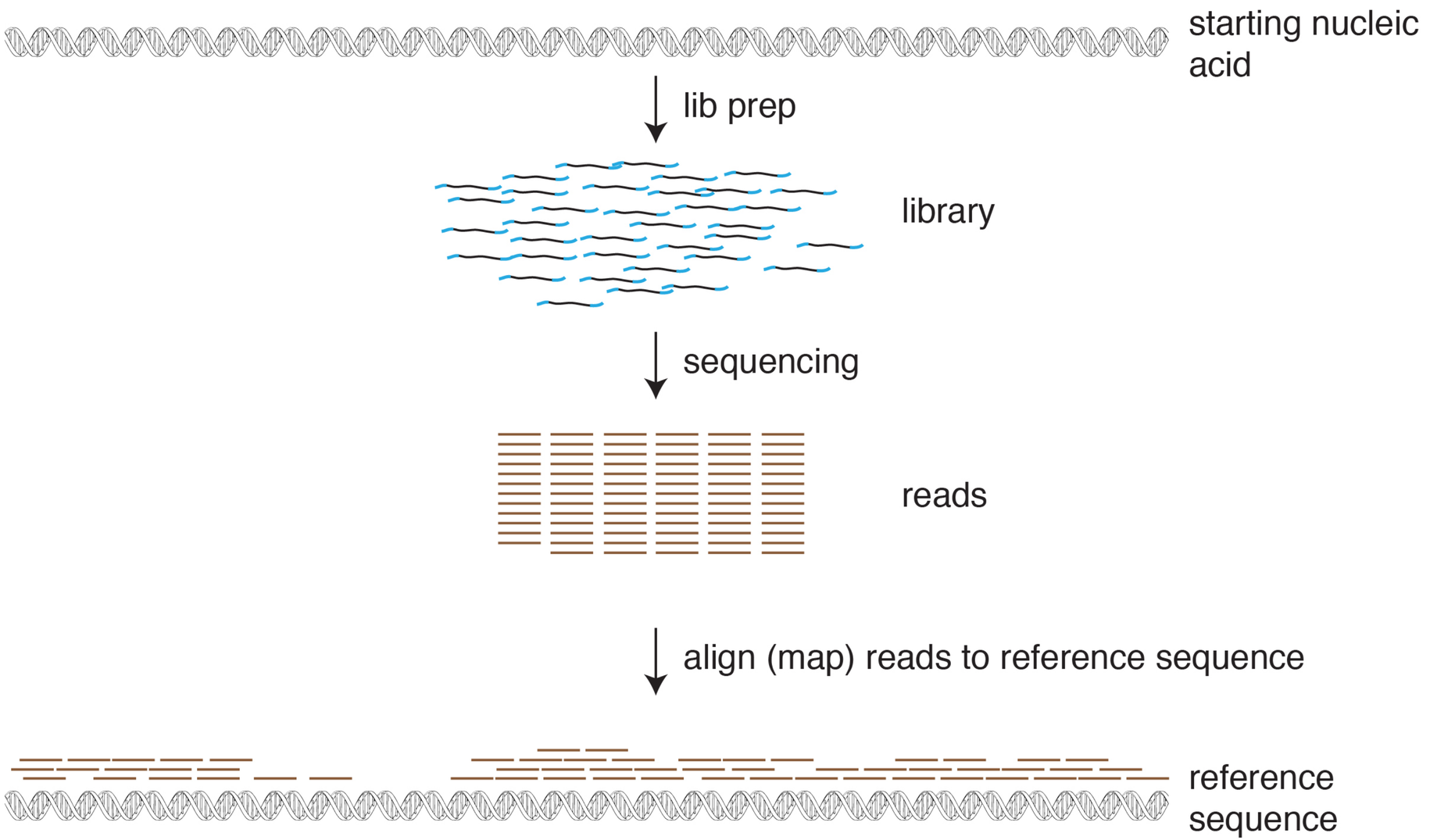
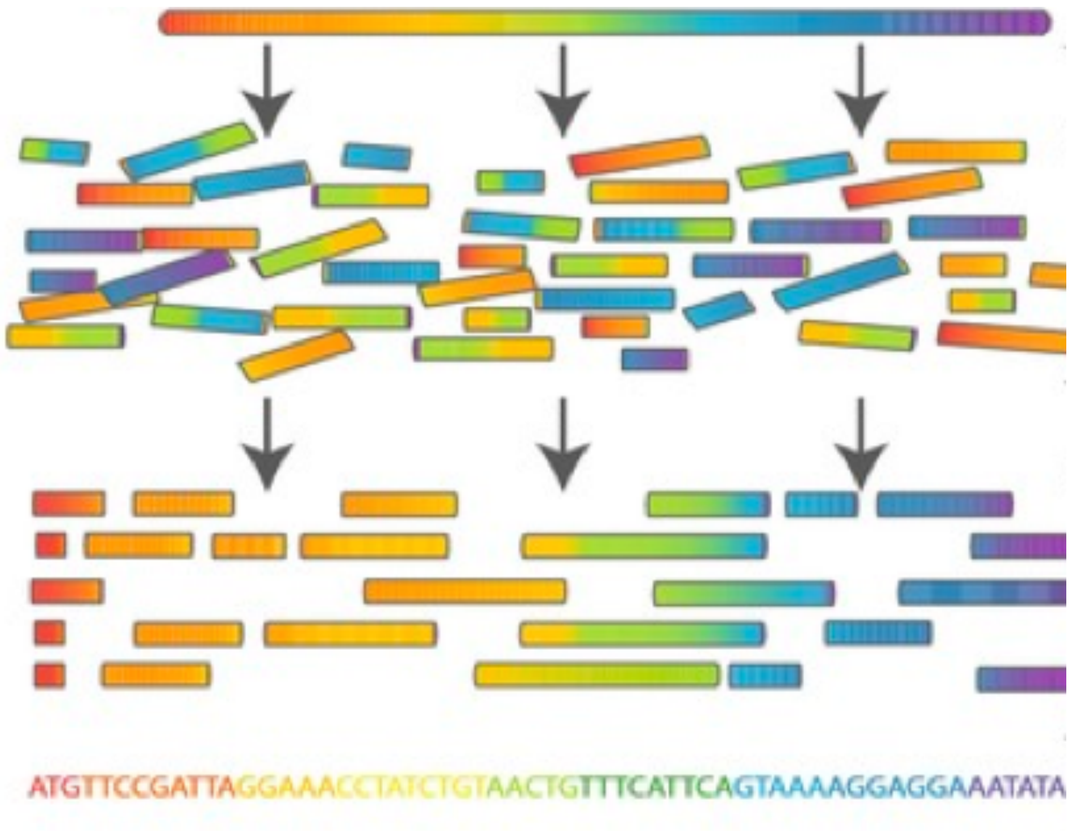
Genome Assembly

