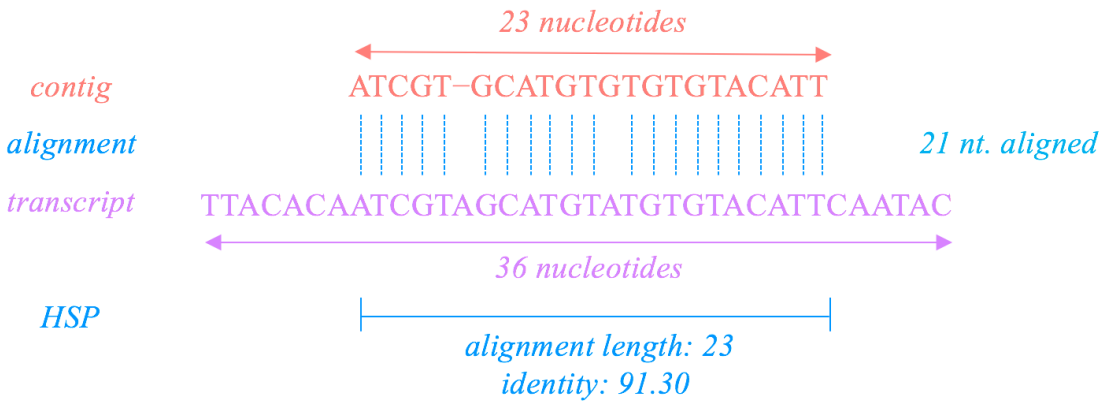
Effect of *de novo* transcriptome assembly on transcript quantification

Additional File 2



$$recovery = \frac{23 \times 91.30}{36} = 0.583 \qquad accuracy = \frac{23 \times 91.30}{23} = 0.913$$

Fig. S1: Recovery and Accuracy.
The diagram illustrates a global alignment between a contig and a transcript. The blue dot lines between sequences represent the matched nucleotides suggested by BLASTN. Based on the global alignment, the recovery is defined as the proportion of nucleotides on transcript that are reconstructed by the contig, while the accuracy is defined as the proportion of nucleotides correctly matched on the contig. Note that these metrics are used to describe the global alignments between contigs and transcripts instead of contig or transcript sequences.

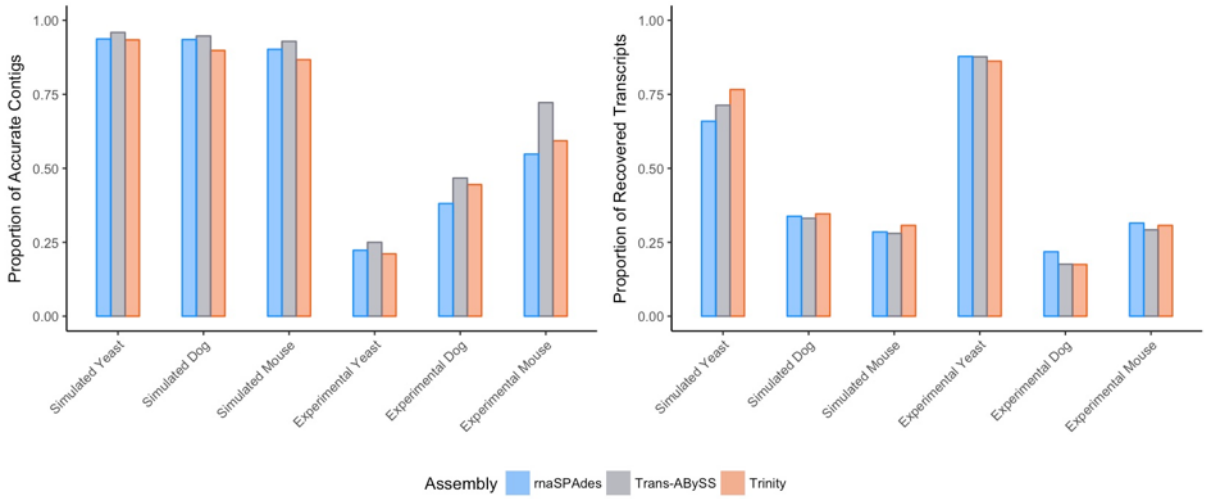


Fig S2: Proportion of Accurate Contigs and Recovered Transcripts

The bar plots illustrate the proportion of accurate contigs (contig aligned with at least one transcript that shows *accuracy* $\geq$ 90) and recovered transcripts (transcript aligned with at least one contig that shows *recovery* $\geq$ 90). In general, the proportion of recovered transcripts is significantly higher for yeast dataset. It appears to be more difficult for the assemblers to properly reconstruct the transcriptome for sequences with higher complexity. Moreover, the proportion of correct contigs for Trans-ABySS is higher in all the datasets and the proportion of recovered transcripts for Trinity is higher in the simulated datasets. However, the performance between different assemblers shows only marginal difference.

