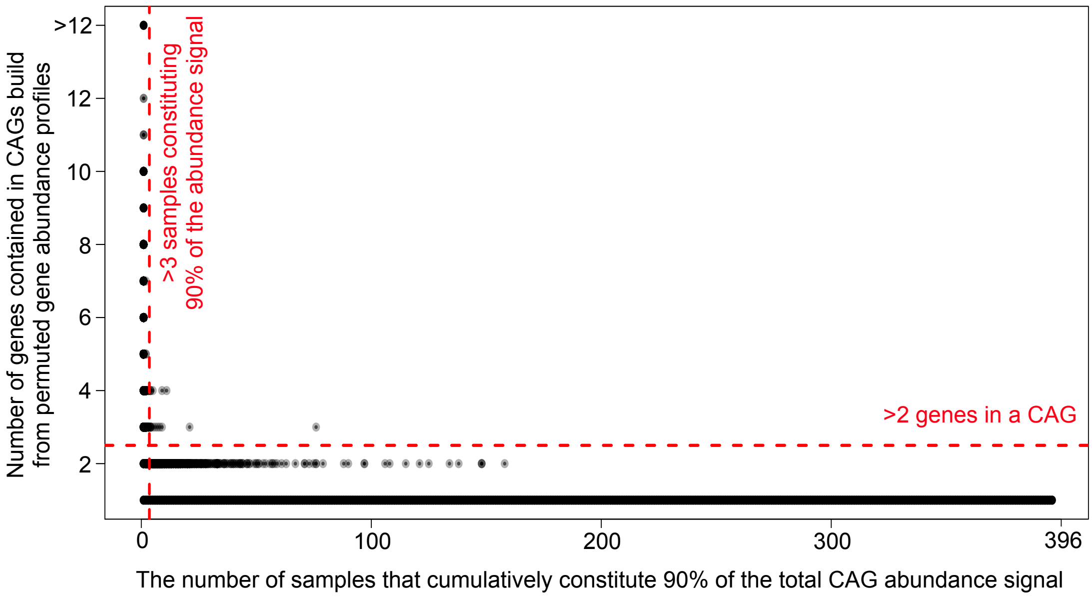
Filtering of canopies after clustering

After the canopy clustering the canopies need to be filtered for: too rare canopies, too small canopies and outlier driven canopies. The canopy clustering C program does not do that.

Reasoning
Supplementary Figure 9 from Nielsen et al. (2014) show the results of a clustering on permuted gene profiles. Hence the result represents random and non-biological clusters of genes. These random clusters contain few genes or are captured in rare canopies (*i.e.* with very few sample wise observations) or canopies that are driven by high outlier measurements. To filter such random canopies we propose the following filters.



Supplementary Figure 9 (from Nielsen et al. 2014). Canopy clustering on permuted abundance profiles. The result of an exhaustive co-abundance binning of a gene-wise shuffled abundance matrix is shown. The size (number of genes) and minimal number of samples that constitute 90% of the total abundance signal from the resulting 1,840,781 random CAGs are shown. Only 18 CAGs escape the QC filter indicated with red dashed lines. All of these contained 3 or 4 genes and were observed in a few samples. 1,539,760 of the random CAGs contained 1 gene and 799 contained more than 12 genes. For all of the latter 90% or more of the abundance signal originated from only one sample. The estimated number of randomly occurring CAGs in the non-permuted canopy clustering (*i.e.* the real data) was very low and only expected among the rare and very small CAGs (FDR ~10% for CAGs with 3 or 4 genes).

Filter against outlier driven canopies and rare CAGs
Canopies with median profiles that are driven by high outliers sample observations may be random or chimeric. Such canopies may be observed in a number of samples, but the Pearson correlation coefficient cannot separate these profiles because it gives too much weight on the high outlier observation. This may happen at the gene level or at the canopy

level. We therefore recommend using two filters to remove outlier driven canopies (see below). The second filter will also remove rare CAGs that are based on observation from less than 3 samples.

Gene outlier filter. All genes must have a spearman correlation coefficient higher than 0.7 to the canopy profile. This filter is intended to remove genes from CAGs.

Canopy outlier filter. 90% of the total canopy profile (*i.e.* the samplewise median gene abundances for each sample, given by the C program) across all samples must originate from more than 3 samples.
Using R:

```
goodCAGs<-rowSums(t(apply(M,1,sort, decreasing=TRUE))[,1:3])/rowSums(M)>0.9
```

This filter removes entire CAGs that are outlier driven as well as GACs that are based on too few observations.

Filter agains too small canopies
All CAGs with less than 3 genes should be discharged. In many cases it may make sense to set this threshold to a higher gene count, but from a purely clustering point of view a filter against smaller than 3 gene seams to capture most random canopies.

We recommend that the outlier filter is applied first, because it may change the size of some canopies.