The battle for reads: evaluating strategies to tackle multi-mapping in RNA-seq quantification in highly repetitive genomes

ADDITIONAL FILE 2: SUPPLEMENTARY FIGURES

Aldana A. Cepeda Dean[1,2], Carlos A. Buscaglia[1,2], Virginia Balouz[1,2], Natalia Rego[3,5*], Luisa Berná[3,4,5*]

[1]Instituto de Investigaciones Biotecnológicas (IIBio), Universidad Nacional de San Martín (UNSAM), and Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires, Argentina.
[2]Escuela de Bio y Nanotecnologías (EByN), UNSAM, Buenos Aires, Argentina.
[3]Bioinformatic Unit, Institut Pasteur de Montevideo, Uruguay.
[4]Laboratory of Apicomplexan Biology. Instutut Pasteur de Montevideo, Uruguay
[5]Laboratorio de Genómica Evolutiva, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay.

[*]Both authors contributed equally to this work and should therefore be considered co-senior authors.

Address correspondence to: Luisa Berná (lberna@pasteur.edu.uy) or Natalia Rego (nrego@pasteur.edu.uy), Institut Pasteur de Montevideo, Mataojo 2020, 11400 Montevideo, Departamento de Montevideo, Uruguay.

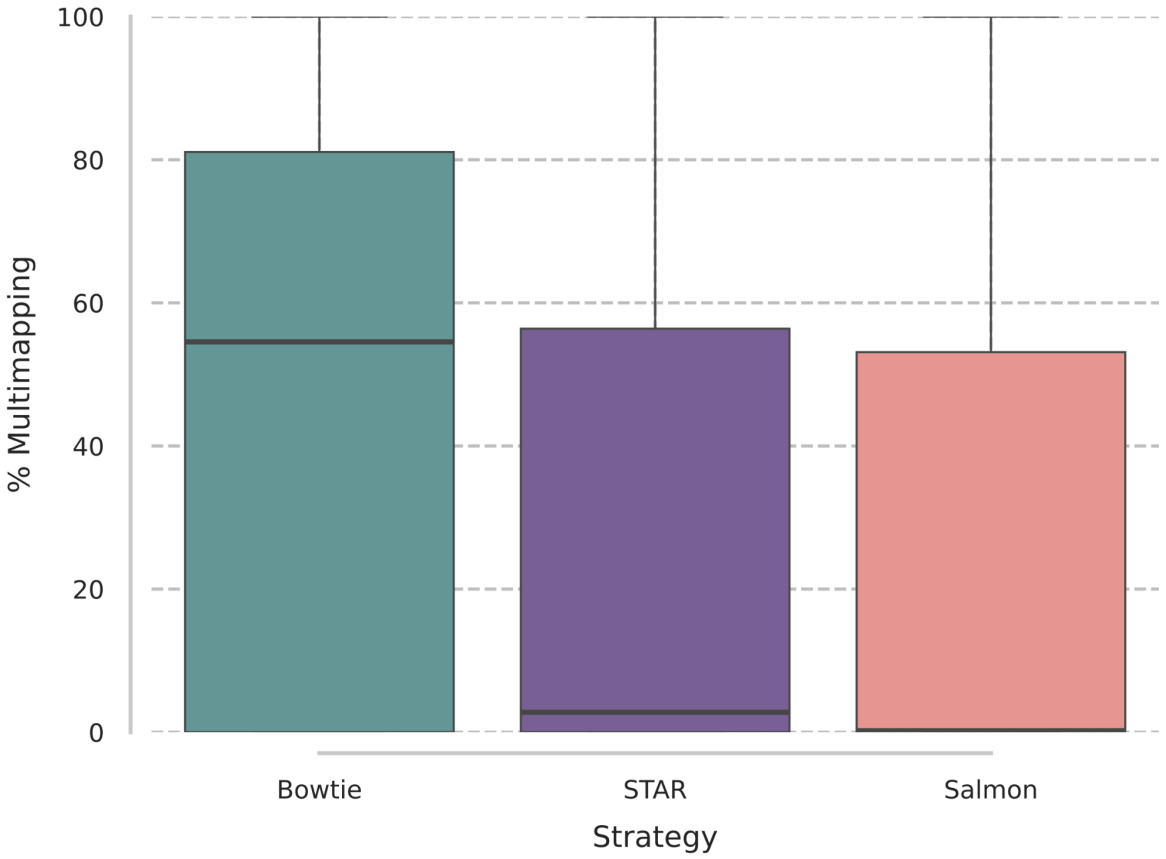

Figure S1. Mean percentage of multi-mapping in *T. cruzi* reads by mapping strategy. Bar plot showing the average percentage of multi-mapping reads for each mapping strategy. Error bars represent the standard deviation, reflecting variability among samples within each group.

