ConStrains identifies microbial strains in metagenomic datasets

Chengwei Luo, Rob Knight, Heli Siljander, Mikael Knip, Ramnik J Xavier & Dirk Gevers

Danielle Miller

10.04.2019

ביה"ס למדעי המחשב ע"ש בלווטניק
הפקולטה למדעים מדויקים
ע"ש ריימונד ובברלי סאקלר
אוניברסיטת תל אביב

Conspecific strains

KINGDOM

PHYLUM

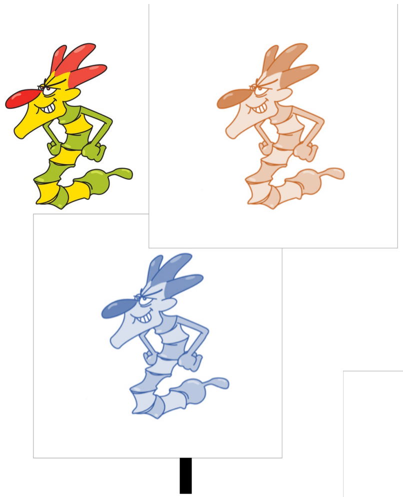
CLASS

ORDER

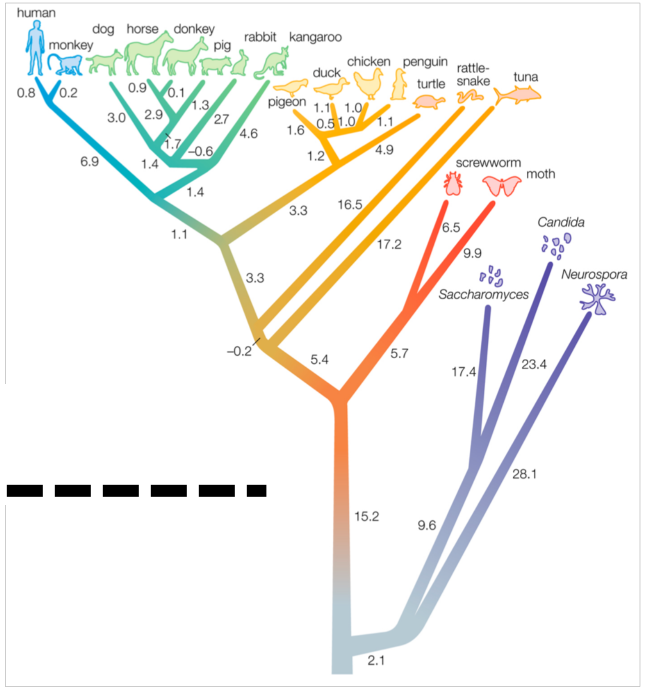
FAMILY

GENUS

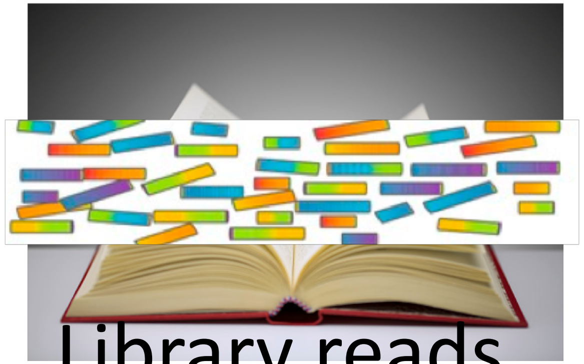
SPECIES



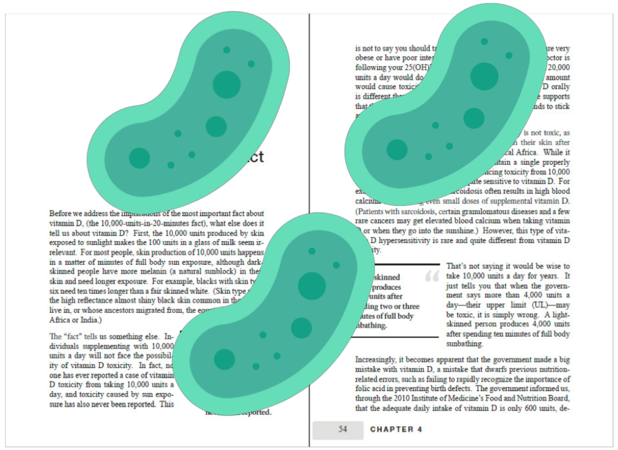
STRAIN



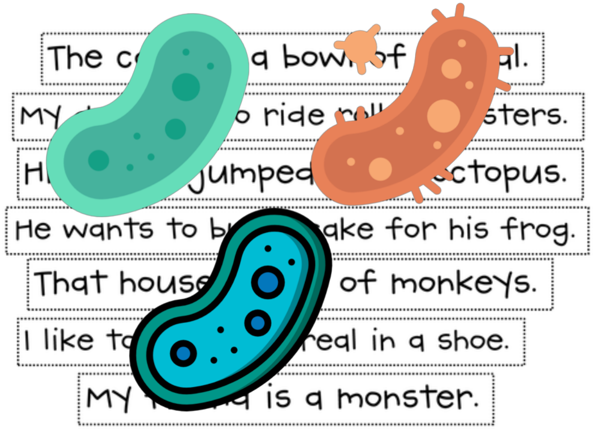
Motivation



Library reads



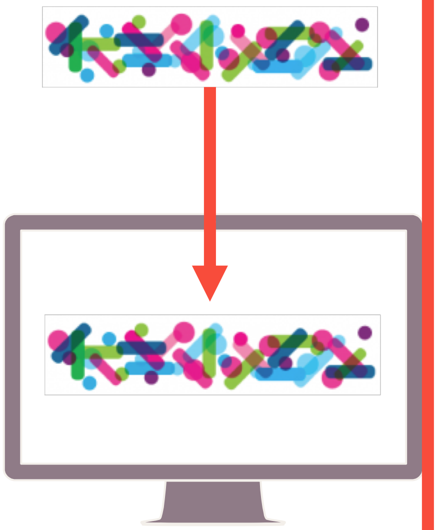
Specie



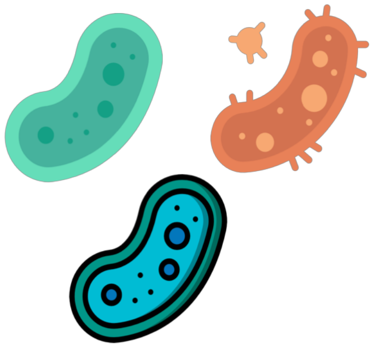
Strains

Outline

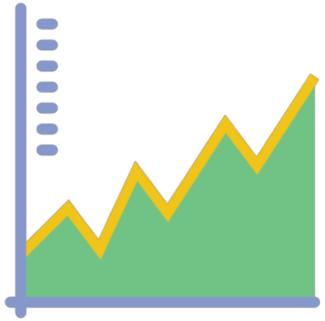
Data processing



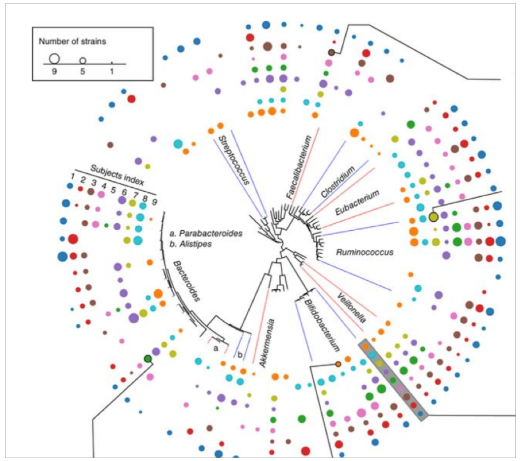
