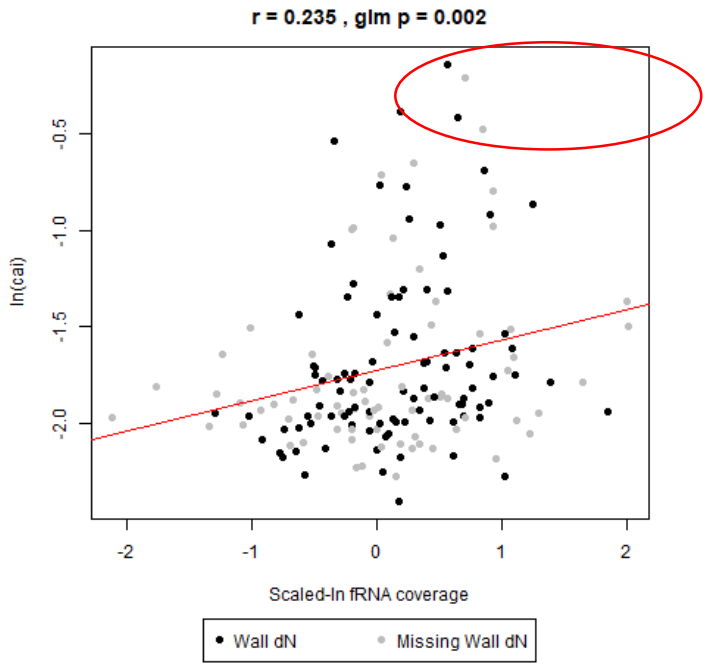
Figure C1: Ribosomal Outliers in fRNA Coverage versus CAI Plot

169 genes (ribosomal removal)



195 genes (WITH ribosomal)

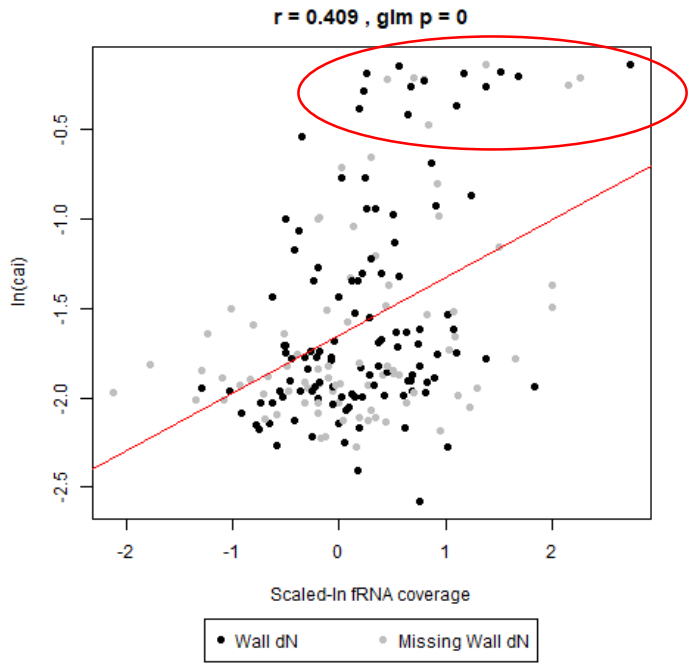
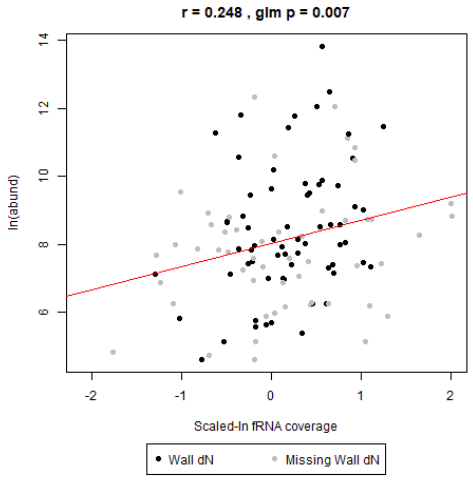
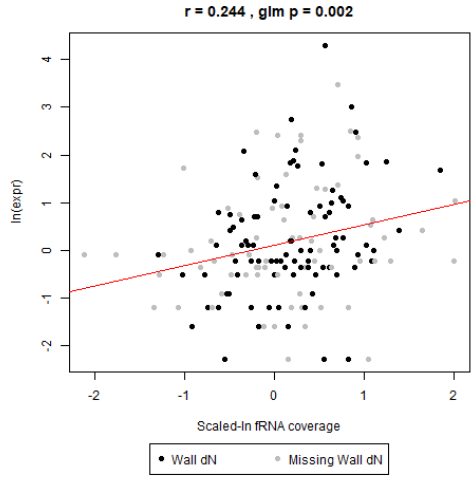
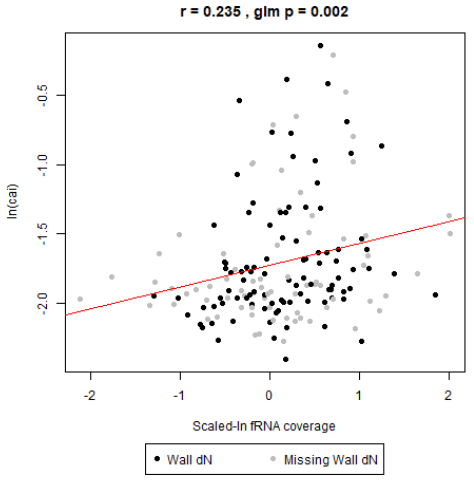
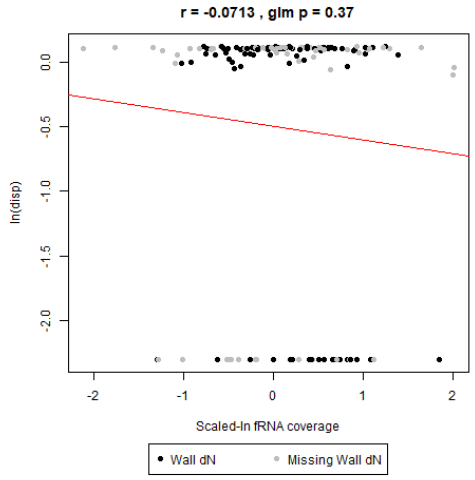


