Experimental SOM clustering based on an 80% quality threshold

2023-08-04

Running the SOM clustering procedure with our given parameters.

```
#Run SOMs, for the full scaled dataset we have selected a 9x8 grid but other grids ar
e being tested
gridrows=10; gridcols=10; gridsize=gridrows*gridcols
initSOM(dimension=c(gridrows,gridcols),maxit=2000,nb.save=10,scaling="none")
```

```
##
##   Parameters of the SOM
##
##     SOM mode                        :  online
##     SOM type                        :  numeric
##     Affectation type                :  standard
##     Grid                            :
##       Self-Organizing Map structure
##
##         Features   :
##            topology     :  square
##            x dimension  :  10
##            y dimension  :  10
##            distance type:  euclidean
##
##     Number of iterations            :  2000
##     Number of intermediate backups  :  10
##     Initializing prototypes method  :  random
##     Data pre-processing type        :  none
##     Neighbourhood type              :  gaussian
```

```
set.seed(123)
#carve.som<-trainSOM(x.data=scaled_matrix)

#Load in variables to save time when knitting
load("Data/all_carve-som-info_experimental-hq.RData")

#Analyze SOMs
#Use superClass with a chosen k value (manually selected) and then index it to get th
e clusters for each genome
num_clusters=8
#carve.sc<-superClass(carve.som,k=num_clusters)

clusters<-carve.sc$cluster
ids<-carve.sc$som$clustering
sample_clusters<-clusters[ids]
```

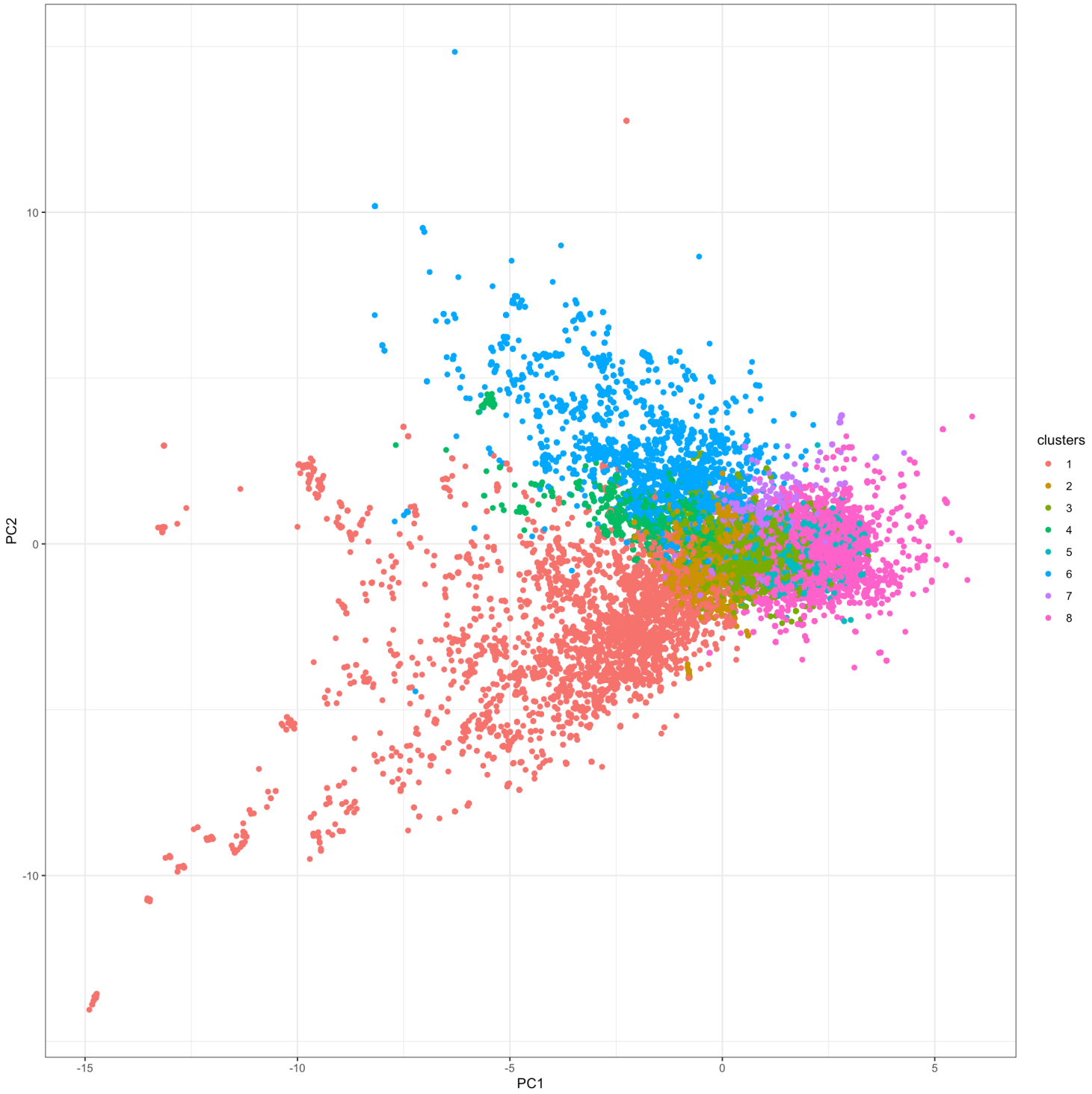
Analyzing the SOM clusters themselves to look for patterns in the data based on the SOM clusters. We also show validations of the clustering here.

```
## [1] "Dupe found in genome GORG_SAMEA6075158_SAGS_AG349D08"
## [1] "Dupe found in genome MARD_SAMN04488092_REFG_MMP04488092"
```

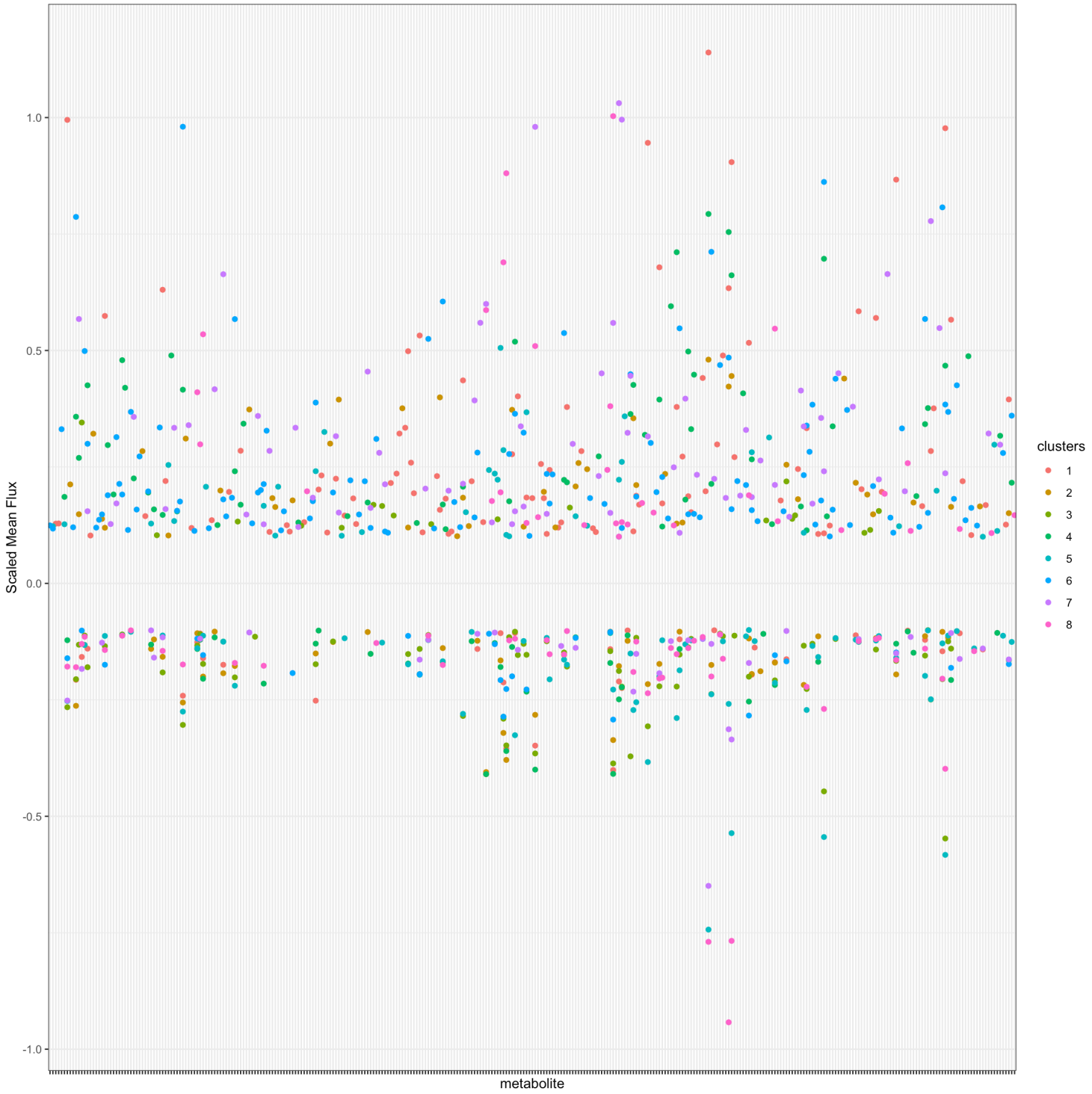
```
## [1] "MARD_SAMN05216212_REFG_MMP05216212 has 2 SOM clusters with 30 models each."
## [1] "MARD_SAMN06948887_REFG_MMP06948887 has 2 SOM clusters with 27 models each."
```

```
## [1] "TARA_SAMEA4398370_METAG_EHKIJLNC has 2 SOM grid points with 30 models each."
## [1] "MARD_SAMN10411098_REFG_MMP10411098 has 2 SOM grid points with 20 models each.
"
## [1] "MARD_SAMN09437463_REFG_MMP09437463 has 2 SOM grid points with 18 models each.
"
## [1] "MARD_SAMN09762594_REFG_MMP09762594 has 2 SOM grid points with 28 models each.
"
## [1] "MARD_SAMN05660413_REFG_MMP05660413 has 2 SOM grid points with 30 models each.
"
## [1] "TARA_SAMEA4397174_METAG_KJLHPIFN has 2 SOM grid points with 15 models each."
## [1] "TARA_SAMEA2591122_METAG_BNKOMGJI has 3 SOM grid points with 12 models each."
```

```
## [1] "Figure: Traditional PCA of data colored by SOM clusters"
```

