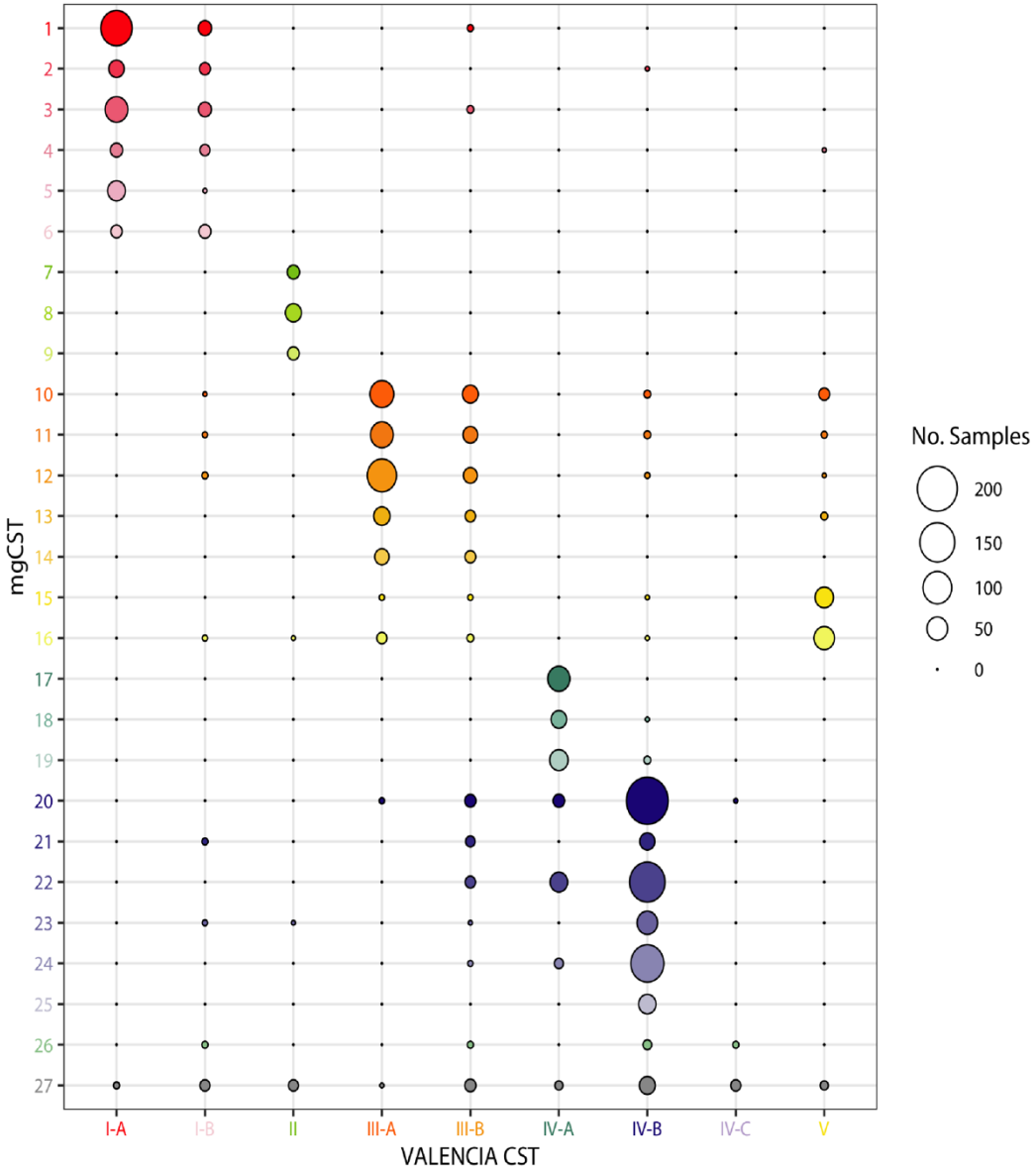


Fig. S1. Metagenomic CST assignments correspond to marker gene-based CST assignments primarily through the predominant taxon. However, dominance by an mgSs is not captured through marker-based CSTs.

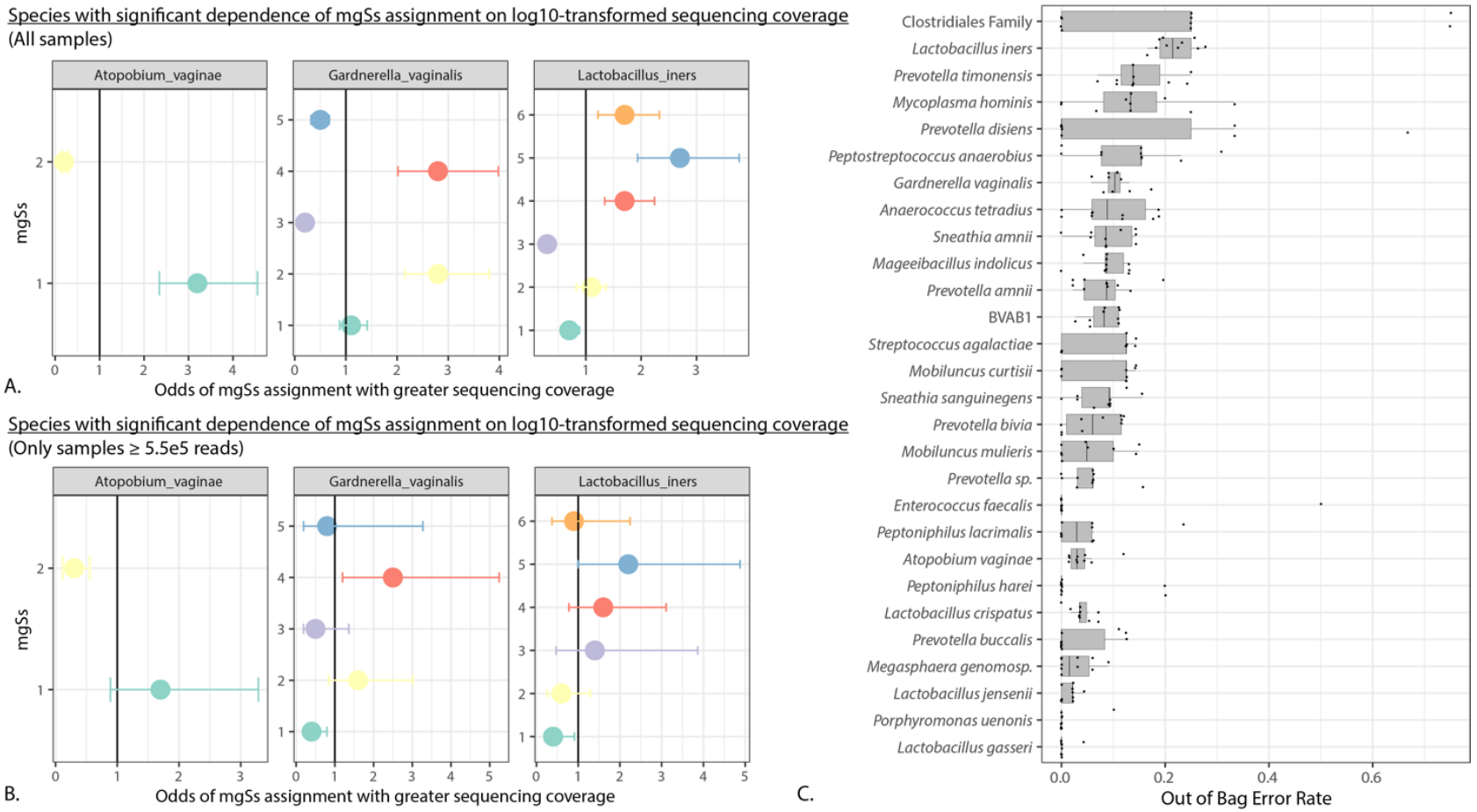


Fig. S2. A) For some species, mgSs assignment was significantly associated with sequencing depth of a sample (see Table S4). B) When only samples with ≥ 5.5e5 reads are used in mgSs classifier random forest tree construction, mgSs assignment is no longer significantly associated with depth of sequencing. C) For each species, 10-fold cross-validation yielded random forest misclassification error estimates.

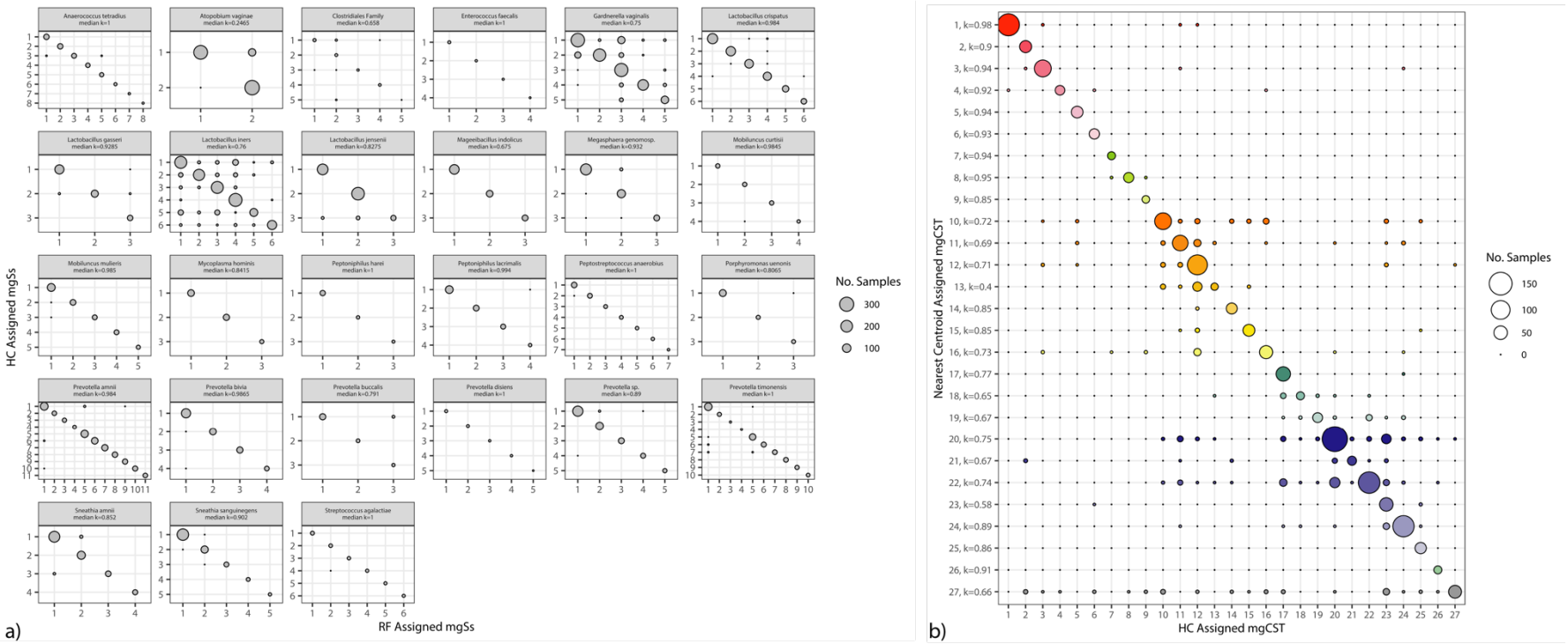


Figure S3. A) MgSs assignment using hierarchical clustering (y-axis) are highly concordant ($\boldsymbol{\kappa} > 0.8$) with those assigned by the random forest classifier (x-axis) for that species. B) MgCST assignment using hierarchical clustering (x-axis) are highly concordant ($\boldsymbol{\kappa} = 0.78$) with those assigned by the nearest centroid classifier (y-axis).

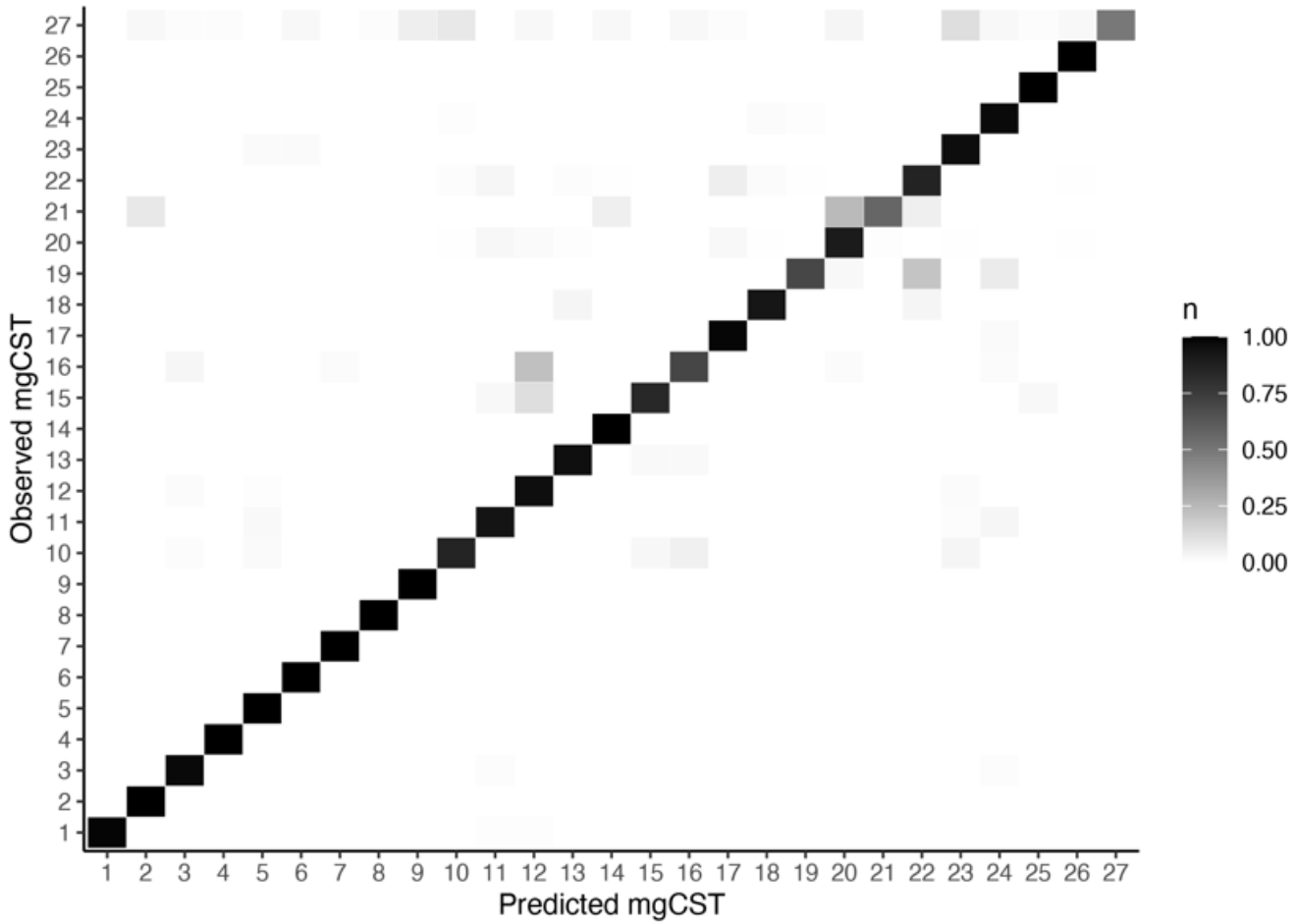
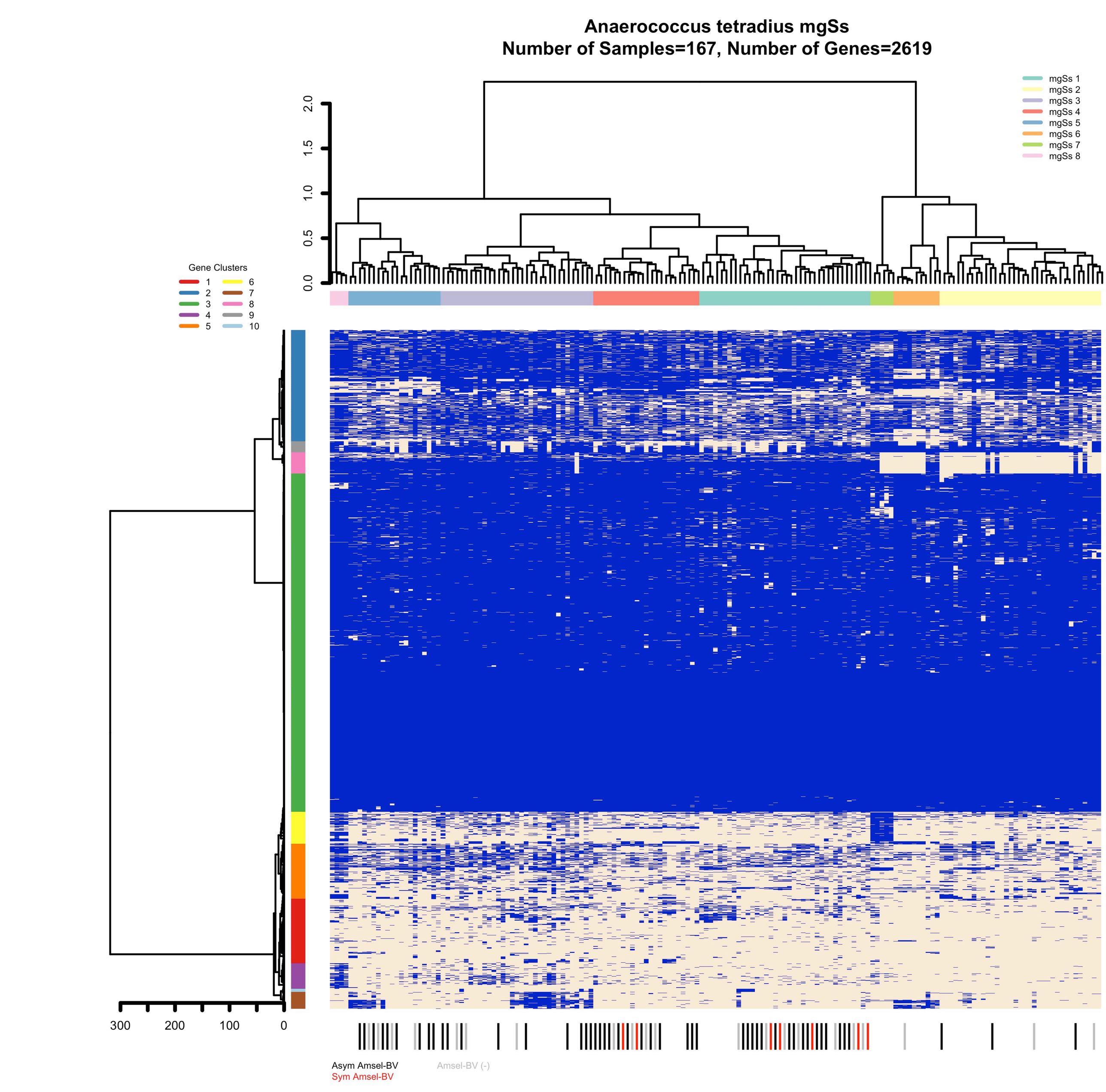


