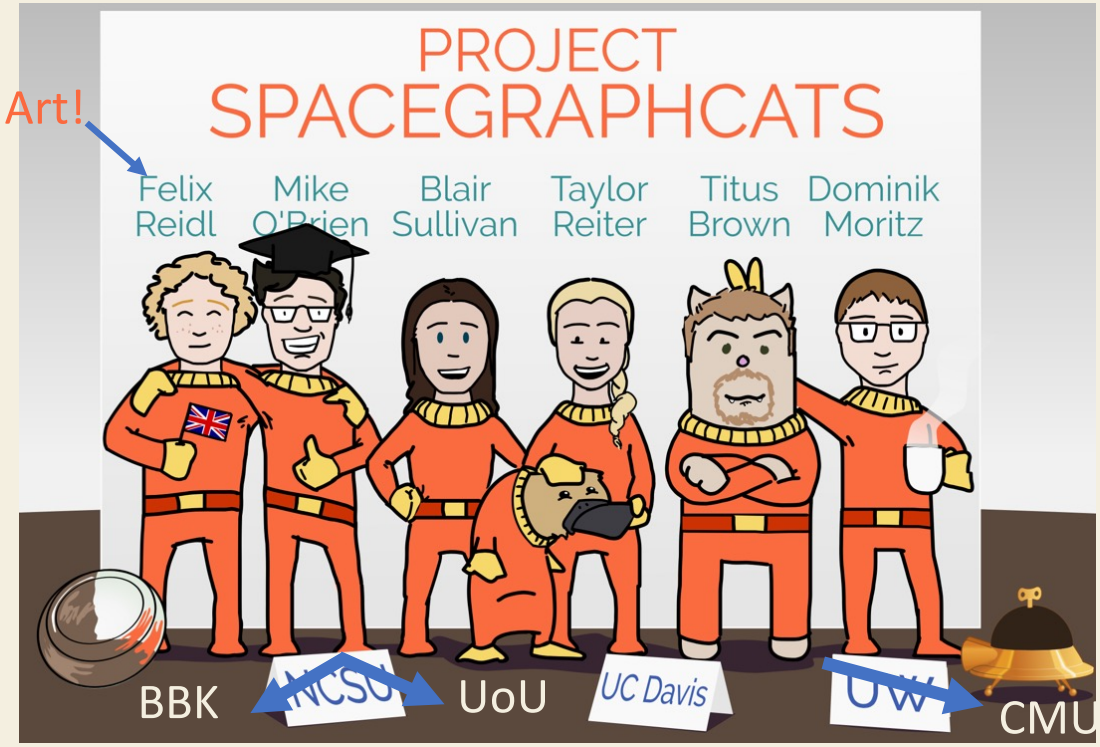
Introduction to spacegraphcats

May 7, 2021

Taylor Reiter, PhD
Lab for Data Intensive Biology
University of California, Davis

@ReiterTaylor
tereiter@ucdavis.edu
taylorreiter



Slides: github.com/taylorreiter/2021-sgc-binder

github.com/spacegraphcats/spacegraphcats

https://spacegraphcats.github.io/spacegraphcats

Why spacegraphcats?

- Can we use graphs to learn more about metagenome sequences?

NB spacegraphcats is a play on words from compounding *sparse graph cuts*

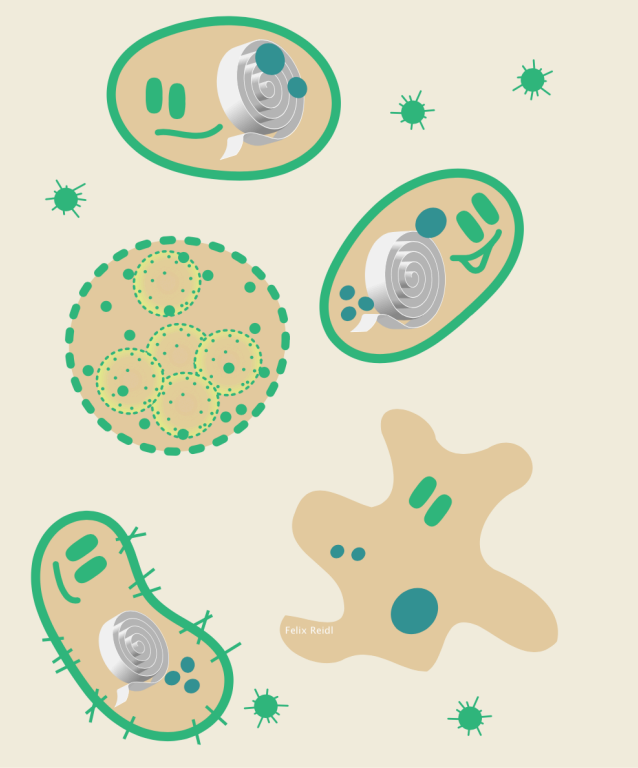
Outline

- Metagenome analysis methods and their problems
- Graphs and why they’re good
- An introduction to spacegraphcats

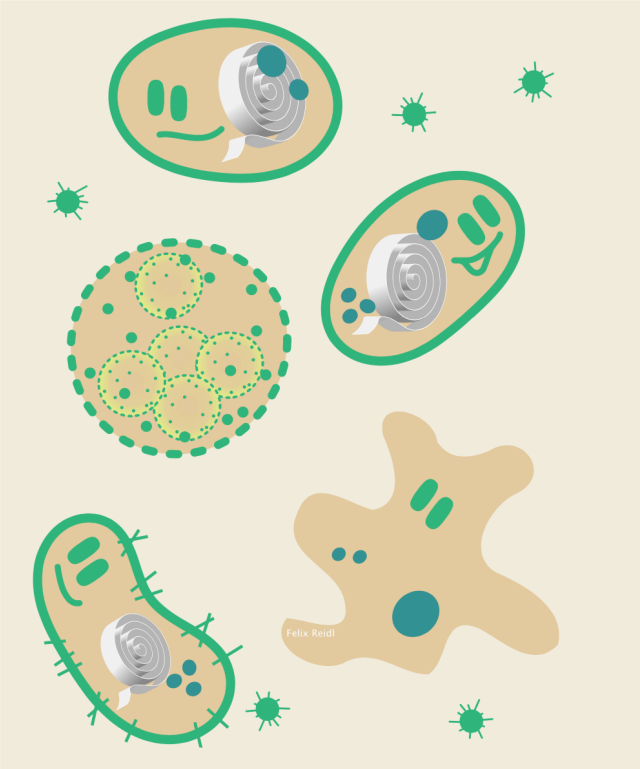
Outline

- Metagenome analysis methods and their problems
- Graphs and why they’re good
- An introduction to spacegraphcats