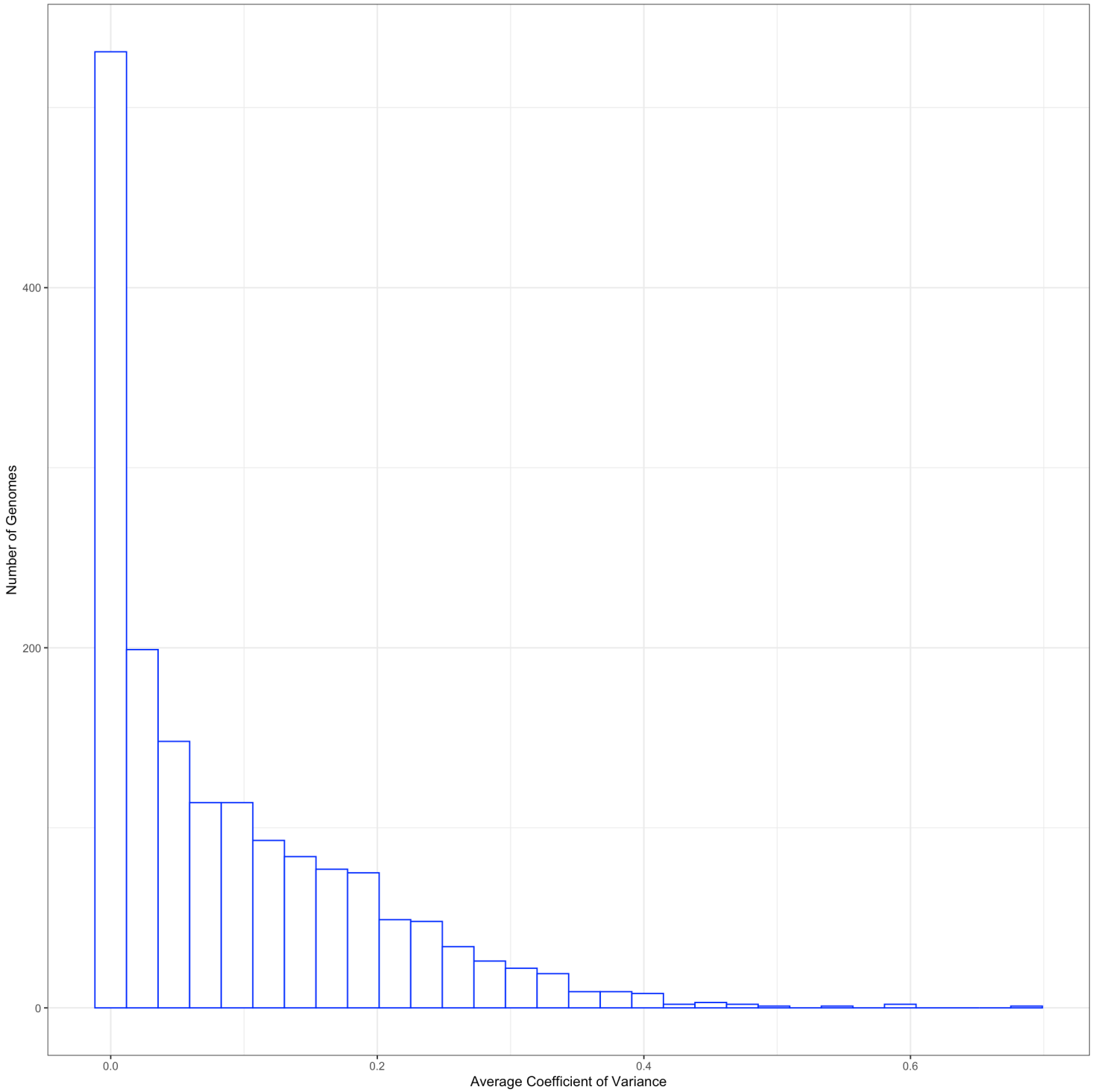



```
## [1] "Figure: Scaled mean flux values for metabolites with >10% flux relative to me
an colored by cluter"
```

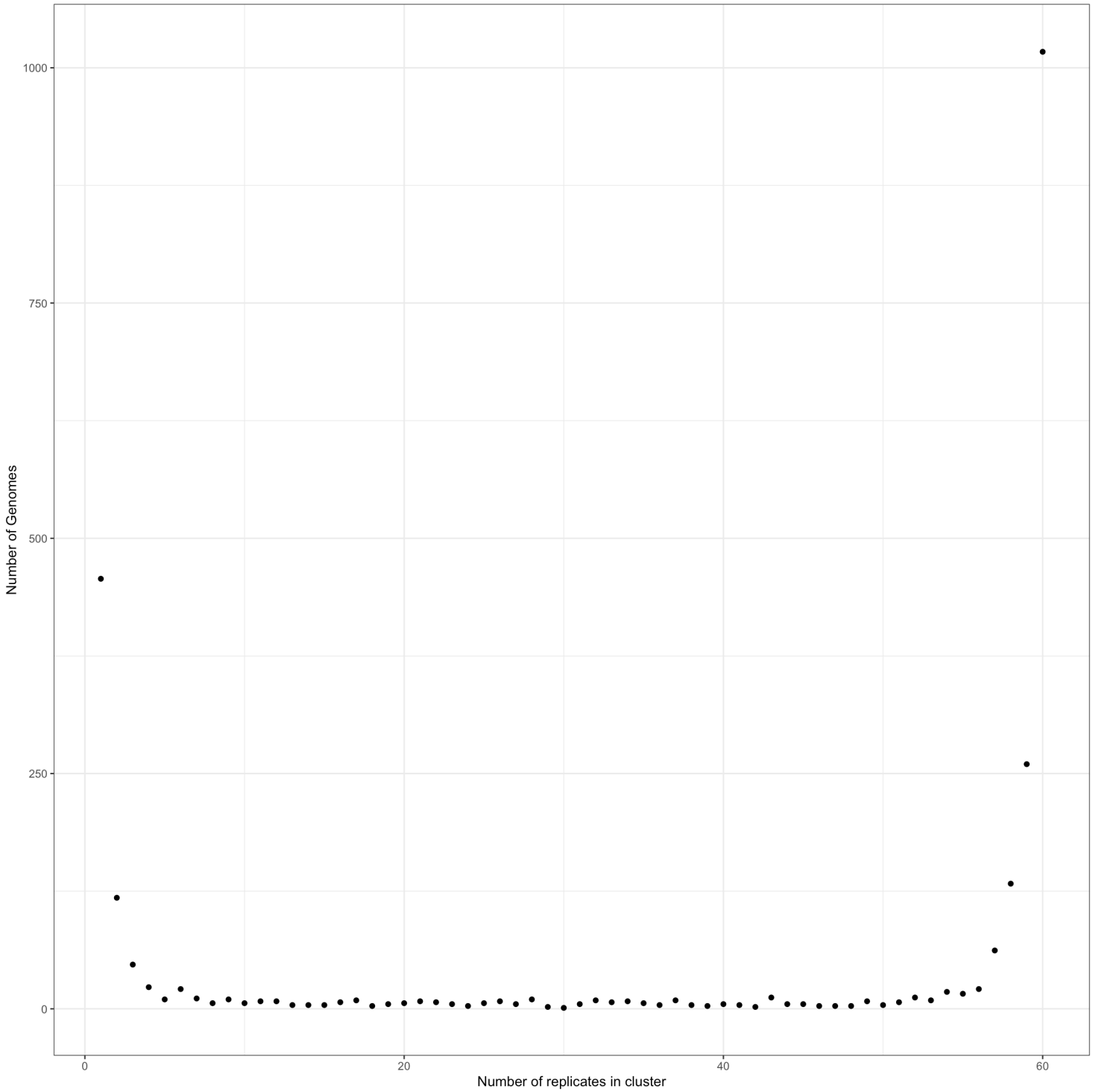



```
## [1] "Figure: Scaled mean flux versus scaled mean standard deviation for metabolite
s with >10% flux relative to mean colored by cluster"
```


```
## [1] "Figure: Distribution of coefficients of variance for each genome ensemble bas
ed on non-zero metabolite flux values for all ensemble models"
```




```
## [1] "Figure: Distribution of counts of models in each cluster per genome."
```




```
## [1] "Figure: Custom defined breadth, consensus, and evenness metrics for the distr
ibution of each genome ensemble's models across clusters"
```

