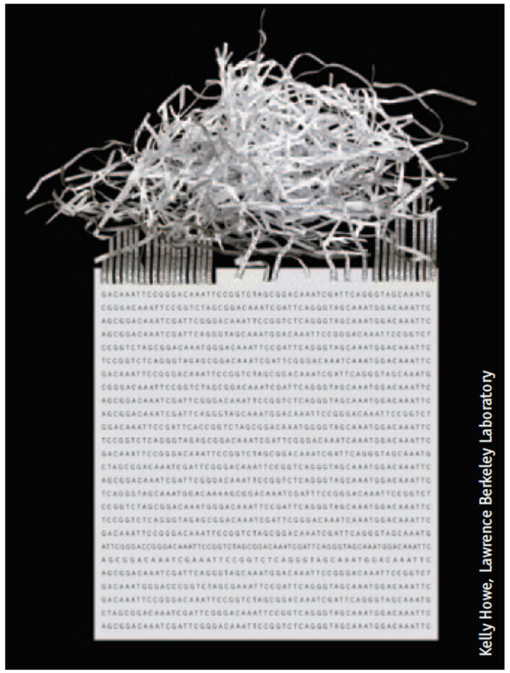
Genome assembly

Mark Stenglein, GDW Workshop

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

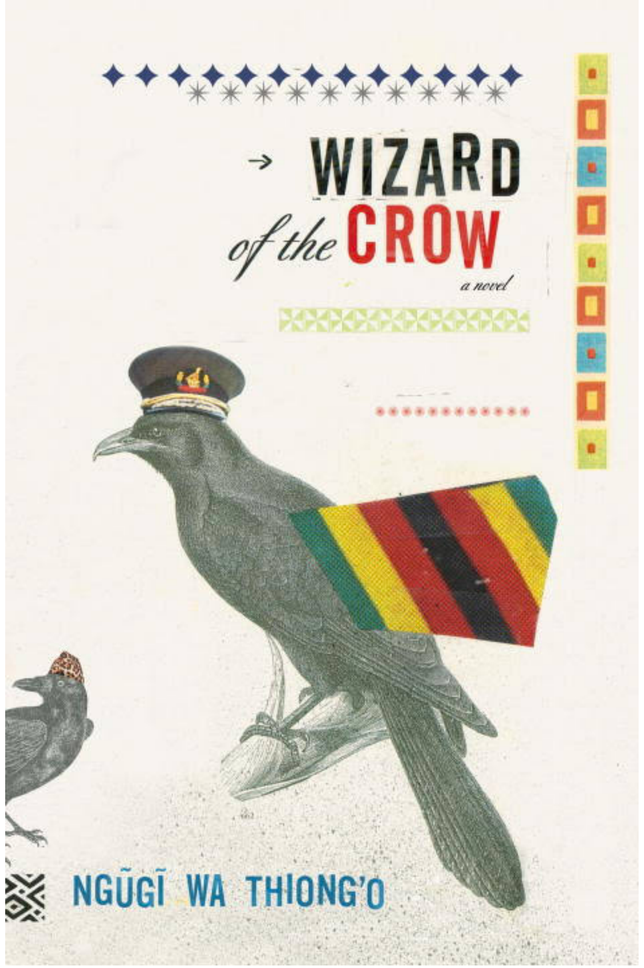
Keith Bradnam, UC Davis

Genome assembly paper exercise

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

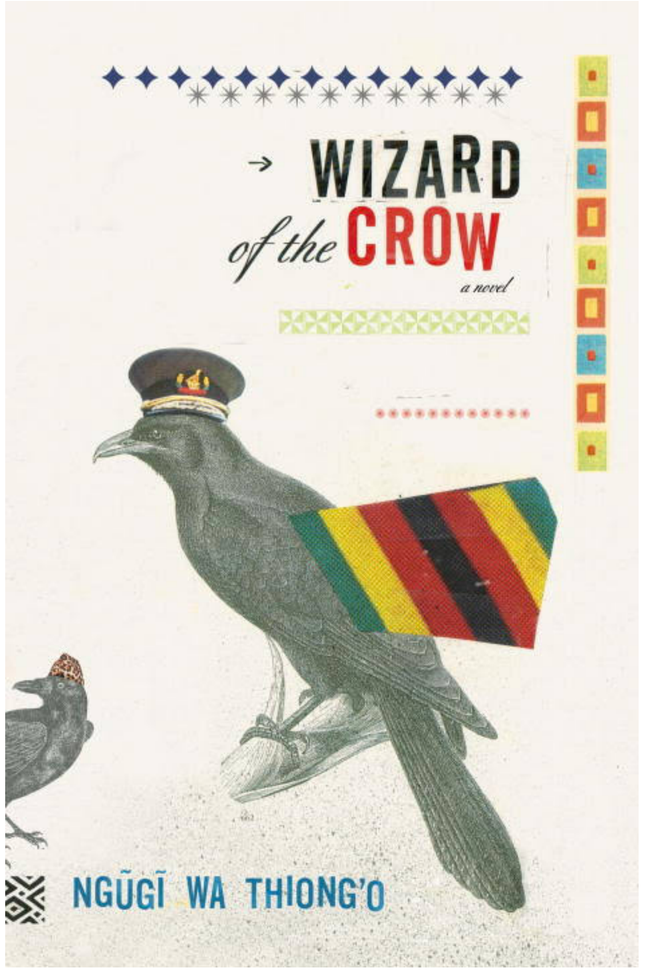
Rules/info:

