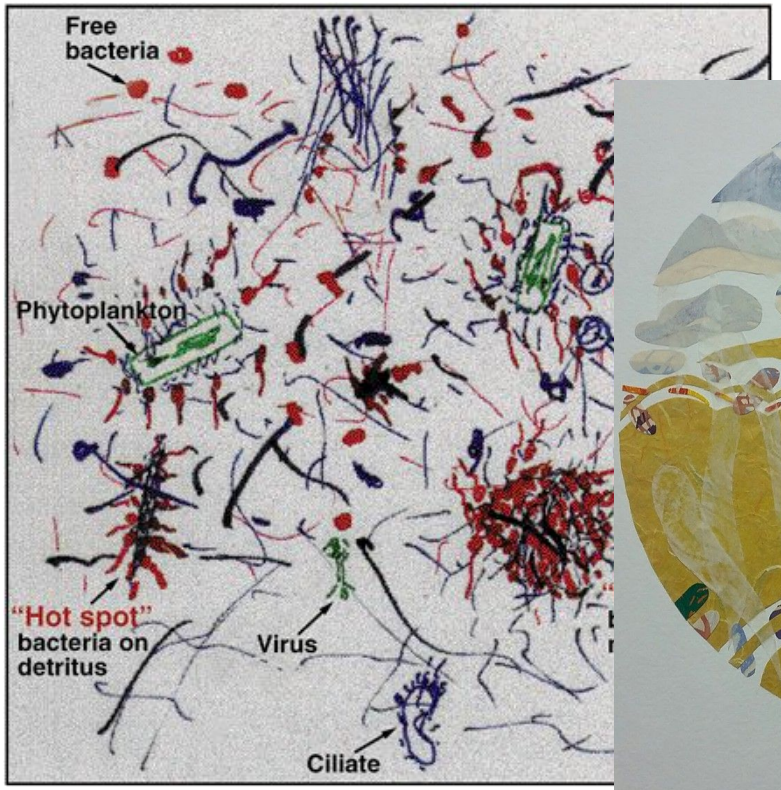
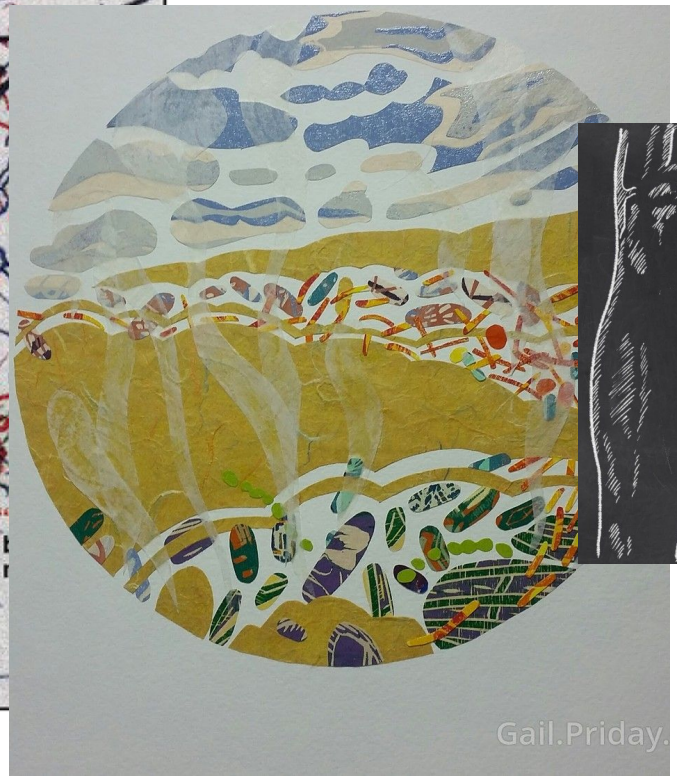
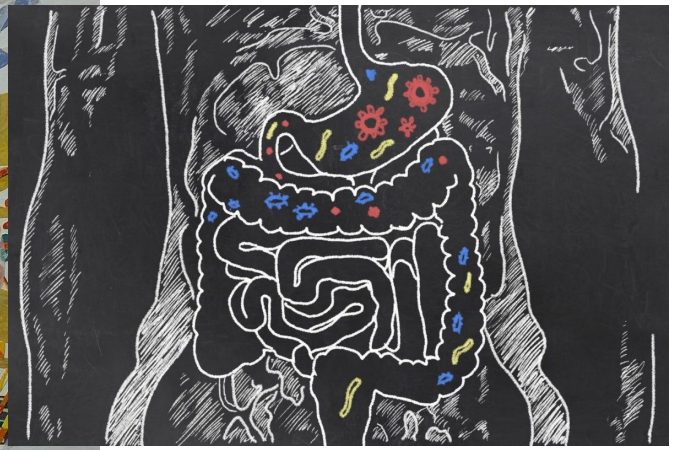
Metagenomics Lesson 5: Sequence Assembly



