BMC Genomics

RESEARCH ARTICLE

Open Access

CrossMark

Limitations of alignment-free tools in total RNA-seq quantification

Douglas C. Wu[1,2] , Jun Yao[1,2], Kevin S. Ho[1,2], Alan M. Lambowitz[1,2] and Claus O. Wilke[1,3*]

Abstract

Background: Alignment-free RNA quantification tools have significantly increased the speed of RNA-seq analysis. However, it is unclear whether these state-of-the-art RNA-seq analysis pipelines can quantify small RNAs as accurately as they do with long RNAs in the context of total RNA quantification.

Result: We comprehensively tested and compared four RNA-seq pipelines for accuracy of gene quantification and fold-change estimation. We used a novel total RNA benchmarking dataset in which small non-coding RNAs are highly represented along with other long RNAs. The four RNA-seq pipelines consisted of two commonly-used alignment-free pipelines and two variants of alignment-based pipelines. We found that all pipelines showed high accuracy for quantifying the expression of long and highly-abundant genes. However, alignment-free pipelines showed systematically poorer performance in quantifying lowly-abundant and small RNAs.

Conclusion: We have shown that alignment-free and traditional alignment-based quantification methods perform similarly for common gene targets, such as protein-coding genes. However, we have identified a potential pitfall in analyzing and quantifying lowly-expressed genes and small RNAs with alignment-free pipelines, especially when these small RNAs contain biological variations.

Keywords: TGIRT-seq, RNA-seq, k-mer

Background

RNA-seq continues to pose great computational and statistical challenges. These challenges range from accurately aligning sequencing reads to accurate inference of gene expression levels [1, 2]. The central computational problem in RNA-seq remains the efficient and accurate assignment of short sequencing reads to the transcripts they originated from and using this information to infer gene expression [3–6]. Conventionally, read assignment is carried out by aligning sequencing reads to a reference genome, such that relative gene expression levels can be inferred by the alignments at annotated gene loci [2, 7]. These alignment-based methods are conceptually simple, but the read-alignment step can be time-consuming and computationally intensive despite recent advancements in fast read aligners [4, 8, 9]. Recently,

several novel tools introduced alignment-free transcript quantification utilizing k-mer-based counting algorithms [4–6]. These alignment-free pipelines are orders of magnitude faster than alignment-based pipelines, and they work by breaking sequencing reads into k-mers and then performing fast matches to pre-indexed transcript databases [4]. To achieve fast transcript quantification without compromising quantification accuracy, different sophisticated algorithms have been implemented in addition to k-mer counting, such as pseudoalignments (Kallisto [5]) or quasi-mapping along with GC- and sequence-bias corrections (Salmon [6]). Given the wide variety of choices in RNA-seq tools, several studies have benchmarked subsets of read aligners and quantification software. These studies generally suggest that most of the current RNA-seq tools display comparable accuracy [10–12].

However, the existing benchmarking studies were generally carried out on either simulated RNA-seq datasets [12] or RNA-seq datasets that focused only on long RNAs, such as messenger RNAs (mRNAs) and long noncoding RNAs (lncRNAs) [10, 11, 13, 14]. Consequently,

*Correspondence: wilke@austin.utexas.edu
[1]Institute for Cellular and Molecular Biology, The University of Texas at Austin, Austin, TX, 78712, USA
[3]Department of Integrative Biology, The University of Texas at Austin, Austin, TX, 78712, USA
Full list of author information is available at the end of the article

Wu *et al. BMC Genomics* (2018) 19:510

Page 2 of 14

they did not evaluate whether these tools are suitable for total RNA quantification in datasets that include small RNAs, such as transfer RNAs (tRNAs) and small nucleolar RNAs (snoRNAs). To some extent, the lack of a comprehensive comparison between small and long RNA quantification may be due to the inability of most current RNA-seq methods to efficiently recover these small RNAs [15]. Recently, however, a novel method has overcome this problem by using a thermostable group II intron reverse transcriptase (TGIRT) during RNA-seq library construction [15]. This method enables more comprehensive profiling of full-length structured small non-coding RNAs (sncRNA) along with long RNAs in a single RNA-seq library workflow [15–17]. Thus, it is now possible to benchmark RNA-seq quantification tools on structured small non-coding RNAs.

To address whether current RNA-seq tools can quantify small RNAs as accurately as they do with long RNAs, we tested four gene quantification pipelines on a previously sequenced TGIRT RNA-seq (TGIRT-seq) dataset [15] obtained from the well-studied microarray/sequencing quality control consortium (MAQC) sample set [18, 19]. Of the four tested pipelines, two are alignment-based and two are alignment-free. We found that all four pipelines are mostly concordant in quantifying common differentially-expressed gene targets, such as mRNAs and mRNA-like spike-ins. However, with respect to quantifying small or lowly-expressed genes, we found that the alignment-based pipelines significantly outperformed the alignment-free pipelines.
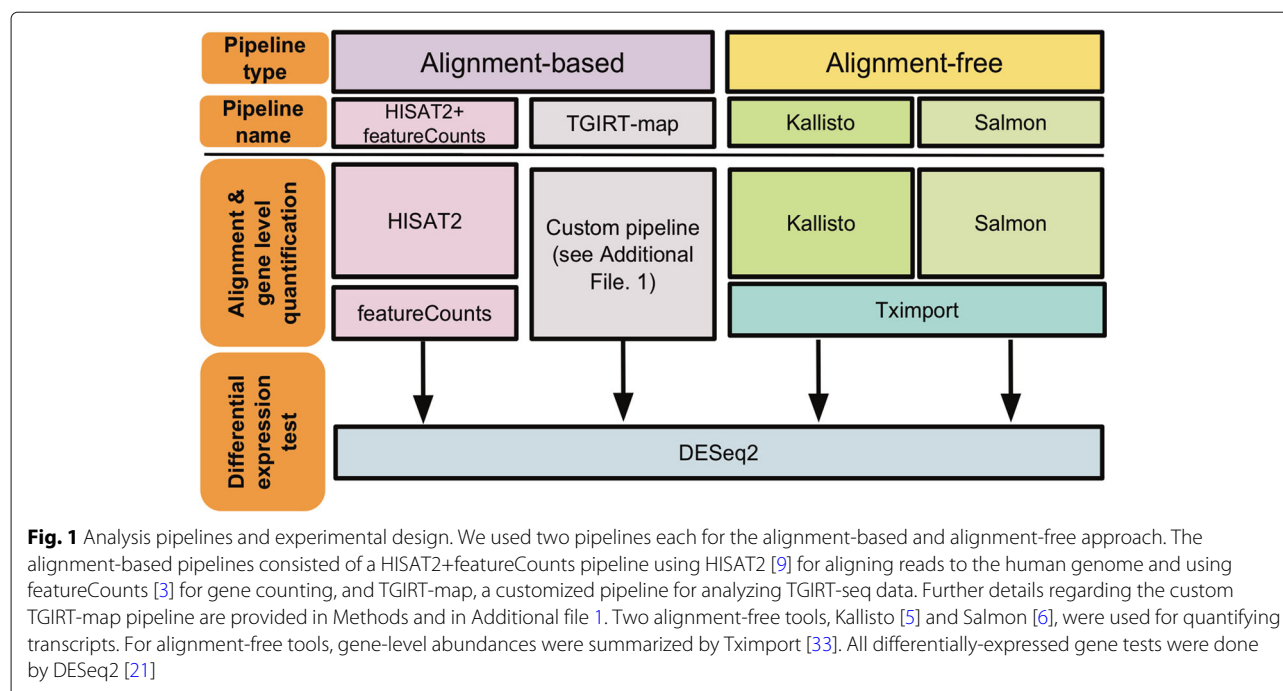
Results
Study design. We tested four RNA-seq quantification pipelines, including two alignment-free and two alignment-based pipelines (Fig. 1): (A) Kallisto, a k-mer counting software that uses pseudoalignments for reducing quantification error and improving speed [5]; (B) Salmon, another k-mer counting software that learns and corrects sequence-specific and GC biases on-the-fly, in addition to using quasi-mapping for further improvement in transcript quantification [6]; (C) HISAT2+featureCounts, a conventional alignment-based pipeline aligning sequencing reads to human genome by a splice-aware aligner, HISAT2 [9], and quantifying genes by featureCounts [3]; and (D) TGIRT-map, a customized alignment-based pipeline using an iterative genome-mapping procedure (Additional file 1).

The benchmarking dataset we used here consists of TGIRT-seq libraries for four well-defined samples (samples A–D) from the microarray/sequencing quality control consortium (MAQC [18, 19]), each obtained in triplicate [15]. The MAQC samples A and B represent universal human reference total RNA and human brain reference total RNA, respectively, that are mixed

with corresponding External RNA Controls Consortium (ERCC) spike-in transcripts. Samples C and D are mixtures of samples A and B at different ratios [18, 19]. The known mix ratios in samples C and D allow for the calculation of expected fold-changes between samples C and D from the measured fold-changes between samples A and B for each gene [18, 19]. Mapping statistics for all pipelines are summarized in Additional file 2: Supplementary Tables S1–4.

