CST Cluster

Weijia Xiong

6/9/2020

Data

- A case-control study of 49 pregnant women, 15 of whom delivered preterm.
- 40 women contributed samples for a discovery dataset (11 of these 40 women delivered preterm) and nine women contributed samples for a validation dataset (four of these nine women delivered preterm).
- From 40 of these women, the authors analyzed bacterial taxonomic composition of 3,767 specimens collected prospectively and weekly during gestation and monthly after delivery from the vagina, distal gut, saliva, and tooth/gum.

1. OTU Table: 1271 taxa and 761 samples (Transform to data proportions)
2. Sample Data: 761 samples by 64 sample variables
3. Taxonomy Table: 1271 taxa by 7 taxonomic ranks
4. Phylogenetic Tree: 1271 tips and 1270 internal nodes

Cluster for CSTs

Distance Matrix

- Calculate the Bray-Curtis distance between all samples.

The general formula for calculating the Bray-Curtis dissimilarity between samples i and i' is as follows, supposing that the counts are denoted by n_{ij} and that their sample (row) totals are n_{i+}

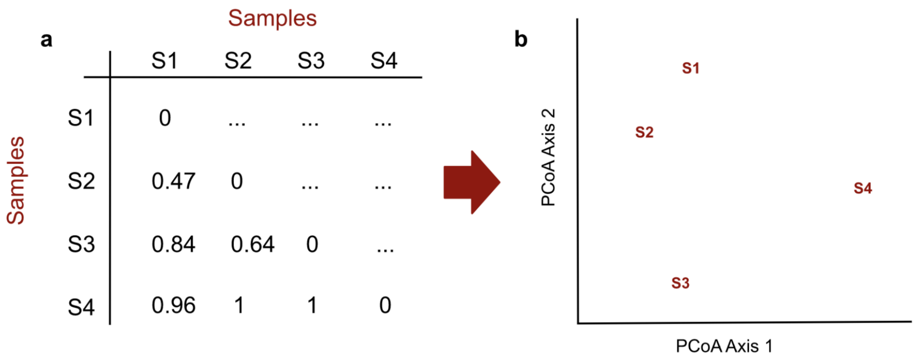
$$d_{ii'} = \frac{\sum_{j=1}^{J} |n_{ij} - n_{ij}|}{n_{i+} + n_{i'+}}$$

PCoA and denoise distance matrix

Principal coordinates analysis

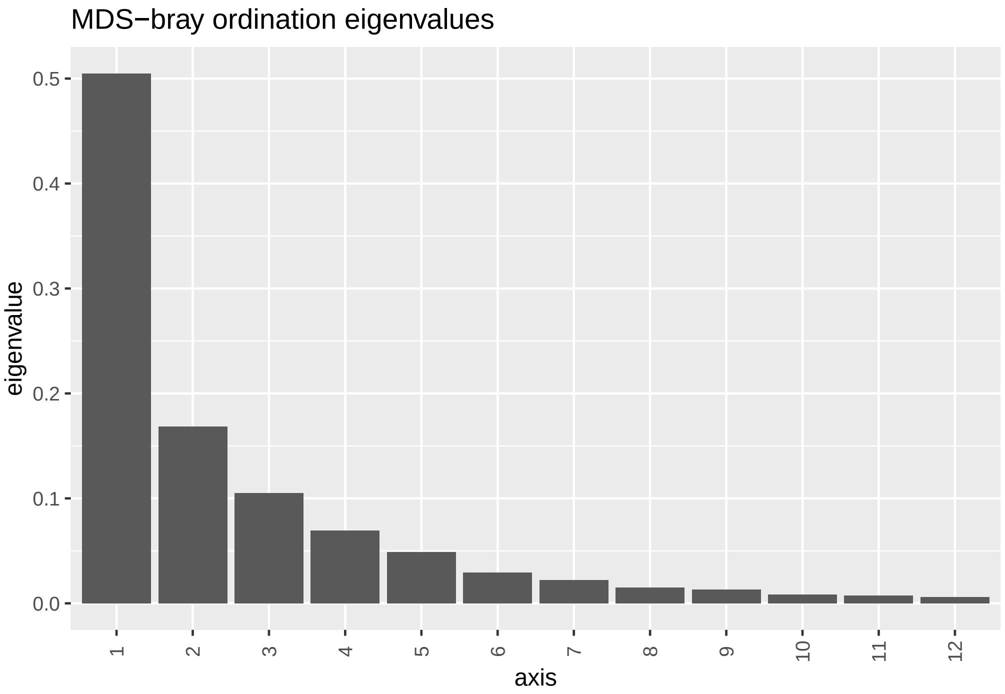
As with other ordination techniques such as PCA and CA, PCoA produces a set of uncorrelated (orthogonal) axes to summarise the variability in the data set. Each axis has an eigenvalue whose magnitude indicates the amount of variation captured in that axis.The proportion of a given eigenvalue to the sum of all eigenvalues reveals the relative 'importance' of each axis. A successful PCoA will generate a few (2-3) axes with relatively large eigenvalues, capturing above 50% of the variation in the input data, with all other axes having small eigenvalues. Each object has a 'score' along each axis. The object scores provide the object coordinates in the ordination plot.

Interpretation of a PCoA plot is straightforward: objects ordinated closer to one another are more similar than those ordinated further away. (Dis)similarity is defined by the measure used in the construction of the (dis)similarity matrix used as input.

