

# K-mer sketching for large-scale analyses

N. Tessa Pierce-Ward  
University of California, Davis



# Managing the Genome Data Deluge

Molecular biologists are turning to computer technology to help them manage the growing flood of sequencing and mapping data their field is producing

In 1980, if you had mentioned the term "bioinformatics" to a typical molecular biologist, you almost certainly would have been met with little more than a blank stare. Plenty of labs had their resident computer nerd, who spent hours crouched over a terminal, agonizing over how the latest batch of data should be stored and analyzed, but few biologists viewed this eccentric activity as a legitimate scientific discipline. "It was O.K. if I worked on computers," recalls James Ostell, who was a biology graduate student at Harvard University in 1980, "as long as it didn't interfere with my benchwork."

Today, however, that's all changed—and not just for Ostell, who's gone on to become chief of the information engineering branch at the National Center for Biotechnology Information (NCBI) at the National Institutes of Health campus in Bethesda, Maryland. Now, molecular biologists everywhere are increasingly turning to computer technology to help them deal with a major challenge: how to manage and interpret the flood of data being generated by the Human Genome Project and its companion efforts on model organisms from roundworms to mice. Entries in nucleotide sequence databases, such as the one run by the Heidelberg-based European Molecular Biology Laboratory (EMBL) data library, are growing exponentially (see figure). And it's a similar story for genetic and physical genome maps, protein structure information—and just about every other type of molecular biology data.

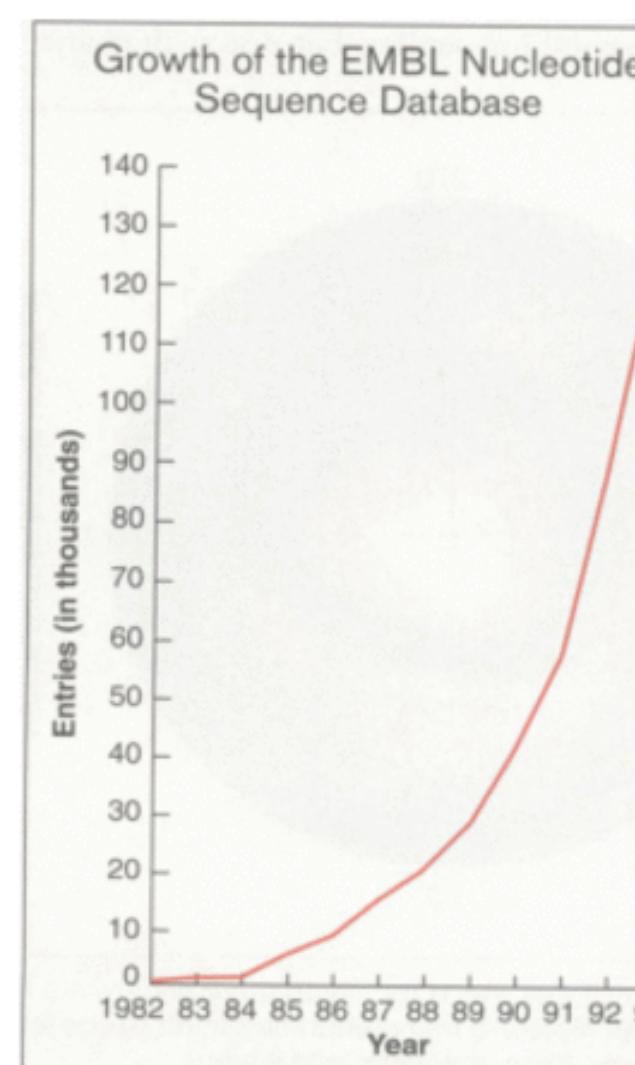
But as the data accumulate, a major problem is emerging. Researchers want instant access to all the information related to the genes they're studying. This would allow them, for instance, to gain clues to the function of a new gene that they've just sequenced by seeing whether other researchers have discovered similar genes and knew what their activities are. But the necessary data are usually spread over several molecular biology databases—there are now more than 50 in all—that don't communicate. It's the classic "Tower of Babel" situation, notes NCBI's Ostell. Moreover, it's an annoying bottleneck for research. When a molecular biologist sequences a new stretch of DNA, and discovers that it's similar to a gene from another organism, days of valuable research time can be wasted tracking down information on the function of this related gene. What's needed, says Cambridge University

## Bioinformatics

The molecular biology data explosion has given rise to the new science of biological computing or "bioinformatics," explored by Peter Aldhous in a story beginning on this page. That the data can be a valuable commodity is also evident in the wrangle between DOE and NIH over GenBank, described by Leslie Roberts on p. 504.

geneticist Michael Ashburner, is an integrated system allowing a researcher to click on boxes on his or her computer screens and summon up all the relevant data instantly.

Producing such a system is a major goal for NCBI and its transatlantic counterpart, the European Bioinformatics Institute (EBI)—an expanded effort based on the EMBL data library, which will open in new quarters at Hinxton near Cambridge, U.K., in 1995 (*Science*, 18 June, p. 1741). In addition to distributing sequence databases to the biology community, both centers will boast major database research efforts that will place



them at the forefront of the field of database integration. About one-third of NCBI's \$7.3 million-a-year budget is currently being spent on research to improve the databases and the software with which to search them, and EBI project leader Graham Cameron hopes to devote up to 20% of EBI's planned annual budget of some \$7.5 million to similar applied database research.

Although NCBI and EBI are similar in overall conception, they are set to tackle the issue of database integration in different ways. NCBI has set about uniting the data from several databases in a central integrated databank. In contrast, EBI plans to weld a multitude of separate databases into a loose "federation," communicating over computer networks—an approach that's also favored by biocomputing experts involved in the U.S. Department of Energy genome project (see p. 504).

Building all of the important biology databases into a centrally integrated system will be a laborious task, but NCBI has already taken a first step down the road toward the goal. Since last fall, researchers using the databases distributed by NCBI on CD-ROM have been able to use a software package called Entrez to browse a central integrated database consisting of three types of data: nucleotide and protein sequences from the leading general sequence databases distributed by NCBI and EMBL, plus abstracts of papers from the Medline biomedical literature database.

To make the system work, the NCBI group first had to build into it cross references that record the connections between data that are biologically related—noting which protein is encoded by a particular genetic sequence, for example. "That's the really critical thing [for any integration project]," says NCBI director David Lipman. Entrez not only recognizes the links between nucleotide and protein sequences and between sequences and the papers that cite them, it also assesses the similarity between the sequences and includes word recognition routines that scan Medline abstracts to identify additional related papers. "We find [Entrez] an enormously useful program," says David Hillis, a regular user who heads a molecular evolution lab at the University of Texas at Austin.

NCBI staff are now working to incorporate three-dimensional structure information

SCIENCE • VOL. 262 • 22 OCTOBER 1993

~130k sequence records

# Managing the Genome Data Deluge

Molecular biologists are turning to computer technology to help them manage the growing flood of sequencing and mapping data their field is producing

In 1980, if you had mentioned the term "bioinformatics" to a typical molecular biologist, you almost certainly would have been met with little more than a blank stare. Plenty of labs had their resident computer nerd, who spent hours crouched over a terminal, agonizing over how the latest batch of data should be stored and analyzed, but few biologists viewed this eccentric activity as a legitimate scientific discipline. "It was O.K. if I worked on computers," recalls James Ostell, who was a biology graduate student at Harvard University in 1980, "as long as it didn't interfere with my benchwork."

Today, however, that's all changed—and not just for Ostell, who's gone on to become chief of the information engineering branch at the National Center for Biotechnology Information (NCBI) at the National Institutes of Health campus in Bethesda, Maryland. Now, molecular biologists everywhere are increasingly turning to computer technology to help them deal with a major challenge: how to manage and interpret the flood of data being generated by the Human Genome Project and its companion efforts on model organisms from roundworms to mice. Entries in nucleotide sequence databases, such as the one run by the Heidelberg-based European Molecular Biology Laboratory (EMBL) data library, are growing exponentially (see figure). And it's a similar story for genetic and physical genome maps, protein structure information—and just about every other type of molecular biology data.

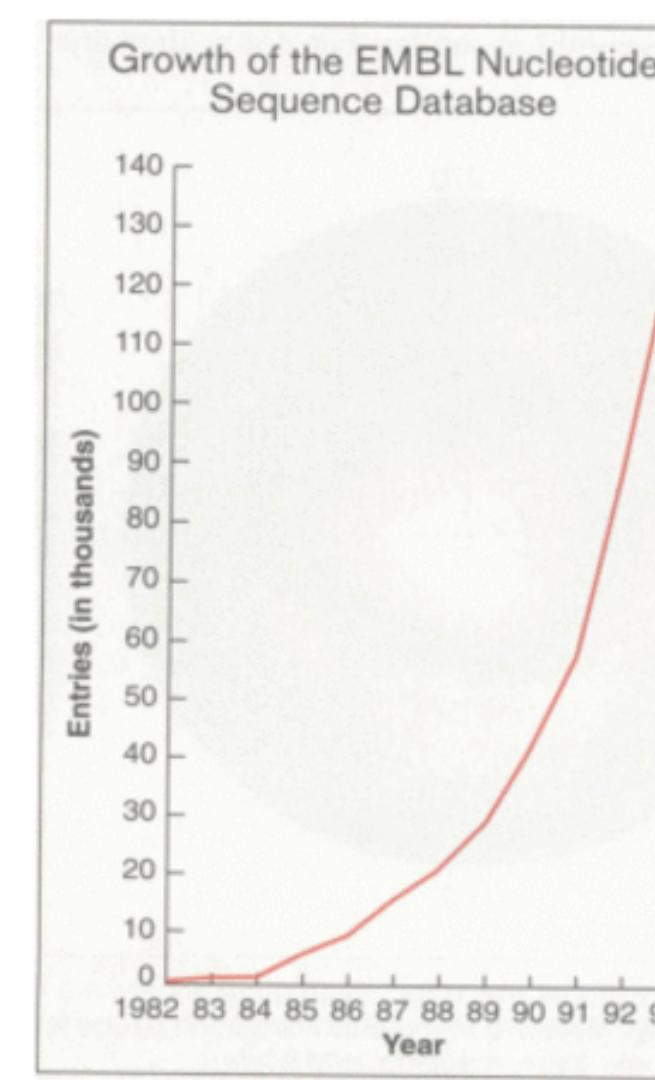
But as the data accumulate, a major problem is emerging. Researchers want instant access to all the information related to the genes they're studying. This would allow them, for instance, to gain clues to the function of a new gene that they've just sequenced by seeing whether other researchers have discovered similar genes and knew what their activities are. But the necessary data are usually spread over several molecular biology databases—there are now more than 50 in all—that don't communicate. It's the classic "Tower of Babel" situation, notes NCBI's Ostell. Moreover, it's an annoying bottleneck for research. When a molecular biologist sequences a new stretch of DNA, and discovers that it's similar to a gene from another organism, days of valuable research time can be wasted tracking down information on the function of this related gene. What's needed, says Cambridge University

## Bioinformatics

The molecular biology data explosion has given rise to the new science of biological computing or "bioinformatics," explored by Peter Aldhous in a story beginning on this page. That the data can be a valuable commodity is also evident in the wrangle between DOE and NIH over GenBank, described by Leslie Roberts on p. 504.

geneticist Michael Ashburner, is an integrated system allowing a researcher to click on boxes on his or her computer screens and summon up all the relevant data instantly.

Producing such a system is a major goal for NCBI and its transatlantic counterpart, the European Bioinformatics Institute (EBI)—an expanded effort based on the EMBL data library, which will open in new quarters at Hinxton near Cambridge, U.K., in 1995 (*Science*, 18 June, p. 1741). In addition to distributing sequence databases to the biology community, both centers will boast major database research efforts that will place



them at the forefront of the field of database integration. About one-third of NCBI's \$7.3 million-a-year budget is currently being spent on research to improve the databases and the software with which to search them, and EBI project leader Graham Cameron hopes to devote up to 20% of EBI's planned annual budget of some \$7.5 million to similar applied database research.

Although NCBI and EBI are similar in overall conception, they are set to tackle the issue of database integration in different ways. NCBI has set about uniting the data from several databases in a central integrated databank. In contrast, EBI plans to weld a multitude of separate databases into a loose "federation," communicating over computer networks—an approach that's also favored by biocomputing experts involved in the U.S. Department of Energy genome project (see p. 504).

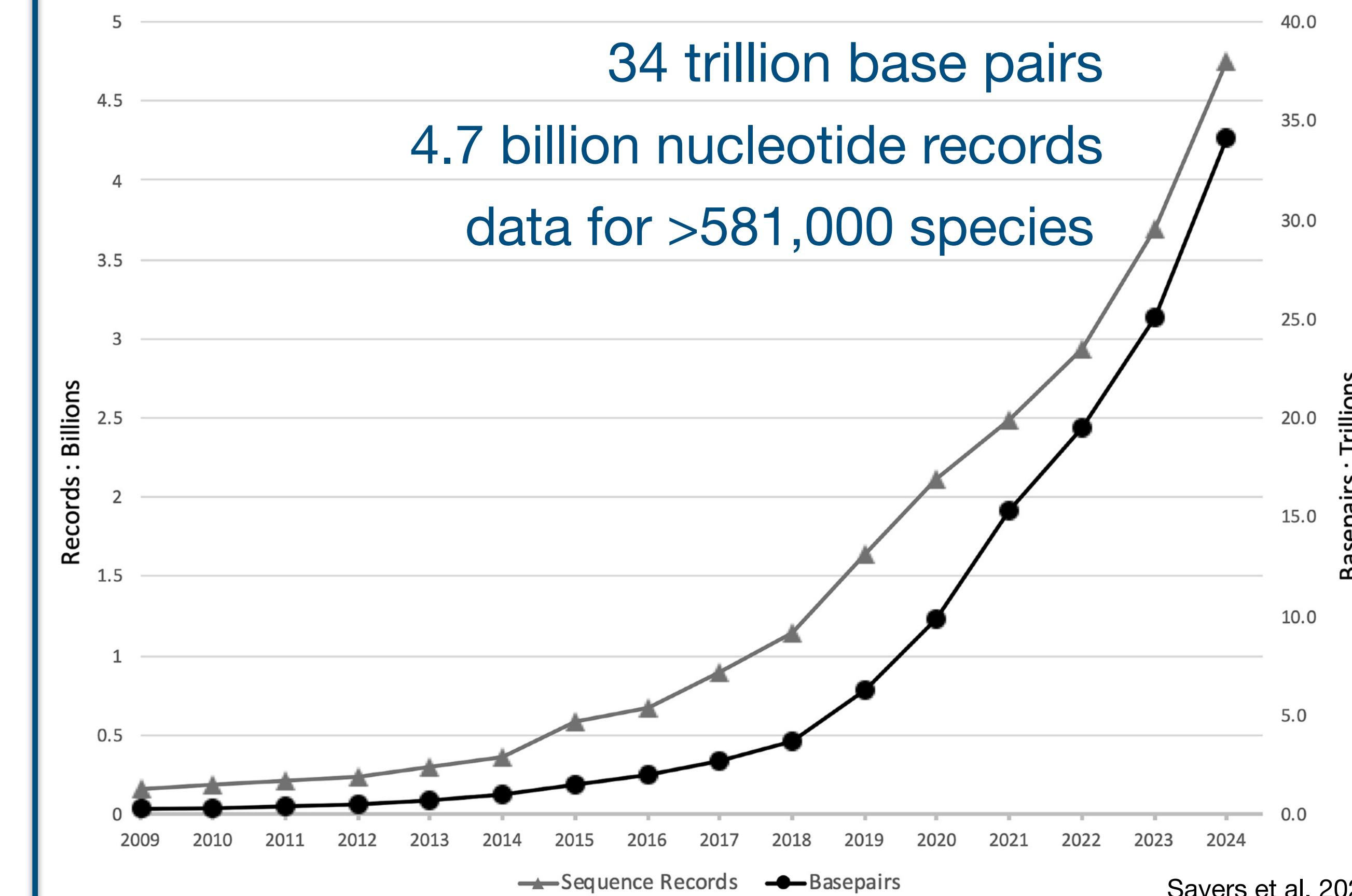
Building all of the important biology databases into a centrally integrated system will be a laborious task, but NCBI has already taken a first step down the road toward the goal. Since last fall, researchers using the databases distributed by NCBI on CD-ROM have been able to use a software package called Entrez to browse a central integrated database consisting of three types of data: nucleotide and protein sequences from the leading general sequence databases distributed by NCBI and EMBL, plus abstracts of papers from the Medline biomedical literature database.

To make the system work, the NCBI group first had to build into it cross references that record the connections between data that are biologically related—noting which protein is encoded by a particular genetic sequence, for example. "That's the really critical thing [for any integration project]," says NCBI director David Lipman. Entrez not only recognizes the links between nucleotide and protein sequences and between sequences and the papers that cite them, it also assesses the similarity between the sequences and includes word recognition routines that scan Medline abstracts to identify additional related papers. "We find [Entrez] an enormously useful program," says David Hillis, a regular user who heads a molecular evolution lab at the University of Texas at Austin.

NCBI staff are now working to incorporate three-dimensional structure information

2025

Annual GenBank Growth



Sayers et al. 2025

<https://doi.org/10.1093/nar/gkae1114>

SCIENCE • VOL. 262 • 22 OCTOBER 1993

~130k sequence records

(GenBank only!)

# Managing the Genome Data Deluge

Molecular biologists are turning to computer technology to help them manage the growing flood of sequencing and mapping data their field is producing

In 1980, if you had mentioned the term "bioinformatics" to a typical molecular biologist, you almost certainly would have been met with little more than a blank stare. Plenty of labs had their resident computer nerd, who spent hours crouched over a terminal, agonizing over how the latest batch of data should be stored and analyzed, but few biologists viewed this eccentric activity as a legitimate scientific discipline. "It was O.K. if I worked on computers," recalls James Ostell, who was a biology graduate student at Harvard University in 1980, "as long as it didn't interfere with my benchwork."

Today, however, that's all changed—and not just for Ostell, who's gone on to become chief of the information engineering branch at the National Center for Biotechnology Information (NCBI) at the National Institutes of Health campus in Bethesda, Maryland. Now, molecular biologists everywhere are increasingly turning to computer technology to help them deal with a major challenge: how to manage and interpret the flood of data being generated by the Human Genome Project and its companion efforts on model organisms from roundworms to mice. Entries in nucleotide sequence databases, such as the one run by the Heidelberg-based European Molecular Biology Laboratory (EMBL) data library, are growing exponentially (see figure). And it's a similar story for genetic and physical genome maps, protein structure information—and just about every other type of molecular biology data.

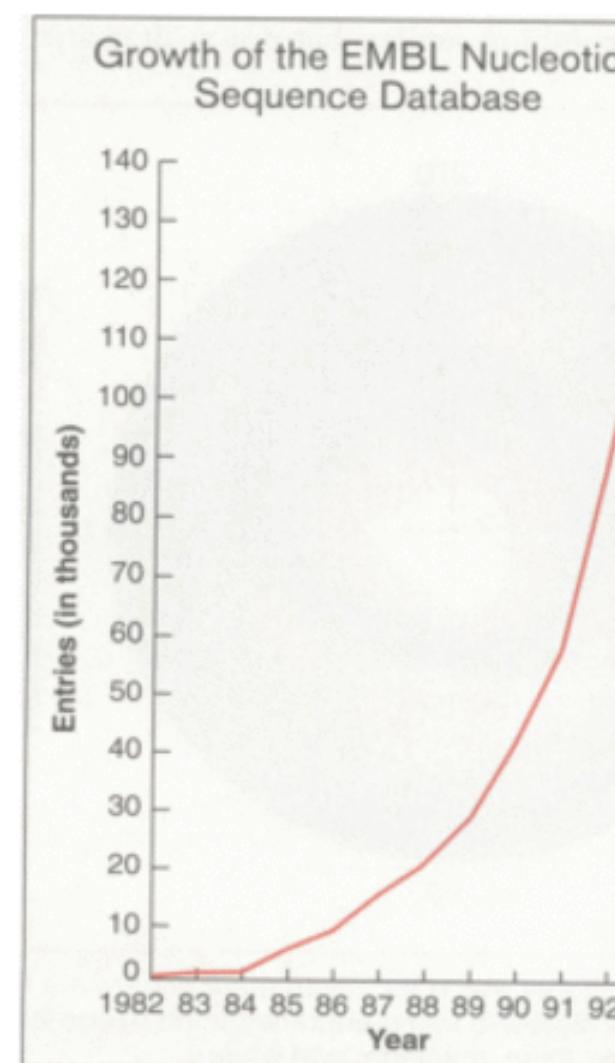
But as the data accumulate, a major problem is emerging. Researchers want instant access to all the information related to the genes they're studying. This would allow them, for instance, to gain clues to the function of a new gene that they've just sequenced by seeing whether other researchers have discovered similar genes and knew what their activities are. But the necessary data are usually spread over several molecular biology databases—there are now more than 50 in all—that don't communicate. It's the classic "Tower of Babel" situation, notes NCBI's Ostell. Moreover, it's an annoying bottleneck for research. When a molecular biologist sequences a new stretch of DNA, and discovers that it's similar to a gene from another organism, days of valuable research time can be wasted tracking down information on the function of this related gene. What's needed, says Cambridge University

## Bioinformatics

The molecular biology data explosion has given rise to the new science of biological computing or "bioinformatics," explored by Peter Aldhous in a story beginning on this page. That the data can be a valuable commodity is also evident in the wrangle between DOE and NIH over GenBank, described by Leslie Roberts on p. 504.

geneticist Michael Ashburner, is an integrated system allowing a researcher to click on boxes on his or her computer screens and summon up all the relevant data instantly.

Producing such a system is a major goal for NCBI and its transatlantic counterpart, the European Bioinformatics Institute (EBI)—an expanded effort based on the EMBL data library, which will open in new quarters at Hinxton near Cambridge, U.K., in 1995 (*Science*, 18 June, p. 1741). In addition to distributing sequence databases to the biology community, both centers will boast major database research efforts that will place



them at the forefront of the field of database integration. About one-third of NCBI's \$7.3 million-a-year budget is currently being spent on research to improve the databases and the software with which to search them, and EBI project leader Graham Cameron hopes to devote up to 20% of EBI's planned annual budget of some \$7.5 million to similar applied database research.

Although NCBI and EBI are similar in overall conception, they are set to tackle the issue of database integration in different ways. NCBI has set about uniting the data from several databases in a central integrated databank. In contrast, EBI plans to weld a multitude of separate databases into a loose "federation," communicating over computer networks—an approach that's also favored by biocomputing experts involved in the U.S. Department of Energy genome project (see p. 504).

Building all of the important biology databases into a centrally integrated system will be a laborious task, but NCBI has already taken a first step down the road toward the goal. Since last fall, researchers using the databases distributed by NCBI on CD-ROM have been able to use a software package called Entrez to browse a central integrated database consisting of three types of data: nucleotide and protein sequences from the leading general sequence databases distributed by NCBI and EMBL, plus abstracts of papers from the Medline biomedical literature database.

To make the system work, the NCBI group first had to build into it cross references that record the connections between data that are biologically related—noting which protein is encoded by a particular genetic sequence, for example. "That's the really critical thing [for any integration project]," says NCBI director David Lipman. Entrez not only recognizes the links between nucleotide and protein sequences and between sequences and the papers that cite them, it also assesses the similarity between the sequences and includes word recognition routines that scan Medline abstracts to identify additional related papers. "We find [Entrez] an enormously useful program," says David Hillis, a regular user who heads a molecular evolution lab at the University of Texas at Austin.

NCBI staff are now working to incorporate three-dimensional structure information

## How to handle effectively infinite data?

SCIENCE • VOL. 262 • 22 OCTOBER 1993

# K-mer sketching

ACTACGCCCTTCATGACTC

ACTA

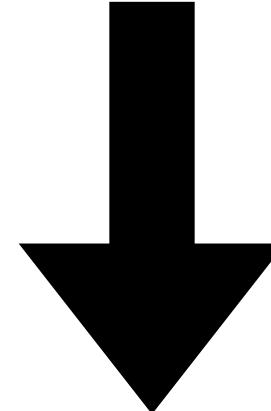
CTAC

TACG

ACGC

CGCT

k-mers of length 4  
(4-mers)



TACG

ACTC

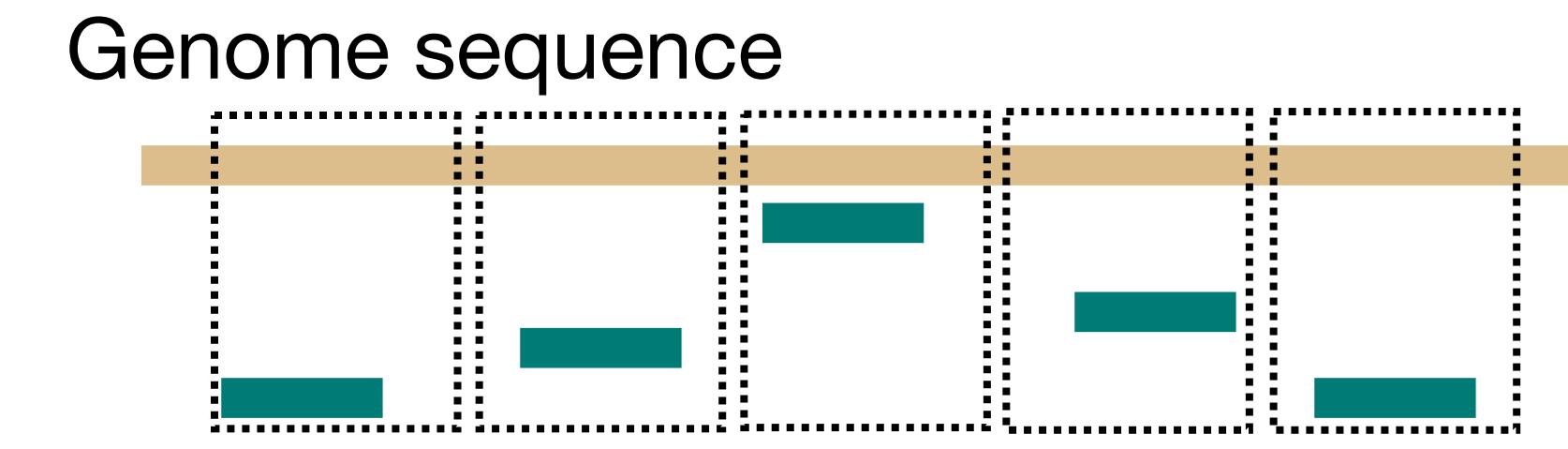
“sketch” of  
subsampled k-mers

- **Sketching:** the process of generating an approximate, compact summary of data (Rowe, 2019)
- Effective data compression while still allowing for ~accurate and ~sensitive query and comparison

# Types of k-mer sketching

- **Local k-mer selection methods**

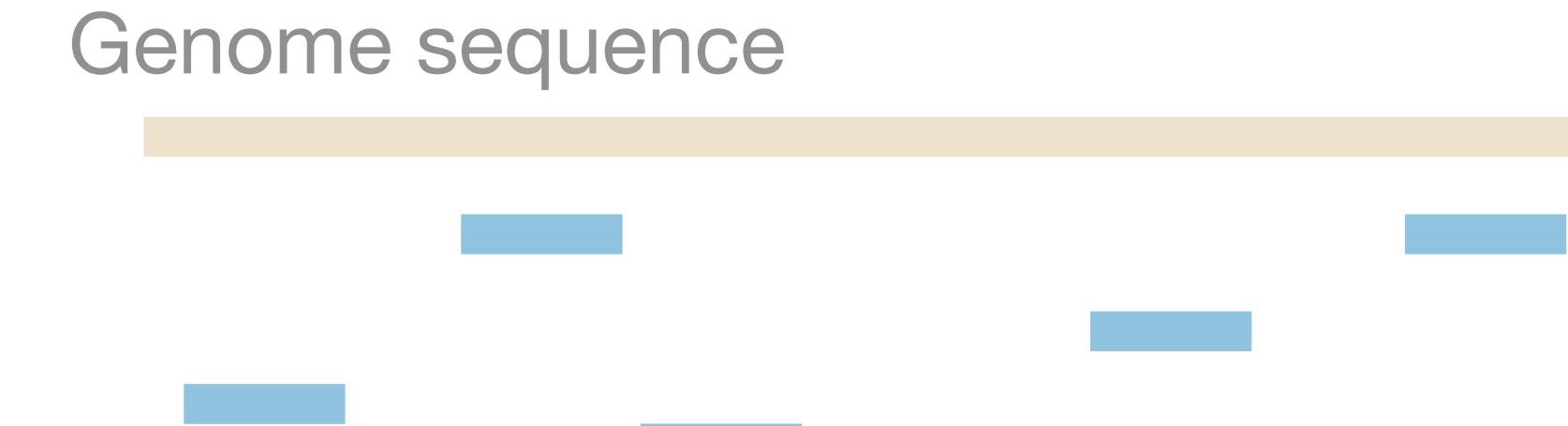
- **Required for location-informed methods (indexing, read mapping, alignment, etc)**  
k-mer selection uses positional information;  
Can provide “window” guarantees.
- Minimizers, Sync-mers...



“window guarantee” = at least one k-mer selected per window

- **Global k-mer selection methods**

- Used primarily for dataset-level comparisons. Selection does not use positional information
- MinHash, FracMinHash, SetSketch...