Mark Stenglein, MIP 280A4

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

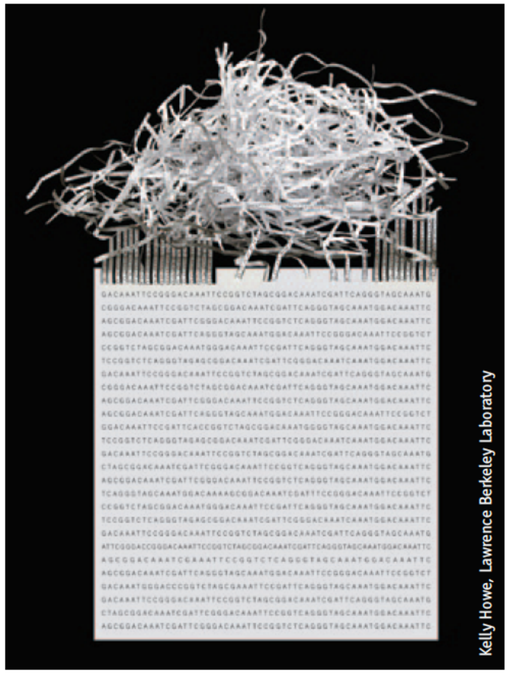
contigs

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

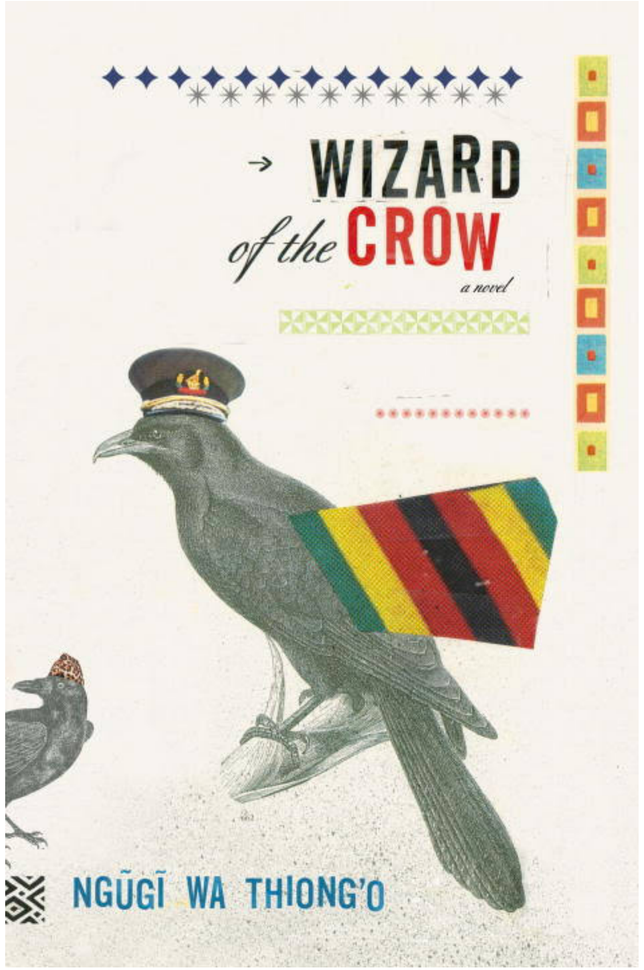
Keith Bradnam, UC Davis

Genome assembly exercise

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

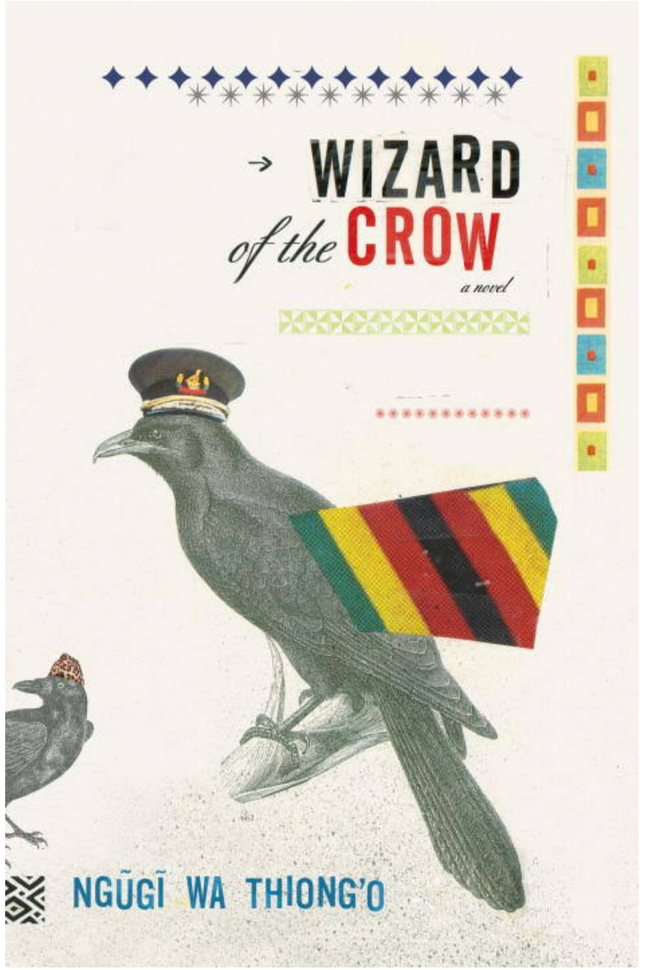
Rules/info:

- 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

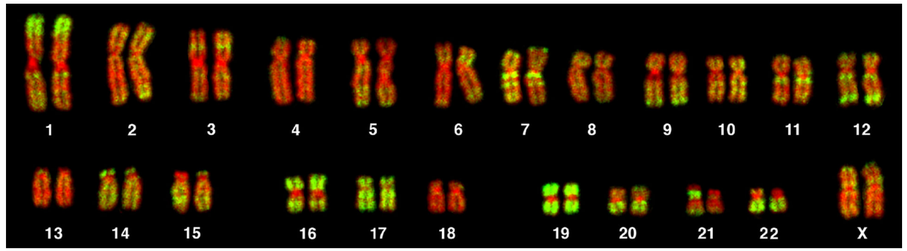
Exercise inspired and enabled by Titus Brown: http://ivory.idyll.org/blog/the-assembly-exercise.html

Conclusion: assembly is not trivial!