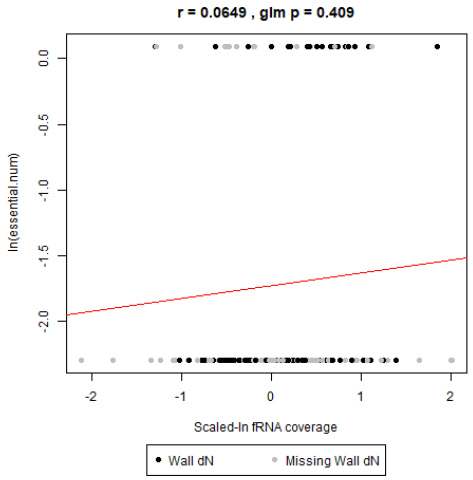
Figure C2: Functional Variable Association with fRNA Coverage

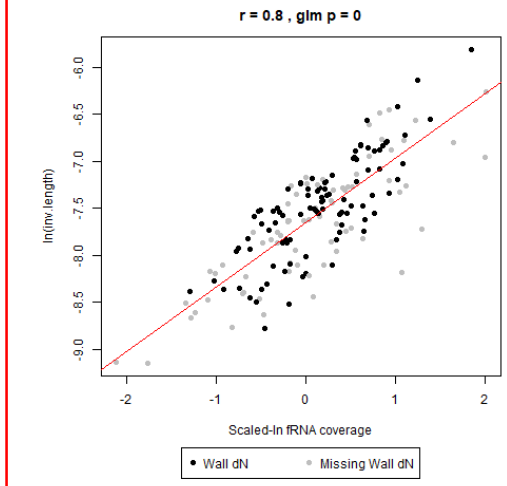


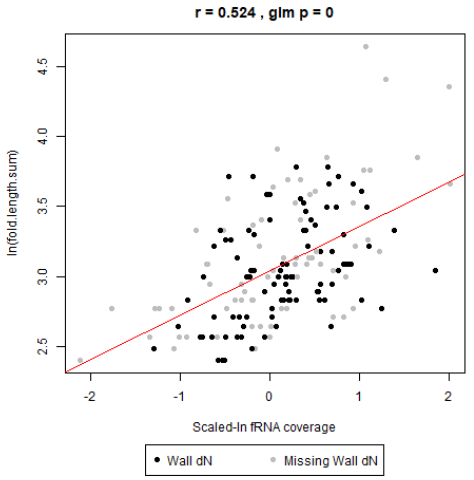








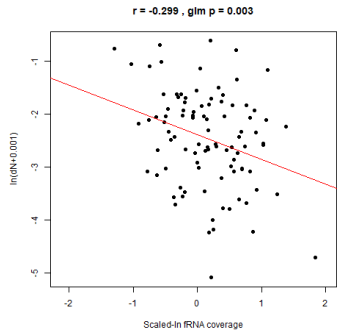


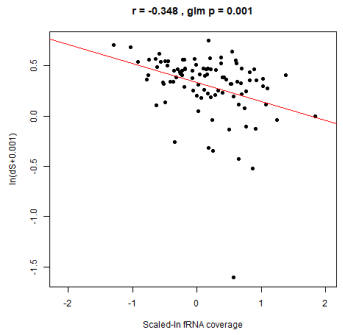


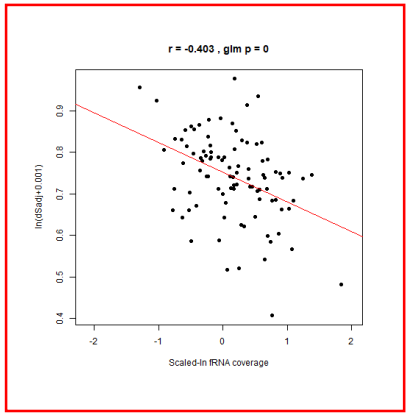
Plot without ribosomal genes. Most significant correlation with 1/(gene length).

