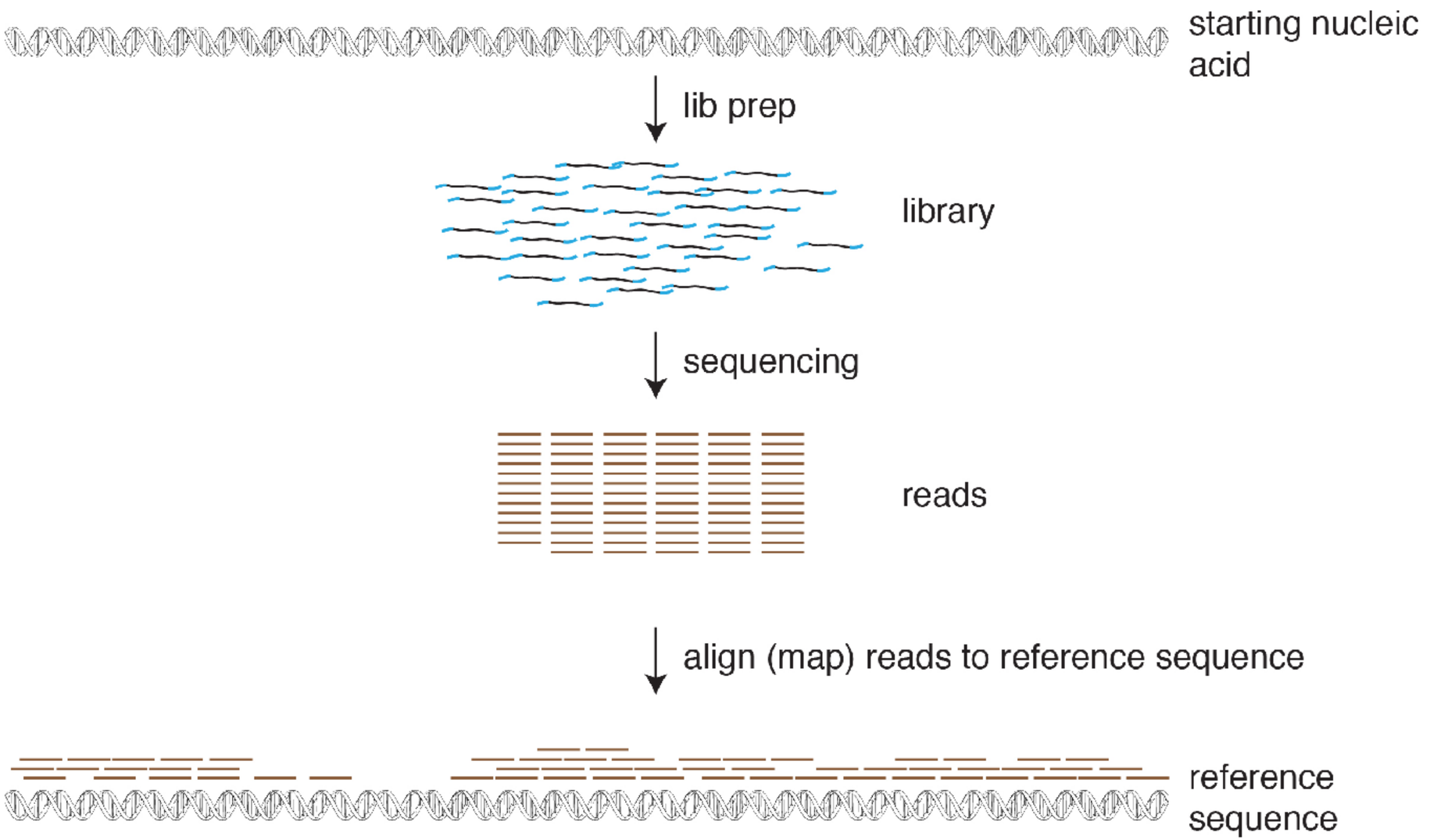
Genome Assembly

Mark Stenglein

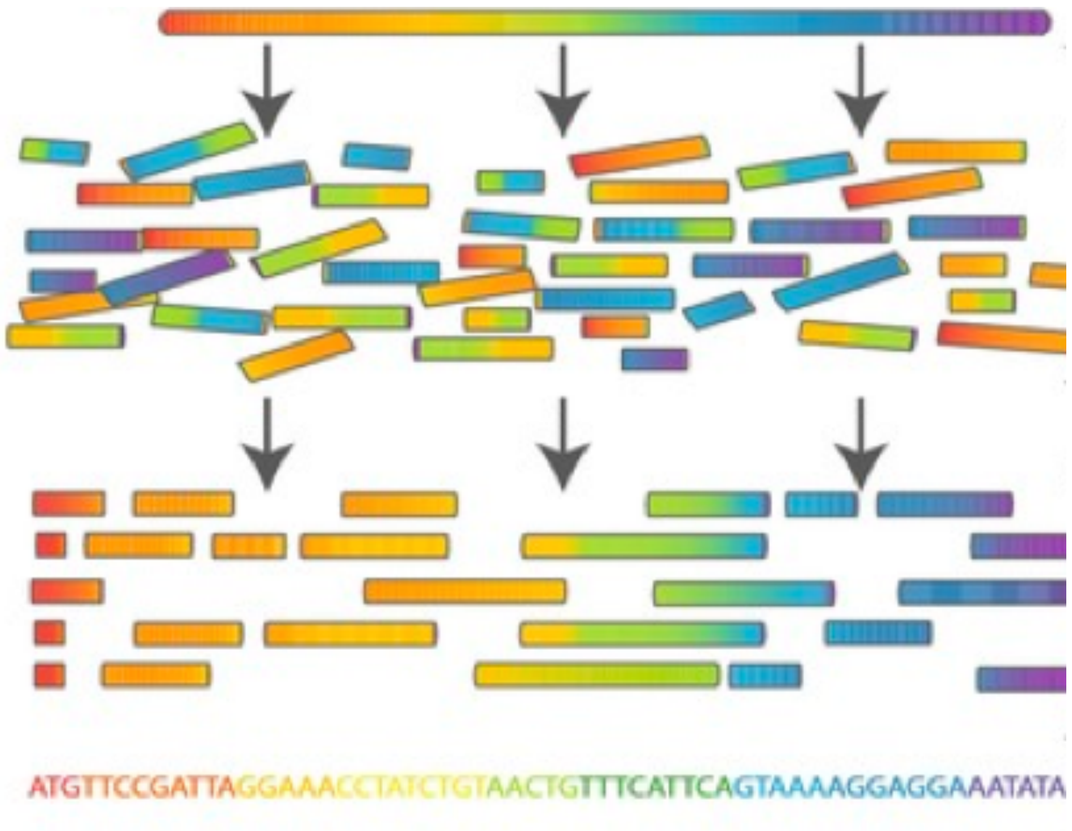
Computational Biology
Workshop

Todos Santos Center
May 9-12, 2022

Mapping is the process by which sequencing reads are aligned to the region of a genome from which they derive.