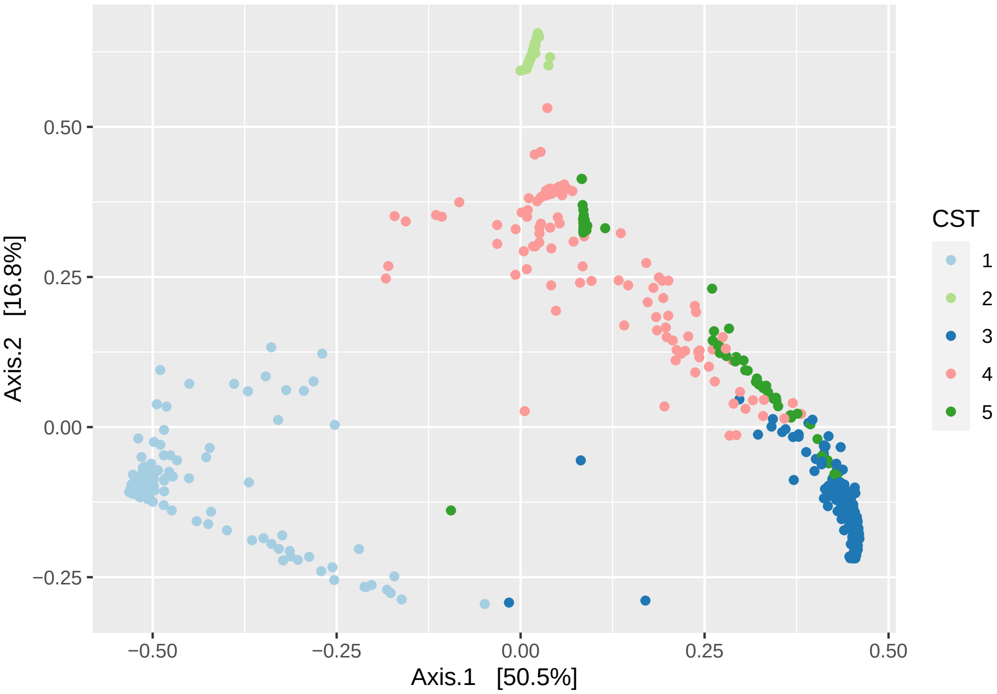


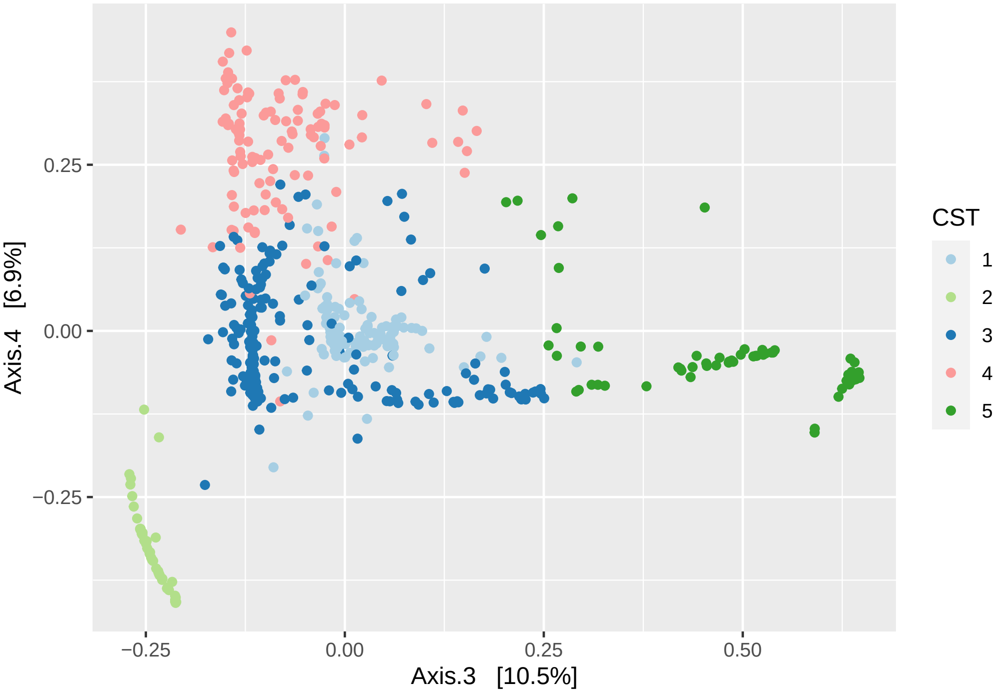
	S1	S2	S3	S4
S1	0	...	...	...
S2	0.47	0	...	...
S3	0.84	0.64	0	...
S4	0.96	1	1	0

The vaginal community is dominated by closely related, but functionally distinct, Lactobacillus species. Therefore it is better to use a non-phylogenetically aware distance measure so as to be able to separate these species. Start with an MDS (or PCoA) ordination.

