

Genome Assembly (pt2)

Michael Schatz

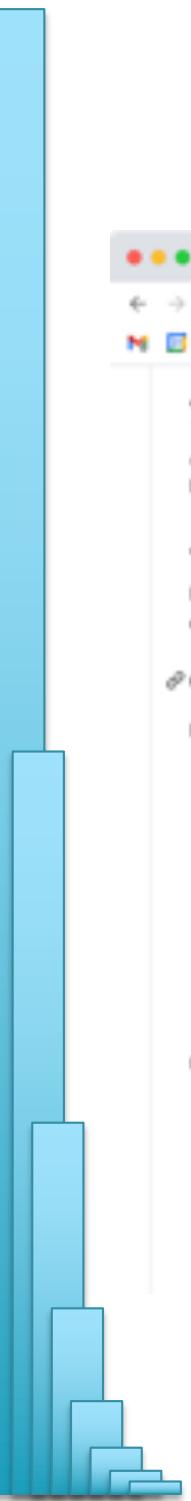
Feb 3, 2021

Lecture 4: Applied Comparative Genomics



Assignment 1: Chromosome Structures

Due Feb 3 @ 11:59pm



A screenshot of a web browser window showing the assignment details. The title "Assignment 1: Chromosome Structures" is at the top. Below it, the assignment date is listed as "Assignment Date: Wednesday, Jan 27, 2021" and the due date as "Due Date: Wednesday, Feb. 3, 2021 @ 11:59pm". A section titled "Assignment Overview" contains text about the assignment's purpose and a reminder to post questions to Piazza. A section titled "Question 1: Chromosome structures" lists eight species with their genome size files and a note to make a table with specific information per species.

Assignment Date: Wednesday, Jan 27, 2021
Due Date: Wednesday, Feb. 3, 2021 @ 11:59pm

Assignment Overview

In this assignment you will profile the overall structure of the genomes of several important species and then study human chromosome 22 in more detail. As a reminder, any questions about the assignment should be posted to [Piazza](#).

Question 1: Chromosome structures

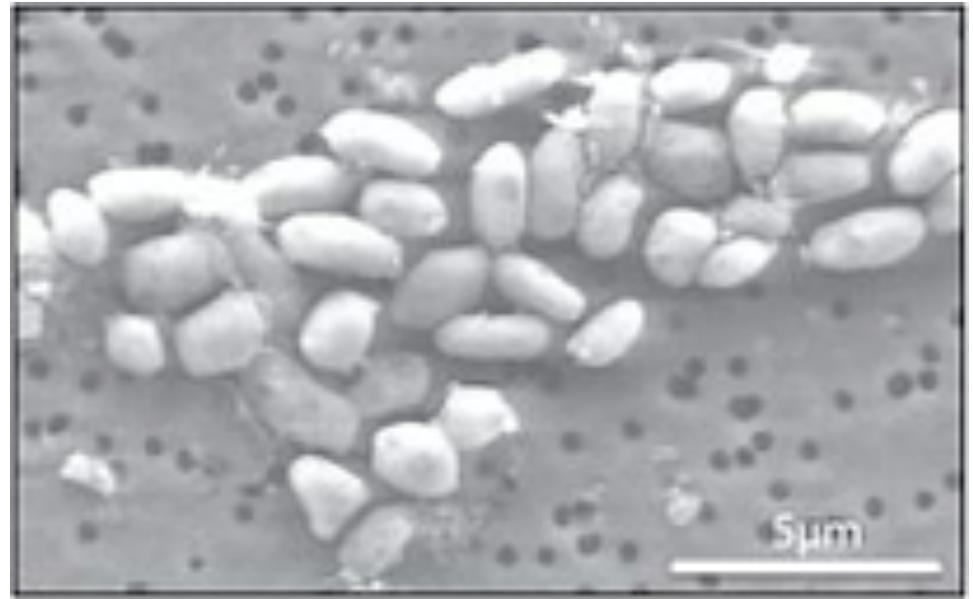
Download the chromosome size files for the following genomes (Note these have been preprocessed to only include main chromosomes):

1. [Arabidopsis thaliana \(TAIR10\)](#) - An important plant model species [\[info\]](#)
2. [Tomato \(Solanum lycopersicum v4.00\)](#) - One of the most important food crops [\[info\]](#)
3. [E. coli \(Escherichia coli K12\)](#) - One of the most commonly studied bacteria [\[info\]](#)
4. [Fruit Fly \(Drosophila melanogaster, dm3\)](#) - One of the most important model species for genetics [\[info\]](#)
5. [Human \(hg38\) - us::\)](#) [\[info\]](#)
6. [Wheat \(Triticum aestivum, IWGSC\)](#) - The food crop which takes up the largest land area [\[info\]](#)
7. [Worm \(Caenorhabditis elegans, ce10\)](#) - One of the most important animal model species [\[info\]](#)
8. [Yeast \(Saccharomyces cerevisiae, sacCer3\)](#) - an important eukaryotic model species, also good for bread and beer [\[info\]](#)

Using these files, make a table with the following information per species:

- Question 1.1. Total genome size
- Question 1.2. Number of chromosomes
- Question 1.3. Largest chromosome size and name
- Question 1.4. Smallest chromosome size and name
- Question 1.5. Mean chromosome length

Halomonas sp. GFAJ-1



Library 1: Fragment

Avg Read length: 100bp

Insert length: 180bp

Library 2: Short jump

Avg Read length: 50bp

Insert length: 2000bp

A Bacterium That Can Grow by Using Arsenic Instead of Phosphorus

Wolfe-Simon et al (2010) *Science*. 332(6034):1163-1166.

Digital Information Storage

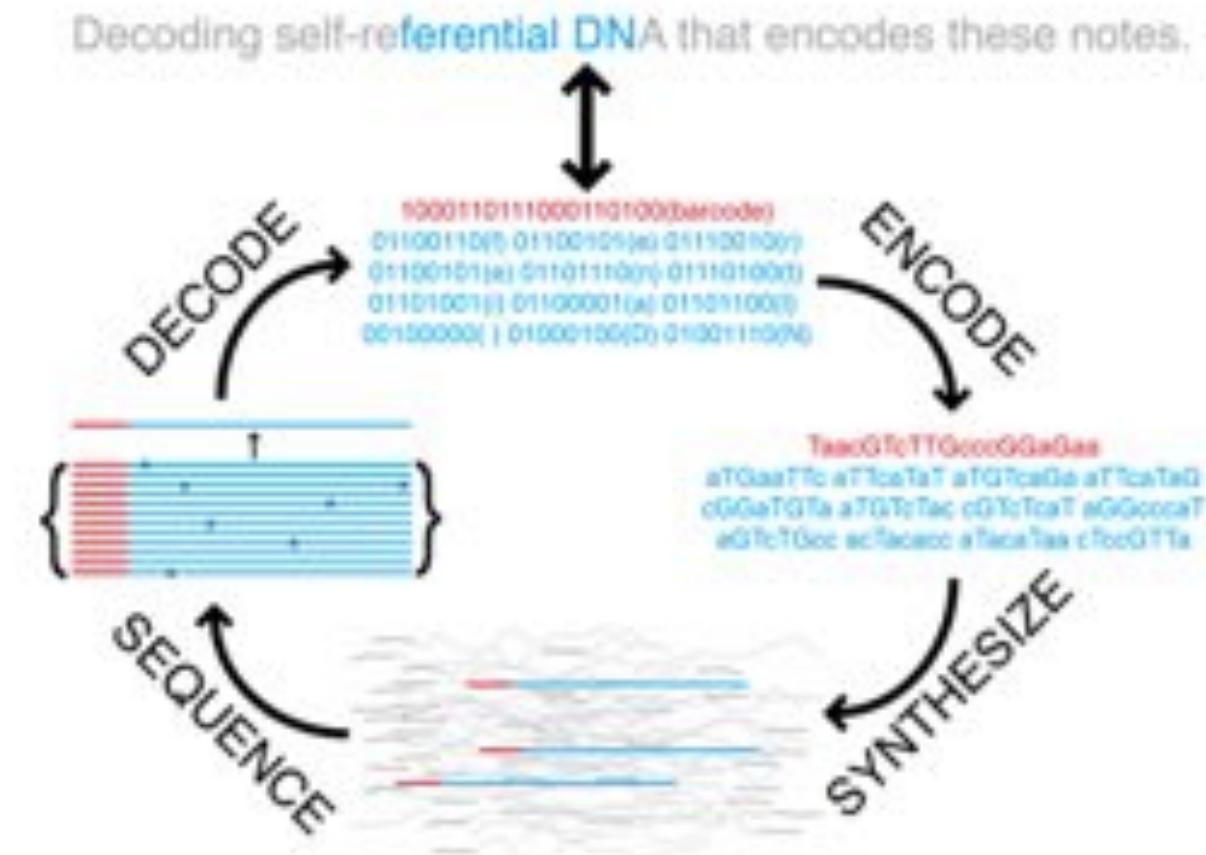


Fig. S1. Schematic of DNA information storage.

Encoding/decoding algorithm implemented in dna-encode.pl from David Dooling.

Next-generation Digital Information Storage in DNA

Church et al (2010) Science. 337(6102)1628

Assignment 2: Genome Assembly

Due Wednesday Feb 10 @ 11:59pm

1. ***Setup Docker/Ubuntu***
2. ***Initialize Tools***
3. ***Download Reference Genome & Reads***
4. ***Decode the secret message***
 1. *Estimate coverage, check read quality*
 2. *Check kmer distribution*
 3. *Assemble the reads with spades*
 4. *Align to reference with MUMmer*
 5. *Extract foreign sequence*
 6. *dna-encode.pl -d*

<https://github.com/schatzlab/appliedgenomics2021/blob/master/assignments/assignment2/README.md>



Find and decode

```
nucmer -maxmatch ref.fasta \
```

```
default/ASSEMBLIES/test/final.contigs.fasta
```

-maxmatch Find maximal exact matches (MEMs) without repeat filtering

-p refctg Set the output prefix for delta file

```
mummerplot --layout --png out.delta
```

--layout Sort the alignments along the diagonal

--png Create a png of the results

```
show-coords -rclo out.delta
```

-r Sort alignments by reference position

-c Show percent coverage

-l Show sequence lengths

-o Annotate each alignment with BEGIN/END/CONTAINS

```
samtools faidx default/ASSEMBLIES/test/final.contigs.fasta
```

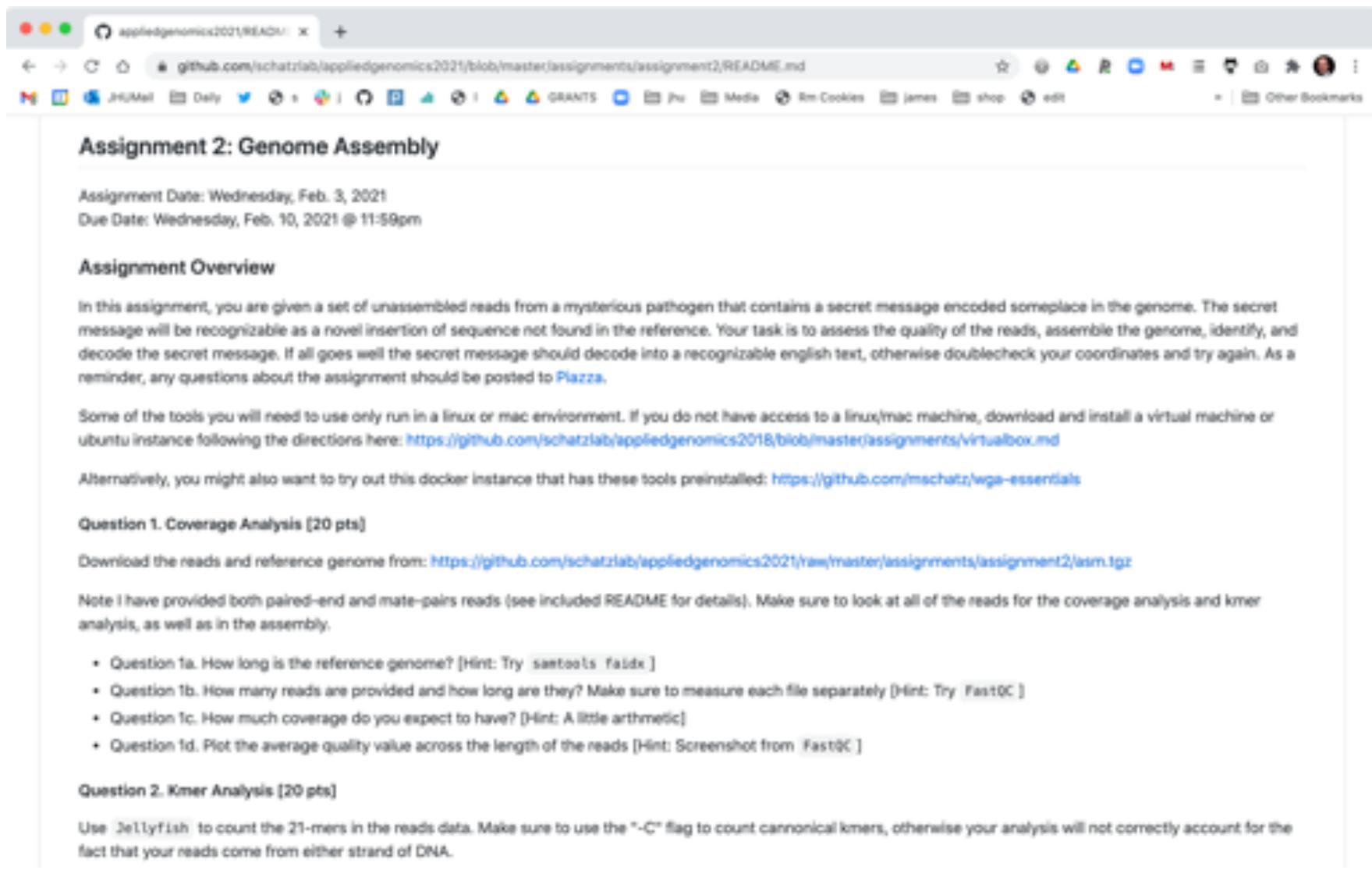
Index the fasta file

```
samtools faidx default/ASSEMBLIES/test/final.contigs.fasta \
```

```
contig_XXX:YYY-ZZZ | ./dna-encode -d
```

Assignment 2: Genome Assembly

Due Feb 10 @ 11:59pm



A screenshot of a web browser window displaying the assignment details. The title bar shows the URL: <https://github.com/schatzlab/appliedgenomics2021/blob/master/assignments/assignment2/README.md>. The page content includes:

Assignment 2: Genome Assembly

Assignment Date: Wednesday, Feb. 3, 2021
Due Date: Wednesday, Feb. 10, 2021 @ 11:59pm

Assignment Overview

In this assignment, you are given a set of unassembled reads from a mysterious pathogen that contains a secret message encoded someplace in the genome. The secret message will be recognizable as a novel insertion of sequence not found in the reference. Your task is to assess the quality of the reads, assemble the genome, identify, and decode the secret message. If all goes well the secret message should decode into a recognizable english text, otherwise doublecheck your coordinates and try again. As a reminder, any questions about the assignment should be posted to Piazza.