Gene detection and quantification. To test if all four pipelines produced concordant gene quantifications, we first compared the numbers of detected genes across all methods. We considered a gene as detected if it was assigned with a transcripts per million (TPM) value > 0.1. We found that the numbers of detected genes were similar among all tested pipelines (Fig. 2a). Moreover, by comparing the identities of the detected genes, we found that the vast majority of the genes were detected by all tested pipelines (Additional file 3). However, we also found that Salmon and TGIRT-map consistently detected more unique genes compared to the other two pipelines (Additional file 3; Friedman test, $p = 4 \times 10^{-7}$; Paired Wilcoxon test, $p < 0.05$ for all pairwise pipeline comparisons).

Even though Salmon and TGIRT-map both detected more unique genes than did the other two pipelines, the additional genes detected were different. Salmon primarily recovered more long RNAs (labeled as antisense, other ncRNAs, and protein-coding genes; Additional file 3). This enrichment in long RNAs could be pipeline-type-specific (when compared to alignment-based pipelines) or algorithm-specific (when compared to Kallisto). Pipeline-type-specific differences could be due to the probabilistic gene quantification methods of Salmon [6]. While Salmon can assign a single fragment to multiple genes, each fragment can only be assigned to a single gene by the alignment-based pipelines under our parameters. In terms of algorithm, the result might be due to how Salmon corrects for GC and sequencing biases or how it handles equivalent classes (i.e. multiply-mapped reads) relative to Kallisto [6]. On the other hand, TGIRT-map recovered more miRNAs (likely to be mis-counting of fragmented exons or unannotated exons in these libraries), some long non-coding RNAs (annotated as other ncRNAs), and small non-coding RNAs (annotated as other sncRNA) (Additional file 3). These enrichments under TGIRT-map could be pipeline-type-specific when compared to alignment-free pipelines, which may be affected by the choice of k-mer size. The differences between TGIRT-map and HISAT2+featureCounts were possibly the result of an additional local-alignment step (BOWTIE2) [20] after the spliced-read mapping step (HISAT2) [9] in TGIRT-map (Additional file 1).

Wu *et al. BMC Genomics*   (2018) 19:510

Page 3 of 14

Fig. 1 Analysis pipelines and experimental design. We used two pipelines each for the alignment-based and alignment-free approach. The alignment-based pipelines consisted of a HISAT2+featureCounts pipeline using HISAT2 [9] for aligning reads to the human genome and using featureCounts [3] for gene counting, and TGIRT-map, a customized pipeline for analyzing TGIRT-seq data. Further details regarding the custom TGIRT-map pipeline are provided in Methods and in Additional file 1. Two alignment-free tools, Kallisto [5] and Salmon [6], were used for quantifying transcripts. For alignment-free tools, gene-level abundances were summarized by Tximport [33]. All differentially-expressed gene tests were done by DESeq2 [21]
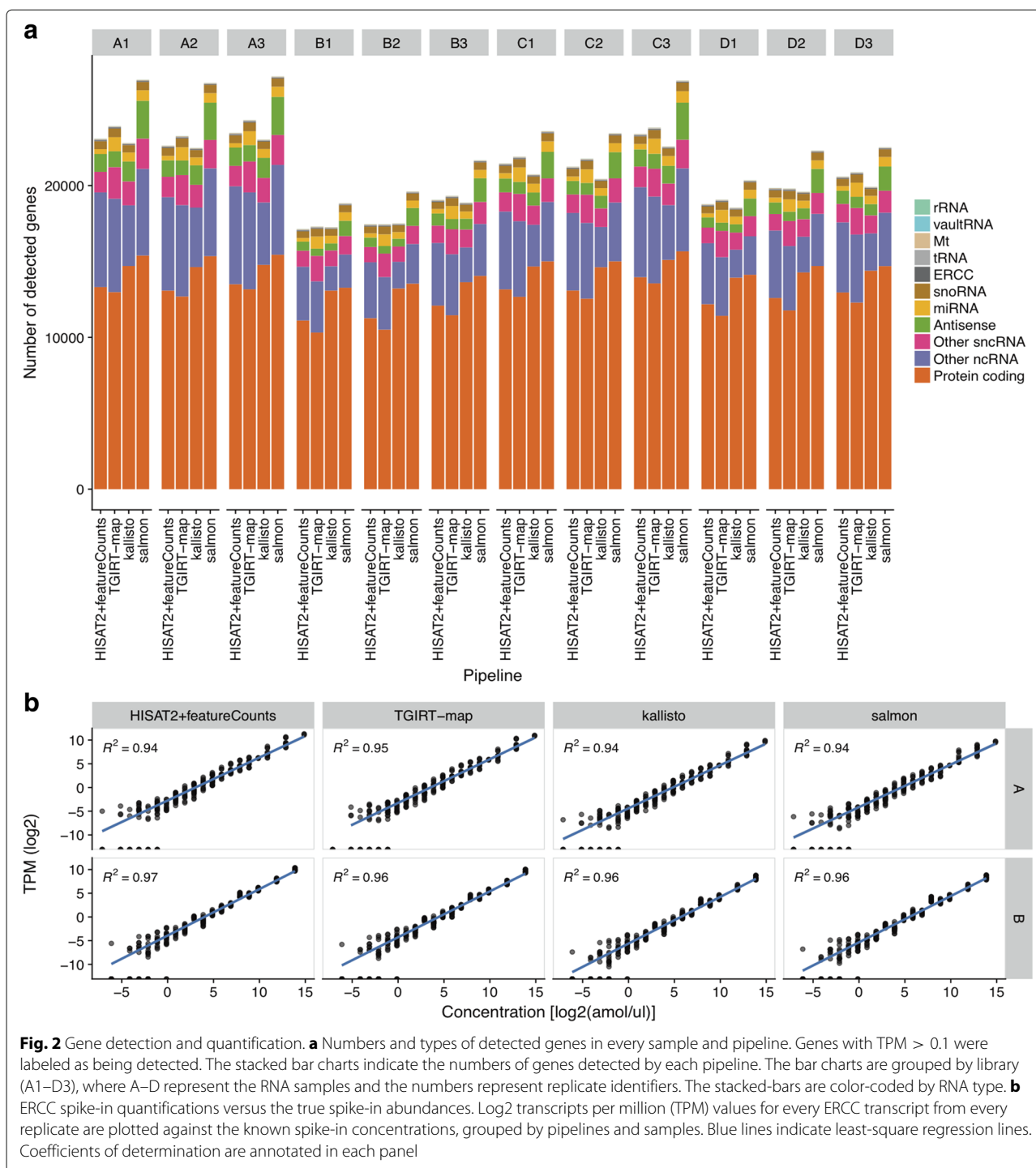
To evaluate if gene expression level estimates were concordant among the tested pipelines, we made pairwise comparisons of these estimates between pipelines (Additional file 4). The gene expression level estimates were generally highly correlated, with Pearson's correlations ranging from 0.68–0.99 (Additional file 4). The Pearson's correlation coefficients were consistently very high for pairwise comparisons between alignment-free pipelines (Kallisto vs Salmon; 0.98–0.99) or between alignment-based pipelines (HISAT2+featureCounts vs TGIRT-map; 0.95–0.96). By contrast, any pairwise correlation between an alignment-free tool and an alignment-based pipeline was generally lower (0.68–0.72; Additional file 4).

Further analyses revealed that different gene types showed distinct variations in gene expression level correlations for every pairwise pipeline comparison (Additional file 4). For instance, ERCC spike-ins, in vitro transcripts that mimic protein-coding transcripts, were recovered with very high correlations for all pairwise pipeline comparisons (Pearson's correlations: 0.99–1; Additional file 4). In comparison to the true abundances that were spiked into the RNA samples [15], the relative expression levels of these ERCC spike-ins were estimated as they were designed and tightly correlated to their true concentrations. We observed a near-perfect linear relationship between inferred TPM values and true concentrations (both log-transformed) for all pipelines (Fig. 2b; $R^2 > 0.94$ for every sample and pipeline; Kruskal-Wallis-test $p = 0.472$). By contrast to ERCC transcripts, the gene expression level estimates of other common gene targets

(antisense, protein-coding genes, etc.) were not as highly correlated among tested pipelines (Additional file 4). To identify the source of this discrepancy, we divided genes into quantile groups of gene lengths or gene expression levels, and found that the abundance estimation inconsistencies among pipelines were largely caused by short gene lengths and low expression levels (Additional file 5), as suggested previously [13].

