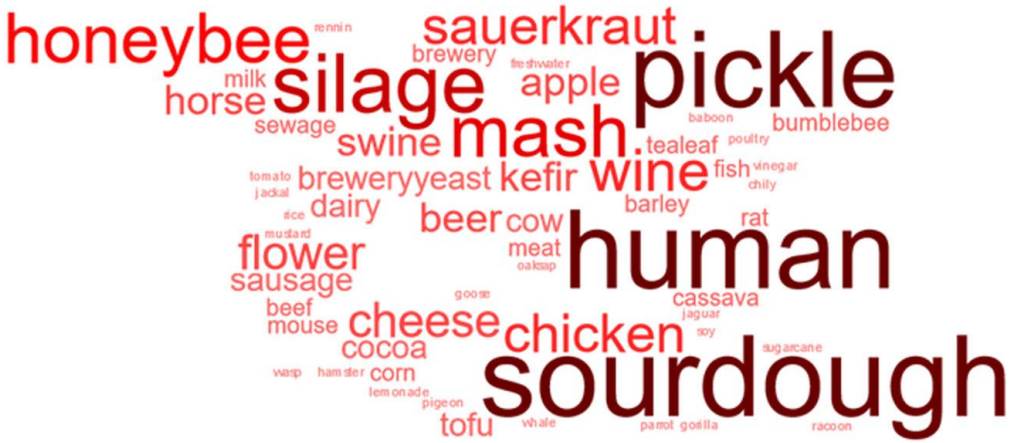
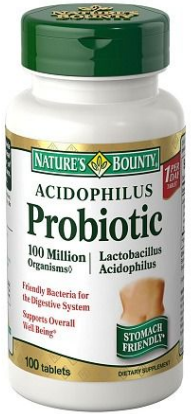
Leveraging public genomes: a case study in *Lactobacillus*

Stijn Wittouck
promoter: Sarah Lebeer (UA)
copromoter: Vera van Noort (KUL)

Background

Goal of PhD: **study evolution of the *Lactobacillus* Genus Complex using public genomes**



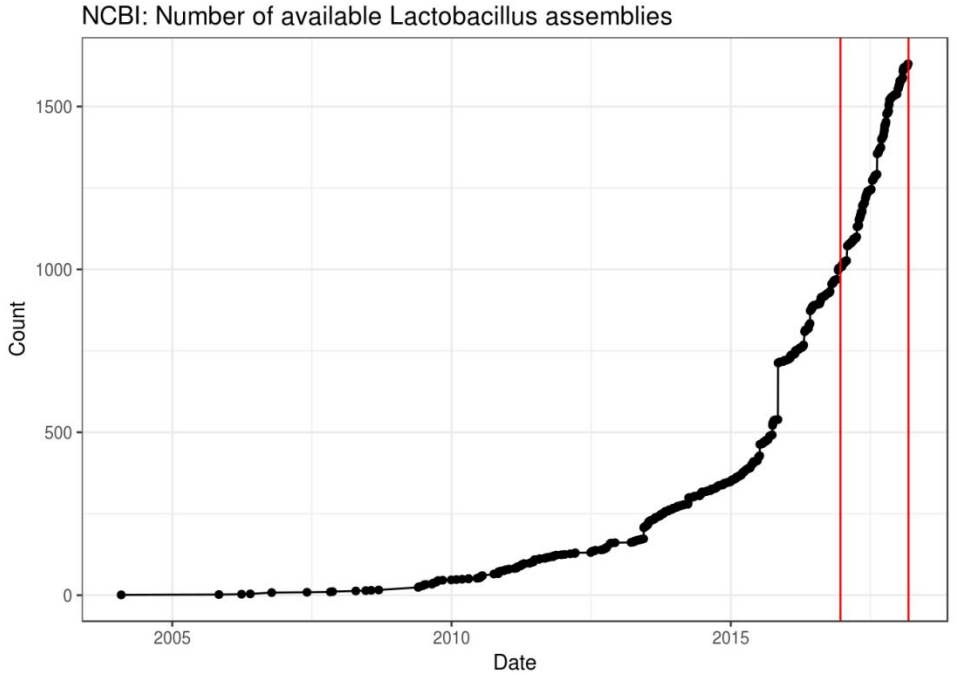


Challenges

1) Data size

2,110 LGC genomes
=> 4,452,100 comparisons

5,438,039 LGC genes
=> 29,572,268,165,521 comparisons



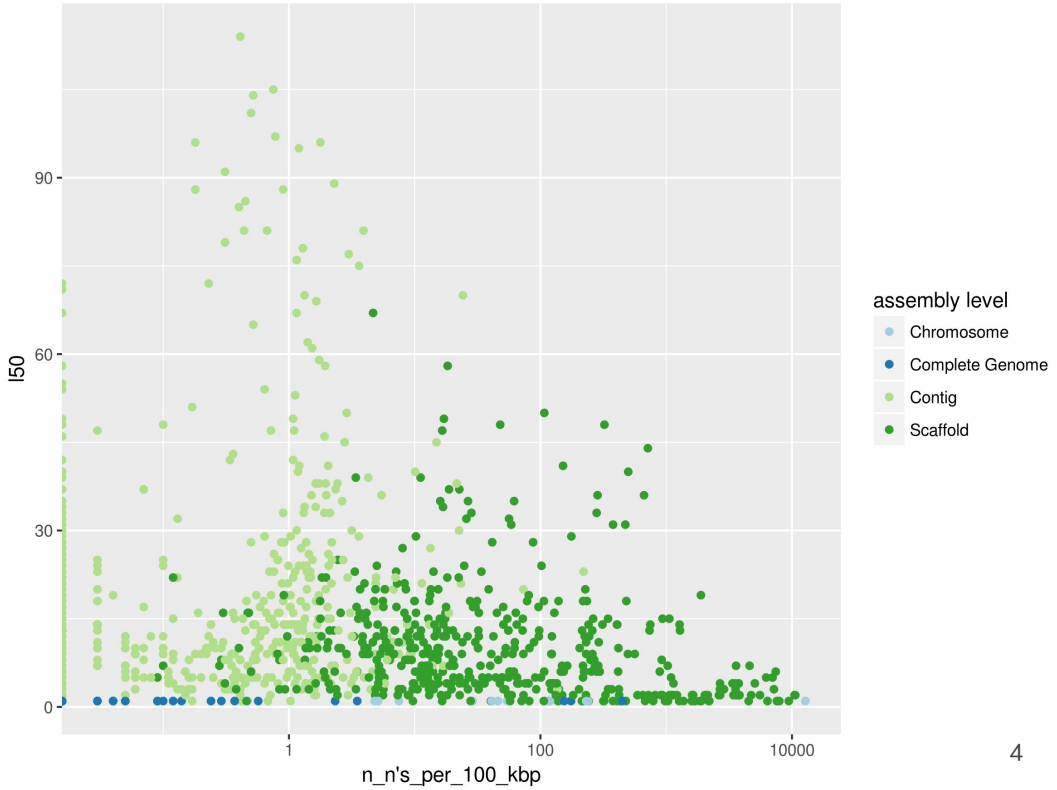
source: blog of Sander Wuyts (swuyts.wordpress.com)