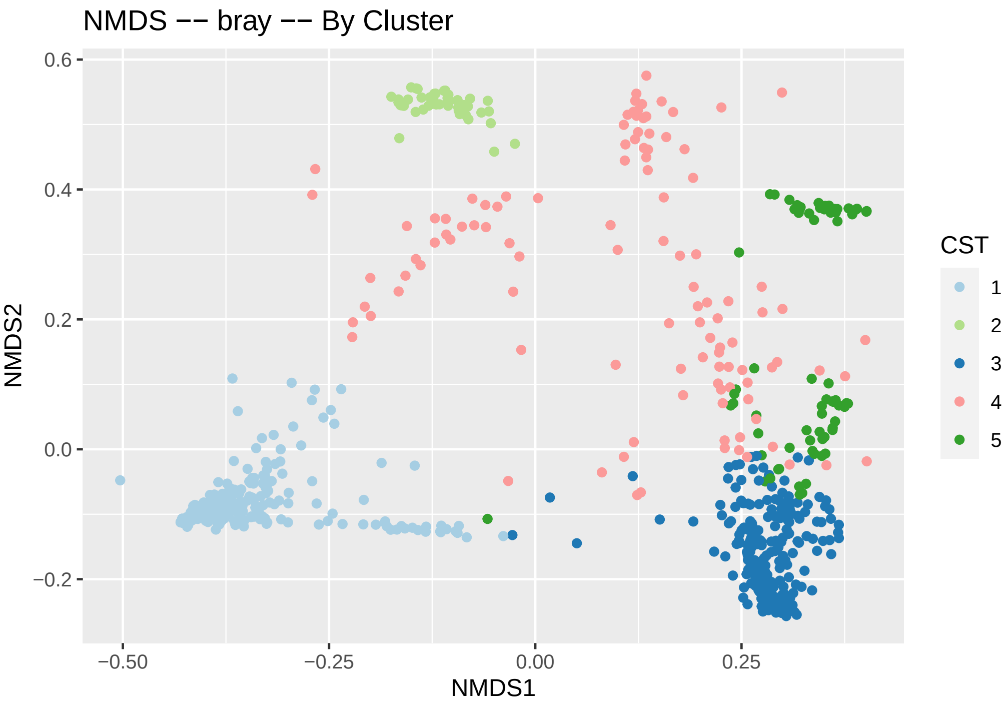


- MDS/PCoA



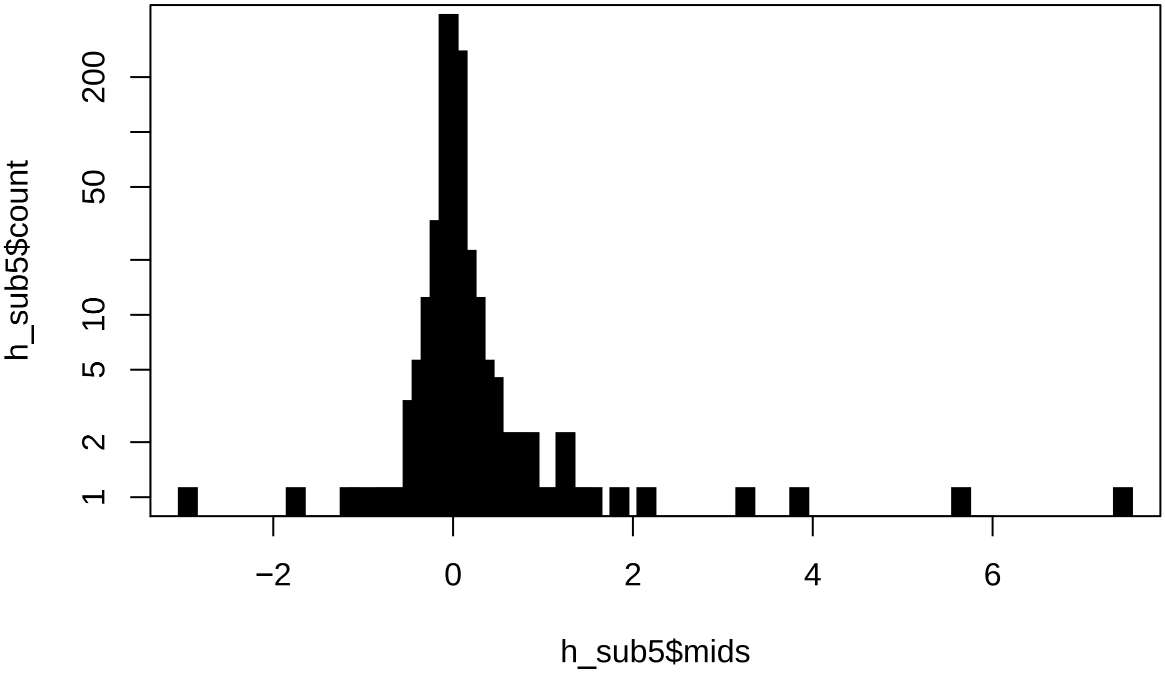


- NMDS(Non-metric MultiDimenstional Scaling)



Denoise distance matrix

The authors would like to clean some of the noise from the data by restricting this to the truly significant dimensions. The top 5 eigenvalues are clearly very significant, but let's keep all the positive eigenvalues that clearly exceed the magnitude of the smallest negative eigenvalues:



Looks like eigenvalues 6 and 7 still stand out, so the authors go with 7 MDS dimensions (Axis 1-7).

Determine the number of clusters

Determine the number of clusters from the gap statistics. Using pam(Partitioning Around Medoids) and clusGap() in R.

Gap statistics

$$D_r = \sum_{i,i' \in C_r} d_{ii'}$$

be the sum of the pairwise distances for all points in cluster r, and set

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r$$

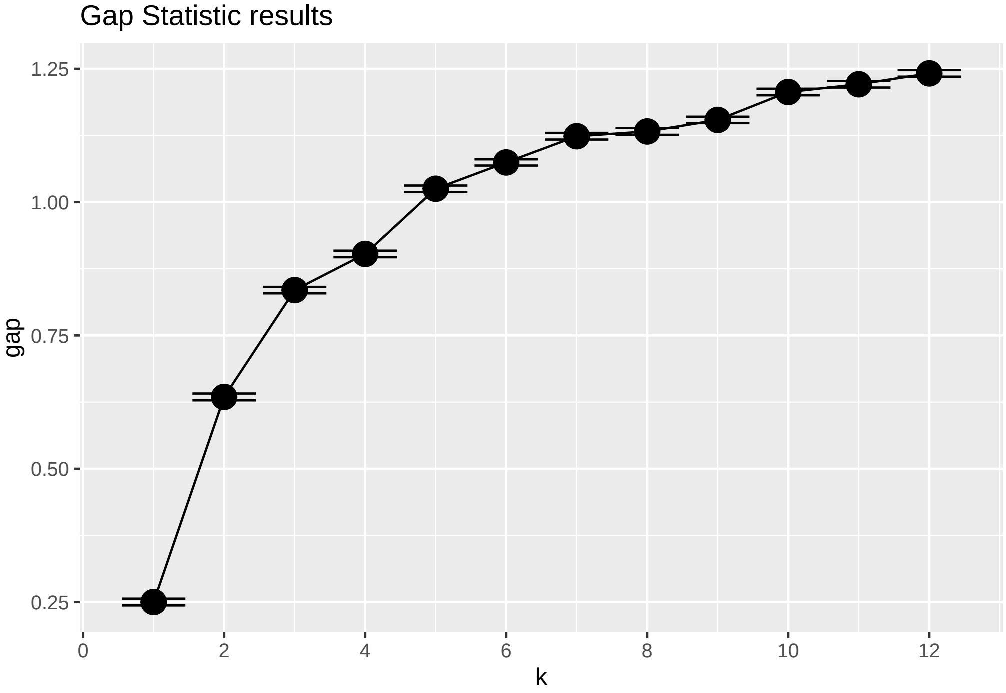
The idea of this approach is to standardize the graph of $\log(W_k)$ by comparing it with its expectation under an appropriate null reference distribution of the data. (The importance of the choice of an appropriate null model is demonstrated in Gordon (1996).) The estimate of the optimal number of clusters is then the value of k for which $\log(W_k)$ falls the farthest below this reference curve. Hence the authors define