Some of the tools you will need to use only run in a linux or mac environment. If you do not have access to a linux/mac machine, download and install a virtual machine or ubuntu instance following the directions here: <https://github.com/schatzlab/appliedgenomics2018/blob/master/assignments/virtualbox.md>

Alternatively, you might also want to try out this docker instance that has these tools preinstalled: <https://github.com/mschatz/wga-essentials>

Question 1. Coverage Analysis [20 pts]

Download the reads and reference genome from: <https://github.com/schatzlab/appliedgenomics2021/raw/master/assignments/assignment2/asm.tgz>

Note I have provided both paired-end and mate-pairs reads (see included README for details). Make sure to look at all of the reads for the coverage analysis and kmer analysis, as well as in the assembly.

- Question 1a. How long is the reference genome? [Hint: Try `samtools faidx`]
- Question 1b. How many reads are provided and how long are they? Make sure to measure each file separately [Hint: Try `FastQC`]
- Question 1c. How much coverage do you expect to have? [Hint: A little arithmetic]
- Question 1d. Plot the average quality value across the length of the reads [Hint: Screenshot from `FastQC`]

Question 2. Kmer Analysis [20 pts]

Use `Jellyfish` to count the 21-mers in the reads data. Make sure to use the "-C" flag to count canonical kmers, otherwise your analysis will not correctly account for the fact that your reads come from either strand of DNA.

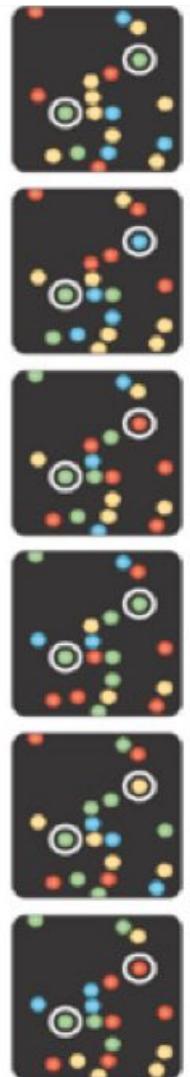
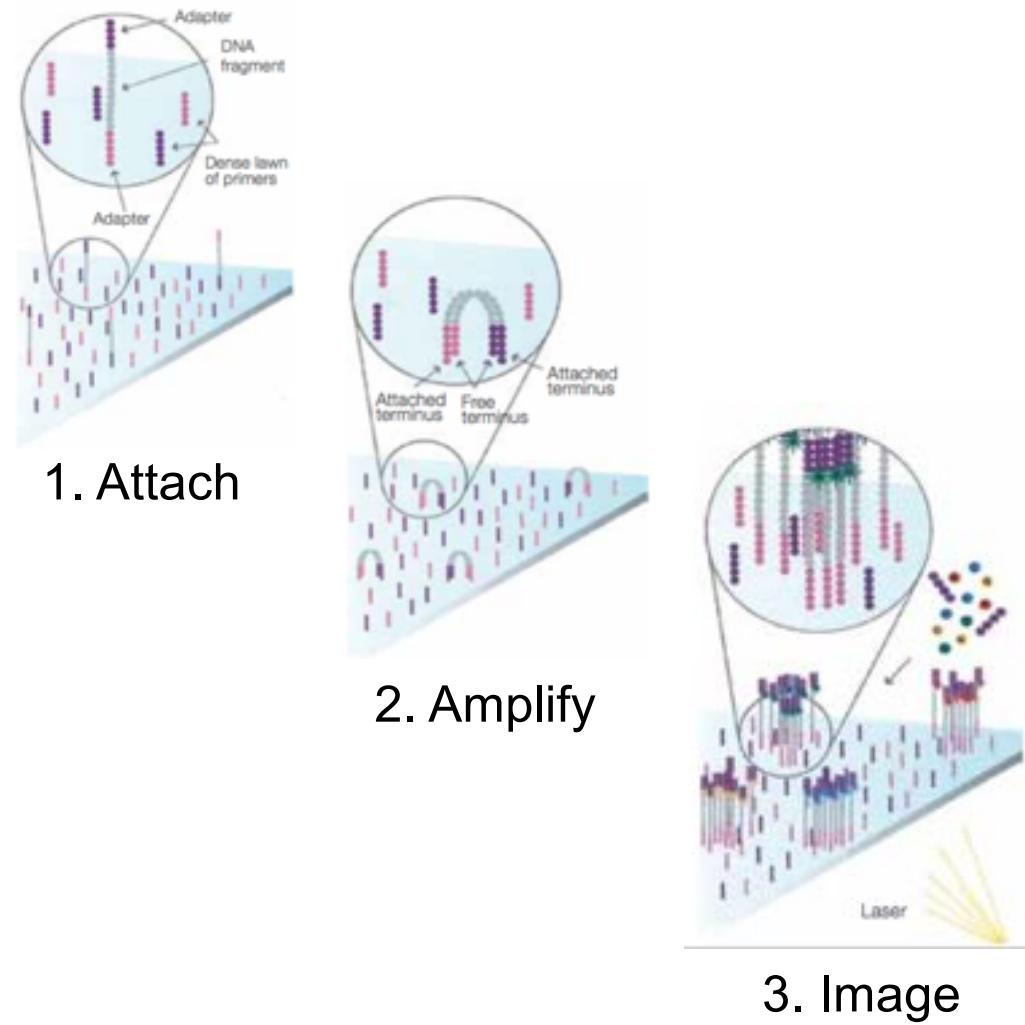
Part I: Recap

Second Generation Sequencing



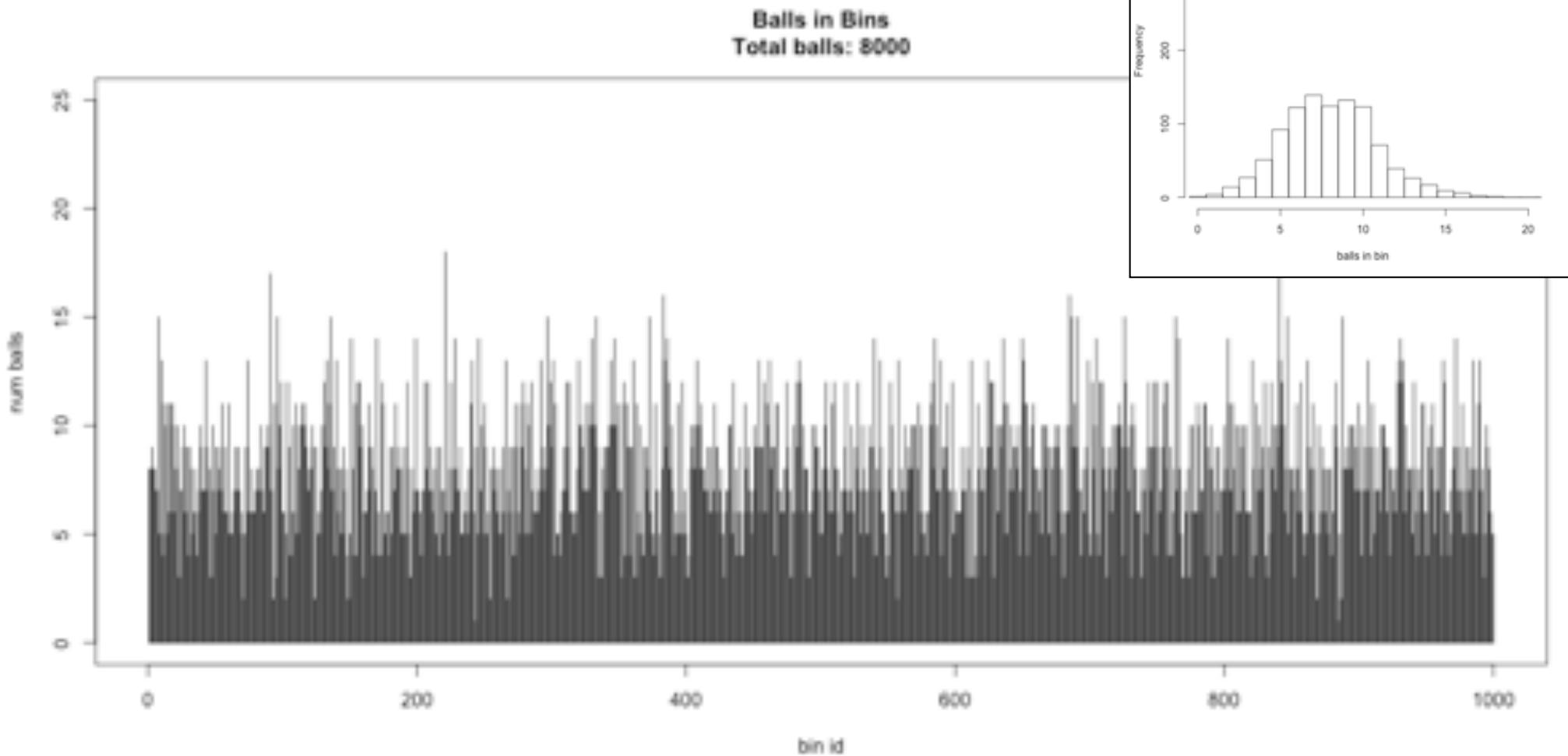
Illumina HiSeq 2000
Sequencing by Synthesis

>60Gbp / day



Metzker (2010) Nature Reviews Genetics 11:31-46
<https://www.youtube.com/watch?v=fCd6B5HRaZ8>

8x sequencing



Paired-end and Mate-pairs

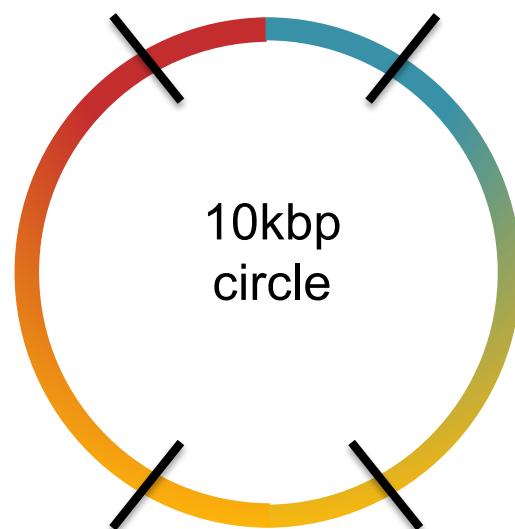
Paired-end sequencing

- Read one end of the molecule, flip, and read the other end
- Generate pair of reads separated by up to 500bp with inward orientation



Mate-pair sequencing

- Circularize long molecules (1-10kbp), shear into fragments, & sequence
- Mate failures create short paired-end reads



2x100 @ ~10kbp (outies)



2x100 @ 300bp (innies)



Pop Quiz!

I want to sequence a 10Mbp genome to 24x coverage.
How many 120bp reads do I need?

I need $10\text{Mbp} \times 24\text{x} = 240\text{Mbp}$ of data
 $240\text{Mbp} / 120\text{bp} / \text{read} = 2\text{M reads}$

I want to sequence a 10Mbp genome so that
>97.5% of the genome has at least 24x coverage.
How many 120bp reads do I need?

Find X such that $X - 2\sqrt{X} = 24$

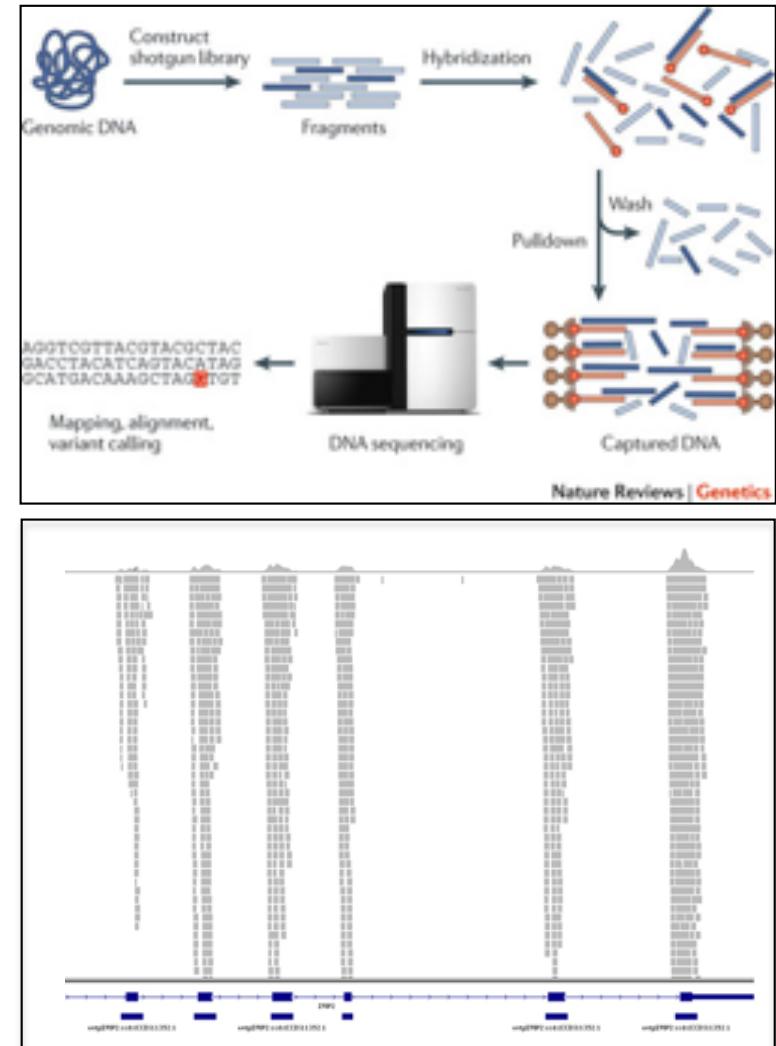
$$36 - 2\sqrt{36} = 24$$

I need $10\text{Mbp} \times 36\text{x} = 360\text{Mbp}$ of data
 $360\text{Mbp} / 120\text{bp} / \text{read} = 3\text{M reads}$

Exome-Capture Sequencing

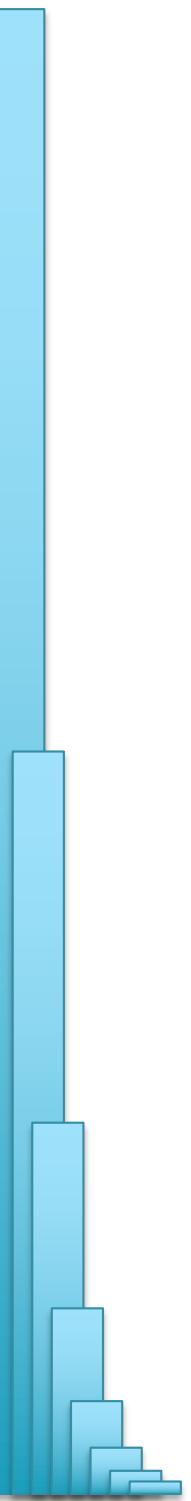
Exome-capture reduces the costs of sequencing

- Currently targets around 50Mbp of sequence: all exons plus flanking regions
- WGS currently costs ~\$1000 per sample, while WES currently costs ~\$250 per sample
- Coverage is highly localized around genes, although will get sparse coverage throughout rest of genome



Exome sequencing as a tool for Mendelian disease gene discovery

Bamshad et al. (2011) Nature Reviews Genetics. 12, 745-755



Part 2: De novo genome assembly



Outline

1. ***Assembly theory***

- Assembly by analogy

2. ***Practical Issues***

- Coverage, read length, errors, and repeats

3. ***Next-next-gen Assembly***

- Canu: recommended for PacBio/ONT project

4. ***Whole Genome Alignment***

- MUMmer recommended

Shredded Book Reconstruction

- Dickens accidentally shreds the first printing of A Tale of Two Cities
 - Text printed on 5 long spools

It was the best of times, it was the worst of times, it was the age of wisdom; it was the age of foolishness, ...