Strain identification



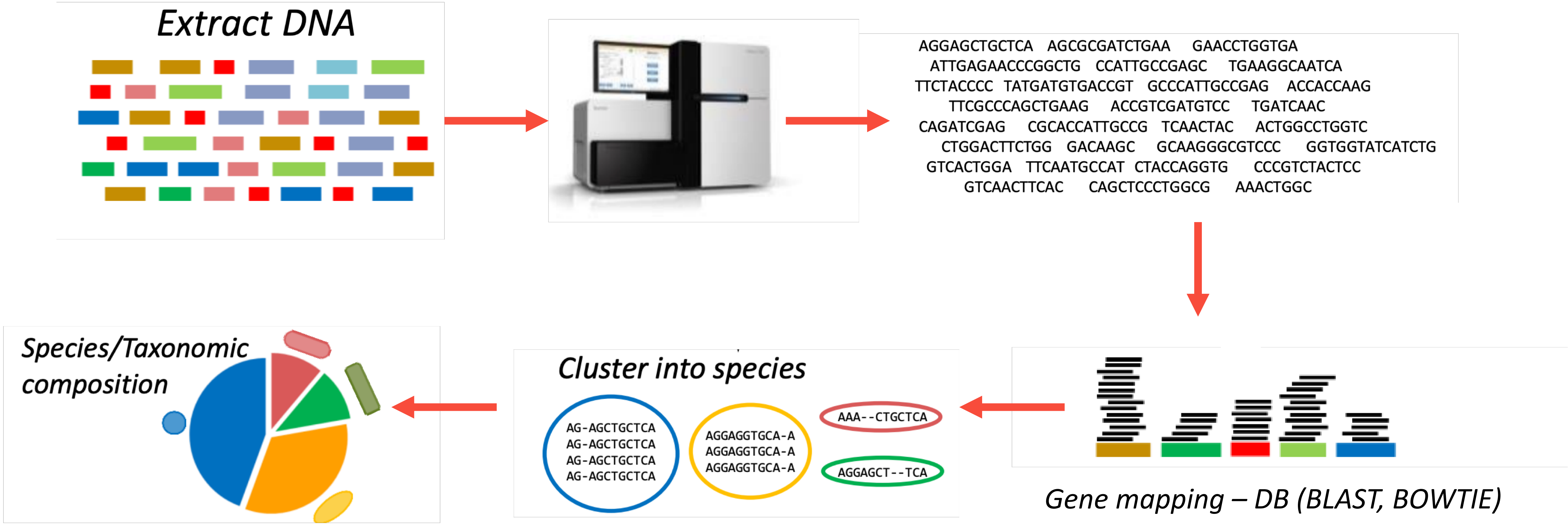
Evaluation



Real data



Shotgun sequencing



Extract DNA

```
AGGAGCTGCTCA   AGCGCGATCTGAA    GAACCTGGTGA
  ATTGAGAACCCGGCTG   CCATTGCCGAGC    TGAAGGCAATCA
TTCTACCCC  TATGATGTGACCGT   GCCCATTGCCGAG    ACCACCAAG
      TTCGCCCAGCTGAAG     ACCGTCGATGTCC     TGATCAAC
CAGATCGAG     CGCACCATTGCCG   TCAACTAC     ACTGGCCTGGTC
    CTGGACTTCTGG   GACAAGC     GCAAGGGCGTCCC     GGTGGTATCATCTG
  GTCACTGGA   TTCAATGCCAT  CTACCAGGTG     CCCGTCTACTCC
       GTCAACTTCAC     CAGCTCCCTGGCG     AAACTGGC
```

Gene mapping – DB (BLAST, BOWTIE)

Cluster into species

```
AG-AGCTGCTCA
AG-AGCTGCTCA
AG-AGCTGCTCA
AG-AGCTGCTCA
```

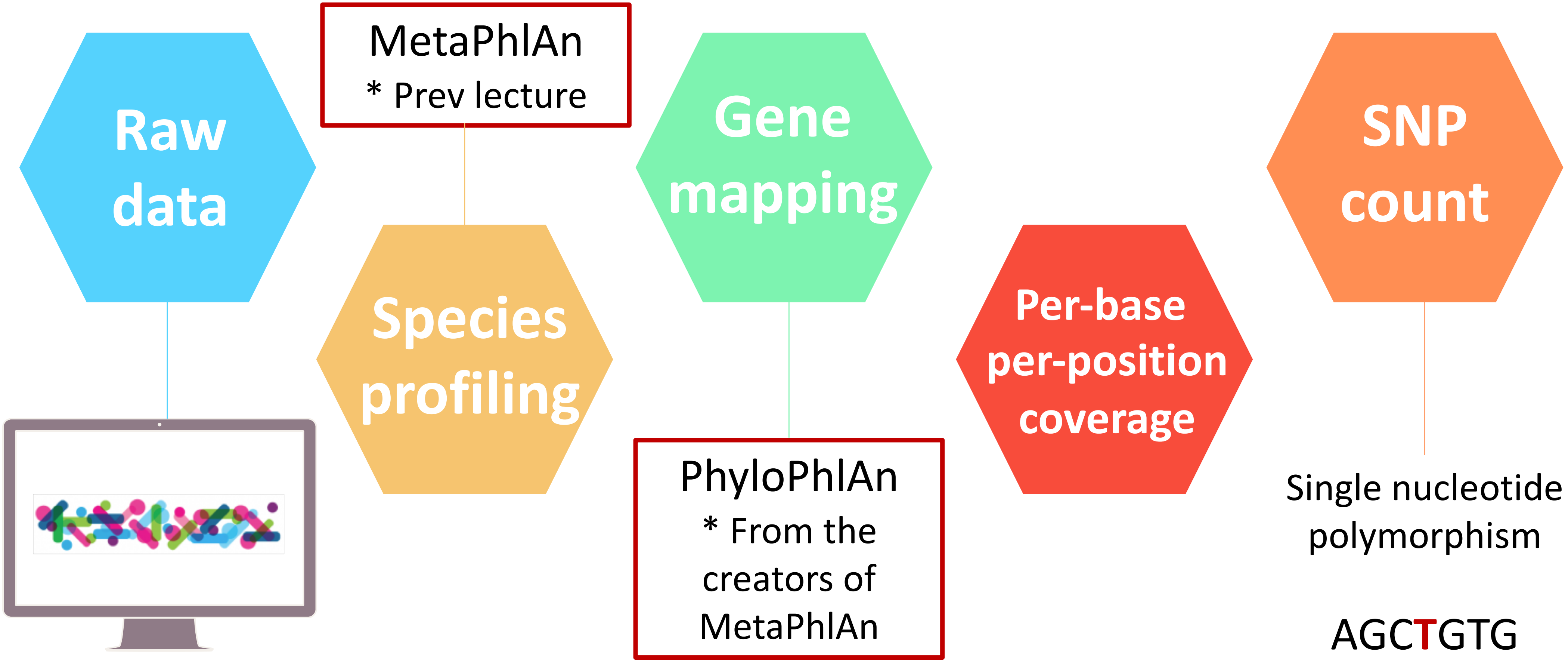
```
AGGAGGTGCA-A
AGGAGGTGCA-A
AGGAGGTGCA-A
```

```
AAA--CTGCTCA
```

```
AGGAGCT--TCA
```