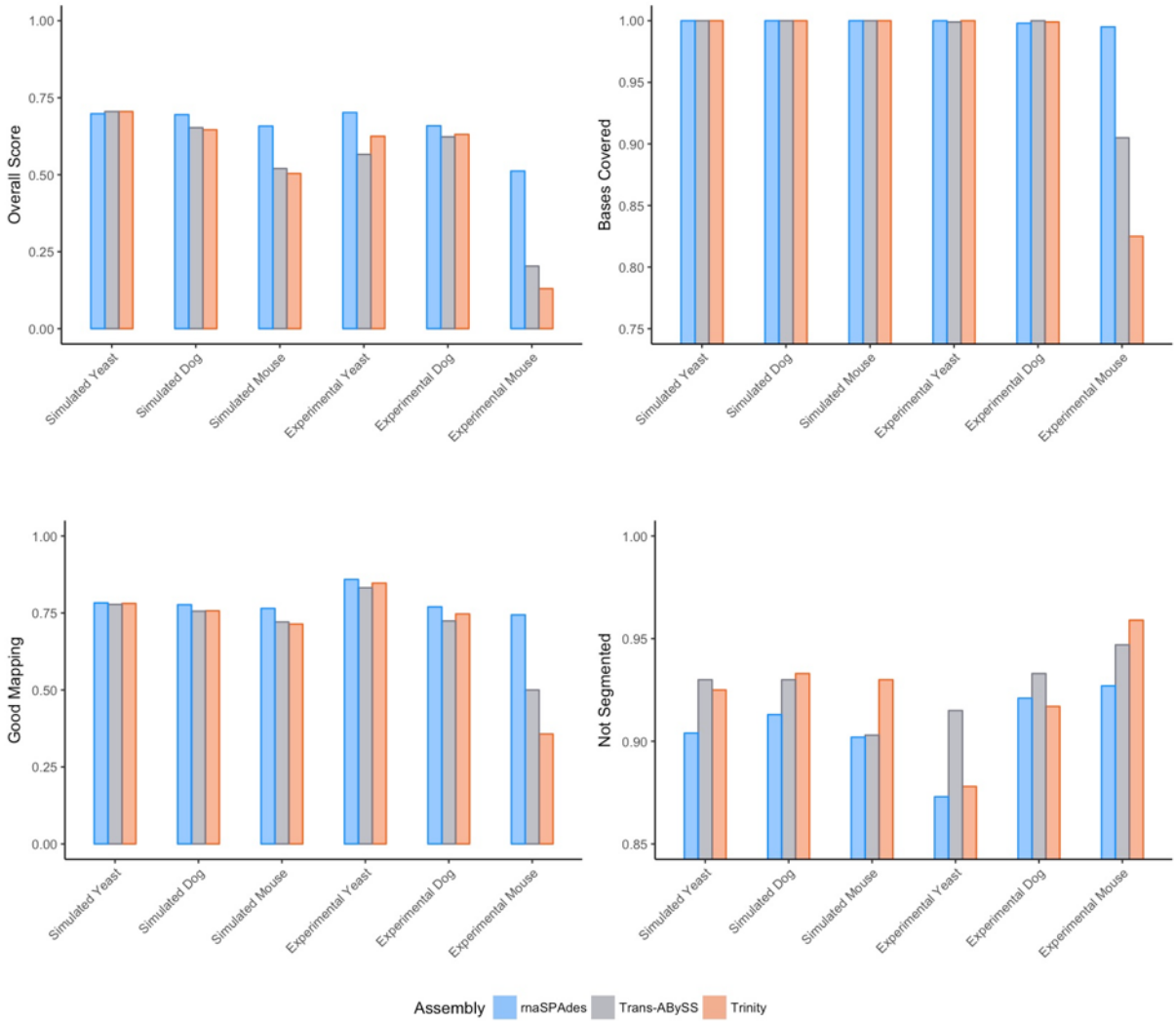


Fig S3: Median of TransRate Scores

The bar plots illustrate the median of TransRate scores for assembled contigs. In general, the *overall TransRate scores* are higher for the contigs constructed by rnaSPAdes. However, the median of TransRate scores of *Bases Covere*d, *Good* and *Not Segmented* varied greatly across different dataset; therefore, it is hard to conclude which assembler outperformed the others based on these metrics.