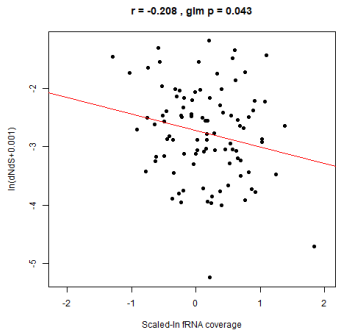
Figure C3: Nucleotide Divergence Association with fRNA Coverage

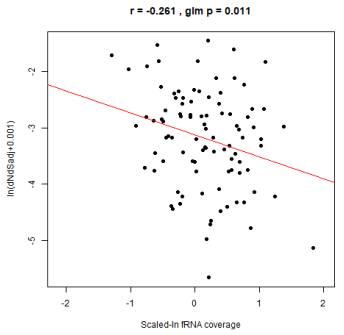
4-species (Wall)



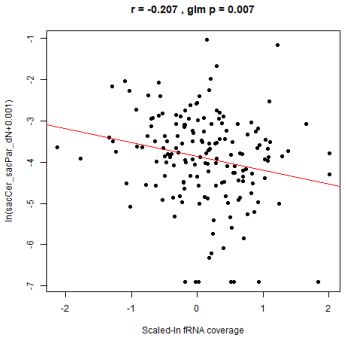


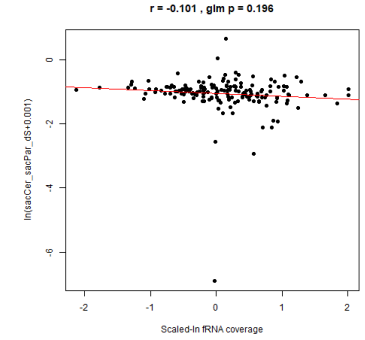


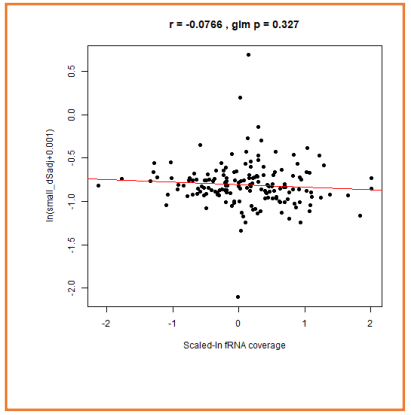


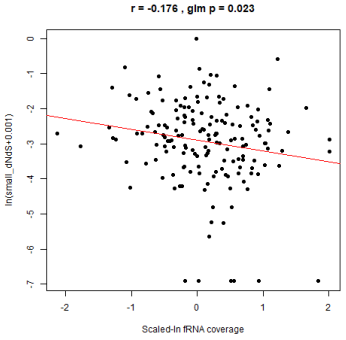


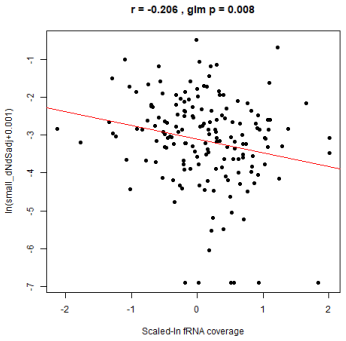
2-species (recalculated)









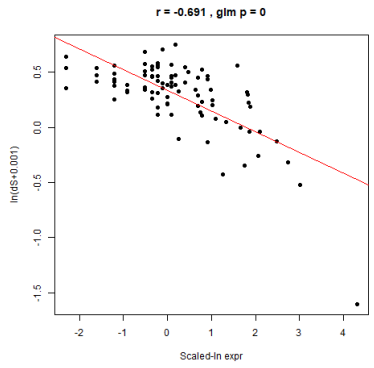


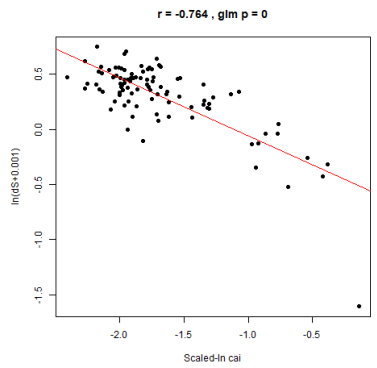
Plot without ribosomal genes

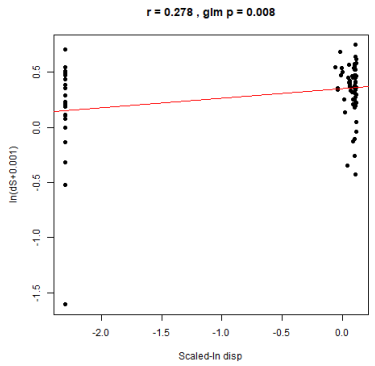
dS`=dS - m*c, where m = -2.02 for 4-species and m=-0.386 for 2-species

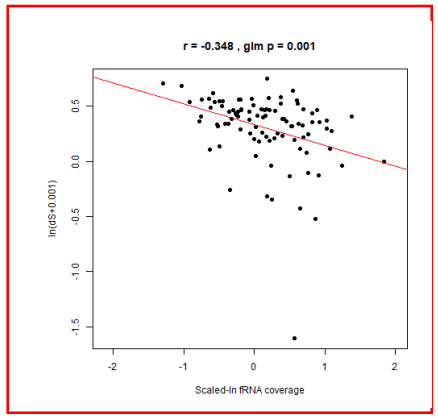
Figure C4: Functional Variable Association with dS` (adjusted dS)

