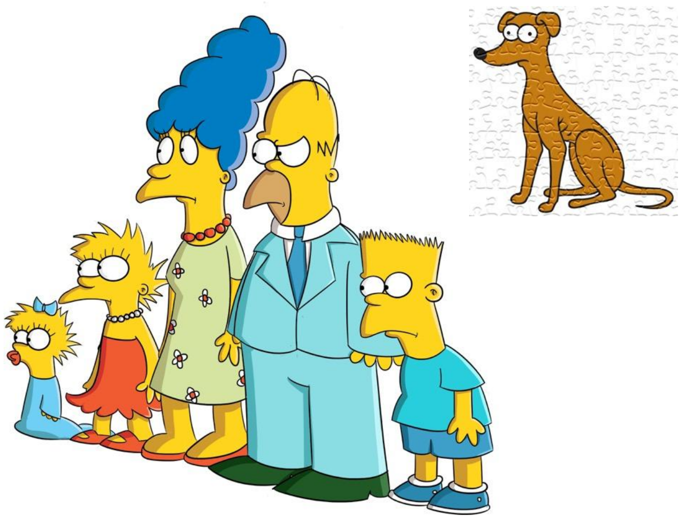
Core and accessory genomes

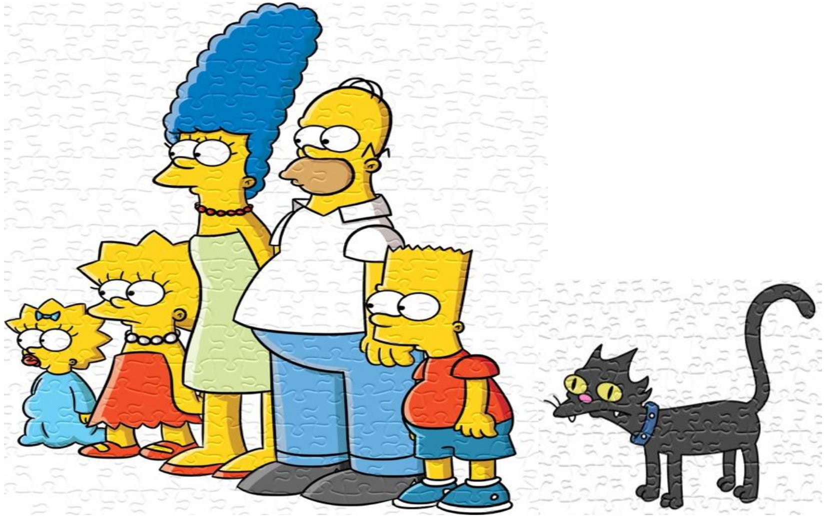
What is the core/accessory/pangenome?

- Bacterial strains have different sets of genes
- Present in all/nearly all strains in a species – **core**
- Present only in some strains – **accessory**
- Everything seen in a species - **pangenome**

