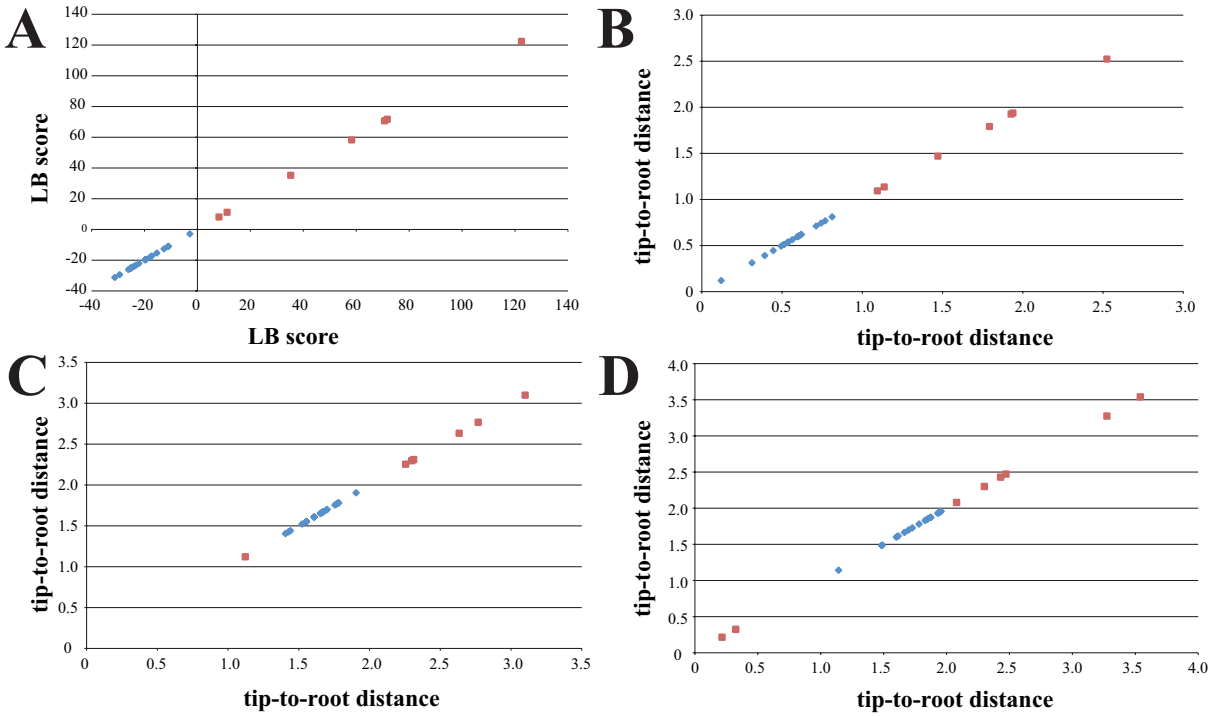

Figure S1. Plot of LB scores or tip-to-root distances for the ML tree of gene 111427 against themselves. (A) LB scores for the three different rooting possibilities. (B) Tip-to-root distance for the ML tree rooted with *Apis*. (C) Tip-to-root distance for the ML tree rooted with *Echinoderes*. (D) Tip-to-root distance for the ML tree rooted with *Priapulus*. Red dots = species with long branches (LB score > 0 or tip-to-root distance in B above 1. Blue dots = all other species.

