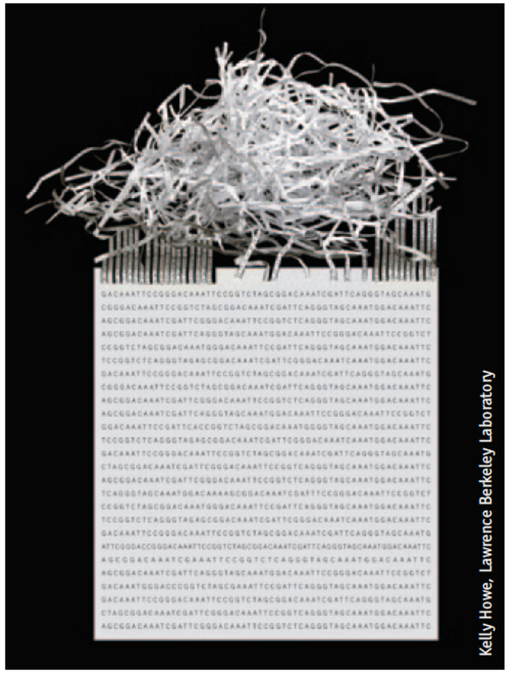
Genome assembly

Mark Stenglein, GDW 2017

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

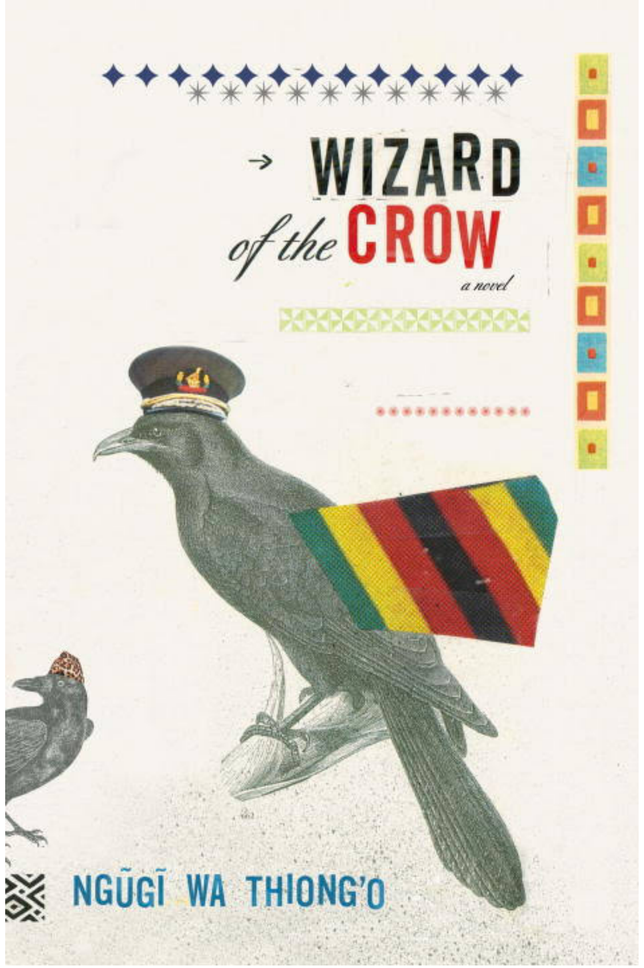
Keith Bradnam, UC Davis

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…

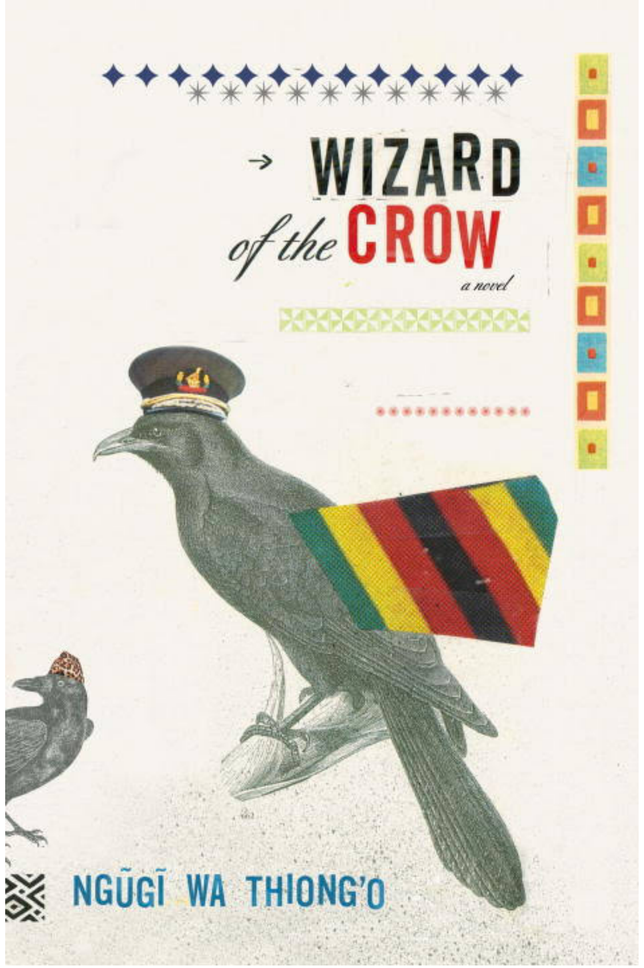


Exercise inspired and enabled by Titus Brown: http://ivory.idyll.org/blog/the-assembly-exercise.html

Genome assembly paper exercise

“Even if they are djinns, I will get djinns that can outdjinn them.”

Ngugi wa Thiong’o, Wizard of the Crow



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