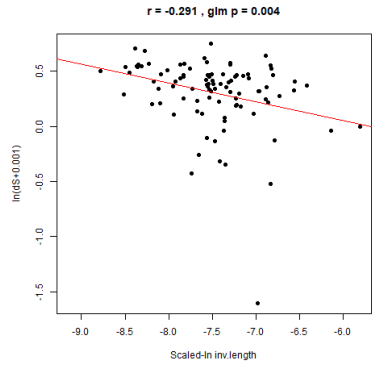
4-species (Wall)



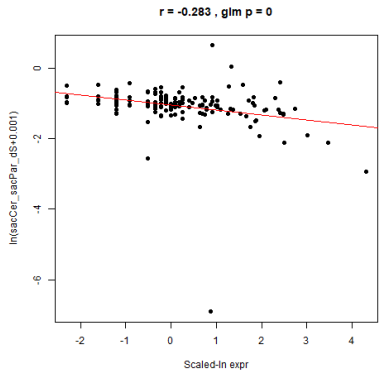


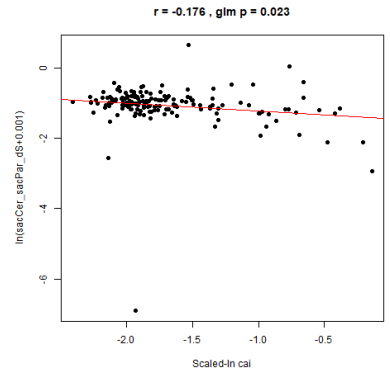


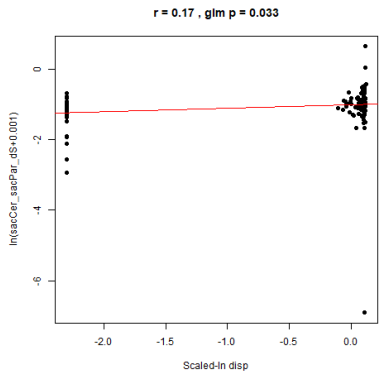


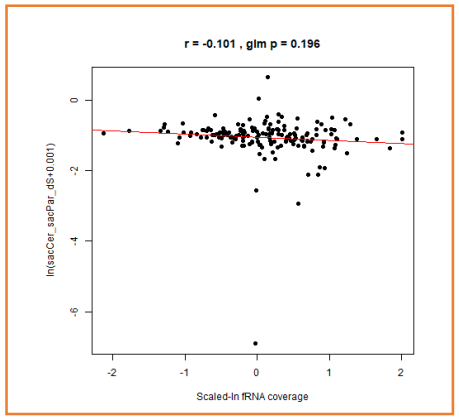


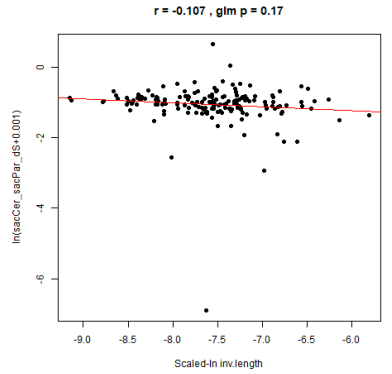
2-species (recalculated)











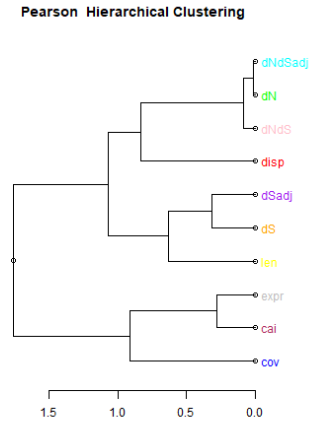
Plot without ribosomal genes

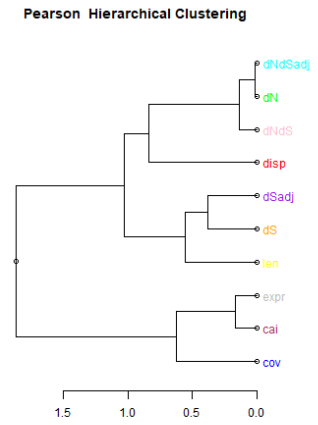
dS`=dS - m*c, where m = -2.02 for 4-species and m=-0.386 for 2-species

Figure C5: Functional Variable Hierarchical Clustering (PCA Samples)

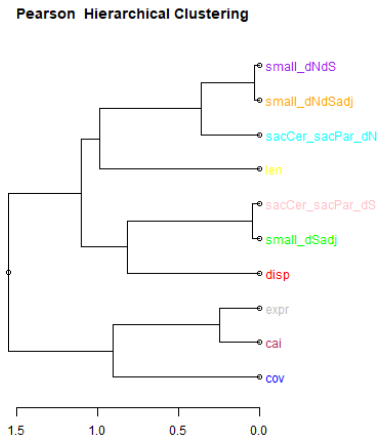
169 genes (ribosomal removal) **195 genes (WITH ribosomal)**