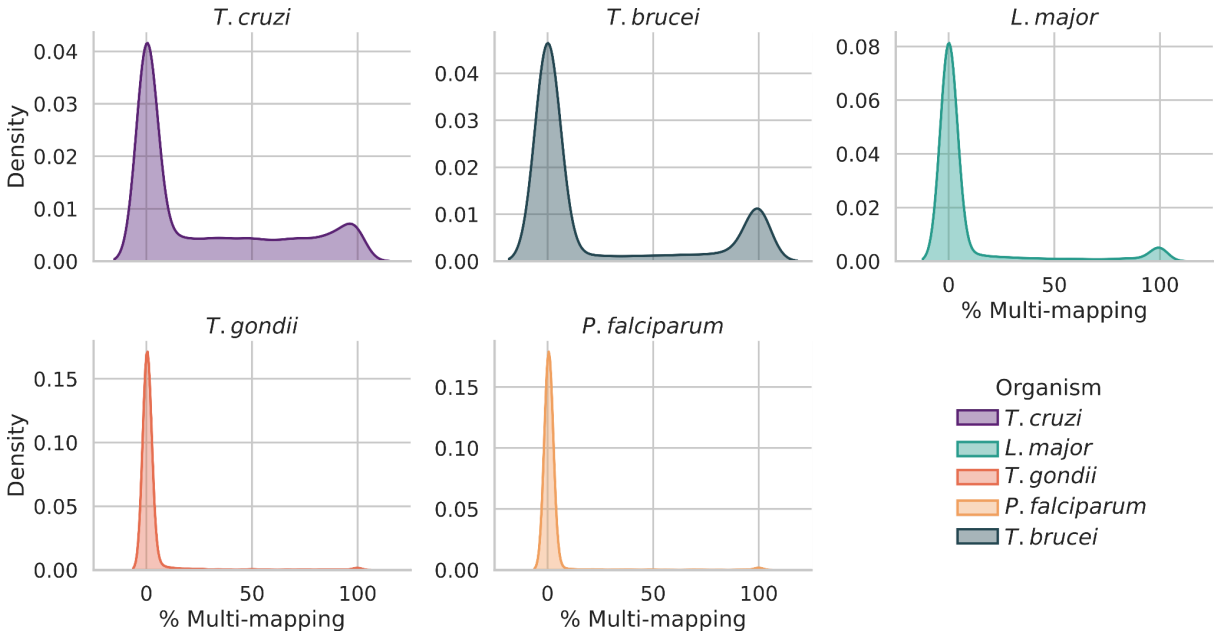

Figure S2. Density distribution of multi-mapping reads in protozoan parasites. Density distribution of the percentage of multi-mapping reads in the indicated organisms, as assessed by STAR.

