Core and accessory genomes

What is the core/accessory/pangenome?

- Bacterial strains have different sets of genes
- Genes present in all/nearly all strains in a species – **core**
- Genes present only in some strains – **accessory**
- All the genes seen in a species - **pangenome**

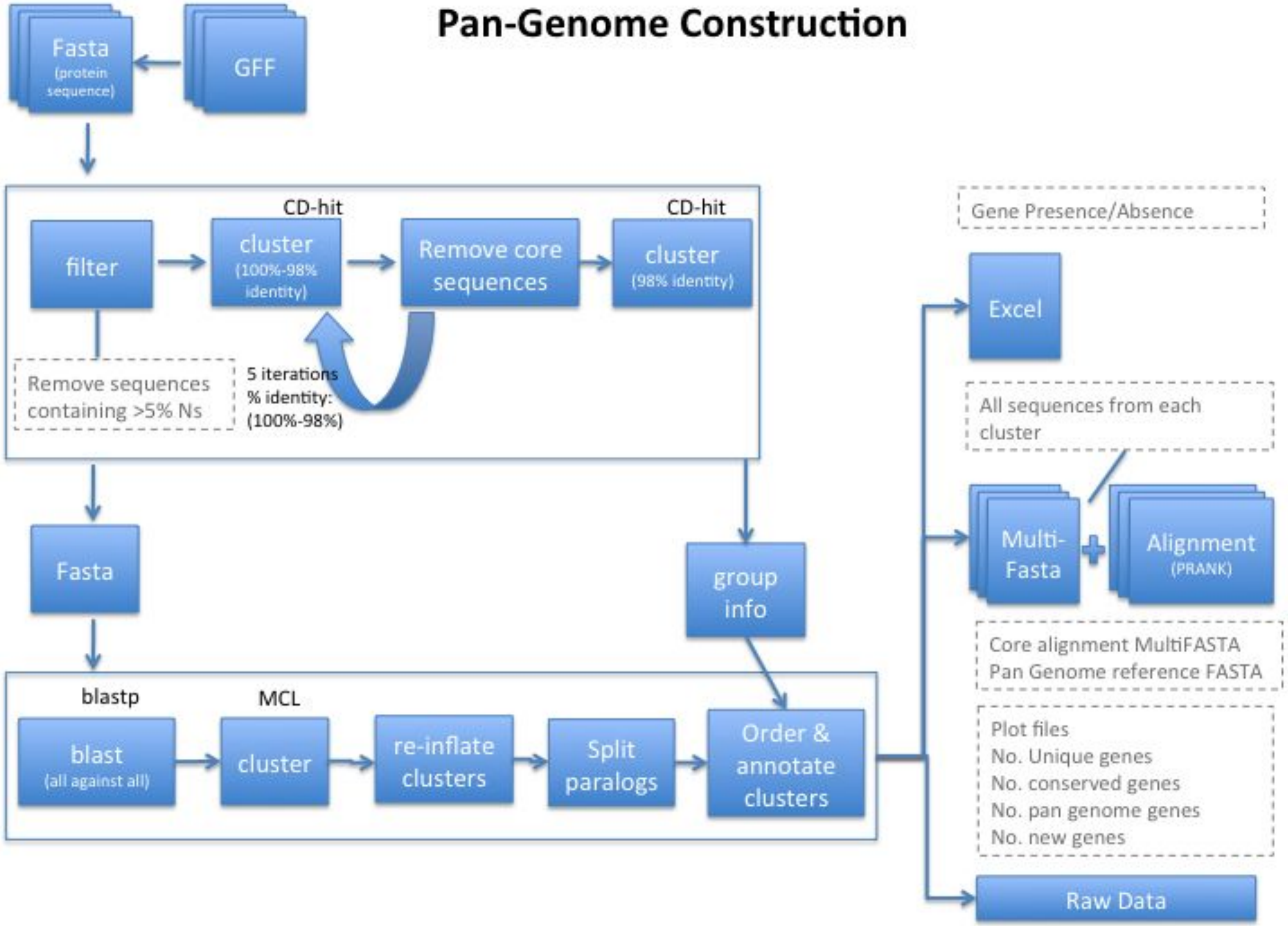
Why look at the pangenome?