It was the best of times, it was the worst of times, it was the age of wisdom; it was the age of foolishness, ...

It was the best of times, it was the worst of times, it was the age of wisdom; it was the age of foolishness, ...

It was the best of times, it was the worst of times, it was the age of wisdom; it was the age of foolishness, ...

It was the best of times, it was the worst of times, it was the age of wisdom; it was the age of foolishness, ...

- How can he reconstruct the text?

- 5 copies x 138,656 words / 5 words per fragment = 138k fragments
- The short fragments from every copy are mixed together
- Some fragments are identical

It was the best of
age of wisdom, it was
best of times, it was
it was the age of
it was the age of
it was the worst of
of times, it was the
of times, it was the
of wisdom, it was the
the age of wisdom, it
the best of times, it
the worst of times, it
times, it was the age
times, it was the worst
was the age of wisdom,
was the age of foolishness,
was the best of times,
was the worst of times,
wisdom, it was the age
worst of times, it was

Greedy Reconstruction

It was the best of
was the best of times,
the best of times, it
best of times, it was
of times, it was the
of times, it was the
times, it was the worst
times, it was the age

The repeated sequence make the correct reconstruction ambiguous

- It was the best of times, it was the [worst/age]

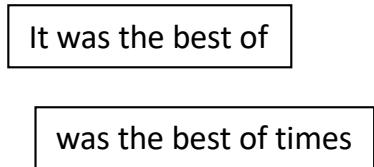
Model the assembly problem as a graph problem

How long will it take to compute the overlaps?

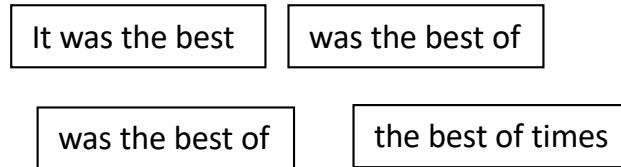
de Bruijn Graph Construction

- $G_k = (V, E)$
 - V = Length- k sub-fragments
 - E = Directed edges between consecutive sub-fragments
 - Sub-fragments overlap by $k-1$ words

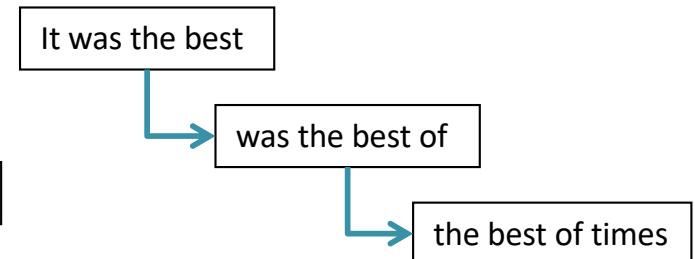
Fragments $|f|=5$



Sub-fragment $k=4$



Directed edges (overlap by $k-1$)



– Overlaps between fragments are implicitly computed

How to pronounce:

https://forvo.com/word/de_briuin/

de Bruijn, 1946
Idury et al., 1995
Pevzner et al., 2001

de Bruijn Graph Assembly

It was the best

was the best of

the best of times,

best of times, it

of times, it was

times, it was the

it was the worst

was the worst of

the worst of times,

worst of times, it

it was the age

was the age of

the age of foolishness

the age of wisdom,

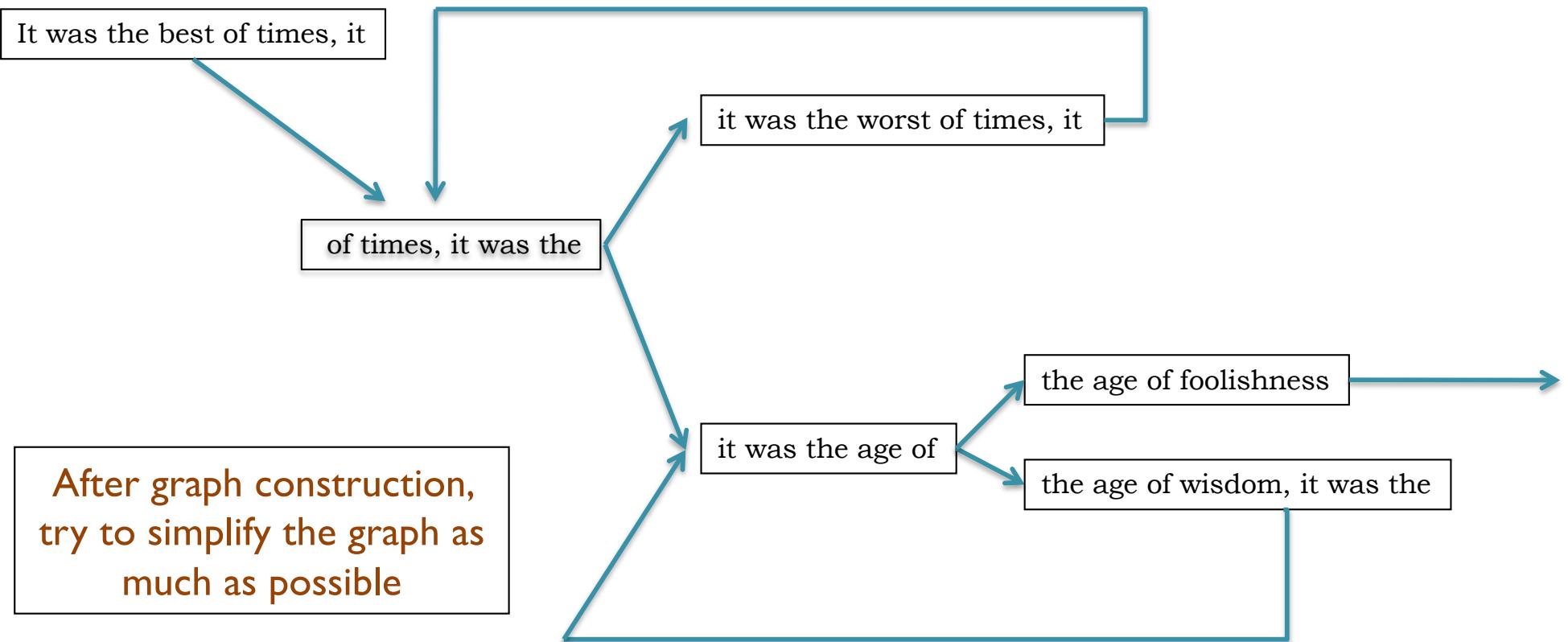
age of wisdom, it

of wisdom, it was

wisdom, it was the

After graph construction,
try to simplify the graph as
much as possible

de Bruijn Graph Assembly



The full tale

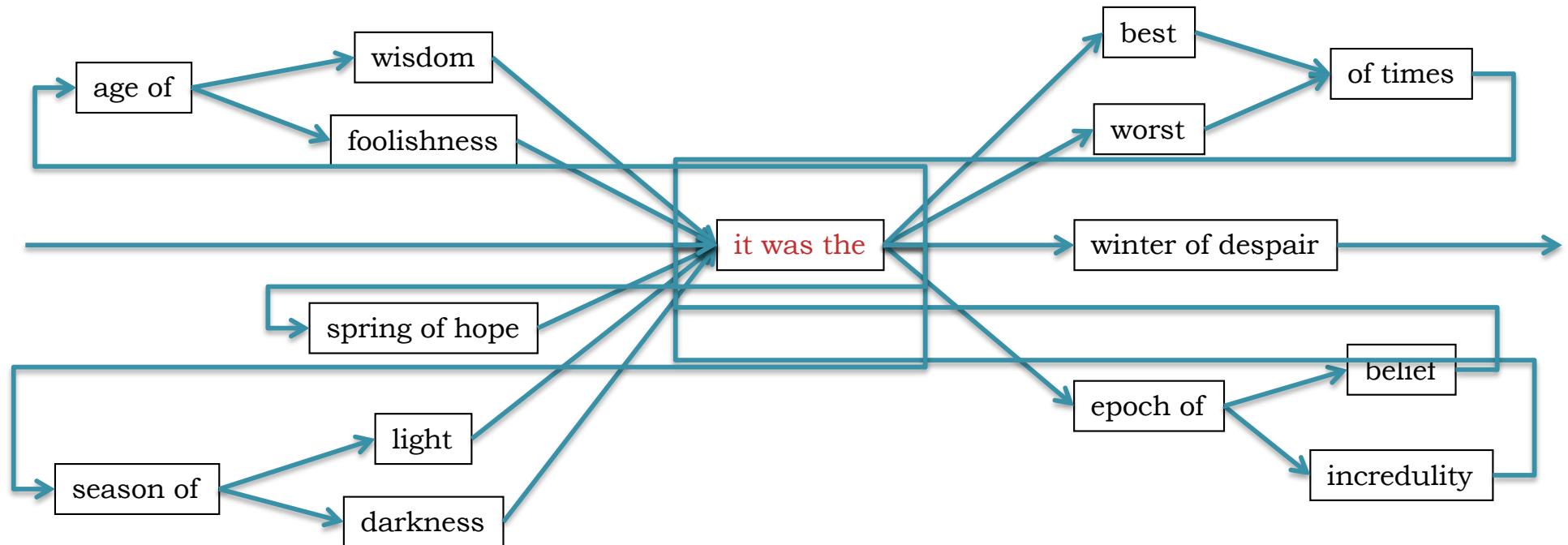
... it was the best of times it was the worst of times ...

... it was the age of wisdom it was the age of foolishness ...

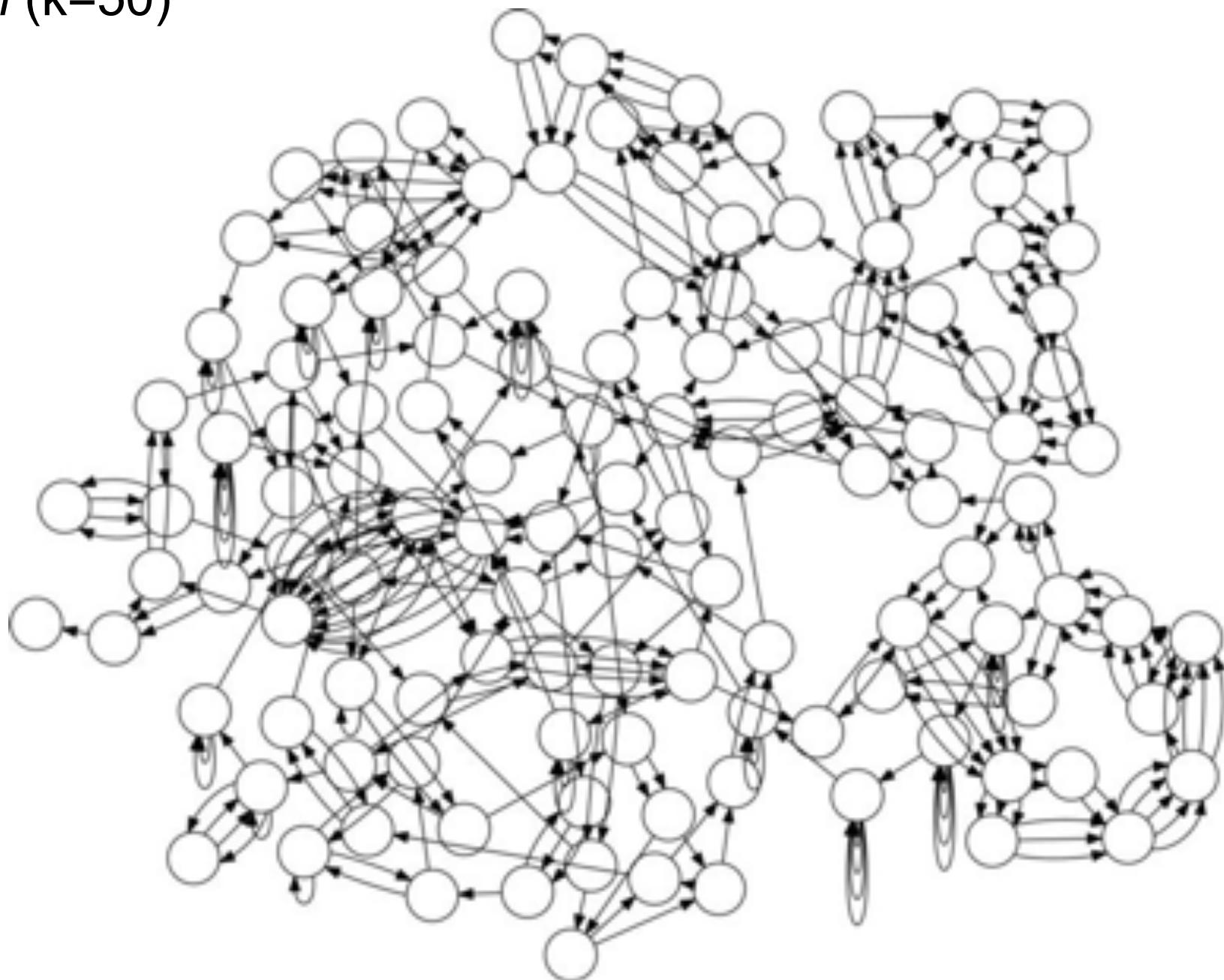
... it was the epoch of belief it was the epoch of incredulity ...

