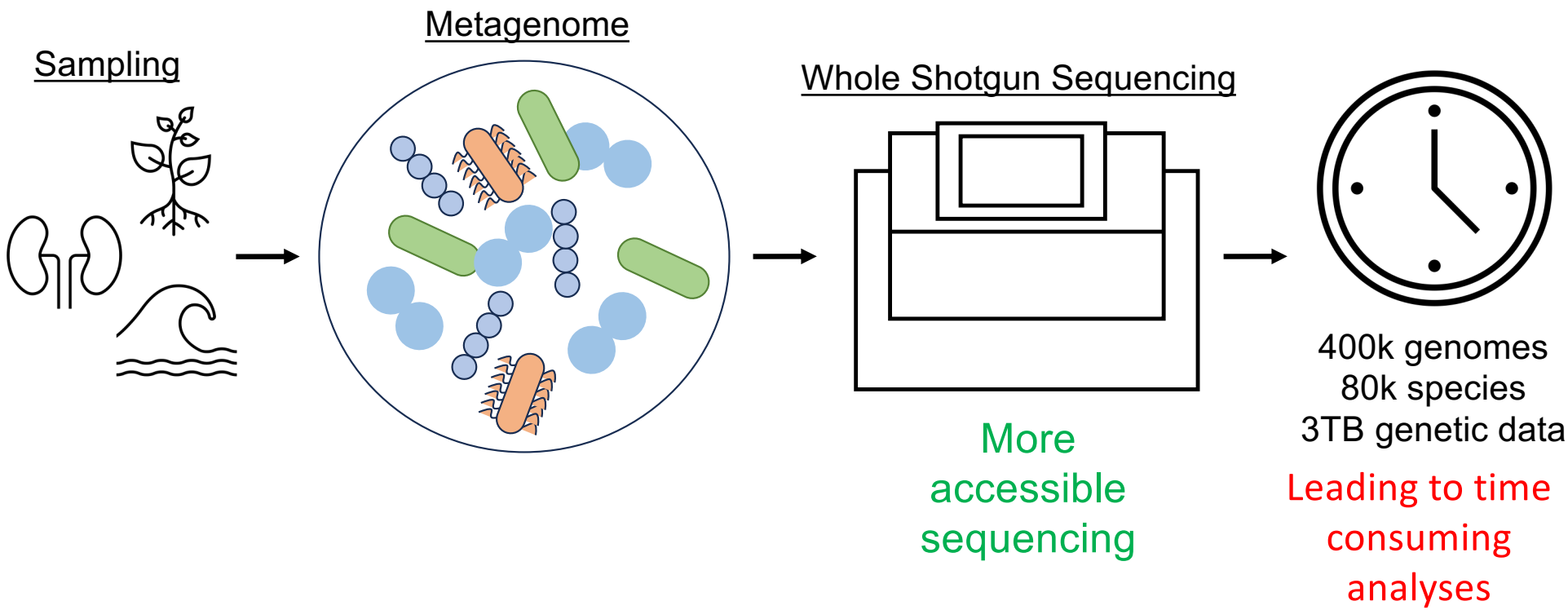
sourmash

a tutorial on the FracMinHash sketch approach for metagenomic data

Judith Rodriguez[1,2,3] and David Koslicki[1,2,3,4]

1. Bioinformatics and Genomics Program, Penn State
2. Life Sciences Huck Institute, Penn State
3. Department of Electrical Engineering and Computer Science, Penn State
4. Department of Biology, Penn State

Genomics and metagenomics research can be **computationally overwhelming**



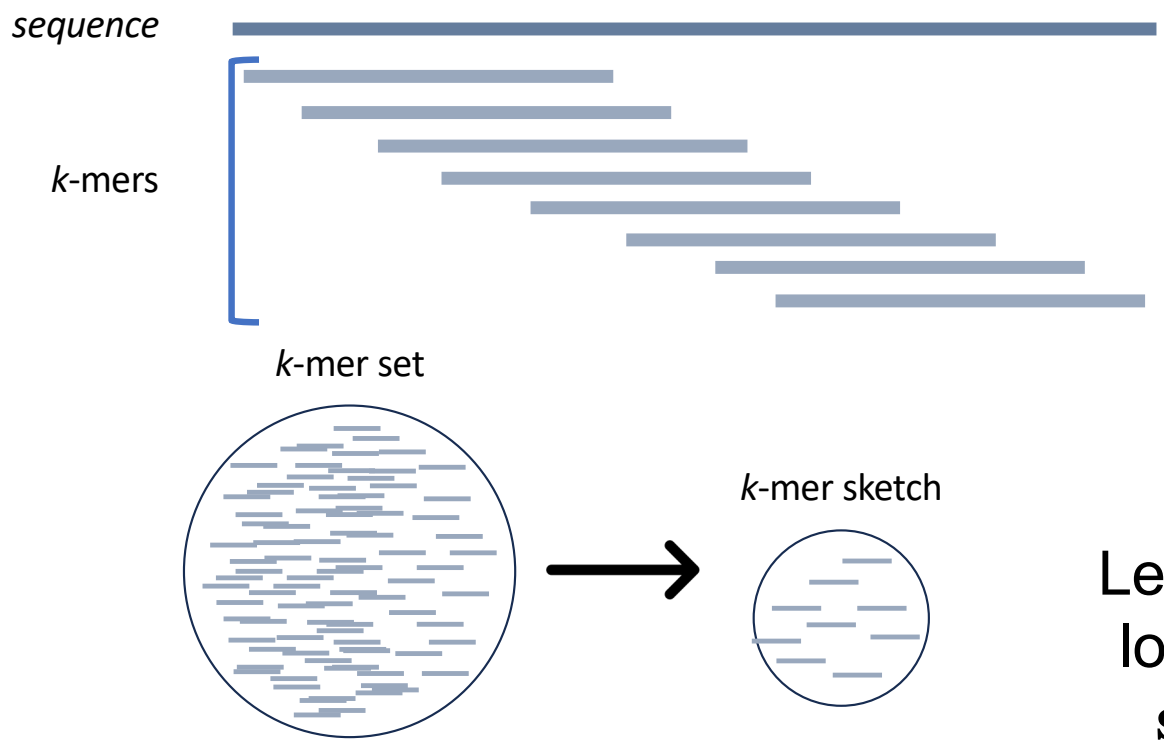
sketching approaches can help with big data

Do we recognize who this is?