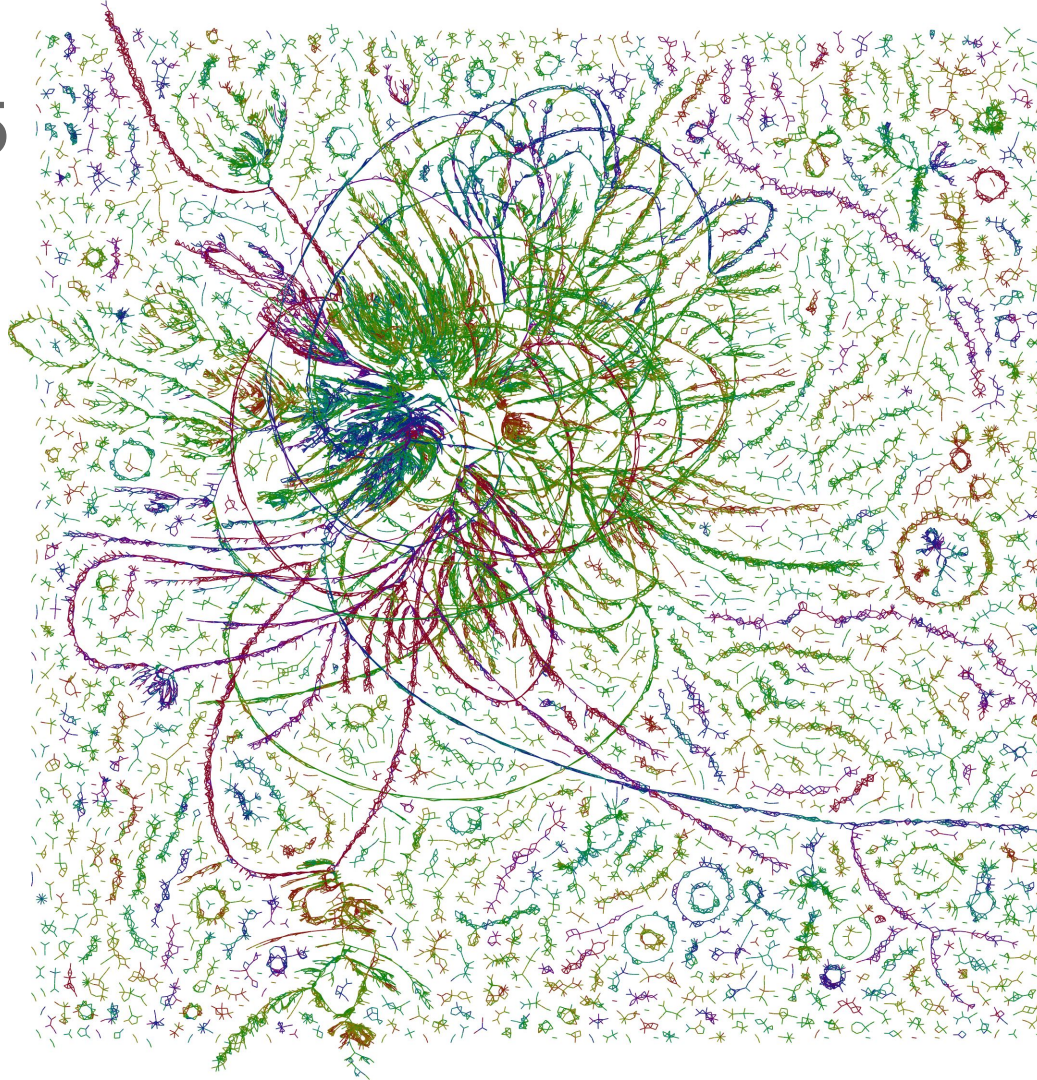




Metagenomics Lesson 5

Sequence Assembly

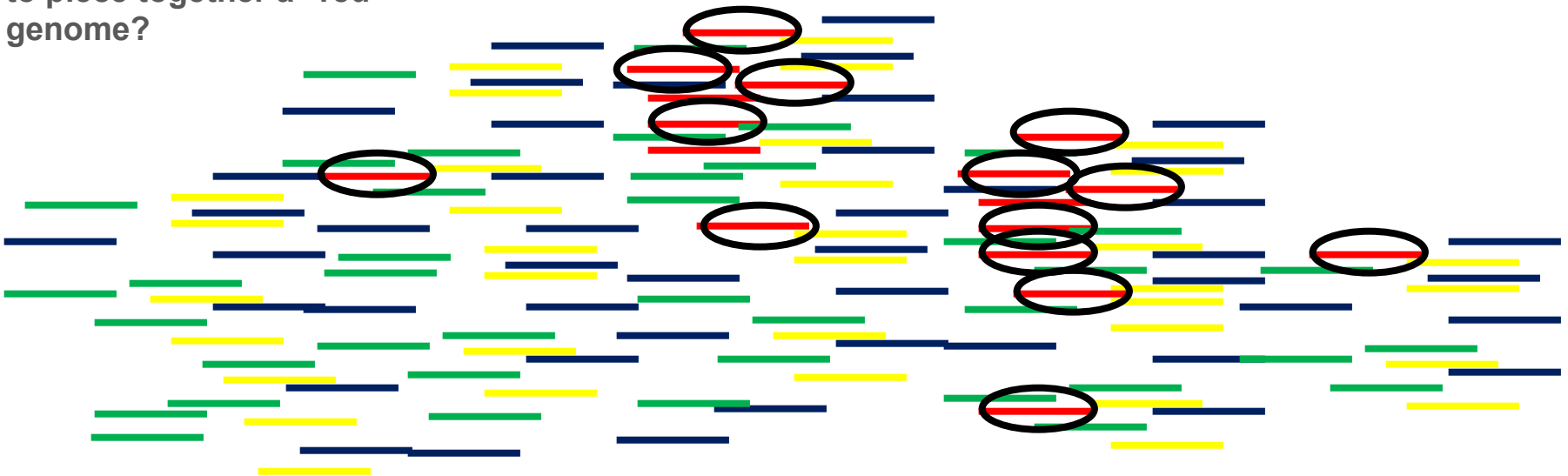
1. Sequencing Diverse Samples
2. Sequencing Statistics
3. Assembly Paradigms
    a. Overlap - Layout - Consensus
    b. de Bruijn Graph
    c. Long-read
    d. Hybrid
4. Assembly Software



Working with Short-read Metagenomic Data

Different colors correspond to different species:

Do we have enough coverage to piece together a “red” genome?



Coverage depends on genome relative abundance

4 dominant species

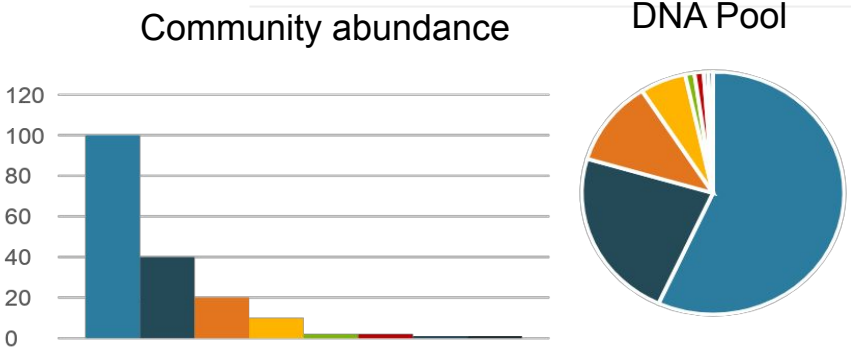
Abundant species >> moderately abundant species > rare species

- **10X coverage of species 1**
- **0.5X coverage of species 4**

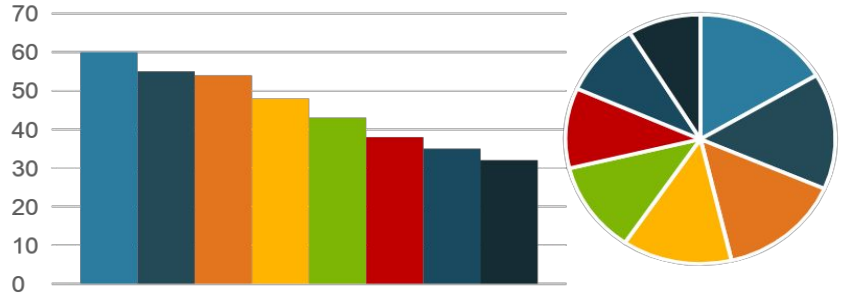
Abundant groups: Draft Genome Assembly Possible
Rare groups: Only Read-Based Analysis Possible

Effects of Evenness on Coverage

Sample 1:
Less Even = Lower sequencing demand for abundant groups



Sample 2: More Even = Higher sequencing demand for abundant groups



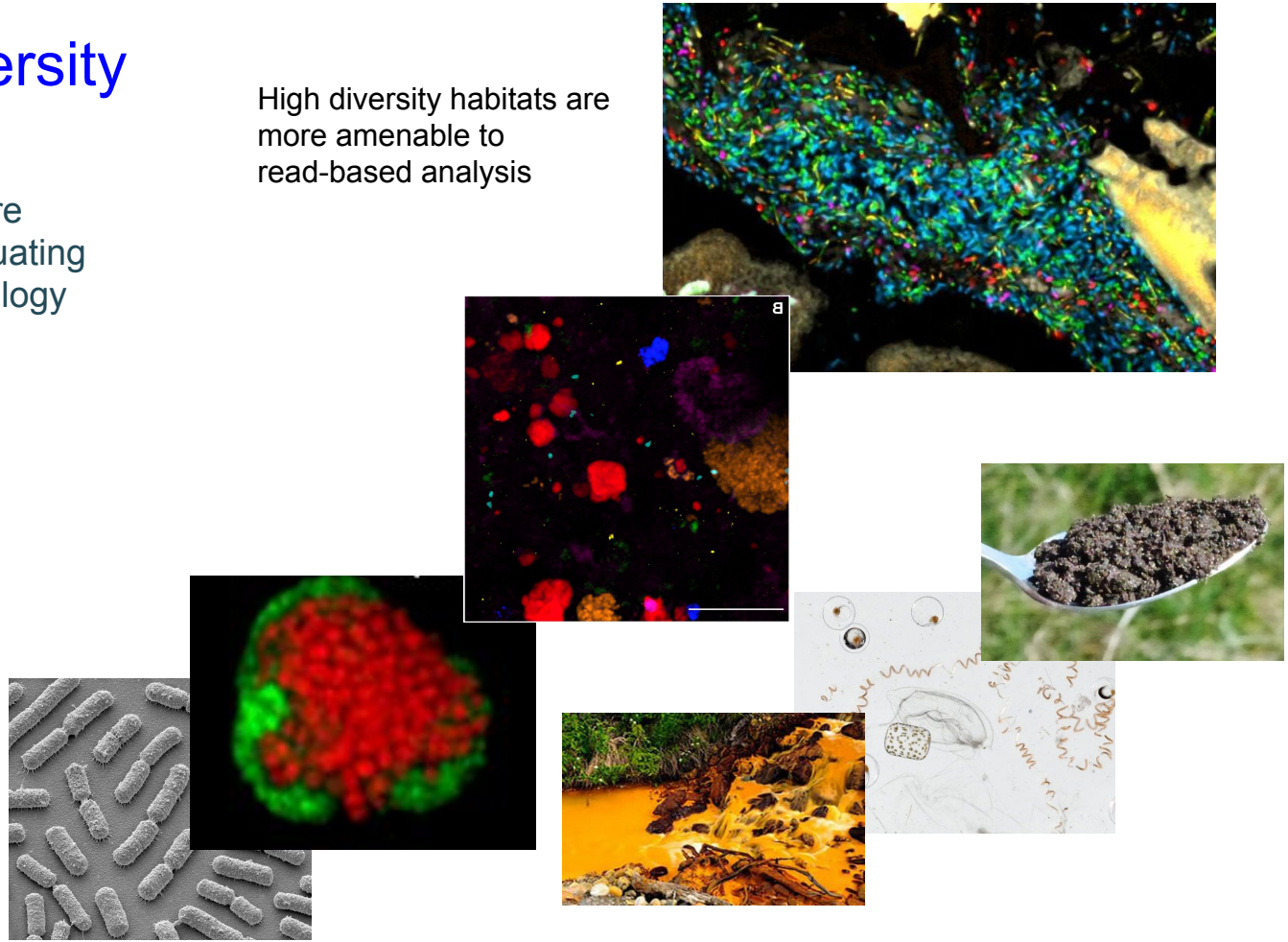
Determining the Rank-Abundance Relationship for a Community:

