Karlsruhe Institute of Technology

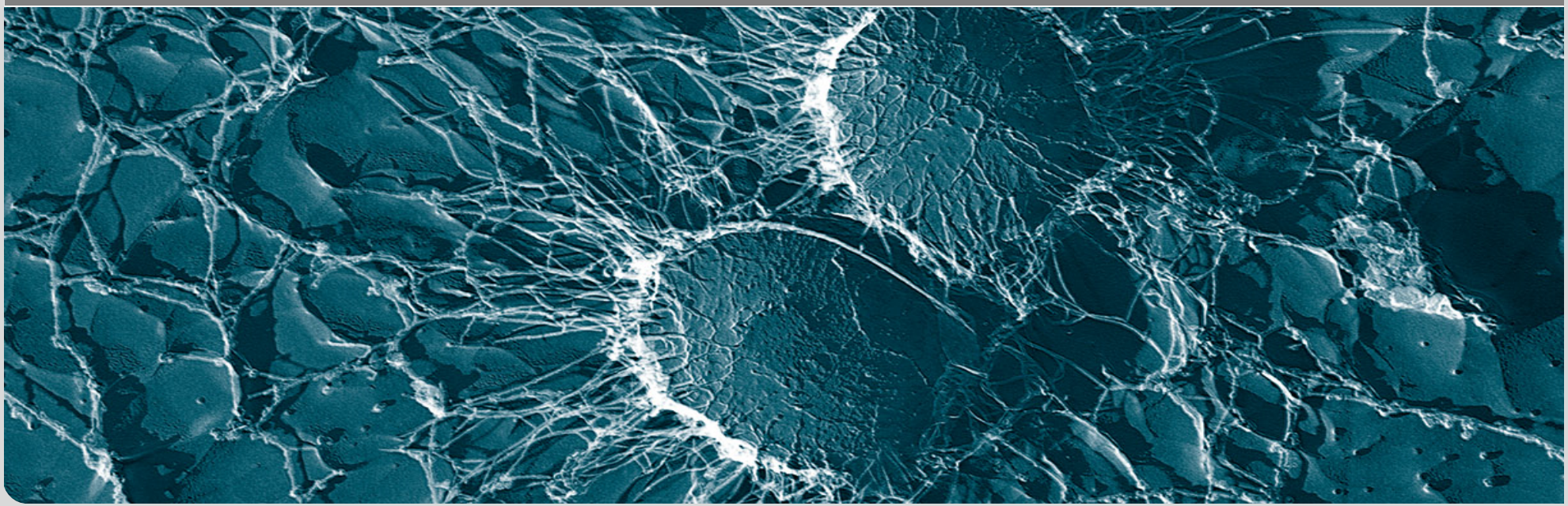
Bacterial Communities in Women with Bacterial Vaginosis: High Resolution Phylogenetic Analyses Reveal Relationships of Microbiota to Clinical Criteria

Seminar presentation

Pierre Barbera

Supervised by: Alexandros Stamatakis, Lucas Czech

CHAIR OF HIGH PERFORMANCE COMPUTING IN THE LIFE SCIENCES

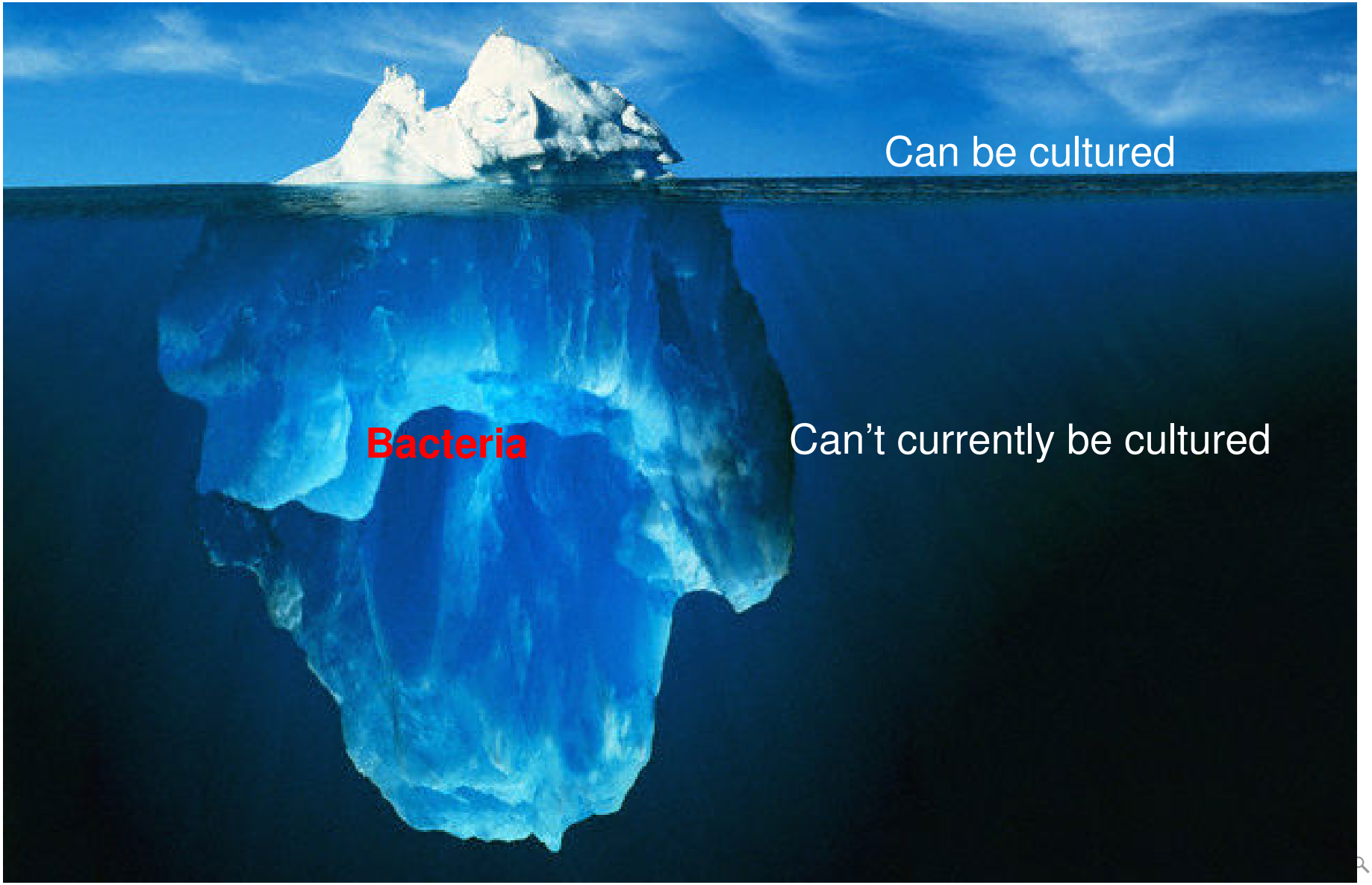


Outline

Outline

How are bacteria studied?

How are bacteria studied?



Metagenomics

- **Microbiome/Microbiota:** collection of microorganisms in environmental niche
- **Metagenomics:** study of collective genetic material from a microbiome
- Interactions and composition of Microbiome centrally important

Bacterial Communities rule your life!

- Human body has roughly 10 trillion cells
- But it houses **10 times that many bacteria**
- Large part of it in the **gut**

Bacterial Vaginosis (BV)

- One of the most common infections of the vagina
- Around **30%** of women in the US are BV-positive
- Cause still unknown (according to CDC)
- Linked to **imbalances in the microbiome** of the vagina

Central questions

- Is there a core BV biome?
- Can novel species be identified?
- Can any synergistic relationships (between bacteria) be identified?
- What is the effect of race on BV prevalence?
- **Can we identify correlations between microbiome composition and clinical features (of BV)?**

Rough Workflow

Take sample
from Patient

Rough Workflow

Take sample from Patient

↓

Isolate and amplify important DNA parts

PCR of 16S

Rough Workflow

Take sample from Patient

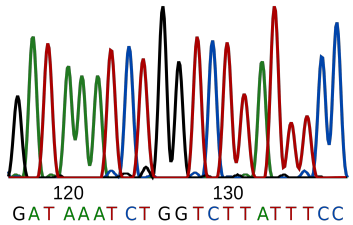
↓

Isolate and amplify important DNA parts

↓

Sequence the DNA

PCR of 16S



Rough Workflow

Take sample from Patient

↓

Isolate and amplify important DNA parts

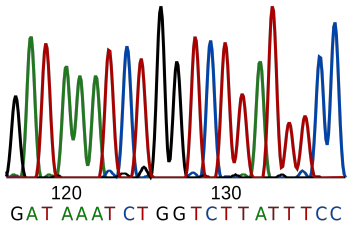
↓

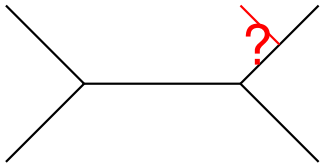
Sequence the DNA

↓

Identify what Species are present

PCR of 16S





Rough Workflow

Take sample from Patient

↓

Isolate and amplify important DNA parts

↓

Sequence the DNA

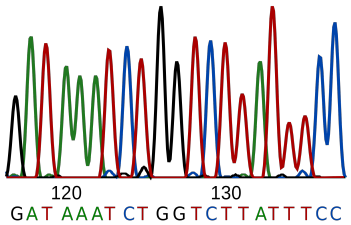
↓

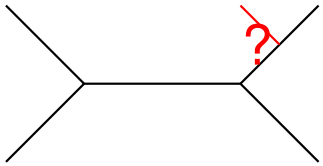
Identify what Species are present

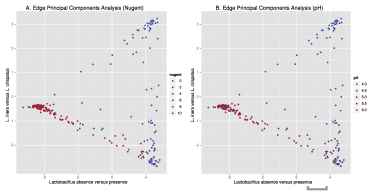
↓

Establish statistical correlations

PCR of 16S







Outline

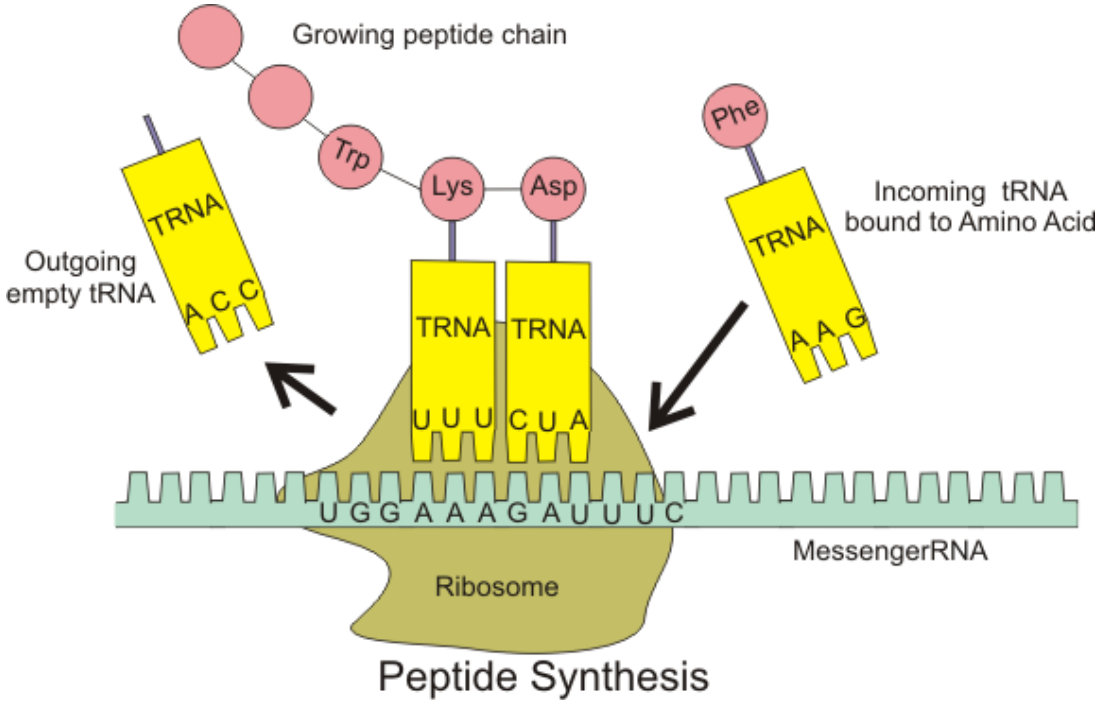
Sample Collection

- Samples from 242 STD clinic patients (Seattle, USA)
- Vaginal swabs, immediately frozen at $-20\,^{\circ}\mathrm{C}$
- 220 samples had sufficient bacterial volume, basis for further methods