Species/Taxonomic composition

Data processing



Raw data

MetaPhlAn
* Prev lecture

Species profiling

Gene mapping

PhyloPhlAn
* From the creators of MetaPhlAn

Per-base per-position coverage

SNP count

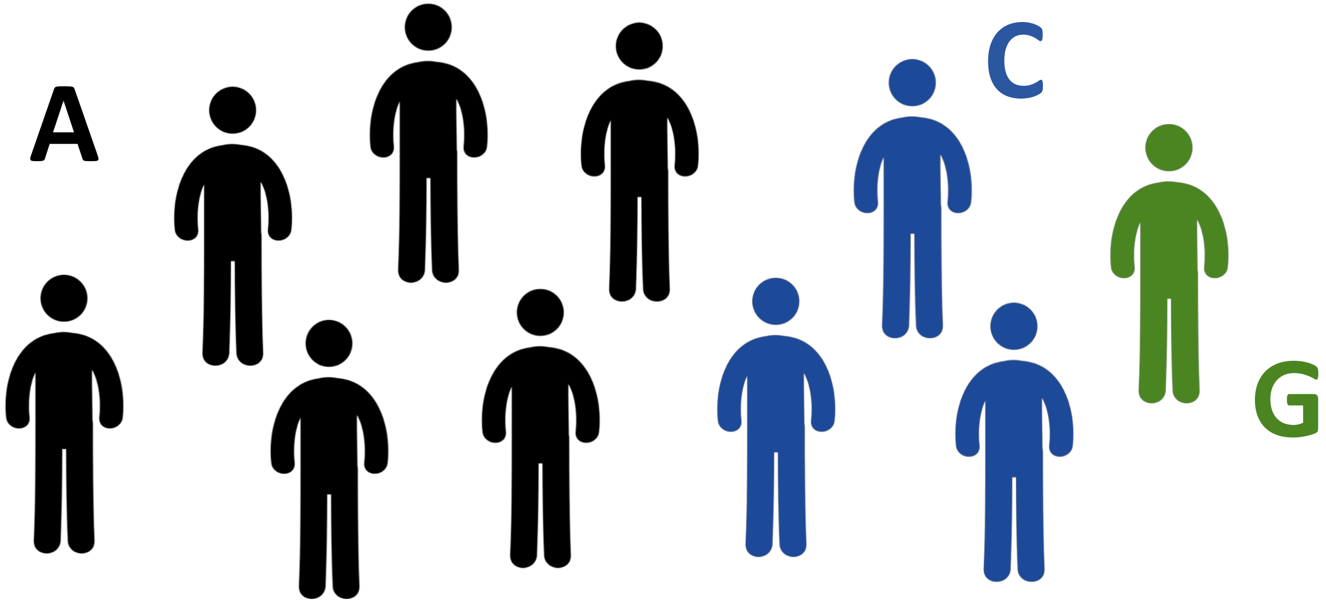
Single nucleotide polymorphism

AGCTGTG
AGCCGTG

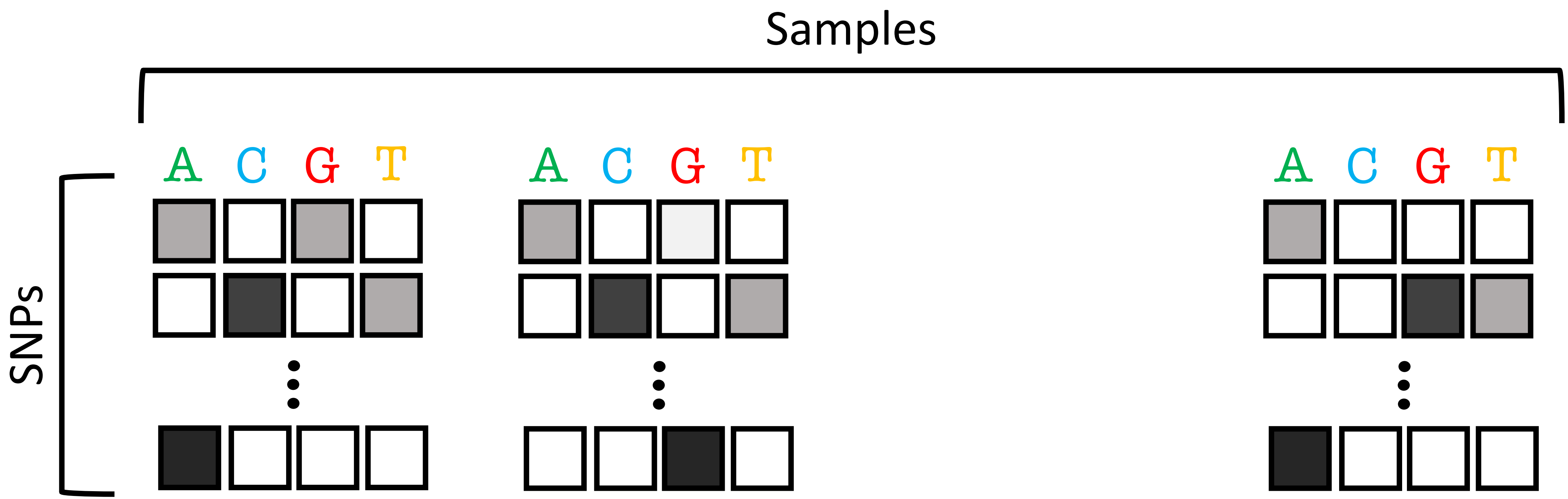
SNPs

SNP count

- A type of genetic variation in a population
- Each SNP represents a difference in a single DNA building block