*Methods have different trade-offs and use cases*

# Local k-mer selection: minimizers

window ( $w$ ) : 5 k-mers

ACTACGCCCTTCATGACTC

ACTA

CTAC

TACG

ACGC

CGCT

k-mers of length 4  
(4-mers)

Apply ordering, e.g. lexicographic

1. ACGC
2. ACTA
3. CGCT
4. CTAC
5. TACG



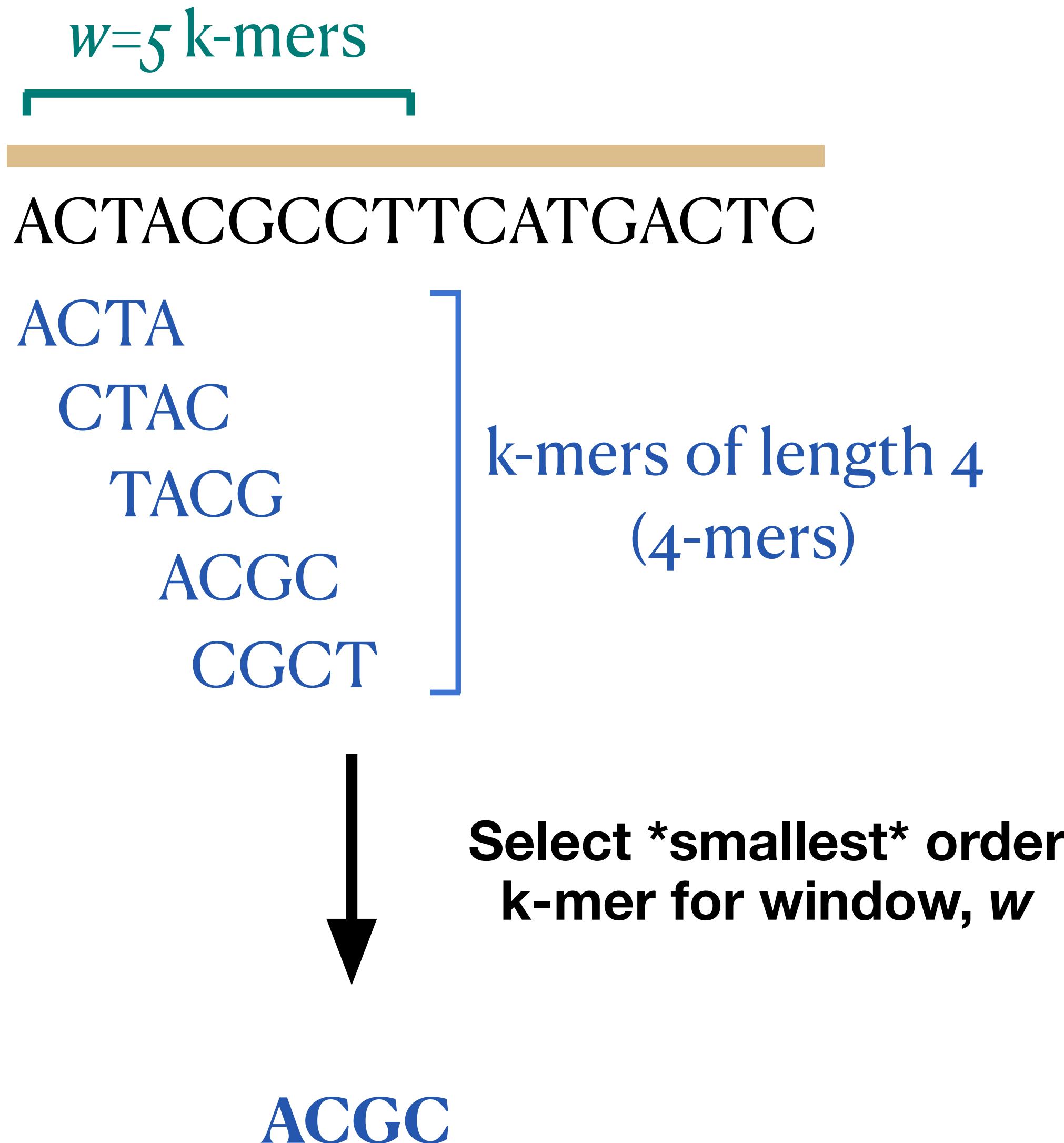
## params:

- k-mer size,  $k$
- window size,  $w$
- ordering method

Select \*smallest\* order  
k-mer for window,  $w$

ACGC

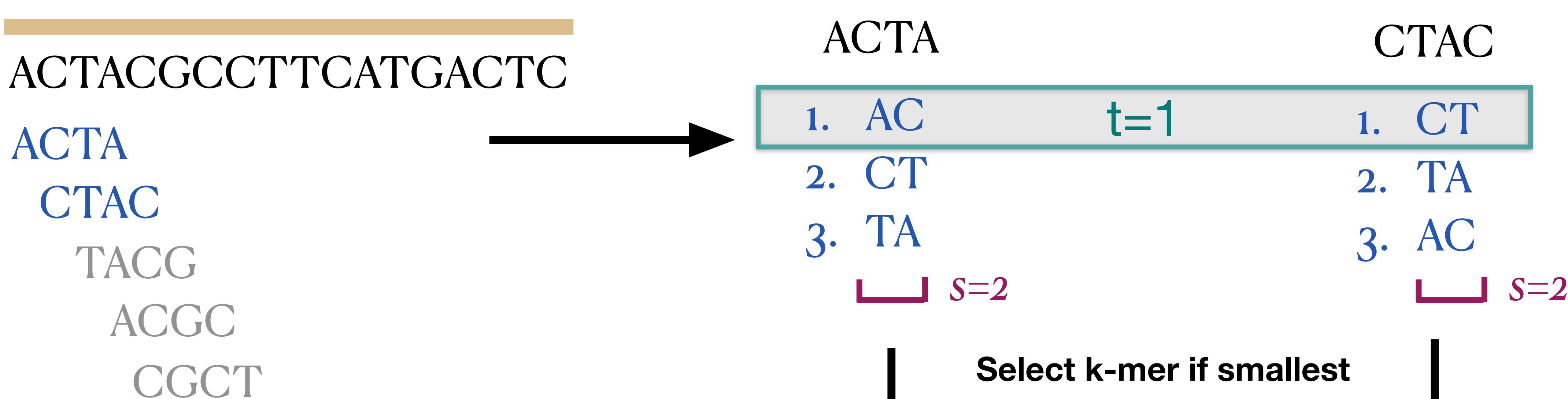
# Local k-mer selection: minimizers



## Properties:

- Dataset compression ~without information loss
  - Fixed compression ratio
- Guarantee k-mers well distributed
- BUT, minimizer selection is impacted by sequence mutations outside of the window

# Local k-mer selection: sync-mers

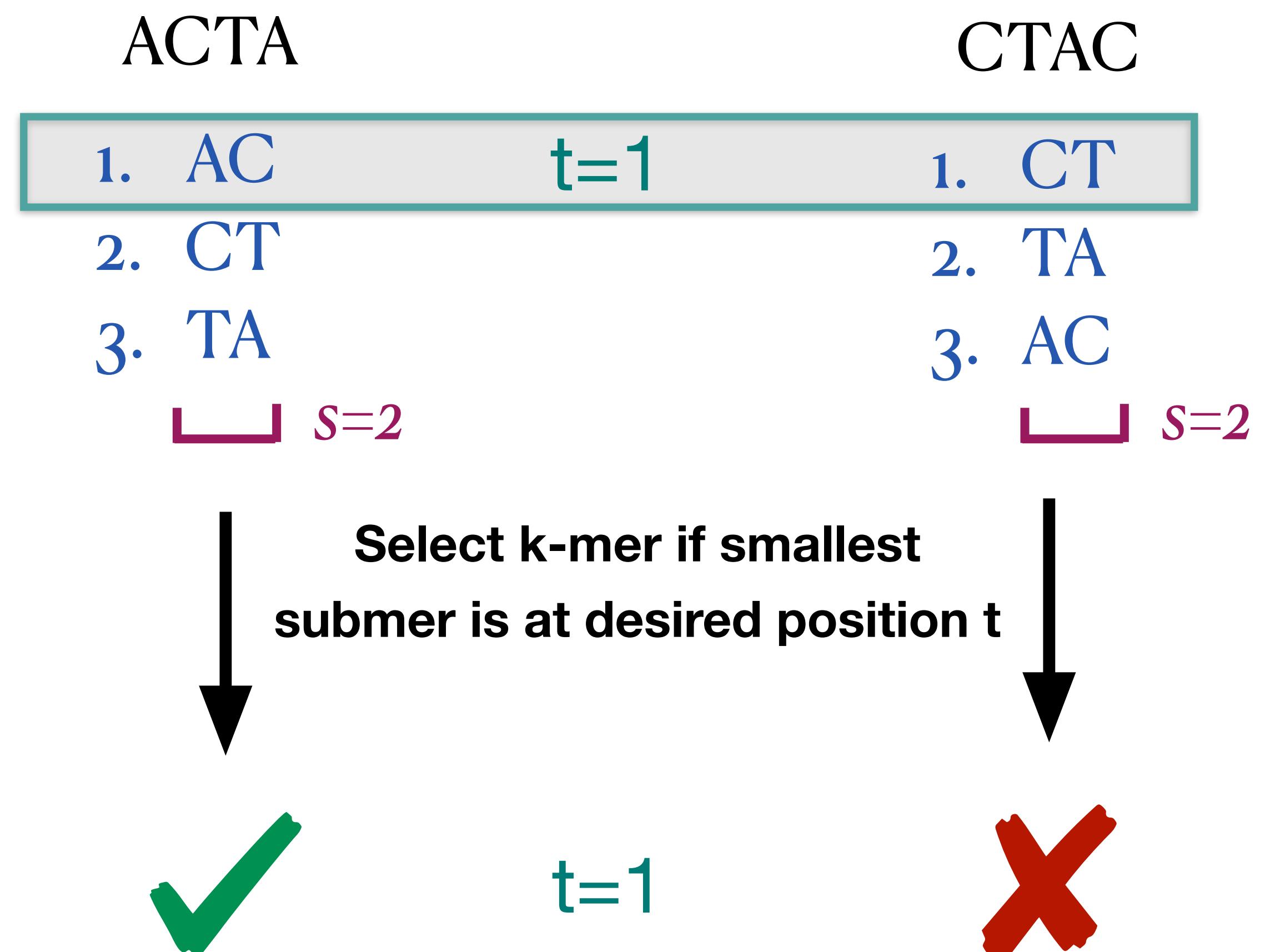


## params:

- *k-mer size, k*
- *submer size, s*
- ordering method
- *submer position, t*

# Local k-mer selection: sync-mers

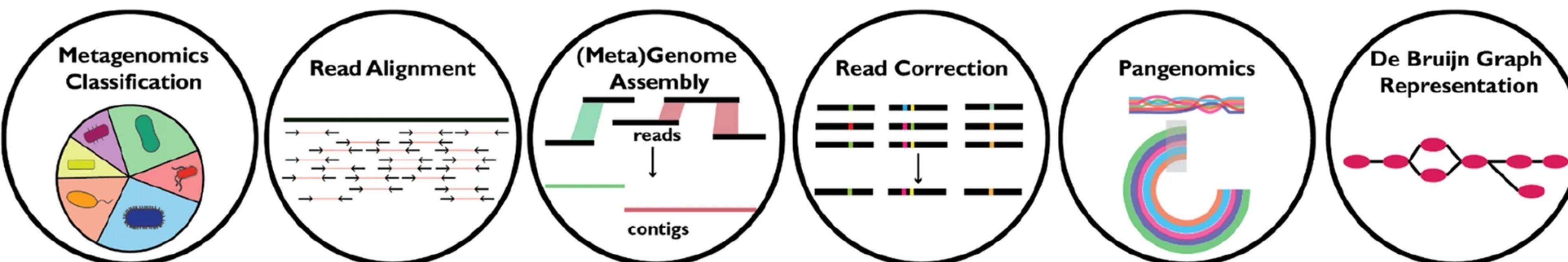
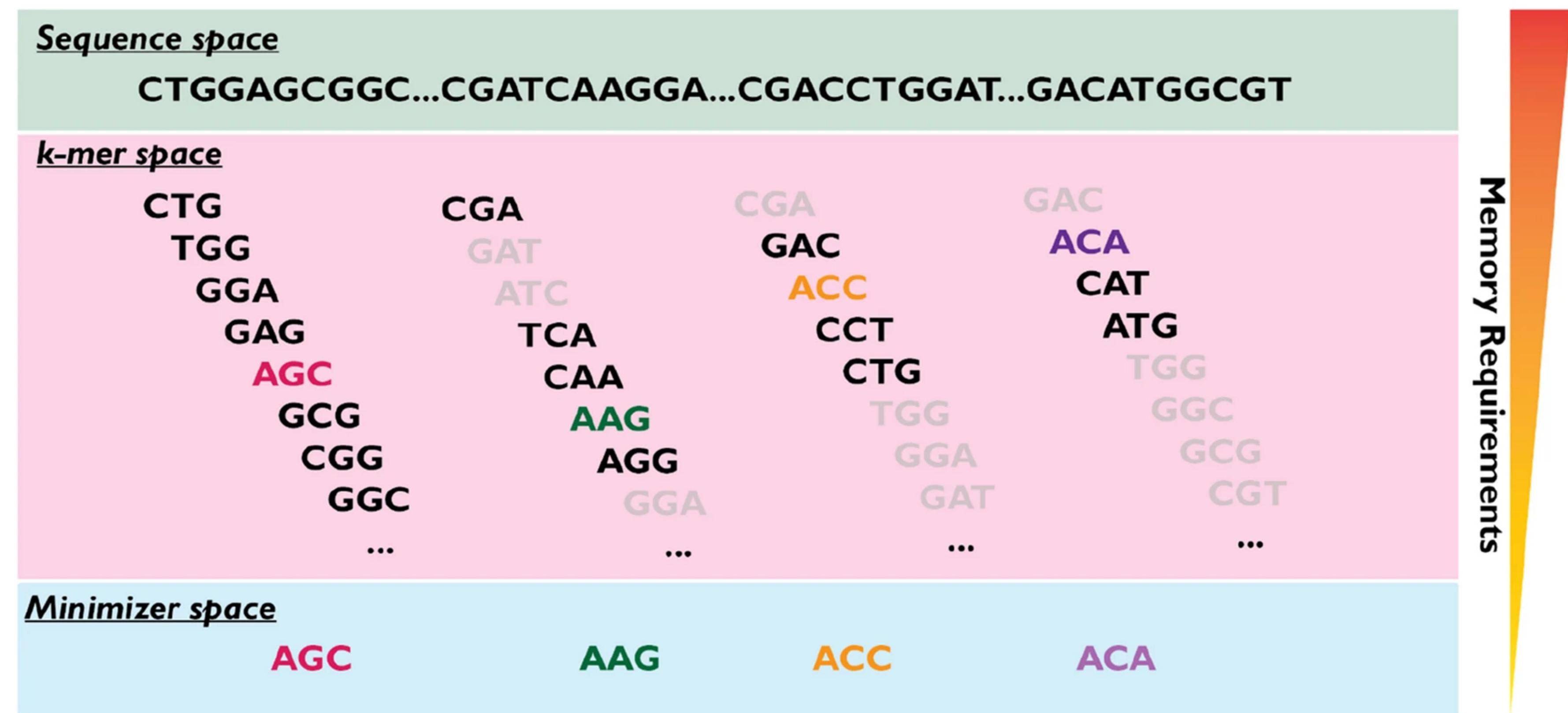
Order submers (s) within each k-mer



Properties:

- Dataset compression ~without information loss
- Fixed compression ratio
- Guarantee k-mers well distributed
- Sync-mer selection is **NOT** impacted by sequence mutations outside of the window

# Common use cases for local k-mer selection



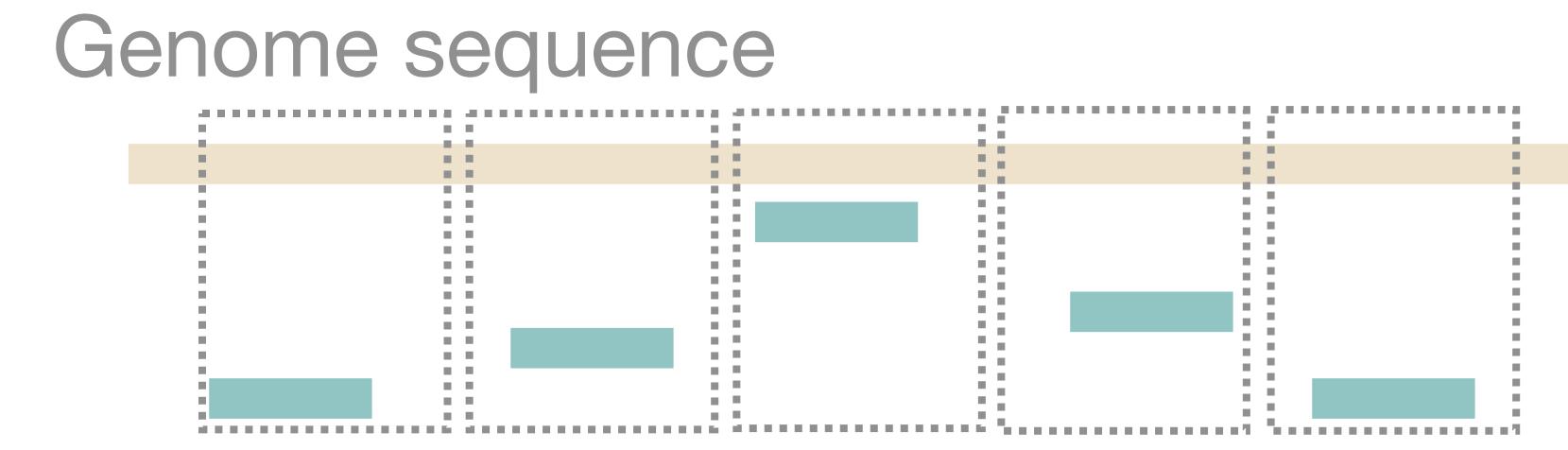
Ndiaye et al 2024

doi: 10.1186/s13059-024-03414-4

# Types of k-mer sketching

- Local k-mer selection methods

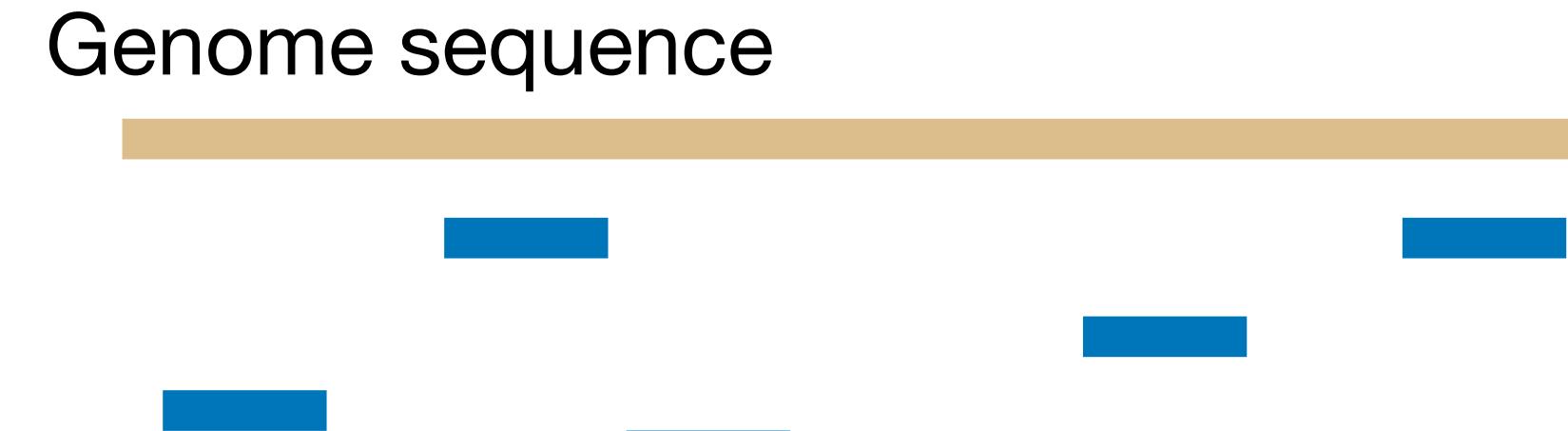
- Required for location-informed methods (indexing, read mapping, alignment, etc)  
K-mer selection uses positional information;  
Can provide “window” guarantees.
- Minimizers, Sync-mers...



“window guarantee” = at least one k-mer selected per window

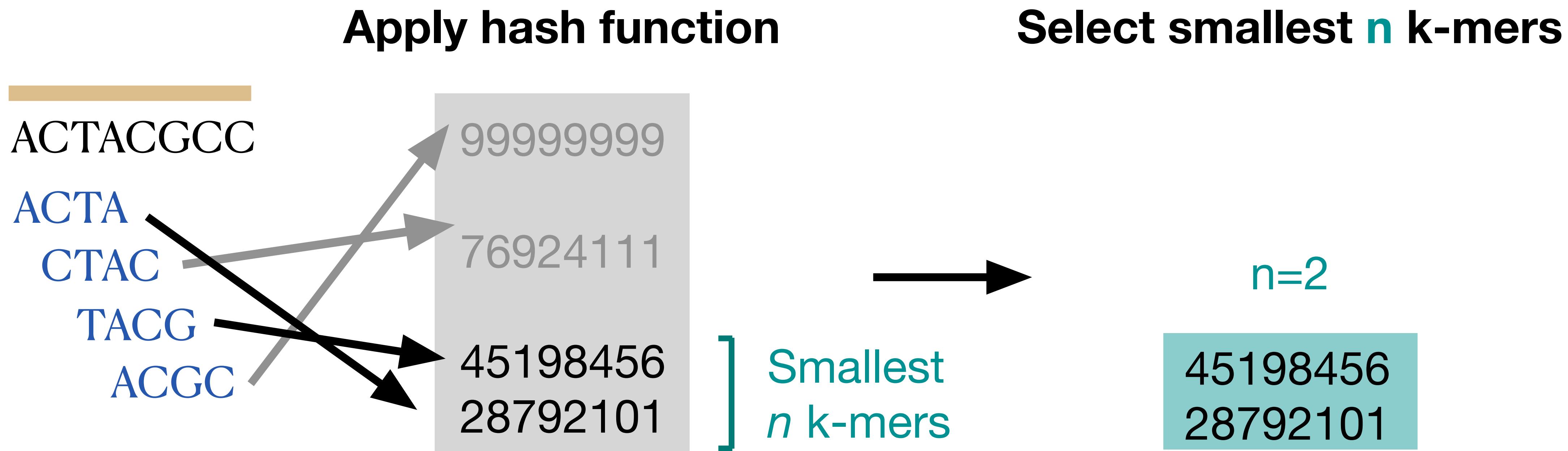
- Global k-mer selection methods

- Used primarily for dataset-level comparisons. K-mer selection does not use positional information
- MinHash, FracMinHash, SetSketch...



*Methods have different trade-offs and use cases*

# Global k-mer selection: MinHash

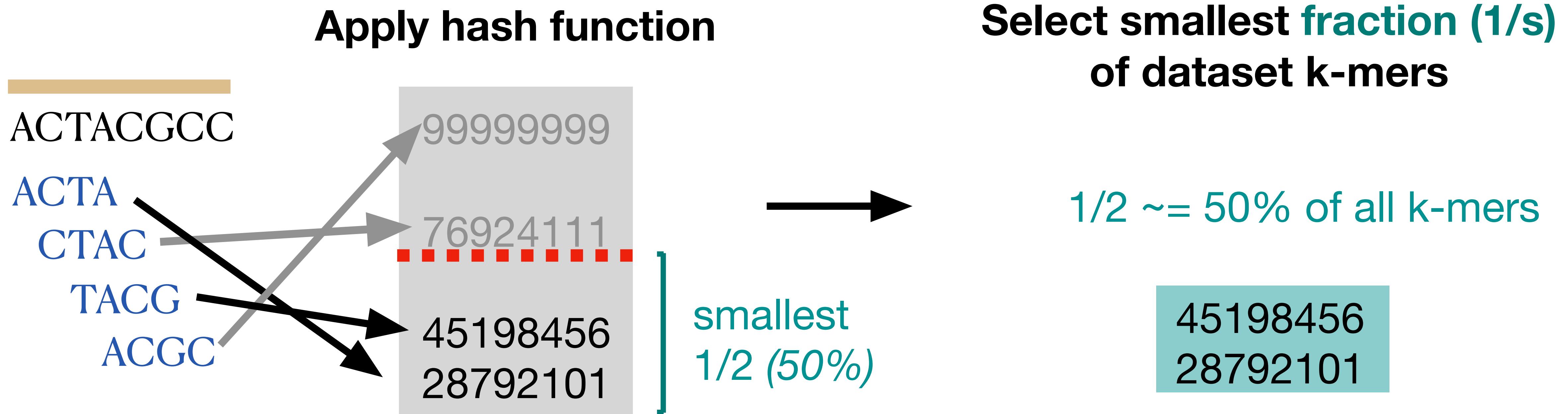


## params:

- k-mer size,  $k$
- number of k-mers to keep,  $n$
- hash function (ordering method)

Common usage: select 2000-10,000 k-mers per dataset

# Global k-mer selection: FracMinHash

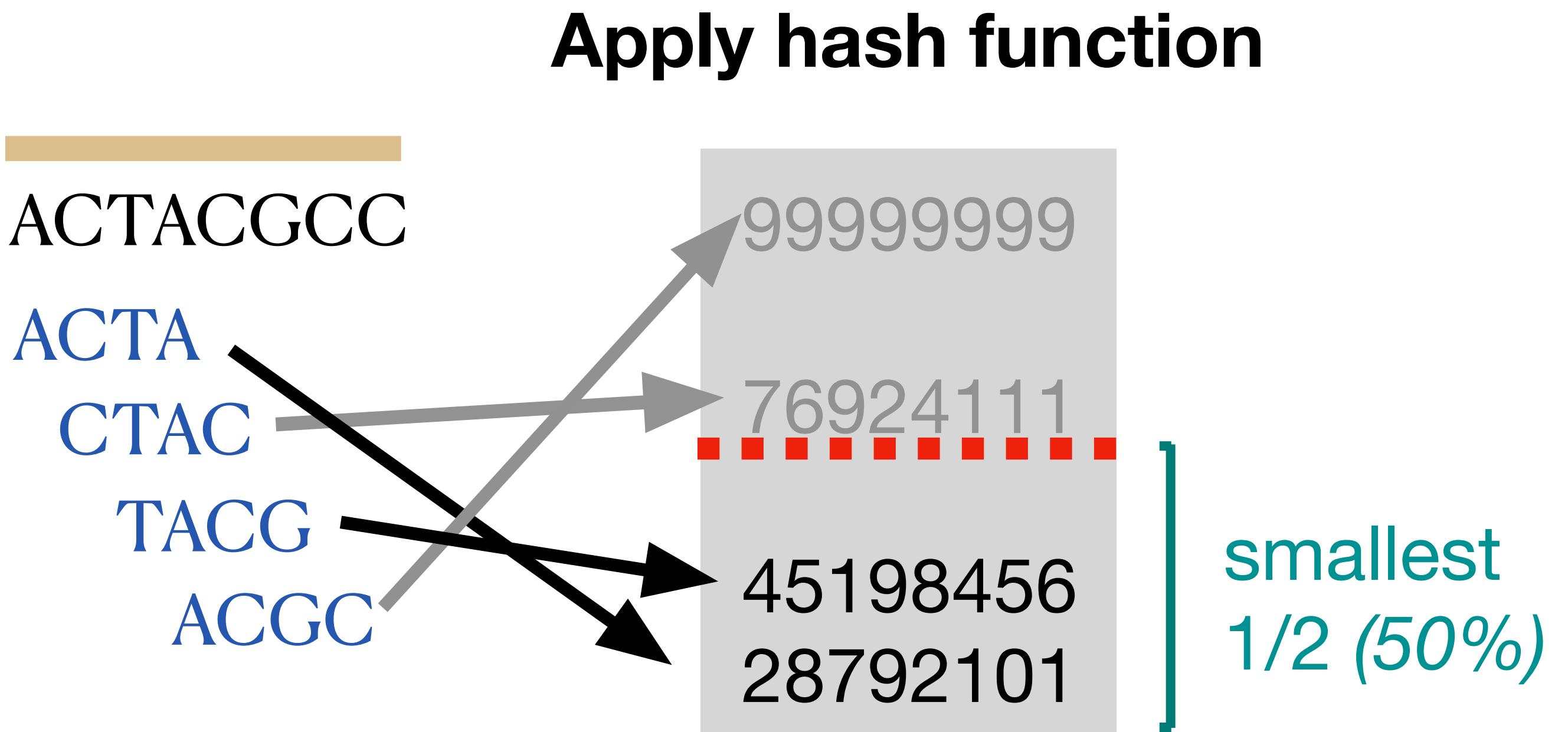


## params:

- k-mer size,  $k$
- fraction of k-mers to keep,  $1/s$
- hash function (ordering method)

**Guarantee:** if a k-mer is selected for one dataset, it will also be selected for all others

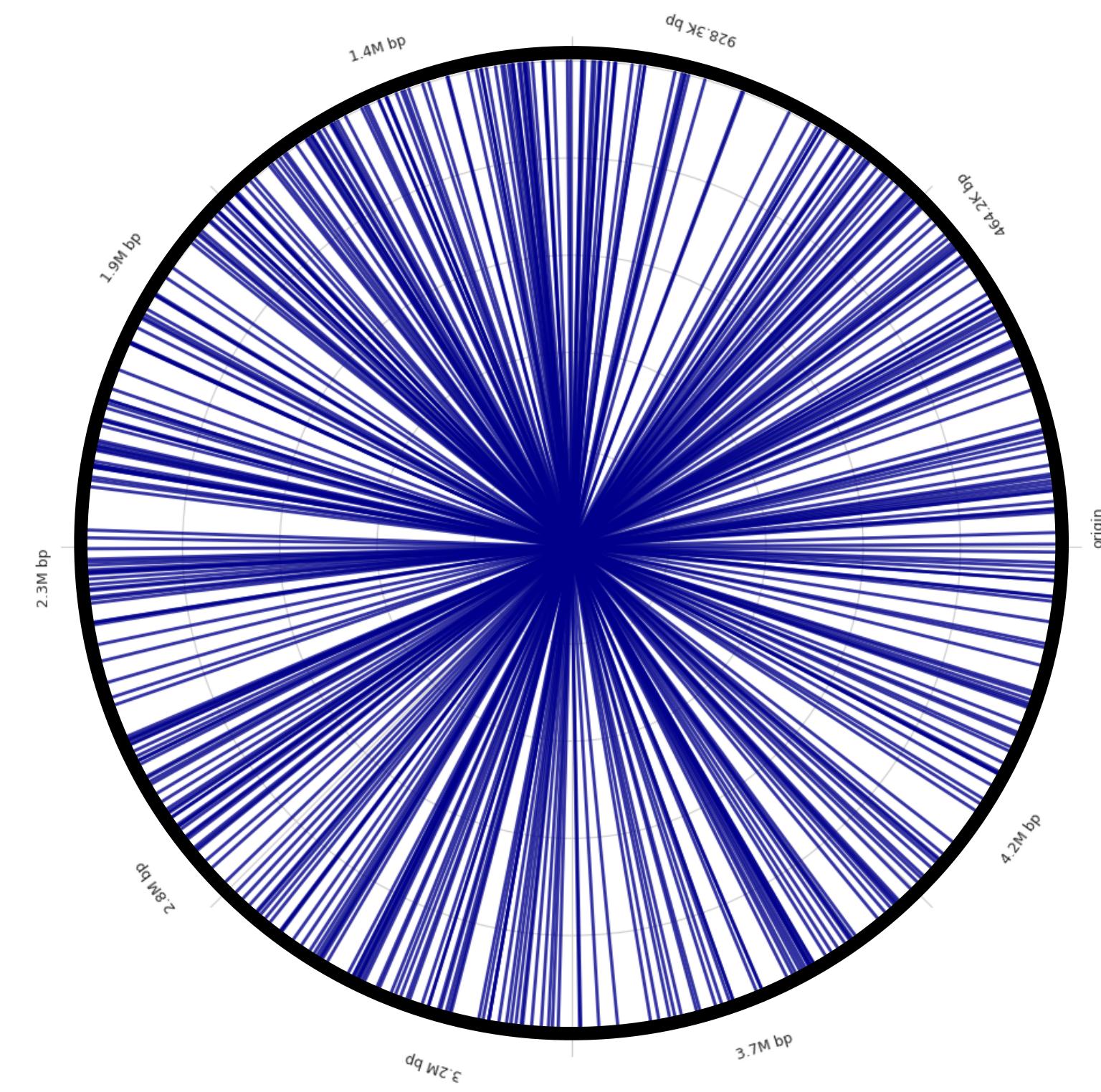
# Global k-mer selection: FracMinHash



## params:

- k-mer size,  $k$
- fraction of k-mers to keep,  $1/s$
- hash function (ordering method)