Isolate and Amplify

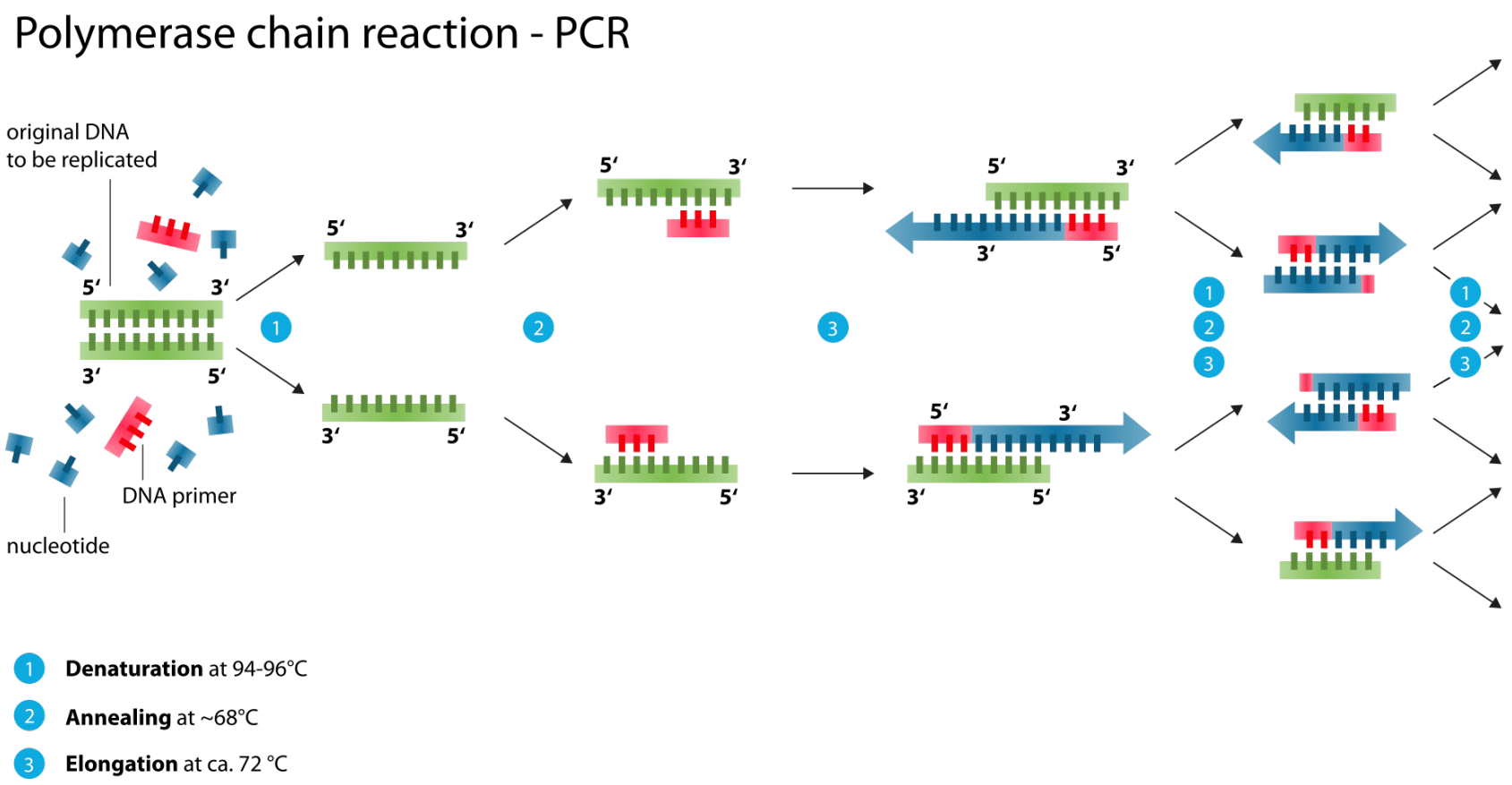
- Only a specific part of the sampled DNA, the **16S rRNA gene**, is needed
- To isolate and amplify this portion the **polymerase chain reaction** lab technique is used

16S ribosomal RNA



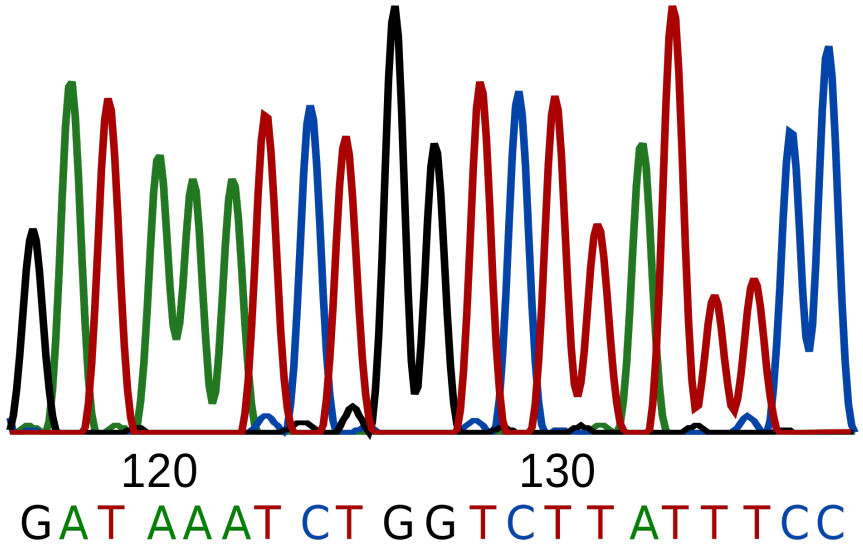
- Part of the ribosome in prokaryotes
- Slow rate of evolution $\Rightarrow$ **highly conserved** between species
- Very good sequence to establish **phylogenies**

Polymerase Chain Reaction (PCR)



- Technique to massively multiply a certain portion of DNA
- Requires **Primer** DNA-strands that will **delimit** the portion to multiply

Sequencing



- Sequence resulting amplified samples
- 454 pyrosequencing
- **Result: many different 16S reads per sample**
- This is where the wet lab work ends and the bioinformatics work begins

Outline

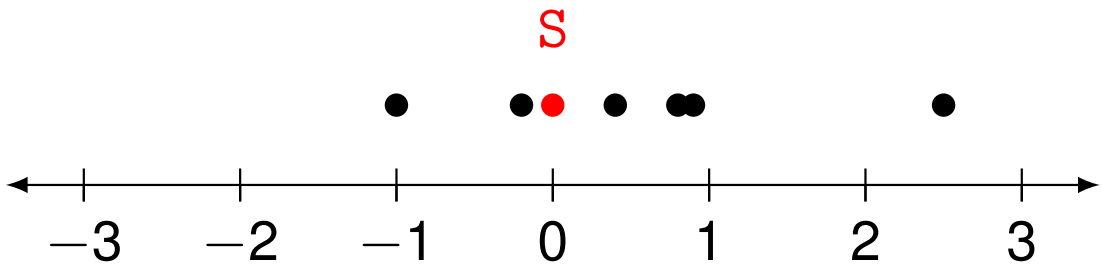
Preprocessing of Reads

- Classify by **barcodes**
- Only keep high quality reads that...
  - start with known barcode
  - contain exact match to used primer
  - are at least 200 base pairs long (excluding primer/barcode)
  - have sufficient quality score
- Trim primer sequences and barcodes
- All done in R, using R/Bioconductor package `microbiome`

Reference Tree Preparations

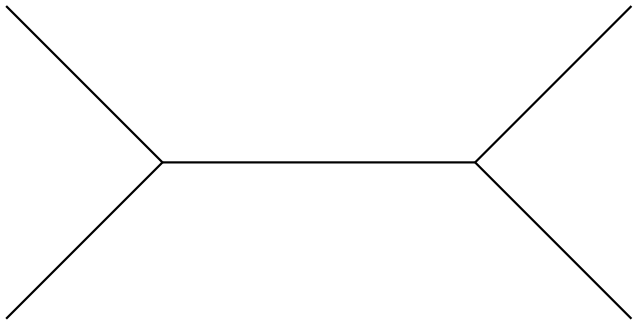
- Take known sequences of bacteria known to reside in vaginal environment
- Trim to 16S region, same as in samples
- Perform **mislabel detection**, as public data is often mislabeled/wrong

Mislabel Detection



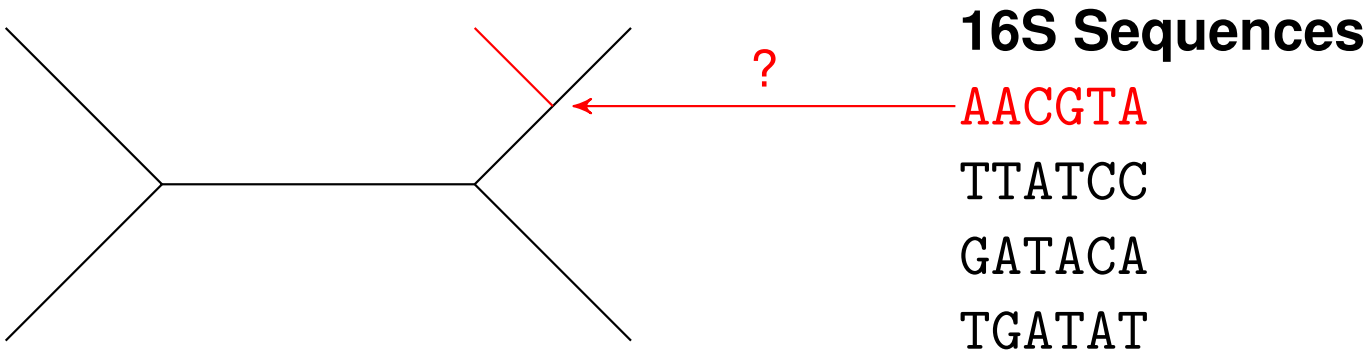
- Compute pairwise distance between all sequences of a taxon
- Select a primary reference sequence$S$ with **smallest median distance** to all others
- Discard sequences that are too far from this reference sequence (by some threshold)

Building the Reference Tree



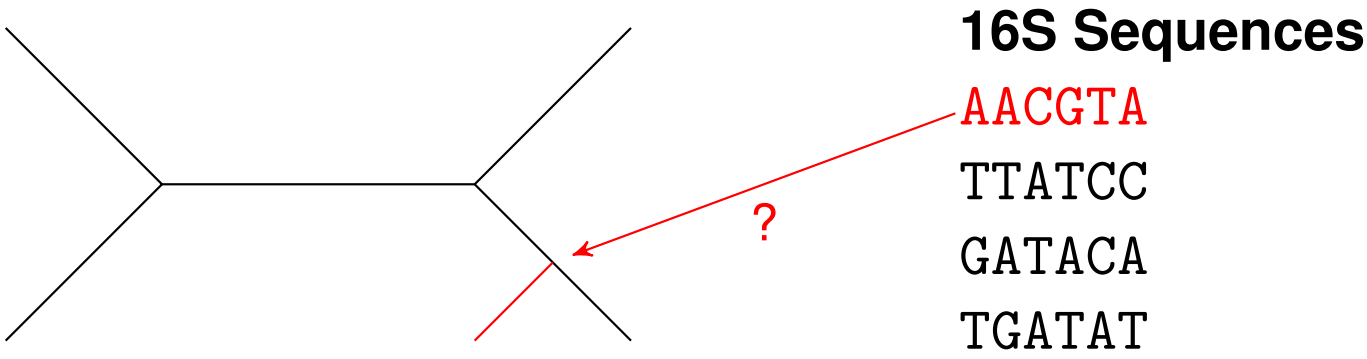
- Build Multiple Sequence Alignment (MSA) using `cmalign`
- Build tree using MSA and `RAxML 7.2.7`, using GTR model

Place Sequence on Tree



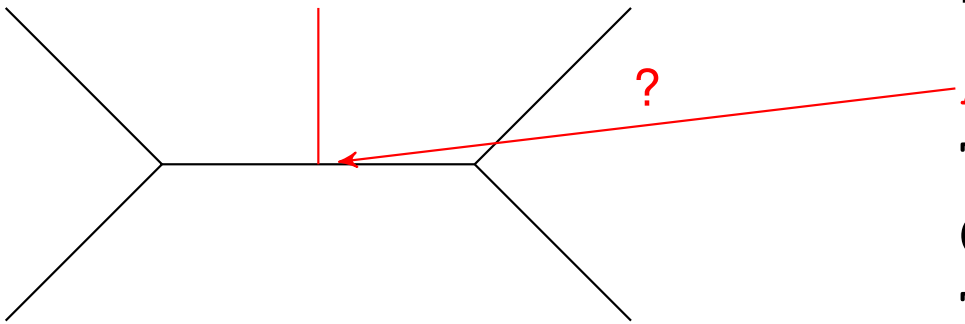
- Take sequence, find **optimal placement** on existing tree
- Optimality meaning Bayesian posterior probability criterion
- Remember where sequence was placed
- Done using the `pplacer` tool

Place Sequence on Tree



- Take sequence, find **optimal placement** on existing tree
- Optimality meaning Bayesian posterior probability criterion
- Remember where sequence was placed
- Done using the `pplacer` tool