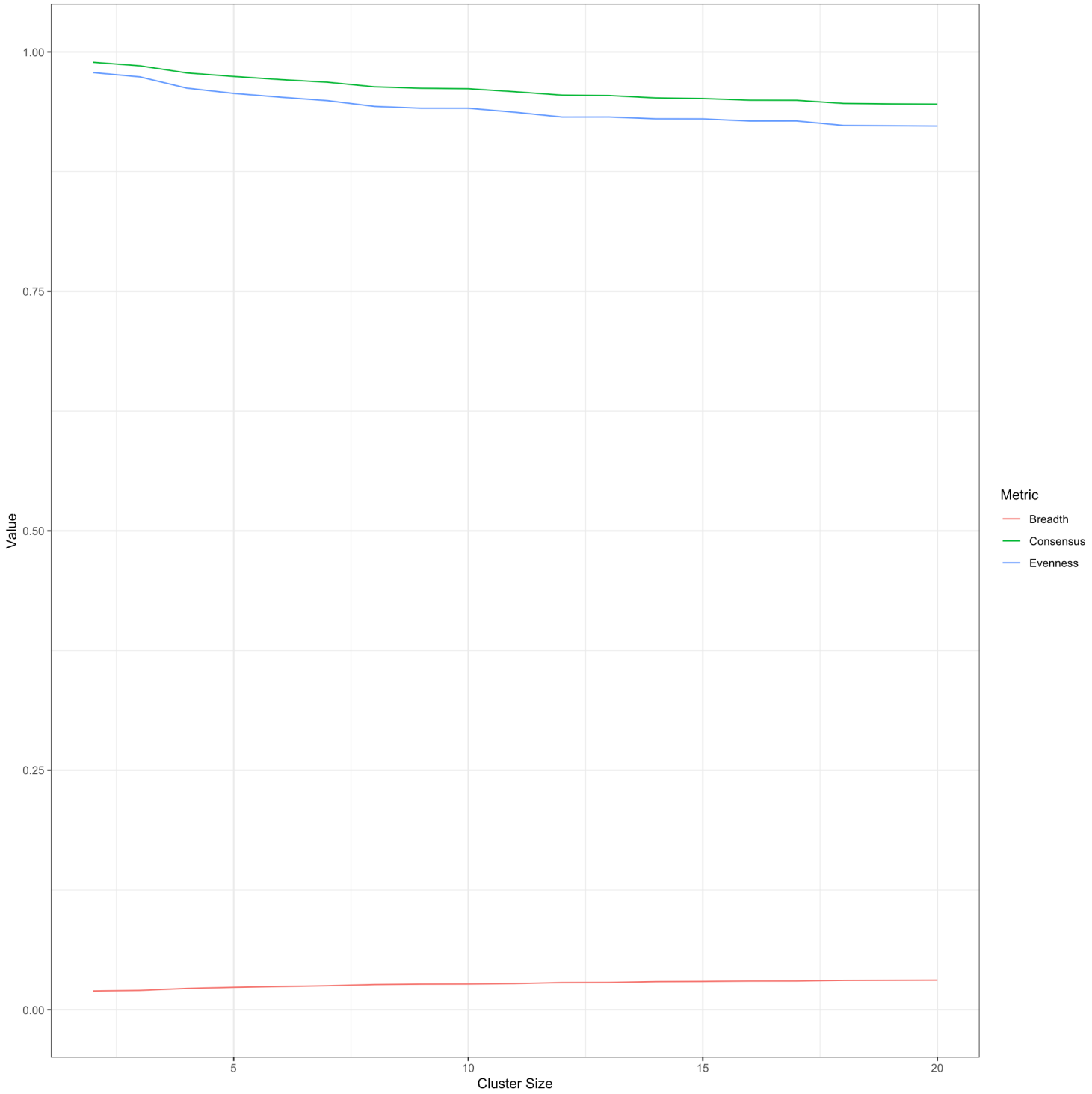
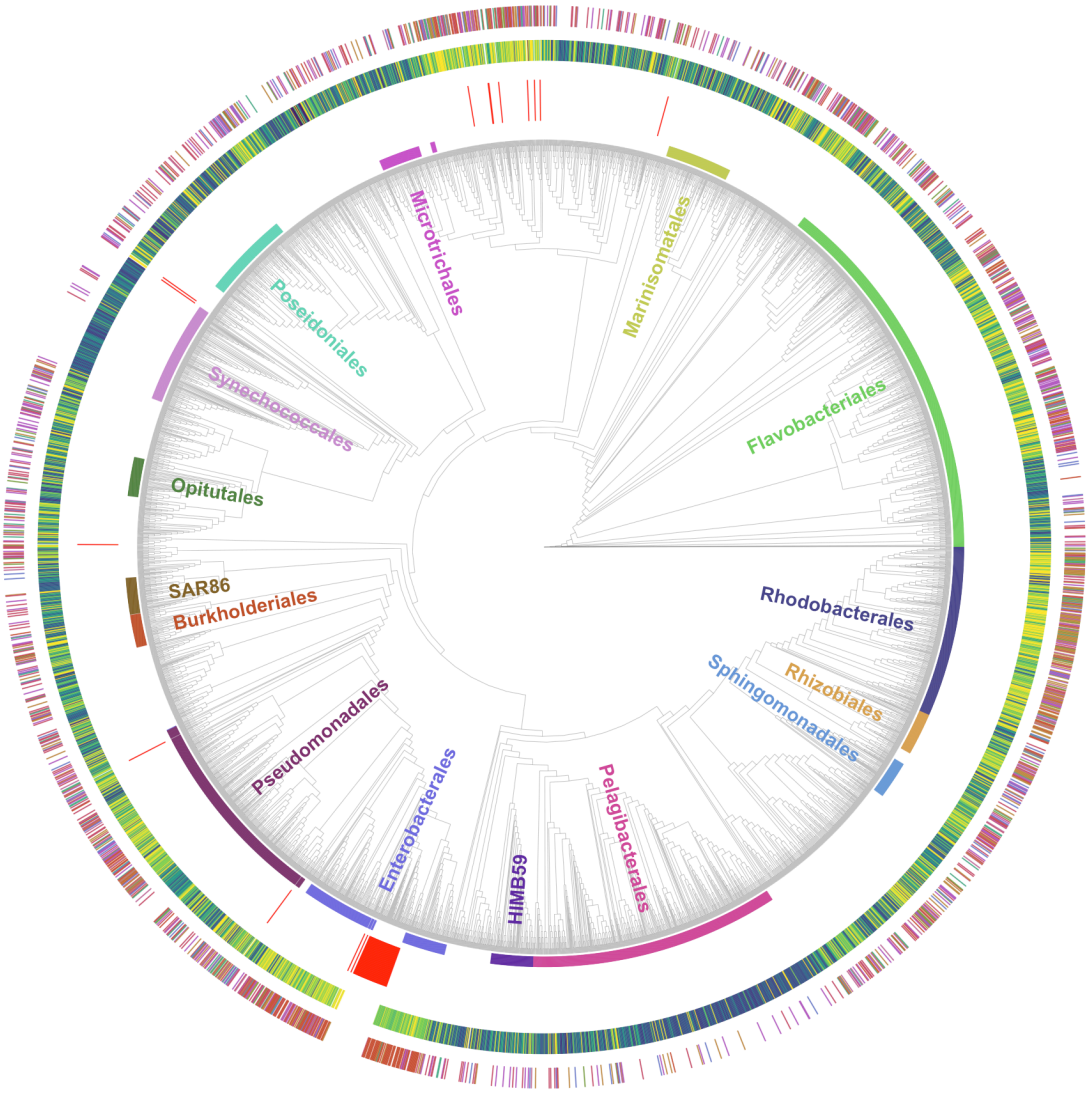
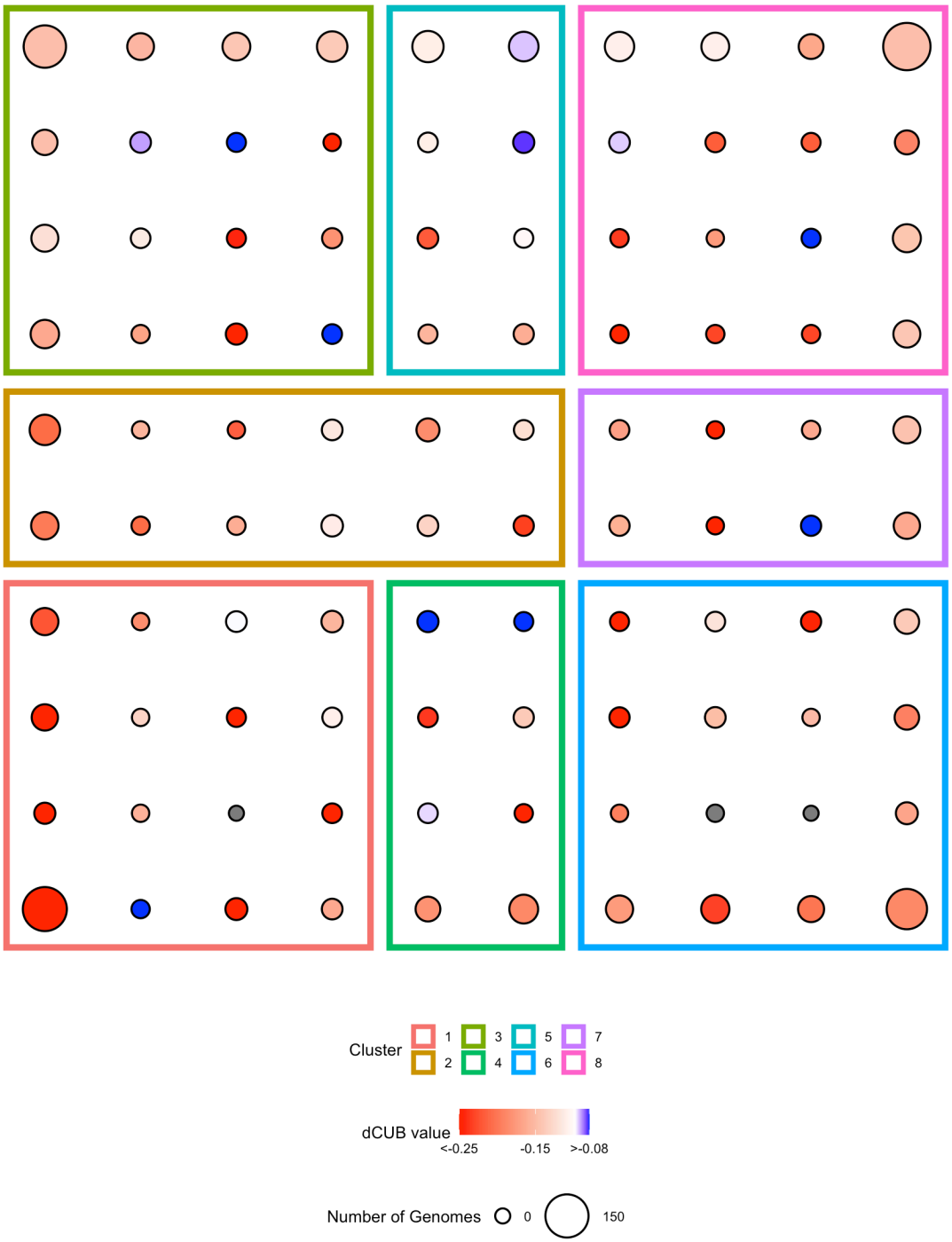