→

Condense data to a simpler representation

sourmash facilitates sketching and analyses of large sequencing data



Let's take a closer look into utilizing **sourmash**…

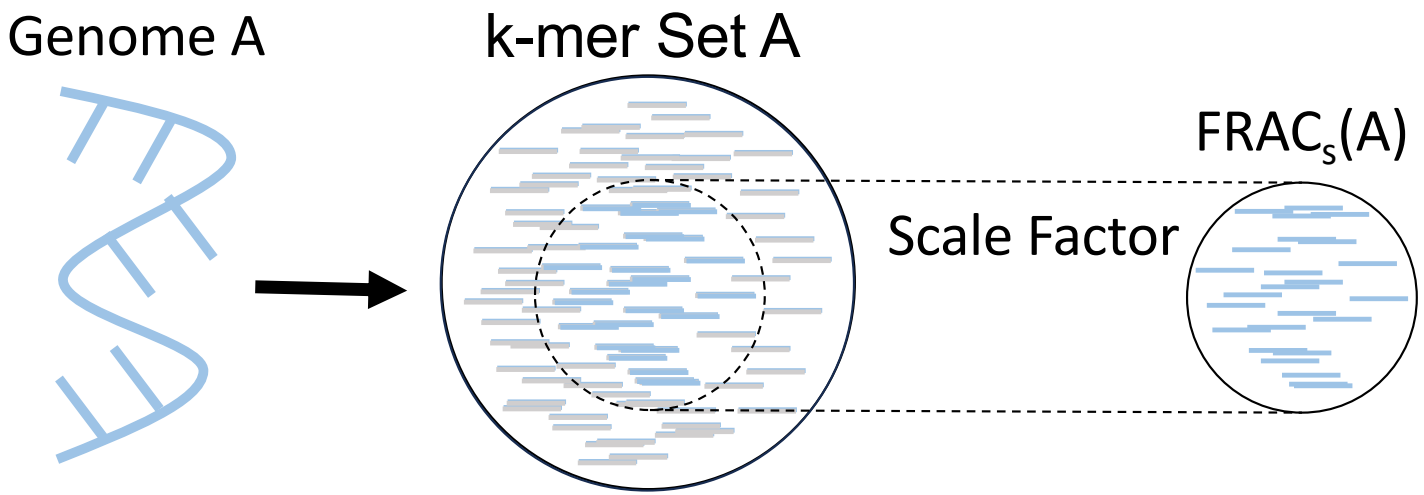
https://sourmash.readthedocs.io/

Brown, C. T., & Irber, L. (2016). sourmash: a library for MinHash sketching of DNA. Journal of open source software, 1(5), 27.

sourmash sketch

first step in
sourmash analyses

The **FracMinHash** sketch

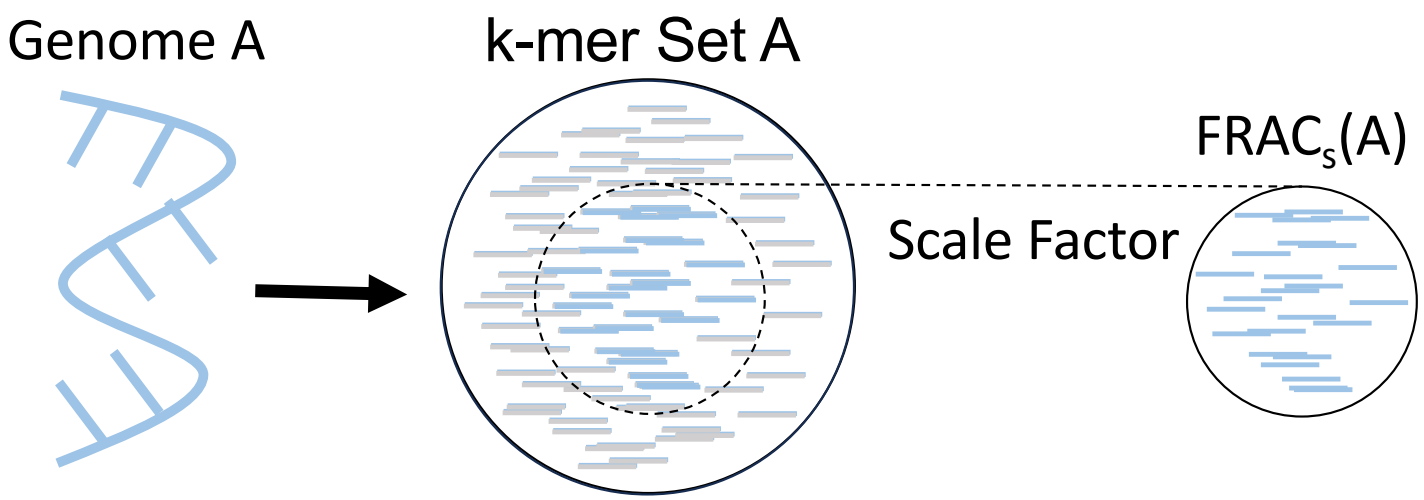


$$FRAC_s(A) = \{h(a) | a \in A \text{ and } h(a) \leq Hs\}$$

Hera, M. R., Pierce-Ward, N. T., & Koslicki, D. (2023). Deriving confidence intervals for mutation rates across a wide range of evolutionary distances using FracMinHash. Genome research, 33(7), 1061-1068.

