Computational Challenges in a Densely Sequenced Tree of Life

Katie Pollard

Gladstone Institutes

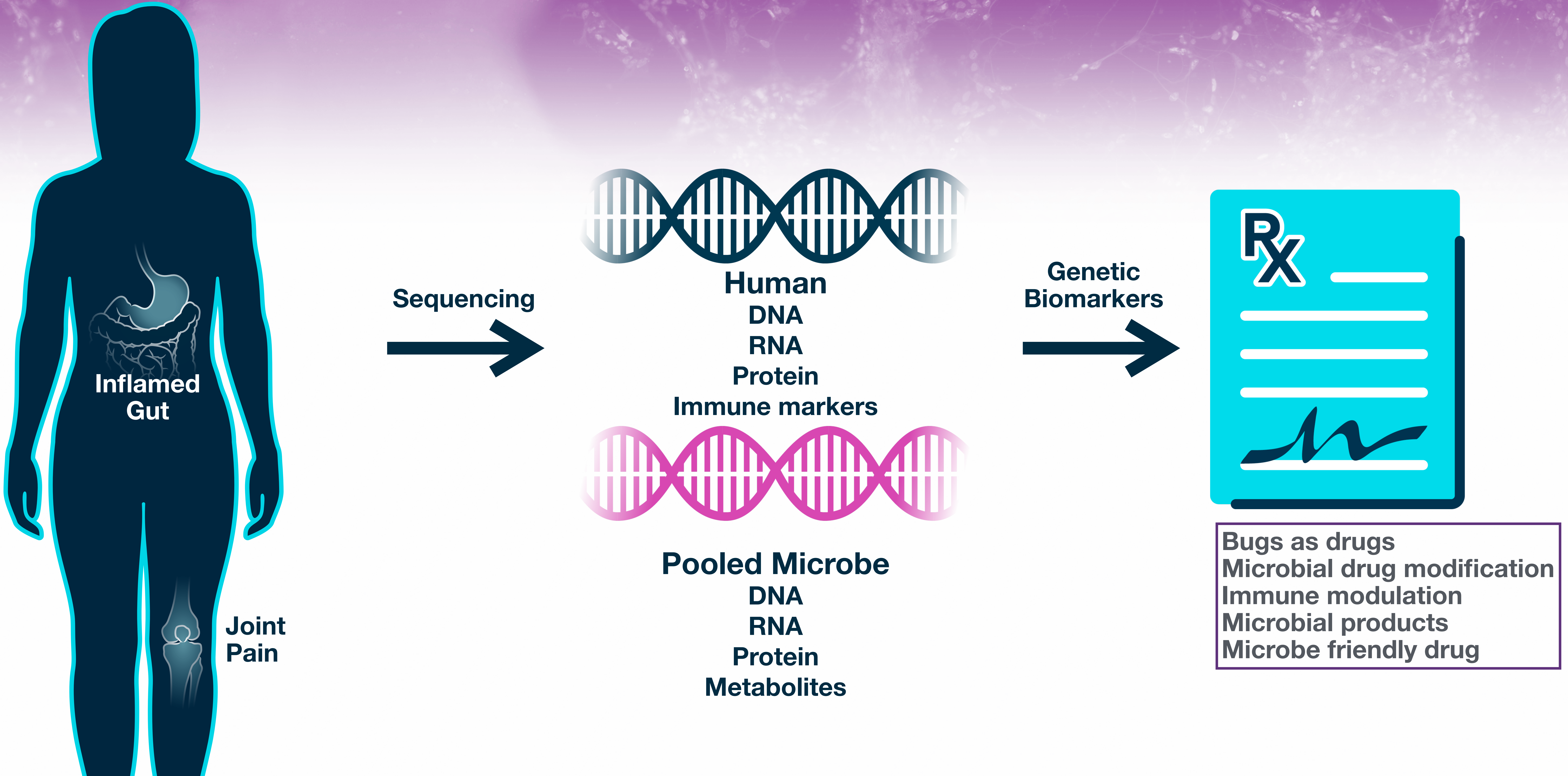
Chan Zuckerberg Biohub

UC San Francisco

Computational Challenges in Very Large-Scale 'Omics Workshop

July 21, 2022 • Simons Institute for the Theory of Computing

Microbiome Precision Medicine



Inflamed Gut

Joint Pain

Sequencing

Human
DNA
RNA
Protein
Immune markers

Pooled Microbe
DNA
RNA
Protein
Metabolites

Genetic Biomarkers

Rx

Bugs as drugs
Microbial drug modification
Immune modulation
Microbial products
Microbe friendly drug

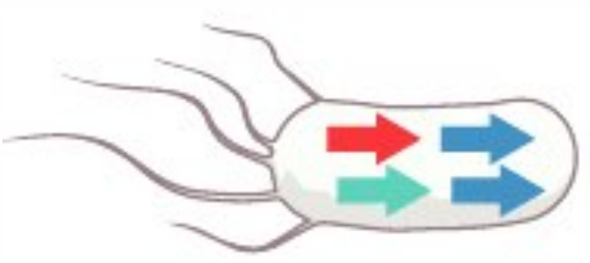
A very large-scale ‘omics problem

ACTGATG
CATCGAT
ATGCTAC
GATCGAT
CGATCTT
ATCGAAG

50 million sequences
300 bp each
from 100s of species mixed

↕

Code to search for matches



ATGCATC
GATCTAC
GATCGAT
TTCGATC
AAATCGA

~300K genomes
~5 million bp each

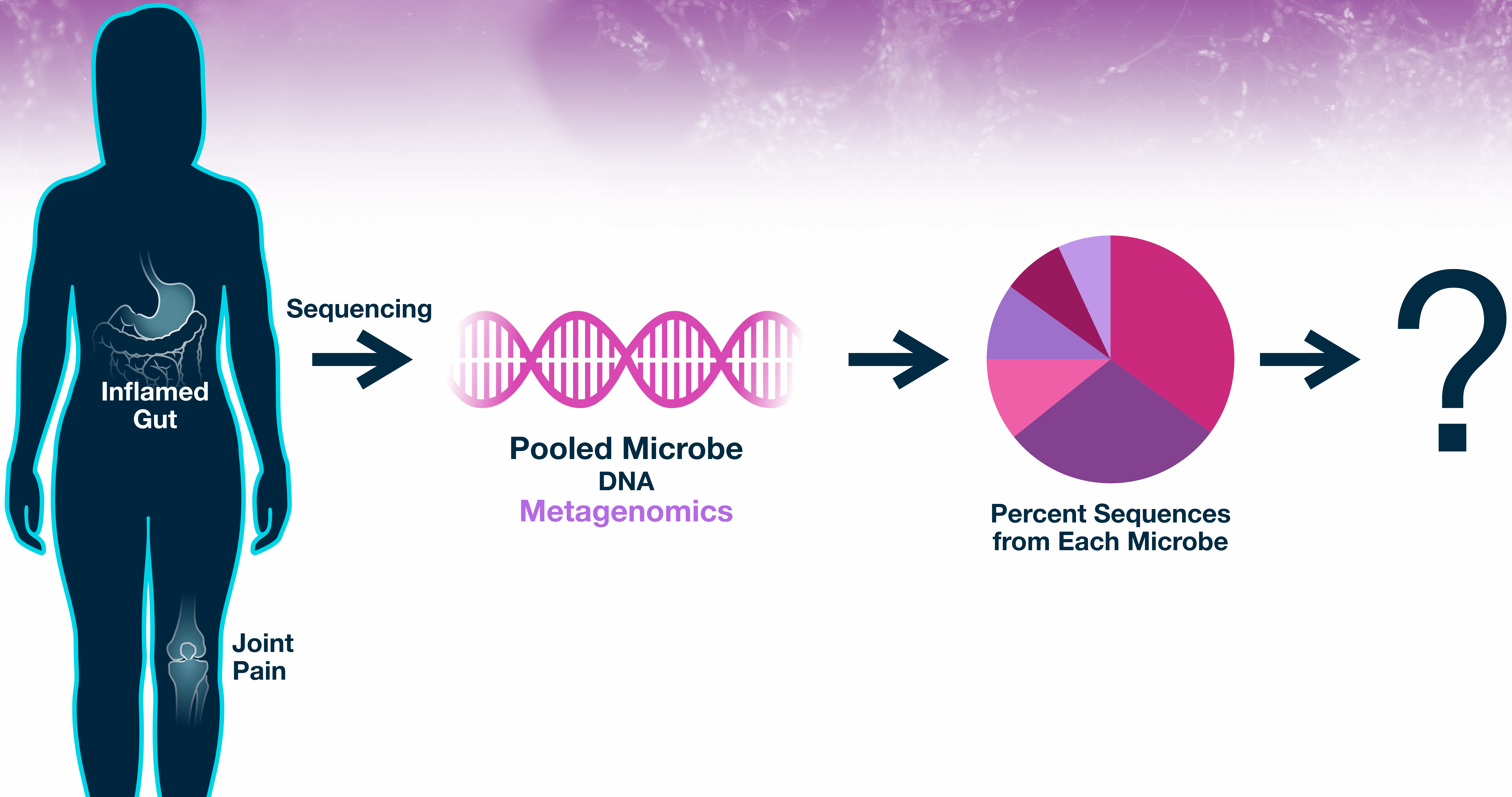
Problems we solved

- 50% of species have no genome: <10% now
- Code takes years to run or costs $10K/month in cloud: runs on laptop
- Closely related species confound alignment and bias downstream statistics: mitigated

Re: our AWS bill
“Let's call up
Jeff Bezos and
talk about this”
Oct 2019

Getty Images

Microbiome Science



Inflamed Gut

Joint Pain

Sequencing

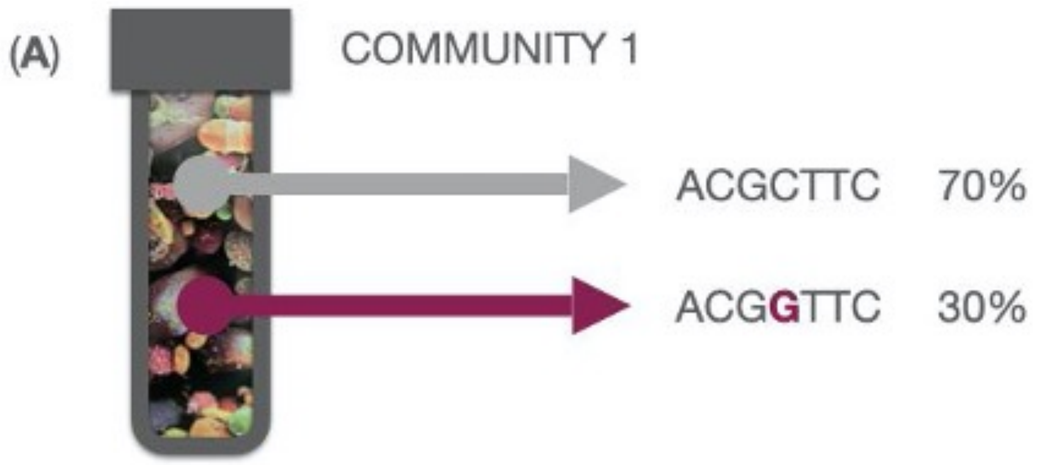
Pooled Microbe DNA

Metagenomics

Percent Sequences from Each Microbe

?

Metagenotyping single nucleotide variants (SNVs)



Similar approach for gene copy number variants (CNVs)

Zhao et al. (2022)

Using Genetic Variation

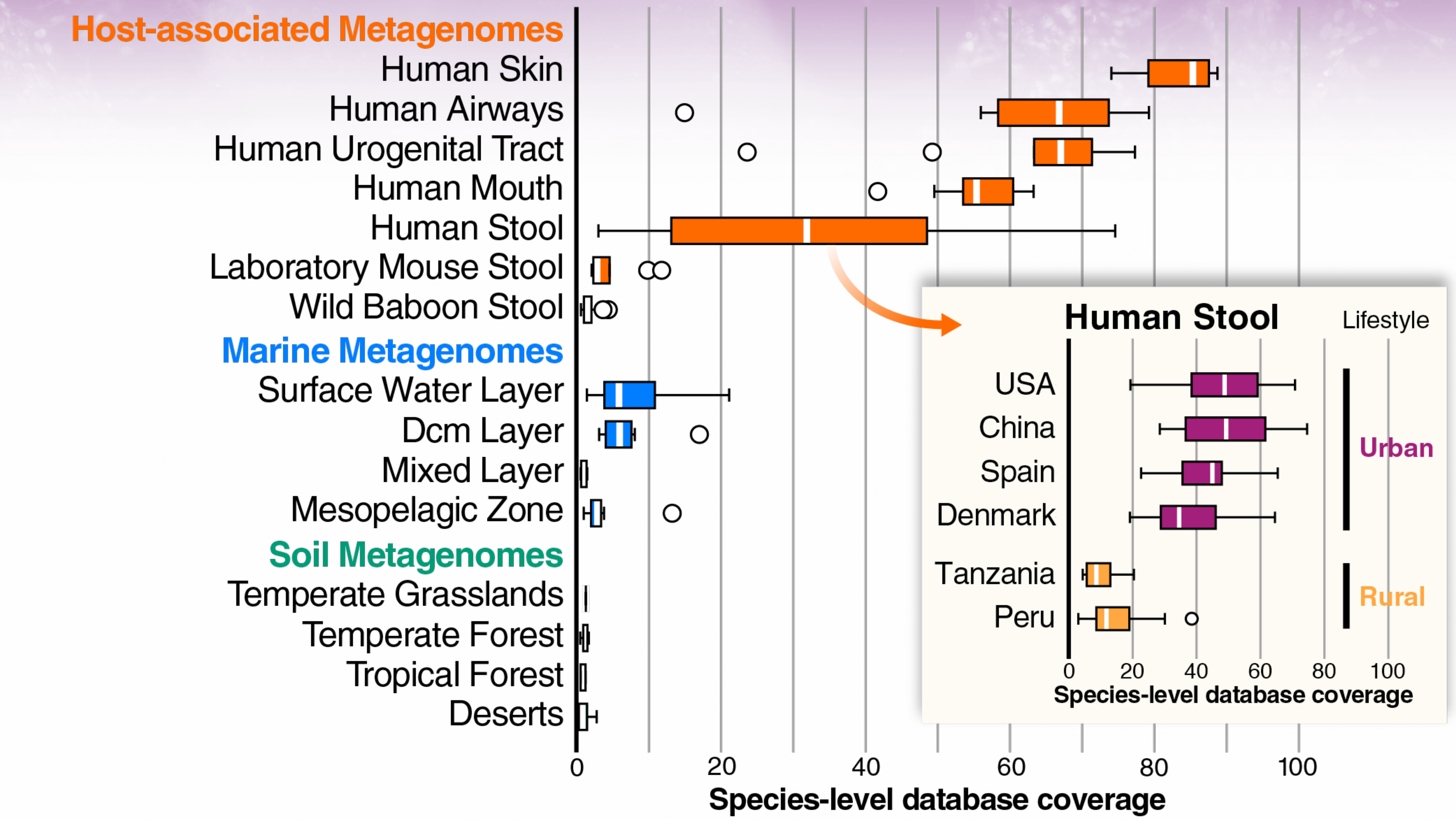
- **Phenotype associations**
  - human traits
  - microbe traits
- **Microbiome evolution**
  - mutation
  - selection
  - recombination
  - demography / ancestry
- **Strain / gene tracking**
- **Human evolution**
- **Genomic technologies**
- **Precision therapies**
- **Clinical decision making**