E. coli genome at scaled=10,000  
(select .01% of k-mers)



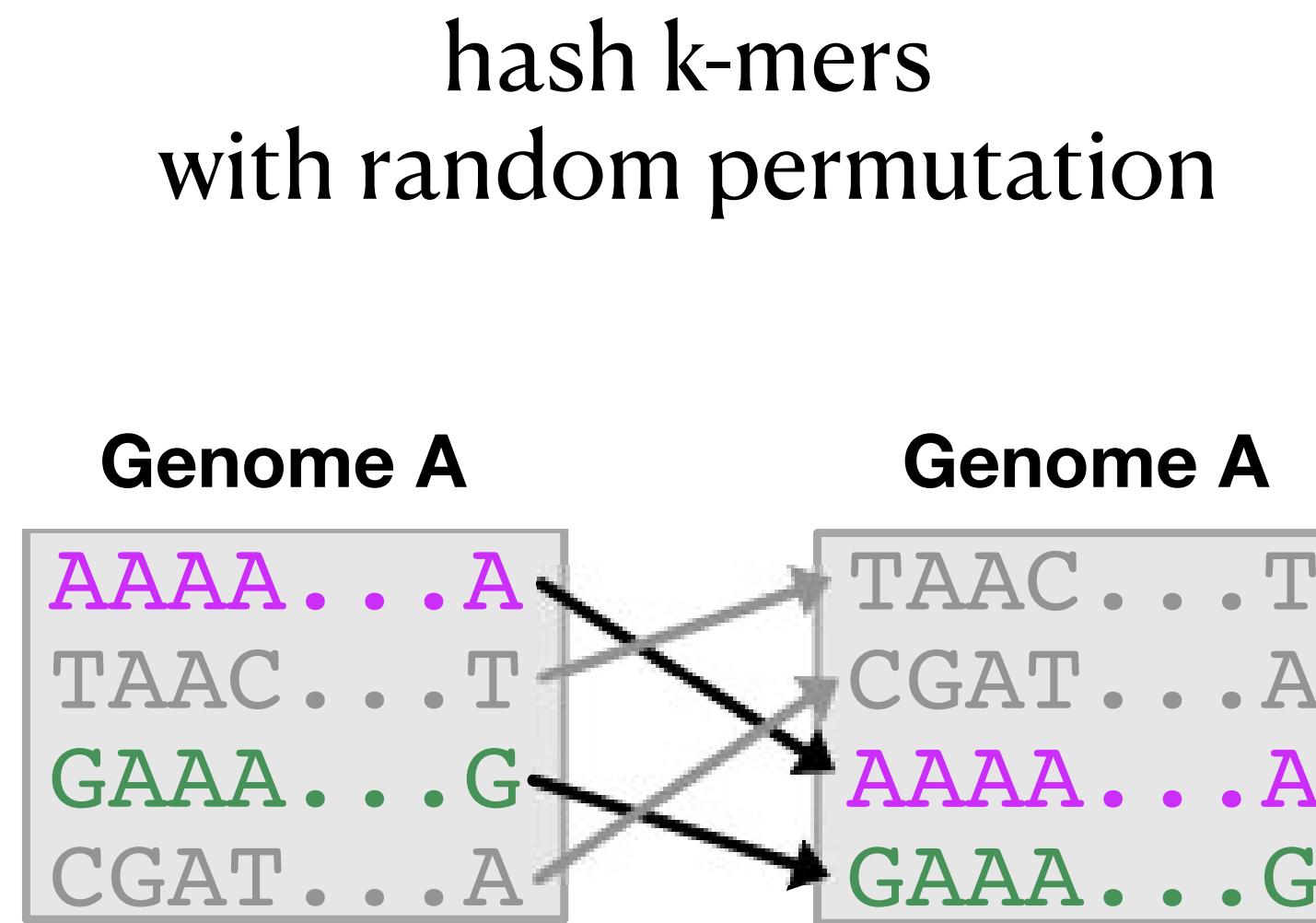
viz by Mo Abuelanin

[github.com/mr-eyes/2024-fracminhash-viz](https://github.com/mr-eyes/2024-fracminhash-viz)

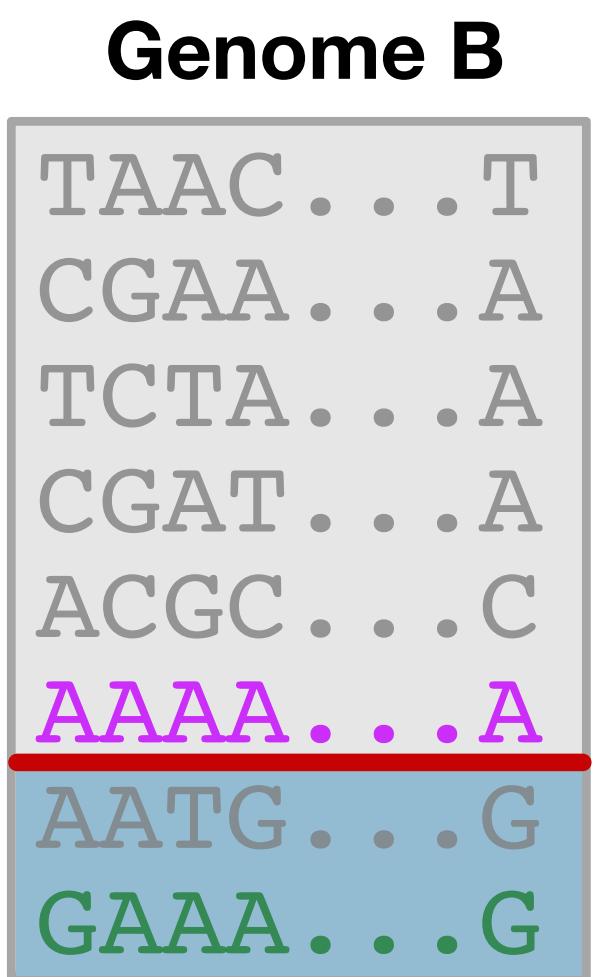
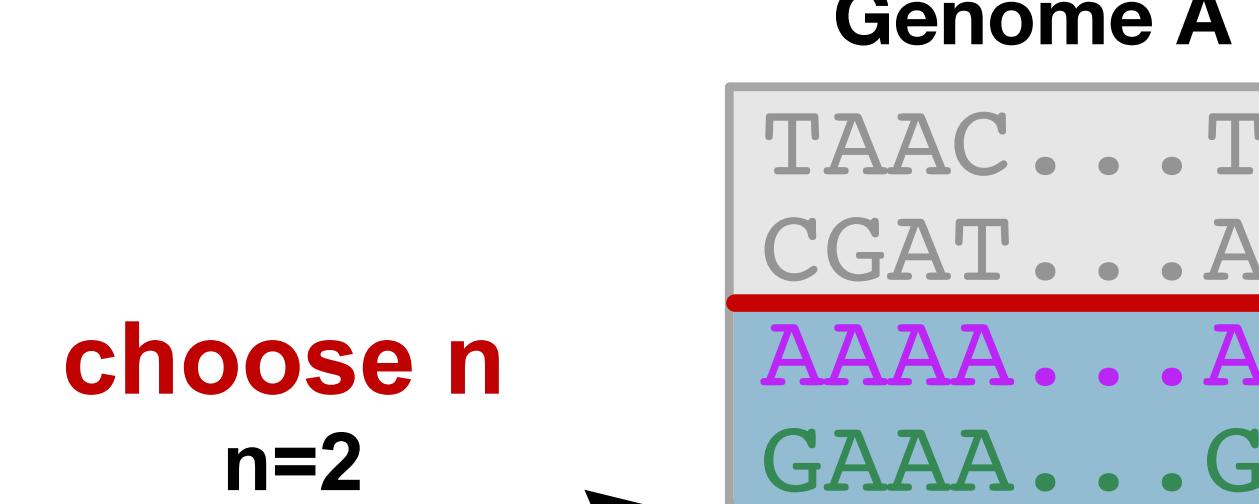
# Features of global k-mer selection methods

- Random sampling, unbiased relative to sequence composition
- **Flexible compression ratio:**
  - **Can tune to be efficient for large-scale comparisons**, e.g. all x all genomic or metagenomic comparisons, database searches, etc
- Known error bounds, variance estimates, etc
- Statistically sound estimates of set similarity

# MinHash vs FracMinHash

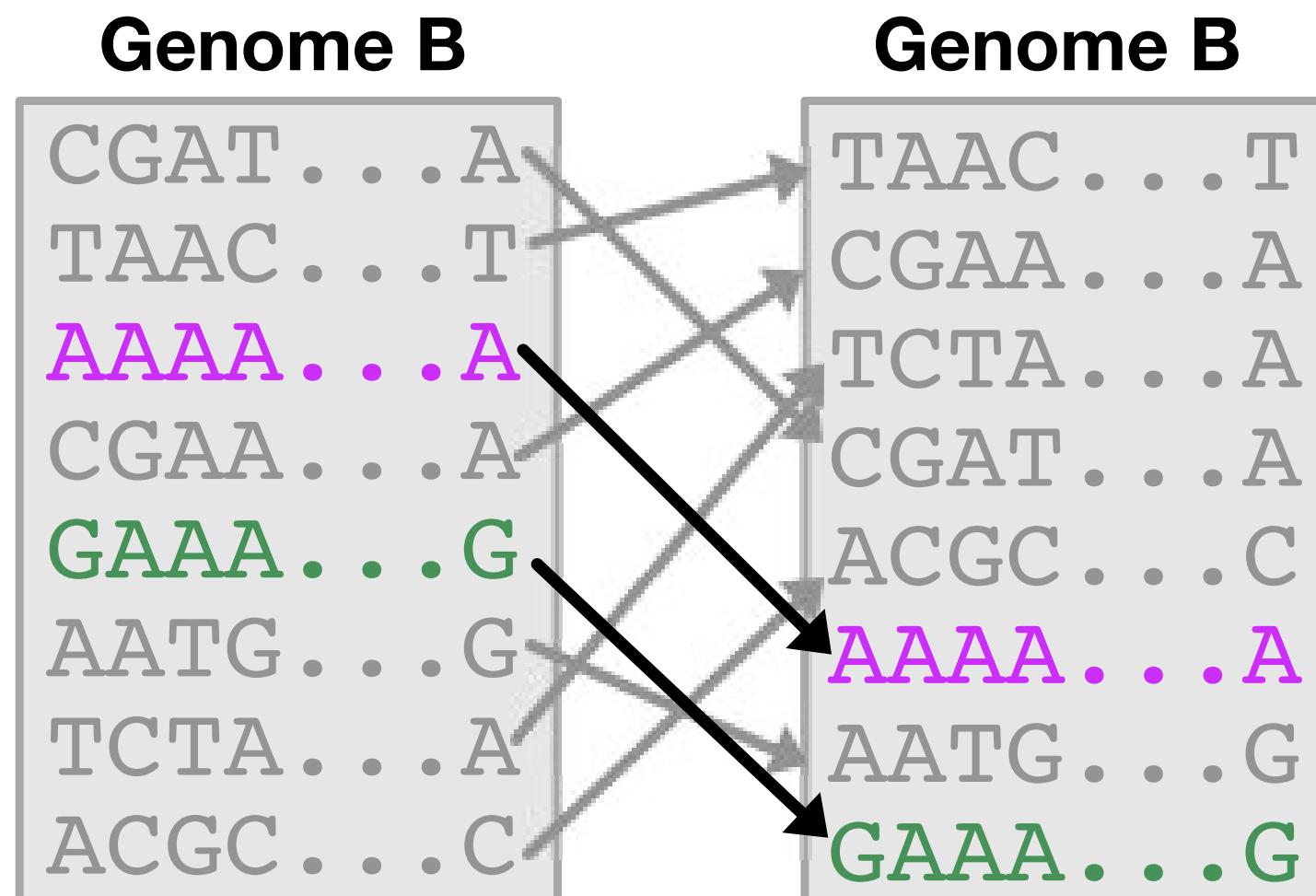


systematically  
subsample

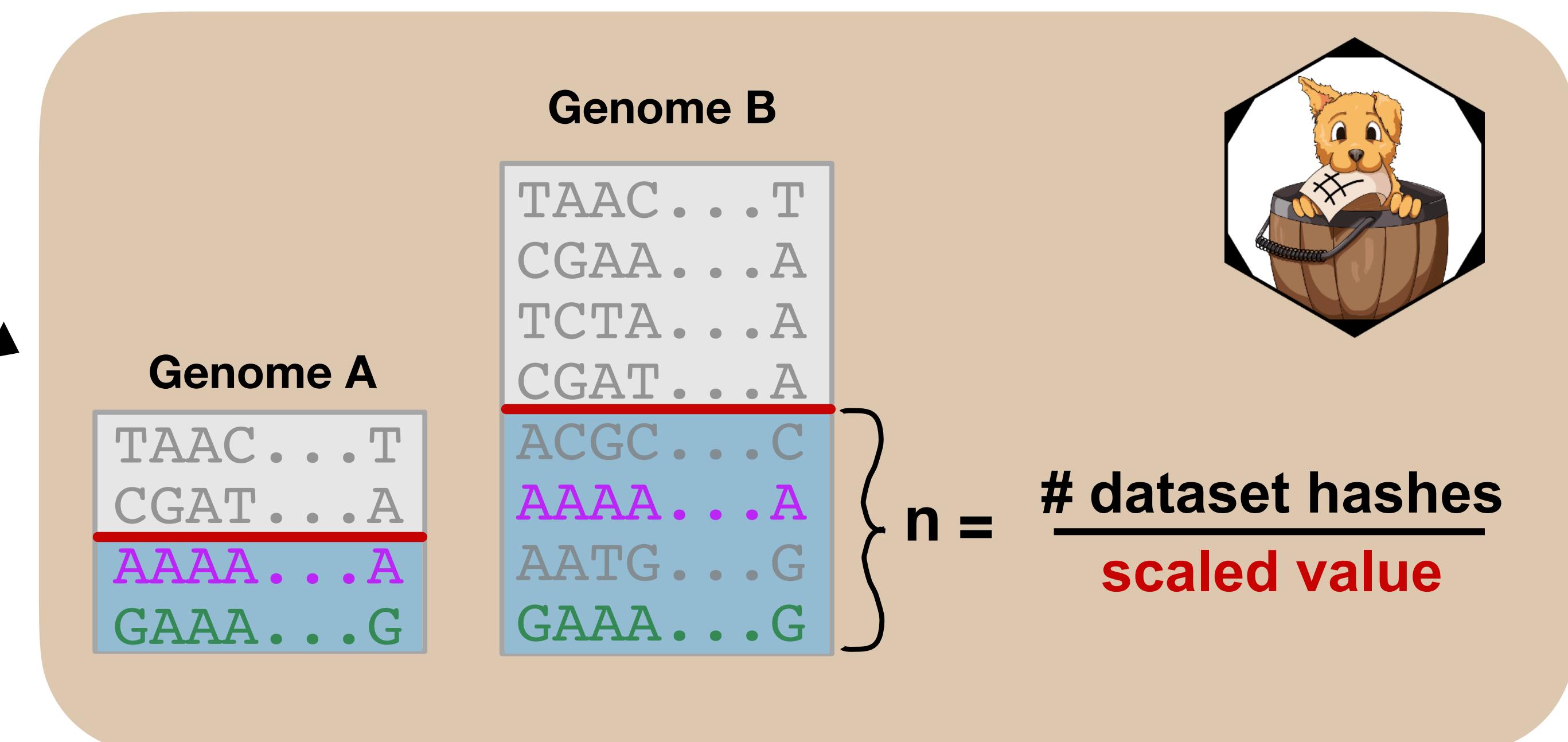


Standard  
MinHash

n = fixed value



choose  
scaled value  
scaled=2



n =  $\frac{\# \text{ dataset hashes}}{\text{scaled value}}$



# (A few) use cases for global k-mer comparisons

1. Genome similarity comparisons
2. Finding genomes in metagenomes
3. Comprehensive metagenome breakdown  
(+ taxonomic profiling)
4. Correlation with mapping (weighted overlap)

# Interactive component: sourmash



**sourmash**



[sourmash.readthedocs.io](https://sourmash.readthedocs.io)

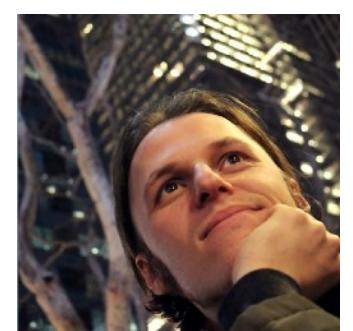


[github.com/sourmash-bio/sourmash](https://github.com/sourmash-bio/sourmash)



[gitter.im/sourmash-bio/community](https://gitter.im/sourmash-bio/community)

- k-mer analysis multitool we'll use to explore these use cases
- Flexible and tunable dataset compression
- Python command line, Rust behind the scenes for faster large-scale analyses

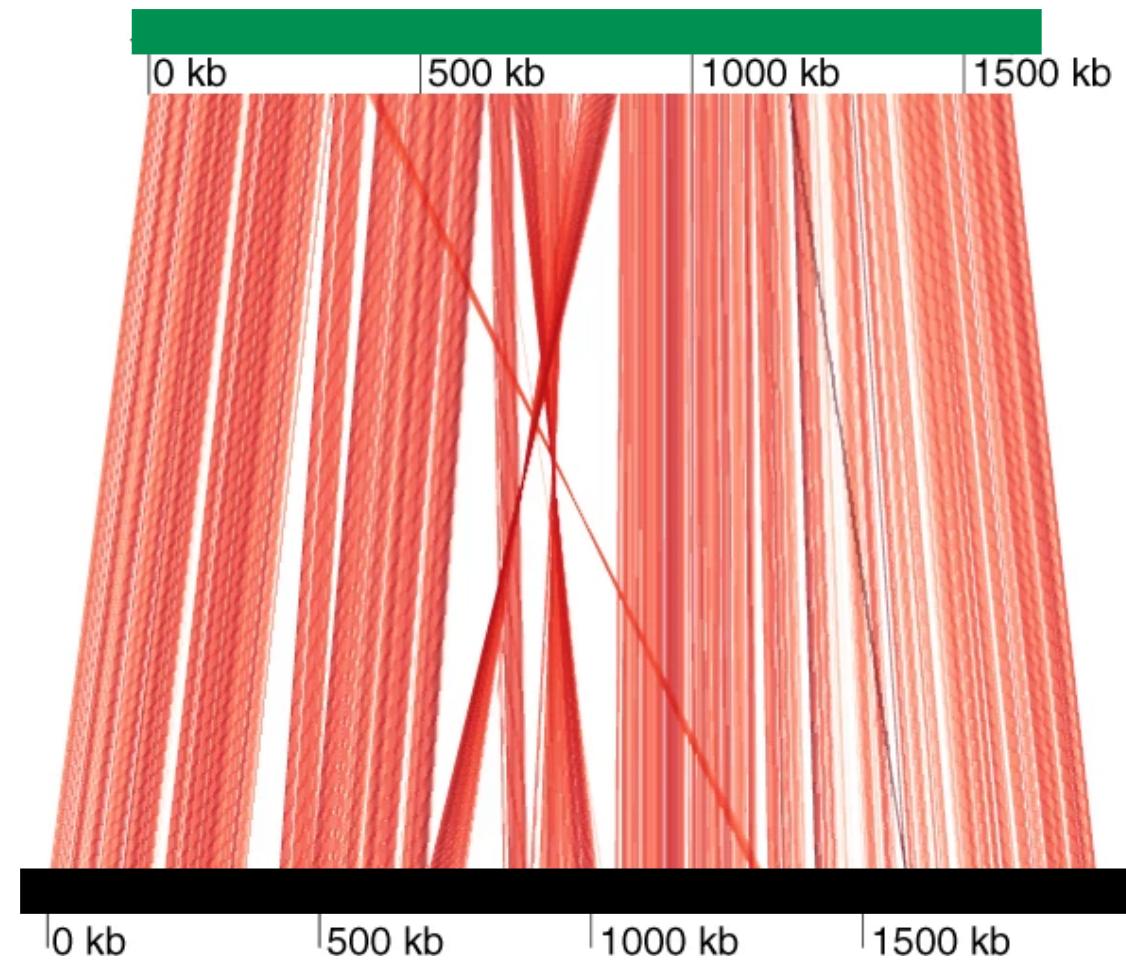


# 1. How similar are two genomes?

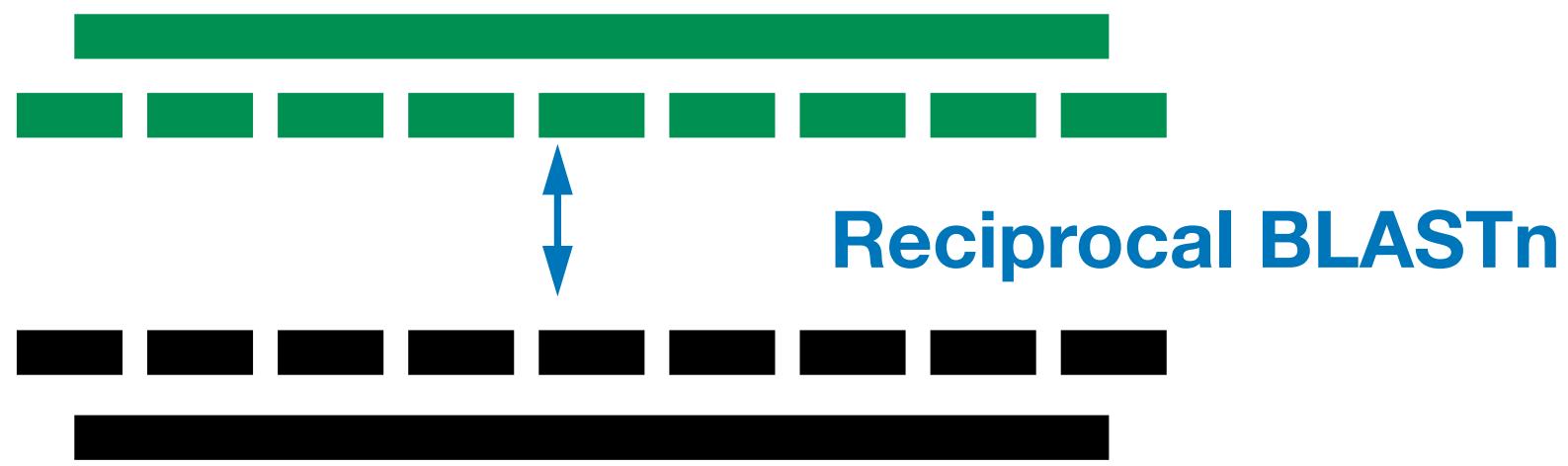
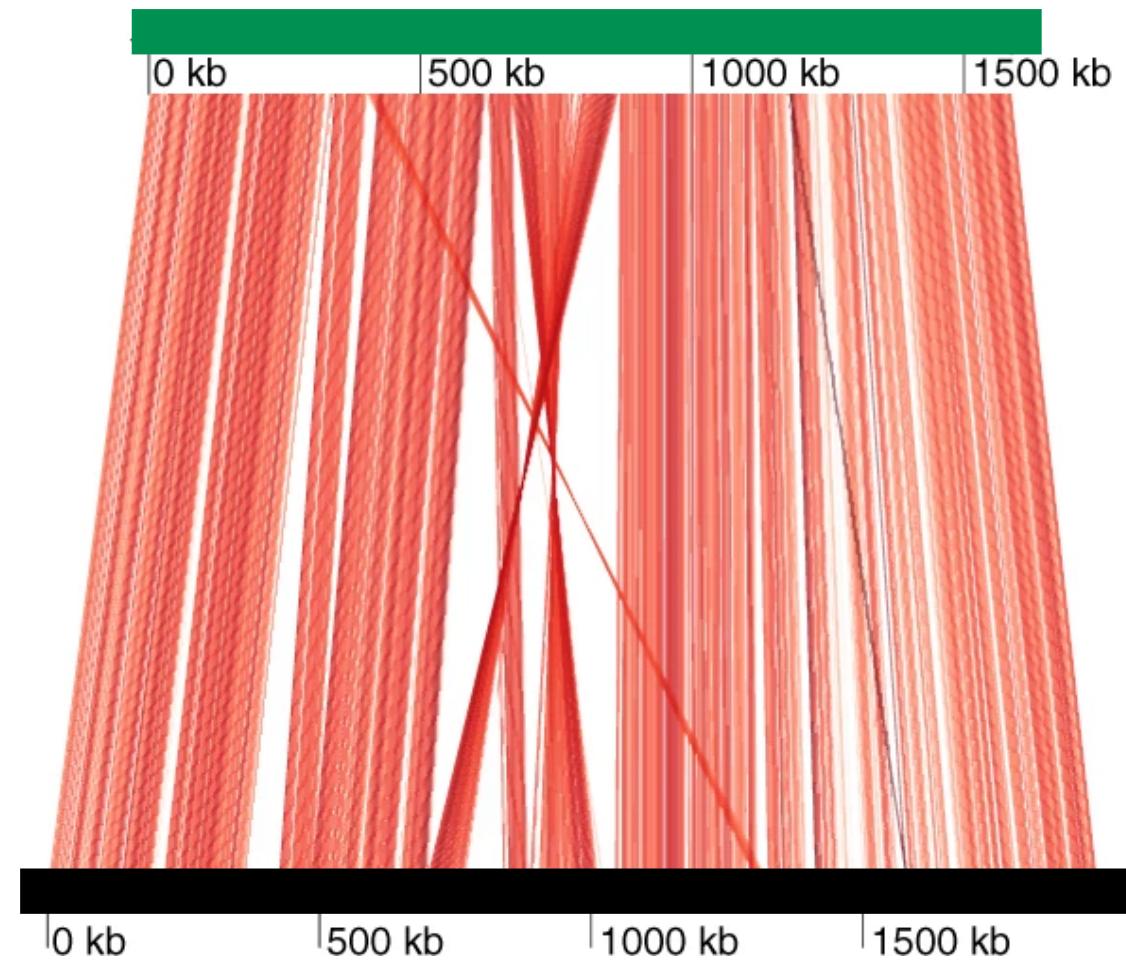
## Average Nucleotide Identity

an average measure of similarity across homologous regions in a pair of genomes

Genome A



Genome B



## Limitation:

**Alignment methods are relatively slow, not feasible for very large-scale analyses**

# Genomic Similarity with Set Comparisons

ACTACGCCCTTCATGACTC

ACTA

CTAC

TACG

ACGC

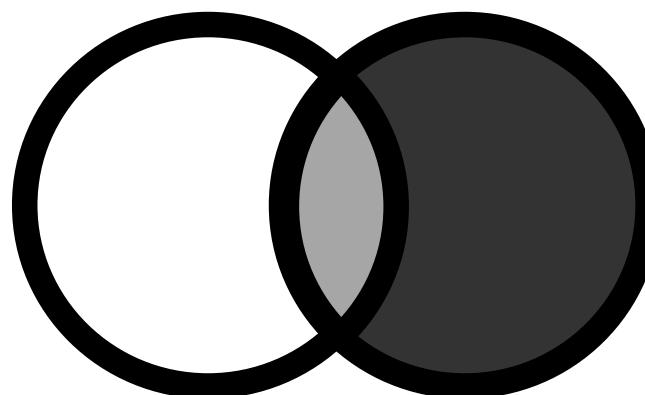
k-mers of length 4  
(4-mers)

Compare datasets  
with set operations

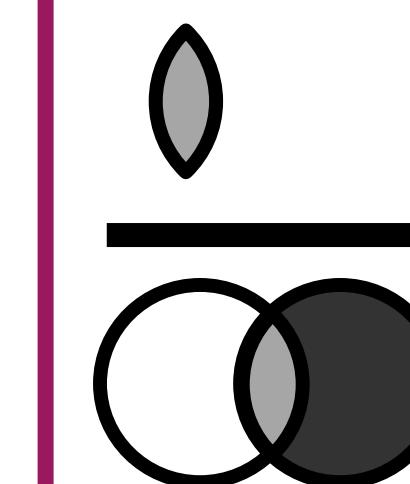
Represent entire dataset as  
collection of k-mers

Genome A

Genome B



e.g.