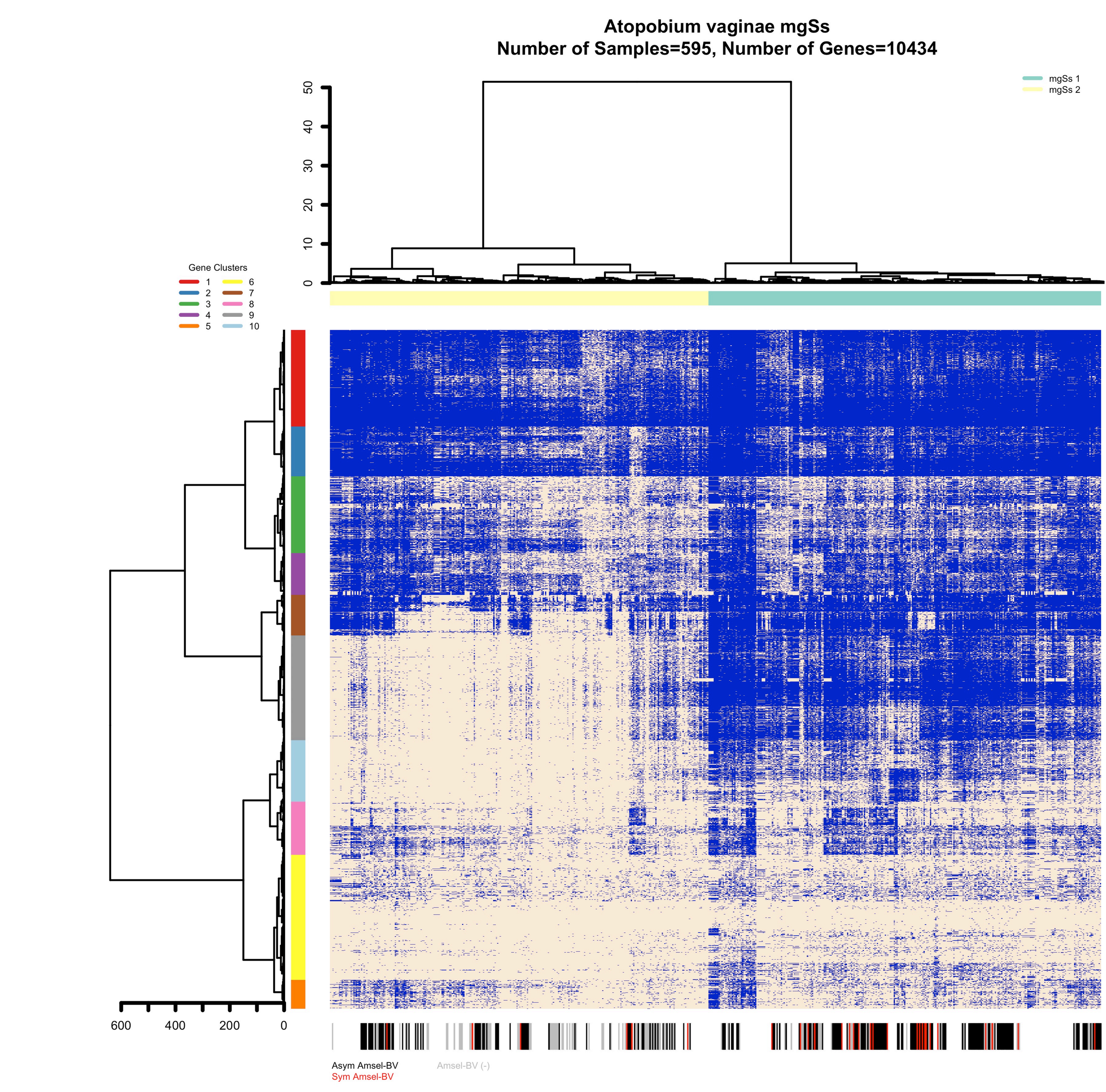
Fig. S4. Confusion matrix and classification error estimates from 10-fold cross-validation of a nearest-centroid classifier for mgCSTs.

Fig. S5. Gene presence heatmaps for all species for which mgSs were constructed. Samples with >75% estimated median number of genes encoded in reference genomes from the Genome Taxonomy Database [64], see Table S3) were used to build mgSs. In each heatmap, samples are in the columns and genes are in the rows. Assigned mgSs are indicated in the column side colors at the top. Gene clusters are colored for each gene on the y-axis. The bottom x-axis indicates Amsel-BV diagnoses, if clinical evaluation data were available for the sample. Dendrograms were built using Ward linkage of Jaccard distances.

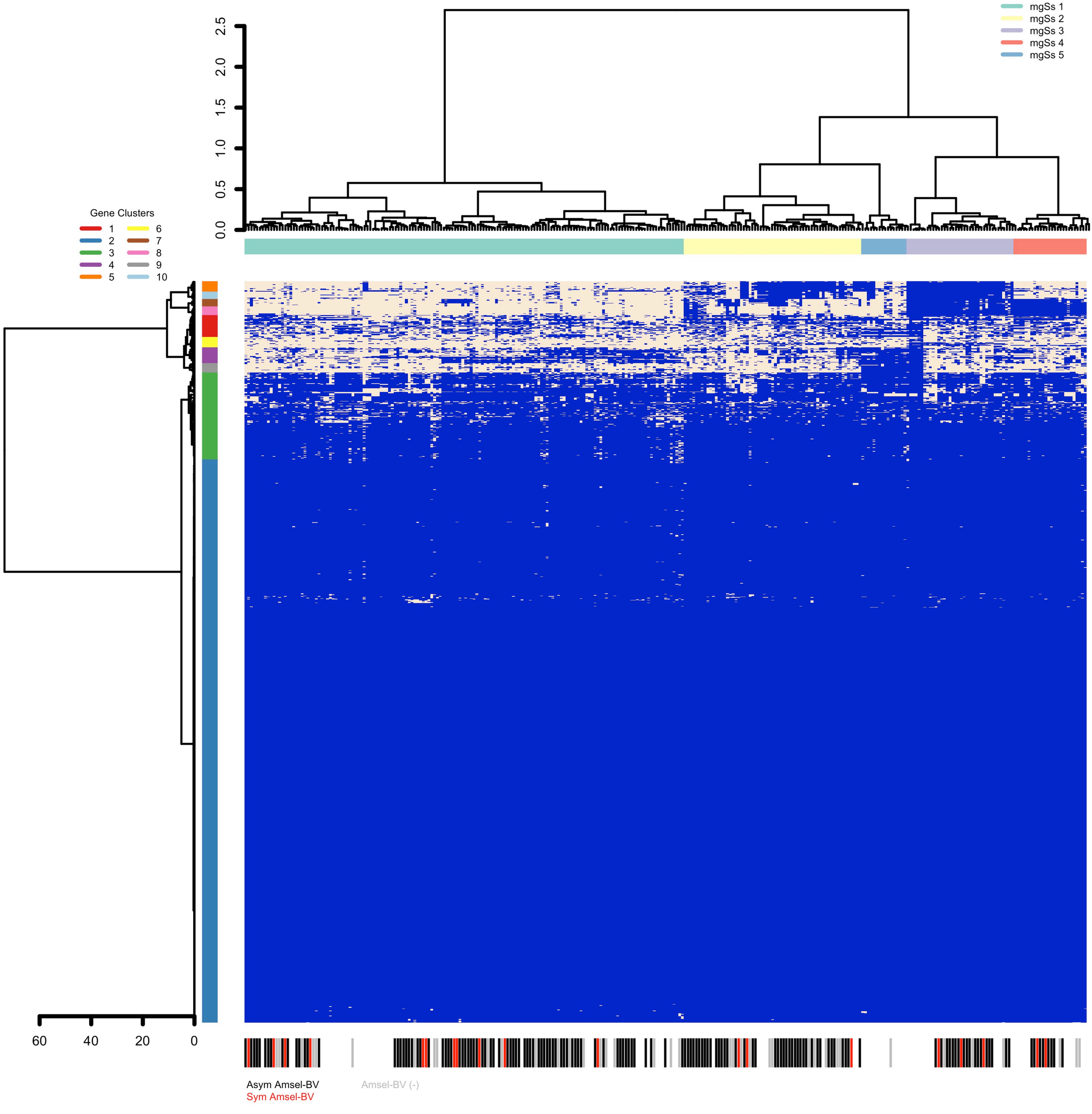
Anaerococcus tetradius mgSs
Number of Samples=167, Number of Genes=2619



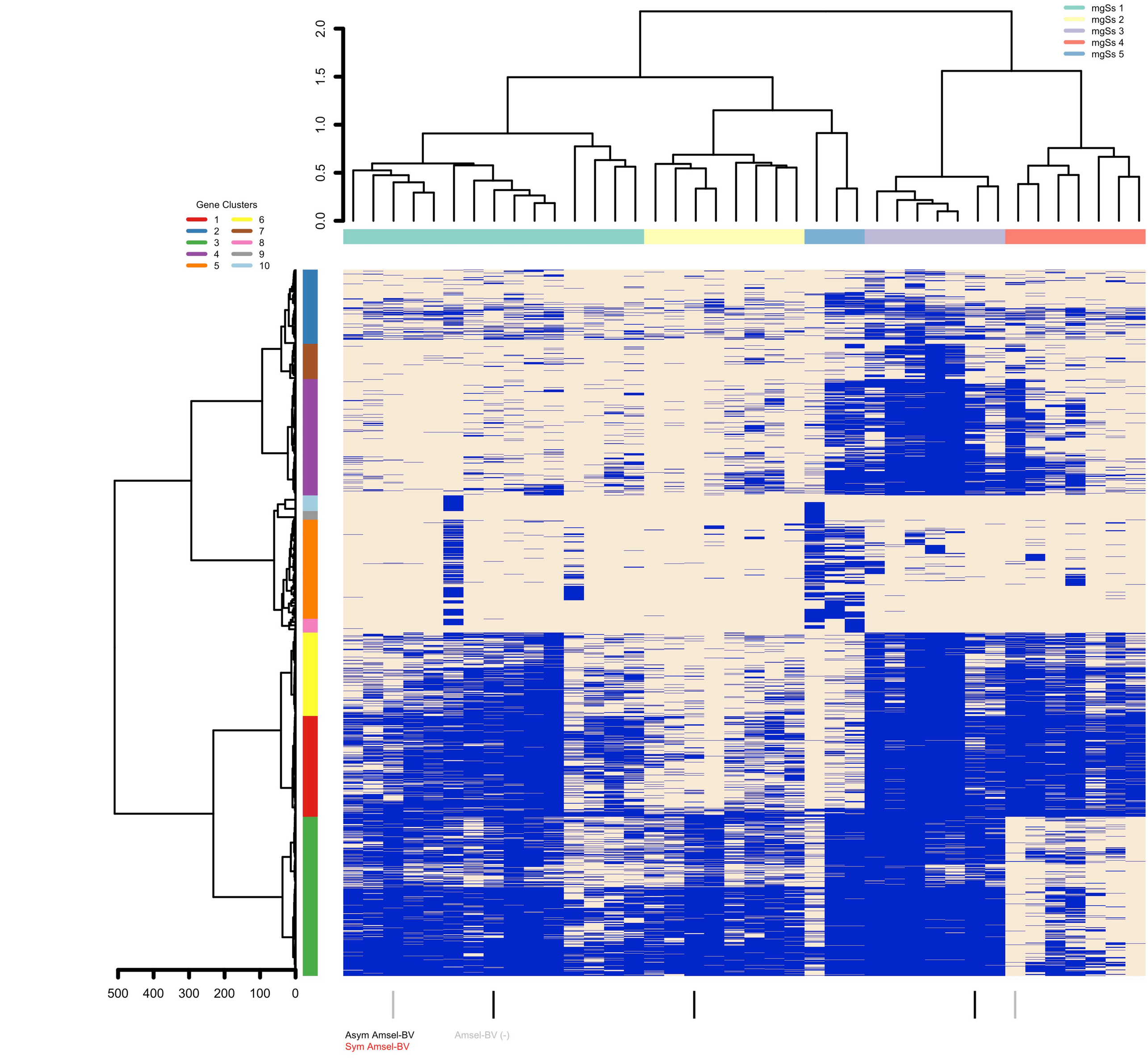
Atopobium vaginae mgSs
Number of Samples=595, Number of Genes=10434