Genome **assembly** is the process of trying to
reconstruct a genome sequence from reads (making a new reference sequence)



not yet sequenced genome

reads

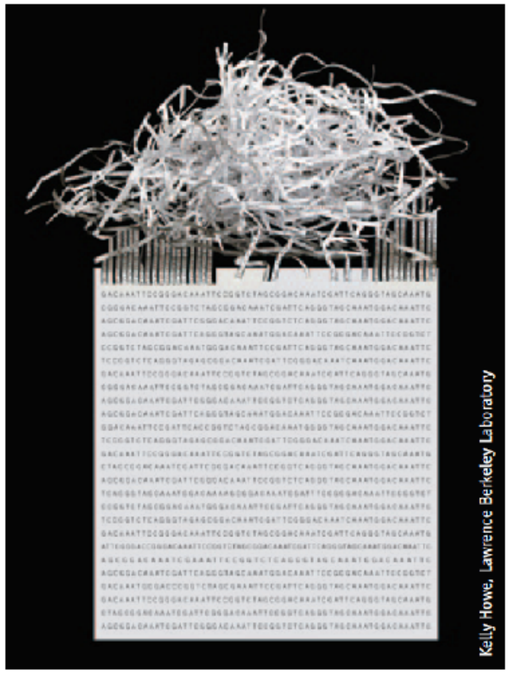
contigs

reconstructed (assembled) genome

Image: Commins, J., Toft, C., Fares, M. A (2009)