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

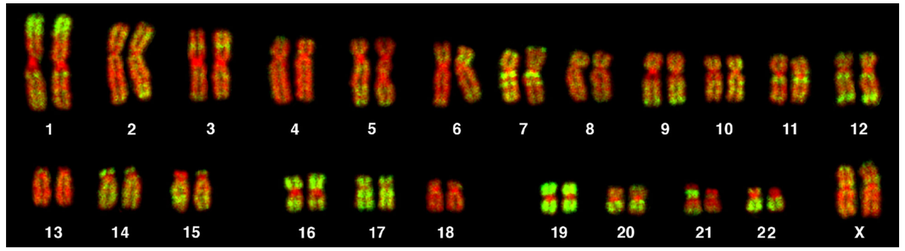


Bolzer et al (2005) PLoS Biol

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

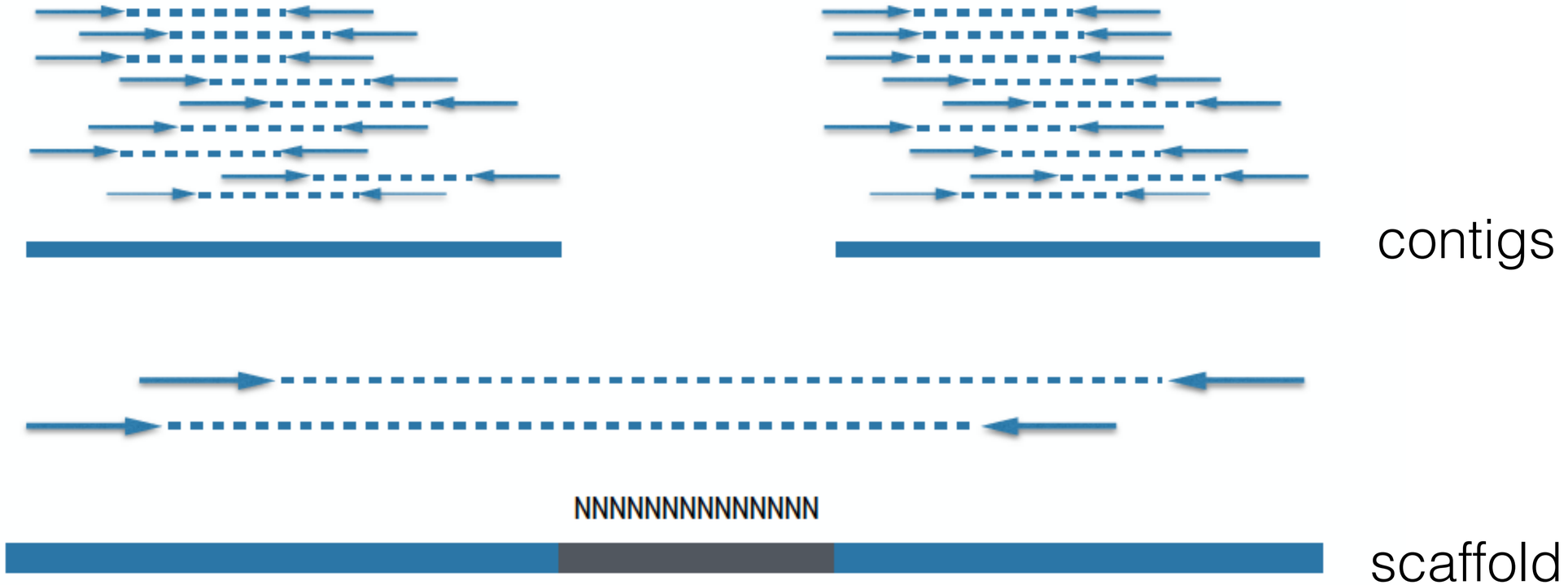


Image: *Keith Bradnam, UC Davis*

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*

Sometimes even 'complete' assemblies contain gaps

human Y chromosome from GRCh38.p12

A truly complete (?) human genome

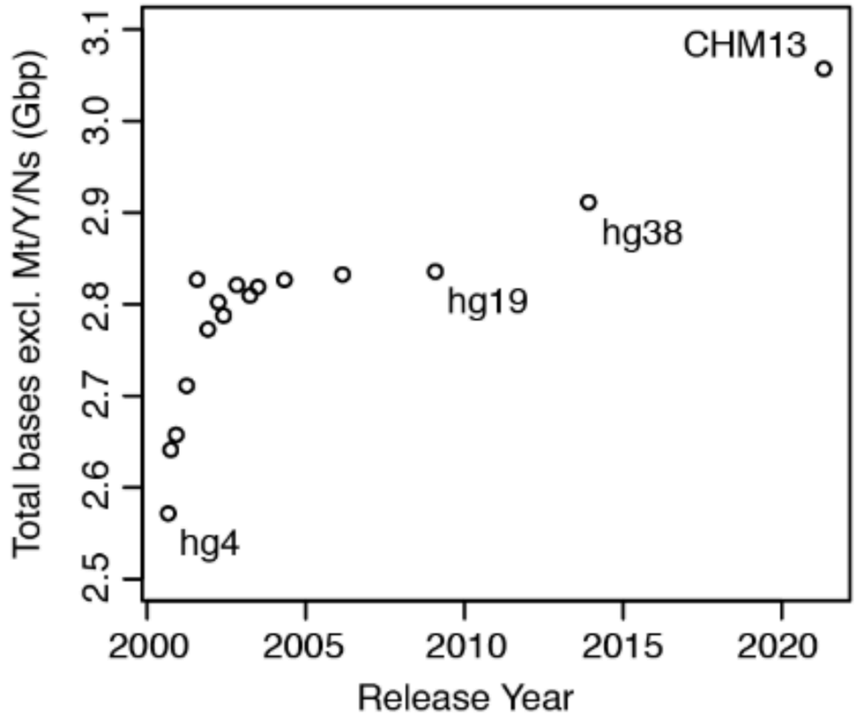
RESEARCH ARTICLE

HUMAN GENOMICS

The complete sequence of a human genome

Sergey Nurk[1]†, Sergey Koren[1]†, Arang Rhie[1]†, Mikko Rautiainen[1]†, Andrey V. Bzikadze[2], Alla Mikheenko[3], Mitchell R. Vollger[4], Nicolas Altemose[5], Lev Uralsky[6,7], Ariel Gershman[8], Sergey Aganezov[9]‡, Savannah J. Hoyt[10], Mark Diekhans[11], Glennis A. Logsdon[4], Michael Alonge[9], Stylianos E. Antonarakis[12], Matthew Borchers[13], Gerard G. Bouffard[14], Shelise Y. Brooks[14], Gina V. Caldas[15], Nae-Chyun Chen[9], Haoyu Cheng[16,17], Chen-Shan Chin[18], William Chow[19], Leonardo G. de Lima[13], Philip C. Dishuck[4], Richard Durbin[19,20], Tatiana Dvorkina[3], Ian T. Fiddes[21], Giulio Formenti[22,23], Robert S. Fulton[24], Arkarachai Fungtammasan[18], Erik Garrison[11,25], Patrick G. S. Grady[10], Tina A. Graves-Lindsay[26], Ira M. Hall[27], Nancy F. Hansen[28], Gabrielle A. Hartley[10], Marina Haukness[11], Kerstin Howe[19], Michael W. Hunkapiller[29], Chirag Jain[1,30], Miten Jain[11], Erich D. Jarvis[22,23], Peter Kerpedjiev[31], Melanie Kirsche[9], Mikhail Kolmogorov[32], Jonas Korlach[29], Milinn Kremitzki[26], Heng