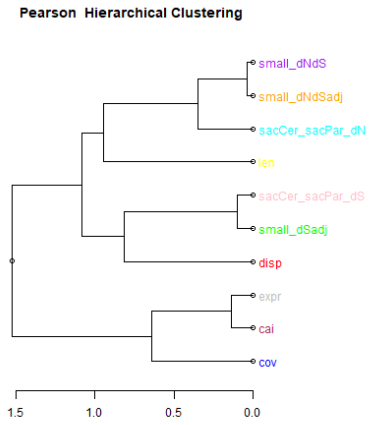
4-species (Wall)





2-species (recalculated)





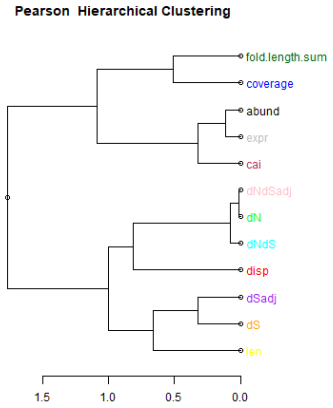
"Complete" linkage used as method for Hierarchical Clustering, with Pearson Dissimilarity used as distance metric.

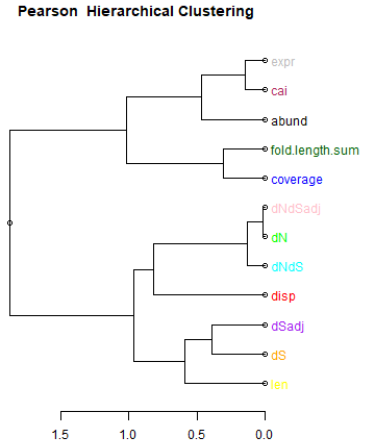
Figure C6: Functional Variable Hierarchical Clustering (All Samples, Pairwise Complete Observations)

169 genes (ribosomal removal)

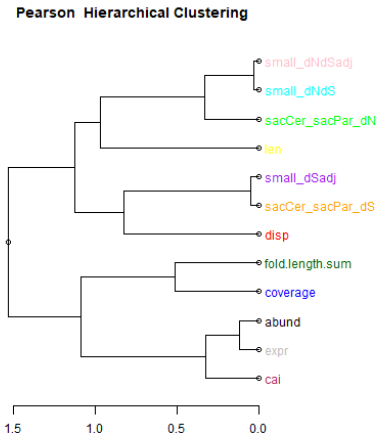
195 genes (WITH ribosomal)

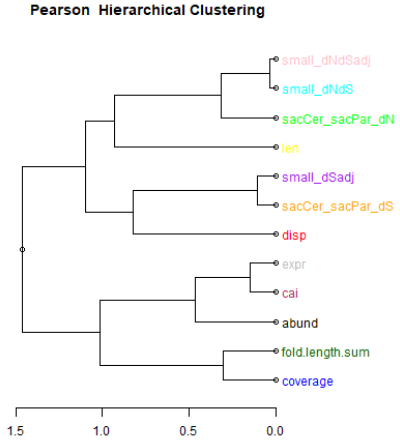
4-species (Wall)





2-species (recalculated)



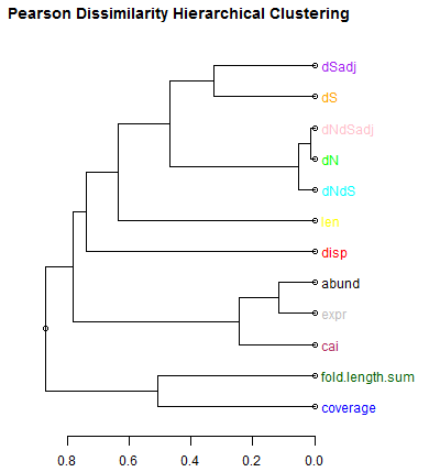


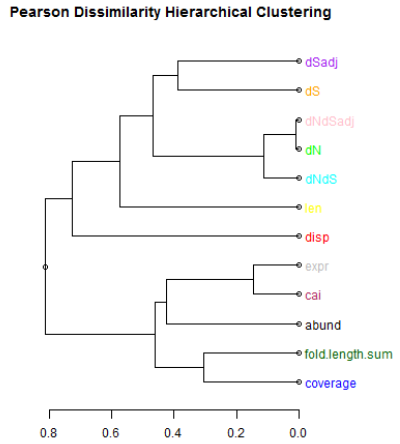
"Complete" linkage used as method for Hierarchical Clustering, with Pearson Dissimilarity used as distance metric.

Figure C7: Pearson Dissimilarity with Different Linkage Method

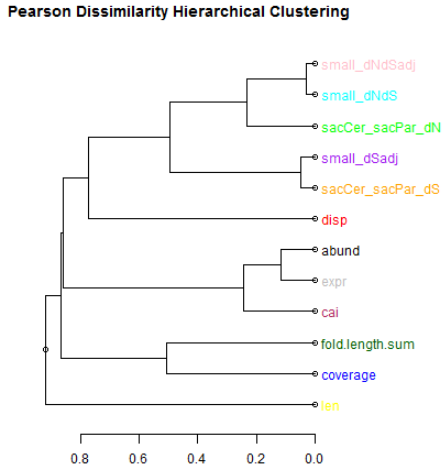
169 genes (ribosomal removal) | **195 genes (WITH ribosomal)**