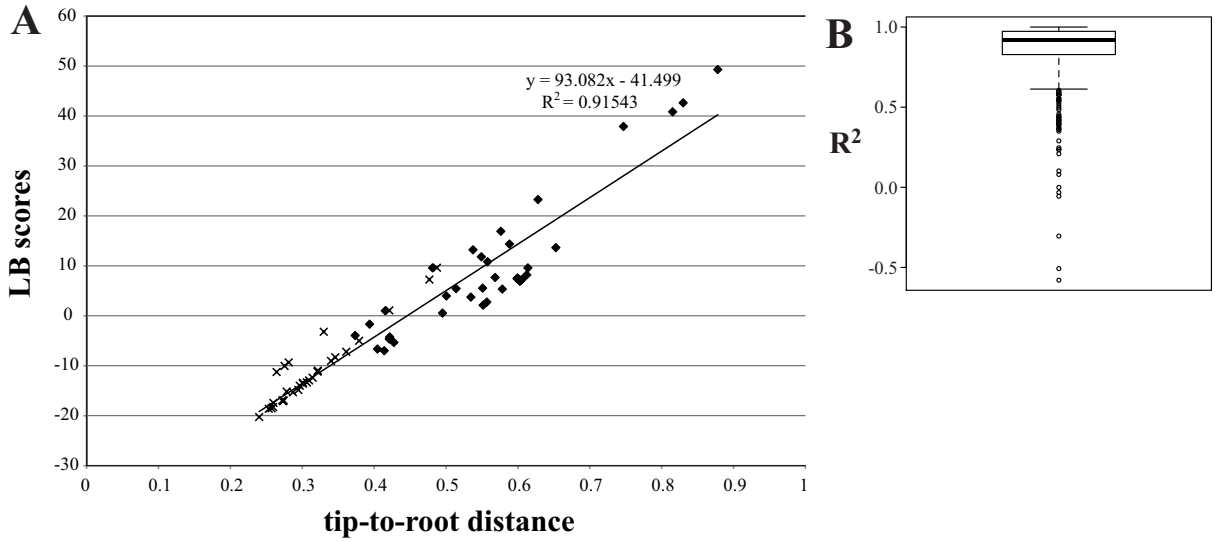

Figure S2. Correlation of LB scores to tip-to-root distances for the ML trees rooted with *Apis*. (A) Plot values against each other for dataset d01. Correlation coefficient is provided. Diamonds = platyzoan species, crosses = other species. (B) Box plot of the correlation coefficients of the analyses based on the 559 individual genes.