Garud & Pollard (2019)

Challenge 0:

Species without a genome in the database are invisible

Most species had no genome

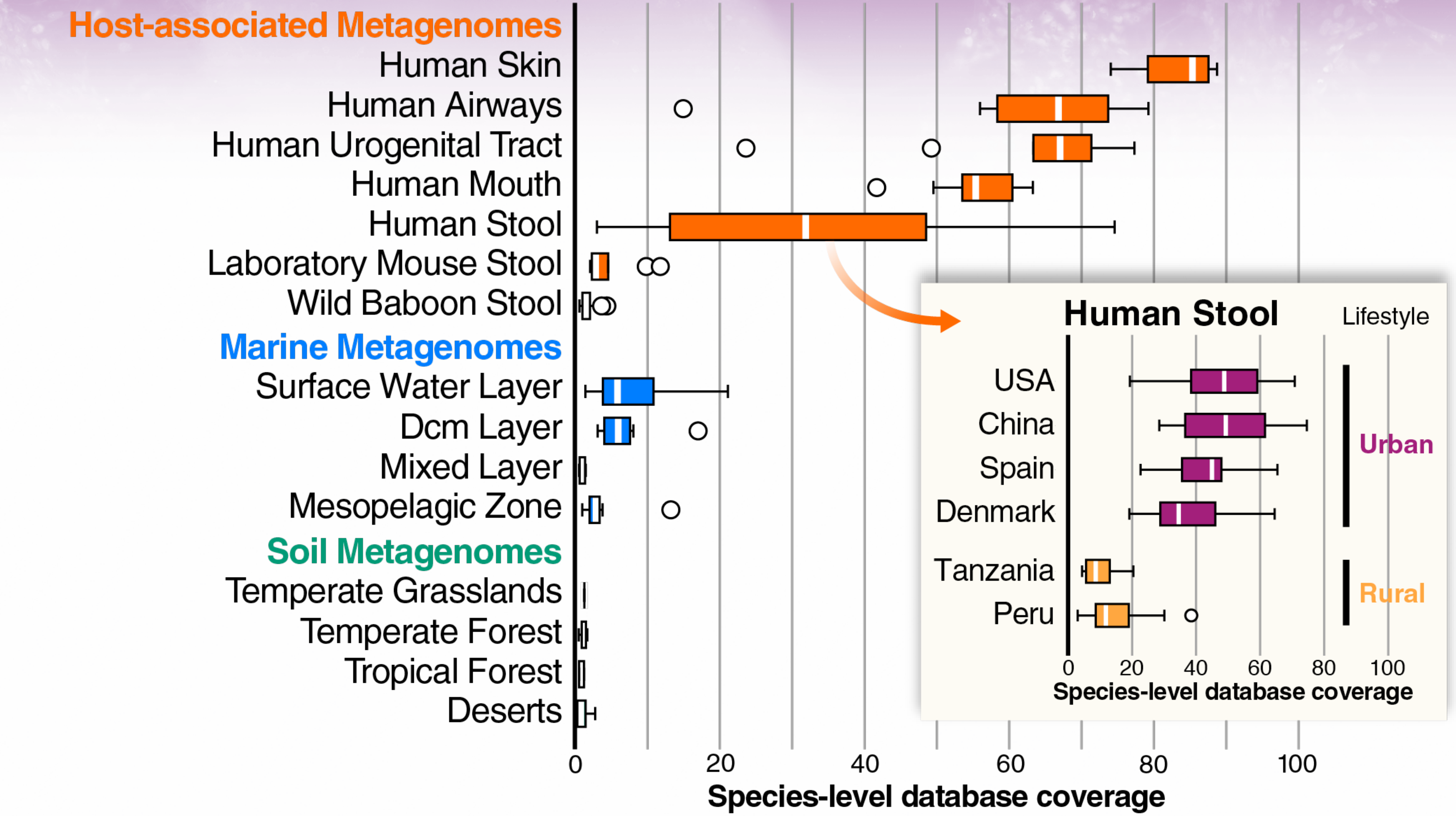


Nayfach et. al (2016)

Data: ~8K metagenomes from SRA, EBI, JGI

Analysis: MicrobeCensus (github/snayfach/MicrobeCensus/)

But this is changing

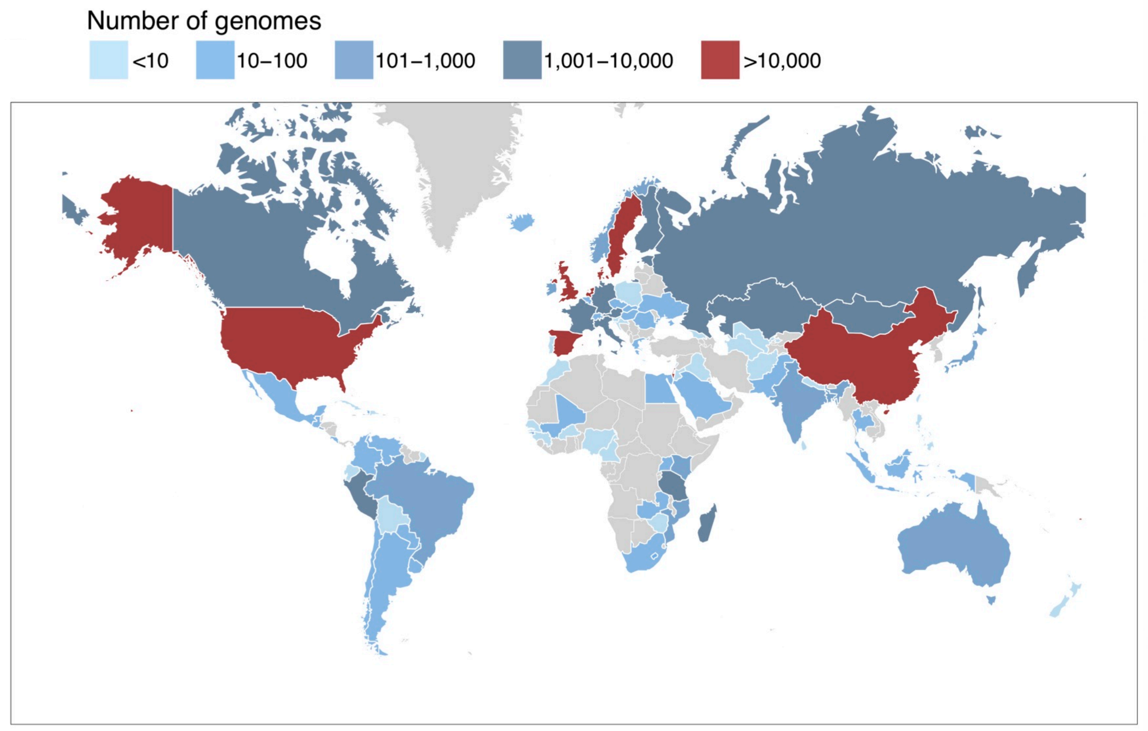


Nayfach et. al (2016)

Data: ~8K metagenomes from SRA, EBI, JGI
Analysis: MicrobeCensus (github/snayfach/MicrobeCensus/)

UHGG Resource

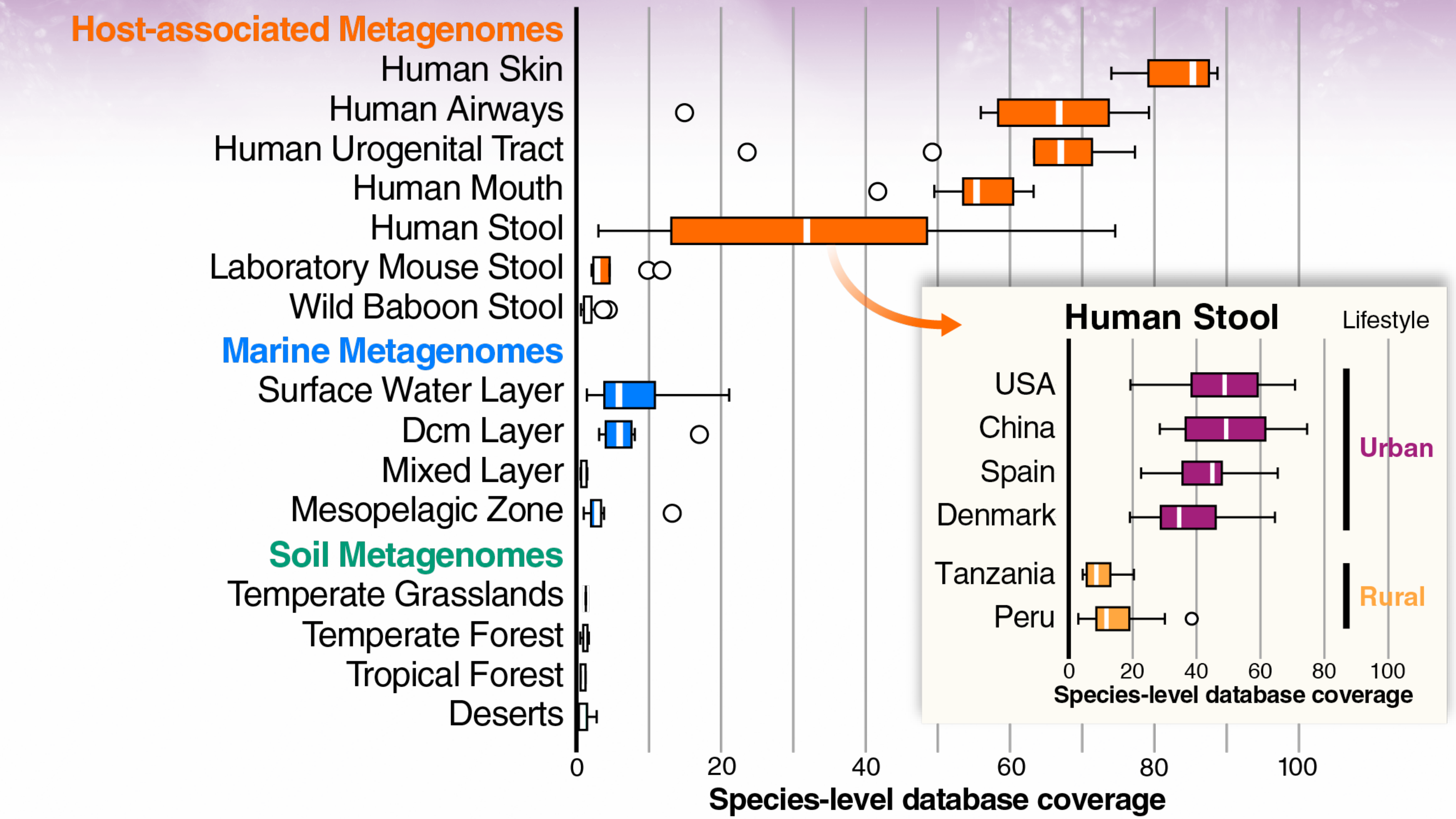
Shotgun metagenomes
- 31 countries, 6 continents
- Different lifestyles & ages



286,799 gut genomes
4,644 species
81% of species MAG-only
50% increase in diversity
>2K disease associations