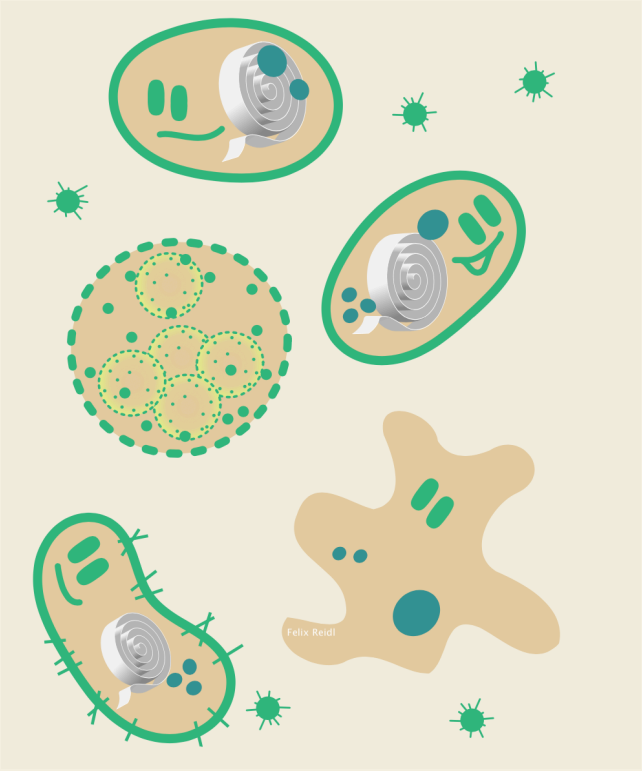
Metagenomics



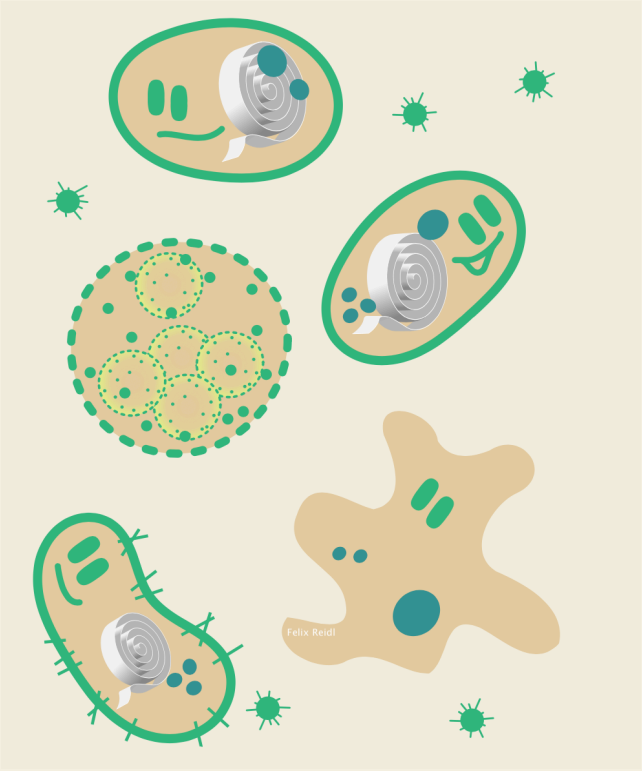
Metagenomics



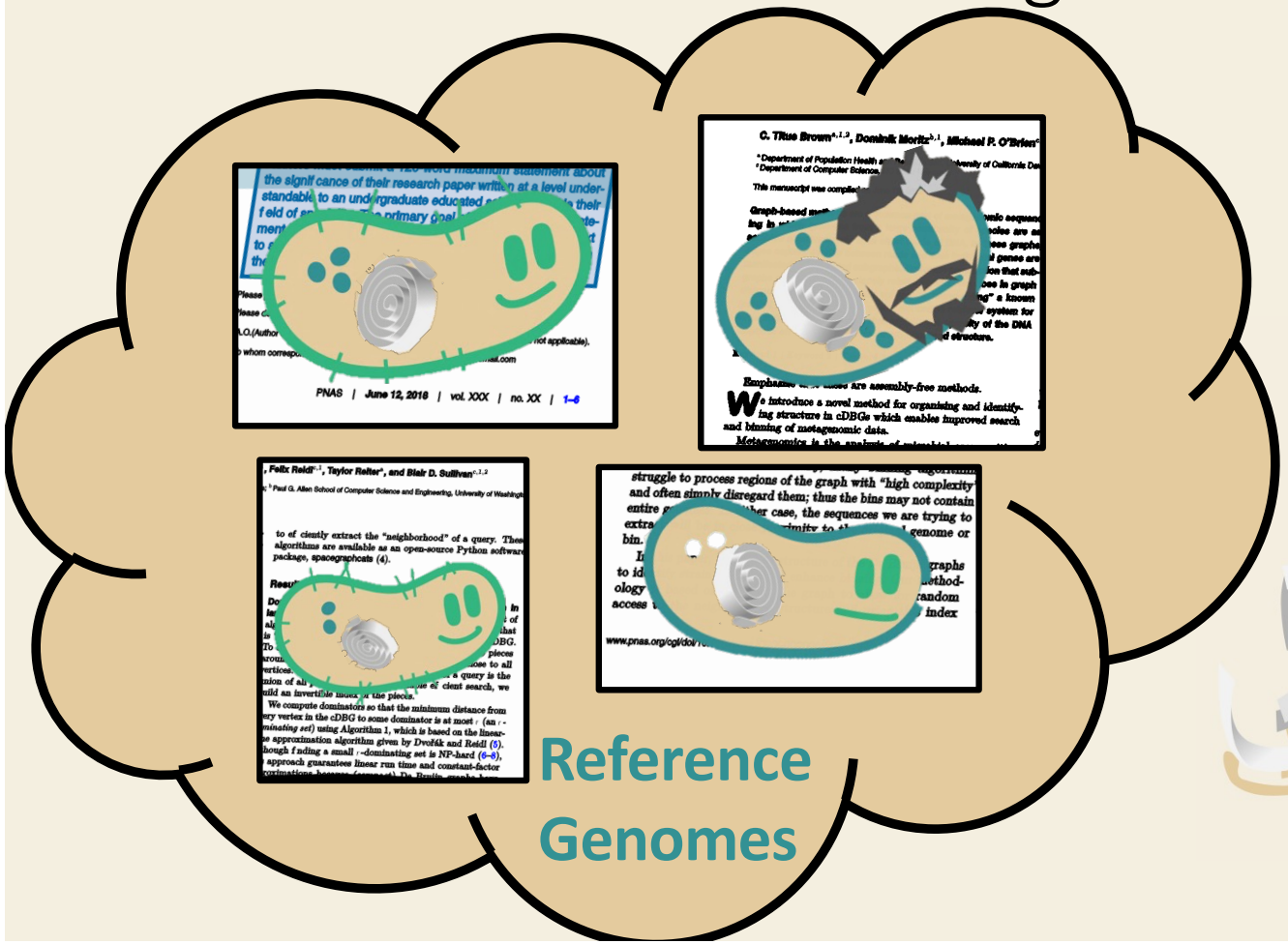
Metagenomics



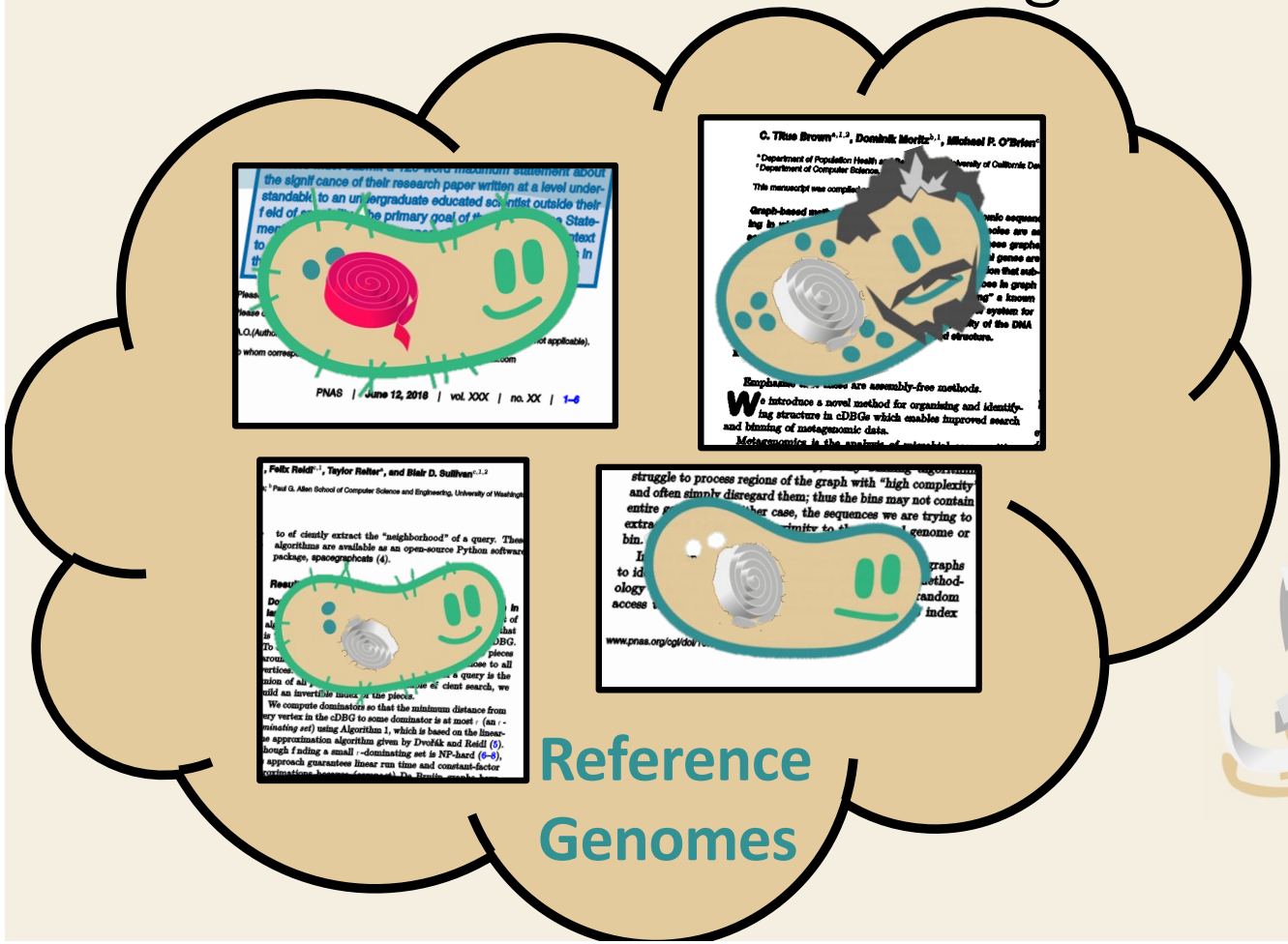
Metagenomics



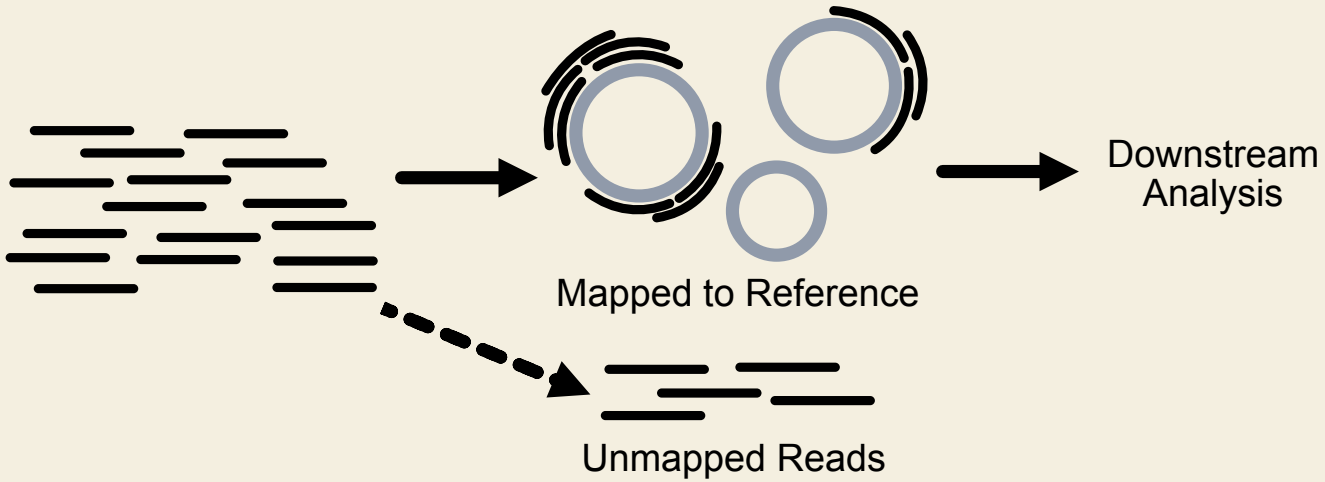
Referenced-based metagenome analysis



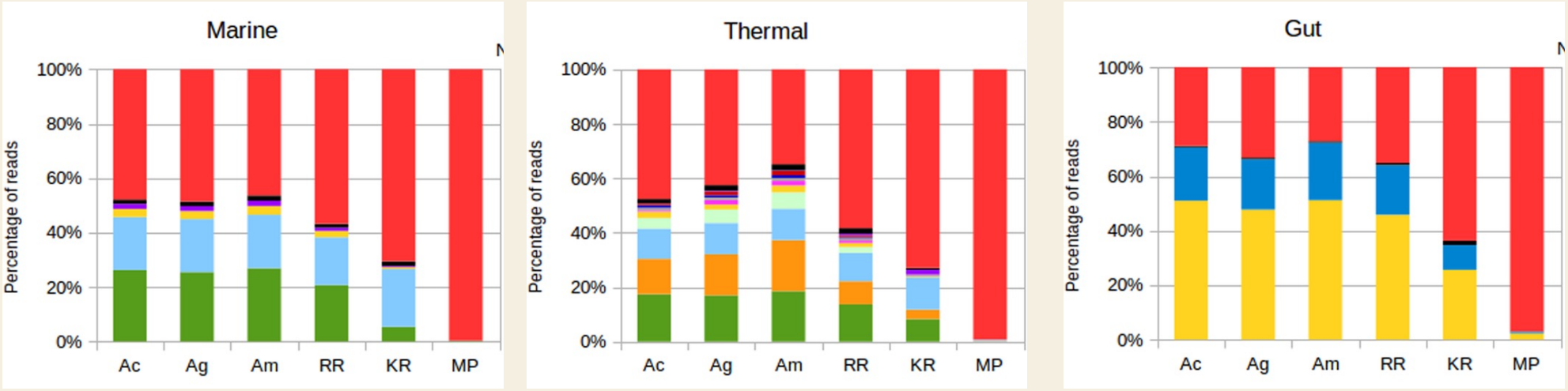
Referenced-based metagenome analysis



Referenced-based metagenome analysis

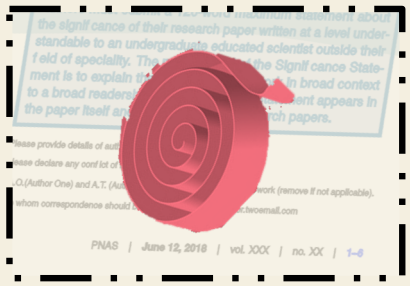


Reference-based methods leave a lot unanalyzed



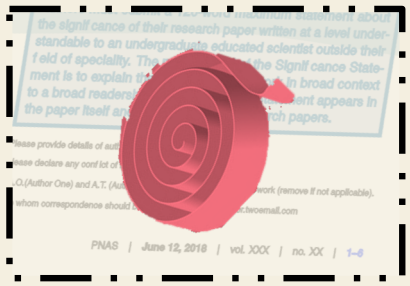
Tamames et al. 2019 https://doi.org/10.1186/s12864-019-6289-6

De novo metagenome analysis



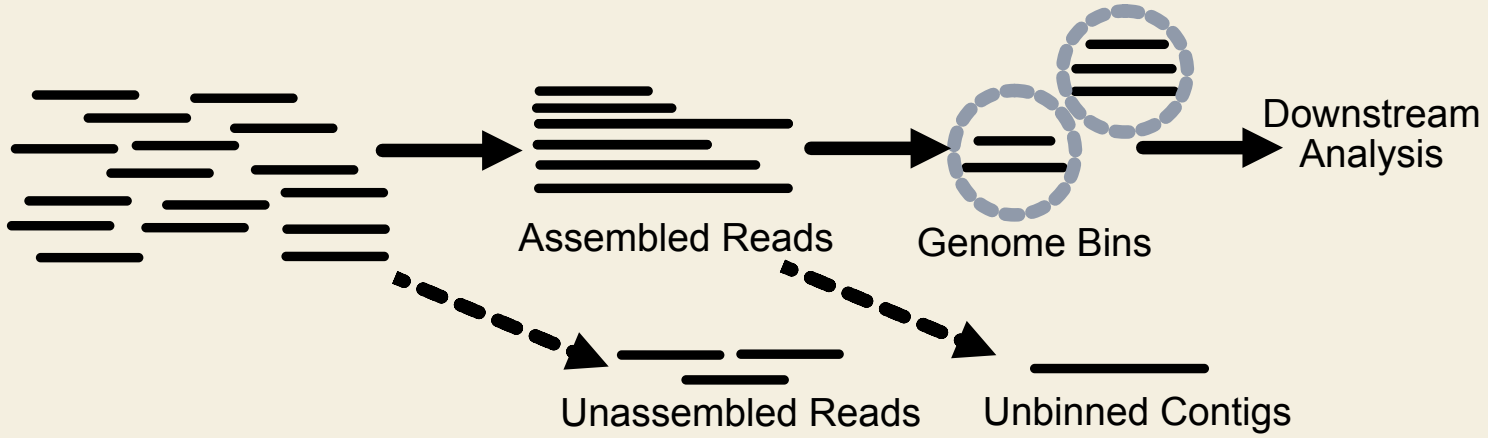
assembly & binning

De novo metagenome analysis

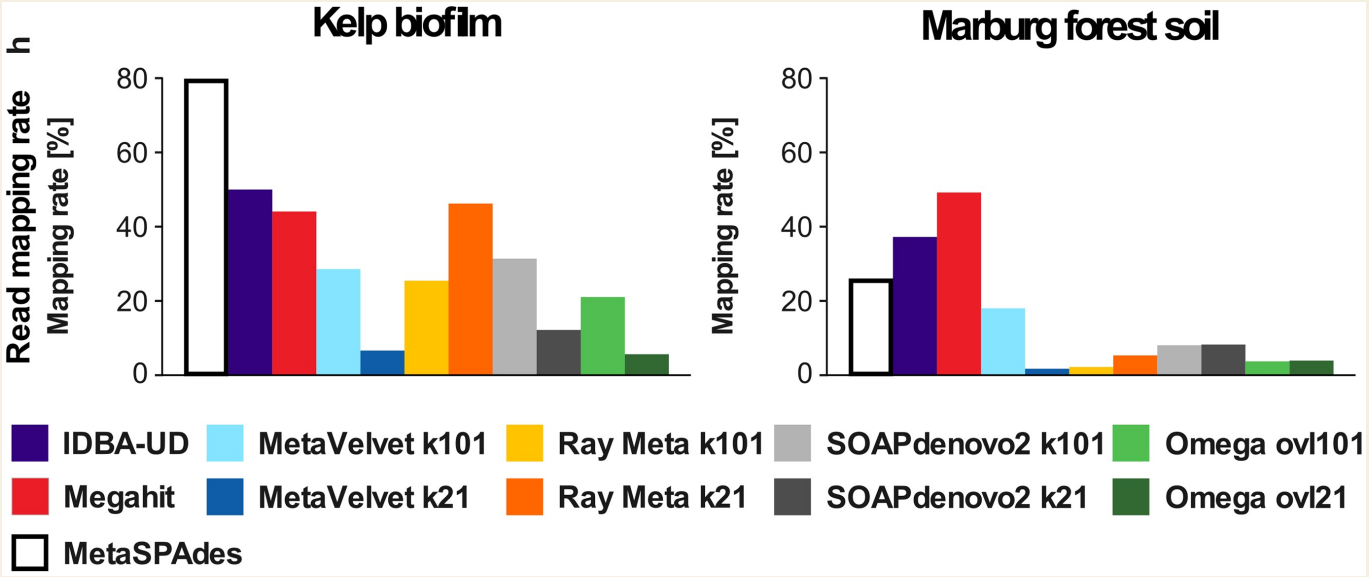


assembly & binning

De novo metagenome analysis



de novo methods leave a lot unanalyzed



Vollmers et al. 2017 https://doi.org/10.1371/journal.pone.0169662

Both reference and *de novo* metagenome analysis techniques leave ~10-90% of (short) sequencing reads unanalyzed