... it was the season of light it was the season of darkness ...

... it was the spring of hope it was the winter of despair ...

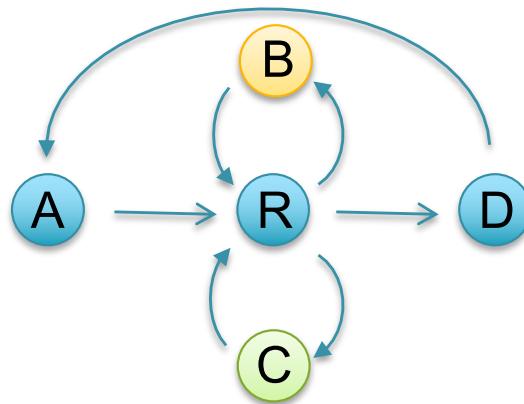


E. coli ($k=50$)



Reducing assembly complexity of microbial genomes with single-molecule sequencing
Koren et al (2013) Genome Biology. 14:R101 <https://doi.org/10.1186/gb-2013-14-9-r101>

Counting Eulerian Cycles



ARBRCRD
or
ARCRBRD

Generally an exponential number of compatible sequences

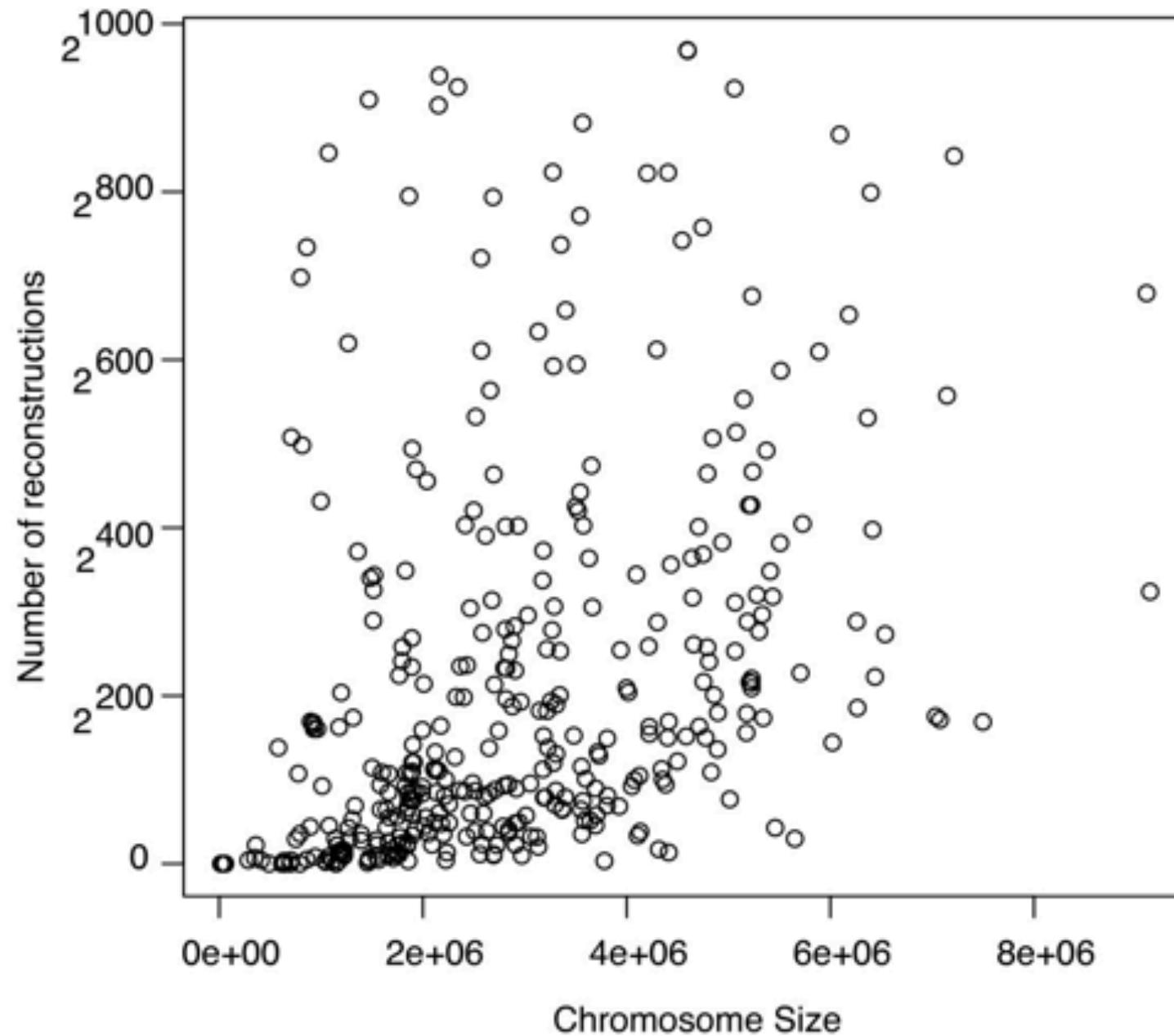
- Value computed by application of the BEST theorem (Hutchinson, 1975)

$$\mathcal{W}(G, t) = (\det L) \left\{ \prod_{u \in V} (r_u - 1)! \right\} \left\{ \prod_{(u,v) \in E} a_{uv}! \right\}^{-1}$$

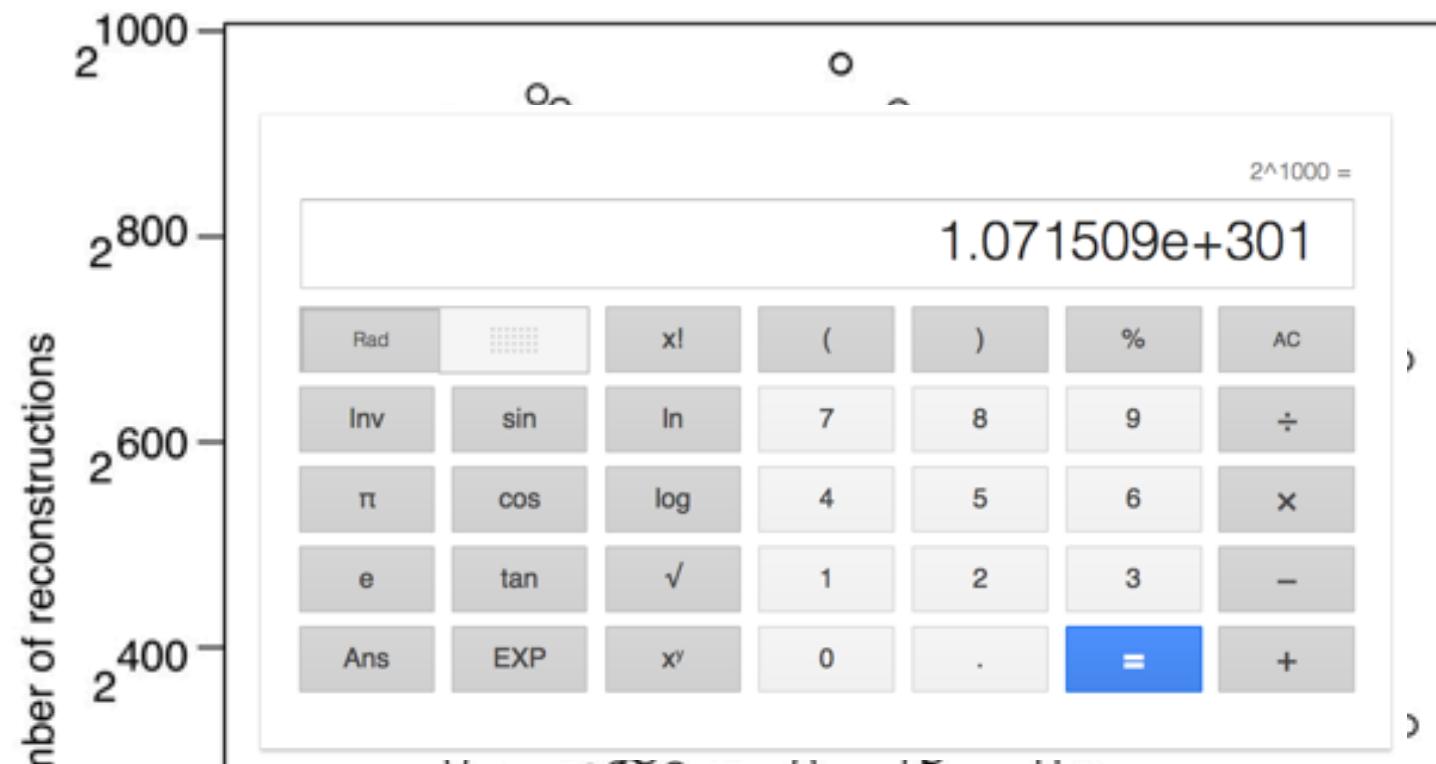
L = $n \times n$ matrix with $r_u - a_{uu}$ along the diagonal and $-a_{uv}$ in entry uv

$r_u = d^+(u) + 1$ if $u=t$, or $d^+(u)$ otherwise

a_{uv} = multiplicity of edge from u to v



Assembly Complexity of Prokaryotic Genomes using Short Reads.
Kingsford C, Schatz MC, Pop M (2010) *BMC Bioinformatics*.

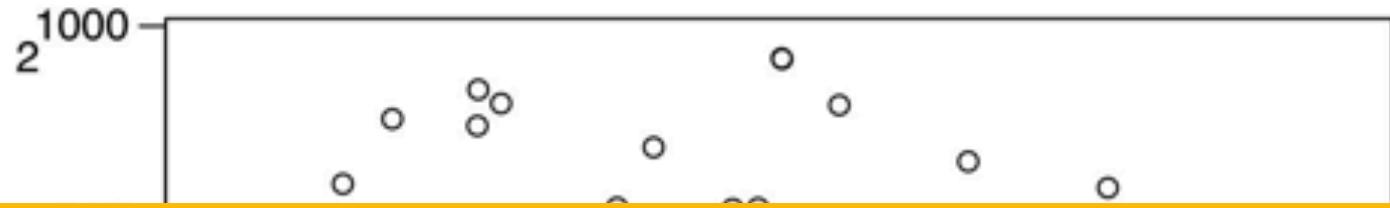


It is believed 74% of the mass of the Milky Way, for example, is in the form of hydrogen atoms. The Sun contains approximately **10⁵⁷ atoms** of hydrogen. If you multiple the number of atoms per star (10⁵⁷) times the estimated number of stars in the universe (10²³), you get a value of **10⁸⁰ atoms** in the known universe. Nov 5, 2017

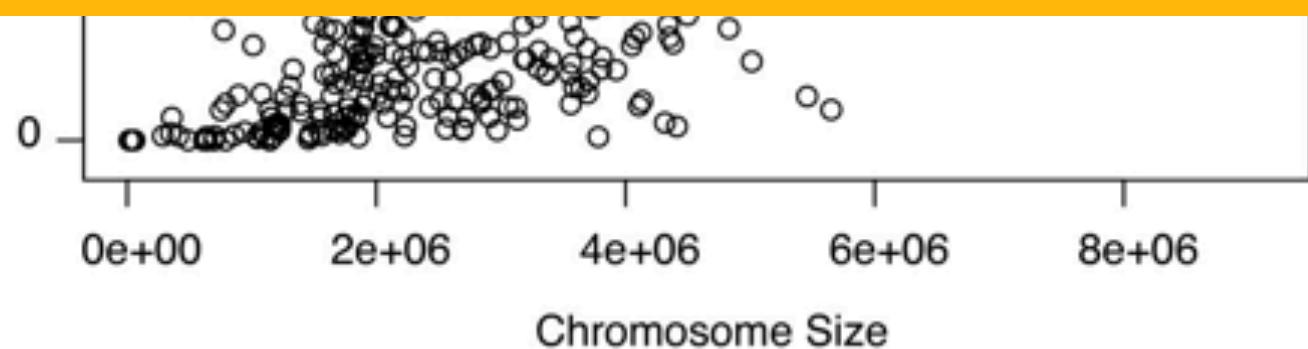


[How Many Atoms Are There in the Universe? - ThoughtCo](https://www.thoughtco.com/number-of-atoms-in-the-universe-603795)
<https://www.thoughtco.com/number-of-atoms-in-the-universe-603795>

Assembly Complexity of Prokaryotic Genomes using Short Reads.
Kingsford C, Schatz MC, Pop M (2010) *BMC Bioinformatics*.



- **Finding possible assemblies is easy!**
- **However, there is an astronomical genomic number of possible paths!**
- **Hopeless to figure out the whole genome/chromosome, figure out the parts that you can**

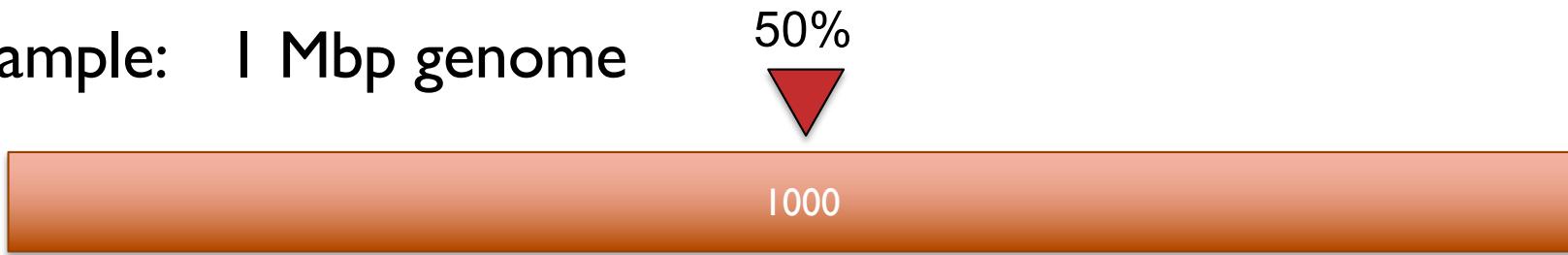


Assembly Complexity of Prokaryotic Genomes using Short Reads.
Kingsford C, Schatz MC, Pop M (2010) *BMC Bioinformatics*.