sourmash sketch

producing a FracMinHash sketch

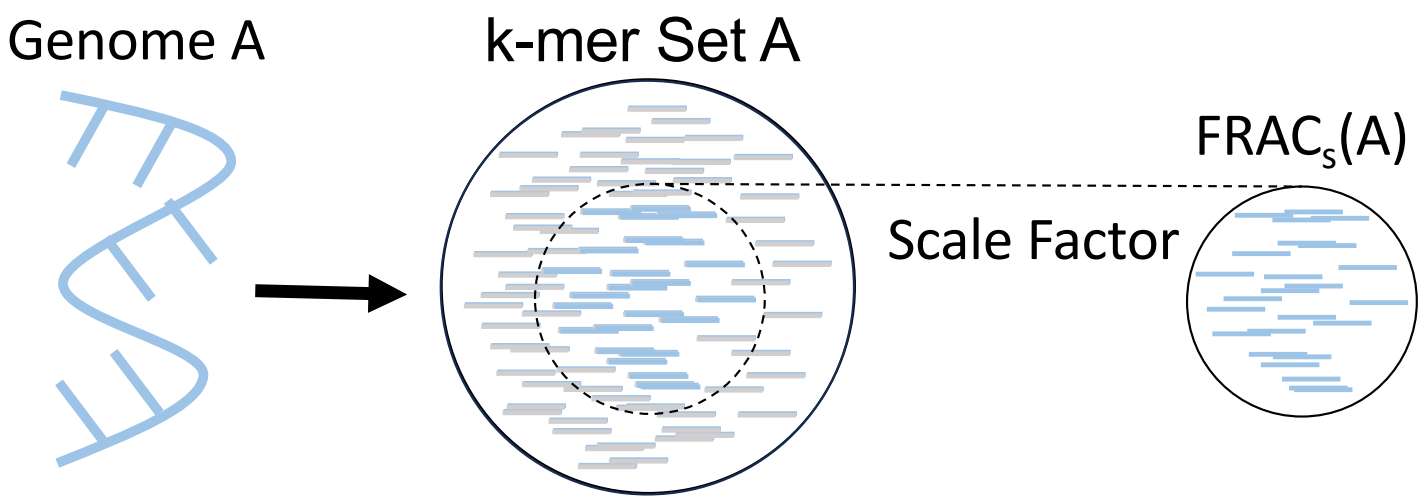



```
sourmash sketch <dna, protein, translate> <FASTA/Q> -p k=<ksize>,scaled=<int>
```

sourmash command

sourmash sketch

producing a FracMinHash sketch

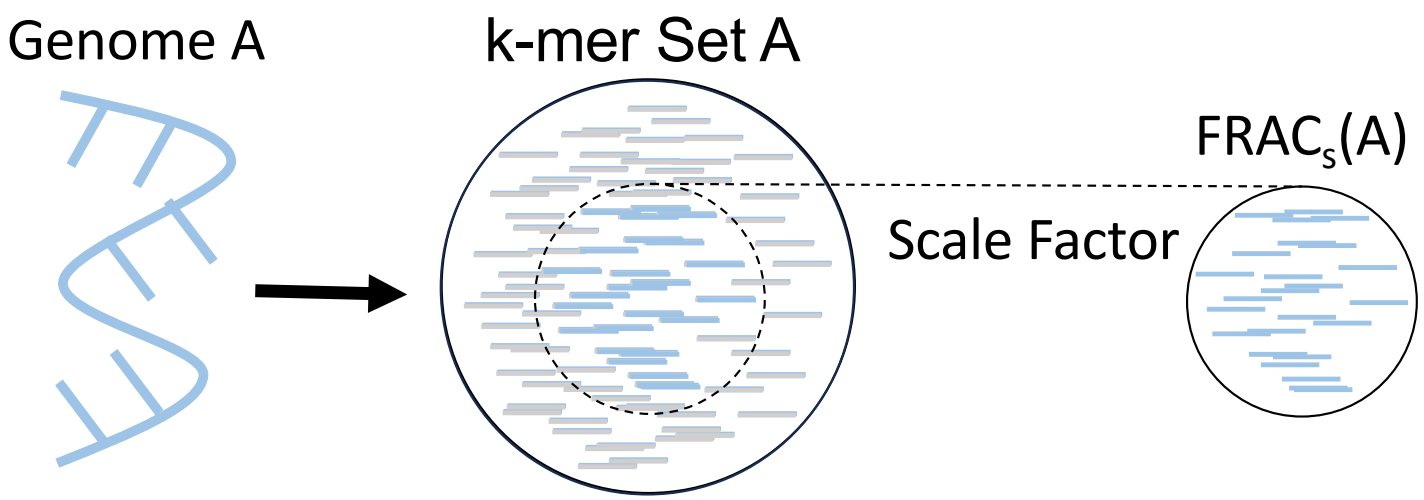



```
sourmash sketch <dna, protein, translate> <FASTA/Q> -p k=<ksize>,scaled=<int>
```

identify the type of input sequences

sourmash sketch

producing a FracMinHash sketch

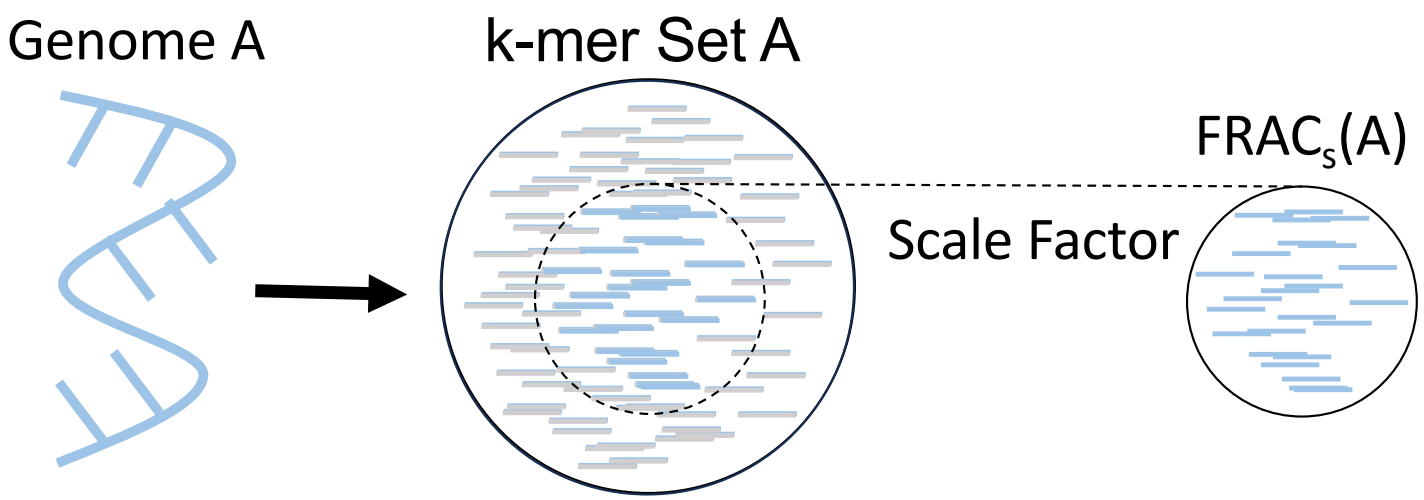



```
sourmash sketch <dna, protein, translate> <FASTA/Q> -p k=<ksize>,scaled=<int>
```