Place Sequence on Tree



16S Sequences

AACGTA

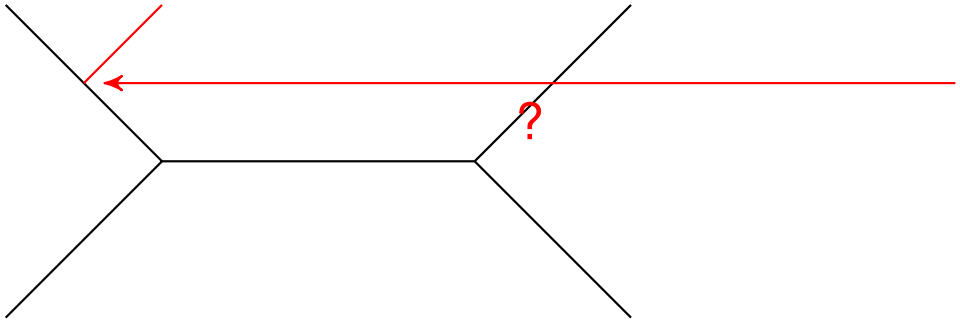
TTATCC

GATACA

TGATAT

- Take sequence, find **optimal placement** on existing tree
- Optimality meaning Bayesian posterior probability criterion
- Remember where sequence was placed
- Done using the `pplacer` tool

Place Sequence on Tree



16S Sequences

AACGTA

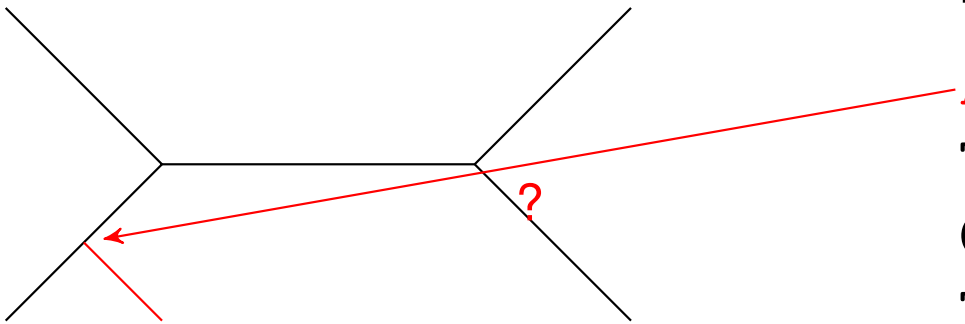
TTATCC

GATACA

TGATAT

- Take sequence, find **optimal placement** on existing tree
- Optimality meaning Bayesian posterior probability criterion
- Remember where sequence was placed
- Done using the `pplacer` tool

Place Sequence on Tree



16S Sequences

AACGTA

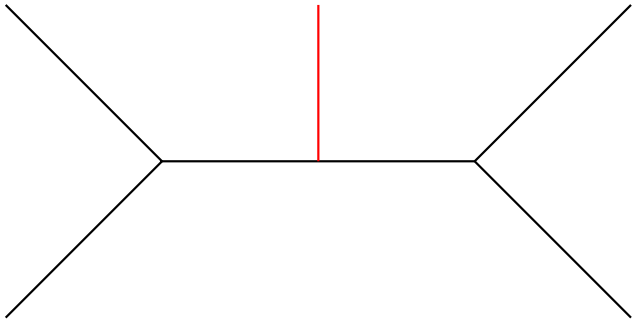
TTATCC

GATACA

TGATAT

- Take sequence, find **optimal placement** on existing tree
- Optimality meaning Bayesian posterior probability criterion
- Remember where sequence was placed
- Done using the `pplacer` tool

Place Sequence on Tree



16S Sequences

AACGTA ✓(< *middle* >)

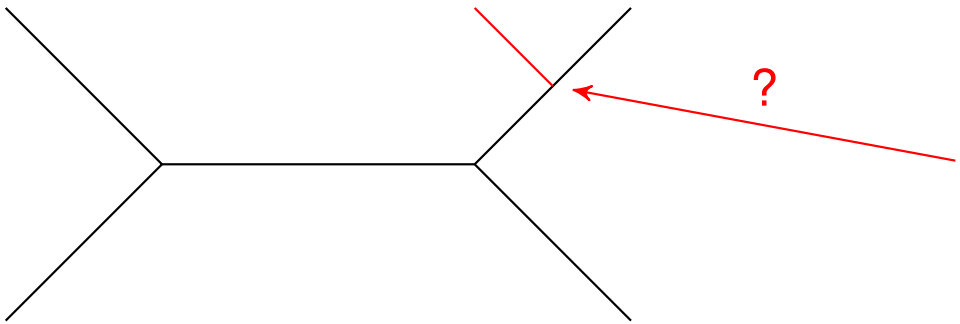
TTATCC

GATACA

TGATAT

- Take sequence, find **optimal placement** on existing tree
- Optimality meaning Bayesian posterior probability criterion
- Remember where sequence was placed
- Done using the `pplacer` tool

Place Sequence on Tree



16S Sequences

AACGTA ✓(< *middle* >)

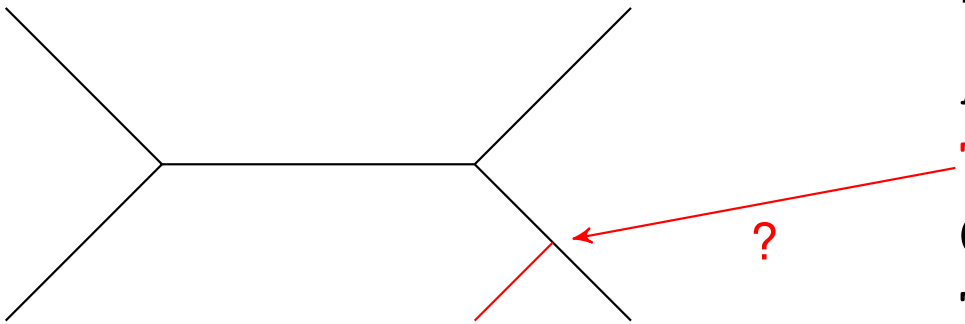
TTATCC

GATACA

TGATAT

- Take sequence, find **optimal placement** on existing tree
- Optimality meaning Bayesian posterior probability criterion
- Remember where sequence was placed
- Done using the `pplacer` tool

Place Sequence on Tree



16S Sequences

AACGTA ✓(< *middle* >)

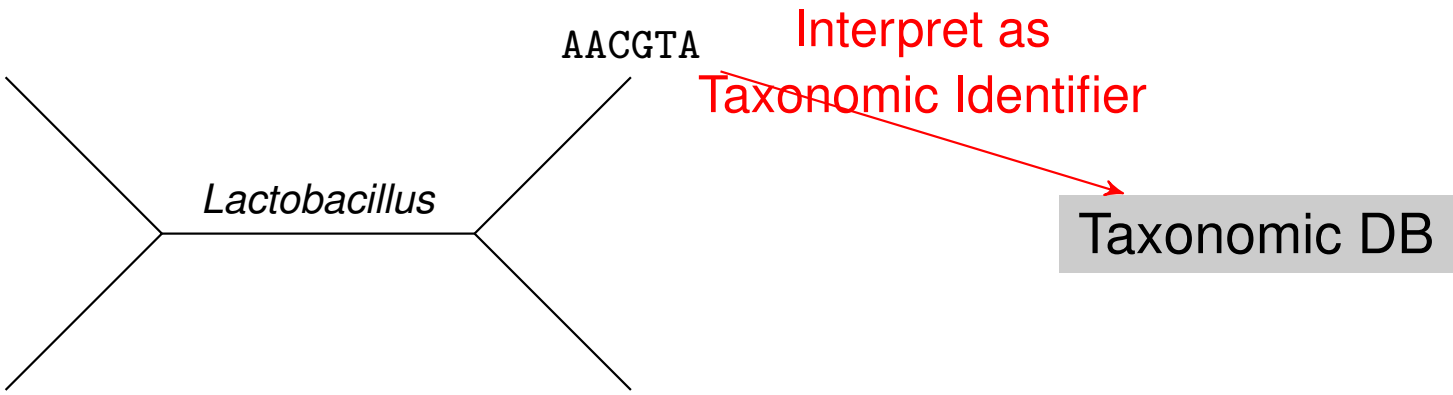
TTATCC

GATACA

TGATAT

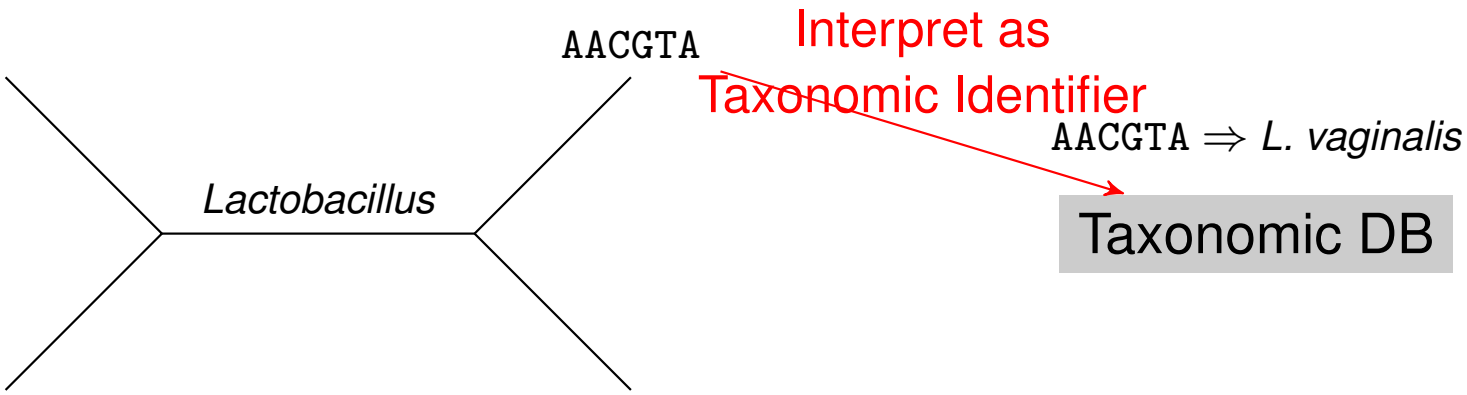
- Take sequence, find **optimal placement** on existing tree
- Optimality meaning Bayesian posterior probability criterion
- Remember where sequence was placed
- Done using the `pplacer` tool

Taxonomic Assignment



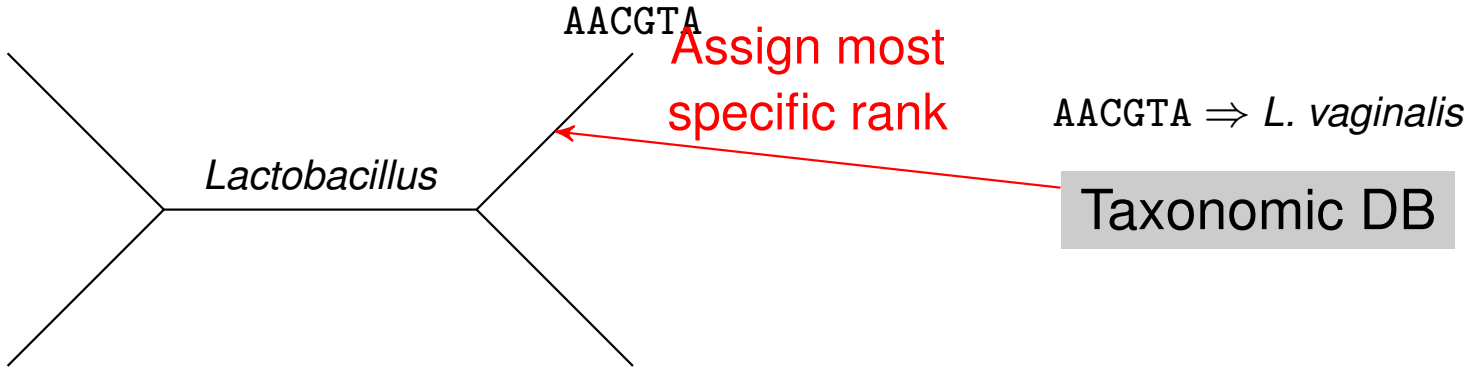
- Assign taxonomic labels to **edges** of the tree
- Such that labels are as specific as possible (species, genus, family etc.)