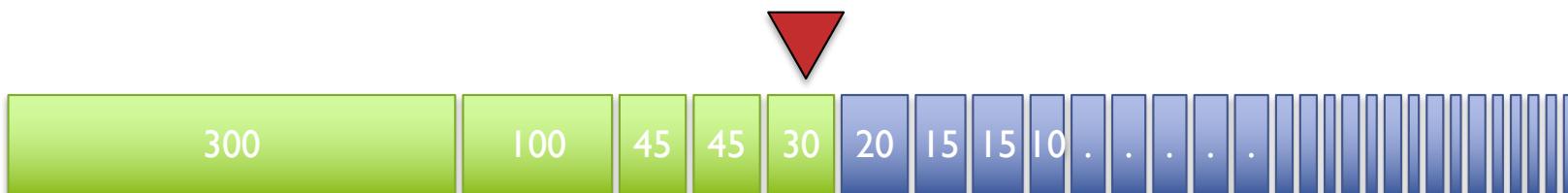
Contig N50

Def: 50% of the genome is in contigs as large as the N50 value

Example: 1 Mbp genome



A



N50 size = 30 kbp

B



N50 size = 3 kbp

Contig N50

Def: 50% of the genome is in contigs as large as the N50 value

50%

Better N50s improves the analysis in every dimension

- Better resolution of genes and flanking regulatory regions
- Better resolution of transposons and other complex sequences
- Better resolution of chromosome organization
- Better sequence for all downstream analysis

Just be careful of N50 inflation!

- A very very very bad assembler in 1 line of bash:
- `cat *.reads.fa > genome.fa`

N50 size = 3 kbp

Pop Quiz I

Assemble these reads using a de Bruijn graph approach (k=3):

ATTA

GATT

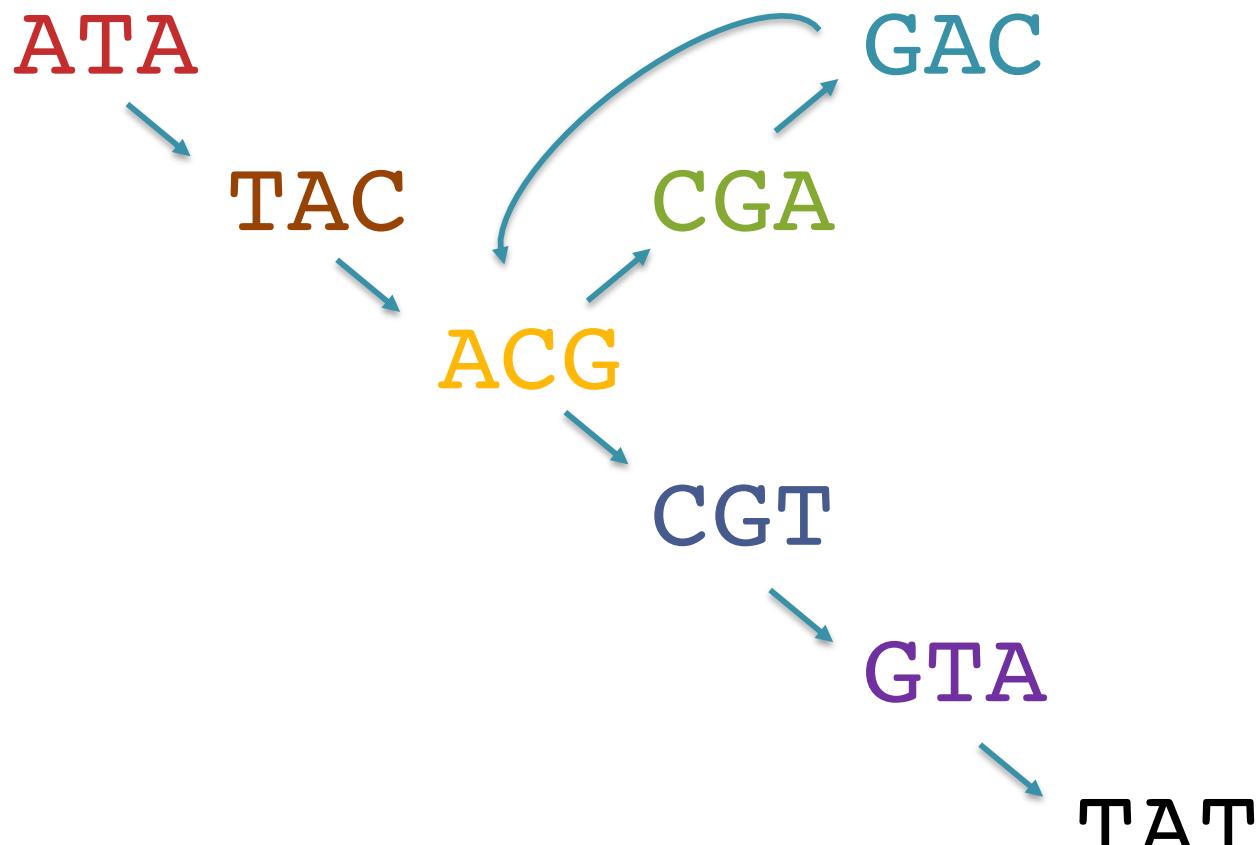
TACA

TTAC

Pop Quiz 2

Assemble these reads using a de Bruijn graph approach (k=3):

~~ACGA~~
~~ACGT~~
~~ATAC~~
~~CGAC~~
~~CGTA~~
~~GACG~~
~~GTAT~~
~~TACG~~



Should we add the edge TAT -> ATA?

ATACGACGTAT



Titus Brown

@ctitusbrown

Following



Wow, this could double as life philosophy, too!

Michael Schatz @mike_schatz

Replying to @ZaminIqbal @nomad421 and 4 others

Yep, very easy to find *a* path, very hard to find *the* path

11:40 AM - 22 Jan 2018

4 Retweets 17 Likes



2

4

17





Outline

1. *Assembly theory*

- Assembly by analogy

2. **Practical Issues**

- Coverage, read length, errors, and repeats

3. *Next-next-gen Assembly*

- Canu: recommended for PacBio/ONT project

4. Whole Genome Alignment

- MUMmer recommended

Assembly Applications

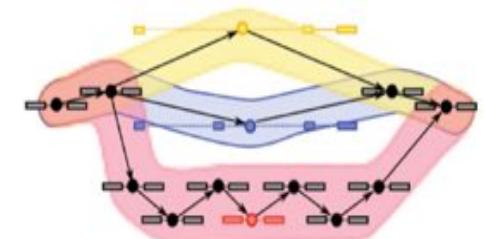
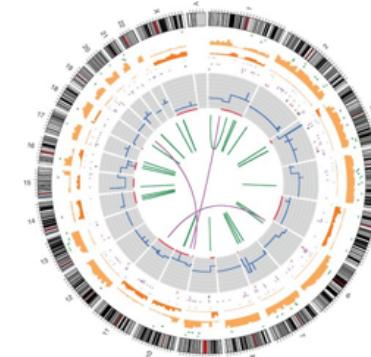
- Novel genomes



- Metagenomes



- Sequencing assays
 - Structural variations
 - Transcript assembly
 - ...



Why are genomes hard to assemble?

1. *Biological:*

- (Very) High ploidy, heterozygosity, repeat content

2. *Sequencing:*

- (Very) large genomes, imperfect sequencing

3. *Computational:*

- (Very) Large genomes, complex structure

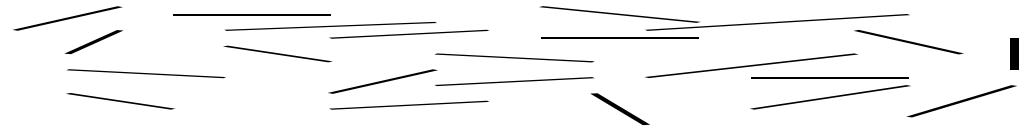
4. *Accuracy:*

- (Very) Hard to assess correctness



Assembling a Genome

I. Shear & Sequence DNA

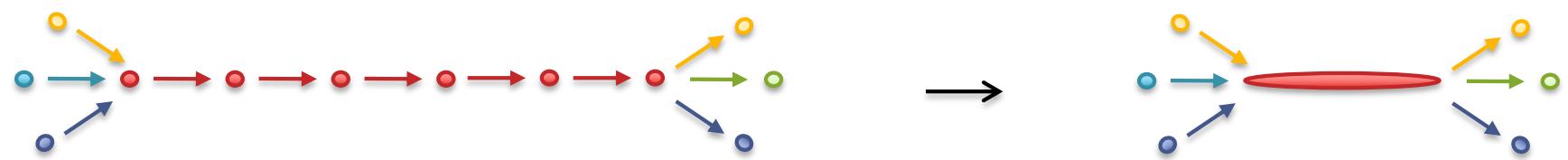


2. Construct assembly graph from reads (de Bruijn / overlap graph)

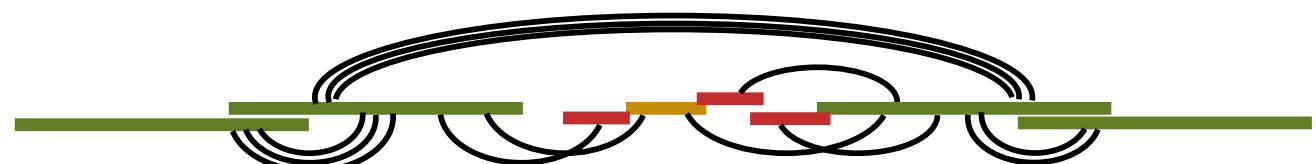
...AGCCTAG**GGATGCGCGACACGT**

GGATGCGCGACACGTCGCATATCCGGTTTGGT**CAACCTCGGACGGAC**
CAACCTCGGACGGACCTCAGCGAA...

3. Simplify assembly graph

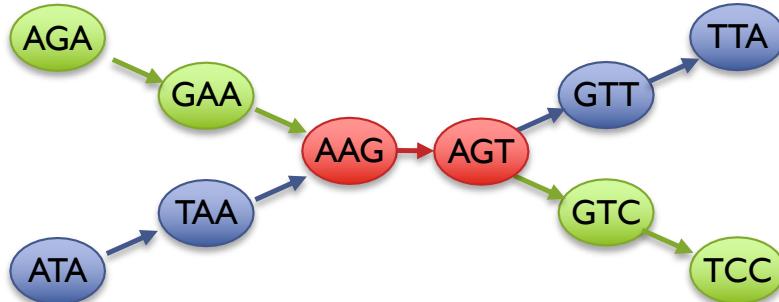


4. Detangle graph with long reads, mates, and other links



Two Paradigms for Assembly

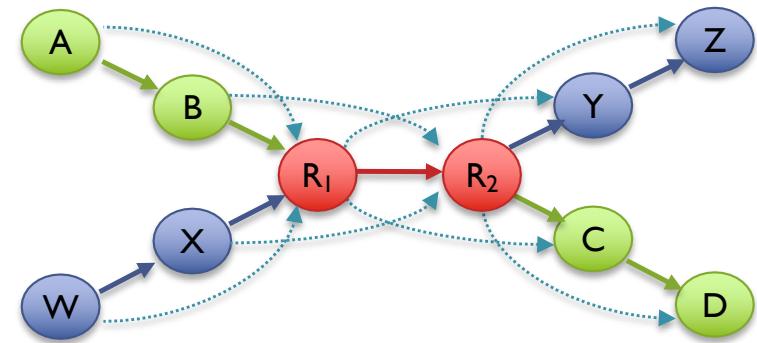
de Bruijn Graph



Short read assemblers

- Repeats depends on word length
- Read coherency, placements lost
- Robust to high coverage

Overlap Graph



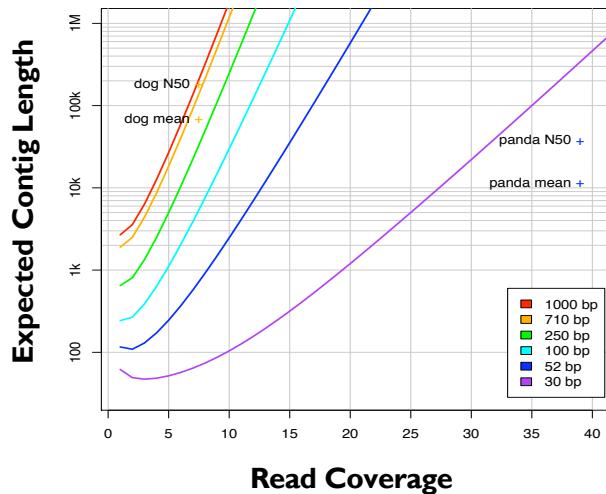
Long read assemblers

- Repeats depends on read length
- Read coherency, placements kept
- Tangled by high coverage

Assembly of Large Genomes using Second Generation Sequencing
Schatz MC, Delcher AL, Salzberg SL (2010) *Genome Research*. 20:1165-1173.

Ingredients for a good assembly

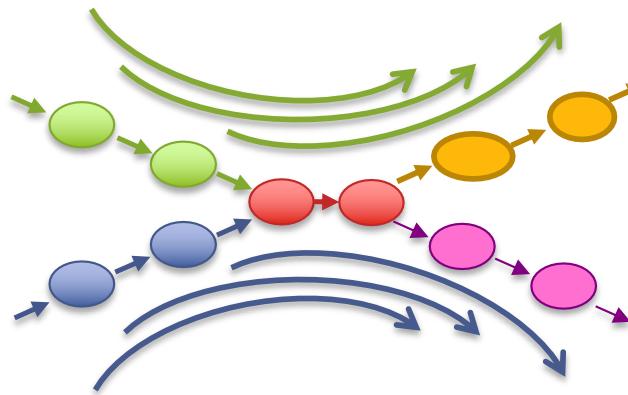
Coverage



High coverage is required

- Oversample the genome to ensure every base is sequenced with long overlaps between reads
- Biased coverage will also fragment assembly

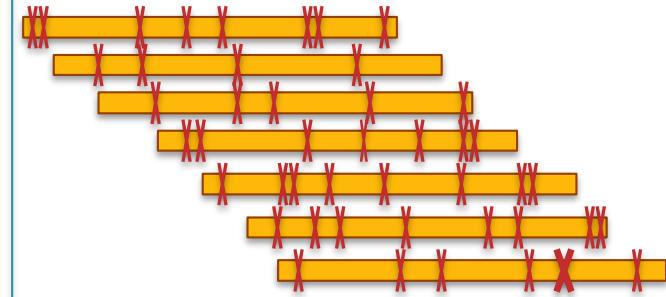
Read Length



Reads & mates must be longer than the repeats

- Short reads will have **false overlaps** forming hairball assembly graphs
- With long enough reads, assemble entire chromosomes into contigs

Quality



Errors obscure overlaps

- Reads are assembled by finding kmers shared in pair of reads
- High error rate requires very short seeds, increasing complexity and forming assembly hairballs

Current challenges in *de novo* plant genome sequencing and assembly

Schatz MC, Witkowski, McCombie, WR (2012) *Genome Biology*. 12:243

Coverage Statistics

$$\text{sequencing_coverage} = \frac{\text{total_bases_sequenced}}{\text{genome_size}}$$

$$\text{genome_size} = \frac{\text{total_bases_sequenced}}{\text{sequencing_coverage}}$$

$$\text{genome_size} = \frac{100\text{Gb}}{50x} = 2\text{Gb}$$

But how can you figure out
the coverage without a genome?