Challenges

2) Uncertain quality

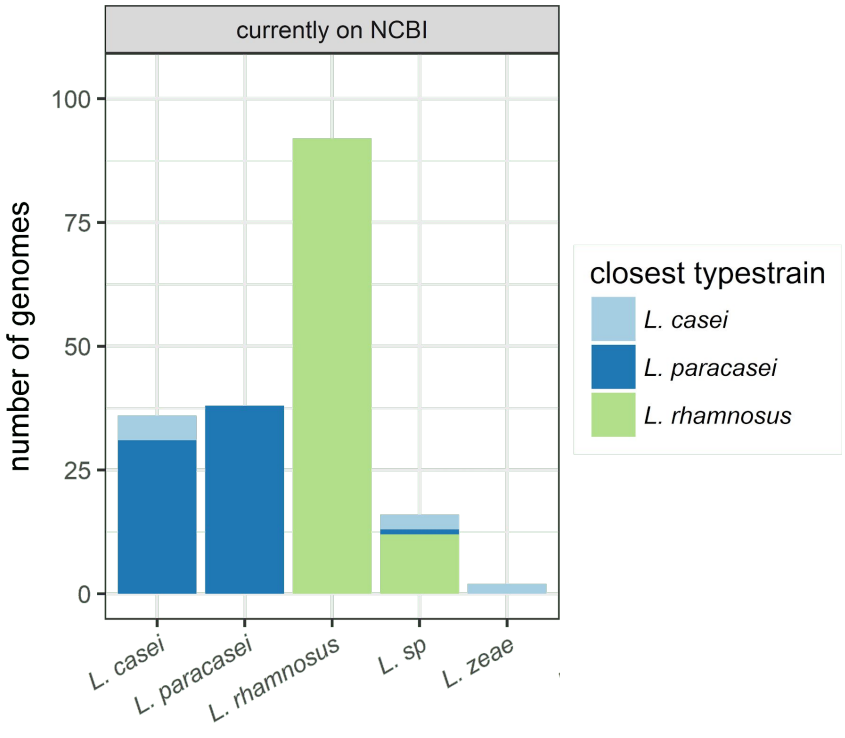
Quality is very variable

How to interpret QC measures?



Challenges

3) Uncertain taxonomy



Rapid single-copy core genes

Why?

- Intuitive genome quality control
  - Interpretation in terms of completeness and contamination
- Fast species delimitation (single-copy core nucleotide identity or SCNI)
  - Alternative to slow ANI calculations
- Phylogenetic tree inference

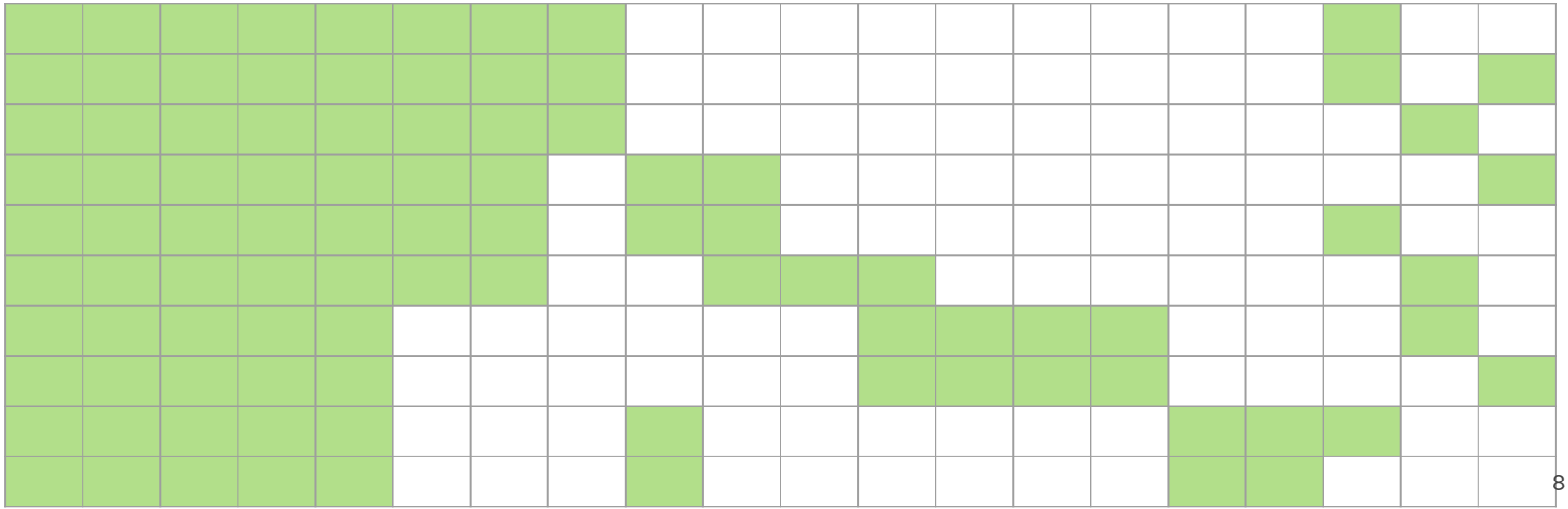
Rapid single-copy core genes

How?

Principle: **every core gene of a set of genomes is, by definition, also a core gene of a subset of those genomes**

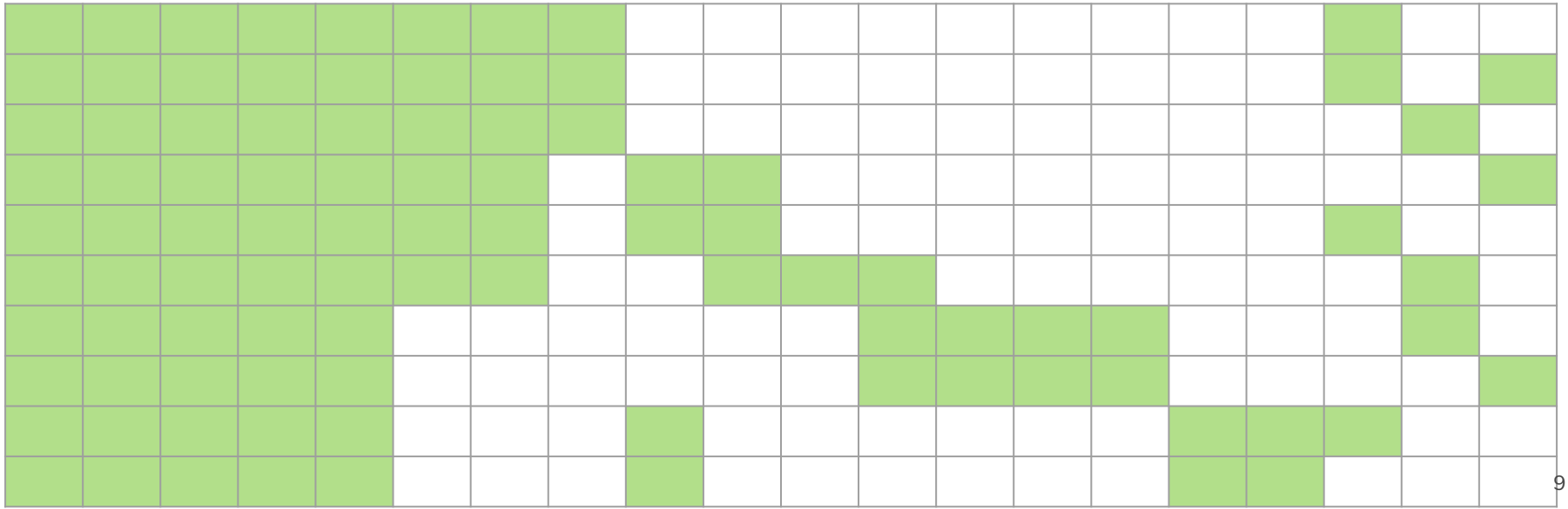
Rapid single-copy core genes

How to get all core genes?