Before sequencing one can estimate community richness/evenness with SSU libraries, FISH, or other techniques

Gradients in diversity

Different Environments Require Different Approaches for Evaluating Microbial Diversity and Physiology

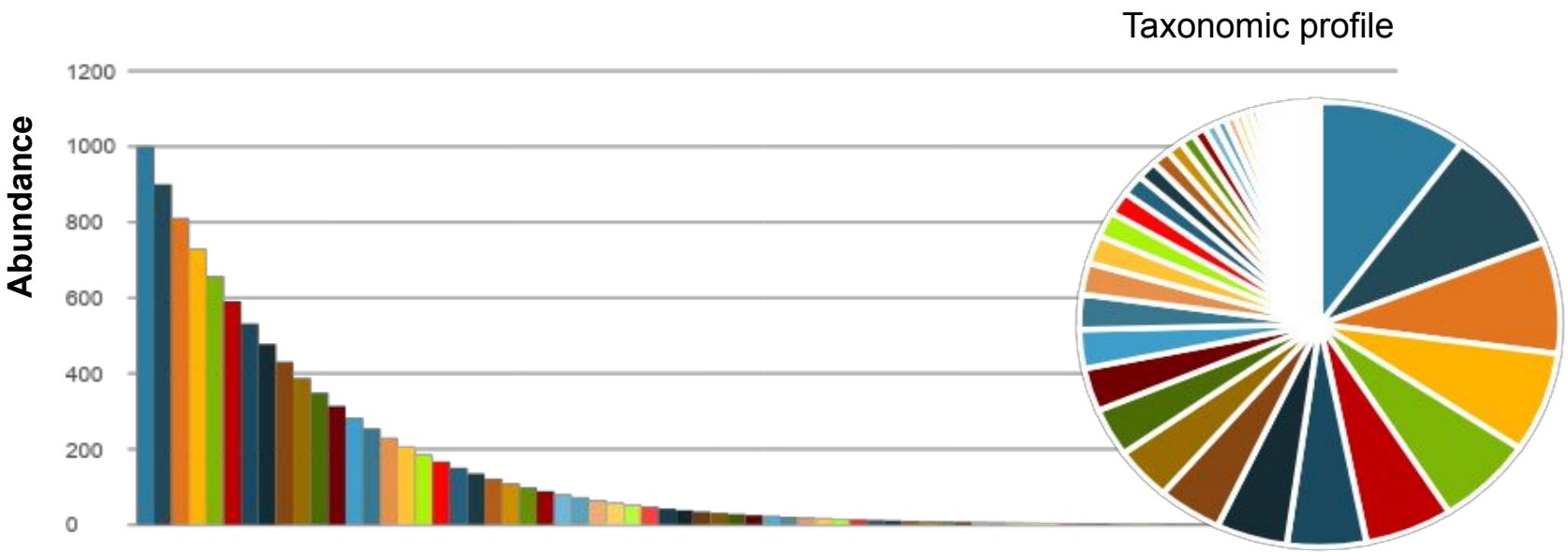
High diversity habitats are more amenable to read-based analysis



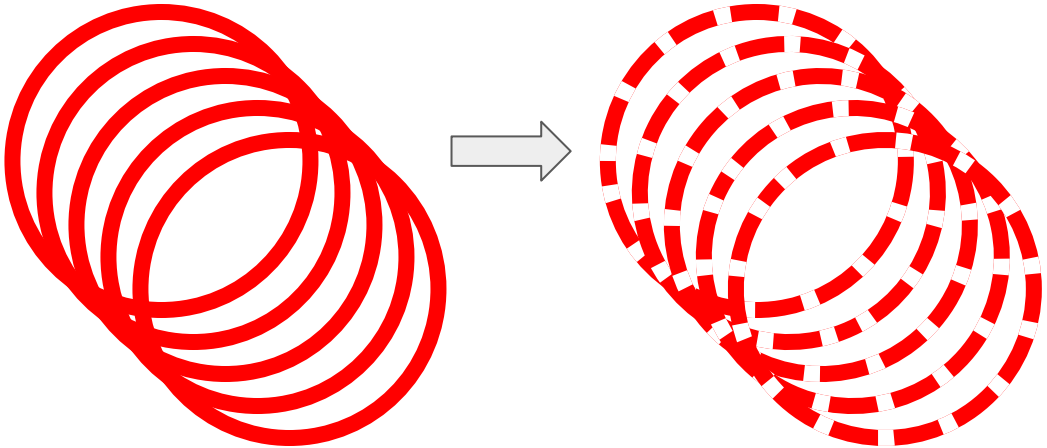
Low diversity habitats are more amenable to assembly based analysis

High-Complexity Environments

Targeting low-abundance populations becomes very difficult without some kind of enrichment (single cell, cell sorting, etc.)



Sequencing Statistics: What is Coverage?



- Genomes are randomly* sheared during DNA extraction and library prep
- During sequencing, a random subset of the fragments are sampled

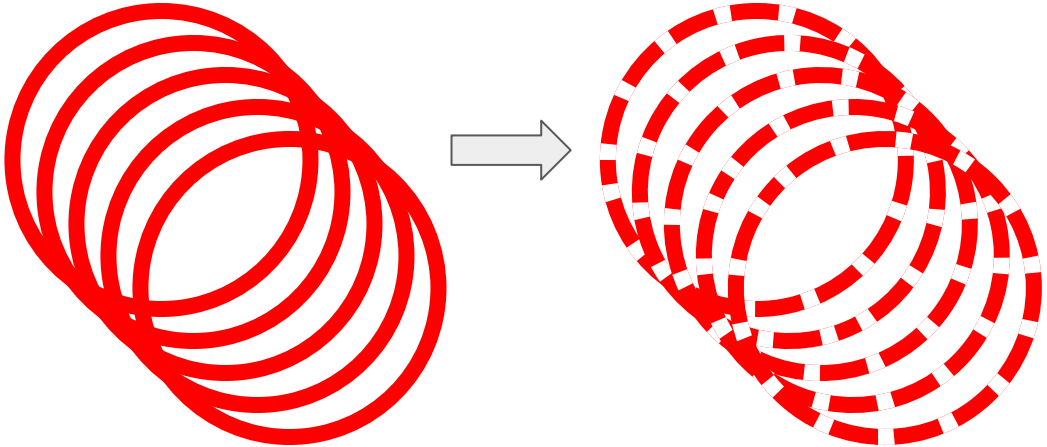
Q: What is the average read coverage?

Coverage C = LN / G
L is the read length
N is the number of reads
G is the haploid genome length

Example: You sequence *E. coli* using a whole MiSeq 2x250bp run

C = 250 * 8M / 4M =
500x coverage ***on average***

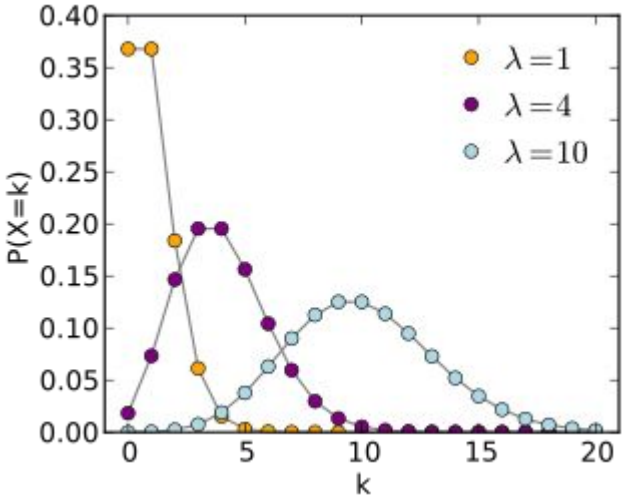
Sequencing Statistics: The Poisson Distribution



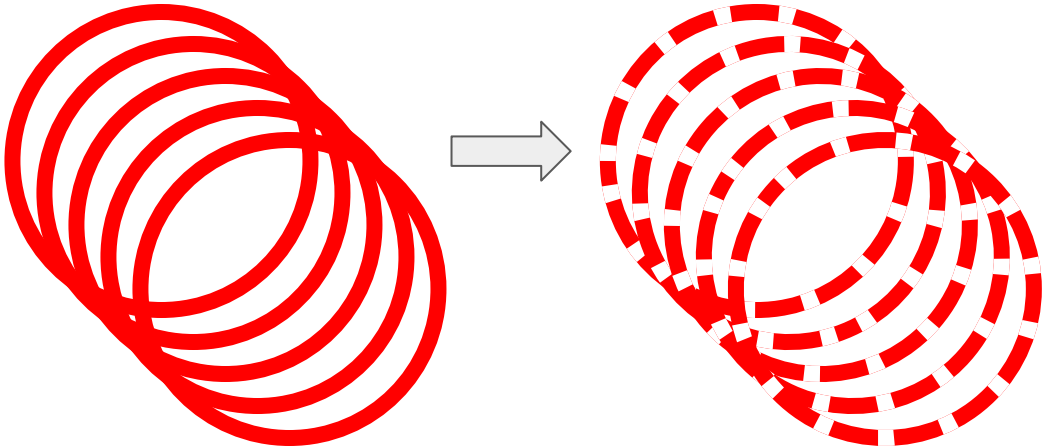
$$\Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- *k* = times base is seen, ƛ = average coverage
- Pr(X=0) = 4^0 * exp(-4)/0! = 0.0183 = 1.83%

- Q: What is the probability that a base will be sequenced exactly 0 times given a coverage of 4?
- A: Ask Poisson!