Taxonomic Assignment



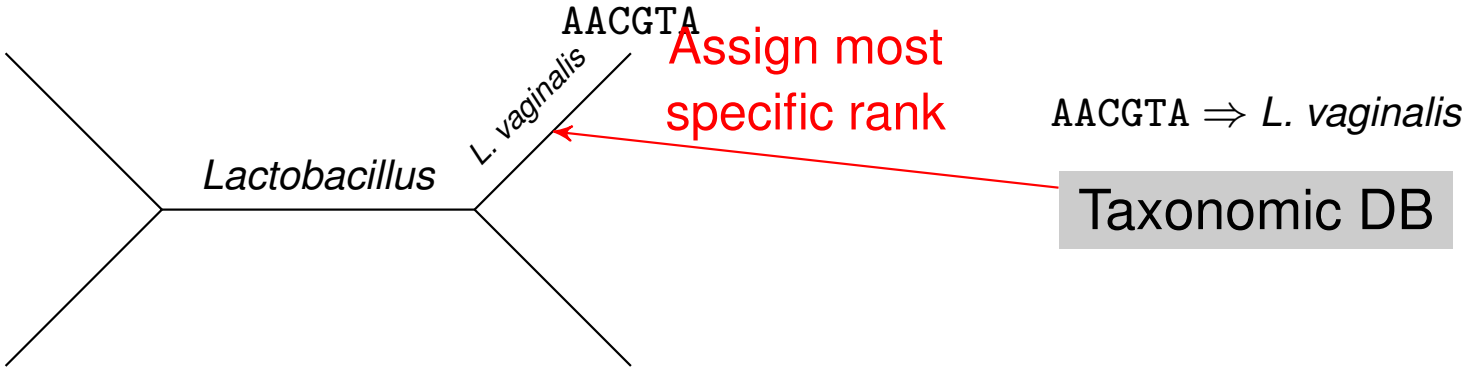
- Assign taxonomic labels to **edges** of the tree
- Such that labels are as specific as possible (species, genus, family etc.)

Taxonomic Assignment



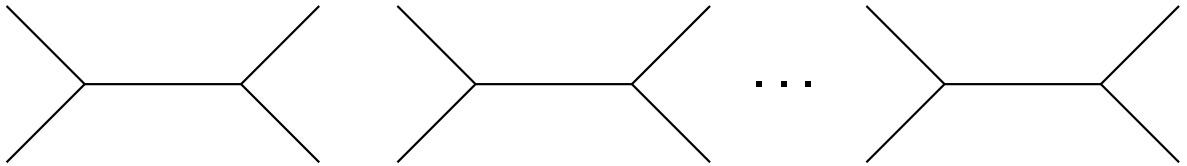
- Assign taxonomic labels to **edges** of the tree
- Such that labels are as specific as possible (species, genus, family etc.)

Taxonomic Assignment



- Assign taxonomic labels to **edges** of the tree
- Such that labels are as specific as possible (species, genus, family etc.)

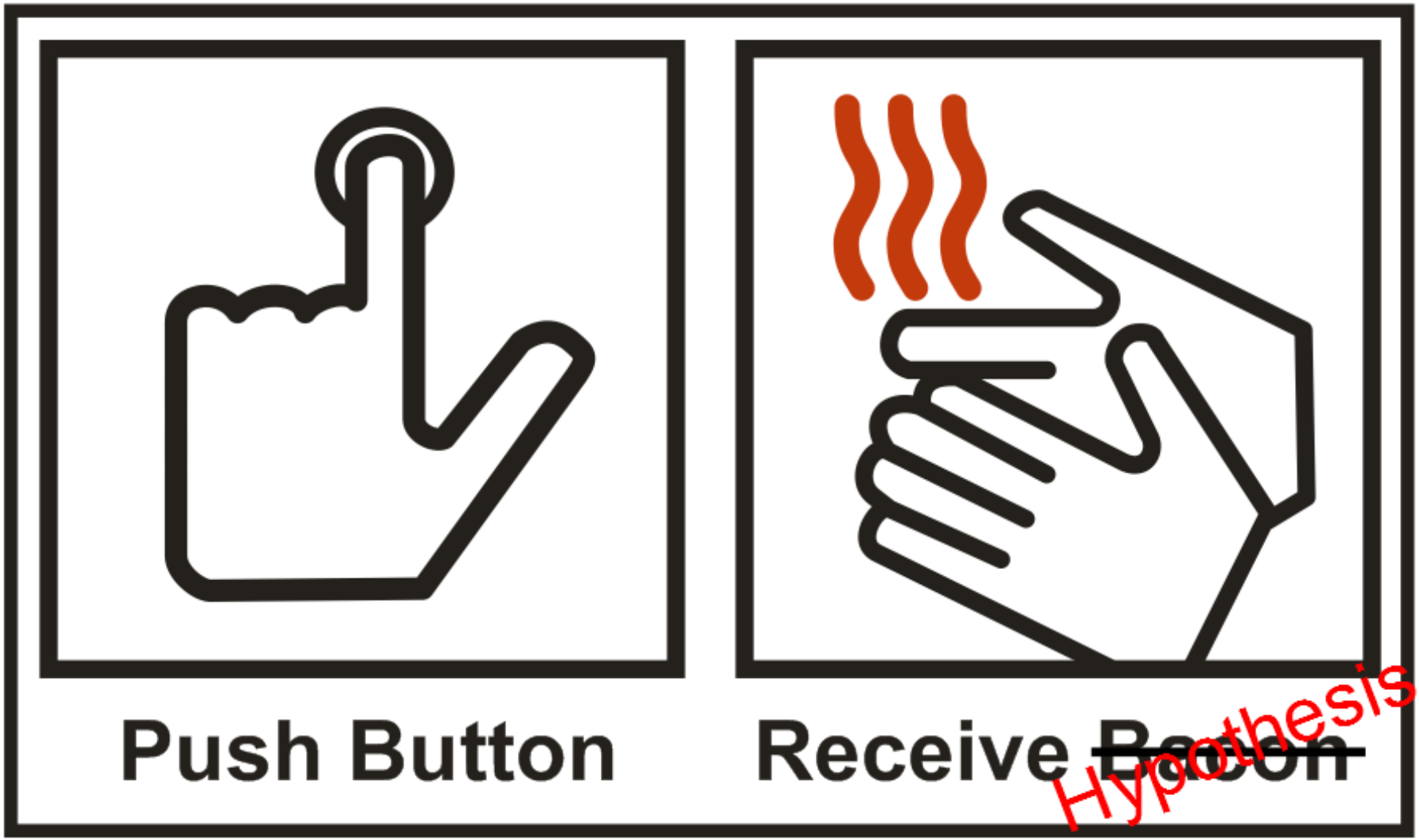
Classification Result



- 220 *virtual* trees, one per sample
  - Virtual as in sequences are **not contained** on one common tree, but are **associated** with the reference tree by coordinates
- Each representing the bacterial composition of a patients vaginal environment
- let's call them **sample-trees**
- Basis for further statistical evaluation

Outline

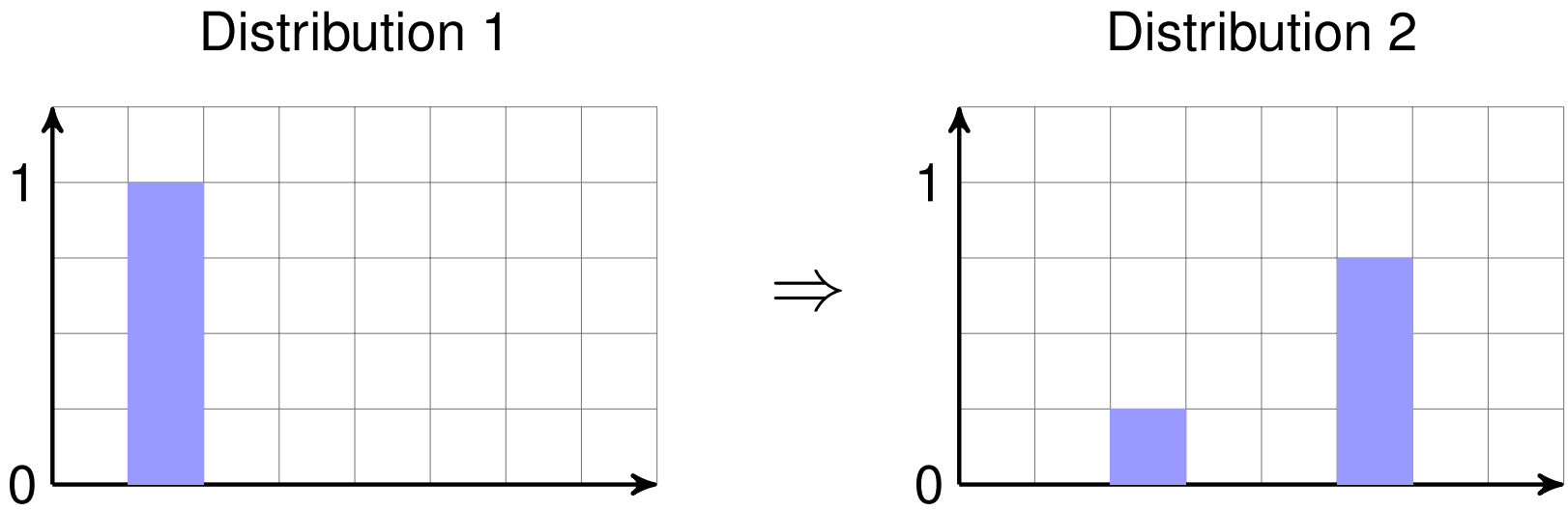
Correlation Analysis



What next?

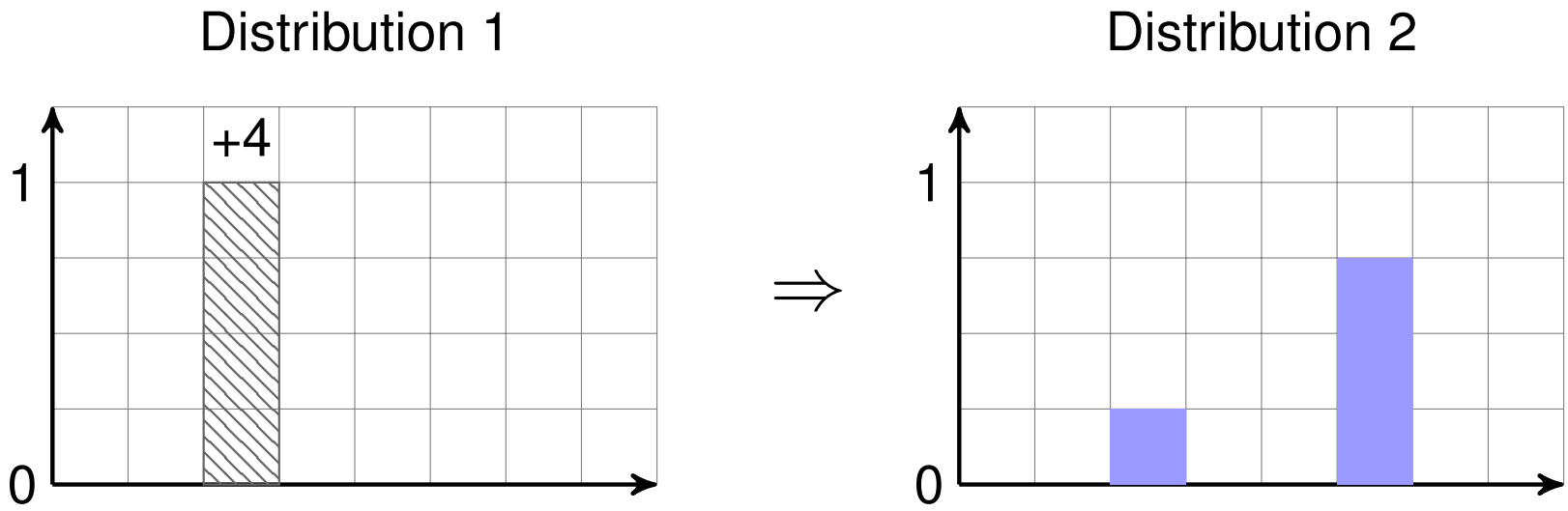
- Now that we have sample-trees we want **compare them**
- In the paper this is done by assembling them into another tree, a **tree of trees**
- To do that we need a **distance metric** and a way to **cluster** the sample-trees

Earth-Mover Distance



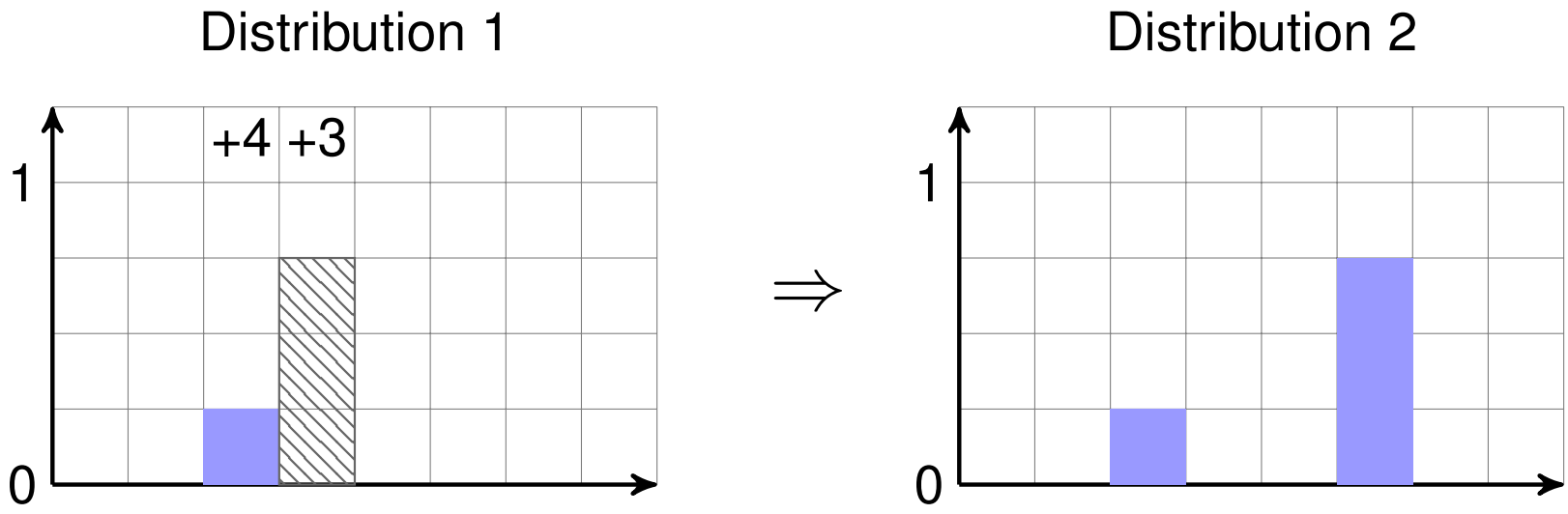
- Distance between two distributions
- Blue = mass
- Distance is the **work required to shift the mass** such that distributions are equal