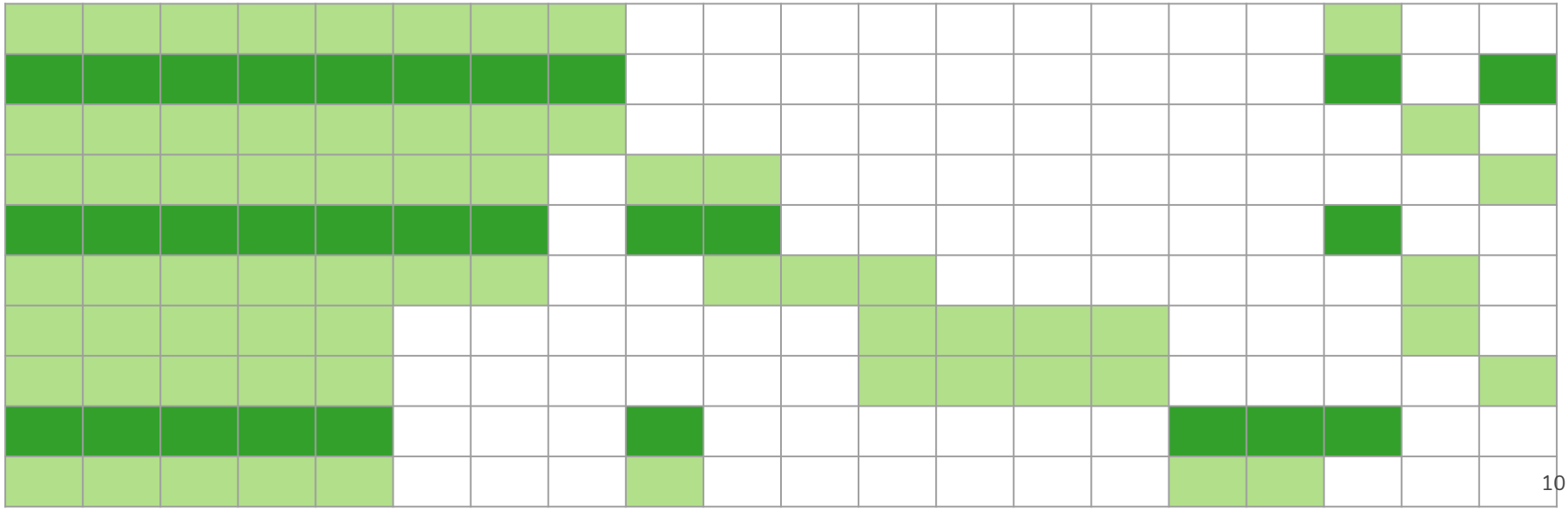
Rapid single-copy core genes

1) Select random “seed genomes” + make gene families



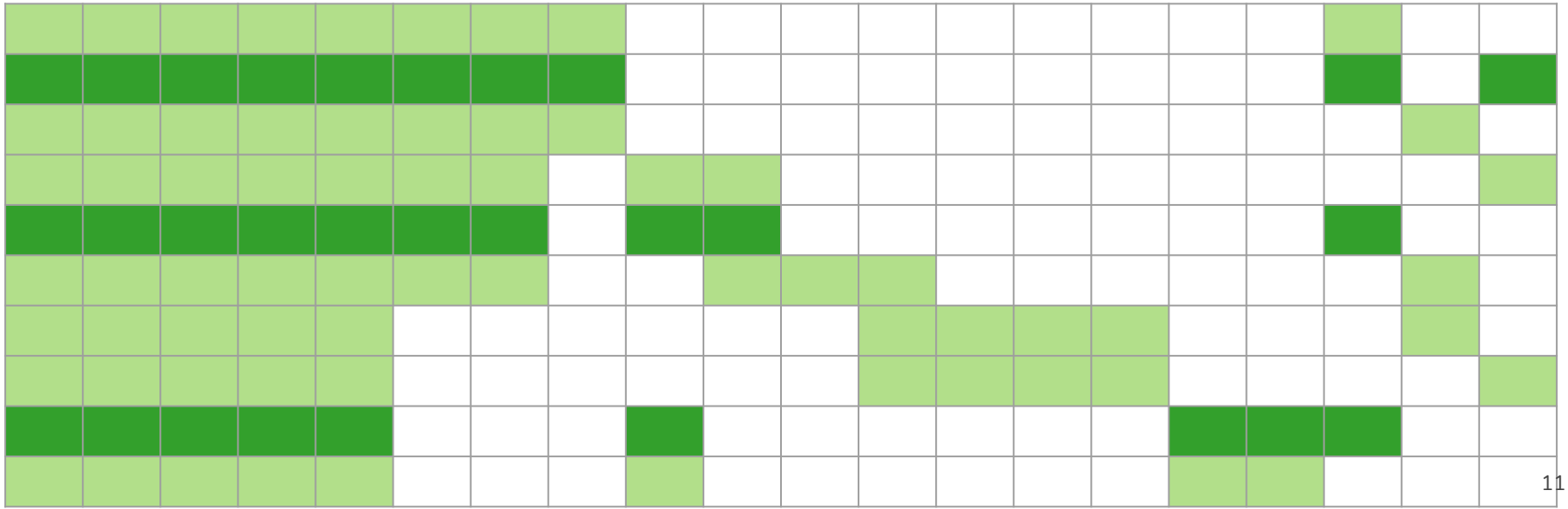
Rapid single-copy core genes

1) Select random “seed genomes” + make gene families



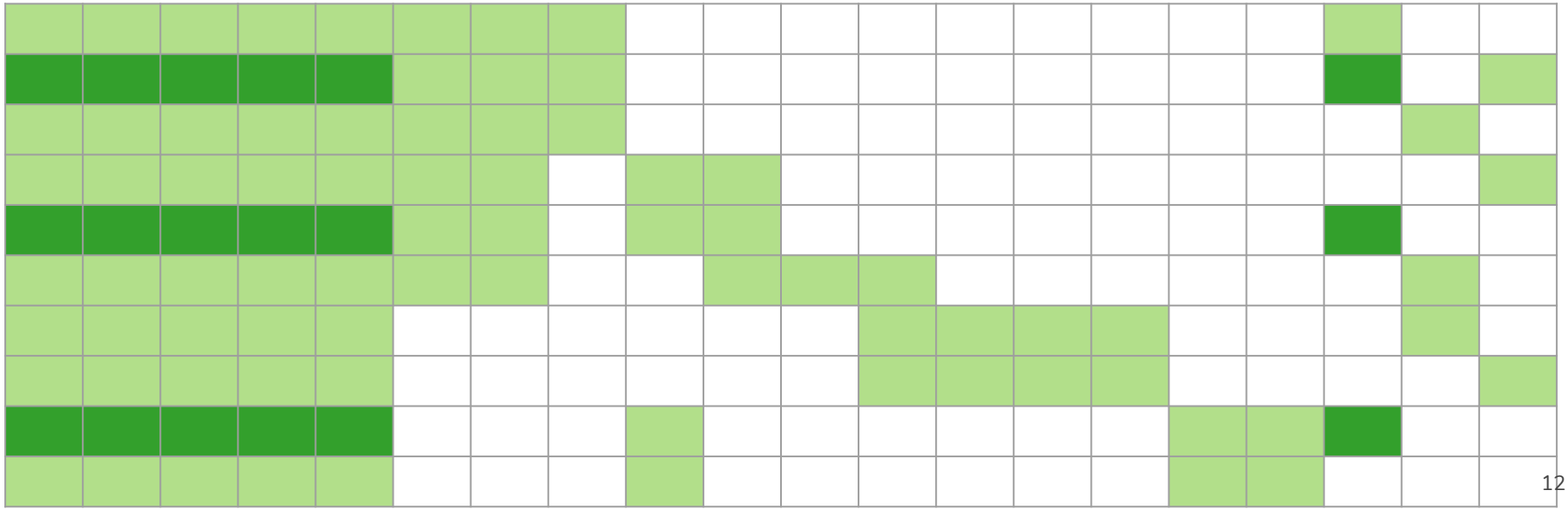
Rapid single-copy core genes

2) Keep only the “seed core genes”