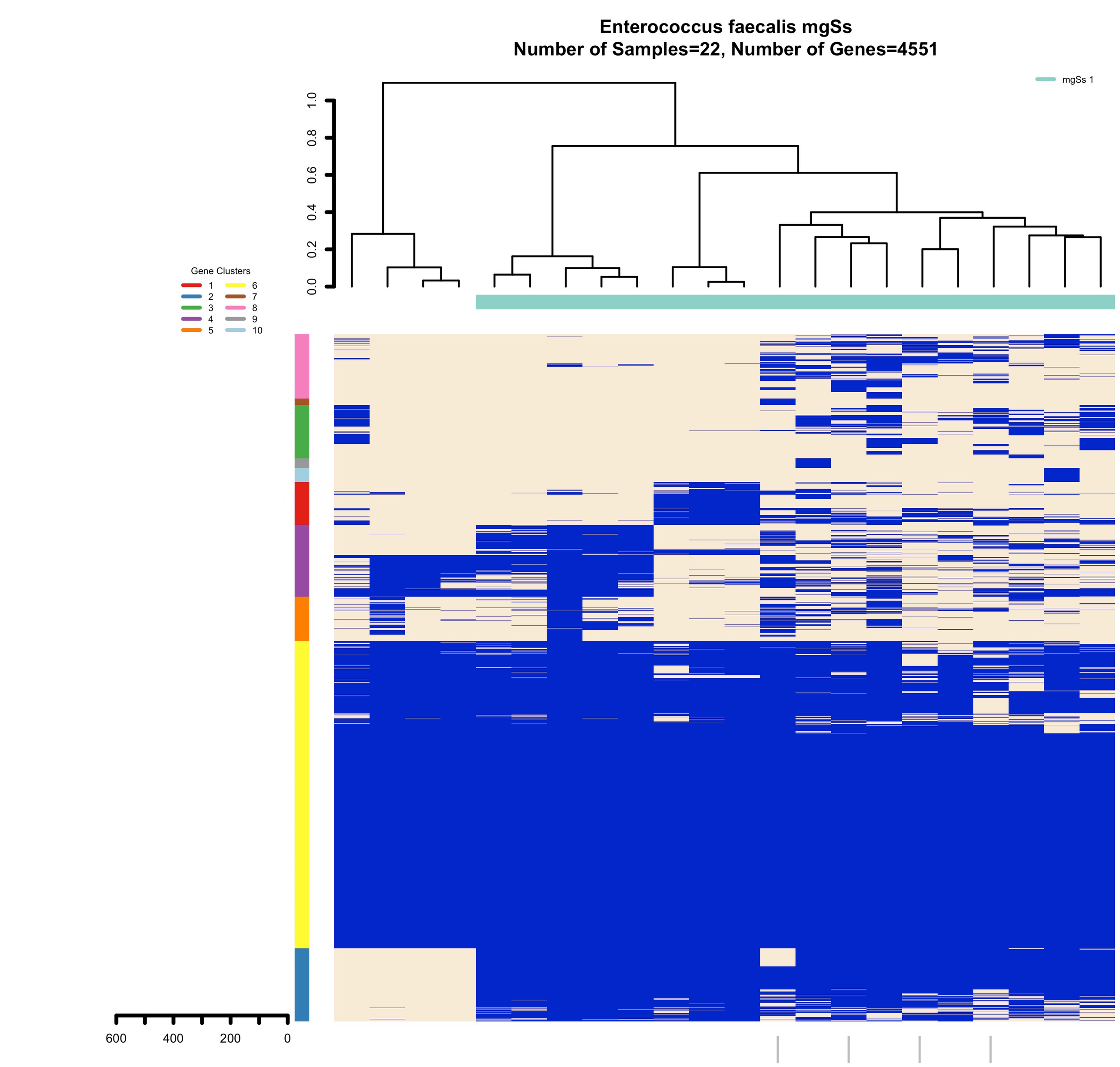
Ca. Lachnocurva vaginae mgSs
Number of Samples=299, Number of Genes=1356



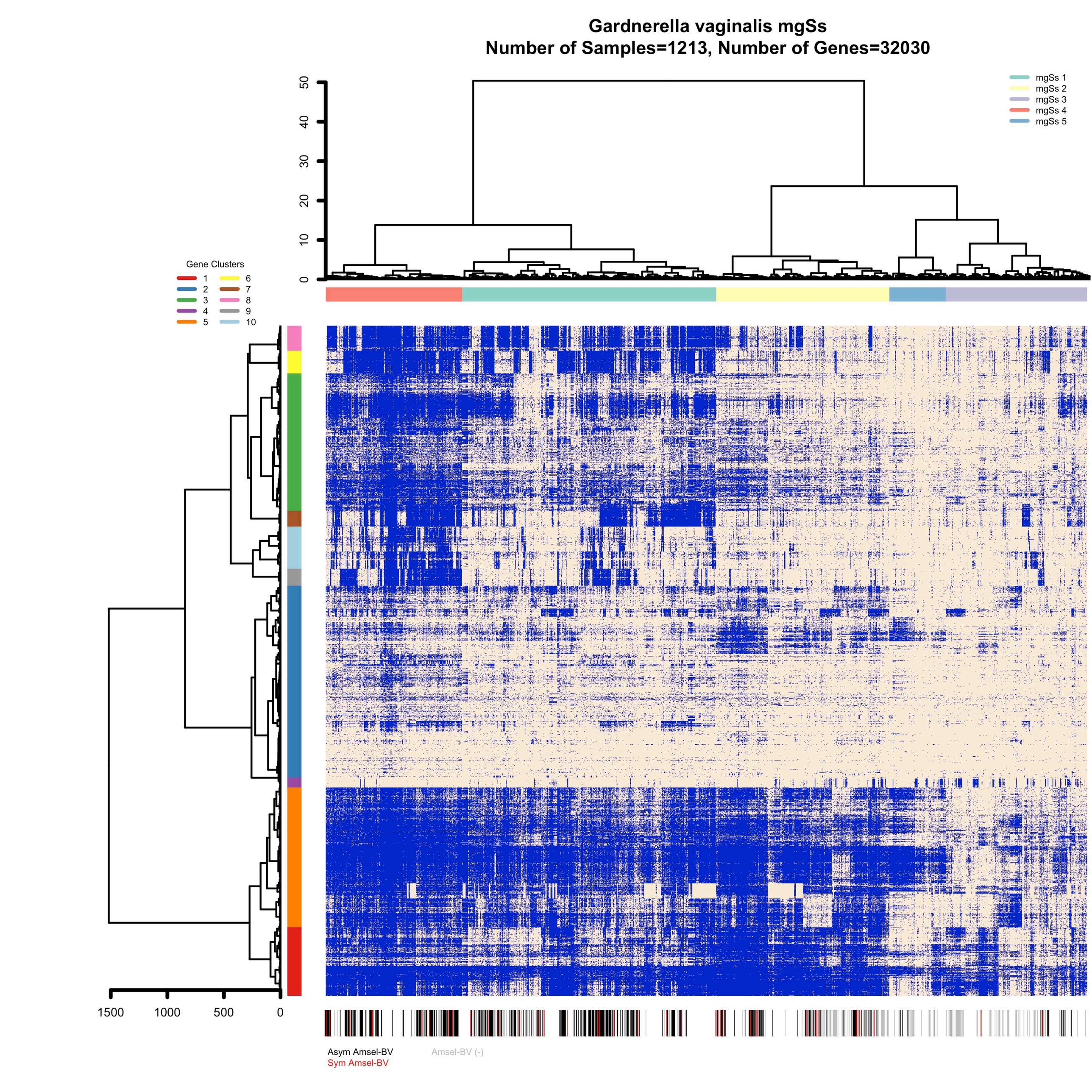
Clostridiales Family mgSs
Number of Samples=40, Number of Genes=5505



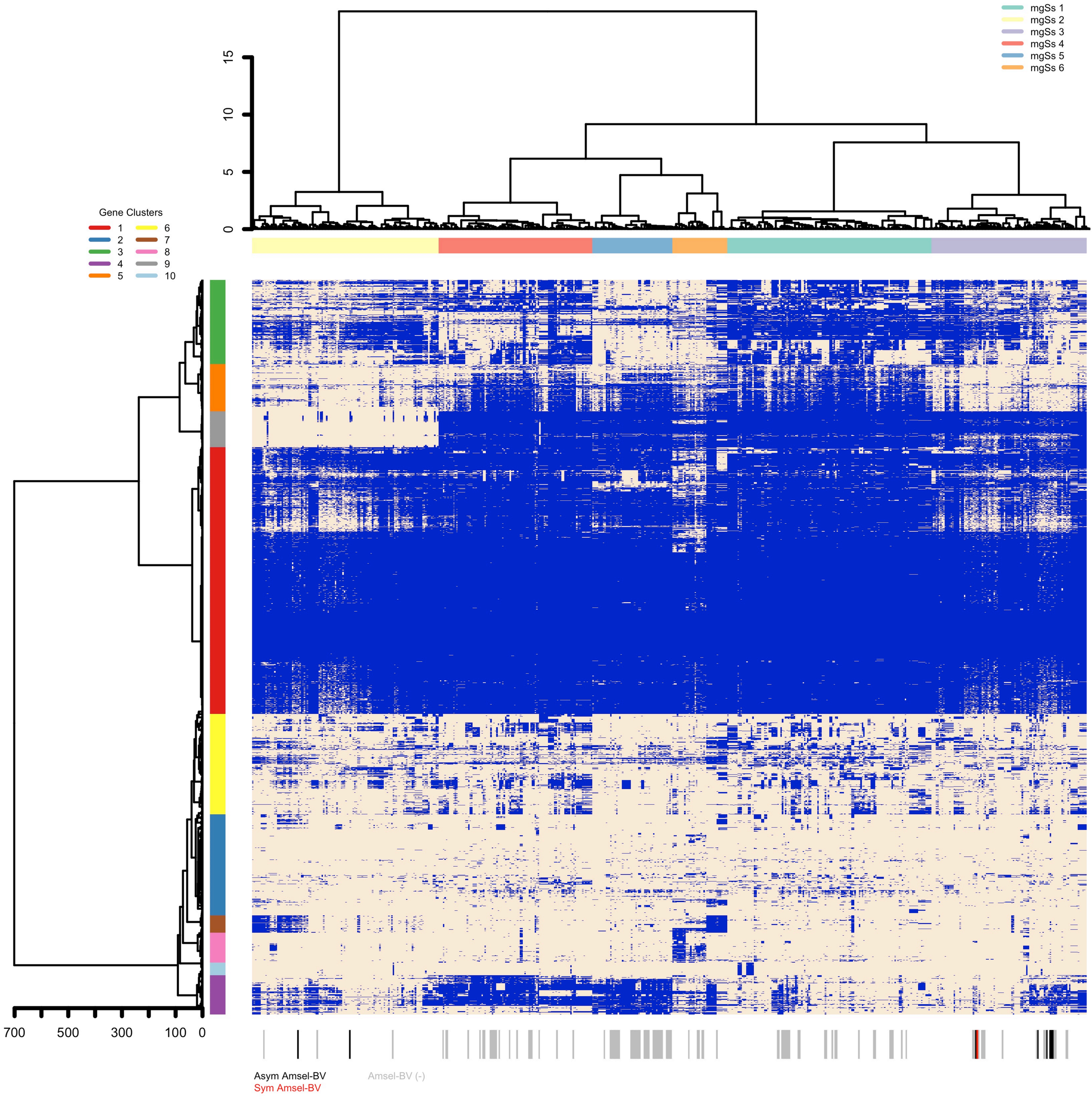
Enterococcus faecalis mgSs
Number of Samples=22, Number of Genes=4551



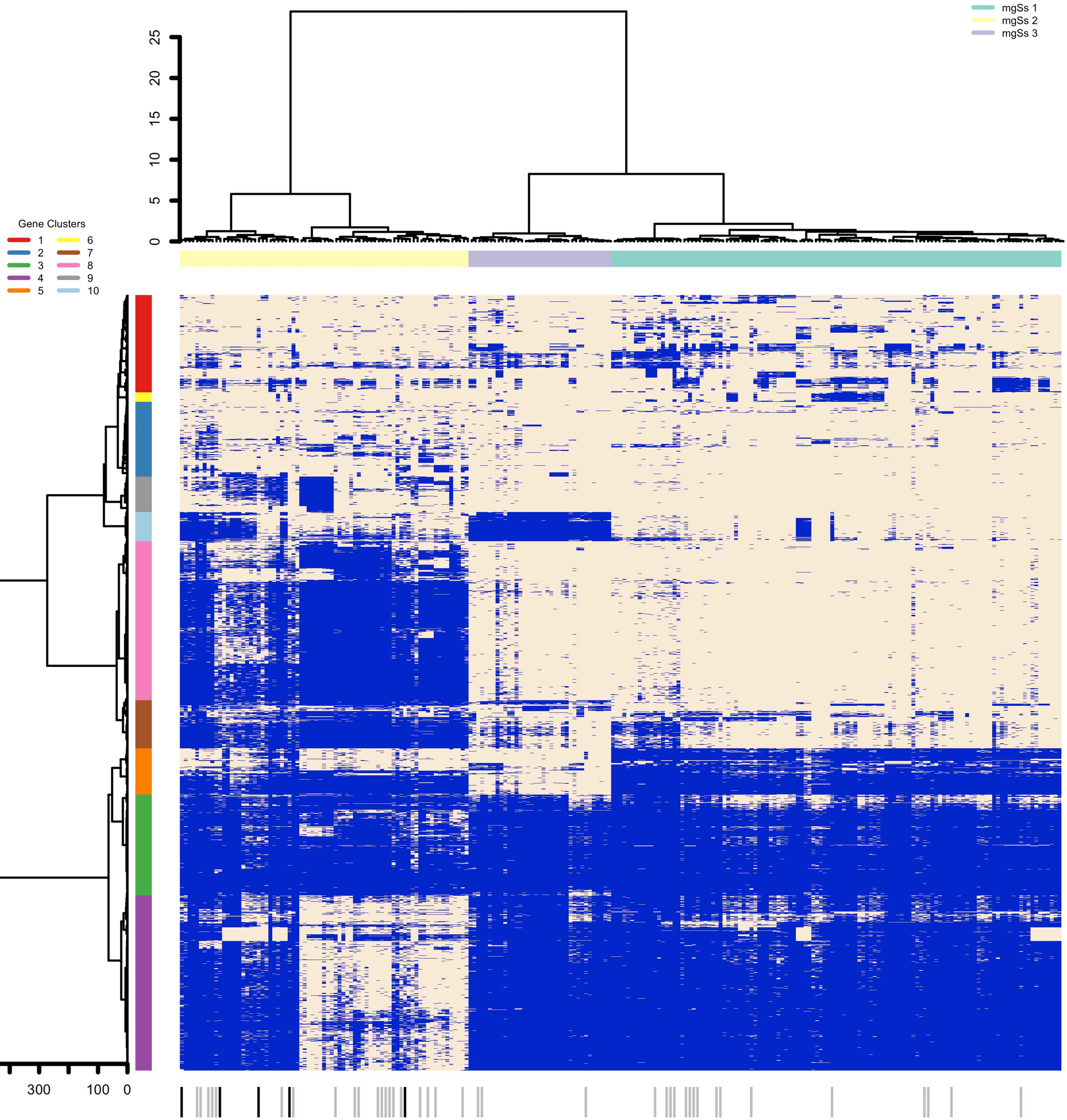
Gardnerella vaginalis mgSs
Number of Samples=1213, Number of Genes=32030