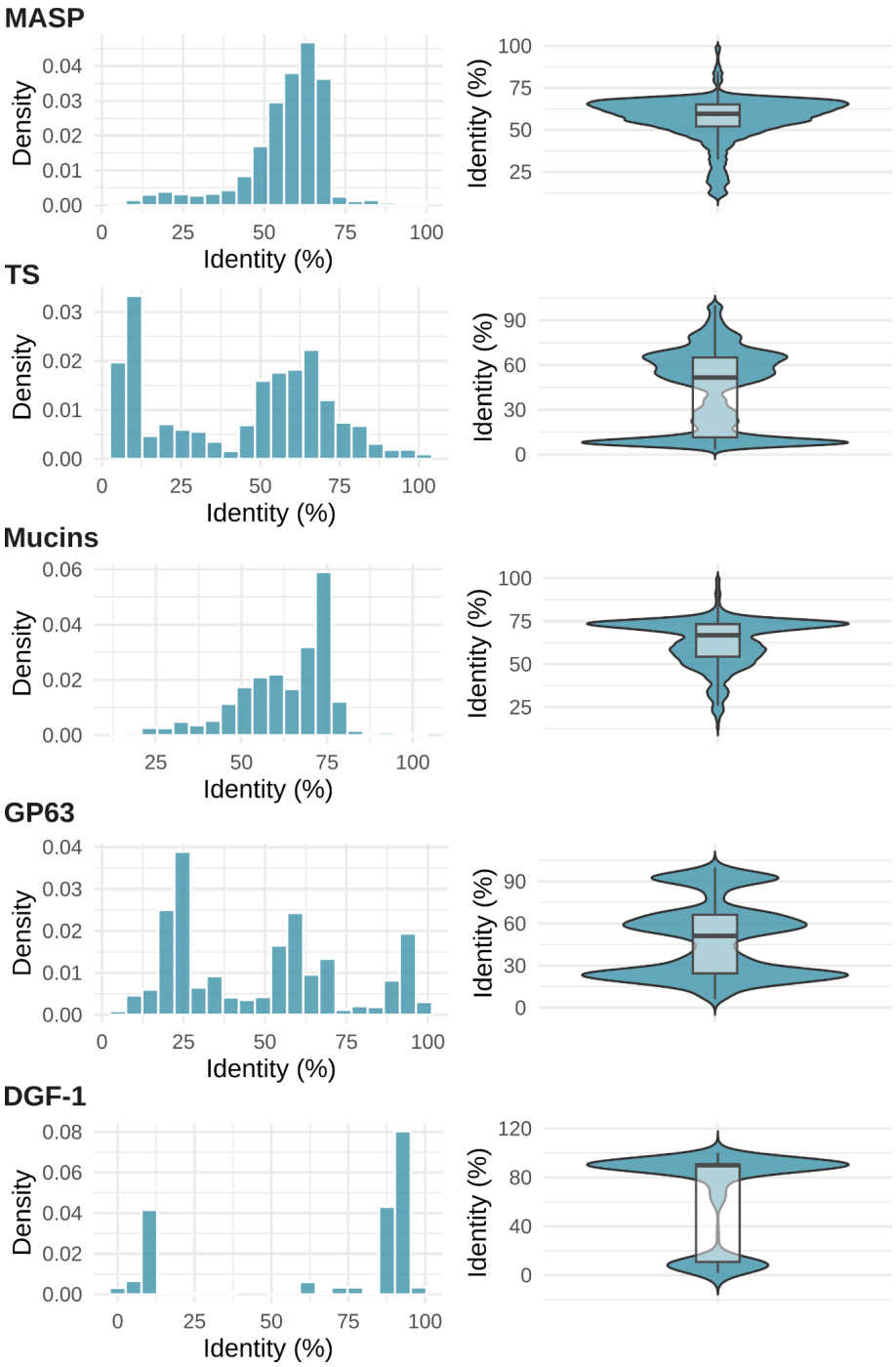

Figure S3. Distribution of pairwise sequence identity among members of selected *T. cruzi* multigene families. (A) Histograms show the density distribution of pairwise sequence identity within each gene family (MASP, TS, Mucins, GP63, and DGF-1). (B) Violin plots represent the distribution and variability of pairwise identity values, highlighting with boxplots the median, interquartile range, and density across the different families.