Differential expression measurements of long genes. The most popular application of RNA-seq is the detection of differentially-expressed genes. To compare the accuracy of differential expression inference among pipelines, we plotted the deviation of measured log2 fold-changes to the expected log2 fold-changes between samples A and B for every ERCC spike-in transcript (Fig 3a; 23 ERCC transcripts in each expected differentially-expressed group). For all pipelines, fold-changes between samples A and B were mostly underestimated (negative Δlog2 fold-changes; Fig. 3a), which may be the result of DESeq2 fold change shrinkage [21]. To quantify the accuracy of fold-change detection for each method, we computed root mean square errors (RMSE) for each ERCC group (Additional file 2). Comparisons of the RMSE values for each ERCC group among pipelines indicated that alignment-free pipelines had comparable performance to alignment-based pipelines in estimating differential expression of ERCC spike-ins (Friedman-test $p = 0.016$; two-sided paired Wilcoxon-test $p > 0.12$ for all pairwise comparisons).

Wu *et al. BMC Genomics* (2018) 19:510

Page 4 of 14

Fig. 2 Gene detection and quantification. **a** Numbers and types of detected genes in every sample and pipeline. Genes with TPM > 0.1 were labeled as being detected. The stacked bar charts indicate the numbers of genes detected by each pipeline. The bar charts are grouped by library (A1–D3), where A–D represent the RNA samples and the numbers represent replicate identifiers. The stacked-bars are color-coded by RNA type. **b** ERCC spike-in quantifications versus the true spike-in abundances. Log2 transcripts per million (TPM) values for every ERCC transcript from every replicate are plotted against the known spike-in concentrations, grouped by pipelines and samples. Blue lines indicate least-square regression lines. Coefficients of determination are annotated in each panel

To test whether the four pipelines provided reliable gene expression level estimates for calling differentially-expressed genes, we called differentially expressed ERCC spike-in transcripts using different p-value cutoffs, compared each call to the known spike-in concentration differences, and plotted the results as receiver operating characteristic (ROC) curves (Fig. 3b). By design, there are 23 spike-ins with same concentrations and 69 spike-ins with different concentrations between samples A and B [22]. Using the areas under the curves (AUCs) of the ROC curves, we found that alignment-free tools (Kallisto and Salmon) performed slightly better than alignment-based pipelines in accurately calling differentially-expressed spike-in transcripts (AUC: 0.65, 0.66, 0.71, and 0.68
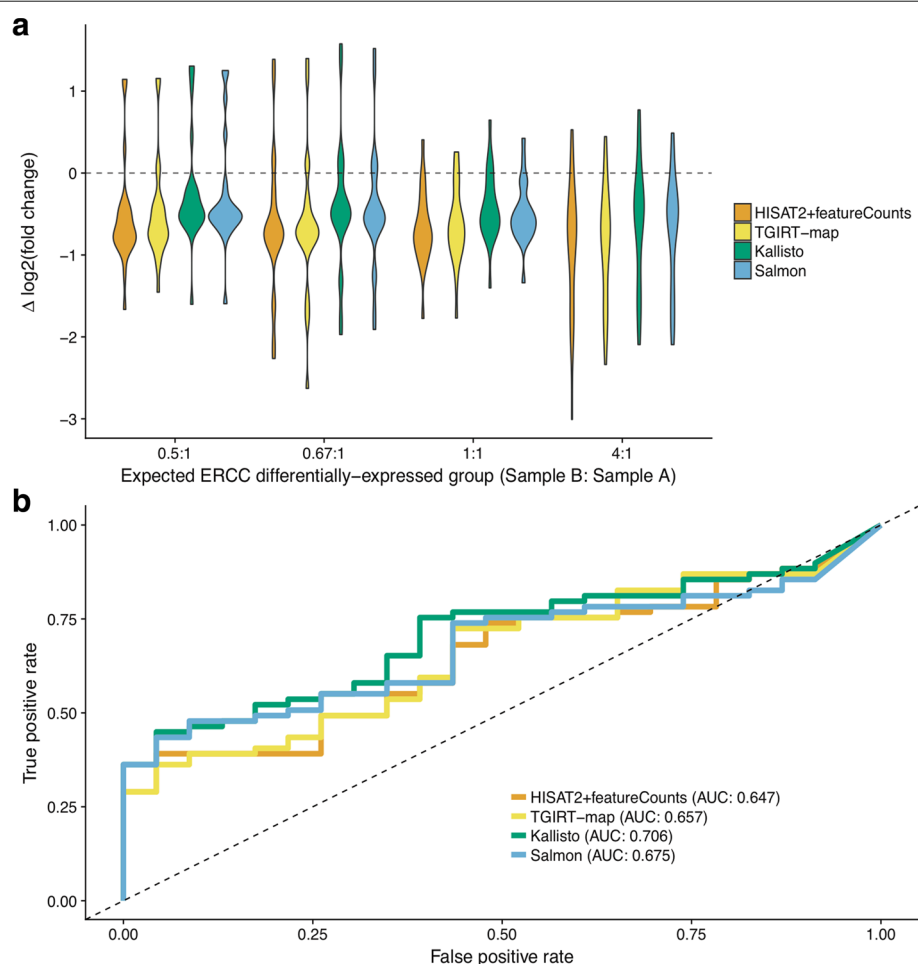
Wu *et al. BMC Genomics*   (2018) 19:510

Page 5 of 14

Fig. 3 Differential expression analysis of ERCC spike-ins. **a** Violin-plots of deviations between measured and expected log2 fold-changes of ERCC transcripts between samples A and B. The distributions of log2 fold-change errors for each ERCC transcript are grouped by their expected differentially-expressed groups and are color-coded by the tested pipelines. The horizontal dashed line indicates no error. **b** Receiver operating characteristic curves for calling ERCC spike-ins as differentially-expressed. Areas under the curve were computed using p-values assigned by the differential expression caller (DESeq2 [21]) on abundance estimations of each ERCC transcript from each pipeline

for HISAT2+featureCounts, TGIRT-map, Kallisto, and Salmon respectively; one non-differentially-expressed and two differentially-expressed ERCC spike-ins had TPM = 0 in all pipelines). In addition to analyzing in vitro ERCC transcripts, we also verified that all pipelines performed nearly identically in terms of quantifying in vivo transcripts, by comparing estimated expression levels to TaqMan qRT-PCR results published previously (Additional file 6) [18].

Whole transcriptome differential expression analysis.
To benchmark the suitability of different gene quantification pipelines for differential expression analysis of all RNA types, we used the known sample mix ratios and the fold-change measurements between samples A and B to construct the expected fold-changes between samples C and D for every gene [18]. By comparing the measured to the expected fold-changes between samples C and D, we found that both alignment-based pipelines showed superior performance over alignment-free pipelines (R^2 = 0.63, 0.62, 0.47, and 0.45 for HISAT2+featureCounts, TGIRT-map, Kallisto, and Salmon, respectively; Fig. 4a). Further analyses showed that low correlations between measured and expected fold-changes were due to lowly-expressed genes (Fig. 4a), as suggested previously [15, 19]. However, to our surprise, Kallisto and Salmon had exceptionally poor fits to the corresponding models for these lowly-expressed genes (Fig. 4a; R^2 = 0.44, 0.43, 0.17, and 0.13 for the lowest 75% expressed genes from HISAT2+featureCounts, TGIRT-map, Kallisto, and Salmon, respectively).

In addition to lowly-expressed genes, we also found that short gene lengths greatly decreased the accuracies in fold-change analyses, particularly for alignment-free
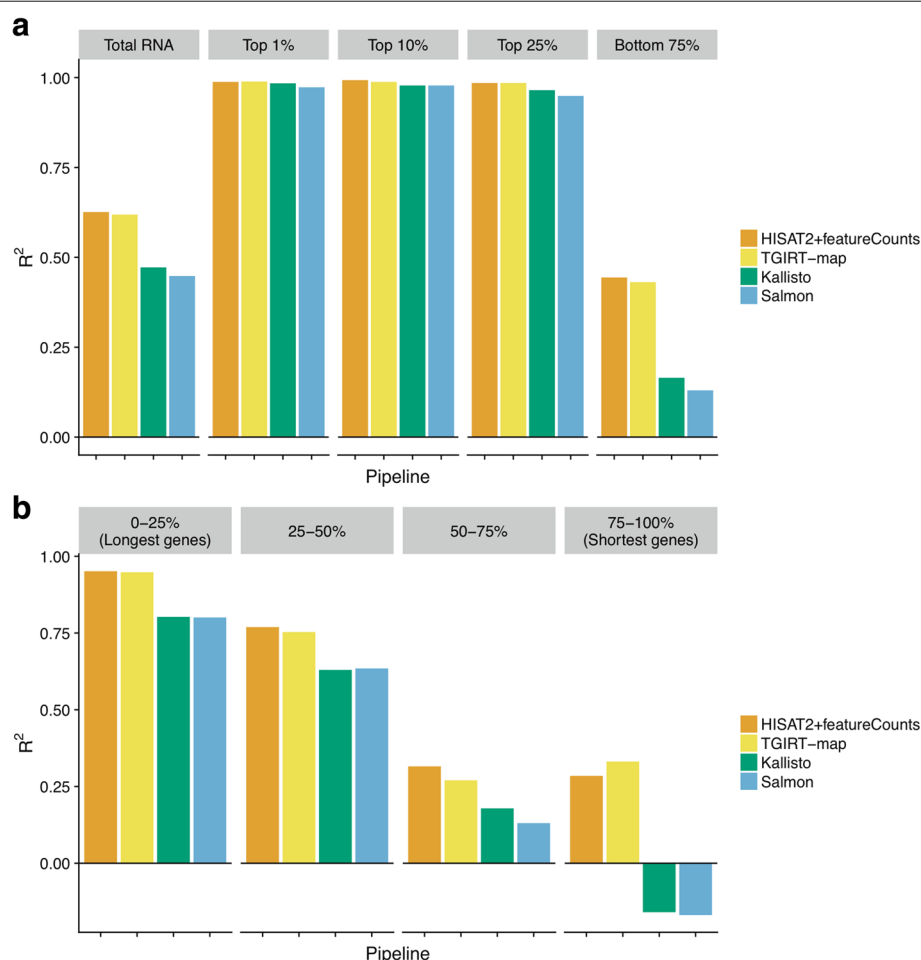
Wu *et al. BMC Genomics*   (2018) 19:510

