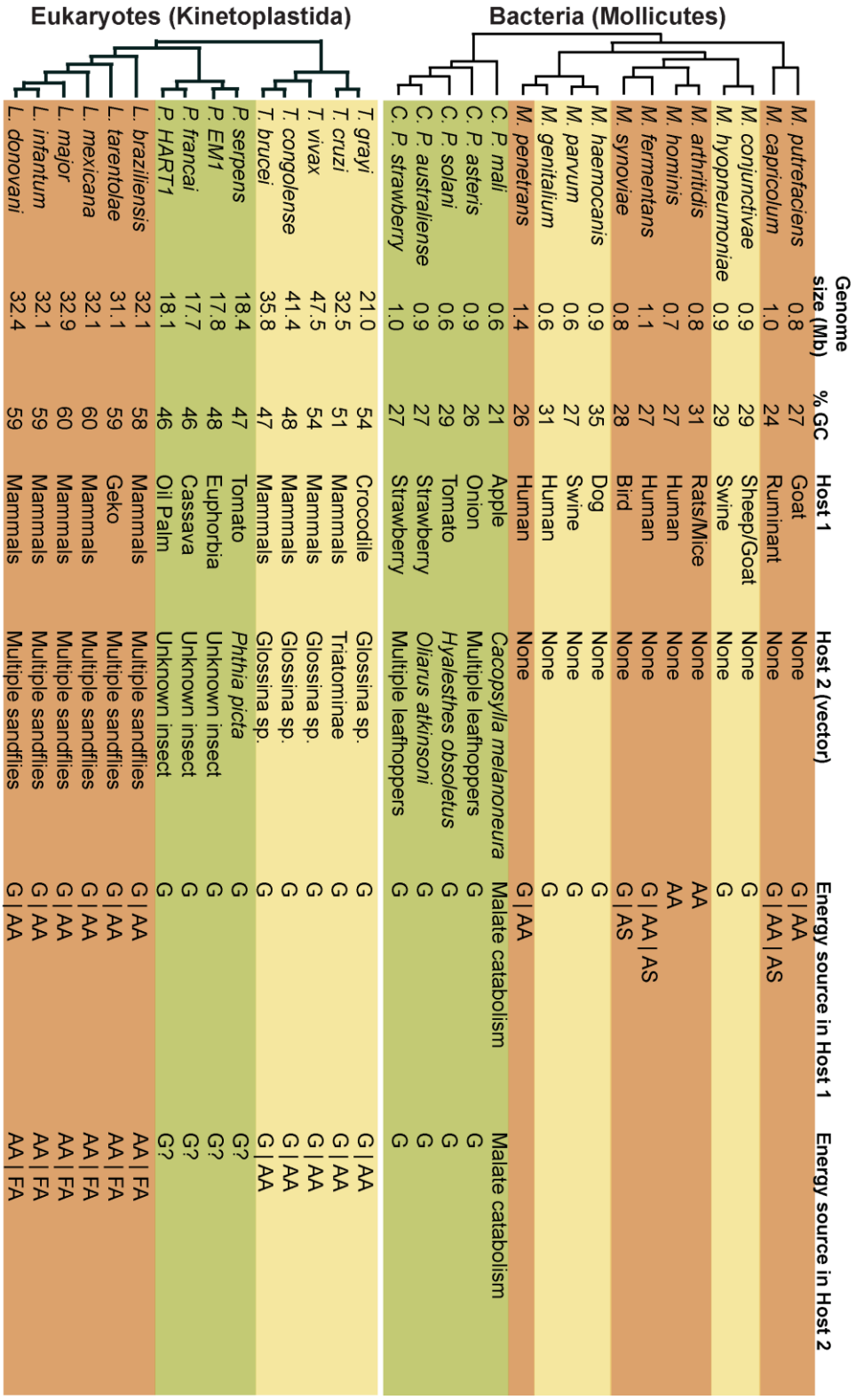
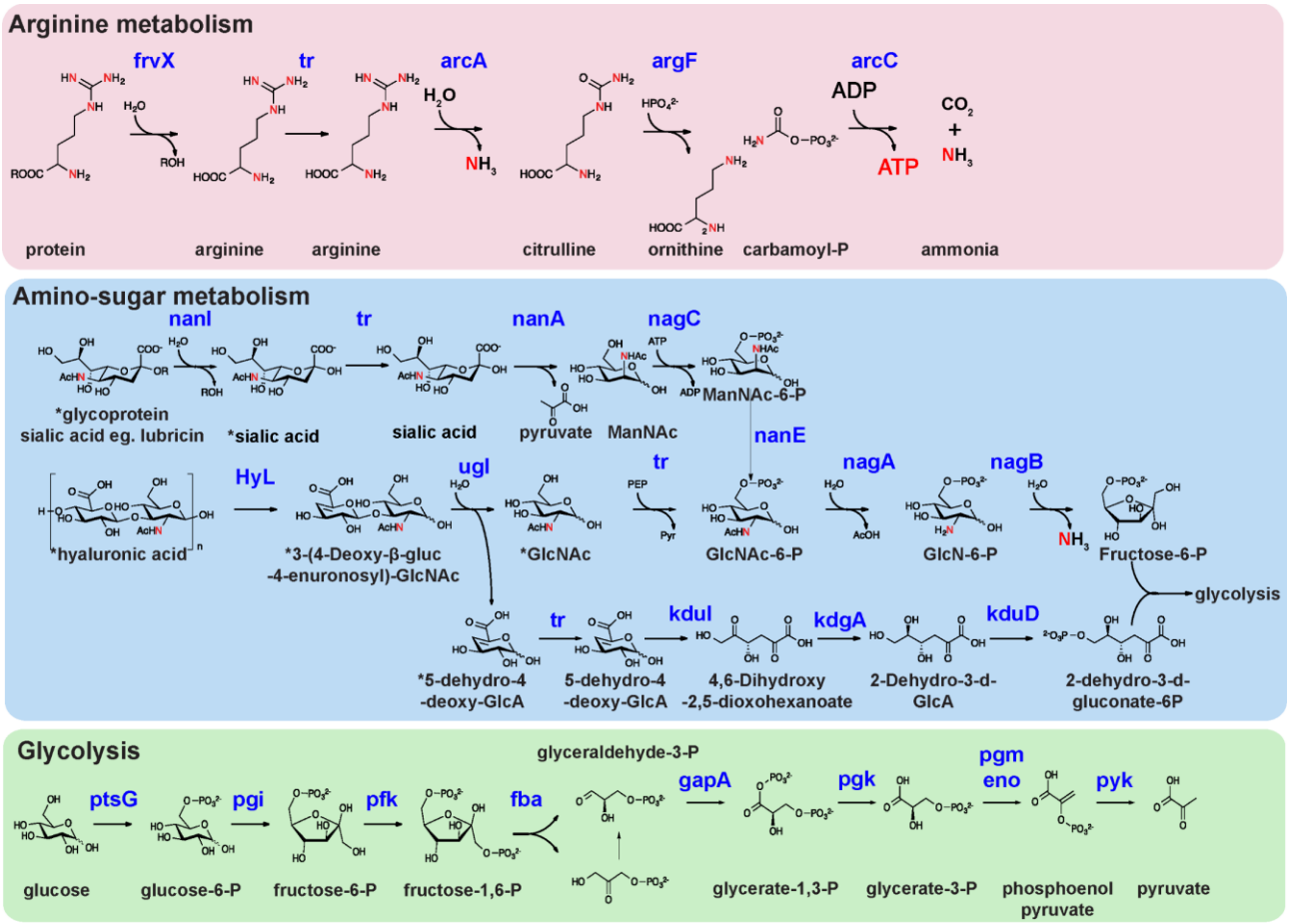
Supplemental Fig. S1.