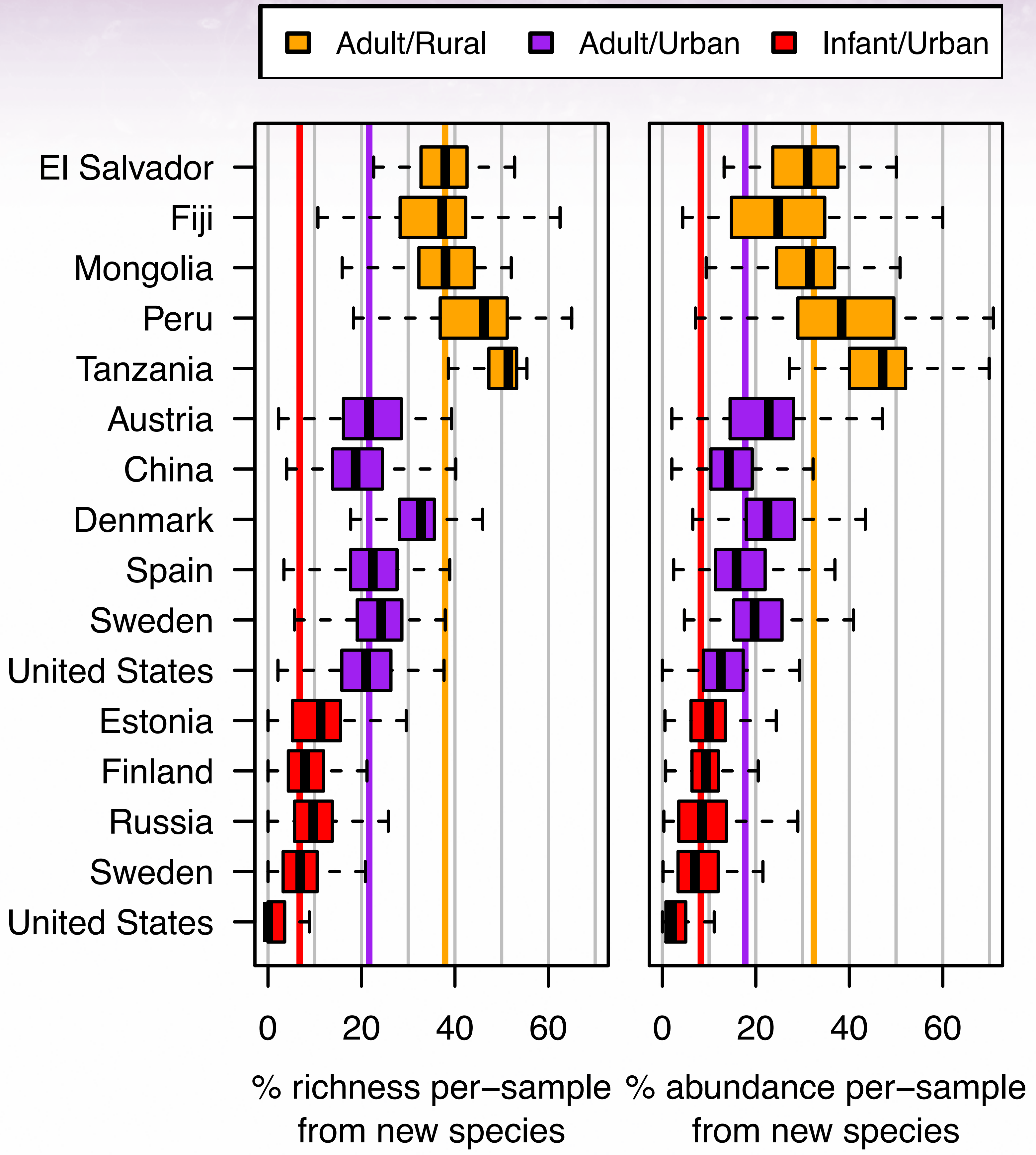
Nayfach et al (2019)
Almeida et al (2019)
Also: Culturomics, single-cell

But this is changing



Nayfach et. al (2016)

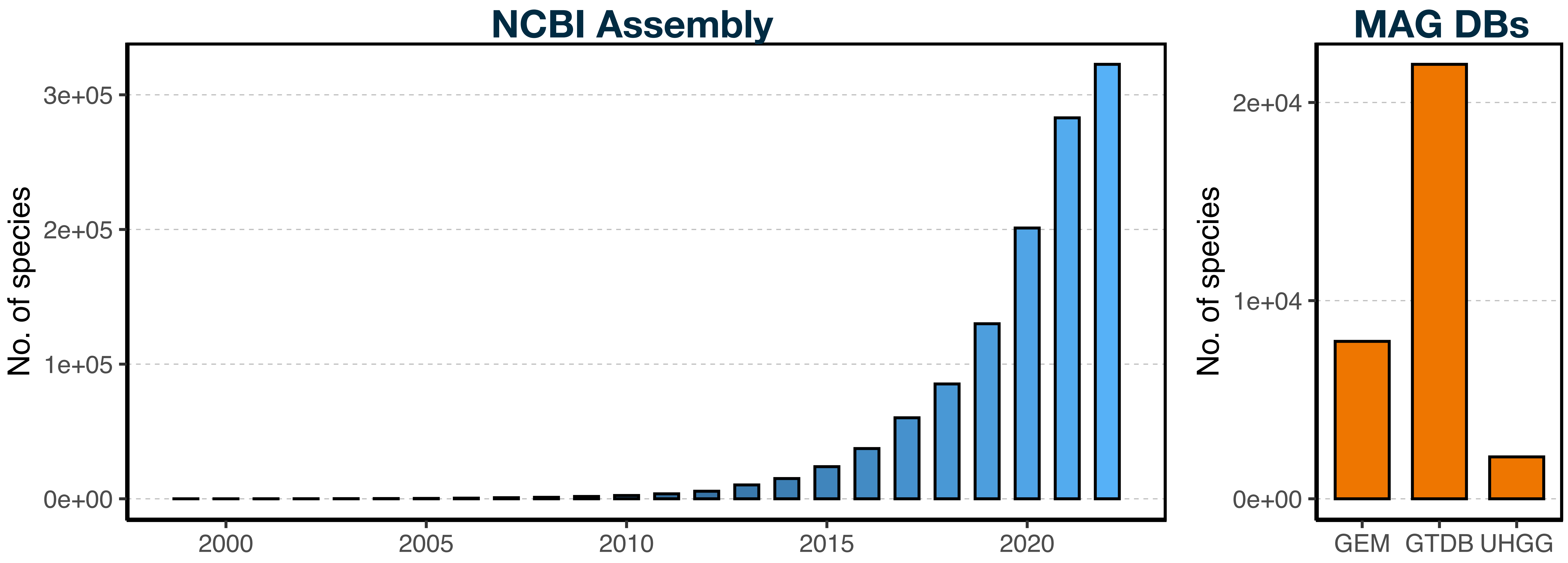
MAGs are closing the gap



Nayfach et. al (2019)
Almeida et al (2019)

Data: ~8K metagenomes from SRA, EBI, JGI
Analysis: MicrobeCensus (github/snayfach/MicrobeCensus/)

Genome explosion



Zhao et. al (2022b)

More Genomes = Good News?

Human gut microbiome alignment rate now > 80%

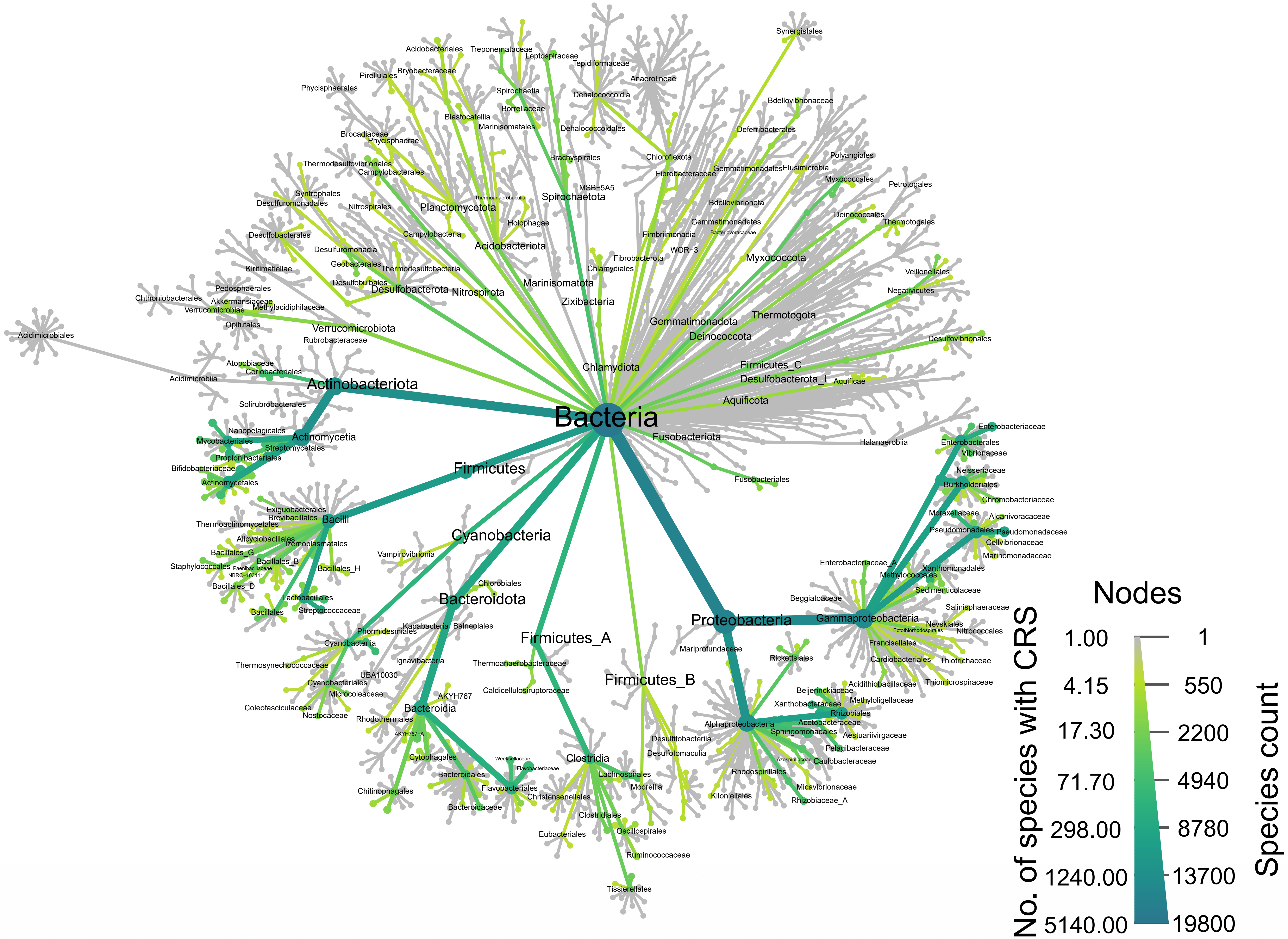
But… new problems arise

Challenge 1:

Closely related species “compete” for reads and bias metagenotypes

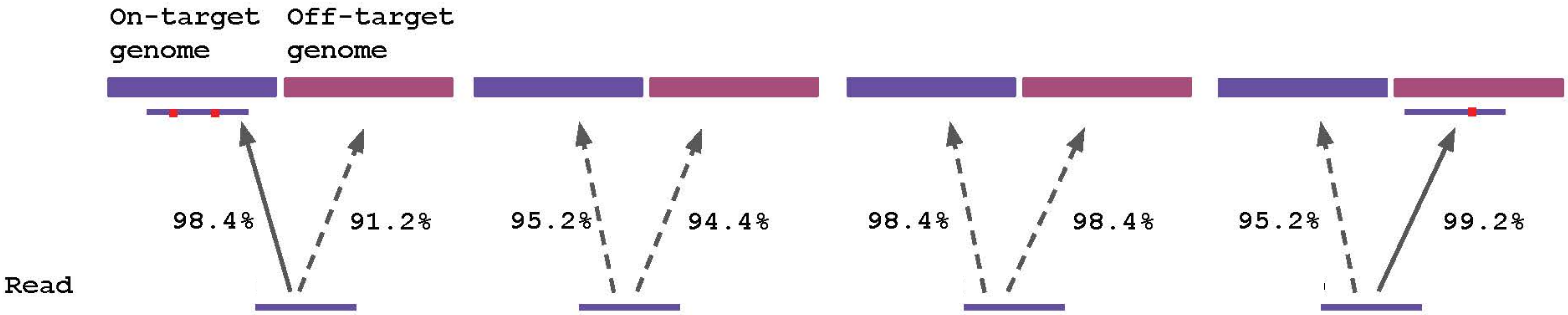
Closely related species are common

CRS = two species with at least one pair of genomes that have average nucleotide identity (ANI) 92%-95%



Zhao et. al (2022b)

