Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation

Published in final edited form as:

Nat Biotechnol. 2016 December ; 34(12): 1287–1291. doi:10.1038/nbt.3682.

Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation

Michael I. Love[1,2], **John B. Hogenesch**[3], and **Rafael A. Irizarry**[1,2,*]

[1] Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA, USA

[2] Department of Biostatistics, Harvard TH Chan School of Public Health, Boston, MA, USA

[3] Department of Pharmacology, Institute for Translational Medicine and Therapeutics, University of Pennsylvania School of Medicine, Philadelphia, PA, USA

We find that current computational methods for estimating transcript abundance from RNA-seq data can lead to hundreds of false-positive results. We show that these systematic errors stem largely from a failure to model fragment GC content bias. Sample-specific biases associated with fragment sequence features lead to mis-identification of transcript isoforms. We introduce alpine, a method for estimating sample-specific bias-corrected transcript abundance. By incorporating fragment sequence features, alpine greatly increases the accuracy of transcript abundance estimates, enabling a fourfold reduction in the number of false positives for reported changes in expression compared with Cufflinks. Using simulated data, we also show that alpine retains the ability to discover true positives, similar to other approache**s.** The method is available as an R/Bioconductor package that includes data visualization tools useful for bias discovery.

Obtaining transcript abundance information from RNA sequencing (RNA-seq) experiments relies on complex methods implemented in software such as Cufflinks [1] and RSEM [2]. However, RNA-seq data can suffer from sample-specific biases as a result of RNA extraction and library preparation steps[3–5]. Methods for estimating gene and transcript abundance attempt to mitigate the effect of technical biases by estimating sample-specific bias parameters. For gene-level expression, common normalization methods use the GC content and length of the gene[6-8], or identify batch effects by detecting structure in expression measurements that are common across genes and not associated with the experimental design[9-11]. At the transcript level, sample-specific biases that current methods correct for include the fragment length distribution induced by size selection, positional bias along the transcript due to RNA degradation and mRNA selection techniques, and sequence-based bias in read start positions arising from the differential binding efficiency of random

* Corresponding author: rafa@jimmy.harvard.edu.

Author Contributions:
MIL and RAI designed the method. MIL, JBH and RAI wrote the manuscript.

Competing Financial Interests Statement:
The authors declare that they have no competing interests.

hexamer primers[2,12-16] (Figure 1a and Supplementary Table 1). Even so, it is common to observe extreme variability in the coverage of RNA-seq fragments along transcripts that is purely technical and sample-specific[17] and not explained by current bias models, which confounds current methods designed to identify and quantify transcripts[18] (Figure 1b).

To investigate the cause of systematic errors in transcript abundance estimates, we used existing data to evaluate currently available software. We downloaded 30 RNA-seq samples from the GEUVADIS Project of lymphoblastoid cell lines derived from the Toscani population, 15 of which were sequenced at one center and 15 at another[19] (Supplementary Table 2). We ran Cufflinks with its sequence bias removal option described in Roberts et al. [14] turned on. Performing a t-test on $\log_2$(FPKM + 1) values from Cufflinks across centers, and filtering on Benjamini-Hochberg adjusted p values for a target 1% false discovery rate (FDR) resulted in 2,510 transcripts out of 25,588 (10%) with average FPKM greater than 0.1 reported as differentially expressed (Figure 2a) However, we expect that nearly all of the 2,510 transcripts reporting differential expression are false positives, as random permutations of samples resulted in no transcripts with differential expression. RSEM had similar rates of differentially expressed transcripts across center (Supplementary Figure 1). When comparing across center, 619 out of 6,761 genes with multiple isoforms and average FPKM greater than 0.1 for one or more isoform had changes in the reported major isoform across centers using FPKM values estimated by Cufflinks (Supplementary Table 3).

We focused on genes with two isoforms to more easily identify what features underlie the difference in estimated expression. Out of 5,716 transcripts from genes with two isoforms in which at least one isoform had FPKM greater than 0.1, 566 transcripts reported differential expression across center according to Cufflinks estimated abundances, at a target FDR of 1%. For genes with one or more isoforms reported differentially expressed, the regions of the isoform that were exclusive to one or the other isoform had much higher GC-content (mode at 70% compared to 50%) than expected by chance (Figure 2b, Wilcoxon $p < 0.0001$).

An example of a gene with differences in expression estimates across center is USF2 (Figure 2c). It is often short sequences that distinguish isoforms of a gene, which in some cases include stretches of high GC content. Because methods such as Cufflinks and RSEM employ a likelihood model that does not account for differences in coverage due to fragment sequence features like GC content, the drop in coverage for samples from center 1 results in a shift in expression estimates from NM003367 to NM207291, which does not include the high GC exon (Figure 2d). The samples from center 1 have dramatically reduced representation of high GC fragments compared to center 2, even after adjusting for differences due to random hexamer priming bias (Figure 2e).

A GC content bias is often observed in high-throughput sequencing data, with fragments of certain GC content under-represented, and can be partially attributed to PCR amplification during library preparation[20]. For DNA-seq, this bias is best corrected by modeling at the scale of the fragment[21]. While one existing method for estimating percent of isoform expression assumed a positive linear relationship between exon counts and exon GC

content[22], we confirmed other findings [21], that the GC content effect was often nonlinear and highly sample-specific, therefore requiring sample-specific bias estimation using smooth functions of fragment GC content. Additionally, stretches of high GC content sequences within a fragment can have an influence on whether a fragment will be amplified and sequenced [23].

To account for this bias, we built a bias modeling framework called alpine, using previously described bias features (fragment length, relative position, and read start sequence) as well as fragment GC content and the presence of long GC stretches within the fragment [23] to model the number of times a potential fragment of a transcript was observed (0,1,2,...). alpine takes as input a set of gene annotations and the RNA-seq reads aligned to the genome. We considered all potential fragments with lengths within the center of the fragment length distribution, at all possible positions consistent with the transcript's beginning and end. alpine employs a Poisson generalized linear model (GLM) for the count of each potential fragment within the transcript using the bias features described above, estimating bias parameters for each sample separately (Supplementary Note). The read start sequence bias parameters are estimated by alpine using the variable length Markov model (VLMM) proposed by Roberts et al. [14] and implemented in Cufflinks. After estimating bias parameters, alpine can predict transcript coverage or produce bias-corrected estimates of transcript abundance.

We used a benchmarking dataset of 1,062 human in vitro transcribed (IVT) and sequenced cDNA clones mixed at various concentrations with mouse total RNA[17]. We focused our analysis on 64 of the IVT transcripts as exhibiting “high unpredictable coverage”[17]. We compared the predictive power of four bias models implemented in alpine: a read start model, a fragment GC content model, a fragment GC content and GC stretches model, and a model including all of these biases (Supplementary Figure 2). Additionally we compared the predictive power of the alpine bias models to the read start bias model mseq [12] that trains multiple additive regression trees (MART) on the local sequence context surrounding read start positions. The mseq method does not itself provide estimates of transcript abundance, but can be used to train a read start bias model and therefore to predict start locations of single-end reads for a test set of transcripts. A good bias model should be able to predict the pattern of fragment coverage along a transcript. For assessing predictions, we used 2-fold cross validation, such that two models were trained on two halves of the data, and always evaluated on transcripts that were not in the training set. For mseq, fragment coverage was extended from predicted start positions from both read pairs using the median fragment length. Predictive power was measured as the percent reduction of mean squared error in explaining raw fragment coverage, compared to a null model that predicted uniform coverage across the transcript. The models that included fragment GC content doubled the predictive power of the read start bias models, both the Cufflinks VLMM read start bias model implemented within alpine and mseq, which performed similarly (Figure 3a). The model that also included the information about GC stretches was more predictive than the model with just the fragment GC content, although only slightly so. The fragment sequence models accurately capture the drops in coverage that were not captured by the read start sequence models (Figure 3b, Supplementary Figures 3-4).

We then used our approach to compare the transcript abundance estimates from one center against the other in the 30 GEUVADIS samples. To more clearly show the performance with respect to differential isoform usage, we focused on 5,676 transcripts from genes with two isoforms, with average FPKM values estimated by Cufflinks greater than 0.1 for one of the two isoforms, and such that the two isoforms had at least one overlapping basepair. We compared $\log_2(\text{FPKM} + 1)$ estimates across center using a t-test. We found that including bias terms for fragment GC and GC stretches resulted in a fourfold decrease in the number of false positives at a target FDR of 1%: Cufflinks reported 562 differentially expressed transcripts, while alpine reported 141 (Figure 3c-d). Using a more conservative Bonferroni correction, Cufflinks reported 157 differentially expressed transcripts across center with FWER of 1%, while alpine reported only 37. In general, alpine greatly reduced across-center significant differences while within-center coefficient of variation of abundance estimates remained the same as for Cufflinks (Supplementary Figures 5-8).

Likewise we observed reduced across-center differences for estimation of isoform percentages within the 2,838 genes with two isoforms. For each gene, we calculated the estimated percent expression of the major isoform for center 1 (a number ranging from 50% to 100% by definition), against the estimated percent expression of that same isoform in center 2. Correcting for fragment sequence bias using alpine reduced the number of extreme predicted changes in isoform percent when comparing across center (Supplementary Figure 9). An example of false positive for isoform switching is the two-isoform gene BASP1 (Figure 3e-f). We further compared the performance of alpine against new lightweight methods for estimating transcript abundance, Sailfish [24], kallisto [25], and Salmon [26]. Note that four of the methods evaluated, Cufflinks, kallisto, Salmon, and Sailfish all were run with read start sequence bias correction turned on (Supplementary Note) and obtained similar results (Supplementary Figure 10). When restricting positives to those transcripts with Benjamini-Hochberg adjusted p value less than 1% and additionally requiring that the $\log_2$ fold change be above a threshold (0.5, 1, or 2), alpine consistently had a lower percent of false positives out of the set of 5,676 transcripts than all other methods (Supplementary Figure 11).

Using only read start bias terms in the alpine model did not provide visible improvements (Supplementary Figure 12). alpine bias estimates for the model including all bias terms (fragment length, read starts, fragment GC and GC stretches, relative position in transcript) are shown in Figure 2e and Supplementary Figure 13. The lightweight quantification methods generated the same kind of isoform-switching errors as Cufflinks and RSEM (Supplementary Figures 14-15), and as MISO [27], a statistical method for estimating isoform percentages within multi-isoform genes (Supplementary Figure 16). While MISO performed better than the existing transcript abundance estimators in consistency of isoform identification, alpine reported less than half of the large isoform switches across center reported by MISO (Supplementary Figure 9). Fitting the fragment sequence model does require more computational effort than other models, though our implementation ran in comparable time to the cuffquant step of the Cufflinks suite (Online Methods).