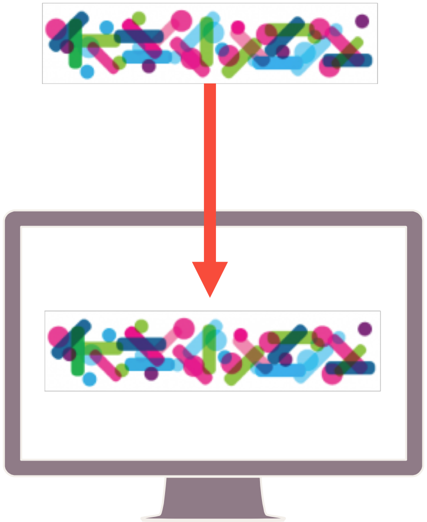


Processing output

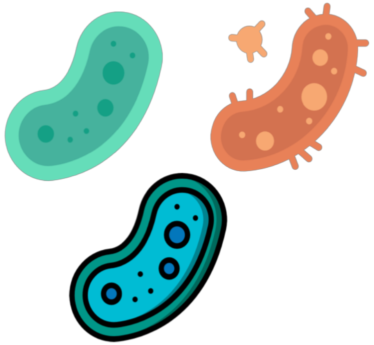


Outline

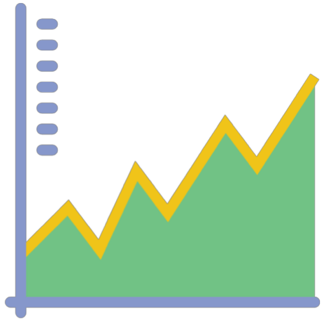
Data process



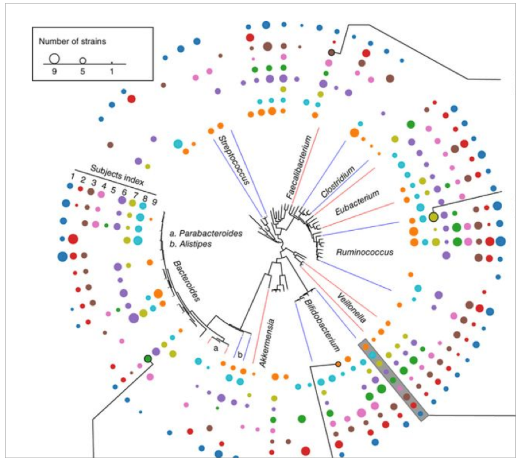
Strain identification



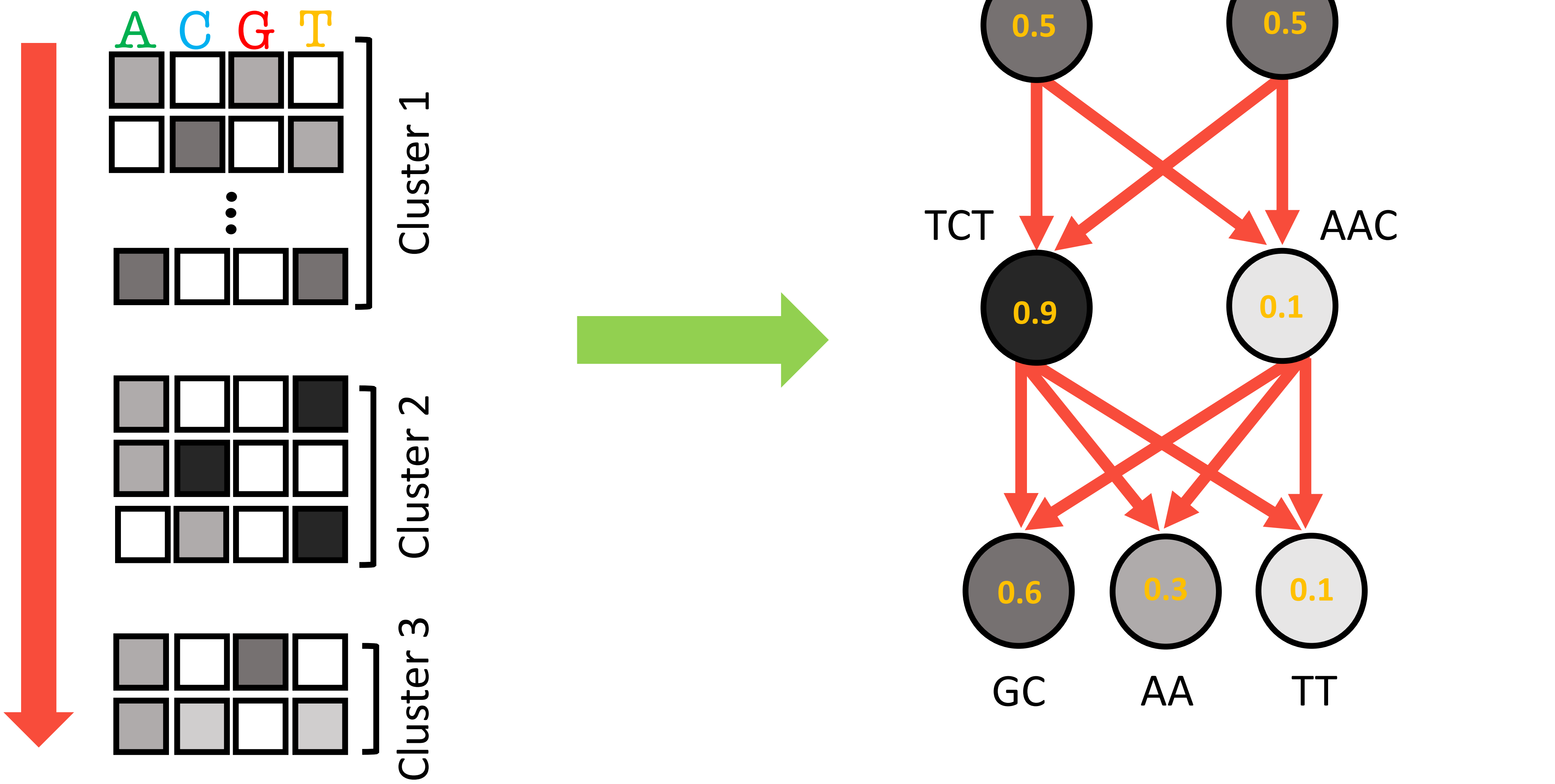
Evaluation



Real data

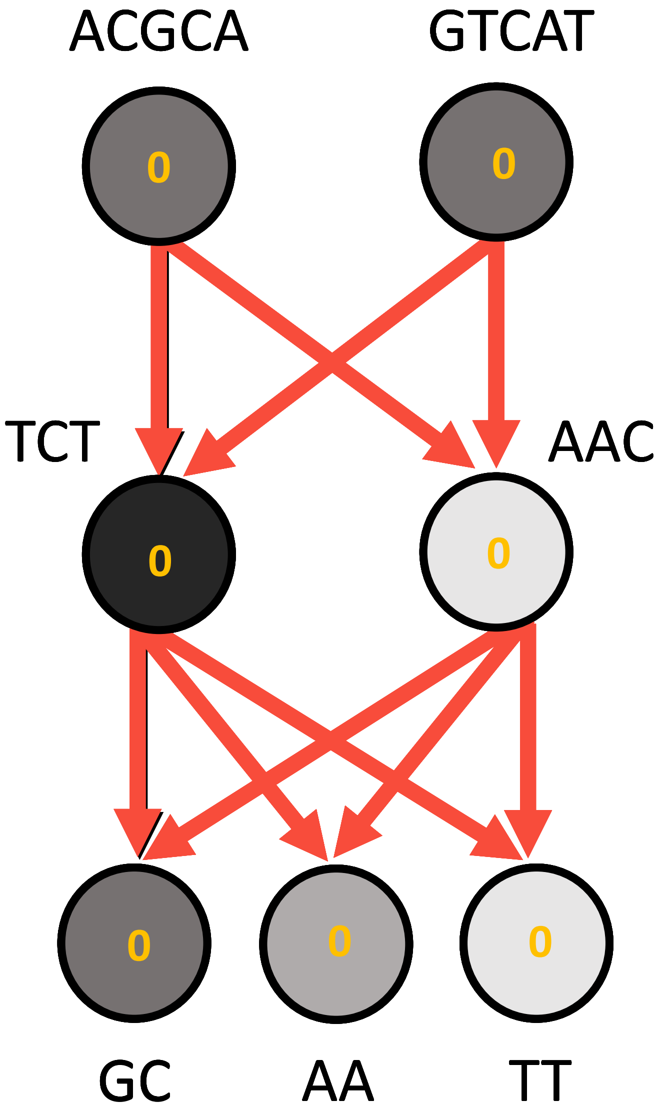


SNP-flow

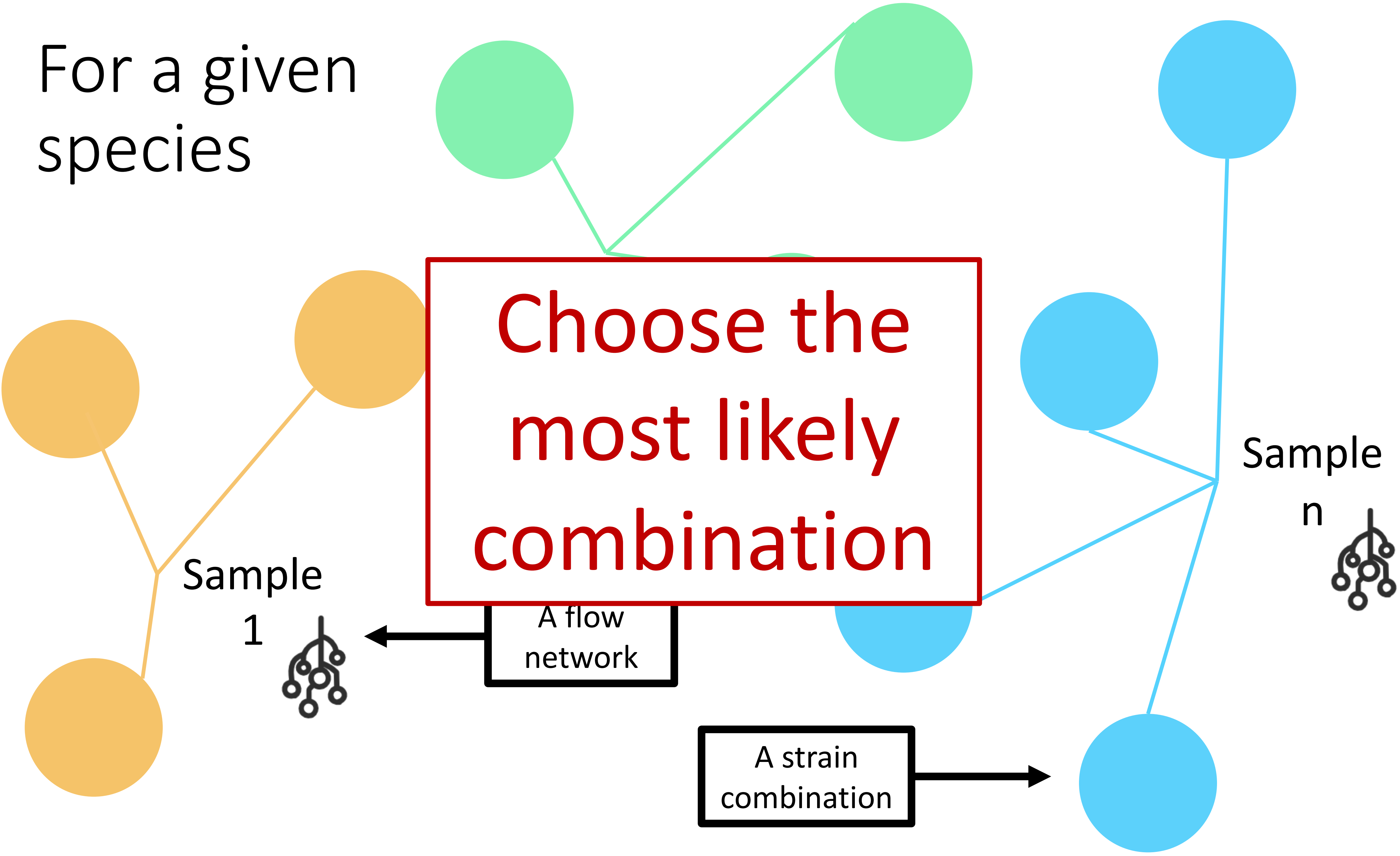

A
C
G
T
Cluster 1
Cluster 2
Cluster 3
ACGCA
GTCAT
0.5
0.5
TCT
AAC
0.9
0.1
0.6
0.3
0.1
GC
AA
TT

Strain identification

Strain combination:		Relative abundance:
Str. 1	ACGCA TCT GC	**0.5**
Str. 2	GTCAT TCT GC	**0.1**
Str. 3	GTCAT TCT AA	**0.3**