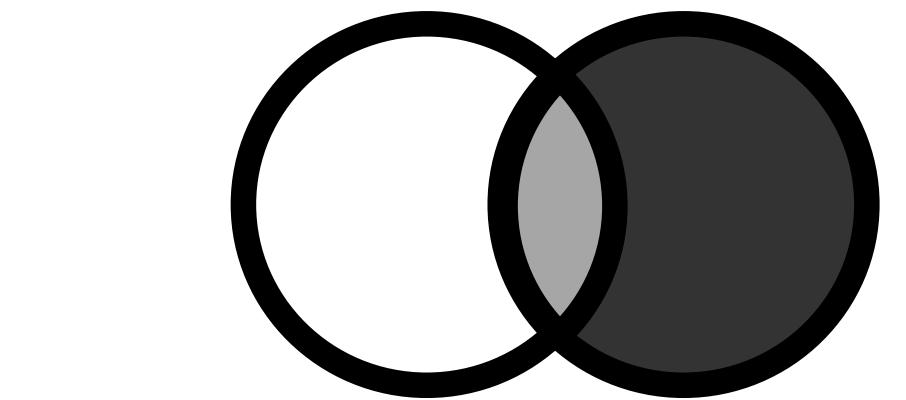


Estimate  
% shared sequence  
between A and B  
**Jaccard Index**

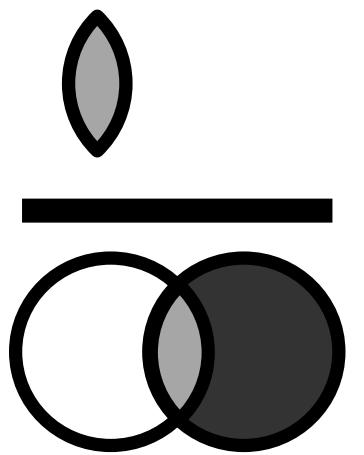
# Genomic similarity

Long k-mers ( $k=21+$ ) capture genomic (& taxonomic) similarity

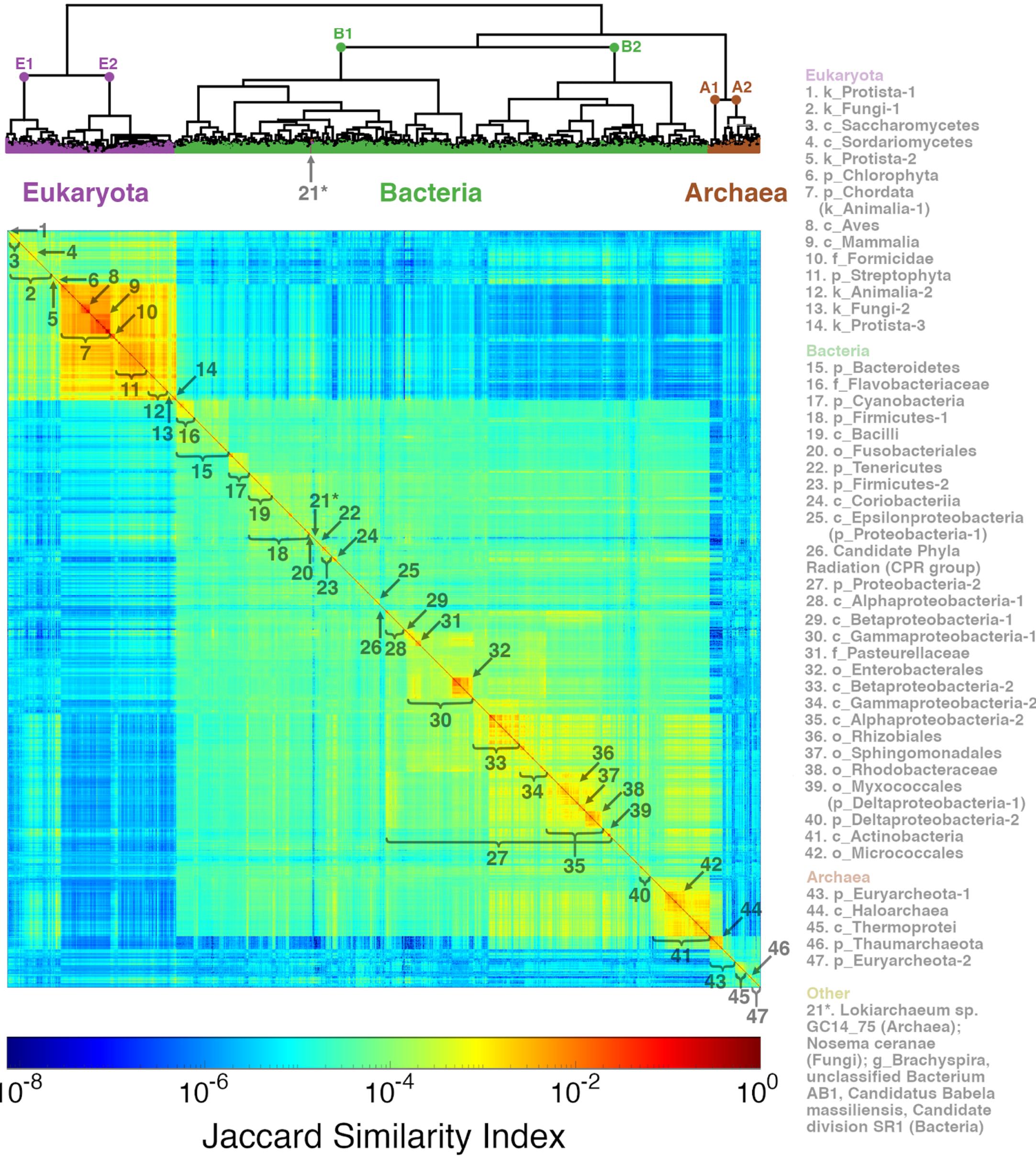
Genome A



Genome B



% shared sequence  
between A and B  
*Jaccard Index*



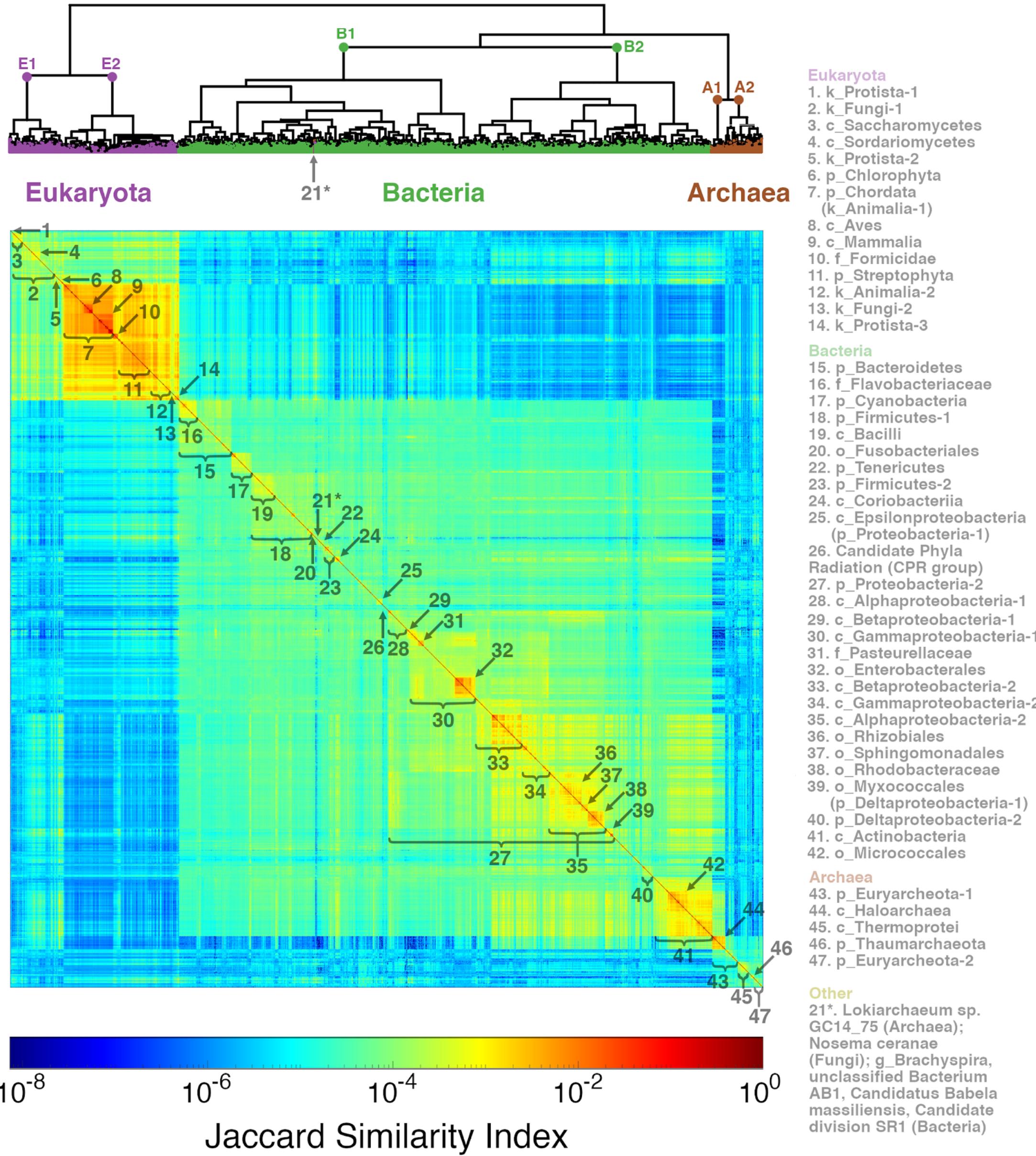
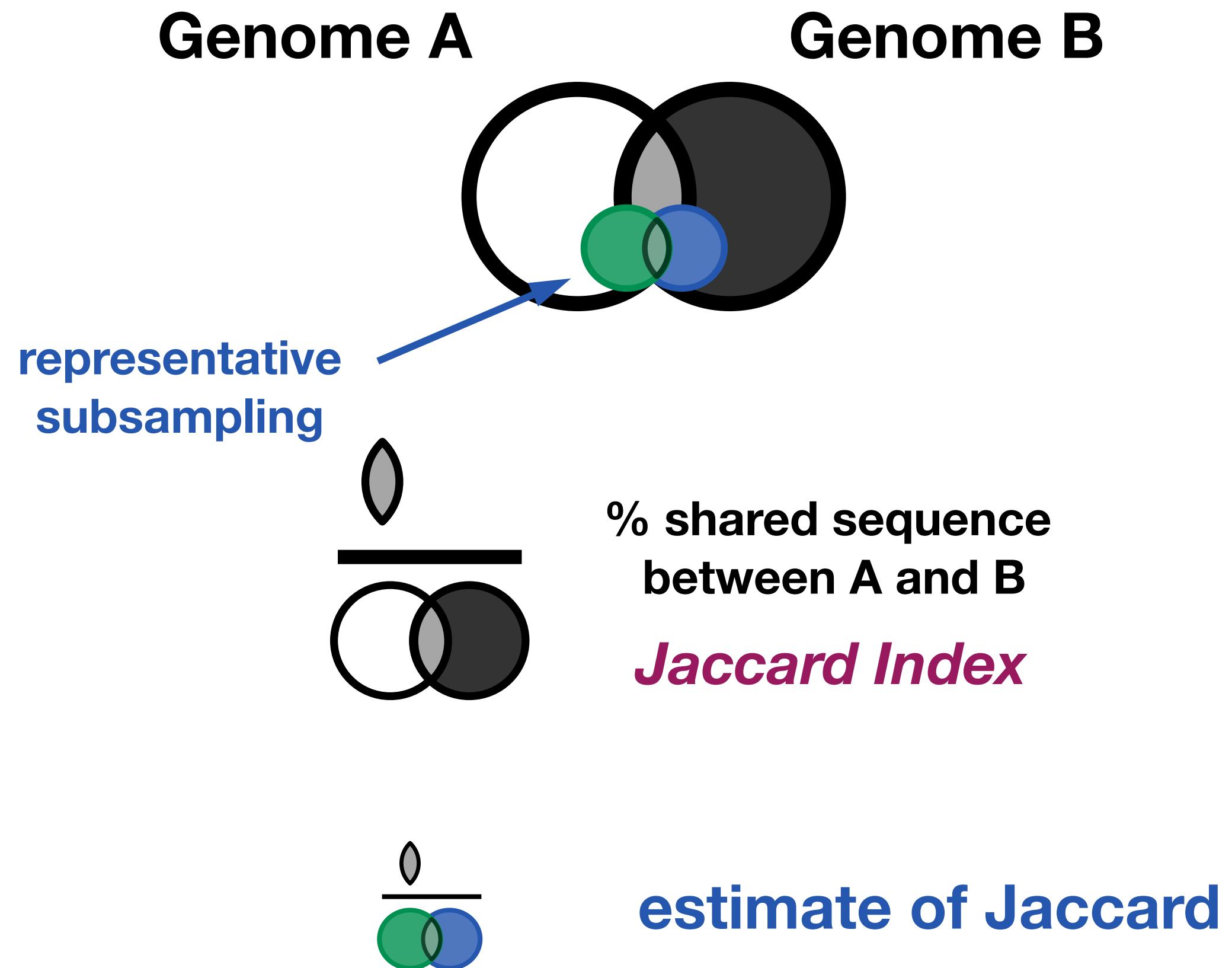
$k=21$

Bussi et al., 2021

doi: [10.1371/journal.pone.0258693](https://doi.org/10.1371/journal.pone.0258693)

# Genomic similarity

Long k-mers ( $k=21+$ ) capture genomic (& taxonomic) similarity



$k=21$

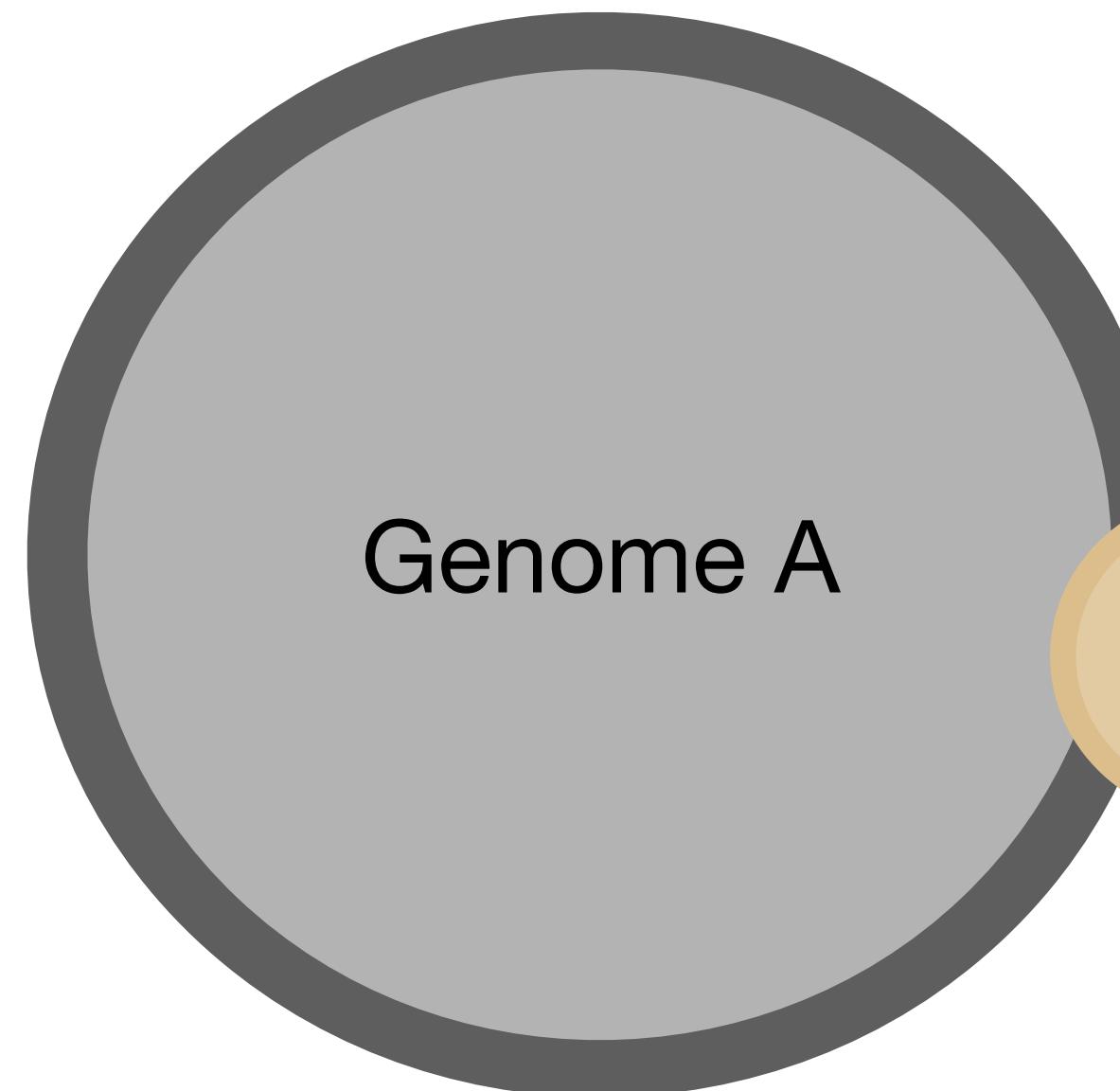
Bussi et al., 2021

doi: [10.1371/journal.pone.0258693](https://doi.org/10.1371/journal.pone.0258693)

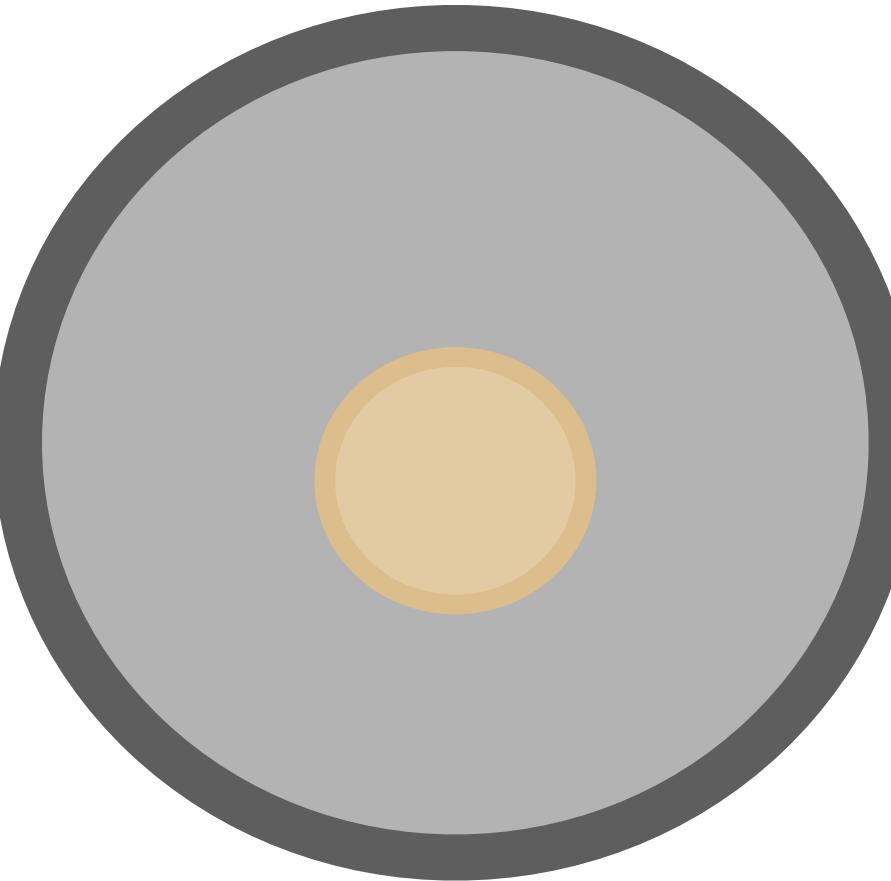
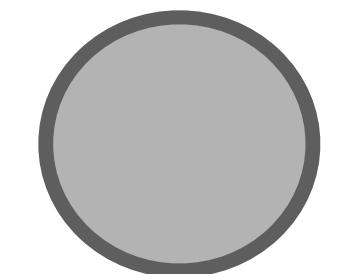
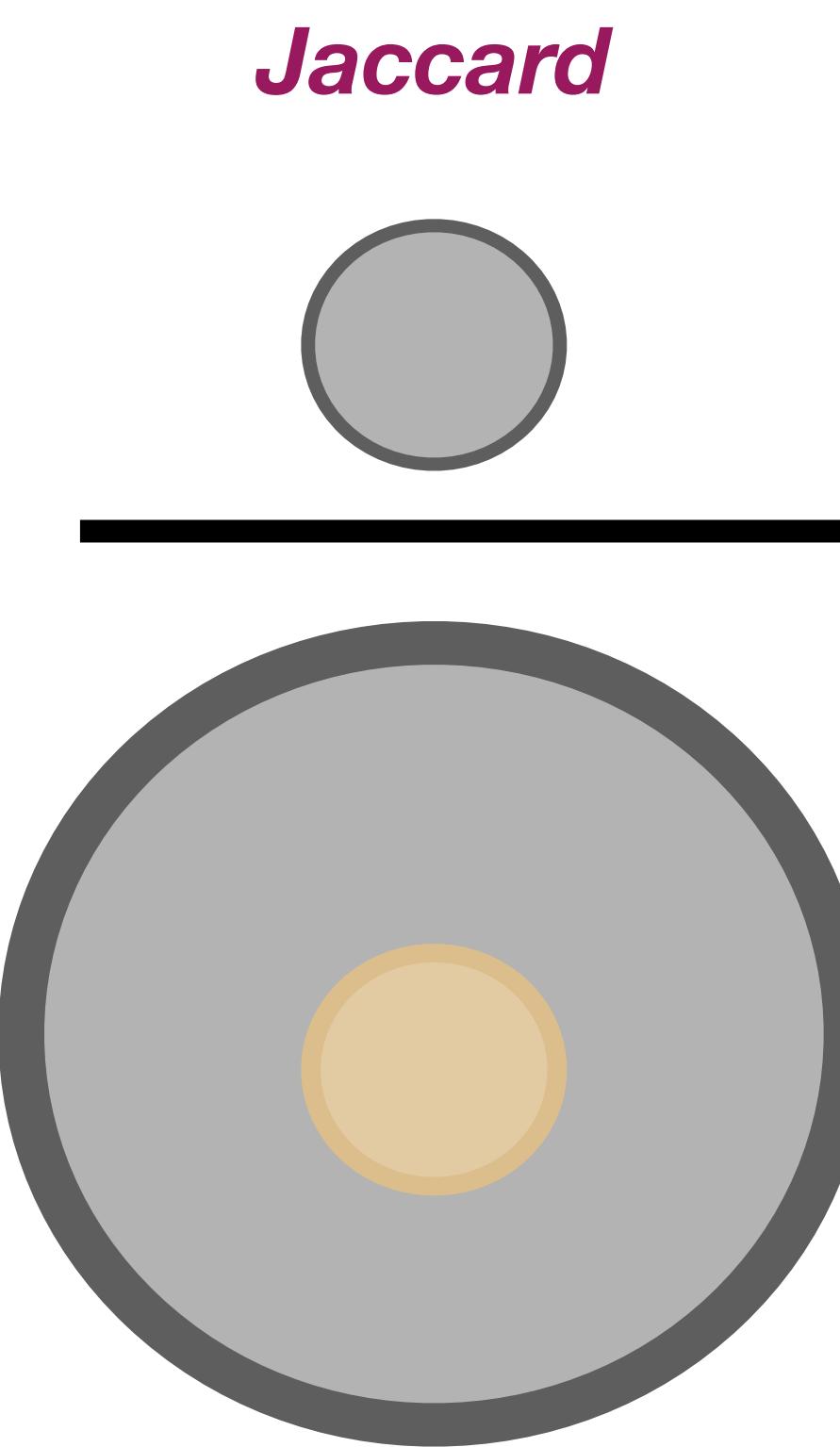
# “Similarity” (Jaccard) is not sufficient for every question

What happens when the sets are of very different sizes?

What happens when set B is entirely contained within A?

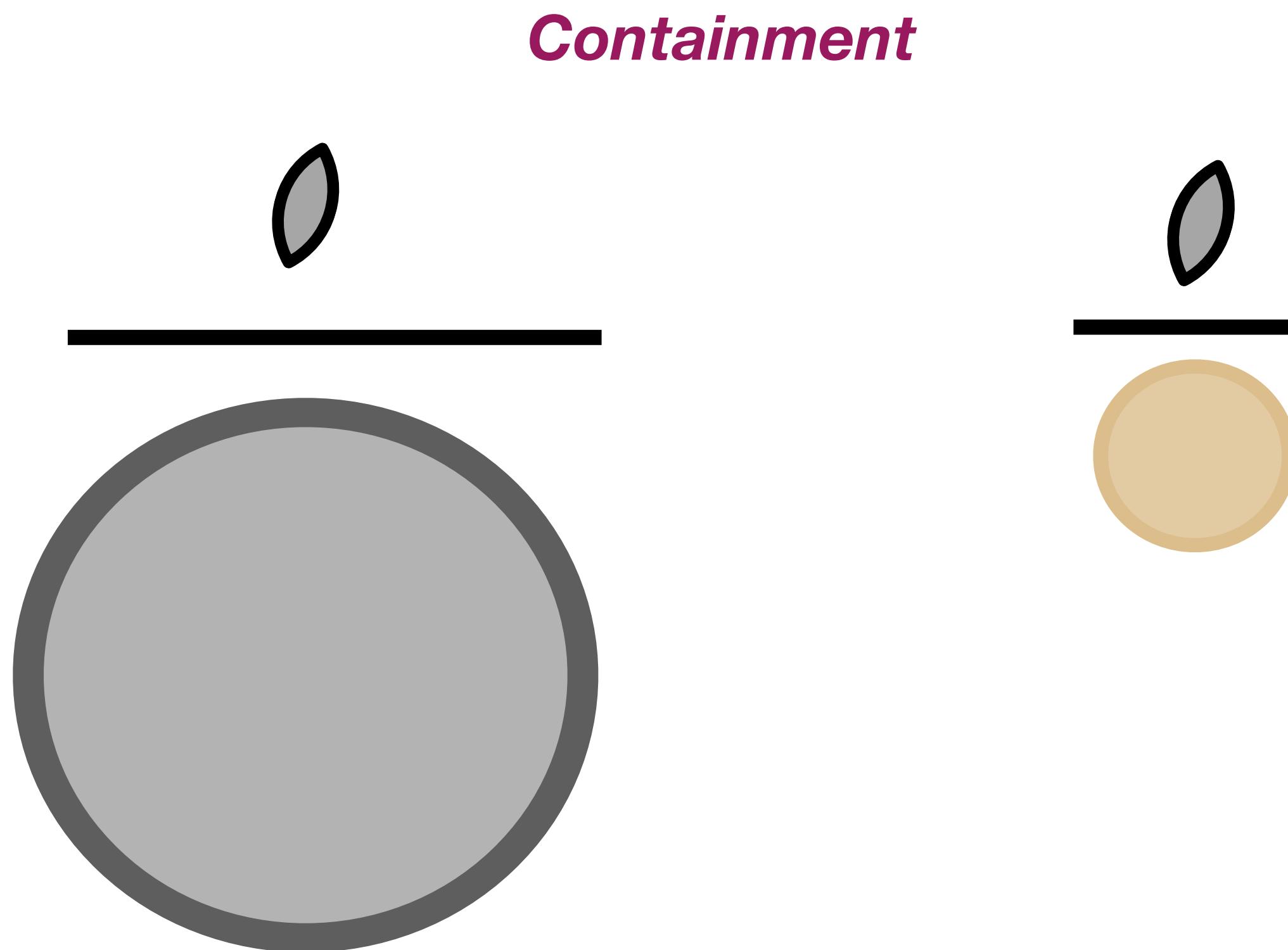


Genome B



“Set similarity” may not be as useful here

# Other set comparisons offer different utility



Containment = % of each dataset matched

- “Overlap” - the shared (intersected) k-mers
- “Containment” provides a method to ask:

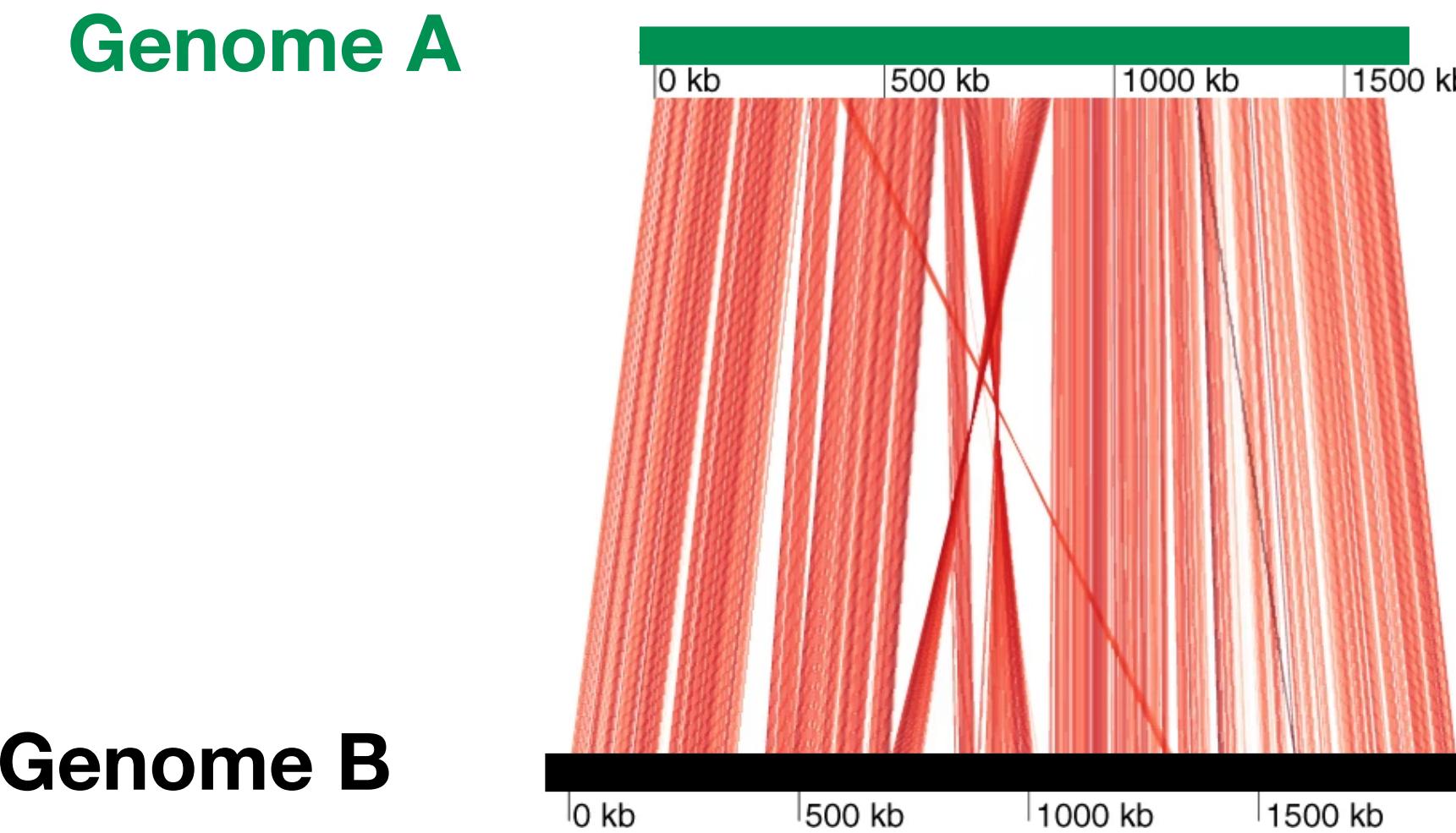
What fraction of this genome is matched?

*and for later-*  
Is this genome in my metagenome?

# Estimating Genome Similarity: ANI

## Average Nucleotide Identity

an average measure of similarity across homologous regions in a pair of genomes



Using k-mers:

ACTGGCTGACTG  
ACTGACTGACTG  
ACTG  
CTGA  
TGAC  
GACT  
ACTG  
CTGA  
TGAC

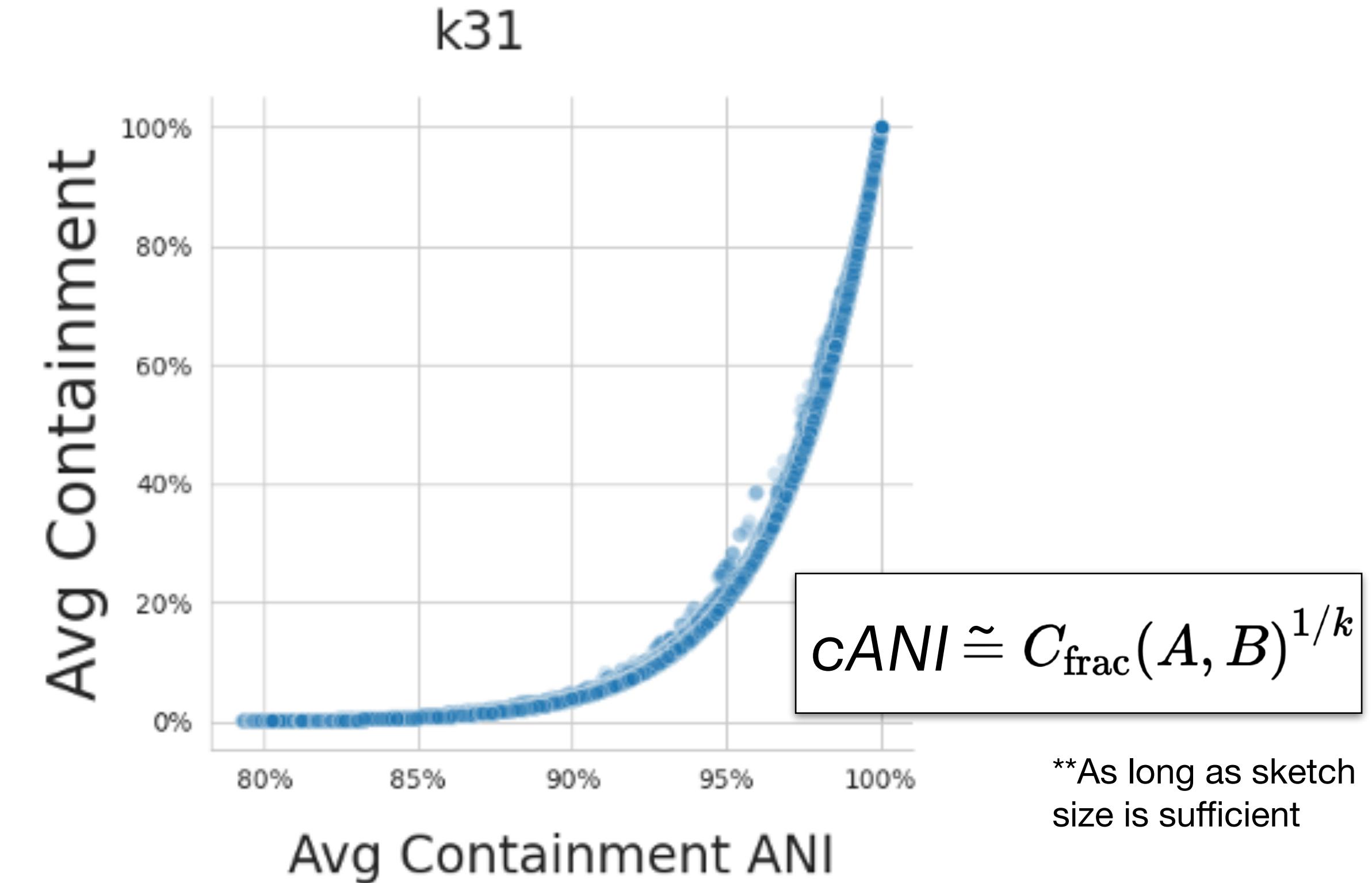
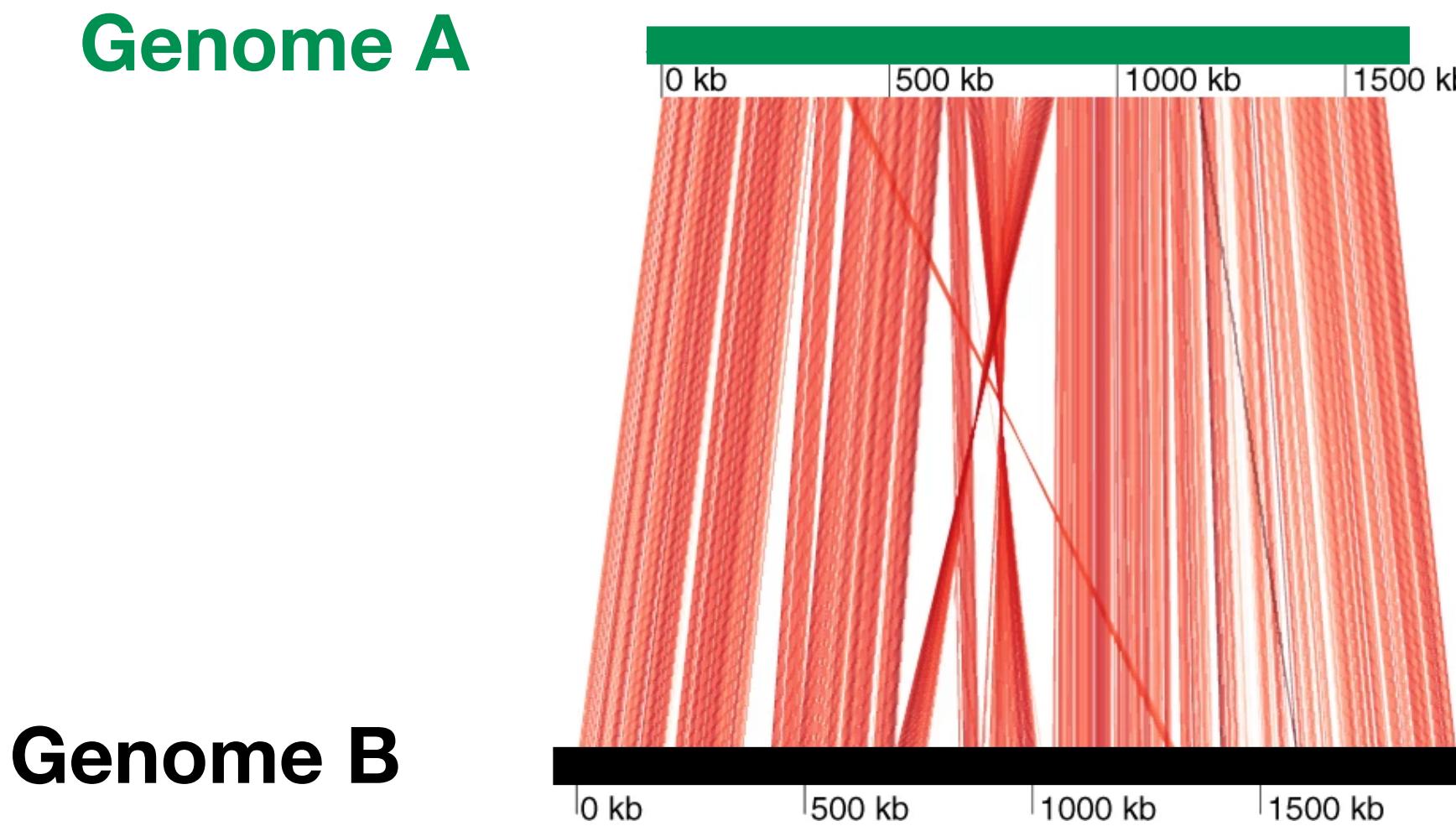
} mutated k-mers