Identify the filename with sequences

sourmash sketch

producing a FracMinHash sketch

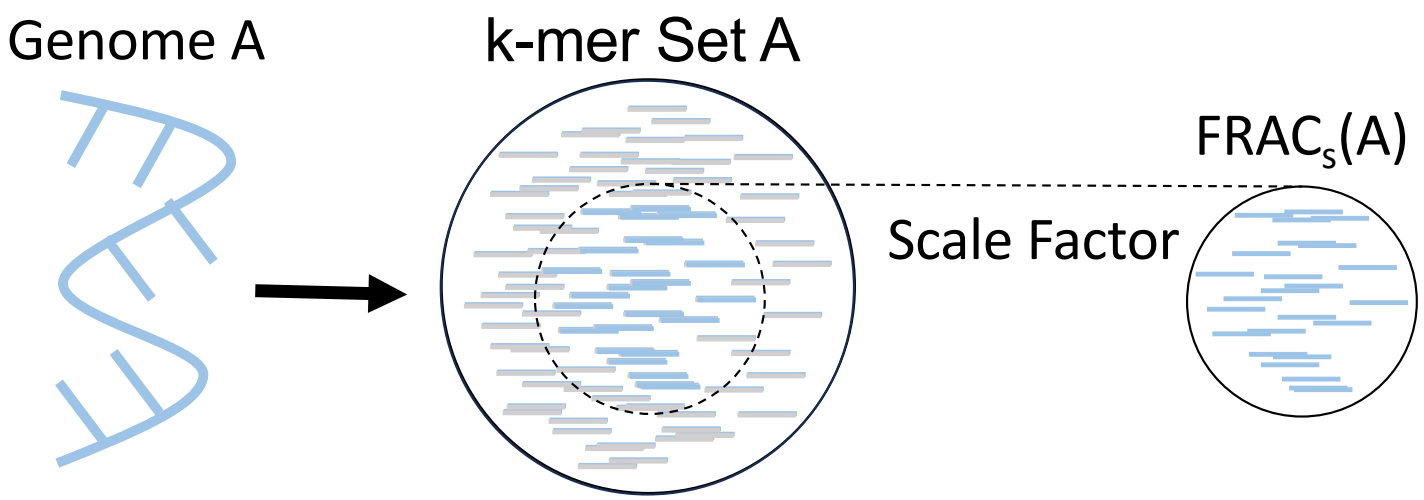



```
sourmash sketch <dna, protein, translate> <FASTA/Q> -p k=<ksize>,scaled=<int>
```

Flag required to tune parameters

sourmash sketch

producing a FracMinHash sketch

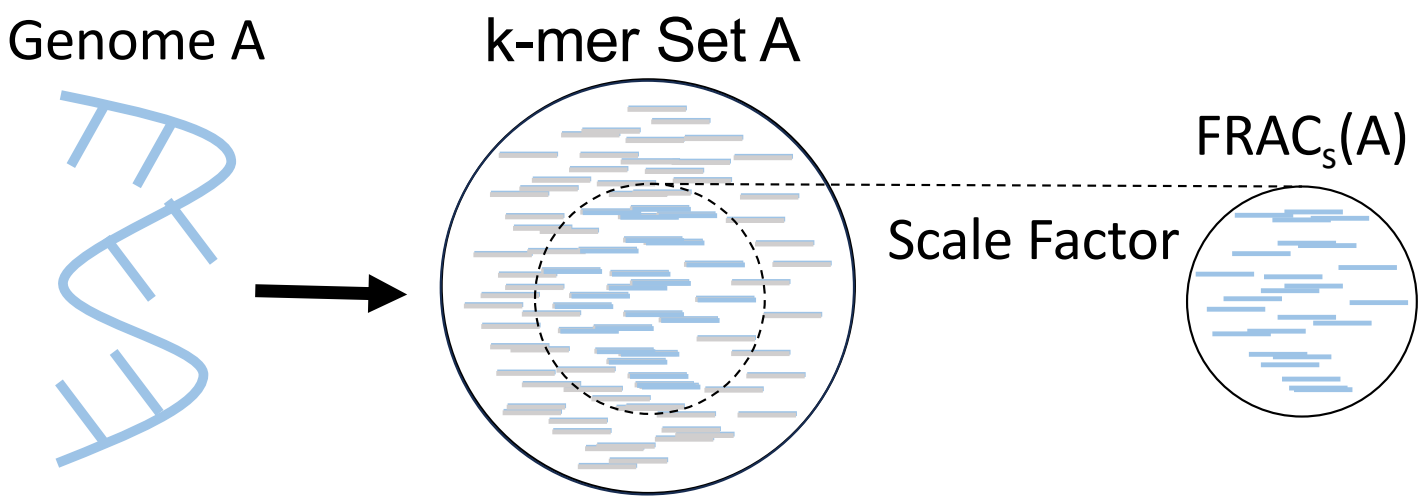



```
sourmash sketch <dna, protein, translate> <FASTA/Q> -p k=<ksize>,scaled=<int>
```