- 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

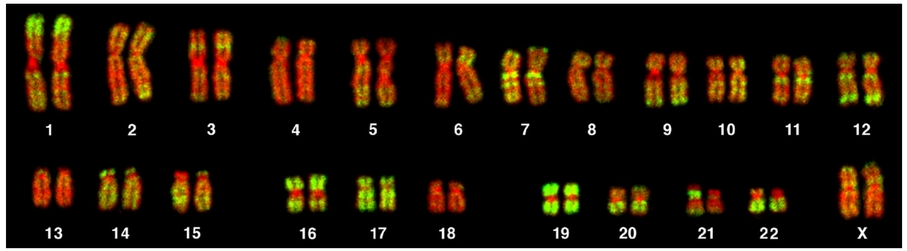
Exercise inspired and enabled by Titus Brown: http://ivory.idyll.org/blog/the-assembly-exercise.html

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

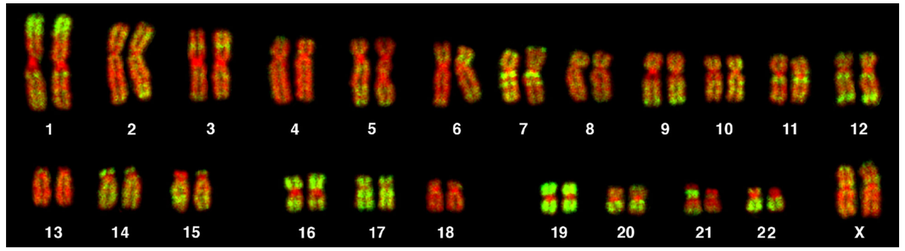


Bolzer et al (2005) PLoS Biol

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

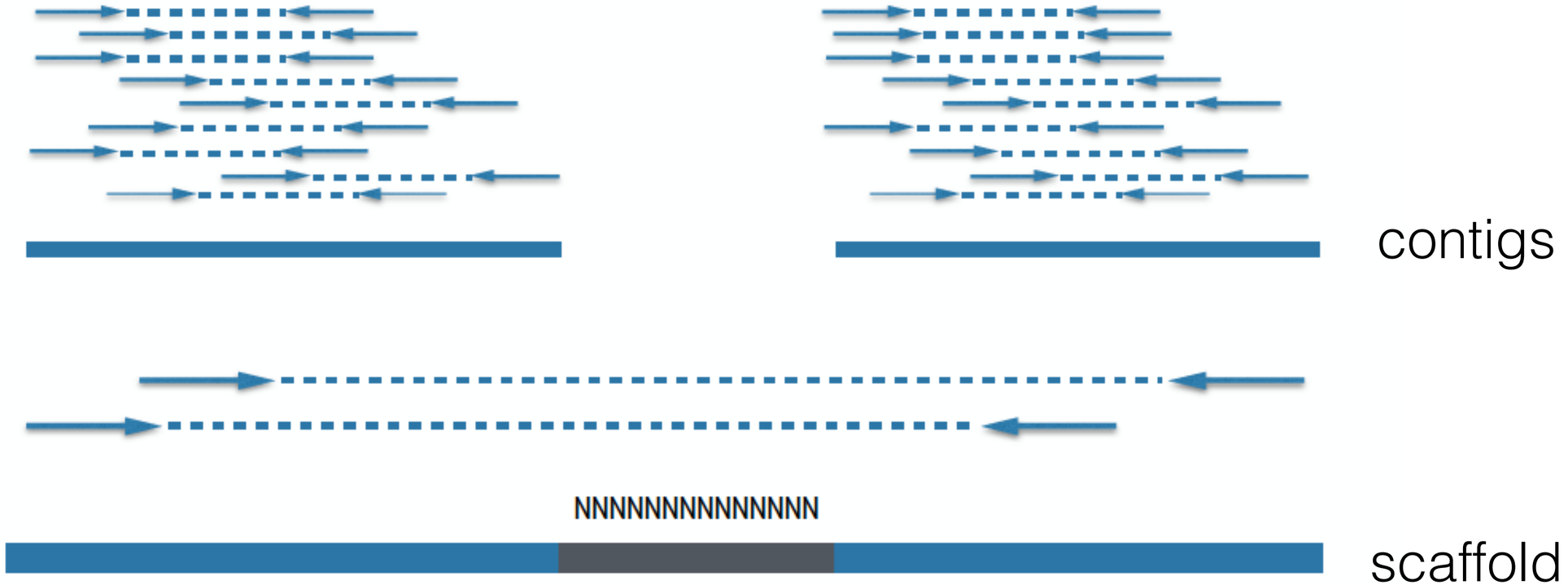


Image: *Keith Bradnam, UC Davis*

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

(We know about the golden gate bridge)