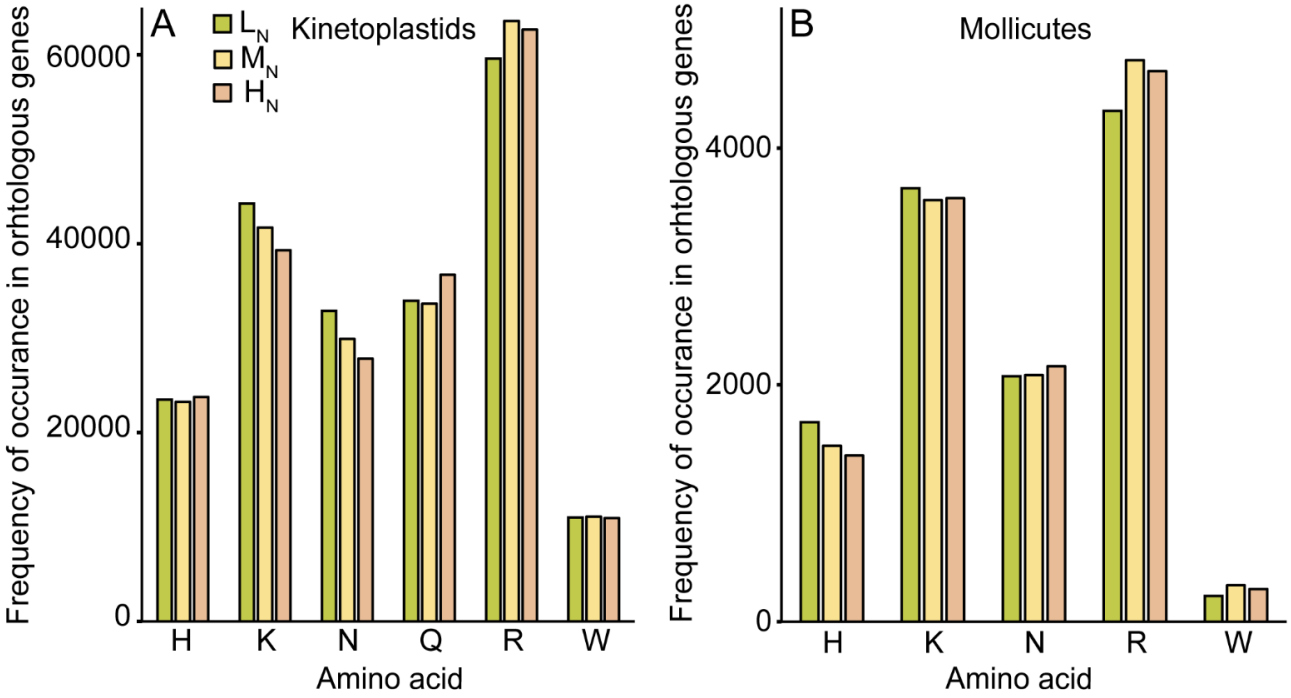
Phylogenetic trees and metabolic information for the parasites used in this study. Hosts and vectors are indicated where known. The major pathways used to generate ATP in each host are provided where G = glycolysis, AA = amino acid metabolism, AS = amino-sugar metabolism, FA = fatty acid metabolism. Genome size and GC content are also provided.