Str. 4	GTCAT AAC TT	**0.1**



For a given species



Choose the most likely combination

Sample 1

A flow network

Sample n

A strain combination

Inferring strain compositions

We have

- Strain combination per-specie per-sample

Str. 1	AACGGTCG	0.6
Str. 2	AATCTGAC	0.4

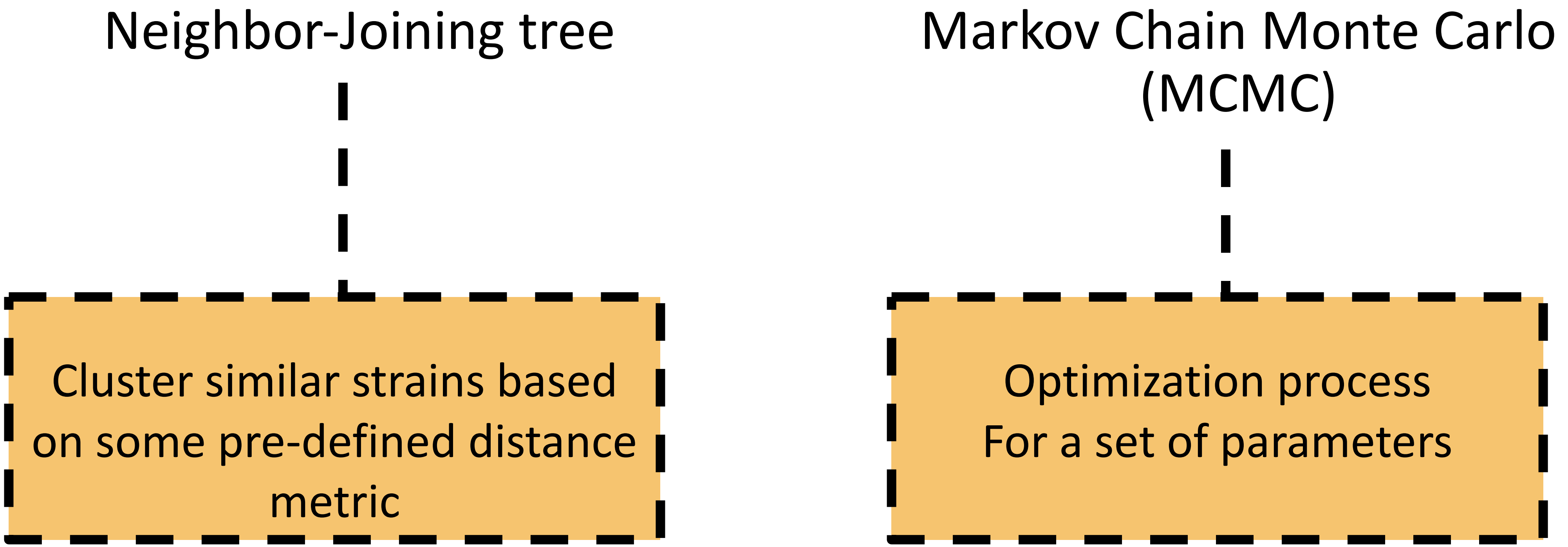
We need

- Optimized strains relative abundance cohort-wise

Inferring strain compositions

Cluster similar strains

Infer strain composition



Neighbor-Joining tree

Cluster similar strains based on some pre-defined distance metric

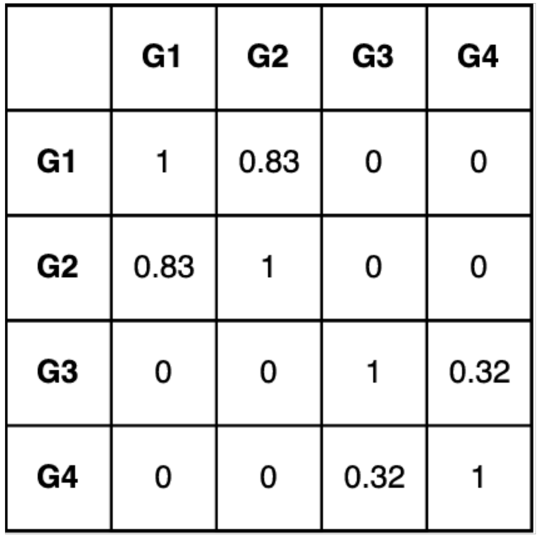
Markov Chain Monte Carlo (MCMC)