Li[16,17], Valerie V. Maduro[33], Tobias Marschall[34], Ann M. McCartney[1], Jennifer McDaniel[35], Danny E. Miller[4,36], James C. Mullikin[14,28], Eugene W. Myers[37], Nathan D. Olson[35], Benedict Paten[11], Paul Peluso[29], Pavel A. Pevzner[32], David Porubsky[4], Tamara Potapova[13], Evgeny I. Rogaev[6,7,38,39], Jeffrey A. Rosenfeld[40], Steven L. Salzberg[9,41], Valerie A. Schneider[42], Fritz J. Sedlazeck[43], Kishwar Shafin[11], Colin J. Shew[44], Alaina Shumate[41], Ying Sims[19], Arian F. A. Smit[45], Daniela C. Soto[44], Ivan Sović[29,46], Jessica M. Storer[45], Aaron Streets[5,47], Beth A. Sullivan[48], Françoise Thibaud-Nissen[42], James Torrance[19], Justin Wagner[35], Brian P. Walenz[1], Aaron Wenger[29], Jonathan M. D. Wood[19], Chunlin Xiao[42], Stephanie M. Yan[49], Alice C. Young[14], Samantha Zarate[9], Urvashi Surti[50], Rajiv C. McCoy[49], Megan Y. Dennis[44], Ivan A. Alexandrov[3,7,51], Jennifer L. Gerton[13,52], Rachel J. O'Neill[10], Winston Timp[8,41], Justin M. Zook[35], Michael C. Schatz[9,49], Evan E. Eichler[4,53]*, Karen H. Miga[11,54]*, Adam M. Phillippy[1]*

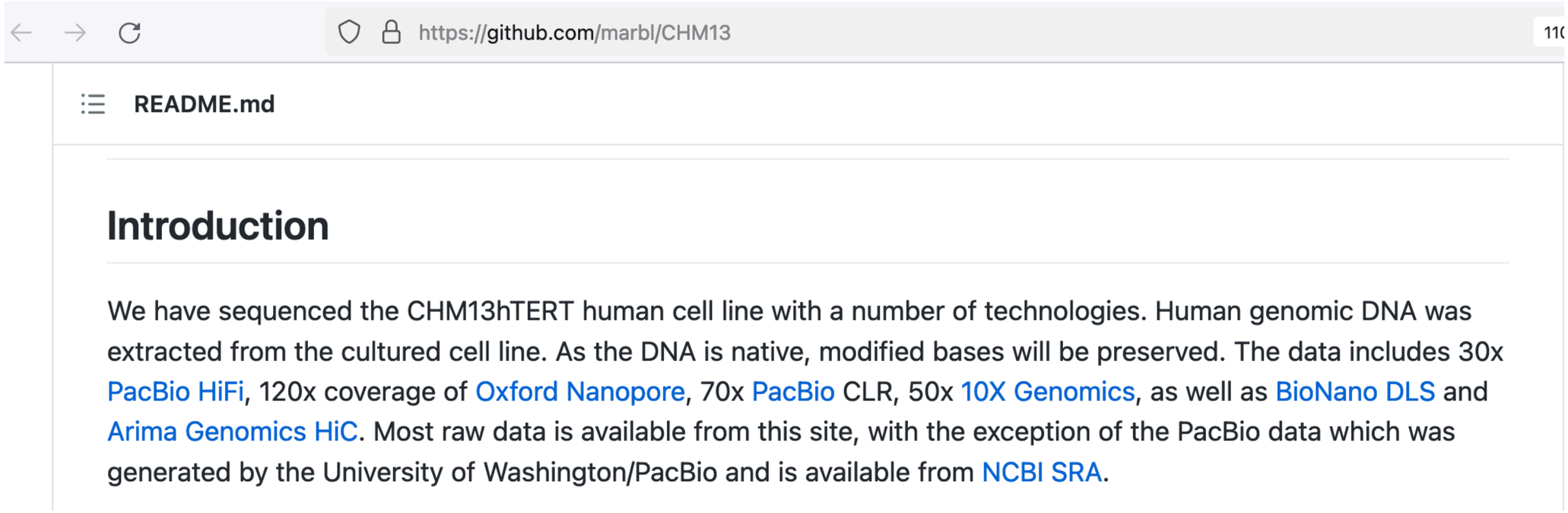
Since its initial release in 2000, the human reference genome has covered only the euchromatic fraction of the genome, leaving important heterochromatic regions unfinished. Addressing the remaining 8% of the genome, the Telomere-to-Telomere (T2T) Consortium presents a complete 3.055 billion–base pair sequence



https://www.science.org/doi/epdf/10.1126/science.abj6987

This “telomere-to-telomere” assembly required a ton of data and many different types of data