Figure 1: tree with SOM info, phylogeny, major groups, and quality score overlaid on top

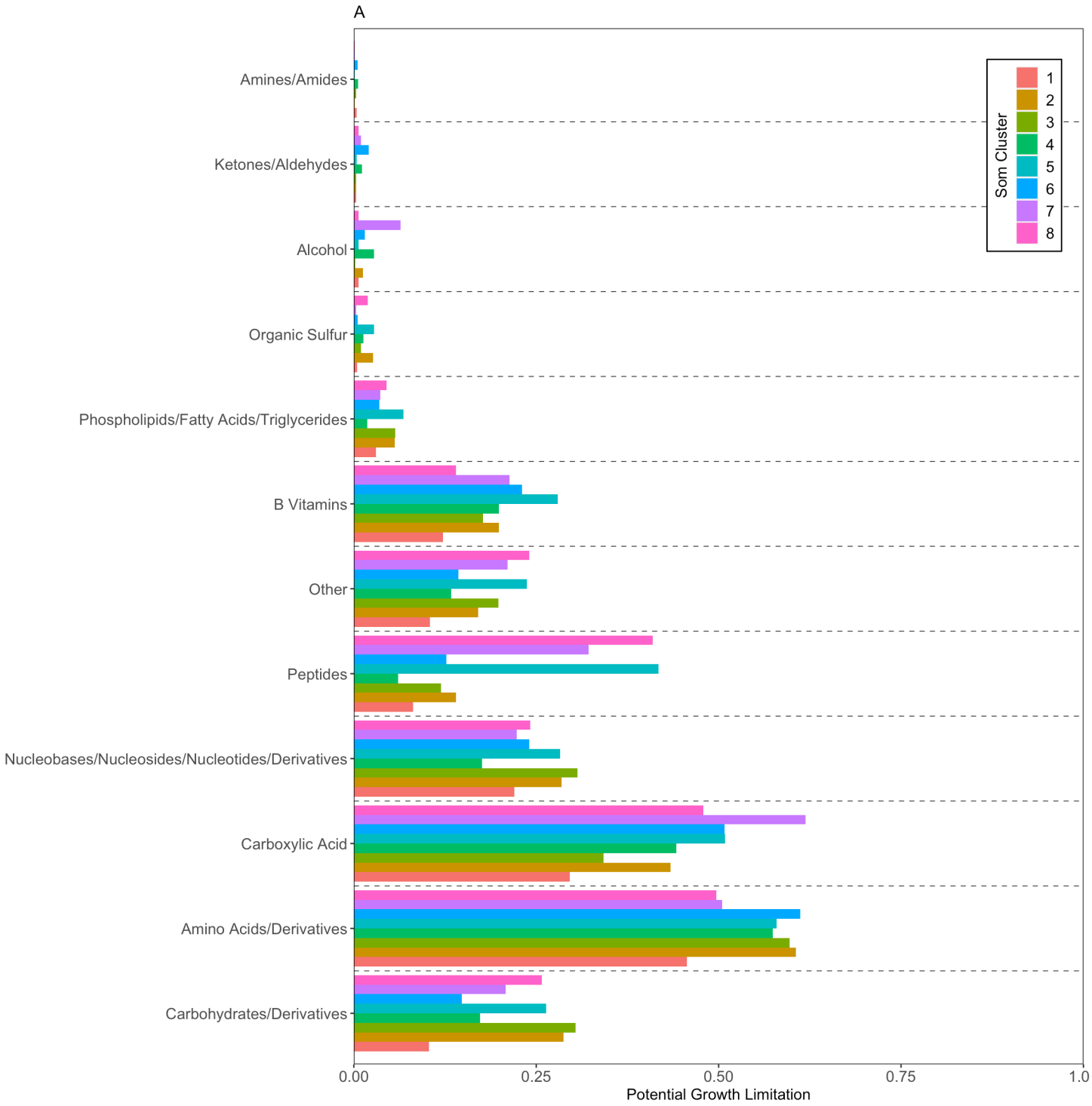


Leveraging the dCUB (growth) data from JL and co. to look for patterns in the SOM clustering based on putative growth rates.

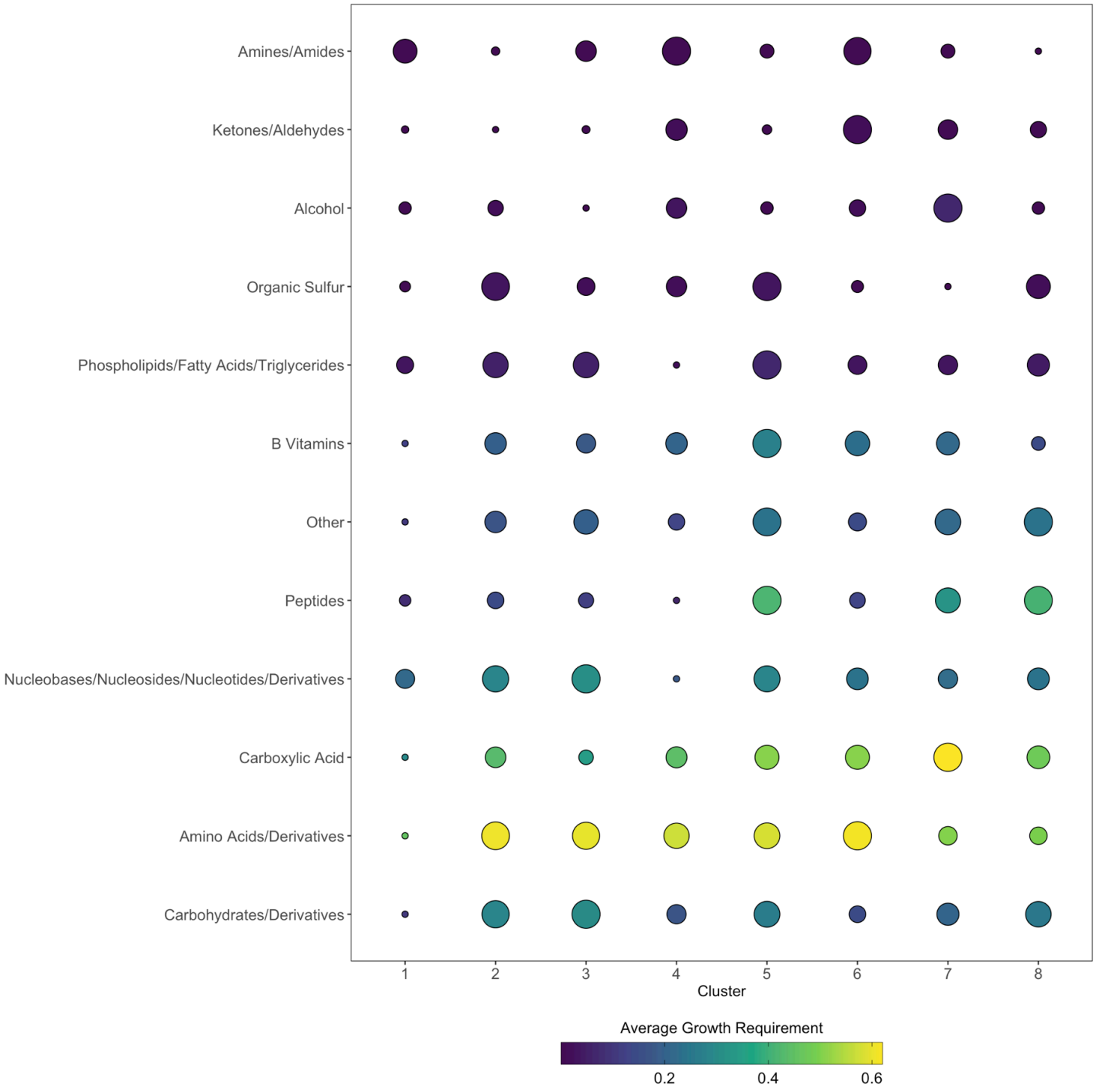
```
## [1] "Figure: SOM grid grouped by cluster (squares), colored by dCUB, and sized bas
ed on number of genomes associated with each grid point."
```