Earth-Mover Distance



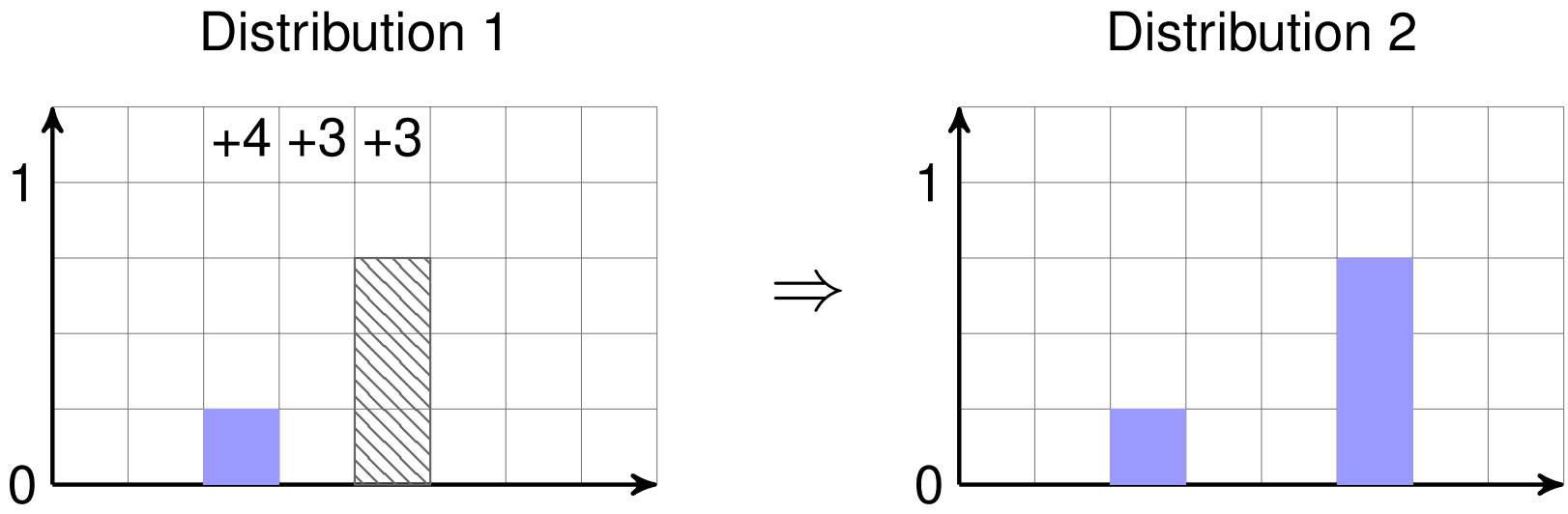
- Distance between two distributions
- Blue = mass
- Distance is the **work required to shift the mass** such that distributions are equal

Earth-Mover Distance



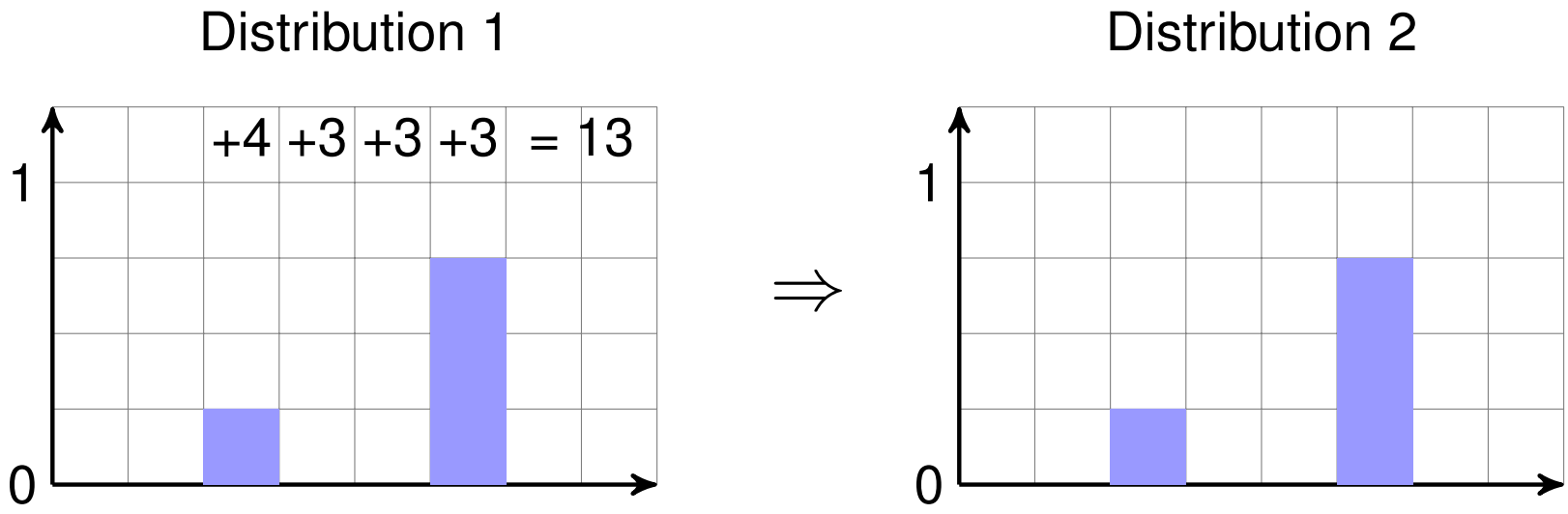
- Distance between two distributions
- Blue = mass
- Distance is the **work required to shift the mass** such that distributions are equal

Earth-Mover Distance



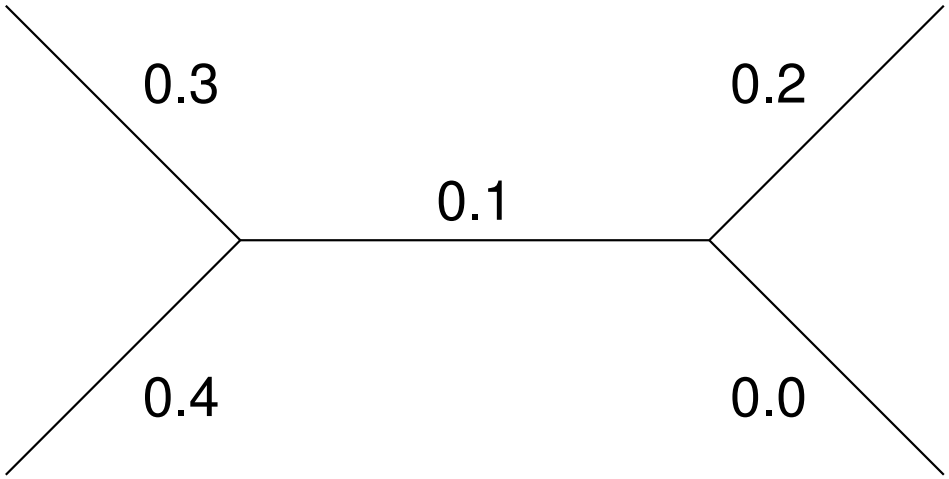
- Distance between two distributions
- Blue = mass
- Distance is the **work required to shift the mass** such that distributions are equal

Earth-Mover Distance



- Distance between two distributions
- Blue = mass
- Distance is the **work required to shift the mass** such that distributions are equal

Samples as Distribution on the Tree

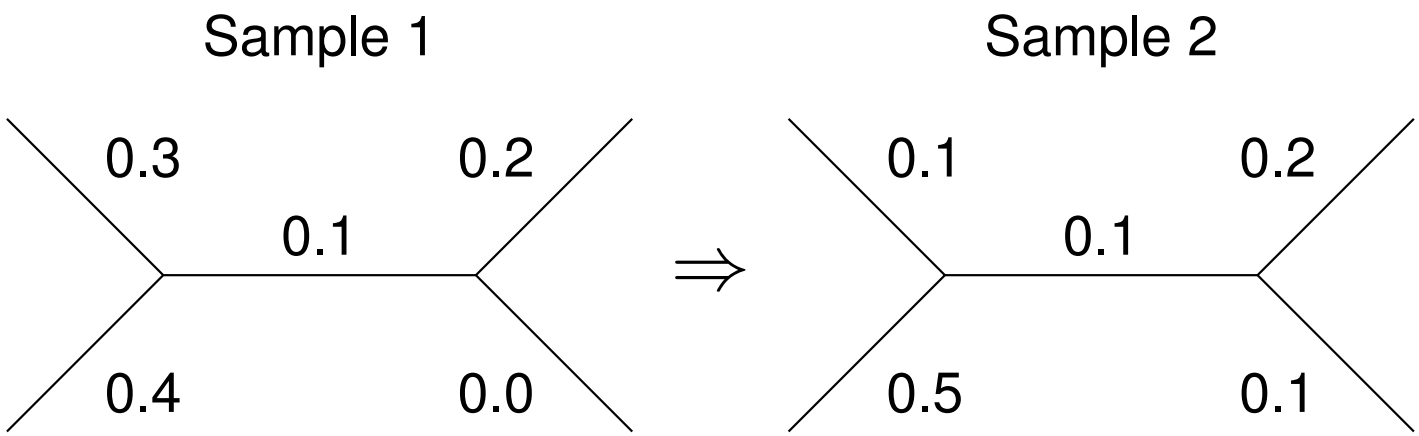


- Edge labels: fraction of total reads that were placed at an edge

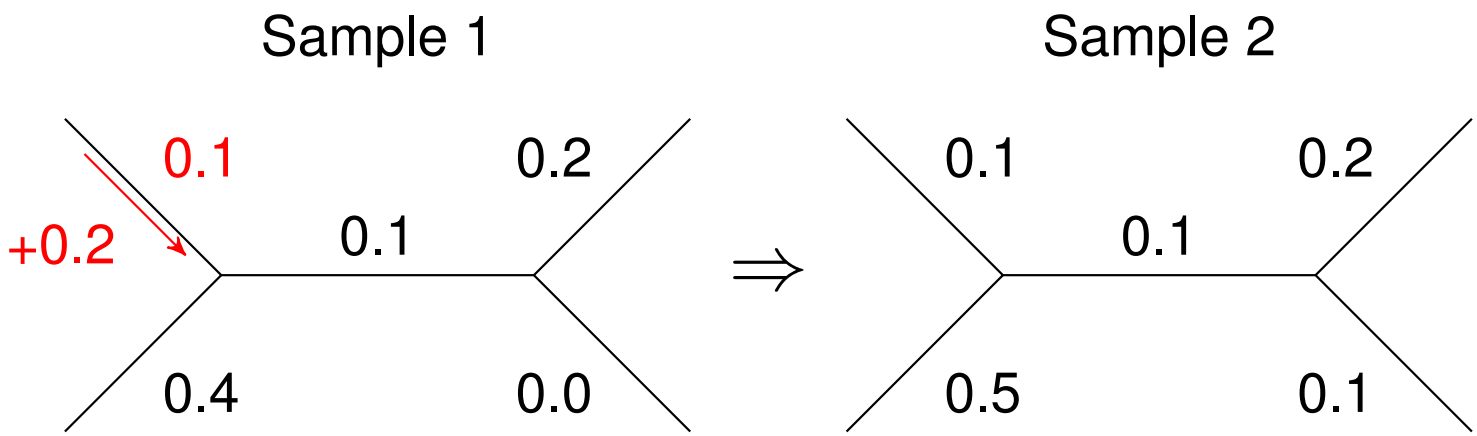
Phylogenetic Kantorovich-Rubinstein

- Combination of **earth-mover distance** and **trees with read distribution**
- Apply earth-mover distance between the edges of two trees
- Distance is **minimal amount of work required** to move mass to match other distribution