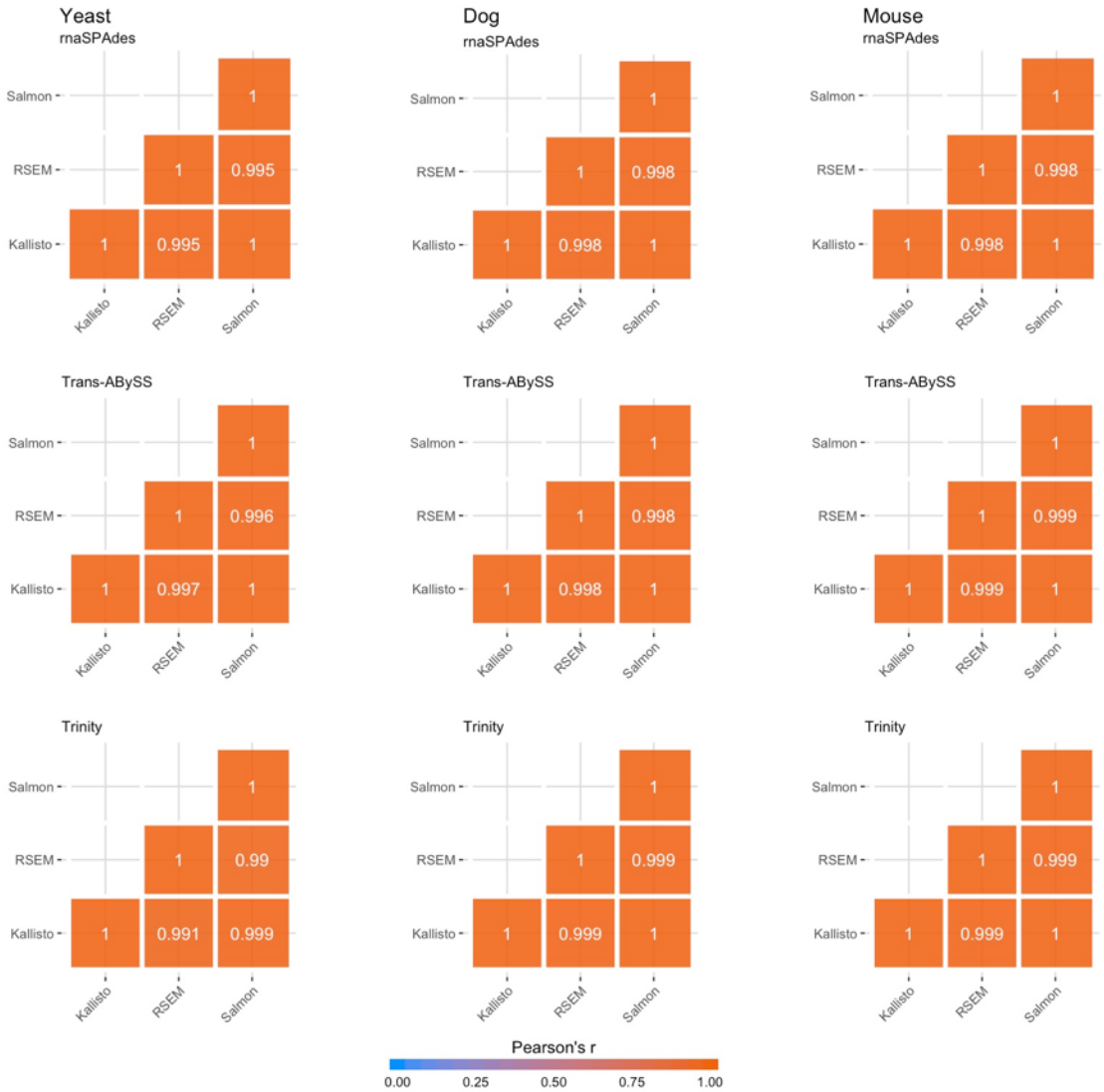


Fig S4-1: Pearson's Correlation Coefficient between Quantifiers (Simulated)
The correlation matrices illustrate the Pearson's correlation coefficient between the estimation made by any of the two quantifiers. The matrices in the left column are the results drawn from yeast dataset, in the middle column are the result from dog dataset and in the right column from mouse dataset. The matrices in the first row are the estimation based on rnaSPAdes assembly, the second row on Trans-ABySS assembly, and the third row on Trinity. In general, the correlation matrices show high consistency for the estimation made by quantifiers on simulated datasets.