Supplemental Fig. S2.



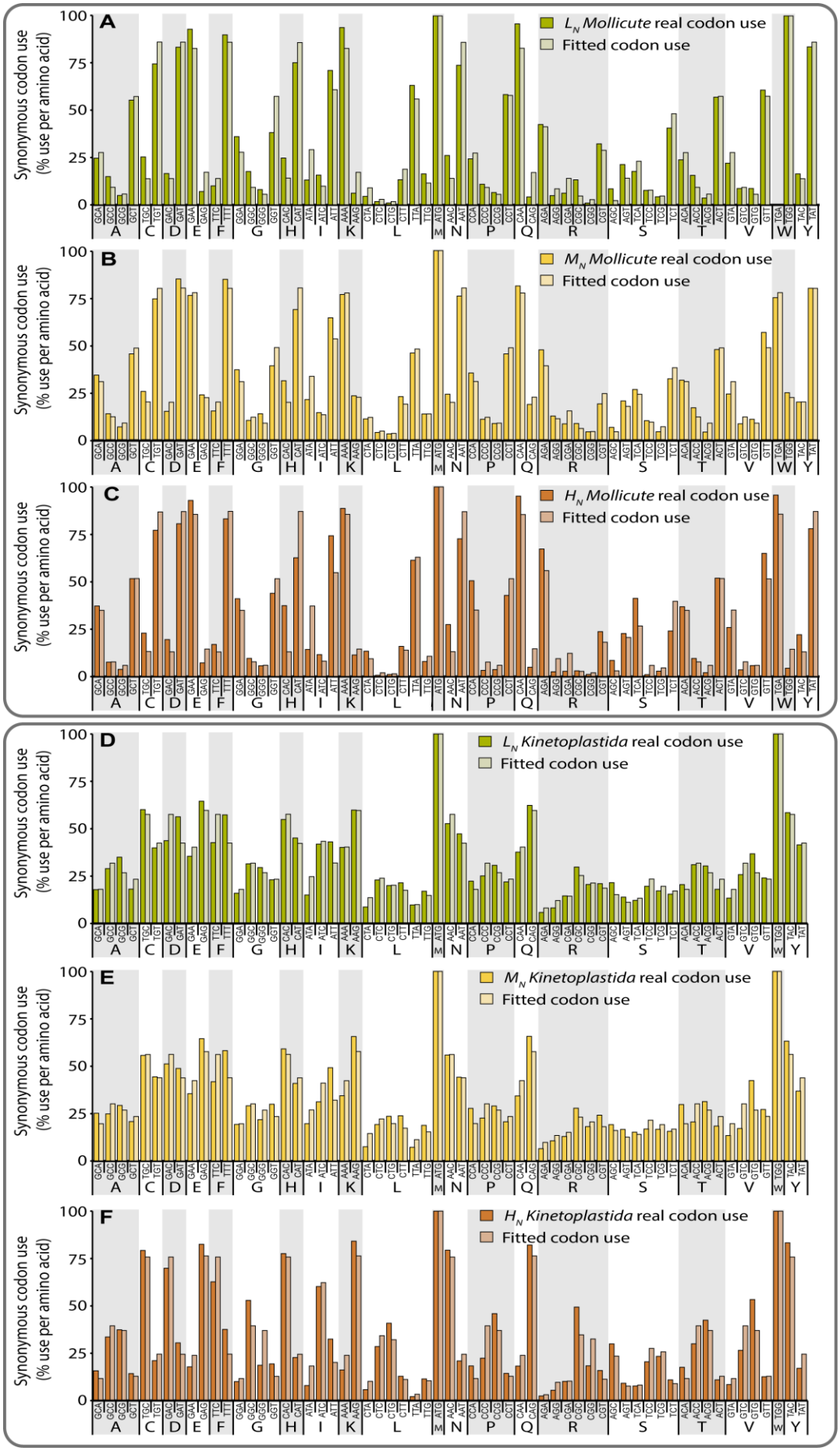
ATP generating metabolic pathways. Nitrogen atoms have been highlighted in red and required genes are indicated in blue. * denotes substrates that are extracellular. tr. is the abbreviation for transporter.