Genome assembly is the process of *attempting* to reconstruct a genome sequence

An assembly is only a “putative reconstruction” of the genome sequence *[Miller, Koren, Sutton (2010)]*



Baker M (2012) Nat Methods

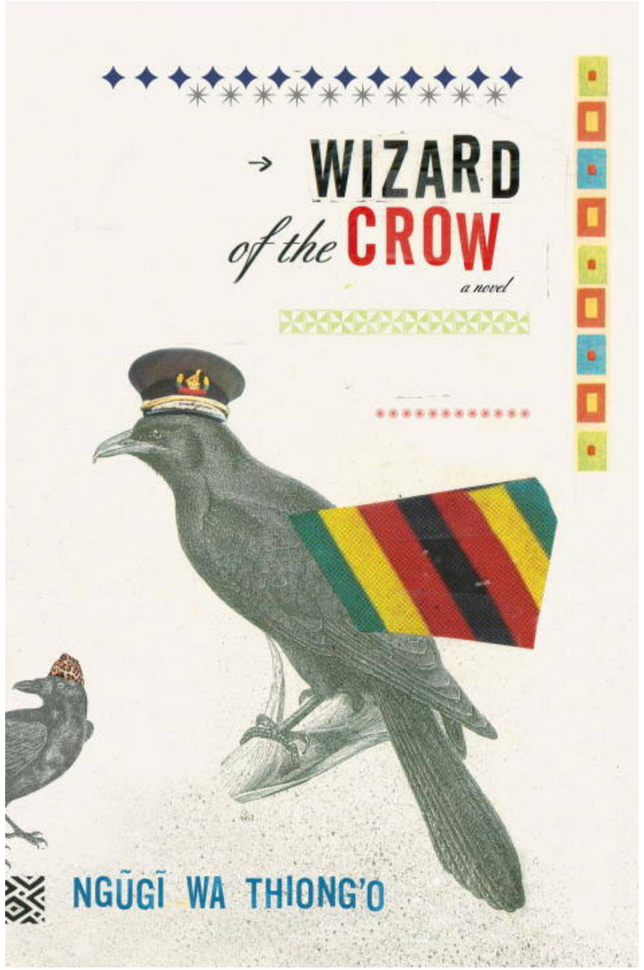
Keith Bradnam, UC Davis

Genome assembly exercise

Your job is to assemble the ‘genome’ from which the ‘reads’ you’ve been given derive.