Each SNP mutates  $k$  k-mers

# FracMinHash containment can estimate ANI

## Average Nucleotide Identity

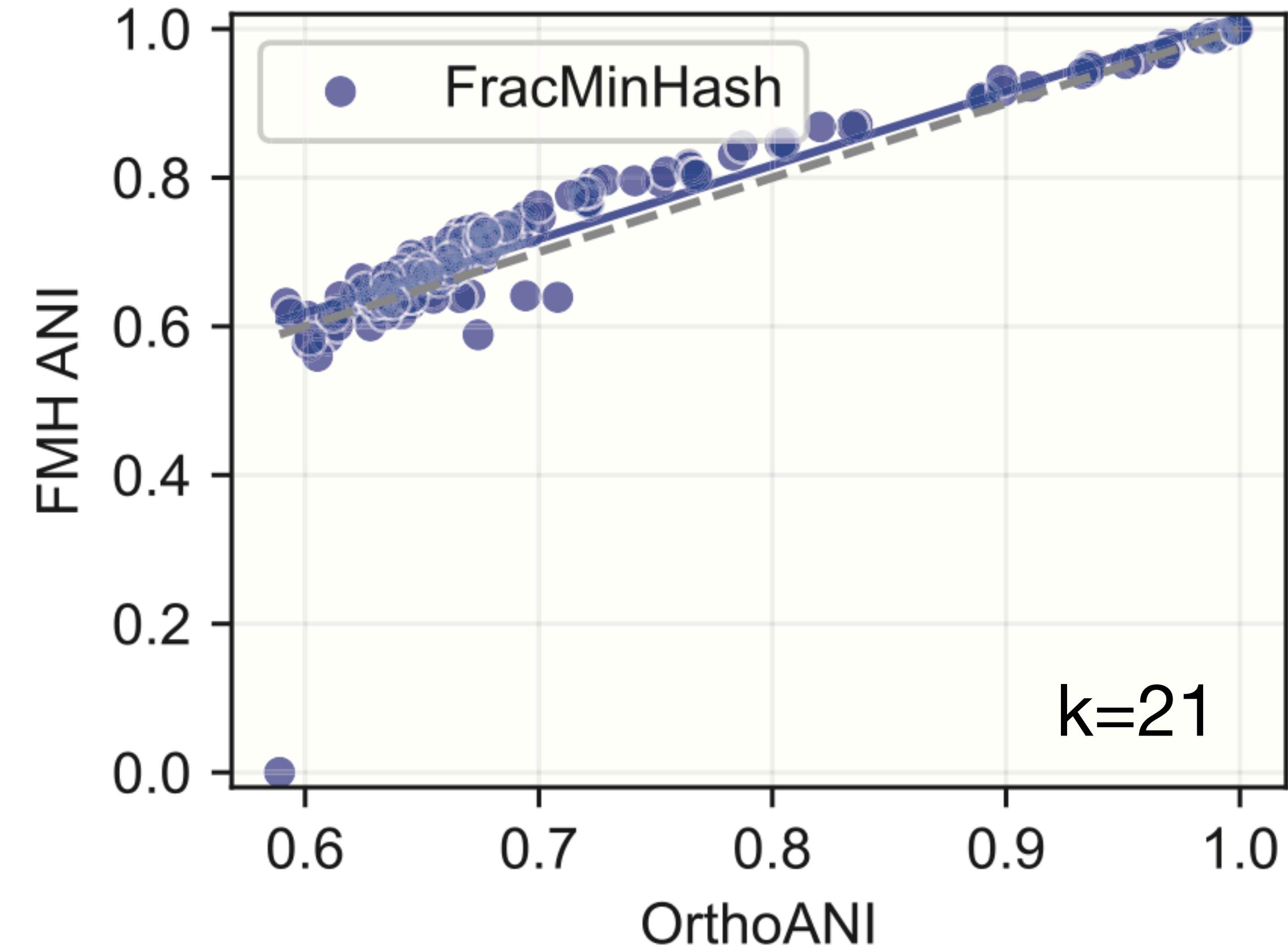
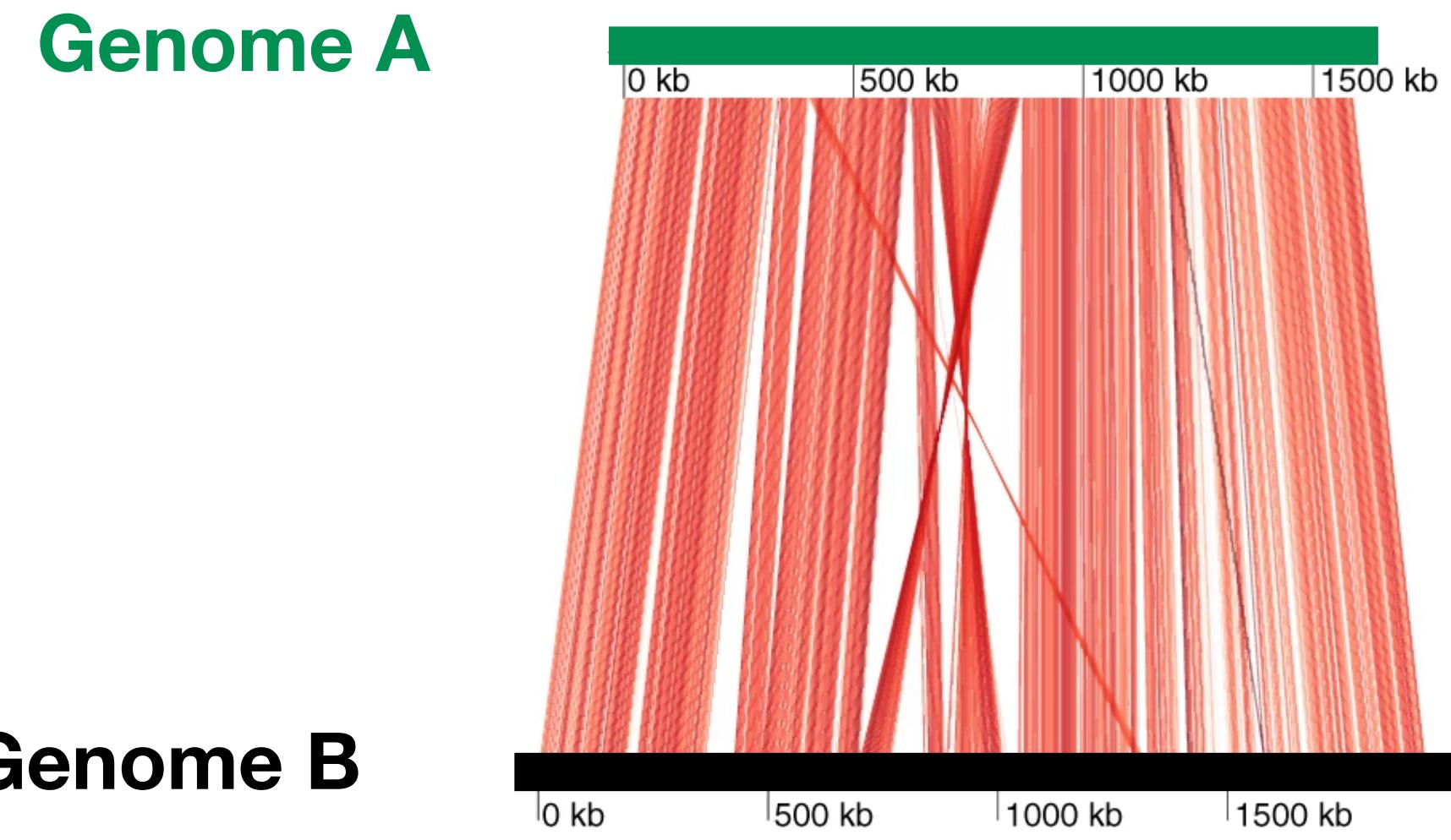
an average measure of similarity across homologous regions in a pair of genomes



# FracMinHash containment can estimate ANI

## Average Nucleotide Identity

an average measure of similarity across homologous regions in a pair of genomes



# Other distances

K-mers (with and without sketching) can be used to estimate a number of other distances

**Table 1** Definition of some classical ecological distances computed by Simka. All quantitative distances can be expressed in terms of  $C_S, f = f(x, y, X, Y)$  and  $g = g(x)$ , using the notations of Eq. (2), and computed in one pass. Qualitative ecological distances (resp. AB-variants of qualitative distances) can also be computed in a single pass over the data by computing first  $a, b$  and  $c$  (resp.  $U$  and  $V$ ). See main text for the definition of  $a, b, c, U$  and  $V$ .

Name	Definition	$C_{S_i}$	$f(x, y, X, Y)$	$g(x)$
<b>Quantitative distances</b>				
Chord	$\sqrt{2 - 2 \sum_w \frac{N_{S_i}(w)N_{S_j}(w)}{C_{S_i}C_{S_j}}}$	$\sqrt{\sum_w N_{S_i}(w)^2}$	$\frac{xy}{XY}$	$\sqrt{2 - 2x}$
Hellinger	$\sqrt{2 - 2 \sum_w \frac{\sqrt{N_{S_i}(w)N_{S_j}(w)}}{\sqrt{C_{S_i}C_{S_j}}}}$	$\sum_w N_{S_i}(w)$	$\frac{\sqrt{xy}}{\sqrt{XY}}$	$\sqrt{2 - 2x}$
Whittaker	$\frac{1}{2} \sum_w \left  \frac{N_{S_i}(w)C_{S_j} - N_{S_j}(w)C_{S_i}}{C_{S_i}C_{S_j}} \right $	$\sum_w N_{S_i}(w)$	$\frac{ xY - yX }{XY}$	$\frac{x}{2}$
Bray–Curtis	$1 - 2 \sum_w \frac{\min(N_{S_i}(w), N_{S_j}(w))}{C_{S_i} + C_{S_j}}$	$\sum_w N_{S_i}(w)$	$\frac{\min(x, y)}{x+y}$	$1 - 2x$
Kulczynski	$1 - \frac{1}{2} \sum_w \frac{(C_{S_i} + C_{S_j})\min(N_{S_i}(w), N_{S_j}(w))}{C_{S_i}C_{S_j}}$	$\sum_w N_{S_i}(w)$	$\frac{(X+Y)\min(x, y)}{XY}$	$1 - \frac{x}{2}$
Jensen–Shannon	$\sqrt{\frac{1}{2} \sum_w \left[ \frac{N_{S_i}(w)}{C_{S_i}} \log \frac{2C_{S_j}N_{S_i}(w)}{C_{S_j}N_{S_i}(w) + C_{S_i}N_{S_j}(w)} + \frac{N_{S_j}(w)}{C_{S_j}} \log \frac{2C_{S_i}N_{S_j}(w)}{C_{S_j}N_{S_i}(w) + C_{S_i}N_{S_j}(w)} \right]}$	$\sum_w N_{S_i}(w)$	$\frac{x}{X} \log \frac{2xY}{xY+yX} + \frac{y}{Y} \log \frac{2yX}{xY+yX}$	$\sqrt{\frac{x}{2}}$
Canberra	$\frac{1}{a+b+c} \sum_w \left  \frac{N_{S_i}(w) - N_{S_j}(w)}{N_{S_i}(w) + N_{S_j}(w)} \right $	—	$\left  \frac{x-y}{x+y} \right $	$\frac{1}{a+b+c}x$
<b>Qualitative distances</b>				
Chord/Hellinger	$\sqrt{2 \left( 1 - \frac{a}{\sqrt{(a+b)(a+c)}} \right)}$	—	—	—
Whittaker	$\frac{1}{2} \left( \frac{b}{a+b} + \frac{c}{a+c} + \left  \frac{a}{a+b} - \frac{a}{a+c} \right  \right)$	—	—	—
Bray–Curtis/Sorensen	$\frac{b+c}{2a+b+c}$	—	—	—
Kulczynski	$1 - \frac{1}{2} \left( \frac{a}{a+b} + \frac{a}{a+c} \right)$	—	—	—
Ochiai	$1 - \frac{a}{\sqrt{(a+b)(a+c)}}$	—	—	—
Jaccard	$\frac{b+c}{a+b+c}$	—	—	—
<b>Abundance-based (AB) variants of qualitative distances</b>				
AB-Jaccard	$1 - \frac{UV}{U+V-UV}$	—	—	—
AB-Ochiai	$1 - \sqrt{UV}$	—	—	—
AB-Sorensen	$1 - \frac{2UV}{U+V}$	—	—	—

# Let's try with sourmash...

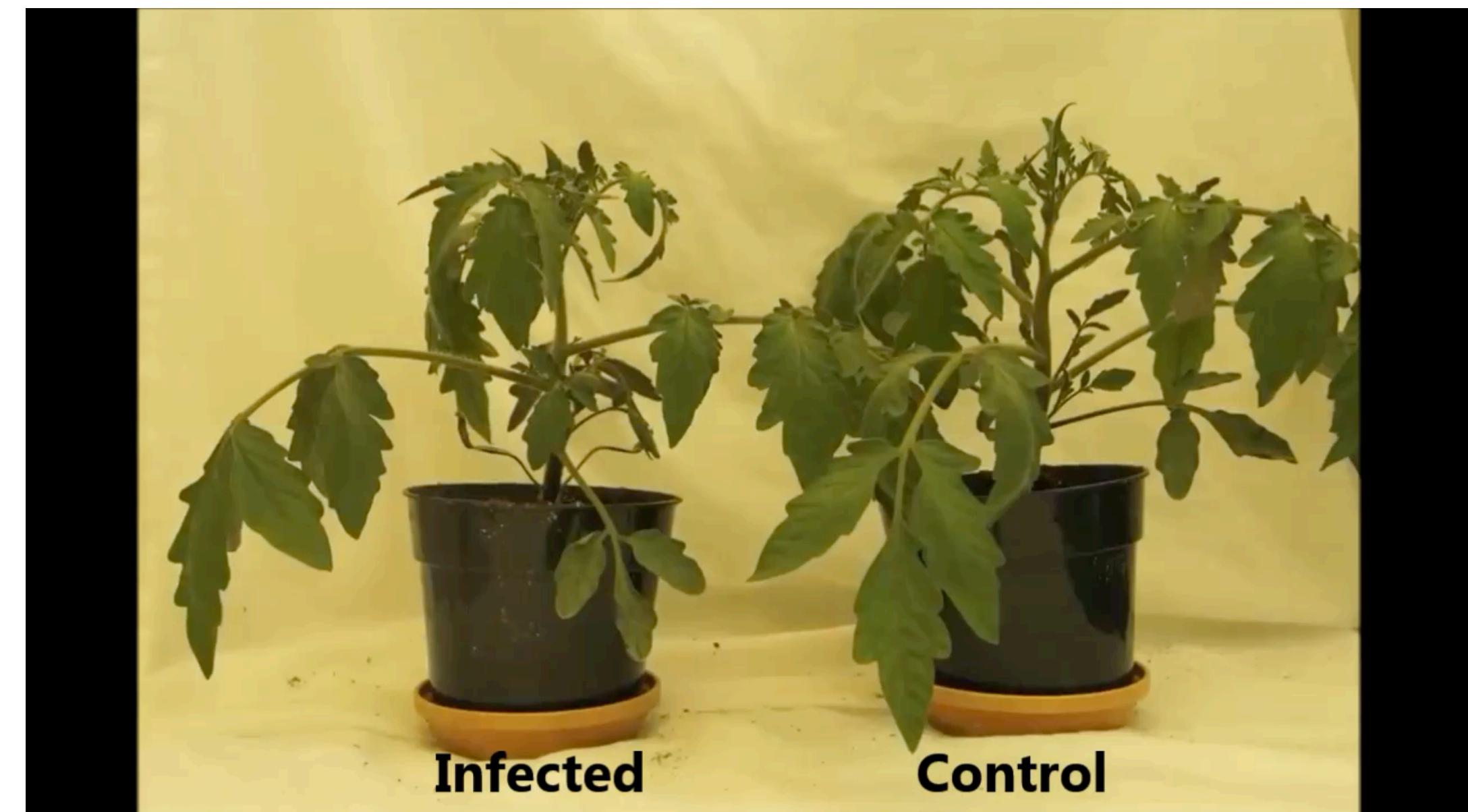
Test genomes: ***Ralstonia solanacearum***  
species complex



Tuan  
Tran

## Plant pathogen

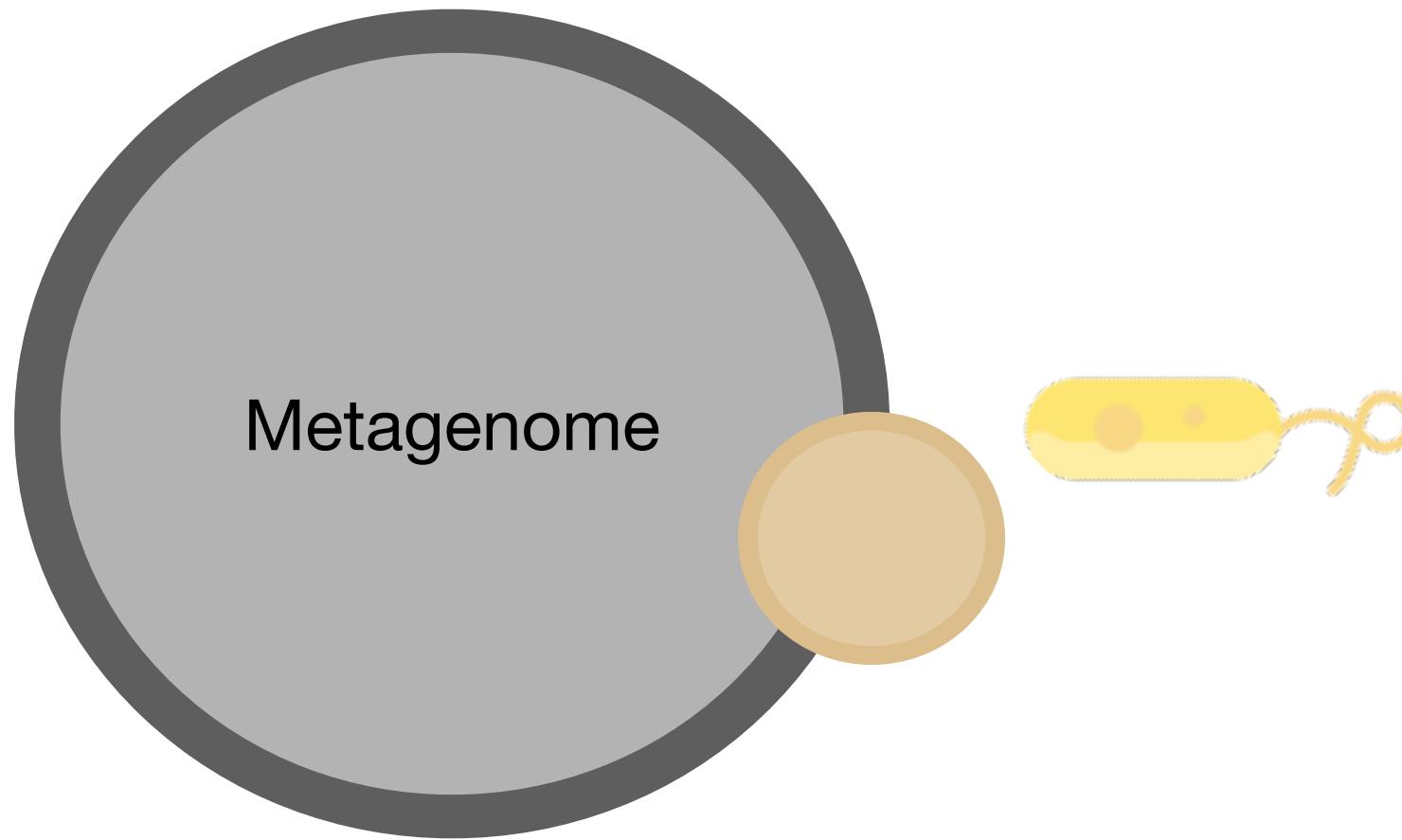
- Bacterial wilt disease/ Brown rot
- Wide host range, >200 plant species
- Global distribution; High socioeconomic cost



Filmed by Jon Jacobs

# 2. Finding genome(s) in metagenomes

Does this metagenome contain my organism of interest?



# Metagenomes are sequence representations of communities

*The study of **the structure and function of entire nucleotide sequences** isolated and analyzed from all the organisms (typically microbes) in a bulk sample (NHGRI).*

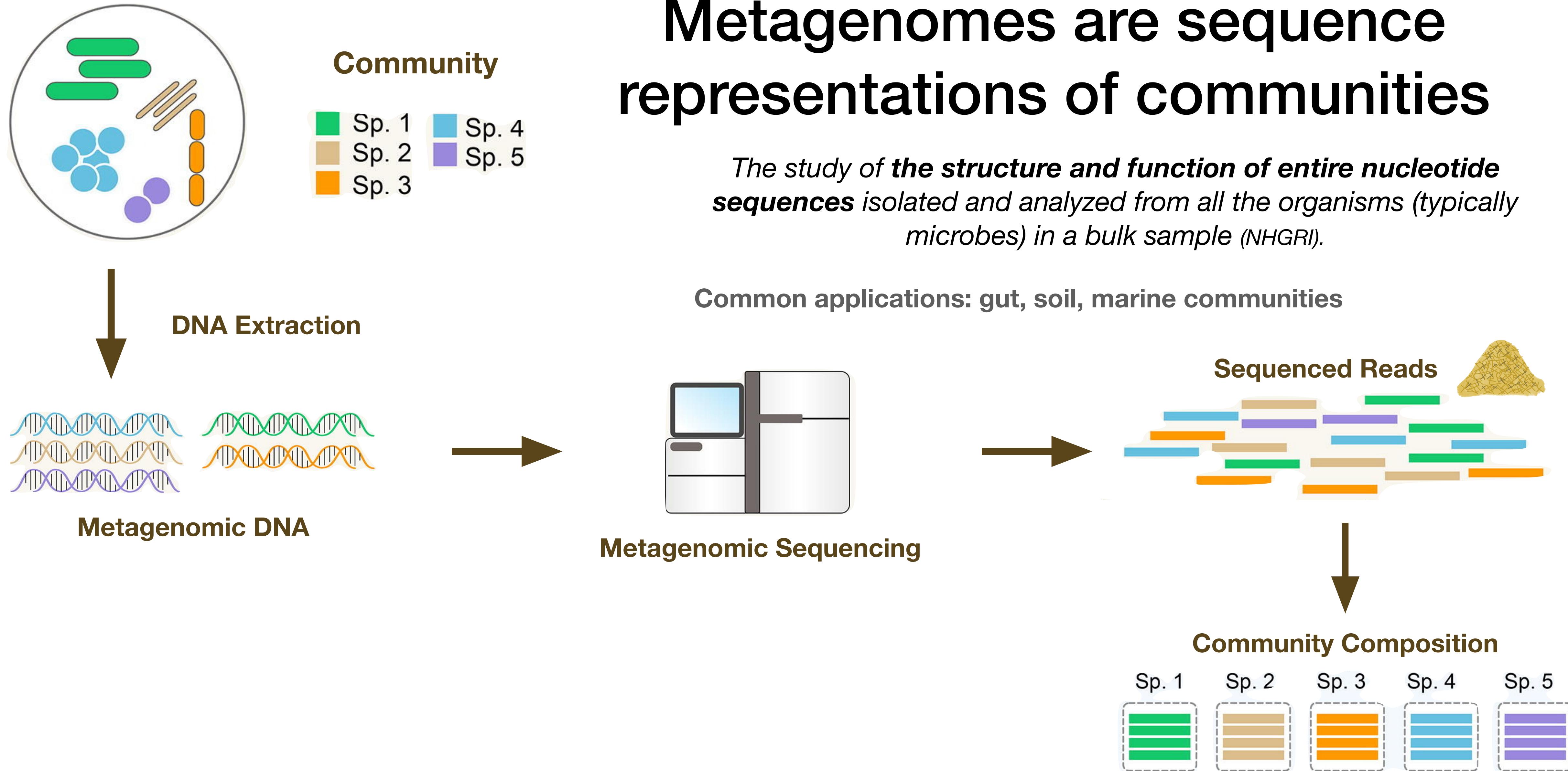
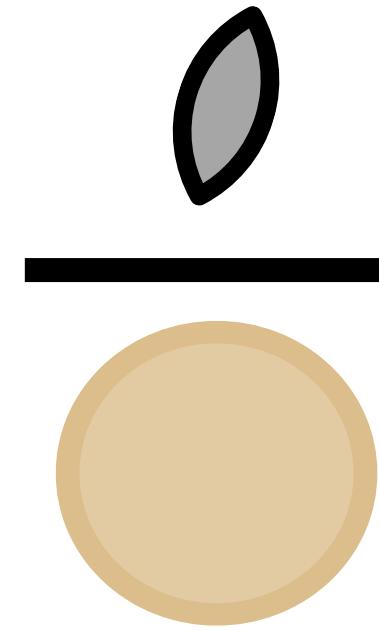


figure modified from  
[10.1016/j.csbj.2021.11.028](https://doi.org/10.1016/j.csbj.2021.11.028)

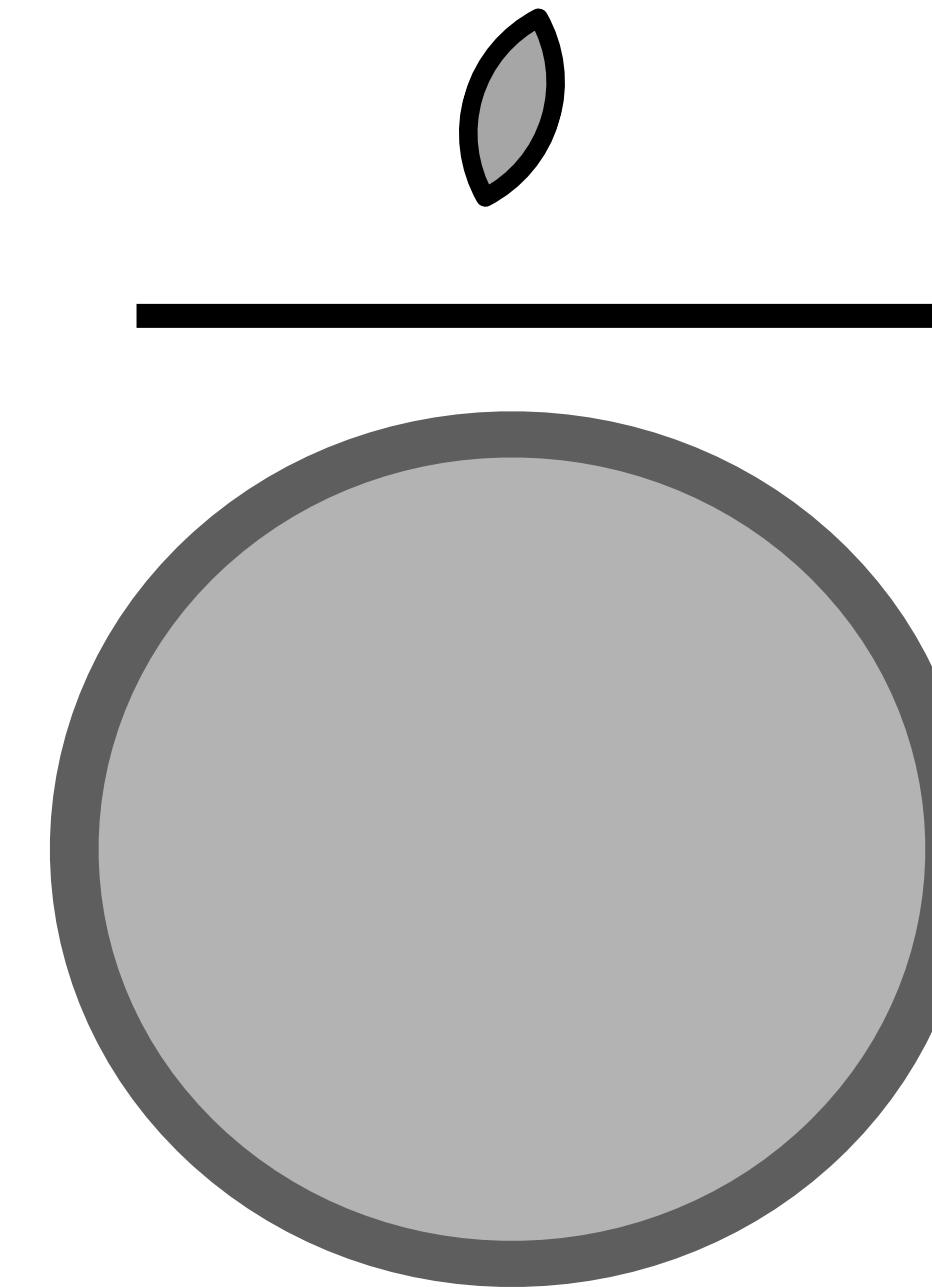
# Containment for genome-metagenome comparisons

## *Containment*



How much of query genome is matched?

(Is my organism in this community?)