Supplemental Fig. S3.



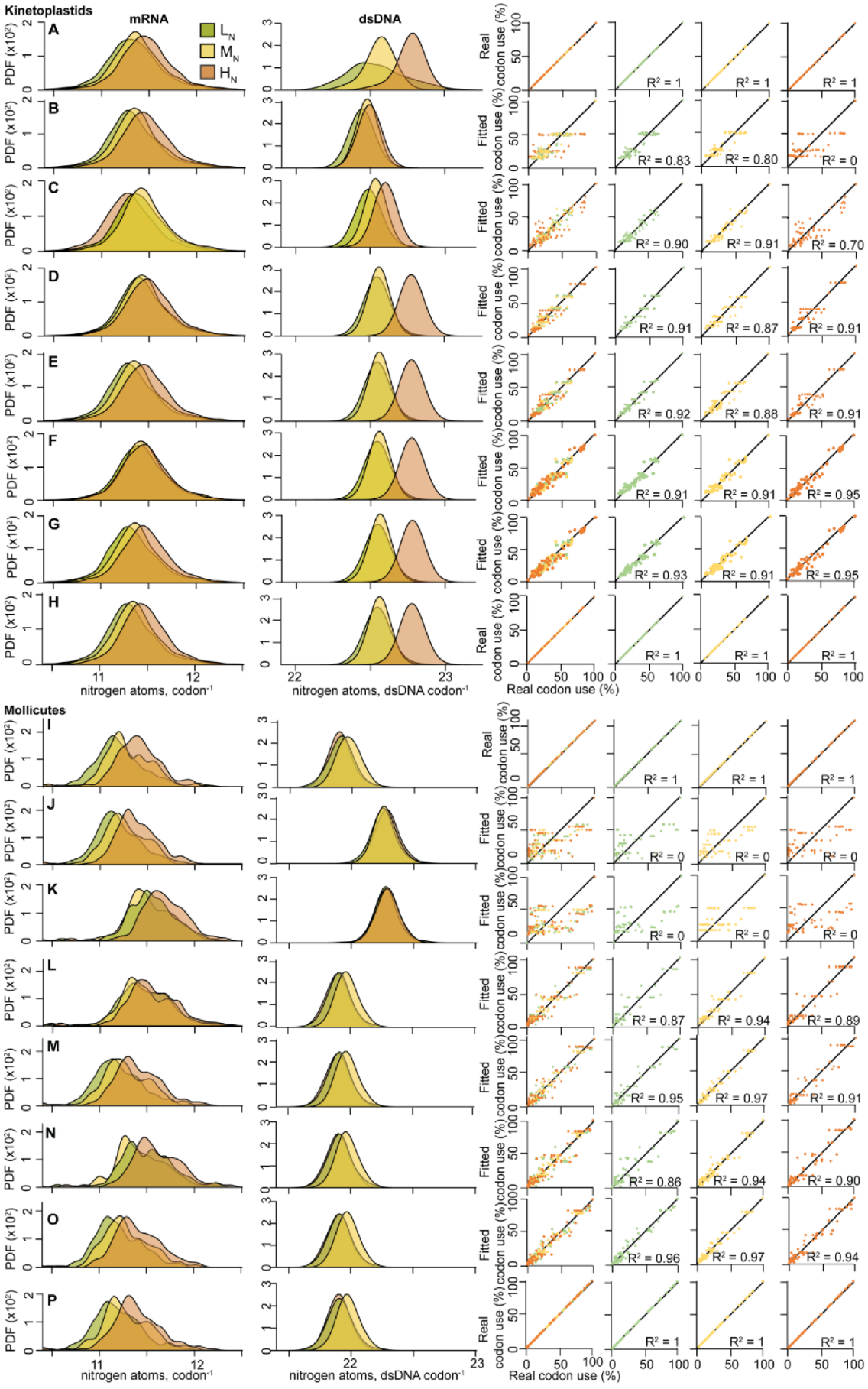
L_N parasites use the least amount of nitrogen in their amino acid side chains compared to M_N and H_N parasites. (*A*) Frequency of occurrence of amino acids with nitrogen in their side chains at orthologous sites in orthologous genes in the kinetoplastid parasites. This corresponds to a total number of nitrogen atoms in amino acids side chains of L_N = 347,789, M_N = 353,574 and H_N = 350,376. (*B*) Frequency of occurrence of amino acids with nitrogen in their side chains at orthologous sites in orthologous genes in the Mollicute parasites. This corresponds to a total number of nitrogen atoms in amino acids side chains of L_N = 11,948, M_N = 12,178 and H_N = 12,063. Glutamine (Q) is not considered for the Mollicutes as the M_N and H_N groups lack glutaminyl-tRNA synthetase (GlnS). Instead a non-discriminating glutamyl-tRNA synthetase (GltX) charges both $tRNA^{Glu}$ and $tRNA^{Gln}$ with Glu. This means use of Q between *Mycoplasma* (M_N and H_N species) and *Phytoplasma* (L_N species) is not comparable. For both the kinetoplastids and the Mollicutes, M_N parasites use more nitrogen in their side chains than H_N parasites. This can be explained by the reduced occurrence of arginine (R) in the H_N parasites, which is expected as these parasites metabolise arginine to generate energy. In both graphs the orthologous sites are the same as for the analysis in the section “Low nitrogen availability parasites have low nitrogen content sequences and vice-versa”.