Optimization process
For a set of parameters

A glance to the NJ algorithm

- Find clusters C_1, C_2 that minimize a function $f(C_1, C_2)$
- Join the two clusters C_1, C_2 into a new cluster C^*
- Add a node to the tree corresponding to C^*
- Assign distances to the new branches

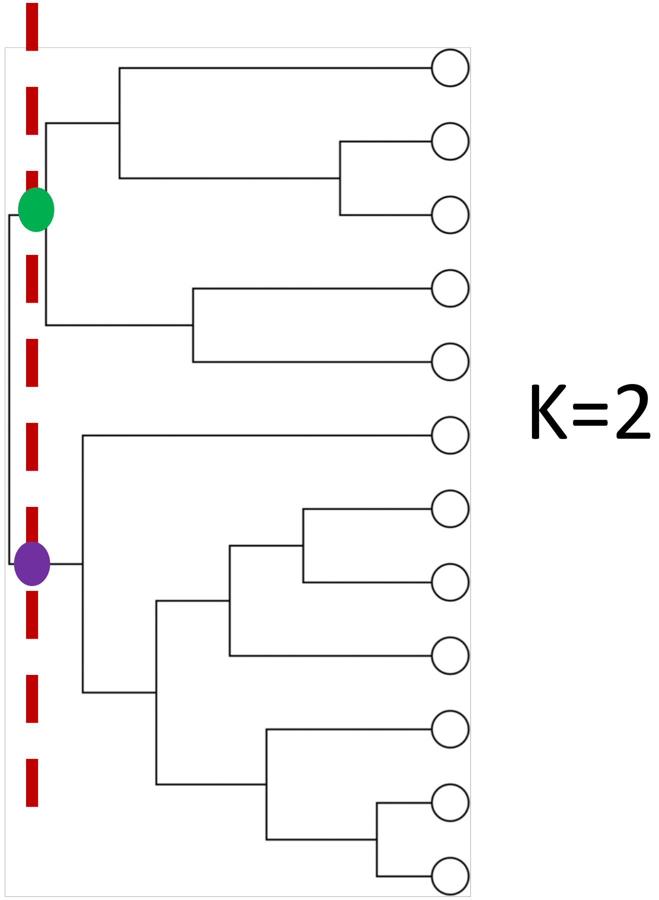
Similarity matrix based on sequence percentage identity

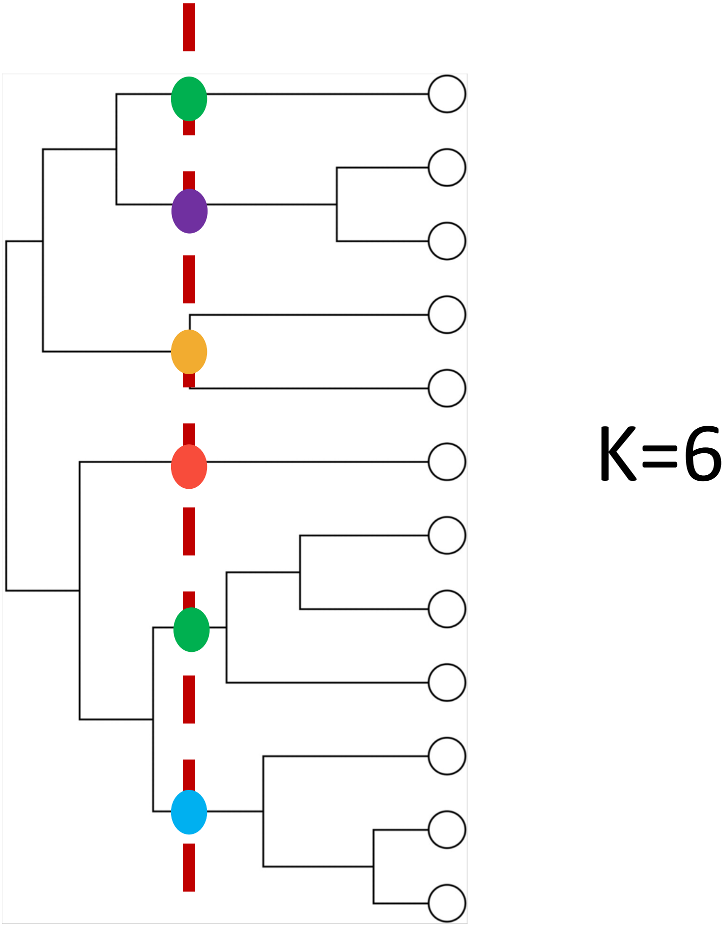


	G1	G2	G3	G4
G1	1	0.83	0	0
G2	0.83	1	0	0
G3	0	0	1	0.32
G4	0	0	0.32	1

Inferring strain composition

Construct NJ tree from all samples





MCMC

A class of algorithms for sampling from a probability distribution

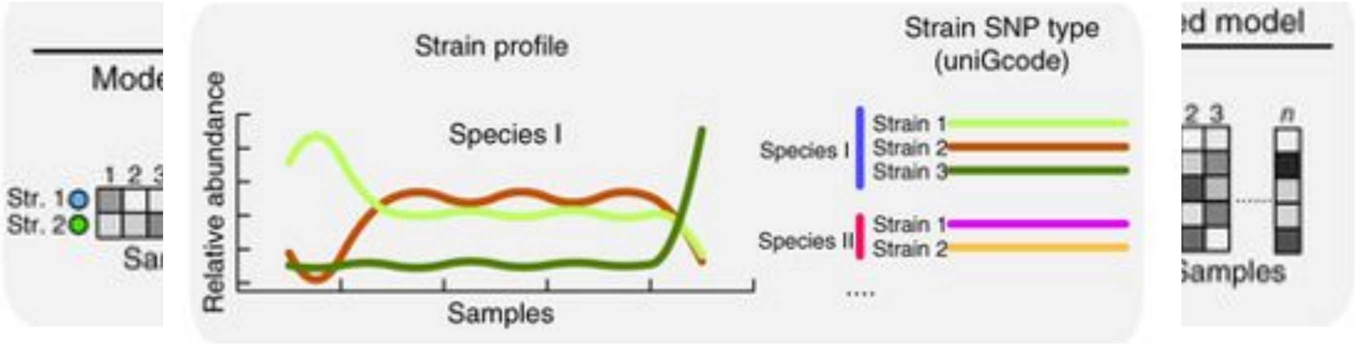
Initialize θ_0 (set of parameters)

Given a current state θ_t:

- Propose a new state θ_{t+1}
- Calculate the probability p - transition to the state $t+1$
- Draw a random number u from $U[0,1]$ - accept new step if $u \leq p$
- Iterate until convergence \ pre defined number of iterations