Sequencing Statistics: Coverage Goals



$$\Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

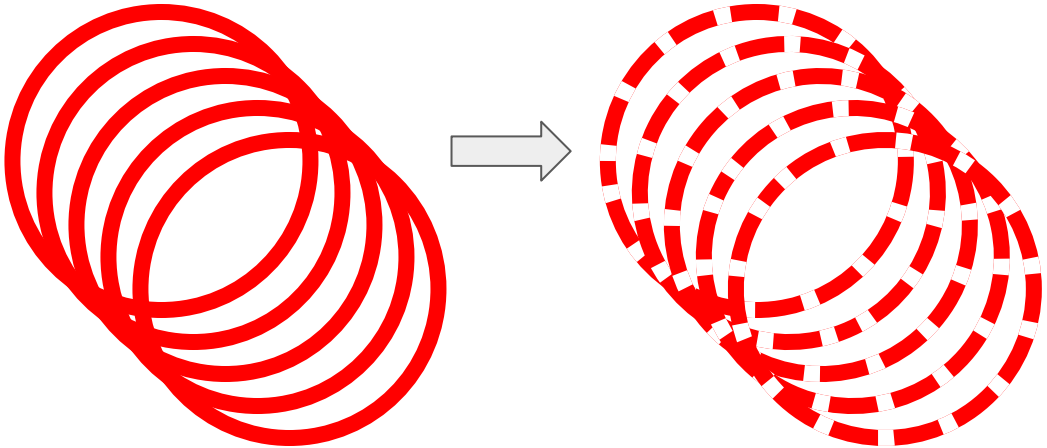
- *k* = times base is seen, ƛ = average coverage

- Q: What is the probability that a base will be sequenced 2 times or less given a coverage of 4?
- A: Pr(X=0) + Pr(X=1) + Pr(X=2)

  = 0.0183 + 0.0732 + 0.1465 = 0.2380 = 23.8%

We'll want to do better than this!

Sequencing Statistics: No Base Left Behind



$$\Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- *k* = times base is seen, ƛ = average coverage

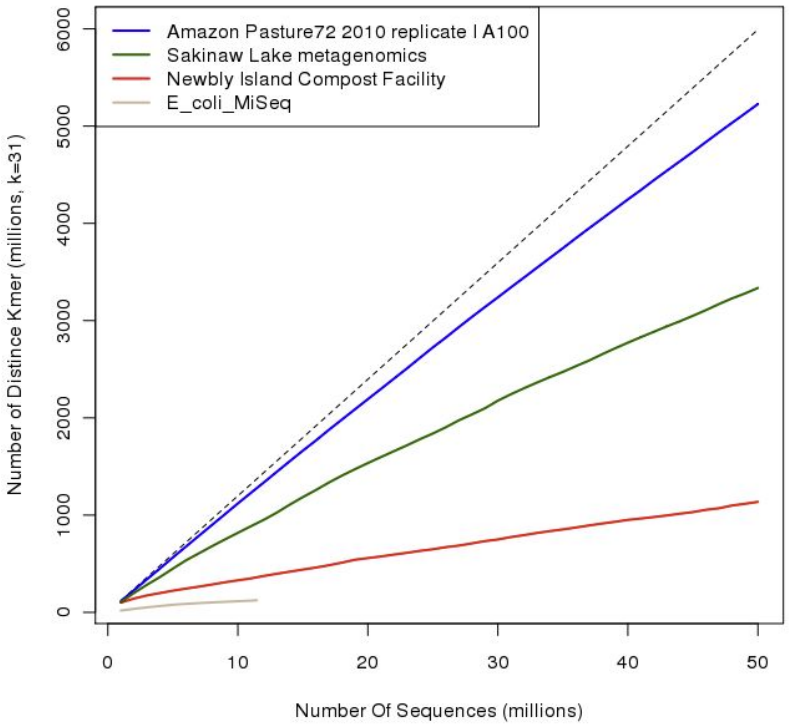
- Q: What is the minimum coverage of *E. coli* needed to ensure (statistically) that no bases are left unsequenced?

- A: Pr(X < 1/4,000,000)

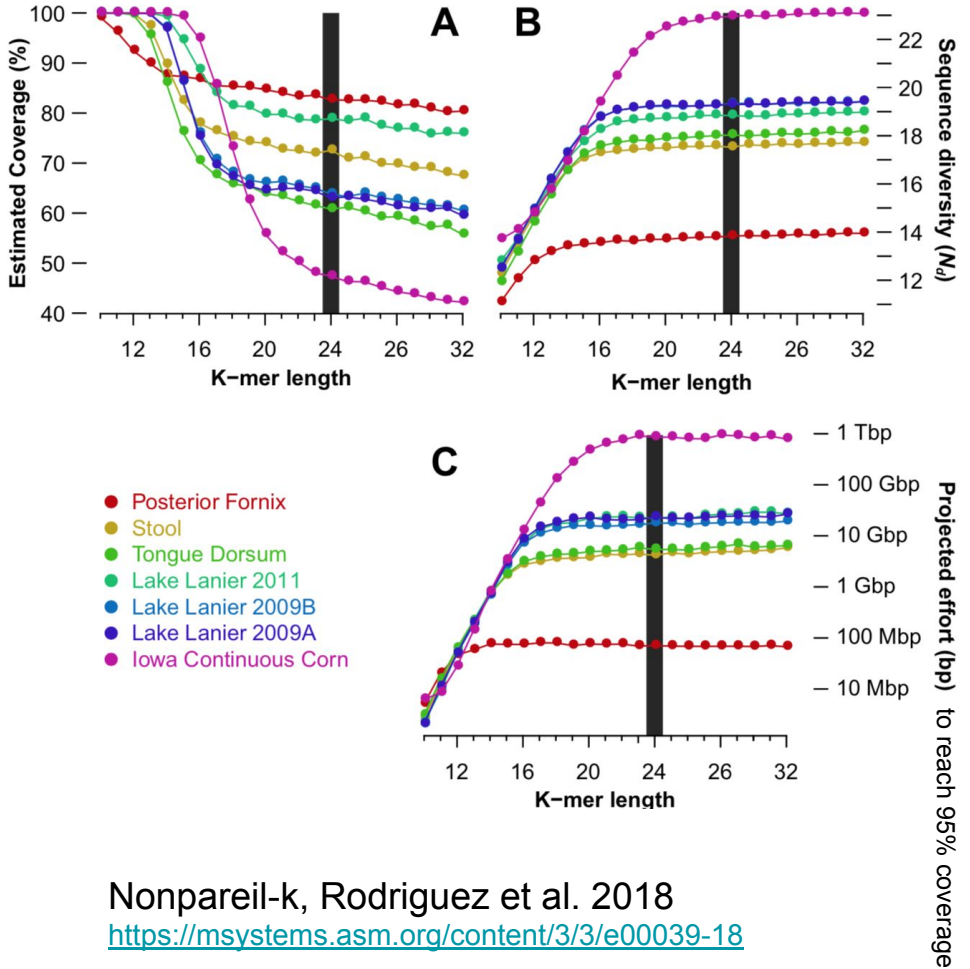
  1/4E6 = C^0 * exp(-C)/0!

  1/4E6 = 1 * exp(-C)/1
  C = -ln(1/4E6)
  C > 15

Metagenome Coverage: There's always more k-mers?