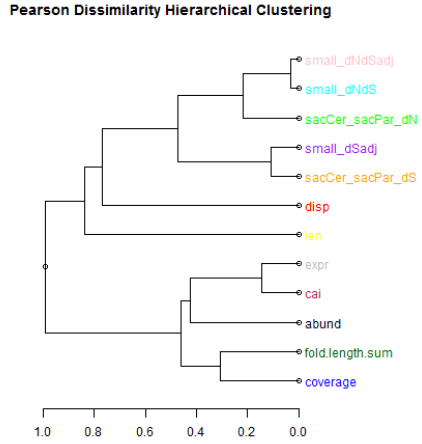
4-species (Wall)





2-species (recalculated)





"Single" linkage used as method for Hierarchical Clustering, with Pearson Dissimilarity used as distance metric. Use all samples, with correlations calculated from Pairwise Complete Observations.

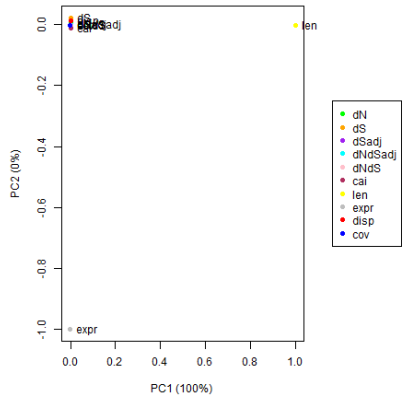
At least for 2-species without ribosomal removal (same set of separate correlations in Figure 2), notice that gene length is greatest outlier (but it is more similar to coverage and the sum of coding fRNA length, which I believe makes sense because coverage = (sum of coding fRNA length) / (gene length).

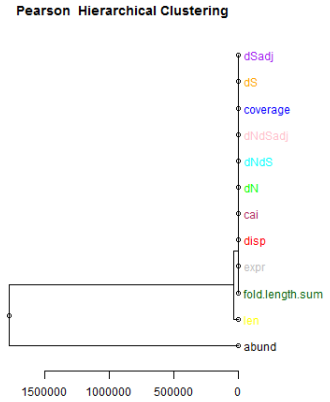
Figure C8: Euclidean Distance Hierarchical Clustering More Similar to R-base PCA

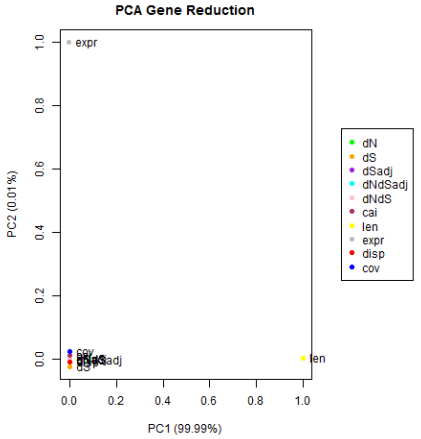
169 genes (ribosomal removal)

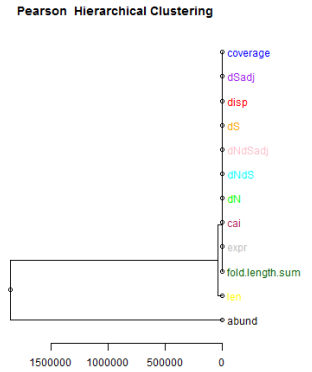
195 genes (WITH ribosomal)

4-species (Wall)

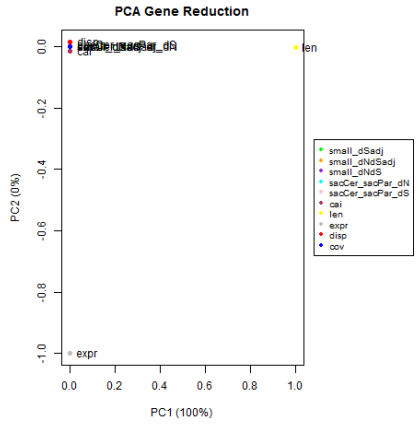


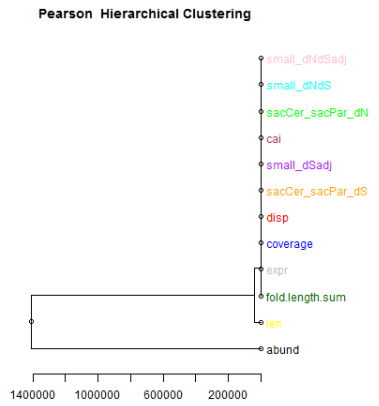


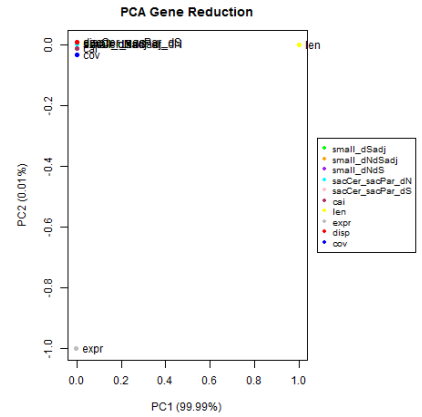


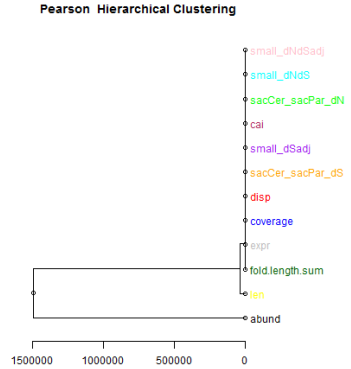


2-species (recalculated)









"Complete" linkage used as method for Hierarchical Clustering. Use all samples.

