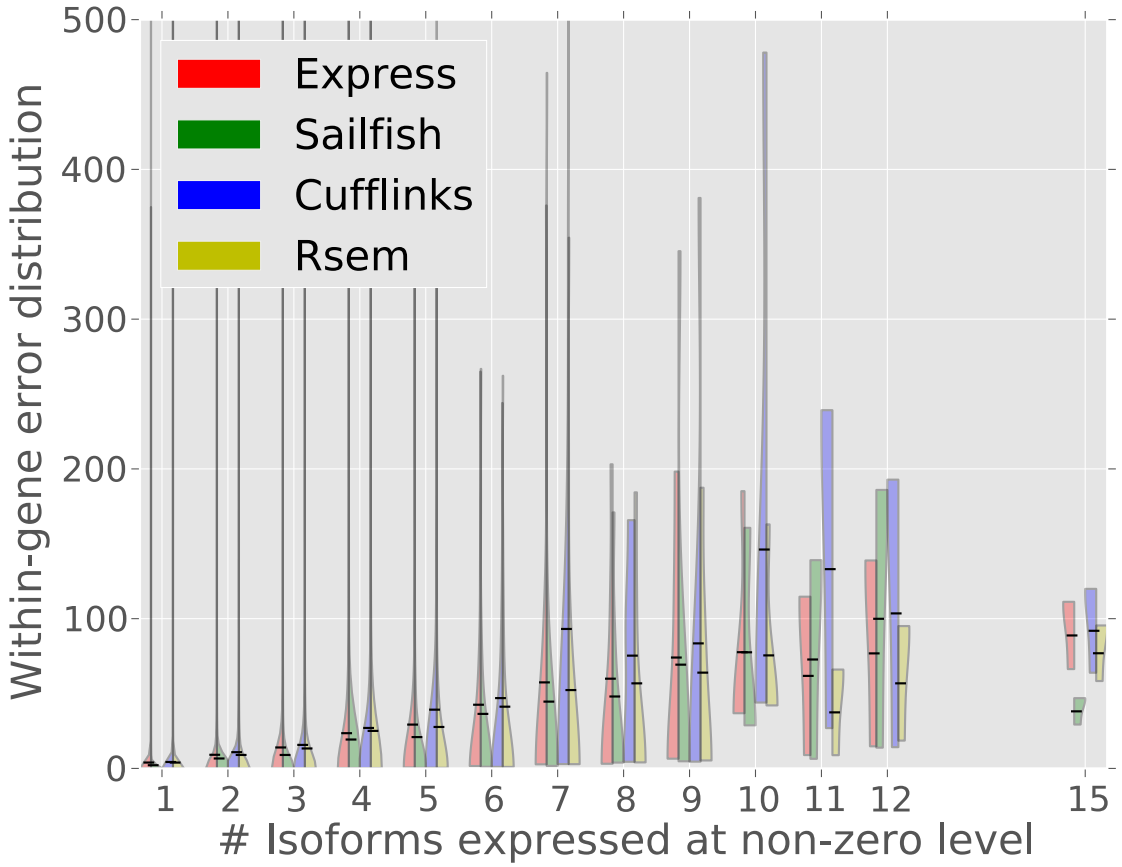
```
### Expression ###
NB_MOLECULES    5000000
REF_FILE_NAME   GRCh37_annotations.gtf
GEN_DIR         GRCh37/chrs
TSS_MEAN        50
POLYA_SCALE     NaN
POLYA_SHAPE     NaN

### Fragmentation ###
FRAG_SUBSTRATE  RNA
FRAG_METHOD     UR
FRAG_UR_ETA     NaN
FRAG_UR_D0      1

### Reverse Transcription ###
RTRANSCRIPTION  YES
RT_PRIMER       RH
RT_LOSSLESS     YES
RT_MIN          500
RT_MAX          5500

### Filtering ####
FILTERING       YES

### Amplification ###
GC_MEAN         NaN
PCR_PROBABILITY 0.05

### Sequencing ###
READ_NUMBER     150000000
READ_LENGTH     76
PAIRED_END      YES
ERR_FILE        76
FASTA           YES
UNIQUE_IDS      NO
```

