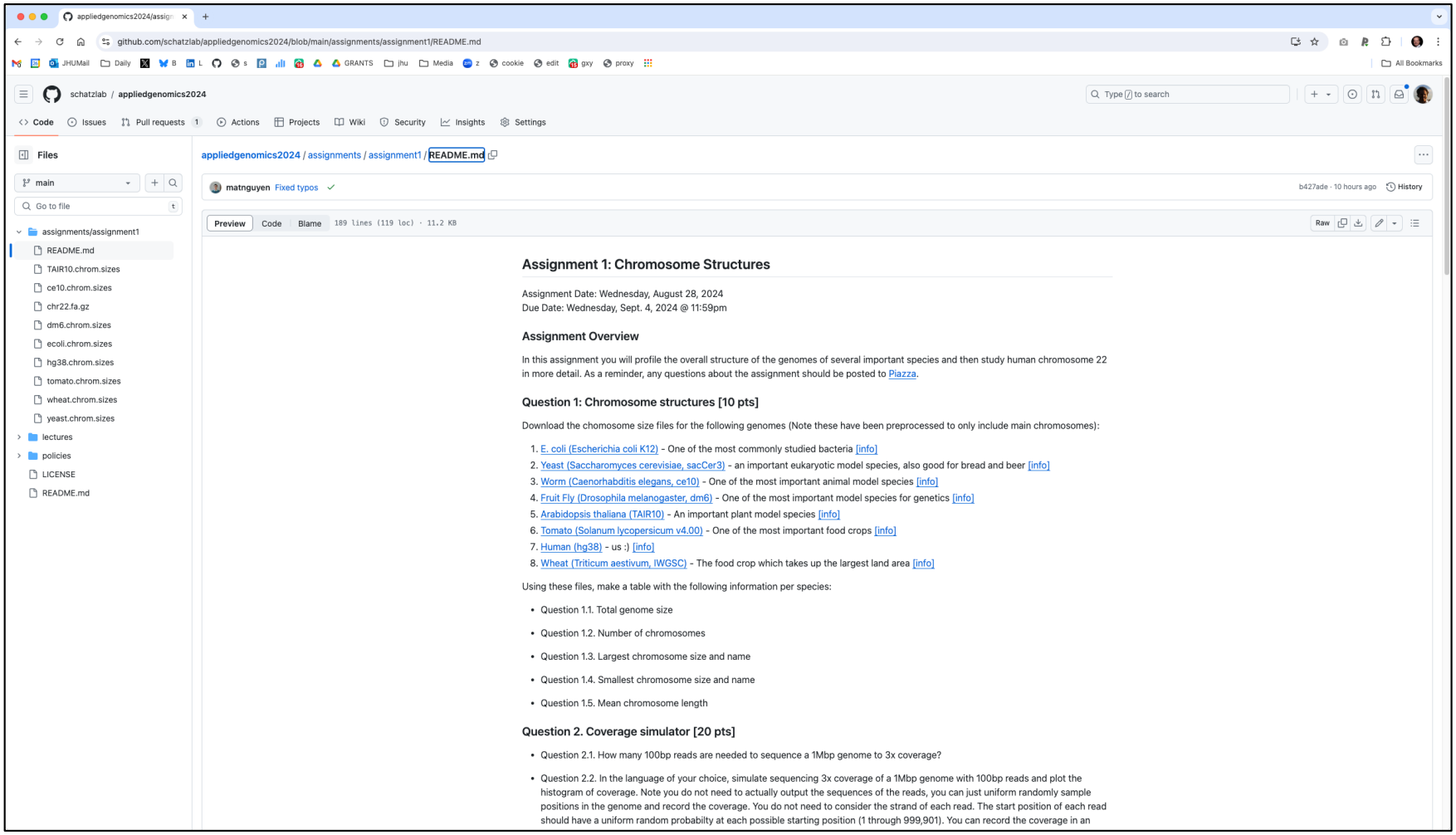
Whole Genome Assembly and Alignment

Michael Schatz

Sept 8, 2024

Lecture 4: Applied Comparative Genomics

Assignment 1



Assignment 1: Chromosome Structures

Assignment Date: Wednesday, August 28, 2024
Due Date: Wednesday, Sept. 4, 2024 @ 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 [10 pts]