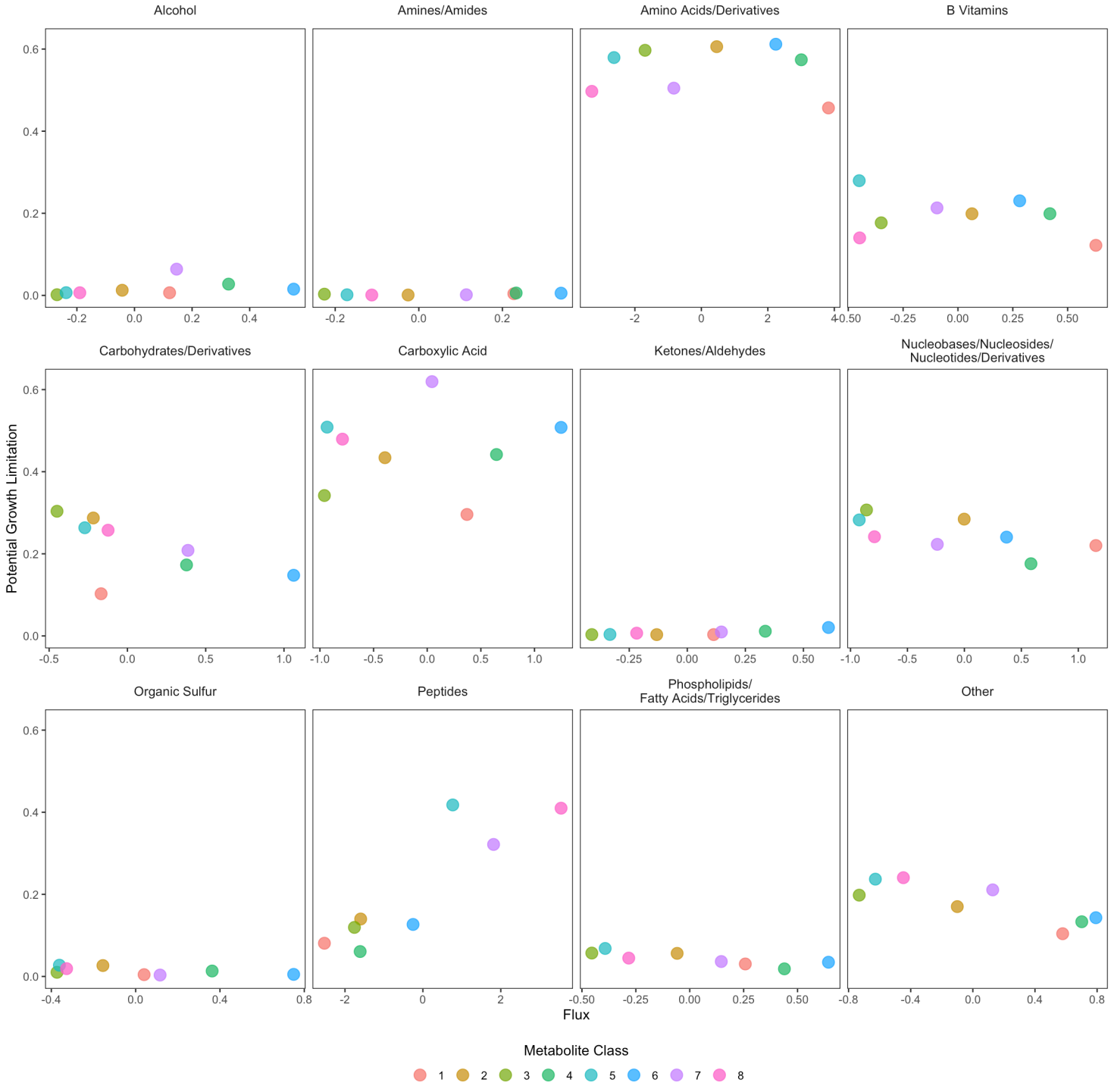

```
## [1] "Figure: Potential growth limitation by each metabolite class colored by SOM c
luster."
```



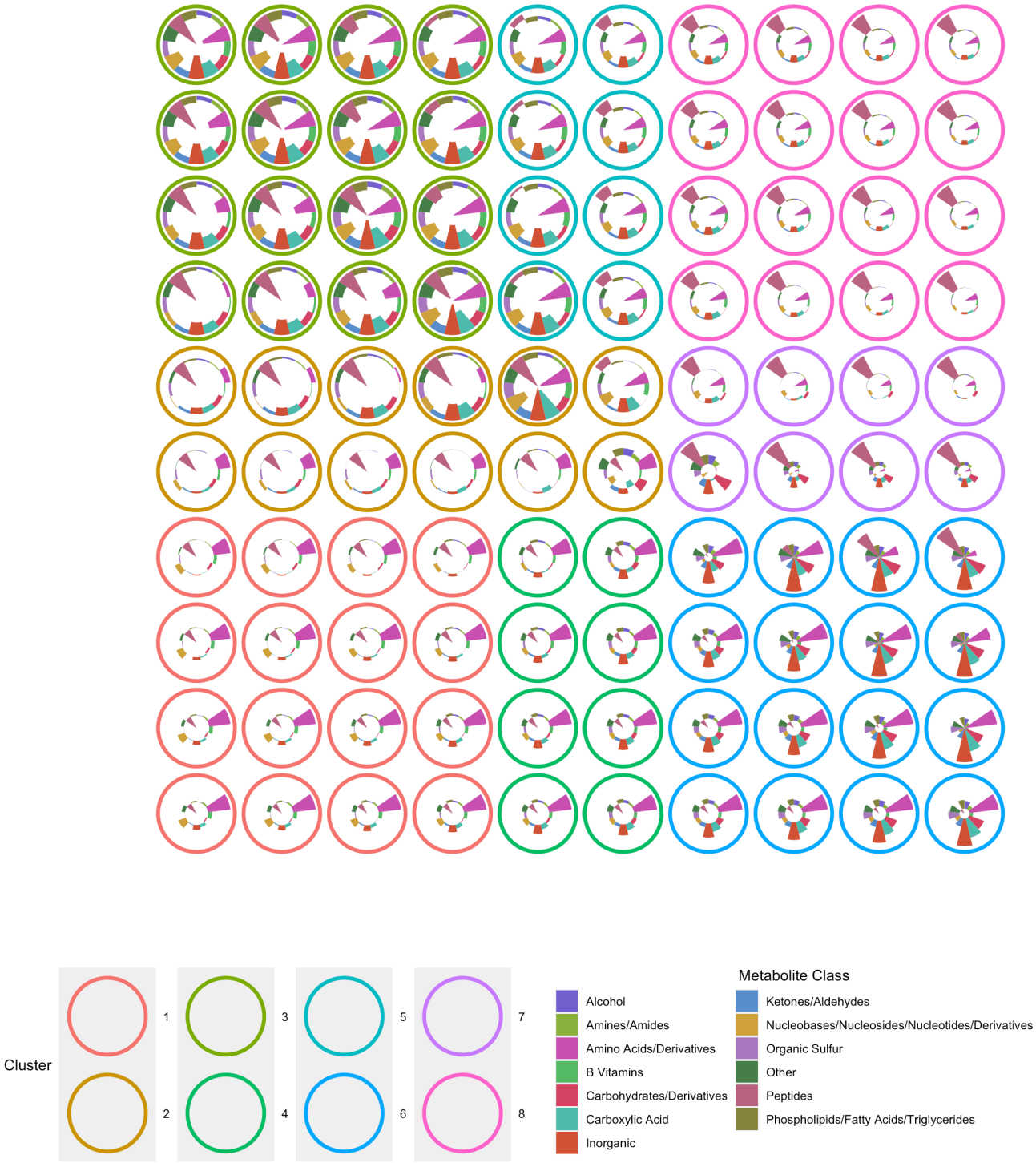

```
## [1] "Figure: Alternative visualization of potential growth limitation (avg growth
req) by metabolite class per cluster. Bubble colors reflect information shown in prev
ious figure, while bubble sizes describe the relative importance of the limitation by
each metabolite class relatively between clusters (i.e., larger bubble implies more i
mportant for that cluster relative to other clusters)."
```



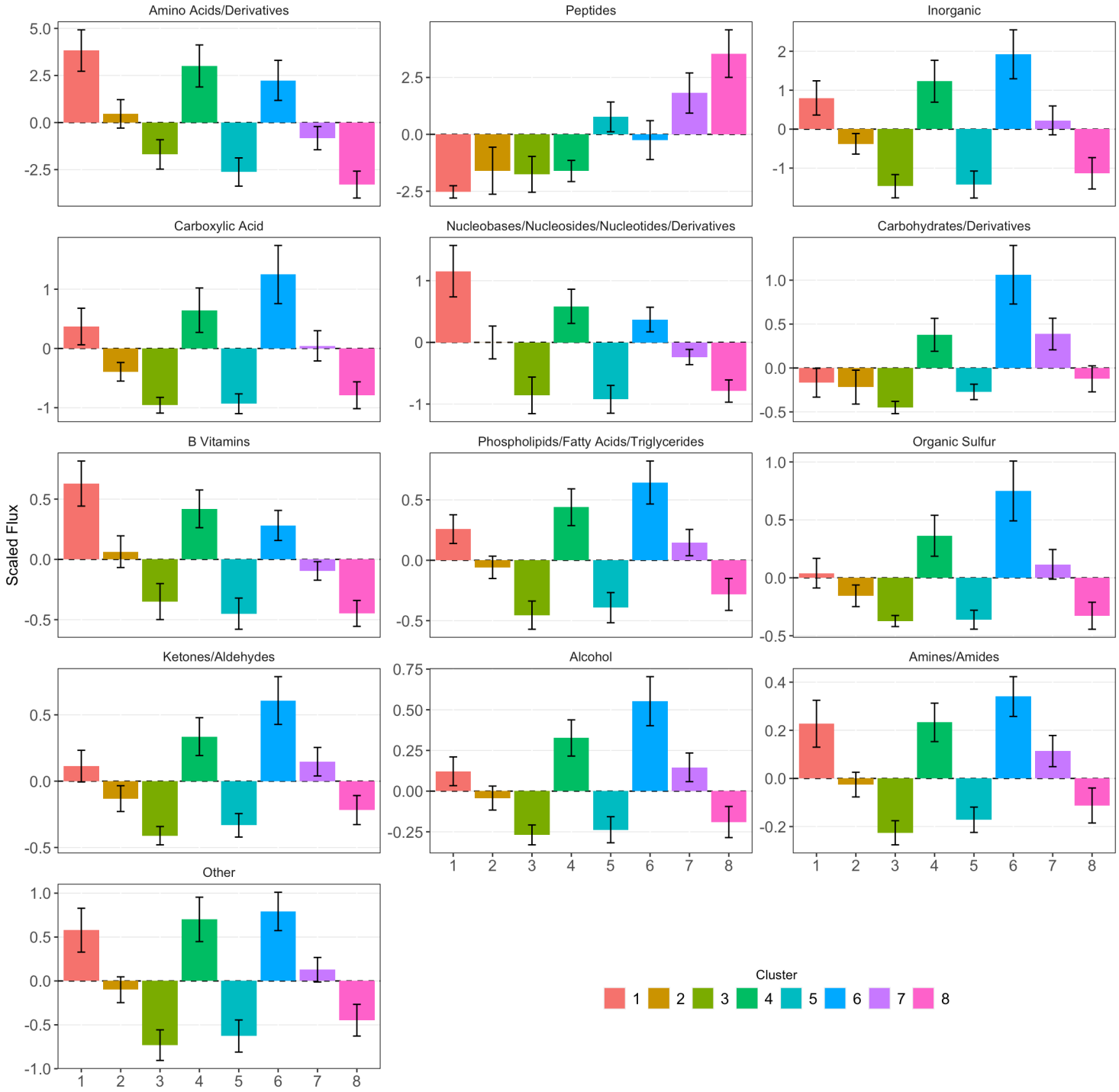

```
## [1] "Figure: Plot of scaled SOM prototype fluxes versus potential growth limitatio
n by cluster for each of the metabolite classes."
```