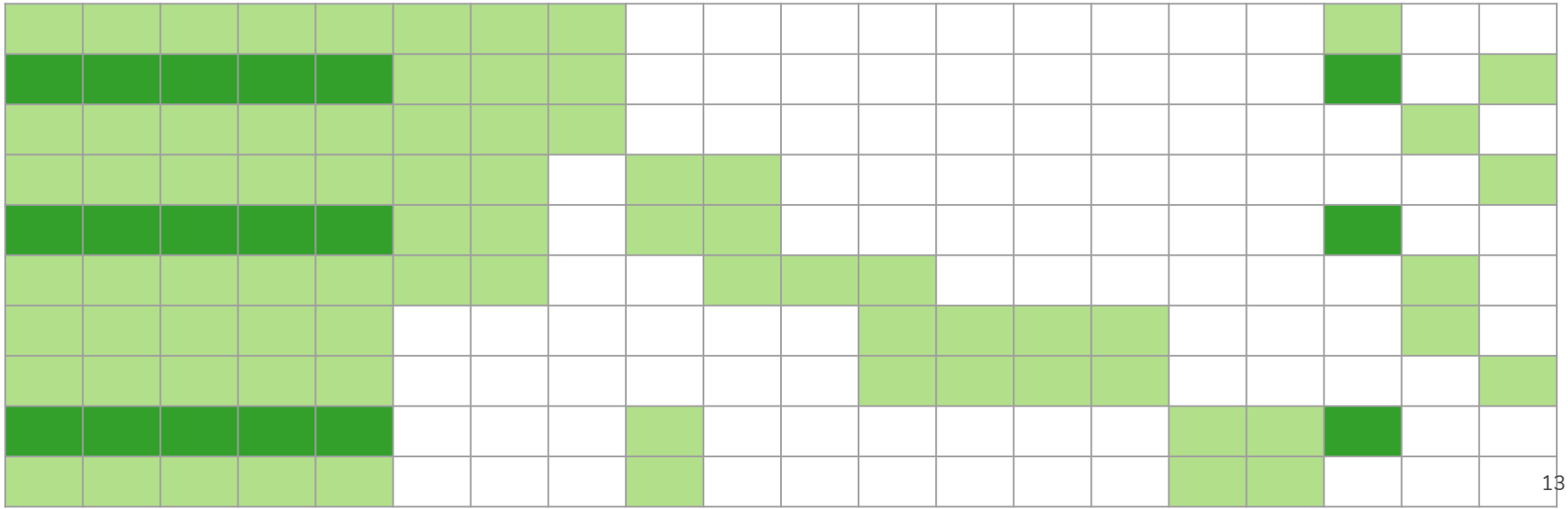
Rapid single-copy core genes

2) Keep only the “seed core genes”



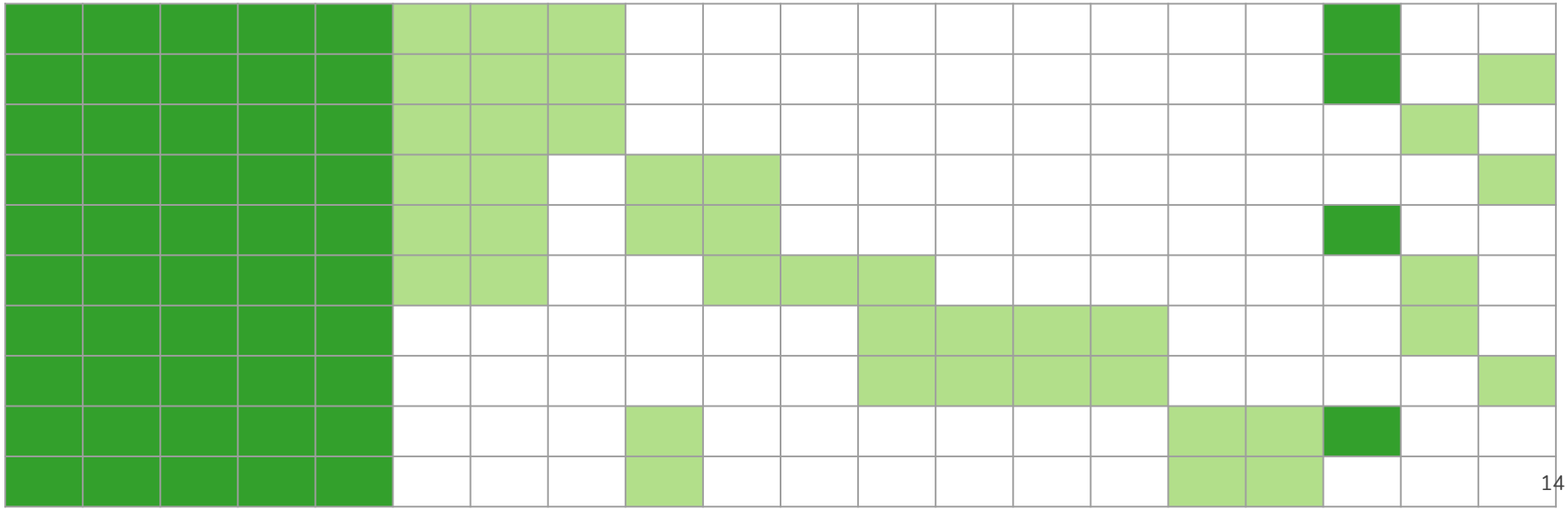
Rapid single-copy core genes

3) Find the seed core genes in all genomes (using profile HMMs)



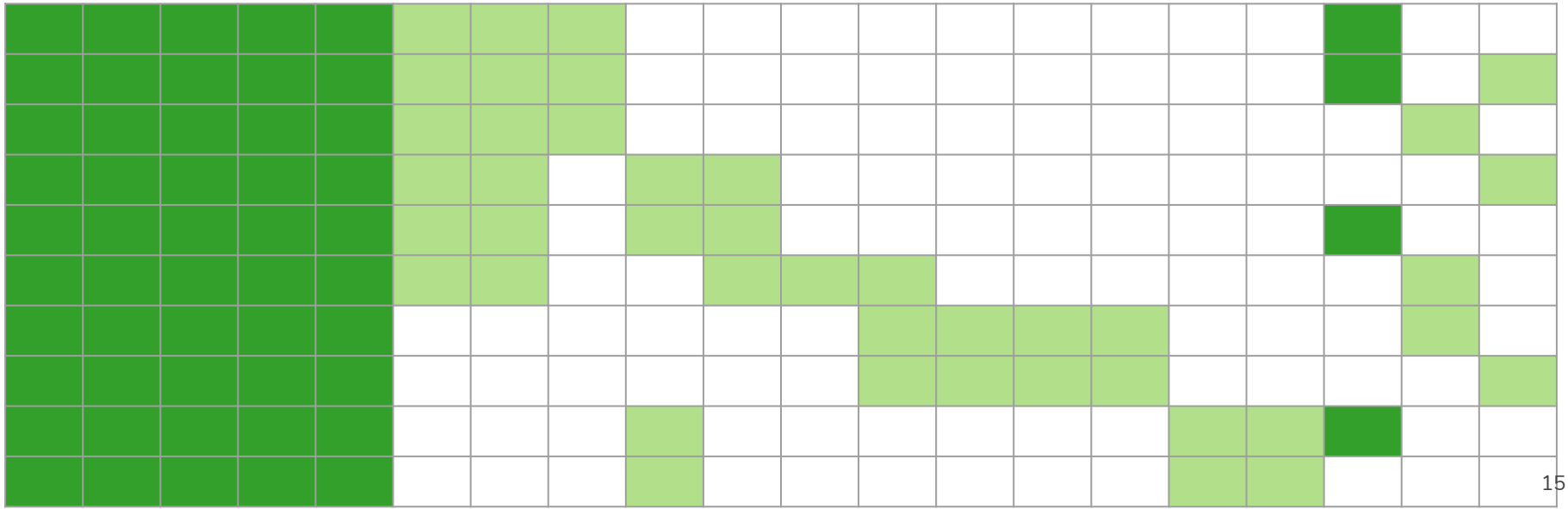
Rapid single-copy core genes

3) Find the seed core genes in all genomes (using profile HMMs)



