Supplementary Materials

Table of Contents

Supplementary Results

1 - Correctly identifying transcript boundaries requires additional data

Biases in RNA-seq read coverage have been widely reported[1 2 3 4] and, although several methods have been developed to attempt to remove such bias[5 6], the methods are typically aimed at correcting transcript expression levels rather than correcting read coverage estimates. As such, local, random changes in read coverage make it difficult to determine whether a particular site is a transcript initiation/termination site, or a random fluctuation in read coverage. To compound this problem, even when we restrict our attention to polyadenylated transcripts so that we can use poly(A) spanning reads to identify transcript ends, sequencing bias makes the poly(A) spanning reads much more rare than other read types [7]. For instance, in the modENCODE poly(A)+ data sets, poly(A) spanning reads were roughly 100 times less likely than should have been given uniform read coverage across transcripts.

To demonstrate the confounding effect of read coverage bias on transcript boundary identification, we use the CAGE and poly(A) data to determine the extent to which one could identify TSS and TES sites purely from RNA-seq data. For each gene in Flybase 5.45 (FB5.45) with a BPKM greater than 10, we found the 10 basepair window with the highest amount of CAGE signal, and recorded the ratio of the net base coverage 50 basepairs upstream of the site to 50 basepairs donwstream of the site. We calculated the same statistic for the furthest poly(A) site in each gene. These two sets gave us our positive control set. Next, for each gene, we uniformily sampled 10 random locations from within annotated transcription regions, and calculated the signal ratio to build the negative control set. Finally, we estimated the posterior probability of a site being a gene boundary by direct application of Bayes theorem, where the marginal probability of a promoter and poly(A) site were taken from the GRIT identified CAGE and poly(A) regions.

On average, the signal enrichment ratios were 19.7 and 9.7 for TSS and TES's respectively, versus 1 for the negative control set. Using the known frequency of promoters in the

genome as an estimate of the probability of a promoter and the estimated enrichment ratios, the maximum posterior probability that a given position is a promoter is 67.9%, and occurs when the upstream to downstream signal ratio is 85.1 Similarly, for poly(A) sites, the maximum posterior probability is 35.5% and occurs when the downstream to upstream ratio is 83.2 (Supplementary Fig 3). Thus, even under ideal conditions, RNA-seq coverage alone is likely insufficient to accurately identify transcript boundaries.

2 - Exon switching may regulate protein localization signals

We find that, of the 1727 genes that have alternative N-terminal coding exons, 701 (40.6%) encode multiple protein localization signals. In comparison, 174 of 1205 (14.4%) genes that have alternative C-terminal coding exons encode multiple protein localization signals. As expected, alternative N-terminal coding exons are more likely to encode alternate localization signals (p < 2.2e-16, binomial test). As the majority of known localization signals are N-terminal, the enrichment relative to other alternative exons makes a useful, albeit conservative negative control. Thus, we conservatively estimate that 26.2% of genes that have alternate first coding exons encode multiple protein localization signals. When we performed the same analysis using FlyBase 5.45 gene models, we found that 127 of 544 (23.3%) of genes that contain alternative N-terminal coding exons encode multiple protein localization signals, versus 44/542 (8.1%) of C-termini (p < 1e-11, binomial test).

Supplementary Notes

1 - Parameter estimation

Maximum likelihood estimation

Maximizing the likelihood equation requires optimizing $lhd(Y;\vec{t}) = \sum_i Y_i \log[\sum_j X_{ij} t_j]$, subject to the constraints $t_j \geq 0$ and $\sum_j t_j = 1$. Although this is convex and can be solved using standard convex solvers like CVX [8], the potentially large number of candidate transcripts makes such approaches too expensive to use routinely. We have found that, in practice, a projected gradient ascent method is the most performant (data not shown). We find a starting location by minimizing $\sum_i \left(\frac{Y_i}{\sum_j Y_j} - \sum_j X_{ij} t_j\right)$ s.t. $t_j \geq 0$ and $\sum_j t_j = 1$ using a QP solver. Then, we use projected gradient ascent with a fast simplex projection method [9] until the update differences are less than machine precision. Since the likelihood surface is smooth and convex, this method always converges to the optimum. We have verified that solutions found by the GRIT software package are equivalent to the CVX solutions (data not shown).