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

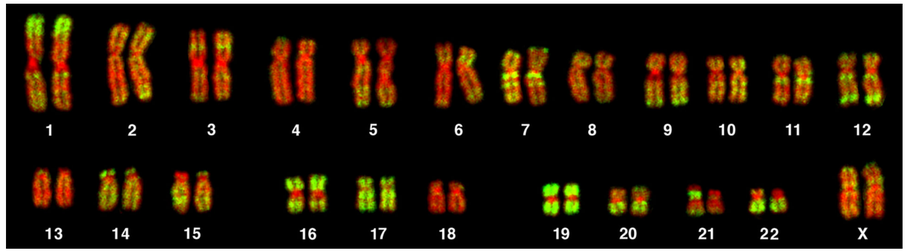
Exercise inspired and enabled by Titus Brown: http://ivory.idyll.org/blog/the-assembly-exercise.html

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

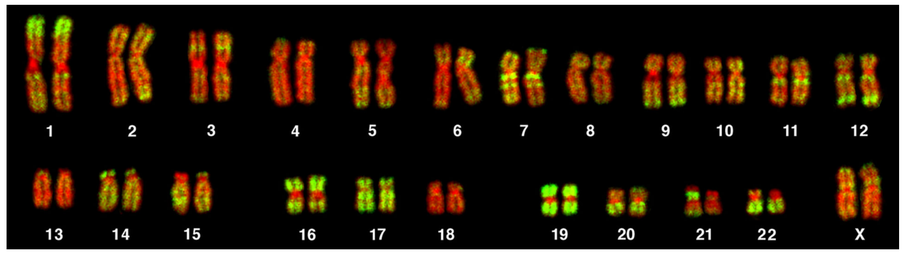


Bolzer et al (2005) PLoS Biol

Some of the main reasons that assembly is difficult

1) Genomes are chock full of repetitive sequences

Alu sequences in the human genome



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

‘Reference-guided assembly’ is a slightly different, easier problem analogous to knowing what the puzzle should generally look like

