Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning (LSA)

Brian Cleary[1,2], Ilana Lauren Brito[2,3,4], Katherine Huang[2], Dirk Gevers[2], Terrance Shea[2], Sarah Young[2], and Eric Alm[2,3,4]

Nature Biotechnology, October 2015

1- Computational and Systems Biology Program, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

2- Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA

3- Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

4- Center for Microbiome Informatics and Therapeutics, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

Dan Coster, 12/06/19

E-Mail: Dancoster@gmail.com

Talk Outline

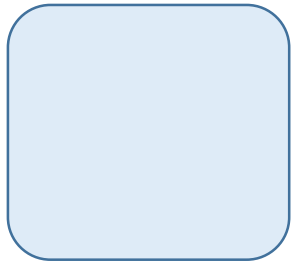
- Background
- Motivation
- LSA - Latent Strain Analysis
- Results
- Pros & cons
- Discussion

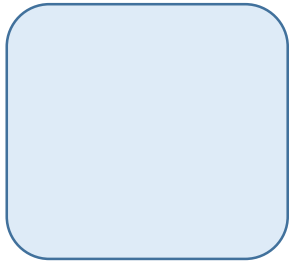
16s Ribosome RNA

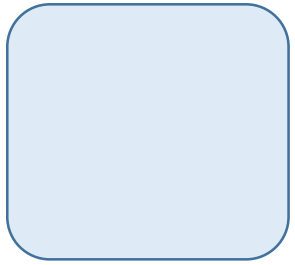
Shotgun Sequencing

What do they do?

Who is there?







SRAIN

SPECIES

GENUS

FAMILY

ORDER

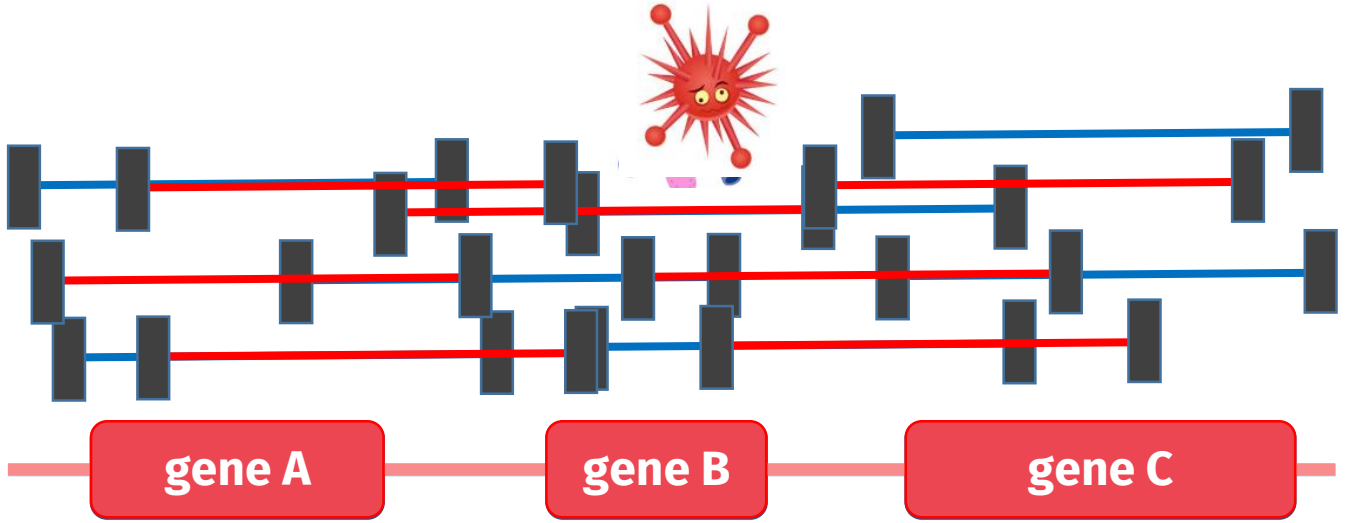
CLASS

PHYLOM

KINGDOM

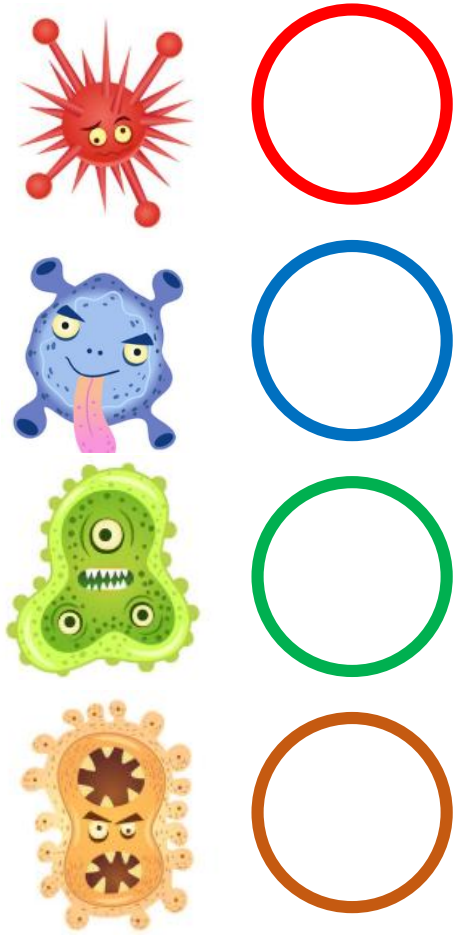
LSA requires fixed amount of memory to detect strain level

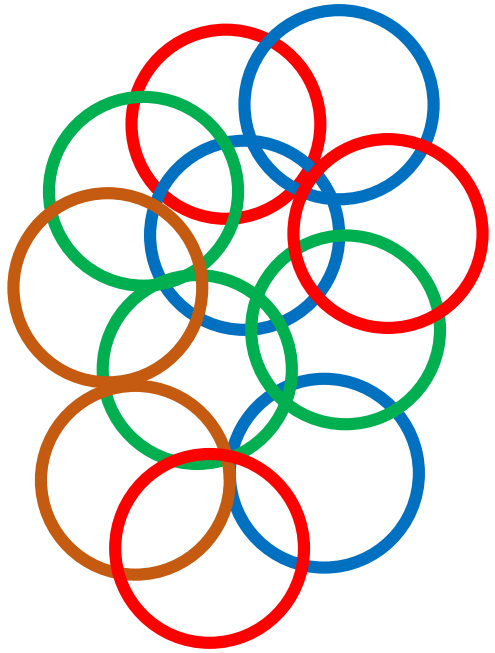
Shotgun Sequencing



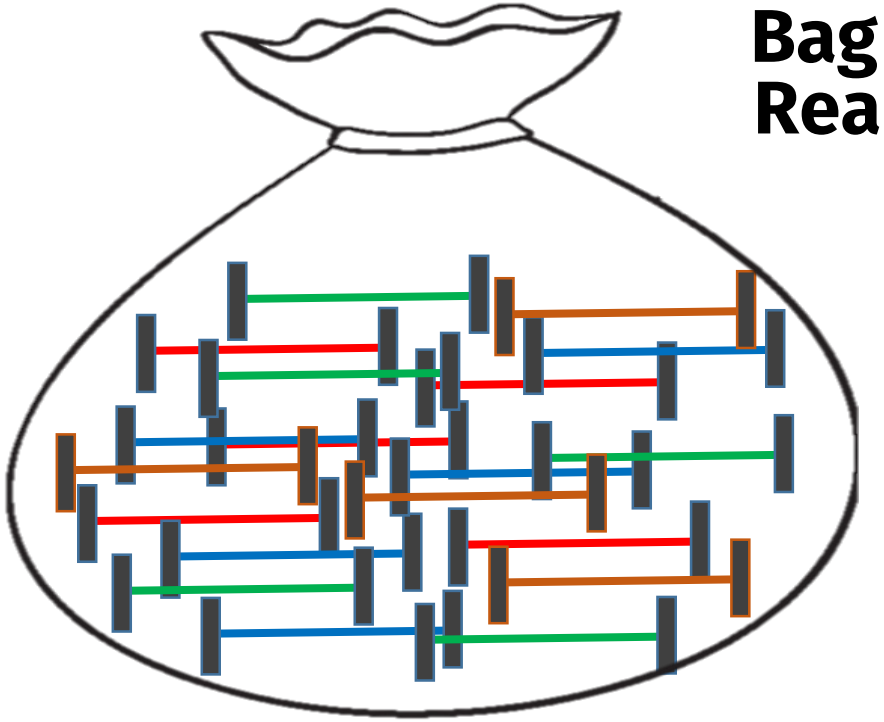
DNA

Sample





Bag of Reads



The Challenge

Deep Sequencing

Detection of Low abundance species
Separate strains of the same species

(Using commodity hardware)