Confidence bound estimation

Finding the lower confidence bound for a given transcript, t_j, involves finding the minimum value of $\vec{t}$ which minimizes the i'th component, subject to the restriction that the log likelihood ratio $lhd(Y;\vec{t}_{mle}) - lhd(Y;\vec{t})$ is sufficiently high. We use the objective $lhd(Y;\vec{t}) = Y_O t_O + \sum_i Y_i \log[\sum_j X_{ij} t_j]$ where t_O and Y_O are the estimated fraction and the count of reads that fall outside the gene of interest. This objective accounts for the fact that the number of reads that originates from a given gene locus is random. Because the maximum likelihood estimate of t_O is $\frac{Y_O}{Y_O + \sum_j Y_j}$, we rescale $\vec{t}_{mle}$ by $1 - \widehat{t_O}$ to calculate $lhd(\vec{t}_{mle})$.

Choosing the sparsity parameter λ

For the sparse objective function, $\max_j \left\{ lhd(Y;\vec{t}) - {\lambda}/{t_j} \right\}$, we wish to choose the largest λ that guarantees that the sparse solution, $\vec{t}$, lies with the confidence region, Δ_R. Formally, we wish to choose λ such that $\sum_i Y_i \log\left[\sum_j X_{ij} t_j\right] - {\lambda}/{\left\|\vec{t}\right\|_\infty} \geq \sum_i Y_i \log\left[\sum_j X_{ij} \widehat{t_{mle}}\right] - \frac{1}{2}\chi^2(2\alpha)$, where $\widehat{t_{mle}}$ refers to the maximum likelihood solution, and we use 2α because the confidence bound test is one-sided. Setting $\left\|\vec{t}\right\|_\infty$ to $\max\{\min \vec{t}\}$, the maximum lower confidence bound, $\lambda \leq \frac{1}{2}\max\{\min t_j \colon \vec{t} \in \Delta_R\}[\ \chi^2(2\alpha) - \ \chi^2(\alpha)]$. Even though λ is typically very close to 0 in the unidentifiable case, in such cases very small values of lambda can change the solution substantially because a large portion of the parameter space has likelihood values very close to the maximum.

2 - Sensitivity to tuning parameters

GRIT has two main tuning parameters: one that governs the threshholding of segments with low read coverage, and one that governs the retention of canonical introns.

Changes to the minimal exon read coverage tuning parameter affects the results very little over reasonable ranges. For instance, in the data set we analyzed for the purposes in this manuscript, changing this parameter from 0.01 BPKM to 1 BPKM reduces the recall by less than 1%, and increases the precision by less than 1%. This is consistent with our observation that the limiting factor for transcript construction is junction reads, rather than read coverage within a gene body.

The other important tuning parameter is the canonical intron retention threshold and, unfortunately, the optimal value is a function of the assay type. For instance, we have applied GRIT to total RNA-seq (data not shown) and find that a threshold of 80% percent is necessary to prevent the routine inclusion of unprocessed introns. However, in the poly(A)+ data that we analyzed for this study, a threshold of 5% was sufficient to exclude the vast majority of unprocessed transcripts. We currently err on the side of conservatism, setting this to 80% by default. This setting has the potential to miss retained introns in poly(A)+ RNASeq, but seems to provide good results over a wide variety of organisms and protocols.