Identify the filename with sequences

sourmash sketch

producing a FracMinHash sketch

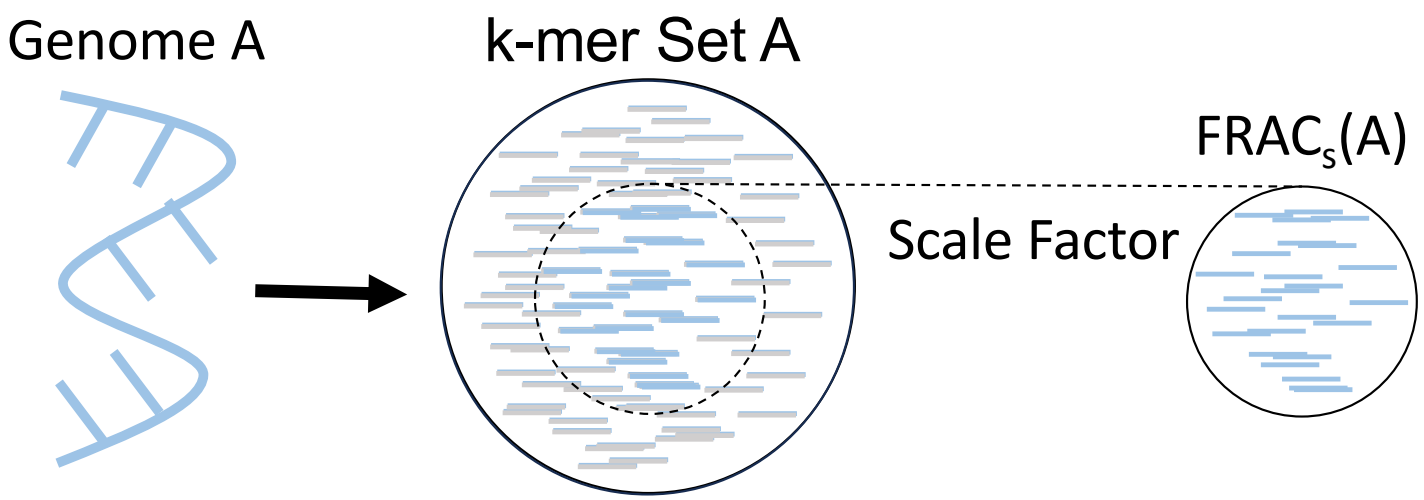



```
sourmash sketch <dna, protein, translate> <FASTA/Q> -p k=<ksize>,scaled=<int>
```

scales down sketch

sourmash sketch

producing a FracMinHash sketch




```
sourmash sketch <dna, protein, translate> <FASTA/Q> -p k=<ksize>,scaled=<int>
```

Example: scaled=10
will keep 1 of 10 *k*-mers

sourmash compare

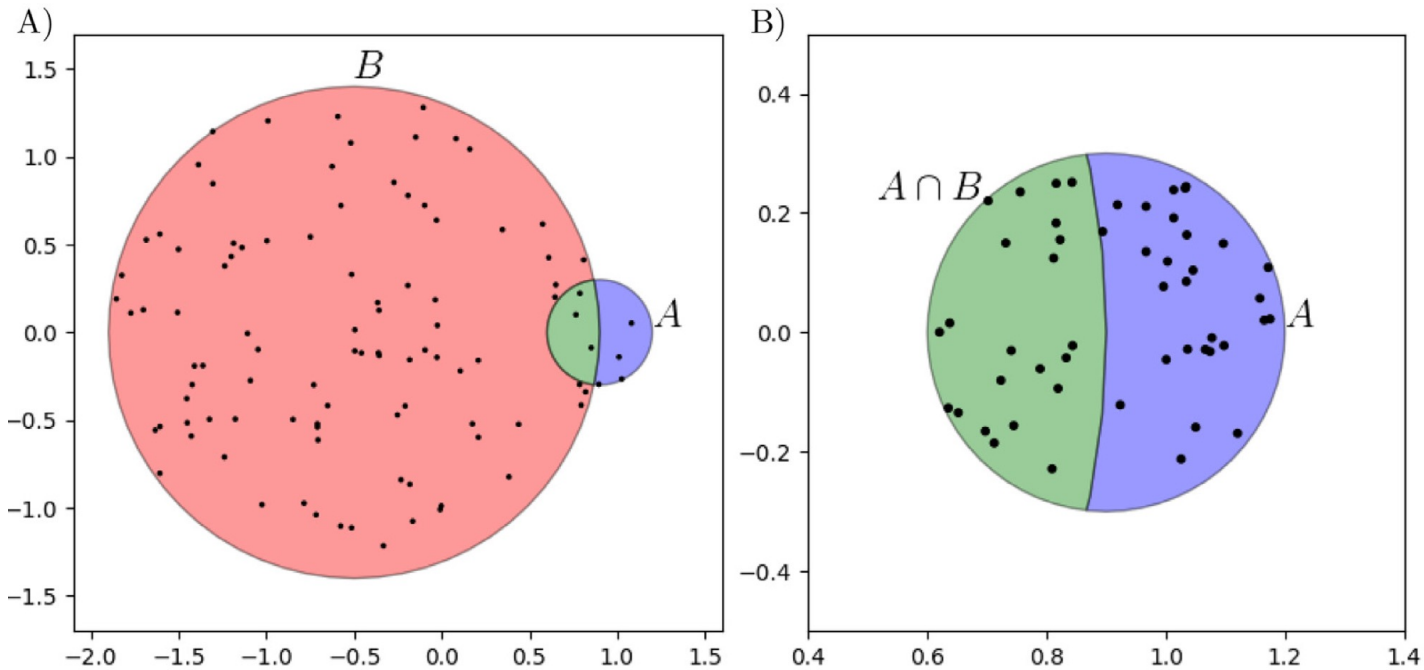
perform pairwise similarity estimations between metagenomic samples

Estimating similarity indexes: **jaccard**

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \approx \frac{|S(A \cup B) \cap S(A) \cap S(B)|}{|S(A \cup B)|}$$