Page 6 of 14

Fig. 4 R^2 of measured versus expected log2 fold-changes between samples C and D. **a** Gene expression level influenced accuracies of fold-change estimation. R^2 values were computed from the expected and measured log2 fold-change values between samples C and D for each pipeline using different gene sets grouped by average gene expression levels. The first gene set, labeled "Total RNA", includes all genes. The subsequent gene sets include only the genes with the top 1%, top 10%, top 25%, or bottom 75% expression levels, as indicated. Bars are color-coded by pipelines. **b** Gene lengths influenced accuracies of fold-change estimation. Genes from each pipeline were grouped by their gene lengths into four quantile groups. For each quantile group, R^2 values were computed from the expected and measured log2 fold-change values between samples C and D. Bars are color-coded by pipelines. Coefficients of determination (R^2) were computed by `R2` function from `R caret` package [35]. Negative R^2 values indicate exceptionally bad fold-change predictions [23] from the software as illustrated in Additional files 7 and 8, where the fold-change prediction do not fit well to the samples mix-ratio

pipelines (Fig. 4b; R^2 = 0.29, 0.33, −0.16, and −0.17 for the shortest 25% genes from HISAT2+featureCounts, TGIRT-map, Kallisto, and Salmon, respectively). As we are testing whether the measured (i.e., observed) fold-changes fit into the expected fold-changes (i.e., the model) constructed by the fold-changes between samples A and B and the known mix-ratios of samples A and B in samples C and D, R^2 values served as a metric of quantifying the measurement errors relative to the known model. Here, we used $R^2 = 1 - \sum_i (y_i - f_i)^2 / \sum_i (y_i - \bar{y})^2$, where y_i is the predicted and f_i the expected $\log_2$ fold change for gene i, and $\bar{y}$ is the mean $\log_2$ fold change across genes. Thus, a negative R^2 indicates highly discordant measurements relative to the expected fold-changes predicted by the known mix-ratios and the measured fold-changes between samples A and B [23], as illustrated in Additional files 7 and 8. These discordant measurements reflect the difficulty the quantification tools have in making accurate fold-change predictions either between samples A and B (f_i) or between samples C and D (y_i) for small genes (Additional file 7).

To evaluate whether the deficiency in small gene quantification was specific to certain gene types, we computed the root mean square errors (RMSE) for the expected versus measured log2 fold-changes between samples C and D for genes grouped by their gene types. Consistent with the gene quantification results, Salmon and Kallisto showed

slightly better fold-change estimation performance on ERCC spike-ins while alignment-based pipelines showed slightly better fold-change estimation on protein-coding genes (Fig. 5a and Table 1). Furthermore, alignment-based and alignment-free pipelines produced similar results for the majority of small RNA types, such as snoR-NAs. For snoRNAs and other small non-coding RNAs (labeled as other sncRNA), Salmon and TGIRT-map recovered more genes than did HISAT2+featureCounts and Kallisto, but Salmon and TGIRT-map also showed similar RMSE values that were comparable to the RMSE values from HISAT2+featureCounts and Kallisto (Fig. 5a and Table 1). We also found that Salmon quantified

ribosomal RNAs (rRNA) better than other pipelines in these rRNA-depleted libraries (Table 1). The reason may be that Salmon handled multiply-mapped reads for these highly repetitive rRNA differently (comparing to alignment-based pipelines) or used bias corrections (comparing to Kallisto). However, for tRNAs, alignment-based pipelines detected higher numbers of tRNA isoacceptors compared to alignment-free pipelines and showed advantages in detecting differential tRNA expression, yielding lower RMSE values (Fig. 5b and Table 1).

Since we found performance differences in tRNA fold-change analysis between alignment-free and alignment-based pipelines and a performance decrease in small
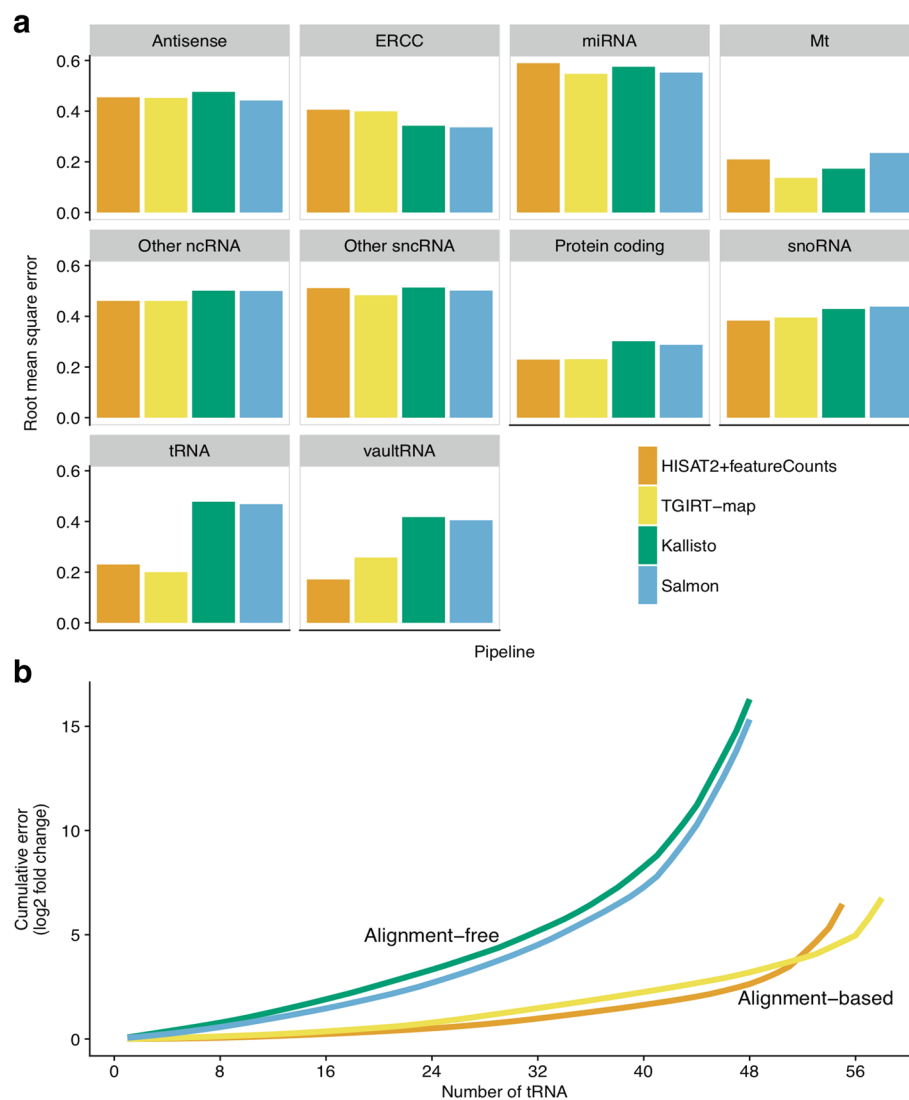

Fig. 5 Pipeline performances of differential expression measurements for different gene types. **a** Root mean square error (RMSE) values between measured and expected log2 fold-change values for each gene type. Bars indicate RMSE values for different gene types from each pipeline and are color-coded by pipelines. **b** Cumulative absolute errors in tRNA log2 fold-change predictions. tRNAs (x-axis) were ordered ascendingly by the absolute errors in log2 fold-change predictions for each pipeline. Cumulative absolute errors for all detected tRNAs are plotted. Lines are color-coded by pipelines