$$\text{Gap}_n(k) = E_n^* \{\log(W_k)\} - \log(W_k)$$

To obtain the estimate $E_n^* \{\log W_k\}$, the authors compute the average of B copies $\log W_k^*$ for $B = 10$, each of which is generated with a Monte Carlo sample from the reference distribution. Those $\log W_k^{*\text{ from the }B}$ Monte Carlo replicates exhibit a standard deviation $\text{sd}(k)$ which, accounting for the simulation error, is turned into the quantity

$$s_k = \sqrt{1 + 1/B}\,\text{sd}(k)$$

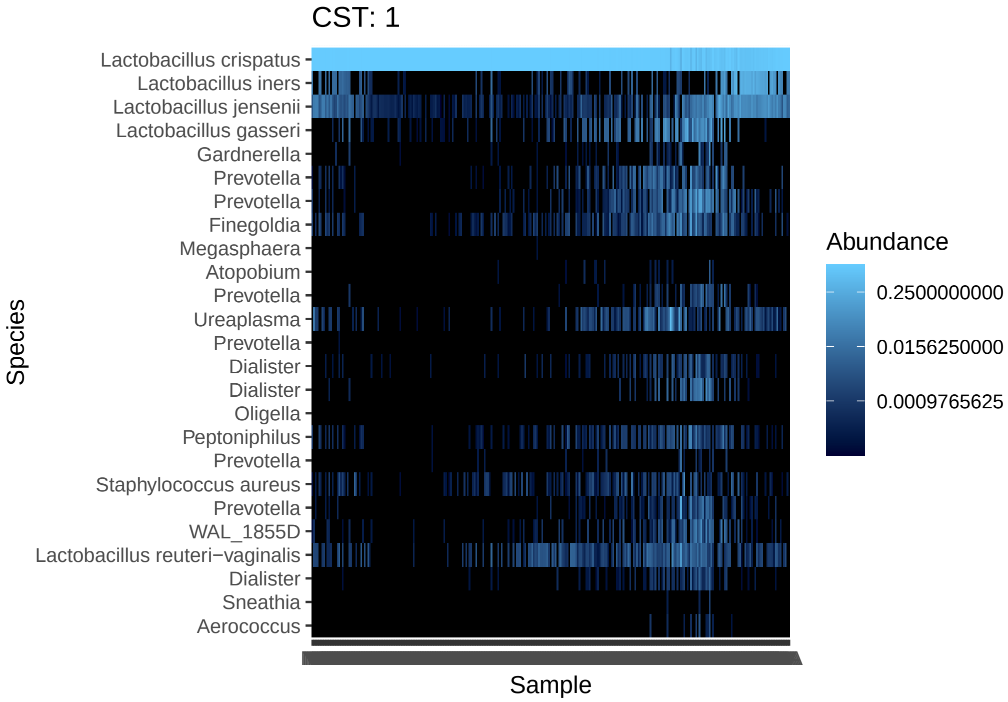
Finally, the optimal number of clusters K is the smallest k such that $\text{Gap}(k) \geq \text{Gap}(k+1) - s_{k+1}$