Image, analogy: *Keith Bradnam, UC Davis*

Sometimes even ‘complete’ assemblies contain gaps

human Y chromosome from GRCh38.p12

A truly complete (?) human genome

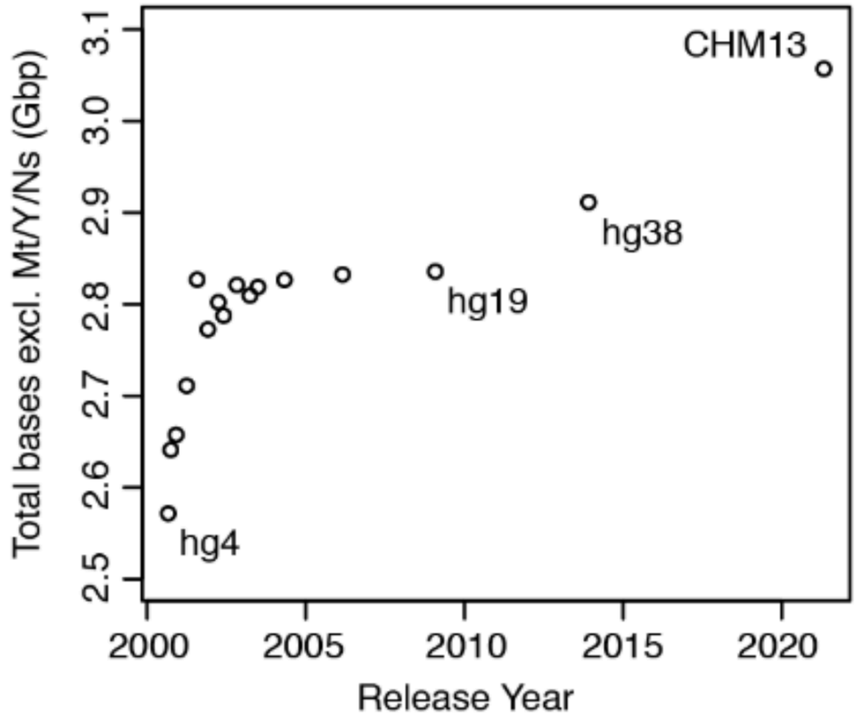
RESEARCH ARTICLE

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]*

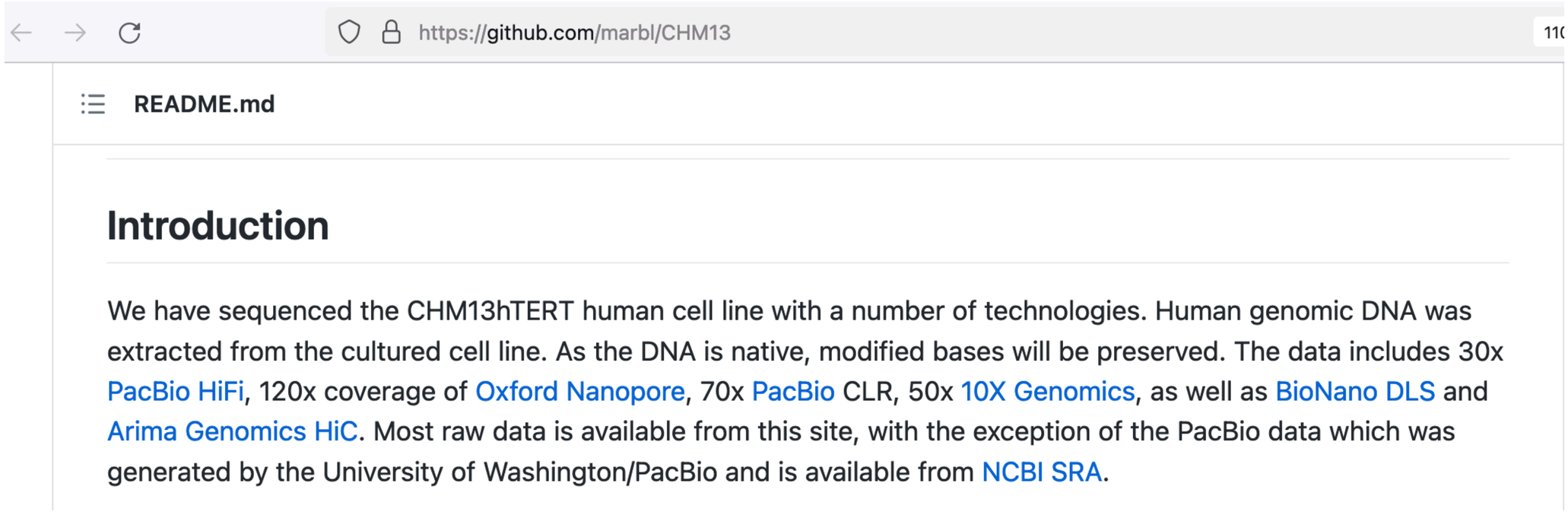
Since its initial release in 2000, the human reference genome has covered only the euchromatic fraction of the genome, leaving important heterochromatic regions unfinished. Addressing the remaining 8% of the genome, the Telomere-to-Telomere (T2T) Consortium presents a complete 3.055 billion–base pair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