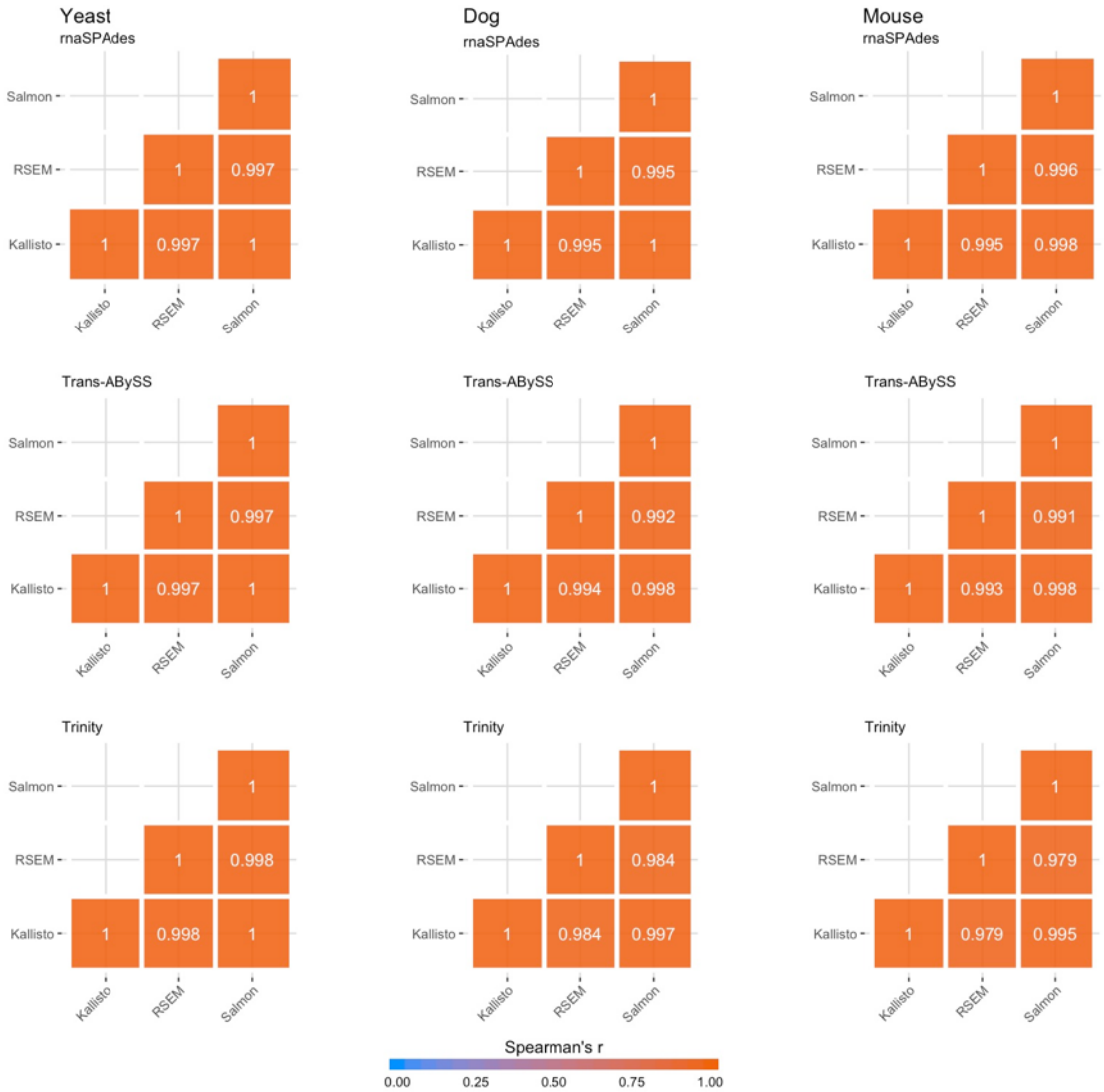


Fig S4-2: Spearman's Correlation Coefficient between Quantifiers (Simulated)

The correlation matrices illustrate the Spearman's correlation coefficient between the estimation made by any of the two quantifiers. The matrices in the left column are the results drawn from yeast dataset, in the middle column are the result from dog dataset and in the right column from mouse dataset. The matrices in the first row are the estimation based on rnaSPAdes assembly, the second row on Trans-ABySS assembly, and the third row on Trinity. In general, the correlation matrices show high consistency for the estimation made by quantifiers on simulated datasets.