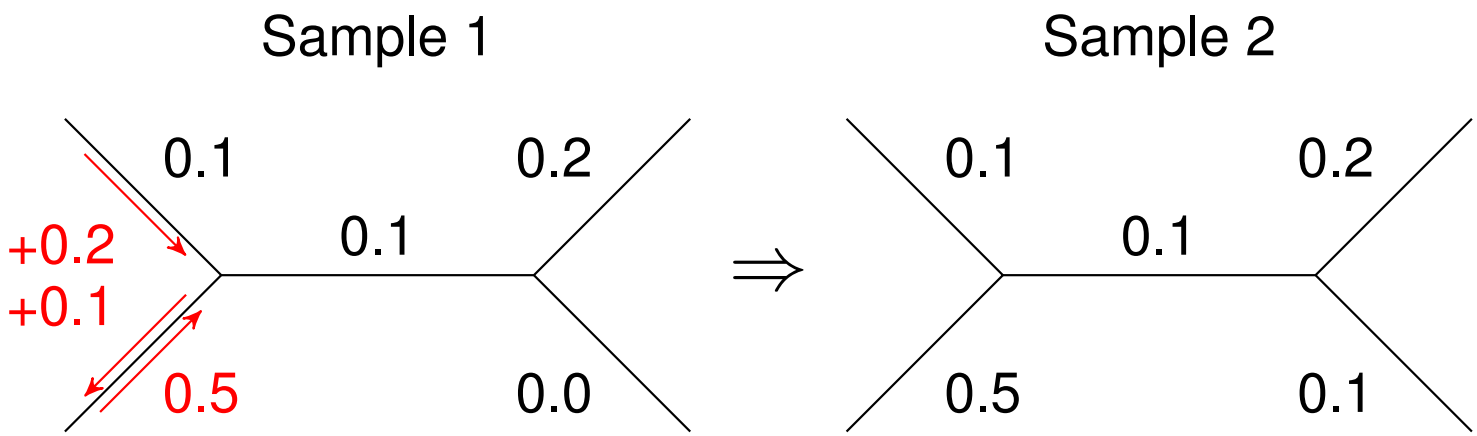
Kantorovich-Rubinstein Visualisation



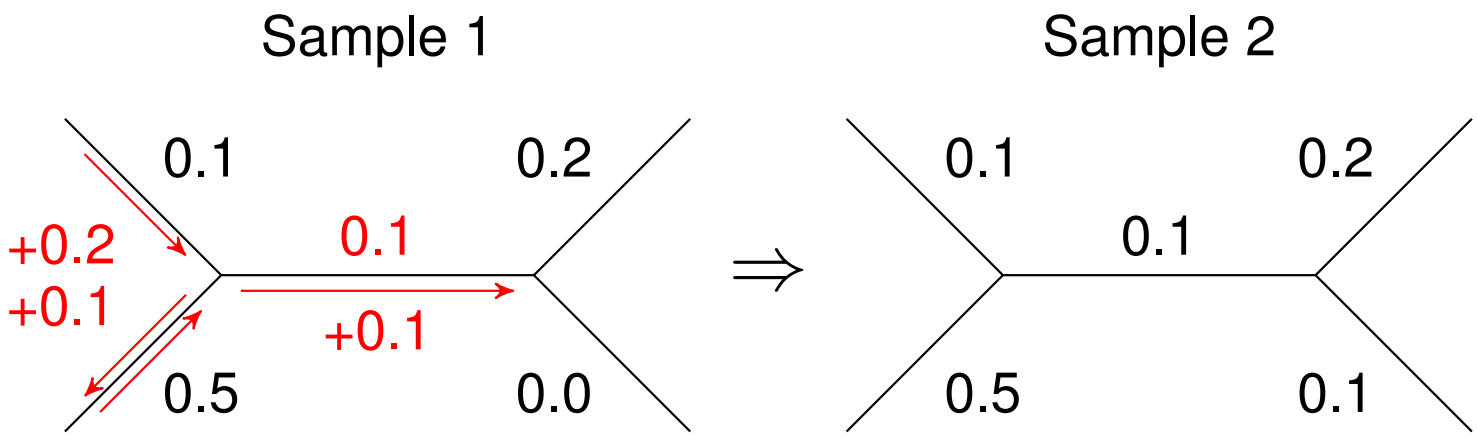
Kantorovich-Rubinstein Visualisation



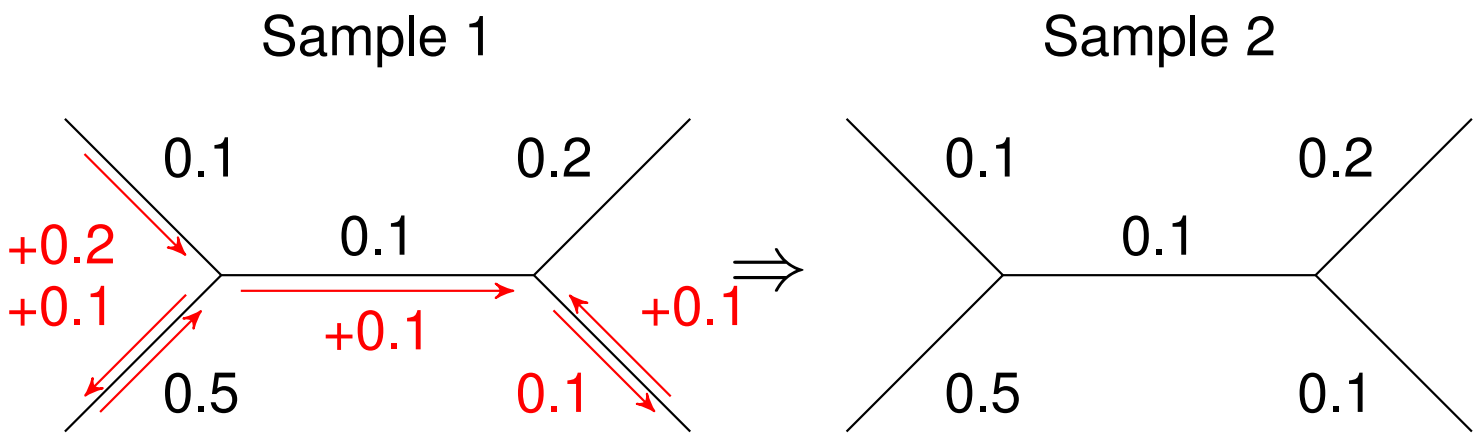
Kantorovich-Rubinstein Visualisation



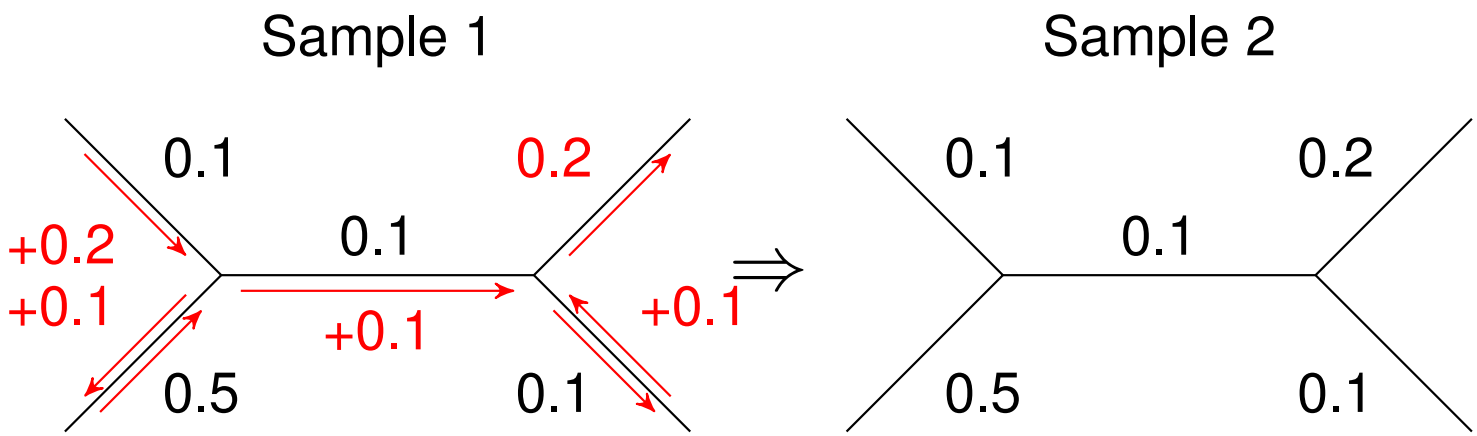
Kantorovich-Rubinstein Visualisation



Kantorovich-Rubinstein Visualisation



Kantorovich-Rubinstein Visualisation



Distance = 0.5

Reminder: Hierarchical Clustering

Pairwise Distance Matrix (PWD)

	A	B	C	D
A		17	21	27
B			12	18
C				14
D				

Reminder: Hierarchical Clustering

Pairwise Distance Matrix (PWD)

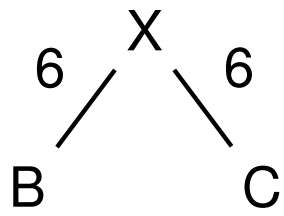
	A	B	C	D
A		17	21	27
B			12	18
C				14
D				

B C

Reminder: Hierarchical Clustering

Pairwise Distance Matrix (PWD)

	A	X	D
A		13	27
X			18
D			

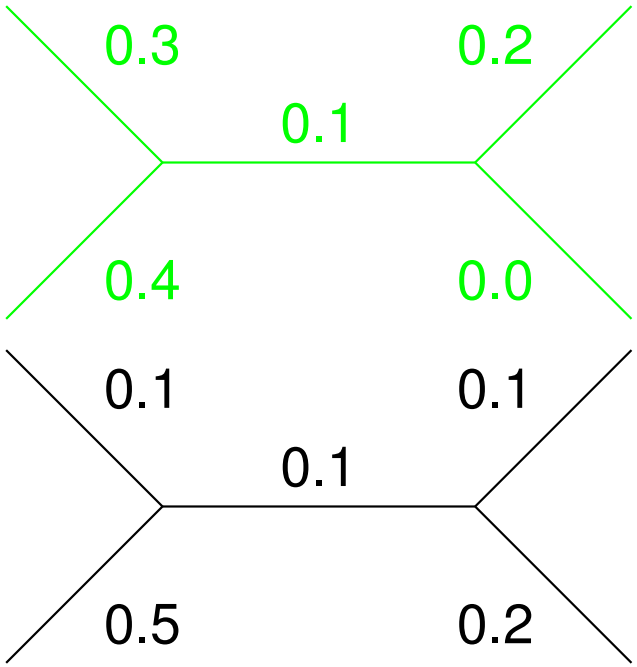


merging and
branch length assignment

Squash Clustering

- Building a tree of trees
- Sample-trees at the tips
- Pairwise distance matrix based on K-R Distance
- Merge by building the **average of the distributions** (*squashing*)
- Branch lengths = K-R Distance between trees at two incident nodes