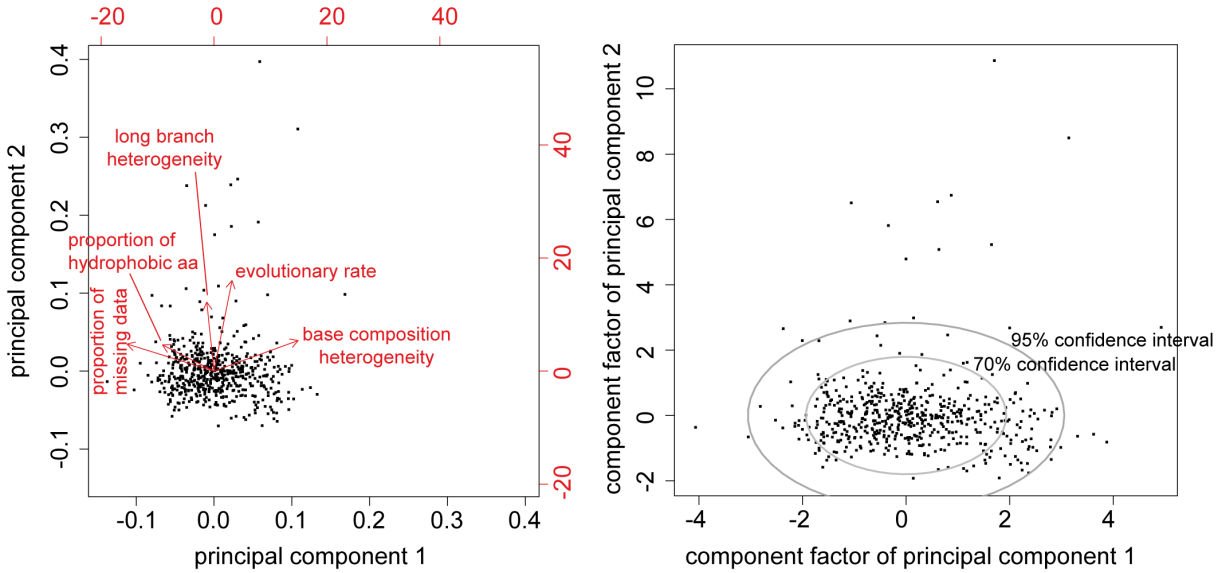

Figure S3. Results for the principal component analyses as biplot of the two first principal components including the eigenvectors of each gene property as well as a plot the components factors of the first two principal components including 95% (2σ) and 70% (1σ) confidence intervals.