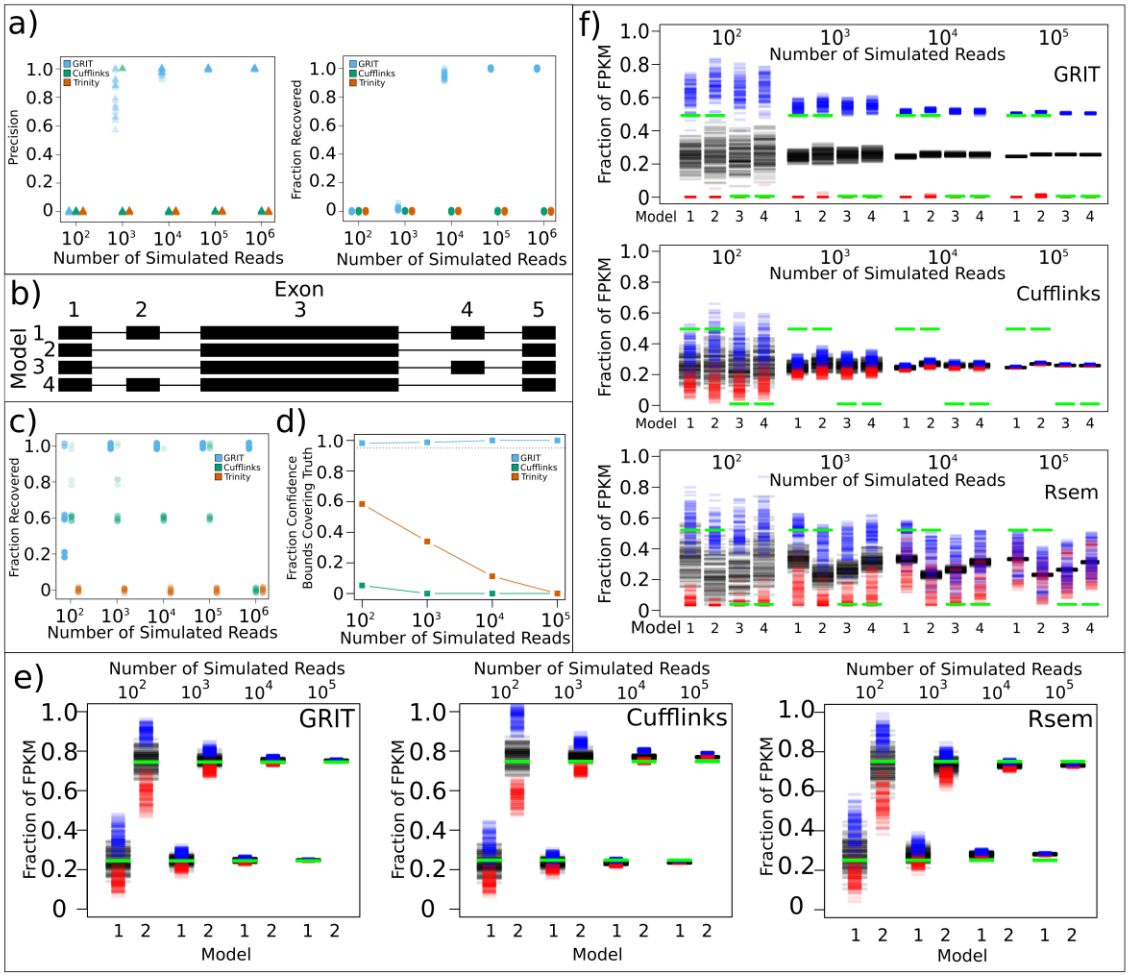
Supplementary Tables

Table 1 – Accession numbers

Data Type	Sample Type	Biological Sample ID	SRA Accession Number
CAGE	Adult Mated Female Heads, 20 days post eclosion	287	SRR488279
CAGE	Adult Mated Male Heads, 20 days post eclosion	290	SRR488280
RNA-seq	Adult Mated Female Heads, 20 days post eclosion	287	SRR070420
RNA-seq	Adult Mated Female Heads, 20 days post eclosion	288	SRR111882
RNA-seq	Adult Mated Male Heads, 20 days post eclosion	290	SRR070421
RNA-seq	Adult Mated Male Heads, 20 days post eclosion	291	SRR070424
poly(A)-site-seq	Adult Mated Female Heads, 20 days post eclosion	288	SRR1151373
poly(A)-site-seq	Adult Mated Male Heads, 20 days post eclosion	291	SRR1151374