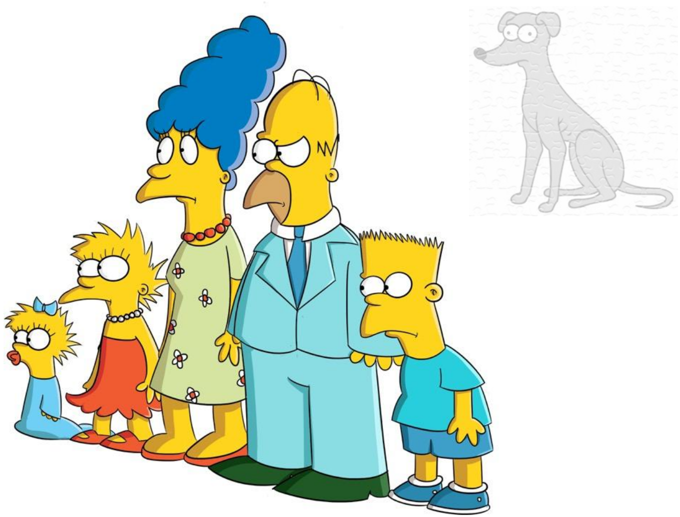
Comparing genomes



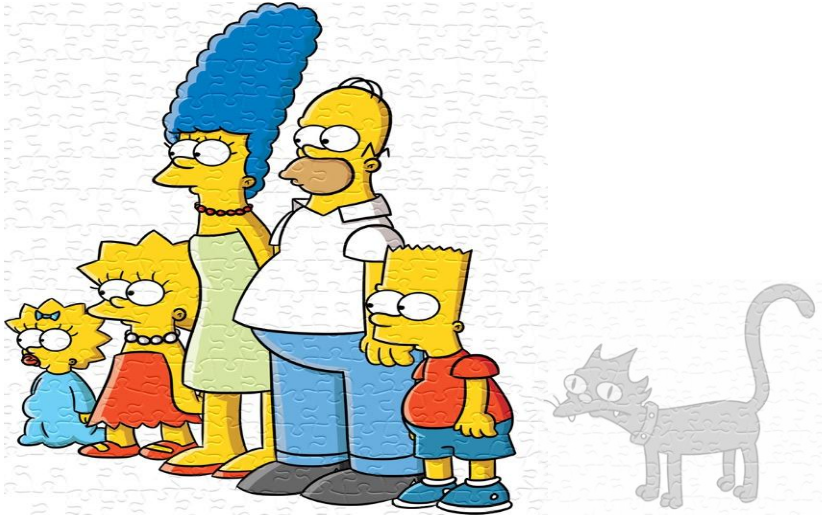
vs.



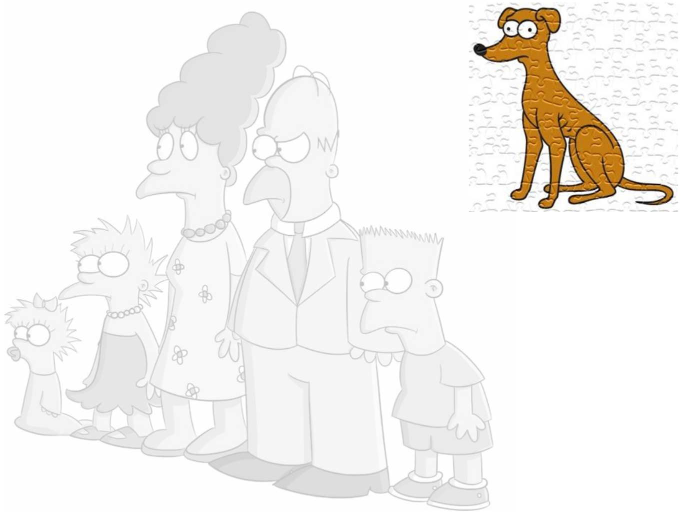
Core



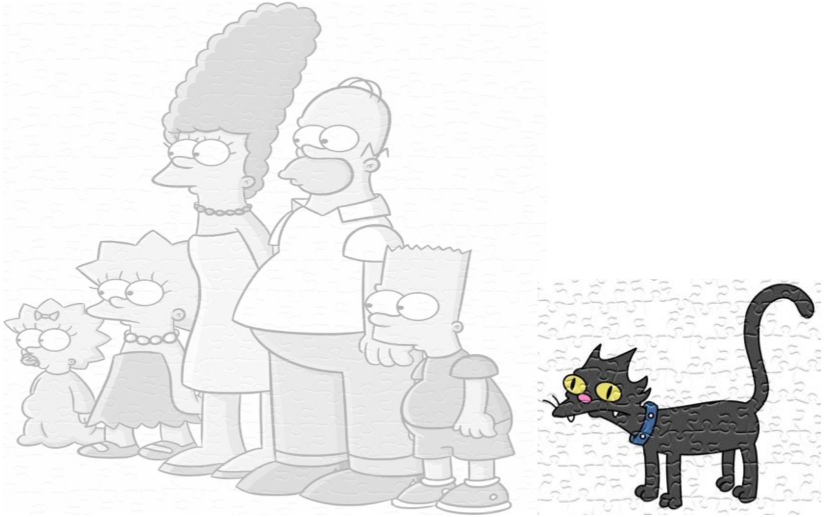
vs.



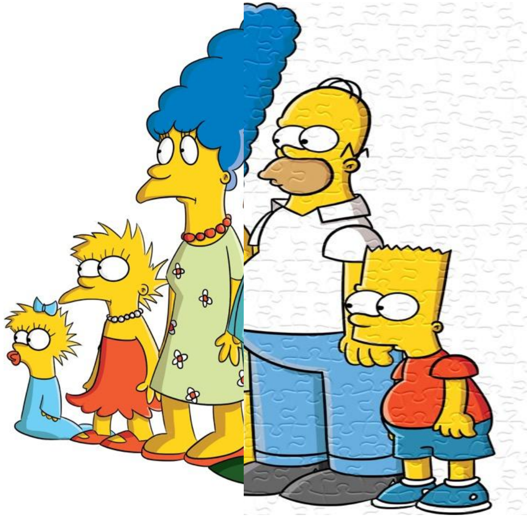
Accessory

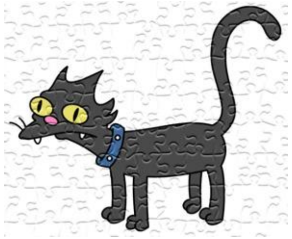


vs.