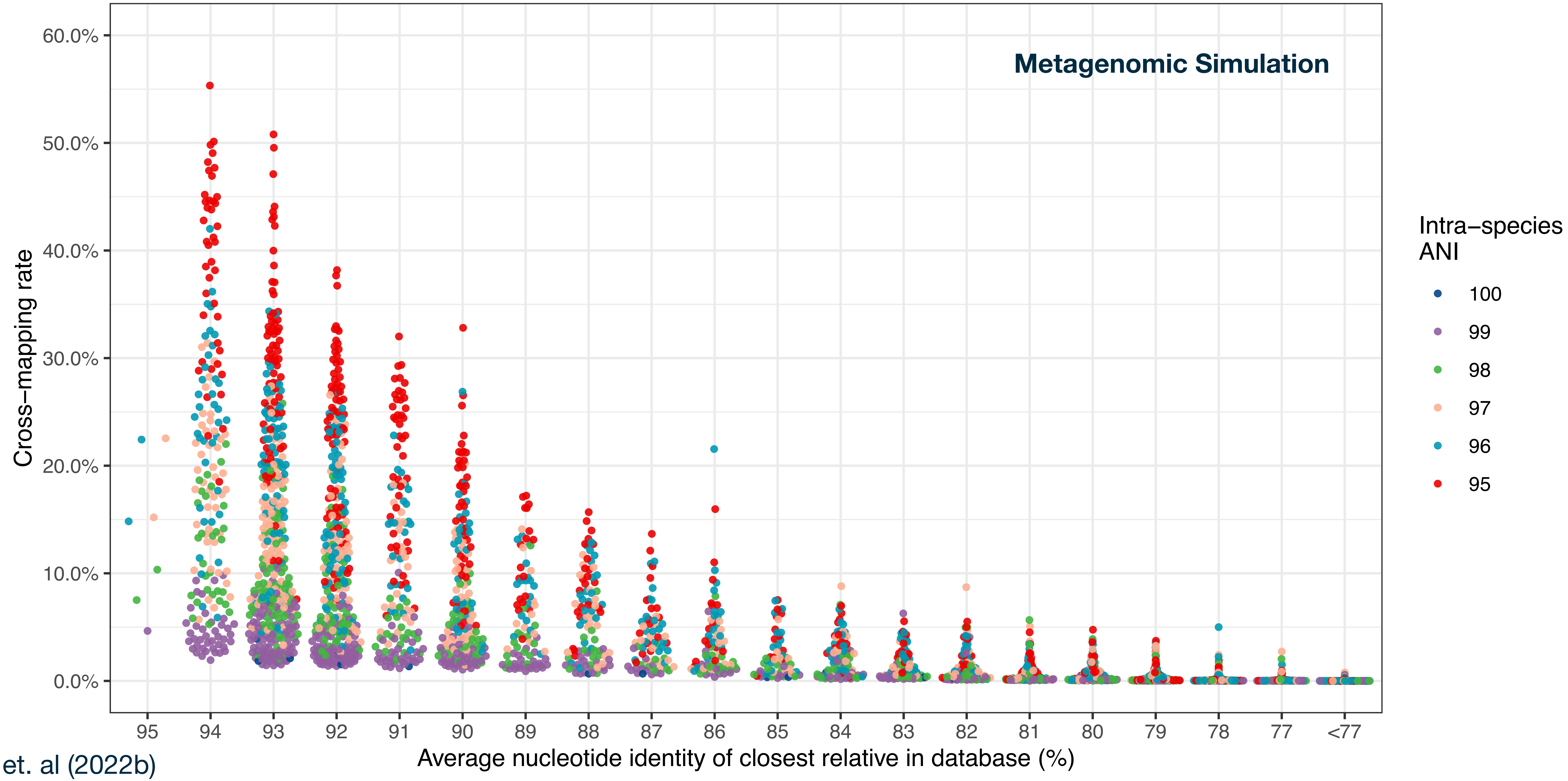
Read competition in dense lineages



	Aligned w/ low uniqueness	Unaligned multi-mapping	Unaligned multi-mapping	Aligned cross-mapping
Alignment similarity	98.4%	95.2%	98.4%	99.2%
Alignment uniqueness	Medium	Low	Low	Medium
Probability correctly aligned	Medium	Low	Low	Medium

Zhao et. al (2022b)

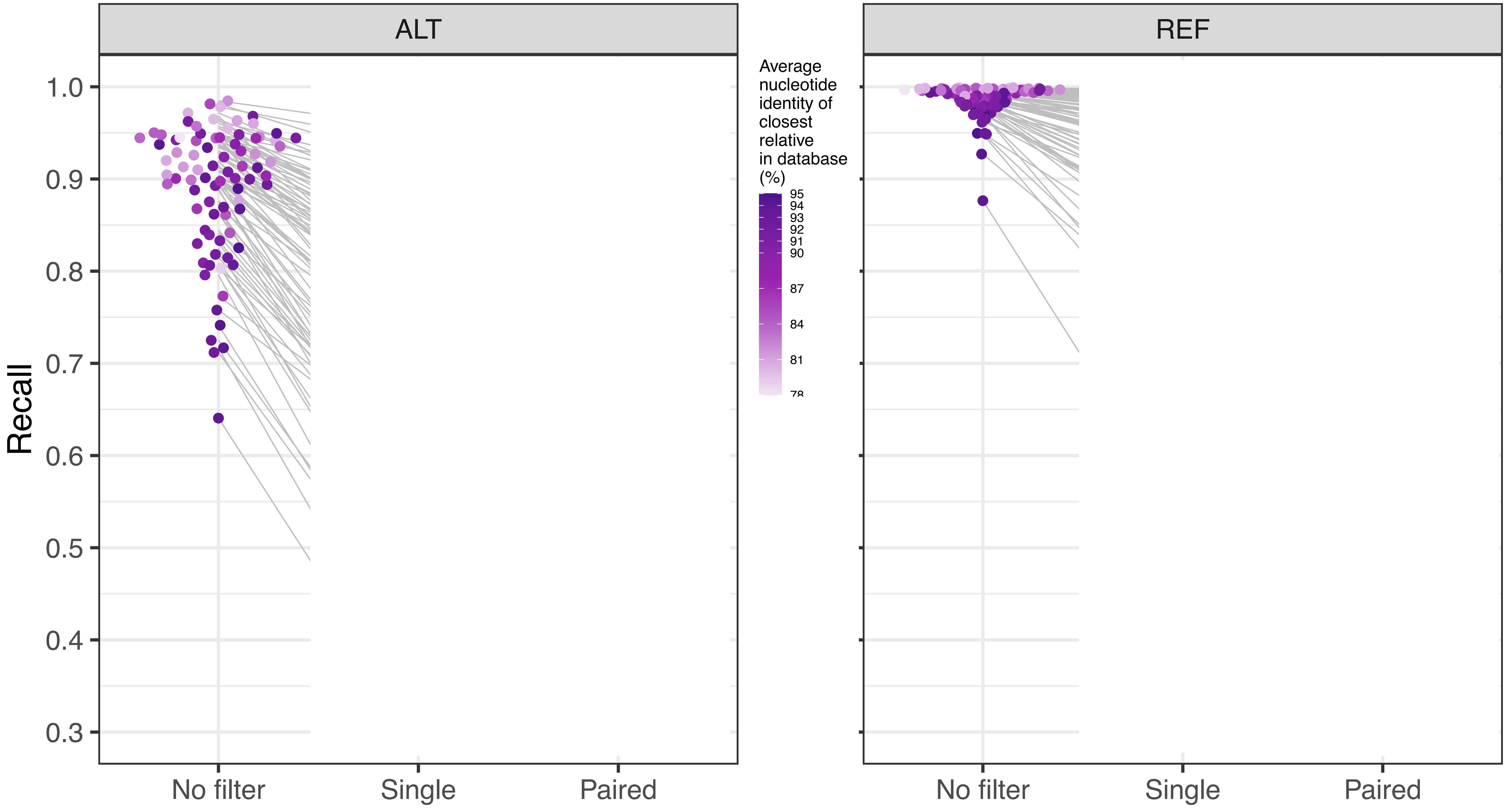
Read competition in dense lineages



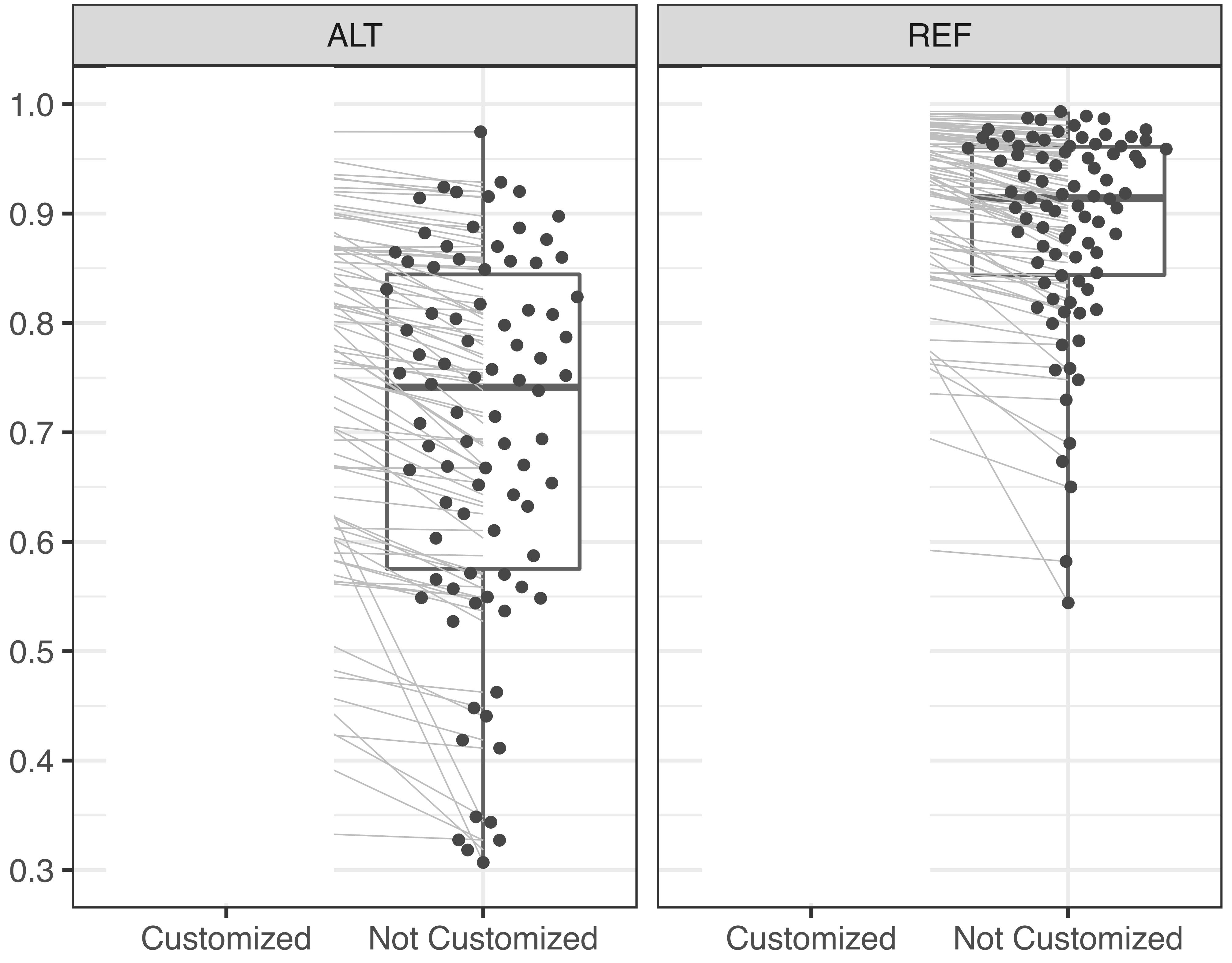
Zhao et. al (2022b)

MIDAS2 mitigates low alignment uniqueness & cross-mapping

Paired-end filtering, MAPQ<10



Dropping undetected species from db



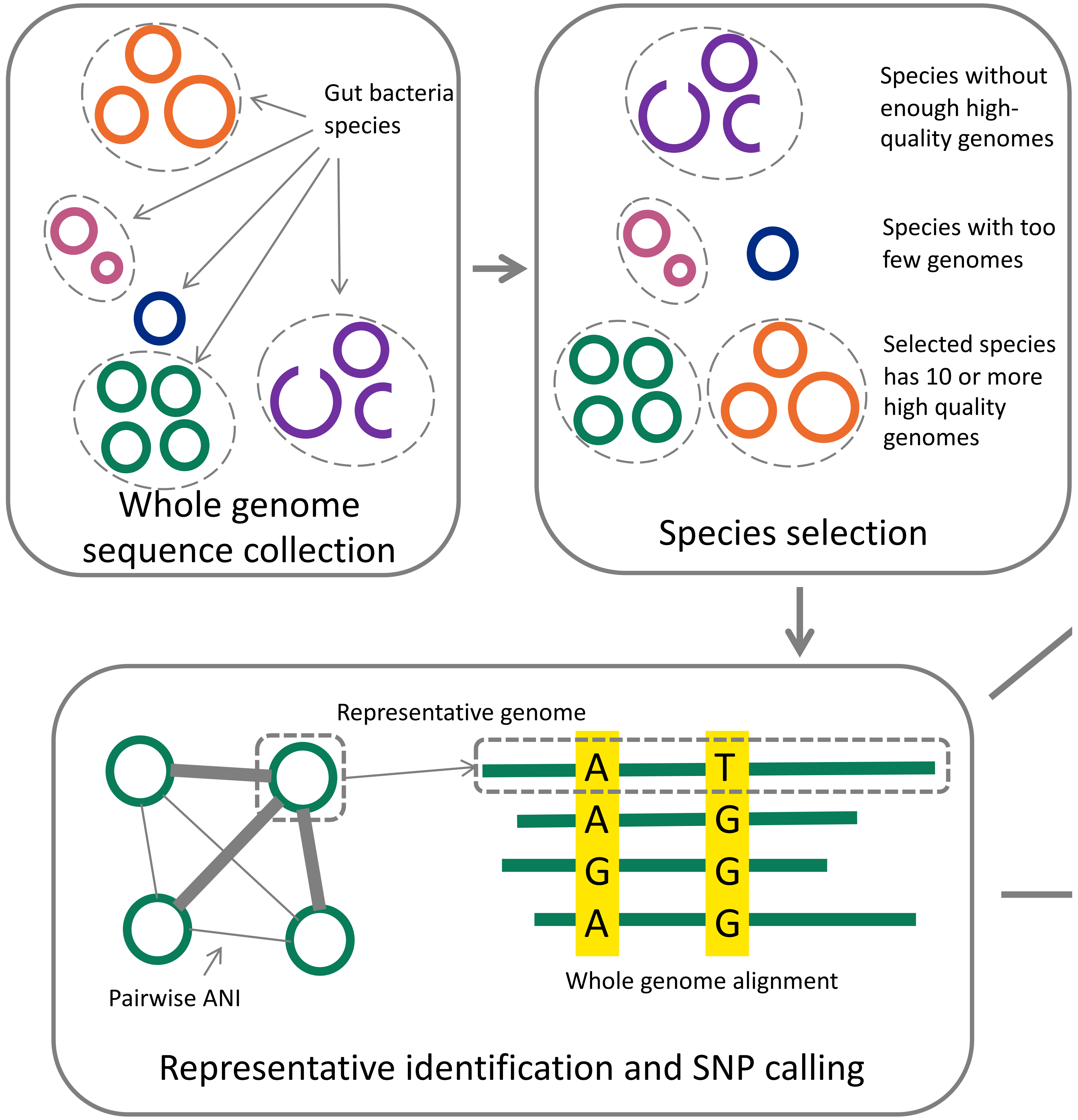
Zhao et. al (2022a)

https://github.com/czbiohub/MIDAS2

Mitigation strategies help…

But can we do better by avoiding alignment?