Supplemental Fig. S4.



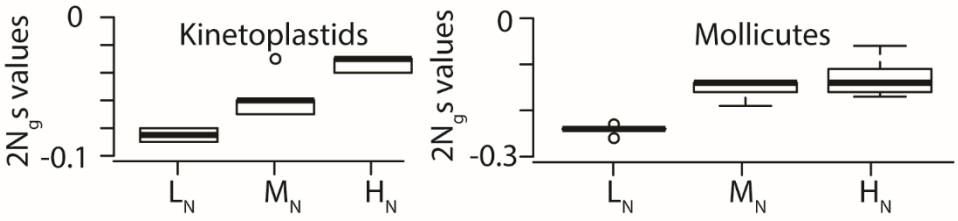
The model for synonymous codon use under the joint pressures of selection acting on nitrogen content and mutation bias fits real codon use with > 90% accuracy. Comparison of observed (dark) and fitted (light) synonymous codon use for (A) Mollicute L_N , $2N_g s$ = -0.24, *m* = 4.8 (B) Mollicute M_N, $2N_g s$ = -0.15, *m* = 3.5 (C) Mollicute H_N, $2N_g s$ = -0.13, *m* = 5.9 (D) Kinetoplastid L_N, $2N_g s$ = -0.09, *m* = 0.7 (E) Kinetoplastid M_N, $2N_g s$ = -0.06, *m* = 0.7 and (F) Kinetoplastid H_N $2N_g s$ = -0.03, *m* = 0.3 parasites.