Images, metaphor: *Keith Bradnam, UC Davis*

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

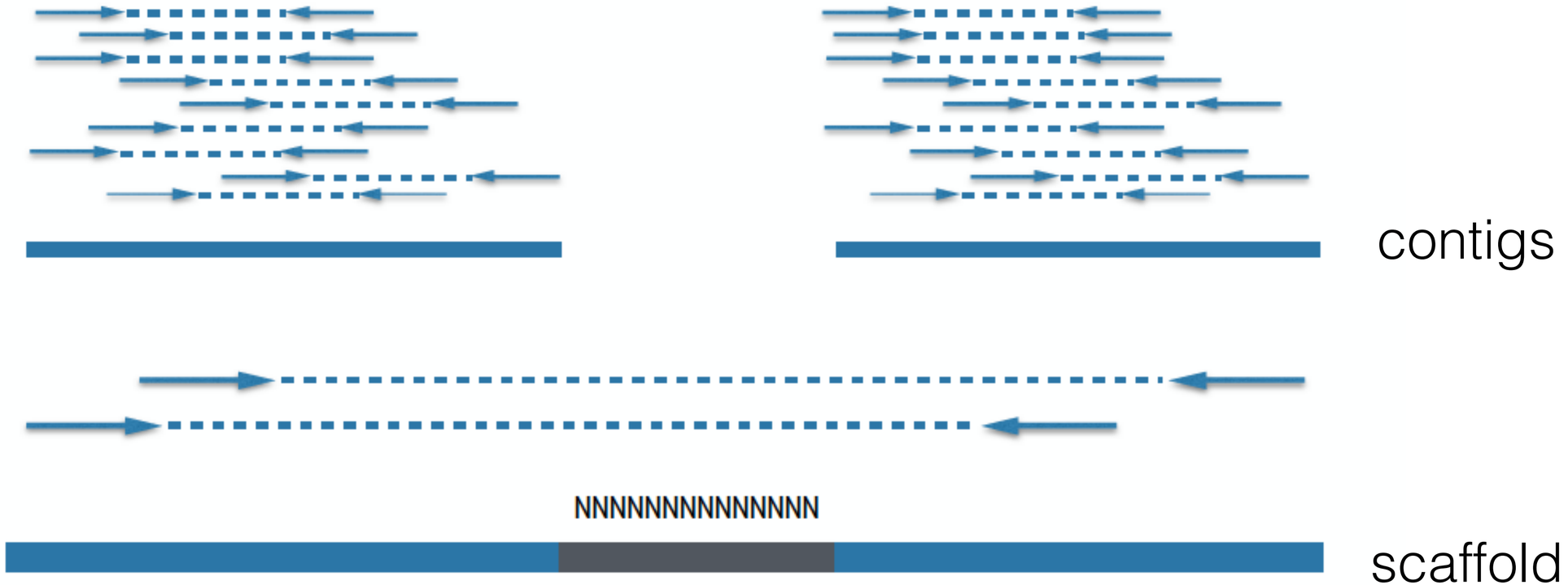
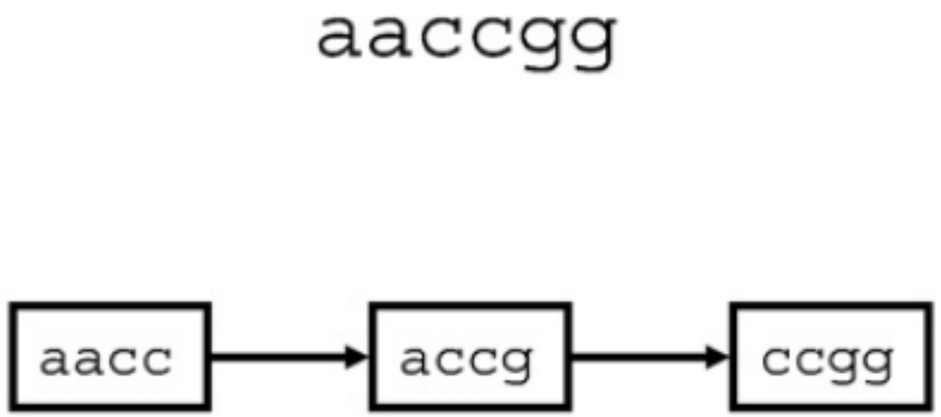