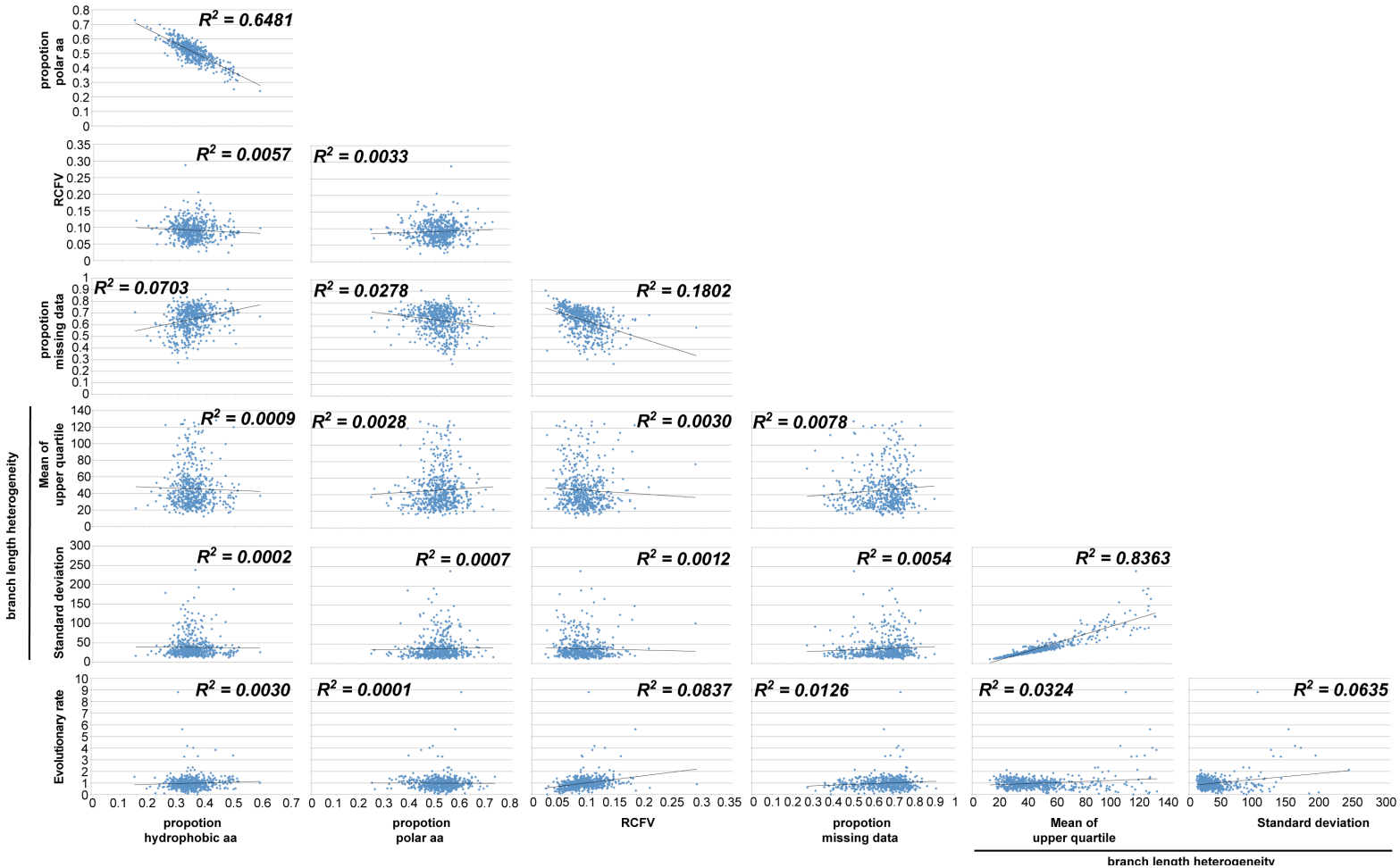

Figure S4. Correlation of the different gene properties such as the proportion of hydrophobic and polar amino acids, proportion of missing data, base composition (i.e., RCFV) and branch length heterogeneity (measured by the mean of upper quartile or the standard deviation of the LB scores) and evolutionary rate.

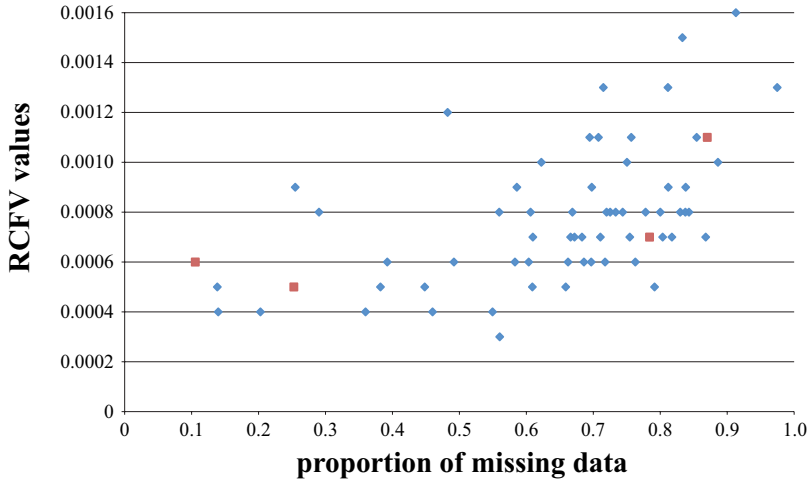

Figure S5. Plot of proportion of missing data against RCFV values measuring base composition heterogeneity for each species of dataset d01. Red = outgroup species, blue = ingroup species.