Supplementary Figures

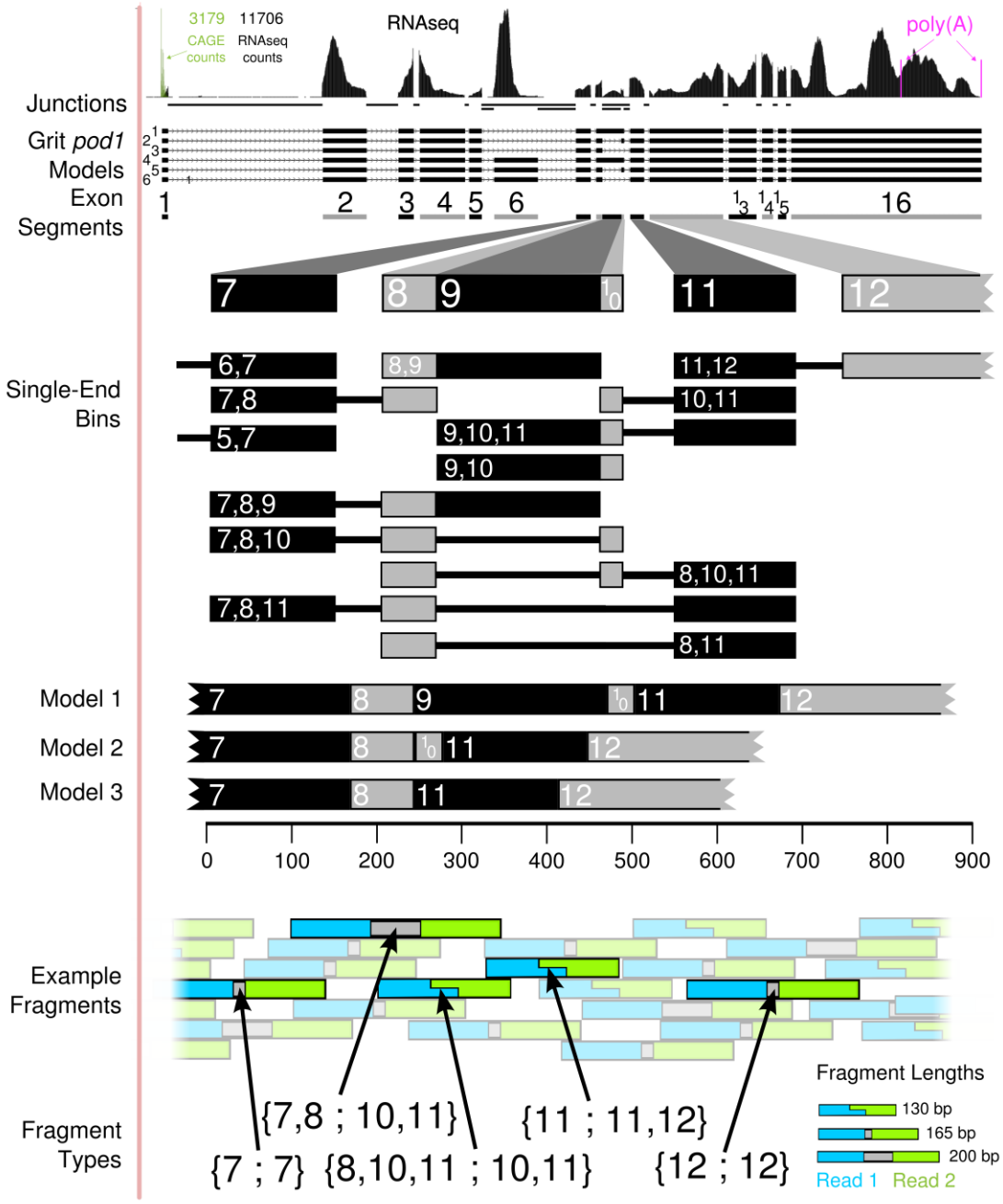
Supplementary Figure 1 - Simulations



Simulations - We used the simulation method described in technique described in Methods – Simulations. **(a) *Dscam1* Simulations:** We simulated from the 38016 potential transcripts identified in Flybase 5.45. Trinity was not able to reconstruct any full length transcripts; Cufflinks was only able to construct a single full length transcript in 1/100 simulations. GRIT recovered most transcripts with high average precision when provided 1000 reads, and was able to reconstruct all 38016 transcripts with perfect precision when provided at 10,000 or more reads. **(b) Simulation Models:** The set of transcript models we simulated from for figure panels **c**, **d, e,** and **f**. Because the middle exon is 600 basepairs - longer than the length of the largest fragment - it is impossible to observe exons 2 and 4 in the same read. Thus the staistical model is not