Supplemental Fig. S5



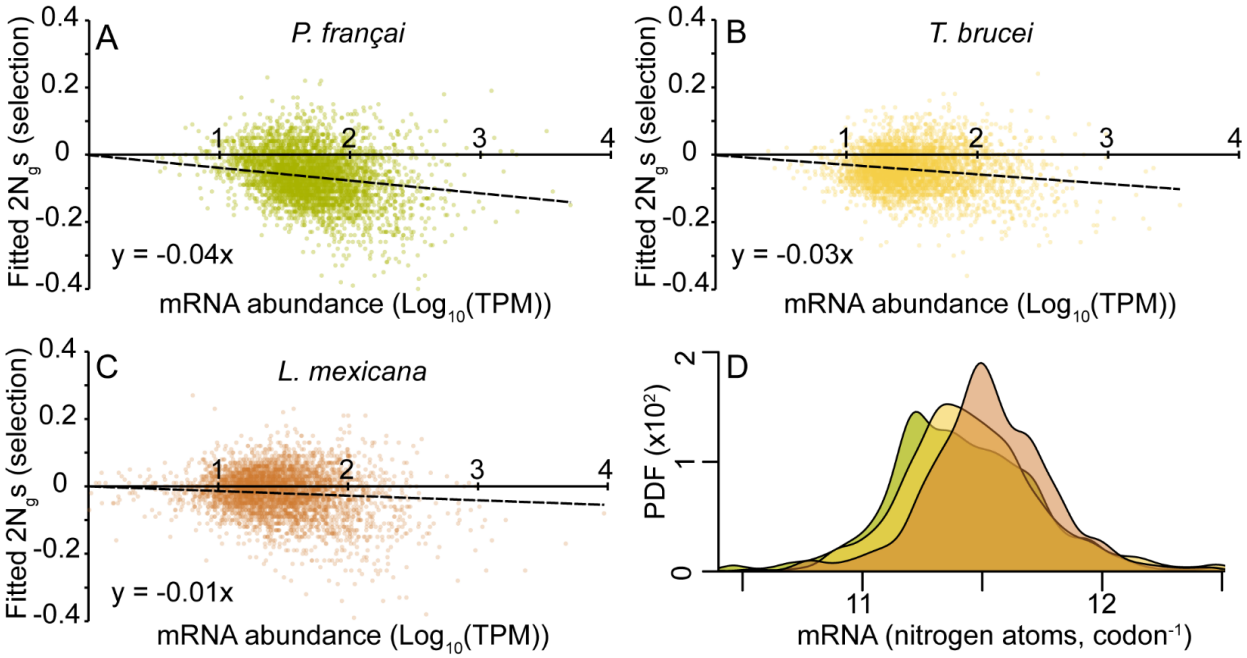
The model that considers both mutation bias and nitrogen content selection in combination provides a better fit than either parameter considered in isolation. (*A*) The average mRNA nitrogen content per codon for 3003 orthologous genes in the kinetoplastida. The average nitrogen content per double stranded codon for the same set of genes. Empirical codon use probabilities (expressed as %) plotted against themselves for all groups, L_N, M_N and H_N respectively. Analogous plots are shown in panels B-H for sequences simulated using fitted values for (*B*) selection acting on codon nitrogen content (Equation 2). (C) selection acting on translational efficiency (tAI) (Equation 10). (D) mutation bias acting on the GC content of the sequences (Equation 5). (E) both selection acting on codon nitrogen content and mutation bias (Equation 7) (F) both selection acting on translational efficiency and mutation bias (Product of equations 5 and 10). (G) all 3 parameters together i.e. Selection acting on codon nitrogen content, mutation bias and selection acting on translational efficiency (Equation 12) (H) empirical codon usage probabilities i.e. the 61 actual codon use frequencies. These simulated sequences produce symmetrical distributions with low variance that do not precisely recapitulate the data found in A. For the kinetoplastids (A-H) the best AIC values for L_N and M_N are 1992812 and 2011196 respectively for the 3 parameter model (G), for H_N it is 1756308 for the two parameter model that considers both translational efficiency and mutation bias. (*I – P)* As for plots A-H but for the Mollicutes. For the Mollicutes the best AIC values for L_N, M_N and H_N are 61440, 71126 and 59288 respectively for the 3 parameter model (O) .Y-axis is the probability density function (PDF) for the distributions.

Supplemental Fig. S6



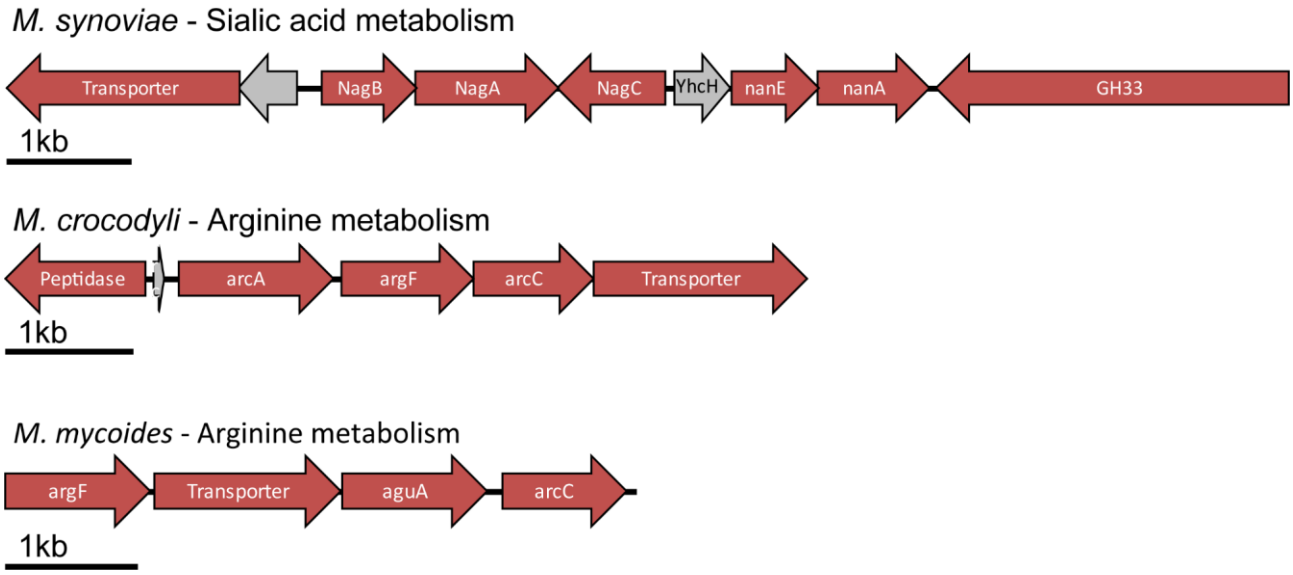
Boxplots showing distribution of $2N_gs$ values for individual species.