While Pearson Dissimilarity looks noticeably different than R-base PCA clustering, notice that Euclidian Distance is more similar to R-base PCA (but that is different than would be inferred from *pcr()* PCA regression from the 'pls' package, using *mysum()* function from Drummond et al. 2006).

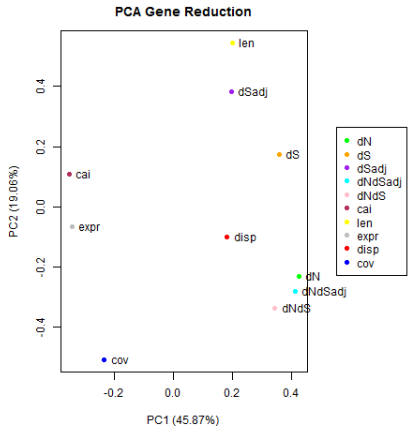
Also, in order for PCs to be in typical regression model, transposed matrix is required (to have one PC1 value per gene, for example; plots to the left reduce genes to number of variables)

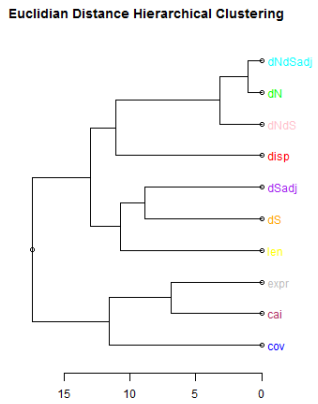
Figure C9: Euclidean Distance Hierarchical Clustering More Similar to R-base PCA (Scaled Variables)

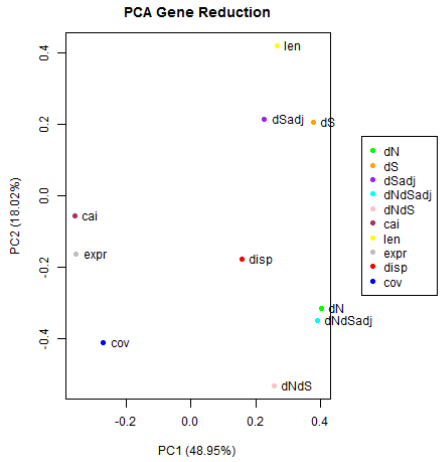
169 genes (ribosomal removal)

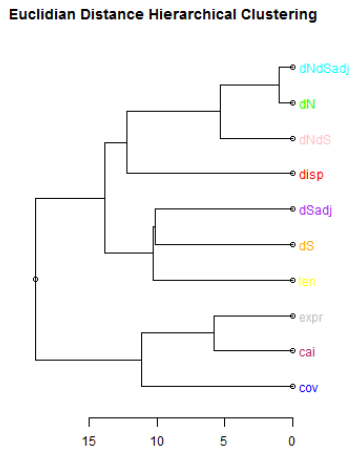
195 genes (WITH ribosomal)

4-species (Wall)

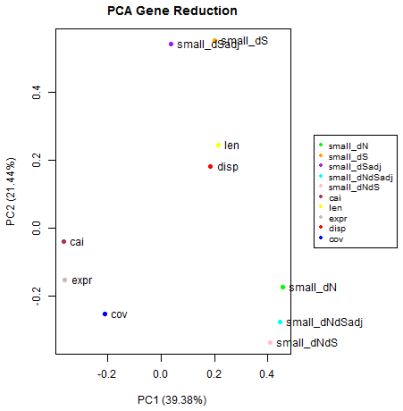


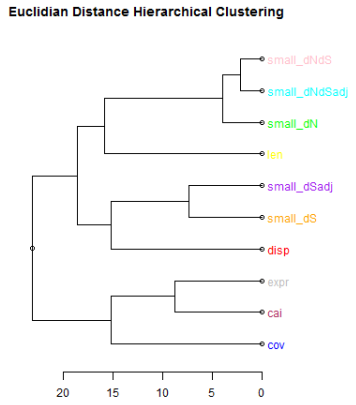


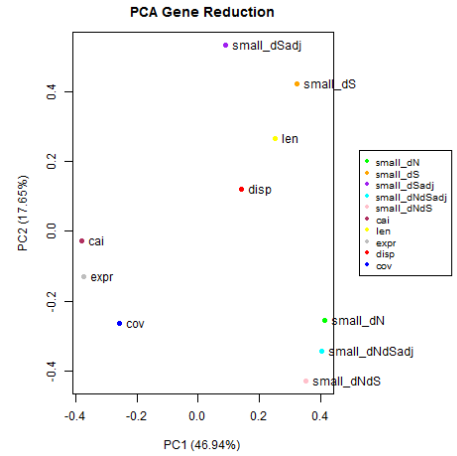


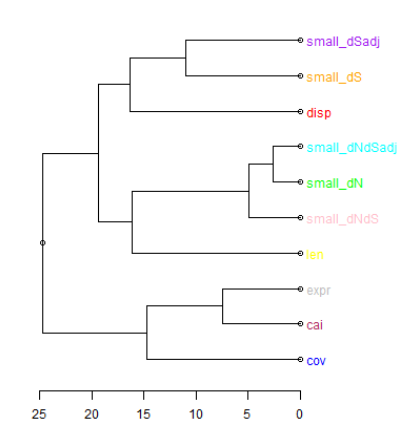


2-species (recalculated)









"Complete" linkage used as method for Hierarchical Clustering. Use all samples, *but scale natural log prior to PCA calculation*.