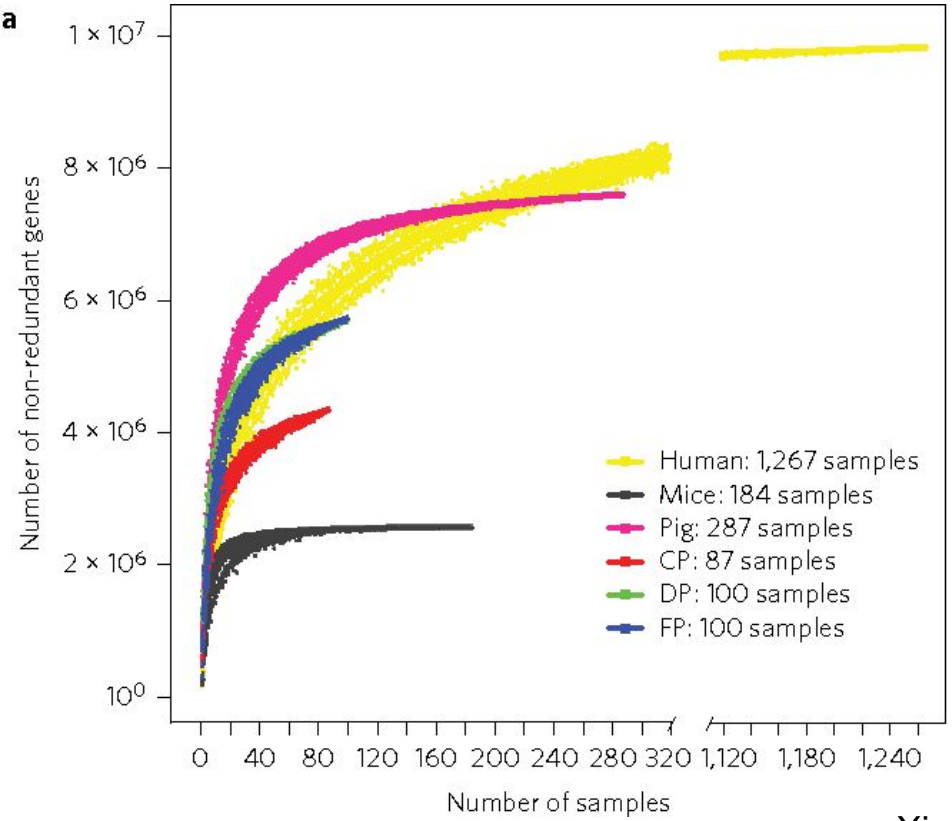


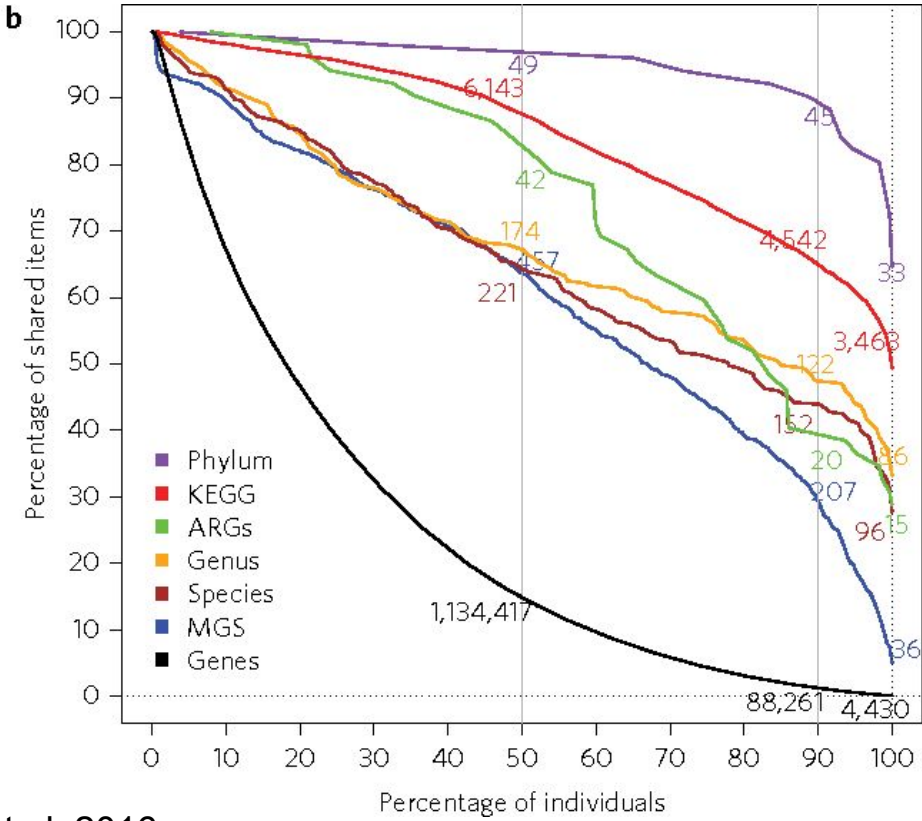
https://ngscc.blogspot.com/2012/06/kmer-rarefaction.html



Nonpareil-k, Rodriguez et al. 2018
https://msystems.asm.org/content/3/3/e00039-18

Metagenome Coverage: Always more genes and taxa?





Xiao et al. 2016 https://www.nature.com/articles/nmicrobiol2016161

Sequence Assembly Methods: Software

OLC: Overlap-Layout-Consensus

Originally designed for high quality long reads from Sanger sequencing

Used mostly for eukaryotes (?)

- MIRA
- Celera
- MaSuRCA
- Newbler
- Allora
- Arachne
- AMOS
- Meraculous
- Opera

de Bruijn Graph

Optimized for high-ish quality short reads from Illumina sequencing

Used for prokaryotes and metagenomes

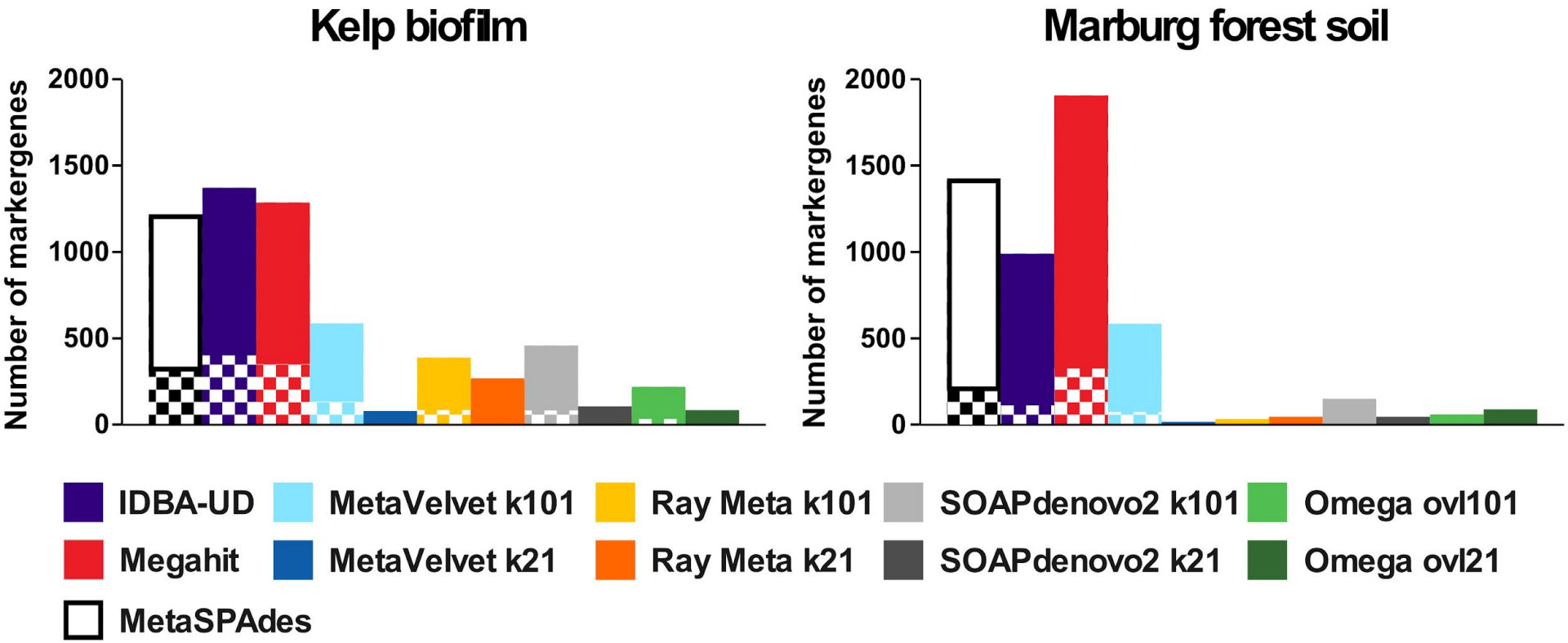
- **SPAdes**
- **MegaHIT**
- **IDBA**
- Minia
- Ray Meta
- SOAPdenovo
- Velvet/Velour
- ABySS