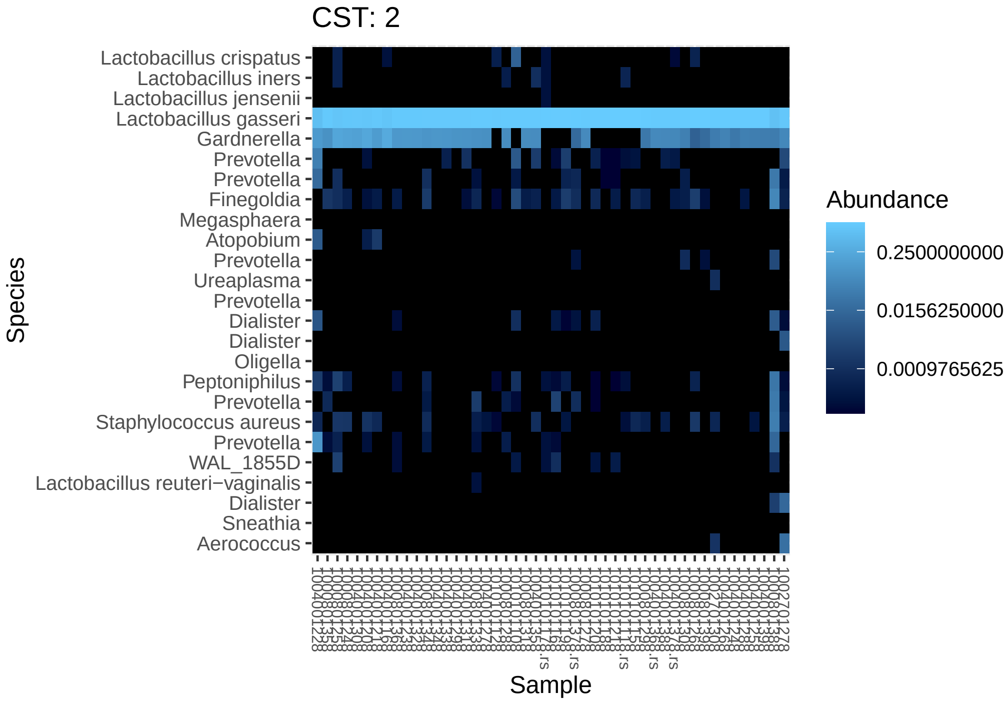


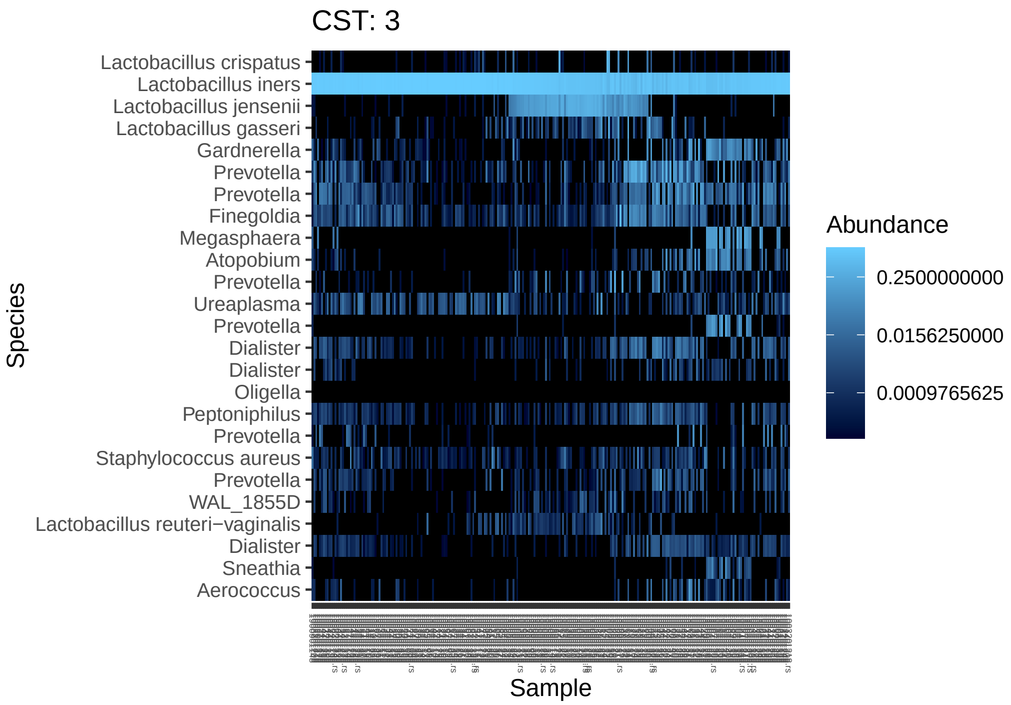
The gap statistic strongly suggests at least three clusters, but makes another big jump at K=5 before the slope gets a lot smaller.