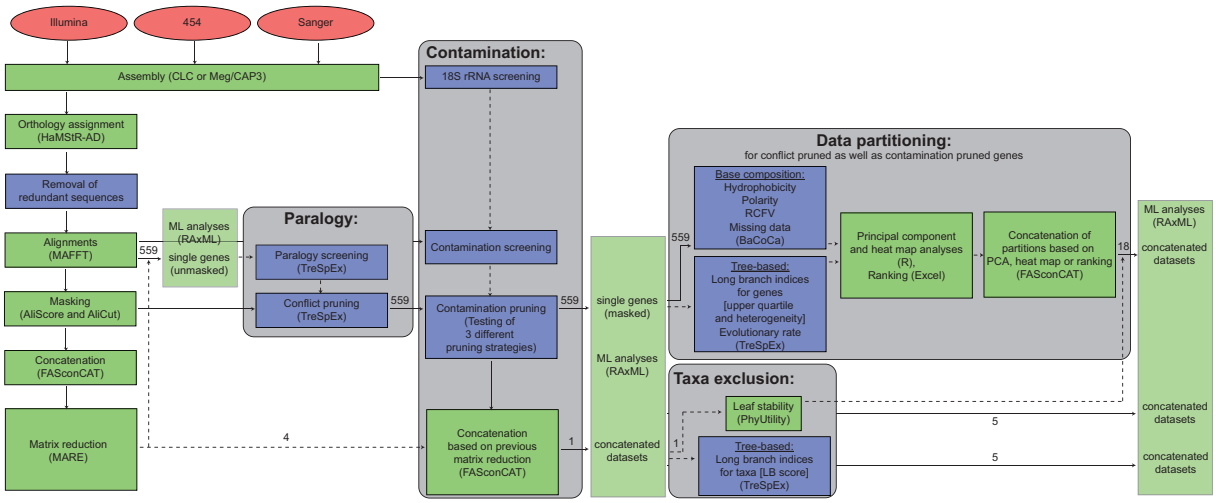

Figure S6. Workflow of the analyses conducted in this study. Grey boxes highlight paralogy and contamination screening as well as data partitioning and taxon exclusion as sensitivity analyses. Green boxes indicate published programs being used, blue boxes scripts and programs developed in the course of this study. Solid lines indicate that sequence information was transferred to the next step and dashed lined that other information from the previous step was taken (e.g., gene IDs to be included in the partition). Numbers of the lines indicate the number of datasets used.

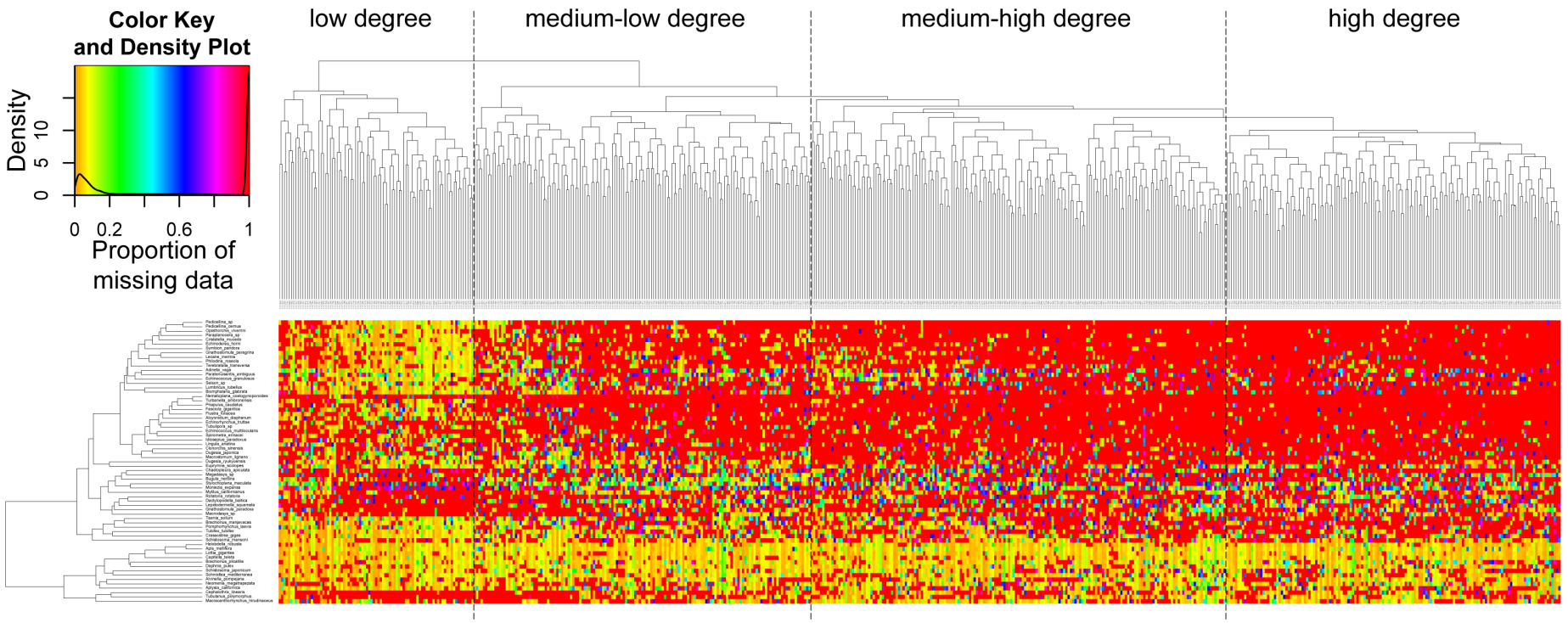

Figure S7. Heatmap of the proportion of missing data for each individual gene and taxon. Hierarchical clustering for both axes. Rows = species, columns = genes.