GT-PRO strategy works for metagenomes, genomes, contigs, unassembled reads


Gut bacteria species
Whole genome sequence collection
Species without enough high-quality genomes
Species with too few genomes
Selected species has 10 or more high quality genomes
Species selection
Representative genome
A
A
G
A
T
G
G
G
Pairwise ANI
Whole genome alignment
Representative identification and SNP calling

Maast

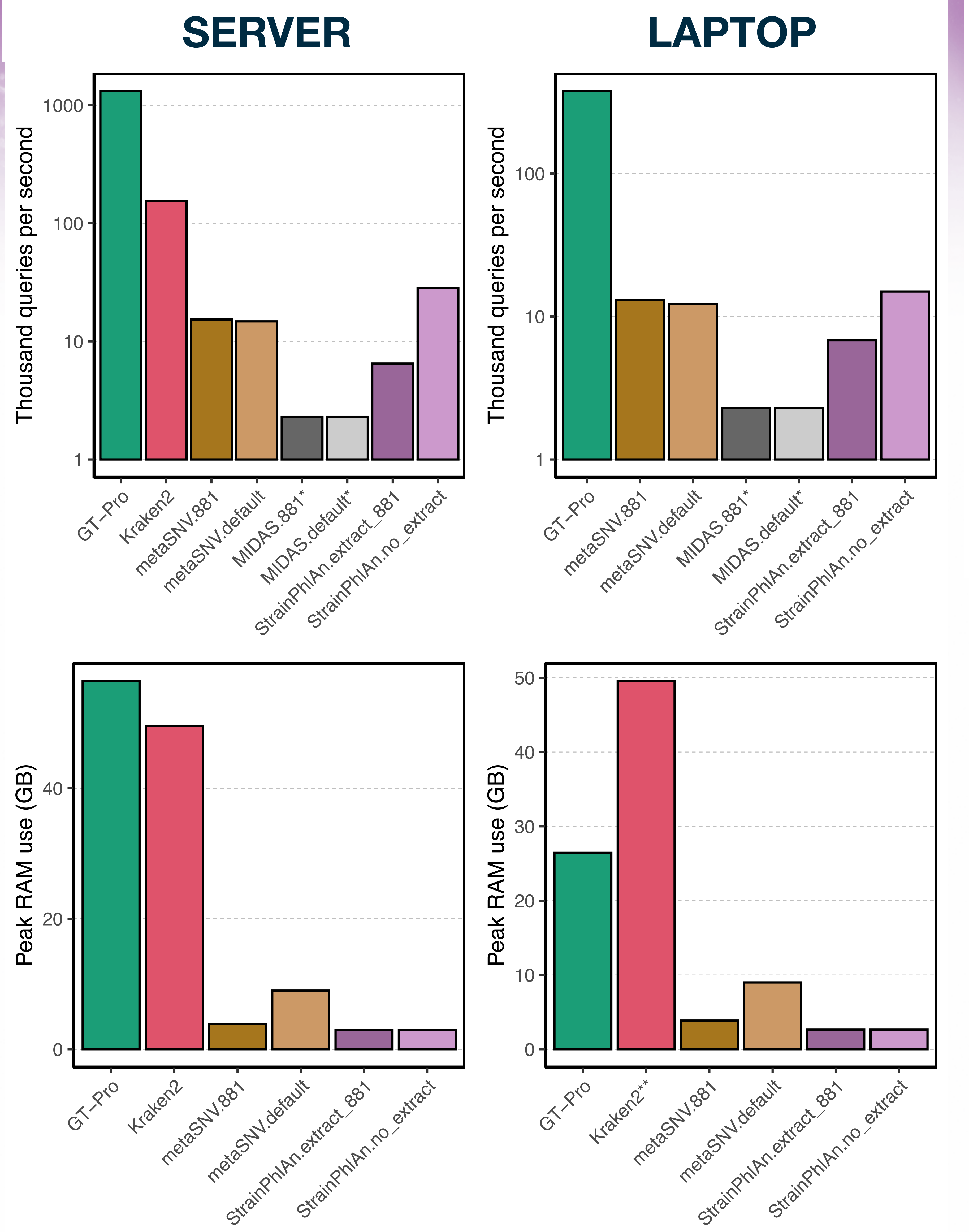
Shi et al. (2021)

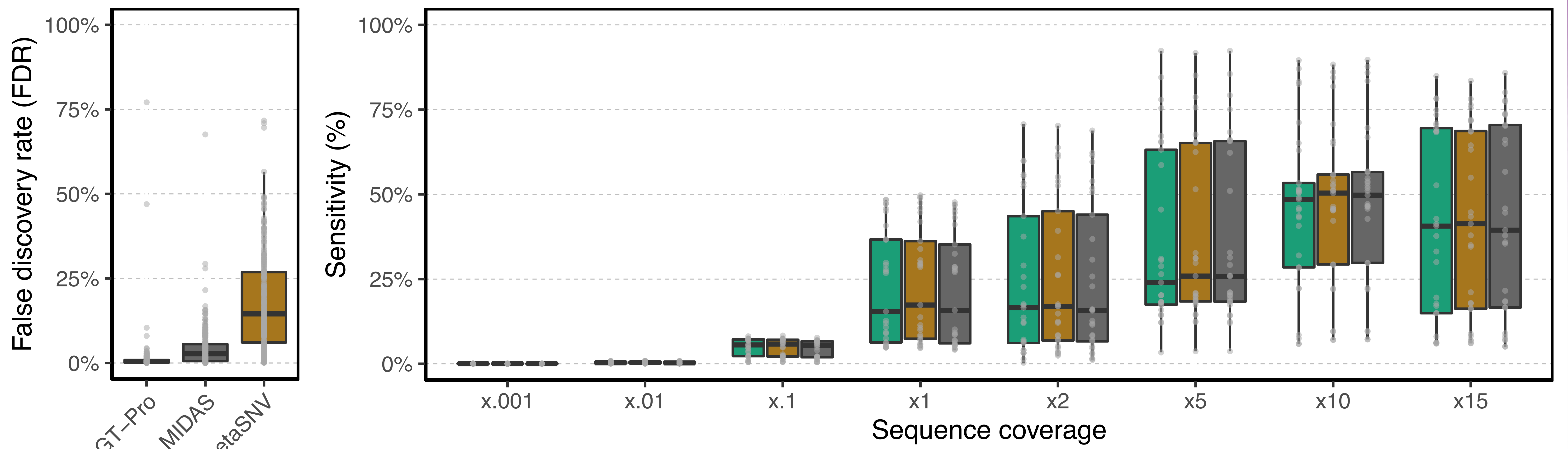
Compression > bzip2, rapid exact matching

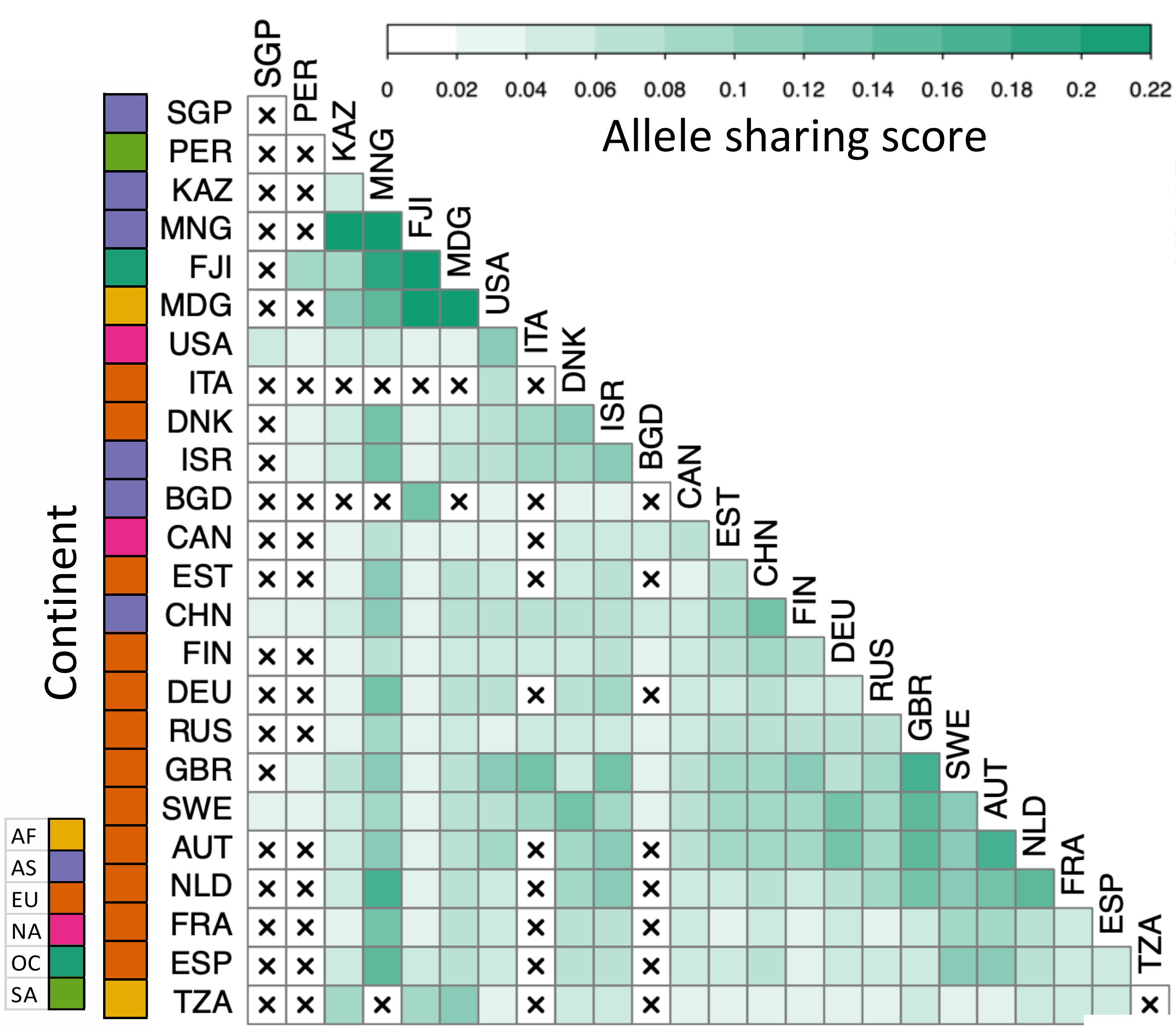
Prefix filter, Suffix array, Colex sort

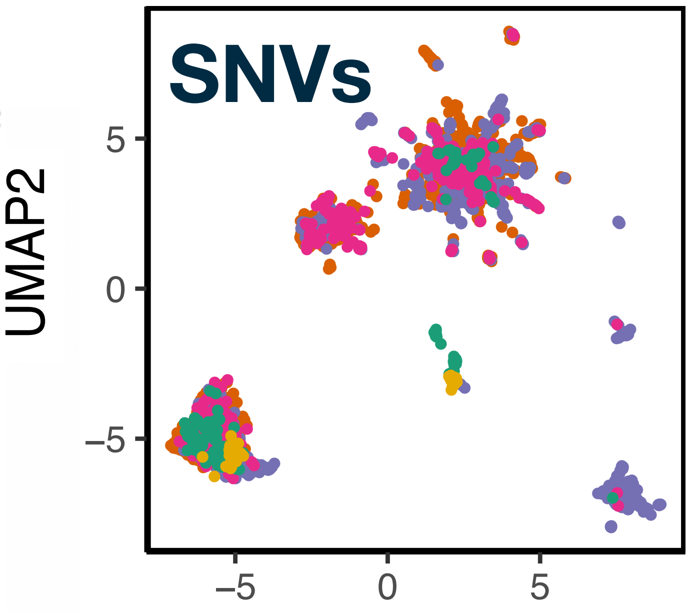
https://github.com/zjshi/gt-pro
https://github.com/zjshi/Maast

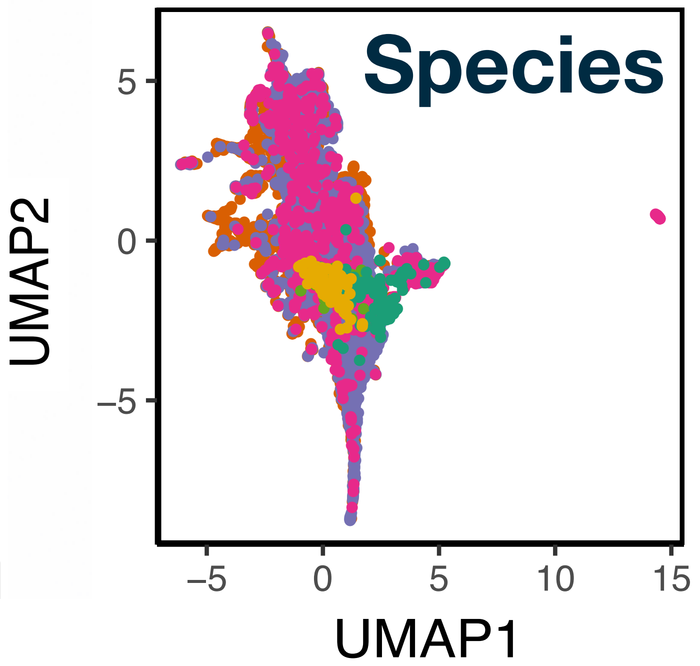
GTPRO: 100x faster, more accurate











Metagenome simulations with varying sequencing coverage

Global stool samples (N=7,459)