identifiable when all four transcript isoforms are present. **(c) Transcript Recovery:** We ran 20 simulations, simulating reads in equal proportions from all four models in panel **b**, and found that only GRIT is able to consistently recover all four models with over a thousand reads. Trinity did not correctly recover any transcript models. Cufflinks recovered 2/20 with 100 reads, 2/20 with 1000 reads, 1/20 with 10k reads, and 6/20 with 10 thousand reads. However, because of the shortest path assumption, each time it built all four models it created an artificial TSS or TTS between 20 and 50 basepairs from the true TSS or TTS. When we restricted the transcripts to be equivalent only when the gene boundaries are within 10 basepairs of the truth, then Cufflinks did not correctly identify more than two models correctly. **(d) Confidence Bound Accuracy:** We simulated reads from all for models in panel **b**, with frequencies of 0.49, 0.49, 0.01, and 0.01 for models 1-4 respectively. For each tool, we plotted the fraction of times that the estimated confidence bounds contained the truth. The dashed black line is at 0.95, the expected fraction of times that the confidence bounds should contain the truth. GRIT's confidence bounds are slightly conservative, covering the truth an average of 99% of the time. Because of the identifiability problem, Cufflinks and Rsem confidence bounds are extremely anti-conservative, never covering the truth for n=10000. This is a summary of the data plotted in panel **f**. Note that, because over 30% of genes have both alternate TSS's and alternate TES's, Cufflinks and Rsem have the potential to produce anti-conservative confidence bounds for a large fraction of annotated gene loci. **(e) Identifiable Simulations:** We simulated from models 1 and 2, with frequencies of 0.75 and 0.25 respectively. The green bar is the true frequency. Blue bars identify estimated upper bounds, black bars represent estimated frequencies, and red bars represent estimated lower bounds. All methods perform reasonably well, although Rsem and Cufflinks estimates exhibit a slight bias. **(f) Unidentifiable Simulations:** We simulated from all four models, with frequencies of 0.49, 0.49, 0.01, and 0.01 for models 1-4 respectively. The green bar is the true frequency. Blue bars identify estimated upper bounds, black bars represent estimated frequencies, and red bars represent estimated lower bounds. Because of the identifiability problem, no methods are able to correctly estimate the transcript frequencies. However, only GRIT is able to properly estimate the confidence bounds. The confidence bound estimate accuracy are plotted in panel **d**.

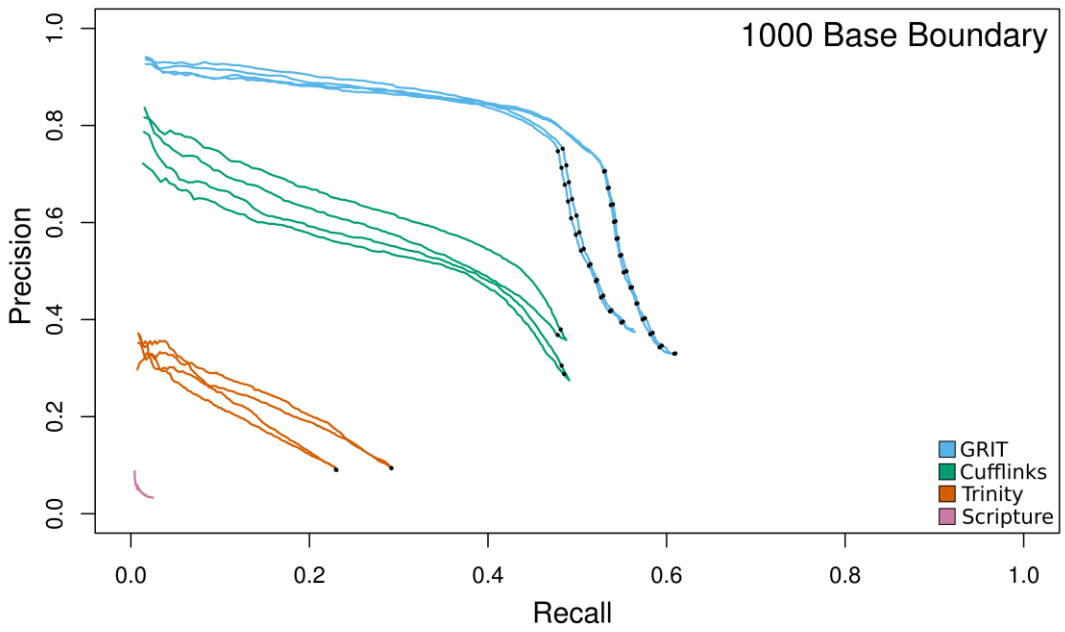
Supplementary Figure 2 - Expression estimation overview



Expression estimation overview - See Section 2.1.4: To identify the set of transcripts in *pod1,* find the set of non-overlapping segments, labeled **exon segments**, with which it possible to reconstruct the transcript set. In the zoomed-in region containing segments 7-12, the possible