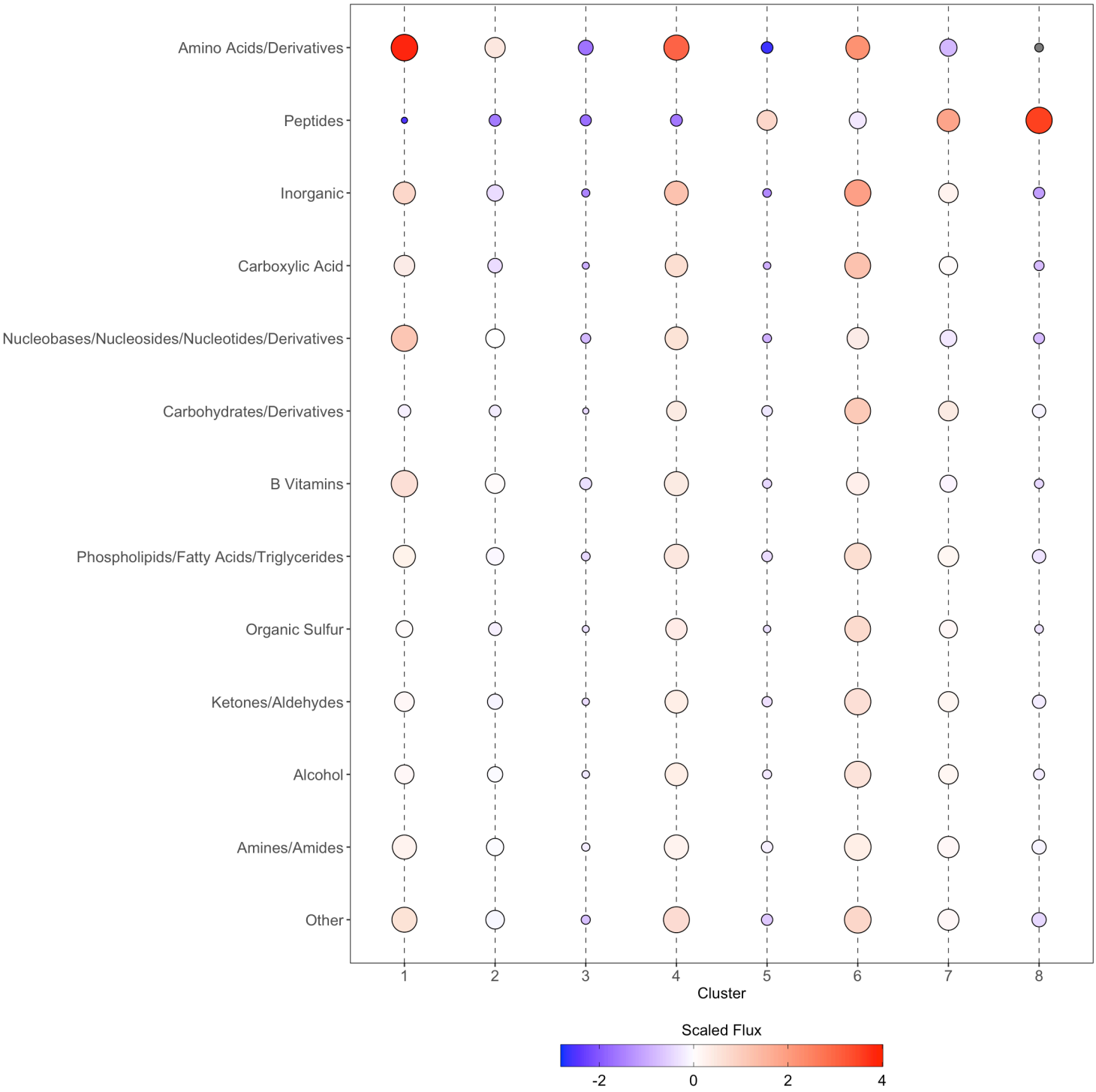

```
## [1] "Figure: Per grid point relative abundances of metabolite classes"
```



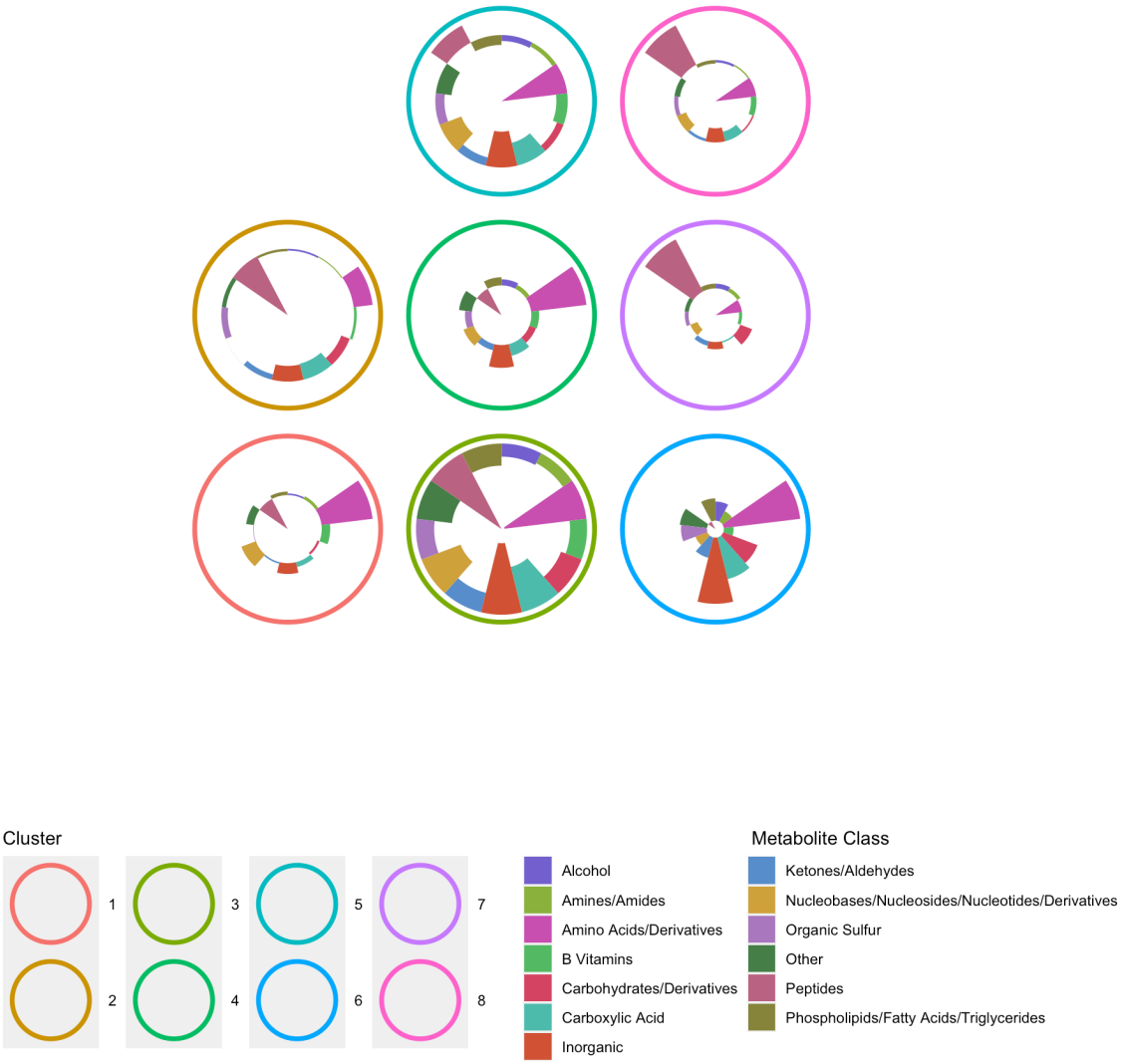

```
## [1] "Figure: Scaled SOM prototype flux values per cluster (with 1 sd error bars) f
or each metabolite class."
```



[1] "Figure: Similar bubble plot to previous plot for growth limitation. Colors describe the information presented in above figure of the scaled SOM prototype fluxes while bubble sizes describe the relative importance of each metabolite class between clusters."