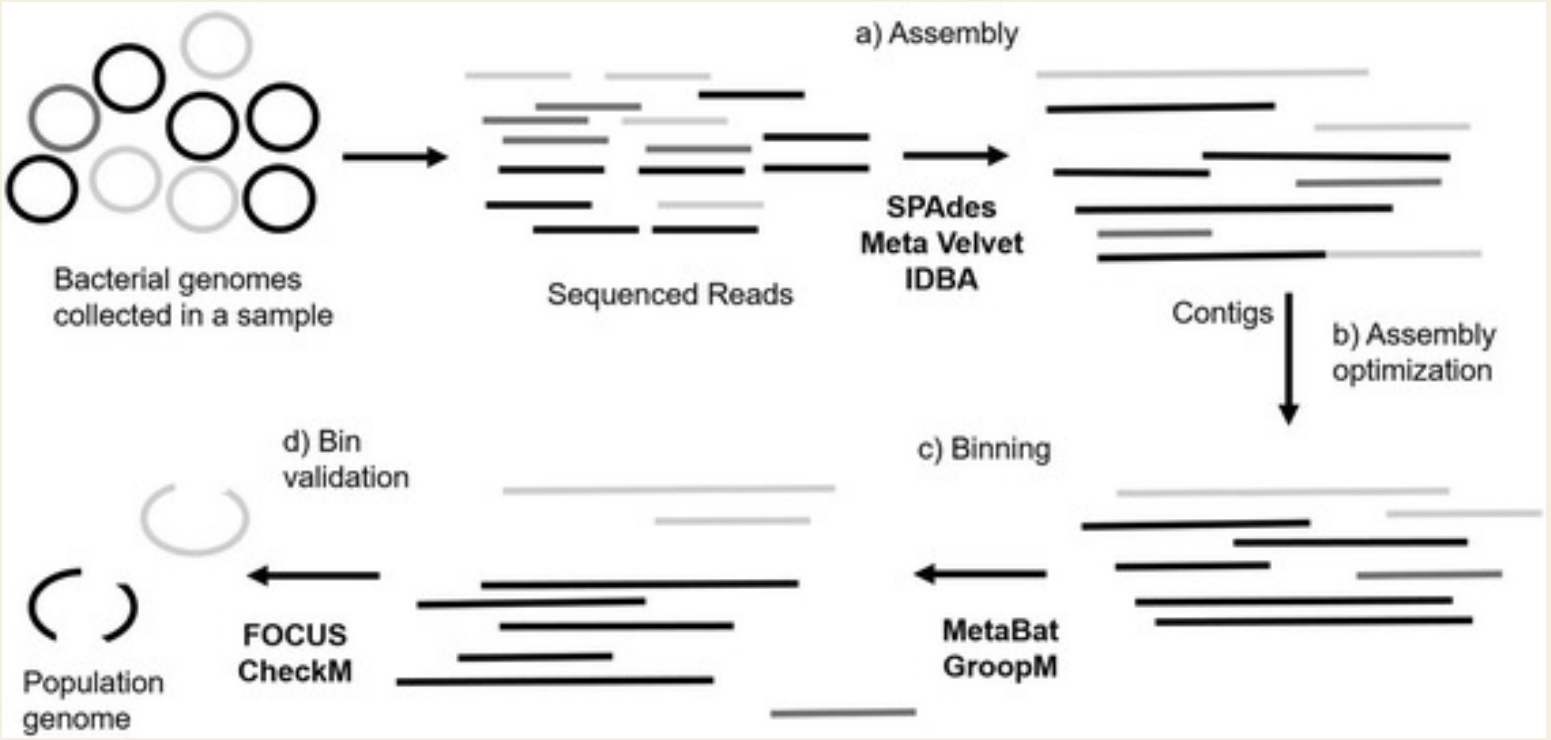
What’s in the leftovers?

- **Spacegraphcats** can help you find out!

Outline

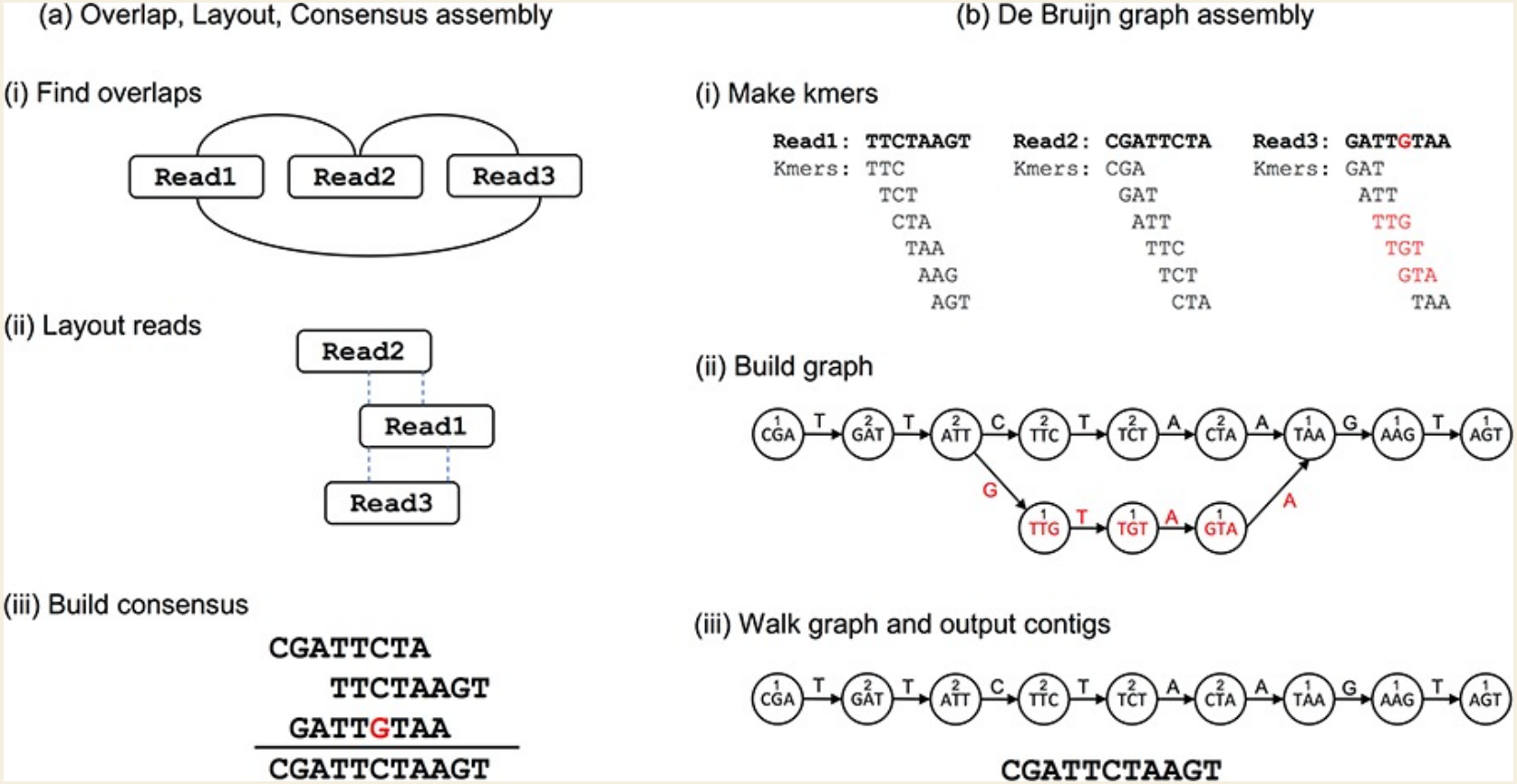
- Metagenome analysis methods and their problems
- Graphs and why they're good
- An introduction to spacegraphcats

Metagenome assembly



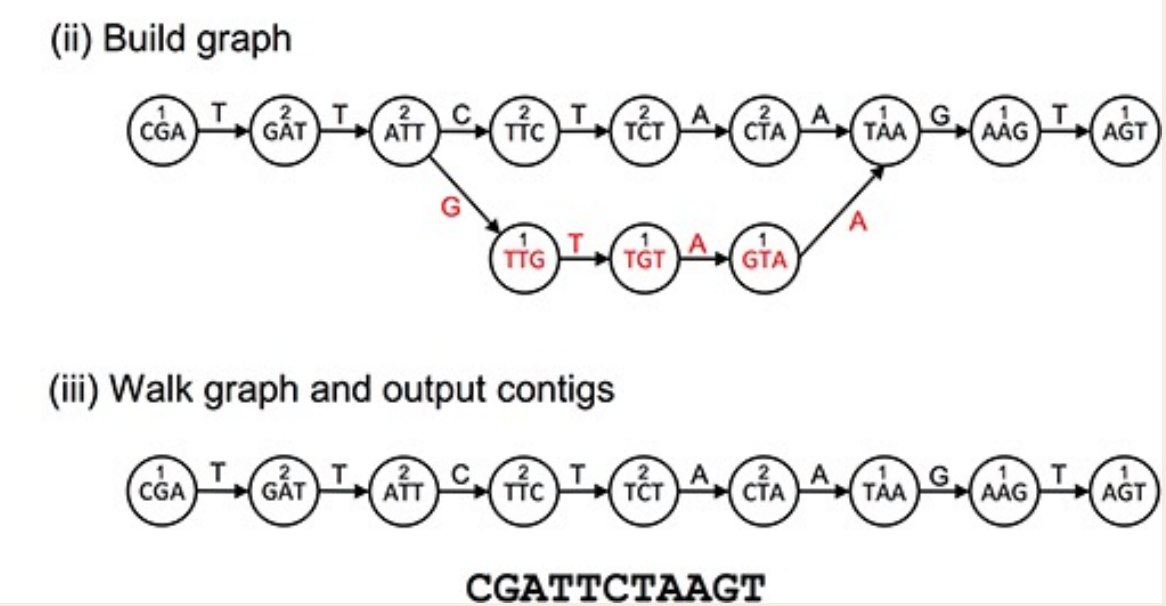
Papudeshi et al. https://doi.org/10.1186/s12864-017-4294-1