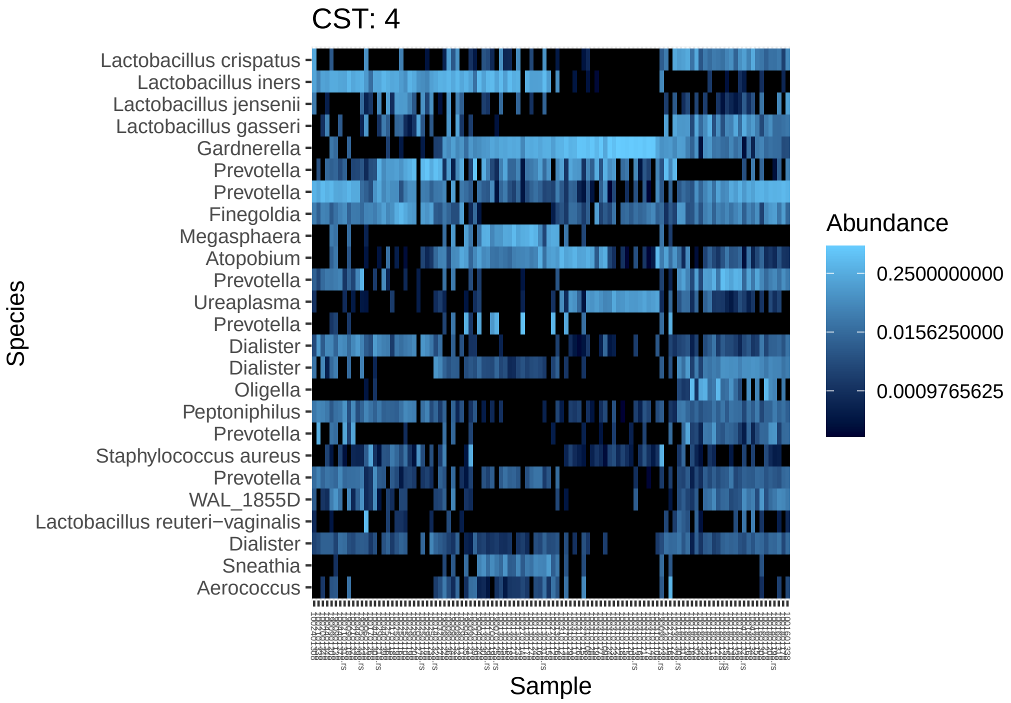
Heatmap

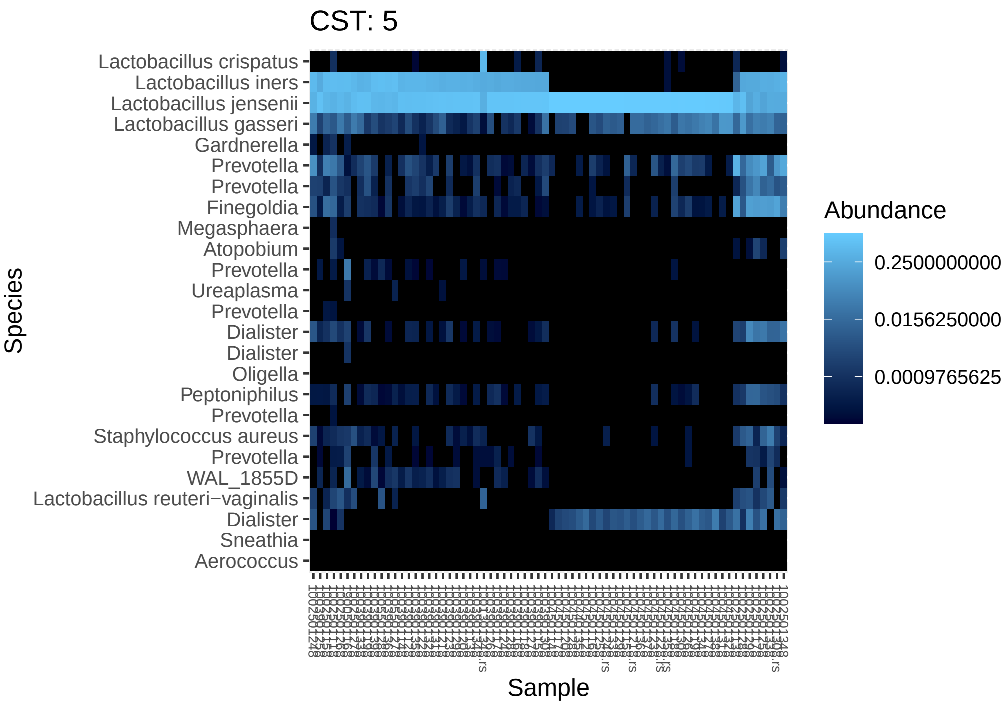
The ordinations offer support for these being legitimate clusters, even if they are not perfect and some samples look like they might be mixtures of two clusters. Let's take a look at the heatmaps of each cluster for additional perspective:





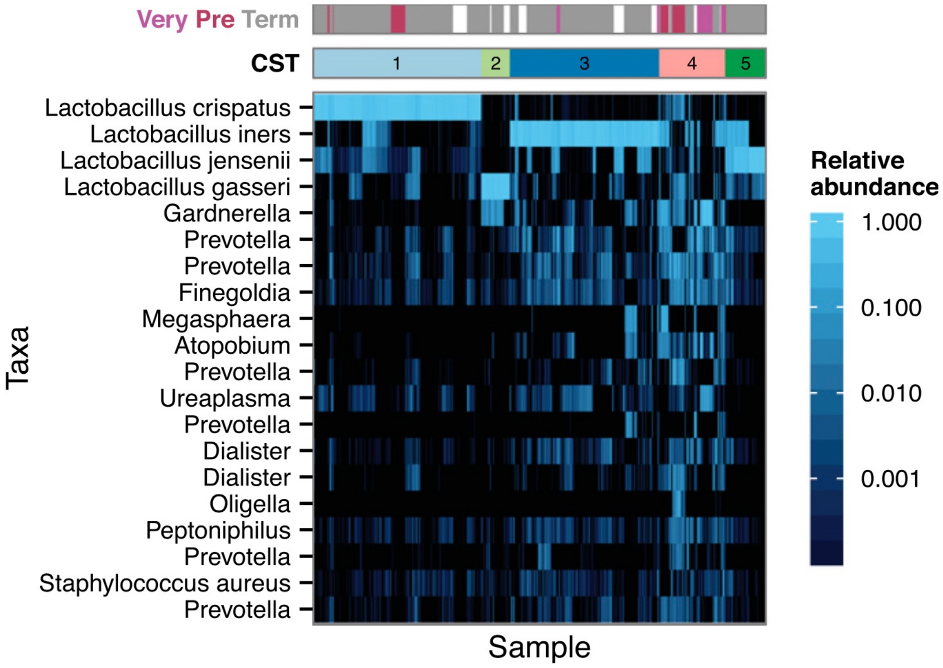

CST: 3
Species
Lactobacillus crispatus
Lactobacillus iners
Lactobacillus jensenii
Lactobacillus gasseri
Gardnerella
Prevotella
Prevotella
Finegoldia
Megasphaera
Atopobium
Prevotella
Ureaplasma
Prevotella
Dialister
Dialister
Oligella
Peptoniphilus
Prevotella
Staphylococcus aureus
Prevotella
WAL_1855D
Lactobacillus reuteri−vaginalis
Dialister
Sneathia
Aerococcus
Sample
Abundance
0.2500000000
0.0156250000
0.0009765625


CST: 4
Species
Lactobacillus crispatus
Lactobacillus iners
Lactobacillus jensenii
Lactobacillus gasseri
Gardnerella
Prevotella
Prevotella
Finegoldia
Megasphaera
Atopobium
Prevotella
Ureaplasma
Prevotella
Dialister
Dialister
Oligella
Peptoniphilus
Prevotella
Staphylococcus aureus
Prevotella
WAL_1855D
Lactobacillus reuteri−vaginalis
Dialister
Sneathia
Aerococcus
Sample
Abundance
0.2500000000
0.0156250000
0.0009765625



These heatmaps show that the clusters have a clear interpretability that further supports the validity of clustering in this context. CSTs 1,2,3 and 5 are dominated by different species of Lactobacillus. CST4 is much more diverse.

Below is a heat map of the relative abundances of the top taxa for all the vaginal samples, with color bars indicating the CST and the preterm Outcome associated with each sample.