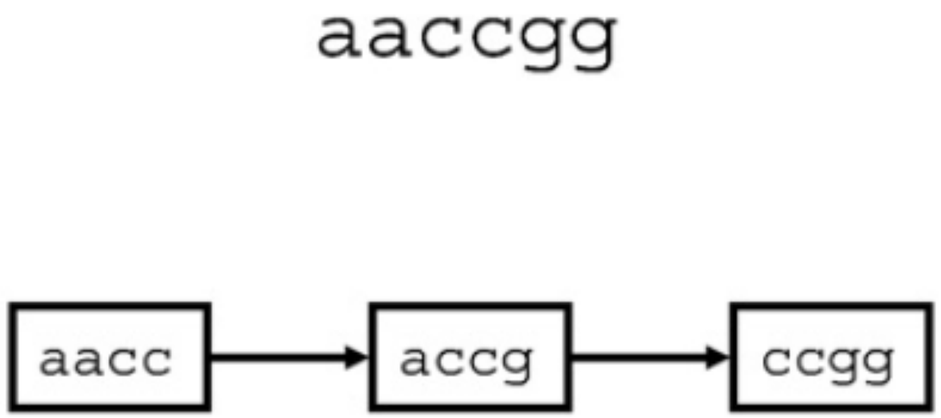
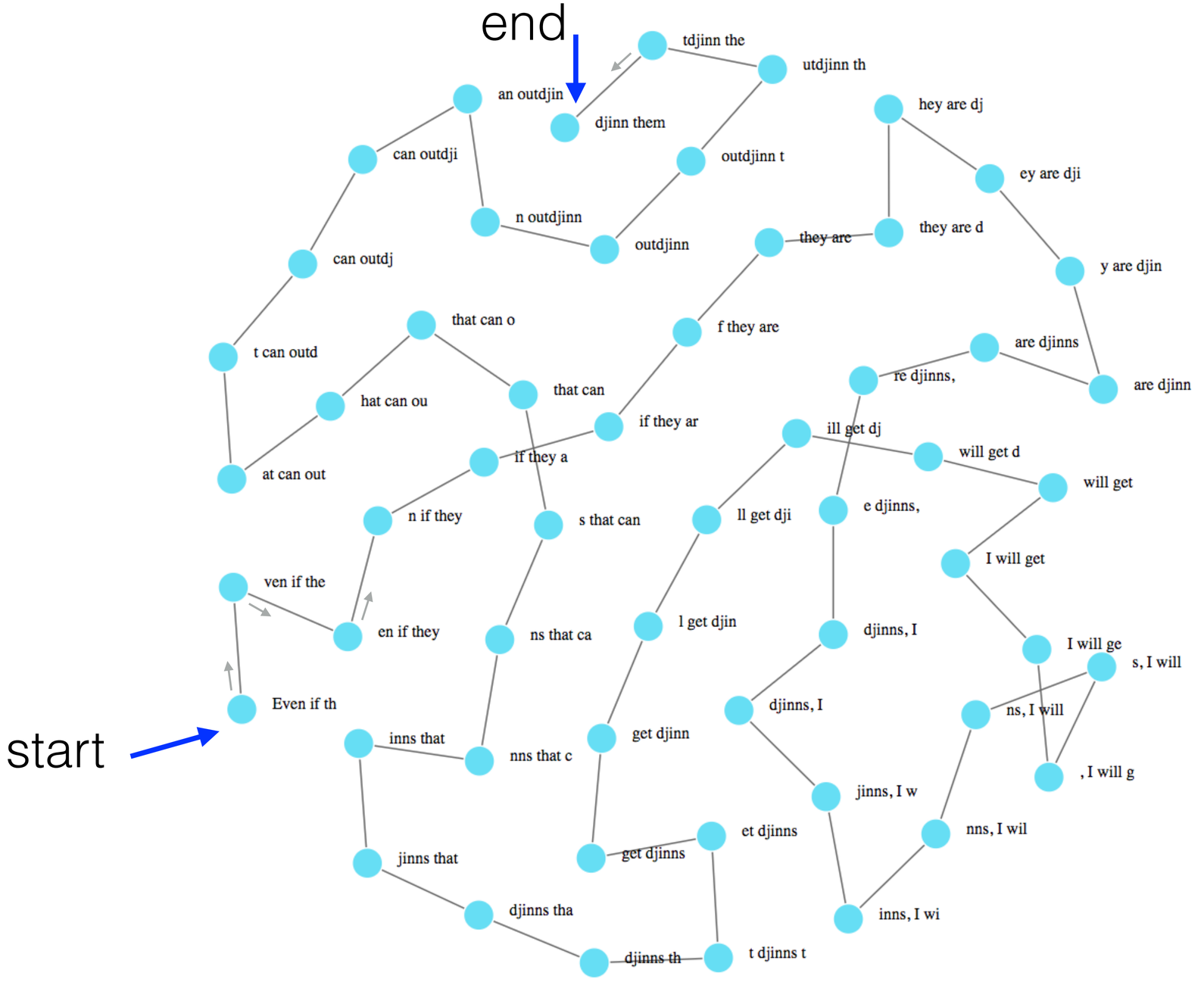


Image: *Miller, Koren, Sutton (2010) Genomics*

Generic simplified strategy:

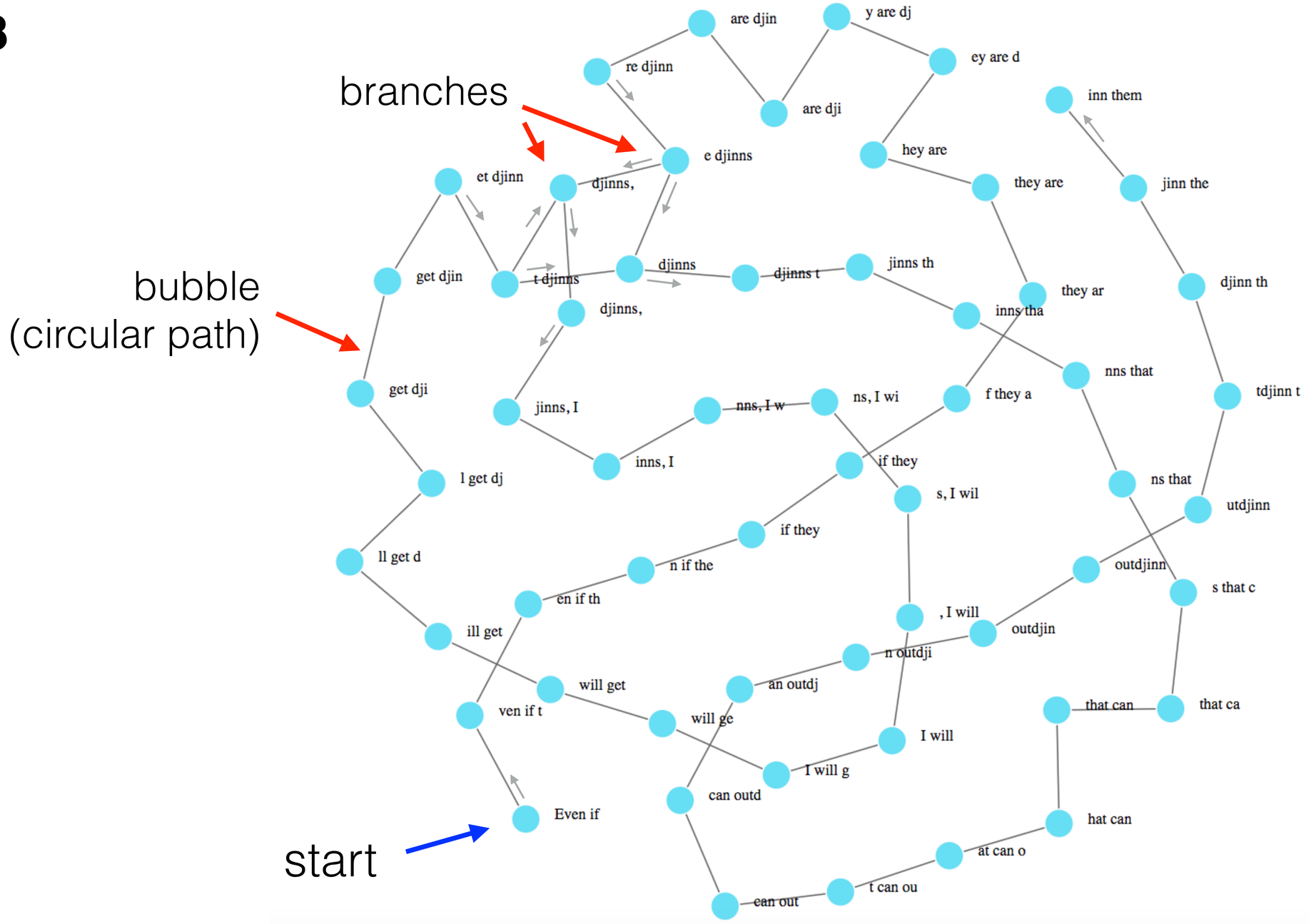
- Attempted error correction
- Break reads into overlapping k-mers (here k = 4)
- Construct de Bruijn graph of k-mers
- Trace path through graph: Tada! Genome sequence