None of this was standard Illumina short read data



https://www.biorxiv.org/content/10.1101/2021.05.26.445798v1

Nearly all short read assemblers use a de Bruijn graph-based algorithm

De bruijn graphs are directed graphs with connected nodes of overlapping k-mers

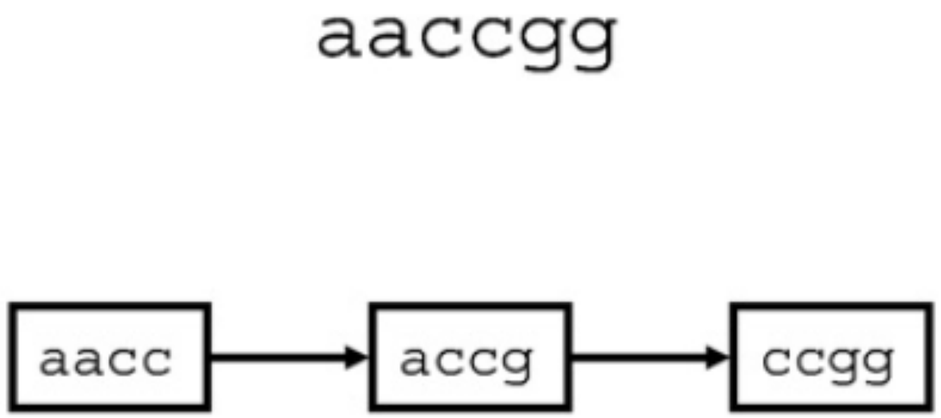


Image: *Miller, Koren, Sutton (2010) Genomics*

Generic simplified strategy:

- Attempted error correction
- Break reads into overlapping k-mers (here k = 4)
- Construct de Bruijn graph of k-mers
- Trace path through graph: Tada! Genome sequence

kmers are just sequences of length k

Ryan Wick
@rrwick

Bioinformatics joke:

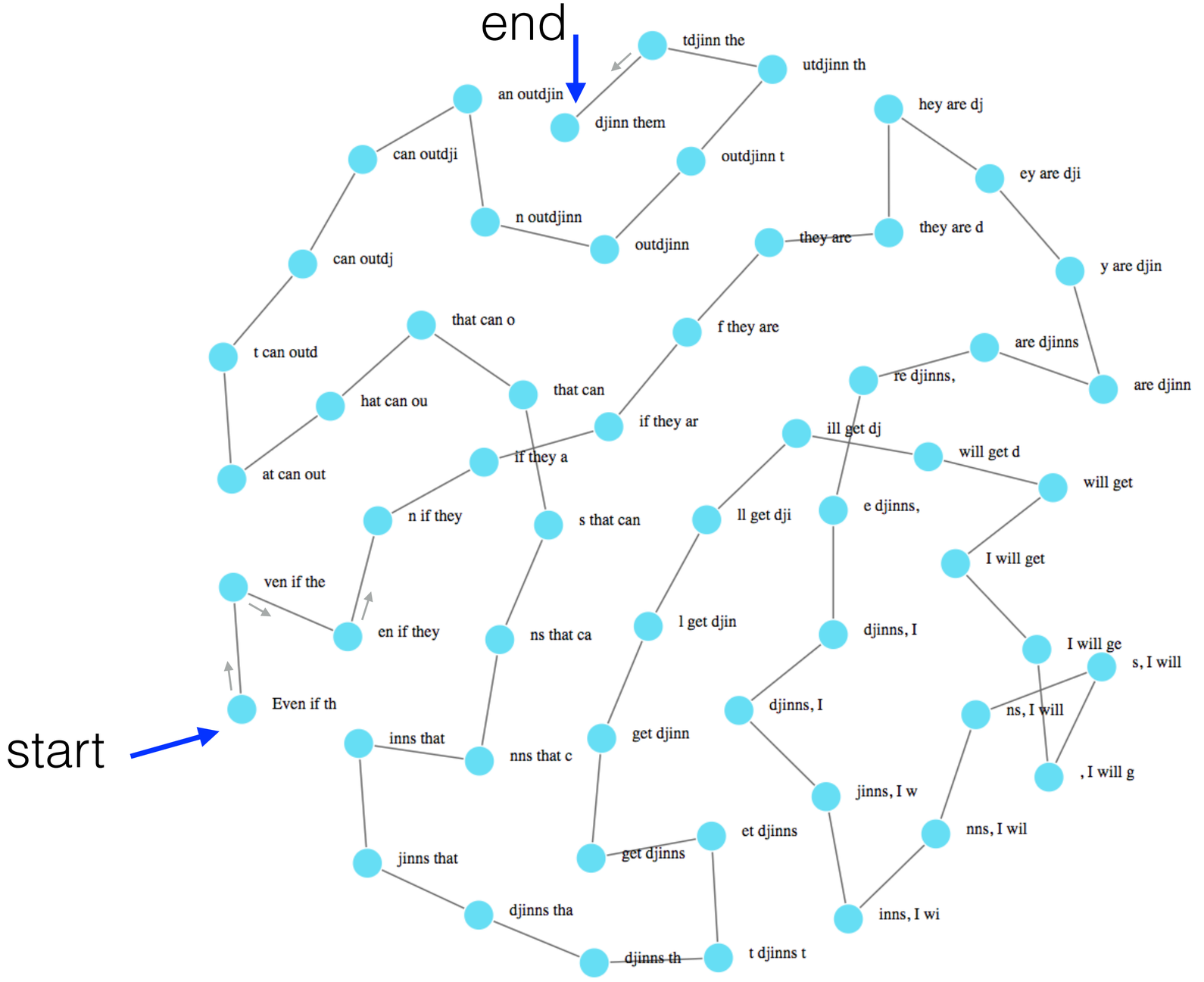
What do you call a tetranucleotide that's had a base added or removed?

A former 4-mer.