The Challenge

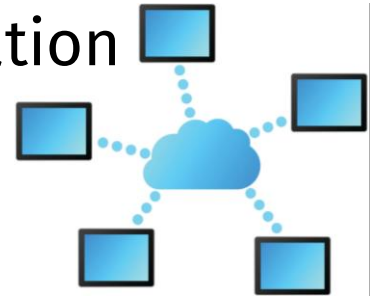
MetAMOS, MetaVelvet, Meta-IDBA

- Relaxing the assumption of single-genome de-Bruign assemblers to allow multiple coverage

Cons: can't scale to terabytes data sets

Ray Meta

- Leverage distributed architectures to parallelize assembly computation

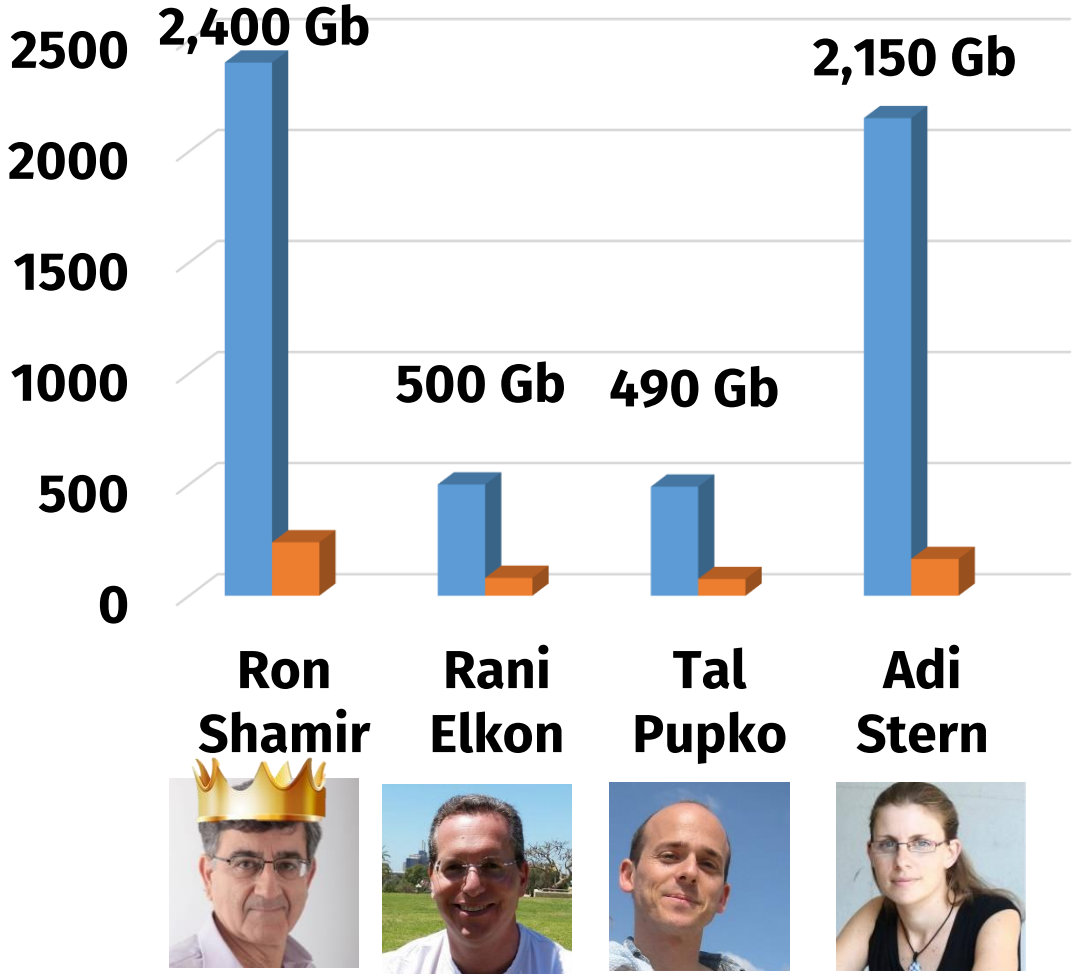


Diginorm & Khmer

- A combination of data reduction + data compression + partioning

Cons: (1) Multiple small contigs
(2) It's not clear which contigs originate from the same species