Long-read assemblers

Optimized for low quality long reads from PacBio/Nanopore sequencing

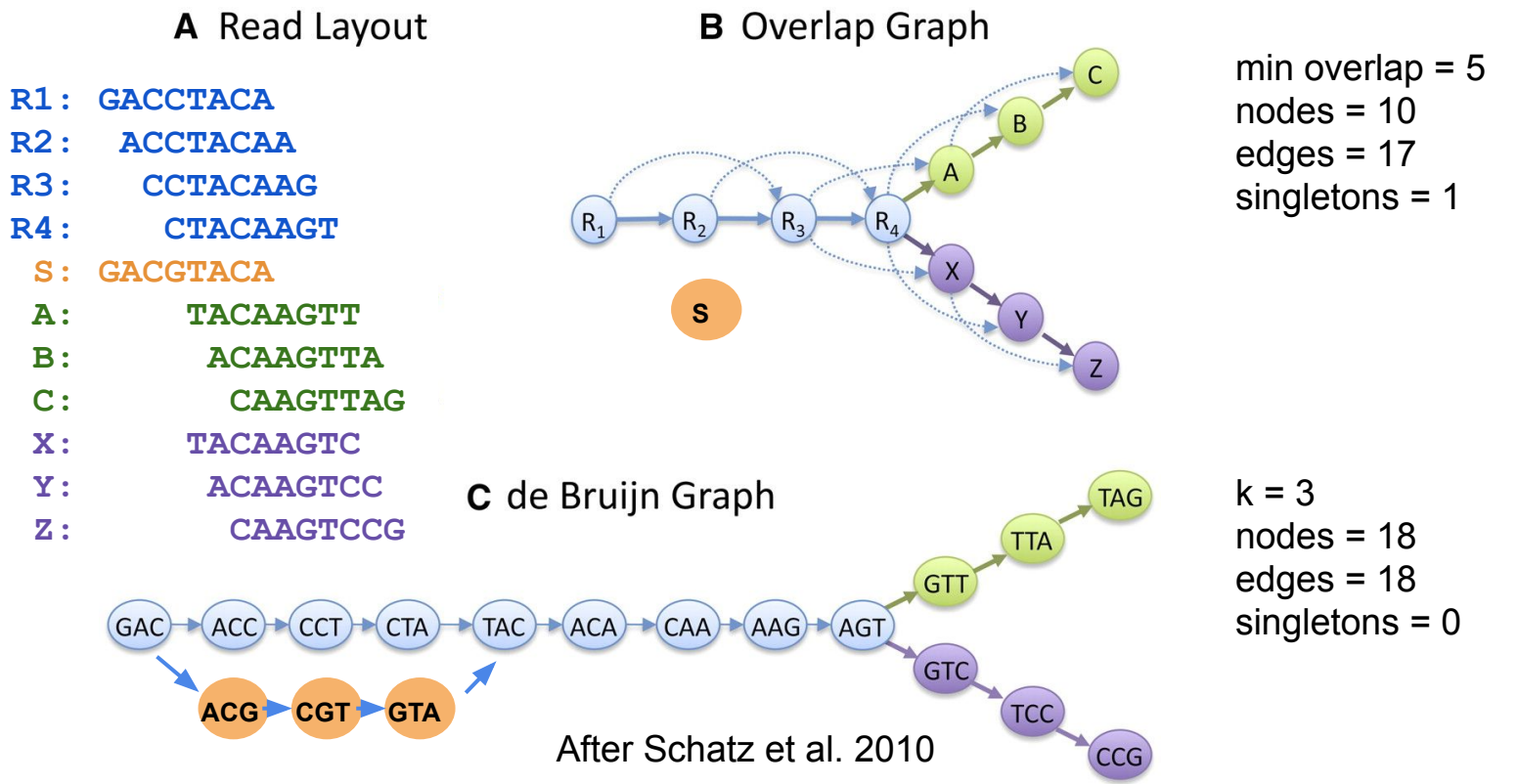
- Canu (Celera fork)
- Flye
- Miniasm/Minipolish
- Raven
- Redbean
- Shasta

Sequence Assembly Methods: Software Comparison

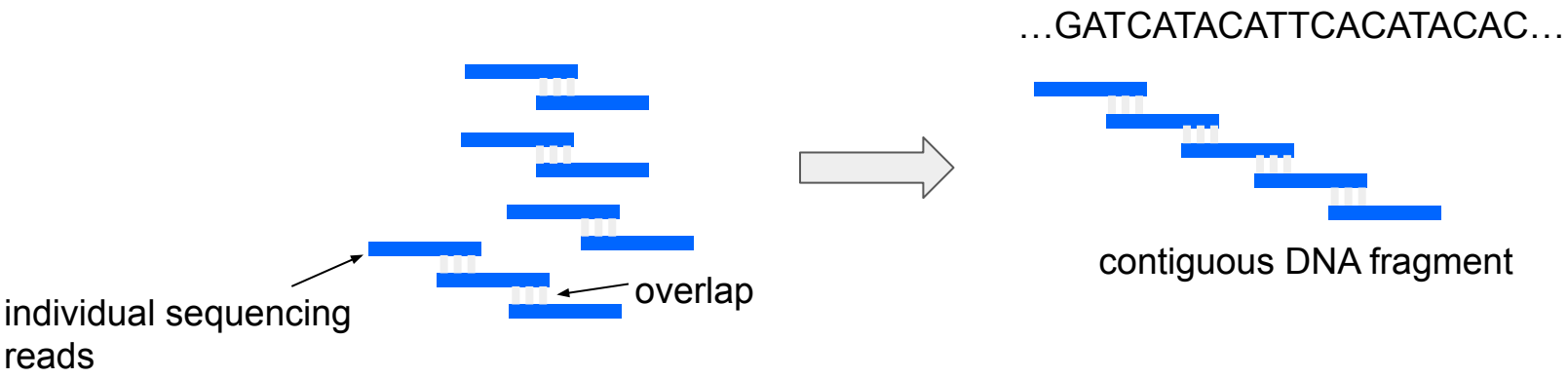


Vollmers et al. 2017

Sequence Assembly Methods: Theory



Overlap-Layout-Consensus Assembly

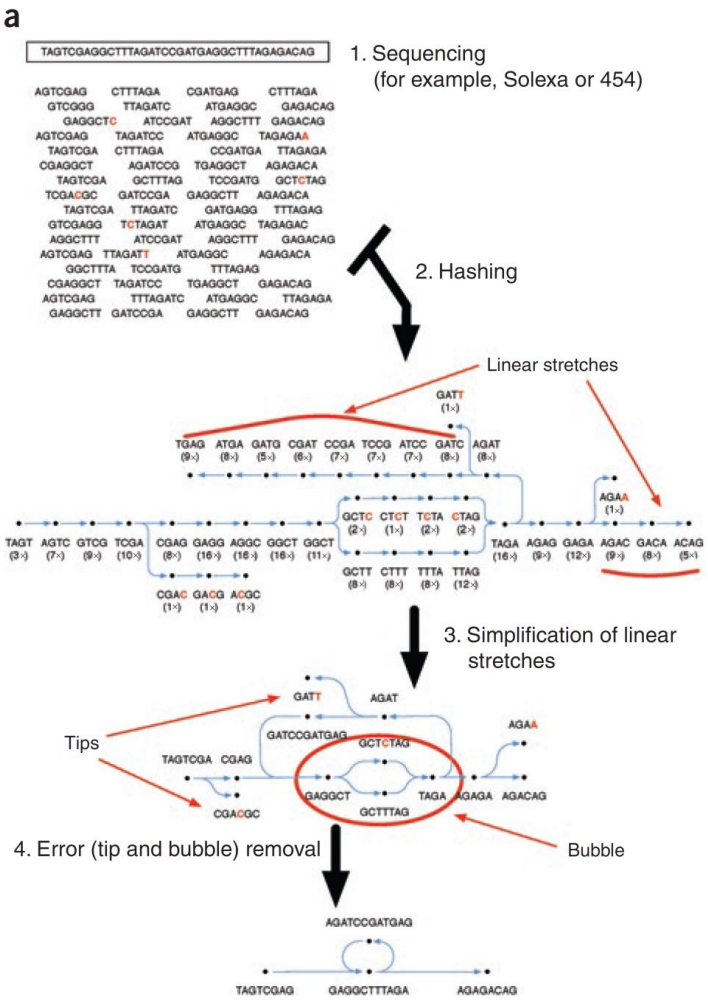


Assembly is greatly complicated by sequence heterogeneity and gaps in coverage due to undersampling.