- 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

Determining the core genome

- Start with a collection of genome assemblies
    - Alignment to a reference genome will miss genes not in that ref
- Annotate the genes in the genome
- Cluster the genes based on strain presence/absence

Pan-Genome Construction



Fasta (protein sequence)

GFF

CD-hit

filter

cluster (100%-98% identity)

Remove core sequences

CD-hit

cluster (98% identity)

Remove sequences containing >5% Ns

5 iterations
% identity:
(100%-98%)

Fasta

group info

blastp

blast (all against all)

MCL

cluster

re-inflate clusters

Split paralogs

Order & annotate clusters

Gene Presence/Absence

Excel

All sequences from each cluster

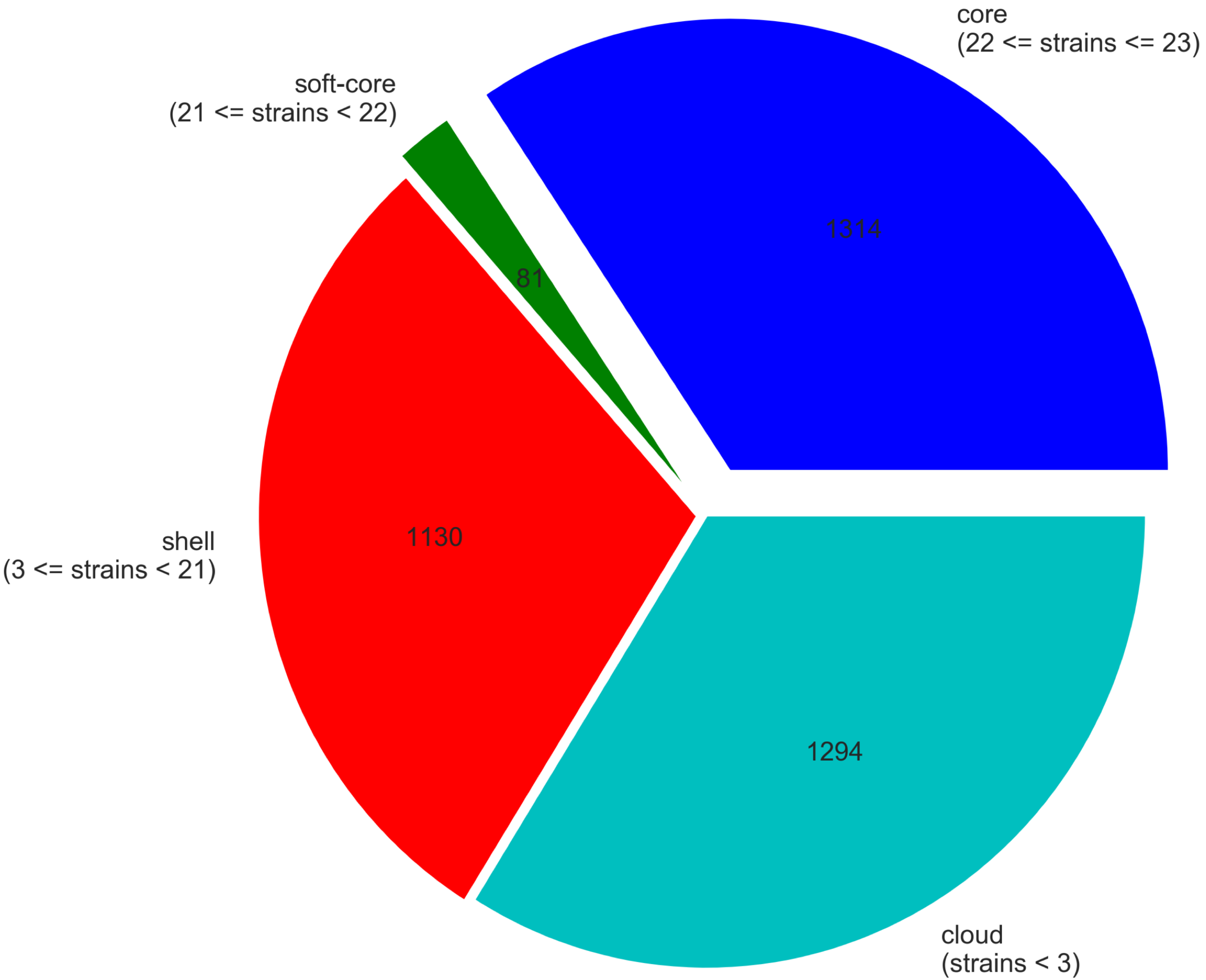
Multi-Fasta

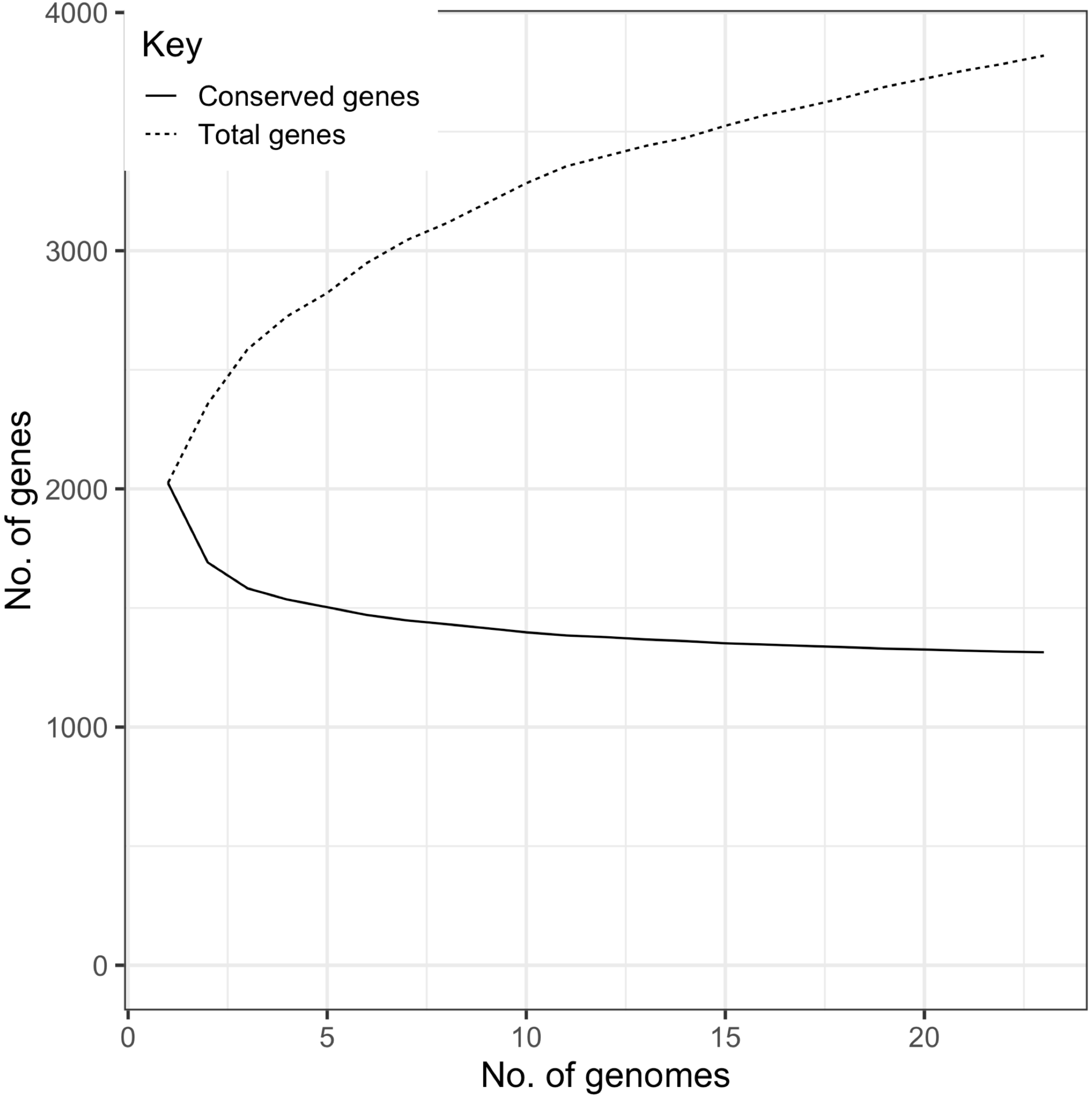
Alignment (PRANK)