Image: *Keith Bradnam, UC Davis*

These “contigs” could be scaffolded

Image: *Keith Bradnam, UC Davis*

Nearly all assemblers use a de Bruijn graph-based algorithm

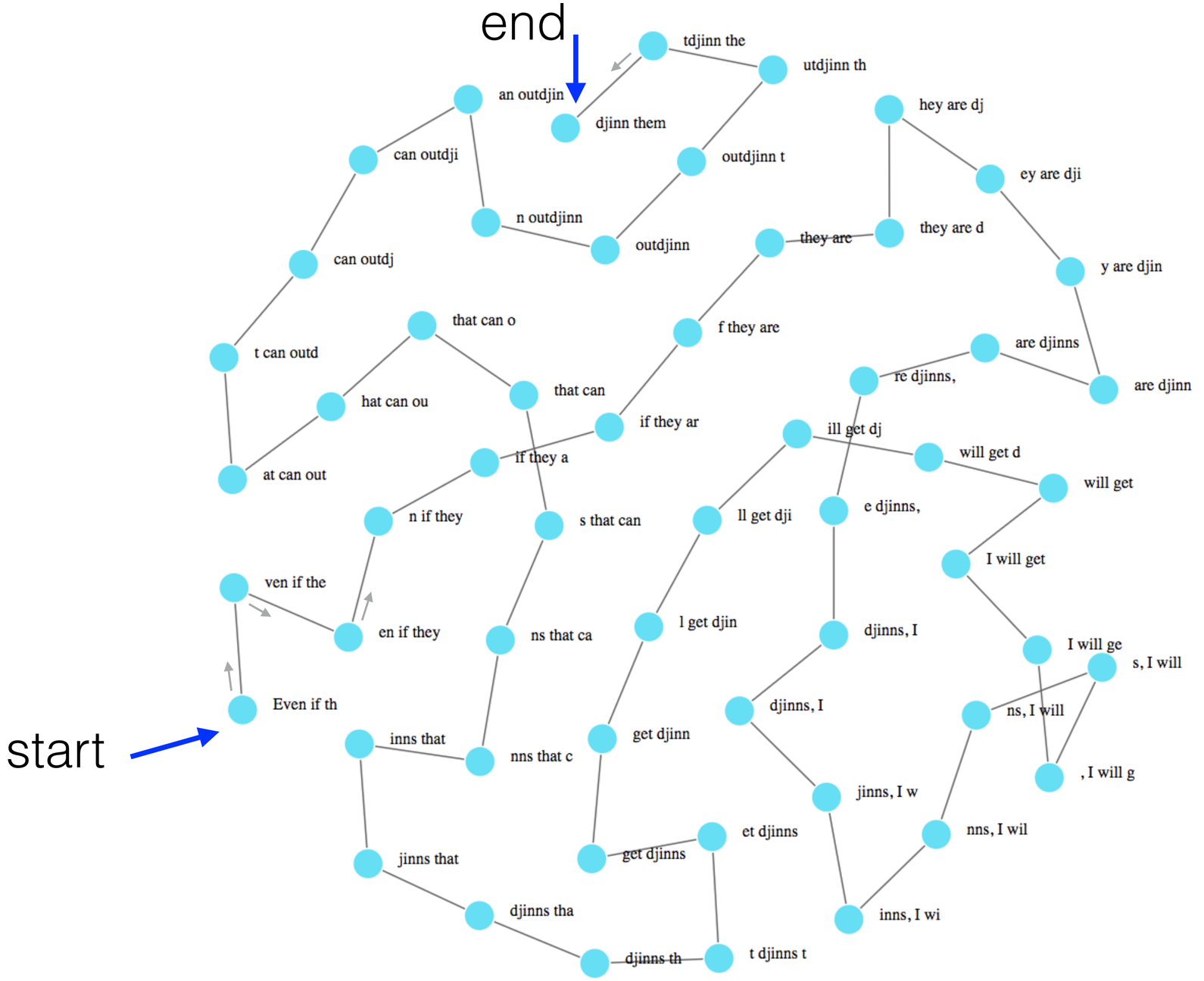


Generic simplified strategy:

- Attempted error correction
- Break reads into overlapping-by-1 ‘k-mers’ (here k = 4)
- Construct de Bruijn graph of k-mers
- Trace path through graph: Tada! Genome sequence