10:52 PM · Jun 24, 2020 · Twitter Web App

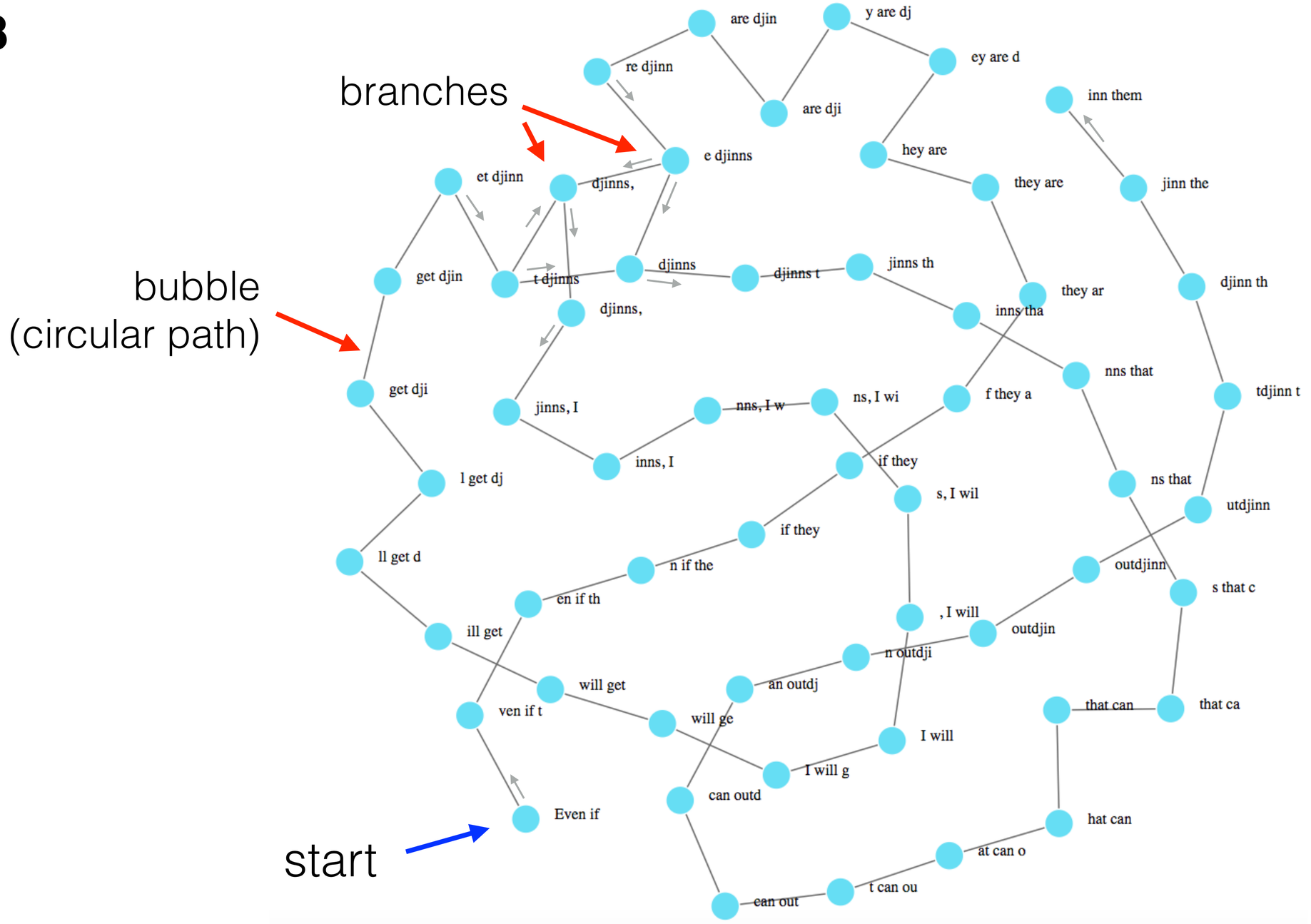
19 Retweets **109** Likes

k=10



http://debruijn.herokuapp.com/graph

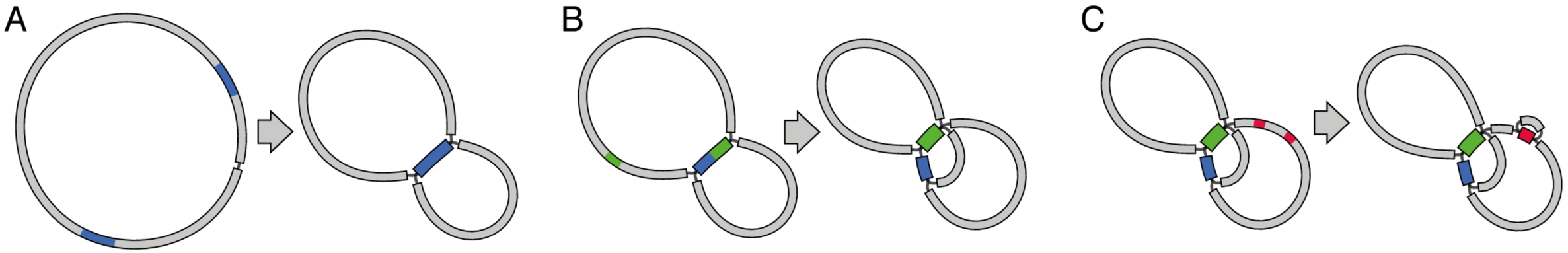
k=8



http://debruijn.herokuapp.com/graph

The impact of additional repeats on graph complexity

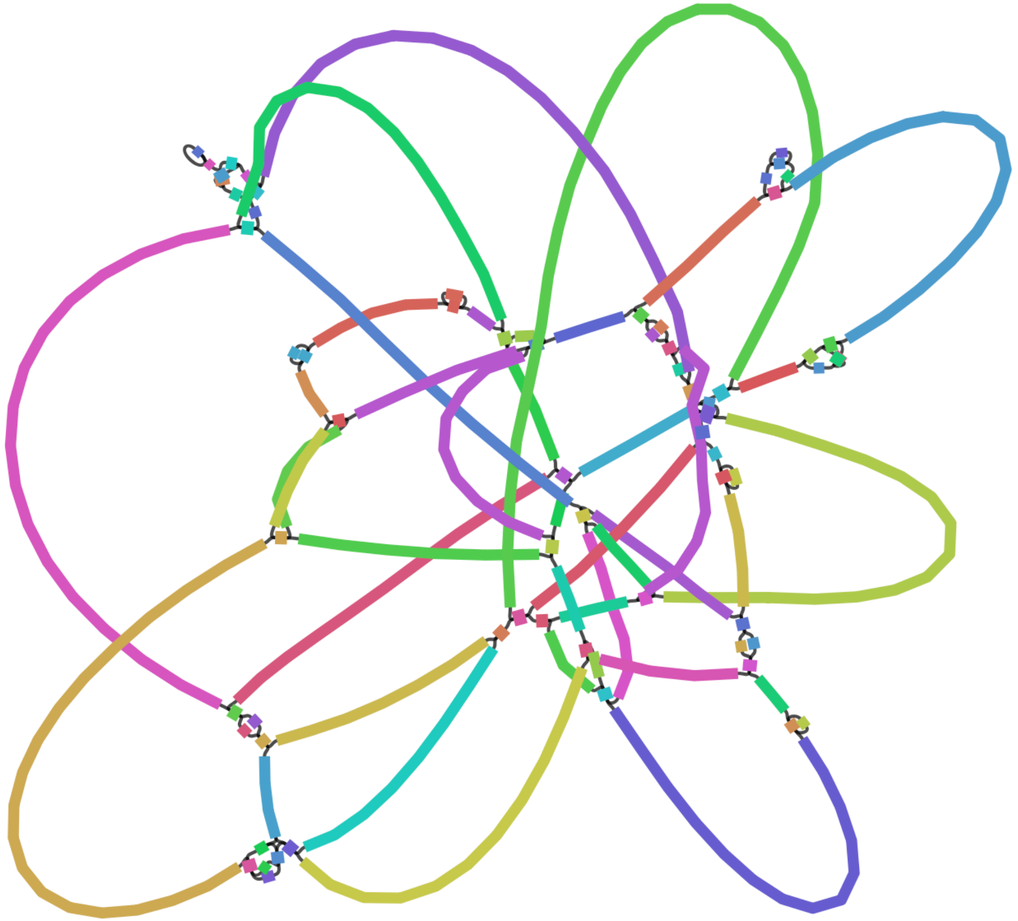
Even bacterial genomes have repeated sequences (e.g. rRNA loci, duplicated genes)