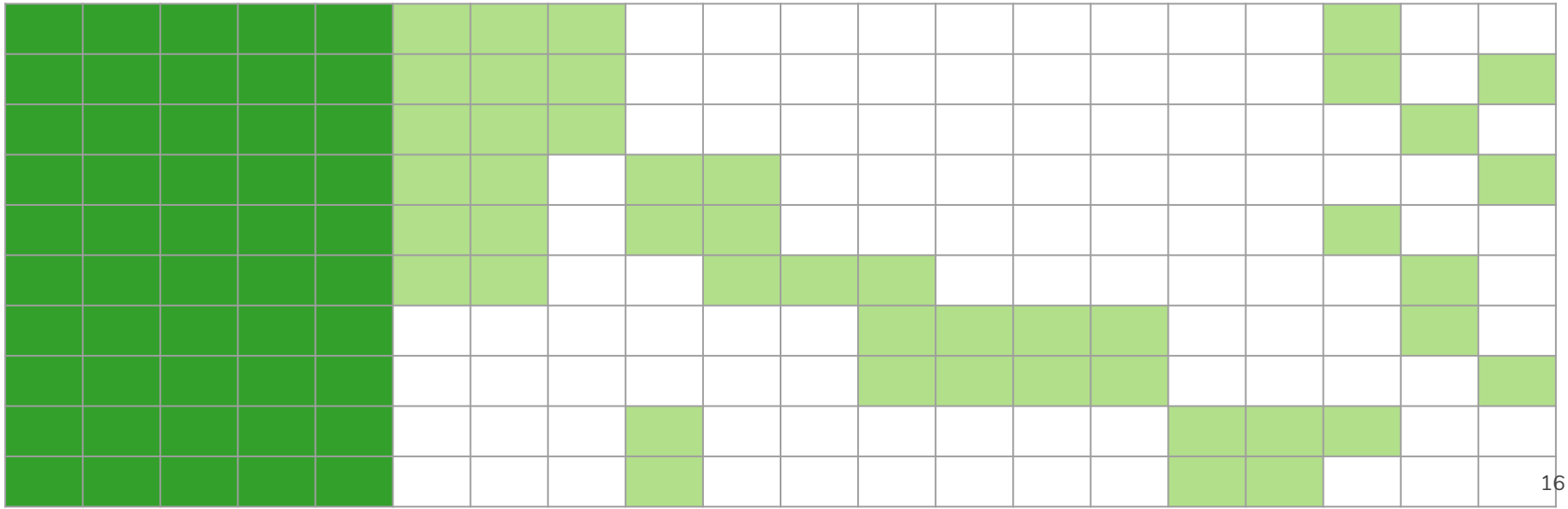
Rapid single-copy core genes

4) Keep only the core genes in all genomes

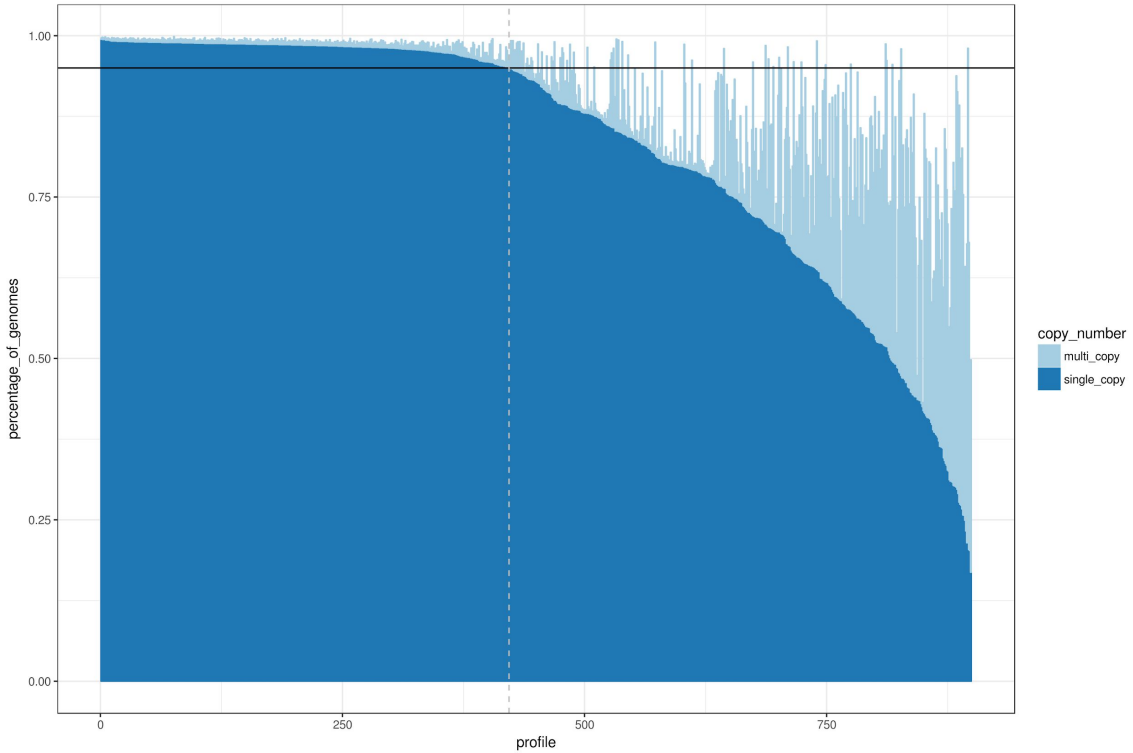


Rapid single-copy core genes

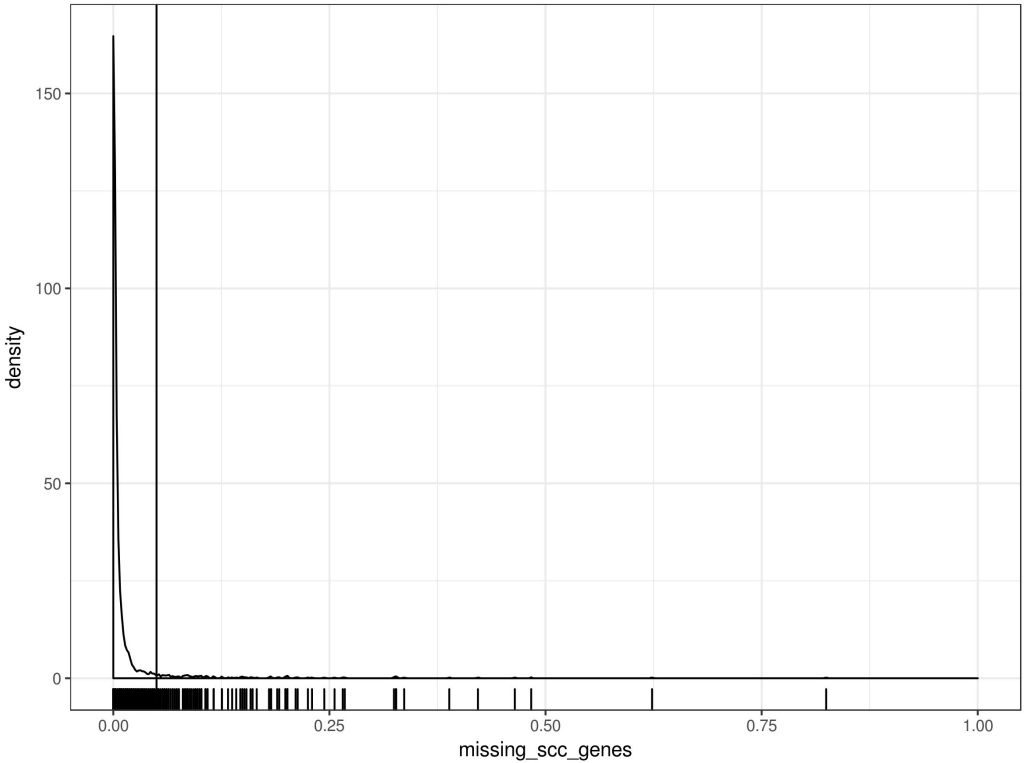
4) Keep only the core genes in all genomes