k=10



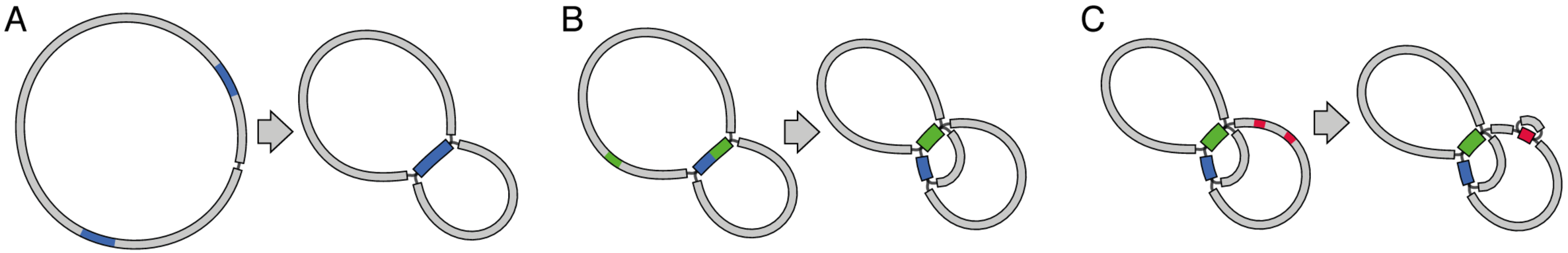
http://debruijn.herokuapp.com/graph

k=8



http://debruijn.herokuapp.com/graph

The impact of additional repeats on graph complexity

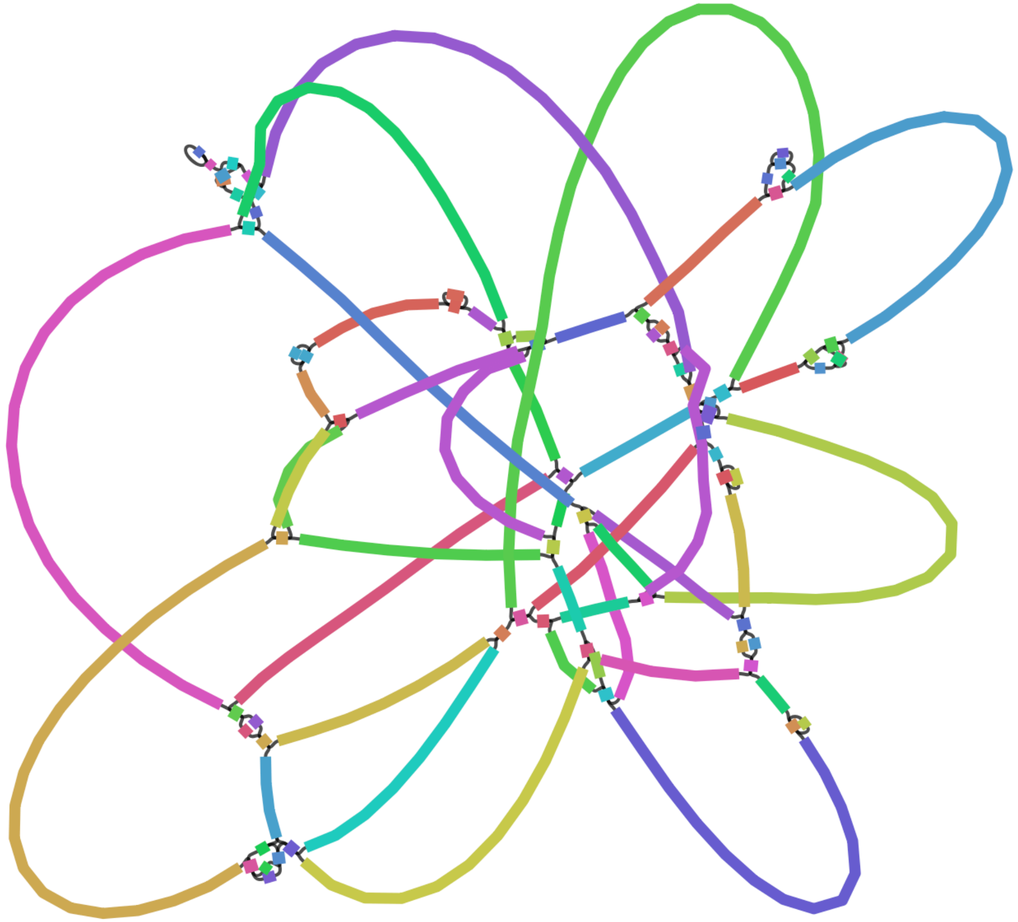


Ryan Wick, Monash University

https://github.com/rrwick/Unicycler

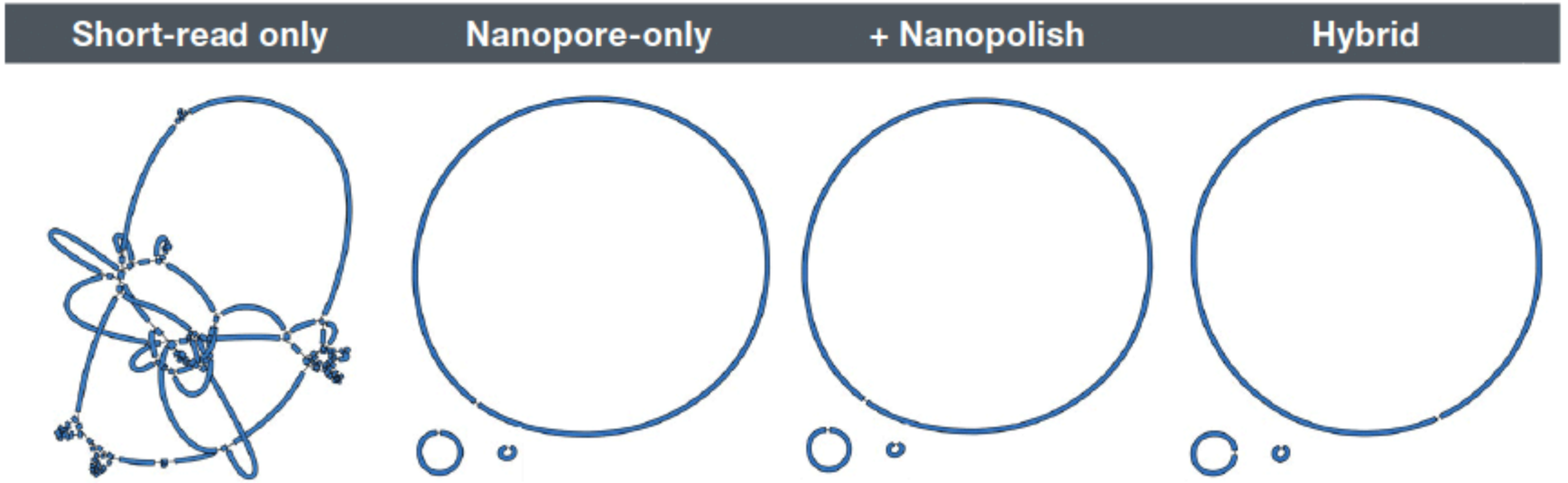
A real life graph from Illumina only data: A *Pseudomonas aeruginosa* isolate Illumina 2x250 paired end data with ~100x average depth of coverage

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



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

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?

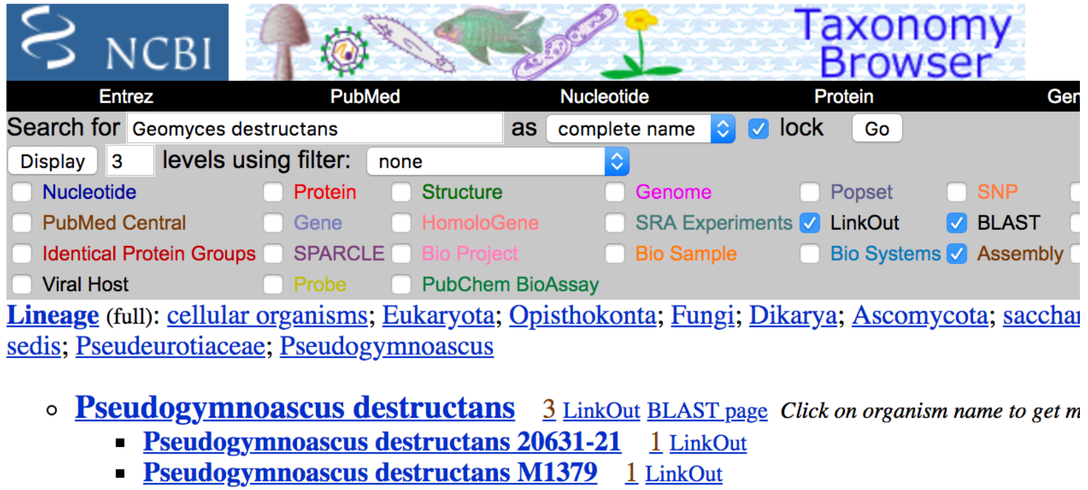
These questions apply to assemblies in databases too.