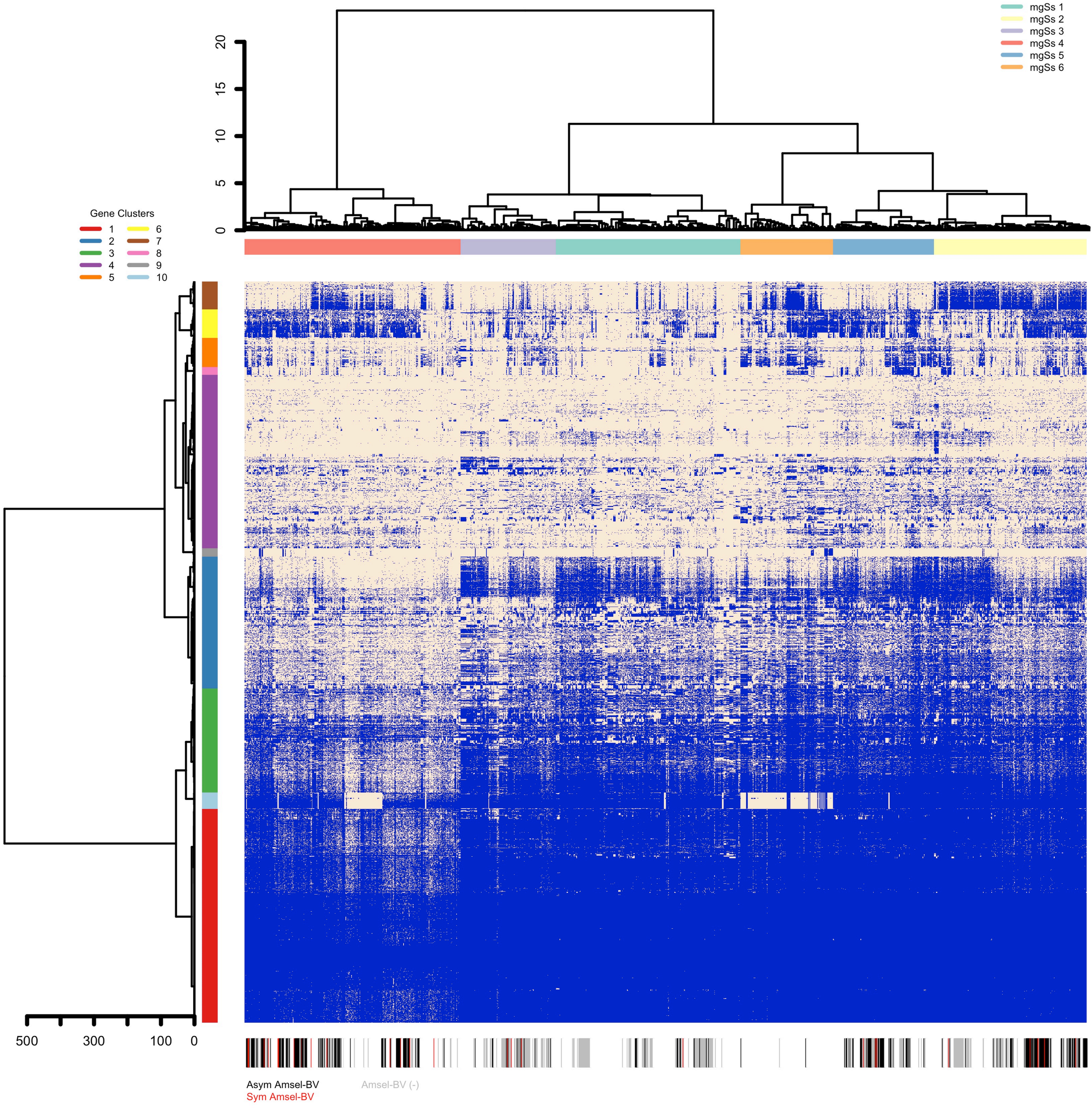
Lactobacillus crispatus mgSs
Number of Samples=564, Number of Genes=6455



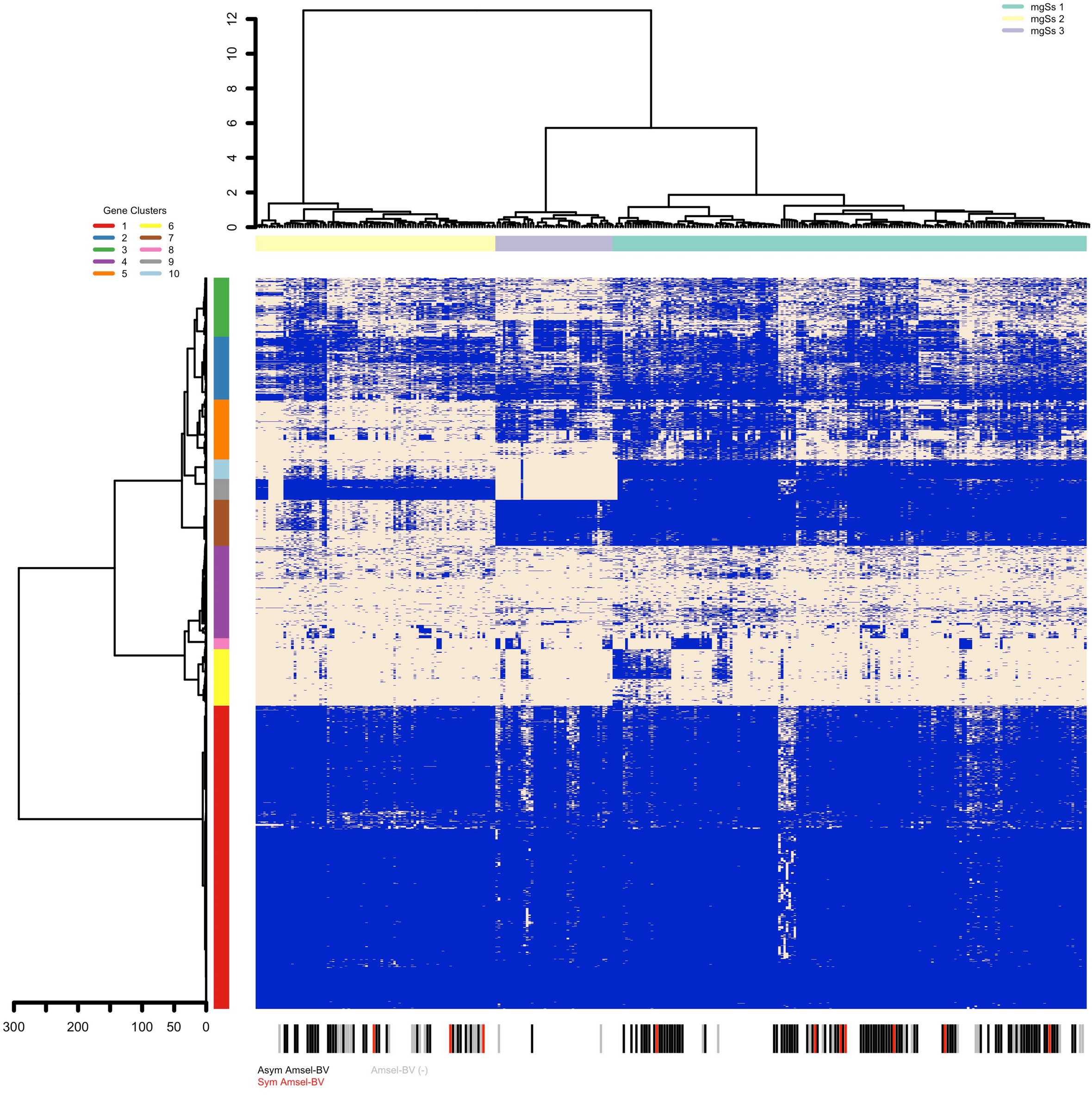
Lactobacillus gasseri mgSs
Number of Samples=229, Number of Genes=4456



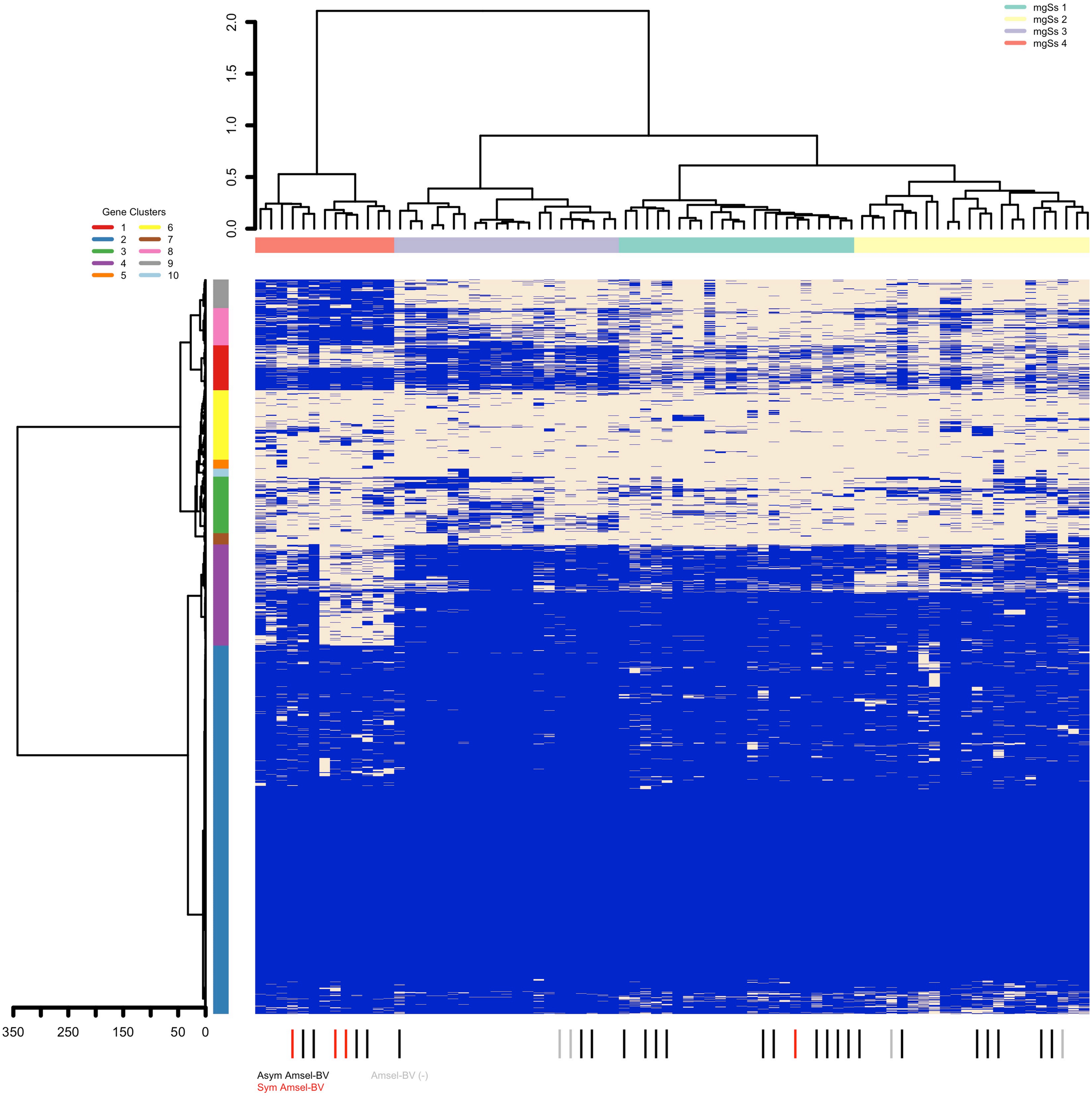
Lactobacillus iners mgSs
Number of Samples=1325, Number of Genes=5011



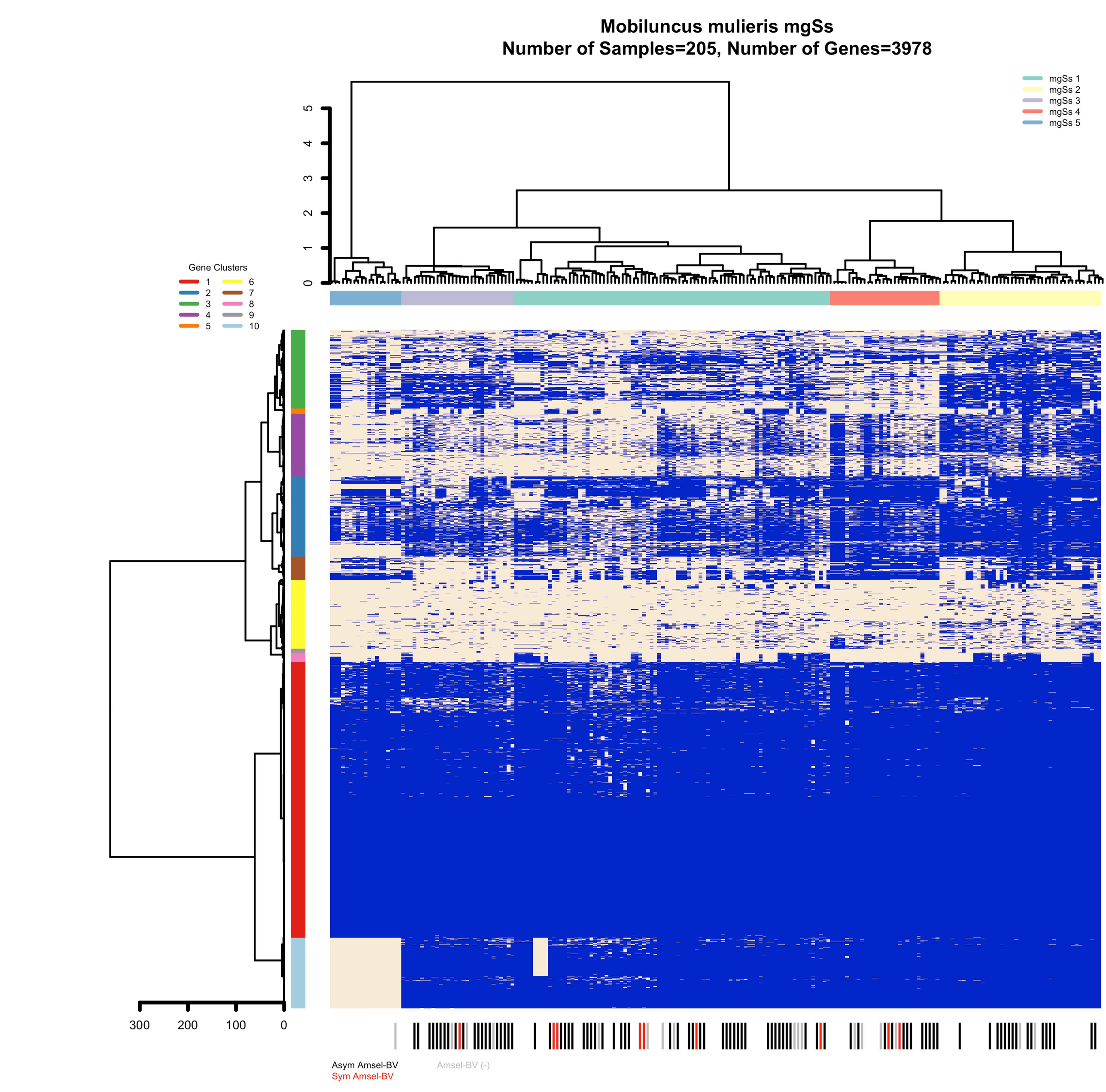
Megasphaera genomosp. mgSs
Number of Samples=326, Number of Genes=3086