Ryan Wick, Monash University

https://github.com/rrwick/Unicycler

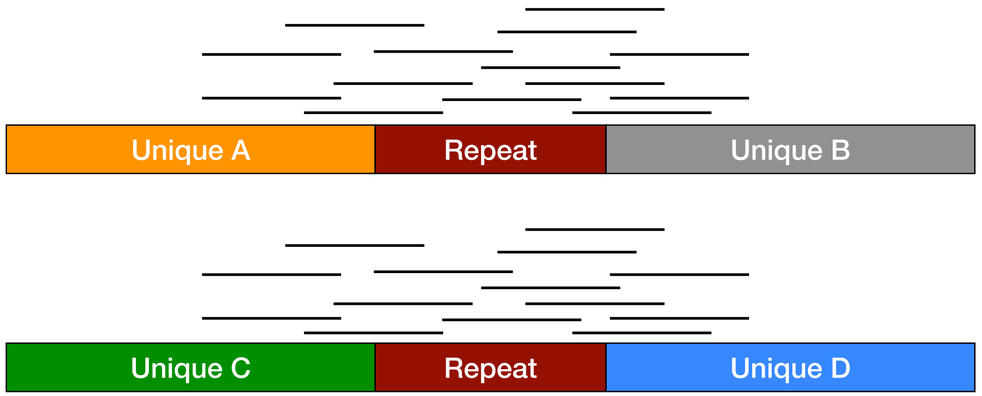
A real life graph from Illumina only data: A *Pseudomonas aeruginosa* isolate Illumina 2x250 paired end data with ~100x average depth of coverage



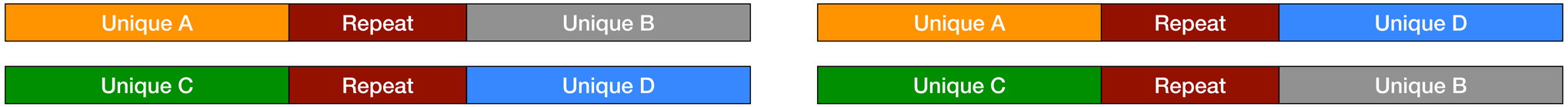
Bandage graph visualization tool
https://github.com/rrwick/Bandage

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

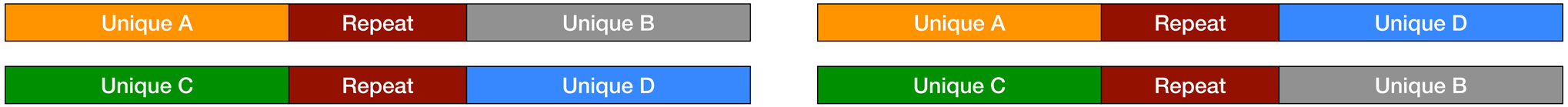


2 equally plausible combinations given short read data

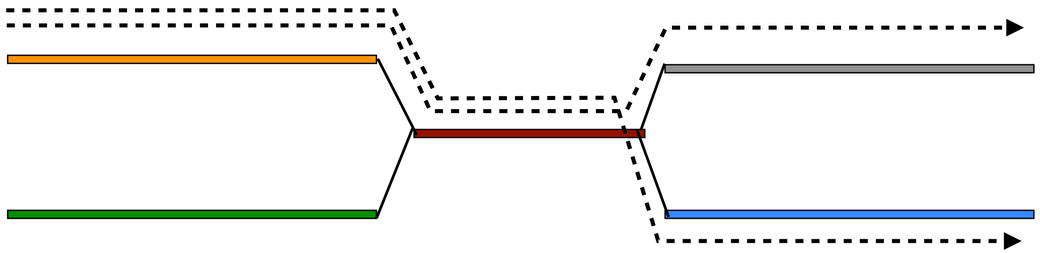


Short reads alone do not contain enough information to resolve repeats

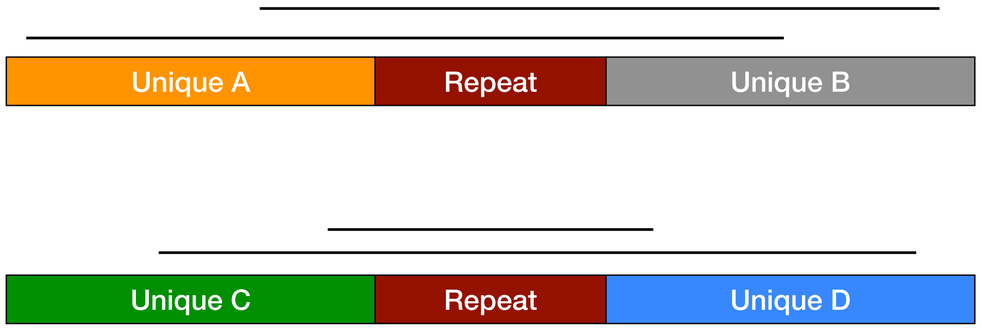
2 equally plausible combinations given short read data