Shi et al. (2021)

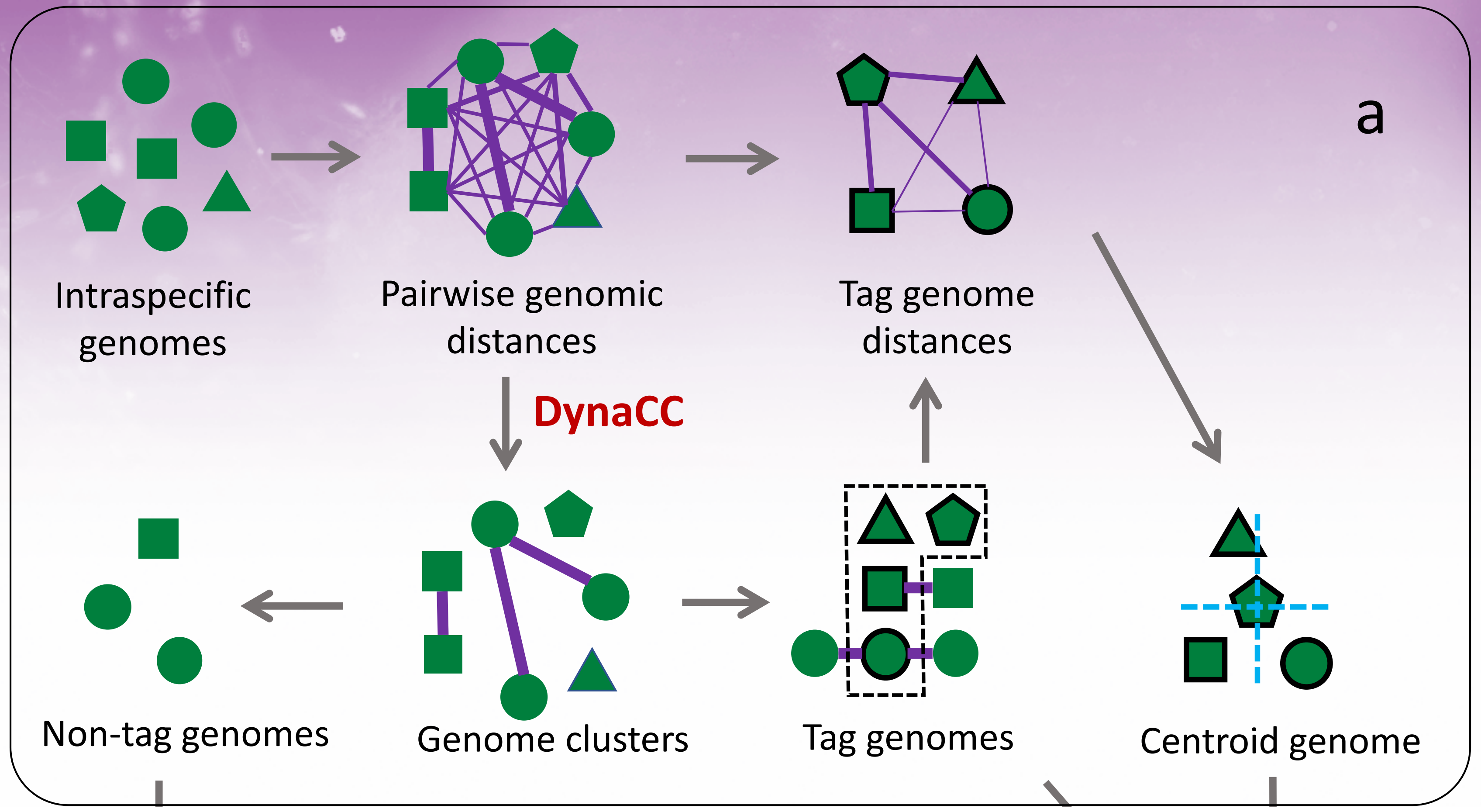
Unique k-mers beat alignment at known SNVs

But current approach only works on SNVs discovered in genomes

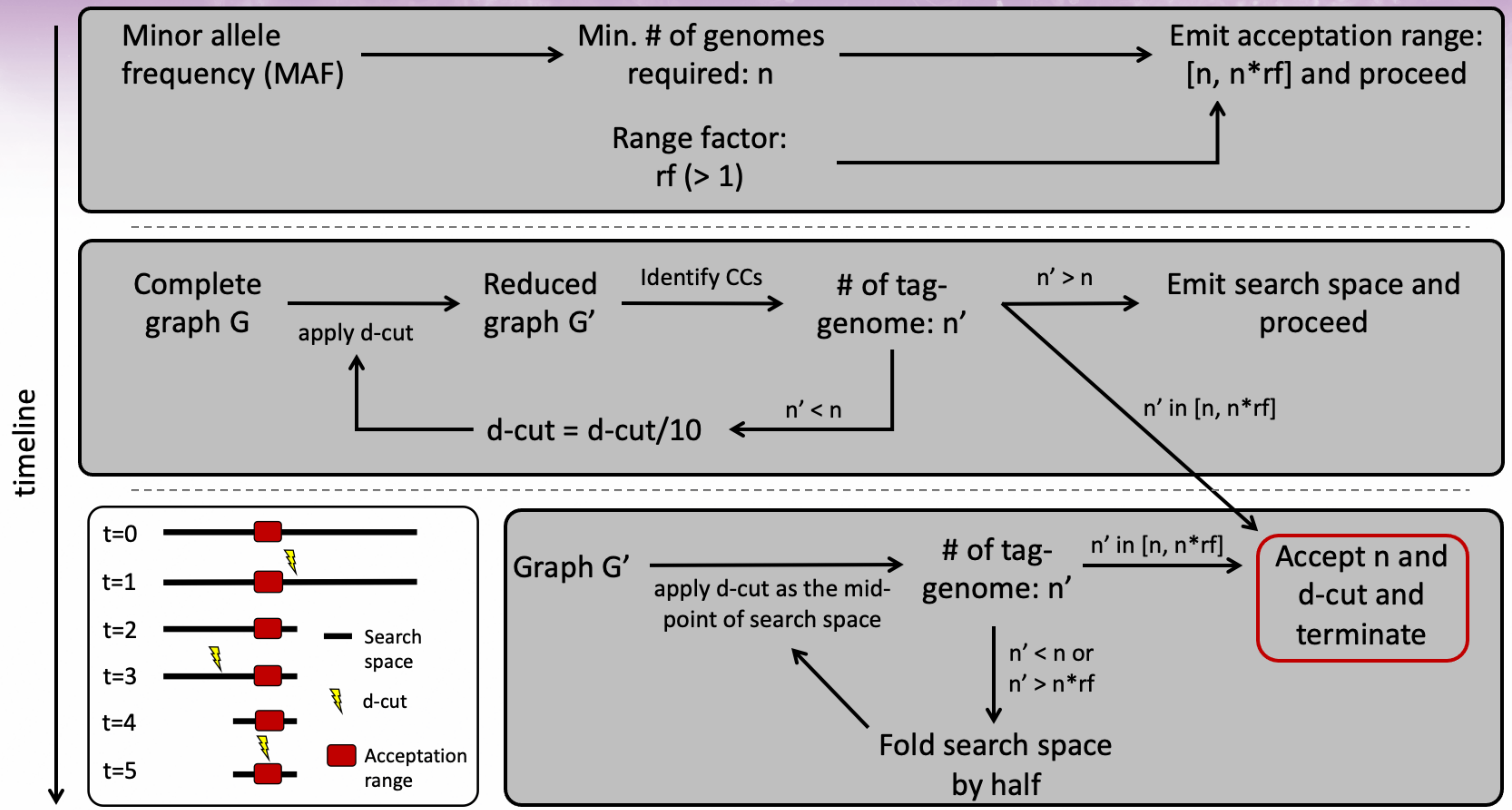
Challenge 2:

How to align and call variants in so many genomes?

Maast: fast variant discovery from genomes



DynaCC algorithm flowchart



timeline

Minor allele frequency (MAF) → Min. # of genomes required: n → Emit acceptation range: [n, n*rf] and proceed

Range factor: rf (> 1)

Complete graph G → apply d-cut → Reduced graph G' → Identify CCs → # of tag-genome: n'

n' > n → Emit search space and proceed

n' < n → d-cut = d-cut/10

n' in [n, n*rf]

t=0
t=1
t=2
t=3
t=4
t=5

Search space

d-cut

Acceptation range

Graph G' → apply d-cut as the mid-point of search space → # of tag-genome: n'

n' in [n, n*rf] → Accept n and d-cut and terminate

n' < n or n' > n*rf → Fold search space by half

https://github.com/zjshi/Maast

Shi et al. (2022)

Genome redundancy offers solution

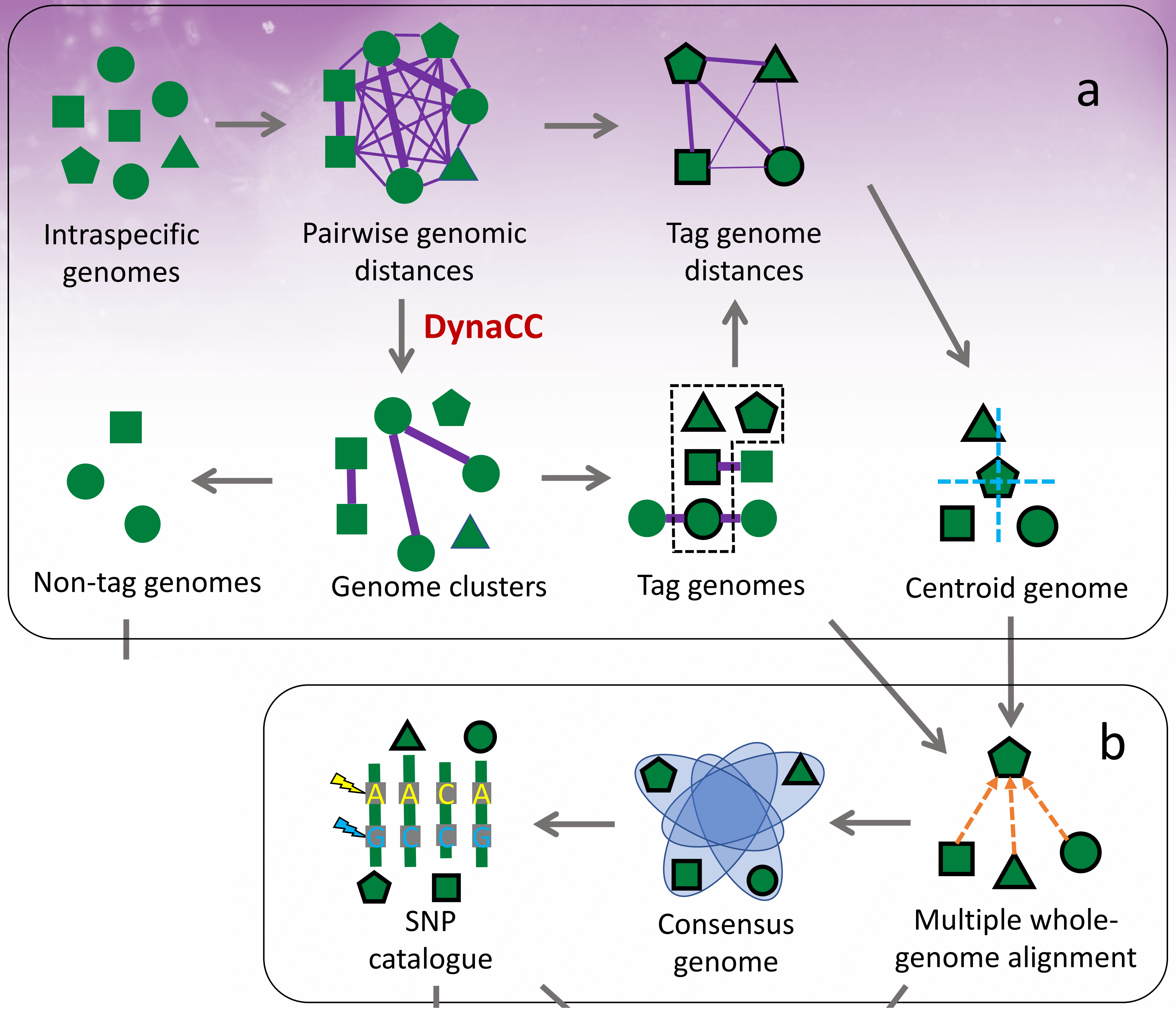
No. of clusters
0
1000
2000
3000
No. of genomes
6000
4000
2000
0
Species
100003
100018
100022
100033
100036
100038
100041
100044
100045
100060
100074
100076
100078
100084
100085
100089
100099
100113
100122
100144
100148
100154
100177
100185
100189
100190
100196
100205
100208
100209
100217
100231
100242
100243
100244
100249
100251
100253
100254
100258
100262
100271
100362
100412
100484
100530
100562
100606
100646
100723
100725
100730
100743
101029
101292
101294
101300
101302
101306
101319
101337
101338
101345
101346
101349
101359
101361
101365
101374
101378
101380
101397
101400
101410
101415
101440
101493
101694
101711
101737
101830
101866
102021
102057
102163
102172
102284
102292
102293
102311
102321
102327
102336
102353
102369
102371
102395
102396
102422
102438
102445
102453
102454
102455
102469
102478
102482
102492
102506
102517
102528
102530
102538
102540
102545
102549
102570
102580
102616
102715
102831
103017
103022
103411
103439
103656
103681
103682
103683
103686
103693
103694
103702
103891
103937
104006
104117
104245
104256
104317
104323
104337
104345
104526
104533