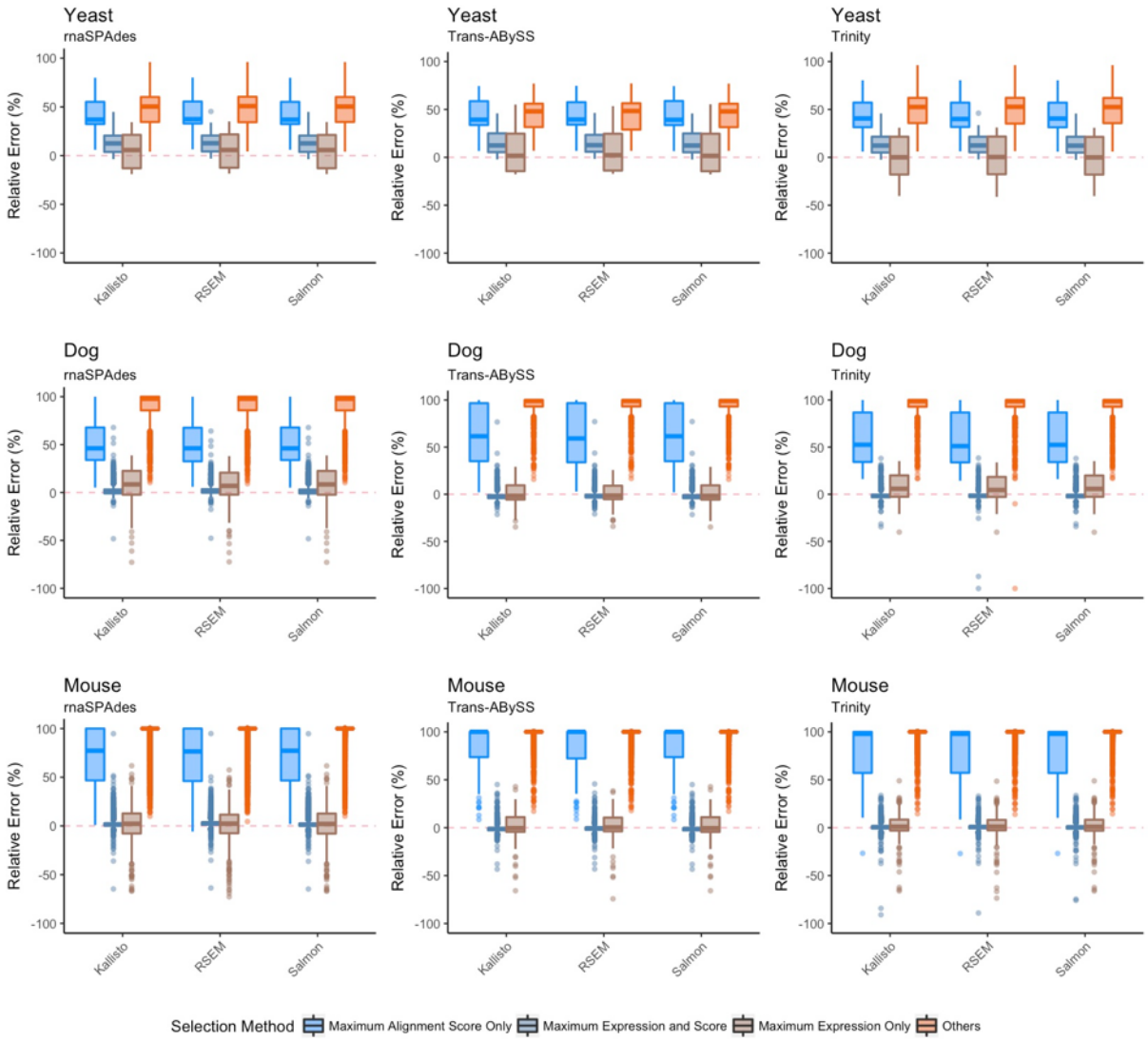


Fig. S5-1: Box Plots for the Relative Error of Unique Contigs (Simulated)

The box plots illustrate the relative quantification error for family-collapse contigs of simulated datasets. Since there are multiple transcripts correspond to one contig, we categorize the expression of corresponding transcript into (1) transcript with the maximum alignment score (but the expression is not the highest), (2) transcript with the highest expression (but the alignment score is not the highest), (3) transcript with the maximum alignment score and also yield the highest expression and (4) others. The box plots in the left column are based on the rnaSPAdes assembly, the second column on Trans-ABySS assembly, and the third column on Trinity. The first row is the result from yeast dataset, while the second and third row depict the that of dog and mouse respectively. Overall, the box plots suggest that the estimation made on family-collapse contigs is closest to the transcript with maximum alignment score.