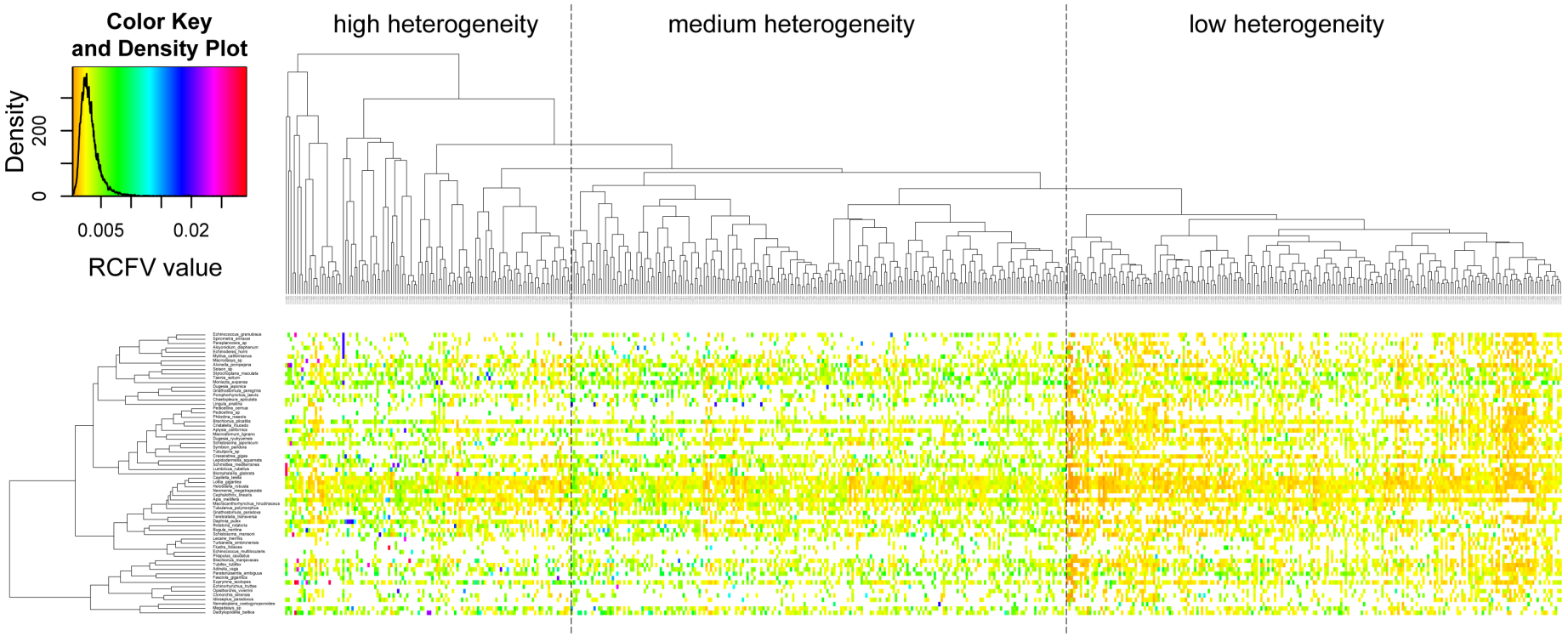

Figure S8. Heatmap of the taxon-specific RCFV values for each individual gene and taxon. Hierarchical clustering for both axes. Rows = species, columns = genes.