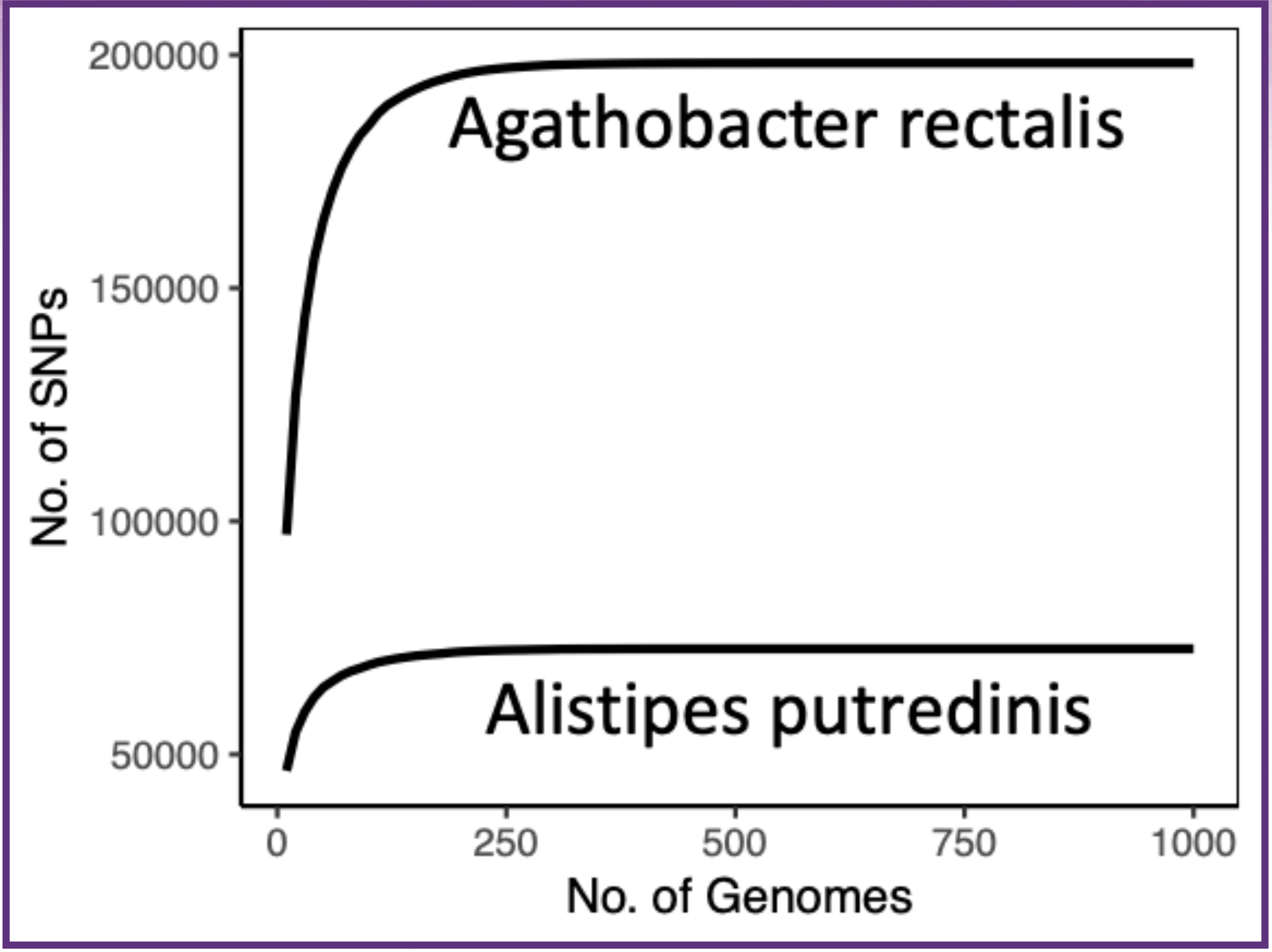
Alistipes putredinis
3,646 genomes
46 clusters

Agathobacter rectalis
5,214 genomes
2,650 clusters

Shi et al. (2022)

Maast: fast variant discovery from genomes



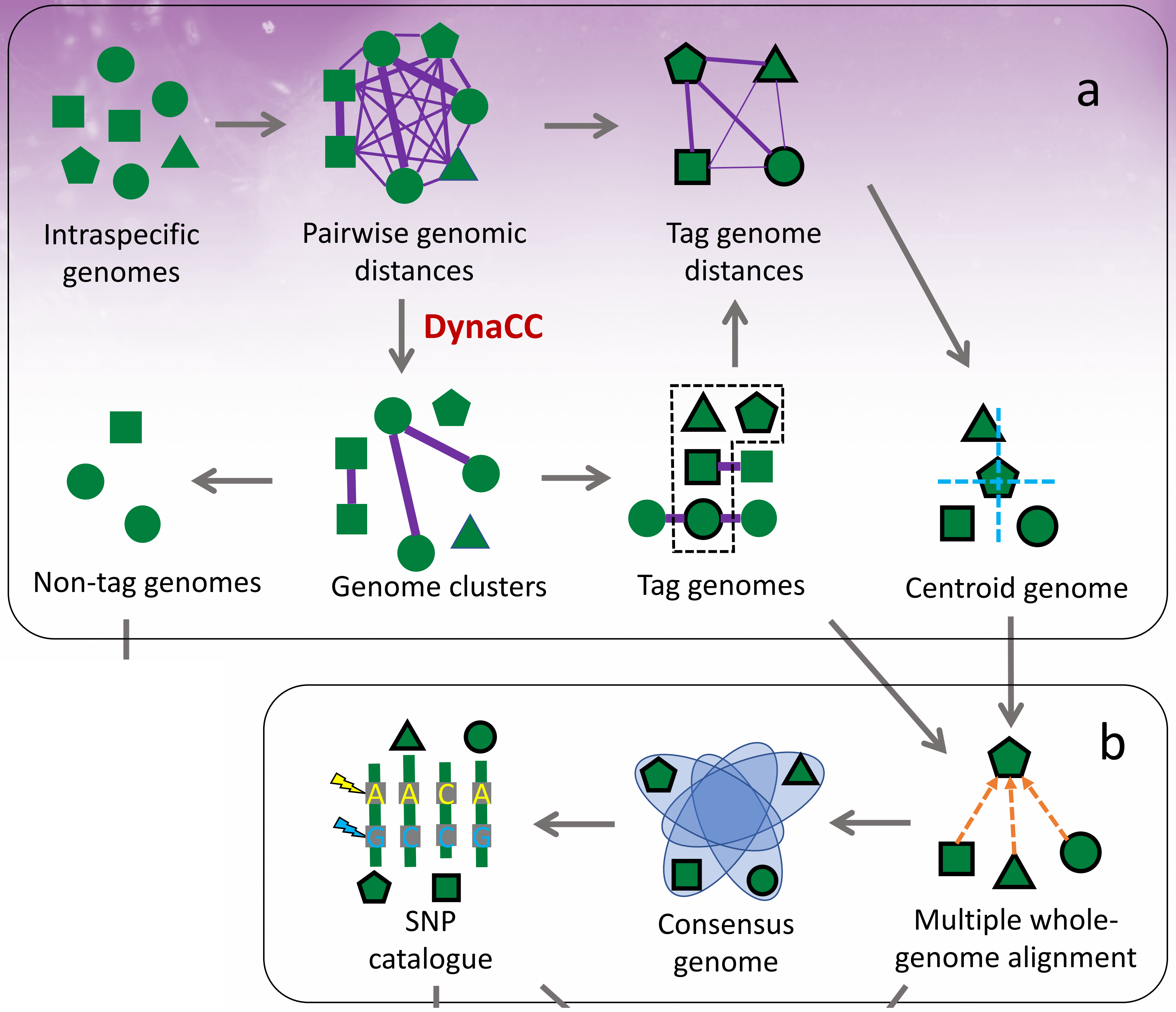


2,650 clusters

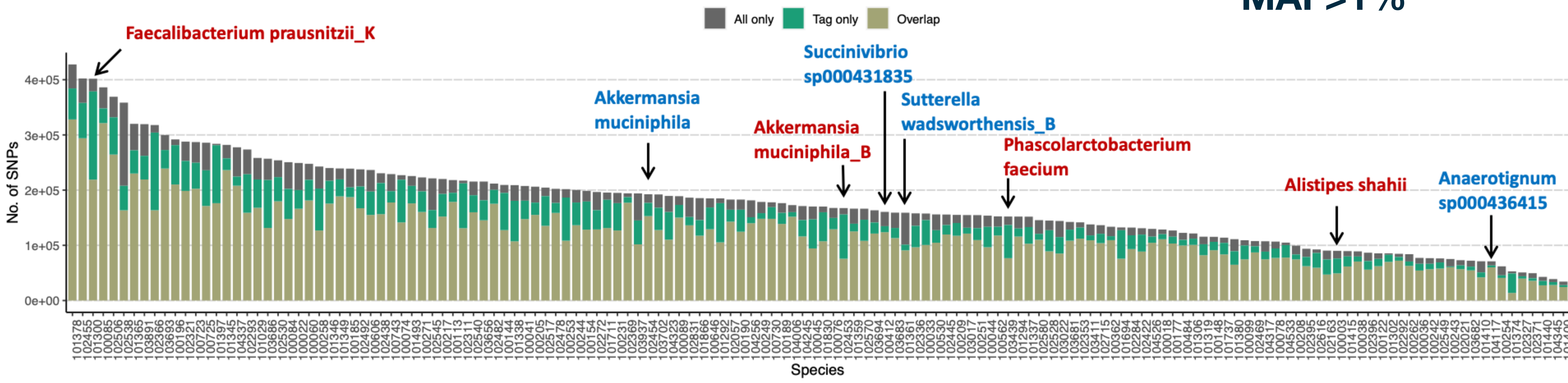
46 clusters

https://github.com/zjshi/Maast Shi et al. (2022)