K-mer counting

Kmer-ize

Read 1: GATTACA => GAT, ATT, TTA, TAC, ACA
Read 2: TACAGAG => TAC, ACA, CAG, AGA, GAG
Read 3: TTACAGA => TTA, TAC, ACA, CAG, AGA

A brown arrow originates from the bottom left, pointing towards the top right. It starts near the word "list" and curves upwards and to the right, ending with an arrowhead pointing at the first item in the list.

GAT	ACA	ACA : 3
ATT	ACA	
TTA	ACA	
TAC	AGA	AGA : 2
ACA	AGA	
TAC	ATT	ATT : 1
ACA	CAG	CAG : 2
CAG	CAG	
AGA	GAG	GAG : 1
GAG	GAT	GAT : 1
TTA	TAC	TAC : 3
TAC	TAC	
ACA	TAC	
CAG	TTA	TTA : 2
AGA	TTA	

3 kmers occur 1x
3 kmers occur 2x
2 kmers occur 3x

• Tunes used on

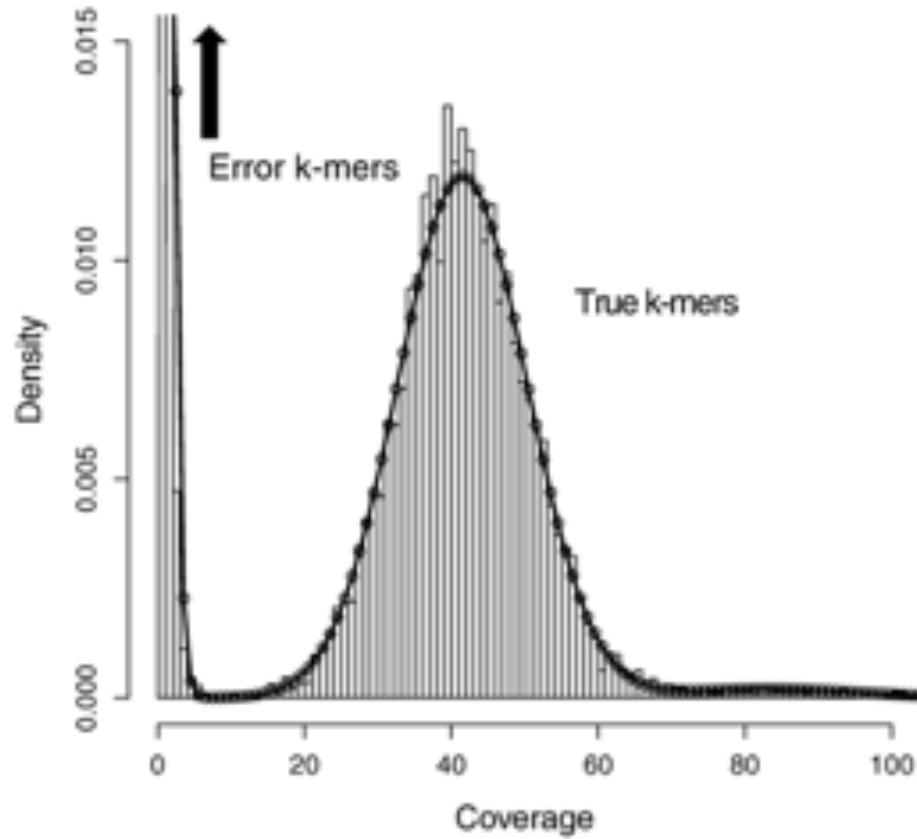
tally

The diagram consists of two brown curved arrows pointing upwards and to the right, positioned above the words "sort" and "count".

From read k-mers alone, can learn something about how frequently different sequences occur (aka coverage)

Fast to compute even over huge datasets

K-mer counting in real genomes



- The tally of k-mer counts in real genomes reveals the coverage distribution.
- Here we sequenced 120Gb of reads from a female human (haploid human genome size is 3Gb), and indeed we see a clear peak centered at 40x coverage
- There are also many kmers that only occur <5 times. These are from errors in the reads
- There are also kmers that occur many times (>>70 times). These are repeats in the genome

K-mer counting in heterozygous genomes

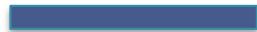
Sequencing read
from homologous
chromosome 1A



Sequencing read
from homologous
chromosome 1B



K-mer counting in heterozygous genomes



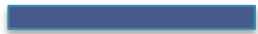
Sequencing read
from homologous
chromosome 1A



Sequencing read
from homologous
chromosome 1B



K-mer counting in heterozygous genomes



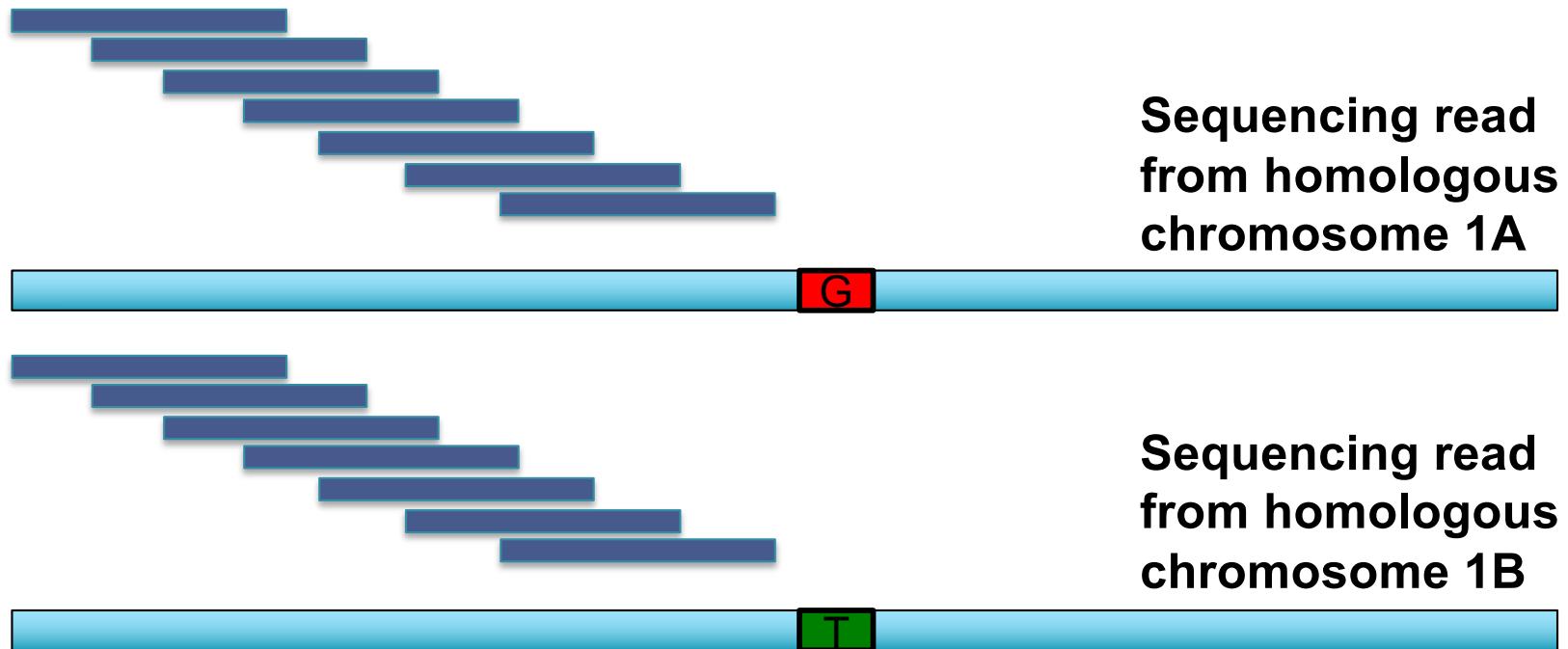
Sequencing read
from homologous
chromosome 1A



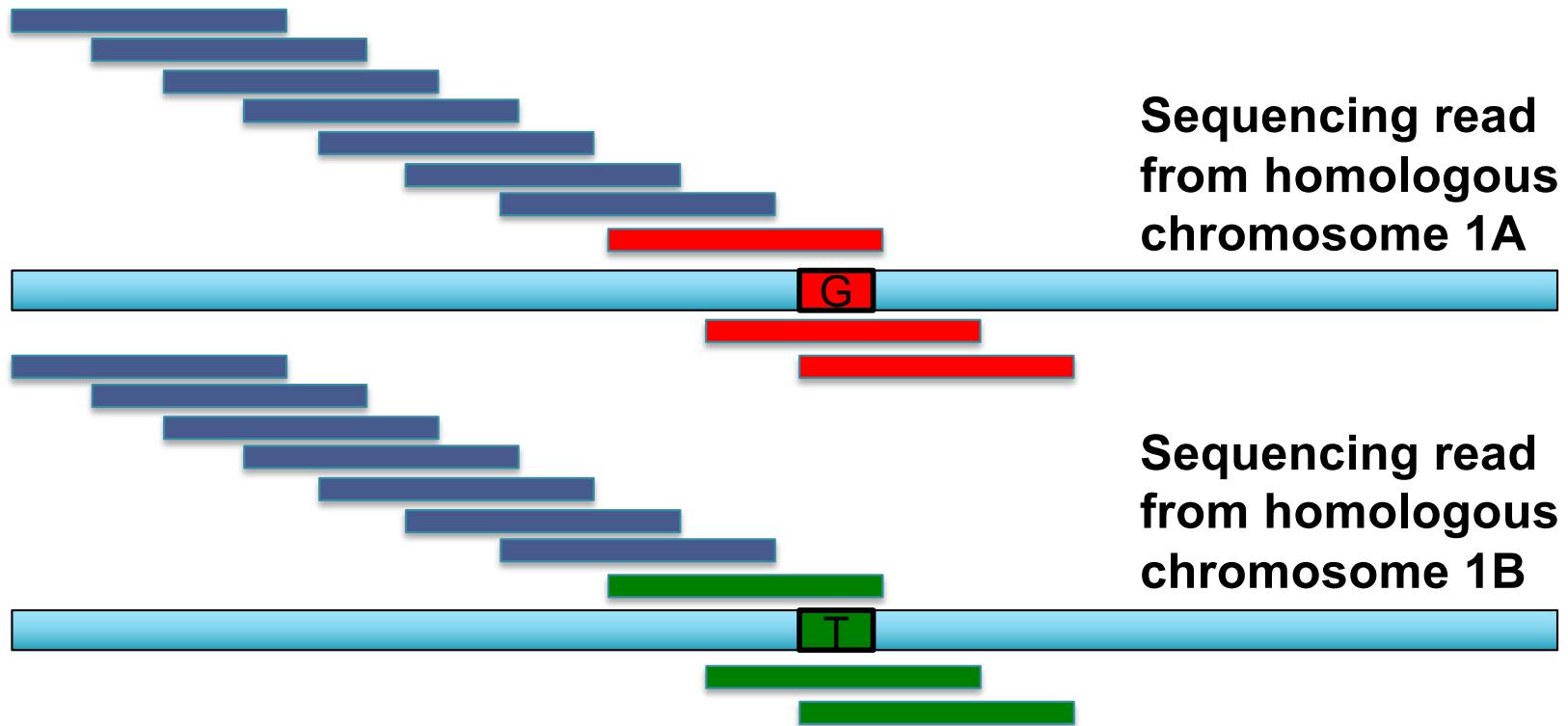
Sequencing read
from homologous
chromosome 1B