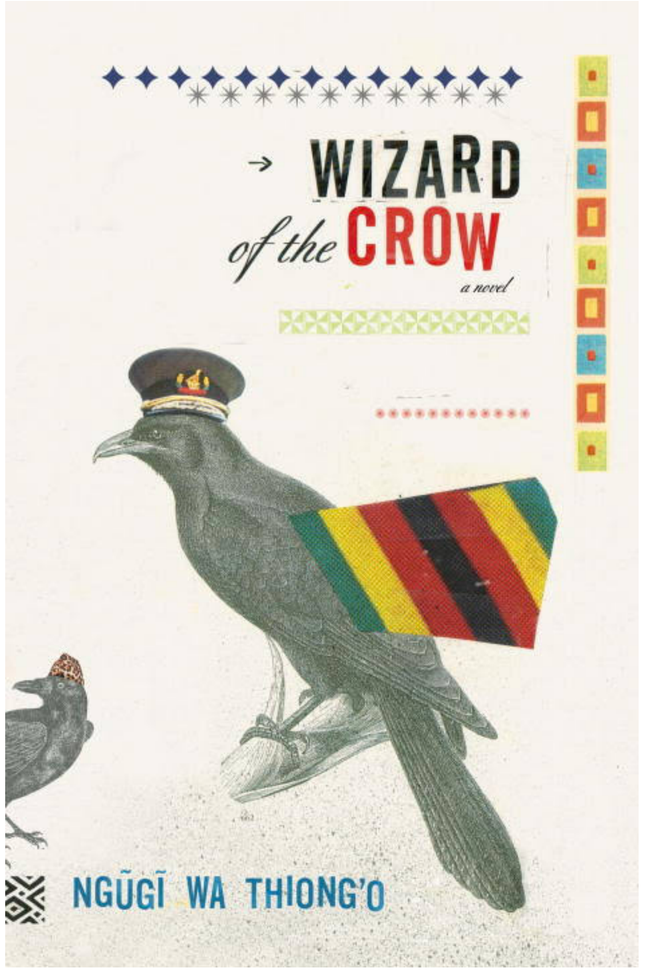
Rules/info:

- Like real sequencing data, these reads contain errors. The error rate is ~2%
- These are single-end 11-base reads
- The average coverage is ~6x
- You’re not allowed to google the answer
- Also: the answer is in the slides: don’t cheat!
- You can use your computers (i.e. word processors or text editors) or paper and whatever strategy you want to do the assembly…



Exercise inspired and enabled by Titus Brown: http://ivory.idyll.org/blog/the-assembly-exercise.html

Genome assembly paper exercise



"Jinn (Arabic), also romanized as djinn … are supernatural creatures in early Arabian and later Islamic mythology and theology."

https://en.wikipedia.org/wiki/Jinn

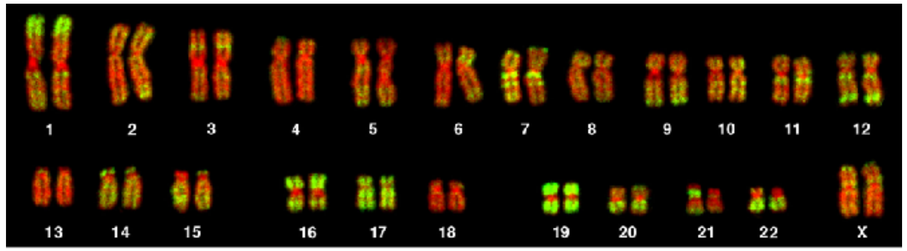
Exercise inspired and enabled by Titus Brown: http://ivory.idyll.org/blog/the-assembly-exercise.html

Conclusion: assembly is difficult!

In this exercise, the ‘genome’ was only 65 positions long, and its alphabet contained 26 ‘bases’ (more information rich)

Eukaryotic genomes can have billions of bases and there are only 4 bases (less information)

the human *haploid* genome is 3 Gb

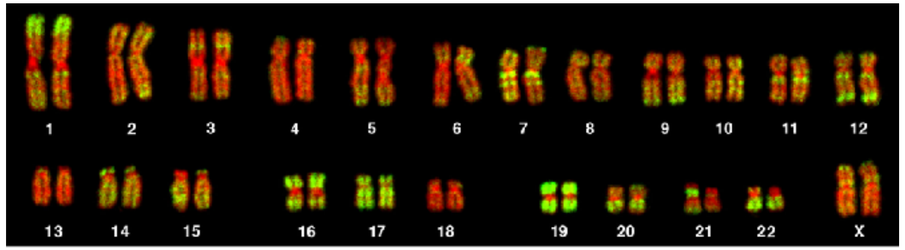


Bolzer et al (2005) PLoS Biol

Some of the reasons that assembly is difficult

1) Genomes are full of repetitive sequences

Alu sequences in the human genome
1 million copies, ~10% of the mass



Bolzer et al (2005) PLoS Biol

2) Reads contain errors

```
_gew_kjinns          get_djinns_          l_get_djinn
```

3) Uneven coverage, including possibly no coverage for particular regions (e.g. GC-rich regions)

4) Even with fast computers, it's still computationally difficult

5) Since you don't know what the 'answer' is, it can be difficult to assess whether your assembly is 'good' or not

6) Polyploidy means you are effectively assembling >1 closely related, but not identical, genome

7) Not to mention annotation, which can be as hard as assembly!