Wu *et al. BMC Genomics* (2018) 19:510

Page 8 of 14

Table 1 RMSE values in differential expression measurements

RNA type	HISAT2+featureCounts	TGIRT-map	Kallisto	Salmon
Antisense	0.455	0.452	0.476	0.442
	(5380)	(5361)	(4855)	(5261)
ERCC	0.406	0.399	0.342	0.336
	(88)	(87)	(88)	(88)
miRNA	0.589	0.547	0.575	0.552
	(558)	(1289)	(952)	(1108)
Mt	0.21	0.137	0.173	0.235
	(36)	(34)	(36)	(36)
Other ncRNA	0.461	0.461	0.501	0.5
	(20611)	(20822)	(18573)	(21614)
Other sncRNA	0.511	0.483	0.514	0.501
	(2236)	(3025)	(2560)	(3077)
Protein coding	0.229	0.231	0.302	0.287
	(19944)	(19965)	(19712)	(19865)
rRNA	0.733	0.608	0.582	0.186
	(4)	(4)	(4)	(4)
snoRNA	0.383	0.395	0.429	0.438
	(620)	(704)	(623)	(742)
tRNA	0.23	0.2	0.478	0.468
	(55)	(58)	(48)	(48)
vaultRNA	0.171	0.258	0.417	0.405
	(6)	(8)	(8)	(9)

Numbers of genes tested by DESeq2 are shown in parentheses. RNA-type annotations were generated from Ensembl [30]. "Other ncRNA" represents the following RNA types: sense intronic, 3prime overlapping ncRNA, processed transcript, sense overlapping, lincRNA, and all pseudogenes. "Other sncRNA" represents the following RNA types: misc RNA, snRNA, scaRNA, sRNA, scRNA. Mt represents all mitochondrial genes, including mitochondrial-encoded tRNAs. vaultRNA represents any Ensembl gene names with Vault or VTRNA

gene quantification for alignment-free pipelines, we anticipated the comparison in tRNA fold-change predictions among the different pipelines would give us insights into the deficiency of alignment-free pipelines in small gene quantification. We hypothesized that the deficiency of alignment-free pipelines in small gene quantification was possibly due to the choice of a long k-mer size (31-mer) relative to the sizes of the small RNA transcripts, such as ∼75 nt for tRNAs. To investigate whether the choice of k-mer size had any effect on small RNA quantification, we tested Salmon with four different k-mer sizes, ranging from 11 to 31 (11, 15, 21, and 31; default is 31 for both Salmon and Kallisto). We found that at $k = 21$, higher R^2 values were observed for total RNA (Additional file 9). Using tRNA as a model for comparisons in small gene quantifications, we also found a performance improvement (detected higher number of tRNA and lower RMSE) using $k = 21$. However, Salmon with this k-mer size still yielded two-fold higher RMSE relative to TGIRT-map for tRNA fold-change prediction (Additional files 2 and 9).

To gain a better understanding to the problem of small gene quantification in alignment-free pipelines, we further inspected the tRNAs that were undetected by both Salmon and Kallisto. Using TGIRT-map alignment results, we found that the read alignments mapped to these tRNAs displayed a high abundance of non-reference bases (Additional file 10). These non-reference bases may be caused by post-transcriptionally modified RNA bases that could introduce reverse-transcription errors by changing base-pairing interfaces [24, 25], as it is known that tRNAs are post-transcriptionally-modified and have abundant base modifications [15, 17, 26]. Thus, we predict that these abundant non-reference bases in small RNAs, tRNA in this case, may have prevented k-mer-based algorithms from successfully counting these reads.

To further test our hypothesis of k-mer assignment being the problem for quantifying small genes, we took advantage of the alignment-based mode of Salmon and compared it to the k-mer-based algorithm implemented in Salmon [6]. Salmon has two different algorithm implementations that accept different input types, either raw sequencing reads (k-mer-based mode) or alignments of sequencing reads pre-assigned to transcripts (alignment-based mode). While the k-mer-based mode performs read assignment and transcript quantification simultaneously, the alignment-based mode only perform transcript quantification on pre-aligned reads [6]. The difference between the two modes allowed us to isolate k-mer matching from transcript quantification in this comparison. Using Salmon alignment-based mode to perform gene quantification on Bowtie2 [20] alignments against the transcriptome, we found that Salmon alignment-based mode is more accurate in quantifying short genes than k-mer-based mode. We observed a higher R^2 value for short gene quantification in Salmon alignment-based method than that in k-mer-based methods with any k value that we have tested (Additional file 9). In additional to short genes, using Salmon alignment-based mode also resulted in a more comparable performance in tRNA quantification relative to TGIRT-map (Additional file 9). These results further supported our hypothesis that k-mer matching algorithms were the sources of inaccurate small RNA quantification.

Discussion

We have performed an in-depth comparison among four RNA-seq pipelines, including two alignment-based and two alignment-free pipelines, to determine the relative performance of these tools for simultaneous quantification of long and small RNAs. The two alignment-based pipelines that we have tested were a widely-adopted align-and-count pipeline (HISAT2+featureCounts) [3, 9] and a custom gene-counting procedure with multi-step iterative alignments (TGIRT-map; Additional file 1); the two

Wu *et al. BMC Genomics*   (2018) 19:510

Page 9 of 14

tested alignment-free pipelines were k-mer counting tools with and without bias corrections (Salmon [6] and Kallisto [5]). We have tested these four pipelines on quantification of both long and small RNAs in a novel benchmarking dataset generated by a thermostable group II intron reverse transcriptase (TGIRT) [15]. This dataset is unique in that small non-coding RNAs are highly represented, while the quality of long RNA quantification is comparable to that of other widely-adopted RNA-seq methods [15]. Using this dataset, we have found that while all four pipelines perform similarly on long and highly-expressed RNAs, alignment-free tools have clear limitations for small and lowly-expressed RNA quantification.

For long gene quantification, we have found that all four pipelines quantify common gene targets (e.g. ERCC spike-ins and protein-coding genes) with similar results, confirming a previously benchmark study on poly-A selected RNA-seq [13]. Generally, gene quantification tends to be more similar between pipelines of the same type (i.e., HISAT2+featureCounts and TGIRT-map, or Kallisto and Salmon) than pipelines of the other type (alignment-based versus alignment-free). This result further supports a previous finding which showed more similar transcript isoforms quantifications were found among alignment-free or among alignment-based pipelines than among pipelines of the different types [14], suggesting that alignment-based pipelines may have somewhat different quantification biases than do the probabilistic models of alignment-free pipelines. Regardless of the difference in gene quantification, our results on differentially-expressed gene detection for long genes showed that all four pipelines performed comparably for ERCC spike-ins and protein-coding genes when compared to their expected fold-changes (ERCC spike-ins) [22] or MAQC TaqMan assay measurements (protein-coding genes) [18]. This result further confirmed previous benchmark studies where alignment-free and alignment-based tools gave similar results for differentially-expressed gene detection [13, 14].

For total gene quantification and differential expression analysis, all tested pipelines generally have performed similarly, with most disagreements occurring between pipeline comparisons of different pipeline types (i.e. an alignment-based pipeline vs an alignment-free pipeline). In the analyses of genes that were inconsistently quantified among pipelines, our results have confirmed that both high gene expression and long gene length were crucial to consistent abundance estimation, as suggested previously [13, 27]. Using fold-change analyses for comparisons in quantification accuracies, we have found that alignment-based tools were more accurate in quantifying lowly-expressed or small genes. This result likely reflects the nature of probabilistic assignments of k-mers and the inferences of TPM values. Although we have found that

alignment-free pipelines were unreliable for quantifying extremely small RNAs (with shorter gene lengths) in total transcriptome analysis, alignment-free tools performed comparably to alignment-based tools for most of the small RNA types, such as snoRNAs or other sncRNAs, in differential expression analyses. This disagreement between small non-coding RNAs and small mRNAs is possibly due to their differences in secondary structures and their different sensitivities to RNA fragmentation prior to RNA-seq library construction, such that overly-fragmented mRNA fragments may be too short to be usable in alignment-free quantifications.

Even though we have found that all pipelines performed similarly on the majority of small non-coding RNAs, our results have revealed that pipelines involving read-alignment steps were superior to alignment-free tools in quantifying tRNAs specifically. We initially hypothesized the differences in performance were due to the choice of k-mer size in the alignment-free pipelines and we have found an improvement in small gene quantifications when a moderately smaller k was chosen ($k = 21$). However, we have also found that the performance of alignment-free quantification at this optimal k was still not comparable to alignment-based pipelines for small genes. Using tRNA as a model, our results suggested that this performance difference was likely due to a combinatorial effect of the choice of k-mer size and misincorporations introduced by post-transcriptionally-modified RNA bases during reverse transcriptions. We reason that a relatively large k-mer size and erroneous sequencing reads likely impede matching to the indexed transcriptome even when these sequencing reads are shredded into k-mers, since all k-mers inherit the same errors or mismatches. Since mismatches on sequencing reads can either be reverse transcription errors or biological variations on small RNAs [28], we expect the same phenomenon may occur if the small RNAs contain single-nucleotide polymorphisms or other biological variations.

Conclusions

In summary, we have shown that different types of pipelines performed similarly for common differentially-expressed gene targets such as protein-coding genes. However, accurate quantification of lowly-expressed or small RNA is difficult to achieve with alignment-free pipelines. Using tRNA as a model, we have also found that k-mer counting algorithms are not compatible for quantifying small genes with abundant biological variations regardless of the choice of k-mer size.