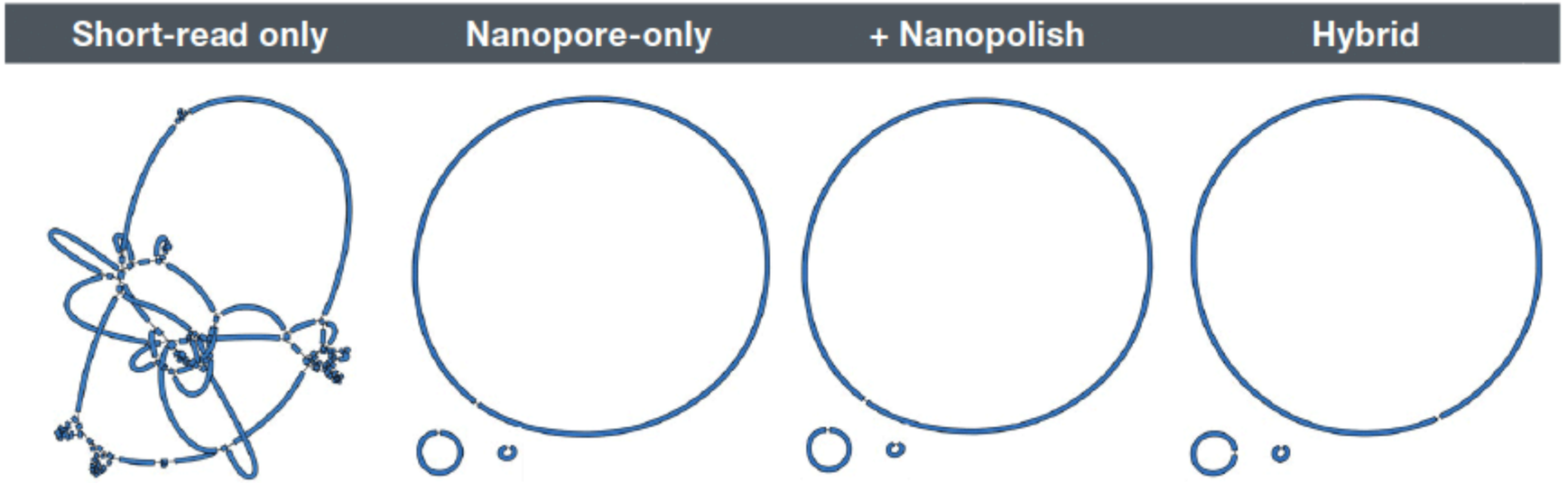
This corresponds to 2 equally plausible paths through the graph



Long reads span repeats to link unique sequence on each side



Combining long + short read produces assemblies with high contiguity and low error rate



This short-read only assembly is fragmented into many contigs

This long-read only assembly has one contig per chromosome or plasmid, but ~1% error rate

This hybrid assembly has one contig per chromosome or plasmid and very low error rate

Image: Ryan Wick, Univ. of Melbourne: https://github.com/rrwick

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?
- 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?

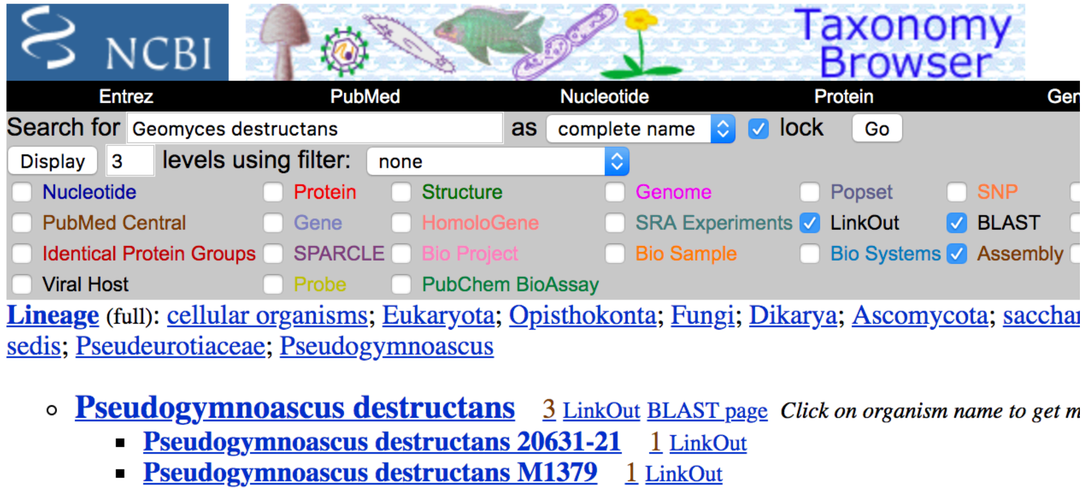
These questions apply to assemblies in databases too.

Mini exercise

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 equally difficult!

tiny ssDNA genome

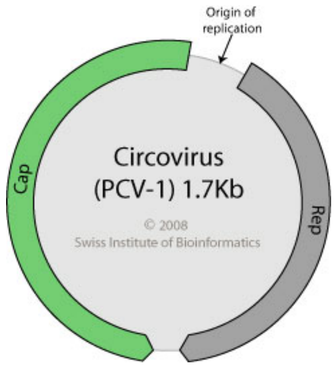
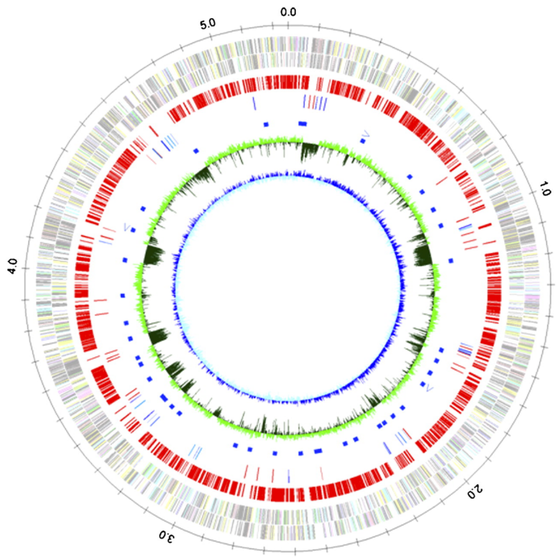


image: viralzone

bacterial genomes ~5 Mbp



Nakazawa et al (2009) Genome Research

Loblloly pine (*Pinus teada*)
22 Gbp genome!

image: Univ of Alabama