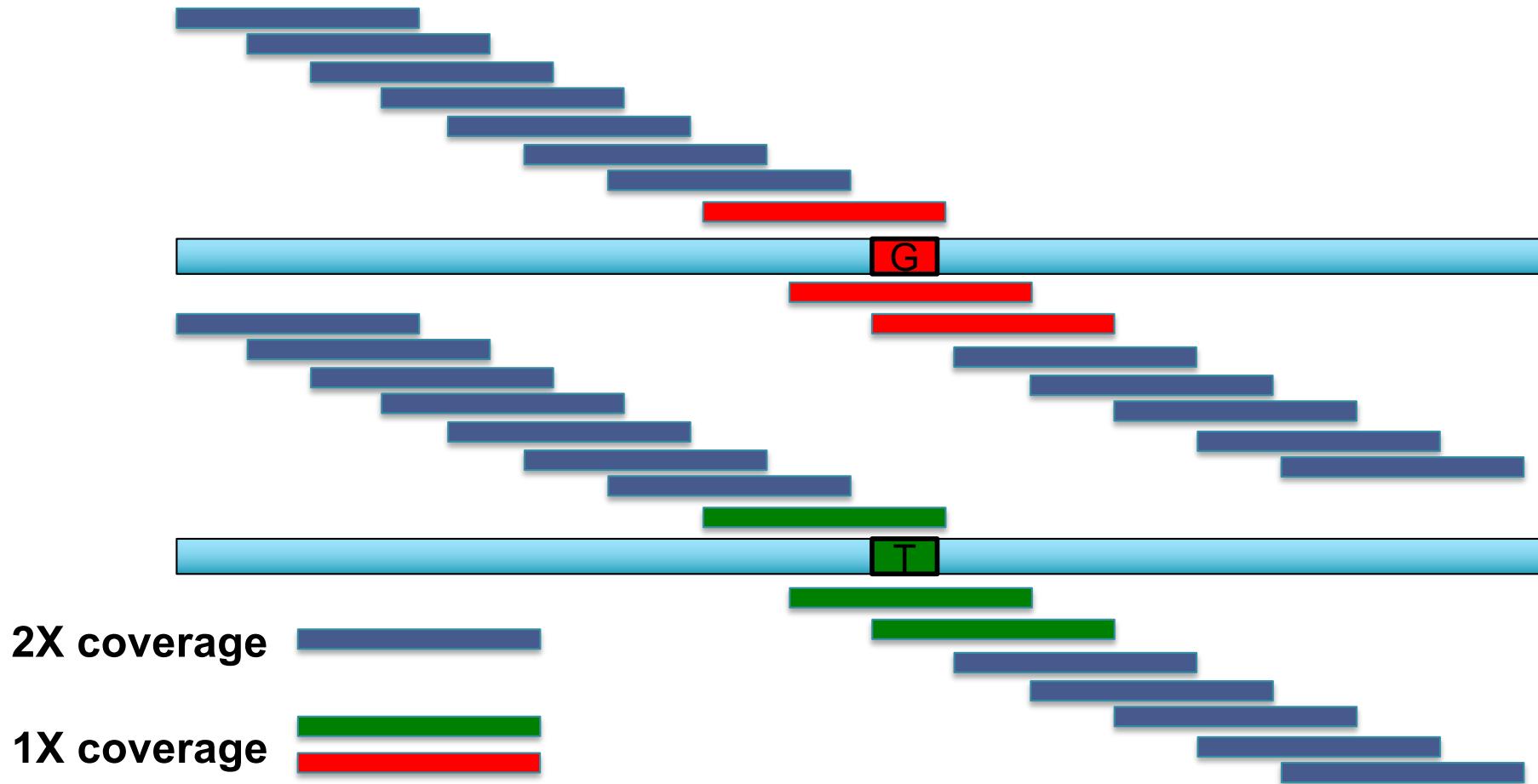
K-mer counting in heterozygous genomes



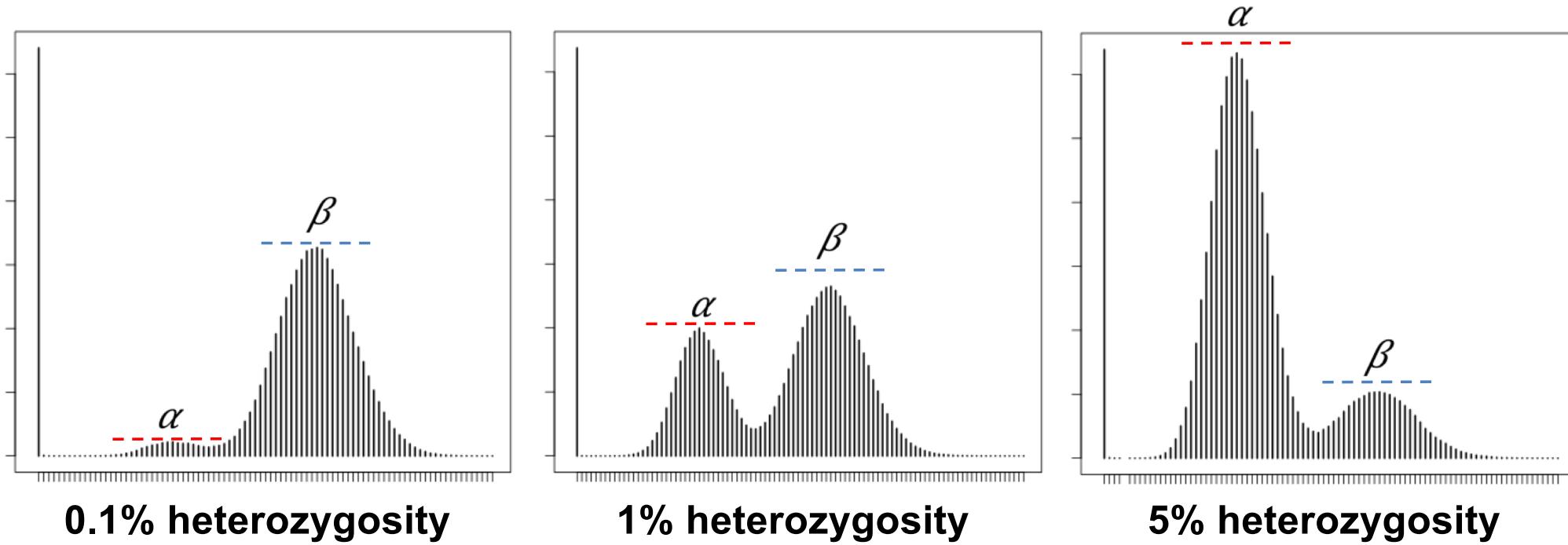
K-mer counting in heterozygous genomes



K-mer counting in heterozygous genomes



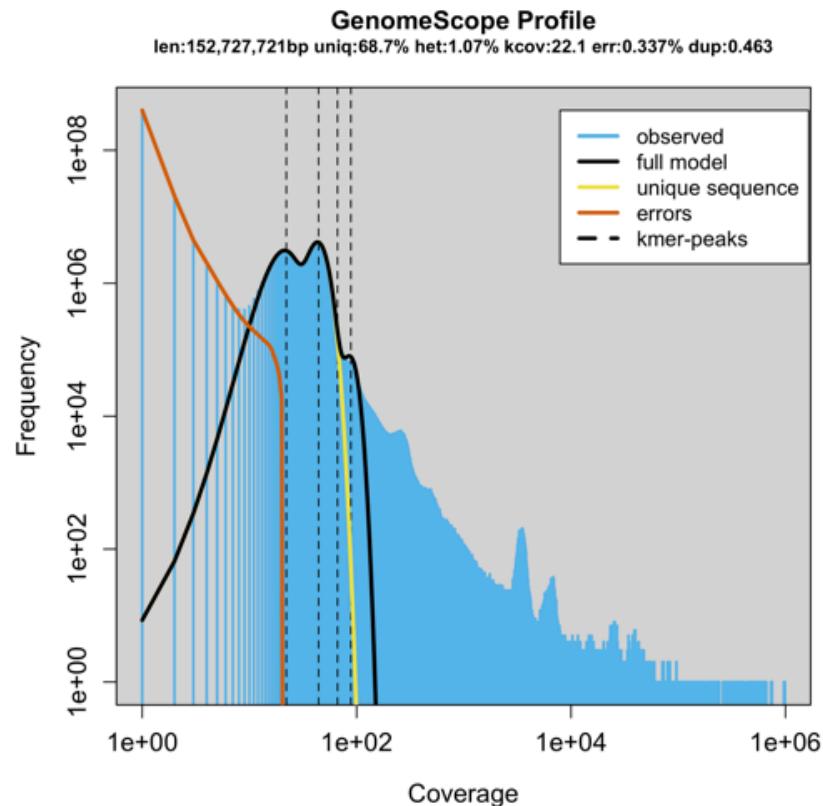
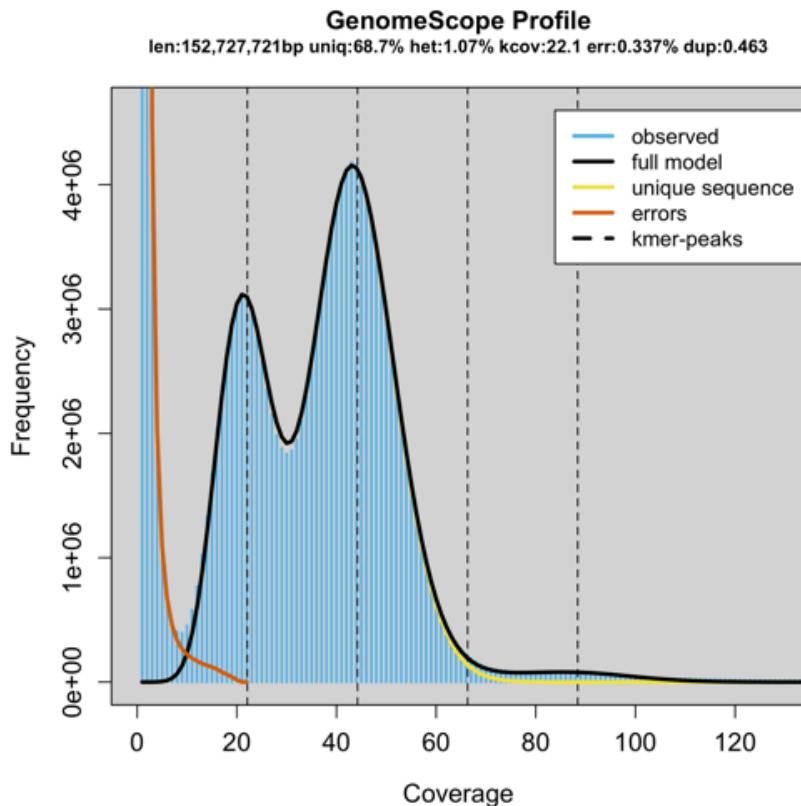
Heterozygous Kmer Profiles



- ***Heterozygosity creates a characteristic “double-peak” in the Kmer profile***
 - Second peak at twice k-mer coverage as the first: heterozygous kmers average 50x coverage, homozygous kmers average 100x coverage
- ***Relative heights of the peaks is directly proportional to the heterozygosity rate***
 - The peaks are balanced at around 1.25% because each heterozygous SNP creates 2^k heterozygous kmers (typically $k = 21$)

GenomeScope: Fast genome analysis from short reads

<http://genomescope.org>



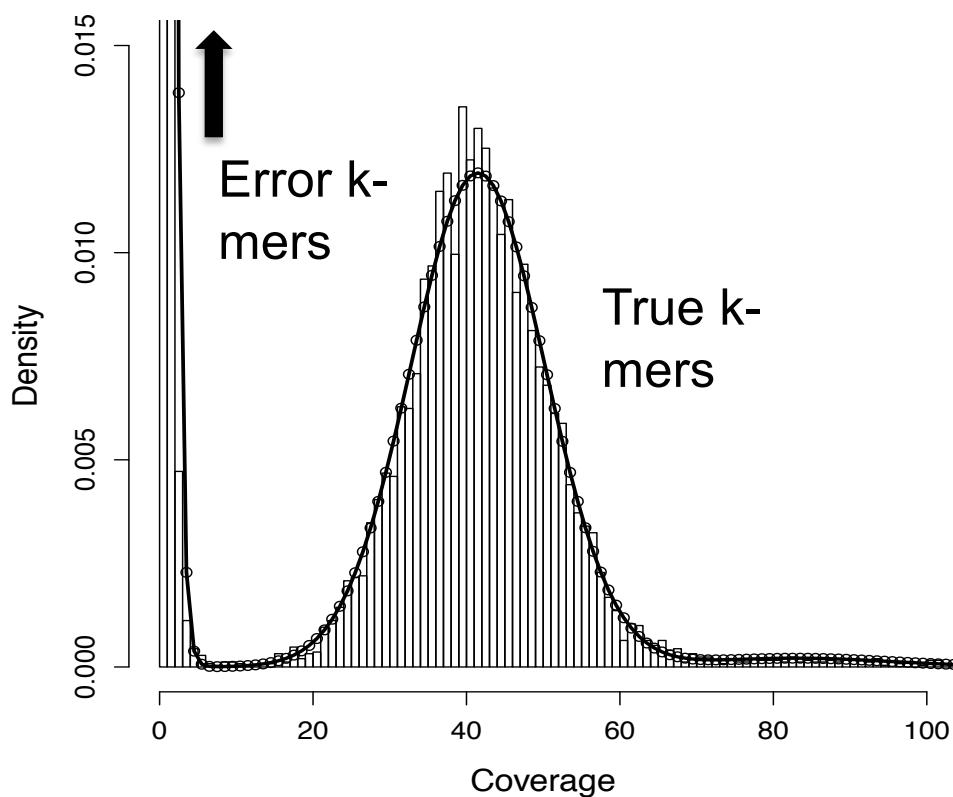
- Theoretical model agrees well with published results:
 - Rate of heterozygosity is higher than reported by other approaches but likely correct.
 - Genome size of plants inflated by organelle sequences (exclude very high freq. kmers)

Vulture, GW*, Sedlazeck FJ*, et al. (2017) *Bioinformatics*
Ranallo-Benavidez, TR. et al. (2020) *Nature Communication*

Error Correction with Quake

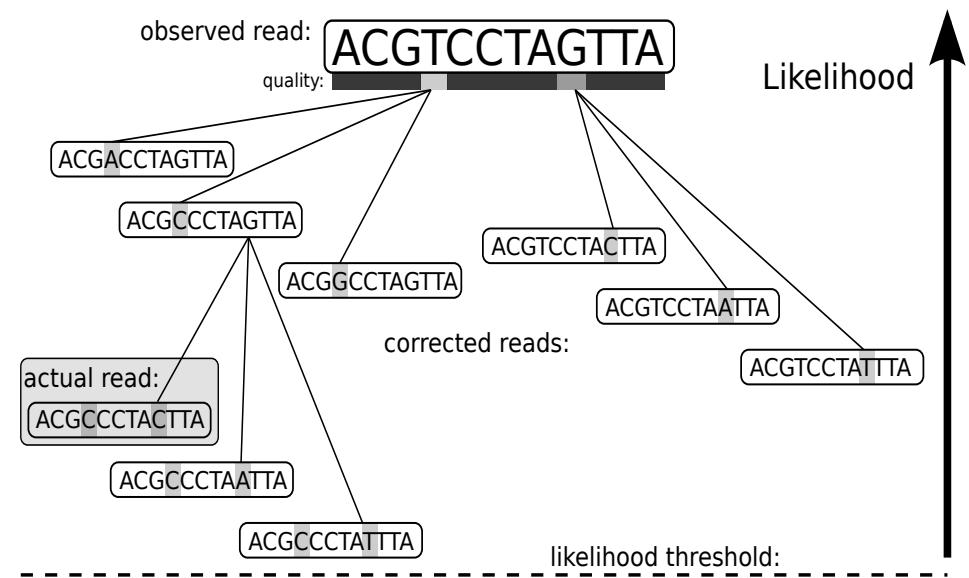
1. Count all “Q-mers” in reads

- Fit coverage distribution to mixture model of errors and regular coverage
- Automatically determines threshold for trusted k-mers