To determine if recovery of true positives was maintained by alpine, we performed a sensitivity analysis using an RNA-seq fragment simulator, Polyester [28], to generate 30

samples with fragment GC content dependence estimated from the 30 GEUVADIS samples, after having removed read start sequence bias (Figure 2e). In one simulation, differential expression of 10% of transcripts was simulated across a condition confounded with sequencing center, while in another simulation the condition was balanced with sequencing center. We then ran alpine, Cufflinks, RSEM, kallisto, Salmon, and Sailfish on the simulated data and compared the true positive rate and false positive rate under the confounded and balanced designs (Online Methods). alpine had the highest sensitivity for a given specificity in the confounded design (Figure 3g-h), while methods performed similarly for the balanced design (Supplementary Figure 17). alpine had the highest accuracy in estimating the true expression values compared to the other methods (Supplementary Figures 18-19), with low median absolute error in estimating the percent of isoform abundance for genes with two isoforms, comparable to RSEM (Supplementary Figure 20).

To confirm that library preparation contributes to the systematic errors we attributed to fragment GC bias, we downloaded a subset of samples from the SEQC Consortium[4], including libraries of the same samples that were prepared and sequenced at three different sites and libraries prepared at a separate site and only sequenced at the three sites (Supplementary Table 4). We found that fragment GC content dependence was strongly associated with the location of library preparation, and not with the location of sequencing (Supplementary Figure 21).

To explore the extent to which different library preparation protocols affect the fragment GC dependence, we downloaded a subset of samples from the ABRF Next-Generation Sequencing Study[29] that were prepared using either ribosomal RNA depletion or poly-A selection (Supplementary Table 5). We observed little change in the shape of fragment GC dependence across protocol, and a strong effect in the positional bias, with poly-A selected samples having highest coverage at the 3' end of transcripts, as reported by the ABRF study authors[29] (Supplementary Figure 22).

Even though the fragment GC content dependence did not differ greatly across protocol in this dataset, we evaluated the extent to which alpine was able to remove systematic bias by modeling positional bias. In a comparison of expression estimates across protocol, alpine reported the fewest number of false positives, controlling at 1% FDR and at 1% FWER (Supplementary Figure 23). The large number of transcripts with false positive differences in abundance reported across protocol, in the range of 10-14,000 at 1% FDR for all methods out of 28,000, suggests that none of the existing methods evaluated including alpine can remove the bulk of systematic bias seen across protocol.

The relationship of the samples in the ABRF dataset allowed an assessment of the measurement accuracy in terms of the methods' recovery of expected mixing ratios (Online Methods). In Supplementary Figures 24-25 we assess the different methods' ability to recover the expected mixing ratio of C/D given measurements of A/B for the transcripts in the top 25% of abundance (as suggested by Li et al. [29]). When quantifying the number of transcripts whose C/D ratio was within 10% of the expected value, RSEM had the highest recovery for both protocols, though alpine also had consistently high recovery (within 5% of

the top method). Overall, the mixing ratio recovery for all methods was higher for poly-A selected samples (65-75%) compared to ribosomal RNA depleted samples (45-65%).

Systematic errors and batch effects are a continuing cause of concern for RNA-seq experiments. Large-scale, block design, and well-documented transcriptome sequencing projects such as those performed by GEUVADIS[3,19], SEQC[4], and ABRF[29] allow the study of technical biases, such as fragment GC content bias, and the creation of computational methods that correct these biases. Indeed, 't Hoen et al.[3] observed that slight differences in average GC content across samples lead to differences in quantification for transcripts. We note that our findings reflect general systematic errors and not just differences induced by batch effects. The problem we identify holds for within-sample transcript abundance estimates for samples from a single center and for samples in small-scale experiments. There are likely to be many incorrectly reported major isoforms and biased abundance estimates for experiments that show strong dependence of the fragment rate on fragment GC content (e.g. Figure 2e), unless these are explicitly corrected for using methods that model fragment sequence bias. Strong GC content bias was found for some samples for all public datasets examined. Fold change thresholds[4] are not an appropriate solution to the particular problem presented here, because fold changes induced by technical bias are often larger than those potentially of biological interest.

Here we demonstrated specificity using data prepared by different sequencing centers, and sensitivity using simulation. New benchmarking experiments would be valuable for further sensitivity analysis, experiments where the true isoform or set of isoforms are known, and in which characteristically highly-variable profiles of transcript coverage are obtained by following as closely to the steps of a standard RNA-seq experiment as possible. While the sequence features we included in our model provided substantial improvements over existing methods, we hypothesize that more variability can be explained by discovering new predictive features. alpine provides a modular framework that facilitates further exploration, which will prove useful for optimization of protocols to reduce fragment GC content bias, preferable to computational corrections.

Online Methods

RNA-seq read alignment and quantification

IVT-seq FASTQ files made publicly available by Lahens et al.[17] were downloaded from the Sequence Read Archive. Paired-end reads were aligned to the human reference genome contained in the Illumina iGenomes UCSC hg19 build, using STAR version 2.5.0[31]. The exons of the GenBank transcripts were read from the feature_quant.txt files posted to GEO, and the list of transcripts with high unpredictable coverage was downloaded from the additional files of Lahens et al.[17].

For the GEUVADIS and ABRF datasets, the same computational pipeline was used. GEUVADIS FASTQ files made publicly available by Lappalainen et al.[19] were downloaded from the European Nucleotide Archive (Supplementary Table 2). ABRF FASTQ files made publicly available by Li et al.[29] were downloaded from the European Nucleotide Archive (Supplementary Table 5). Paired-end reads were aligned to the human reference genome

contained in the Illumina iGenomes UCSC hg19 build, using TopHat version 2.0.11[32] (for Cufflinks) and STAR version 2.5.0 (for alpine). The genes.gtf file contained in the Illumina iGenomes build containing RefSeq transcripts was filtered to genes on chromosomes 1-22, X, Y and M, and provided to Cufflinks, RSEM and alpine as gene annotation.

For the SEQC dataset, the bias estimation steps of alpine were run, using STAR version 2.5.0 and the same gene annotation as above. The SEQC FASTQ files made publicly available by Su et al.[4] were downloaded from the European Nucleotide Archive (Supplementary Table 4).

Transcript quantification details, including the description of the alpine model and the commands used for running other software are provided in the Supplementary Note. Generating the bias coefficients for the 30 GEUVADIS samples using the alpine software required 40 minutes using 6 cores and 25 Gb of memory. Using alpine to estimating the transcript abundances for 5,676 transcripts from two-isoform genes for 30 samples required 4 hours using 30 cores and 55 Gb of memory.

Training and testing mseq on IVT-seq data

mseq version 1.2 Li et al.[12] was used to model read start bias on the IVT-seq dataset, using the same training- and test-set splits as used by alpine. martTrain was run with interaction depth of 10 and 2000 trees (the defaults) using 15 basepairs to the left and right of the position to be modeled. martPred and getPredCount were used to predict expected read start counts for positions in the test transcripts. As suggested in the mseq documentation, read starts from both ends of the fragment were modeled by providing forward and reverse strand data for each transcript. To generate predicted fragment coverage, the median fragment length was used to extend fragment coverage from mseq predicted read starts.

Simulation

Polyester[28] was used for the RNA-seq fragment simulator, which models variability in expression levels across biological replicates and which was easily extended to include the exact fragment sequence bias obtained from the GEUVADIS dataset. The default simulation parameters were used except the size parameter for the overdispersion was set to 100 (corresponding to a negative binomial dispersion parameter of 0.01). 300 genes with a single isoform, 300 genes with two isoforms, and 300 genes with three to five isoforms were simulated for 30 samples. Average gene-level FPKM estimates across the GEUVADIS samples as estimated by Cufflinks were used for the simulation. Expression levels for the isoforms of genes were determined by multiplying the gene-level FPKM value by a random vector from a flat Dirichlet distribution.

Reads were assigned to the transcripts according to the FPKM values and assuming an experiment with a total of 60 million paired-end reads. Paired-end reads were generated and randomly kept or discarded according to a probability derived from the GC content bias curves for 30 GEUVADIS samples in Figure 2e. As the process of discarding pairs of reads would result in unequal final library size, the simulation process was repeated, after first scaling the initial target library size higher such that the final library size for the experiment would be equal to 60 million paired-end reads in expectation. Note that the GC content

curves used in the simulation (Figure 2e) were estimated after removing the read start bias using the Cufflinks VLMM, and so the fragment sequence effect observed is not a proxy for read start bias. Simulated paired-end reads were shuffled before being used as input to quantification methods.

10% of transcripts were selected to be differentially expressed across condition, with equal chance of twofold up- or down-regulation. In one simulation, the condition was confounded with the sequencing center (15 samples against 15 samples), and in another simulation, condition was balanced across sequencing center. A t-test was performed on $\log_2$(FPKM + 1) values. For kallisto, Salmon, and Sailfish, $\log_2$(TPM + PC) was used, with a pseudocount corresponding to 1 on the FPKM scale. For the balanced design, sequencing center was added as a blocking term to a linear model for differential expression. Note that the balanced design with blocking factor represents the best-case scenario for the competing methods, as the batches were known exactly and the residual degrees of freedom was high, such that the batch effects on transcript abundance estimates could be precisely estimated and removed by the linear model. In contrast, scenarios with total or partial confounding, unknown batches, and sample-specific deviations within batches are typical in RNA-seq experiments and not represented by the balanced design simulation.

Specificity analysis in ABRF samples

Testing specificity on the ABRF dataset was performed in a similar manner as in the GEUVADIS dataset, by performing a t-test on log2(FPKM + 1) abundance estimates (or using TPM for kallisto, Salmon, and Sailfish with a pseudocount corresponding to 1 on the FPKM scale). Only the coding transcripts, as annotated by RefSeq, were used for the analysis, as the poly-A selection protocol is not designed to capture all the non-coding transcripts. The alpine software was run with bias correction terms for fragment length, relative position, and fragment GC content, and other methods were run with their bias correction arguments turned on, including a positional bias correction term for RSEM (Supplementary Note). A single scaling factor for FPKM or TPM values was calculated for each method to adjust for the removal of the non-coding transcripts. This factor was calculated by dividing, for each transcript, the average of abundance estimates for poly-A selected samples by the average for ribosomal RNA depleted samples. The median of this ratio over all coding transcripts was used to scale the ribosomal RNA depleted samples, similar to the median-ratio method for library normalization[33].