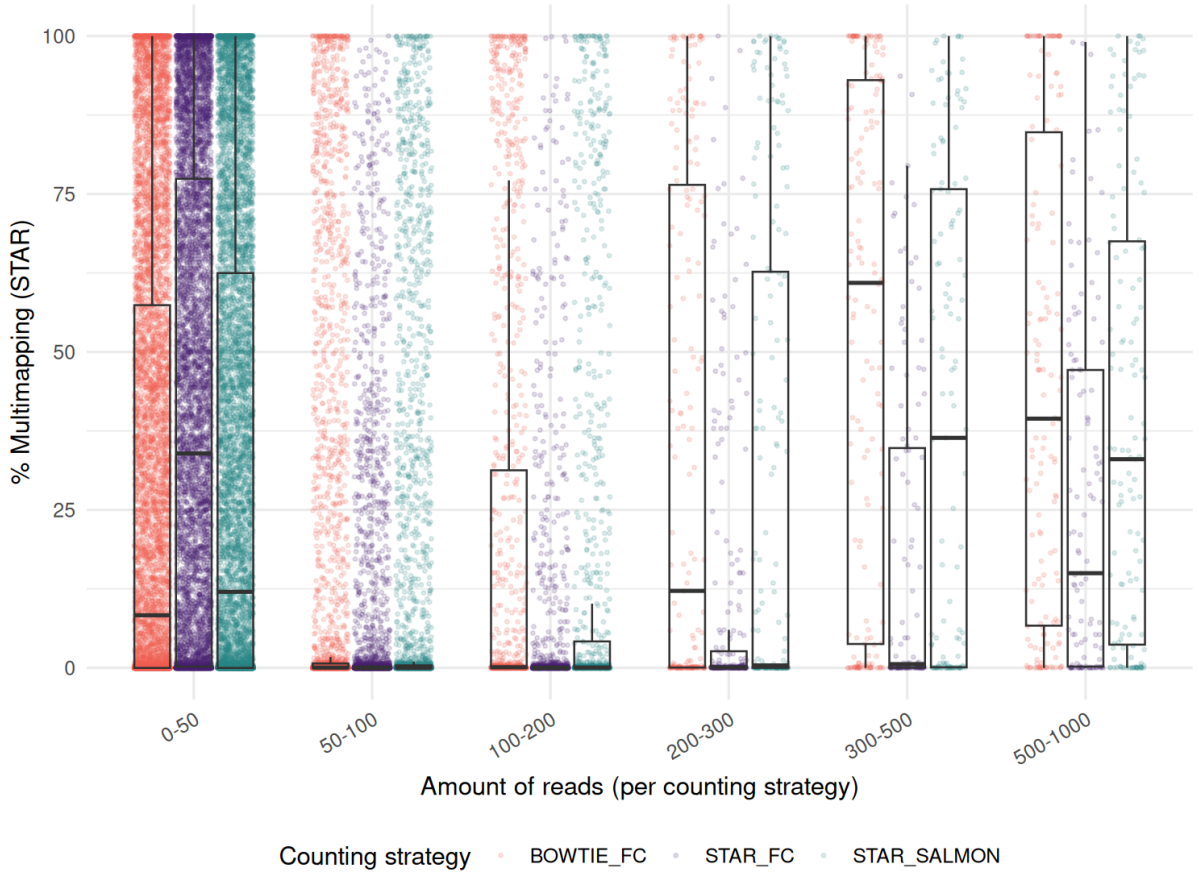

Figure S4: Multi-mapping rates across expression bins for different quantification strategies. *T. cruzi* genes were stratified into expression bins, and the percentage of multi-mapped reads was compared across the three alignment/quantification methods. Boxplots show the distribution within each bin, with individual genes represented as points. No clear association between expression level and multi-mapping was observed, although lowly expressed genes displayed higher variability, consistent with stochastic noise.