Analyzing CST dynamics

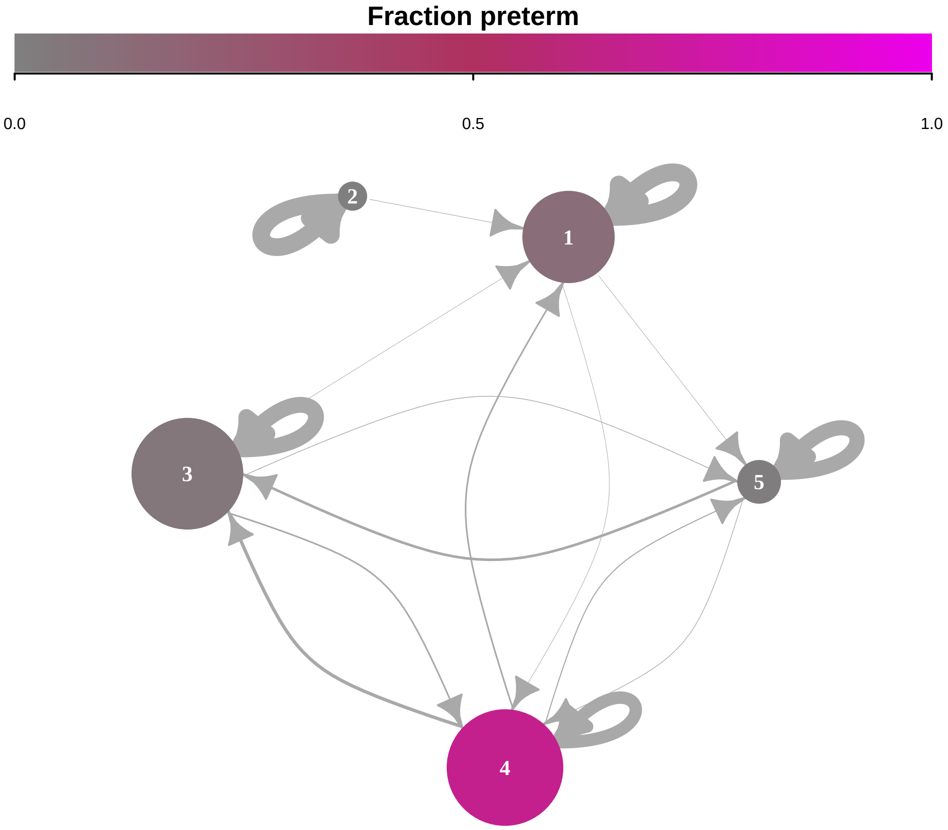
The dynamics of the vaginal CSTs, in particular the transition rates between them, are of interest. The authors will take each pair of sequential samples separated by one week (4-10 days) and calculate the MLE estimate of the transition matrix between CSTs from the list of transitions observed, which is MLE(t__ij) = n__ij/n__i. This lacks error bars, and is not using all the data as it drops the information contained in transitions interrupted by missing data, but the authors have enough sequential data samples for this to be a reasonable estimate:

Transition between CSTs

```
## PregCST
##  A  5 - dimensional discrete Markov Chain defined by the following states:
##  1, 2, 3, 4, 5
##  The transition matrix  (by rows)  is defined as follows:
##             1         2         3          4           5
## 1 0.979423868 0.0000000 0.0000000 0.01646091 0.004115226
## 2 0.024390244 0.9756098 0.0000000 0.00000000 0.000000000
## 3 0.009615385 0.0000000 0.8750000 0.08173077 0.033653846
## 4 0.084210526 0.0000000 0.1789474 0.68421053 0.052631579
## 5 0.000000000 0.0000000 0.1384615 0.03076923 0.830769231
```

- CST and preterm outcome

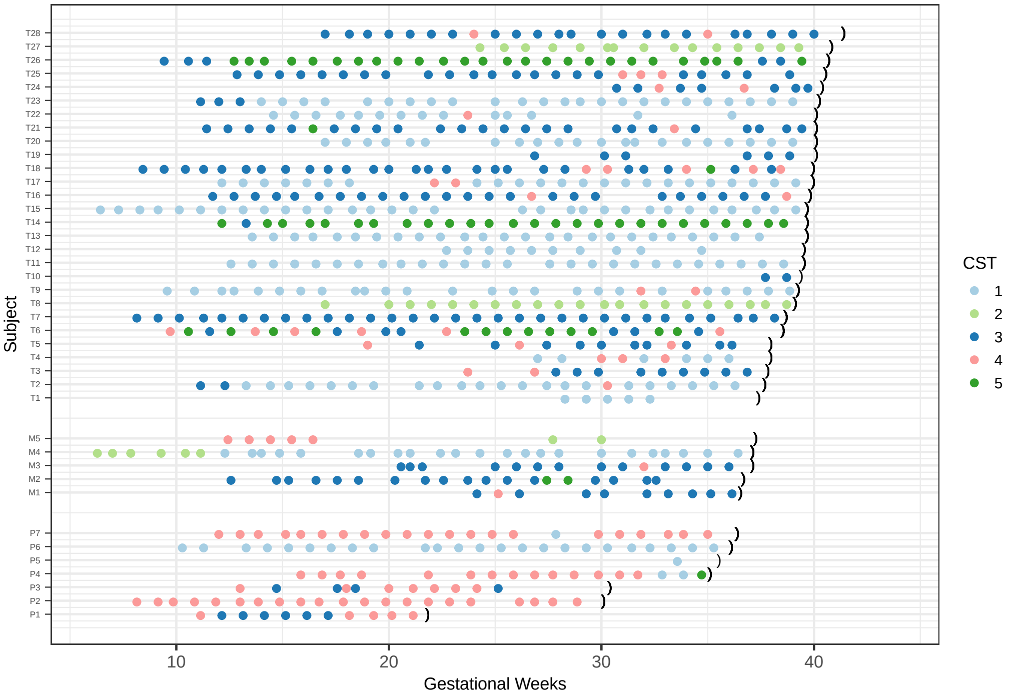
```
##
##           Term    Preterm
##   1 0.89230769 0.10769231
##   2 1.00000000 0.00000000
##   3 0.95260664 0.04739336
##   4 0.33980583 0.66019417
##   5 0.98529412 0.01470588
```



Color indicates the fraction of those CSTs samples from preterm births (excluding marginal subjects). Size indicates the number of subjects in which that CST was observed. The much, much higher association of CST 4 with preterm birth suggests that transitions into that state might be an important warning sign to watch out for.

Plot Sampling Time Course

It is also useful to look at the sampling trajectory of our subjects colored by the CST.