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

1. E. coli (Escherichia coli K12) - One of the most commonly studied bacteria [info]
2. Yeast (Saccharomyces cerevisiae, sacCer3) - an important eukaryotic model species, also good for bread and beer [info]
3. Worm (Caenorhabditis elegans, ce10) - One of the most important animal model species [info]
4. Fruit Fly (Drosophila melanogaster, dm6) - One of the most important model species for genetics [info]
5. Arabidopsis thaliana (TAIR10) - An important plant model species [info]
6. Tomato (Solanum lycopersicum v4.00) - One of the most important food crops [info]
7. Human (hg38) - us :) [info]
8. Wheat (Triticum aestivum, IWGSC) - The food crop which takes up the largest land area [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

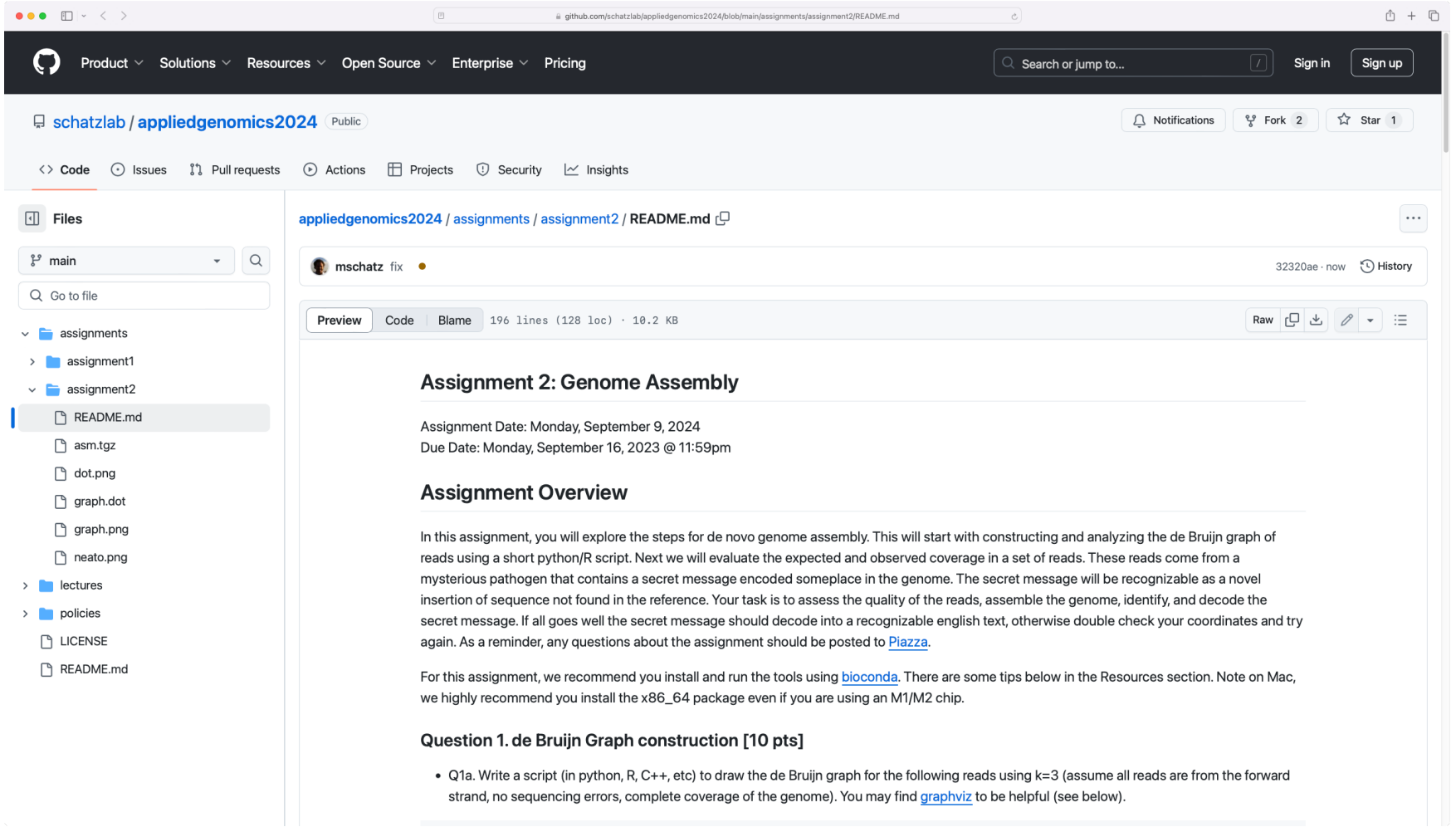
Question 2. Coverage simulator [20 pts]