Methods

Data and reference preparation. Raw sequencing reads for TGIRT-seq data generated from the well-studied MAQC samples were downloaded from the NCBI

Wu *et al. BMC Genomics* (2018) 19:510

Page 10 of 14

Sequence Read Archive, accession number SRP066009 [15]. In brief, this dataset includes triplicates of four different human total RNA samples (called A–D) spiked with External RNA Control Consortium (ERCC) transcripts. ERCC spike-in transcripts are 92 *in vitro* transcripts with 250–2000-nt long [22]. Two ERCC spike-in mixes (mixes 1 and 2) with different concentrations for each transcript were spiked into RNA samples A (universal human RNA) and B (human total brain RNA), respectively, to provide known fold-changes of these spike-in transcripts between these two samples [18, 22]. By design, these two spike-in mixes establish four different differentially-expressed gene groups with relative ratios of 0.67:1, 1:1, 1:2, and 1:4 between samples A and B. Samples C and D are mixes of samples A and B in ratios of 3:1 and 1:3, respectively. For detailed library preparations, please refer to Nottingham, et al. [15].

ERCC spike-in sequences were downloaded from the vendor's website (ThermoFisher; https://tools.thermofisher.com/content/sfs/manuals/cms_095047.txt). Transfer RNA (tRNA) annotations and sequences were downloaded from the Genomic tRNA Database (http://gtrnadb.ucsc.edu/GtRNAdb2/genomes/eukaryota/Hsapi38/hg38-tRNAs.tar.gz) [29]. For tRNA sequences, introns were removed, and all sequences were de-duplicated at the sequence level to reduce multiply-mapped reads and improve tRNA counting. 5S rDNA (GenBank accession: X12811.1) and complete rDNA repeat unit (GenBank accession: U13369.1) sequences were downloaded from GenBank (NCBI). Human genome sequences and annotations were downloaded as FASTA and GTF files from Ensembl (human genome build version GRCh38.88) [30], combined with ERCC spike-in sequences and rDNA sequences, and indexed by HISAT2 v2.1.0 [9]. A HISAT2 splice-site file was created using the human genome build GRCh38.88 GTF file [9]. ERCC and rRNA records were then added to the GTF file and human transcriptome reference sequences were extracted using gffread (https://github.com/gpertea/gffread). After appending mature tRNA sequences to the human transcriptome reference, the transcriptome reference sequence file was indexed with Kallisto v0.43.0 [5] and Salmon v0.8.2 [6] using the default k-mer size (k = 31). Transcriptomic sequences and annotations were deposited to Zenodo (https://doi.org/10.5281/zenodo.1219203).

Adaptor trimmings. Raw reads were adapter- and quality- trimmed by atropos v1.1.16 [31] via:

```
atropos trim -m 15 -O 5 -n 3 -q 20 \
--minimum-length=15 \
--threads=${THREADS} \
--error-rate=0.1 \
--overlap 5 \
--quality-cutoff=20 \
--aligner insert \
-b AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC \
-B GATCGTCGGACTGTAGAACTCTGAACGTGTAGA \
-o ${TRIMMED_FQ1} -p ${TRIMMED_FQ2} \
${FQ1} ${FQ2}
```

HISAT2 mapping. Reads were aligned to the genome using HISAT2 v2.1.0 [9] via:

```
hisat2 -k 10 --no-mixed \
-no-discordant \
-p $THREADS \
--known-splicesite-infile=${SPLICESITE_
FILE} \
-1 ${TRIMMED_FQ1} -2 ${TRIMMED_FQ2}
```

HISAT2+featureCounts. Trimmed reads were aligned using HISAT2 mapping. Gene counts from HISAT2 mapping were generated by featureCounts v1.5.3 [3] via:

```
featureCounts -F SAF -O -s 1 -M -T 24 \
--largestOverlap --minOverlap 10 \
--primary -p -P -d 10 -D 10000 -B \
-C -donotsort \
-o ${OUTPUT_PREFIX} \
${LIST_OF_BAM_FILES}
```

TGIRT-map. Our custom pipleine TGIRT-map first filters out all trimmed reads that can be aligned to tRNA or rRNA references to reduce multiply-mapped reads (Additional file 1). The remaining unaligned reads are then sequentially aligned to the human genome using a splice-aware aligner (HISAT2) [9] and a sensitive local aligner (BOWTIE2) [20]. A single alignment locus is picked from the multiply-mapped fragments by asserting the assumptions that (A) RNA-seq fragments are small (smallest insert size), (B) ribosomal genes are abundant in the genome (ribosomal gene loci), and (C) fragments are unlikely to be originated from haplotype or patch sequences (as defined by Ensembl annotations) of the artificially assembled genome. Finally, gene quantification is done on genomic loci, except for tRNA and rRNA which require an additional step of re-aligning and counting.

Trimmed reads were first aligned to rRNA (GenBank accession numbers: X12811.1 and U13369.1) and tRNA sequences with BOWTIE2 v2.3.3.1 [20] via:

```
bowtie2 -p ${THREADS}\
--local --score-min G,1,10 -D 20 \
-R 3 -N 0 -L 8 -i S,1,0.50 \
--no-mixed --norc --no-discordant --dovetail \
```

Wu *et al. BMC Genomics* (2018) 19:510

Page 11 of 14

```
-x ${tRNA_rRNA_INDEX} \
-1 ${TRIMMED_FQ1} -2 ${TRIMMED_FQ2}
```

We define the mapped reads as tRNA and rRNA pre-mapped reads. Unmapped reads were then aligned to the genome with HISAT2 as described above in HISAT2 mapping [9]. Unaligned reads from HISAT2 mapping were then rescued by re-aligning locally to the genome with BOWTIE2 [20] via:

```
bowtie2 -p ${THREADS} \
--local --score-min G,1,10 \
-D 20 -R 3 -N 0 -L 8 -i S,1,0.50 \
-k 10 --no-mixed --dovetail --no-discordant \
-x ${GENOME_INDEX} \
-1 ${FQ1} -2 ${FQ2}
```

All alignment pairs with > 10 nucleotides of soft-clipped bases on either reads or were discordant pairs were discarded. Multiply-mapped reads from the two genome-mapping steps were grouped and a pair of best alignment was chosen by the following ordered criteria: the alignment pair (A) had the smallest insert size, (B) was mapped to ribosomal gene locus, or (C) was mapped to chromosome 1–23, X or Y. If none of these criteria was met by a single alignment pair, a random alignment pair was chosen from the multiply-mapped loci. These filtered alignments were then merged with the uniquely-mapped alignments. All reads that mapped to tRNA or rRNA loci were extracted by BEDtools v2.26 using `pairtobed` command with options `-s -f 0.01` [32], combined with tRNA and rRNA pre-mapped reads and re-aligned to tRNA and rRNA references. Counts were generated from the aligned BAM file. Counts for each anticodon were aggregated. Other gene counts were calculated by converting the genome alignments to fragment coordinates in a BED file using the BEDtools `bamtobed` command and counted using the BEDtools `coverage` command [32]. TGIRT-map pipeline is available at: https://github.com/wckdouglas/tgirt_map.

Kallisto and Salmon. We used Kallisto v0.43.0 [5] and Salmon v0.8.2 [6] for our alignment-free pipelines. In both cases, adaptor-trimmed reads were used as input.

We called Salmon with the following command-line arguments:

```
salmon quant \
--seqBias --gcBias \
--index ${INDEX} \
--libType ISF \
--writeMappings \
--threads=${THREADS} \
--auxDir ${RESULTPATH} \
--numBootstraps 100 \
--mates1 ${TRIMMED_FQ1} \
```

```
--mates2 ${TRIMMED_FQ2} \
--output ${RESULTPATH}
```

We called Kallisto with the following command-line arguments:

```
kallisto quant \
-i ${INDEX} \
-o ${OUTPUT_PATH} \
--fr-stranded \
--threads=${THREADS}\
${TRIMMED_FQ1} ${TRIMMED_FQ2}
```

Salmon alignment mode Trimmed reads were first aligned to the transcriptome using Bowtie2 [20] via:

```
bowtie2 --threads ${THREADS} \
--score-min G,1,10 \
-k 30 -D 20 -R 3 -N 0 -L 8 -i S,1,0.50 \
--no-mixed --no-discordant --dovetail \
--fr --local \
-x ${TRANSCRIPTOME_INDEX} \
-1 ${TRIMMED_FQ1} -2 ${TRIMMED_FQ2}
```

Transcript quantification was done by Salmon [6] via:

```
salmon quant \
--seqBias --gcBias \
--libType ISF \
--targets ${TRANSCRIPTOME_FASTA} \
--threads=${THREADS} \
--auxDir ${RESULTPATH}/${SAMPLENAME} \
--numBootstraps 100 \
--useErrorModel \
--alignments ${BAM_FILE} \
--output ${RESULTPATH}/${SAMPLENAME}
```

Differential expression analysis. DESeq2 v1.14.1 [21] was used for all differential expression analyses.