Supplemental Fig. S7



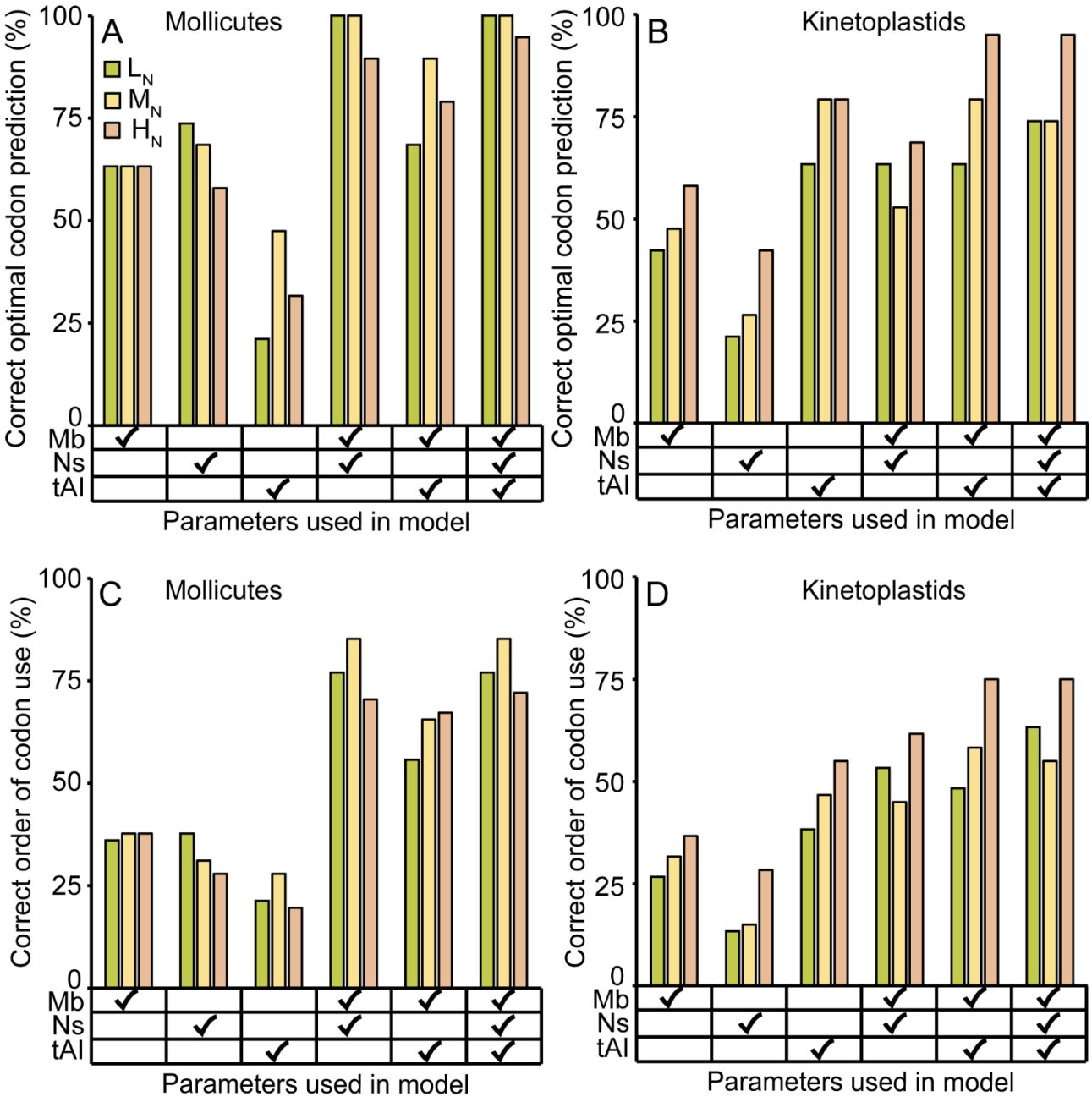
Gene expression negatively correlates with selection acting on mRNA nitrogen content. (*A-C*) Using the selection-mutation model, $2N_gs$ values were calculated for 4083 orthologous genes for one species from each of the L_N , M_N and H_N kinetoplastid groups (*P. français, T. brucei* and *L. mexicana* respectively). Mutation bias values were fixed as constant at the overall value for each species (L_N = 0.61, M_N = 0.92 H_N = 0.34). These values were plotted against mRNA expression data from equivalent lifecycle stages (procyclic) and data points were set to an opacity value of 50% to help judge density. Linear regressions were fit to the data for each species. The slope of the L_N line was the most negative, M_N was intermediate and H_N was the closest to 0. This is consistent with our other results and shows that selection to minimise nitrogen in mRNA is strongest for the species that are the most nitrogen limited. ie a gene with a TPM value of 100 would be predicted to have a $2N_gs$ value of L_N = -0.08, M_N = -0.06 and H_N = -0.02. Two-tailed t-tests comparing the slopes showed that all of them were significantly different from one another ($p < 0.05$). R^2 = 0.07, 0.01 and 0.02 respectively. (*D*) Comparison of the nitrogen use of the 4083 orthologous genes found in all of all of *P. français* (green), *T. brucei* (yellow) and *L. mexicana* (orange) taking expression into account. ie. If a gene had a TPM value of 10, it is represented in the distribution 10 times. All distributions are significantly different using a Wilcoxon Signed-Ranks ($p < 0.05$).