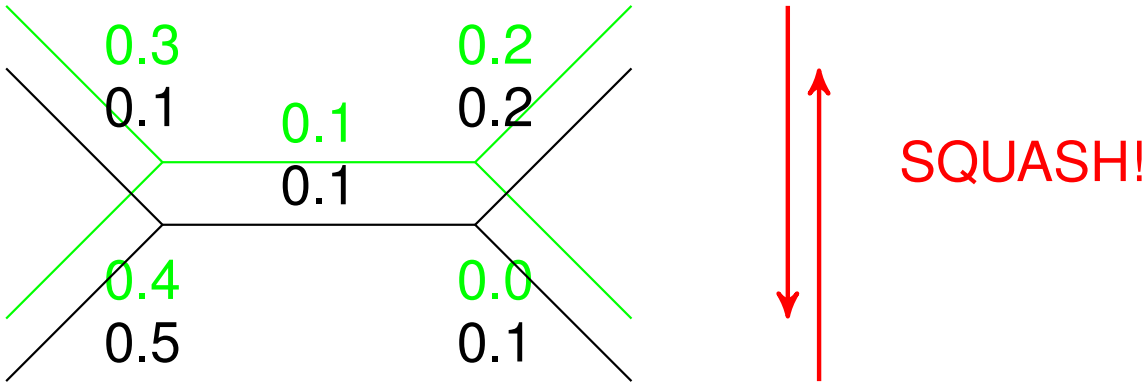
Squashing Visualised



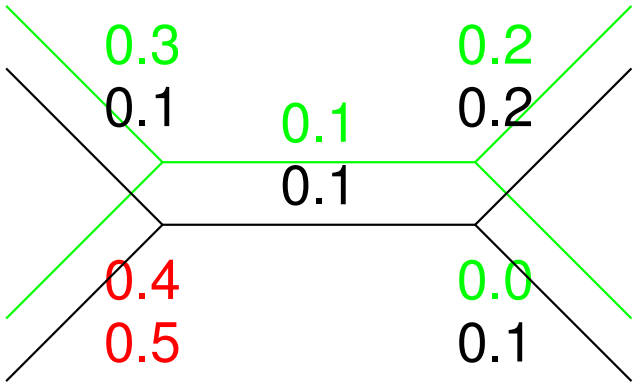
Sample 1

Sample 2

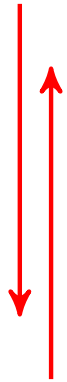
Squashing Visualised



Squashing Visualised

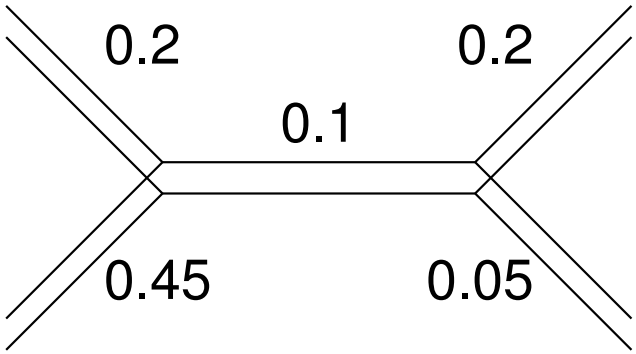


$(0.4 + 0.5)/2 = 0.45$

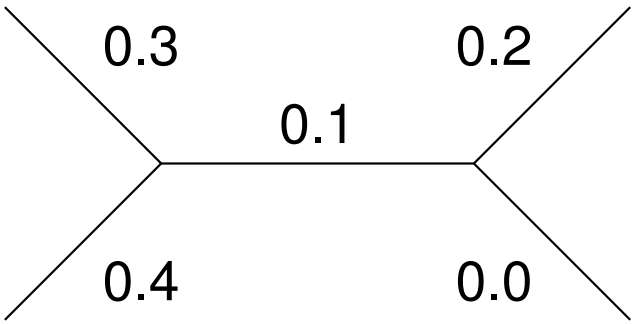


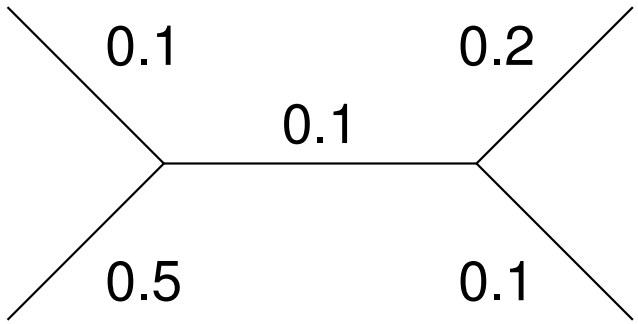
SQUASH!