Pan

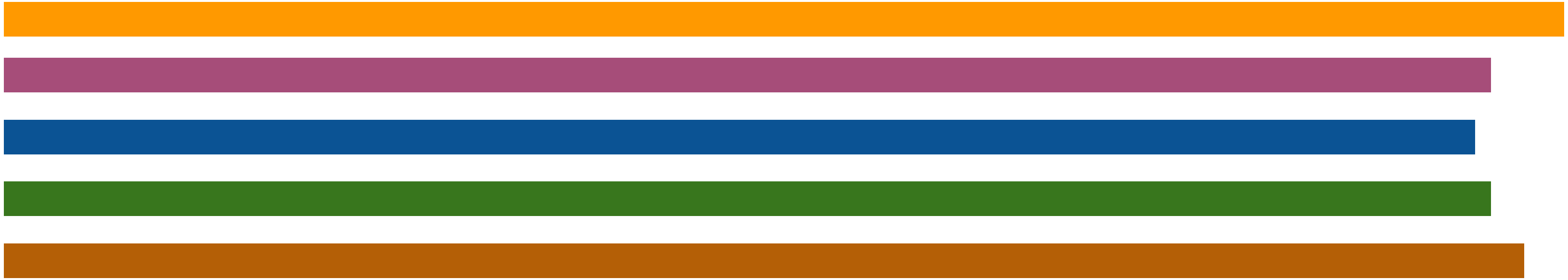




Why look at the pangenome?

- Choice of reference genome biases the gene content
- Find pathways/genes present in only some strains
    - virulence genes, antibiotic resistant genes, metabolic pathways
- Association of genes with other characteristics
    - virulence, environmental vs clinical isolates, disease severity
- Core genome are in every strain, represent essential genes