Because DESeq2 does not accept TPM values as input, transcript TPM values from Salmon and Kallisto were converted to gene-level counts using Tximport v1.4.0 [33] before any differential-expression analyses.

Predicted $\log_2$ fold change between samples C and D. Predicted fold changes between samples C and D were computed using the following equation as described in Su, et al. [19]:

$$\log_2 \frac{C}{D} = \log_2(k_1 \frac{A}{B} + (1 - k_1)) - \log_2(k_2 \frac{A}{B} + (1 - k_2))$$

where $\frac{C}{D}$ indicates predicted fold change between samples C and D, $\frac{A}{B}$ indicates measured fold change between samples A and B, $k_1 = \frac{3z}{3z+1}$, and $k_2 = \frac{z}{z+3}$. In this work, $z = 1.43$ was used as suggested previously [19].

Wu *et al. BMC Genomics* (2018) 19:510

Page 12 of 14

Additional files

Additional file 1: Flow chart of the custom pipeline TGIRT-map. As TGIRT-seq simultaneously recovers a large fraction of small structural RNAs in addition to regular long RNAs [15, 17], TGIRT-map was designed to optimize the quantification of these small genes. The pipeline is similar to the Genobee-exceRpt small RNA-seq pipeline [34], where reads are first aligned against the tRNA and rRNA sequences to avoid ambiguous assignments in later steps. Unaligned reads (red arrow) are iteratively aligned to the human genome by HISAT2 [9] and BOWTIE2 [20] to minimize unassigned reads. Multiply-mapped reads are extracted and assigned as uniquely-mapped at loci selected by the following criteria in order: the pair of alignments (A) has the smallest insert size, (B) is mapped to ribosomal gene loci, and (C) is mapped to chromosome 1-23, X or Y, otherwise a random pair of alignments is selected. From these genome alignments, reads that aligned to tRNA or rRNA loci are extracted and combined with tRNA and rRNA reads from the first step. Finally, these combined reads are realigned to the tRNA and rRNA sequences to generate gene counts. (PNG 247 kb)

Additional file 2: Zip file containing several data tables in tab-separated format, as well as a readme file that explains the contents of each data file. (ZIP 5 kb)

Additional file 3: Intersections of detected genes among tested pipelines. Genes with TPM > 0.1 were defined as being detected. (a–d) UpSet graphs of intersections of detected genes among pipelines. An UpSet graph describes intersections of detected genes among subsets of tested pipelines for samples A–D. Each UpSet graph contains a bar chart and a sample-comparison matrix. Bars indicate the numbers of detected genes in the specific intersection specified in the sample-comparison matrix. Each bar indicates the number of genes detected only in the intersection of the set of compared pipelines. The sample-comparison matrix at the bottom describes the set of pipelines within which the comparison is made. Grey boxes in the matrix annotate the pipelines included in the comparison. (e) Uniquely-detected genes. The heights of the stacked bars indicate the numbers of genes uniquely detected by the corresponding pipeline. Stacked bars are color-coded with gene types. (PNG 428 kb)

Additional file 4: Pearson's correlation coefficients for pairwise pipeline comparisons in gene quantifications. For every pairwise pipeline comparison of each sample (A–D), Pearson's correlation coefficients were computed using the average log2 TPM of each gene across the triplicates for each sample. Bar heights indicate average correlations among the four samples A–D. Error bars represent standard deviations among the four samples A–D. Red bars represent total RNA correlation coefficients (all genes). Grey bars indicate correlation coefficients grouped by gene type. Each panel represents a single pairwise pipeline comparison. (PNG 207 kb)

Additional file 5: Impact of gene lengths and expression levels on gene quantification correlations among pipelines. (a) Gene lengths influence gene expression estimation correlations. Genes are separated into quantile groups according to their lengths. For every quantile group from each sample, Pearson's correlation coefficients of average TPMs for each gene across replicates were computed for every pairwise pipeline comparison and then averaged across the four samples. Plotted are the average correlations versus the quantile groups. Each line represents a pairwise comparison and is color-coded accordingly. (b) Gene expression levels influence gene expression estimate correlations. Genes were separated into quantile groups according to their average expression values (TPMs). For every quantile group from each sample, Pearson's correlations were computed in the same way as in panel (a). Each line represents a pairwise comparison and is color-coded accordingly. (PNG 412 kb)

Additional file 6: Differential expression analysis of protein-coding genes. **(a)** Comparison of measured fold-changes between samples A and B from the TGIRT-seq analysis and the MAQC TaqMan assay. MAQC TaqMan assay data for the same samples were downloaded from the NCBI Gene Expression Omnibus (GEO), accession number GSE5350 [18]. 972 unique protein-coding genes were compared. Log2 fold-changes were calculated from MAQC data and compared to the DESeq2 results for each pipeline. Distributions of the deviations between RNA-seq and MAQC data log2 fold-changes were plotted as violin-plots. Distributions are color-coded by pipelines. **(b)** ROC curves for calling differentially-expressed (DE)

protein-coding genes. MAQC TaqMan assay genes were binary-labeled as DE if absolute log2 fold-change > 0.5 and not DE otherwise (739 DE and 225 non-DE), as suggested by a previous study [10]. ROC curves for DESeq2 results from abundance estimates computed by each pipeline were plotted. AUCs were computed for each ROC curve to quantify performances of the differential expression caller (DESeq2) [21] on abundance estimates from each pipeline. ROC curves are color-coded by pipelines. The dotted diagonal line indicates random guessing (AUC=0.5). (PNG 352 kb)

Additional file 7: Measured versus expected log2 fold-changes between samples C and D for Kallisto. Expected log2 fold-changes for each gene (denoted by red line) were constructed from (A) the Kallisto-measured log2 fold-changes between samples A and B, and (B) the known sample mix ratio of samples A and B in samples C and D, using a previously-described algorithm [19], see method. Each panel shows a two-dimensional kernel density estimation of the distribution of measured vs expected log2 fold-changes between samples C and D for genes grouped by quantile groups of gene lengths. Coefficients of determination for each group of genes are annotated as R^2. Perfectly correlated log2 fold-changes are annotated by the diagonal red lines. Panels with higher R^2 generally show denser trends along the red line, indicating the log2 fold-changes between samples C and D were accurately predicted by the software that recapitulated the sample-mixing ratio and thus, fitted better to the model (red line). For groups with lower R^2, a denser (lighter color) horizontal trend was observed on the right side of the red line, indicating the software either over-estimated log2 fold-changes between samples A and B that led to construction of an incorrect model or under-estimated log2 fold-changes between samples C and D that led to a poor fit to the model. (PNG 471 kb)

Additional file 8: Measured versus expected log2 fold-changes between samples C and D for the shortest 25% genes from each pipeline. Expected log2 fold-changes for each gene (denoted by red line) were constructed from (A) the measured log2 fold-changes between samples A and B, and (B) the known sample mix ratio of samples A and B in samples C and D, using a previously-described algorithm [19]. Each panel shows a two-dimensional kernel density estimation of the distribution of measured vs expected log2 fold-changes between samples C and D for genes in the lowest gene length quantile. Coefficients of determination for each group of genes are annotated as R^2. Perfectly correlated log2 fold-changes are annotated by the diagonal red lines. (PNG 428 kb)

Additional file 9: Performances of Salmon using alignment-mode and different k-mer sizes of 11, 15, 21, and 31 (default). **(a)** Accuracies of fold-change estimation in quantile groups of gene lengths. All detected genes from TGIRT-map or Salmon with different k-mer sizes were grouped by their gene lengths into four quantile groups. For each quantile group, R^2 values were computed from the expected and measured log2 fold-change values between samples C and D. Bars are color-coded by pipelines. Coefficients of determination (R^2) were computed by `R2` function from `R caret` package [35]. **(b)** Cumulative absolute errors in tRNA log2 fold-change predictions for each k-mer size for Salmon and TGIRT-map. tRNAs were sorted in ascending order by their absolute errors in log2 fold-change predictions. Cumulative absolute errors for all detected tRNA were plotted. Lines are color-coded by pipeline. (PNG 286 kb)

Additional file 10: Erroneous sequencing reads may have impeded k-mer counting for tRNAs. IGV genome browser screen-shots of three tRNA alignments from Sample A1 aligned via TGIRT-map were shown. The three tRNA alignments being shown are (a) TRR-CCG, (b) TRS-ACT, and (c) TRT-AGT. Arrows on the top of each panel indicate the corresponding tRNA transcript strands (5′ to 3′) and sizes. Bar charts below the arrows represent coverages at every position along the tRNA transcripts. Read alignments from Sample A1 aligned by TGIRT-map are shown below the bar charts. Grey colors represent read bases that match the reference bases. Other colors indicate bases on the read alignments that do not match the reference bases (purple, red, green, blue, and brown colors indicate insertions, thymidines, adenosines, cytidines, and guanosines on the read alignments, respectively). These errors may reflect post-transcriptionally-modified RNA bases, which interfere with canonical base-pairings and lead to misincorporations during TGIRT reverse transcriptions [15, 17, 26]. Errors due to misincorporations at known and unknown sites of post-transcriptionally modified RNA bases are highlighted with boxes (m$_2^2$G26: N2,N2-dimethylguanosine at position 26;