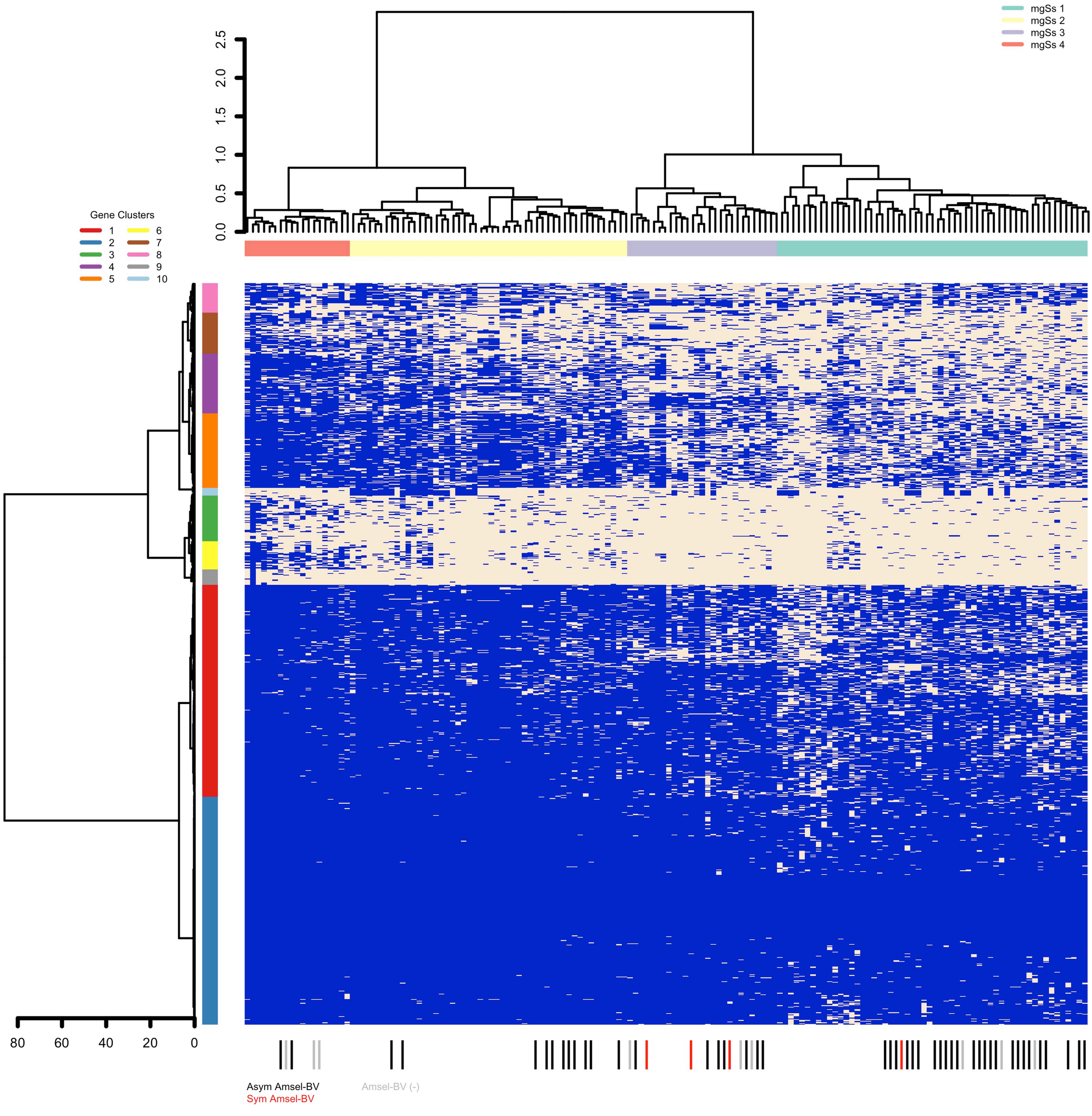
Mobiluncus curtisii mgSs
Number of Samples=78, Number of Genes=2646



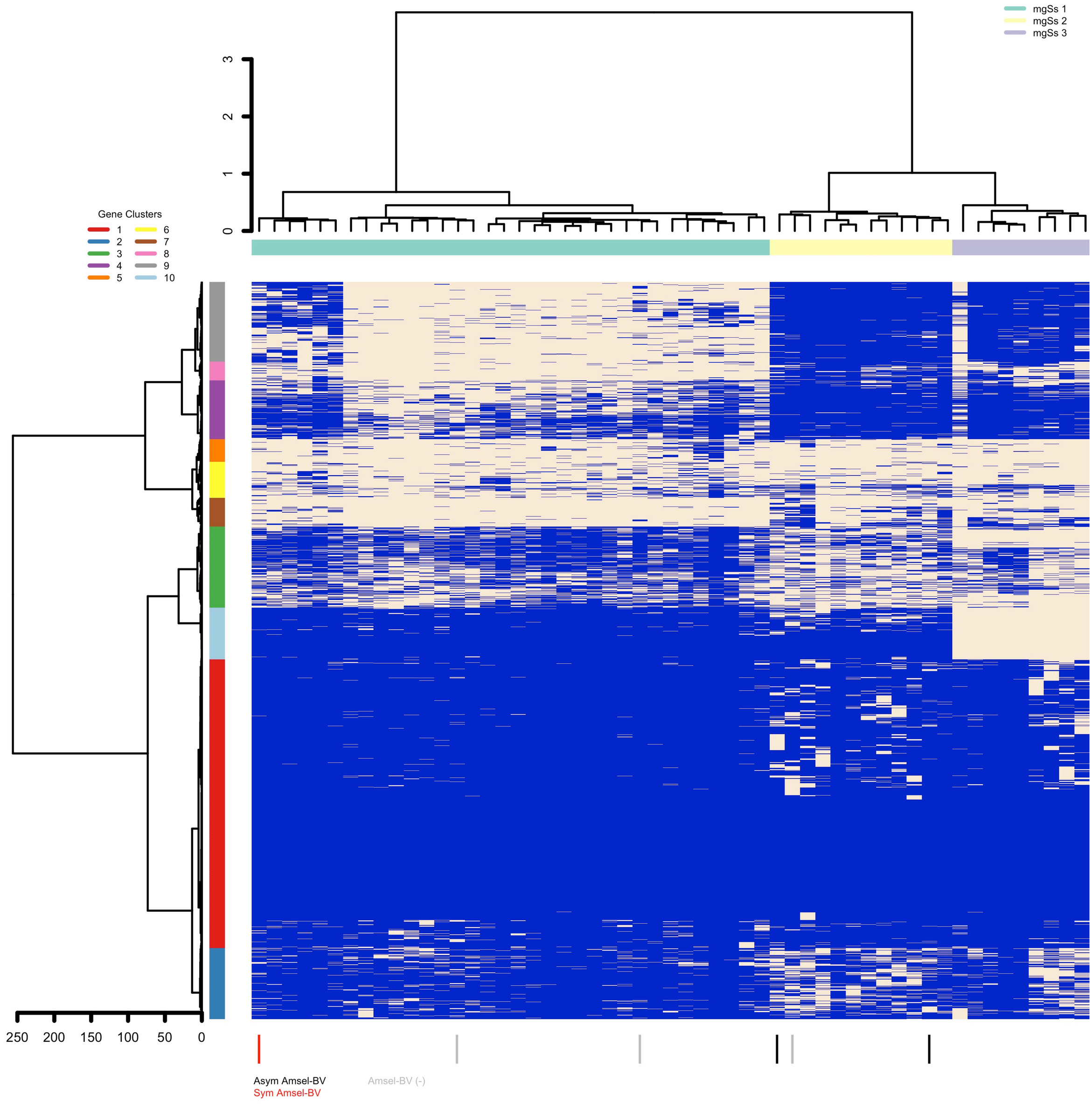
Mobiluncus mulieris mgSs
Number of Samples=205, Number of Genes=3978



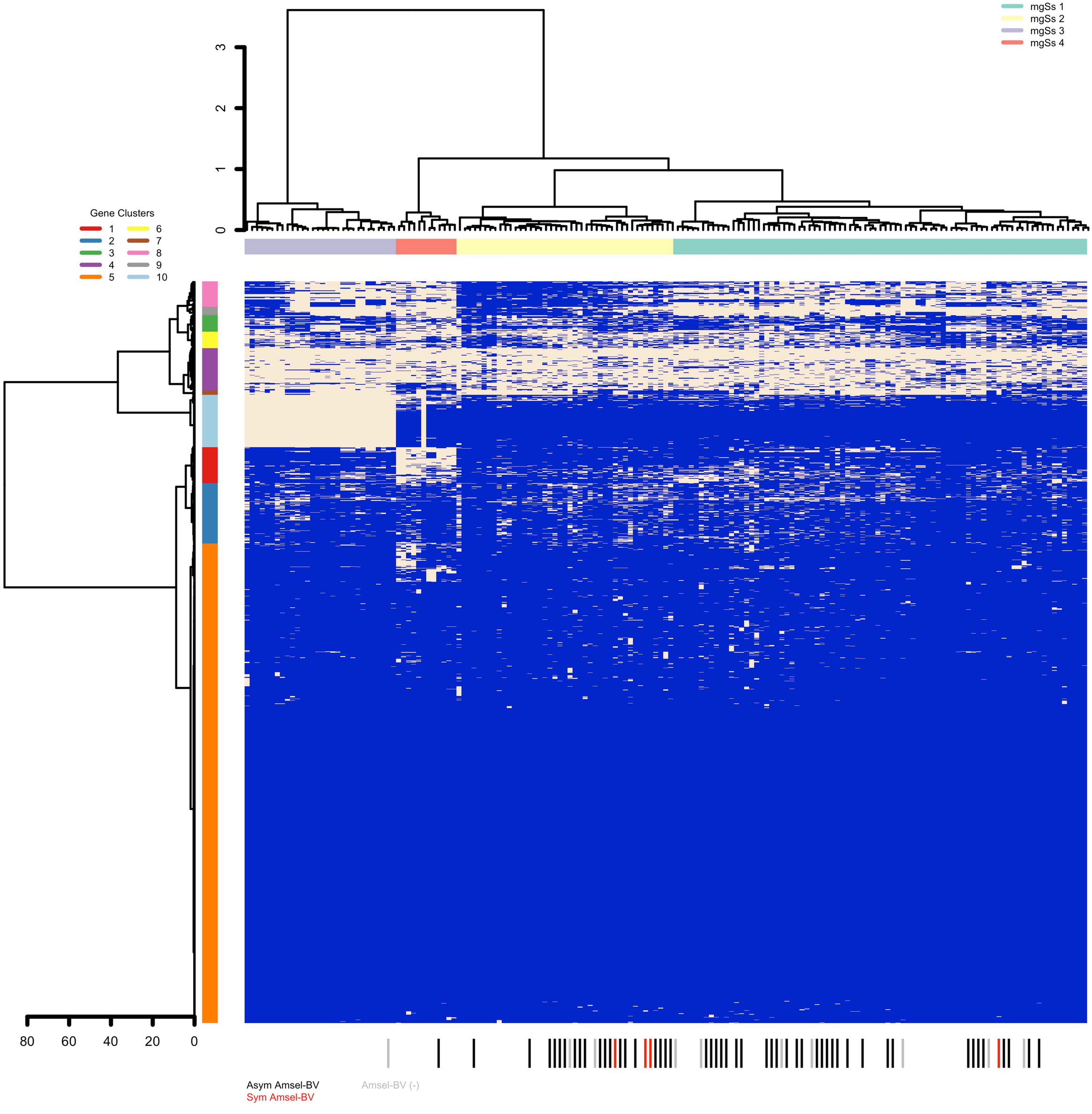
Mycoplasma hominis mgSs
Number of Samples=152, Number of Genes=956



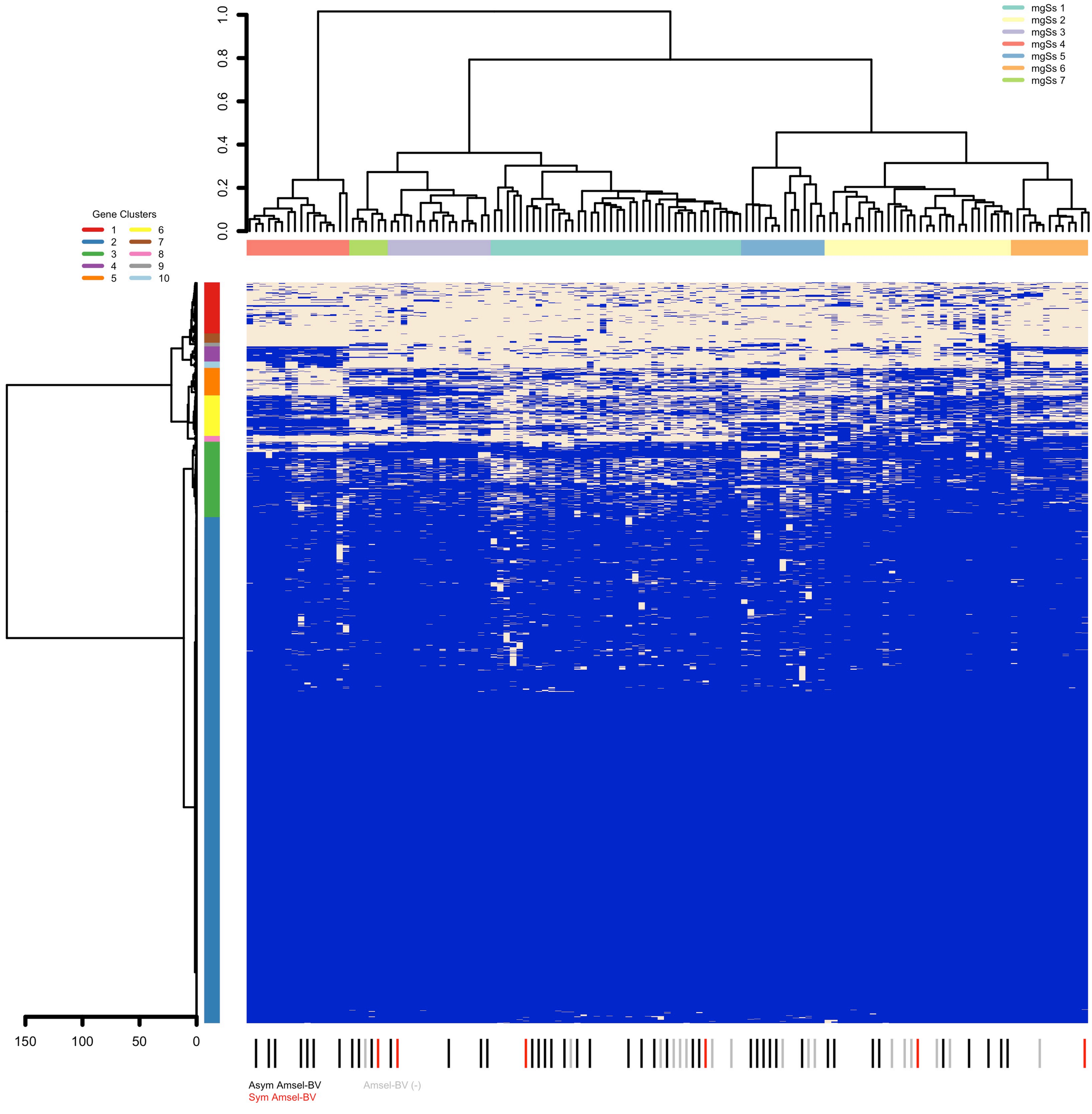
Peptoniphilus harei mgSs
Number of Samples=55, Number of Genes=2604