2. Correction Algorithm

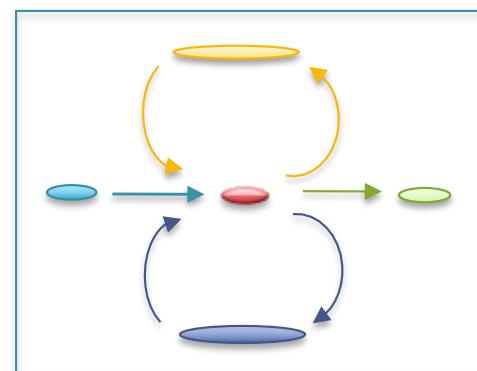
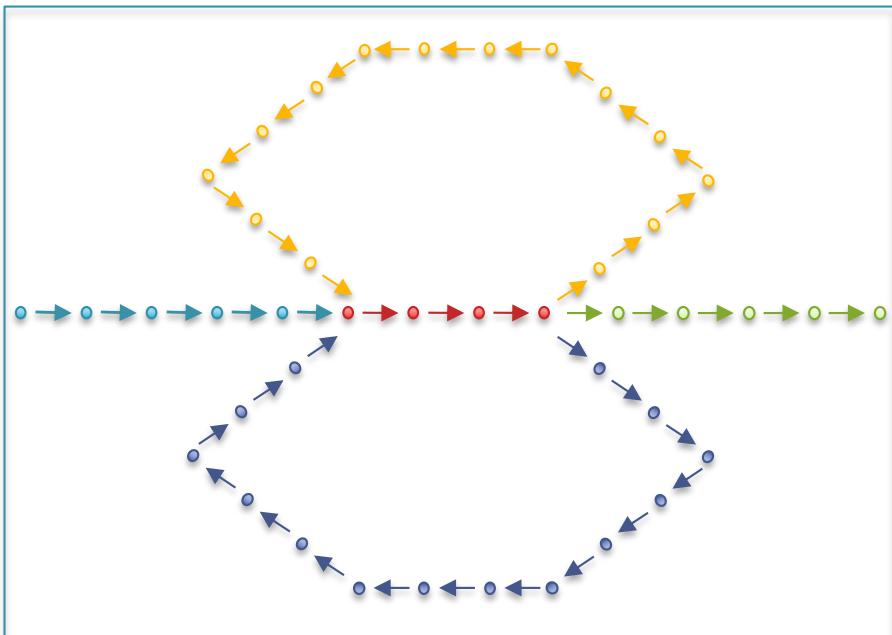
- Considers editing erroneous kmers into trusted kmers in decreasing likelihood
- Includes quality values, nucleotide/nucleotide substitution rate



Quake: quality-aware detection and correction of sequencing reads.
Kelley, DR, Schatz, MC, Salzberg SL (2010) *Genome Biology*. 11:R116

Unitigging / Unipathing

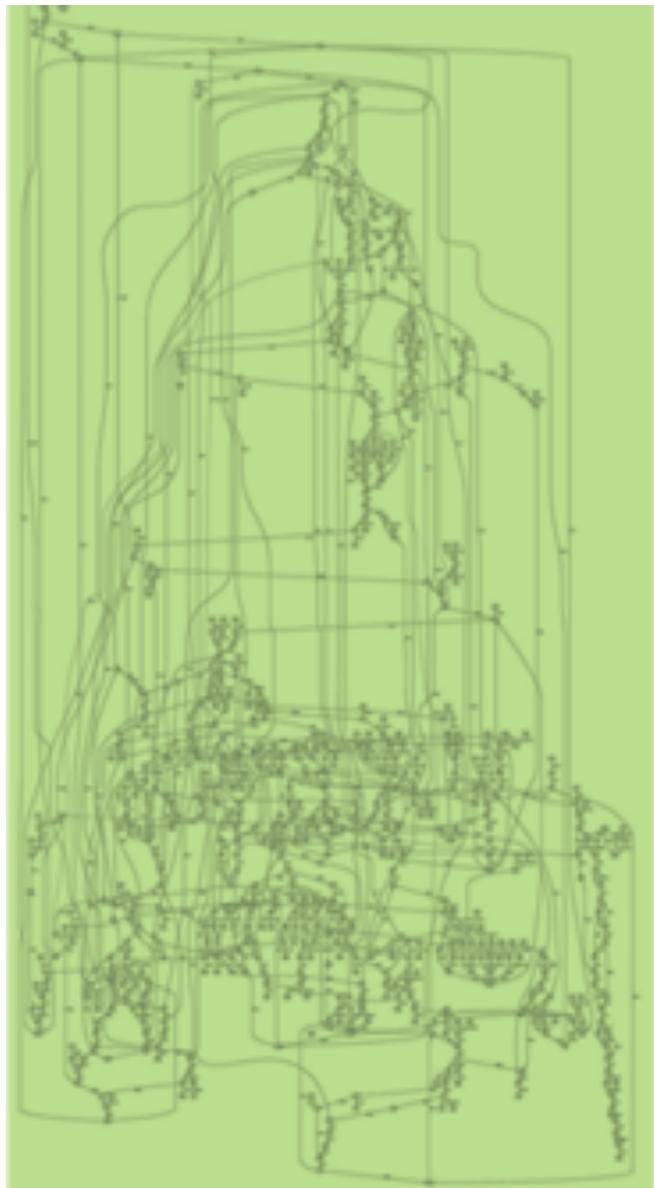
- After simplification and correction, compress graph down to its non-branching initial contigs
 - Aka “unitigs”, “unipaths”



Why do contigs end?

- (1) End of chromosome! ☺, (2) lack of coverage, (3) errors, (4) heterozygosity and (5) repeats

Errors in the graph



(Chaisson, 2009)

Clip Tips

was the worst of times,

was the worst of **tymes**,

the worst of times, it

Pop Bubbles

was the worst of times,

was the worst of **tymes**,

times, it was the age

tymes, it was the age

the worst of **tymes**,

was the worst of

the worst of times,

worst of times, it

tymes,

was the worst of

it was the age

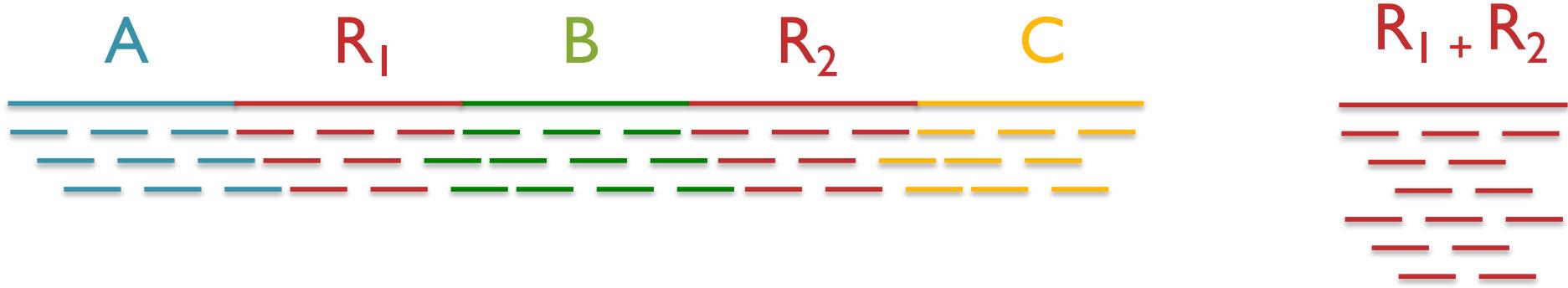
times,

Repetitive regions

Repeat Type	Definition / Example	Prevalence
Low-complexity DNA / Microsatellites	$(b_1 b_2 \dots b_k)^N$ where $1 \leq k \leq 6$ CACACACACACACACACA	2%
SINEs (Short Interspersed Nuclear Elements)	<i>Alu</i> sequence (~280 bp) <i>Mariner</i> elements (~80 bp)	13%
LINEs (Long Interspersed Nuclear Elements)	~500 – 5,000 bp	21%
LTR (long terminal repeat) retrotransposons	Ty1-copia, Ty3-gypsy, Pao-BEL (~100 – 5,000 bp)	8%
Other DNA transposons		3%
Gene families & segmental duplications		4%

- Over 50% of mammalian genomes are repetitive
 - Large plant genomes tend to be even worse
 - Wheat: 16 Gbp; Pine: 24 Gbp

Repeats and Coverage Statistics



- If n reads are a uniform random sample of the genome of length G , we expect $k = n \Delta/G$ reads to start in a region of length Δ .
 - If we see many more reads than k (if the arrival rate is $> A$) , it is likely to be a collapsed repeat

$$\Pr(X - copy) = \binom{n}{k} \left(\frac{X\Delta}{G} \right)^k \left(\frac{G - X\Delta}{G} \right)^{n-k}$$

$$A(\Delta, k) = \ln \left(\frac{\Pr(1 - copy)}{\Pr(2 - copy)} \right) = \ln \left(\frac{\frac{(\Delta n / G)^k e^{-\Delta n}}{k!}}{\frac{(2\Delta n / G)^k e^{-2\Delta n}}{k!}} \right) = \frac{n\Delta}{G} - k \ln 2$$

The fragment assembly string graph

Myers, EW (2005) Bioinformatics. 21(suppl 2): ii79-85.

Paired-end and Mate-pairs

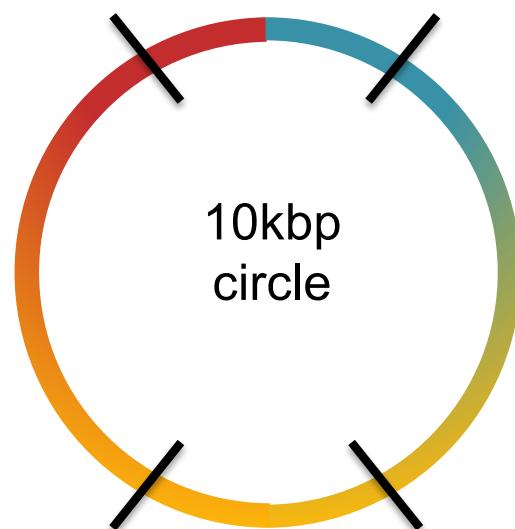
Paired-end sequencing

- Read one end of the molecule, flip, and read the other end
- Generate pair of reads separated by up to 500bp with inward orientation



Mate-pair sequencing

- Circularize long molecules (1-10kbp), shear into fragments, & sequence
- Mate failures create short paired-end reads



2x100 @ ~10kbp (outies)

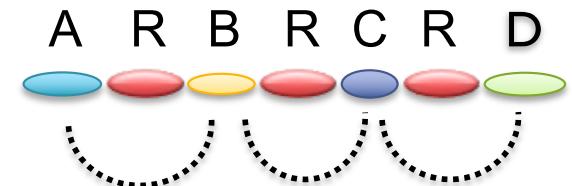
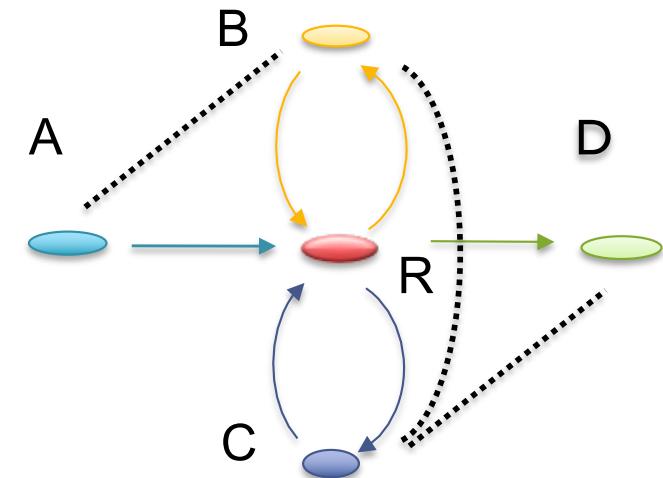


2x100 @ 300bp (innies)



Scaffolding

- Initial contigs (*aka* unipaths, unitigs) terminate at
 - Coverage gaps: especially extreme GC
 - Conflicts: errors, repeat boundaries
- Use mate-pairs to resolve correct order through assembly graph
 - Place sequence to satisfy the mate constraints
 - Mates through repeat nodes are tangled
- Final scaffold may have internal gaps called sequencing gaps
 - We know the order, orientation, and spacing, but just not the bases. Fill with Ns instead



Why do scaffolds end?

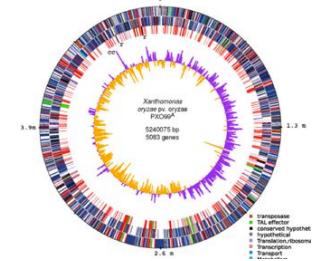
Assemblathon Results

ID	Overall	CPNG50	SPNG50	Struct.	CC50	Subs.	Copy. Num.	Cov. Tot.	Cov. CDS
BGI	36	★					★	★	★
Broad	37	★	★	★	★				
WTSI-S	46		★	★	★	★			
CSHL	52	★							★
BCCGSC	53						★	★	
DOEJGI	56		★	★	★	★			
RHUL	58								

- SOAPdenovo and ALLPATHS came out neck-and-neck followed closely behind by SGA, Celera Assembler, ABySS
- My recommendation for “typical” short read assembly is to use ALLPATHS or Spades

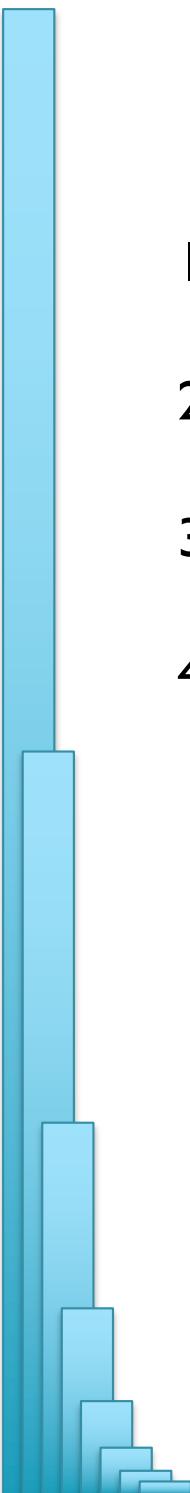
Assemblathon I: A competitive assessment of de novo short read assembly methods
Earl et al. (2011) Genome Research. 21: 2224-2241

Assembly Summary



Assembly quality depends on

1. **Coverage:** low coverage is mathematically hopeless
 2. **Repeat composition:** high repeat content is challenging
 3. **Read length:** longer reads help resolve repeats
 4. **Error rate:** errors reduce coverage, obscure true overlaps
-
- Assembly is a hierarchical, starting from individual reads, build high confidence contigs/unitigs, incorporate the mates to build scaffolds
 - Extensive error correction is the key to getting the best assembly possible from a given data set
 - Watch out for collapsed repeats & other misassemblies
 - Globally/Locally reassemble data from scratch with better parameters & stitch the 2 assemblies together



Next Steps

1. Reflect on the magic and power of DNA 😊
2. Check out the course webpage
3. Register on Piazza
4. Work on Assignments 1 & 2
 1. Set up Linux, set up Virtual Machine
 2. Set up Dropbox for yourself!
 3. Get comfortable on the command line