Indexes close to 1 are interpreted as more similar

However, the **jaccard** estimate becomes worse with differing set sizes

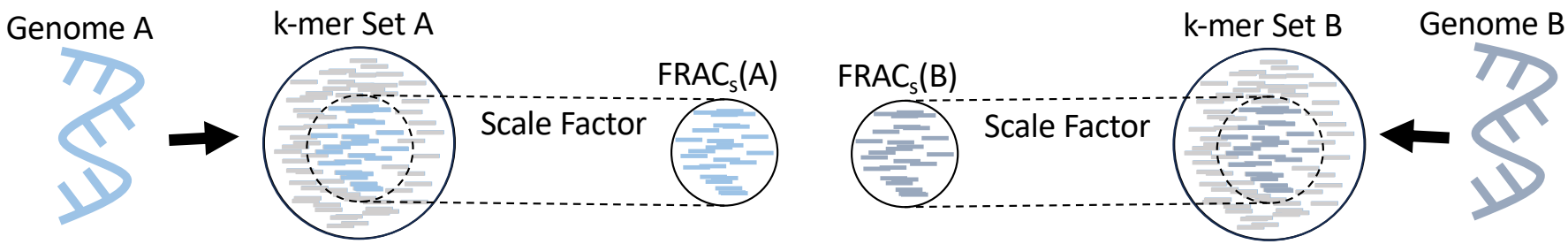


…further, bloom filters are limited to smaller datasets

Koslicki and Zabeti 2019

Estimating similarity indexes: **containment**

…addresses issues using in jaccard

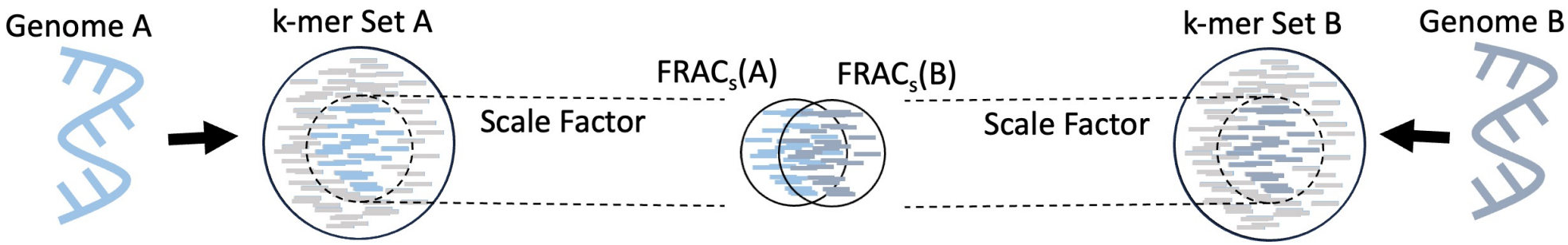


$$C_{frac}(A,B) := \frac{|\boldsymbol{FRAC_s}(A) \cap \boldsymbol{FRAC_s}(B)|}{|\boldsymbol{FRAC_s}(A)|(1-(1-s)^{|A|})} \mathbb{1}_{|\boldsymbol{FRAC_s}(A)|>0}$$

Hera and colleagues 2023

sourmash compare

Estimate the containment between two genomes

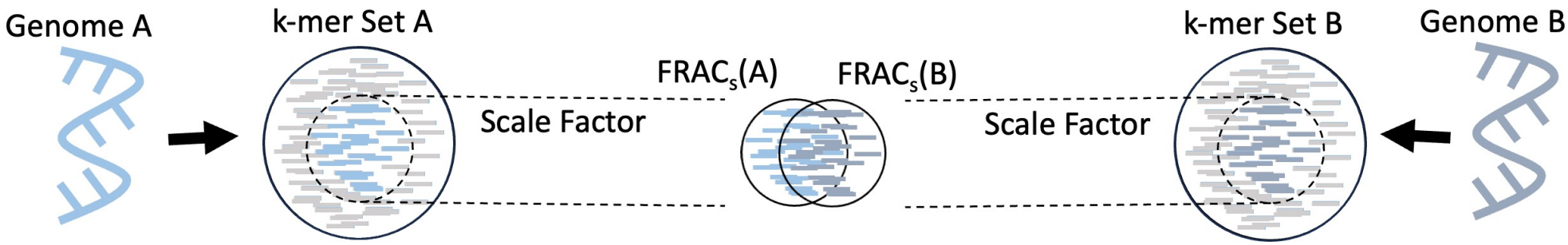



```
sourmash compare <ref signature> <query signature> --containment --ksize <ksize>
```

sourmash command

sourmash compare

Estimate the containment between two genomes

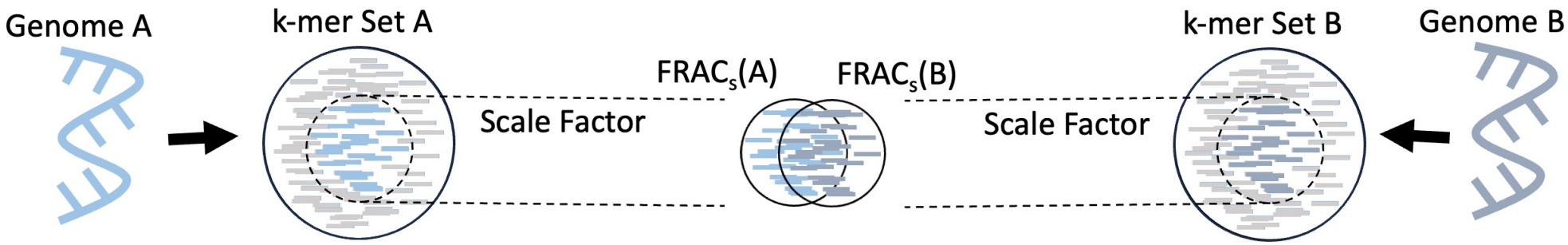



```
sourmash compare <query signature> <ref signature> --containment --ksize <ksize>
```