Wu *et al. BMC Genomics* (2018) 19:510

Page 13 of 14

m^{2}G26: N2-methylguanosine at position 26; I34: Inosine at position 34; m^{1}A58: 1-methyladenosine at position 58; see also [15, 17, 36]). Unannotated mismatches from both ends can be untrimmed adapters or tRNA-precusor sequences. We observe that on these tRNAs, the errors induced by RNA-base-modification occurred at positions ~20, 30, or 50 on ~75 nucleotide long tRNA transcripts. Therefore, we expect that every k-mer originating from an erroneous sequencing read inherits at least one of these errors, and therefore k-mers cannot be matched accurately to the transcript database. The effect is worse when a large k-mer size is selected, such as the default $k = 31$. (PNG 337 kb)

Abbreviations
AUC: area under the curve; ERCC: External RNA controls consortium; MAQC: microarray quality control consortium; RMSE: root mean square error; ROC: receiver operating characteristic; TGIRT: Thermostable group II intron reverse transcriptase; TPM: transcripts per million

Availability of data and materials
All scripts used for data processing are deposited on GitHub at: https://github.com/wckdouglas/tgirt_benchmark, and archived at Zenodo (https://doi.org/10.5281/zenodo.1239242). Raw RNA-seq dataset can be accessed through NCBI Sequence Read Archive, accession number SRP066009.

Authors' contributions
DCW, AML, COW designed the experiments. DCW, KSH, JY, AML and COW wrote the manuscript. DCW, KSH and JY conducted the experiments and analyzed the data with input from AML and COW. All authors read and approved the final manuscript.

Ethics approval and consent to participate
This study has not involved any material directly derived from humans, animals or plants, though is based on pre-existing, publicly available datasets.

Competing interests
Thermostable group II intron reverse transcriptase (TGIRT) enzymes and methods for their use are the subject of patents and patent applications that have been licensed by The University of Texas and East Tennessee State University to InGex, LLC. A.M.L. and The University of Texas are minority equity holders in InGex, LLC, and A.M.L. and some present and former Lambowitz laboratory members receive royalty payments from the sale of TGIRT enzymes and kits and from the licensing of intellectual property.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]Institute for Cellular and Molecular Biology, The University of Texas at Austin, Austin, TX, 78712, USA. [2]Department of Molecular Biosciences, The University of Texas at Austin, Austin, TX, 78712, USA. [3]Department of Integrative Biology, The University of Texas at Austin, Austin, TX, 78712, USA.

References
1. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Meth. 2008;5(7):621–8.
2. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protocol. 2012;7(3):562–78.
3. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics (Oxford, England). 2014;30(7):923–30.
4. Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat Biotech. 2014;32(5):462–4.
5. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotech. 2016;34(5):525–7.
6. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Meth. 2017;advance online publication.
7. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protocol. 2016;11(9):1650–67.
8. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.
9. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12(4):357–60.
10. Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, Mason CE, Socci ND, Betel D. Comprehensive evaluation of differential gene expression analysis methods for. Genome Biol. 2013;14(9):95.
11. Teng M, Love MI, Davis CA, Djebali S, Dobin A, Graveley BR, Li S, Mason CE, Olson S, Pervouchine D, Sloan CA, Wei X, Zhan L, Irizarry RA. A benchmark for RNA-seq quantification pipelines. Genome Biol. 2016;17(1):74.
12. Baruzzo G, Hayer KE, Kim EJ, Di Camillo B, FitzGerald GA, Grant GR. Simulation-based comprehensive benchmarking of RNA-seq aligners. Nat Meth. 2017;14(2):135–9.
13. Everaert C, Luypaert M, Maag JLV, Cheng QX, Dinger ME, Hellemans J, Mestdagh P. Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data. Sci Rep. 2017;7(1):1559.
14. Sahraeian SME, Mohiyuddin M, Sebra R, Tilgner H, Afshar PT, Au KF, Asadi NB, Gerstein MB, Wong WH, Snyder MP, Schadt E, Lam HYK. Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. Nat Commun. 2017;8(1):59.
15. Nottingham RM, Wu DC, Qin Y, Yao J, Hunicke-Smith S, Lambowitz AM. RNA-seq of human reference RNA samples using a thermostable group II intron reverse transcriptase. RNA. 2016;22(4):597–613.
16. Mohr S, Ghanem E, Smith W, Sheeter D, Qin Y, King O, Polioudakis D, Iyer VR, Hunicke-Smith S, Swamy S, Kuersten S, Lambowitz AM. Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing. RNA. 2013;19(7):958–70.
17. Qin Y, Yao J, Wu DC, Nottingham RM, Mohr S, Hunicke-Smith S, Lambowitz AM. High-throughput sequencing of human plasma RNA by using thermostable group II intron reverse transcriptases. RNA. 2016;22(1):111–28.
18. Consortium M. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nat Biotech. 2010;28(8):827–38.
19. SEQC/MAQC-III Consortium: A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat Biotech. 2014;32(9):903–14.
20. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Meth. 2012;9(4):357–9.
21. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
22. Jiang L, Schlesinger F, Davis CA, Zhang Y, Li R, Salit M, Gingeras TR, Oliver B. Synthetic spike-in standards for RNA-seq experiments. Genome Res. 2011;21(9):1543–51.

Wu *et al. BMC Genomics* (2018) 19:510

Page 14 of 14

23. Alexander DLJ, Tropsha A, Winkler DA. Beware of R2: simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. J Chem Inf Model. 2015;55(7):1316–22.
24. Ryvkin P, Leung YY, Silverman IM, Childress M, Valladares O, Dragomir I, Gregory BD, Wang L-S. HAMR: high-throughput annotation of modified ribonucleotides. RNA. 2013;19(12):1684–92.
25. Clark WC, Evans ME, Dominissini D, Zheng G, Pan T. tRNA base methylation identification and quantification via high-throughput sequencing. RNA. 2016;19:1684–92.
26. Katibah GE, Qin Y, Sidote DJ, Yao J, Lambowitz AM, Collins K. Broad and adaptable RNA structure recognition by the human interferon-induced tetratricopeptide repeat protein IFIT5. Proc Natl Acad Sci USA. 2014;111(33):12025–30.
27. Robert C, Watson M. Errors in RNA-Seq quantification affect genes of relevance to human disease. Genome Biol. 2015;16:177.
28. Hrdlickova B, de Almeida RC, Borek Z, Withoff S. Genetic variation in the non-coding genome: Involvement of micro-RNAs and long non-coding RNAs in disease. Biochim Biophys Acta (BBA) - Mol Basis Dis. 2014;1842(10):1910–22.
29. Chan PP, Lowe TM. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. Nucleic Acids Res. 2016;44(D1):184–9.
30. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, Gil L, Girón CG, Gordon L, Hourlier T, Hunt SE, Janacek SH, Johnson N, Juettemann T, Kähäri AK, Keenan S, Martin FJ, Maurel T, McLaren W, Murphy DN, Nag R, Overduin B, Parker A, Patricio M, Perry E, Pignatelli M, Riat HS, Sheppard D, Taylor K, Thormann A, Vullo A, Wilder SP, Zadissa A, Aken BL, Birney E, Harrow J, Kinsella R, Muffato M, Ruffier M, Searle SMJ, Spudich G, Trevanion SJ, Yates A, Zerbino DR, Flicek P. Ensembl 2015. Nucleic Acids Res. 2015;43(D1):662–9.
31. Didion JP, Martin M, Collins FS. Atropos: specific, sensitive, and speedy trimming of sequencing reads. PeerJ. 2017;5:3720.
32. Quinlan AR. BEDTools: the swiss-army tool for genome feature analysis. Curr Protocol Bioinform. 2014;47:11–12134.
33. Soneson C, Love M, Robinson M. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences [version 1; referees: 2 approved]. F1000Research. 2015;4(1521):1521.
34. Freedman JE, Gerstein M, Mick E, Rozowsky J, Levy D, Kitchen R, Das S, Shah R, Danielson K, Beaulieu L, Navarro FCP, Wang Y, Galeev TR, Holman A, Kwong RY, Murthy V, Tanriverdi SE, Koupenova M, Mikhalev E, Tanriverdi K. Diverse human extracellular RNAs are widely detected in human plasma. Nat Commun. 2016;7:11106.
35. Kuhn M. Building Predictive Models in R Using the caret Package. J Stat Softw. 2008;28(5):1–26.
36. Hopper AK, Phizicky EM. tRNA transfers to the limelight. Genes Dev. 2003;17(2):162–80.