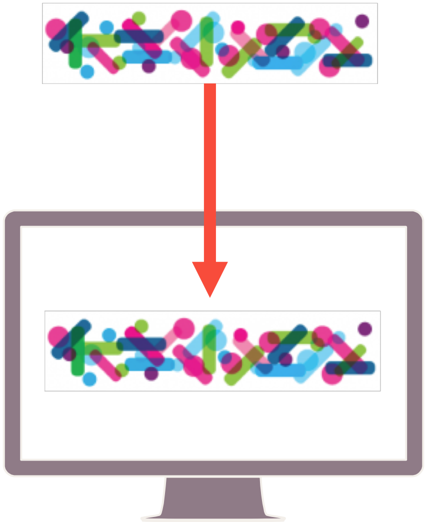
MCMC for composition detection

- For each model (k) find a composition $\alpha^* = (\alpha_1^*, \alpha_2^*, \ldots, \alpha_k^*)$ using the MH MCMC algorithm.
- Minimize expected SNPs frequencies and observed SNPs frequencies
- Model selection using a corrected AICc

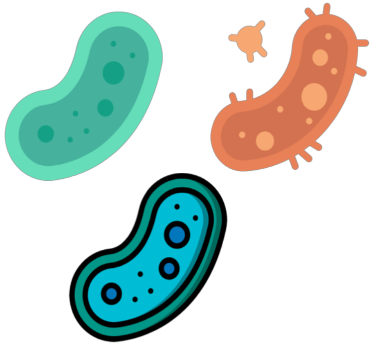


Outline

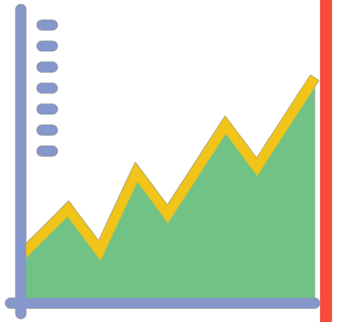
Data process



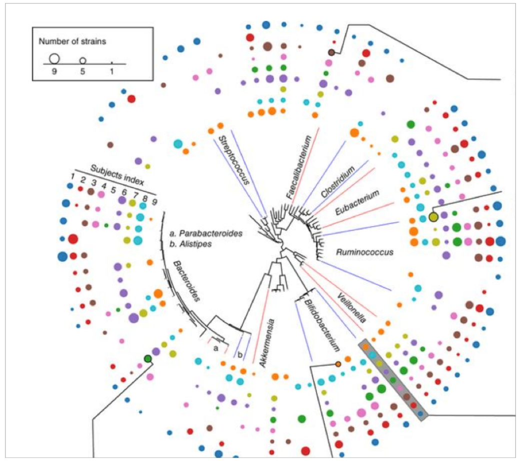
Strain identification



Evaluation

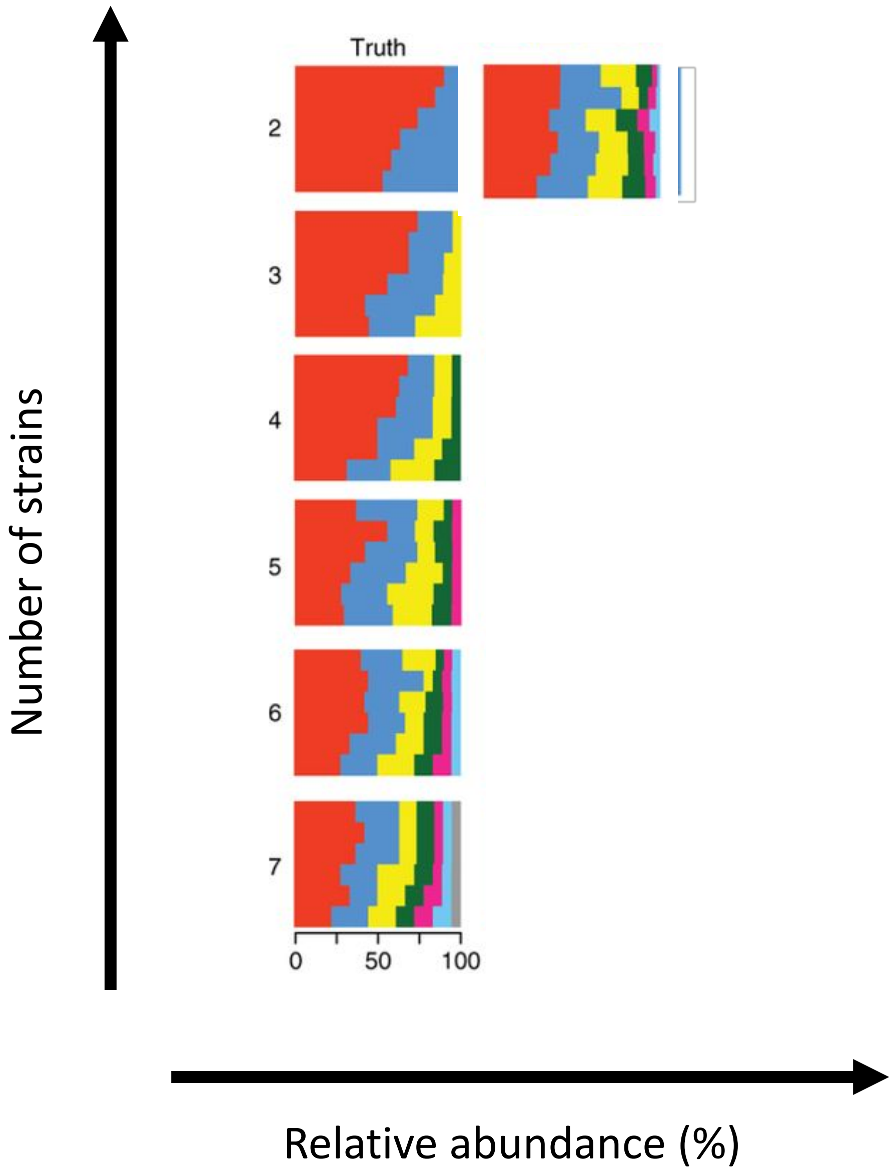


Real data



Evaluation

- 36 simulated datasets with ranging k-strain combinations



What about the composition?
Use Jensen-Shannon divergence!