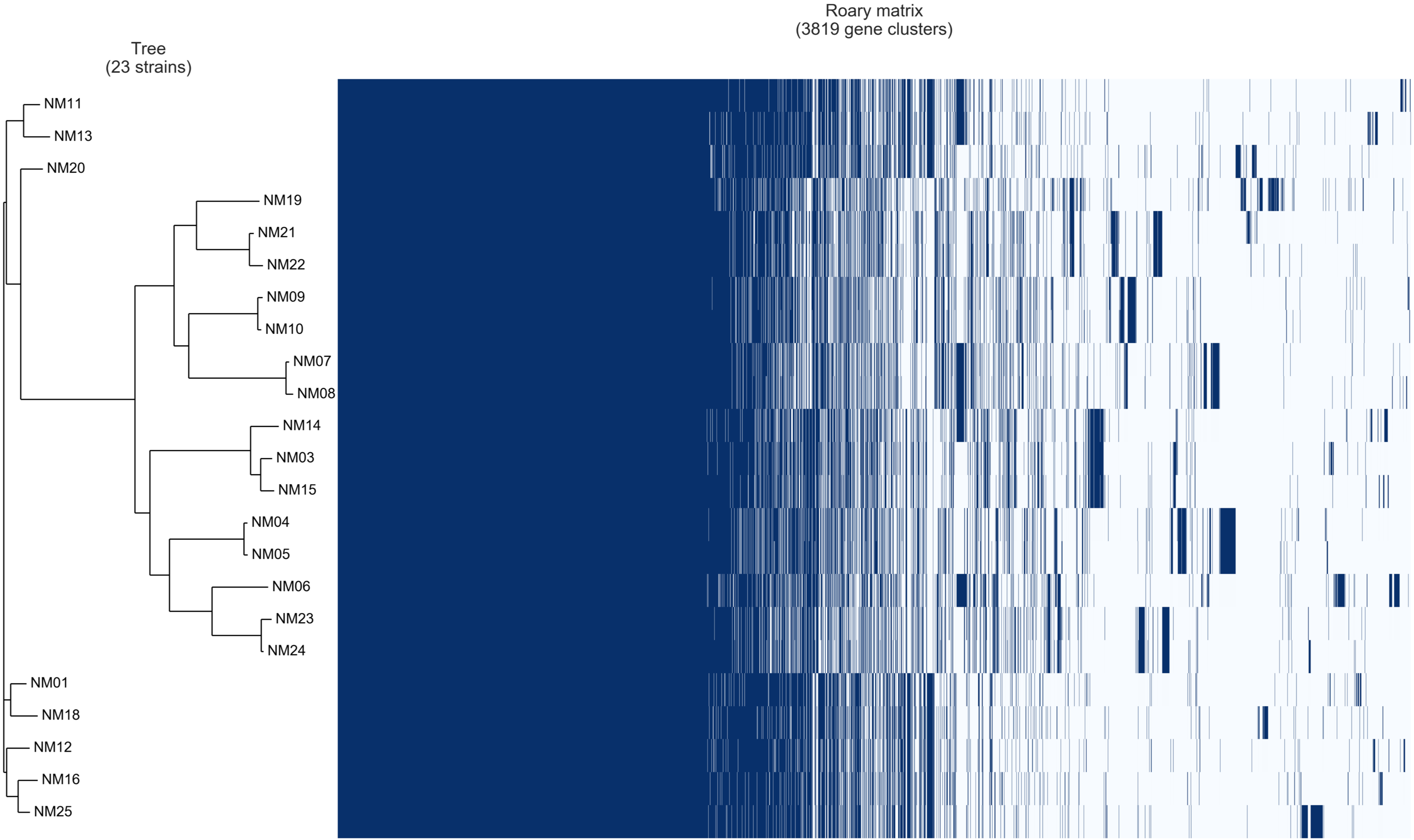
Core alignment MultiFASTA
Pan Genome reference FASTA

Plot files
No. Unique genes
No. conserved genes
No. pan genome genes
No. new genes

Raw Data


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