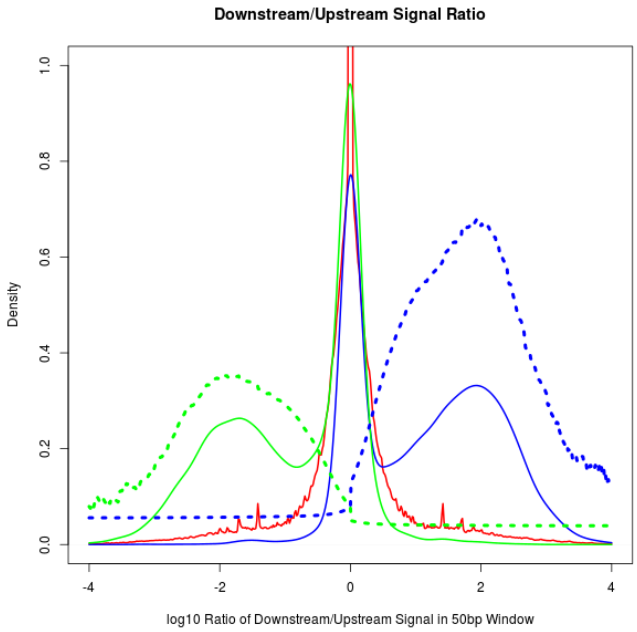
bins, labeled **Single-End Bins**, that can be observed from 75 basepair reads are shown. Next, estimate the fragment length distribution, and then identify the sets of pseudo exons that can be overlapped by paired end reads. The blue and green fragments are possible fragments taken from transcript model 2. For example, in the 200 basepair fragment labeled {7,8;10,11}, read 1 (in blue) overlaps exon segments 7 and 8, while pair 2 (in green) overlaps segments 10 and 11. The fact that read 1 overlaps segments 7 and 8 doesn't give us any additional information about the transcript isoform from which it originated, but the fact that read 2 overlaps 10 and 11 implies that it must have come from either model 2 or 5.

Supplementary Figure 3- Internal structure recall-precision curve



Internal structure recall-precision curve: This is the same plot described in Fig 2, but we increased the boundary match condition to ±1000 basepairs. We do not believe that this is a good measure of tool's performance, but rather gives an upper bound to how well the various tools can perform at predicting internal structure. Even when we ignore gene bounds for the purposes of evaluation, GRIT outperforms other methods.

Supplementary Figure 4 - Identifying gene boundaries solely from RNA-seq



Identifying gene boundaries solely from RNA-seq: The dark red line indicates the marginal distribution of RNA-seq signal across exonic regions. The dark blue and dark green lines indicate the distribution of RNA-seq signal ratios over CAGE peaks and poly(A) sites, respectively. The dashed blue and green lines indicate the posterior probability that a location is a TSS or TES, based solely upon its RNA-seq signal ratio. For instance, the dashed blue line peaking at 0.65 indicates that it is impossible to identify a CAGE site from RNA-seq signal ratio alone with greater than 65% certainty.

Bibliography

1. Risso, D., Schwartz, K., Sherlock, G. & Dudoit, S., GC-content normalization for RNA-Seq data. *BMC bioinformatics* **12** (1), 480 (2011).
2. Bullard, J., Purdom, E., Hansen, K. & Dudoit, S., Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC bioinformatics* **11** (1), 94 (2010).
3. Hansen, K. D., Brenner, S. E. & Dudoit, S., Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic acids research* **38** (12), e131--e131 (2010).
4. Wang, Z., Gerstein, M. & Snyder, M., RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10** (1), 57-63 (2009).
5. Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A. & Dewey, C. N., RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* **26** (4), 493-500 (2010).
6. Zheng, W., Chung, L. M. & Zhao, H., Bias detection and correction in RNA-Sequencing data. *BMC bioinformatics* **12** (1), 290 (2011).
7. Cui, P. *et al.*, A comparison between ribo-minus RNA-sequencing and polyA-selected RNA-sequencing. *Genomics* **96** (5), 259-265 (2010).
8. Grant, M., Boyd, S. & Ye, Y., in *CVX: Matlab software for disciplined convex programming* (2008).
9. Duchi, J., Shalev-Shwartz, S., Singer, Y. & Chandra, T., *Efficient projections onto the l 1-ball for learning in high dimensions*, presented at Proceedings of the 25th international conference on Machine learning, 2008 (unpublished).