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

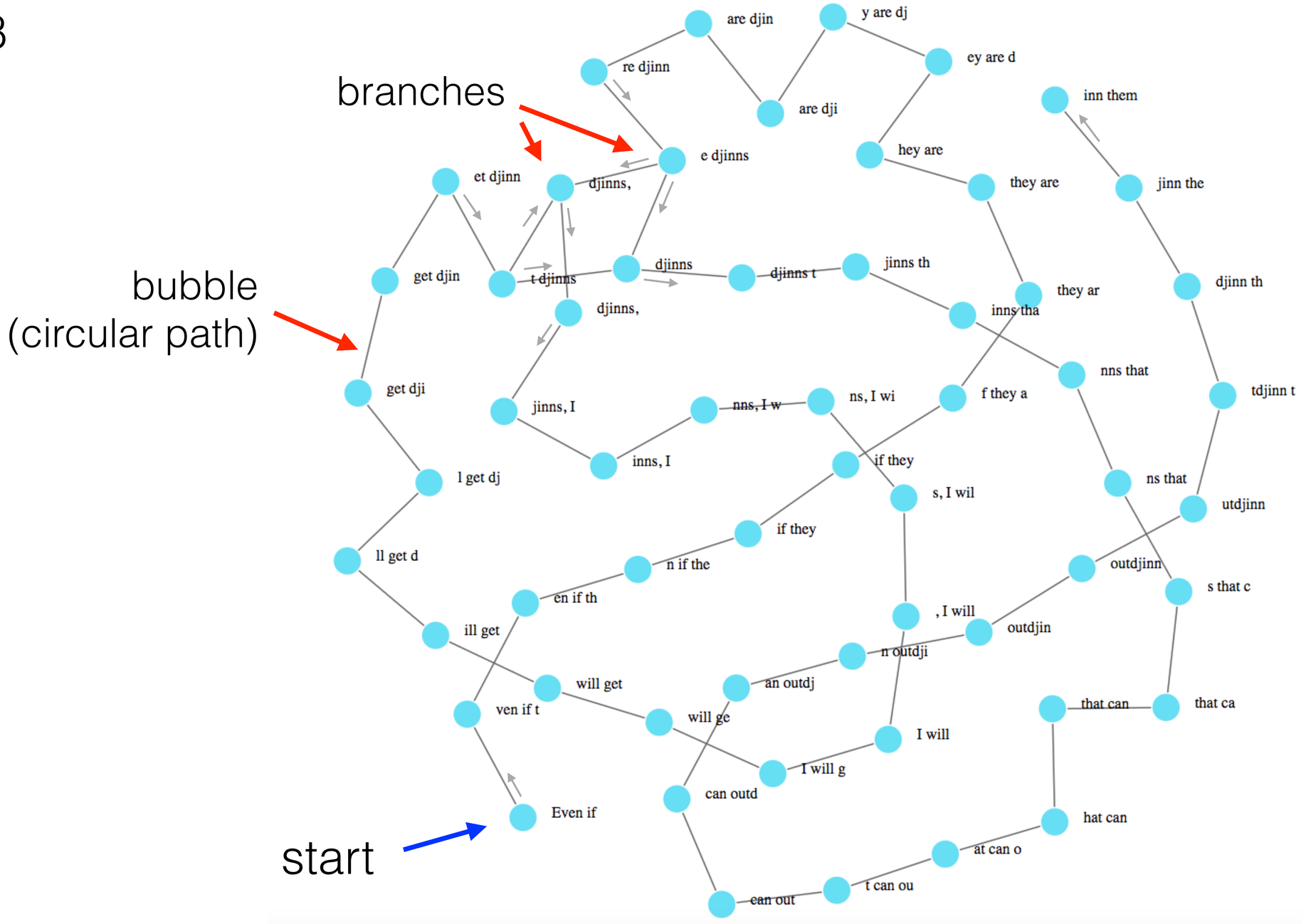
k=10



http://debruijn.herokuapp.com/graph

Even if they are djinns, I will get djinns that can outdjinn them

k=8



http://debruijn.herokuapp.com/graph

Even if they are djinns, I will get djinns that can outdjinn them

Assemblers use a variety of strategies to try to resolve graph complexity

To read more about these strategies:

- Miller JR, Koren S, Sutton G. Assembly algorithms for nextg eneration sequencing data. Genomics 2010;95:315–27.
- Compeau PE, Pevzner PA, Tesler G. How to apply de Bruijn graphs to genome assembly. Nat Biotechnol 2011;29:987–91.
- Nagarajan N, Pop M. Sequence assembly demystified. Nat Rev Genet 2013;14:157–67.
- Sohn JI, Nam JW. The present and future of de novo whole-genome assembly. Brief Bioinform. 2016 Oct 14. pii: bbw096.