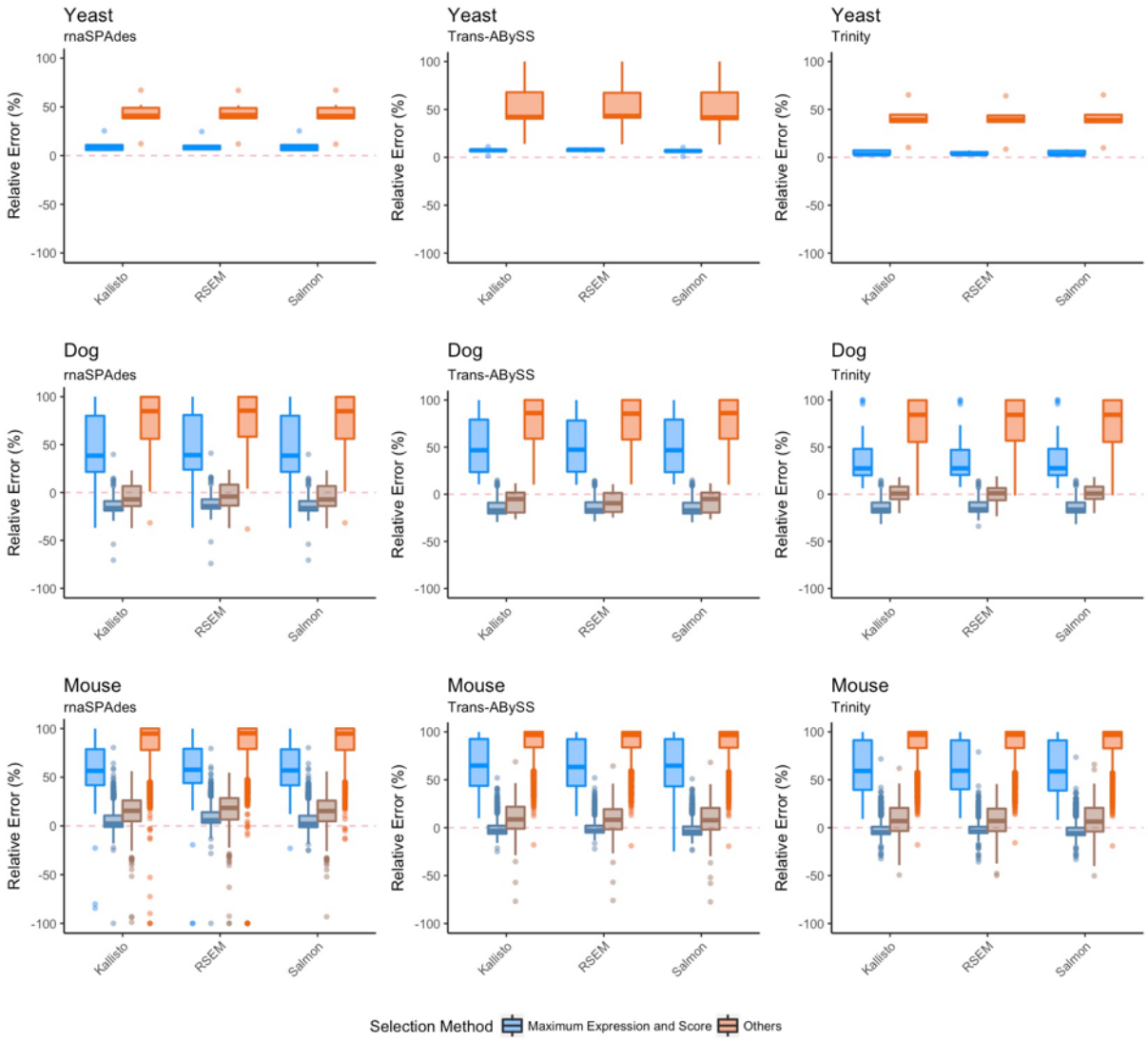


Fig. S5-2: Box Plots for the Relative Error of Unique Contigs (Experimental)
The box plots illustrate the relative quantification error for family-collapse contigs of simulated datasets. Since there are multiple transcripts correspond to one contig, we categorize the expression of corresponding transcript into (1) transcript with the maximum alignment score (but the expression is not the highest), (2) transcript with the highest expression (but the alignment score is not the highest), (3) transcript with the maximum alignment score and also yield the highest expression and (4) others. The box plots in the left column are based on the rnaSPAdes assembly, the second column on Trans-ABySS assembly, and the third column on Trinity. The first row is the result from yeast dataset, while the second and third row depict the that of dog and mouse respectively. Overall, the box plots suggest that the estimation made on family-collapse contigs is closest to the transcript with maximum alignment score.