- Question 2.1. How many 100bp reads are needed to sequence a 1Mbp genome to 3x coverage?
- Question 2.2. In the language of your choice, simulate sequencing 3x coverage of a 1Mbp genome with 100bp reads and plot the histogram of coverage. Note you do not need to actually output the sequences of the reads, you can just uniform randomly sample positions in the genome and record the coverage. You do not need to consider the strand of each read. The start position of each read should have a uniform random probabilty at each possible starting position (1 through 999,901). You can record the coverage in an

https://github.com/schatzlab/appliedgenomics2024/tree/main/assignments/assignment1

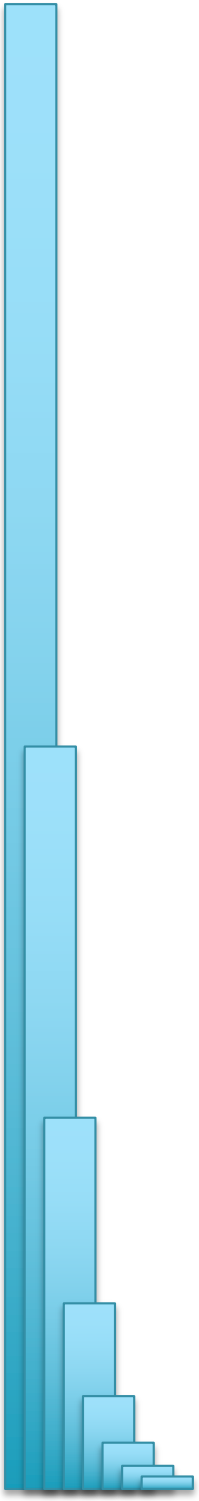
Due end of day on Sept 4 (right before midnight)

Assignment 2: Genome Assembly
Due Monday Sept 16 by 11:59pm



https://github.com/schatzlab/appliedgenomics2024/tree/main/assignments/assignment2

Check Piazza for questions!

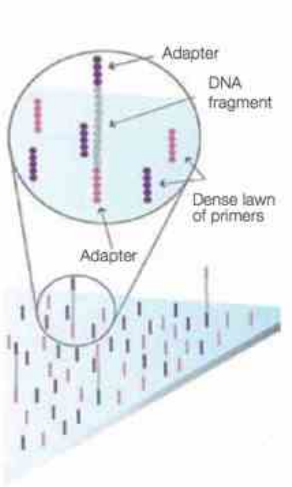


Part 1: Recap

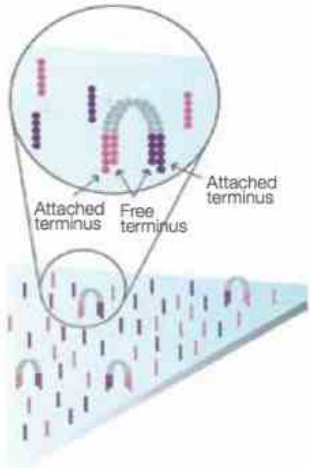
Second Generation Sequencing

Illumina NovaSeq 6000
Sequencing by Synthesis

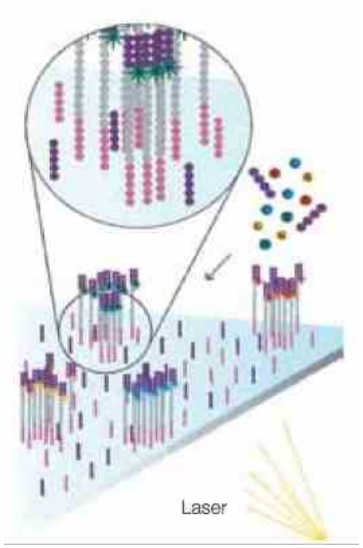
>3Tbp / day
(JHU has 4 of these!)