Note that the as long read sequencing continues to improve and gain ground, these issues may become moot.

Assemblies that mix long and short reads are called ‘hybrid’ assemblies, and they are increasingly the norm.

How do you know if your assembly is any 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?)

The Assemblathon: a de novo assembly competition

THE ASSEMBLATHON

Bradnam *et al. GigaScience* 2013, **2**:10
http://www.gigasciencejournal.com/content/2/1/10

(GIGA)^n SCIENCE

RESEARCH Open Access

Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species

Keith R Bradnam[1*†], Joseph N Fass[1†], Anton Alexandrov[36], Paul Baranay[2], Michael Bechner[39], Inanç Birol[33],

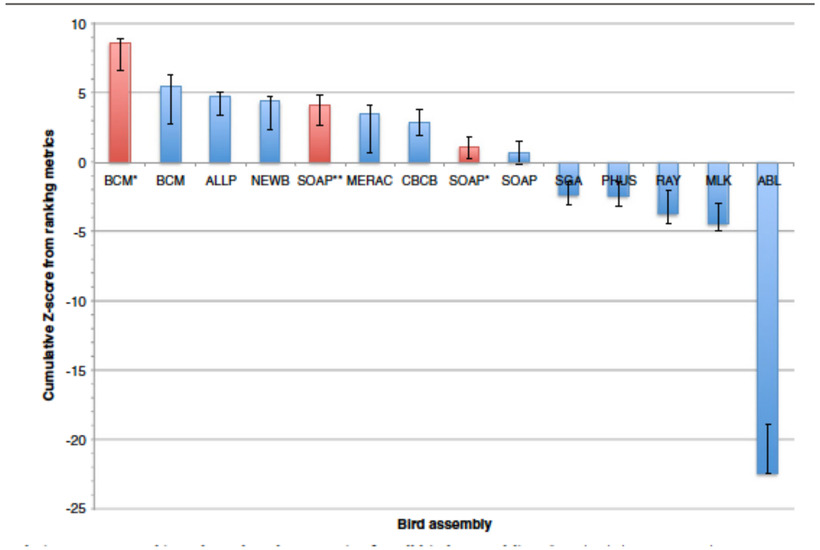
Melopsittacus undulatus (budgie)

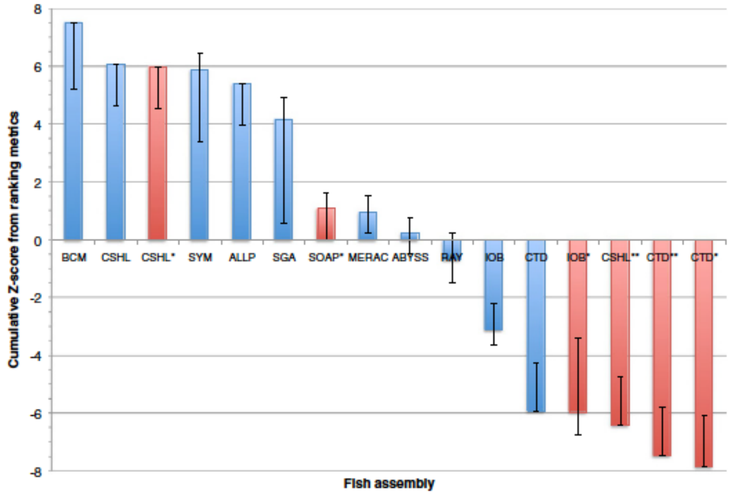
Maylandia zebra (zebra mbuna)