Supplemental Fig. S8



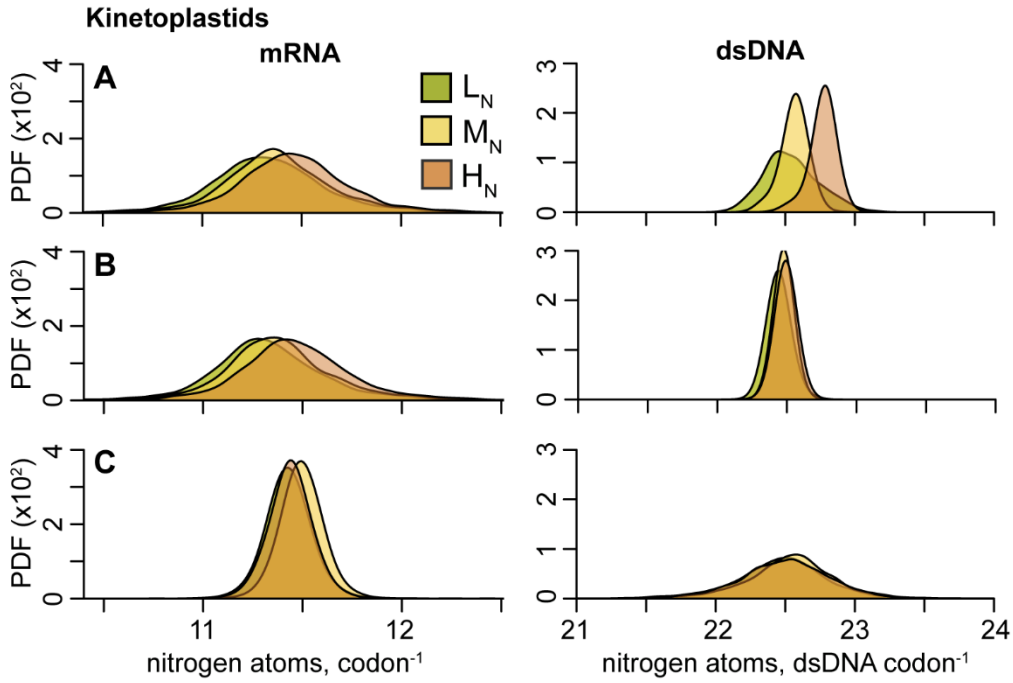
Gene clusters for nitrogen liberating metabolic pathways in H_N Mollicute parasites.

Supplemental Fig. S9



The model parameters which provide the best fit between observed and predicted codon use also provide the best percentage of correctly predicted optimal codons and correctly ordered codon use. (A) Percentage of correctly predicted optimal codons for Mollicute parasites using different parameter combinations. (B) As in (A) but for kinetoplastid parasites. (C) Percentage of correctly ordered relative codon use for different parameter combinations. (D) As in (C) but for kinetoplastid parasites. Mb = Mutation bias, Ns = selection acting on nitrogen content, tAI = selection acting on translational efficiency.

Supplemental Fig. S10



Example distribution of the model when run with shuffled codon nitrogen content. (*A*) Left, the average mRNA nitrogen content per codon for 3003 orthologous genes in the kinetoplastida (observed data). Right, the average nitrogen content per double stranded codon for the same set of genes. Analogous plots are shown in panels B and C for sequences simulated using fitted values for (*B*) selection acting on codon nitrogen content (Equation 2). (C) selection acting on codon nitrogen content where the nitrogen content of the codons has been shuffled (as described in the methods). Shuffled codon nitrogen cost (C) provides distributions that do not fit observed data (A) as well as distributions that use real nitrogen content (B). $2N_gs$ values for the shuffled example were $L_N = 0.01$, $M_N = 0.02$ and $H_N = 0$ compared to -0.06, -0.04 and 0.03 for the real nitrogen content. Thus when codon nitrogen content is shuffled, the model is unable to recapitulate the observed distribution of nitrogen content in gene sequences and the influence of the model parameter is reduced towards zero.