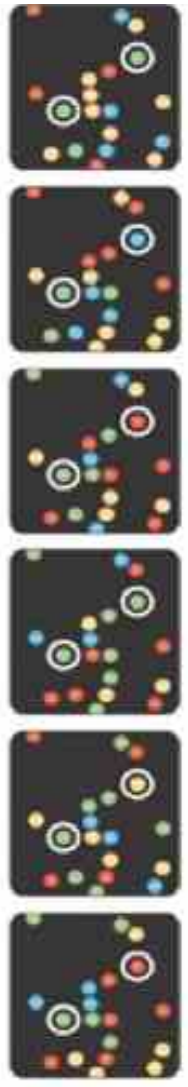
1. Attach



2. Amplify

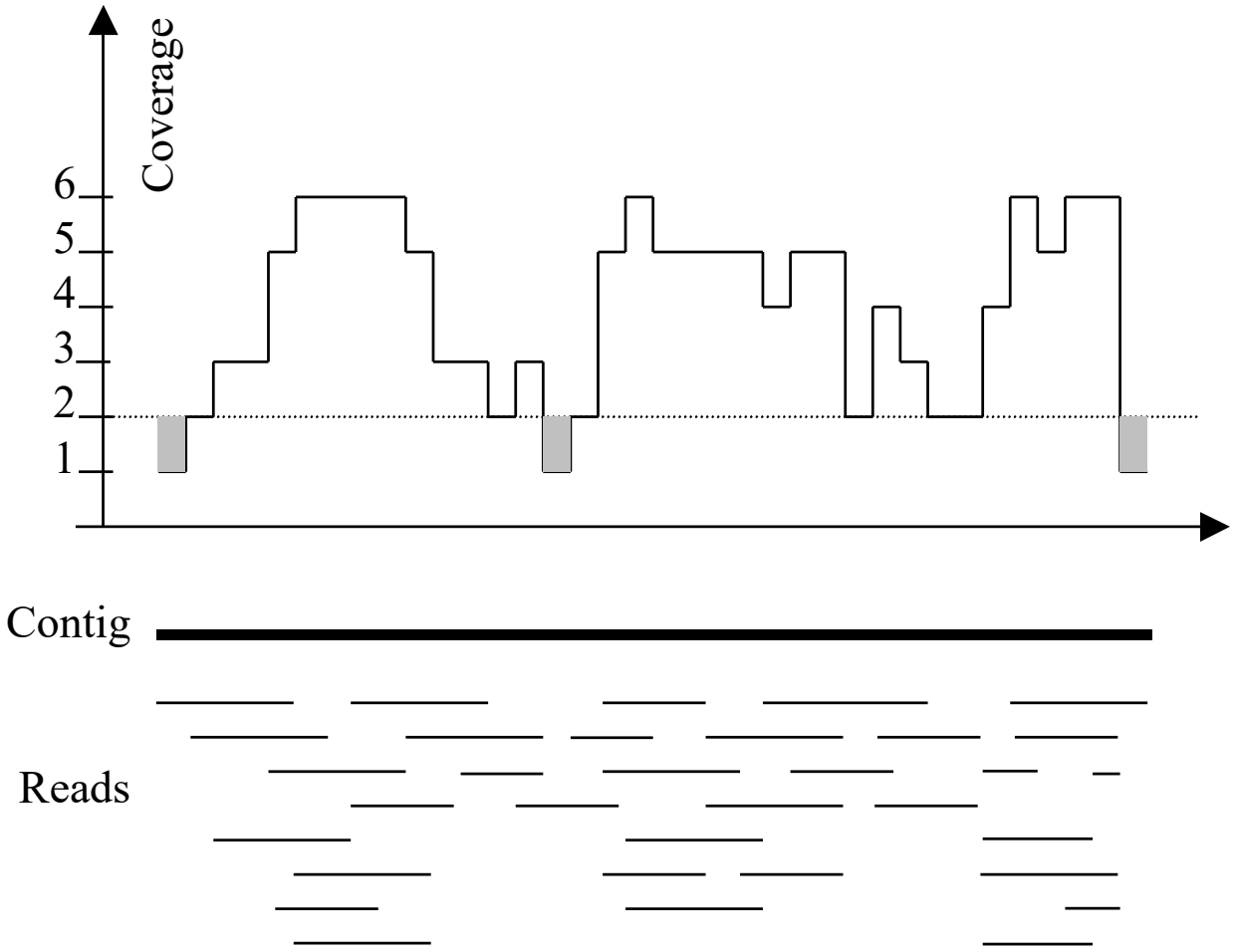


3. Image



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

Typical sequencing coverage



Imagine raindrops on a sidewalk

We want to cover the entire sidewalk but each drop costs $1

If the genome is 10 Mbp, should we sequence 100k 100bp reads?

Poisson Distribution

The probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and independently of the time since the last event.