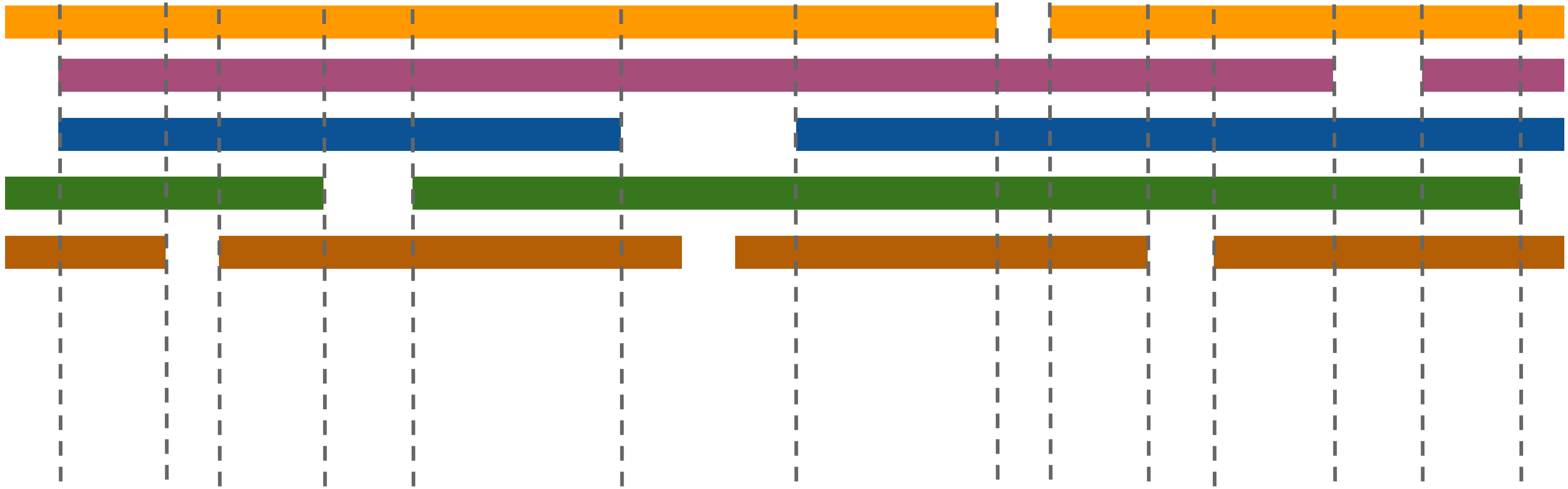
Five genomes



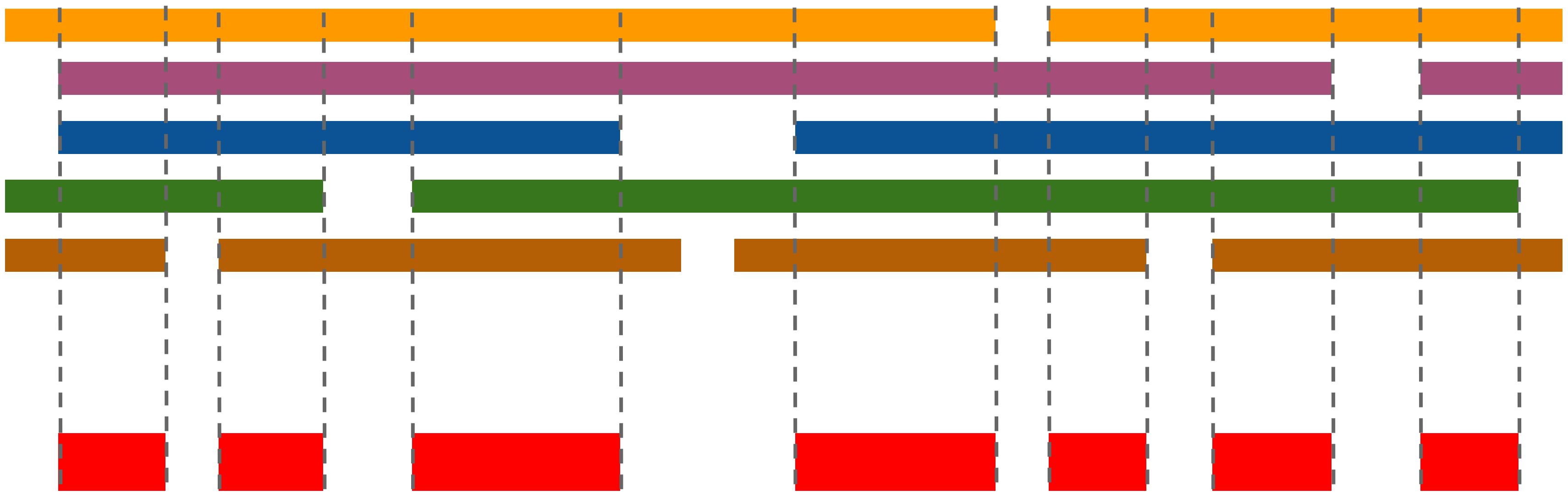
Whole genome multiple alignment ^

^ this problem is intractable

Find “common” segments

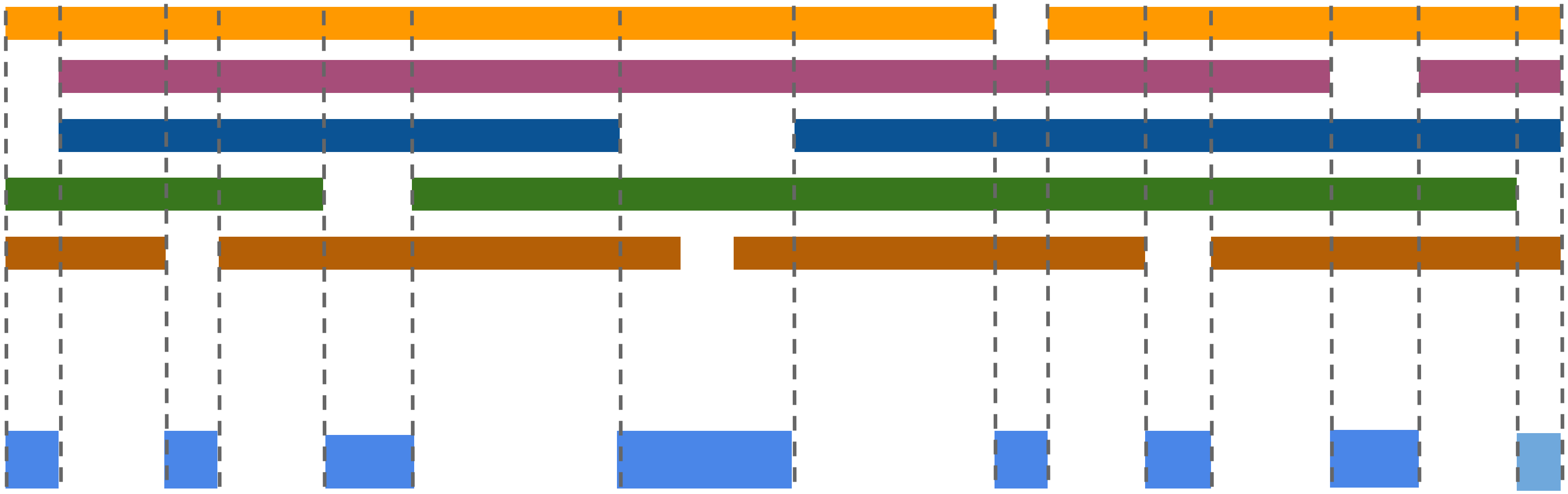


The core genome



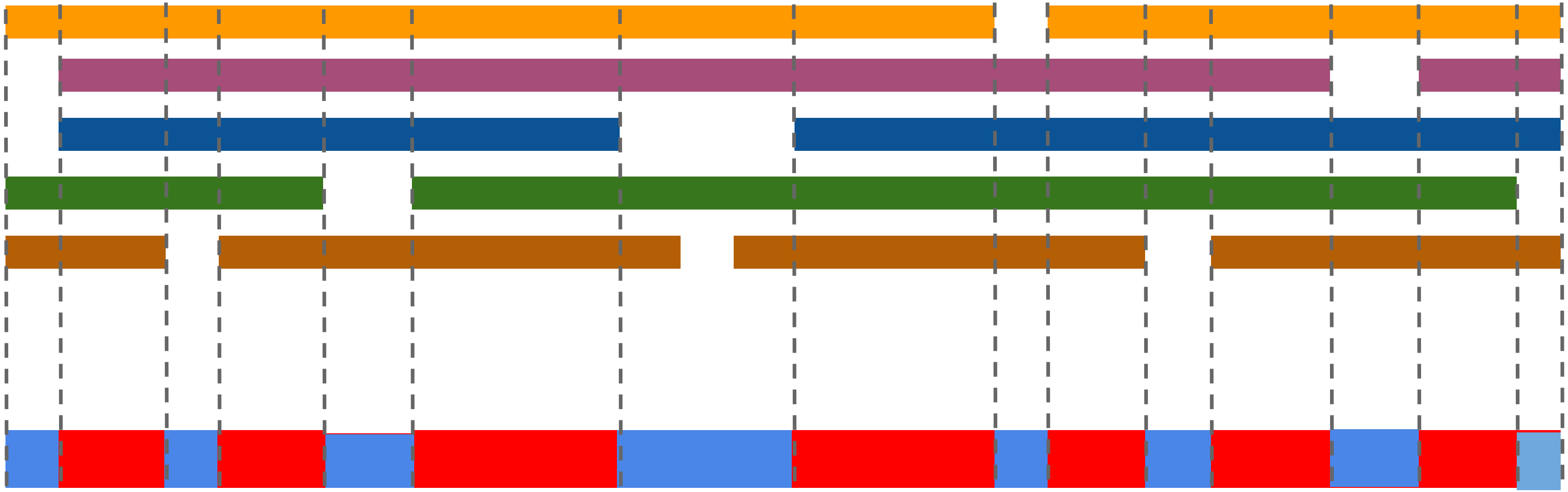
Core is common to all & has similar sequence.