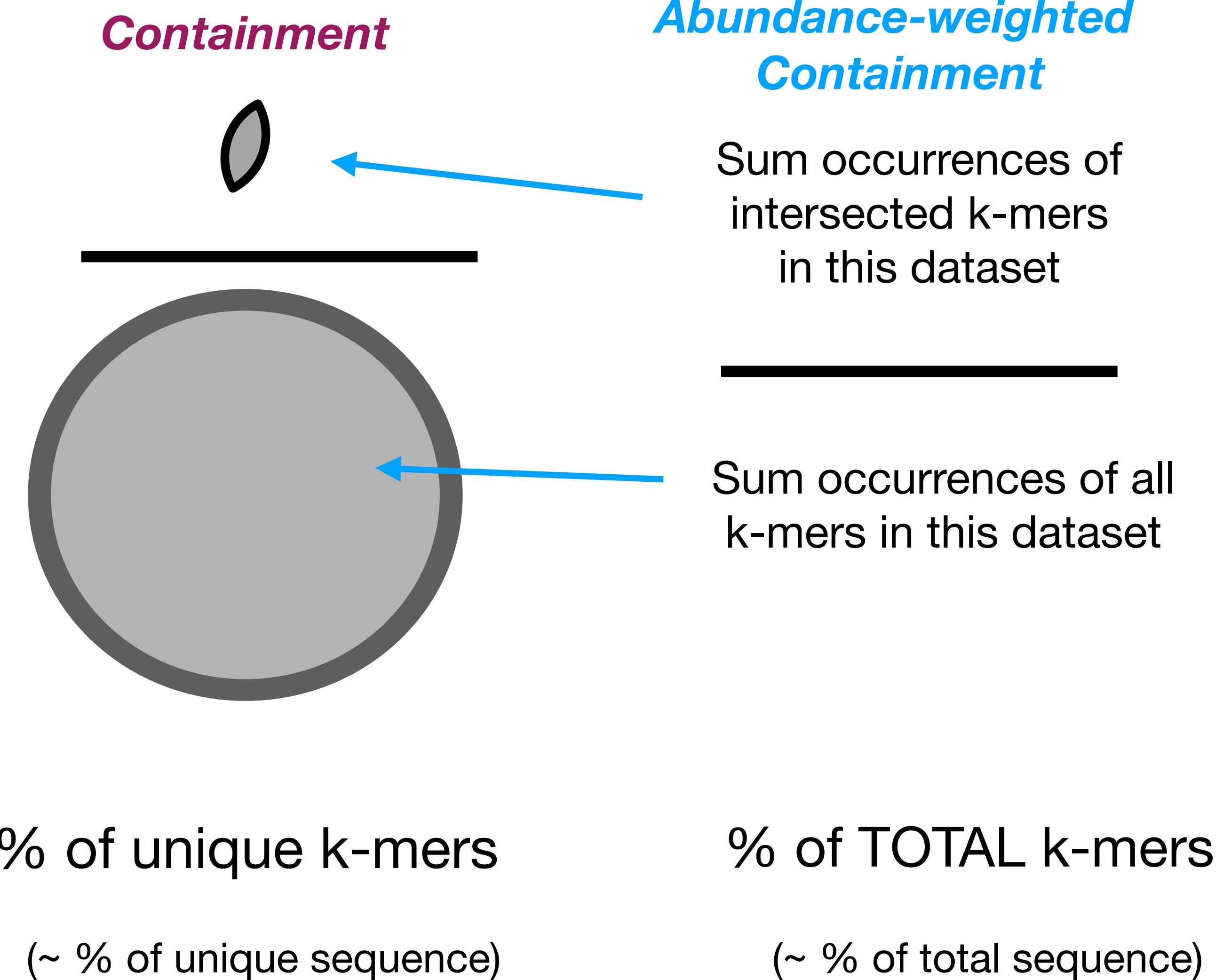


What fraction of the community does it make up?

**“Containment”:** % of each dataset matched

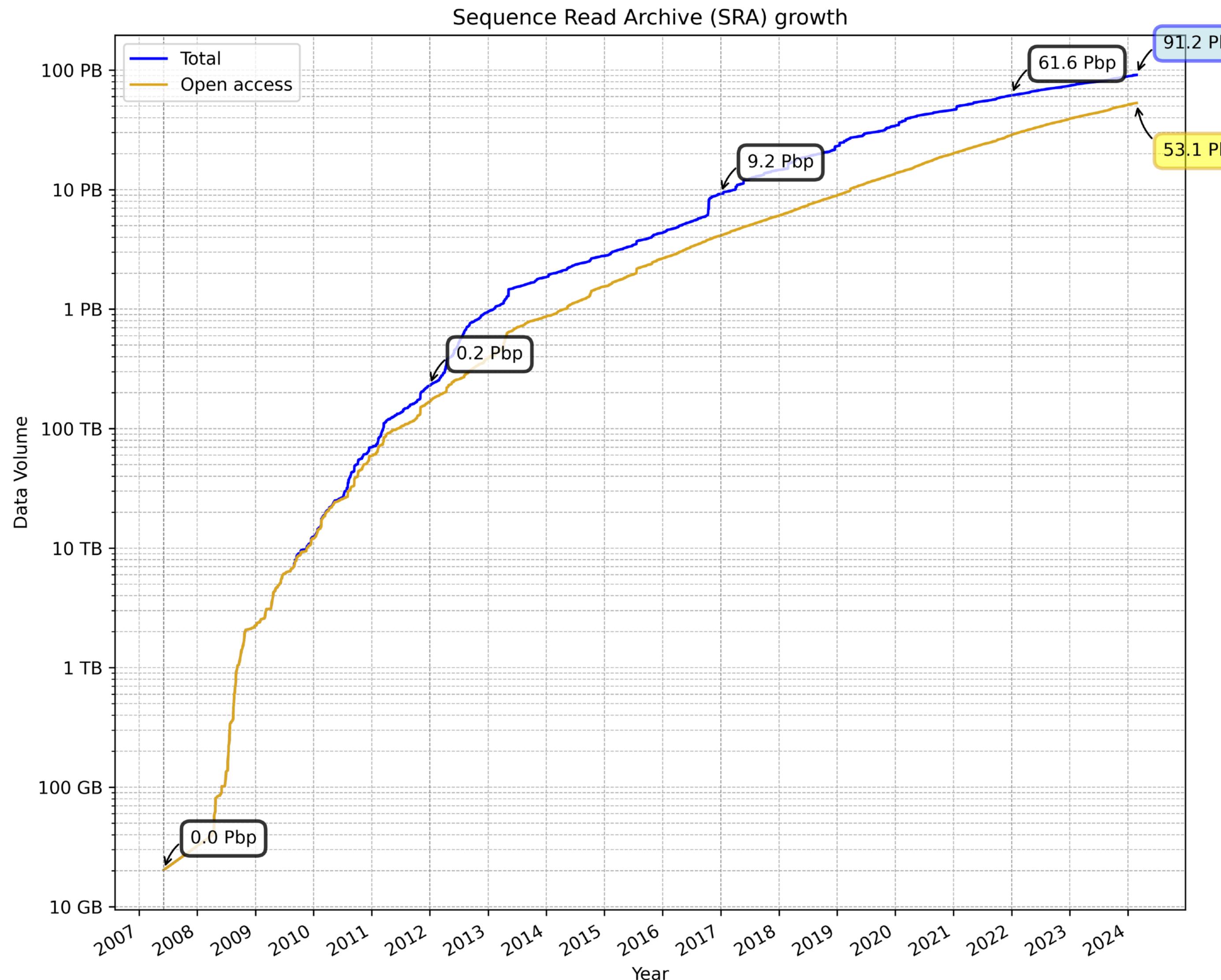
# Abundance-weighted comparisons

- Jaccard, Containment, etc use k-mer identities - i.e. what fraction of the unique k-mer sequence did we match?
- But in metagenomes, metatranscriptomes, transcriptomes, etc: k-mer *multiplicity*, or *abundance*, is also important



# **Hands-on part 2: finding a genome in metagenome**

# Finding genome(s) in metagenomes ... at scale!



~53 Pb Open Access  
(May 2025)

Includes > 1.2 million  
metagenomes

( and many added daily!)

# Branchwater Metagenome Query

Real-time search for a genome within metagenomes in the SRA.

[Home](#) [Advanced](#) [About](#) [Examples](#) [Contact](#)

## Try out the search!

Submit one of the genomes below to examine its potential SRA metagenome matches and explore the default metadata options.

Refresh the page to try a different genome.

### *Ralstonia solanacearum*

- Ralstonia solanacearum*: a soilborne bacterial pathogen that causes bacterial wilt in several crops.  
[RefSeq:GCF\\_021117135.1](#)
- Salmonella enterica* subsp. *enterica*: a widespread bacterial pathogen that causes salmonellosis in humans.  
[\(RefSeq:GCF\\_000006945.2\)](#)
- Prochlorococcus marinus*: a widespread and abundant marine cyanobacteria. ([RefSeq:GCF\\_000015665.1](#))
- Candida albicans* : a fungi common in the human gut and other parts of the body. It is an opportunistic pathogen and causes infection under certain conditions. ([RefSeq:GCF\\_000182965.3](#))
- Aspergillus sydowii* : a fungal pathogen that can cause disease in humans and sea fan corals.  
[\(RefSeq:GCF\\_001890705.1\)](#)
- Candidatus Pelagibacter ubique* : ubiquitous marine bacterium SAR11, strain HTCC1062. ([RefSeq:GCF\\_000012345.1](#))

[Submit](#)

# Finding *Ralstonia* in public microbiome data

<https://branchwater.jgi.doe.gov>

**Search >1.16million SRA metagenomes**

**Results in < 30 seconds**

web server  
collaborators

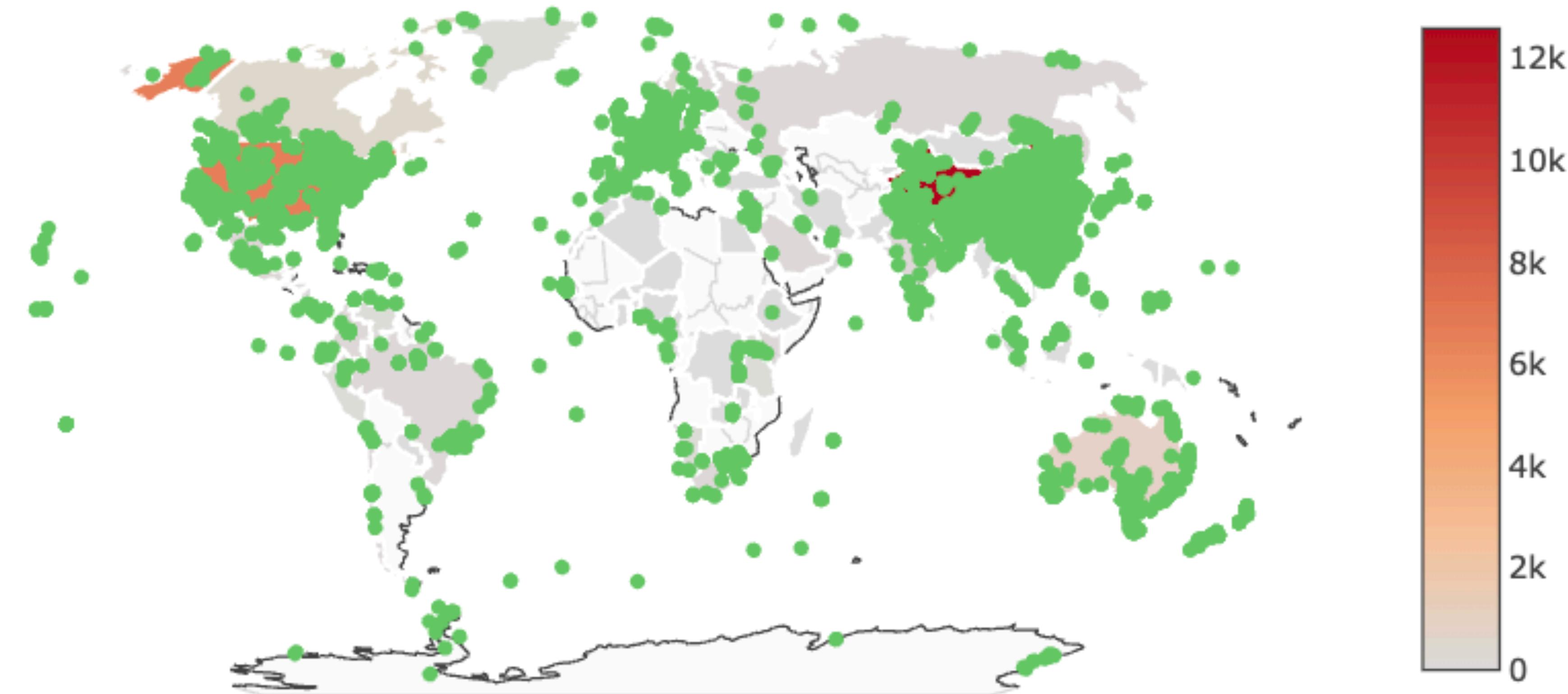


**UCDAVIS**  
UNIVERSITY OF CALIFORNIA

**JGI**  
JOINT GENOME INSTITUTE

**USDA**

# 30,369 Metagenomes contain *Ralstonia* sequence



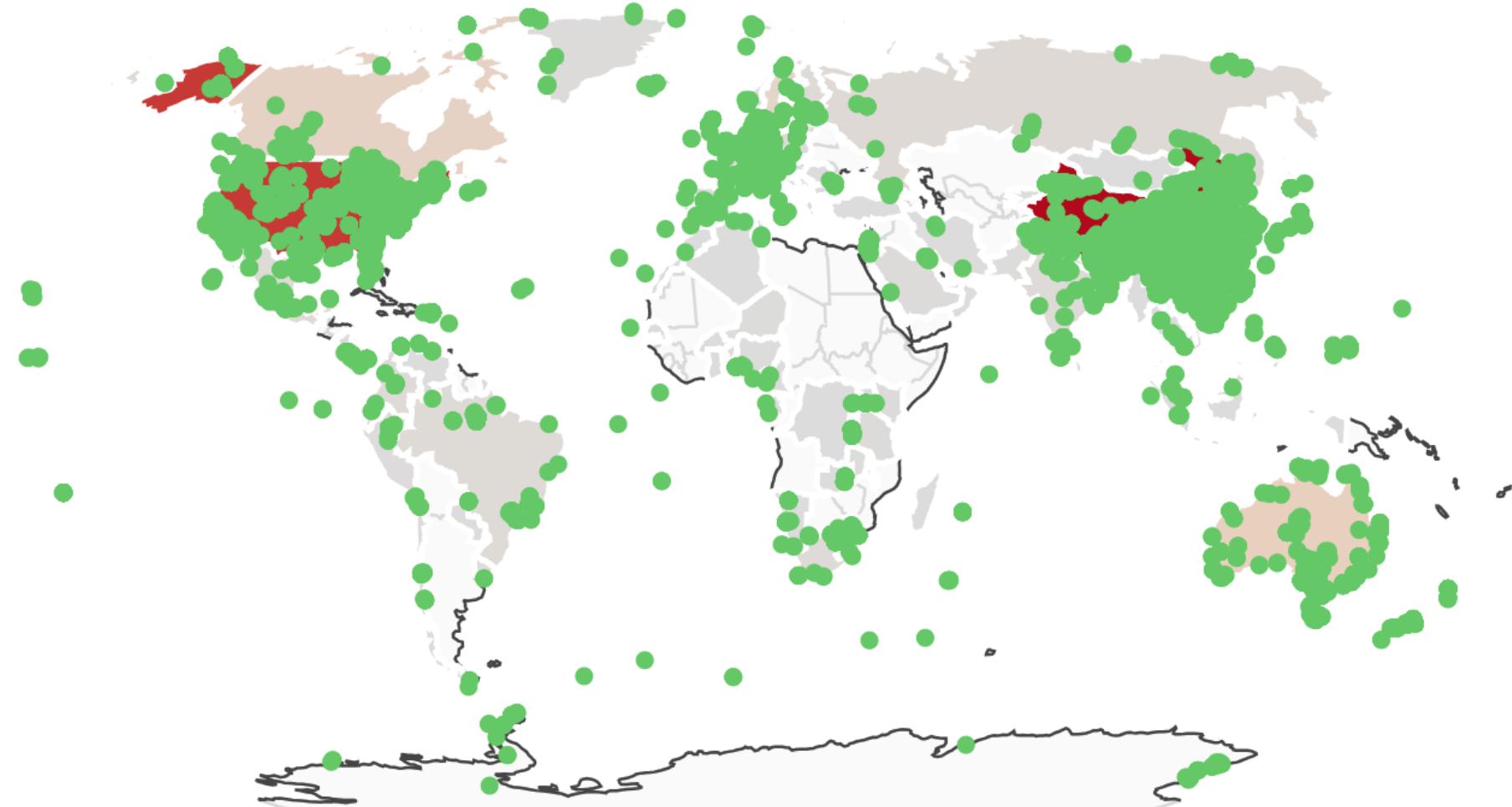
(1.16m searched)

<https://branchwater.jgi.doe.gov>

currently: k=21, scaled=1000 searches

# What kinds of metagenomes was it found in?

Query genome:  
GCF\_002251655.1 – IIB-1 UW551



k=21, scaled=1000 → ~genus-level

SRA metagenomes (top 11 categories)

<b>soil metagenome</b>	8864
metagenome	2013
sediment metagenome	1138
wastewater metagenome	1074
rhizosphere metagenome	1058
freshwater metagenome	658
activated sludge metagenome	567
freshwater sediment metagenome	345
peat metagenome	335
plant metagenome	318
root metagenome	245

# Other real-time search using the sourmash library

genome similarity and classification

**ACCESS MICROBIOLOGY**  
an open research platform

Volume 4, Issue 5

Meeting Report | Open Access

**genomeRxiv: a microbial whole-genome database and diagnostic marker design resource for classification, identification, and data sharing** 

Leighton Pritchard<sup>1</sup>, Parul Sharma<sup>2</sup>, Reza Mazloom<sup>2</sup>, Tessa Pierce<sup>3</sup>, Luiz Irber<sup>3</sup>, Bailey Harrington<sup>1</sup>, Lenwood Heath<sup>2</sup>, C Titus Brown<sup>3</sup> and Boris Vinatzer<sup>2</sup>

 View Affiliations

Published: 27 May 2022 | <https://doi.org/10.1099/acmi.ac2021.po0165>

metagenomic analysis

JOURNAL ARTICLE

**Mibianto: ultra-efficient online microbiome analysis through k-mer based metagenomics** 

Pascal Hirsch, Leidy-Alejandra G Molano, Annika Engel, Jens Zentgraf, Sven Rahmann, Matthias Hannig, Rolf Müller, Fabian Kern, Andreas Keller , Georges P Schmartz

[Author Notes](#)

*Nucleic Acids Research*, gkae364, <https://doi.org/10.1093/nar/gkae364>

Published: 08 May 2024 Article history ▾

Metagenome profiling against GTDB representatives

**greyhound gather**

Choose a FASTA/Q file to upload. File can be gzip-compressed.

Choose Files No file chosen

 Download

This is a demo for a system running **gather**, an algorithm for decomposing a query into reference datasets.

**greyhound** is an optimized approach for running **gather** based on an Inverted Index containing a mapping of hashes to datasets containing them. In this demo the datasets are Scaled MinHash sketches ( $k=21$ , scaled=1000) calculated from the [85,205 species clusters in the GTDB rs214 release](#).

Branchwater also in use in MGNify, RKI MetagenomeWatch

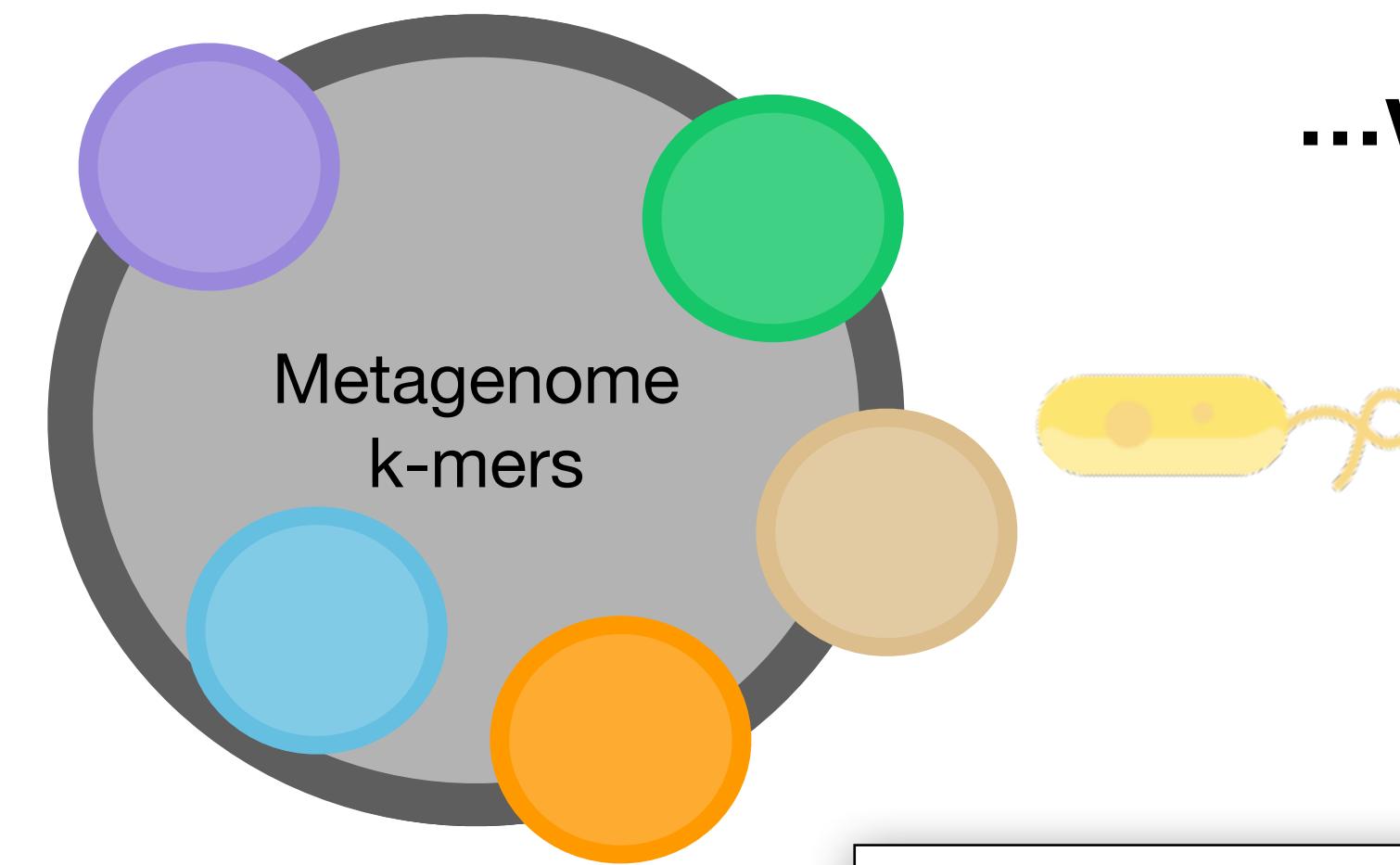
Open source software provides myriad opportunities for collaboration and extension

### **3. Comprehensive Metagenomic Breakdown**

# What *genomes* are in my sample?



Sketch metagenome  
into k-mer collection



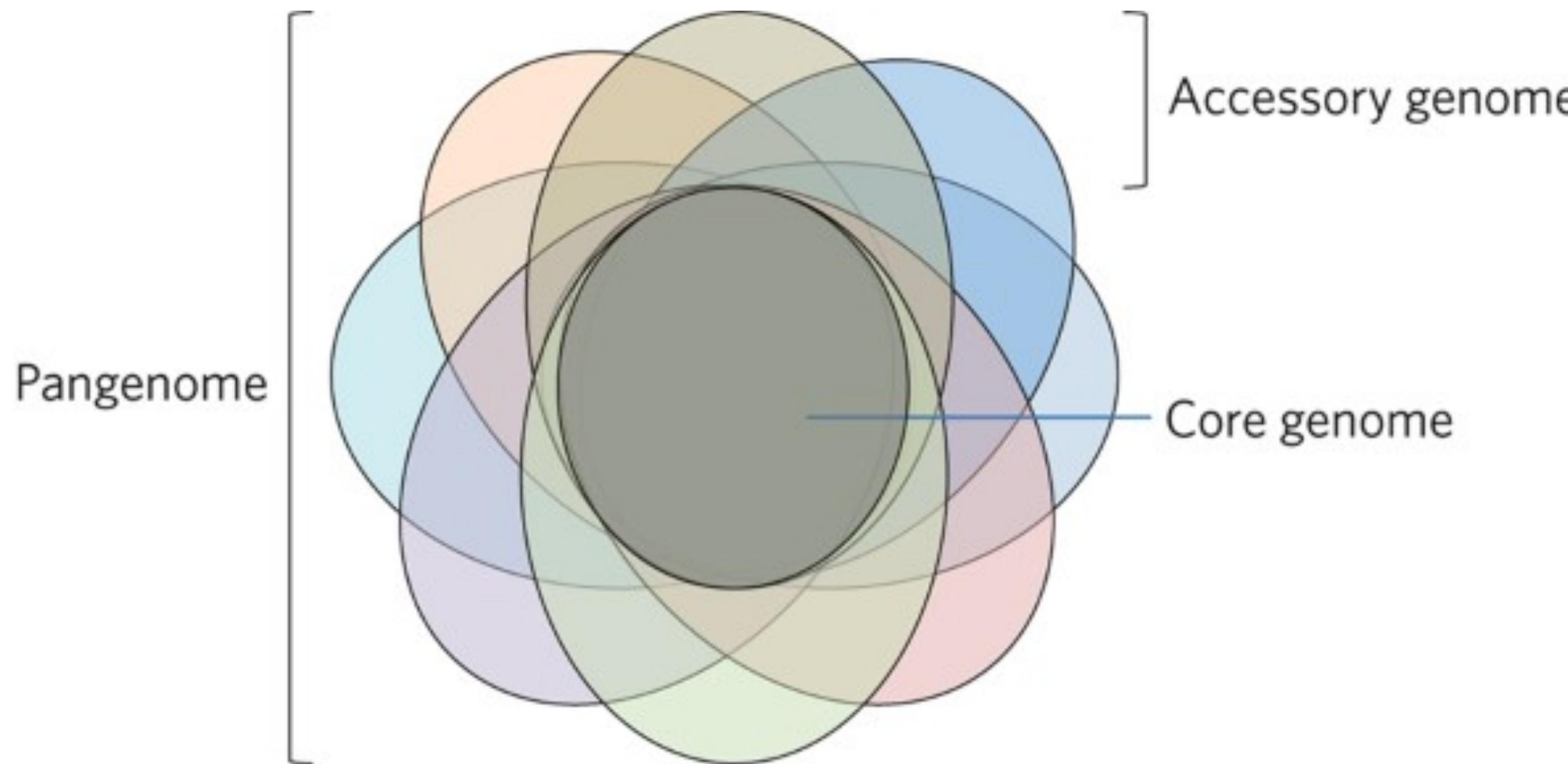
**Screen large databases  
for genome matches**

Community Composition



What challenges does this present?

# Large databases with significant shared content



- Databases are growing: GenBank contains 2.64 million bacteria + archaea genomes
- Many strains of well-studied species, e.g. 588k *Salmonella enterica*

**How do you assign k-mer content in this case?**

# Strain-resolved metagenomics is challenging

- Metagenomes often contain *mixtures* of strains and/or contain strains that are not found in the reference database
- Strains always have significant overlap (~99% ANI)
- Different strains of the same species may have very different function — e.g. harmless or pathogenic *E. coli*

# One approach:

What genomes are in my metagenome?



What is the shortest list of genomes containing  
all *known* content in my metagenome?

# ....now this is a known CS problem - “min set cover”

What is the shortest list of genomes containing all **known** content in my metagenome?

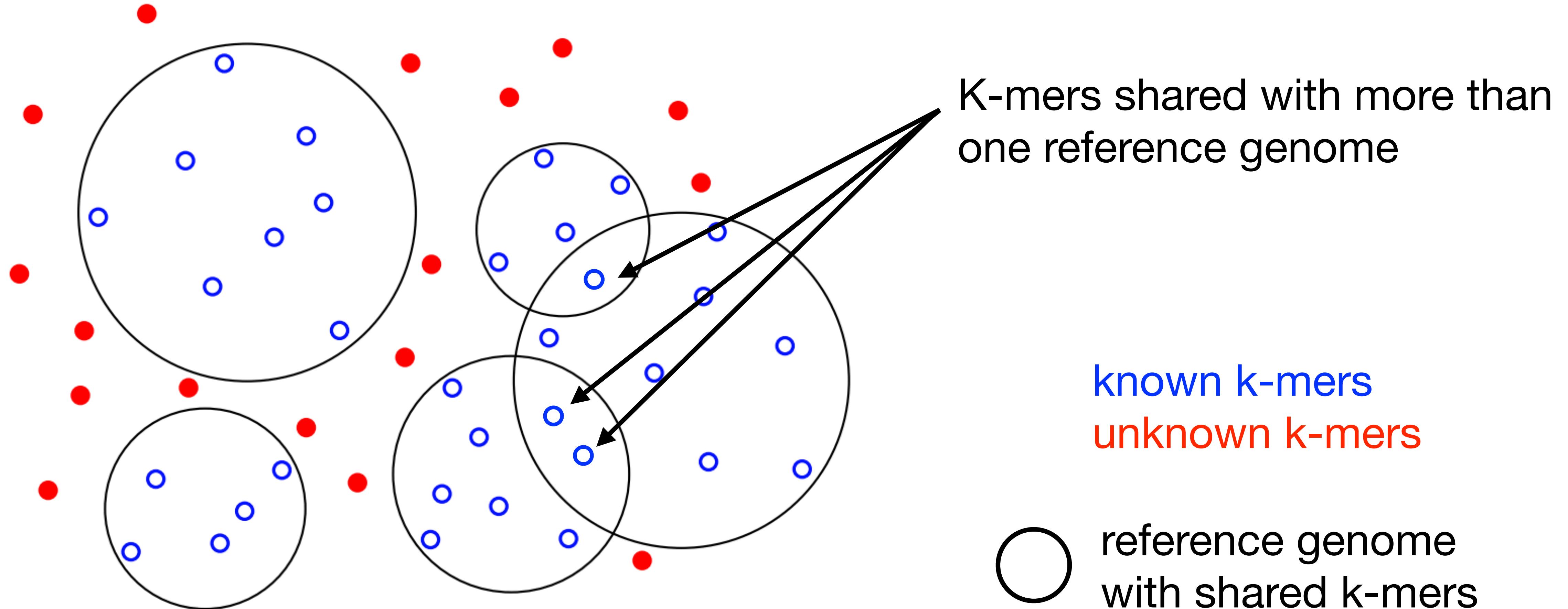
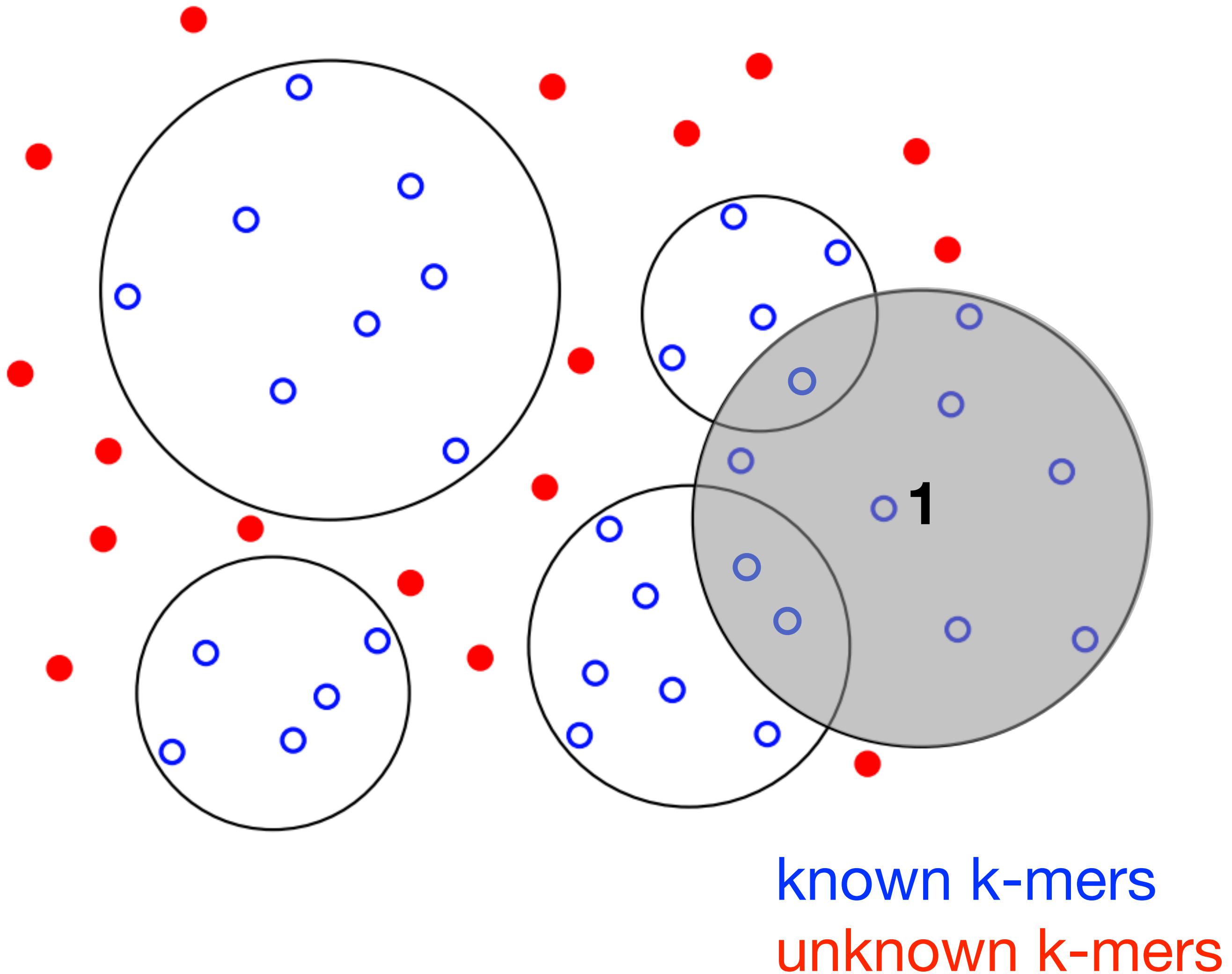