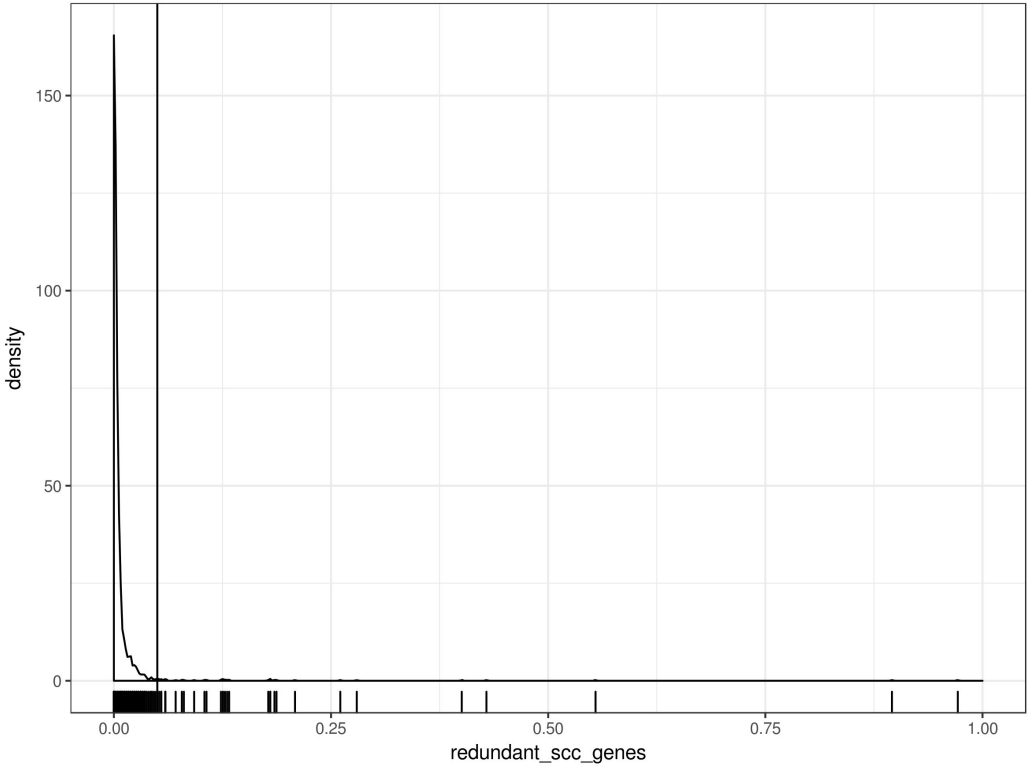
Results in the LGC



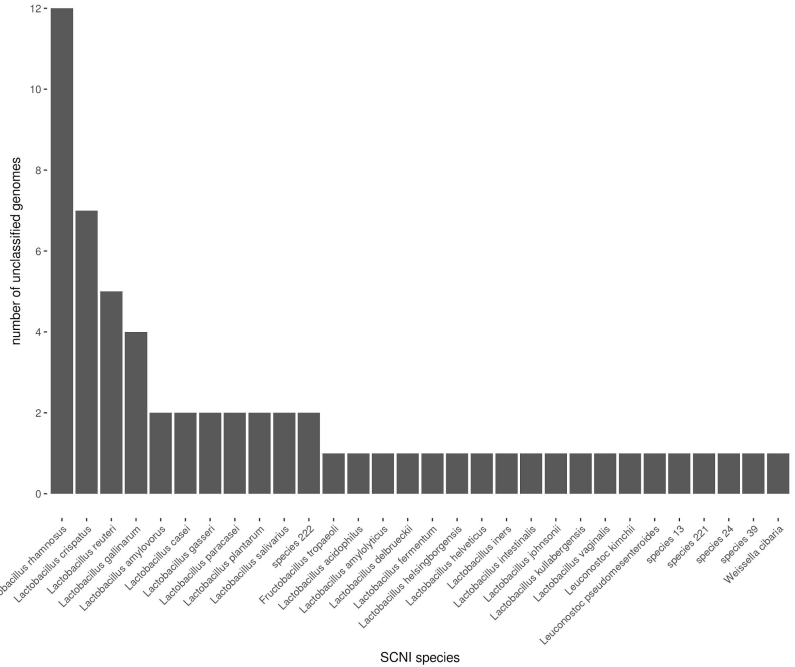
Results in the LGC

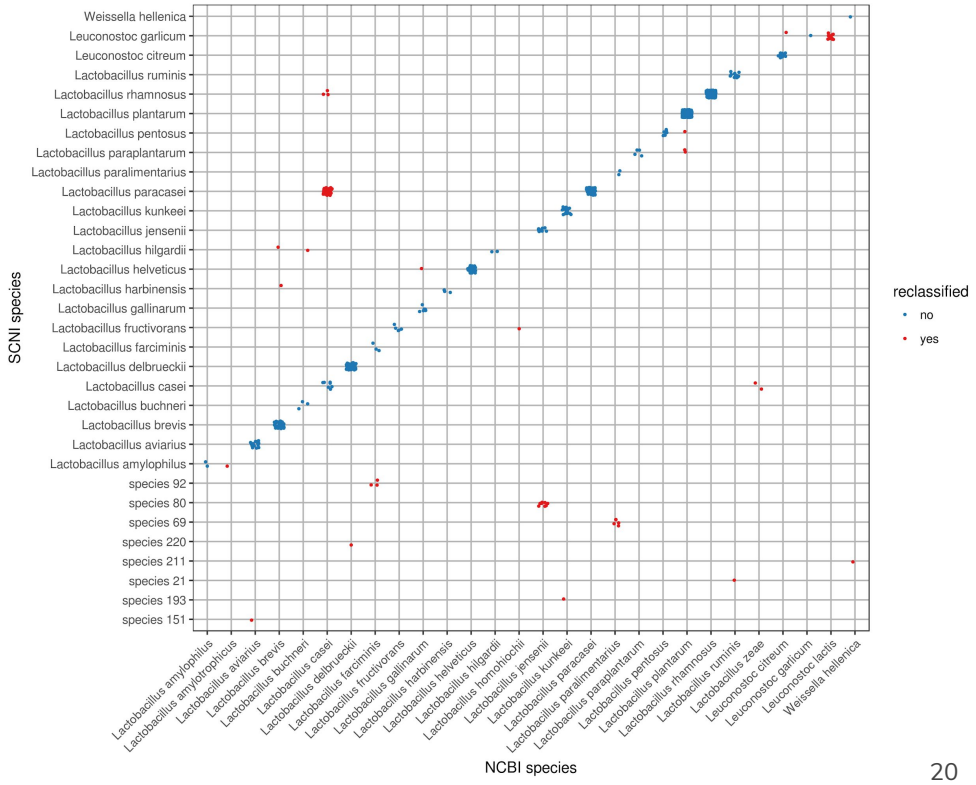


Results in the LGC



Results in the LGC





Implementation

Rapid bActerial Cross-species COmparative geNOmics

→ **RACCOON**

- In development
- Mostly python
- Temporarily some R
- Dependencies: orthofinder, hmmer, mafft

github.com/SWittouck/raccoon

Thank you!