Metagenome assembly



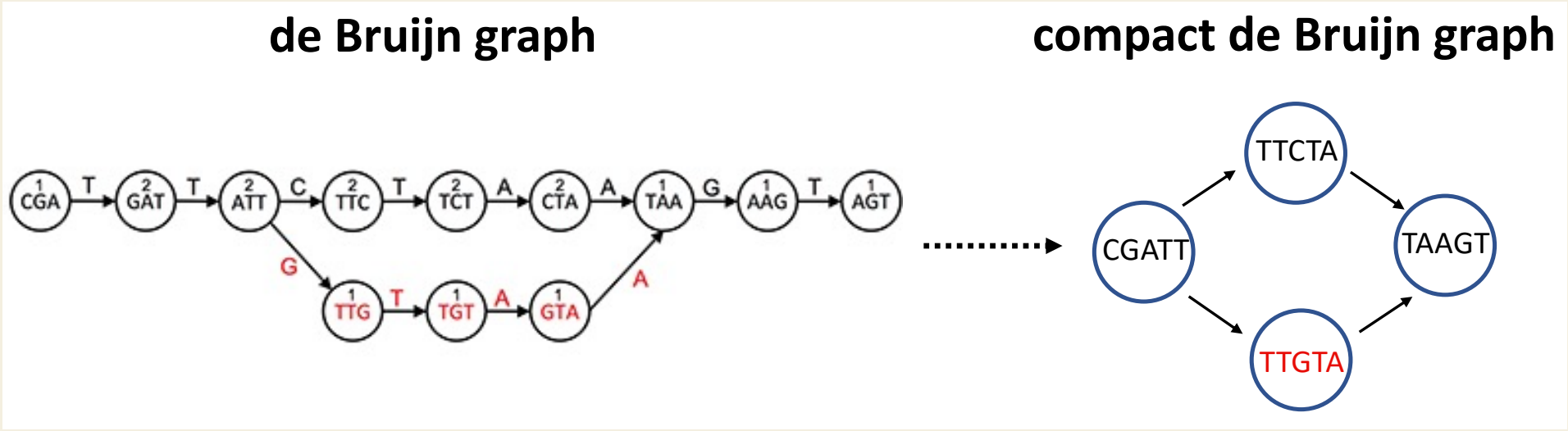
Ayling et al. 2020 https://doi.org/10.1093/bib/bbz020

Unlike assemblies which are forced to make “choices” in regions of sequence complexity, graphs represent all sequencing data



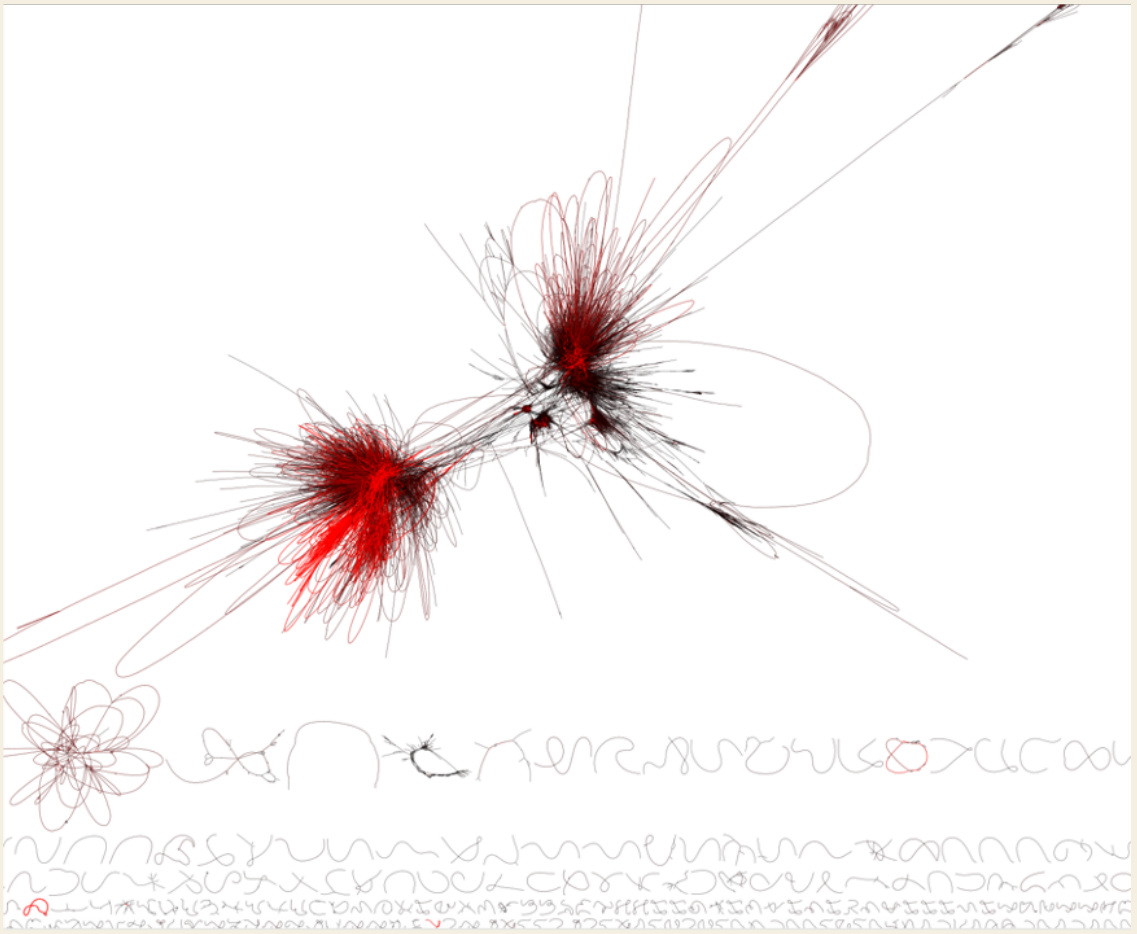
Ayling et al. 2020 https://doi.org/10.1093/bib/bbz020

Assembly graphs



Assembly graphs in the wild

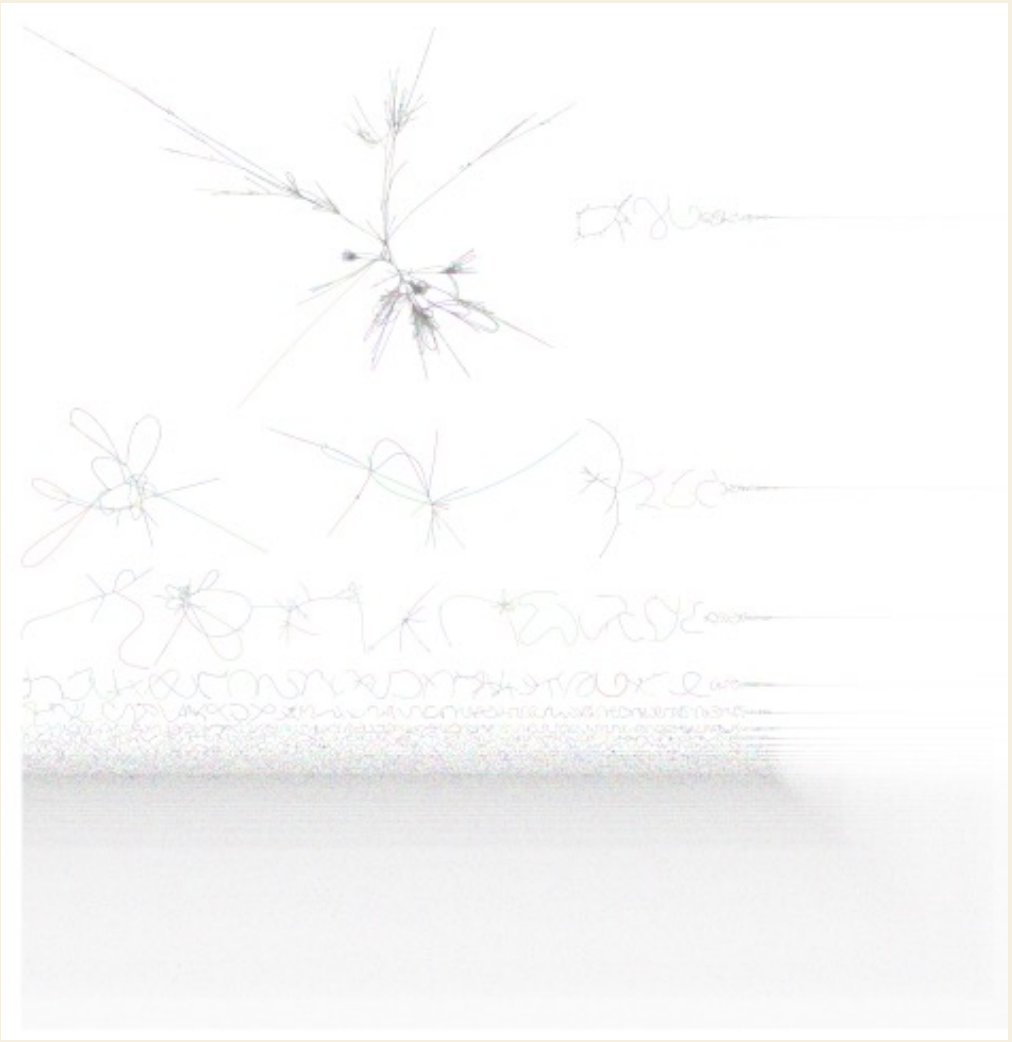
Mouse gut metagenome



@SilasKieser https://twitter.com/SilasKieser/status/1308752555795779585/photo/1

Assembly graphs in the wild

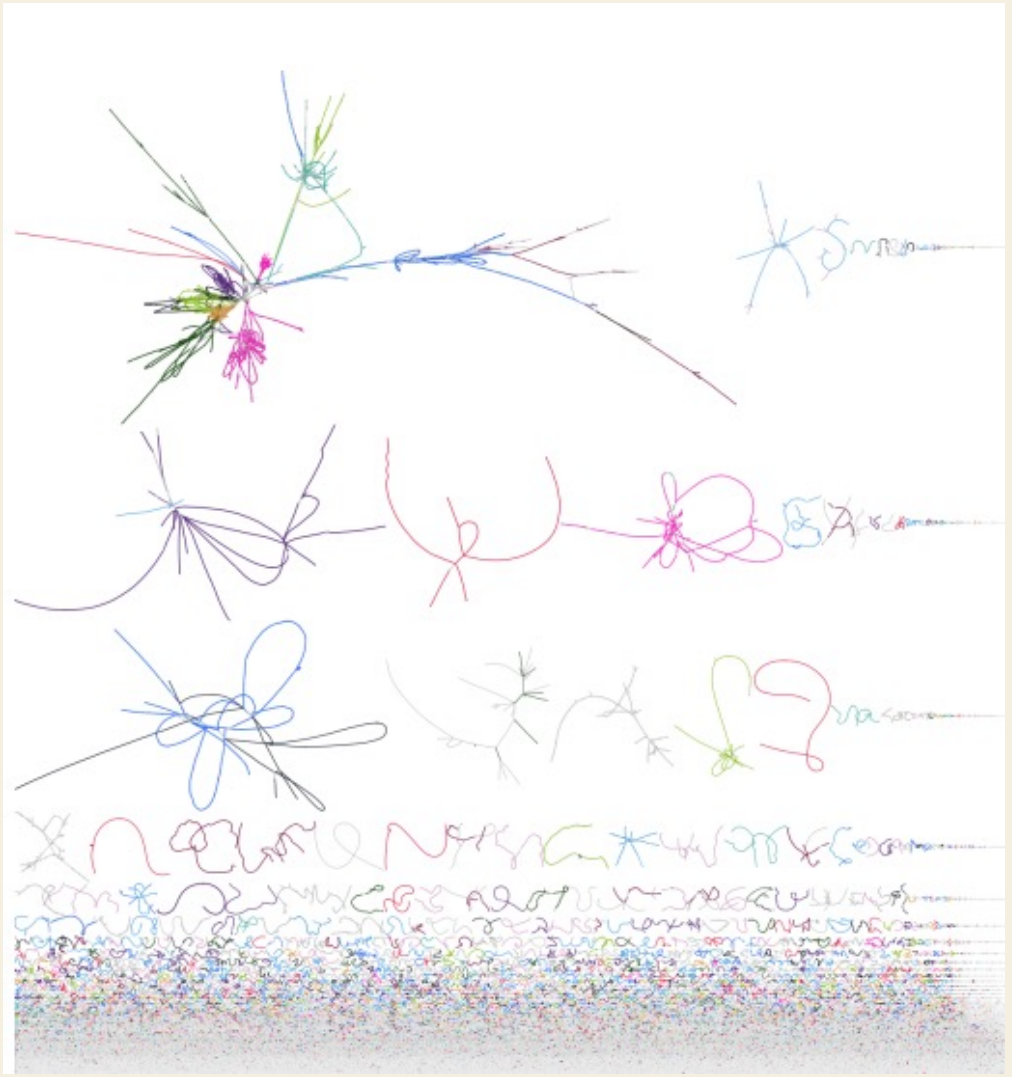
Percholate-reducing metagenome



https://tylerbarnum.com/2018/02/26/how-to-use-assembly-graphs-with-metagenomic-datasets/