Figure: Bereg et al., 2012  
doi: 10.1016/j.comgeo.2012.01.014

# ....now this is a known CS problem - “min set cover”

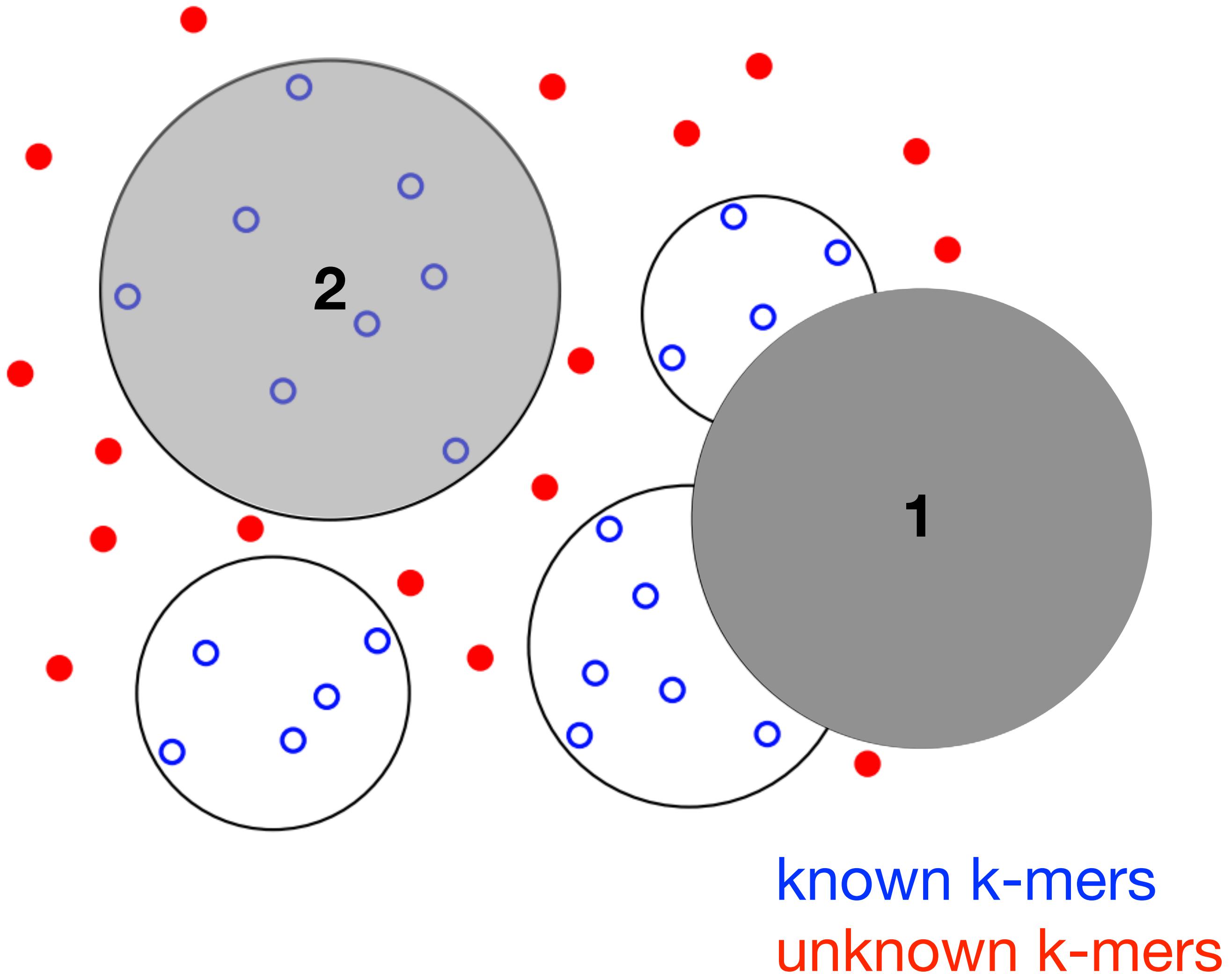


What is the shortest list of genomes containing all **known** metagenome content?

Greedy approach:

- Find circle that contains the most points; assign all included points to that circle.
- Repeat

# ....now this is a known CS problem - “min set cover”

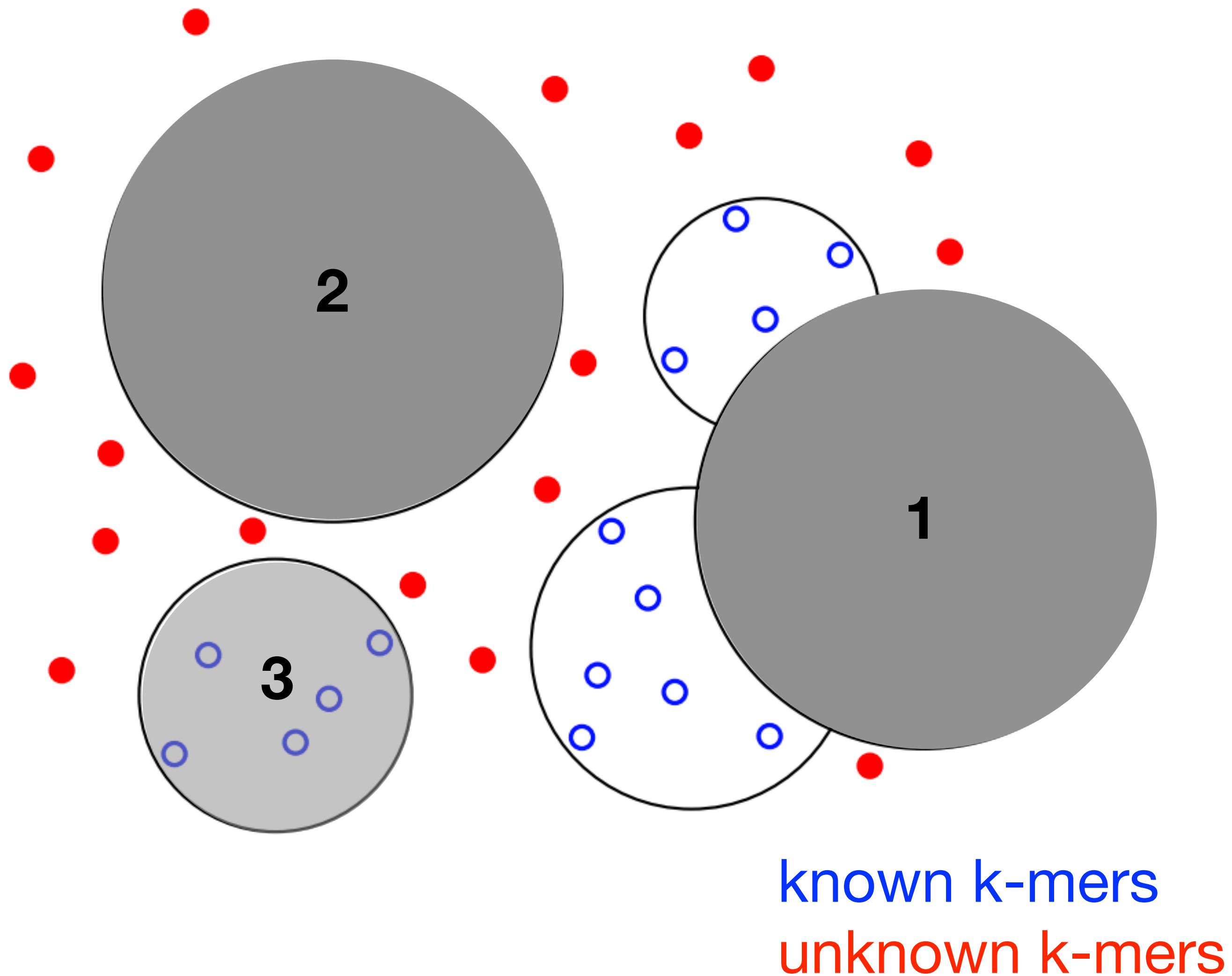


What is the shortest list of genomes containing all **known** metagenome content?

Greedy approach:

- Find circle that contains the most points; assign all included points to that circle.
- Repeat

# ....now this is a known CS problem - “min set cover”

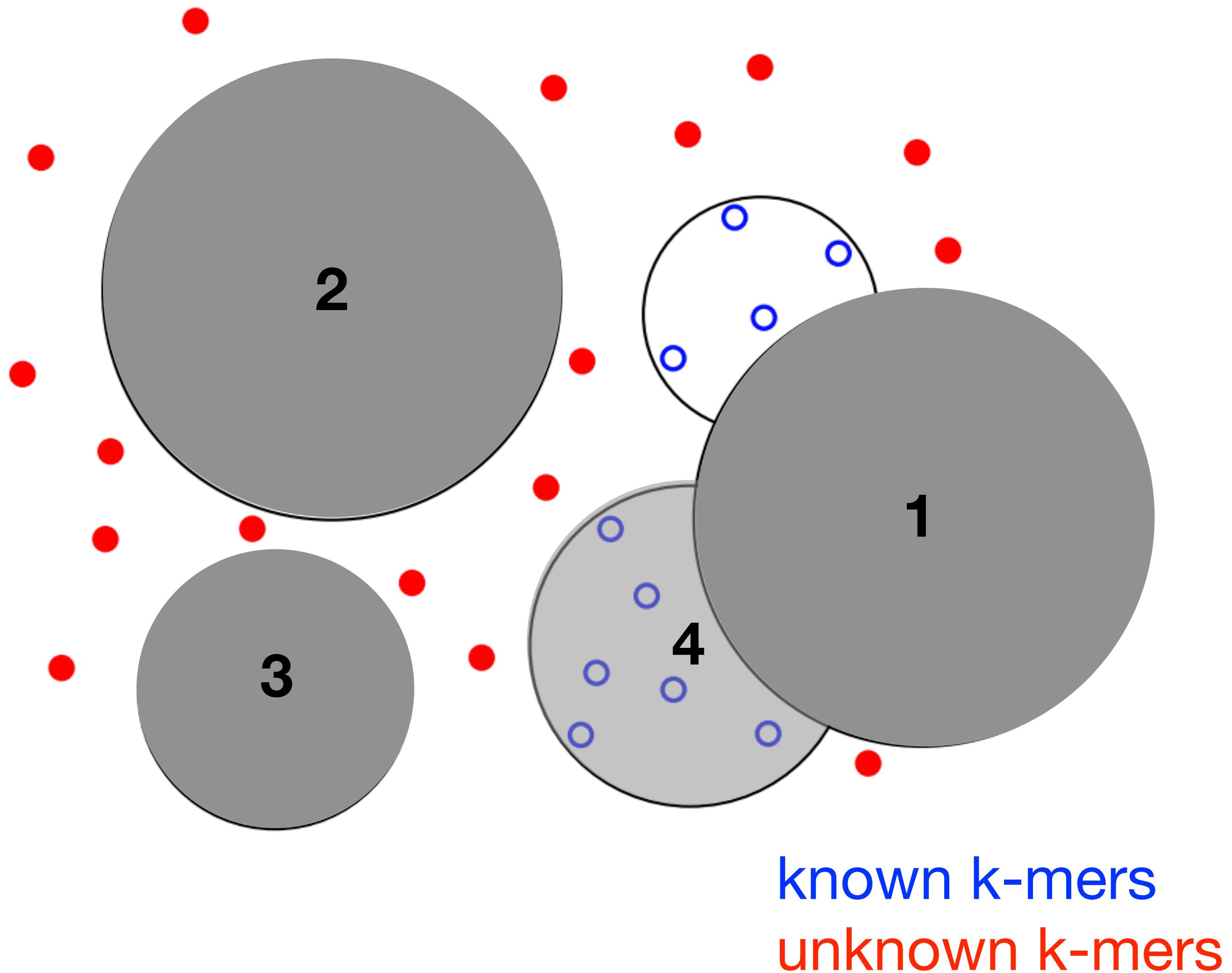


What is the shortest list of genomes containing all **known** metagenome content?

Greedy approach:

- Find circle that contains the most points; assign all included points to that circle.
- Repeat

# ....now this is a known CS problem - “min set cover”

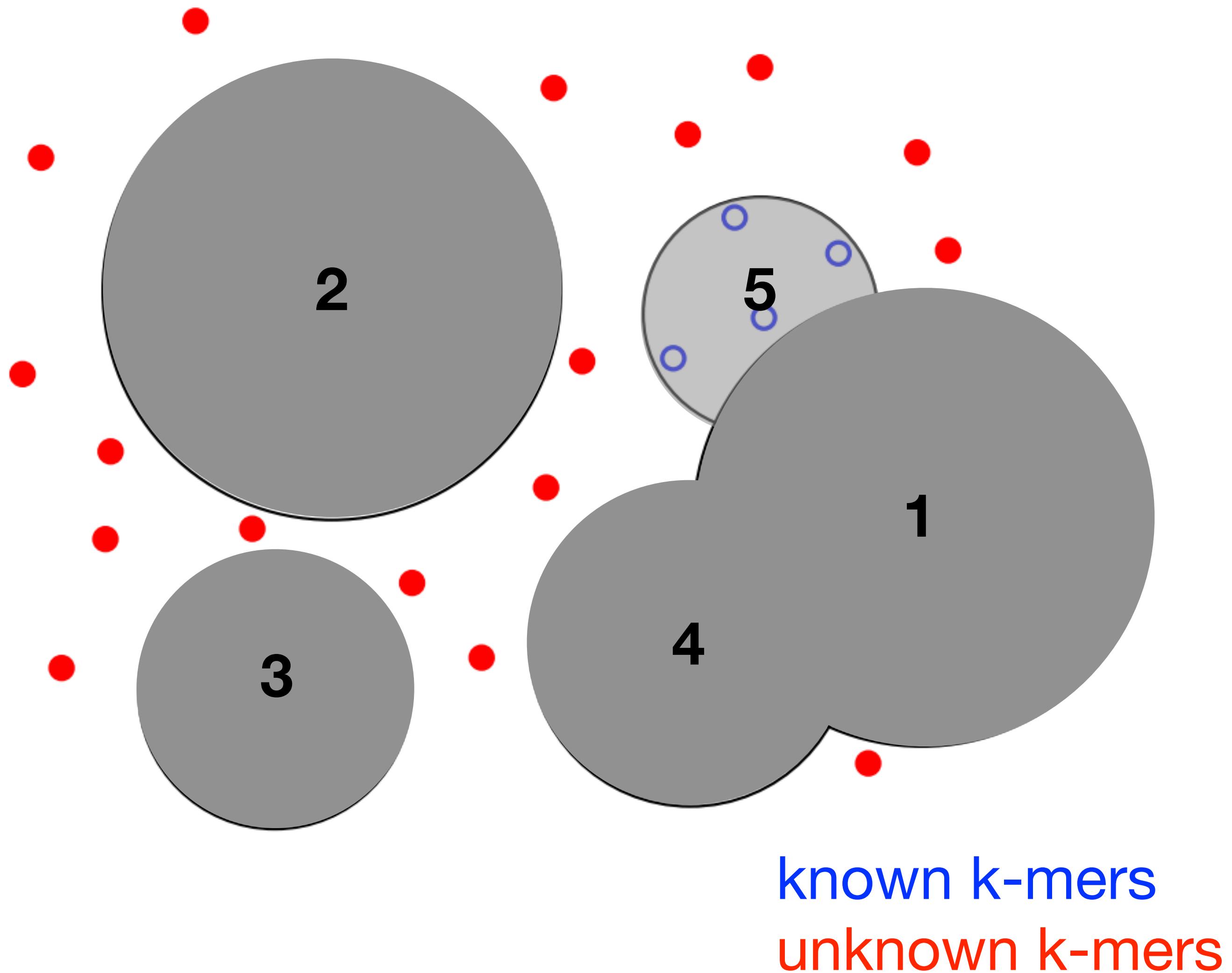


What is the shortest list of genomes containing all **known** metagenome content?

Greedy approach:

- Find circle that contains the most points; assign all included points to that circle.
- Repeat

# ....now this is a known CS problem - “min set cover”

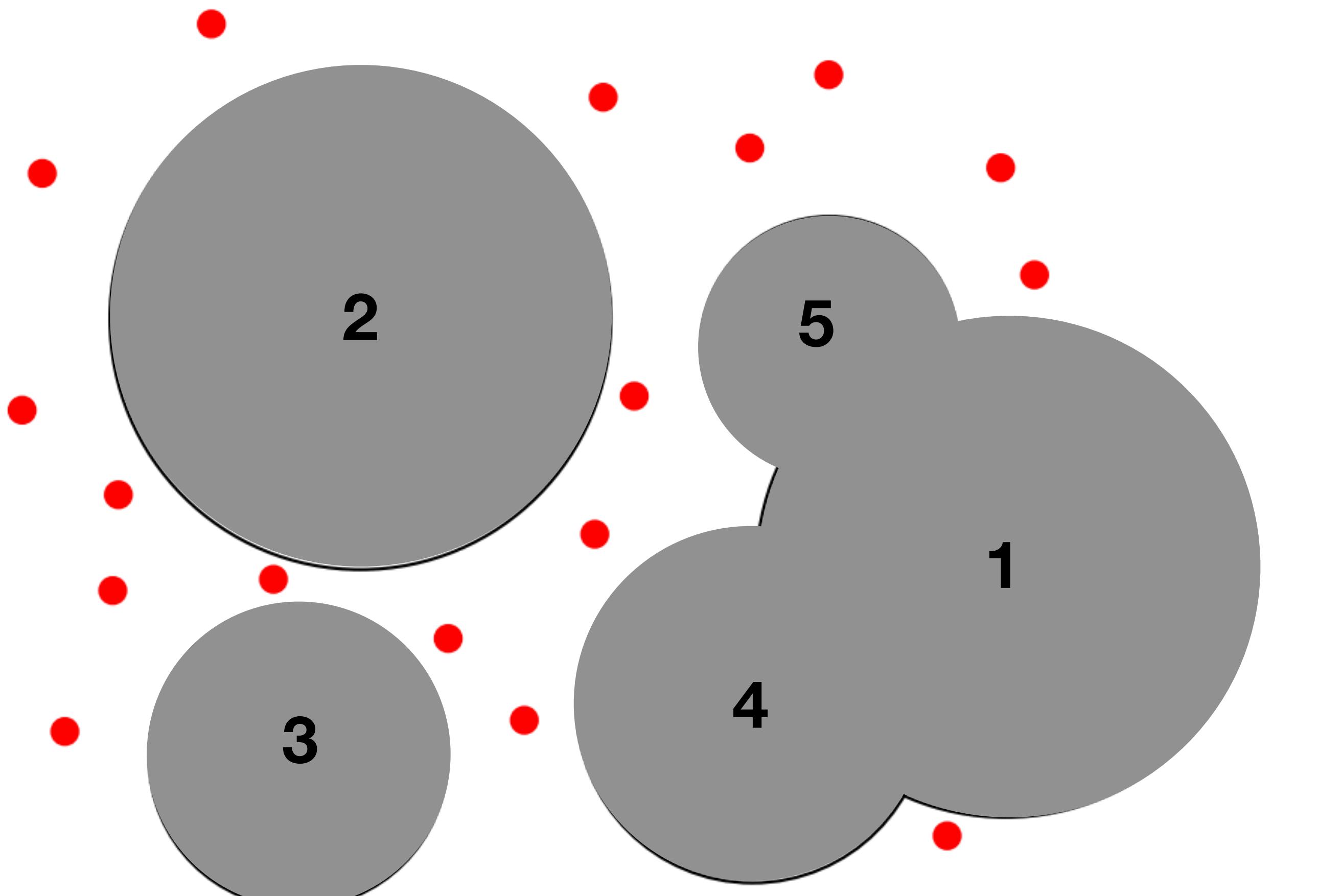


What is the shortest list of genomes containing all **known** metagenome content?

Greedy approach:

- Find circle that contains the most points; assign all included points to that circle.
- Repeat

# ....now this is a known CS problem - “min set cover”



What is the shortest list of genomes containing all **known** metagenome content?

known k-mers  
unknown k-mers

# We can think about this with sequence content:

Metagenome



## Strains in the Reference Database

---

Genome A

shared content

Accessory 3

Acc. 4

Genome B

shared content

Accessory 2

Genome C

shared content

Accessory 1

Accessory 3



Genome D

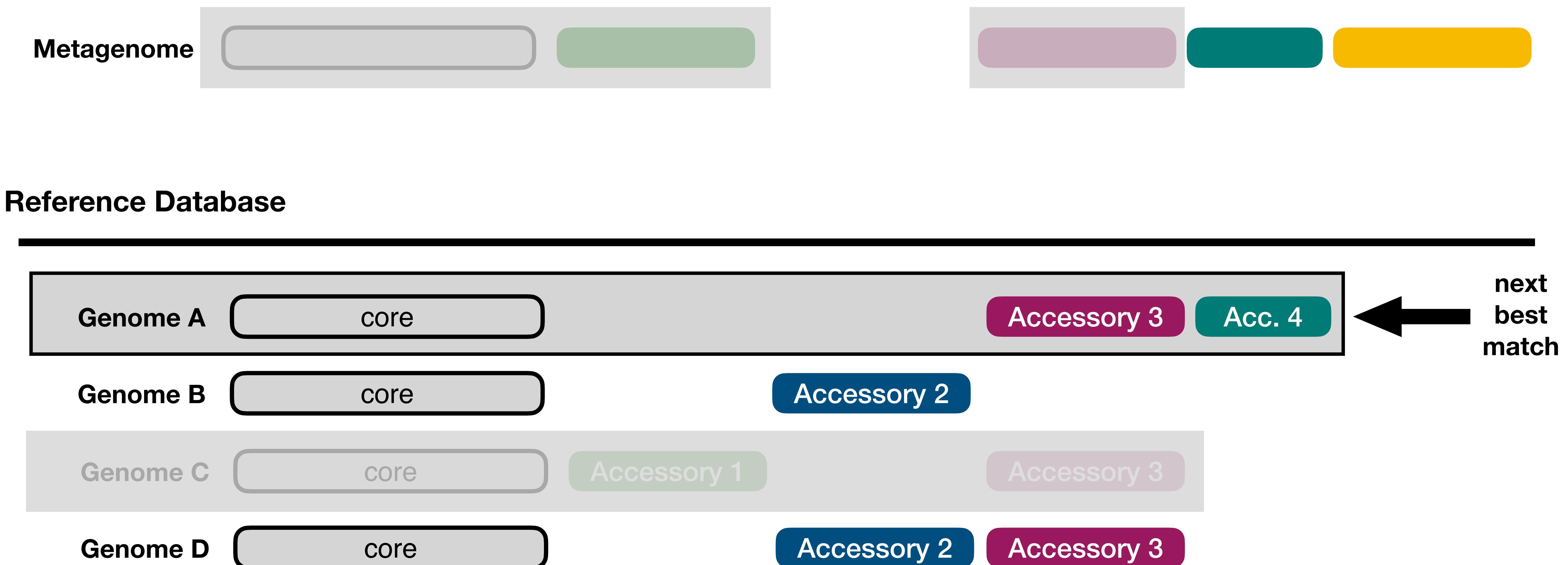
shared content

Accessory 2

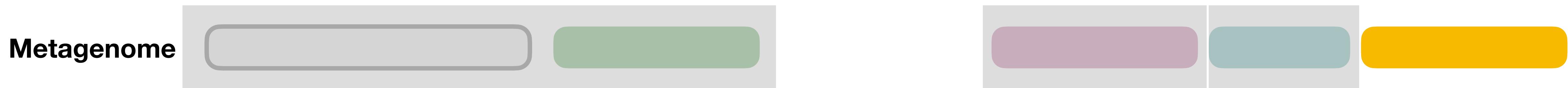
Accessory 3

~ core genome

# What genomes are present in the metagenome?

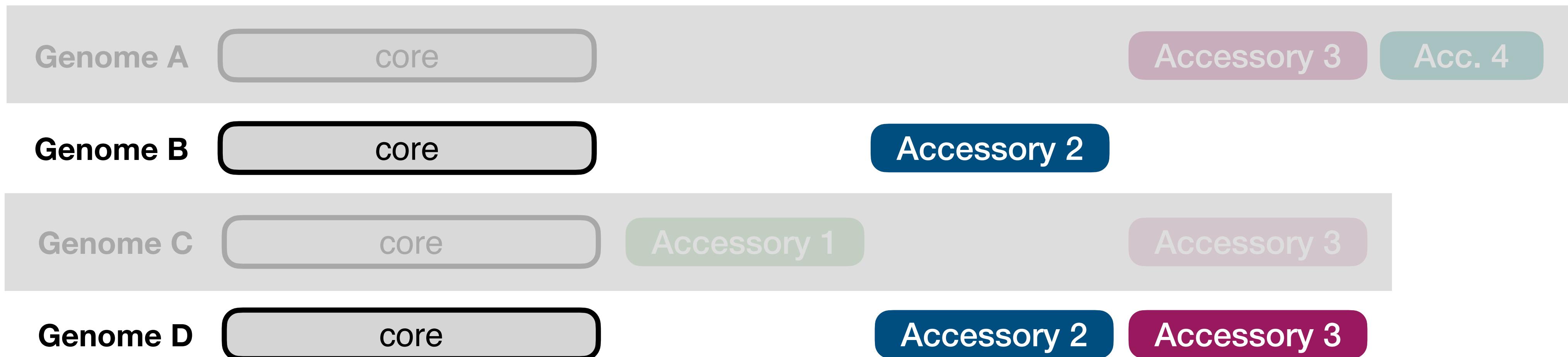


# What genomes are present in the metagenome?

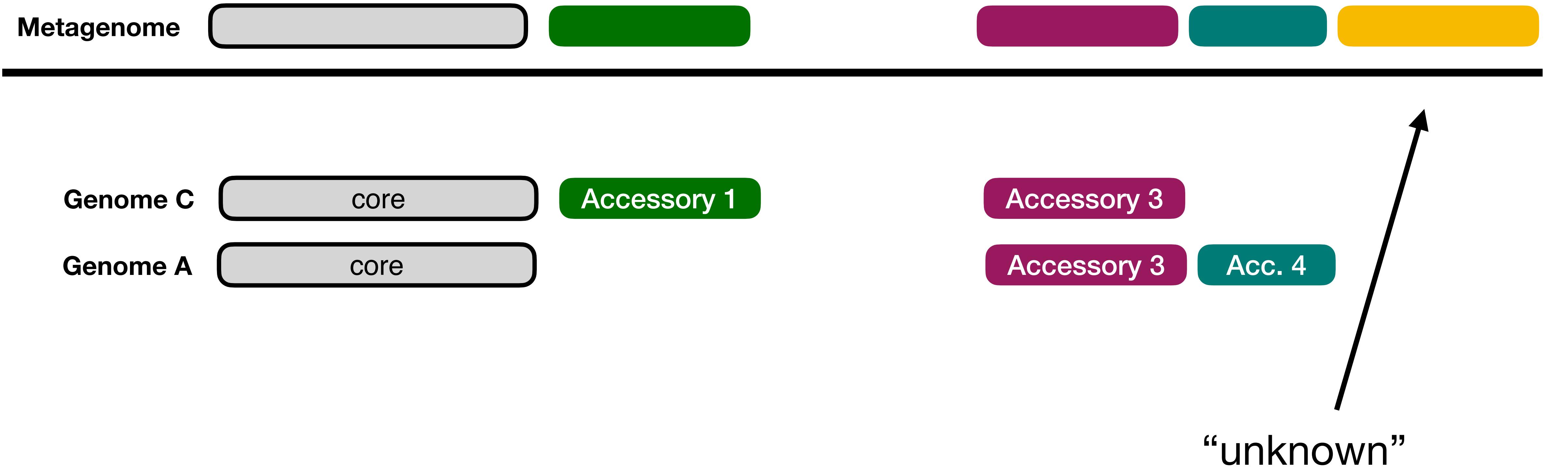


## Reference Database

---

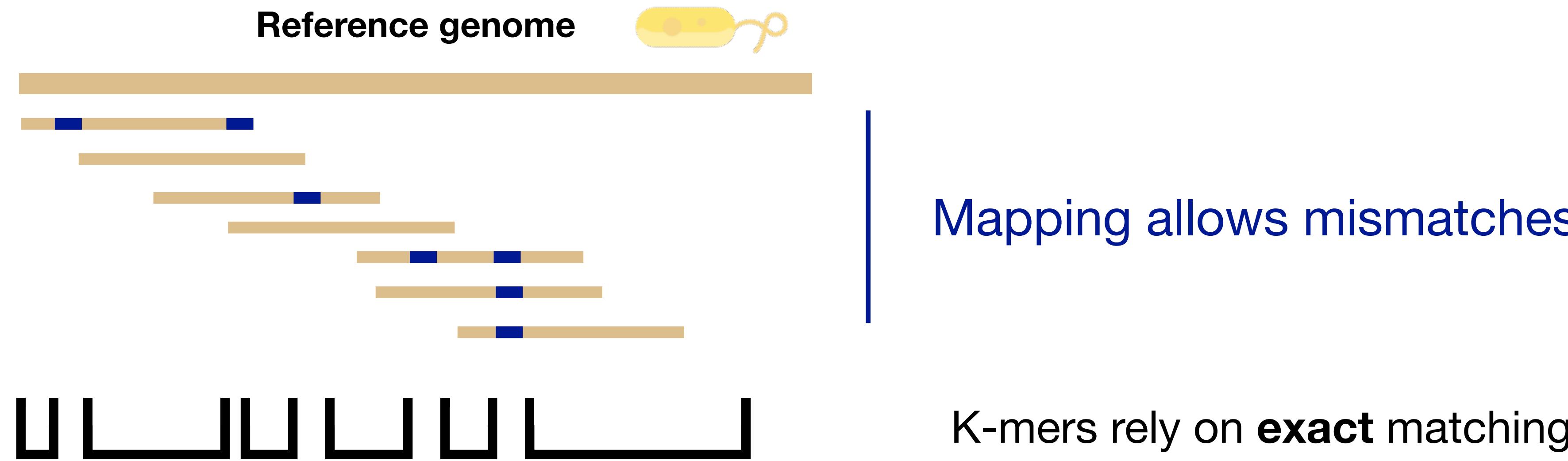


# minimum set of genomes that contain all known query sequence



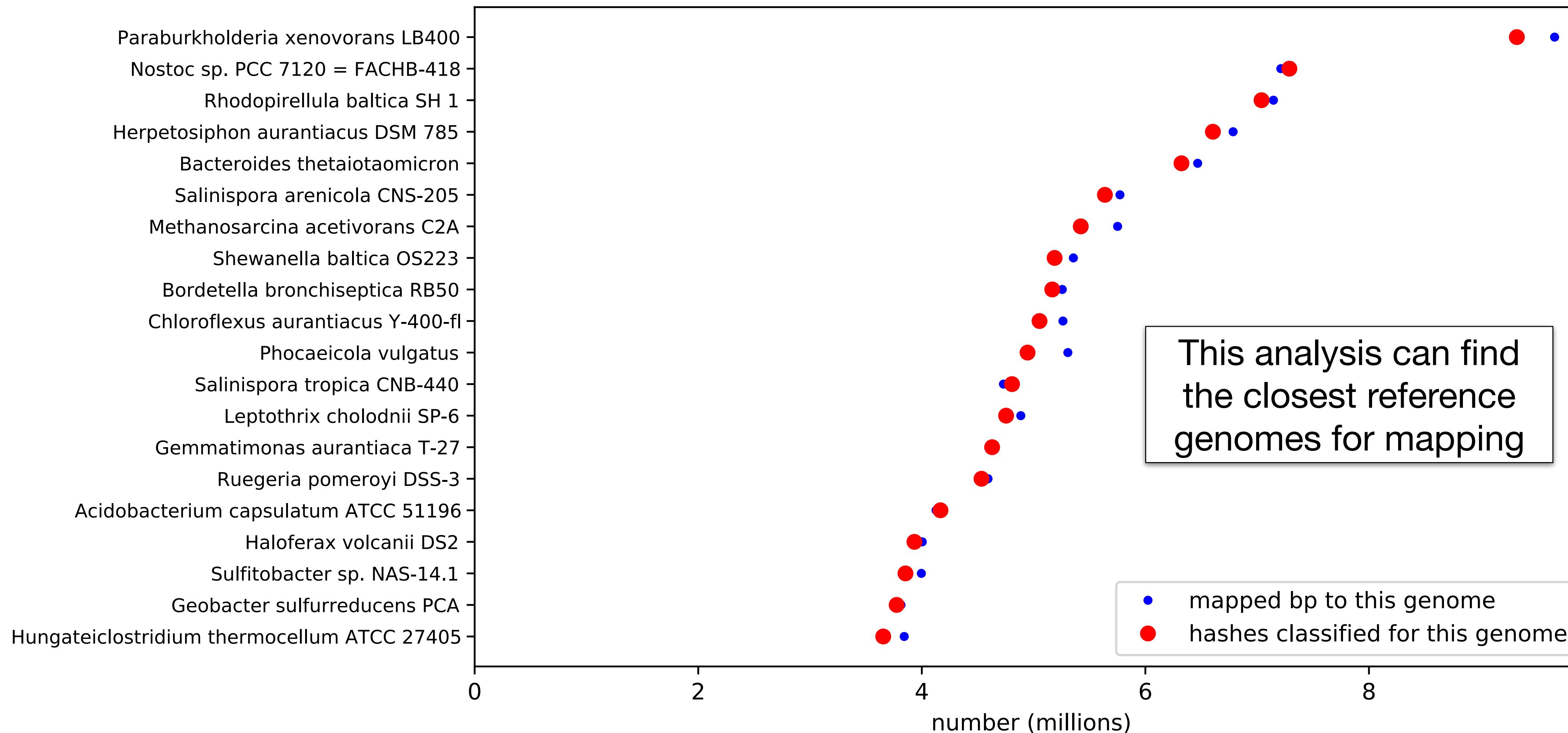
(did not match any genomes  
in reference database)