sample of interest

sourmash compare

Estimate the containment between two genomes

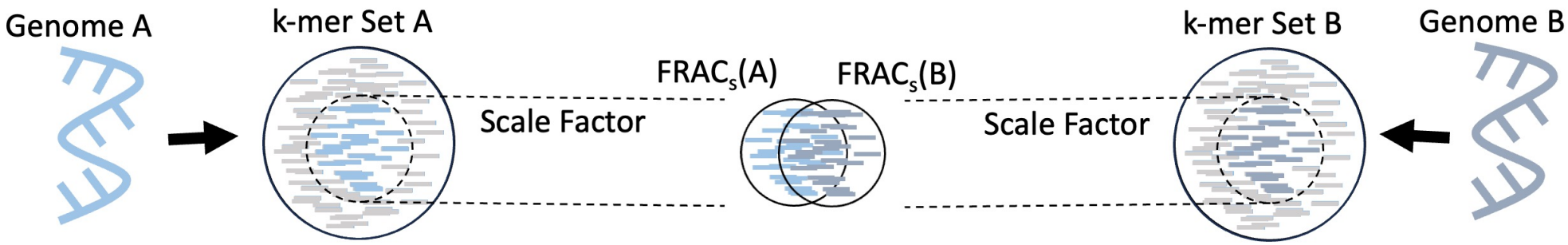



```
sourmash compare <query signature> <ref signature> --containment --ksize <ksize>
```

compared to another sample

sourmash compare

Estimate the containment between two genomes

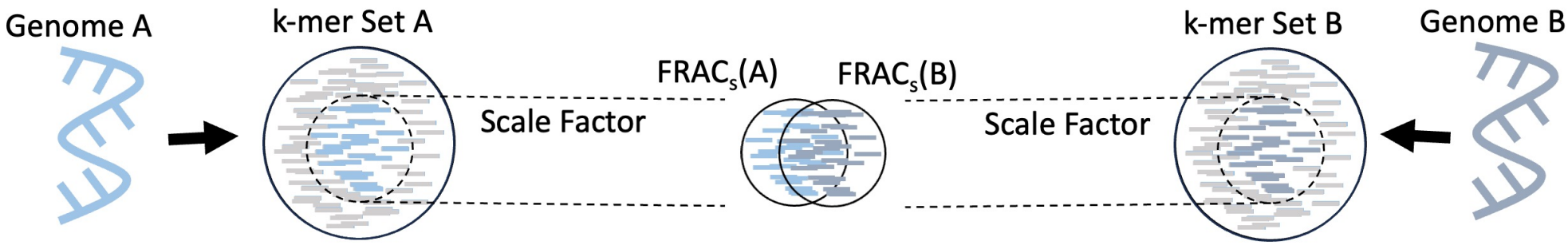



```
sourmash compare <query signature> <ref signature> --containment --ksize <ksize>
```

indicate the similarity index

sourmash compare

Estimate the containment between two genomes




```
sourmash compare <query signature> <ref signature> --containment --ksize <ksize>
```

sourmash search

report highly similar sequences data between samples

sourmash search

First, produce a MANIFEST.csv to search using an md5 identifier as the query

```
sourmash sig manifest <signature file> -o MANIFEST.csv
```

command for further options to information from signatures

sourmash search

First, produce a MANIFEST.csv to search using an md5 identifier as the query

```
sourmash sig manifest <signature file> -o MANIFEST.csv
```

option to create MANIFEST

sourmash search

First, produce a MANIFEST.csv to search using an md5 identifier as the query

```
sourmash sig manifest <signature file> -o MANIFEST.csv
```

input signature query of interest

sourmash search

First, produce a MANIFEST.csv to search using an md5 identifier as the query

```
sourmash sig manifest <signature file> -o MANIFEST.csv
```

output name flag

sourmash search

First, produce a MANIFEST.csv to search using an md5 identifier as the query

```
sourmash sig manifest <signature file> -o sample_001.manifest.csv
```

```
# SOURMASH-MANIFEST-VERSION: 1.0
internal_location,md5,md5short,ksize,moltype,num,scaled,n_hashes,with_abundance,name,filename
sample_001.fna.singleton.sig,d3513280b35b2a918a7181875c0683c8,d3513280,31,DNA,0,500,22,0,genome_A,sample_001.fna
sample_001.fna.singleton.sig,2d96ee330b6b295a06b70fdbfb75af34,2d96ee33,31,DNA,0,500,21,0,genome_B,sample_001.fna
sample_001.fna.singleton.sig,364c20a1ca43d3ae3e9ae7d9e9a6c837,364c20a1,31,DNA,0,500,19,0,genome_C,sample_001.fna
```

Column Name	Description
internal_location	Name of signature file
md5	Name for sketch
ksize	Size used to produce *k*-mers
moltype	Type of sequence: DNA, Protein
scaled	Parameter used for sourmash sketch to reduce *k*-mer set
n_hashes	Total k-mers in final sketch
name	Sequence name in FASTA/Q
filename	Input FASTA/Q name

sourmash search

First, produce a MANIFEST.csv to search using an md5 identifier as the query

```
sourmash sig manifest <signature file> -o sample_001.manifest.csv
```

```
# SOURMASH-MANIFEST-VERSION: 1.0
internal_location,md5,md5short,ksize,moltype,num,scaled,n_hashes,with_abundance,name,filename
sample_001.fna.singleton.sig,d3513280b35b2a918a7181875c0683c8,d3513280,31,DNA,0,500,22,0,genome_A,sample_001.fna
sample_001.fna.singleton.sig,2d96ee330b6b295a06b70fdbfb75af34,2d96ee33,31,DNA,0,500,21,0,genome_B,sample_001.fna
sample_001.fna.singleton.sig,364c20a1ca43d3ae3e9ae7d9e9a6c837,364c20a1,31,DNA,0,500,19,0,genome_C,sample_001.fna
```

```
sourmash search <signature query> <signature ref> --md5 d3513280b35b2a918a7181875c0683c8 --containment
```