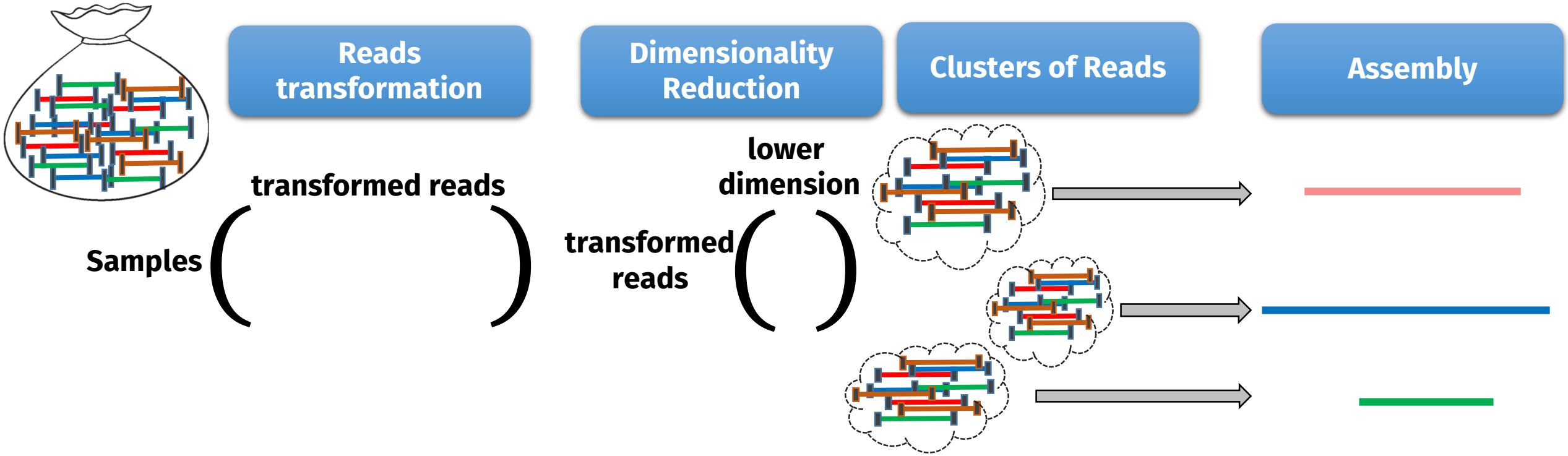
RAM



100-200 Gb per student

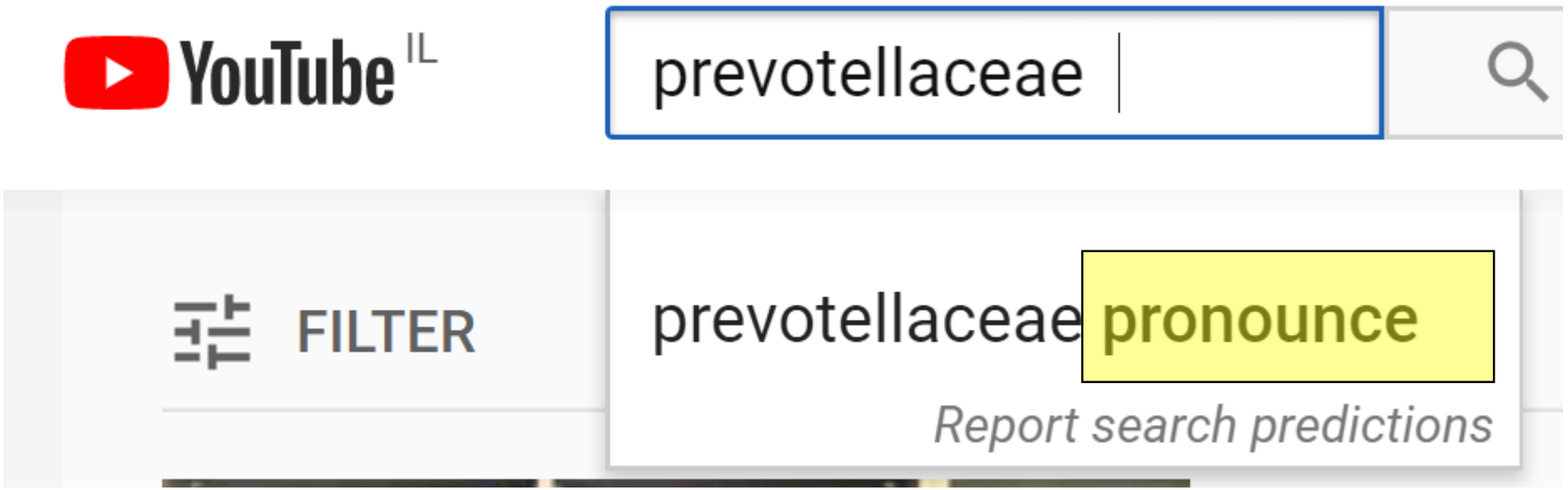
LSA – Latent Strain Analysis: High Level Description

A scalable ‘de novo’ **pre-assembly method**, that **separate reads into biological informed partition** and thereby **enables assembly of individual genomes**

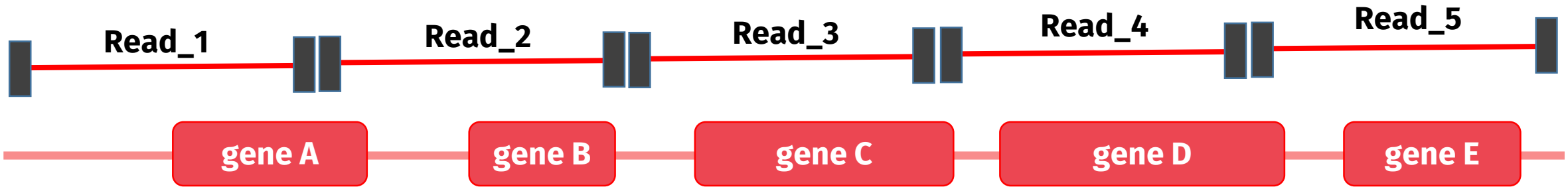


Intuition

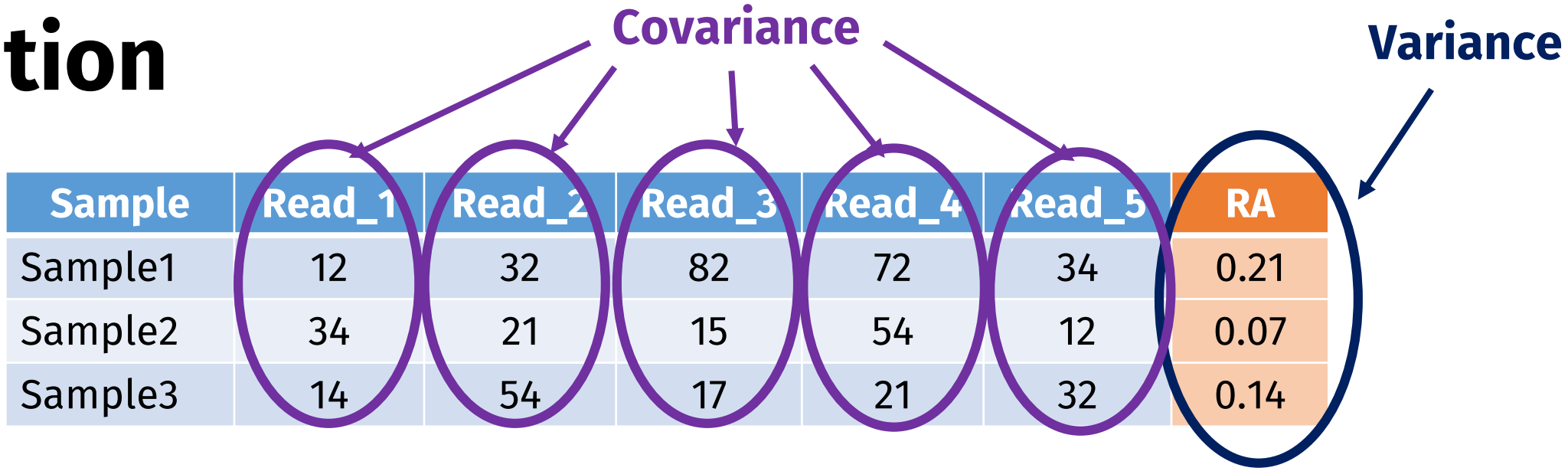
Relative Abundance (RA) of **Prevotellacea**



Let's assume that the Relative Abundance of Prevotellacea can be explained by a finite set of 5 reads



Intuition



Main Assumption

"The variance of a given species' abundance across samples imparts a covariance to the read depth at **every read** in that species' genome"

Reads transformation: From Read to K-mer

K-Mer - All the possible substrings of length k that are contained in a read

K=4	{ 1 , 2 , 1 , 1 , 1 , 1 }
ACTCTCTGAT	{ACTC, CTCT, TCTC, TCTG, CTGA, TGAT}

Assumption: The observed frequency of every k-mer in a sample is a function of the abundance of each distinct DNA fragment (read) containing that k-mer

Read (100-500 bp) → **A vector of k-mer counts (K=33 bp)**

Reads transformation: K-mer hashing & Normalization

K = 33 → 4^{33} **optional K-mers!** $O(10^{19})$

Hashing!

$4^{33} \rightarrow 2^{31}$ $O(10^{9})$

Main Assumption

The variance of a given species' abundance across samples imparts a covariance to the read depth at **every K-mer** in that species' genome"

TF-IDF Normalization ('global weight')

Sample	Hash1	Hash2	Hash3	...	Hash 2B
Sample1	0.1	0.28	0.75	...	0.4
Sample2	0.3	0.12	0.23	...	0.2
Sample3	0.7	0.45	0.09	...	0.014

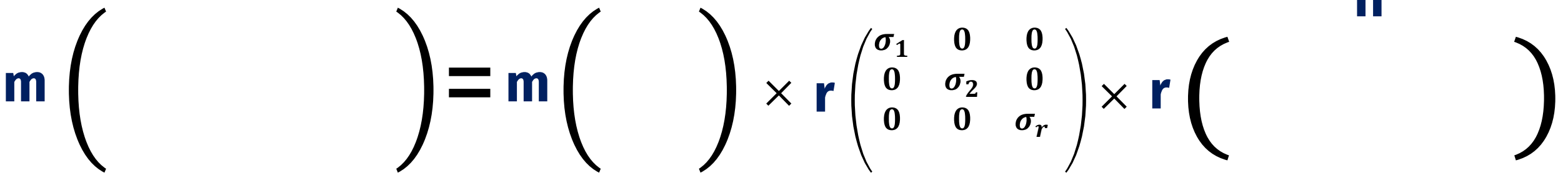
Abundance Matrix

Dimensionality Reduction : what is SVD?