prof. dr. ir. Sarah Lebeer

prof. dr. ir. Vera van Noort

dr. Conor Meehan

ENdEMIC
Environmental Ecology and Applied Microbiology
University of Antwerp

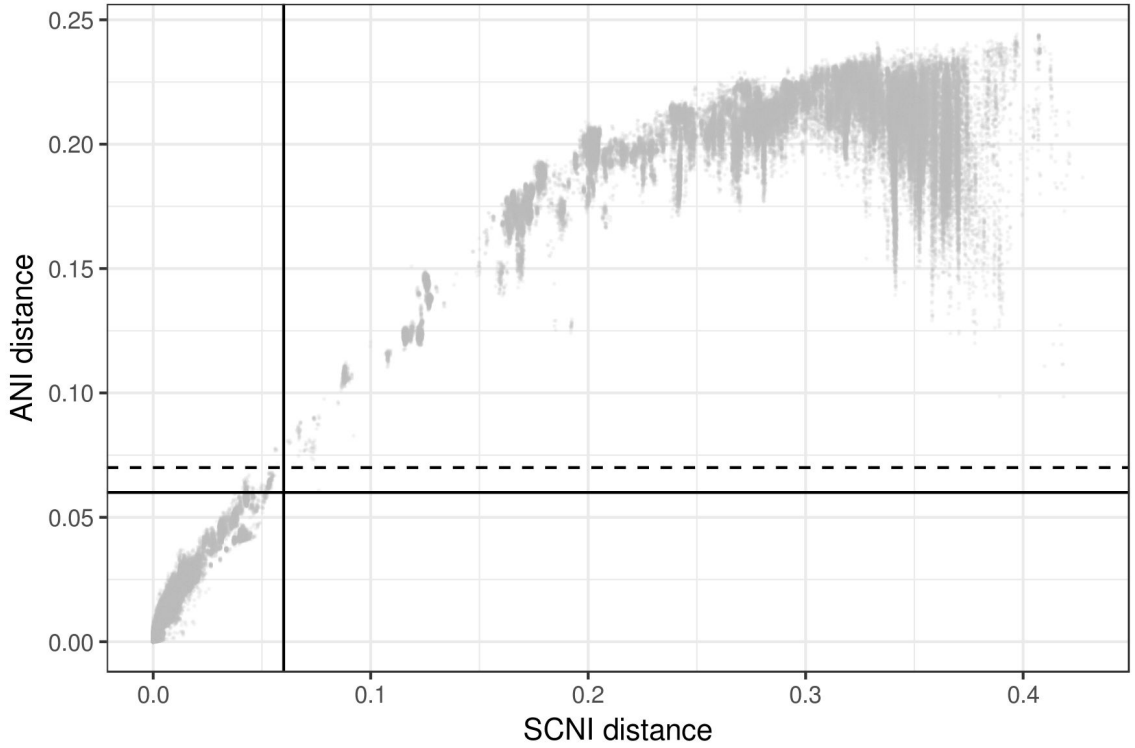
fwo
Research Foundation
Flanders
Opening new horizons

Background

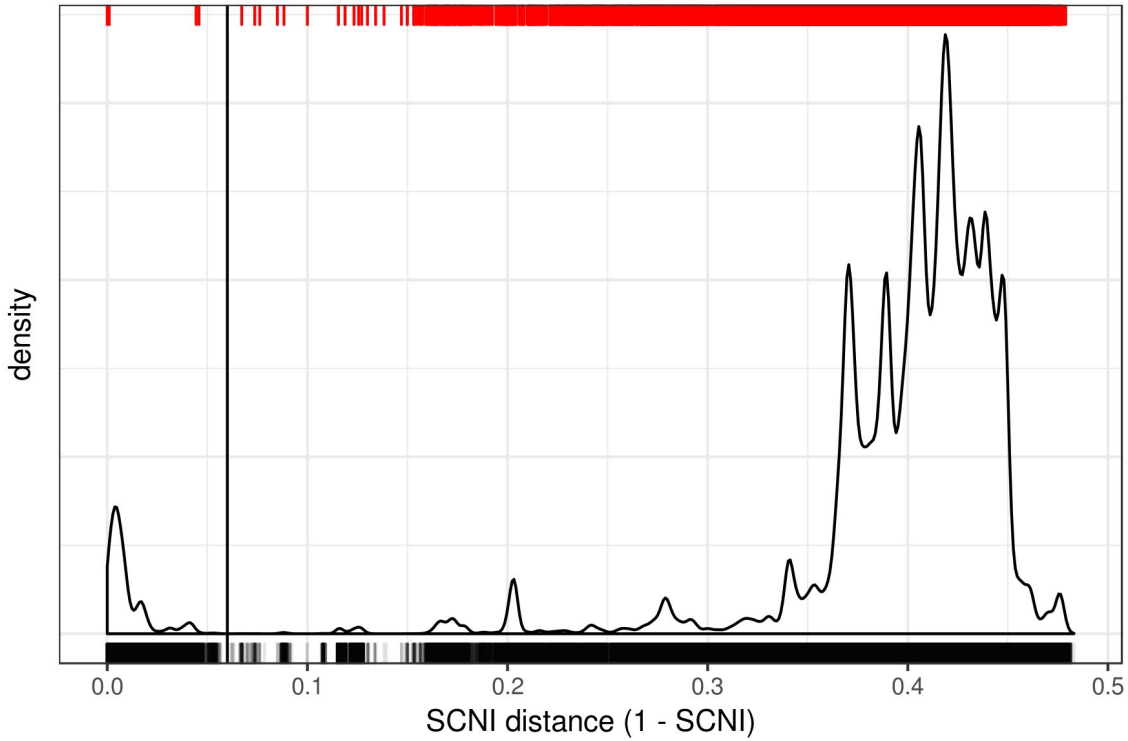
Why study Lactobacillus genome evolution?

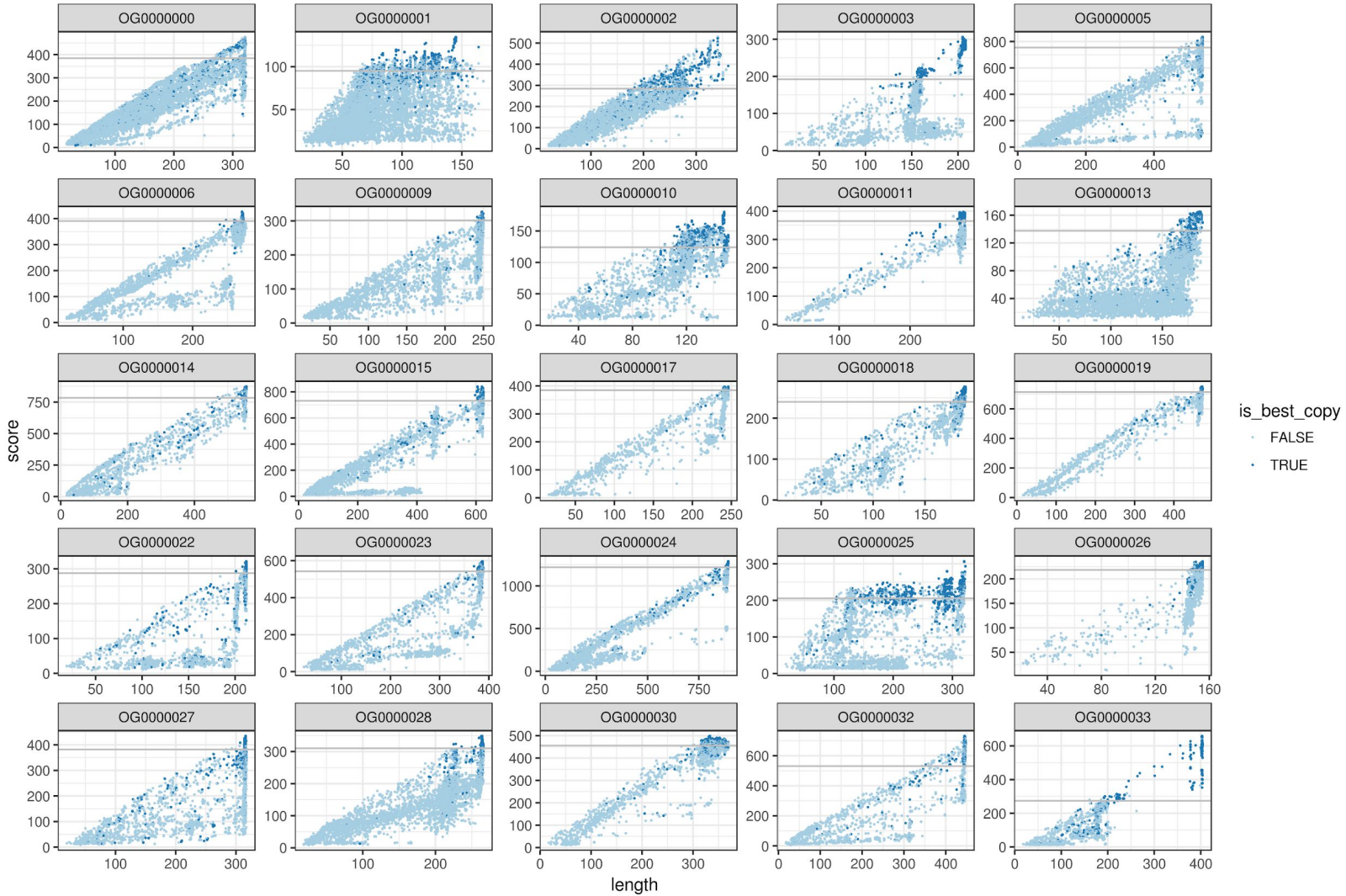
- Characterize known probiotic genes
- Characterize niche-specificity → where to find new probiotic strains?
- Characterize niche-adaptation → where to find new probiotic genes?
- Use LGC as model system for microbial evolution in general