Unassembled reads that do not find overlap are “singletons”

de Bruijn Graph Assembly

- Index k-mers in reads
  - k = 21 - 127 is typical
- Collect k-mers into nodes
- Record coverage of each node
- Rescue “mercy” k-mers that would be lost by coverage limits
- Flag repeats and errors by anomalous coverage
- Assemble prefix & suffix nodes to create edges
- Simplify graph by collapsing nodes into linear stretches
- Find a path or trail that visits each edge once
- Traverse optimal path to reconstruct consensus sequence



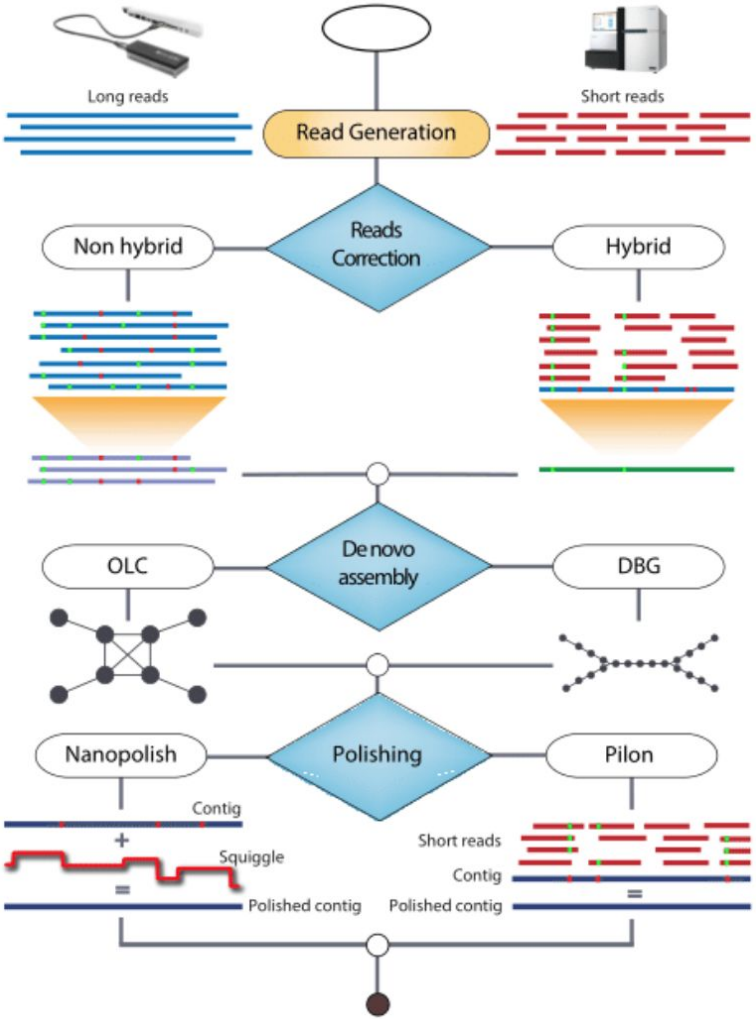
Flicek & Birney (2009) Nature Methods 6:S6

Long-read Assembly

Long-read (multi-kbp) sequences from Nanopore and PacBio suffer from high error rates, so two options for good assemblies:

1) Hybrid: Use long reads as scaffolds to resolve repeats from the short-read assembly

2) De novo: These methods use high coverage of long reads to resolve errors. Can also use short reads for ‘polishing’ completed assembly.



Magi et al (2018) Briefings in Bioinformatics

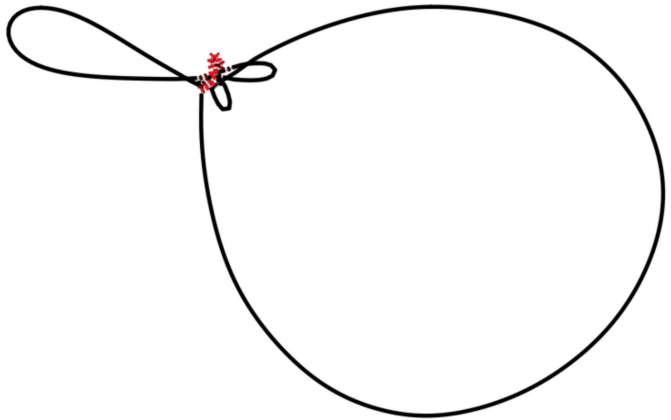
Meta/Genome Assembly

Parameterization + Iteration = Success!

- minimum length of overlap
- k-mer length
- mismatch penalties
- gap penalties
- minimum quality score
- contig merge stringency
- …and many more…

The output of assembly programs include contigs, scaffolds, and assembly graphs

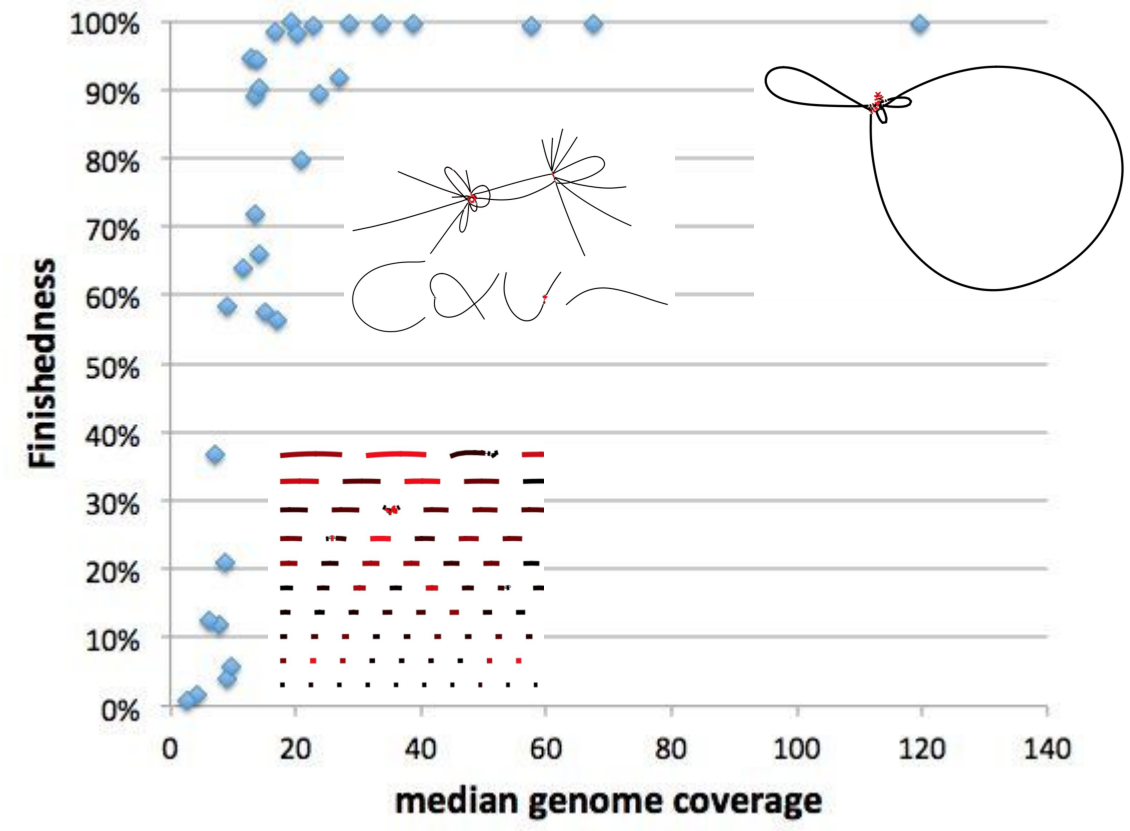
If you're doing genome sequencing, ***always look at your assembly graphs!***