In order to test for differences due to protocol, for each transcript, a linear model was fit with the coefficients

$$Y_s = \beta_A A_s + \beta_B B_s + \beta_C C_s + \beta_D D_s + \beta_{polyA} P_s + \varepsilon_s$$

where Y_s is a $\log_2$ transformed expression estimate for a single sample s. A_s is an indicator variable taking a value 0 if sample s is not reference sample A and 1 if sample s is A, and similarly for B_s, C_s, and D_s. P_s is an indicator variable indicating if sample s was produced using the poly-A selection protocol. The polyA term then provides the difference in $\log_2$ abundance values between poly-A selected and ribosomal RNA depleted samples. Null

hypothesis tests for this coefficient being equal to zero were performed generating two-tailed Wald test *p* values. *p* values from each transcript were adjusted using the Benjamini-Hochberg method controlling the FDR and the more conservative Bonferroni method controlling the FWER.

Accuracy in mixing ratios for ABRF samples

As the C and D samples in the ABRF dataset were created by mixing 3:1 and 1:3 ratios of the A and B samples respectively, a comparison of C/D and A/B for the different methods can be used to generate a measure of accuracy in estimating transcript abundance. The same calculation was used as described by Su et al.[4] in the Methods section of the SEQC paper. The expected mixing ratio of C/D can be calculated as a ratio of polynomials given the value of A/B from the abundance estimates. The same adjustment as used by Su et al. [4] was used to account for the different ratios of mRNA to total RNA in A and B samples. Mixing ratios were evaluated for the transcripts in the top 25%, and for which both samples A and B had positive abundance estimates (> 0.1 on the FPKM scale).

Code availability

alpine version 0.1.2 was used in this manuscript. The alpine software is made available at:

http://bioconductor.org/packages/alpine

The modified branch of Polyester including fragment sequence bias used here is made available at:

https://github.com/mikelove/polyesterAlpineMs

Code for producing the simulation is made available at:

https://github.com/mikelove/fragmentBiasSimulation

alpine is implemented in the R language using core Bioconductor[34] packages. Information for all of the fragment features is generated using the packages GenomicAlignments, Biostrings, and GenomicRanges[35].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors are grateful for helpful suggestions from Yuval Benjamini, Wolfgang Huber, Nicholas Lahens, Luca Pinello, Clifford Meyer, Rob Patro, Zhonghui Xu, and Yun Li. MIL was supported by NIH grant 5T32CA009337-35. JBH was supported by NIH R01 grant HG005220, the National Institute of Neurological Disorders and Stroke (5R01NS054794-08 to JBH), the Defense Advanced Research Projects Agency (DARPA-D12AP00025, to John Harer, Duke University). RAI was supported by NIH R01 grant HG005220.

References

1. Trapnell, Cole; Williams, Brian A.; Pertea, Geo; Mortazavi, Ali; Kwan, Gordon; van Baren, Marijke J.; Salzberg, Steven L.; Wold, Barbara J.; Pachter, Lior. Transcript assembly and quantification by

RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010; 28(5):511–515. [PubMed: 20436464]

2. Li, Bo; Ruotti, Victor; Stewart, Ron M.; Thomson, James A.; Dewey, Colin N. RNA-seq gene expression estimation with read mapping uncertainty. Bioinformatics. 2010; 26(4):493–500. [PubMed: 20022975]
3. 't Hoen, Peter A. C.; Friedlander, Marc R.; Almlof, Jonas; Sammeth, Michael; Pulyakhina, Irina; Anvar, Seyed Y.; Laros, Jeroen F. J.; Buermans, Henk P. J.; Karlberg, Olof; Brannvall, Mathias; van Ommen, Gert-Jan B.; Estivill, Xavier; Guigo, Roderic; Syvanen, Ann-Christine; Gut, Ivo G.; Dermitzakis, Emmanouil T.; Antonarakis, Stylianos E.; Brazma, Alvis; Flicek, Paul; Schreiber, Stefan; Rosenstiel, Philip; Meitinger, Thomas; Strom, Tim M.; Lehrach, Hans; Sudbrak, Ralf; Carracedo, Angel; 't Hoen, Peter A. C.; Pulyakhina, Irina; Anvar, Seyed Y.; Laros, Jeroen F. J.; Buermans, Henk P. J.; van Iterson, Maarten; Friedlander, Marc R.; Monlong, Jean; Lizano, Esther; Bertier, Gabrielle; Ferreira, Pedro G.; Sammeth, Michael; Almlof, Jonas; Karlberg, Olof; Brannvall, Mathias; Ribeca, Paolo; Griebel, Thasso; Beltran, Sergi; Gut, Marta; Kahlem, Katja; Lappalainen, Tuuli; Giger, Thomas; Ongen, Halit; Padioleau, Ismael; Kilpinen, Helena; Gonzalez-Porta, Mar; Kurbatova, Natalja; Tikhonov, Andrew; Greger, Liliana; Barann, Matthias; Esser, Daniela; Hasler, Robert; Wieland, Thomas; Schwarzmayr, Thomas; Sultan, Marc; Amstislavskiy, Vyacheslav; den Dunnen, Johan T.; van Ommen, Gert-Jan B.; Gut, Ivo G.; Guigo, Roderic; Estivill, Xavier; Syvanen, Ann-Christine; Dermitzakis, Emmanouil T.; Lappalainen, Tuuli. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. Nat Biotechnol. 2013; 31(11):1015–1022. [PubMed: 24037425]
4. Su, Zhenqiang; Labaj, Pawel P.; Li, Sheng; Thierry-Mieg, Jean; Thierry-Mieg, Danielle; Shi, Wei; Wang, Charles; Schroth, Gary P.; Setterquist, Robert A.; Thompson, John F.; Jones, Wendell D.; Xiao, Wenzhong; Xu, Weihong; Jensen, Roderick V.; Kelly, Reagan; Xu, Joshua; Conesa, Ana; Furlanello, Cesare; Gao, Hanlin; Hong, Huixiao; Jafari, Nadereh; Letovsky, Stan; Liao, Yang; Lu, Fei; Oakeley, Edward J.; Peng, Zhiyu; Praul, Craig A.; Santoyo-Lopez, Javier; Scherer, Andreas; Shi, Tieliu; Smyth, Gordon K.; Staedtler, Frank; Sykacek, Peter; Tan, Xin-Xing; Aubrey Thompson, E.; Vandesompele, Jo; Wang, May D.; Wang, Jian; Wolfinger, Russell D.; Zavadil, Jiri; Auerbach, Scott S.; Bao, Wenjun; Binder, Hans; Blomquist, Thomas; Brilliant, Murray H.; Bushel, Pierre R.; Cai, Weimin; Catalano, Jennifer G.; Chang, Ching-Wei; Chen, Tao; Chen, Geng; Chen, Rong; Chierici, Marco; Chu, Tzu-Ming; Clevert, Djork-Arne; Deng, Youping; Derti, Adnan; Devanarayan, Viswanath; Dong, Zirui; Dopazo, Joaquin; Du, Tingting; Fang, Hong; Fang, Yongxiang; Fasold, Mario; Fernandez, Anita; Fischer, Matthias; Furio-Tari, Pedro; Fuscoe, James C.; Caimet, Florian; Gaj, Stan; Gandara, Jorge; Gao, Huan; Ge, Weigong; Gondo, Yoichi; Gong, Binsheng; Gong, Meihua; Gong, Zhuolin; Green, Bridgett; Guo, Chao; Guo, Lei; Guo, Li-Wu; Hadfield, James; Hellemans, Jan; Hochreiter, Sepp; Jia, Meiwen; Jian, Min; Johnson, Charles D.; Kay, Suzanne; Kleinjans, Jos; Lababidi, Samir; Levy, Shawn; Li, Quan-Zhen; Li, Li; Li, Li; Li, Peng; Li, Yan; Li, Haiqing; Li, Jianying; Li, Shiyong; Lin, Simon M.; Lopez, Francisco J.; Lu, Xin; Luo, Heng; Ma, Xiwen; Meehan, Joseph; Megherbi, Dalila B.; Mei, Nan; Mu, Bing; Ning, Baitang; Pandey, Akhilesh; Perez-Florido, Javier; Perkins, Roger G.; Peters, Ryan; Phan, John H.; Pirooznia, Mehdi; Qian, Feng; Qing, Tao; Rainbow, Lucille; Rocca-Serra, Philippe; Sambourg, Laure; Sansone, Susanna-Assunta; Schwartz, Scott; Shah, Ruchir; Shen, Jie; Smith, Todd M.; Stegle, Oliver; Stralis-Pavese, Nancy; Stupka, Elia; Suzuki, Yutaka; Szkotnicki, Lee T.; Tinning, Matthew; Tu, Bimeng; van Delft, Joost; Vela-Boza, Alicia; Venturini, Elisa; Walker, Stephen J.; Wan, Liqing; Wang, Wei; Wang, Jinhui; Wang, Jun; Wieben, Eric D.; Willey, James C.; Wu, Po-Yen; Xuan, Jiekun; Yang, Yong; Ye, Zhan; Yin, Ye; Yu, Ying; Yuan, Yate-Ching; Zhang, John; Zhang, Ke K.; Zhang, Wenqian; Zhang, Wenwei; Zhang, Yanyan; Zhao, Chen; Zheng, Yuanting; Zhou, Yiming; Zumbo, Paul; Tong, Weida; Kreil, David P.; Mason, Christopher E.; Shi, Leming. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the sequencing quality control consortium. Nat Biotechnol. 2014; 32(9):903–914. [PubMed: 25150838]
5. Li, Sheng; Labaj, Pawel P.; Zumbo, Paul; Sykacek, Peter; Shi, Wei; Shi, Leming; Phan, John; Wu, Po-Yen; Wang, May; Wang, Charles; Thierry-Mieg, Danielle; Thierry-Mieg, Jean; Kreil, David P.; Mason, Christopher E. Detecting and correcting systematic variation in large-scale RNA sequencing data. Nat Biotechnol. 2014; 32(9):888–895. [PubMed: 25150837]
6. Hansen, Kasper D.; Irizarry, Rafael A.; Wu, Zhijin. Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics. 2012; 13(2):204–216. [PubMed: 22285995]