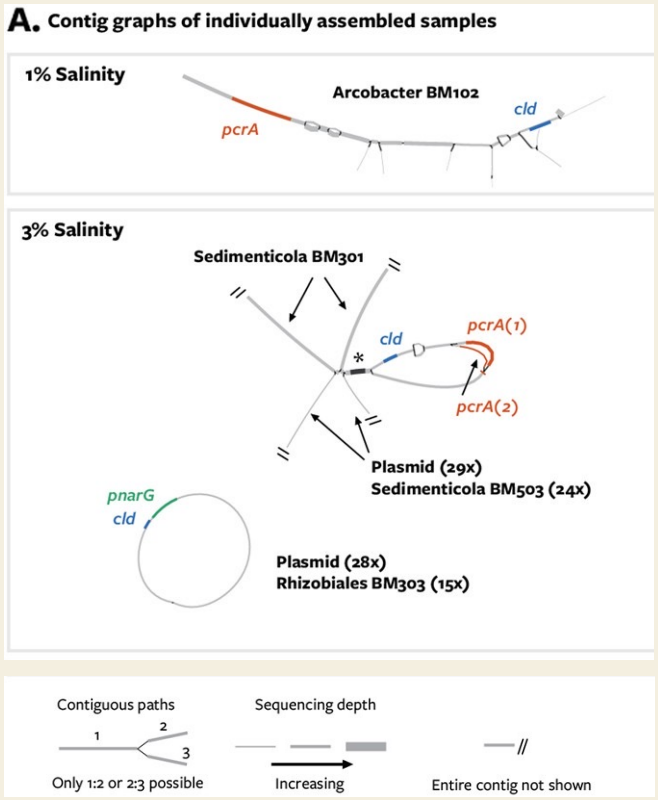
Assembly graphs organize sequences

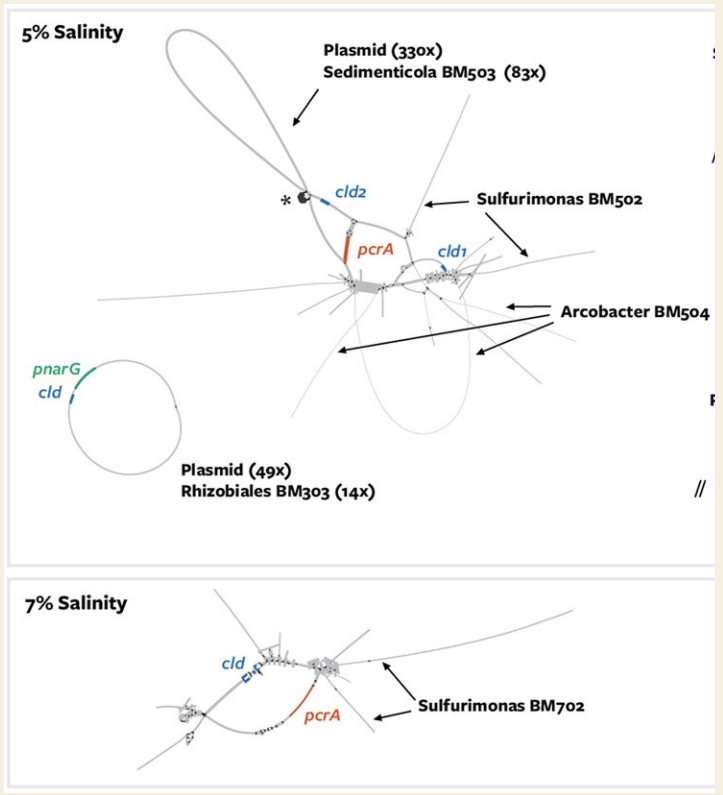
The combined assembly used in the publication Barnum et al. 2018 (*ISME Journal*). **Each of the 48 bins is one of 12 colors**, every color is used 4 times. **Unbinned sequences are in gray**; these bins were below 70% complete (as assessed by CheckM) and not included in the publication.



https://tylerbarnum.com/2018/02/26/how-to-use-assembly-graphs-with-metagenomic-datasets/

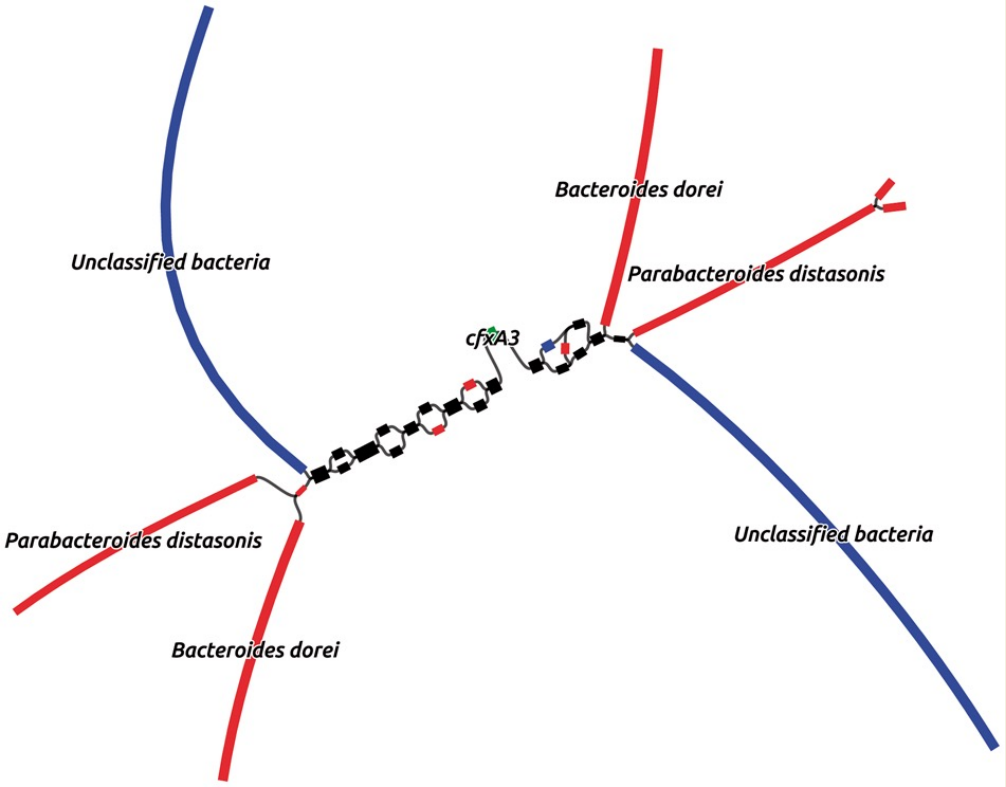
Assembly graphs organize sequences





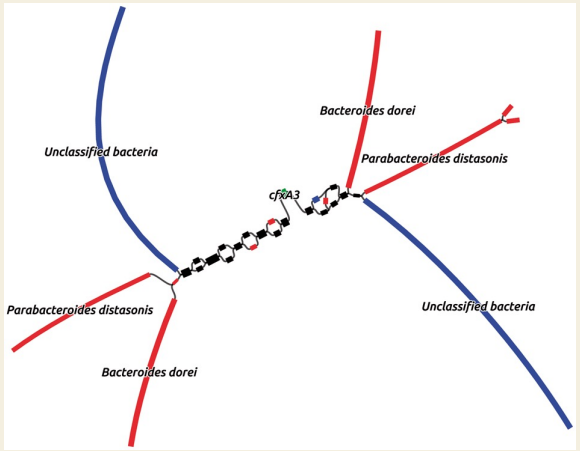
Barnum et al. 2018 https://doi.org/10.1038/s41396-018-0081-5

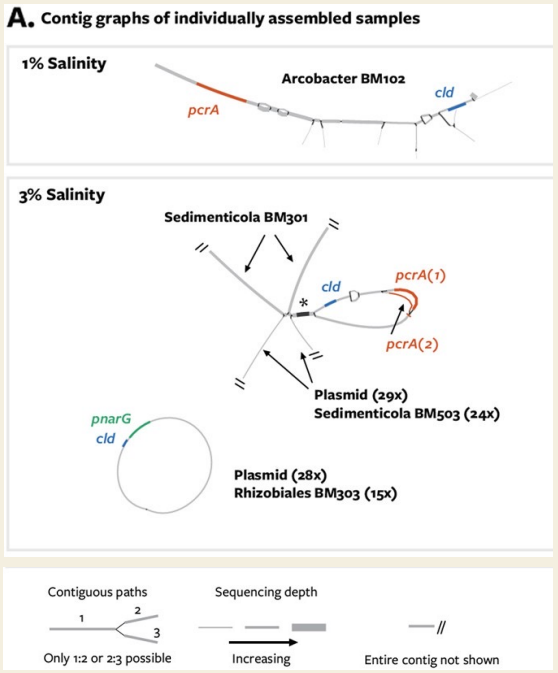
Assembly graphs organize sequences

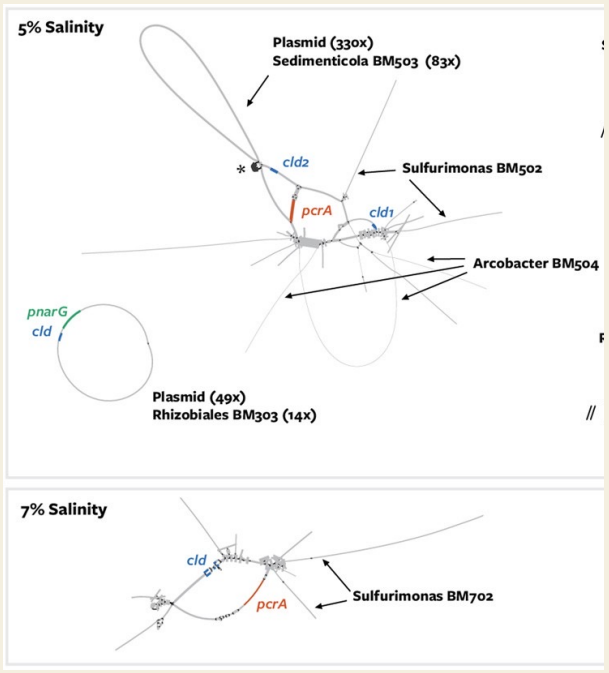


MetaCherchant tool. Olekhnovich et al. 2018 https://doi.org/10.1093/bioinformatics/btx681

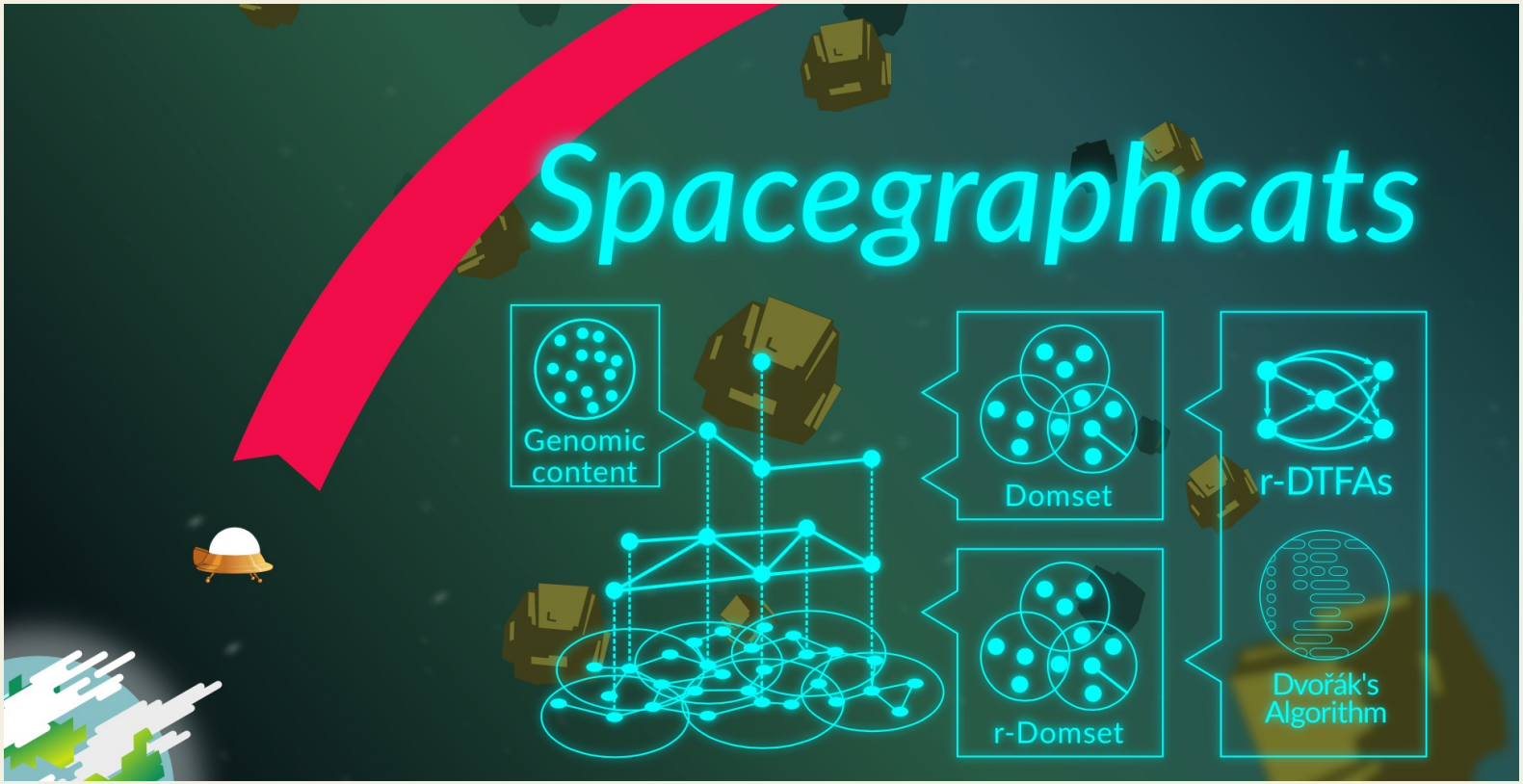
Querying assembly graphs for sequences of interest is computationally prohibitive at scale