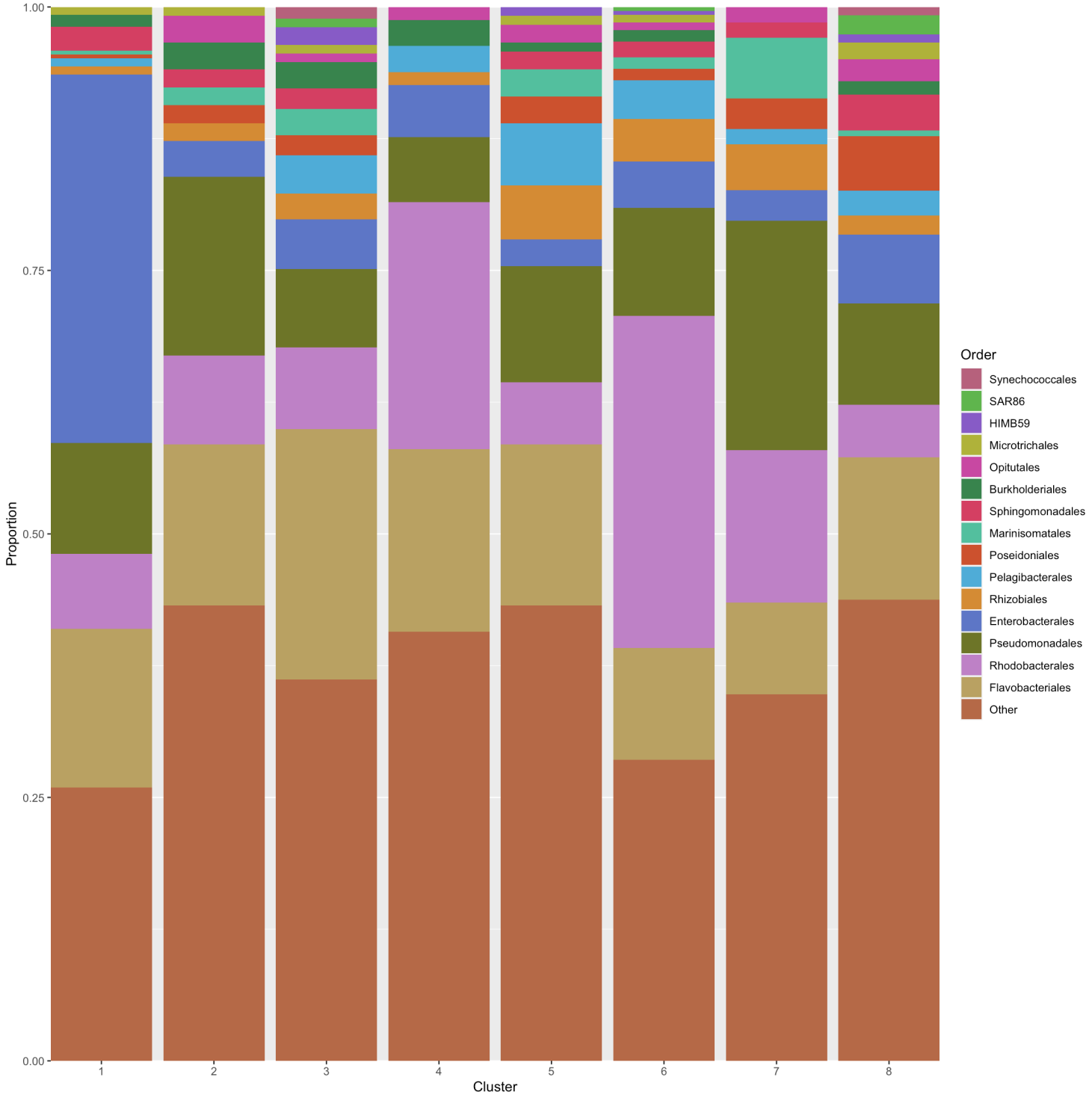



```
## [1] "Figure: Per cluster relative abundances of metabolite classes"
```

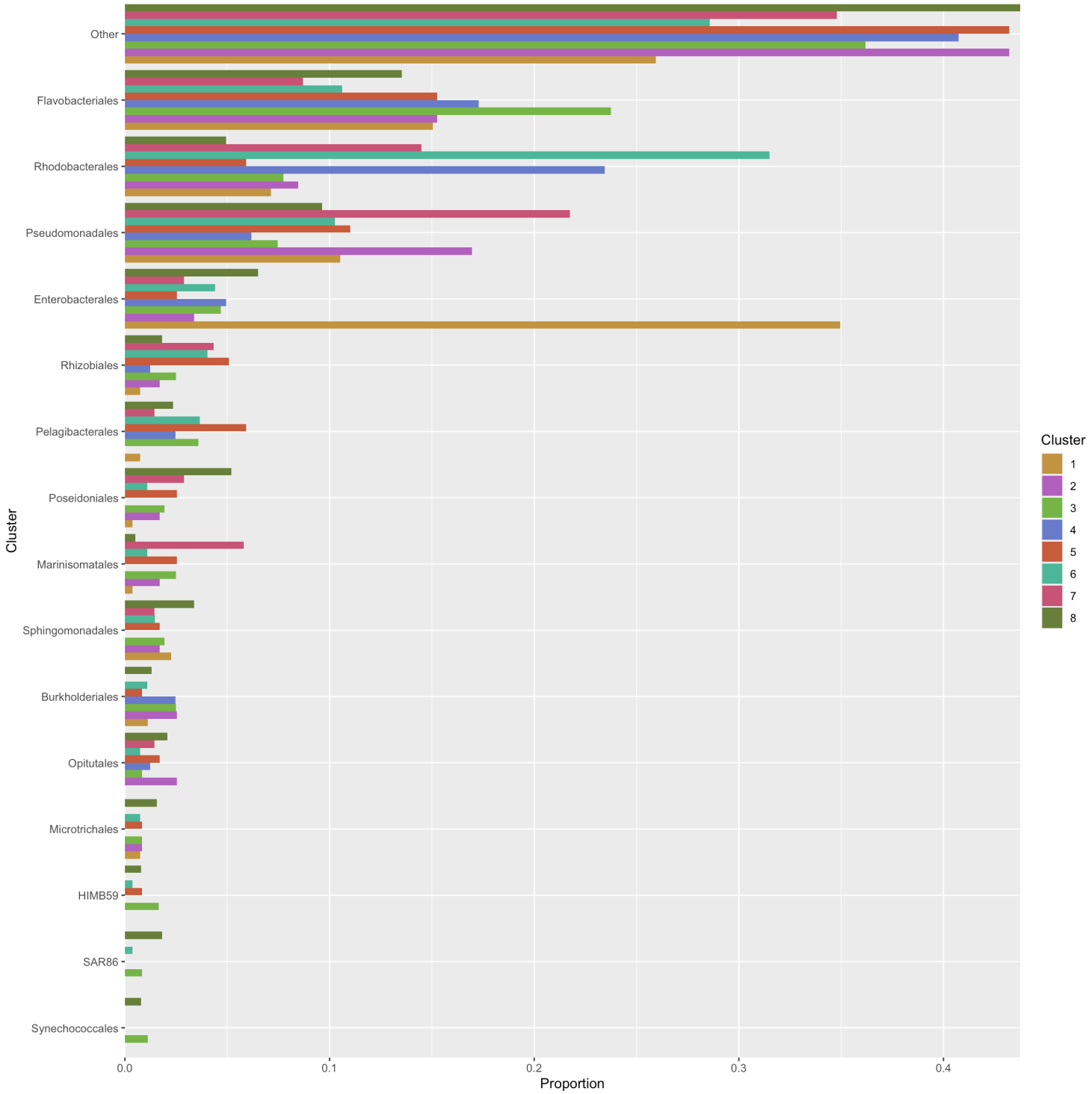


Phylogenetic analysis of the dataset based on whole data relative and per cluster relative abundances.

```
## [1] "Figure: Within cluster relative abundances of top 15 phylogenetic orders (in total dataset) plus 'Other' category."
```

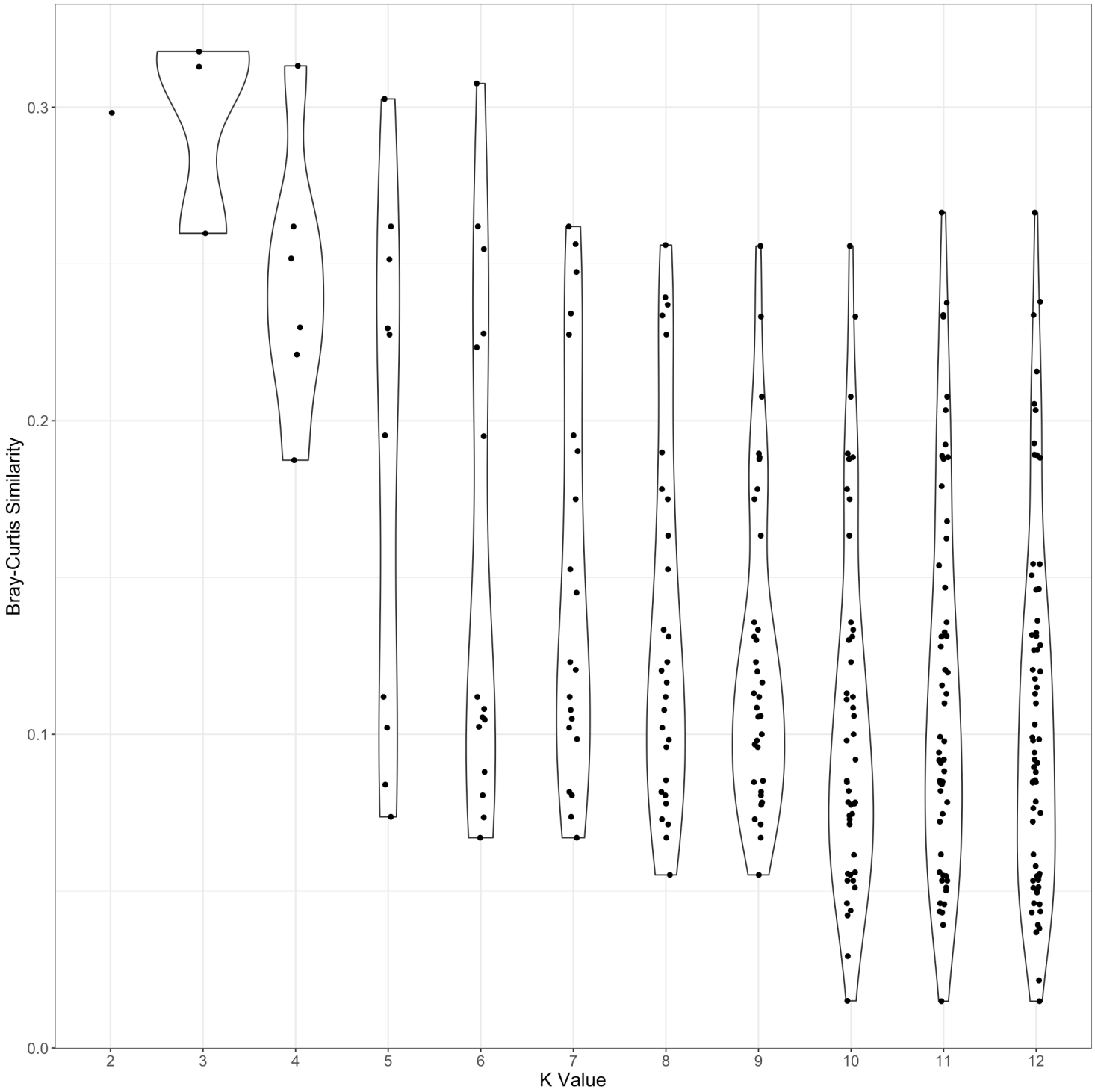



```
## [1] "Figure: Per cluster relative abundances of top 15 orders (in total dataset) grouped by order."
```