7. Risso, Davide; Schwartz, Katja; Sherlock, Gavin; Dudoit, Sandrine. GC-content normalization for RNA-seq data. BMC Bioinformatics. 2011; 12(1):480. [PubMed: 22177264]
8. Zheng, Wei; Chung, Lisa M.; Zhao, Hongyu. Bias detection and correction in RNA-sequencing data. BMC Bioinformatics. 2011; 12(1):290. [PubMed: 21771300]
9. Stegle, Oliver; Parts, Leopold; Piipari, Matias; Winn, John; Durbin, Richard. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat Protoc. 2012; 7(3):500–507. [PubMed: 22343431]
10. Risso, Davide; Ngai, John; Speed, Terence P.; Dudoit, Sandrine. Normalization of RNA-seq data using factor analysis of control genes or samples. Nat Biotechnol. 2014; 32(9):896–902. [PubMed: 25150836]
11. Leek, Jeffrey T. svaseq: removing batch effects and other unwanted noise from sequencing data. Nucleic Acids Res. 2014; 42(21):000.
12. Li, Jun; Jiang, Hui; Wong, Wing. Modeling non-uniformity in short-read rates in RNA-seq data. Genome Biol. 2010; 11(5):R50. [PubMed: 20459815]
13. Hansen KD, Brenner SE, Dudoit S. Biases in illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res. 2010; 38(12):gkq224–e131.
14. Roberts, Adam; Trapnell, Cole; Donaghey, Julie; Rinn, John; Pachter, Lior. Improving RNA-seq expression estimates by correcting for fragment bias. Genome Biol. 2011; 12(3):R22–14. [PubMed: 21410973]
15. Nicolae, Marius; Mangul, Serghei; Mandoiu, Ion I.; Zelikovsky, Alex. Estimation of alternative splicing isoform frequencies from RNA-seq data. Algorithms Mol Biol. 2011; 6(1):9. [PubMed: 21504602]
16. Li, Wei; Jiang, Tao. Transcriptome assembly and isoform expression level estimation from biased RNA-seq reads. Bioinformatics. 2012; 28(22):2914–2921. [PubMed: 23060617]
17. Lahens, Nicholas F.; Kavakli, Ibrahim H.; Zhang, Ray; Hayer, Katharina; Black, Michael B.; Dueck, Hannah; Pizarro, Angel; Kim, Junhyong; Irizarry, Rafael; Thomas, Russell S.; Grant, Gregory R.; Hogenesch, John B. IVT-seq reveals extreme bias in RNA sequencing. Genome Biol. 2014; 15(6):R86. [PubMed: 24981968]
18. Hayer, Katharina; Pizzaro, Angel; Lahens, Nicholas L.; Hogenesch, John B.; Grant, Gregory R. Benchmark analysis of algorithms for determining and quantifying full-length mRNA splice forms from RNA-seq data. Bioinformatics. 2015; 31(24):3938–3945. [PubMed: 26338770]
19. Lappalainen, Tuuli; Sammeth, Michael; Friedlander, Marc R.; /`t Hoen, Peter A. C.; Monlong, Jean; Rivas, Manuel A.; Gonzalez-Porta, Mar; Kurbatova, Natalja; Griebel, Thasso; Ferreira, Pedro G.; Barann, Matthias; Wieland, Thomas; Greger, Liliana; van Iterson, Maarten; Almlof, Jonas; Ribeca, Paolo; Pulyakhina, Irina; Esser, Daniela; Giger, Thomas; Tikhonov, Andrew; Sultan, Marc; Bertier, Gabrielle; MacArthur, Daniel G.; Lek, Monkol; Lizano, Esther; Buermans, Henk P. J.; Padioleau, Ismael; Schwarzmayr, Thomas; Karlberg, Olof; Ongen, Halit; Kilpinen, Helena; Beltran, Sergi; Gut, Marta; Kahlem, Katja; Amstislavskiy, Vyacheslav; Stegle, Oliver; Pirinen, Matti; Montgomery, Stephen B.; Donnelly, Peter; McCarthy, Mark I.; Flicek, Paul; Strom, Tim M.; The Geuvadis Consortium. Lehrach, Hans; Schreiber, Stefan; Sudbrak, Ralf; Carracedo, Angel; Antonarakis, Stylianos E.; Hasler, Robert; Syvanen, Ann-Christine; van Ommen, Gert-Jan; Brazma, Alvis; Meitinger, Thomas; Rosenstiel, Philip; Guigo, Roderic; Gut, Ivo G.; Estivill, Xavier; Dermitzakis, Emmanouil T. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013; 501(7468):506–511. [PubMed: 24037378]
20. Aird, Daniel; Ross, Michael G.; Chen, Wei-Sheng; Danielsson, Maxwell; Fennell, Timothy; Russ, Carsten; Jaffe, David B.; Nusbaum, Chad; Gnirke, Andreas. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. 2011; 12(2):R18. [PubMed: 21338519]
21. Benjamini, Yuval; Speed, Terence P. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 2012; 40(10):e72. [PubMed: 22323520]
22. Jessica, Jingyi; Li, J.; Jiang, Ci-Ren R.; Brown, James B.; Huang, Haiyan; Bickel, Peter J. Sparse linear modeling of next-generation mRNA sequencing (RNA-seq) data for isoform discovery and abundance estimation. Proc Natl Acad Sci U S A. 2011; 108(50):19867–19872. [PubMed: 22135461]

23. Hron, Tomas; Pajer, Petr; Paces, Jan; Bartunek, Petr; Elleder, Daniel. Hidden genes in birds. Genome Biol. 2015; 16(1):164. [PubMed: 26283656]
24. Patro, Rob; Mount, Stephen M.; Kingsford, Carl. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat Biotechnol. 2014; 32(5): 462–464. [PubMed: 24752080]
25. Bray, Nicolas; Pimentel, Harold; Melsted, Pall; Pachter, Lior. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016; 34(5):525–527. [PubMed: 27043002]
26. Patro, Rob; Duggal, Geet; Kingsford, Carl. Accurate, fast, and model-aware transcript expression quantification with Salmon. bioRxiv. 2015
27. Katz, Yarden; Wang, Eric T.; Airoldi, Edoardo M.; Burge, Christopher B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods. 2010; 7(12):1009–1015. [PubMed: 21057496]
28. Frazee, Alyssa C.; Jaffe, Andrew E.; Langmead, Ben; Leek, Jeffrey T. Polyester: simulating RNA-seq datasets with differential transcript expression. Bioinformatics. 2015; 31(17):2778–2784. [PubMed: 25926345]
29. Li, Sheng; Tighe, Scott W.; Nicolet, Charles M.; Grove, Deborah; Levy, Shawn; Farmerie, William; Viale, Agnes; Wright, Chris; Schweitzer, Peter A.; Gao, Yuan; Kim, Dewey; Boland, Joe; Hicks, Belynda; Kim, Ryan; Chhangawala, Sagar; Jafari, Nadereh; Raghavachari, Nalini; Gandara, Jorge; Garcia-Reyero, Natalia; Hendrickson, Cynthia; Roberson, David; Rosenfeld, Jeffrey; Smith, Todd; Underwood, Jason G.; Wang, May; Zumbo, Paul; Baldwin, Don A.; Grills, George S.; Mason, Christopher E. Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. Nat Biotechnol. 2014; 32(9):915–925. [PubMed: 25150835]
30. Katz, Yarden; Wang, Eric T.; Silterra, Jacob; Schwartz, Schraga; Wong, Bang; Thorvaldsdottir, Helga; Robinson, James T.; Mesirov, Jill P.; Airoldi, Edoardo M.; Burge, Christopher B. Quantitative visualization of alternative exon expression from RNA-seq data. Bioinformatics. 2015; 31(14):2400–2402. [PubMed: 25617416]
31. Dobin, Alexander; Davis, Carrie A.; Schlesinger, Felix; Drenkow, Jorg; Zaleski, Chris; Jha, Sonali; Batut, Philippe; Chaisson, Mark; Gingeras, Thomas R. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013; 29(1):15–21. [PubMed: 23104886]
32. Kim, Daehwan; Pertea, Geo; Trapnell, Cole; Pimentel, Harold; Kelley, Ryan; Salzberg, Steven L. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013; 14(4):R36. [PubMed: 23618408]
33. Anders, Simon; Huber, Wolfgang. Differential expression analysis for sequence count data. Genome Biol. 2010; 11(10):R106. [PubMed: 20979621]
34. Huber, Wolfgang; Carey, Vincent J.; Gentleman, Robert; Anders, Simon; Carlson, Marc; Carvalho, Benilton S.; Bravo, Hector C.; Davis, Sean; Gatto, Laurent; Girke, Thomas; Gottardo, Raphael; Hahne, Florian; Hansen, Kasper D.; Irizarry, Rafael A.; Lawrence, Michael; Love, Michael I.; MacDonald, James; Obenchain, Valerie; Oles, Andrzej K.; Pages, Herve; Reyes, Alejandro; Shannon, Paul; Smyth, Gordon K.; Tenenbaum, Dan; Waldron, Levi; Morgan, Martin. Orchestrating high-throughput genomic analysis with bioconductor. Nat Methods. 2015; 12(2): 115–121. [PubMed: 25633503]
35. Lawrence, Michael; Huber, Wolfgang; Pages, Herve; Aboyoun, Patrick; Carlson, Marc; Gentleman, Robert; Morgan, Martin T.; Carey, Vincent J. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013; 9(8):e1003118. [PubMed: 23950696]