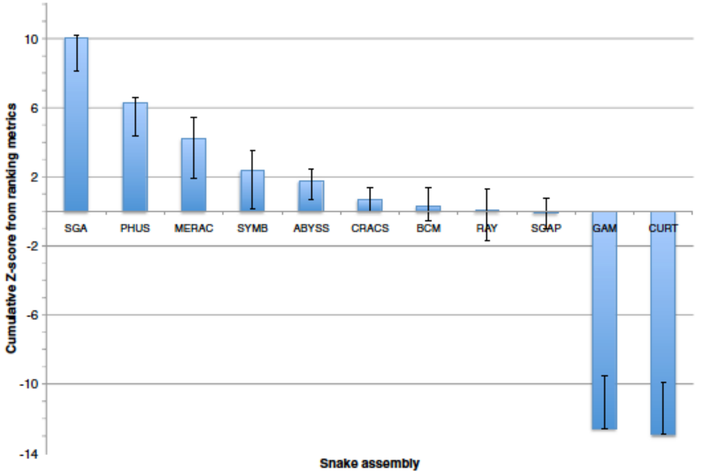
Boa constrictor (boa constrictor)

Images: *wikipedia*

There was not one obviously best assembler







Melopsittacus undulatus (budgie)

Maylandia zebra (zebra mbuna)

Boa constrictor (boa constrictor)

Lessons learned from assemblathon 2

The clear take-home message from this exercise is the lack of consistency between assemblies in terms of interspecific as well as intraspecific comparisons. An assembler may produce an excellent assembly when judged by one approach, but a much poorer assembly when judged by another. The SGA snake assembly ranked 1st overall, but only ranked 1st in one individual key metric, and ranked 5th and 7th in others. Even when an assembler performs well across a range of metrics in one species, it is no guarantee that this assembler will work as well with a different genome. The BCM-HGSC team produced the top ranking assembly for bird and fish, but a much lower-ranked assembly for snake. Comparisons between the performance of the same assembler in different species are confounded by the different nature of the input sequence data that was provided for each species.

Lack of consistency:

- No assembler did the best for all 3 species.
- Individual assemblers did better on some metrics than on others
  - e.g.: an assembler might produce longer scaffolds, but with more errors

Assembler performance depended on:

- Characteristics of genome: repeat content, size, etc.
- Characteristics of the input data: read length, coverage, error rates

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

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*

‘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

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

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

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?

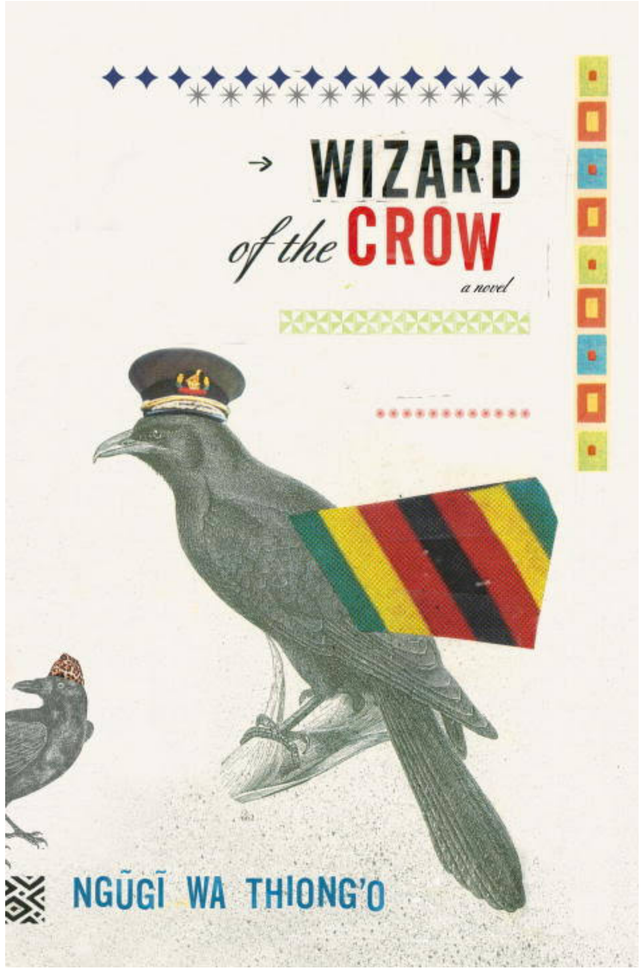


Image: *Keith Bradnam, UC Davis*