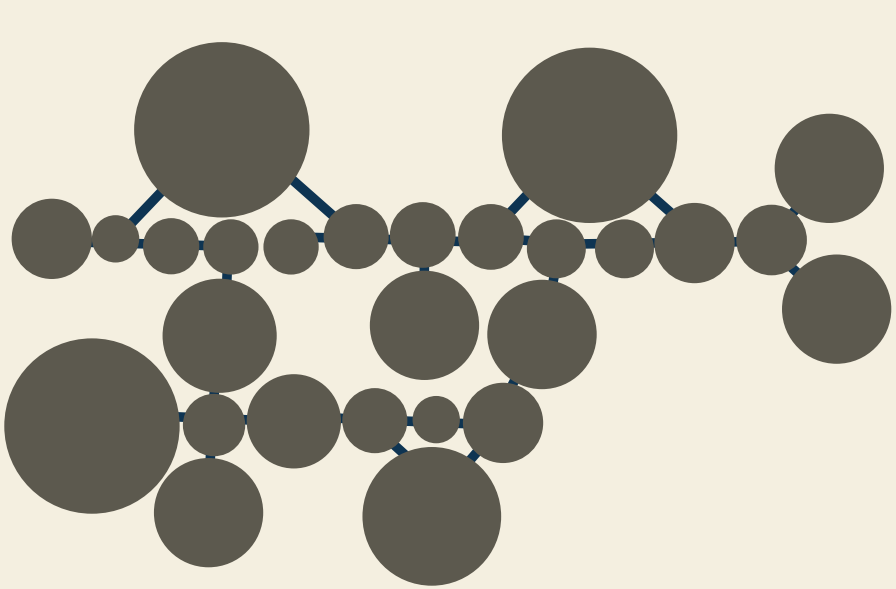
Querying assembly graphs for sequences of interest ~~is~~ was computationally prohibitive at scale

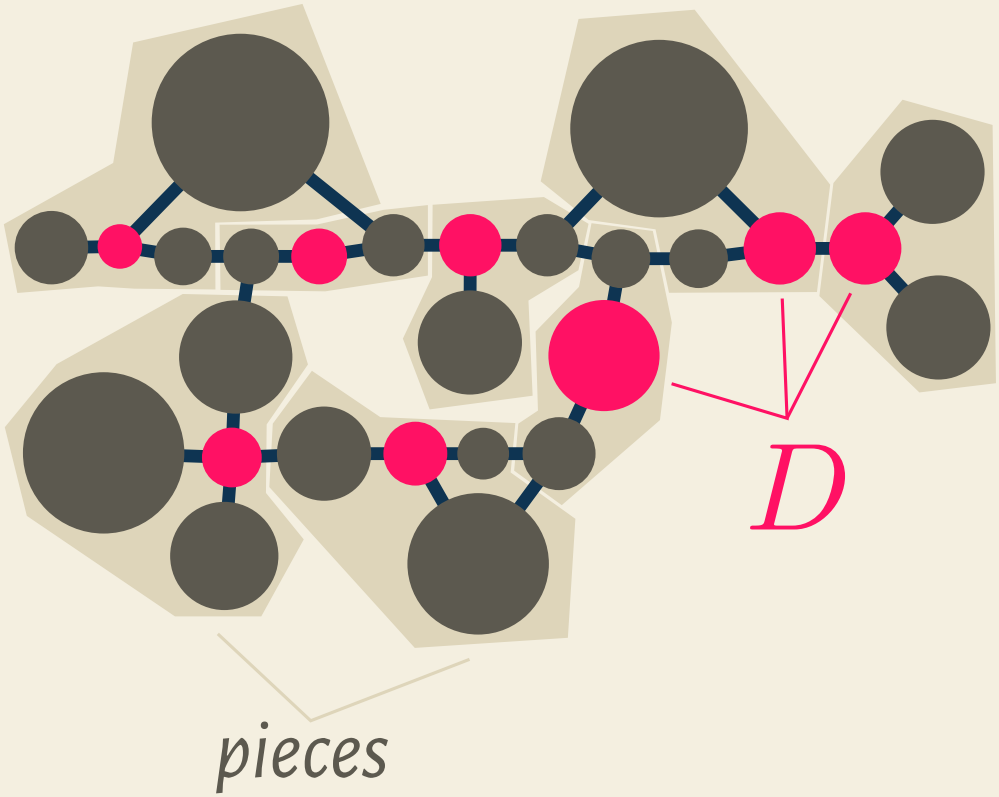


Outline

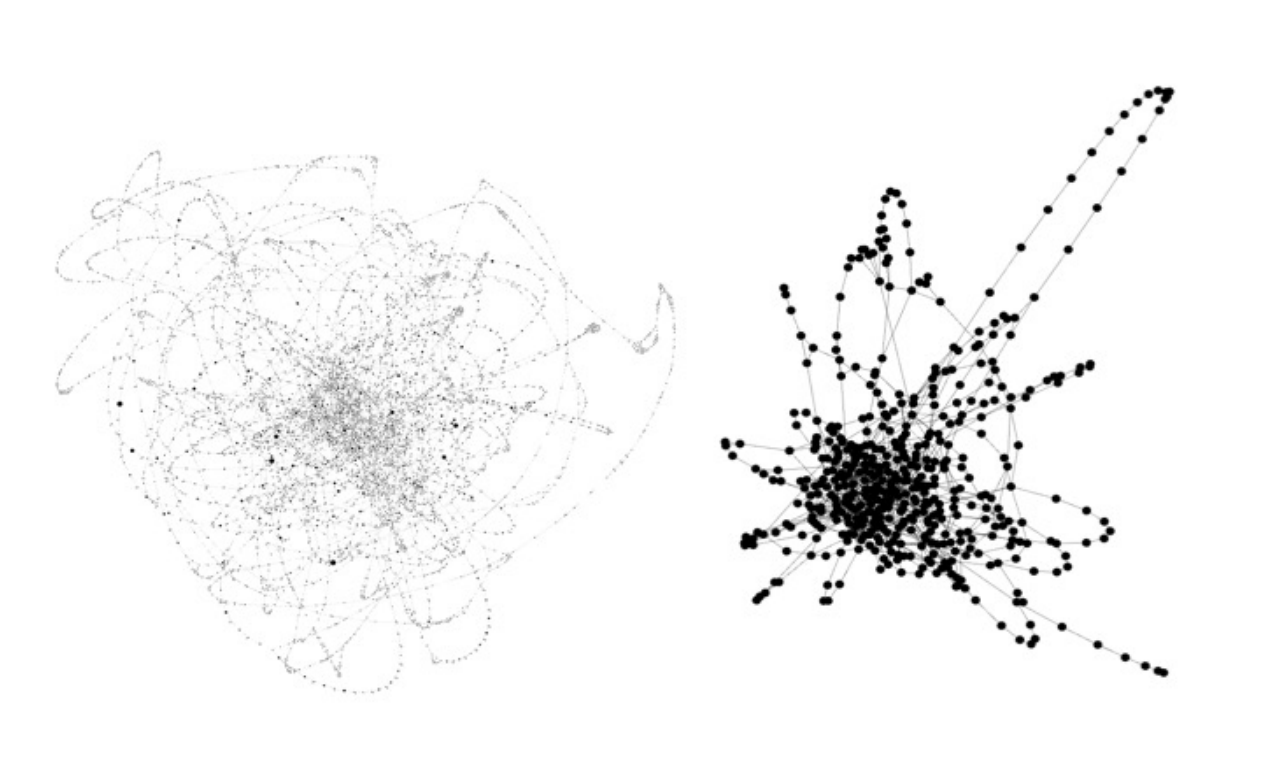
- Metagenome analysis methods and their problems
- Graphs and why they’re good
- An introduction to spacegraphcats