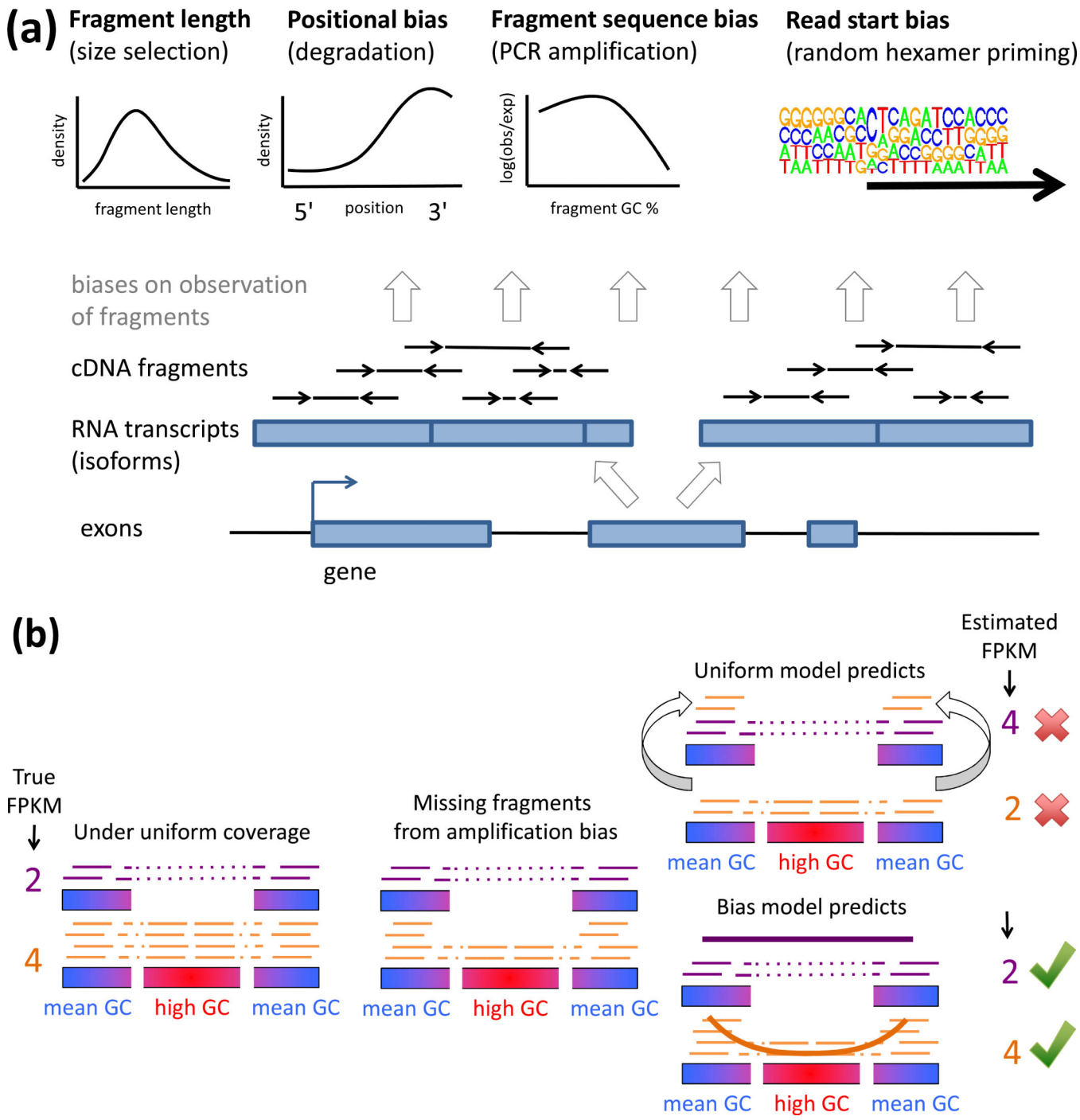
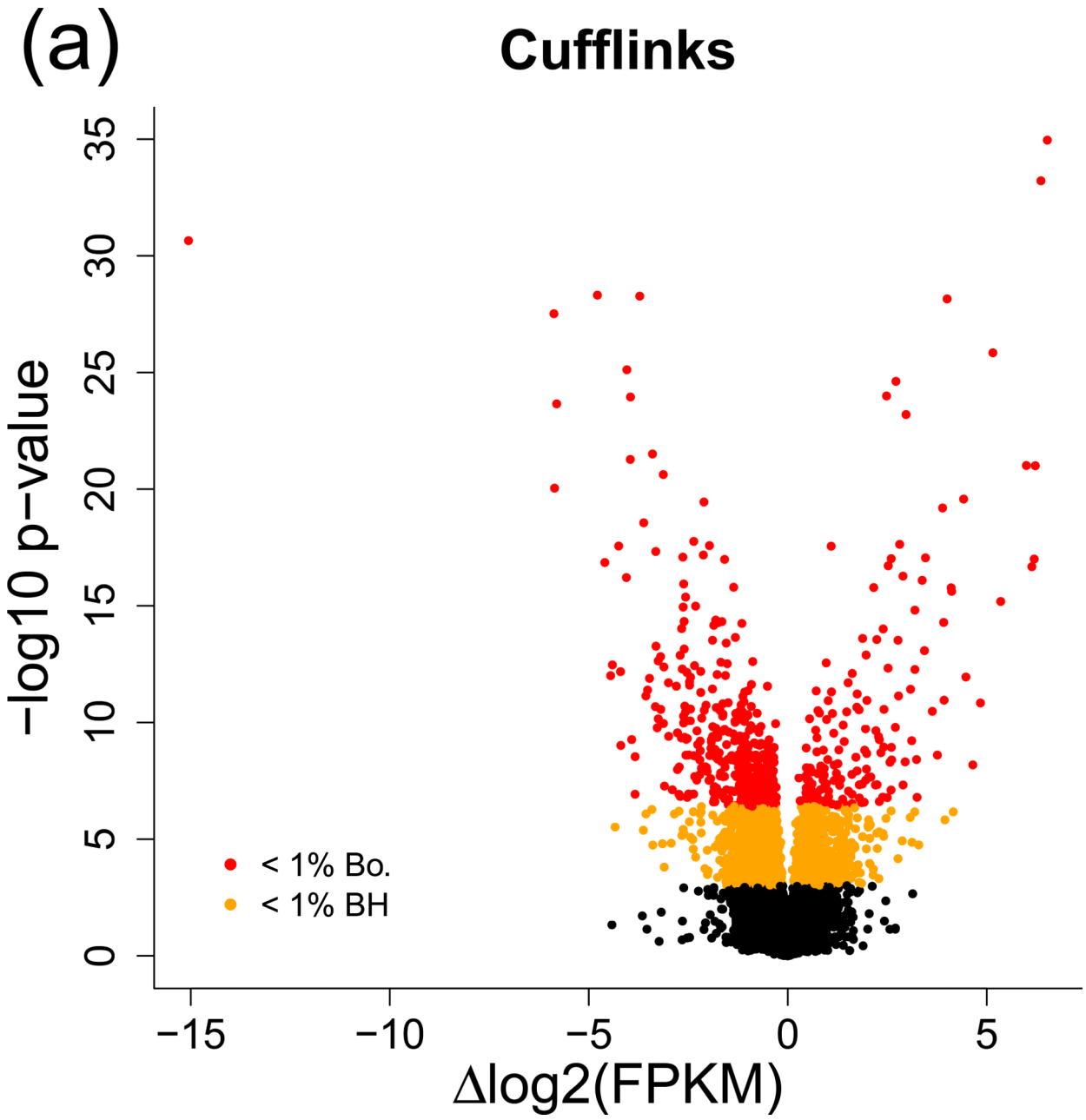
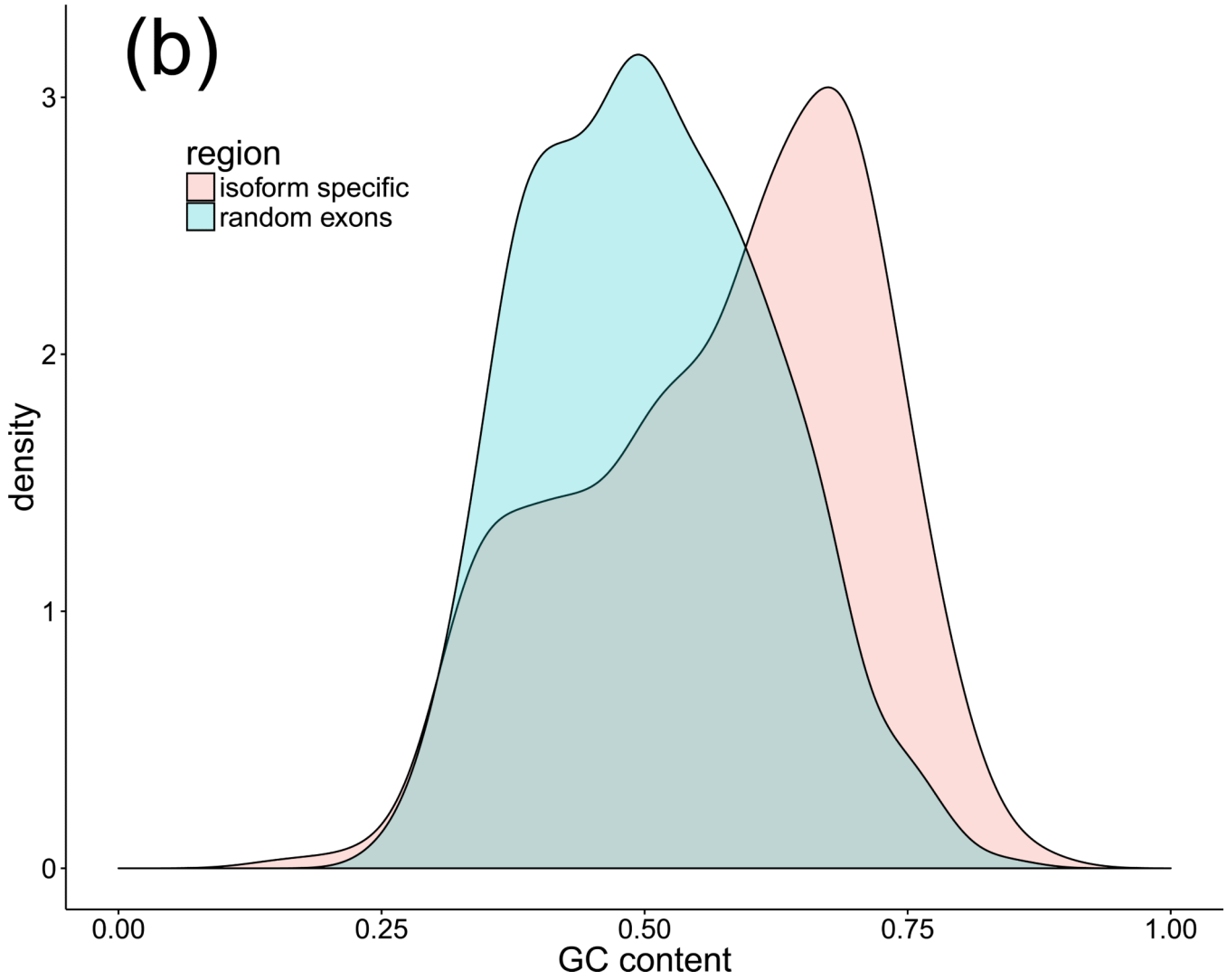
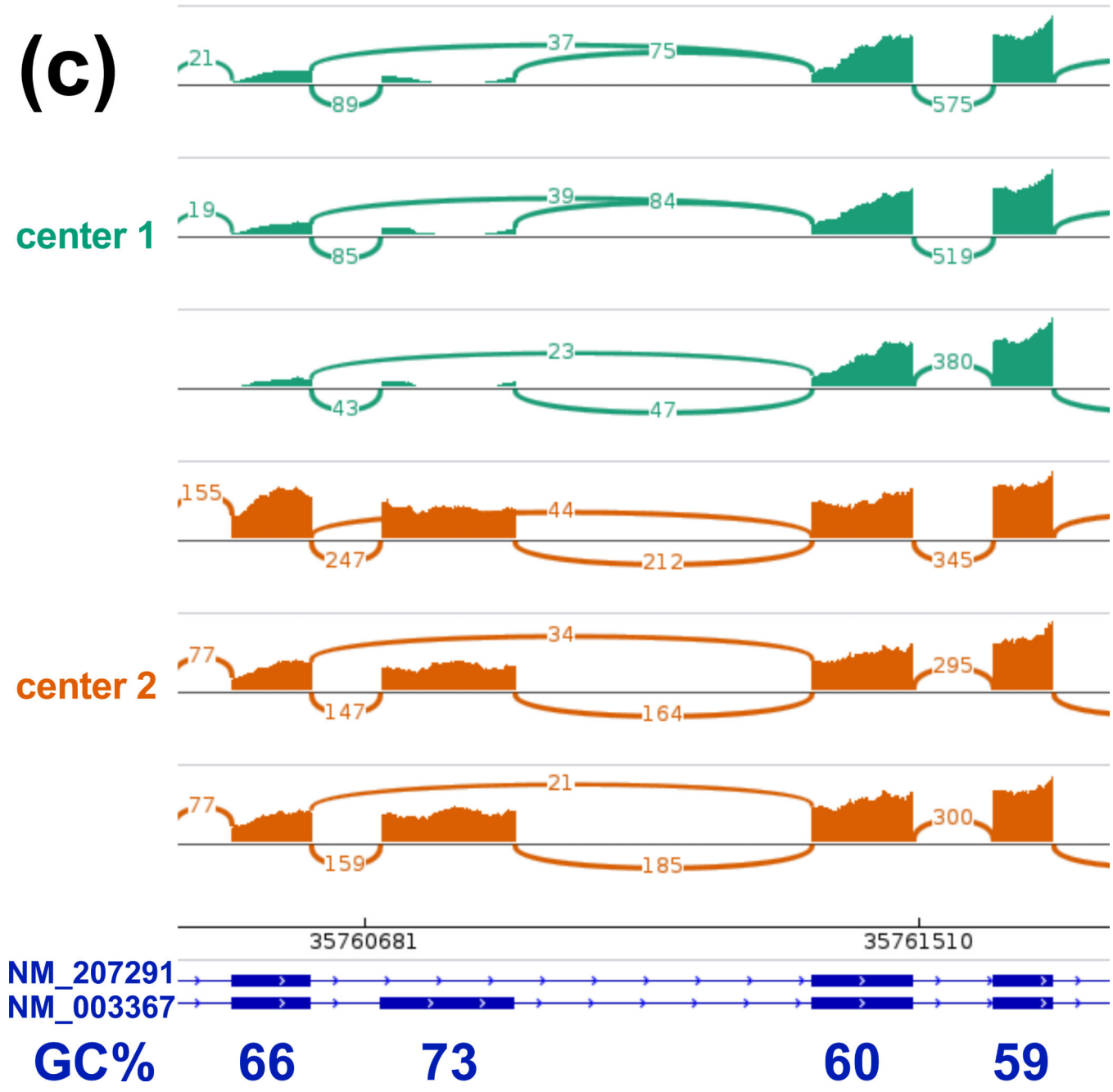