command to searching signatures

sourmash search

First, produce a MANIFEST.csv to search using an md5 identifier as the query

```
sourmash sig manifest <signature file> -o sample_001.manifest.csv
```

```
# SOURMASH-MANIFEST-VERSION: 1.0
internal_location,md5,md5short,ksize,moltype,num,scaled,n_hashes,with_abundance,name,filename
sample_001.fna.singleton.sig,d3513280b35b2a918a7181875c0683c8,d3513280,31,DNA,0,500,22,0,genome_A,sample_001.fna
sample_001.fna.singleton.sig,2d96ee330b6b295a06b70fdbfb75af34,2d96ee33,31,DNA,0,500,21,0,genome_B,sample_001.fna
sample_001.fna.singleton.sig,364c20a1ca43d3ae3e9ae7d9e9a6c837,364c20a1,31,DNA,0,500,19,0,genome_C,sample_001.fna
```

```
sourmash search <signature query> <signature ref> --md5 d3513280b35b2a918a7181875c0683c8 --containment
```

signature file where md5 of interest is found

sourmash search

First, produce a MANIFEST.csv to search using an md5 identifier as the query

```
sourmash sig manifest <signature file> -o sample_001.manifest.csv
```

```
# SOURMASH-MANIFEST-VERSION: 1.0
internal_location,md5,md5short,ksize,moltype,num,scaled,n_hashes,with_abundance,name,filename
sample_001.fna.singleton.sig,d3513280b35b2a918a7181875c0683c8,d3513280,31,DNA,0,500,22,0,genome_A,sample_001.fna
sample_001.fna.singleton.sig,2d96ee330b6b295a06b70fdbfb75af34,2d96ee33,31,DNA,0,500,21,0,genome_B,sample_001.fna
sample_001.fna.singleton.sig,364c20a1ca43d3ae3e9ae7d9e9a6c837,364c20a1,31,DNA,0,500,19,0,genome_C,sample_001.fna
```

```
sourmash search <signature query> <signature ref> --md5 d3513280b35b2a918a7181875c0683c8 --containment
```

signature file that will be searched against

sourmash search

First, produce a MANIFEST.csv to search using an md5 identifier as the query

```
sourmash sig manifest <signature file> -o sample_001.manifest.csv
```

```
# SOURMASH-MANIFEST-VERSION: 1.0
internal_location,md5,md5short,ksize,moltype,num,scaled,n_hashes,with_abundance,name,filename
sample_001.fna.singleton.sig,d3513280b35b2a918a7181875c0683c8,d3513280,31,DNA,0,500,22,0,genome_A,sample_001.fna
sample_001.fna.singleton.sig,2d96ee330b6b295a06b70fdbfb75af34,2d96ee33,31,DNA,0,500,21,0,genome_B,sample_001.fna
sample_001.fna.singleton.sig,364c20a1ca43d3ae3e9ae7d9e9a6c837,364c20a1,31,DNA,0,500,19,0,genome_C,sample_001.fna
```

```
sourmash search <signature query> <signature ref> --md5 d3513280b35b2a918a7181875c0683c8 --containment
```

md5 identifier of interest indicated

sourmash search

First, produce a MANIFEST.csv to search using an md5 identifier as the query

```
sourmash sig manifest <signature file> -o sample_001.manifest.csv
```

```
# SOURMASH-MANIFEST-VERSION: 1.0
internal_location,md5,md5short,ksize,moltype,num,scaled,n_hashes,with_abundance,name,filename
sample_001.fna.singleton.sig,d3513280b35b2a918a7181875c0683c8,d3513280,31,DNA,0,500,22,0,genome_A,sample_001.fna
sample_001.fna.singleton.sig,2d96ee330b6b295a06b70fdbfb75af34,2d96ee33,31,DNA,0,500,21,0,genome_B,sample_001.fna
sample_001.fna.singleton.sig,364c20a1ca43d3ae3e9ae7d9e9a6c837,364c20a1,31,DNA,0,500,19,0,genome_C,sample_001.fna
```

```
sourmash search <signature query> <signature ref> --md5 d3513280b35b2a918a7181875c0683c8 --containment
```

indicate the similarity index

sourmash search

First, produce a MANIFEST.csv to search using an md5 identifier as the query

```
sourmash sig manifest <signature file> -o sample_001.manifest.csv
```

```
# SOURMASH-MANIFEST-VERSION: 1.0
internal_location,md5,md5short,ksize,moltype,num,scaled,n_hashes,with_abundance,name,filename
sample_001.fna.singleton.sig,d3513280b35b2a918a7181875c0683c8,d3513280,31,DNA,0,500,22,0,genome_A,sample_001.fna
sample_001.fna.singleton.sig,2d96ee330b6b295a06b70fdbfb75af34,2d96ee33,31,DNA,0,500,21,0,genome_B,sample_001.fna
sample_001.fna.singleton.sig,364c20a1ca43d3ae3e9ae7d9e9a6c837,364c20a1,31,DNA,0,500,19,0,genome_C,sample_001.fna
```

```
sourmash search <signature query> <signature ref> --md5 d3513280b35b2a918a7181875c0683c8 --containment
```

```
select query k=31 automatically.
loaded query: genome_A (k=31, DNA)
-
loaded 10 total signatures from 1 locations. after selecting signatures compatible with
search, 10 remain.

9 matches above threshold 0.080; showing first 3:
similarity match
---------- -----
100.0% ref_gene
90.9% genome_A_2
86.4% genome_A_1
```

sourmash overview

sourmash overview

Task
sketch a signature file

Commands
```
sourmash sketch
```

Output
```
sketch.sig
```

Produce a scaled down representation of your sequencing data

Task
similarity index matrix

Commands
```
sourmash compare
```

Output
```
similarity_matrix.csv
```

Report similarity between sequencing data

Task
report similarity highly similar sketches

Commands
```
sourmash search
```

Output
```
similarity match
---------- -----
100.0% ref_gene
90.9% genome_A_2
86.4% genome_A_1
```

Match highly similar sequences data

sourmash tutorial

https://github.com/KoslickiLab/sourmash_tutorial