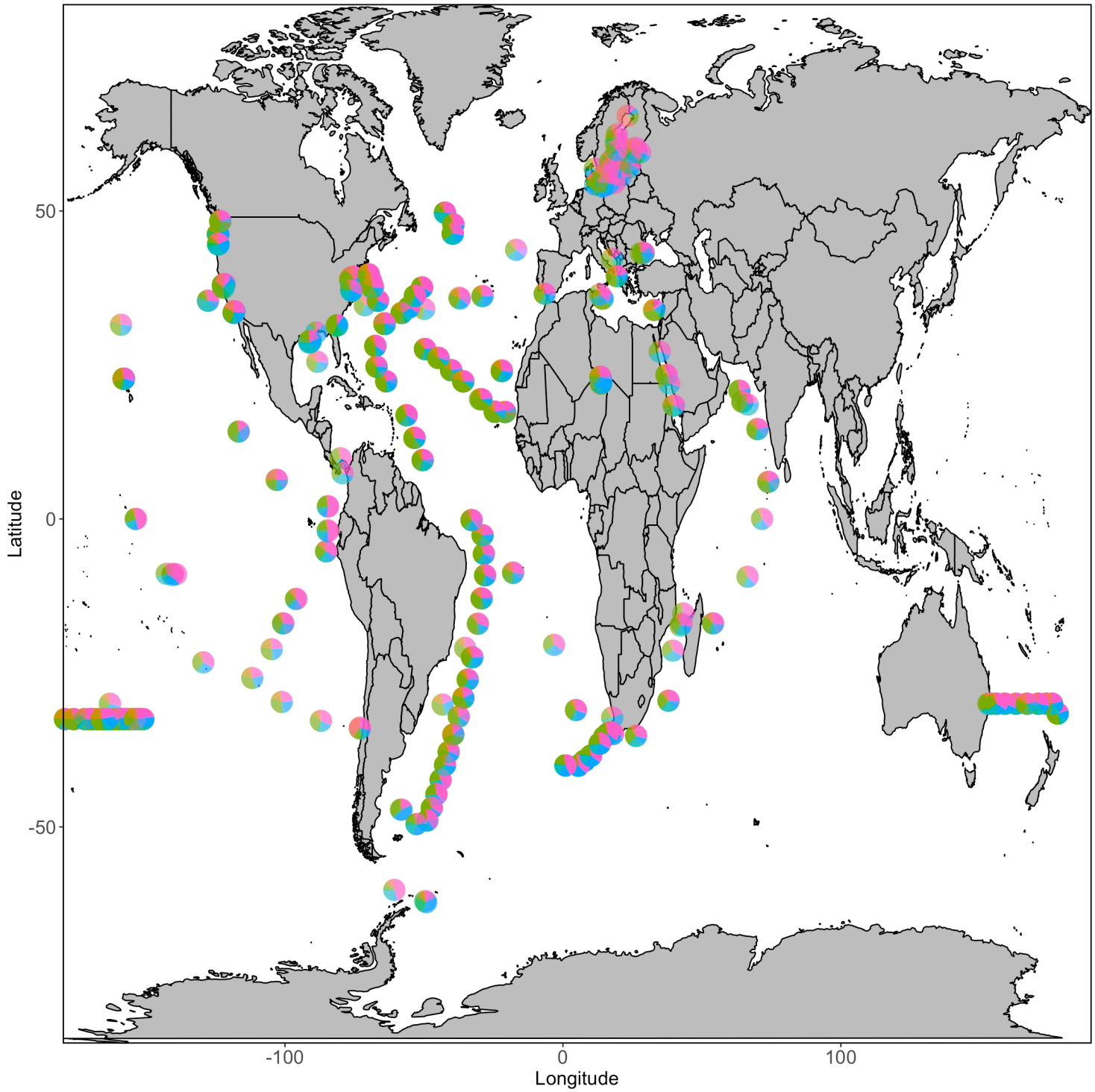

```
## [1] "Figure: Relative difference between within cluster and full dataset abundance
of top 15 taxonomic orders grouped by order."
```


```
## [1] "Figure: Bray-Curtis dissimilarity between the relative abundances of unique g
enuses in the dataset as cluster size is increased from 2 to 12 (NOTE: undefined genu
ses were defined as `order`_other)."
```

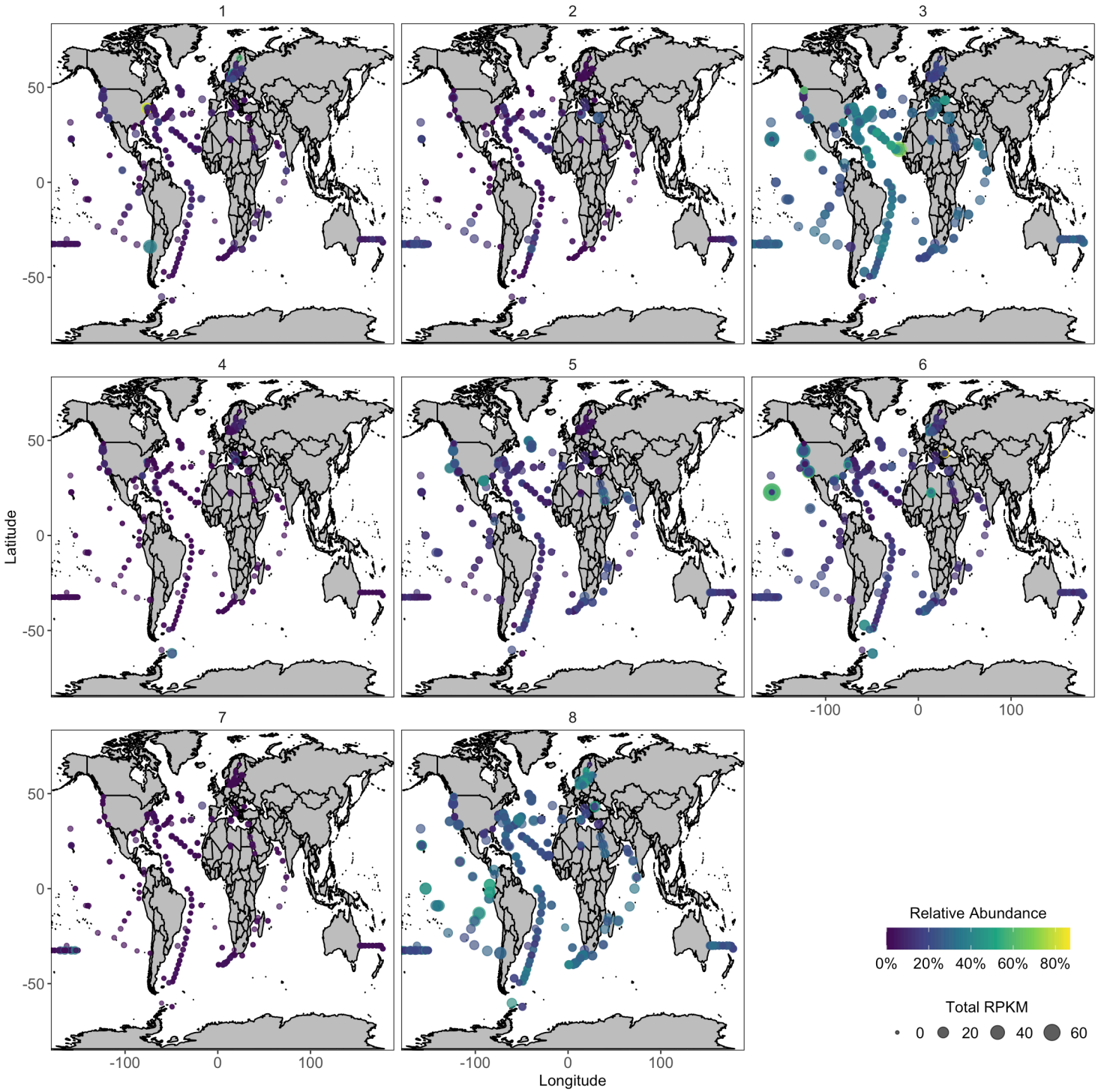


Read recruitment was performed on 1,425 ocean metagenomes and then the resulting RPKM values were broken out by cluster to determine geographic distribution of clusters (NOTE: this is still the 50% threshold, 7 cluster read recruitment, new plots coming soon).

```
## [1] "Figure: Read recruitment map showing the relative abundance of RPKM attribute
d to each cluster at each station."
```




```
## [1] "Figure: Per cluster read recruitment maps colored by relative contribution to
total RPKM at each station. Bubble sizes reflect total RPKM value at each site."
```


1
2
3
4
5
6
7
8
50
0
-50
Latitude
-100
0
100
Longitude
Relative Abundance
0%
20%
40%
60%
80%
Total RPKM
0
20
40
60