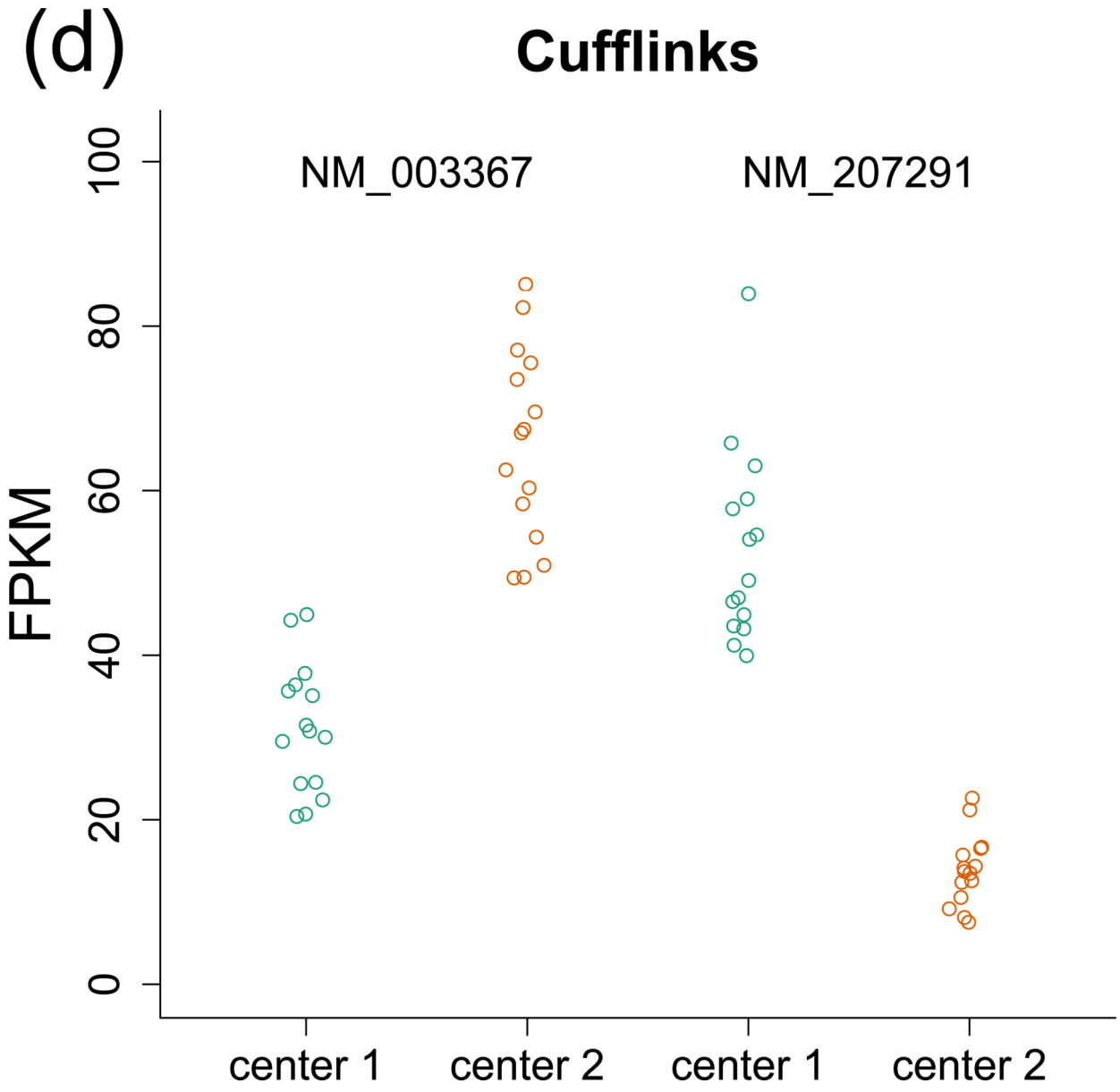
Figure 1.

Quantification of transcript abundance from RNA-seq experiments. (a) RNA-seq biases. (b) Ignoring fragment sequence bias impairs transcript abundance estimation when isoform-specific regions have GC content or sequence features that lead to under-representation of fragments, resulting in false positives of predicted expression of isoforms that are lowly or not expressed.