Squashing Visualised

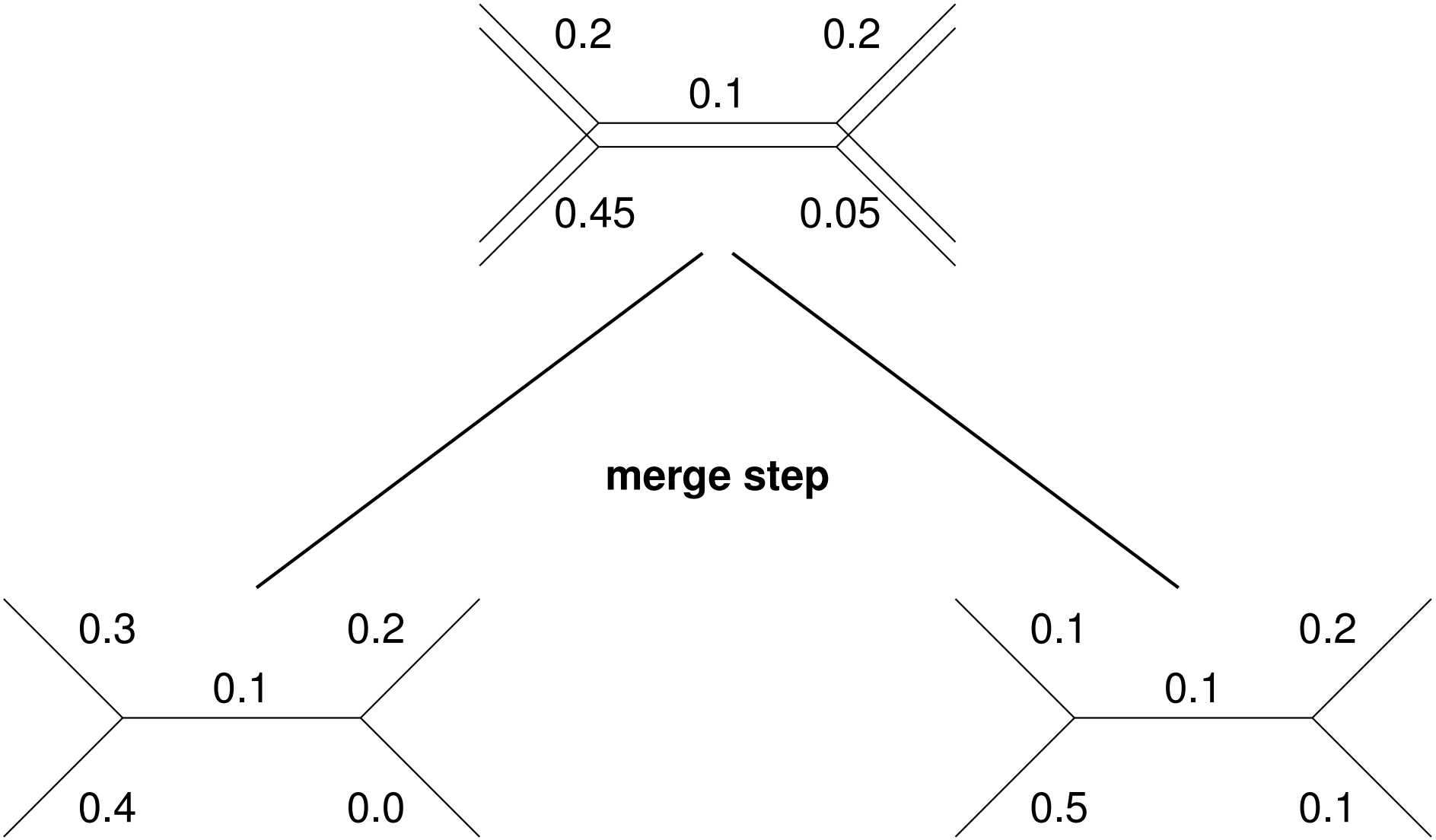


Squash Clustering Visualised

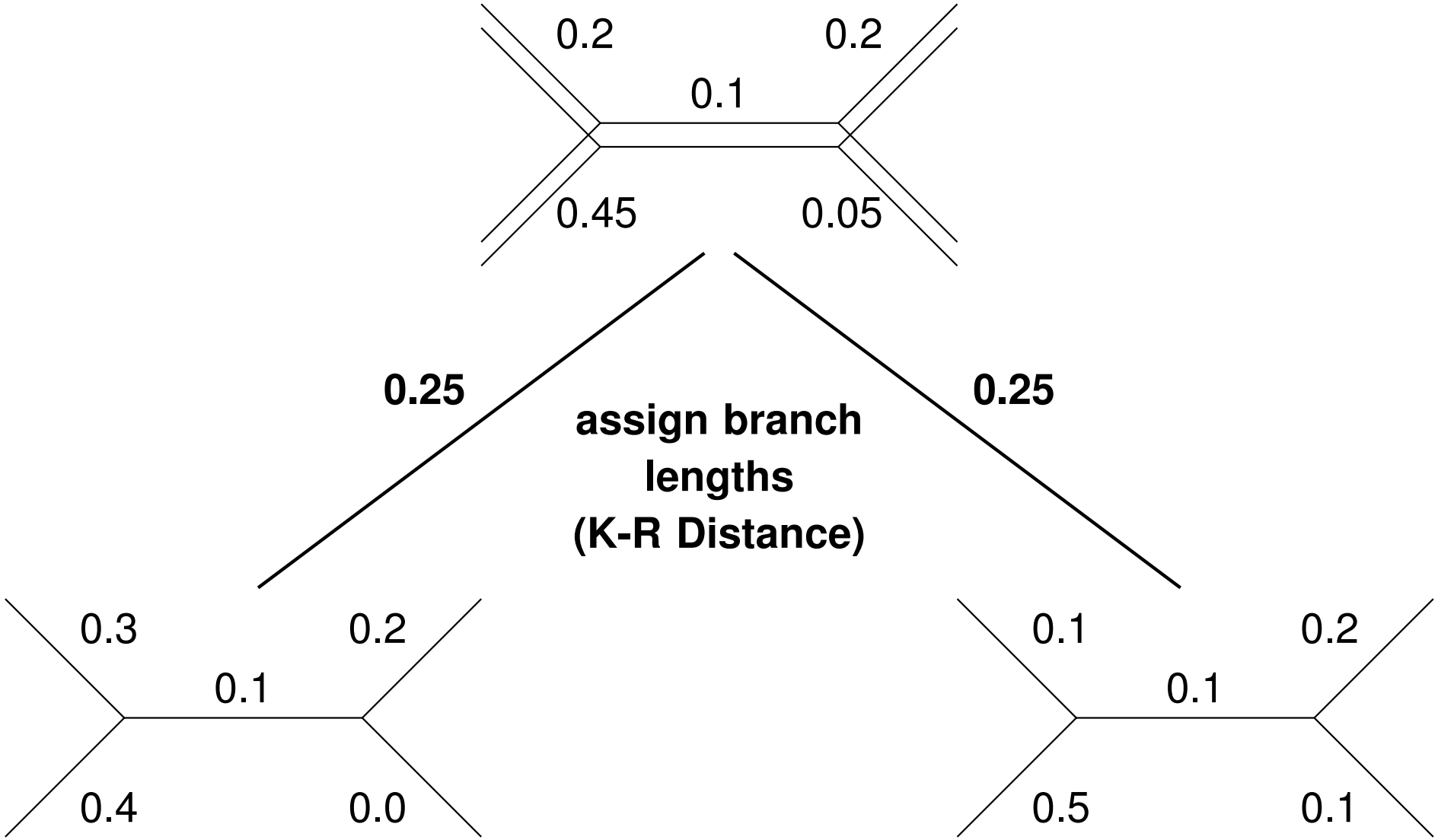




Squash Clustering Visualised

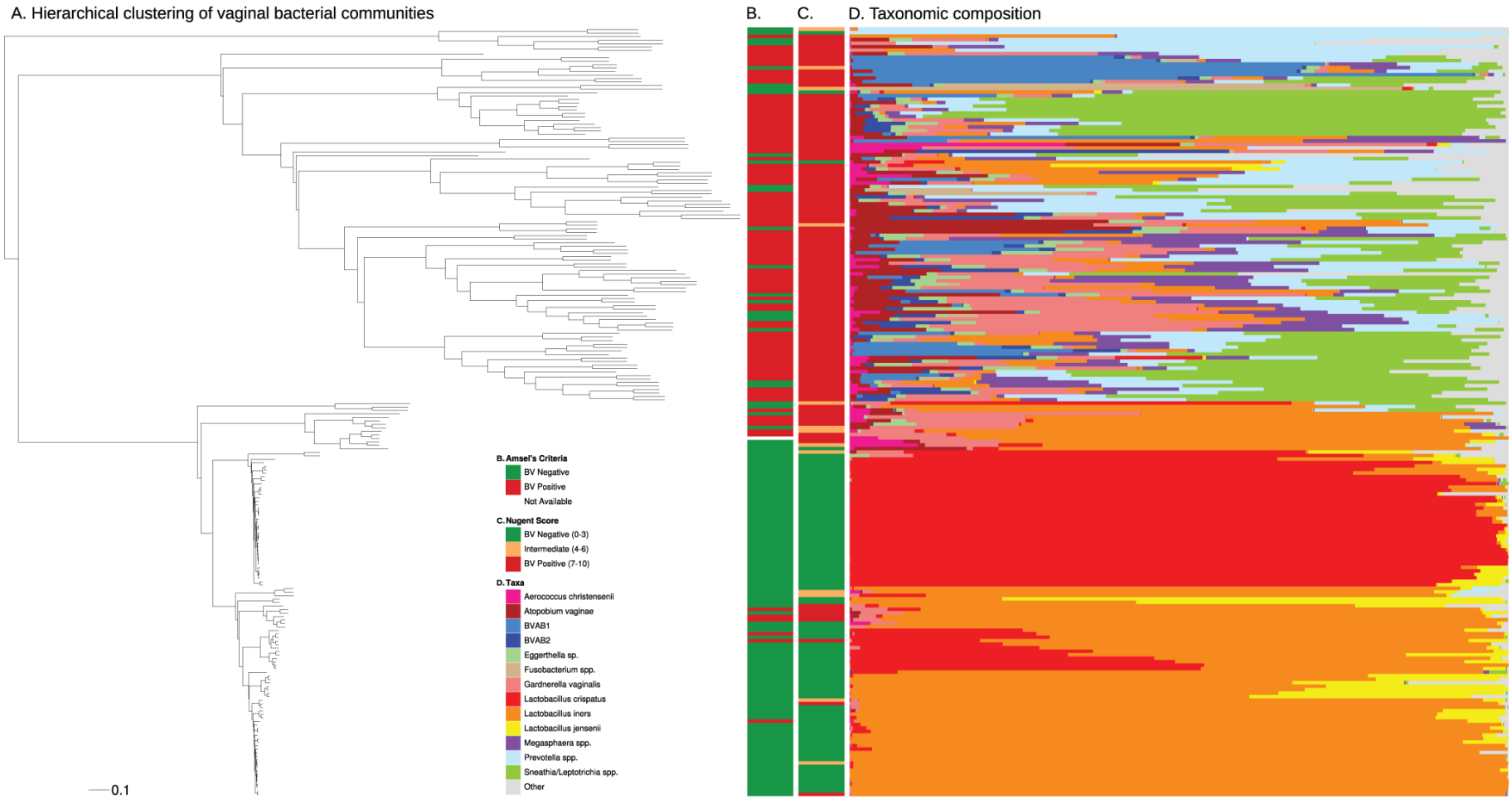


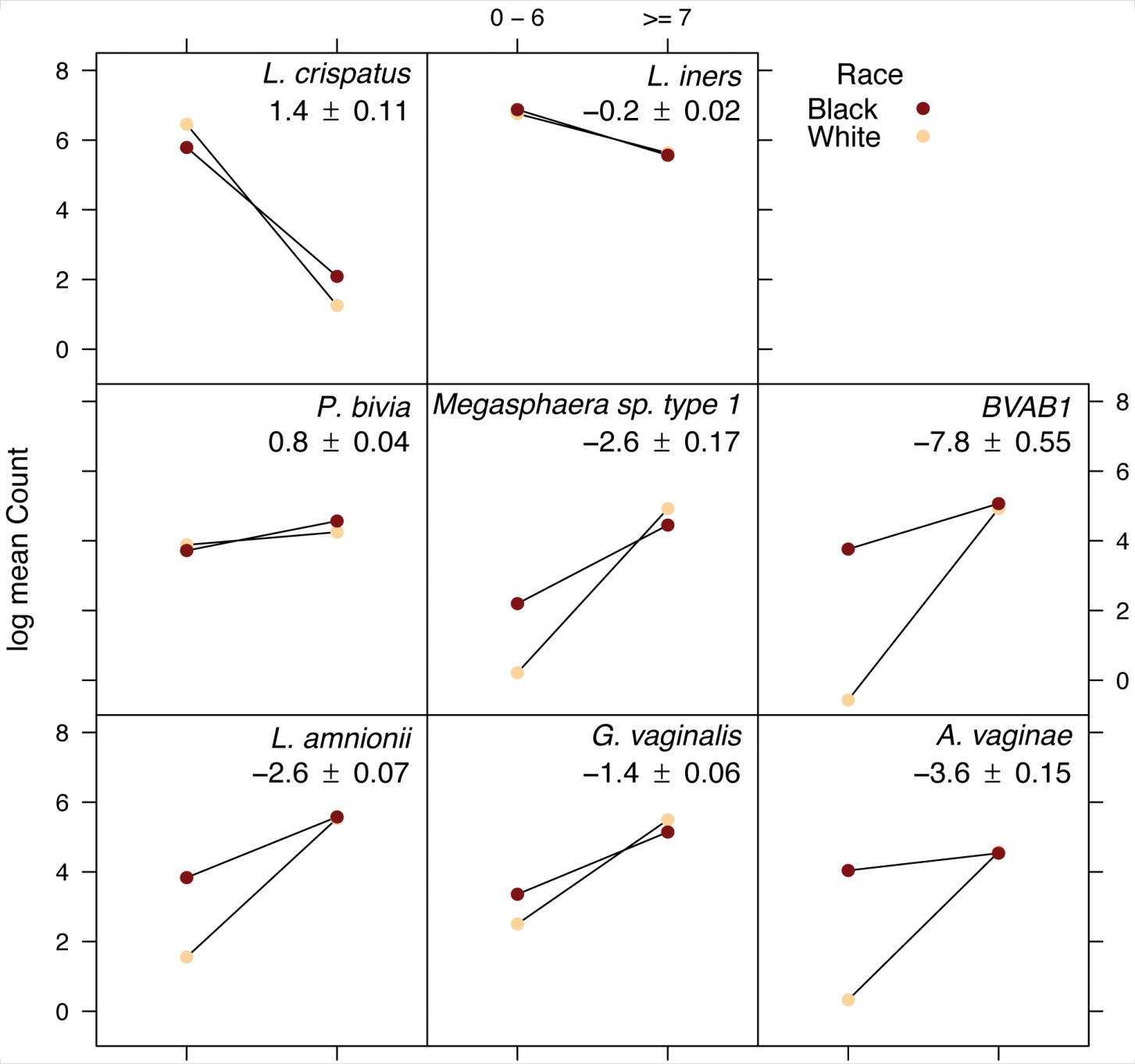
Squash Clustering Visualised

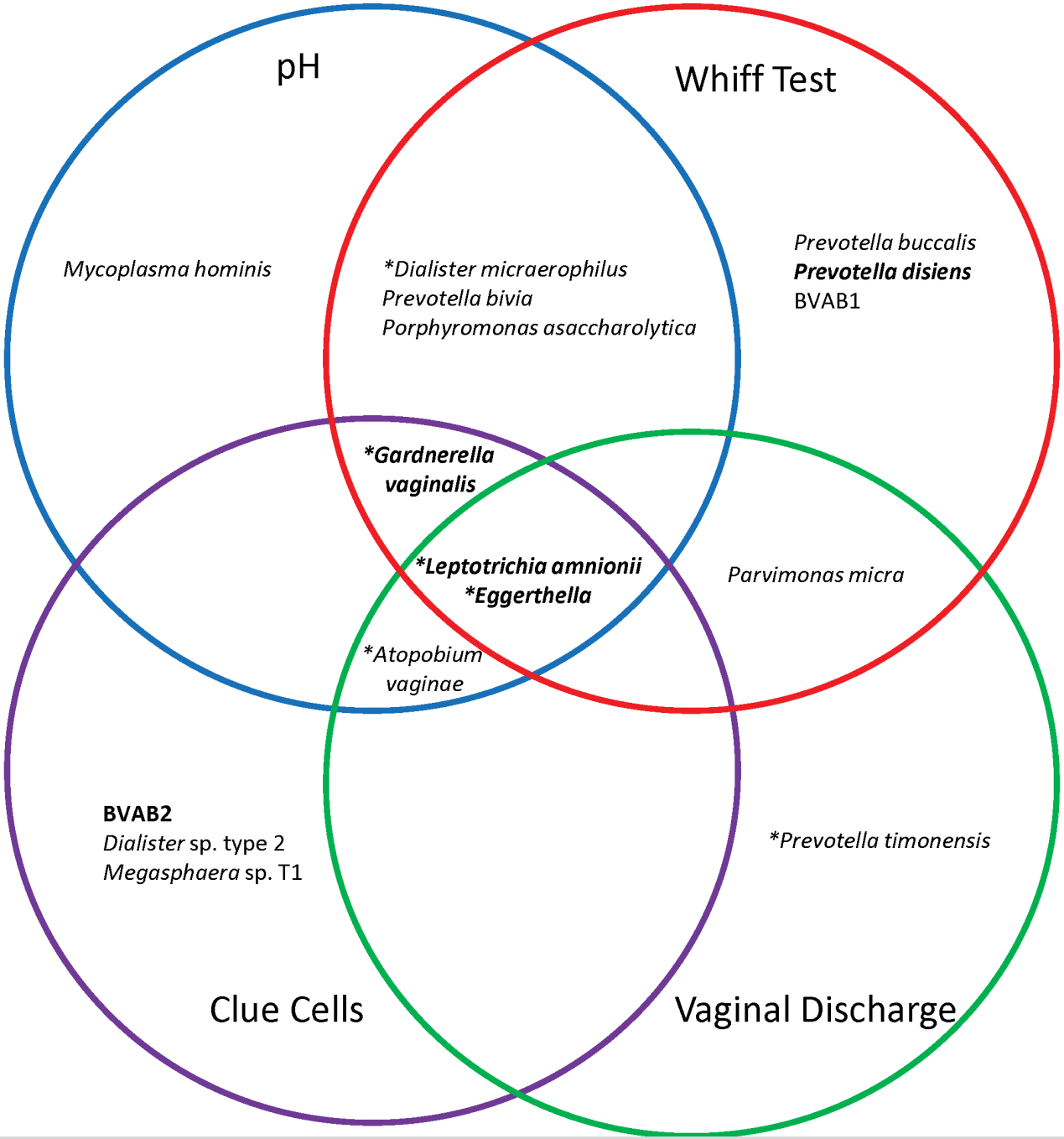


Outline

Results of Clustering




0 – 6
>= 7
L. crispatus
1.4 ± 0.11
L. iners
−0.2 ± 0.02
Race
Black
White
P. bivia
0.8 ± 0.04
Megasphaera sp. type 1
−2.6 ± 0.17
BVAB1
−7.8 ± 0.55
L. amnionii
−2.6 ± 0.07
G. vaginalis
−1.4 ± 0.06
A. vaginae
−3.6 ± 0.15
log mean Count
0
2
4
6
8


pH
Whiff Test
Clue Cells
Vaginal Discharge
Mycoplasma hominis
**Dialister micraerophilus*
Prevotella bivia
Porphyromonas asaccharolytica
Prevotella buccalis
Prevotella disiens
BVAB1
****Gardnerella vaginalis***
****Leptotrichia amnionii***
****Eggerthella***
**Atopobium vaginae*
Parvimonas micra
BVAB2
Dialister sp. type 2
Megasphaera sp. T1
**Prevotella timonensis*

Results

- Healthy vaginal microbiome dominated by *Lactobacillus* spp.
- Women with BV have highly diverse vaginal microbiome
- Clinical tests of BV correlate differently well to bacteria
- Race appears to have influence on whether some bacteria contribute to BV

References I

Sujatha Srinivasan et al. “Bacterial communities in women with bacterial vaginosis: High resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria”. In: *PLoS ONE* 7.6 (2012). ISSN: 19326203. DOI: 10.1371/journal.pone.0037818.

Image sources

- Slide 4:
  - https://en.wikipedia.org/wiki/Petri_dish#/media/File:Agar_plate_with_colonies.jpg, Wikipedia user Phyzome
- Slide 5:
  - https://www.flickr.com/photos/pere/523019984, flickr user pere, with modification
- Slide 10:
  - https://en.wikipedia.org/wiki/Cotton_swab#/media/File:White_menbo.jpg, Wikipedia user Aney
  - https://commons.wikimedia.org/wiki/File:DNA_sequence.svg, Wikimedia user Sjef
  - Scatterplots taken from Srinivasan et al. 2012
- Slide 15:
  - https://en.wikipedia.org/wiki/Ribosome#/media/File:Peptide_syn.png, Wikipedia user Boumphreyfr
- Slide 15:
  - https://en.wikipedia.org/wiki/Polymerase_chain_reaction#/media/File:Polymerase_chain_reaction.svg, Wikipedia user Enzoklop
- Slide 26:
  - https://www.flickr.com/photos/darkuncle/4421756078/, flickr user darkuncle, modification by me

Identifying Novel Bacteria

- Take all sequenced reads from interesting bacterial order (here: *Clostridiales*)
- Place on tree, cluster into *islands* with distance cutoff 0.02
- Throw away islands that have reads only from one individual
- Choose representative from reads arbitrarily, `BLAST` it to find appropriate island label
- Display islands as leaves on Ref. tree

Results Novel Bacteria

- 11 novel bacteria identified
- Less than 97% identity to known bacteria
- Range: 91% to 96%

Some Numbers

- 425775 sequence reads from 220 samples
- Median read length: 225bp
- Mean number of reads per subject: 1620
- 99.1% of reads were classified at species level

Kantorovic-Rubinstein Formula

$$Z(P, Q) = \int_T |P(\tau(y)) - Q(\tau(y))| \lambda(dy)$$

Z is the resulting distance

P, Q are the distributions

T is the tree

τ signifies the sub tree below vertex y

λ is the length measure

Classifier Validation

- Assemble validation set of 16S sequences belonging to *Lactobacillus* genus (source: RDP)
- Consisting of subspecies found in human vagina
- Also met previous distance metric from mislabel detection
- Trim sequences to match V4 16S region

Pearson Correlation Coefficient

- Measure of the linear correlation (dependence) between two variables X and Y
- Gives a value between +1 and 1 inclusive
- 1 is total positive correlation
- 0 is no correlation
- 1 is total negative correlation

p-value

- Qualifies the significance of a result
- Tells you whether your data rejects your *null hypothesis* (NH)
- $0 \leq p \leq 1$
- $p \leq 0.05$: strong evidence against NH $\Rightarrow$ reject NH
- $p > 0.05$: weak evidence against NH $\Rightarrow$ failed to reject NH