Singular Value Decomposition

$$\mathbf{A} = \mathbf{U} \times \mathbf{\Sigma} \times \mathbf{V}^{\mathbf{T}}$$

input data — Left singular values — Singular values — Right singular values

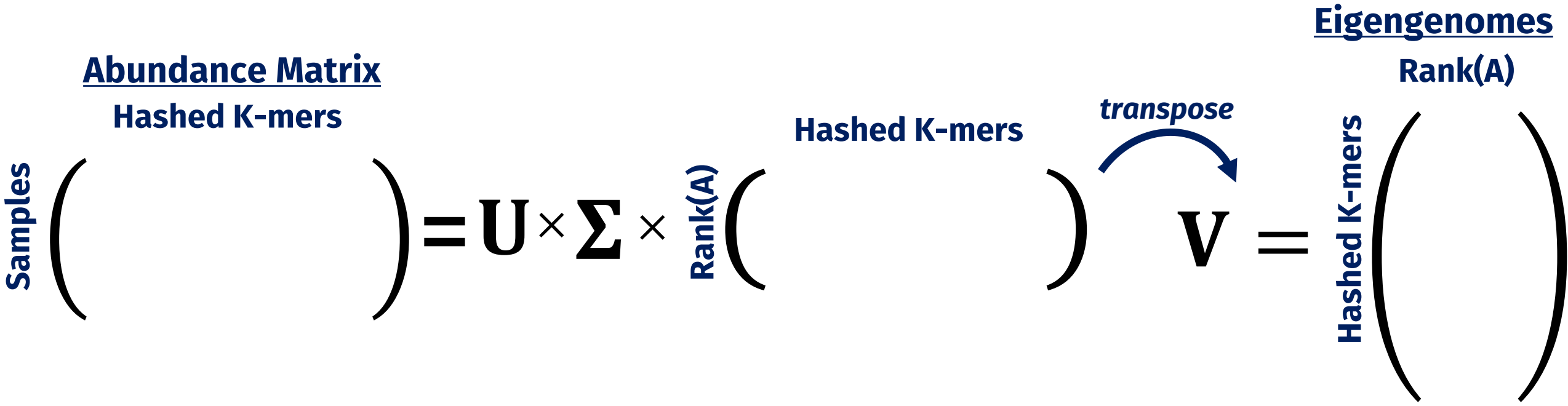


Claim: It is always possible to decompose A real matrix $\mathbf{A}$ to $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathbf{T}}$:

I. $\mathbf{U}, \mathbf{\Sigma}, V$ – unique
II. $\mathbf{U}, \mathbf{V}$ – column orthonormal $(U^T U = V^T V = I)$
III. $\mathbf{\Sigma}$ diagonal and its entries (singular values) are positive and sorted in decreasing order ($\sigma_1 >= \sigma_2 .. \geq 0$) while **r** is the rank of matrix A.

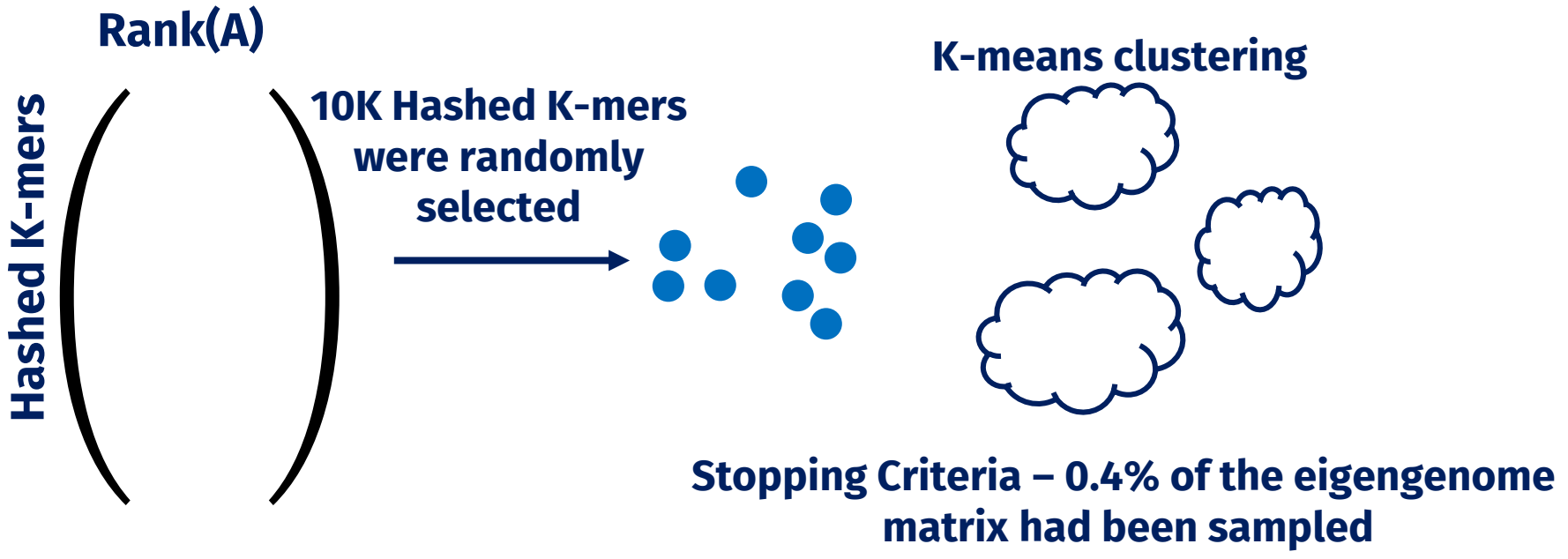
Dimensionality Reduction : Eigengenomes ?

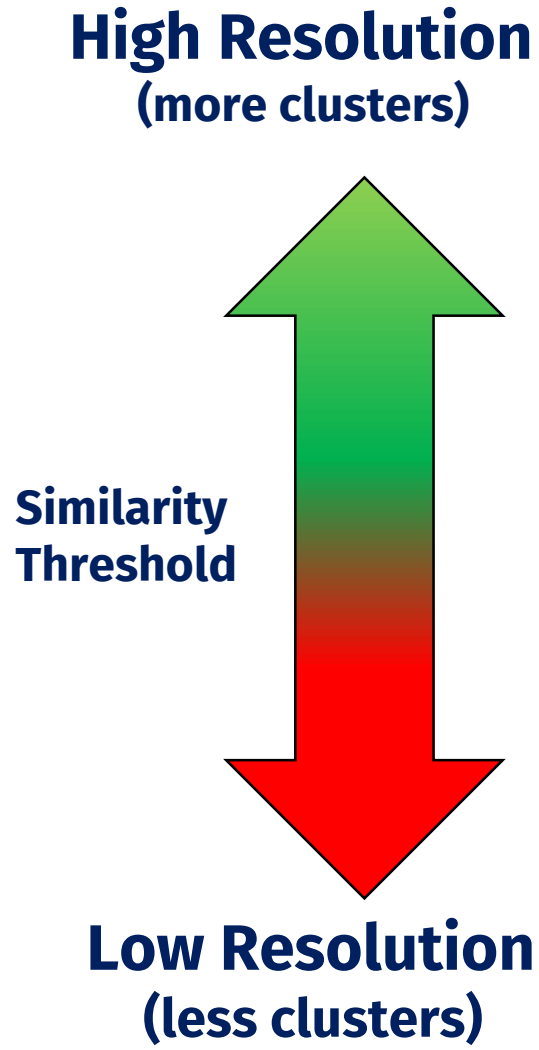
$$\mathbf{A} = \mathbf{U} \times \mathbf{\Sigma} \times \mathbf{V}^{\mathbf{T}}$$