Maast: fast variant discovery from genomes

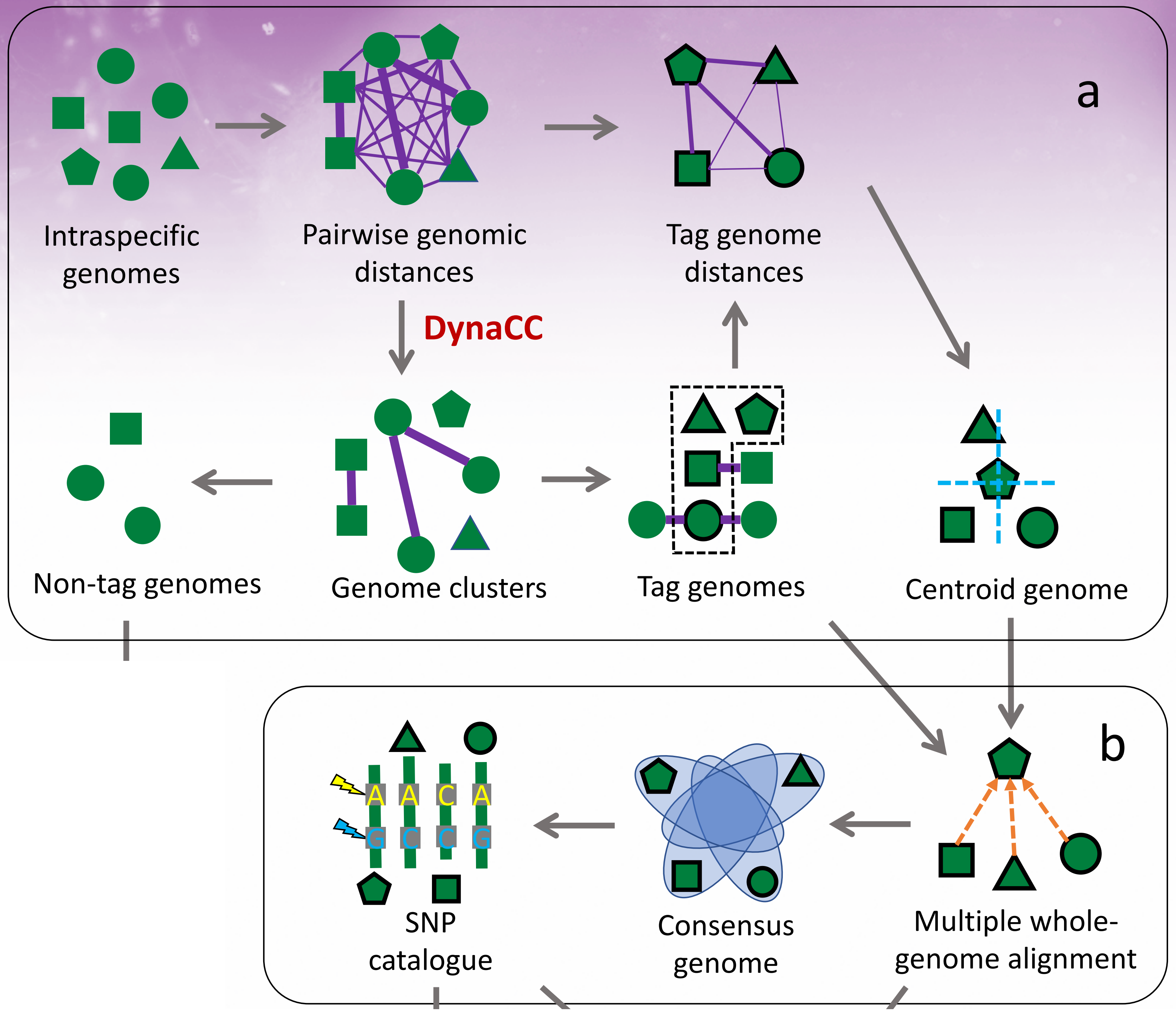


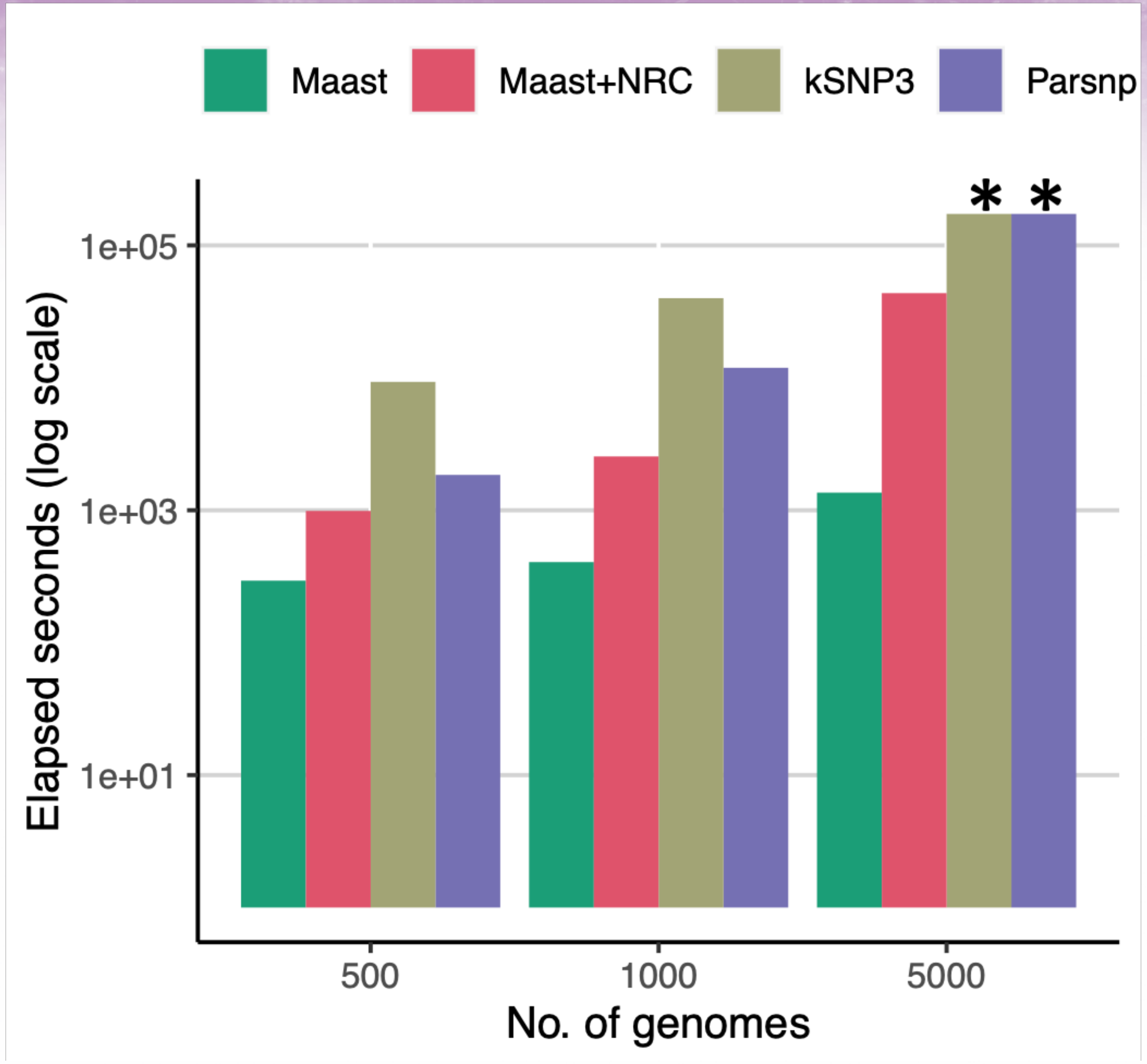
MAF>1%



https://github.com/zjshi/Maast Shi et al. (2022)

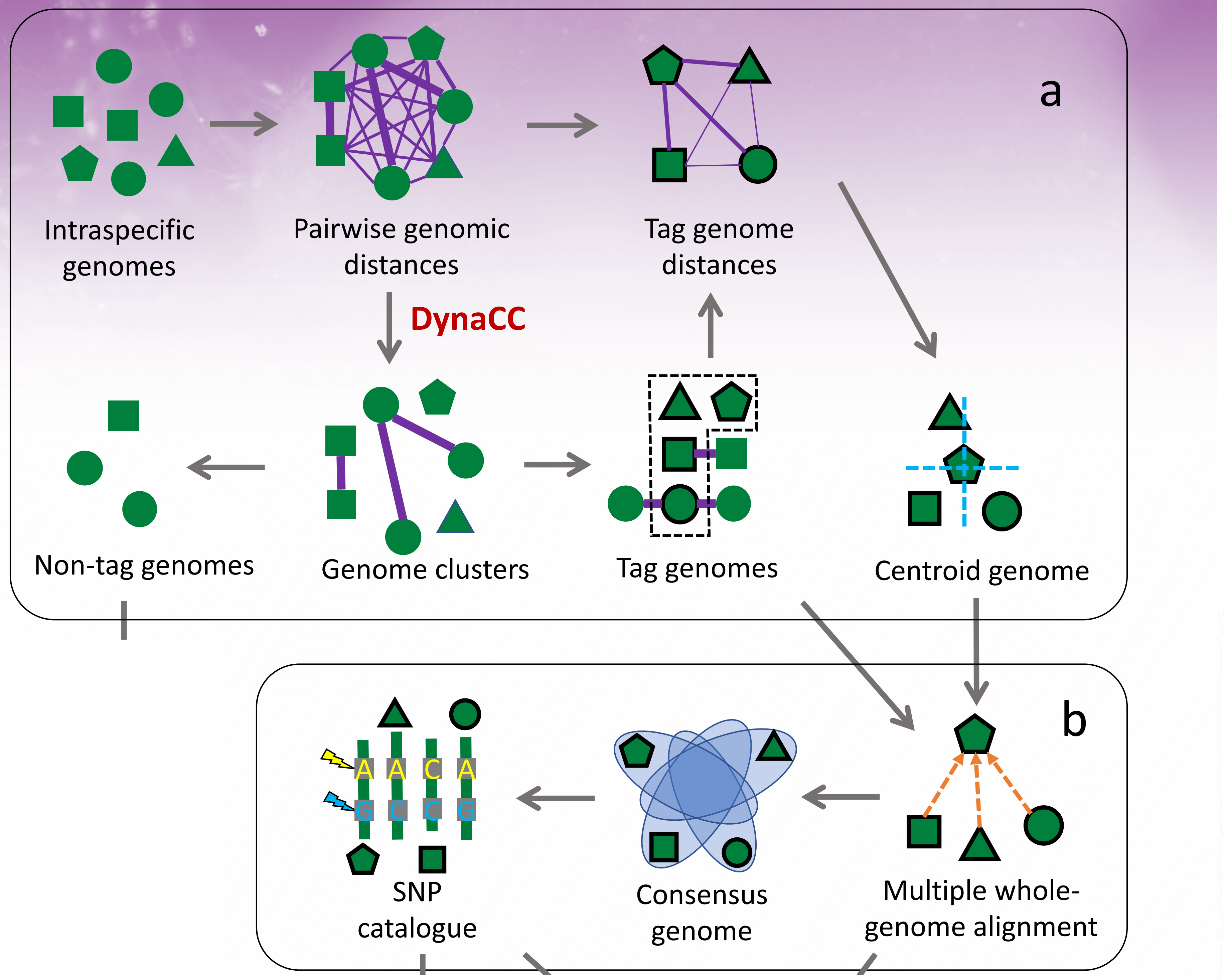
Maast: fast variant discovery from genomes

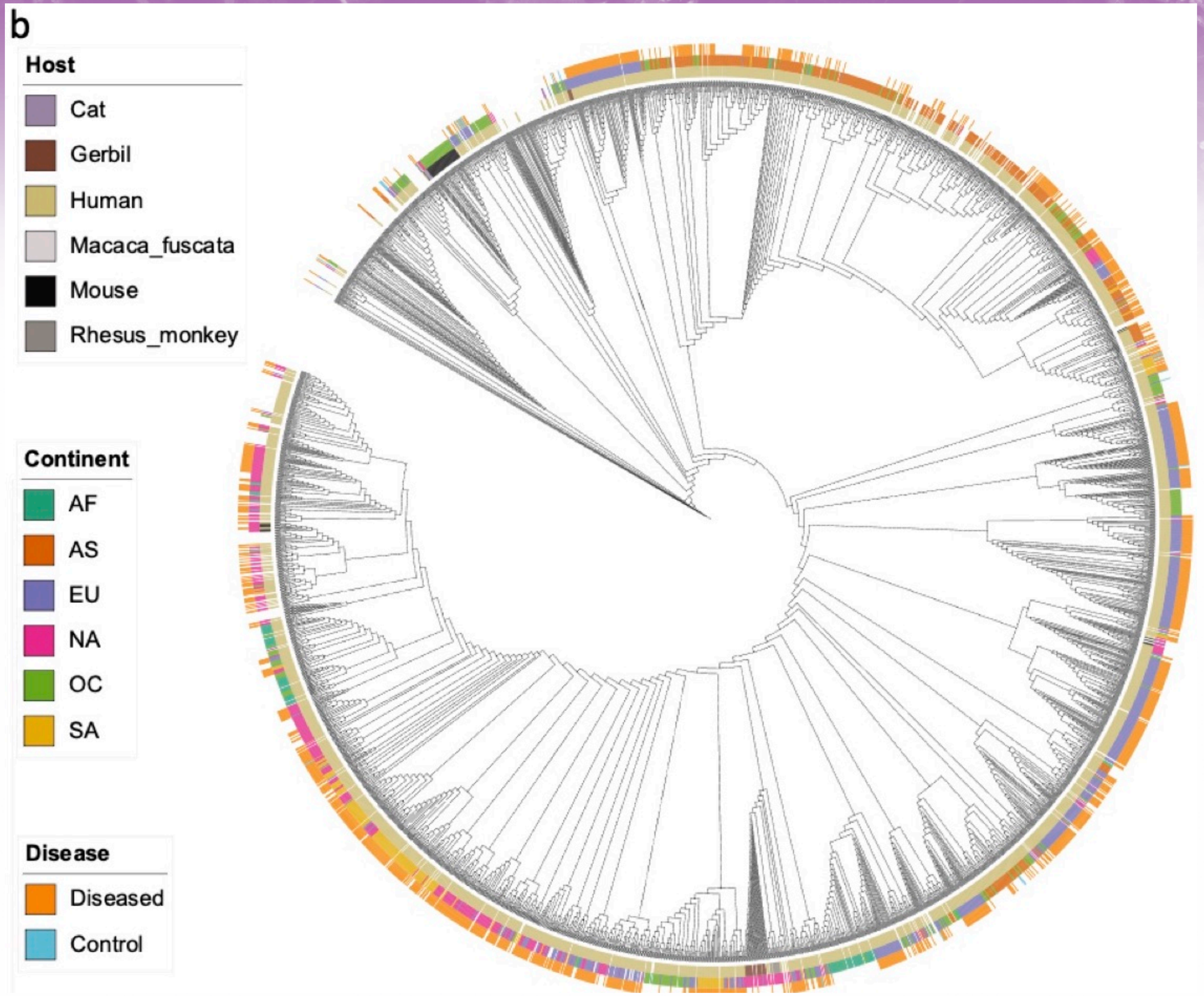




https://github.com/zjshi/Maast Shi et al. (2022)

Maast: fast variant discovery from genomes





3,068 *H. pylori* strains
Also: 37,096 SARS-CoV-2 strains

https://github.com/zjshi/Maast Shi et al. (2022)

Tag genomes speed up variant discovery and improve accuracy

Sequencing effort should focus on new lineages not redundant ones

Future Prospects

- Strategies beyond short-read aligners are needed, e.g.,
  - faster genome graph algorithms
  - probabilistic read mapping
  - read-to-read comparisons (reference databases for interpretation)
  - long reads / haplotypes
- Tools that use reference databases need to be flexibly implemented so that the algorithms and database can be tailored to the community

Future Prospects

- Not just problems for bacterial communities.
  - CRS and redundant genomes in some lineages of archaea, eukaryotes, and viruses.
- These challenges affect all bioinformatics methods that compare reads to databases, not just metagenotyping.
- Democratizing large-scale bioinformatics is critical!

Acknowledgements

pollard lab

Alumni: **Stephen Nayfach** (JGI), **Nandita Garud** (UCLA)

Collaborators: **Boris Dimitrov** (CZI), Ben Good (Stanford), Oskar Hallatschek (UCB), Nikos Kiripides (JGI)

Funded by NIH, NSF, Gladstone, UCSF, Chan-Zuckerberg Biohub