The accessory genome



Accessory = not core (but still similar within)

The pan genome



Pan = Core + Accessory

Determining the pan genome

Whole genome alignment is difficult !

Rearrangements.

Sequence divergence.

Duplications.

Does not scale computationally.

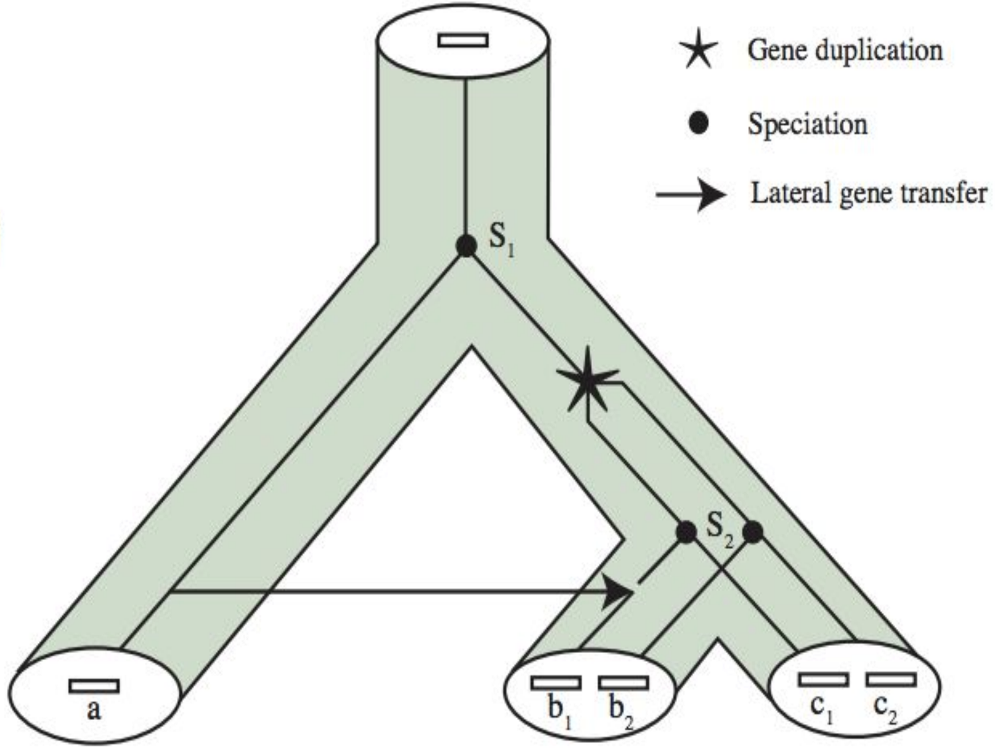
Reframing the problem

Align whole genomes
(DNA)