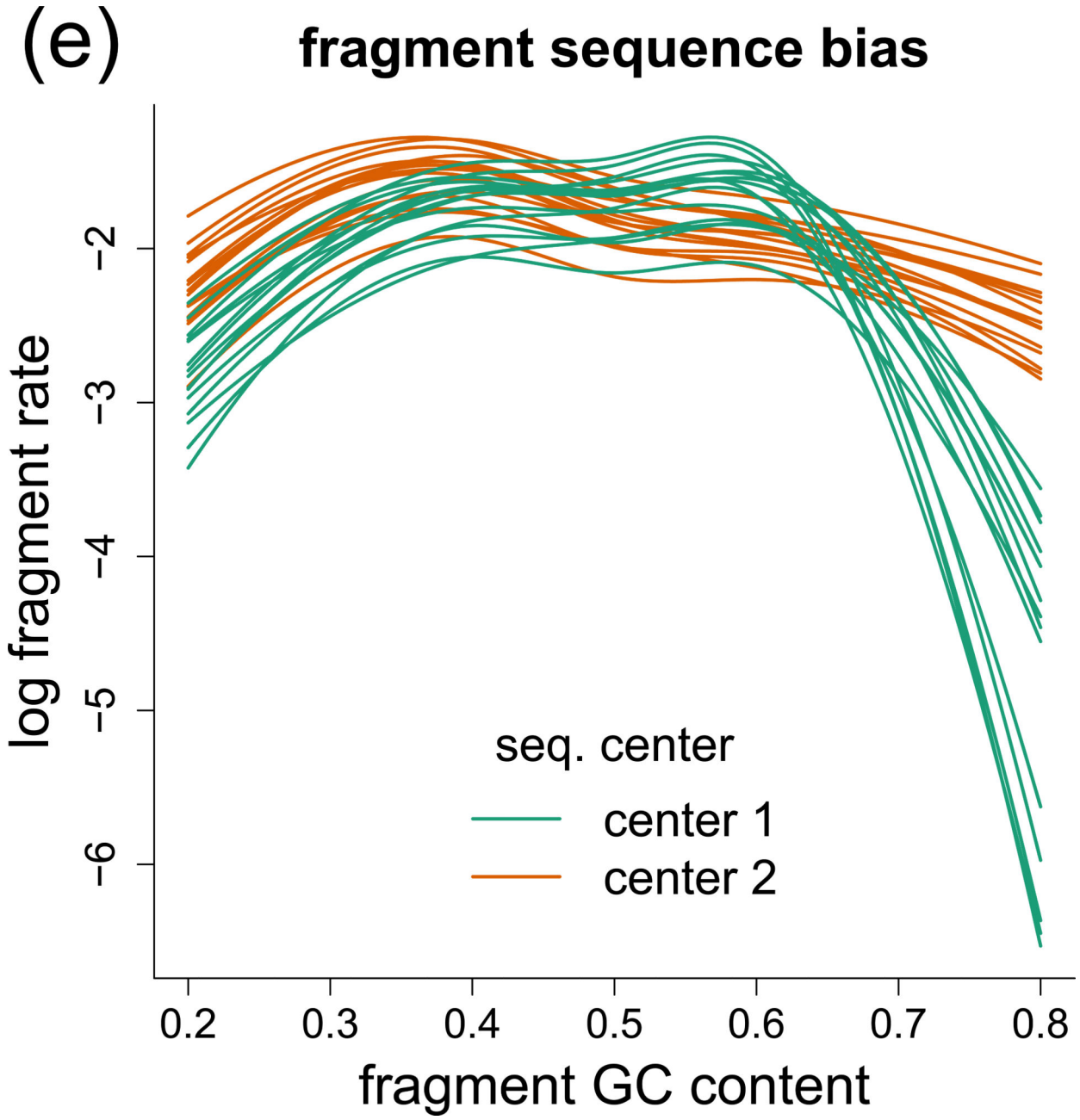
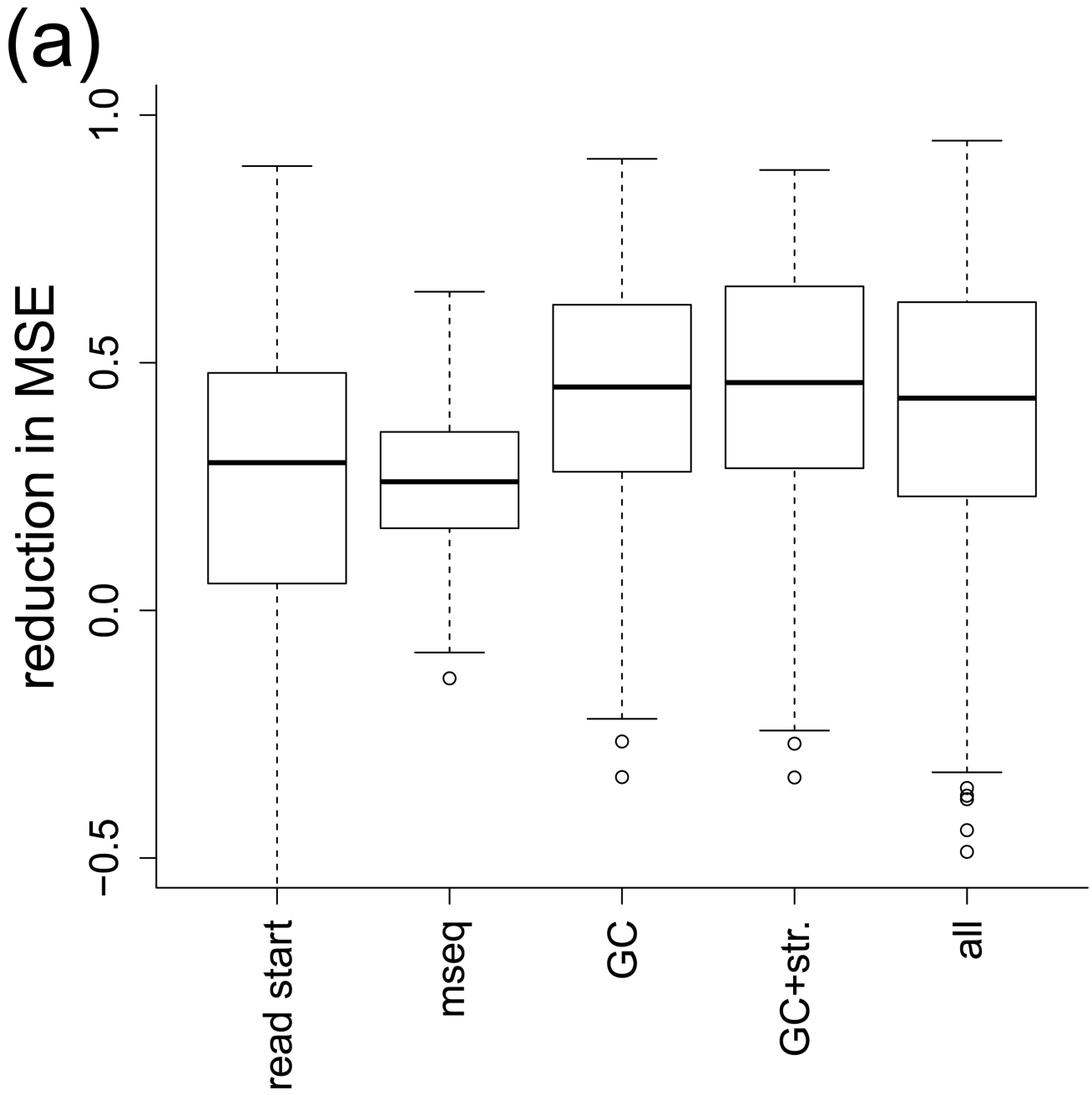
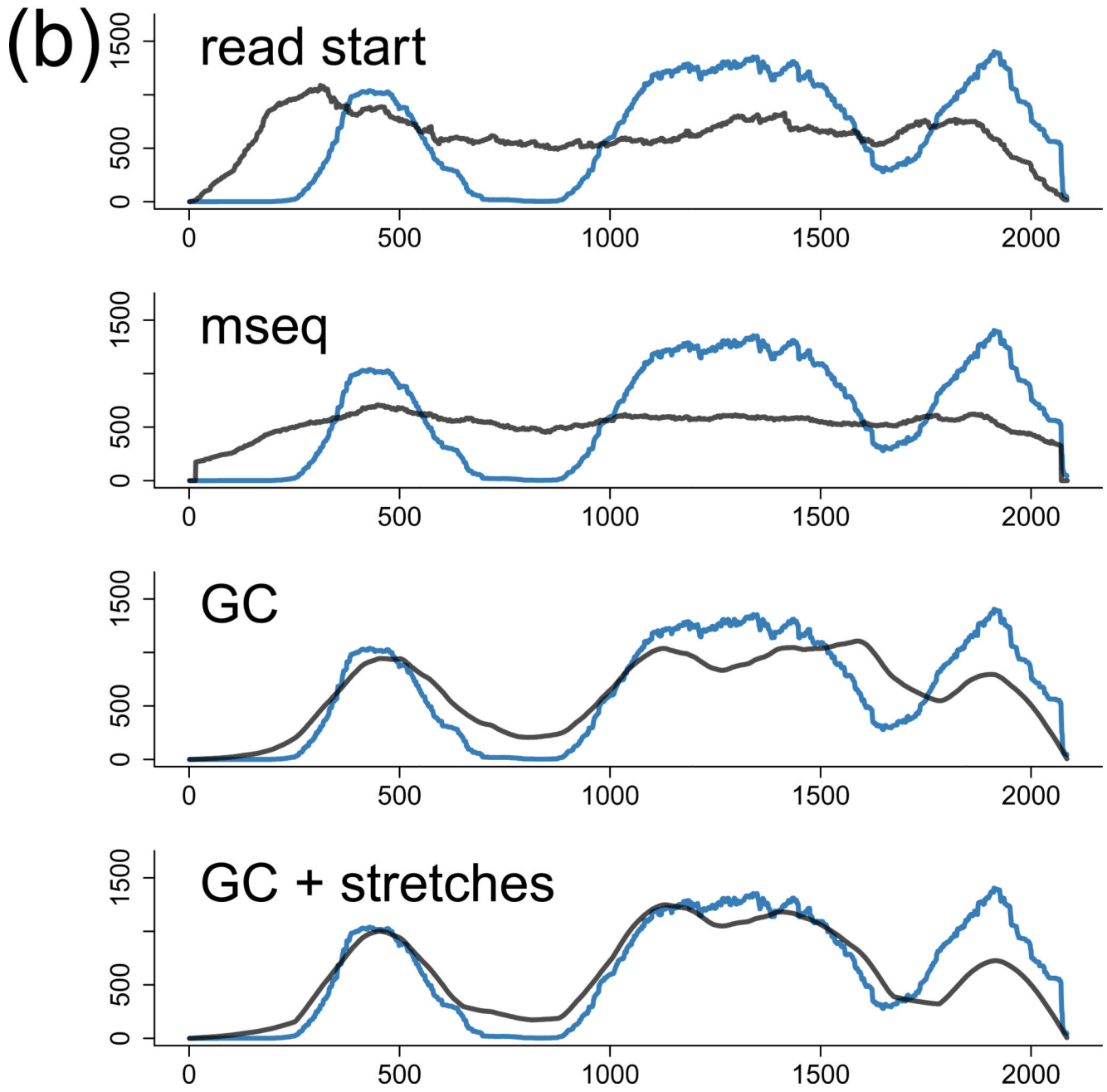


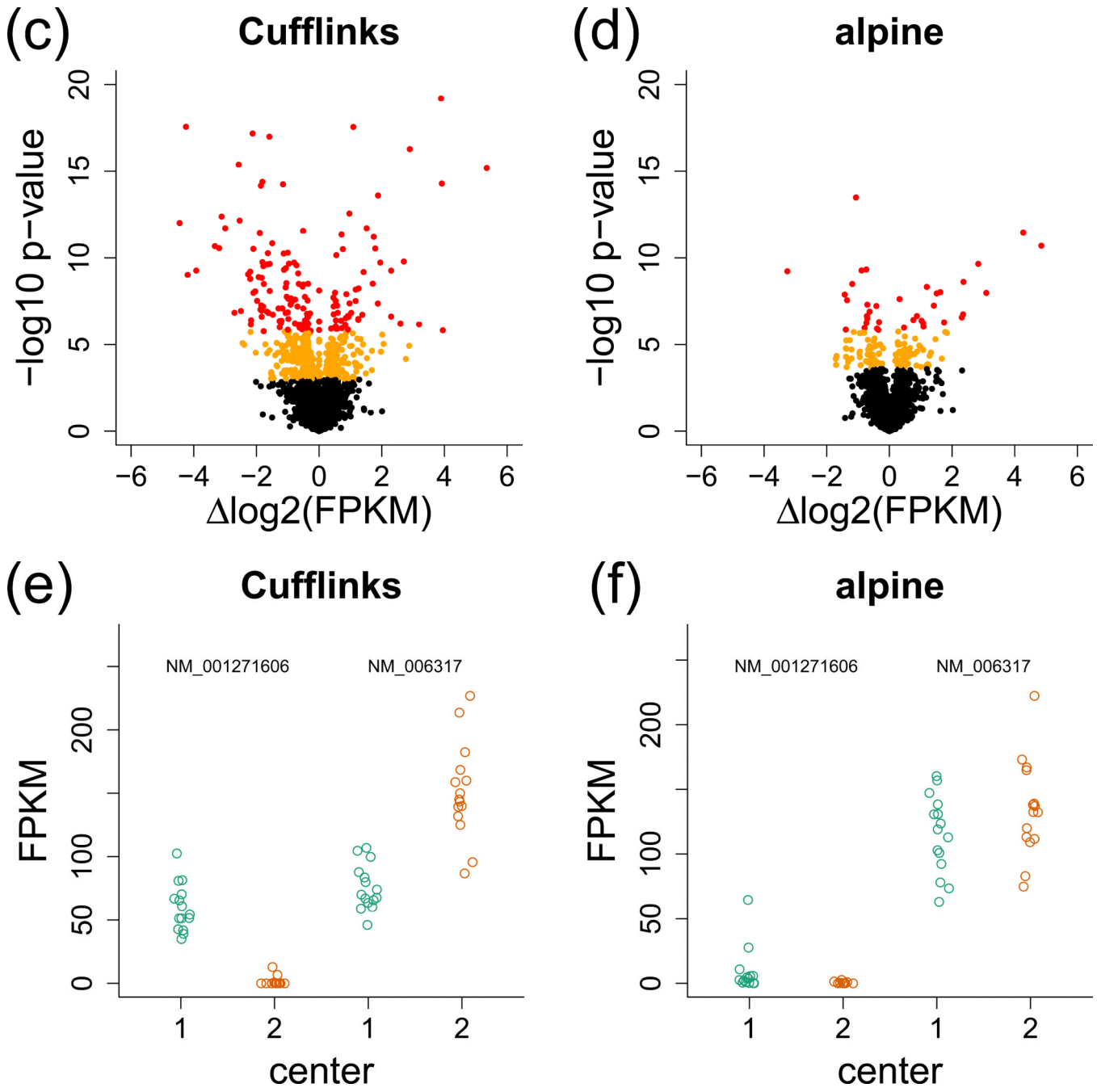
Figure 2.