Assemblies 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

I’m painting a bleak picture, but don’t be too intimidated: genome sequencing and assembly *is* possible.

Not all assembly problems are equally difficult!

tiny ssDNA genome

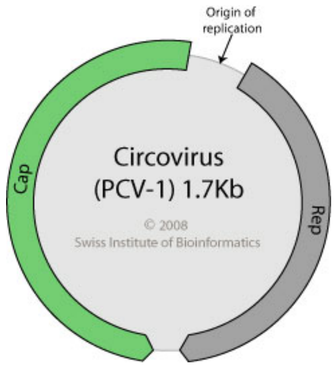
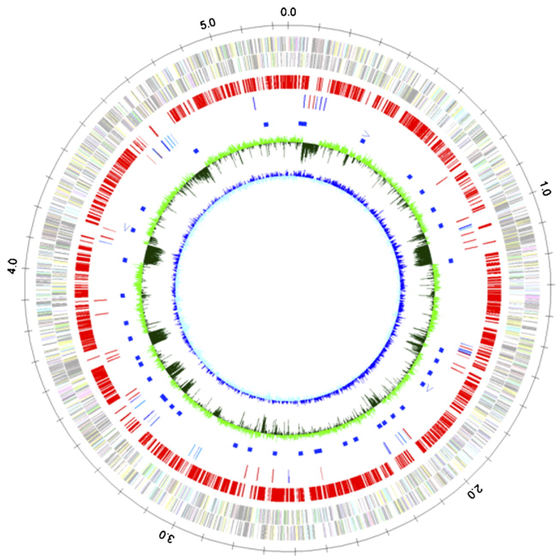


image: viralzone

bacterial genomes ~5 Mbp



Nakazawa et al (2009) Genome Research

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

image: Univ of Alabama

Also: most of the time, you don’t need a perfect assembly to answer your biological question

Reading what others have done is a great way to figure out what *you* could do

MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2016) **16**, 314–324 doi: 10.1111/1755-0998.12443

The *de novo* genome assembly and annotation of a female domestic dromedary of North African origin

ROBERT R. FITAK,*[1] ELMIRA MOHANDESAN,* JUKKA CORANDER† and PAMELA A. BURGER*
**Institut für Populationsgenetik, Vetmeduni Vienna, Veterinärplatz 1, Vienna 1210, Austria, †Department of Mathematics and Statistics, University of Helsinki, Helsinki, FIN-0014, Finland*

You could call these ‘bioinformatics protocols’

mina, USA). Preprocessing of the sequence reads included the removal of adapter sequences and removal of reads with >10% uncalled bases and/or >50% of bases with a Phred-scaled quality score <4. After preprocessing, all 100-bp (paired-end) and 50-bp (mate-pair) reads were retained as the set of ‘raw’ reads. We trimmed the 3′ end of all raw reads using a modified Mott algorithm in POPOOLATION v1.2.2 (Kofler *et al*. 2011) to a minimum quality score of 20 and a minimum length threshold of 50 bp and 30 bp for the paired-end and mate-pair reads, respectively.

We corrected the trimmed, paired-end reads for substitution sequencing errors using QUAKE v0.3.5 (Kelley *et al*. 2010). Salzberg *et al*. (2012) showed previously that the error correction of sequencing reads can greatly improve the *de novo* assembly of genomes, including genomes assembled using the program ABYSS (Simpson *et al*. 2009). QUAKE uses the distributions of infrequent and abundant *k*-mers to model the nucleotide error rates and subsequently corrects substitution errors. As input to QUAKE and again after error correction, we counted the frequency of 20-mers in the paired-end reads using DSK v1.6066 (Rizk *et al*. 2013). To estimate genome size, we divided the total number of error-free 20-mers by their peak coverage depth.

We assembled the genome using the trimmed and error-corrected paired-end reads with ABYSS v1.3.6. To determine the optimal *k*-mer length, we repeated the assembly using $k = 40$–88 in 8-bp increments. All scaffolding steps were performed using the trimmed mate-pair reads also in ABYSS, and only scaffolds longer than

Fitak et al (2016) Mol Ecol Resources

Read and synthesize a bunch of these like you would ‘wet lab’ protocols

4. Sequence assembly

All cleaned sequences were assembled using the Newbler Assembler (*25*) v2.6 (build version 20110517_1502) with the following parameters “-scaffold -het -large -cpu 3 -siod -noinfo”. Our decision to use Newbler was influenced by the large proportion of 454 sequences used and the ability for Newbler to handle multiple data, which allowed BACends, Illumina, and 454 data to be combined. Assemblies were run on a 16-processor node with 256 GB of RAM. Our current assembly consists of 43,234 contigs with an average size of 15,456 bp (min= 436 bp; max=287,935 bp), an N50 size of 29,456 bp, and an N50 count of 6,448. Scaffolding by virtue of the cleaned paired-end reads resulted in 5,745 scaffolds, with an average size of 123 kb (min= 1,732 bp; max= 15.98 Mb), an N50 size of 4.93 Mb, and an N50 count of 50. Based on the N90 statistics, 90% of our assembled sequence resides within 155 scaffolds, each of which is 1.16 Mb

Chamala et al (2016) Science

Bioinformatics protocols are analogous to any lab protocol

mina, USA). Preprocessing of the sequence reads included the removal of adapter sequences and removal of reads with >10% uncalled bases and/or >50% of bases with a Phred-scaled quality score <4. After preprocessing, all 100-bp (paired-end) and 50-bp (mate-pair) reads were retained as the set of 'raw' reads. We trimmed the 3′ end of all raw reads using a modified Mott algorithm in POPOOLATION v1.2.2 (Kofler *et al*. 2011) to a minimum quality score of 20 and a minimum length threshold of 50 bp and 30 bp for the paired-end and mate-pair reads, respectively.