↓

Cluster homologous genes
(DNA or AA)

Homologs = common ancestor



<u>Ortholog</u>

Speciation

<u>Paralog</u>

Duplication

<u>Xenolog</u>

Lateral transfer

Homolog clustering

- Group homologous proteins together
  - exploit sequence similarity + synteny + operons
  - all versus all sequence comparison (not scalable)
    - DNA or amino acid (fast heuristics)
  - difficulty increases with taxa distance
- Depends on annotation quality
    - Missing genes
    - False genes

Typical workflow

:: *De novo* assembly - SPAdes

:: Annotation - Prokka

:: Pan-genome - Roary

:: Visualise - Phandango

SPAdes

Roary: the Pan Genome Pipeline

phandango

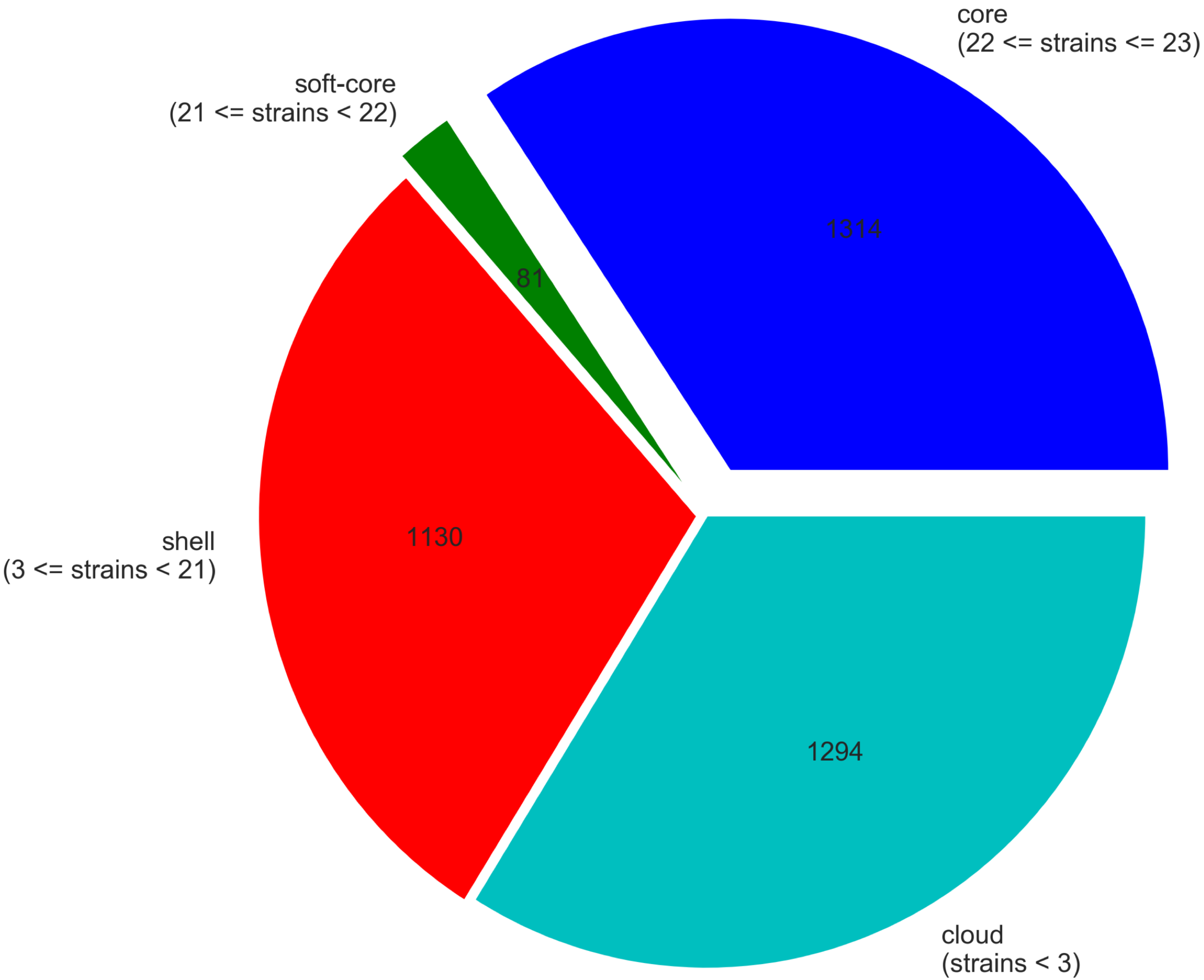
Roary → matrix / spreadsheet

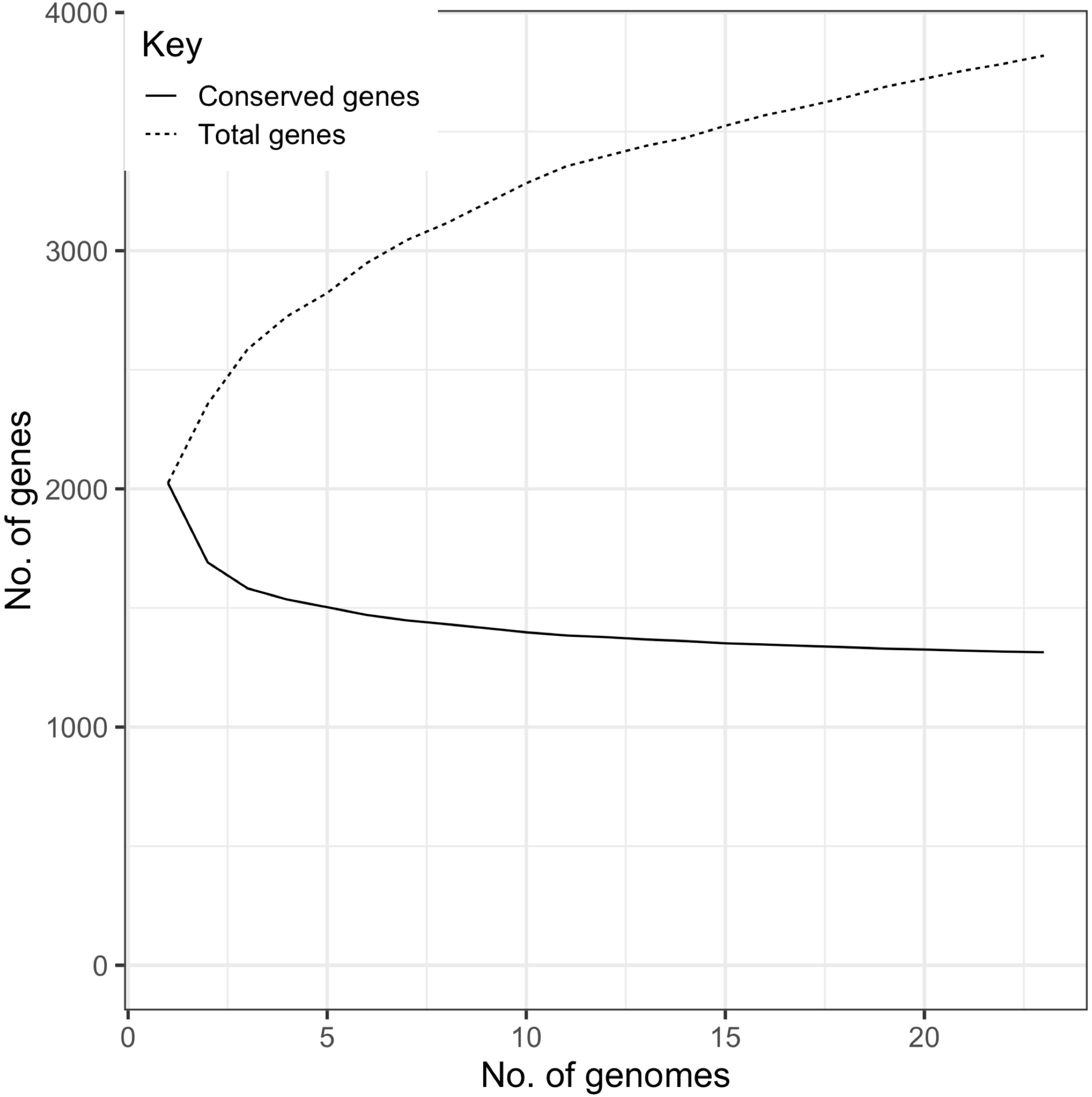
CLUSTER	STRAIN1	STRAIN2	STRAIN3
00001	DNO1000	EHEC1000	MRSA_1000
00002	DNO1001	EHEC1002	MRSA_1001
00003	DNO1002	EHEC1003	MRSA_1002
00004	DNO1003	EHEC1004	MRSA_1003
00005	DNO1004	EHEC1005	MRSA_1022
:	:	:	:
02314	DNO1005	na	MRSA_1023
02315	DNO1451	EHEC3215	na
02316	na	EHEC3216	MRSA_1923
:	:	:	:
04197	DNO1456	na	na
04198	na	EHEC3877	na
04199	na	na	MRSA_0533