# How does this approach compare with mapping?

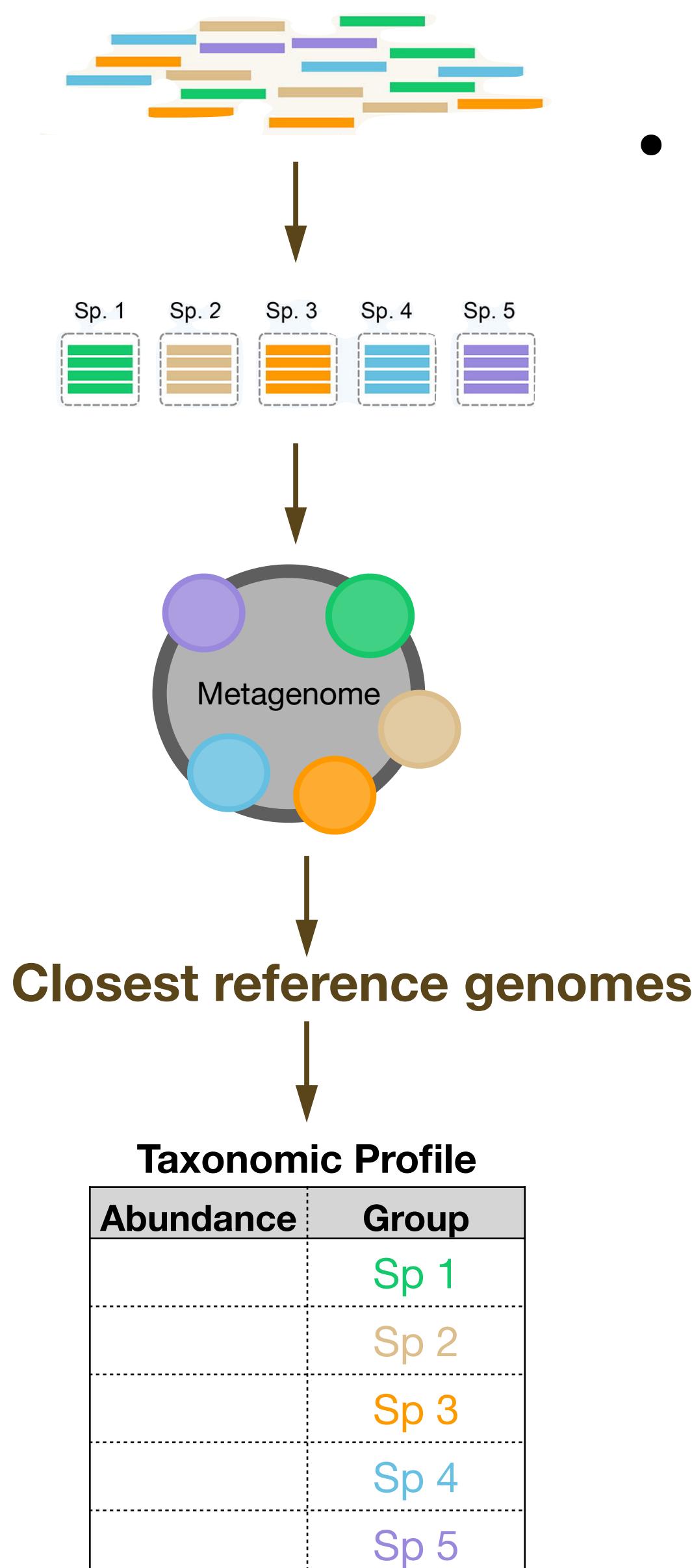


# Mapping metagenome reads to genomes closely matches gather k-mer estimates

SRR606249: k-mers vs mapped bp



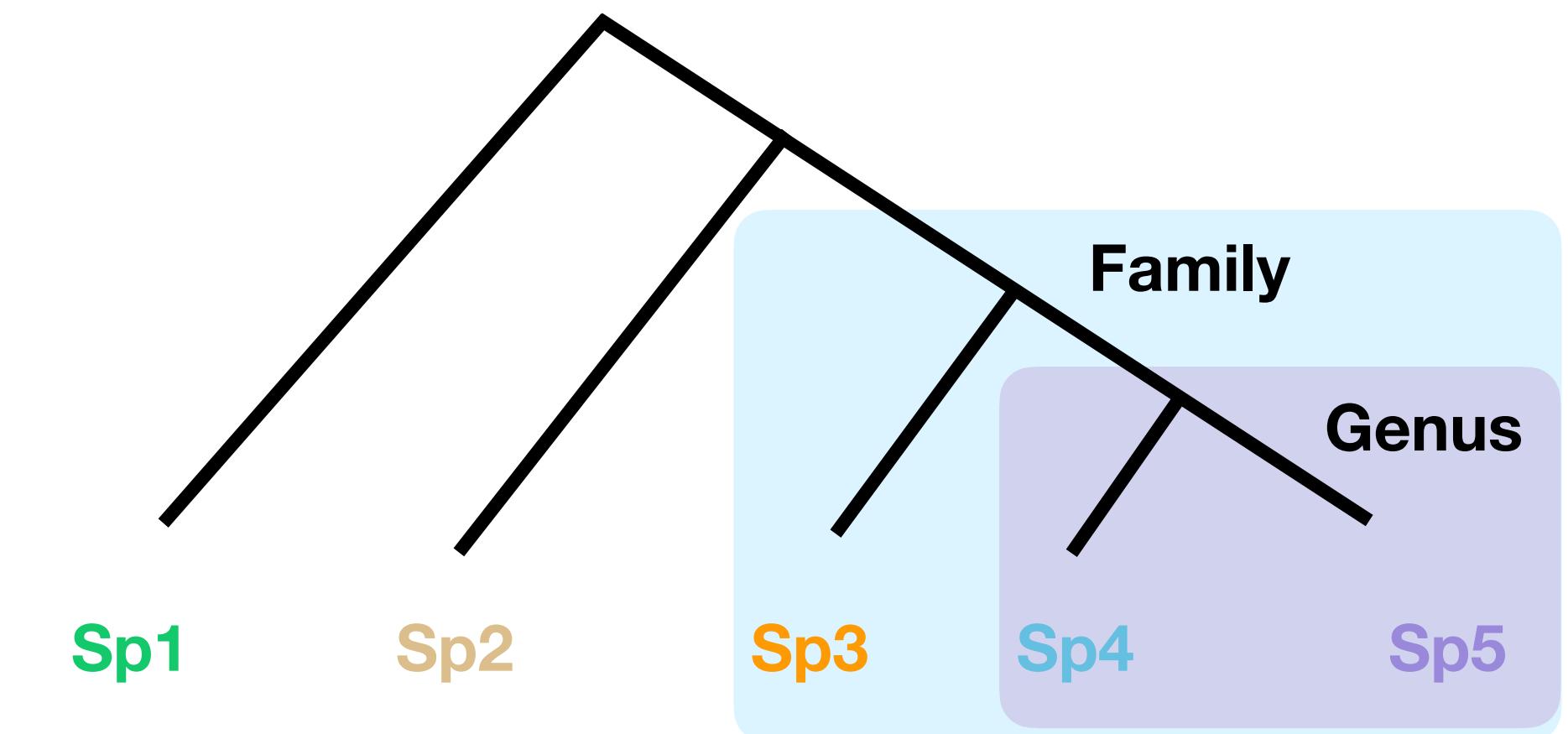
# Metagenome breakdown to taxonomic profiling



- **sourmash taxonomy:**

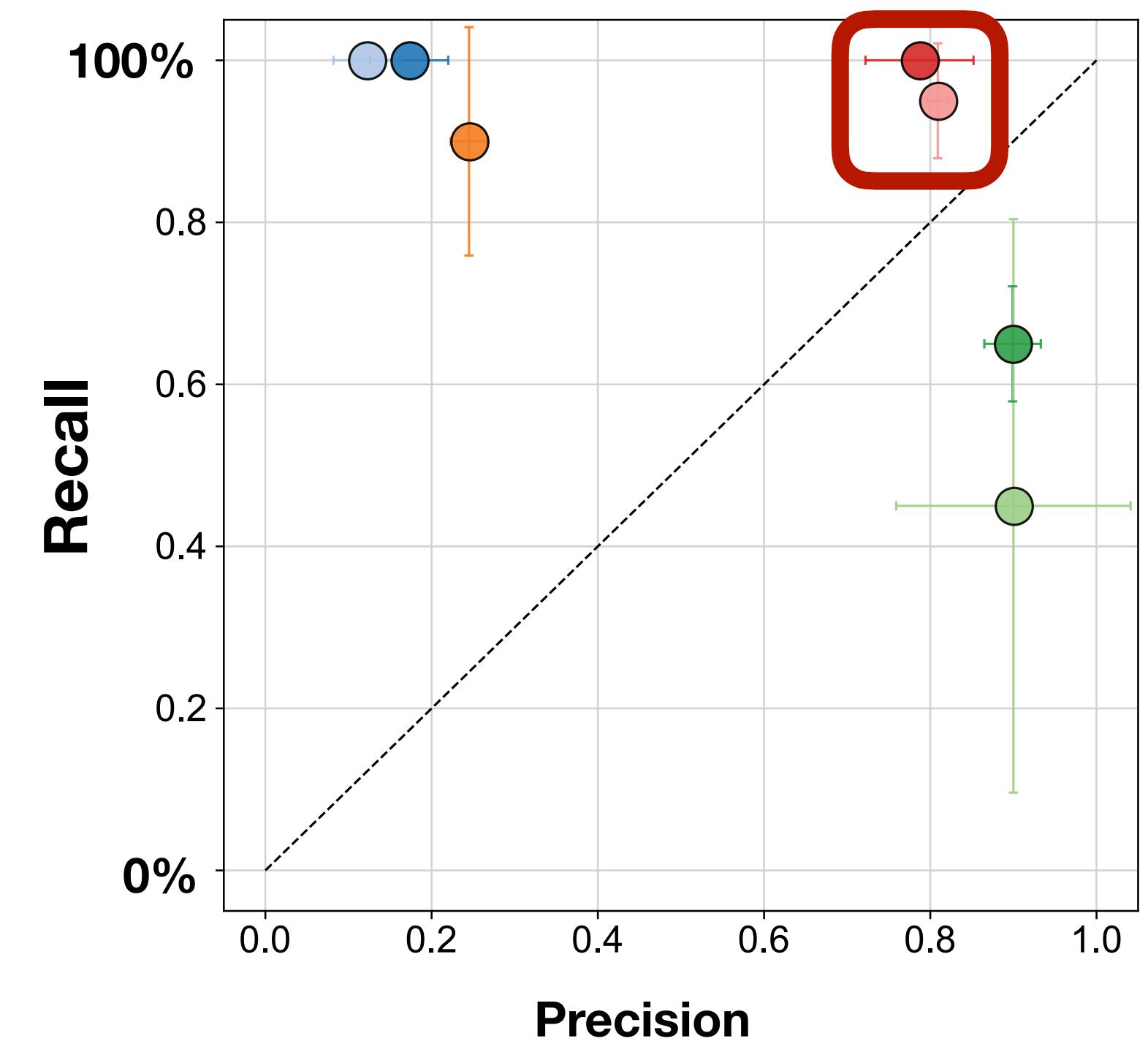
- use **gather's** non-overlapping genome matches to add taxonomic information
- Aggregate with LCA if needed

## Lowest Common Ancestor (LCA) Approach

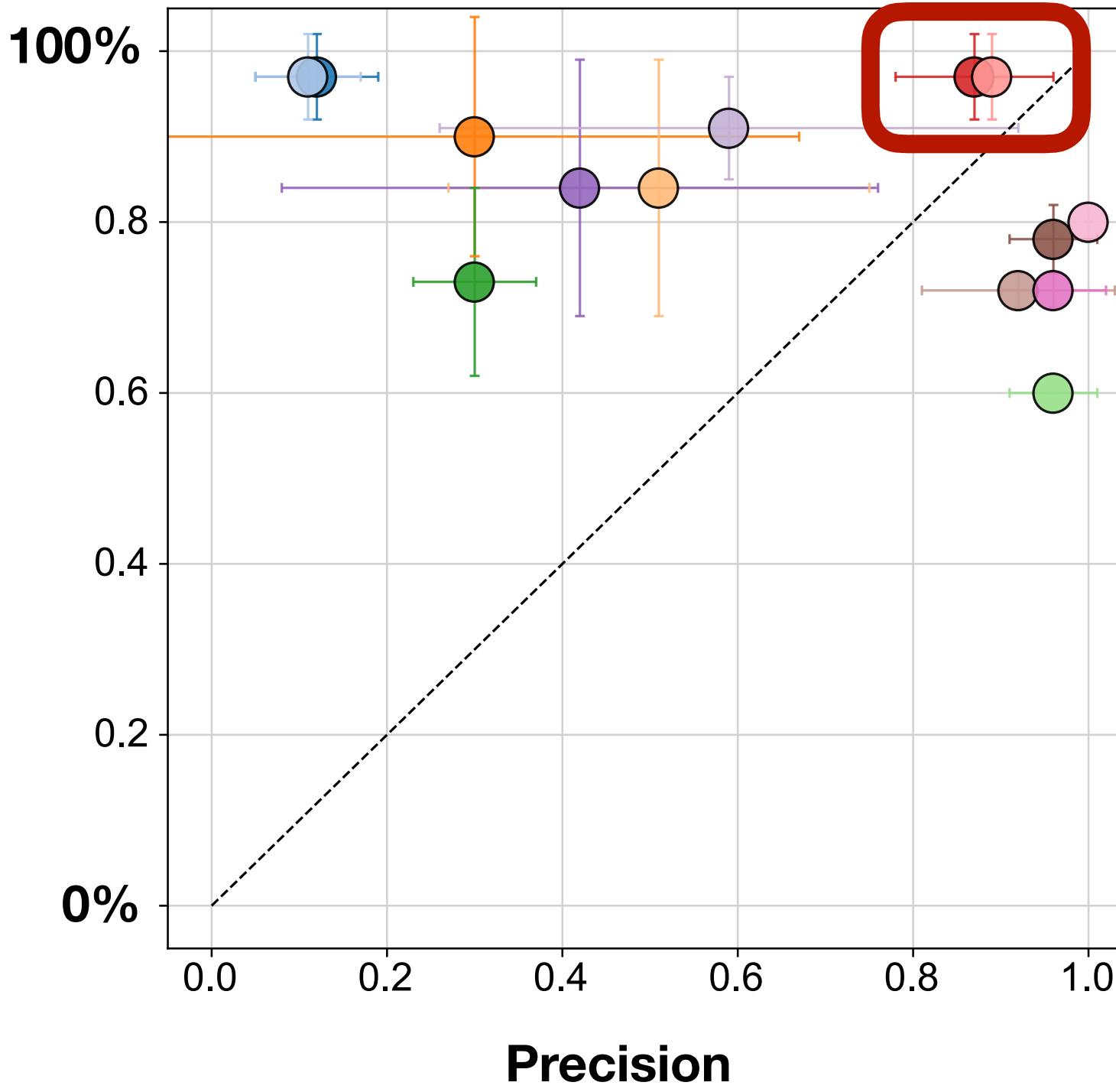


# This approach performs very well for species-level taxonomic profiling

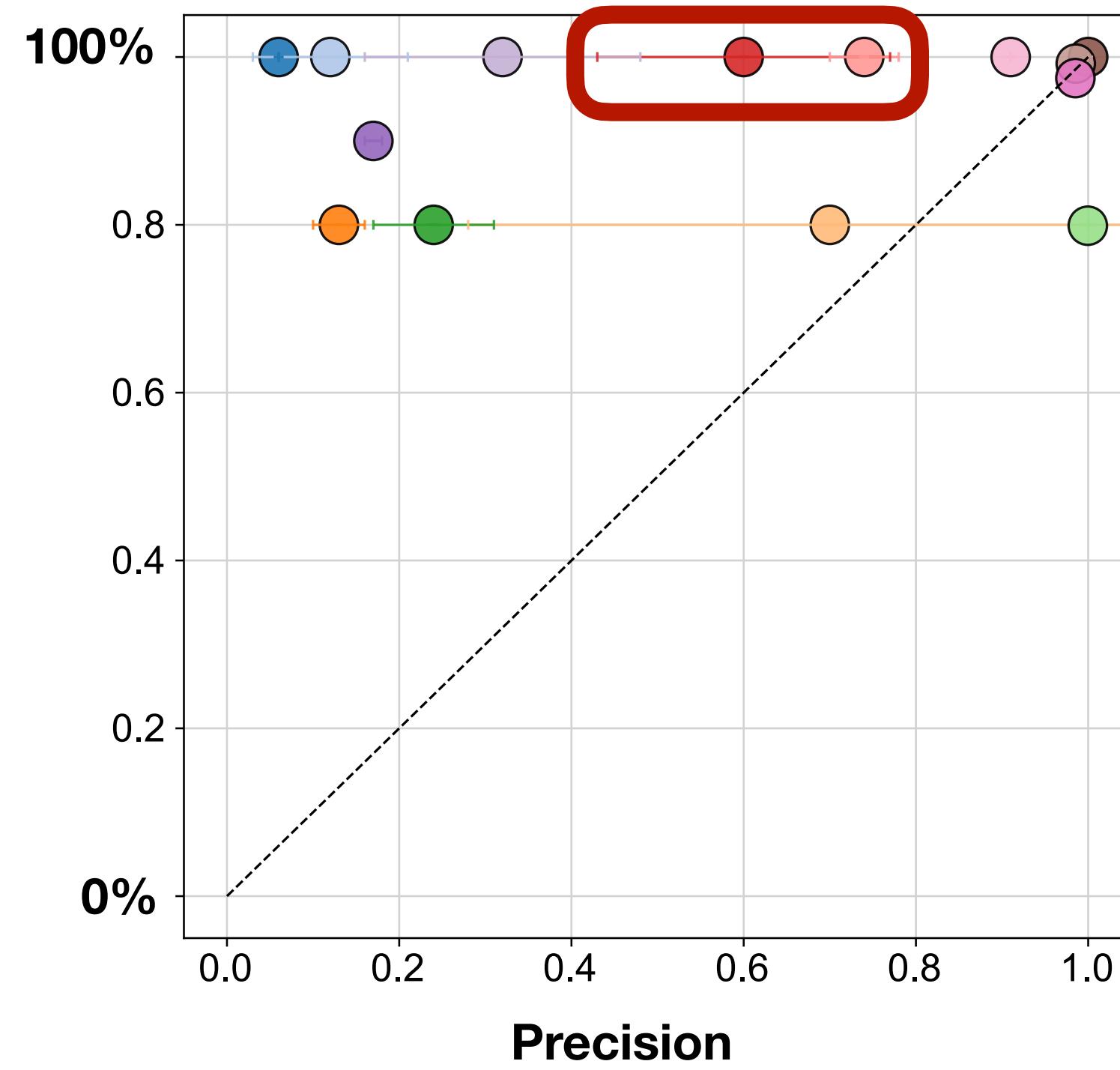
Illumina Short Reads



PacBio HiFi



Oxford Nanopore (R10.3, Q20)



● Kraken2  
● Bracken  
● Centrifuge-h22  
● Centrifuge-h500

● Metaphlan3  
● mOTUs  
● Sourmash-k31  
● Sourmash-k51

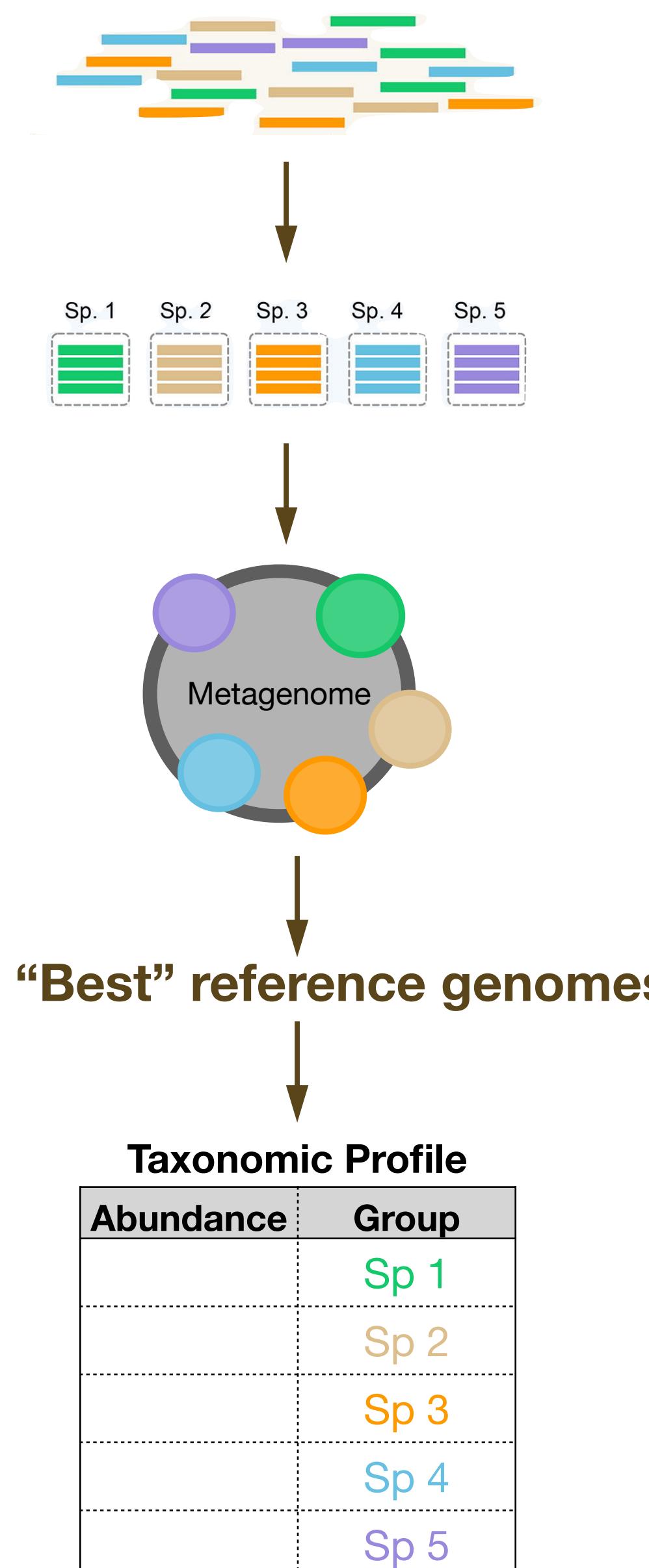
● Metamaps  
● MMseqs2  
● MEGAN-LR-Prot  
● MEGAN-LR-Nuc-HiFi

(mock community empirical datasets)

Portik, Brown, and Pierce-Ward (2022)

[10.1186/s12859-022-05103-0](https://doi.org/10.1186/s12859-022-05103-0)

# Sourmash Genome-Resolved Taxonomic profiling

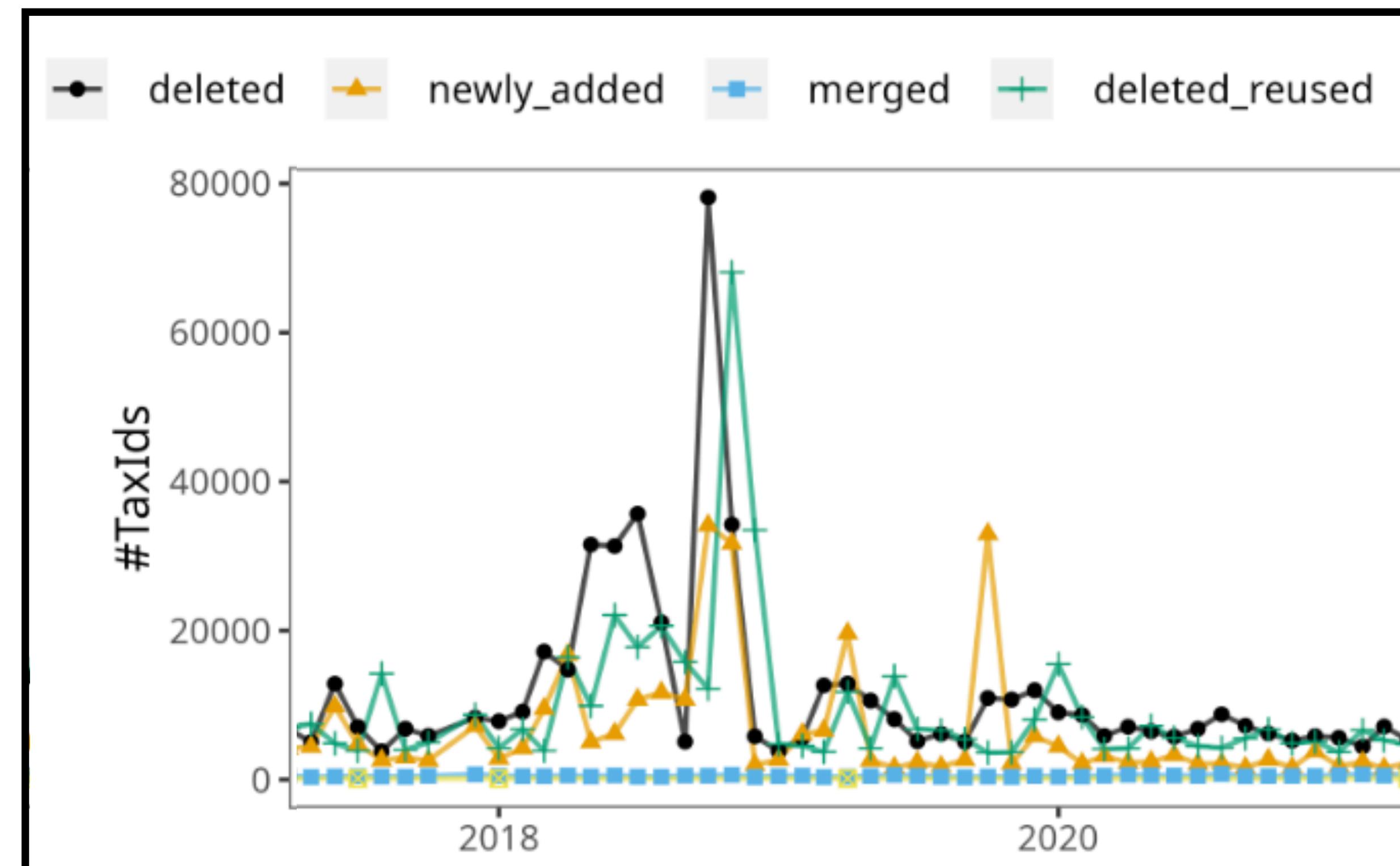


- Rather than considering each read independently, sourmash can use information from the full file
  - Larger combinations of k-mers can more robustly distinguish between closely-related genomes
    - particularly helpful for short reads
    - Enables faster search of large databases (e.g. GenBank, AllTheBacteria)

# Genome-anchored taxonomic integration has benefits

Reference genome sequences are ~stable; taxonomic groupings are not

- Sequence-associated stable identifiers (e.g. GCA\_000961135.2) allow for **persistent identification** despite taxonomy and metadata updates over time
- By anchoring metagenome profiling to reference genomes, **we can use stable identifiers to integrate different/multiple taxonomies: NCBI, GTDB, ICTV, LINs**



<https://github.com/shenwei356/taxid-changelog>

Hands-on part 3:  
metagenome breakdown  
& taxonomic profiling with sourmash

# Other sourmash applications

- Many other use cases!
  - Transcriptome, metatranscriptome, metagenome comparisons
  - Protein (and reduced Amino Acid alphabet) k-mer comparisons
    - proteome AAI, family-level taxonomic profiling, functional comparisons
  - Search public and private sequence collections, find transposable elements, find and remove contamination, compare and validate binning, pangenome k-mer comparisons, tetramer nucleotide clustering, classify strain variants, k-mer co-occurrence networks, input for ML approaches ...

# When and why to use sketching?

Datasets are not getting any smaller – sketching is an excellent way to narrow down your search space and/or get initial comparisons, even if you follow up with detailed analyses

- **Data exploration & early stage analysis**
  - Rapidly identify relevant datasets, prioritize analysis, find contamination, compare and cluster samples
- **Large-scale search, comparison, and characterization**
  - Estimate similarity at massive-scale with reasonable compute
- **Fast and (computationally) cheap streaming analysis, monitoring**
  - (e.g. across wastewater or SRA samples)

Advancements in sketching and storage techniques are ongoing!

# Questions?

Please contact me at [ntpierce@ucdavis.edu](mailto:ntpierce@ucdavis.edu)!  
(and/or talk to Cassie :)