Supplementary Figure 5: Within-gene relative isoform abundance accuracy



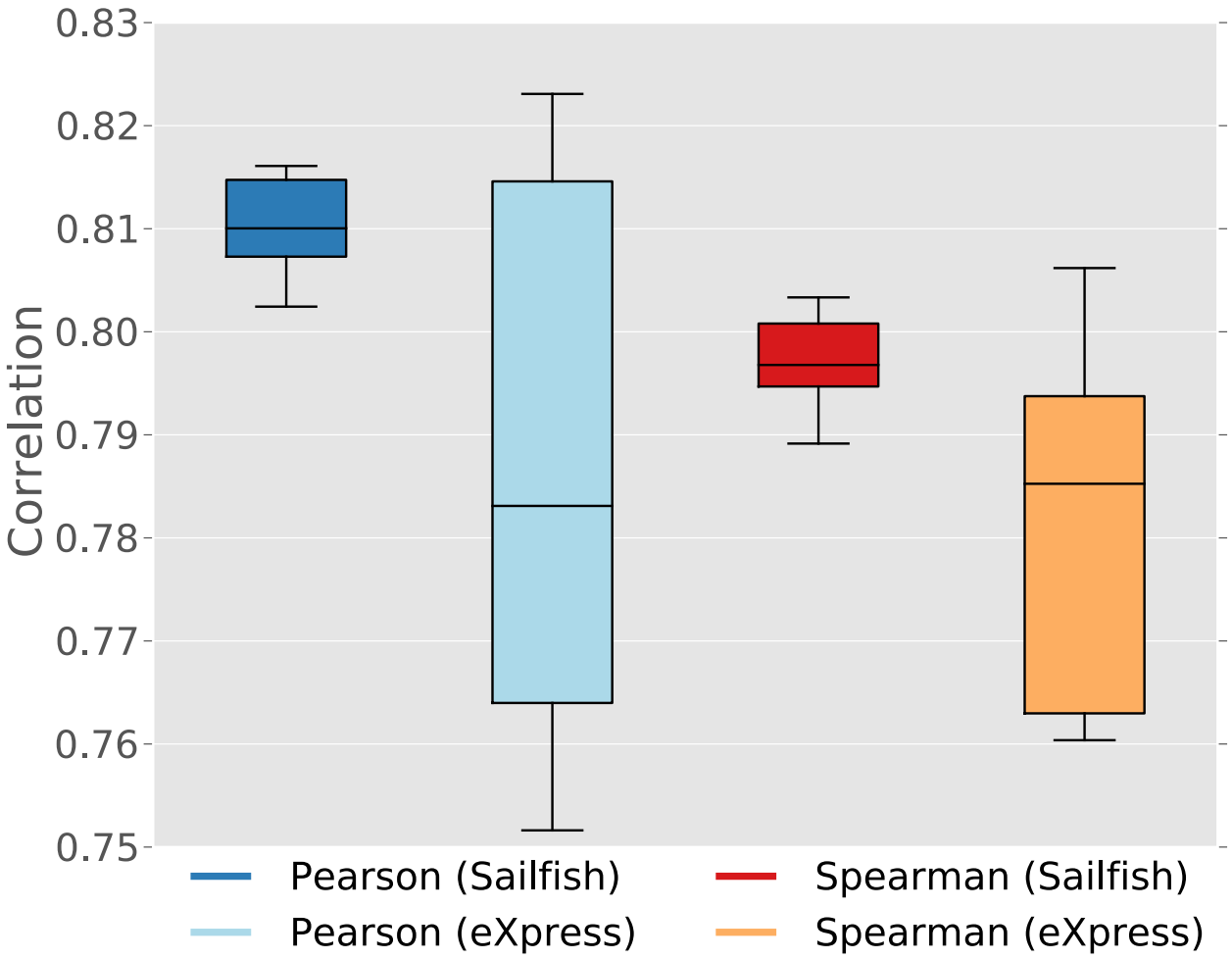
Supplementary Figure 5: Violin plots showing the distribution of relative error of the true vs. estimated within-gene expressed isoform fractions, stratified by the number of isoforms expressed at a non-trivial level (K/FPKM > 0.01 in the simulated data). All transcripts were first grouped by gene, and each gene was placed into a bin labeled by the number of different isoforms expressed with a true K/FPKM > 0.01. Each gene was converted into a vector of isoform fractions by dividing the expression (K/FPKM) of each isoform by the total gene expression. Given the true and estimated isoform fractions of each gene — denoted by $\mathbf{g}$ and $\hat{\mathbf{g}}$ respectively — the relative error was computed as $\varepsilon_g = \|\mathbf{g} - \hat{\mathbf{g}}\|_1$. The violin plots then show the distribution of ε_g for each method and for each gene category. The black bar overlaid on each violin plot denotes the median of the distribution. Even within genes expressing multiple isoforms simultaneously, Sailfish is able to quantify the mixture of isoforms as well as the read-mapping-based approaches.