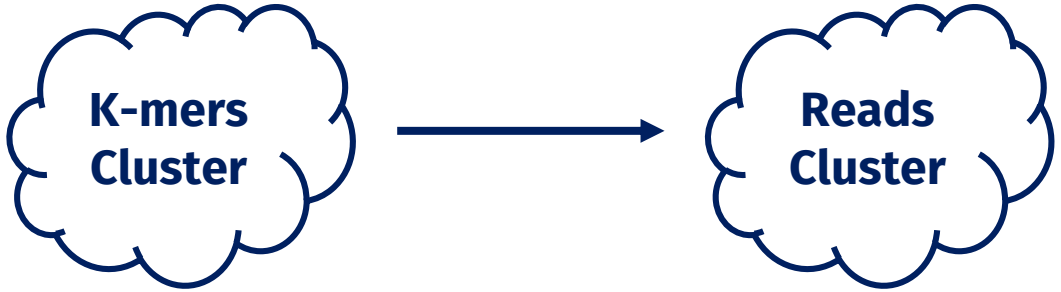
Eigengenome - Analog to the Principle Component of the sequence space. (the columns of V collectively as the set of eigengenomes).

Reads' Clustering: K-mers clustering





Reads' Clustering: From k-mers cluster to reads

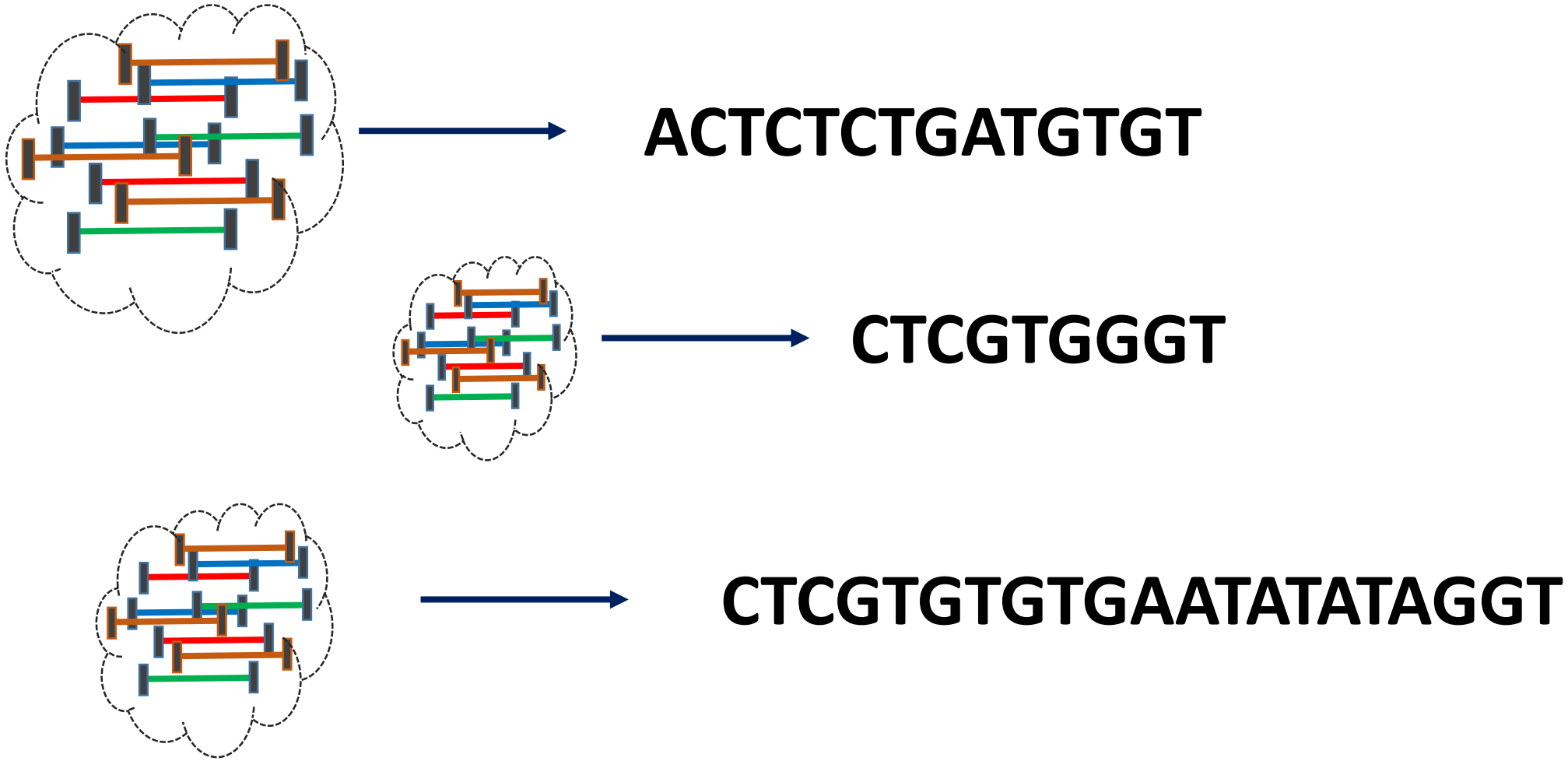


log-likelihood:

- The size of each of the k-mer clusters
- The intersection of the k-mers in the read with each of the clusters
- The global weight of each of the intersecting k-mers

Unique read in each cluster

Assembly: From reads' clustering to assembly



Results: Sanity check - I

<u>Question</u>: Can <u>LSA</u> partition mock reads from <u>single genome</u> mixed with other genomes?