We corrected the trimmed, paired-end reads for substitution sequencing errors using QUAKE v0.3.5 (Kelley *et al*. 2010). Salzberg *et al*. (2012) showed previously that the error correction of sequencing reads can greatly improve the *de novo* assembly of genomes, including genomes assembled using the program ABYSS (Simpson *et al*. 2009). QUAKE uses the distributions of infrequent and abundant *k*-mers to model the nucleotide error rates and subsequently corrects substitution errors. As input to QUAKE and again after error correction, we counted the frequency of 20-mers in the paired-end reads using DSK v1.6066 (Rizk *et al*. 2013). To estimate genome size, we divided the total number of error-free 20-mers by their peak coverage depth.

We assembled the genome using the trimmed and error-corrected paired-end reads with ABYSS v1.3.6. To determine the optimal *k*-mer length, we repeated the assembly using $k = 40–88$ in 8-bp increments. All scaffolding steps were performed using the trimmed mate-pair reads also in ABYSS, and only scaffolds longer than

Fitak et al (2016) Mol Ecol Resources

Cells were analyzed using a Cell Lab Quanta SC flow cytometer (Beckman Coulter). CD14-positive cells were stained with CD14-FITC (Miltenyi Biotec). Cells were incubated with propidium iodide to assess cell viability.

Immunoblotting and antibodies. Cells were harvested and total protein extracted in a buffer containing 25 mM HEPES (pH 7.4), 10% glycerol, 150 mM NaCl, 0.5% Triton X-100, 1 mM EDTA, 1 mM $MgCl_2$, 1 mM $ZnCl_2$, and protease inhibitors. The extracts were clarified by centrifugation for 10 minutes at 20,800g at 4˚C. The extracted proteins (15 µg) were fractionated by SDS-PAGE, transferred to a polyvinylidene difluoride membrane (Millipore), and probed with an anti-A3A polyclonal antiserum, an anti-GFP monoclonal antibody (Clontech), or an anti-eEF1alpha monoclonal antibody (Upstate). The anti-A3A polyclonal serum was generated by immunizing a rabbit with a peptide corresponding to A3A residues 171-199 (CPFQPWDGLEEHSQALSGRLRAILQNQGN) mixed with TiterMax Gold adjuvant (Sigma). Primary antibodies were detected by incubation with fluorescently labeled secondary antibodies and imaging on an Odyssey imaging device (LI-COR Biosciences).

DNA cytidine deaminase activity assays. PBMC or transfected HEK-293T cell lysates were prepared as above for immunoblotting. The deaminase activity in the lysates was determined using a FRET-based assay essentially as described[59]. Briefly, serial dilutions of lysates were incubated for 2h at 37˚C with a DNA oligonucleotide 5'-(6-FAM)-AAA-TTC-TAA-TAG-ATA-ATG-TGA-(TAMRA). FRET occurs between the fluorophores, decreasing FAM fluorescence. If

Questions?

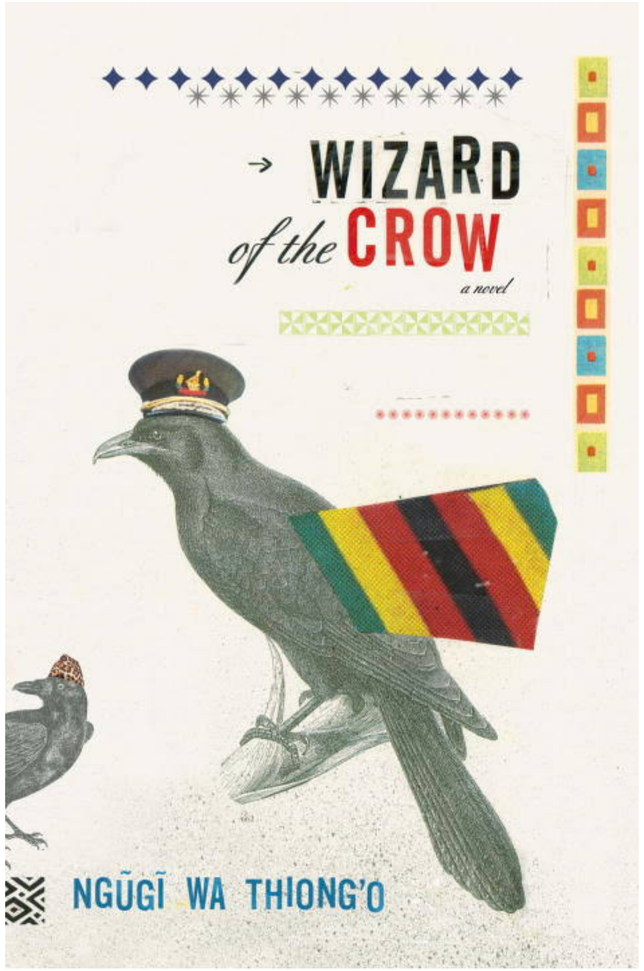


Image: *Keith Bradnam, UC Davis*