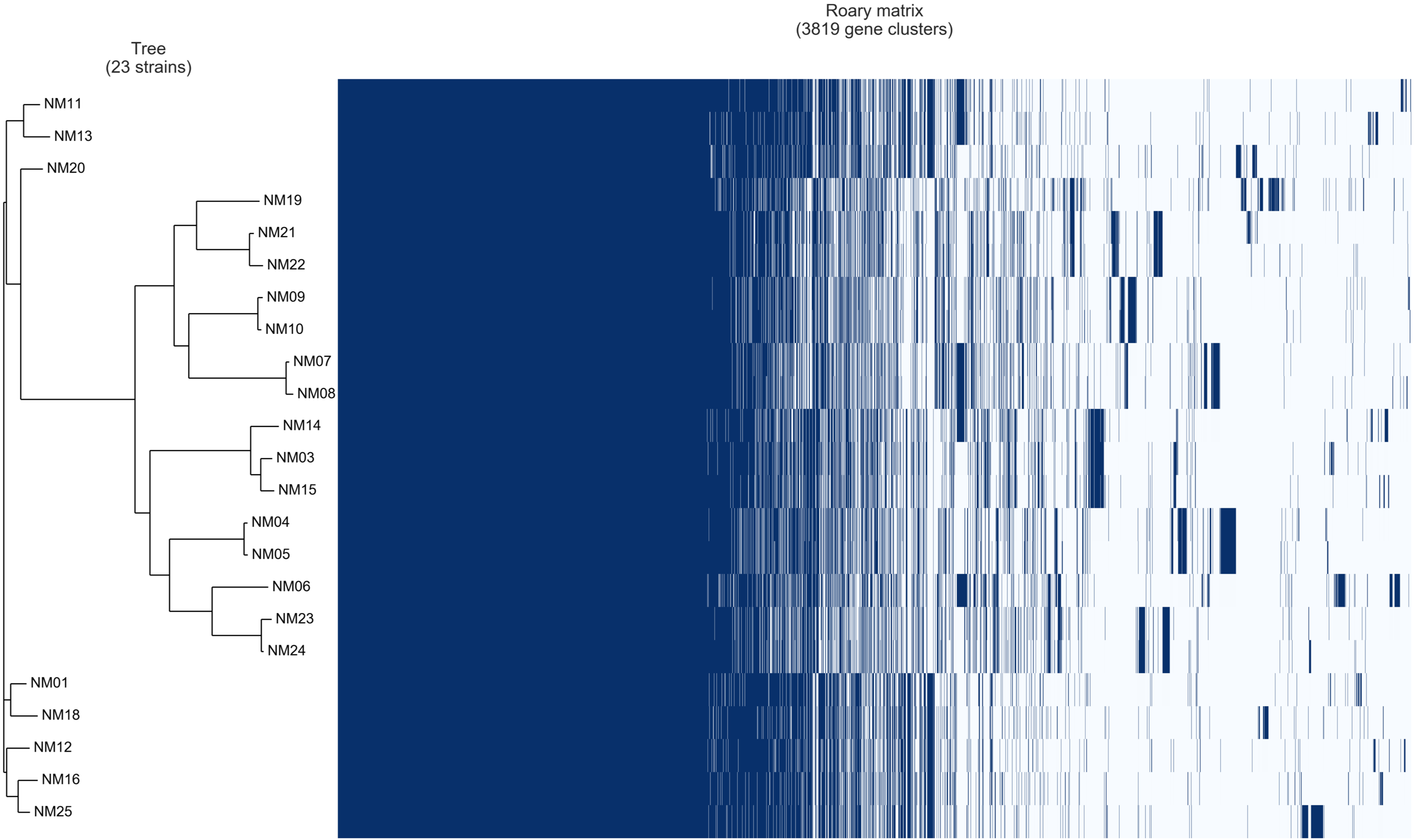
Core

Dispensable

Isolate-specific


core
(22 <= strains <= 23)
1314
soft-core
(21 <= strains < 22)
81
shell
(3 <= strains < 21)
1130
cloud
(strains < 3)
1294


Key
— Conserved genes
--- Total genes
4000
3000
2000
1000
0
No. of genes
0
5
10
15
20
No. of genomes


Tree
(23 strains)
Roary matrix
(3819 gene clusters)
NM11
NM13
NM20
NM19
NM21
NM22
NM09
NM10
NM07
NM08
NM14
NM03
NM15
NM04
NM05
NM06
NM23
NM24
NM01
NM18
NM12
NM16
NM25

Visualisation tools

- Phandango demo