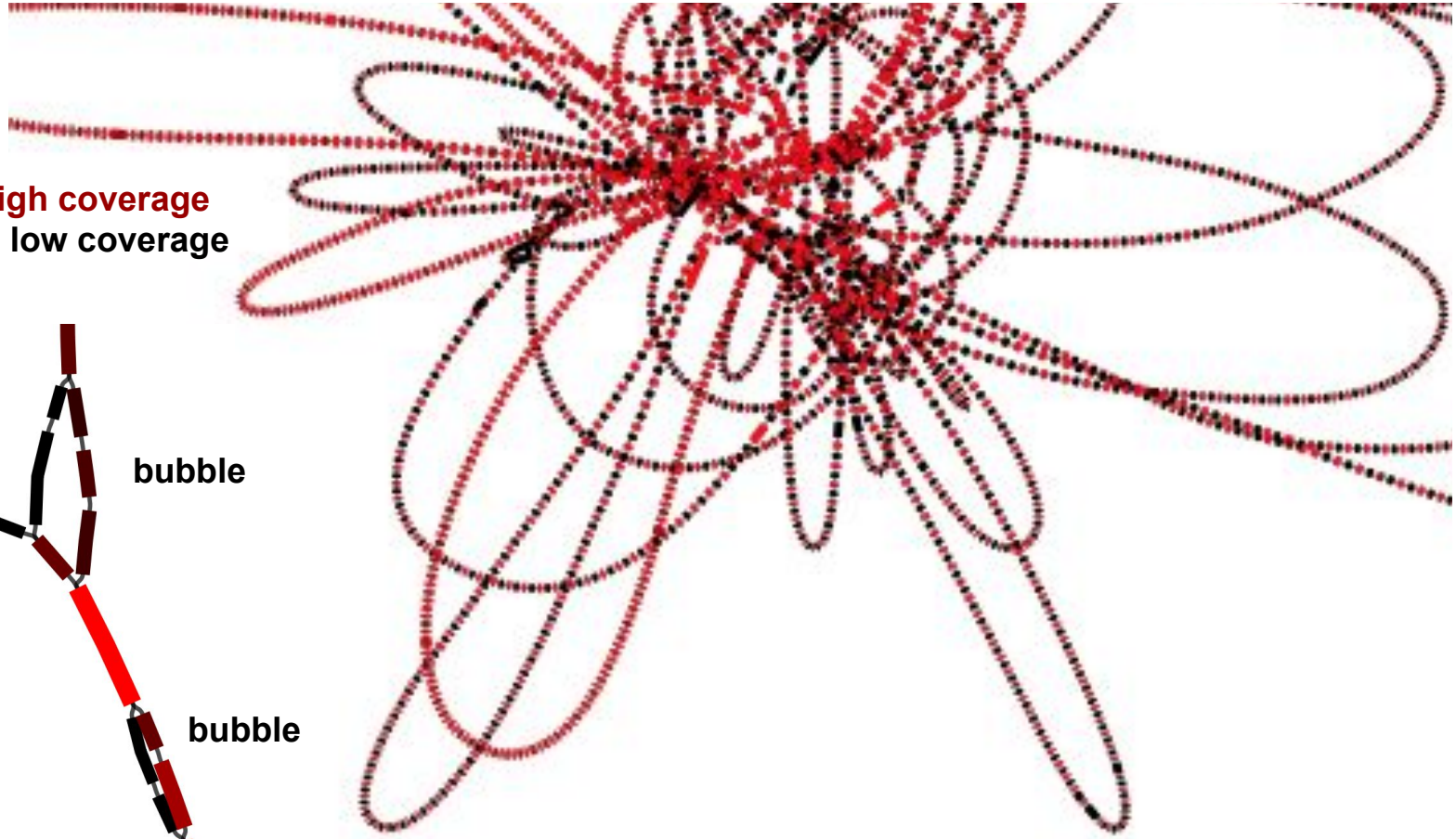
Genome Assembly: “Finishedness”

“Finishedness” (a term I made up) = size of largest connected component/total size of assembly

In real data, 30-40x coverage can be needed to get “perfect” assemblies due to sequencing error, biased library preparation methods, variations in DNA content due to exponential growth, or other mysterious reasons.

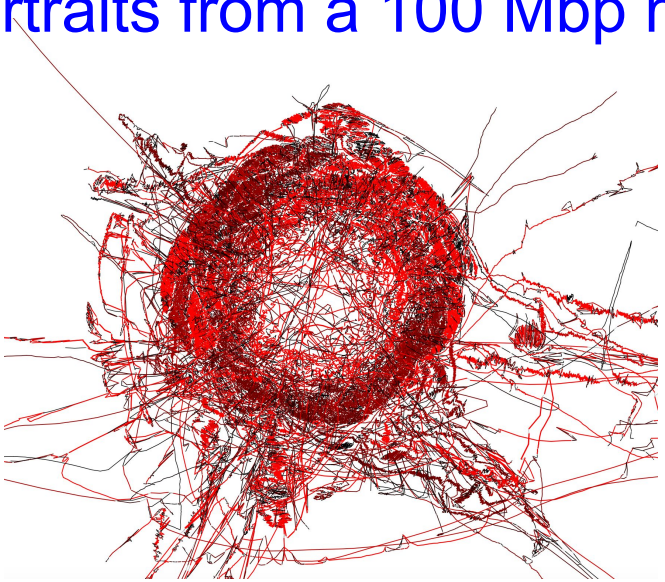


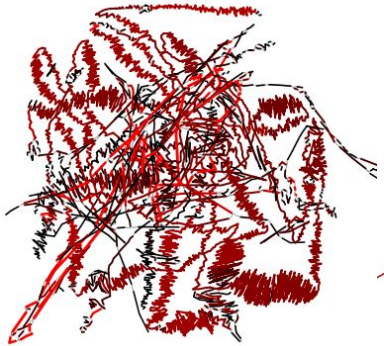
de Bruijn Graph Co-Assembly of Two Closely Related Genomes

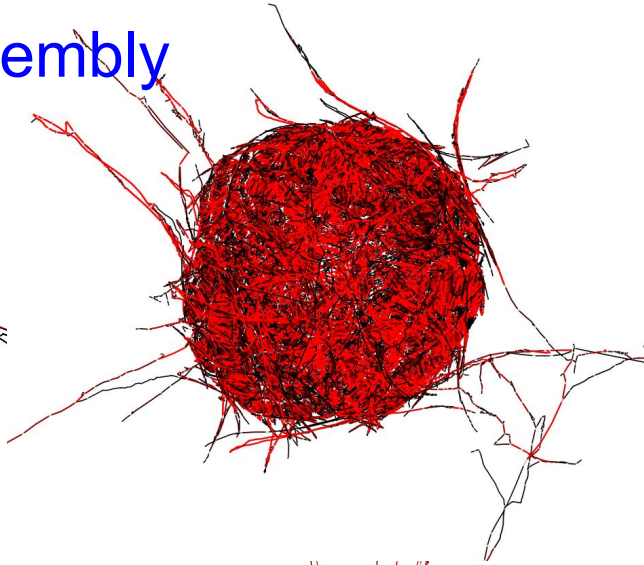


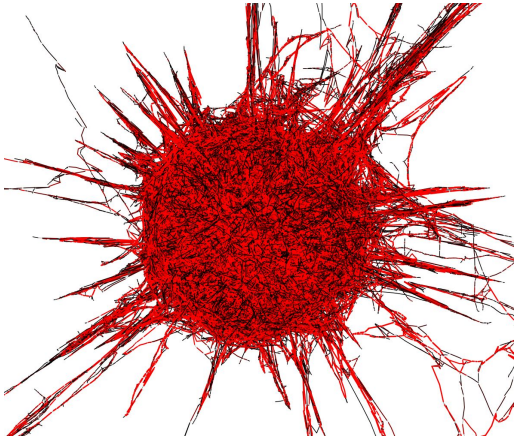
Visualized in Bandage (Wick et al. 2015)

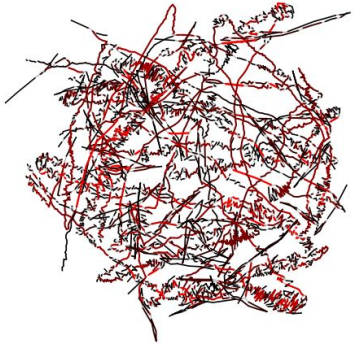
Portraits from a 100 Mbp metagenome assembly

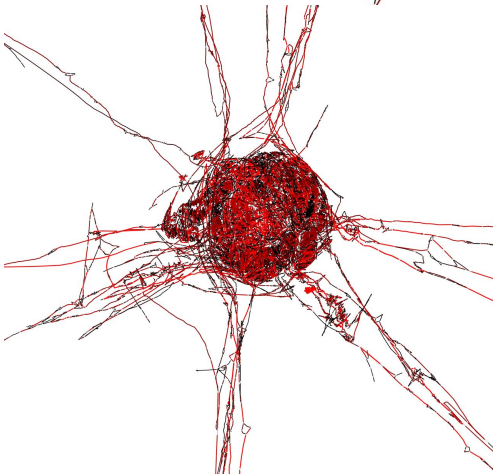












Resources

- Will Trimble: All kmers are not created equal
  - https://www.slideshare.net/wltrimbl/all-kmers-are-not-created-equal-recognizing-the-signal-from-the-noise-in-largescale-metagenomes
- Rob Edwards Genome Assembly videos
  - https://www.youtube.com/playlist?list=PLbHoaEmUsmfUXxTTGsL163rmeR8_qCY9F
- Megahit definition of mercy k-mers
  - https://github.com/voutcn/megahit/issues/86
- CAMI: Critical Assessment of Metagenome Interpretation
  - https://data.cami-challenge.org/
  - https://www.nature.com/articles/nmeth.4458

Demo 5.1 Genome Assembly using SPAdes

- Binder for demo
  - https://mybinder.org/v2/gh/biovcnet/metagenomics-binder-assembly/master?urlpath=lab
- Bash script
  - https://github.com/biovcnet/topic-metagenomics/blob/master/Lesson-5/Demo5.1_spades.sh