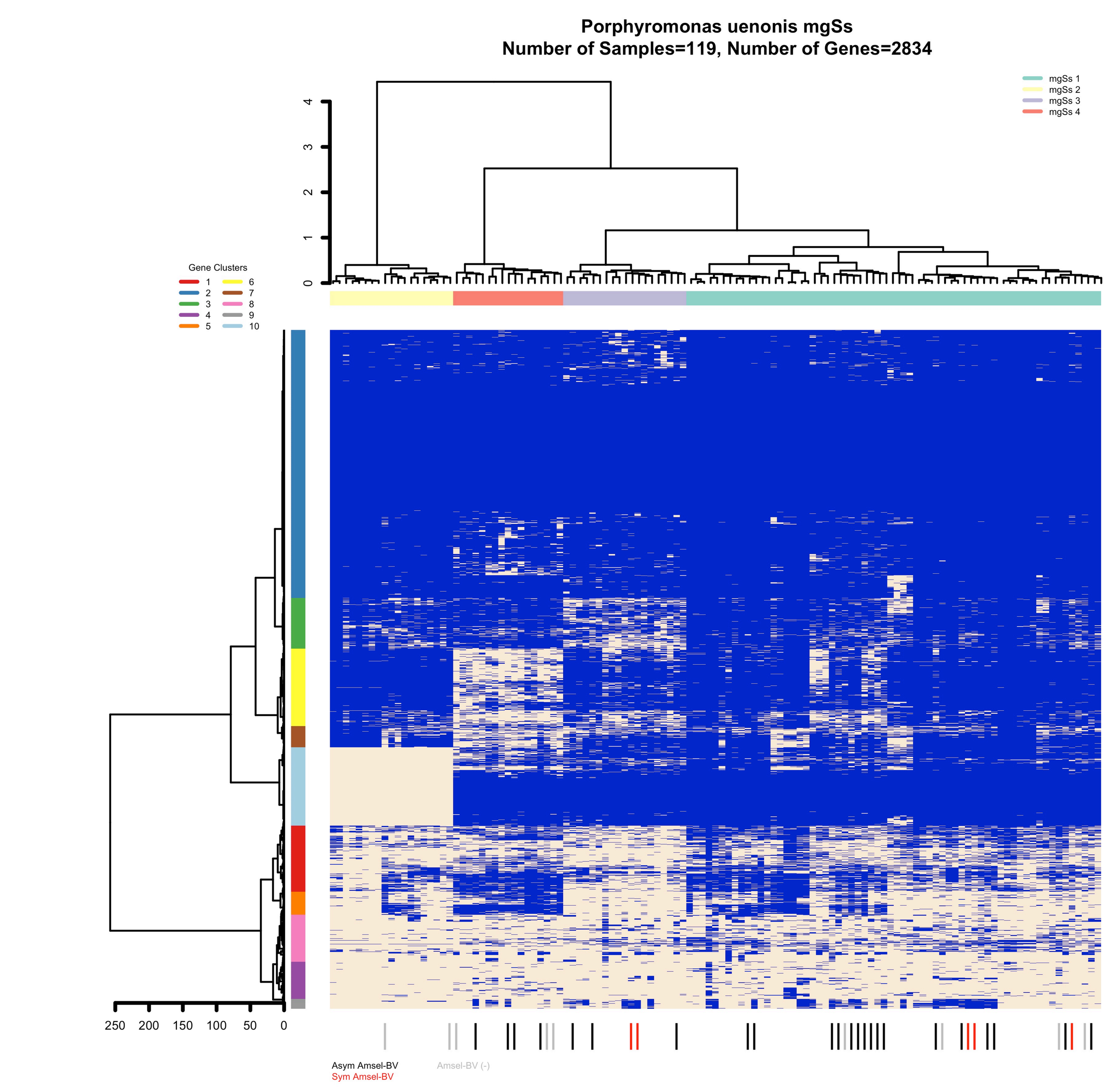
Peptoniphilus lacrimalis mgSs
Number of Samples=167, Number of Genes=1676



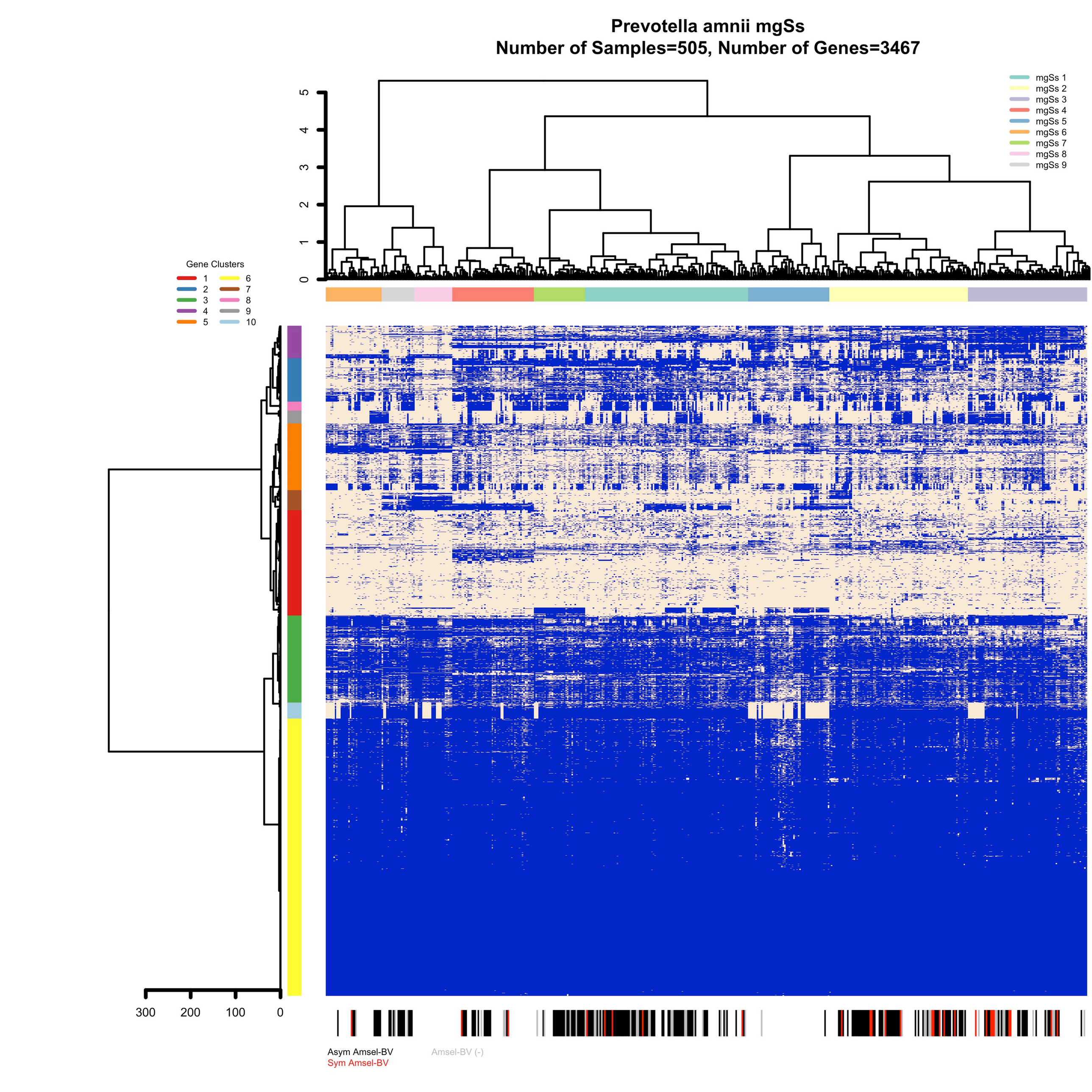
Peptostreptococcus anaerobius mgSs
Number of Samples=131, Number of Genes=2018



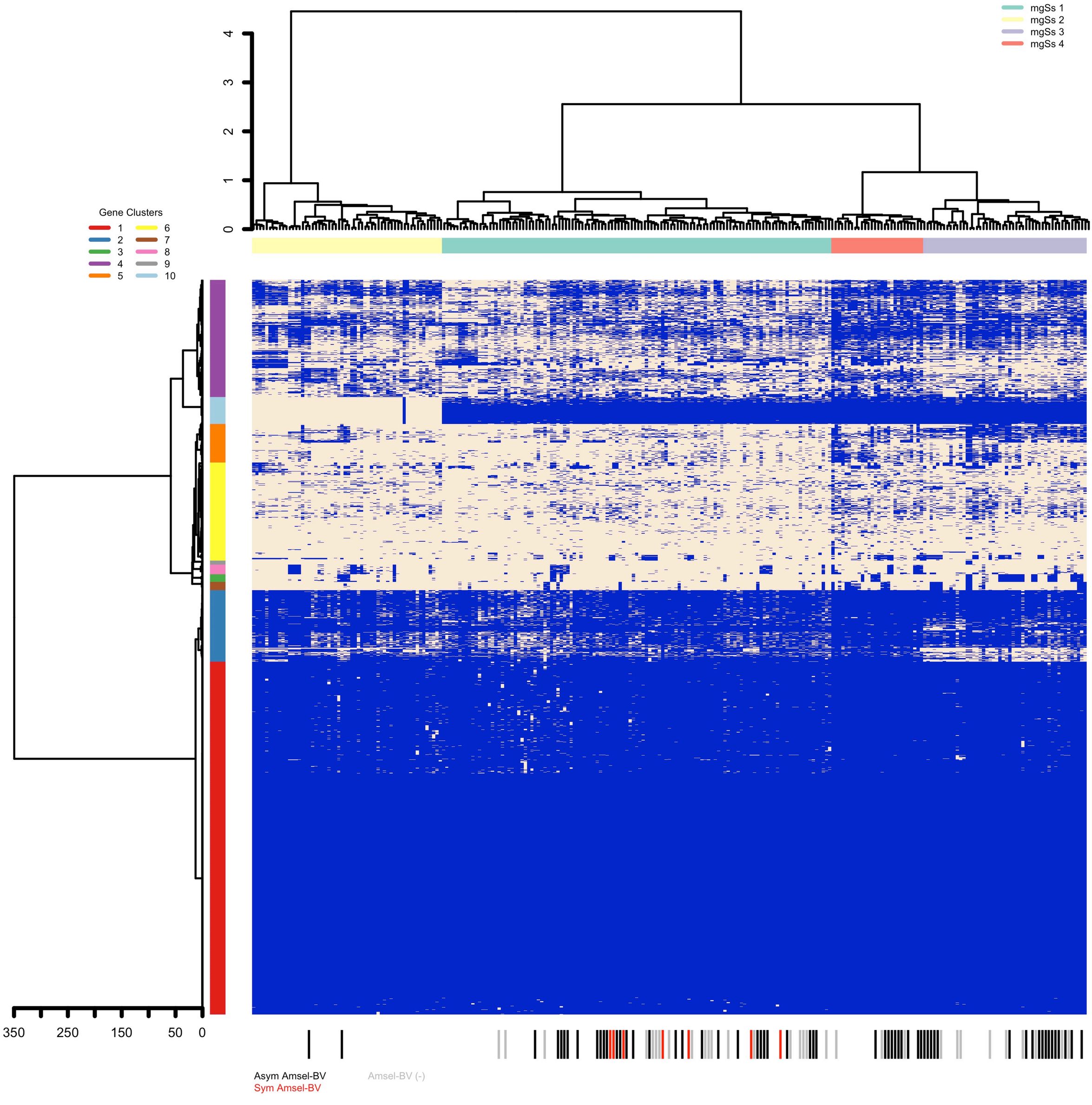
Porphyromonas uenonis mgSs
Number of Samples=119, Number of Genes=2834



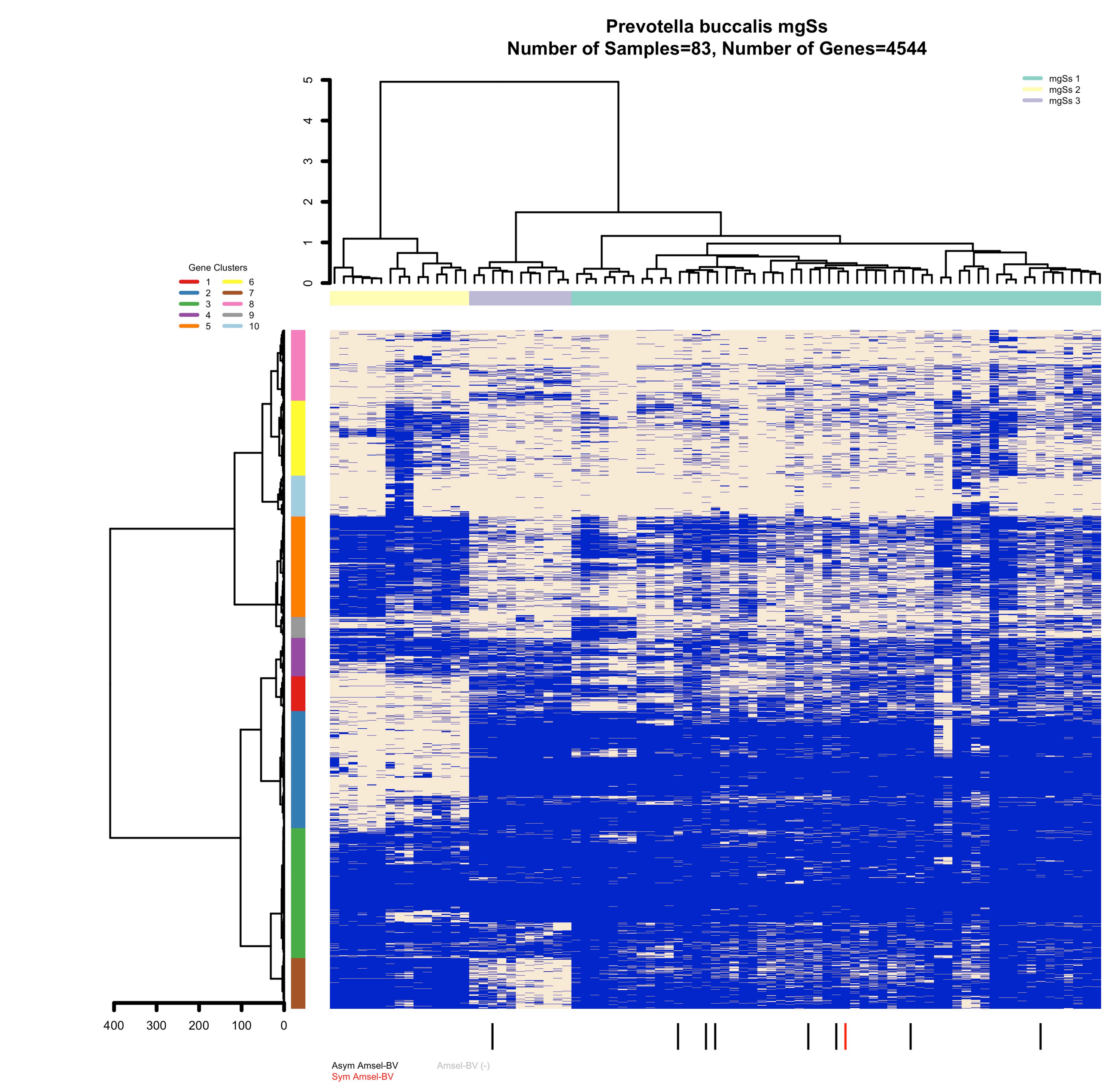
Prevotella amnii mgSs
Number of Samples=505, Number of Genes=3467