De novo assembly is like doing a jigsaw puzzle without the picture on the box

Images, metaphor: *Keith Bradnam, UC Davis*

Reads are assembled into **contigs**, contigs into **scaffolds**, and scaffolds into chromosomes or genomes

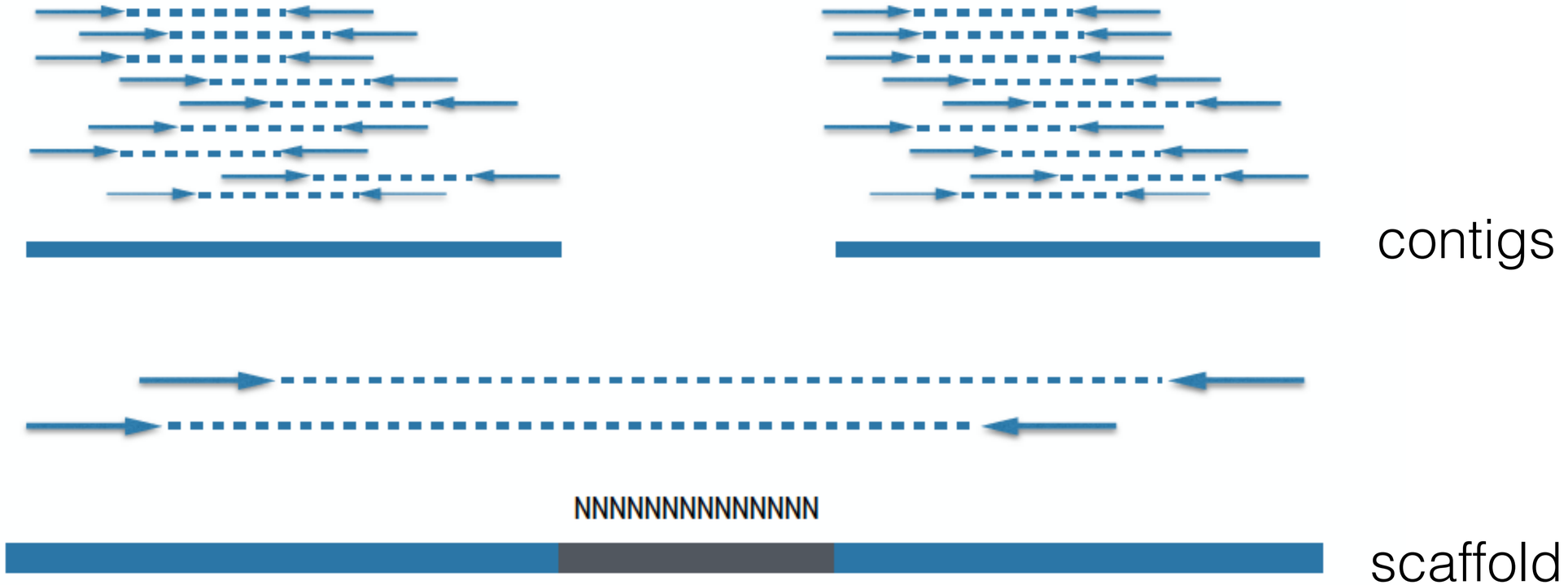


Image: *Keith Bradnam, UC Davis*

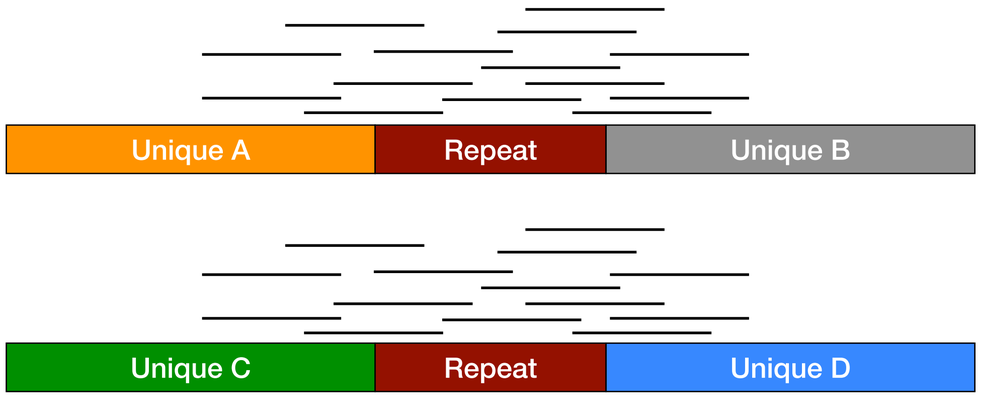
These “contigs” could be scaffolded because we have additional information.

(We know that the two halves of the golden gate bridge should go together)

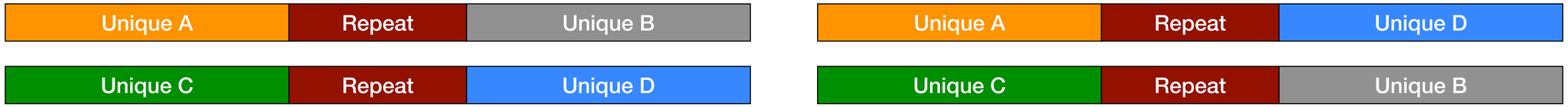
Image, analogy: *Keith Bradnam, UC Davis*

Short reads alone do not contain enough information to resolve repeats

No read with read length < repeat length can bridge the repeat to link unique sequence on each side



2 equally plausible combinations given short read data



Short and Long Read Sequencing Technology

Short read sequencing