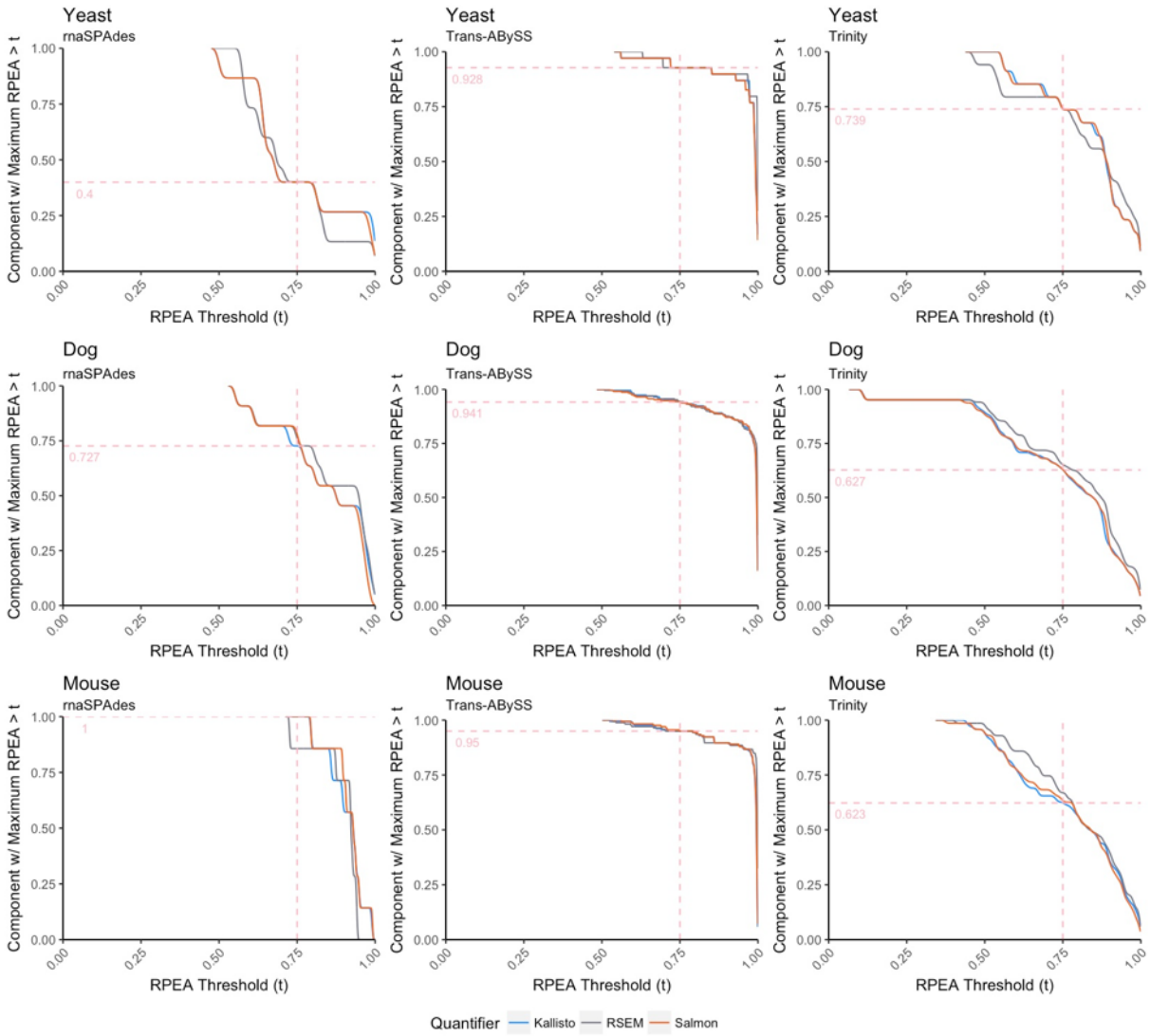


Fig. S6-1: Proportion of the Duplicated Connected Component with Highest Read Proportion (Simulated)

The line graphs illustrate the proportion of connected components of duplicated contigs. The X-axis is the threshold of RPEA (t) while the Y-axis is the proportion of connected component with the maximum RPEA > t. By this mean, we are allowed to examine how quantifiers allocate the RNA reads for duplicated contigs. Based on our result, most of the connected components have at least one contig that contribute over 0.75 of its expression, which suggest that the quantifiers tend to allocate most of the RNA reads to a single contig in the connected component instead of distributing evenly.