Supplementary Figure 6: Within-gene absolute isoform abundance accuracy



Supplementary Figure 6: This plot was generated in a fashion similar to supplementary Fig. 5 above, except that the errors in expression estimates are measured in more absolute terms. Again, the genes are stratified based on the number of isoforms expressed with a true FPKM > 0.01. However, we now consider $\mathbf{g}$ and $\hat{\mathbf{g}}$ to be, respectively, the true and estimated un-normalized vectors of expression values (in K/FPKM) for all isoforms belonging to gene g. We then compute $\varepsilon_g = \|\mathbf{g} - \hat{\mathbf{g}}\|_1$, and plot the distributions of ε_g for each method and each gene category. The black bar overlaid on each violin plot denotes the median of the distribution. For complex, multi-isoform genes, Sailfish is able to quantify the relative expressions (in terms of KPKM), not just the mixture, of different isoforms as well as read-mapping-based methods.

Supplementary Figure 7: Robustness to mutations in reference transcripts



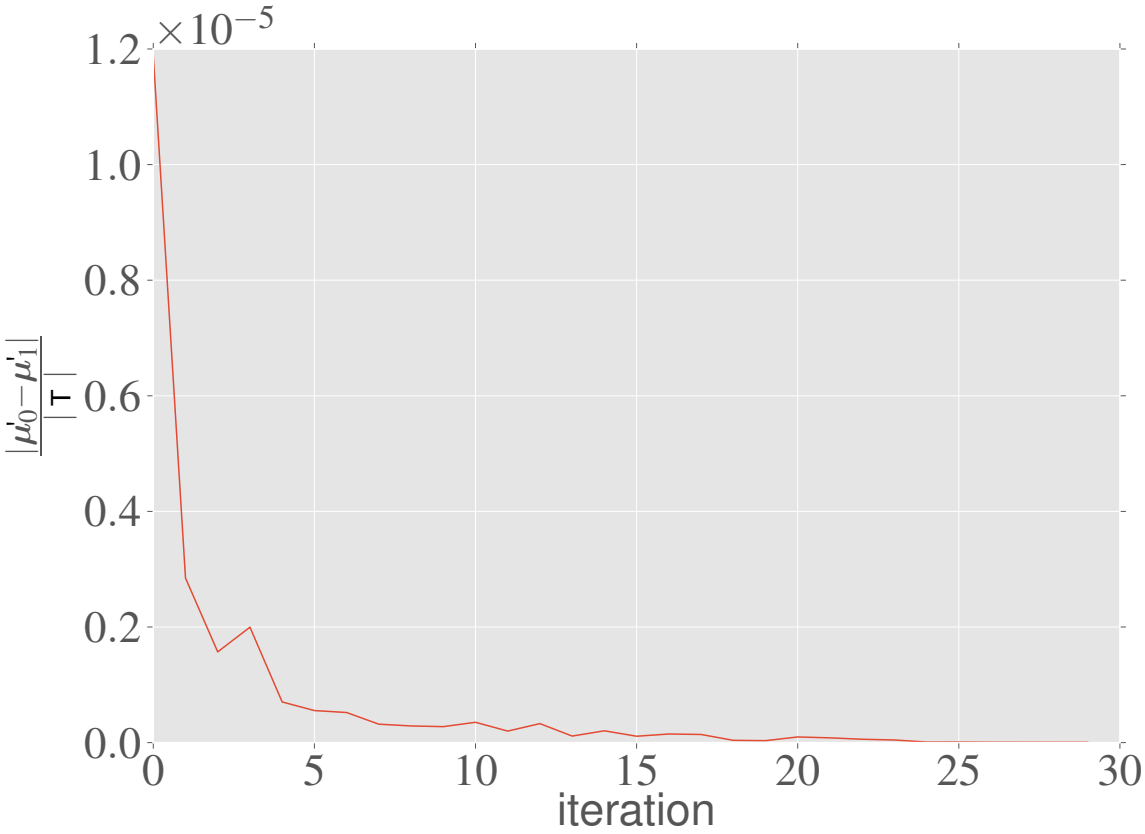
Supplementary Figure 7: For each of 10 independent tests, a mutated transcript set was produced by altering nucleotides randomly in the underlying genome according to a Poisson process with rate parameter 0.05. The transcript reference sequences were then extracted from the mutated genome (using the `gffread` tool, which is part of Cufflinks), and used as targets for expression quantification by Sailfish and eXpress. The box plots show the distribution of Pearson (log-transformed) and Spearman correlation between the qPCR expression estimates and expression estimates produced by Sailfish and eXpress using the `SRX016366` reads, but with reference transcript sets extracted from 10 independently mutated copies of the genome. We observe that, as expected under such a significant mutation rate, the accuracy of the quantification estimates produced by both Sailfish and eXpress are diminished with respect to their corresponding error-free counterparts (see Fig. 2d). However, even under such a significant mutation rate, both methods are able to infer estimates that correlate well with the qPCR-based expression values. We observe that the median accuracy for Sailfish is higher than that of eXpress in these experiments, and the variance of its accuracy over the 10 trials is noticeably smaller.