- Millions of reads
- Relatively short: ~50-300 nt (Illumina)
- Relative low error rates: ~0.1%
- Illumina has virtually all of the market share

MiSeq

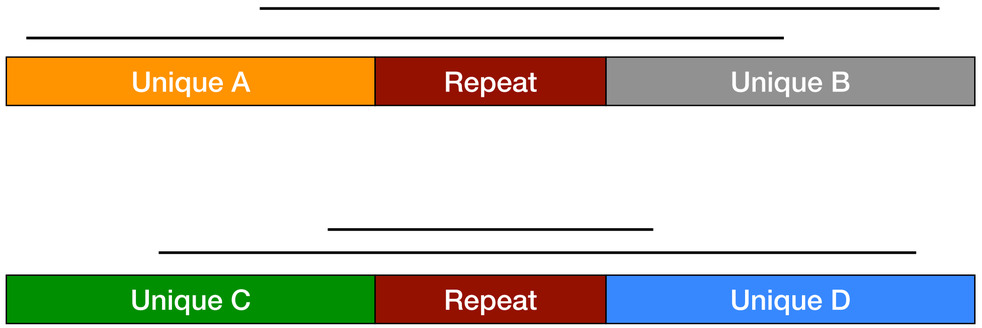
Long read sequencing

- Fewer, longer reads
- >1 kb (PacBio), up to 100s of kb (Oxford Nanopore)
- Relative high error rates: ~10%

Oxford Nanopore MinION

PacBio RS-II

Long reads span repeats to link unique sequence on each side



How do you know if your assembly is good?

- Size of the assembly: does it match estimates from other means?
- Size of the contigs/scaffolds: are they reasonably long?
- Are the expected ‘core genes’ present in the assembly?
- What fraction of reads map to the assembly?
- Does the assembly contain sequences of contaminating organisms?
- Is the assembly consistent with independently derived data? (optical mapping, transcriptome sequencing, genomes of related organisms?)