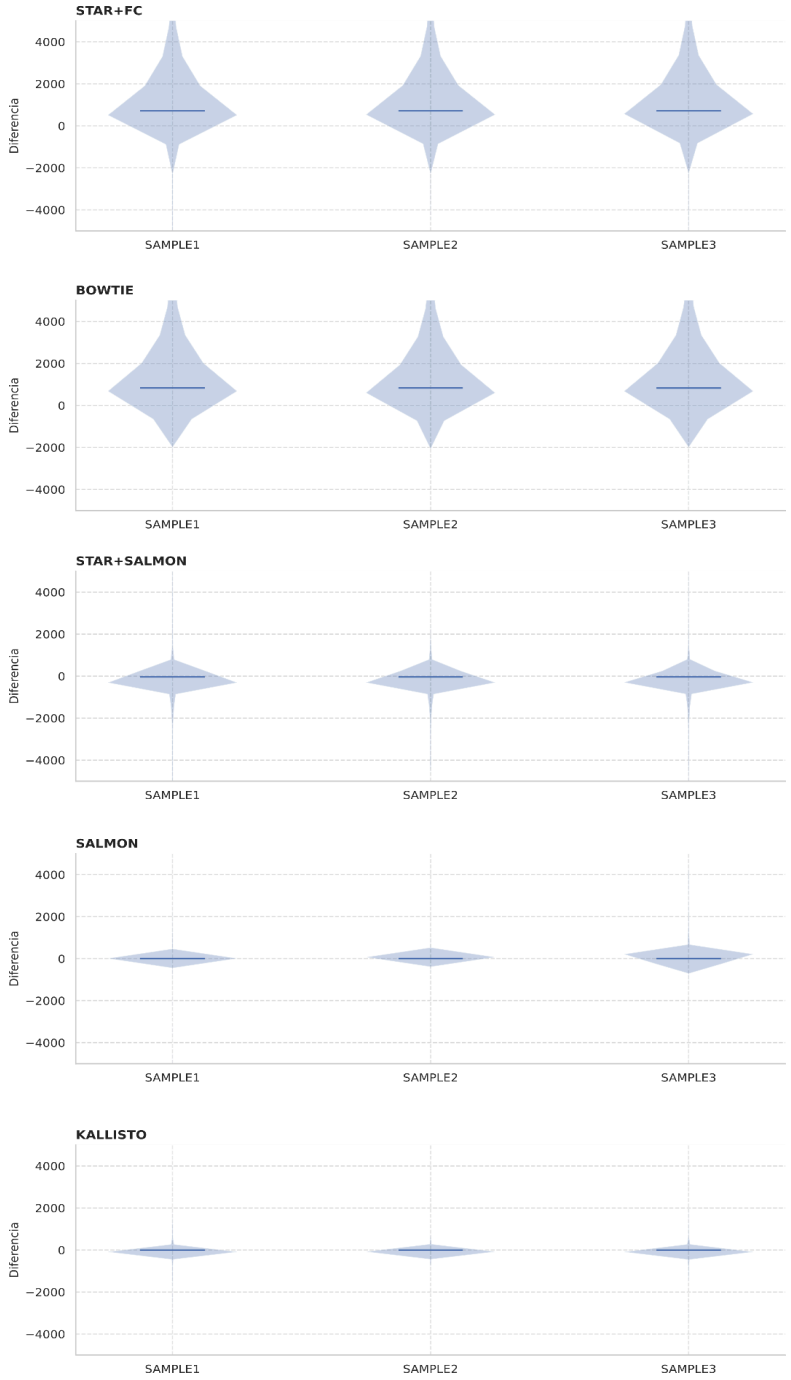

Figure S5: Distribution of differences between observed and simulated counts across quantification methods. Violin plots showing the distribution of the differences between simulated and observed counts across three replicate samples for five quantification methods: STAR+FC, BOWTIE2, STAR+Salmon, Salmon, and KALLISTO. Each panel corresponds to one method. The y-axis represents the difference between the observed and simulated counts (Observed − Simulated) for each gene. The width of the violin indicates the density of values at each difference level. Median values are shown as horizontal blue lines within each violin.

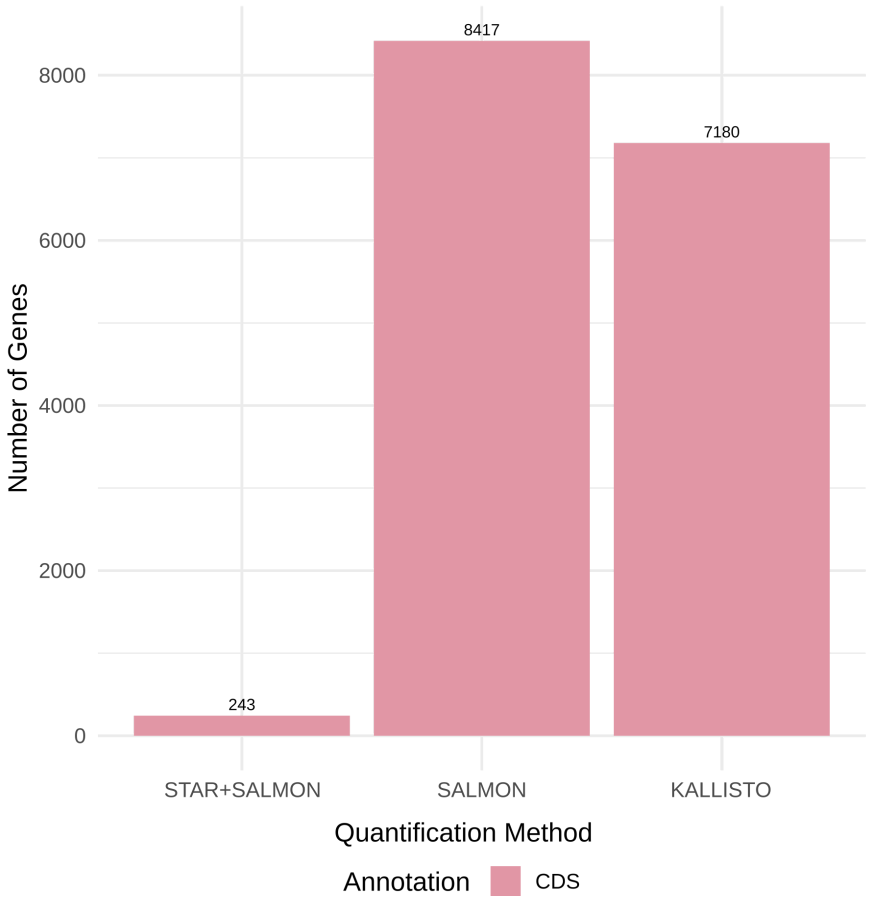

Figure S6. Number of genes with identical read counts between observed and simulated quantification. Bar chart indicating the number of genes where quantified counts match the simulated counts within a tolerance of 0.01, for the three selected strategies.