spacegraphcats 🥳





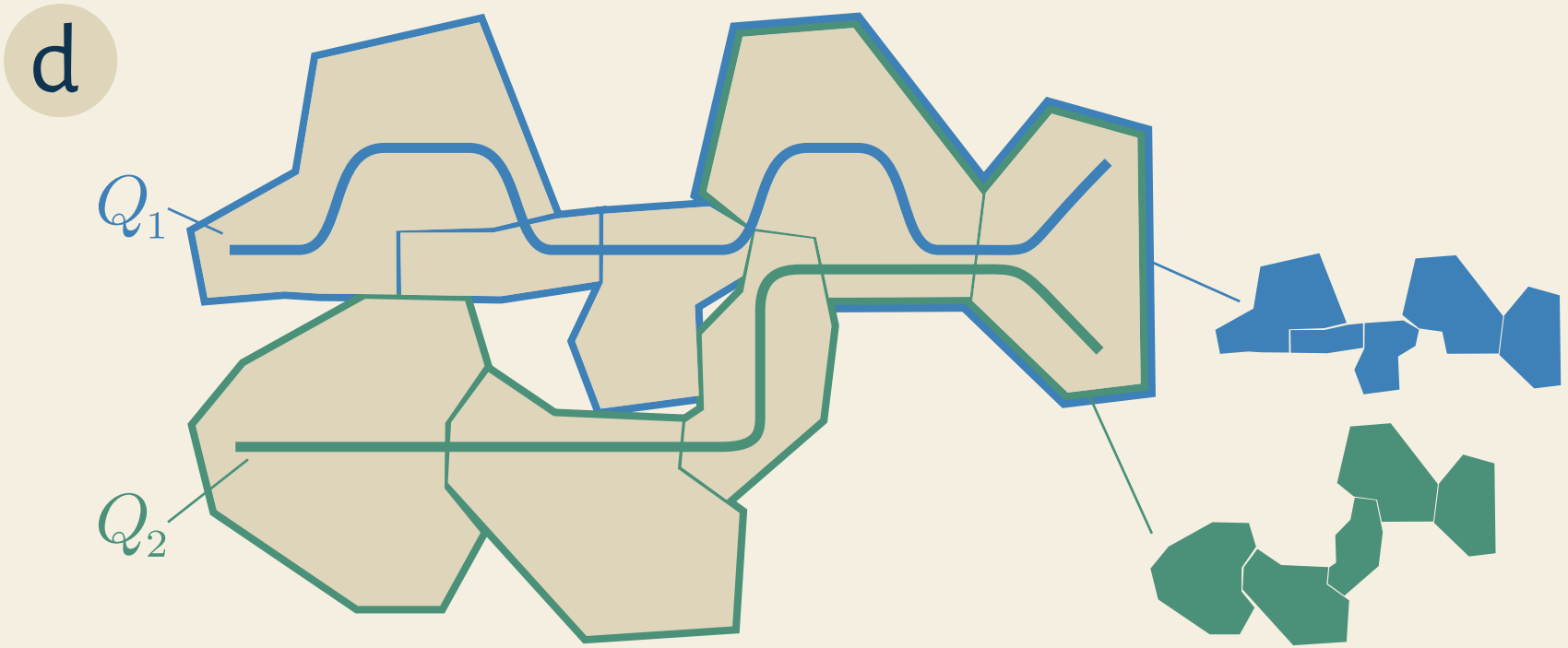
spacegraphcats 🥳



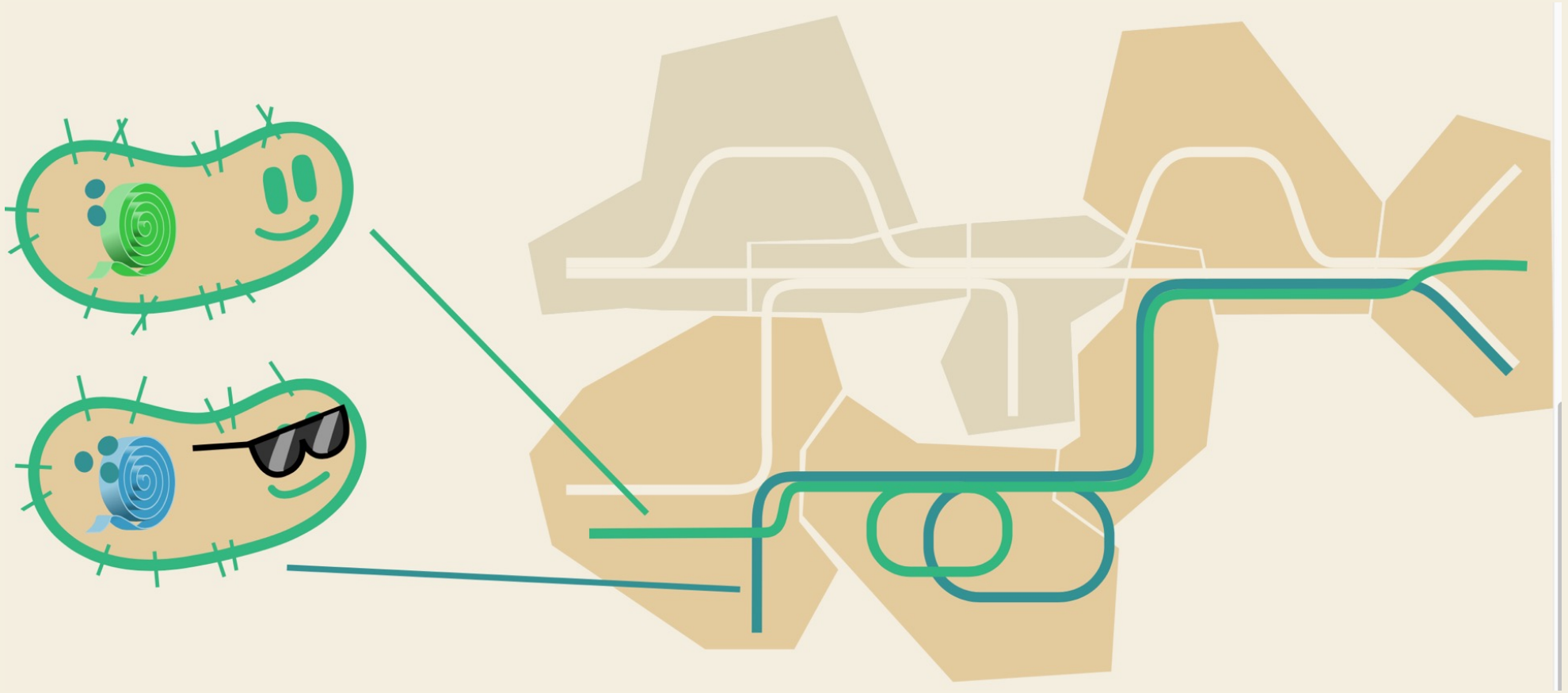
Escherichia coli cDBG

Escherichia coli simplified
by spacegraphcats

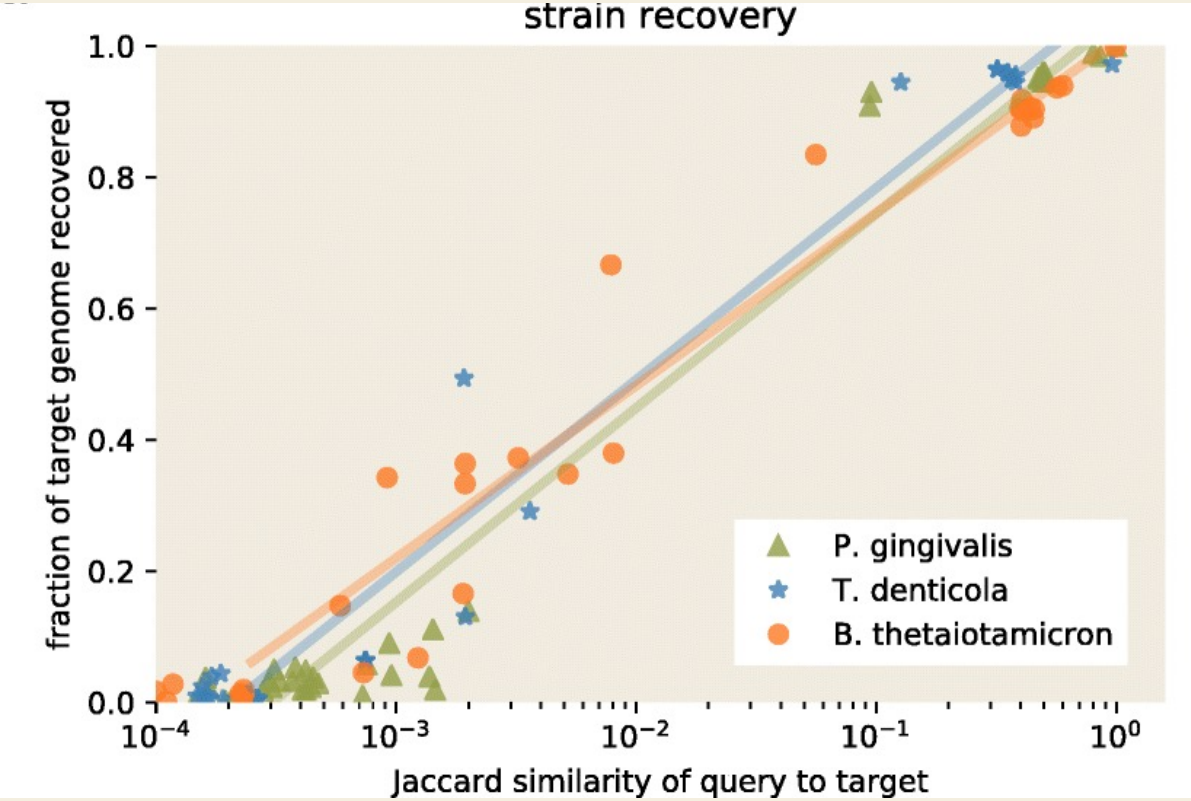
spacegraphcats 🥳



spacegraphcats 🥳

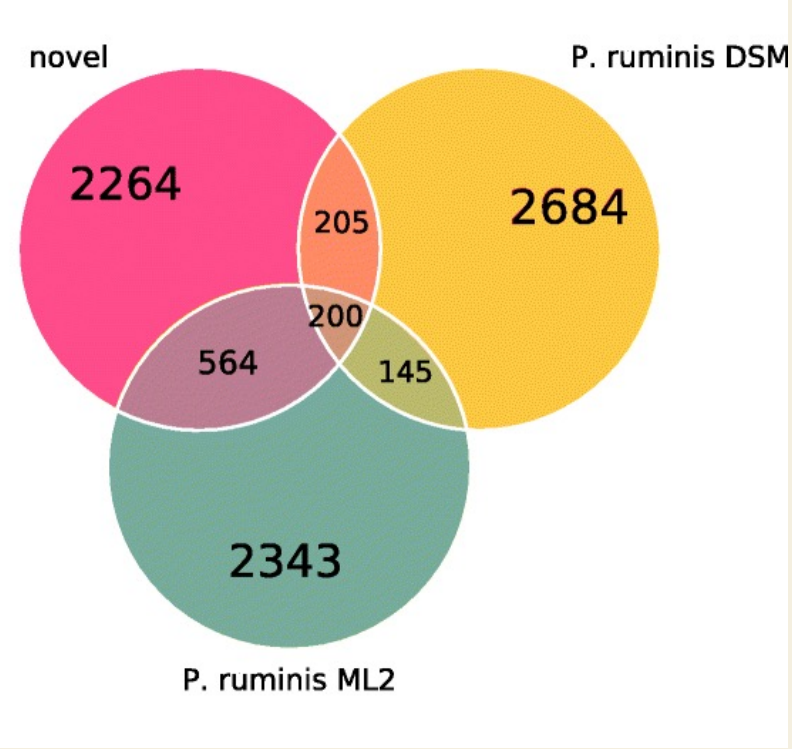


spacegraphcats recovers hidden sequence diversity



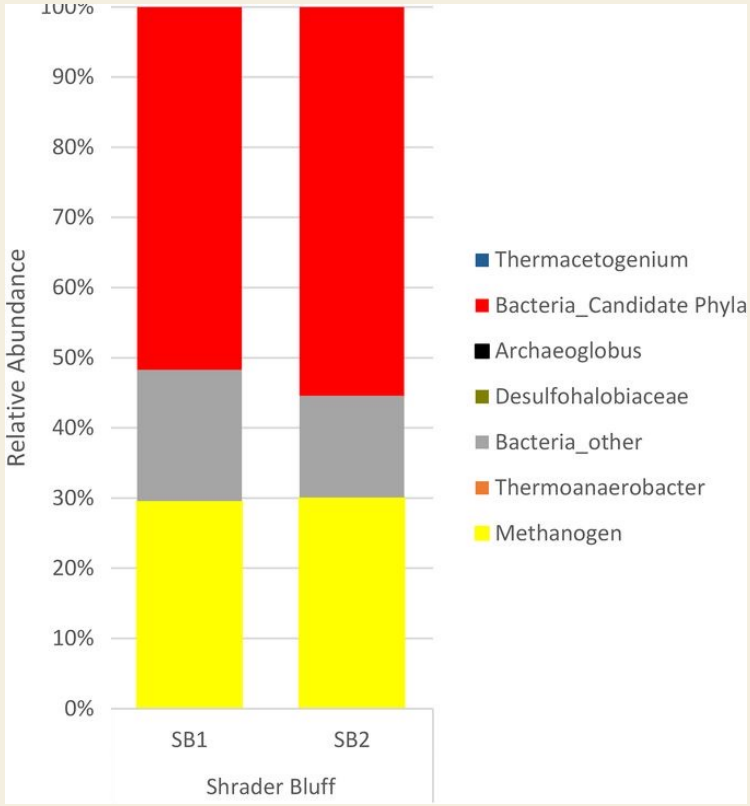
Brown et al. 2020 https://doi.org/10.1186/s13059-020-02066-4

spacegraphcats recovers hidden sequence diversity



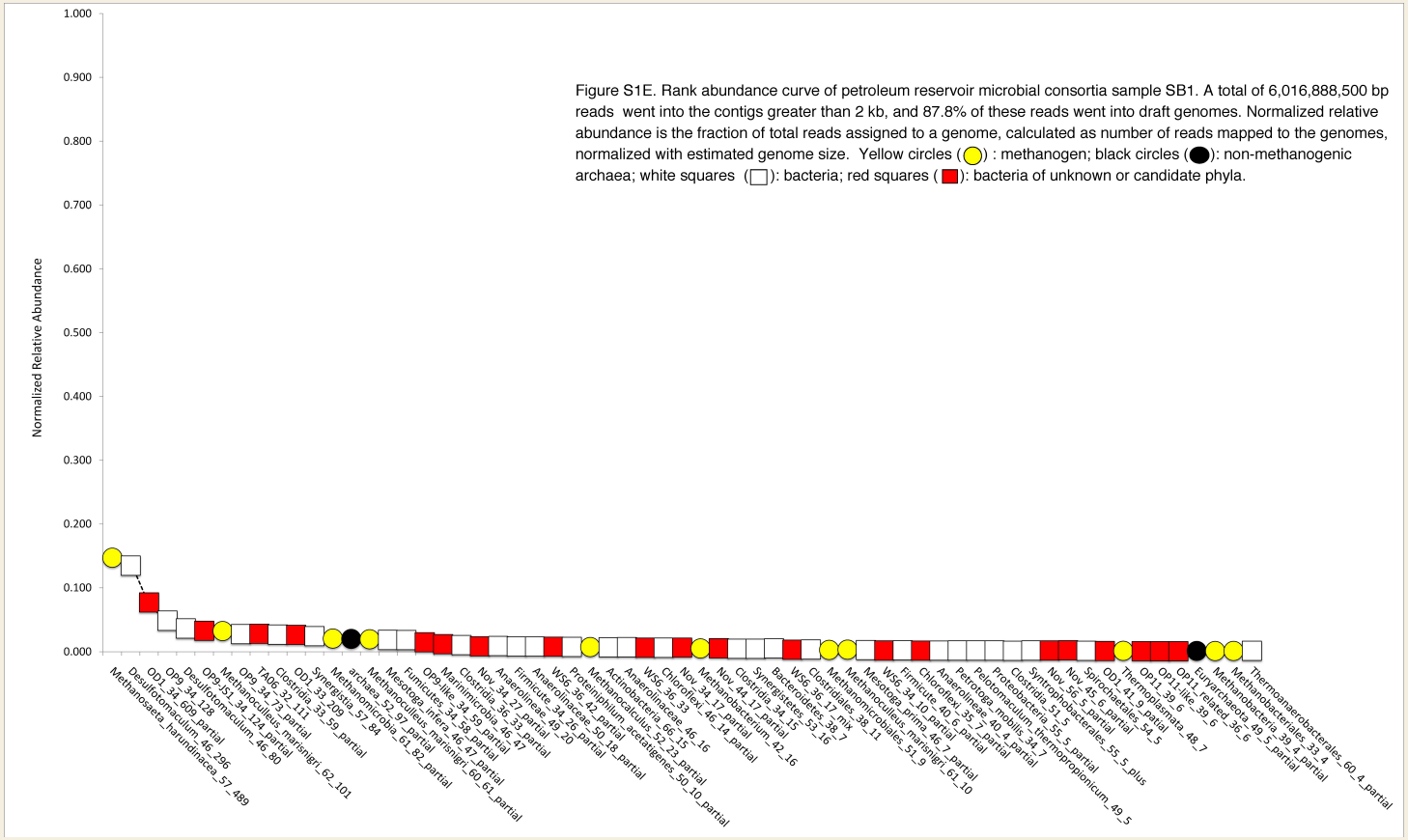
Brown et al. 2020 https://doi.org/10.1186/s13059-020-02066-4