JSD

P – predicted composition; Q – true composition

$\boldsymbol{P = Q}$

0

$$M = \frac{1}{2}(P + Q)$$

$$JSD = \frac{1}{2}D(P||M) + \frac{1}{2}D(Q||M)$$

$\boldsymbol{P \neq Q}$

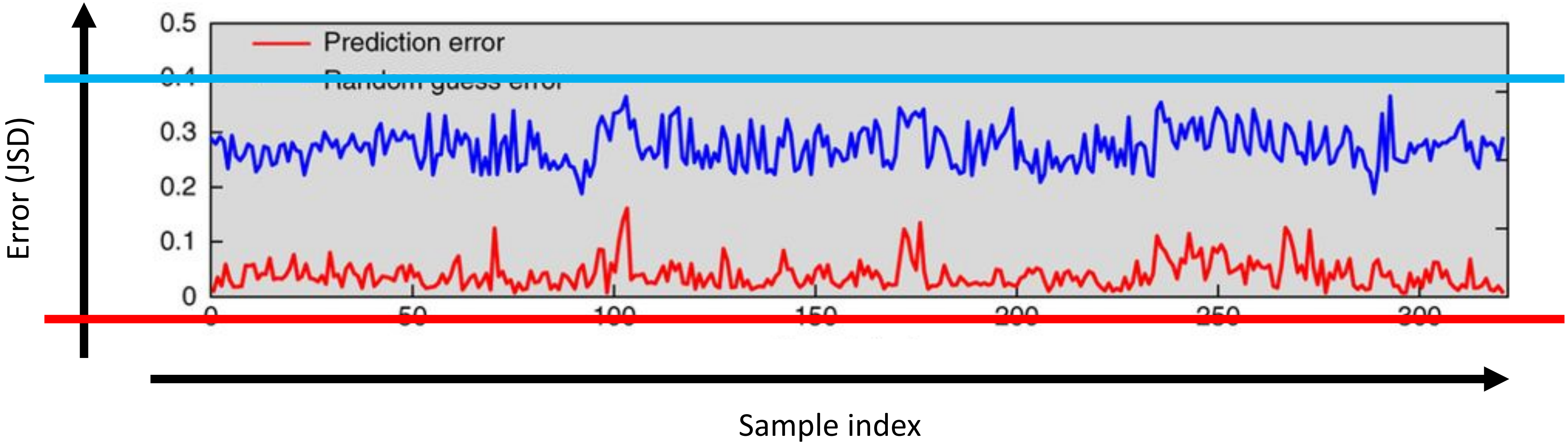
1

Were $D(X||Y)$ is the KL divergence $\sum X(i) \log \frac{X(i)}{Y(i)}$

How Y describes X

Evaluation

- Simulated shotgun sequencing data
  - 91 species across 322 *in silico* samples
- Jensen-Shannon divergence – the lower the better



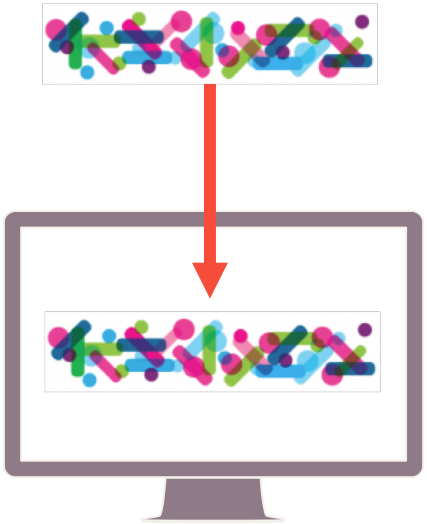
Outline

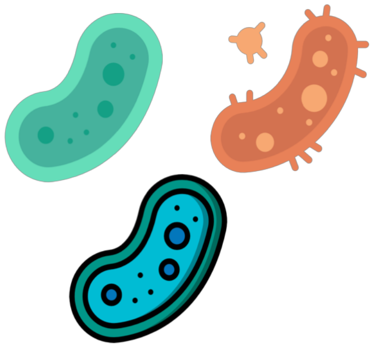
Data process

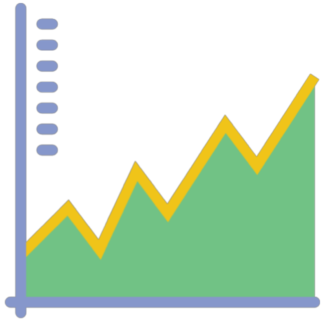
Strain identification

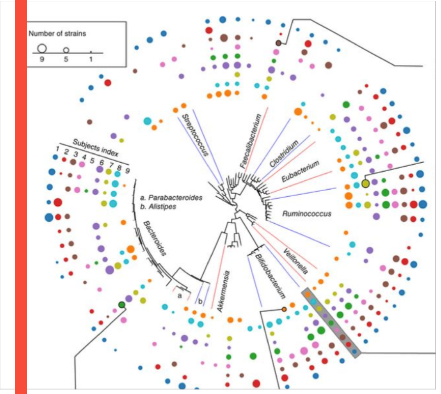
Evaluation

Real data





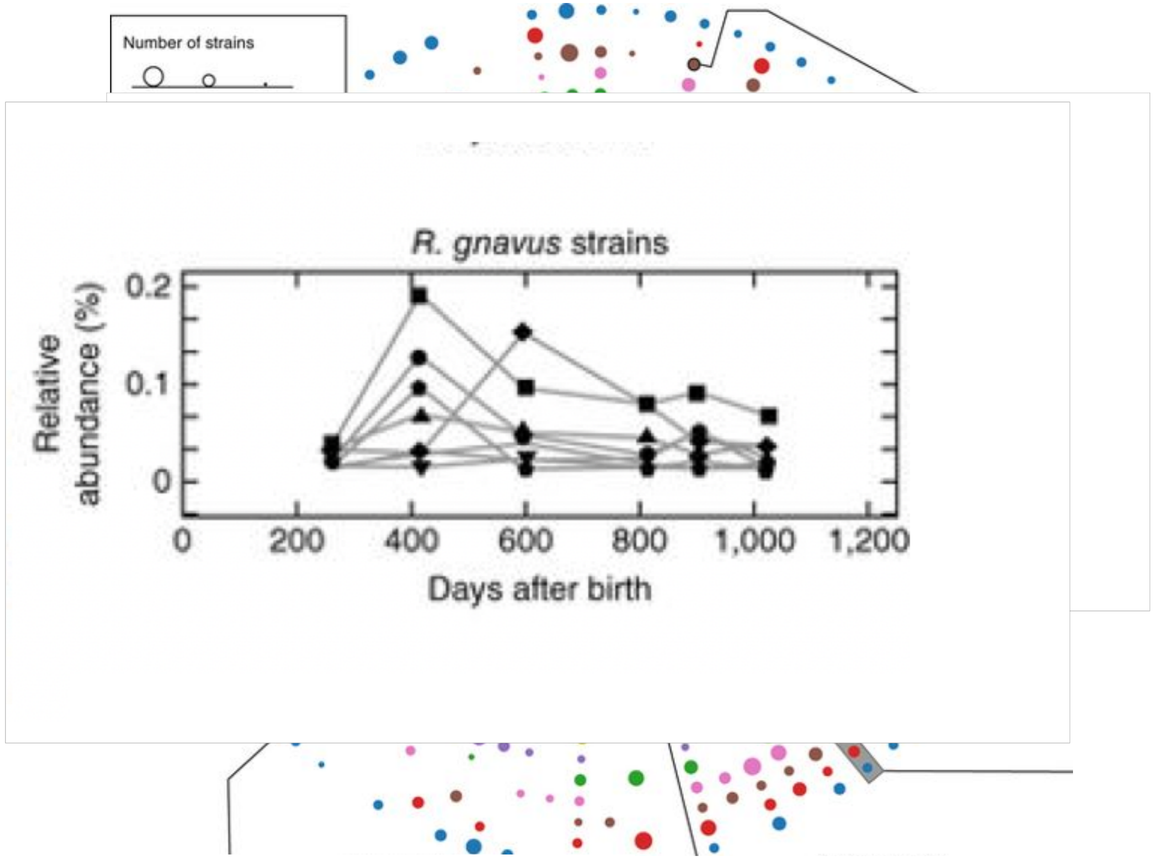




Uncovering strain dynamics in infant gut development

- 54 samples from 9 different subjects

- Samples were taken from the first 3 years of the subjects life



Summary

- A greedy algorithm for inferring strain composition and type using SNPs
- Strain reference Independent
- Minimal resource requirement
- Open source (Yay!!)

Discussion points

- The first step of the algorithm is species mapping, however the number of known bacterial species is miniscule. Is this good enough for healthcare based applications?

- Simplicity vs. complexity – what do we prefer?