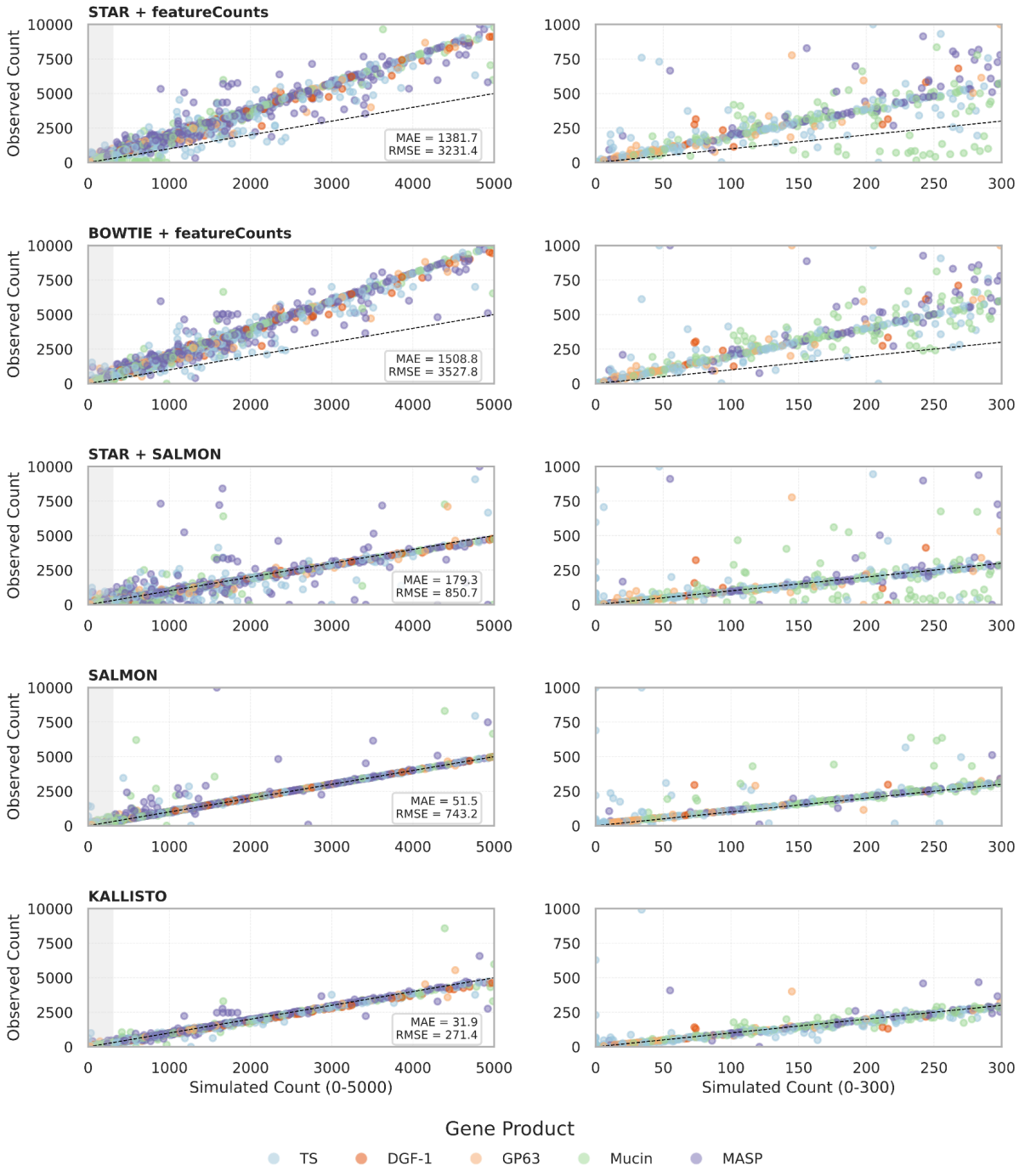

Figure S7: Performance of the quantification strategies in the simulated dataset across selected multigene families. Comparison of gene counts between simulated and observed values across all gene families using different approaches. The left column shows the range of 0-5,000 simulated counts, while the right column focuses on the range between 0 and 300 counts, which is highlighted with a gray box in the left panels. The black dashed line corresponds to the identity line.