<u>Test:</u> 30 human gut samples + *'Salmonella Bongori'* mock reads

Human Microbiome Project

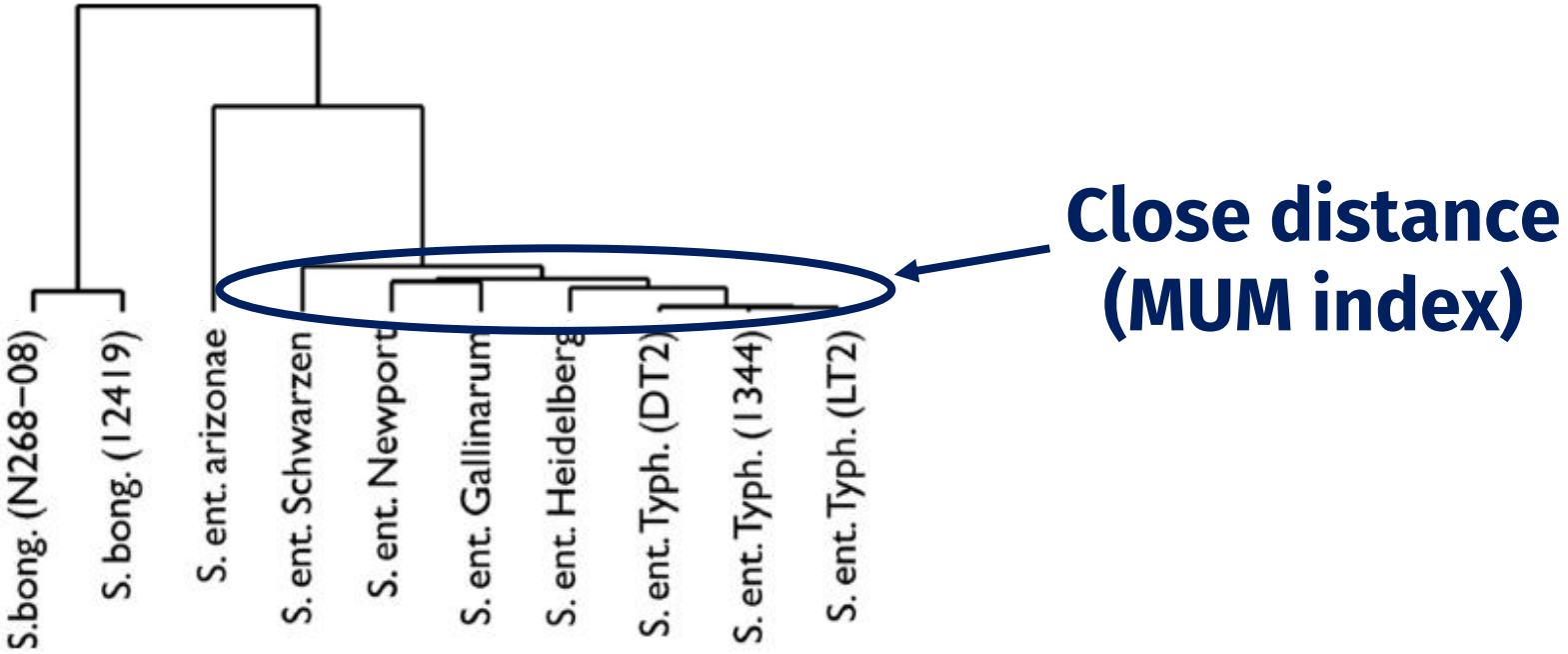
<u>Result:</u>

- LSA produced **451 partitions** using **25 Gb**
- **Out of a total of ~20 million** spiked '*S. bongori*' reads, **more than 99% ended up in a single partition**.

Results: Sanity check - II

Question: Can LSA Separate reads from closely related strains into different partitions?

Test: 50 human gut samples + 2 strains of '**S. Bongori**' mock reads + 8 strains of '**S. Enterica**' mock reads



Results: Sanity check - II

Accuracy - percentage of partition's assembly covered by reads simulated from a given reference genome (RG).

RG Reads' coverage = 16

Assembly length = 25

Accuracy = 64%

Completeness- percentage of a given reference genome covered by each partition .

Partition reads' coverage = 22

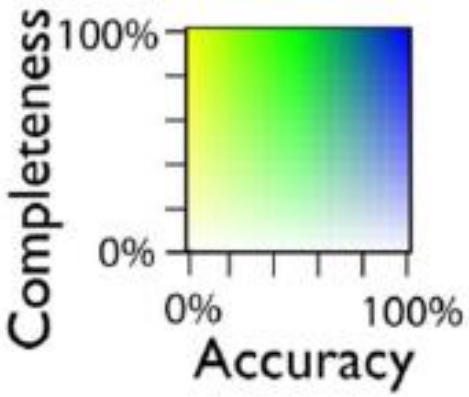
RG length = 32

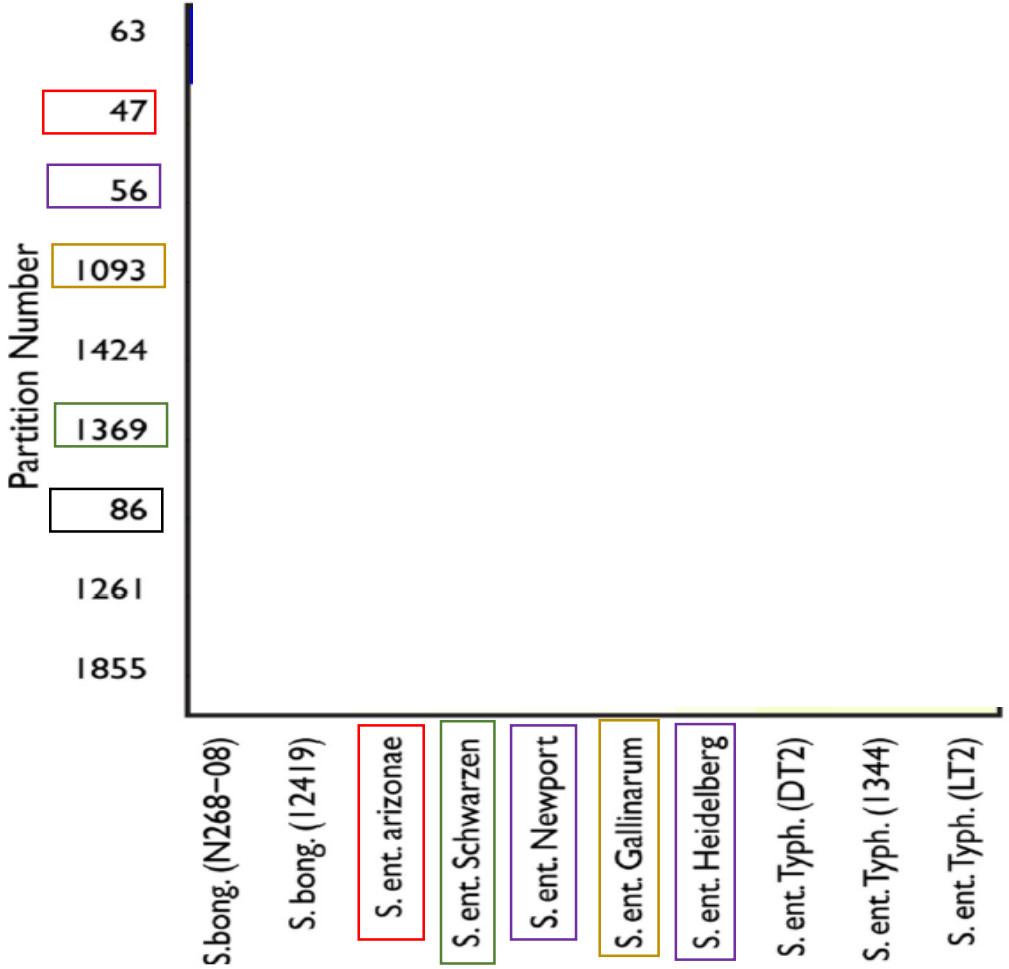
Completeness = 69%

Results: Sanity check - II

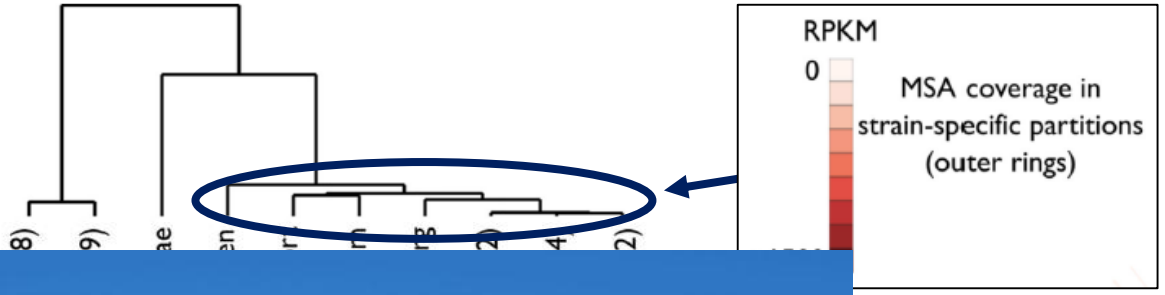
- LSA produced **2,543 partitions, 'S. bongori'** partition:
  - **Accuracy = 99.52% (of the reads are from 'S. bongori')**
  - **Completness = 95.79% (**of all the '**S. bongori' reads)**

near perfect separation!