For what purpose do you need the assembly?

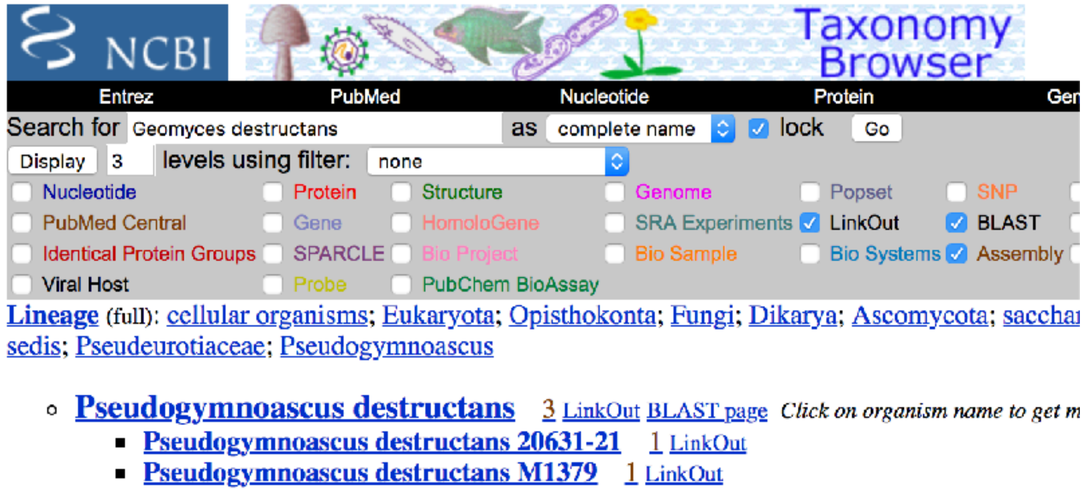
These questions apply to assemblies in databases too.

Exercise: compare 3 genome assemblies

Pseudogymnoascus destructans
cause of white nose syndrome

image: Marvin Moriarty/USFWS

Visit the pages for the 3 assemblies.
How were they made? What type of data?
Is one obviously better? Which would you use?



a common assembly metric:
N50: a measure of the average size of contigs & scaffolds

Not all assembly problems are difficult!

tiny viral genome: easy

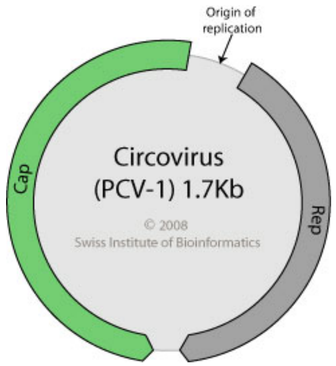
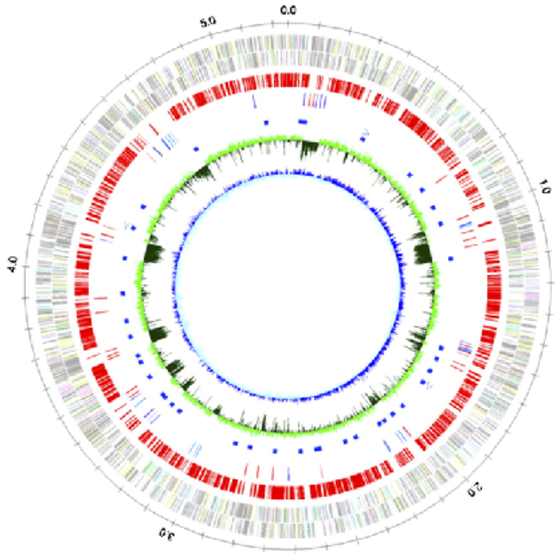


image: viralzone

bacterial genomes ~5 Mbp
A little more difficult



Nakazawa et al (2009) Genome Research

Loblloly pine (*Pinus teada*)

22 Gbp genome!
Very difficult

image: Univ of Alabama