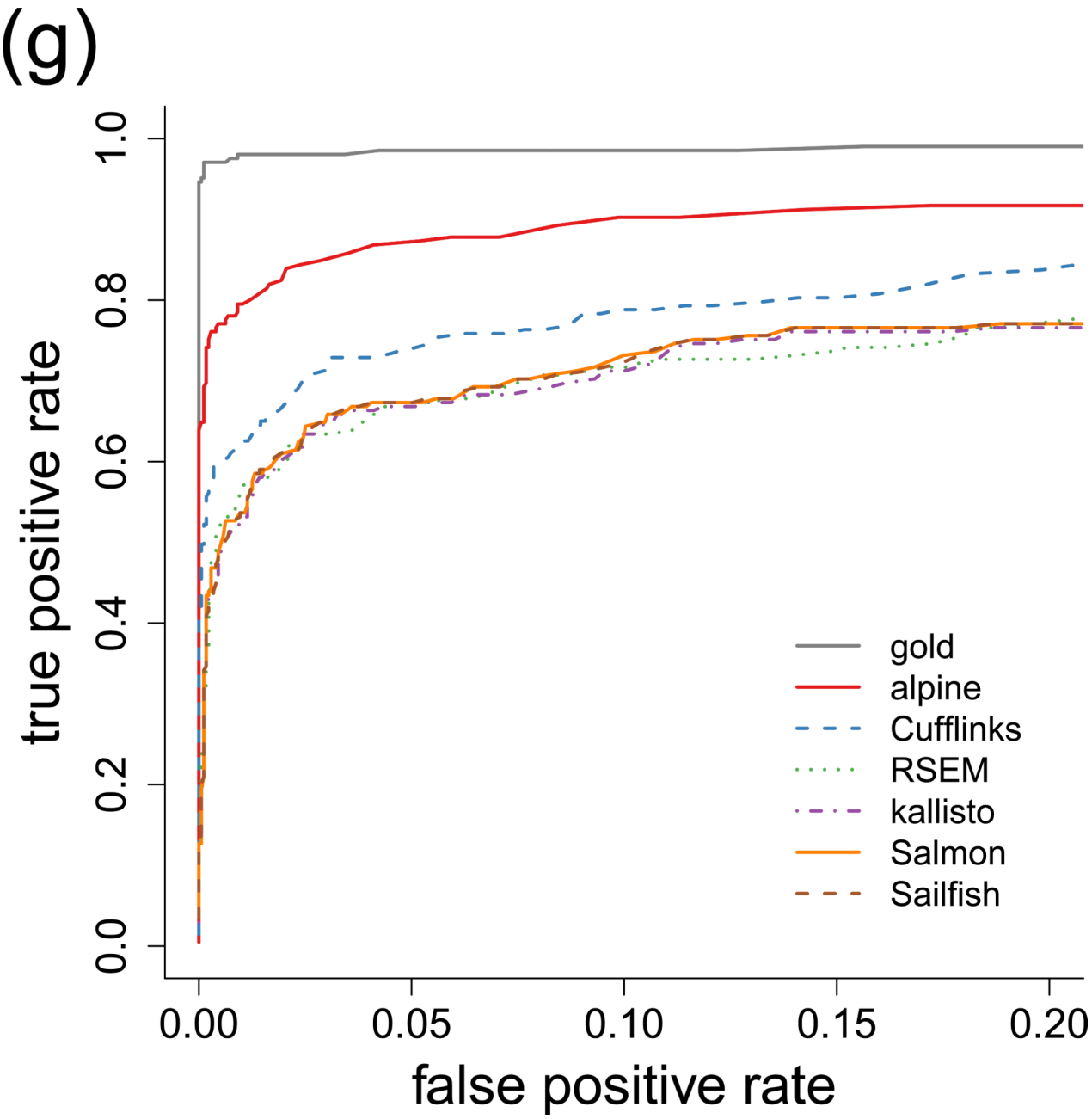
Problems with current transcript abundance estimation methods. (a) Volcano plot of a comparison of Cufflinks transcript estimates from center 1 against center 2, with 2,510 transcripts reported differentially expressed at a target FDR of 1% and 515 with family-wise error rate (FWER) of 1% using a more conservative Bonferroni correction. (b) Densities of GC content of isoform-specific regions from genes with two isoforms when one or more reported differential expression, compared to GC content of random exons. (c) Sashimi plot [30] for GEUVADIS samples in a region of the USF2 gene containing the alternative exon distinguishing two isoforms, with the GC content of each exon listed below the gene model. Samples from center 1 had a drop-out in coverage on the high GC exons, including the

alternative exon. The curved lines in the Sashimi plot represent RNA-seq reads spanning exon-exon junctions with numbers indicating number of supporting reads. (d) Cufflinks FPKM estimates for the two isoforms of USF2 where technical artifacts in coverage lead to discordant estimates of abundance across center. (e) Curves fit by alpine estimating dependence of fragment rate on GC content after controlling for random hexamer priming bias of read starts (see main text and Supplementary Note).









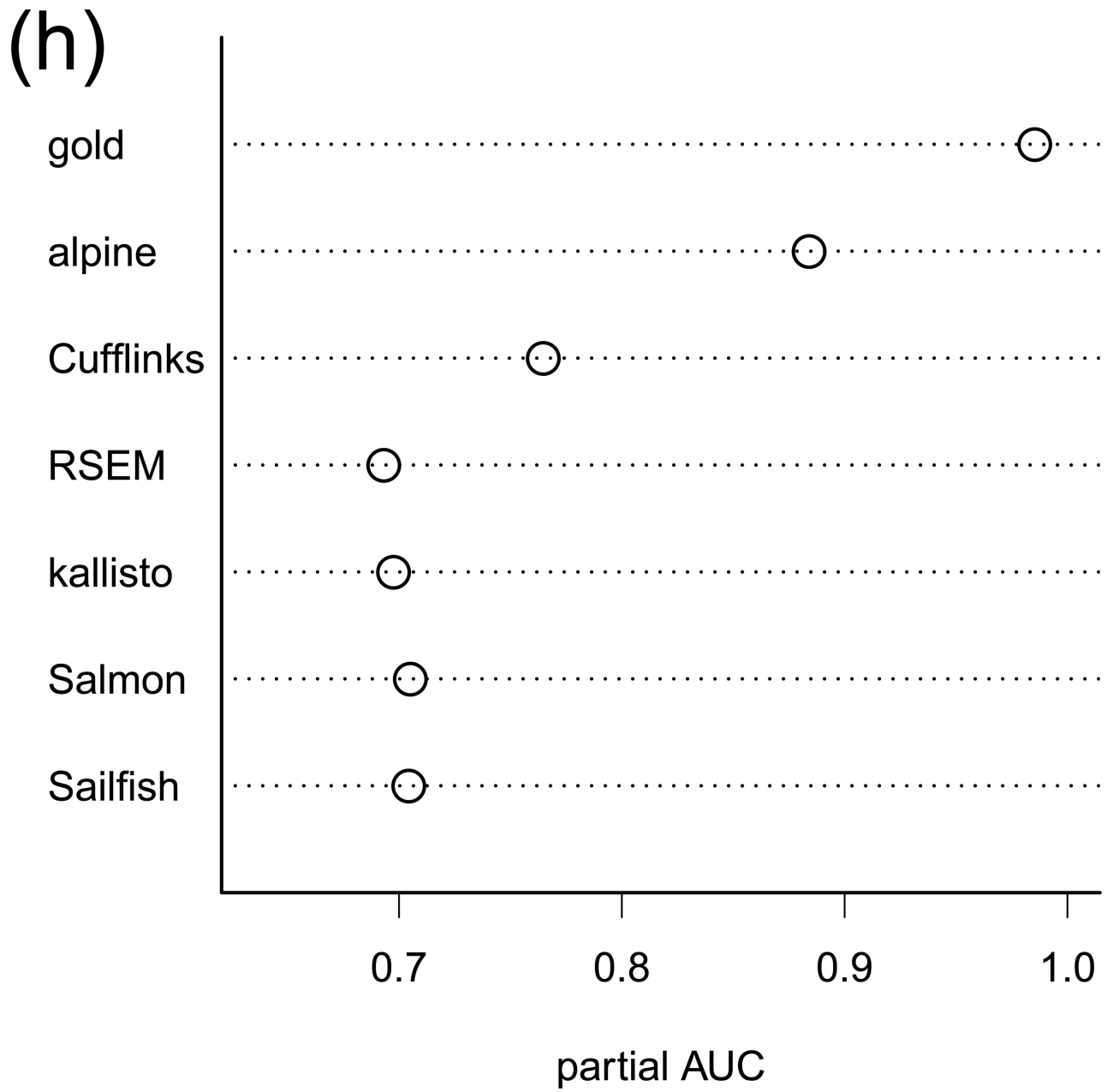


Figure 3.

Modeling and correcting fragment sequence bias. (a) Boxplot comparing reduction in mean squared error (MSE) in predicting coverage for different bias models for all 8 samples and 64 transcripts (n=512). The models are “read start”: the Cufflinks VLMM for read starts, the *mseq* model for read starts, “GC”: a model using fragment GC content, “GC+str”: as in “GC” plus additional terms for stretches of high GC, “all”: the VLMM for read starts in addition to the terms in “GC+str”. The models including fragment GC doubled the reduction in MSE as the read start models. (b) Predicted coverage plots for bias models on GenBank BC011380 (raw coverage in blue, test-set predicted coverage in black). (c) and (d) Volcano plots of differential transcript expression across centers for genes with two isoforms (orange:

Benjamini-Hochberg adjusted p values less than 1%; red: Bonferroni FWER rate less than 1%). (e) and (f) FPKM estimates for two isoforms of BASP1 across center. (g) ROC curves for a simulation of a confounded design, with fragment sequence bias drawn from 30 GEUVADIS samples. The “gold” ROC curve indicates the sensitivity and specificity using the true underlying fragment counts, without fragment sequence bias and with known transcript assignment for fragments (h) The partial area under the curve (AUC) for panel (g), considering false positive rate in [0, 0.2], scaled to take values between 0 and 1.