Results: Sanity check - II



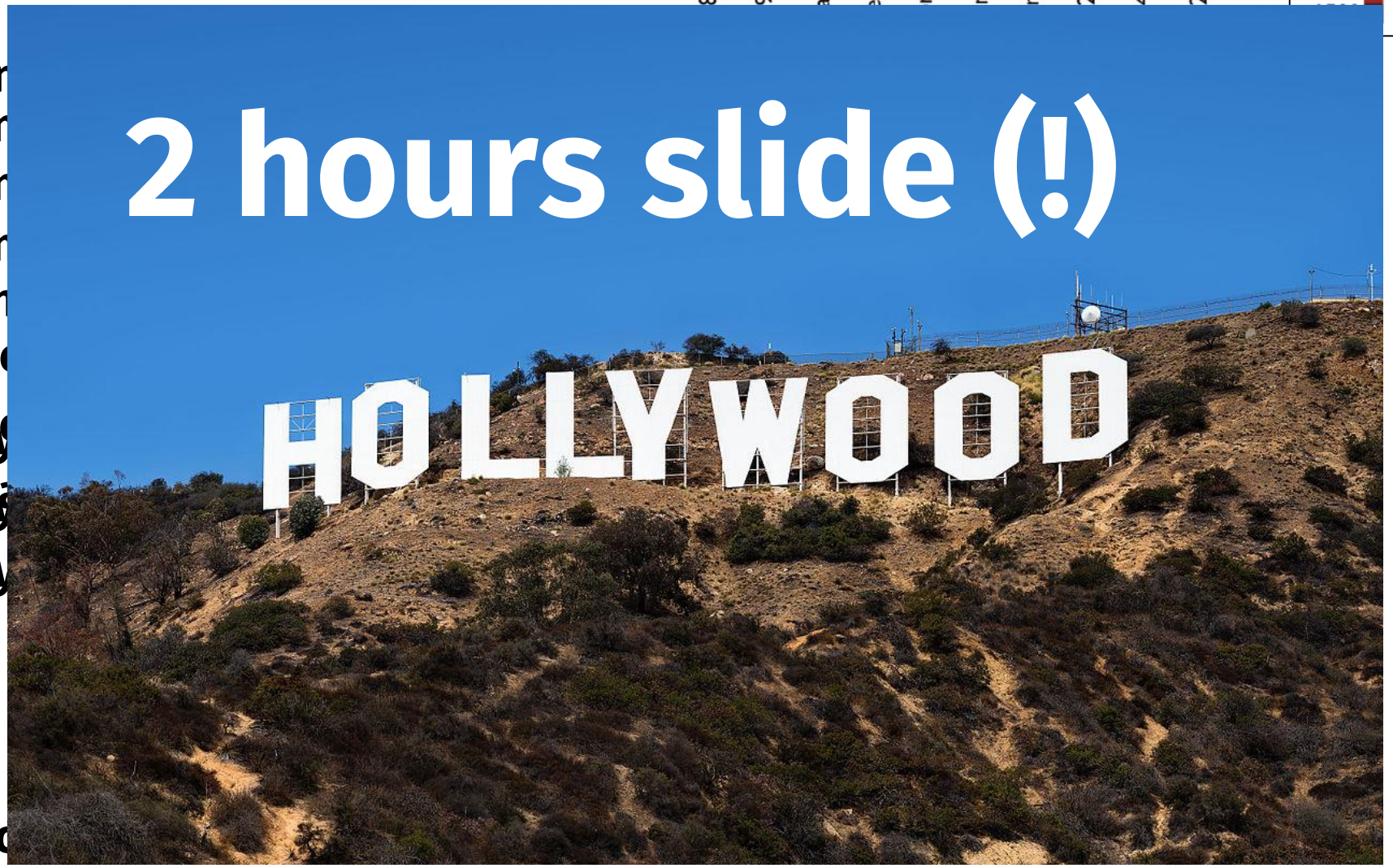
Outer Rings

Partitio
Partitio
Partitio
Partitio
Partitio
MSA
Partitio
S. ent. H
S. ent. N
S. ent. Ty
S. ent. G
S. ent. Ty
S. ent. Ty