While Pearson Dissimilarity looks noticeably different than R-base PCA clustering, notice that Euclidian Distance is more similar to R-base PCA (but that is different than would be inferred from *pcr()* PCA regression from the 'pls' package, using *mysum()* function from Drummond et al. 2006). **Nevertheless, fRNA Coverage (cov) and dNdS (or small_dNdS) are now closer on PC2 in both plots (with CAI/expr/cov clustered on PC1).**

Also, in order for PCs to be in typical regression model, transposed matrix is required (to have one PC1 value per gene, for example; plots to the left reduce genes to number of variables)

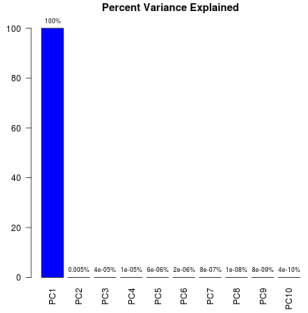
Figure C10: R-base PCA Percent Variance Explained

Figure C8
(Original, Linear Variables)

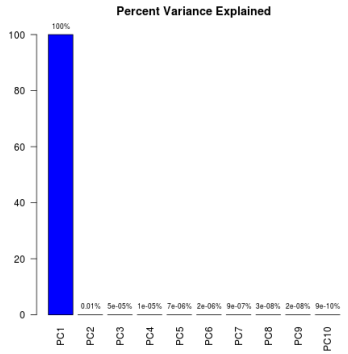
Figure C9
(Scaled Variables)

4-species (Wall)

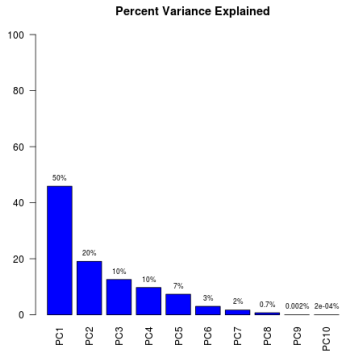
169 genes



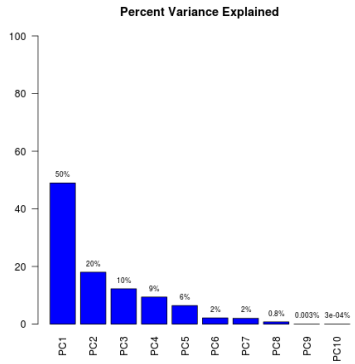
195 genes



169 genes

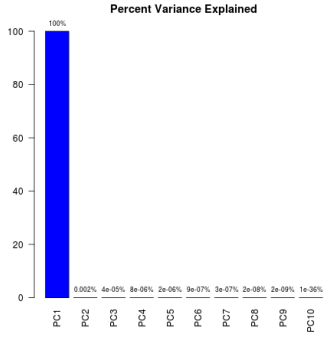


195 genes

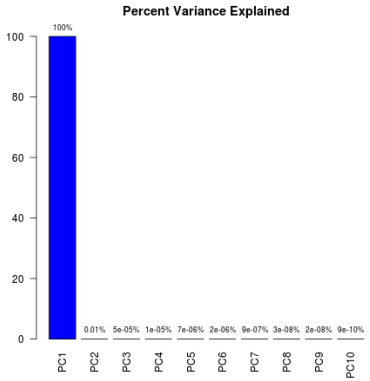


2-species (recalculated)

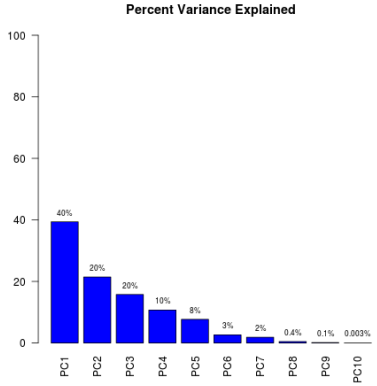
169 genes



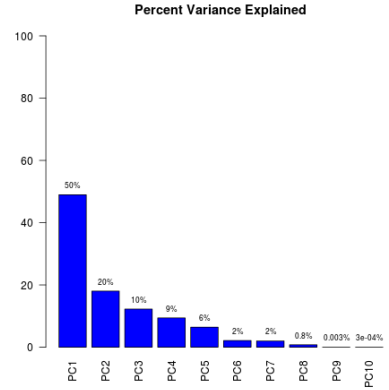
195 genes



169 genes



195 genes



- As I would expect, percent variance explained decreases for later principal components (and PC2 explains larger amount of variance in scaled data)