Figure S8. Quantification accuracy on extremely similar gene paralogues: Comparison between simulated and estimated read counts for a subset (2 to 6) of genes of either TS, Gp63, Mucin or MASP family with >95% nucleotide identity. The y-axis represents read counts. The left panel shows a uniform simulation scenario where the baseline expressions of each gene were obtained by averaging gene-level counts from previous analysis. The right panel shows a differential simulation where only gene #1 was assigned 1,000 reads, while the remaining genes from this family retained their original expression levels. Gray bars indicate simulated values, while colored bars correspond to different quantification strategies (STAR + Salmon, Salmon, and Kallisto).

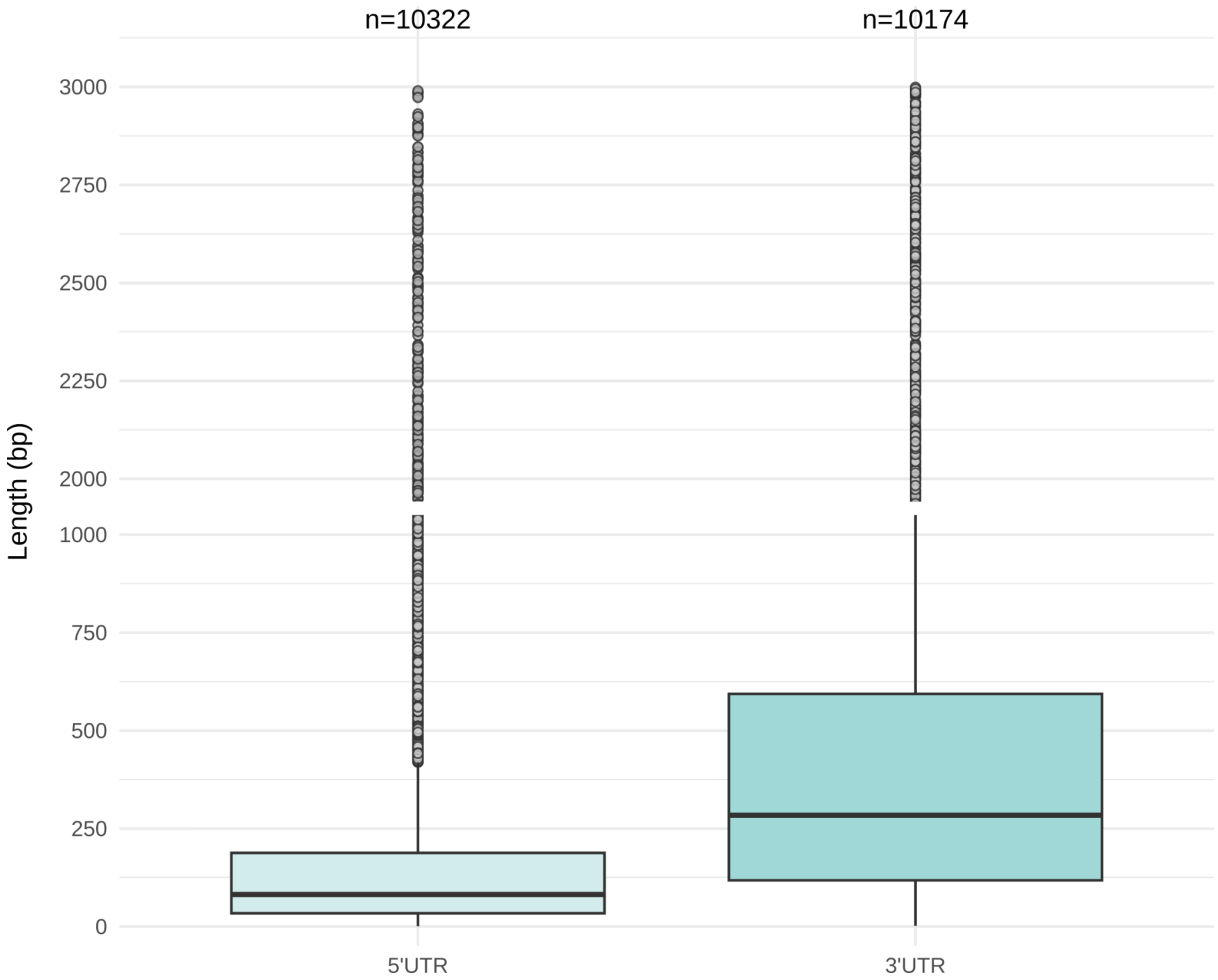

Figure S9: Structure of *T. cruzi* untranslated regions (UTRs) as predicted by UTRme. The box plots show the length distribution for each UTR, with the median (black line) and outliers (gray dots).