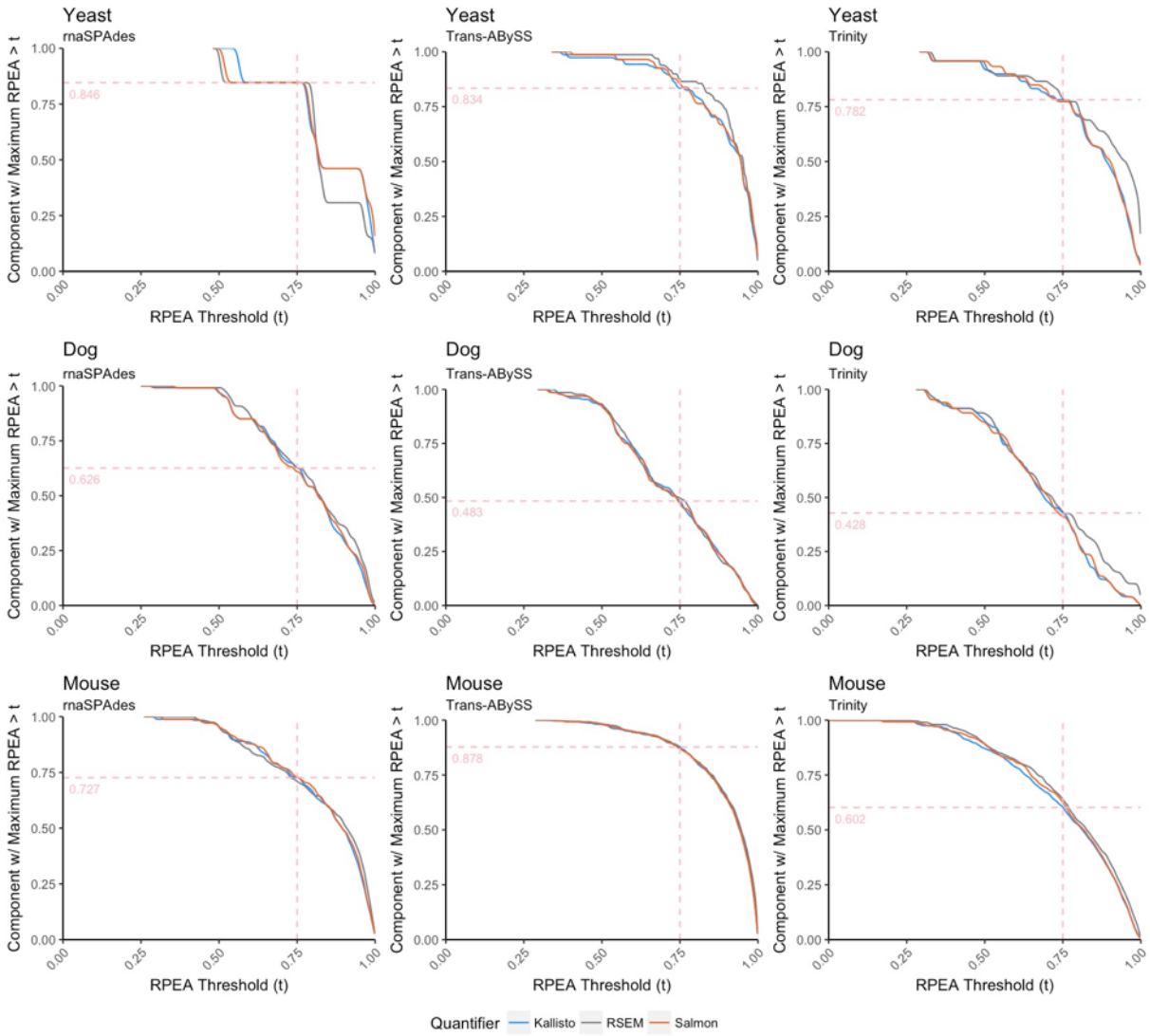


Fig. S6-2: Proportion of the Duplicated Connected Component with Highest Read Proportion (Experimental)

The line graphs illustrate the proportion of connected components of duplicated contigs. The X-axis is the threshold of RPEA (t) while the Y-axis is the proportion of connected component with the maximum RPEA > t. By this mean, we are allowed to examine how quantifiers allocate the RNA reads for duplicated contigs. Based on our result, most of the connected components have at least one contig that contribute over 0.75 of its expression, which suggest that the quantifiers tend to allocate most of the RNA reads to a single contig in the connected component instead of distributing evenly.