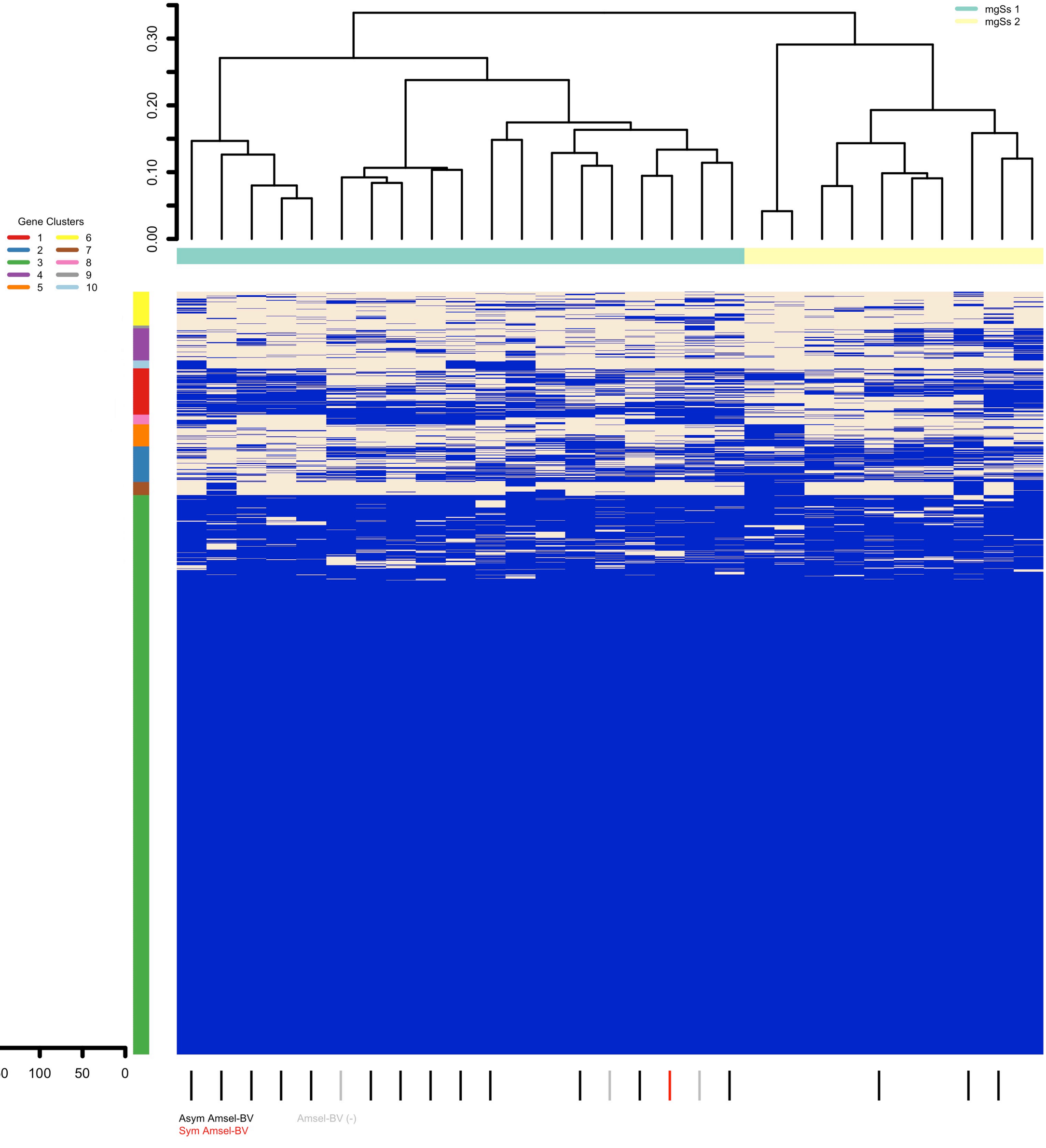
Prevotella bivia mgSs
Number of Samples=255, Number of Genes=2820



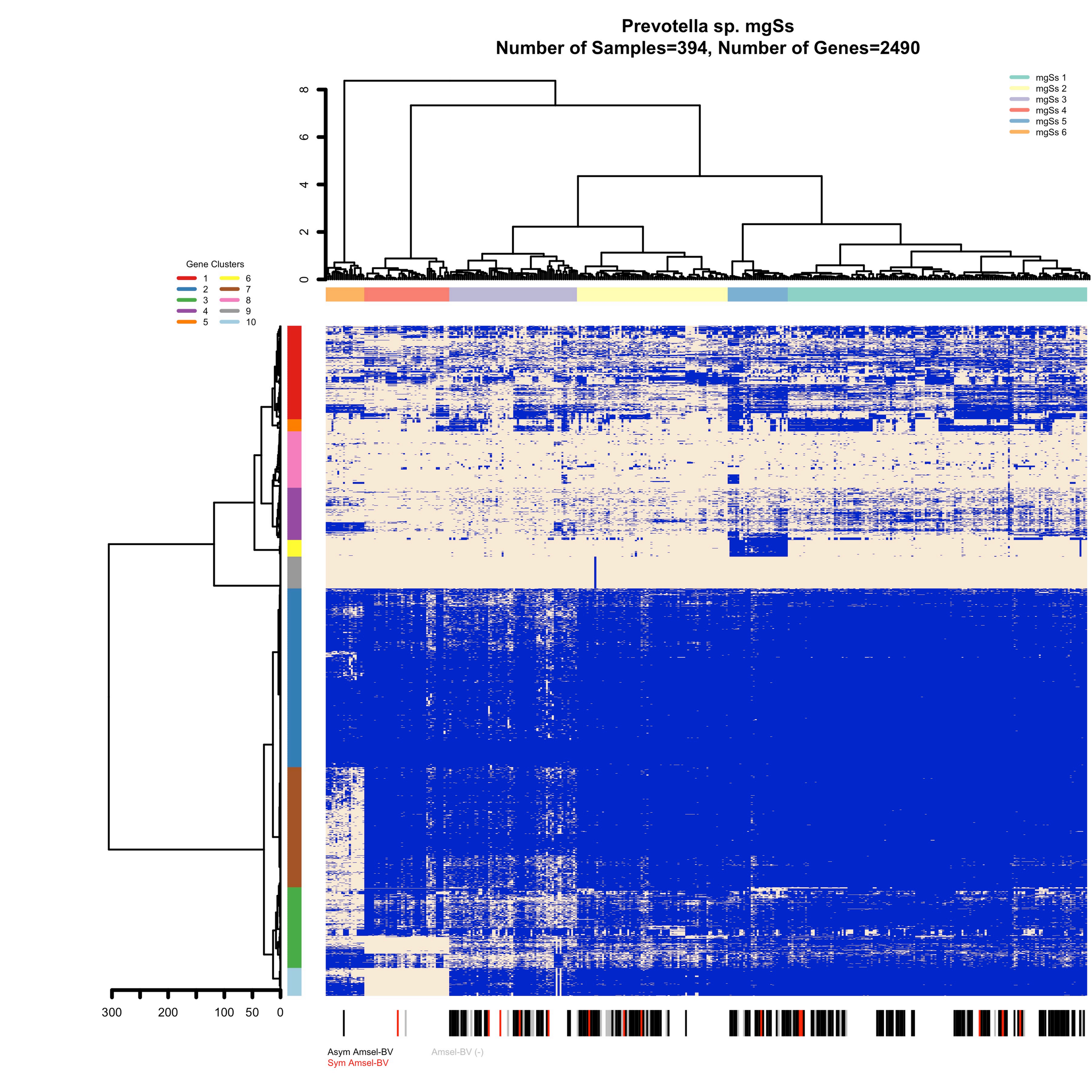
Prevotella buccalis mgSs
Number of Samples=83, Number of Genes=4544



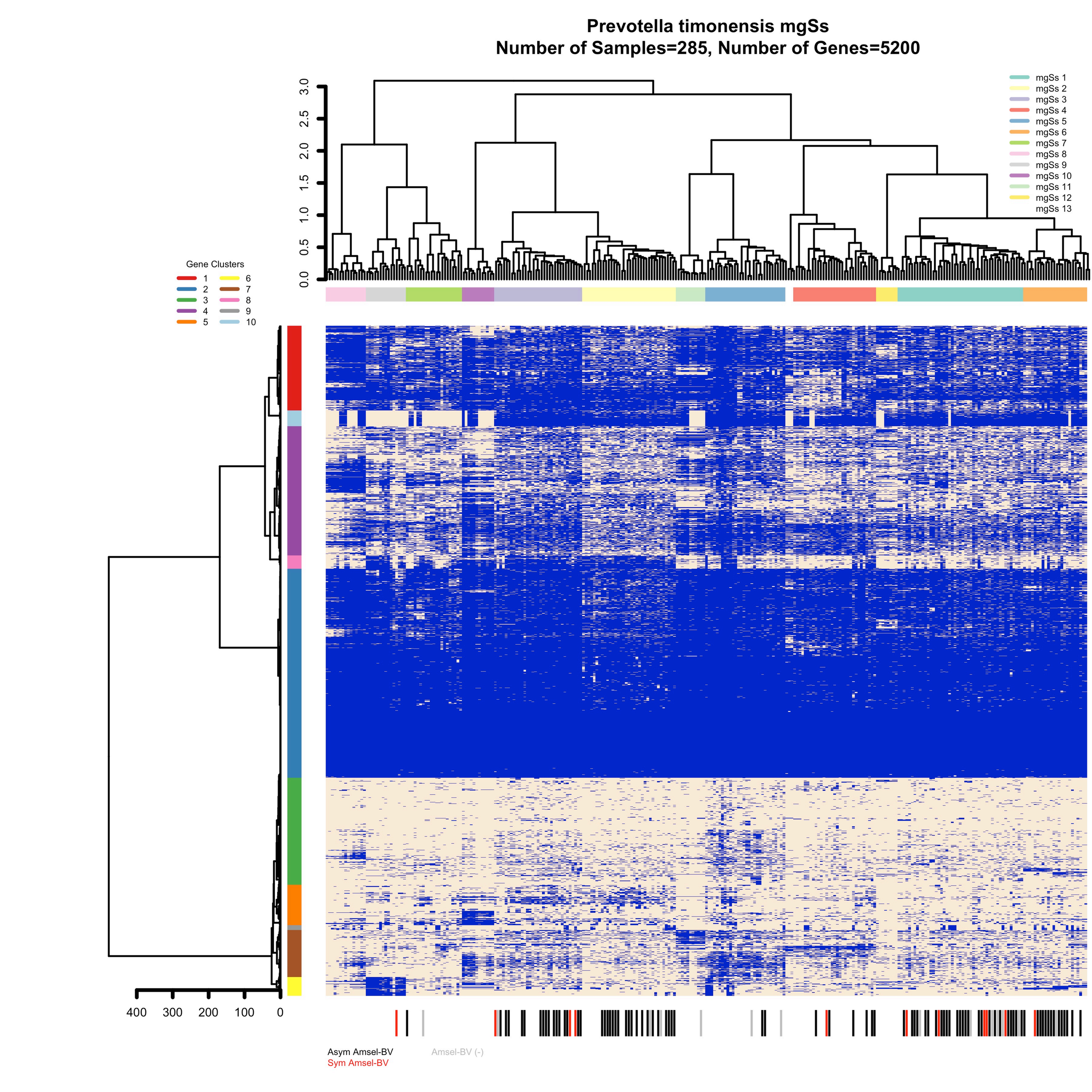
Prevotella disiens mgSs
Number of Samples=29, Number of Genes=2550



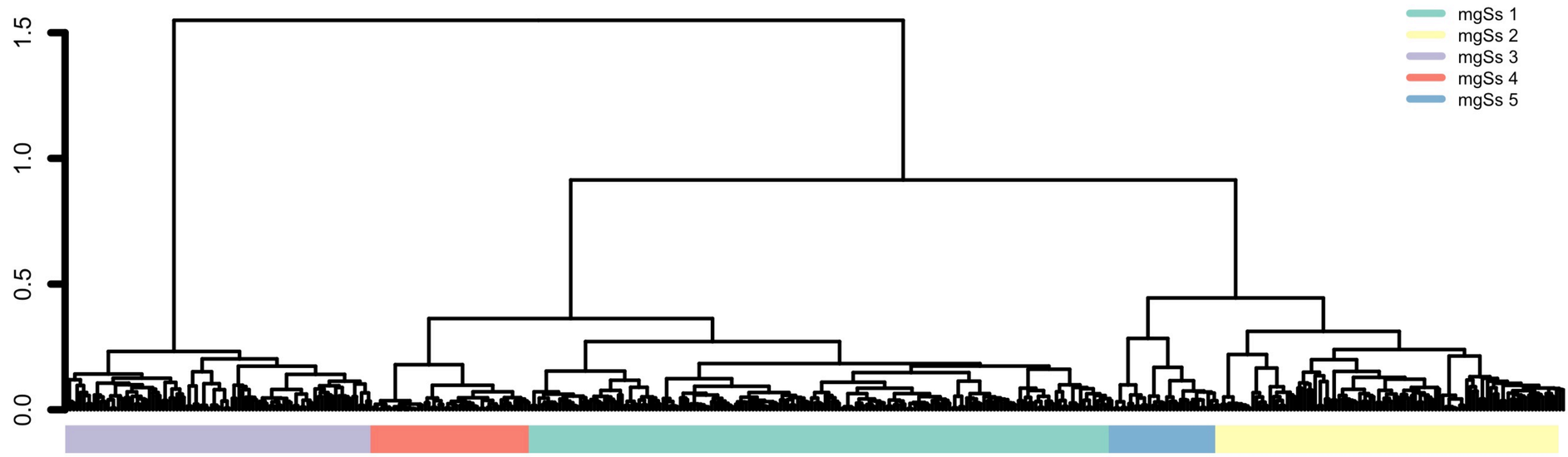
Prevotella sp. mgSs
Number of Samples=394, Number of Genes=2490



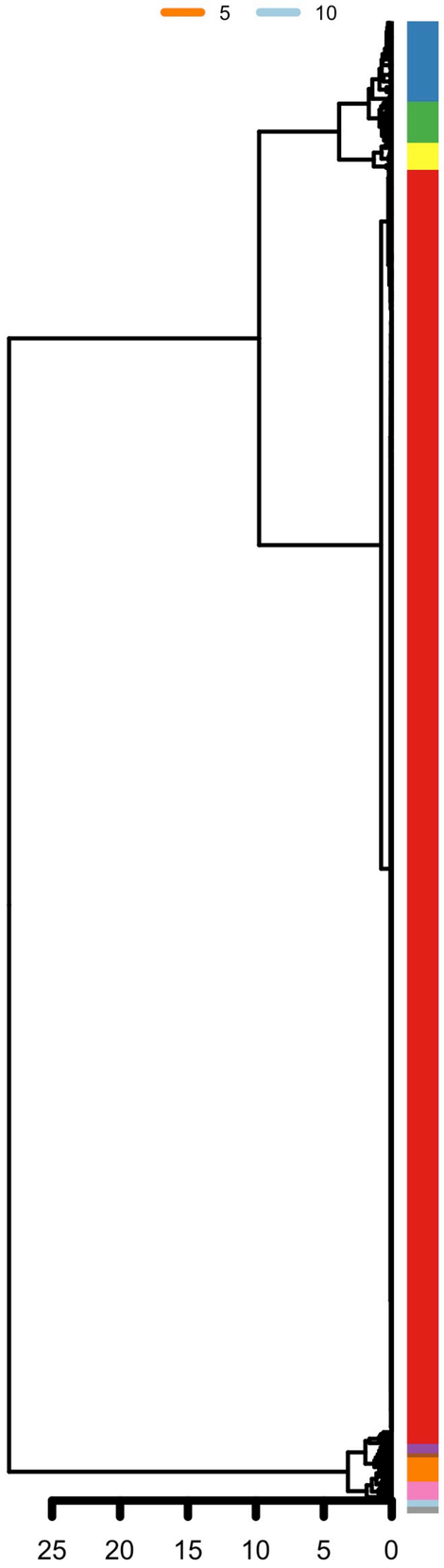
Prevotella timonensis mgSs
Number of Samples=285, Number of Genes=5200