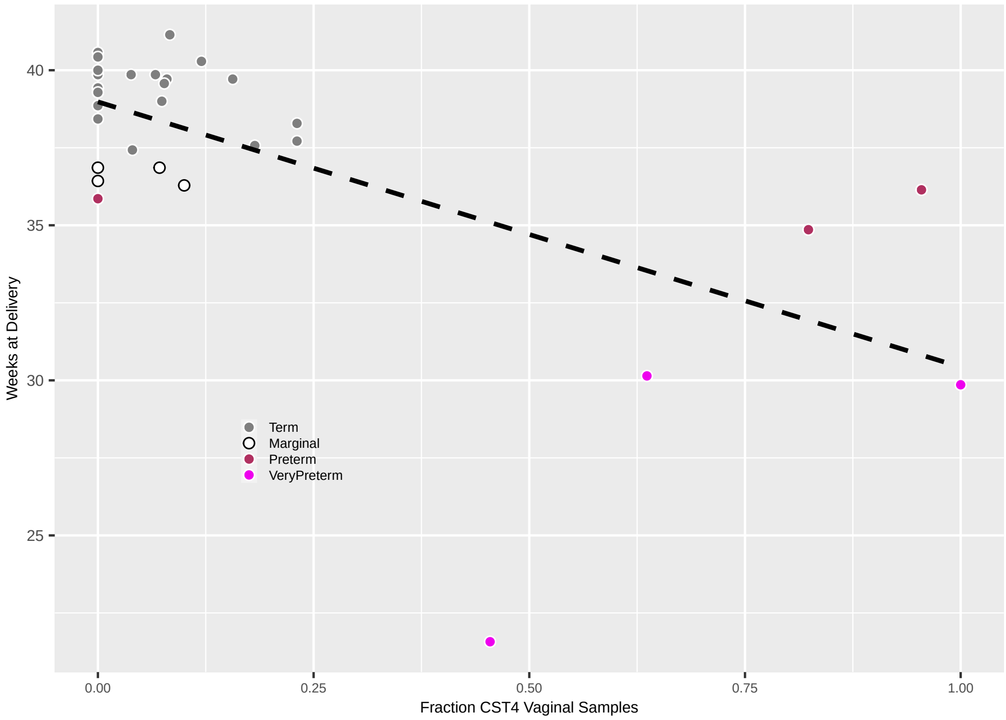
This shows pretty clearly that CST4 samples are more prevalent in the shorter, preterm pregnancies than they are in the term pregnancies.

Test for associations between pregnancy outcome and CST

Correlation between CST4 prevalence and gestational time at delivery

First the authors evaluate whether there is a relationship between the prevalence of CST4 during pregnancy and the length of gestation. They restrict this analysis to **those subjects with at least 10 samples** so that the independent variable (the proportion of CST4 samples) is not unduly influenced by randomness associated with a very small number of samplings.

- Plot Length of Delivery vs. the Fraction of CST4 samples



- Test the association

```
##
##  Pearson's product-moment correlation
##
## data:  subdf$FracDiv and subdf$GDDel
## t = -4.4177, df = 31, p-value = 0.0001131
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7951846 -0.3537287
## sample estimates:
##        cor
## -0.6215569
```

```
##
##  Spearman's rank correlation rho
##
## data:  subdf$FracDiv and subdf$GDDel
## S = 8503.6, p-value = 0.01468
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.4210613
```

- Correct for white/non-white then test

```
##
##  Pearson's product-moment correlation
##
## data:  regdf$FracDivCor and regdf$GDDel
## t = -4.1334, df = 31, p-value = 0.0002518
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
```

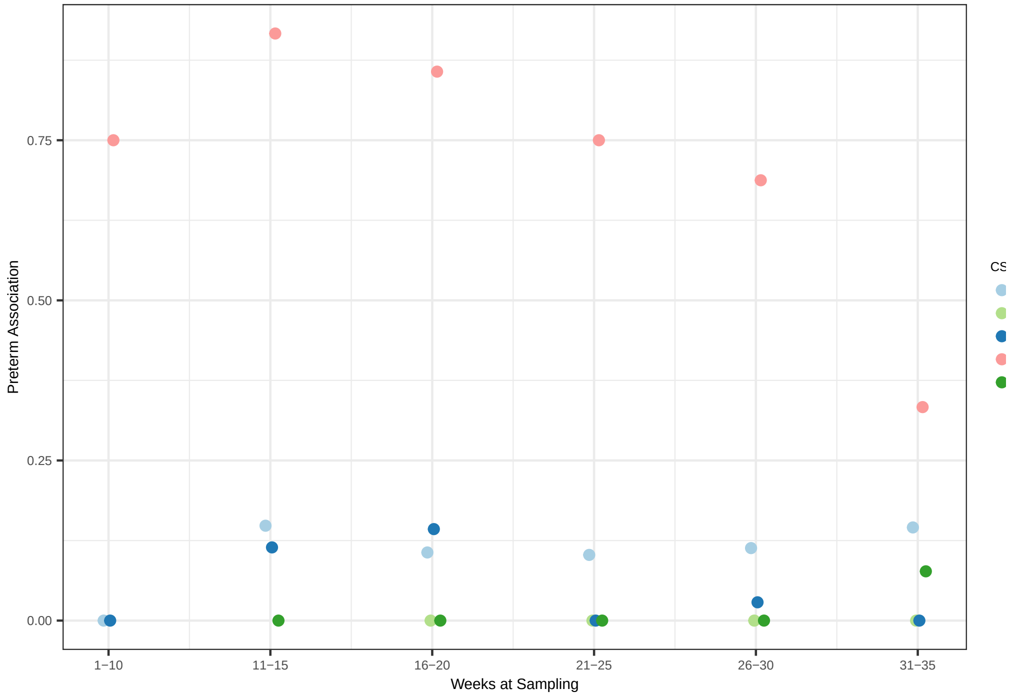
```
##  -0.7798064 -0.3178006
## sample estimates:
##        cor
## -0.5960755
```

```
##
##  Spearman's rank correlation rho
##
## data:  regdf$FracDivCor and regdf$GDDel
## S = 8081.6, p-value = 0.0455
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.3505329
```

There is a significant negative association with the prevalance of CST4 and gestational time at delivery.

Show the association of CST4 with preterm outcome at different periods during pregnancy

Restrict to <= 35 gestational weeks and exclude marginal deliveries



```
##
##     Marginal Preterm Term VeryPreterm
##   1       23      28  232           0
##   2       12       0   36           0
##   3       39       0  201          10
##   4        7      35   35          33
##   5        2       1   67           0
```

CST4 specimens are associated with preterm outcomes even when obtained relatively early in pregnancy.