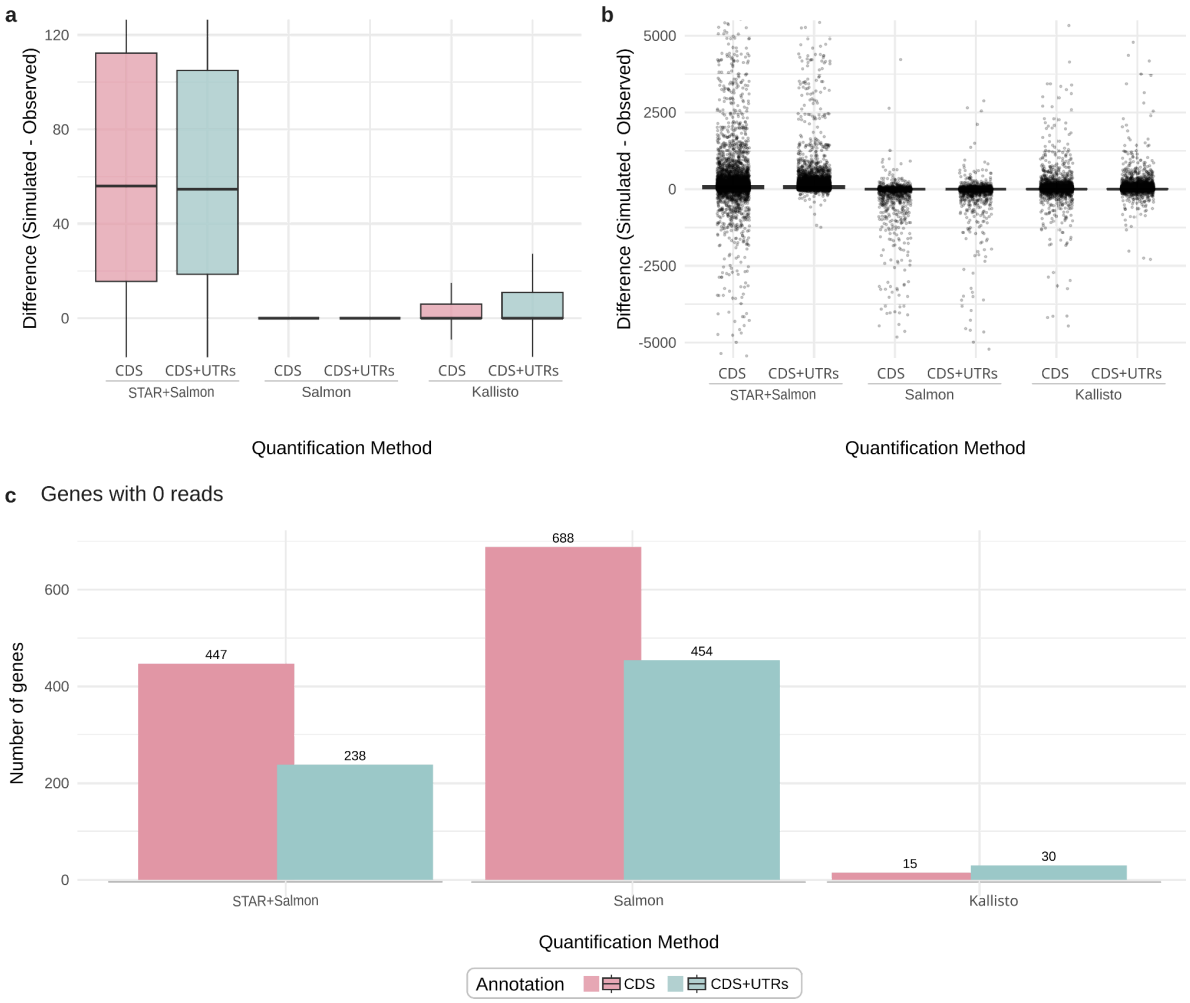

Figure S10: Comparison between results obtained with CDS and CDS+UTRs. (a) Boxplots showing differences between simulated and observed counts within the 0–120 range. (b) Boxplots showing differences within the -5,000 to 5,000 range, with individual points overlaid in black to illustrate data dispersion. (c) This bar chart displays the number of genes where the quantified read count is zero while the simulated count was different from zero. All panels include three quantification methods (STAR+SALMON, Salmon, and Kallisto); and includes two gene annotation schemes (CDS and CDS+UTRs)