Supplementary Table 1: Abundance estimation in sequence-redundant human genes

	Sailfish	RSEM	eXpress	Cufflinks
Pearson	0.93	0.95	0.92	0.76
Spearman	0.88	0.89	0.88	0.80
RMSE	21.33	21.71	22.76	41.31
medPE	6.27	9.28	12.05	81.76

Table 1: This table shows the accuracy of Sailfish, eXpress, Cufflinks and RSEM on a sequence-redundant subset of genes using the synthetic human expression data. To construct this sequence-redundant set of genes, we first performed a global pairwise alignment on all of the reference (GRCh37.73) transcripts using the `ggsearch36` global alignment program, which is part of the `FASTA` suite of tools. Let $i(t,t')$ be the percent identity between transcripts t and t' as computed by `ggsearch36`; we define the gene-level similarity between genes g and g' as $i(g,g') = \max i(t,t')$ where t is a transcript of gene g and t' is a transcript of gene g'. Given these gene-level similarities, we define two genes g and g' to be highly sequence-redundant if $i(g,g') \geq 80$. We then extracted all pairs of highly sequence-redundant genes according to this metric and the resulting set of genes constitutes our sequence-redundant subset. This subset of genes has the property that for every gene g in this subset, there exists at least one other gene g' in this subset such that $i(g,g') \geq 80$, though there may be more. We find that on this subset of genes, Sailfish, eXpress and RSEM remain fairly accurate, while the accuracy of Cufflinks suffers somewhat. However, even when quantifying abundance on significantly sequence-redundant genes, Sailfish is able to produce accurate estimates, comparable to or better than those of other read-mapping-based methods.

Supplementary Figure 8: Convergence of relative abundance estimates



Supplementary Figure 8: The average difference between the relative abundance as estimated by two successive applications of the EM step (Algo. 2 lines 1–2) versus iterations of the SQUAREM algorithm (in the Universal Human Reference tissue experiment). We can see that the residual drops off quickly, and appears to have converged before 30 iterations of the SQUAREM procedure have been performed.

References

[1] Ben Langmead, Cole Trapnell, Mihai Pop, Steven L Salzberg, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, 2009.

[2] 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, et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, 24(9):1151–1161, 2006.

[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*, 40(20):10073–10083, 2012.