Rapid single-copy core genes



Results in the LGC




OG0000000
OG0000001
OG0000002
OG0000003
OG0000005
OG0000006
OG0000009
OG0000010
OG0000011
OG0000013
OG0000014
OG0000015
OG0000017
OG0000018
OG0000019
OG0000022
OG0000023
OG0000024
OG0000025
OG0000026
OG0000027
OG0000028
OG0000030
OG0000032
OG0000033
score
length
is_best_copy
FALSE
TRUE

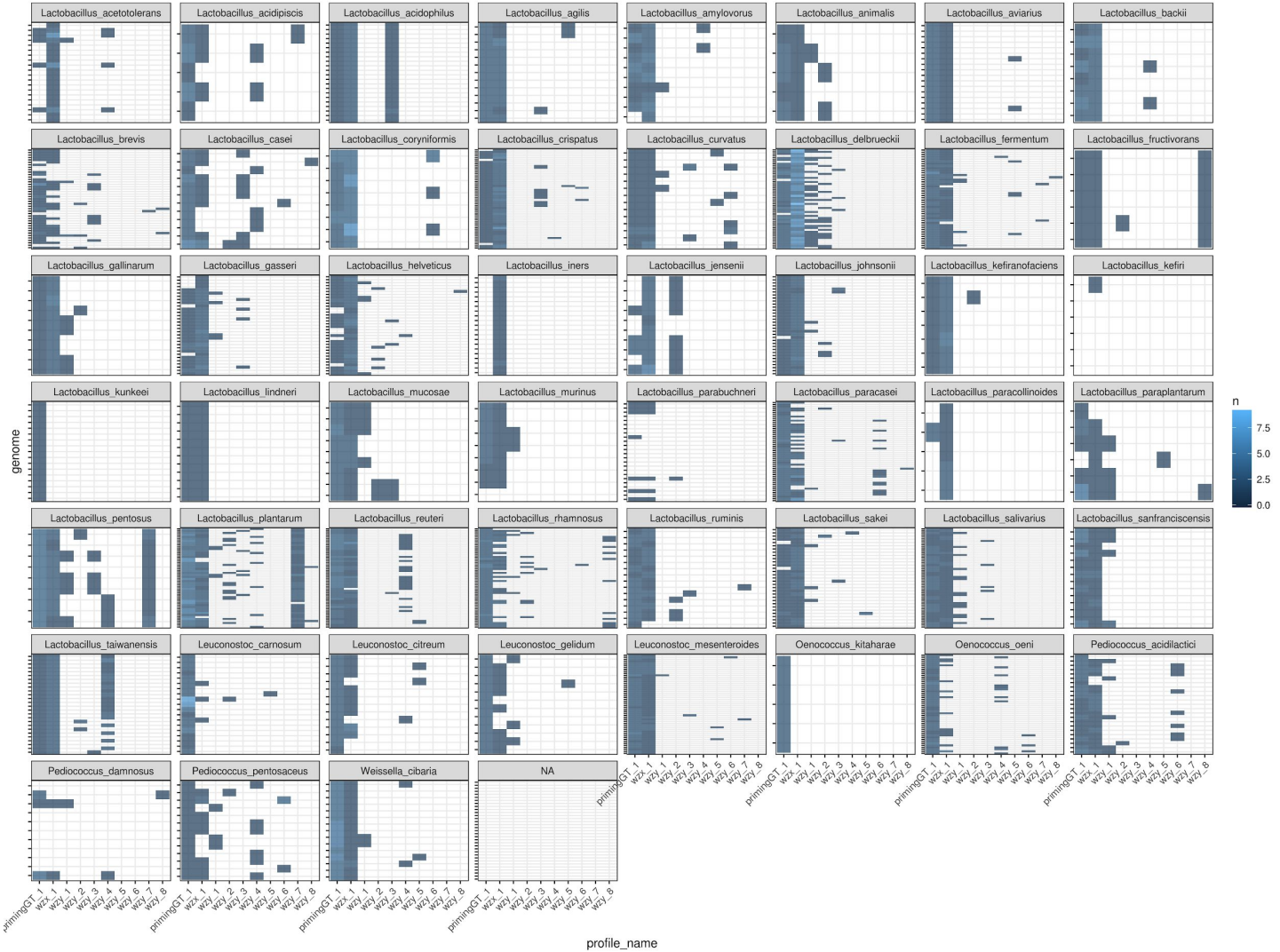
Hierarchical gene family clustering

Idea:

1. Cluster gene families independently for each species → pangenomes
2. Gather the pangenome of each species
3. Cluster gene families on these pangenomes

Why?

- Higher resolution species-level gene families
- Faster


Lactobacillus_acetotolerans
Lactobacillus_acidipiscis
Lactobacillus_acidophilus
Lactobacillus_agilis
Lactobacillus_amylovorus
Lactobacillus_animalis
Lactobacillus_aviarius
Lactobacillus_backii
Lactobacillus_brevis
Lactobacillus_casei
Lactobacillus_coryniformis
Lactobacillus_crispatus
Lactobacillus_curvatus
Lactobacillus_delbrueckii
Lactobacillus_fermentum
Lactobacillus_fructivorans
Lactobacillus_gallinarum
Lactobacillus_gasseri
Lactobacillus_helveticus
Lactobacillus_iners
Lactobacillus_jensenii
Lactobacillus_johnsonii
Lactobacillus_kefiranofaciens
Lactobacillus_kefiri
Lactobacillus_kunkeei
Lactobacillus_lindneri
Lactobacillus_mucosae
Lactobacillus_murinus
Lactobacillus_parabuchneri
Lactobacillus_paracasei
Lactobacillus_paracollinoides
Lactobacillus_paraplantarum
Lactobacillus_pentosus
Lactobacillus_plantarum
Lactobacillus_reuteri
Lactobacillus_rhamnosus
Lactobacillus_ruminis
Lactobacillus_sakei
Lactobacillus_salivarius
Lactobacillus_sanfranciscensis
Lactobacillus_taiwanensis
Leuconostoc_carnosum
Leuconostoc_citreum
Leuconostoc_gelidum
Leuconostoc_mesenteroides
Oenococcus_kitaharae
Oenococcus_oeni
Pediococcus_acidilactici
Pediococcus_damnosus
Pediococcus_pentosaceus
Weissella_cibaria
NA
genome
profile_name
n
7.5
5.0
2.5
0.0
primingGT_1
wzx_1
wzy_1
wzy_2
wzy_3
wzy_4
wzy_5
wzy_6
wzy_7
wzy_8