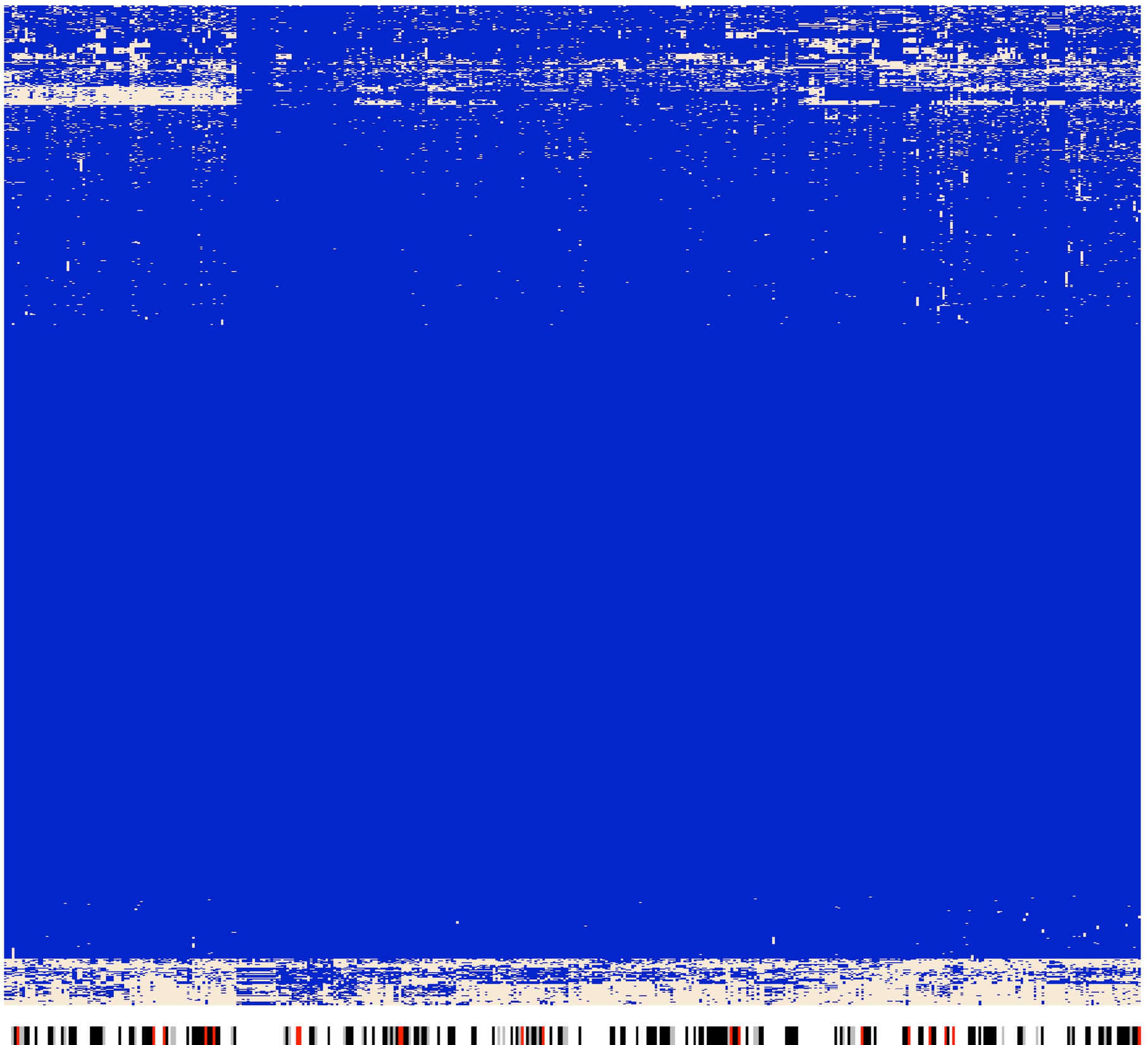


Sneathia amnii mgSs
Number of Samples=435, Number of Genes=1114



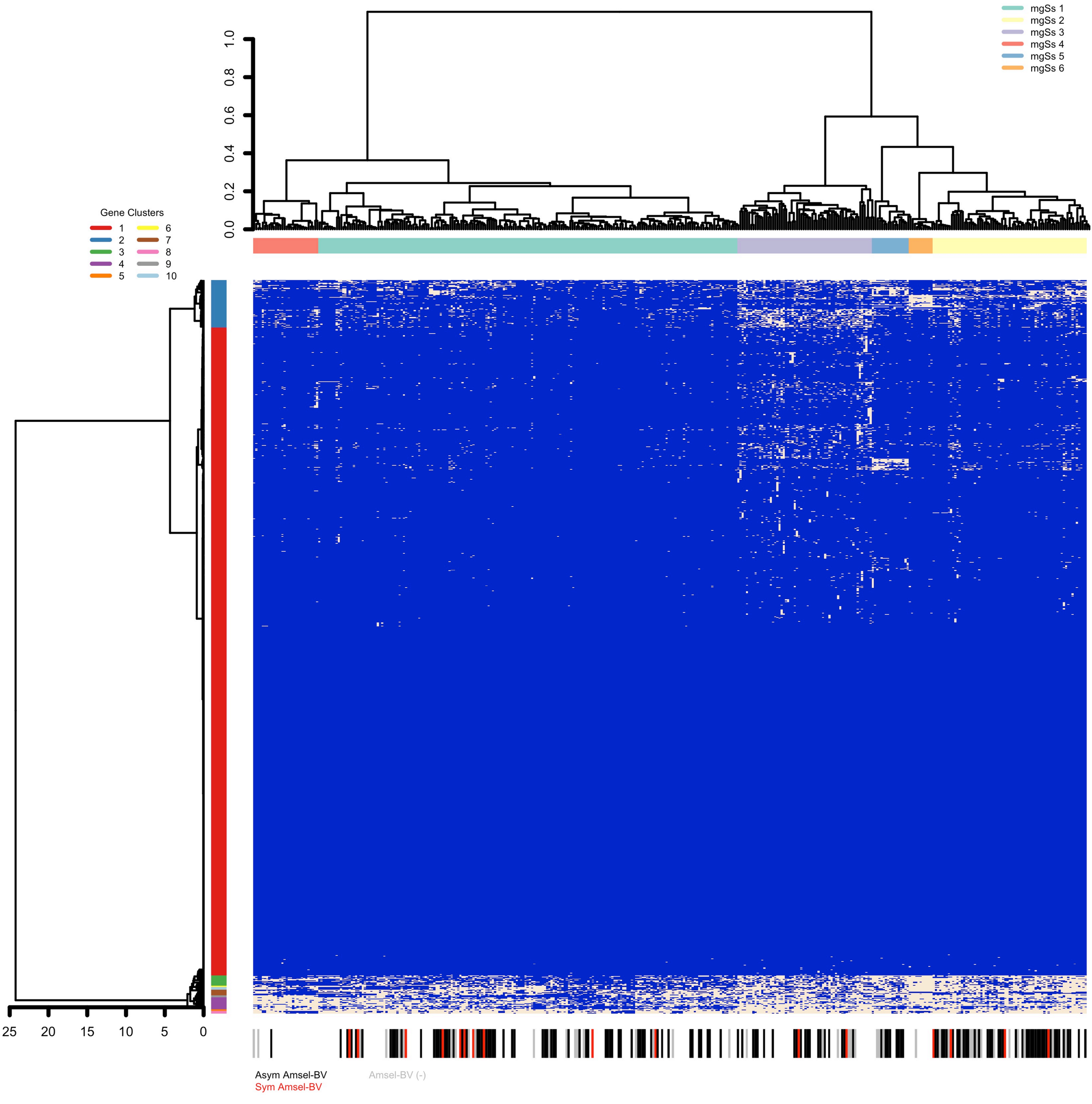
Gene Clusters
1 6
2 7
3 8
4 9
5 10



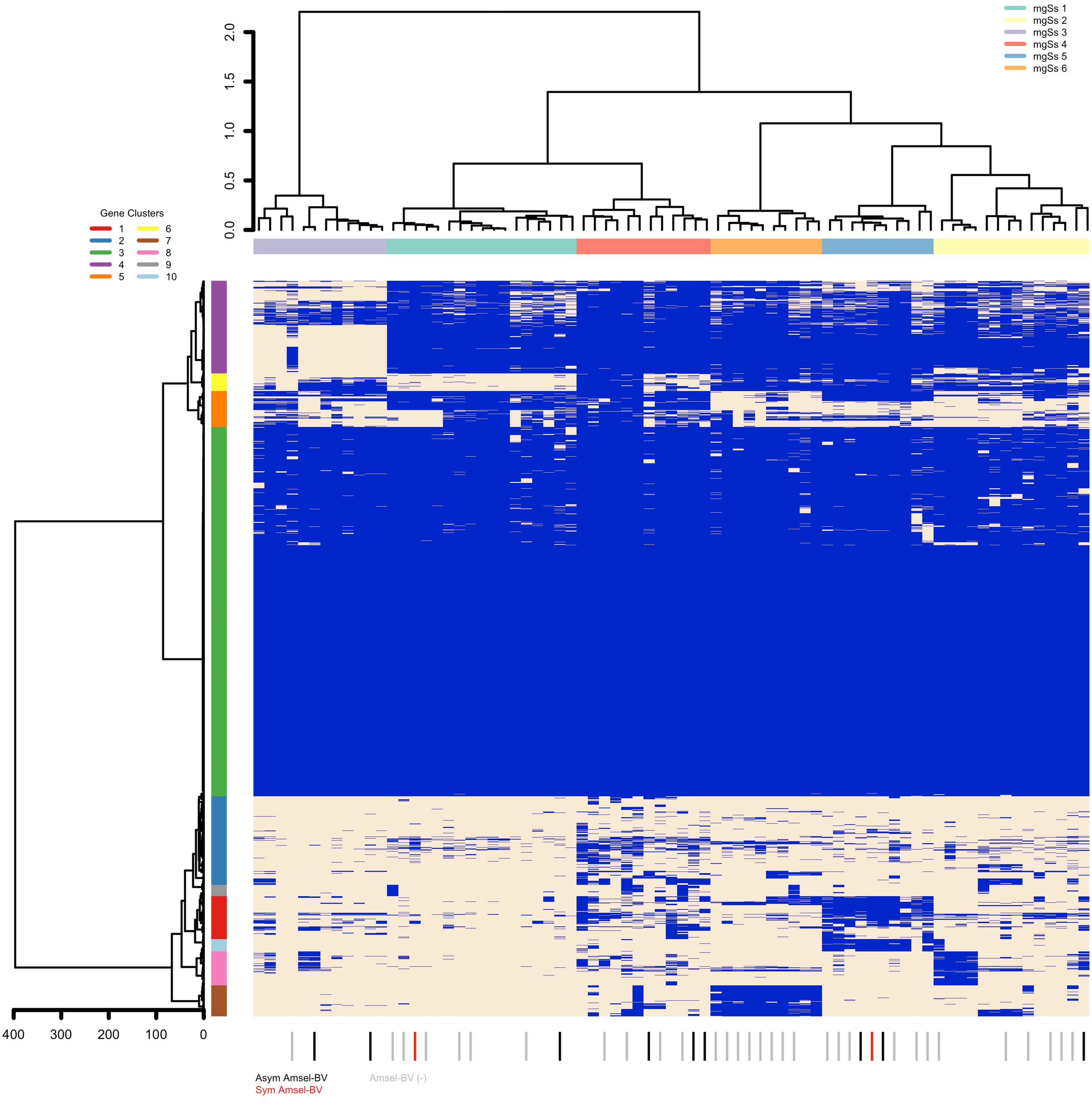


Asym Amsel-BV
Sym Amsel-BV
Amsel-BV (-)

Sneathia sanguinegens mgSs
Number of Samples=384, Number of Genes=1129



Streptococcus agalactiae mgSs
Number of Samples=75, Number of Genes=2905



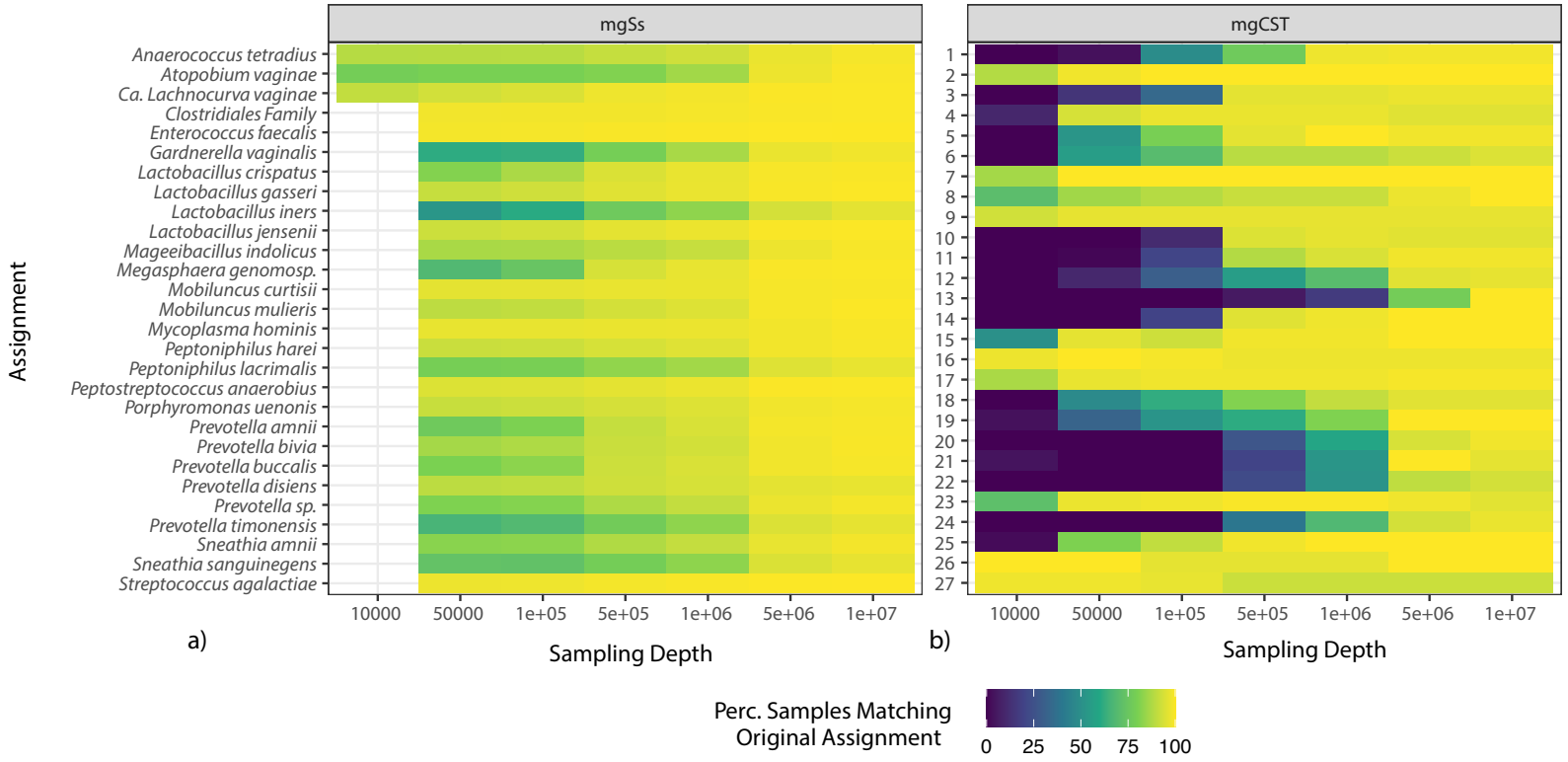


Fig. S10. Metagenomic subspecies (a, mgSs) and metagenomic community state types (b, mgCSTs) may be impacted with sampling (sequencing) depth. A minimum of 1 x 10^6 reads per sample is recommended for mgCST assignment.