spacegraphcats on a real metagenome



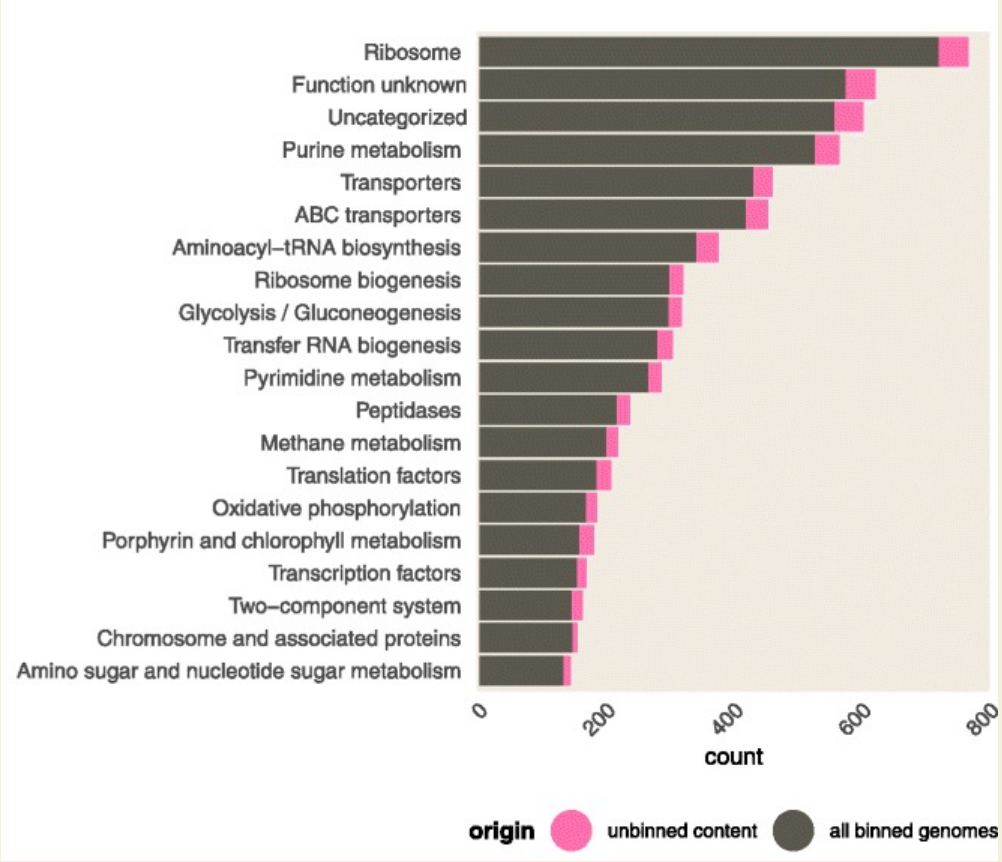
Hu et al. 2016 https://doi.org/10.1128/mBio.01669-15

spacegraphcats with *de novo* bins as queries



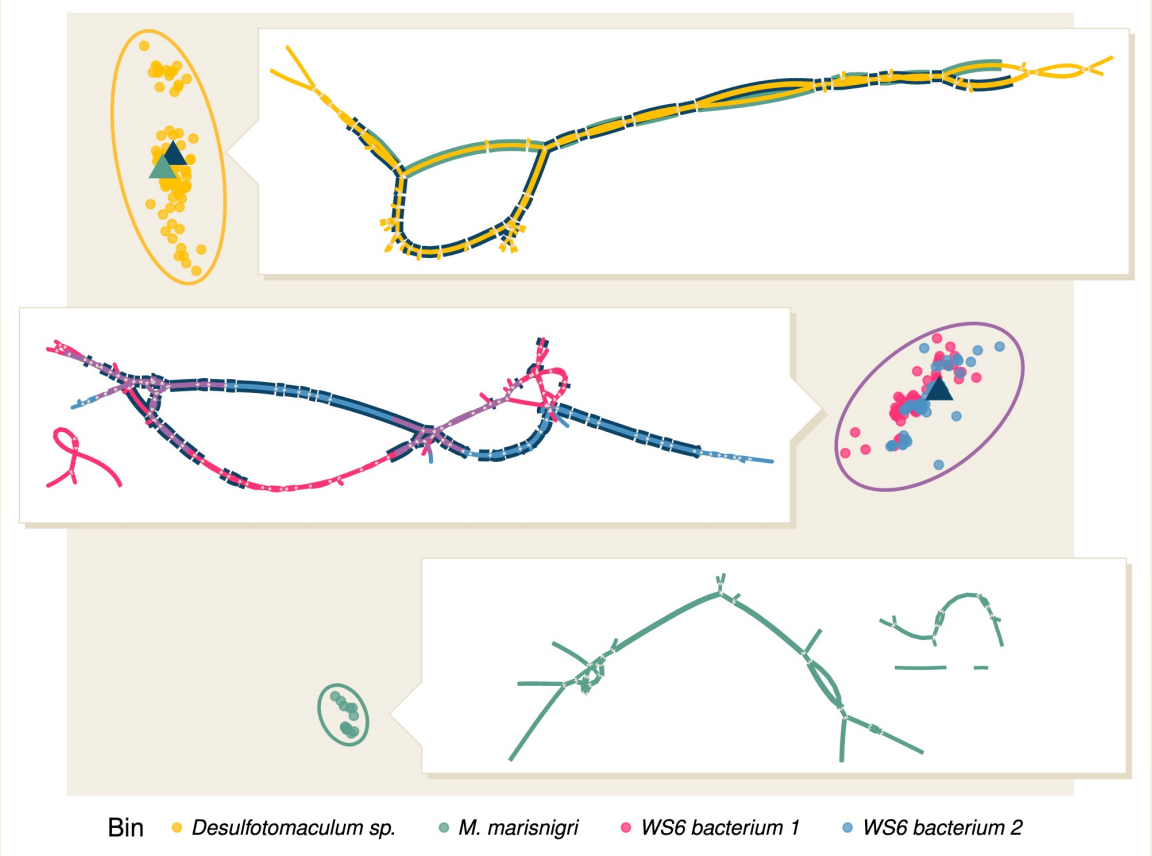
Hu et al. 2016 https://doi.org/10.1128/mBio.01669-15

spacegraphcats uncovers hidden functional potential



Brown et al. 2020 https://doi.org/10.1186/s13059-020-02066-4

spacegraphcats uncovers hidden sequence variation



Brown et al. 2020 https://doi.org/10.1186/s13059-020-02066-4

Use case: metagenome bin completion

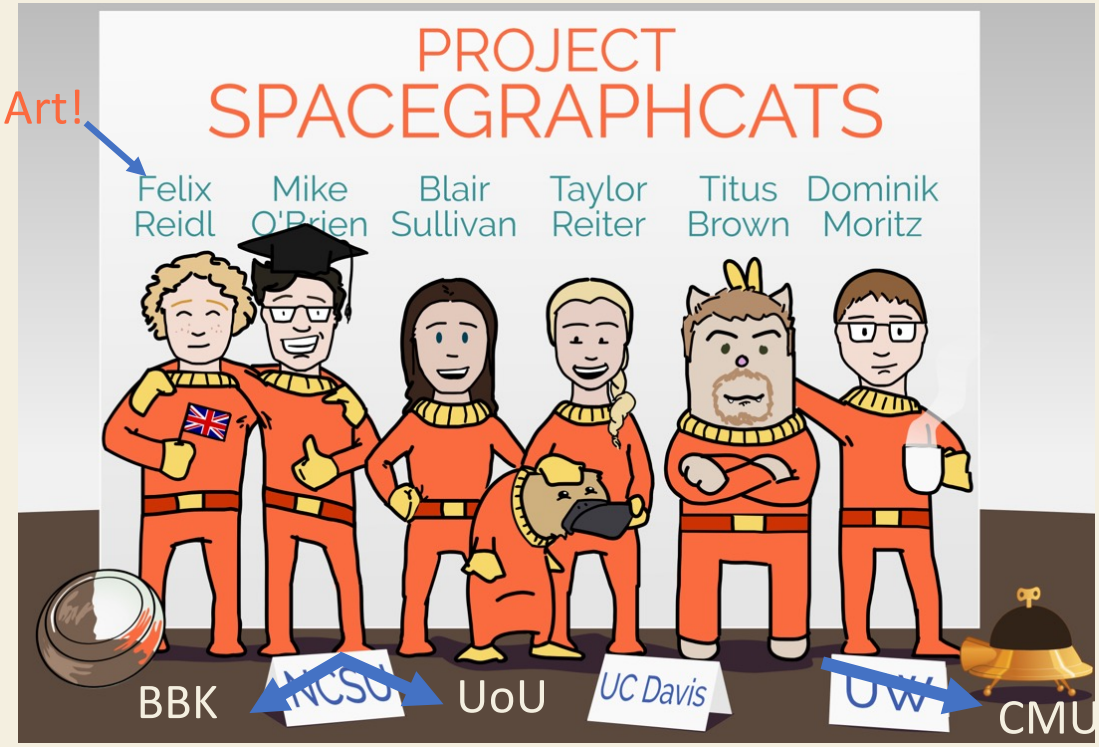
github.com/taylorreiter/2021-sgc-binder

https://binder.pangeo.io/v2/gh/taylorreiter/2021-sgc-binder/main

More use cases and documentation:

- https://spacegraphcats.github.io/spacegraphcats/

Thank you!



@ReiterTaylor

tereiter@ucdavis.edu

taylorreiter

Slides: github.com/taylorreiter/2021-sgc-binder

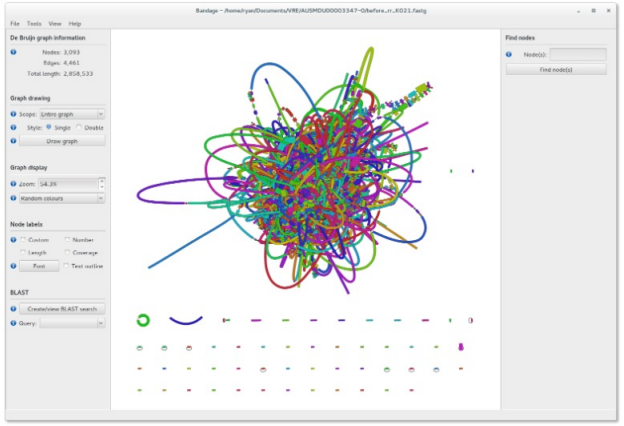
github.com/spacegraphcats/spacegraphcats

https://spacegraphcats.github.io/spacegraphcats

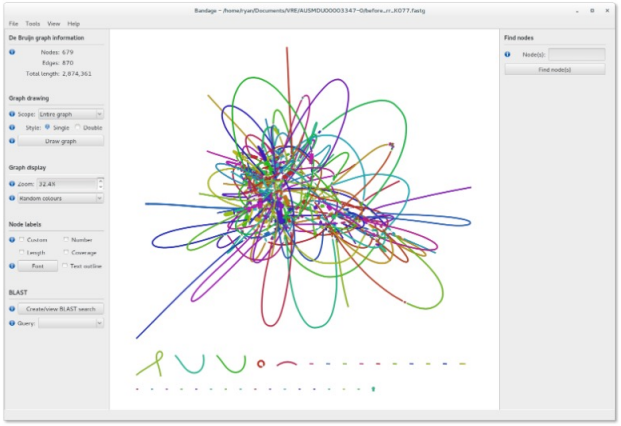
Assembly graphs are parameter-sensitive

5 SCREENSHOTS

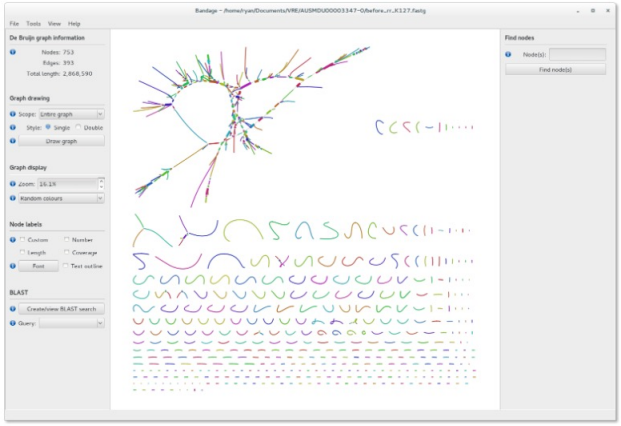
These graphs are the same bacterial isolate assembled in SPAdes using three different k-mer values:



K-mer of 21. This value is too small, resulting in short contigs and many connections, giving a dense tangled graph.



K-mer of 77. This is a good balance, giving a smaller number of long contigs that are well connected.



K-mer of 127. This value is too large, resulting in the graph breaking into many discontinuous pieces.

Wick et al. 2015 doi: 10.1093/bioinformatics/btv383