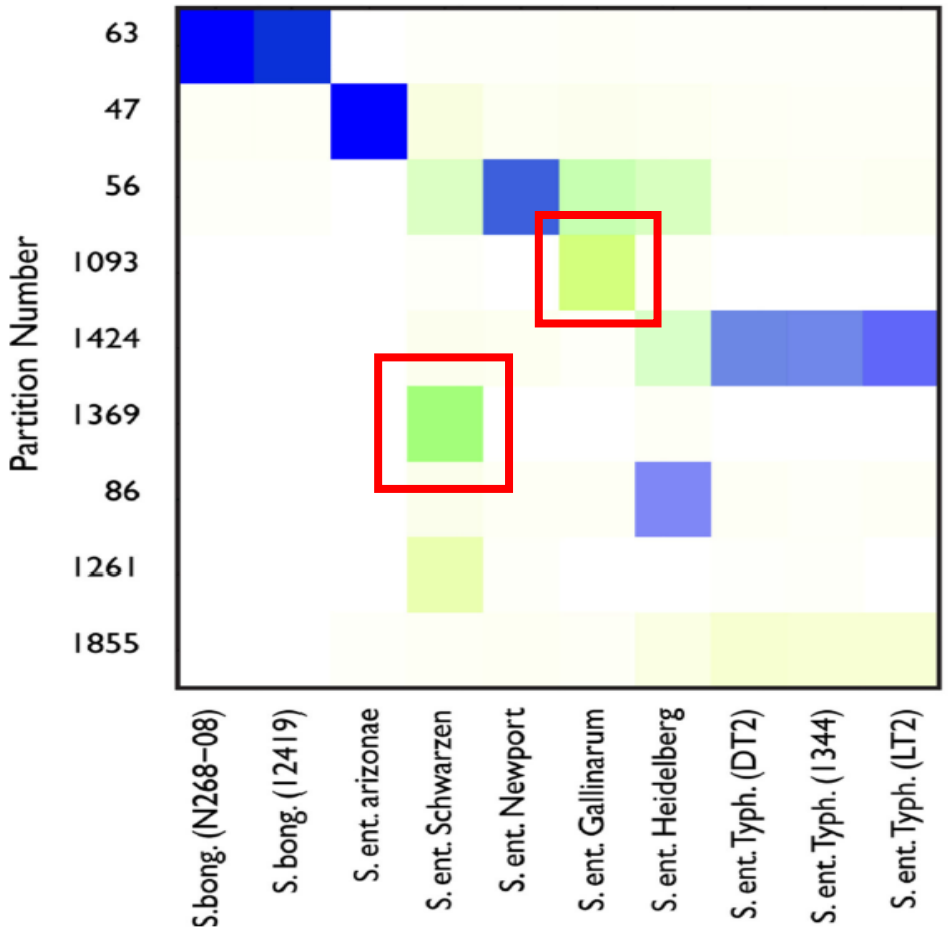
Inner Rings

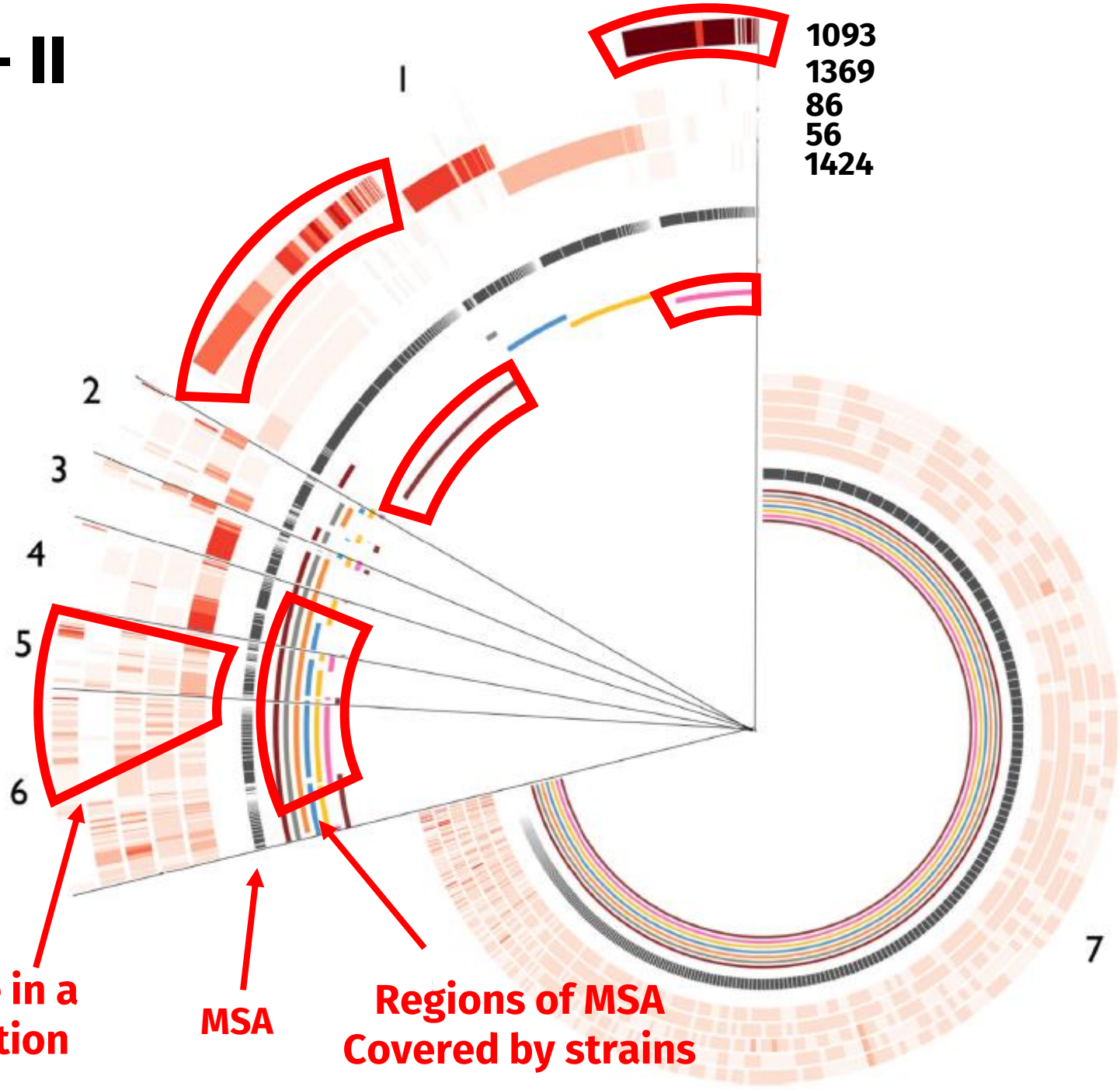
S. ent. Sc

2 hours slide (!)



Results: Sanity check - II

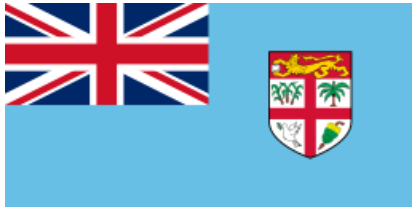




Results: Low Abundance Species

<u>Question</u>: Can assembled partitions can be aligned to reference genomes?

<u>Test:</u> 176 human stool samples from FijiCoMP (4 Tb)



Completeness – assessed by by AMPHORA set (31 house keeping genes)
N50 - the minimum contig length needed to cover 50% of the assembled genome

<u>Result:</u>

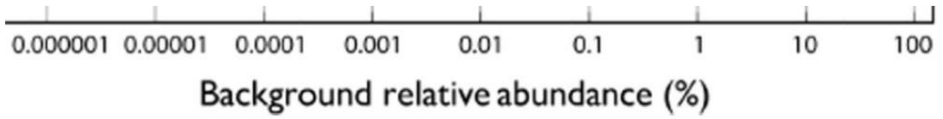
- LSA produced **4,306 partitions**
- **Considering only contigs greater in length than the N50 of a given partition == 344 partitions which are relative specific (>50% of total alignment)**

Results: Low Abundance Species

MetaPhyler – a taxonomic classifier uses phylogentic marker genes.

Result:

- **Out of 344 partitions, 93 contains all 31 AMPHORA genes.**
- **16s sequencing detected >70 bacterial families with low abundance ($4 \times 10^{-6\%}$)**

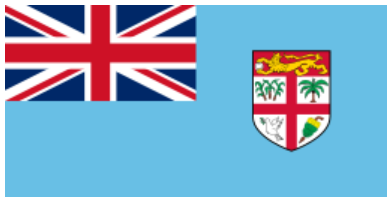


Most Enriched Families

Abundant	Uncommon
a. Enterobacteriaceae (99.05%) b. Prevotellaceae (97.44%) c. Lachnospiraceae (95.58%) d. Ruminococcaceae (92.32%)	e. Eubacteriaceae (96.39%) f. Streptococcaceae (91.9%) g. Rikenellaceae (66.59%) h. Coriobacteriaceae (65.85%)
Rare	**Very Rare**
i. Lactobacillaceae (30.11%) j. Flavobacteriaceae (19.67%) k. Sphingobacteriaceae (10.96%) l. Comamonadaceae (10.72%)	m. Spirochaetaceae (17.4%) n. Peptococcaceae (11.93%) o. Burkholderiaceae (10.61%) p. Mycoplasmataceae (8.27%)

Results: Memory Consumption

Data Set	#Samples	Type	Size (Gb)
FijiCoMP	176	Human Stool	4,000
HUGE	32	Human Stool	300
Sharon et .al	18	premature infant gut	20



Results: Memory

streaming SVD - **Gensin** packge (python) operates in fixed memory

Pipeline Step	Number of tasks	Time per task (hrs)	RAM / task	(1) 176 samples	(2) 32 samples	(3) 18 samples	(1) 4Tb	(2) 300Gb	(3) 20Gb
Hyperplane Hashing	number of reads / 1m	1.2	3Gb	17516	3184	561	3Gb	3Gb	3Gb
Hashed K-mer Counting	number of samples	2	4Gb	176	32	18	4Gb	4Gb	2Gb
Global K-mer Weighting	1	1.7	25Gb	1	1	1	25Gb	25Gb	12Gb
K-mer Abundance Matrix	number of samples	0.45	32Gb	176	32	18	32Gb	32Gb	16Gb
Streaming SVD	1	*	4Gb	1	1	1	4Gb	4Gb	2Gb
K-mer ClusterIndex	1	24	1Gb	1	1	1	1Gb	1Gb	1Gb
K-mer ClusterMap	k-mer hash size / 1m	2.2	1Gb	2148	2148	1074	1Gb	1Gb	1Gb
K-mer ClusterReduce	1	1.1	50Gb	1	1	1	**50Gb**	**50Gb**	**25Gb**
Read Partitioning	number of reads / 1m	6.2	1Gb	17516	3148	561	1Gb	1Gb	1Gb
	General computational requirements			Number of tasks per collection			Peak memory use per collection		

Pros & Cons

- **Pros:**
  - Open Source

- **Cons:**
  - Tables / Figures mismatch
  - No comparison to other methods
  - One data set
  - Fancy algorithm, compare to random?
- **Pros:**
  - Fixed Memory
  - Integration of concepts

Discussion

- Validation of novel & specific new computational methods
- Article name Vs. its actual value?

Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning (LSA)

Thanks!