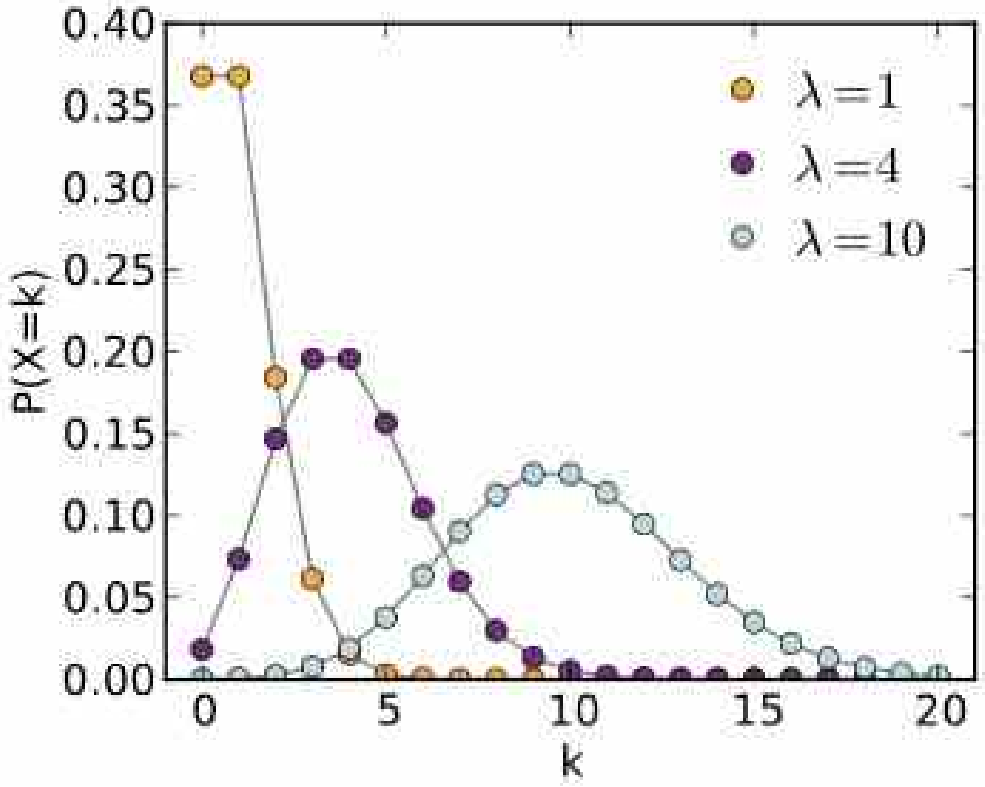
$$P(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Formulation comes from the limit of the binomial equation

Resembles a normal distribution, but over the positive values, and with only a single parameter.

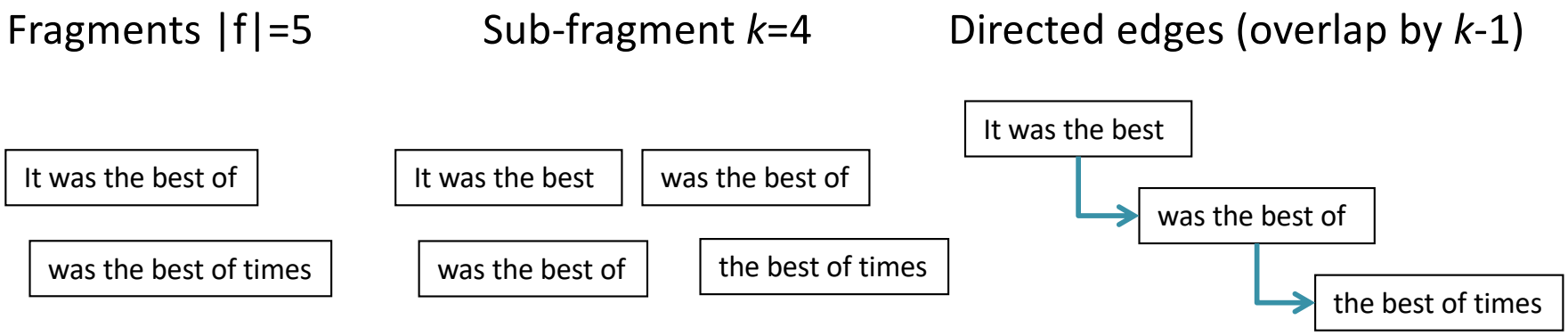
Key properties:

- ***The standard deviation is the square root of the mean.***
- ***For mean > 5, well approximated by a normal distribution***



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

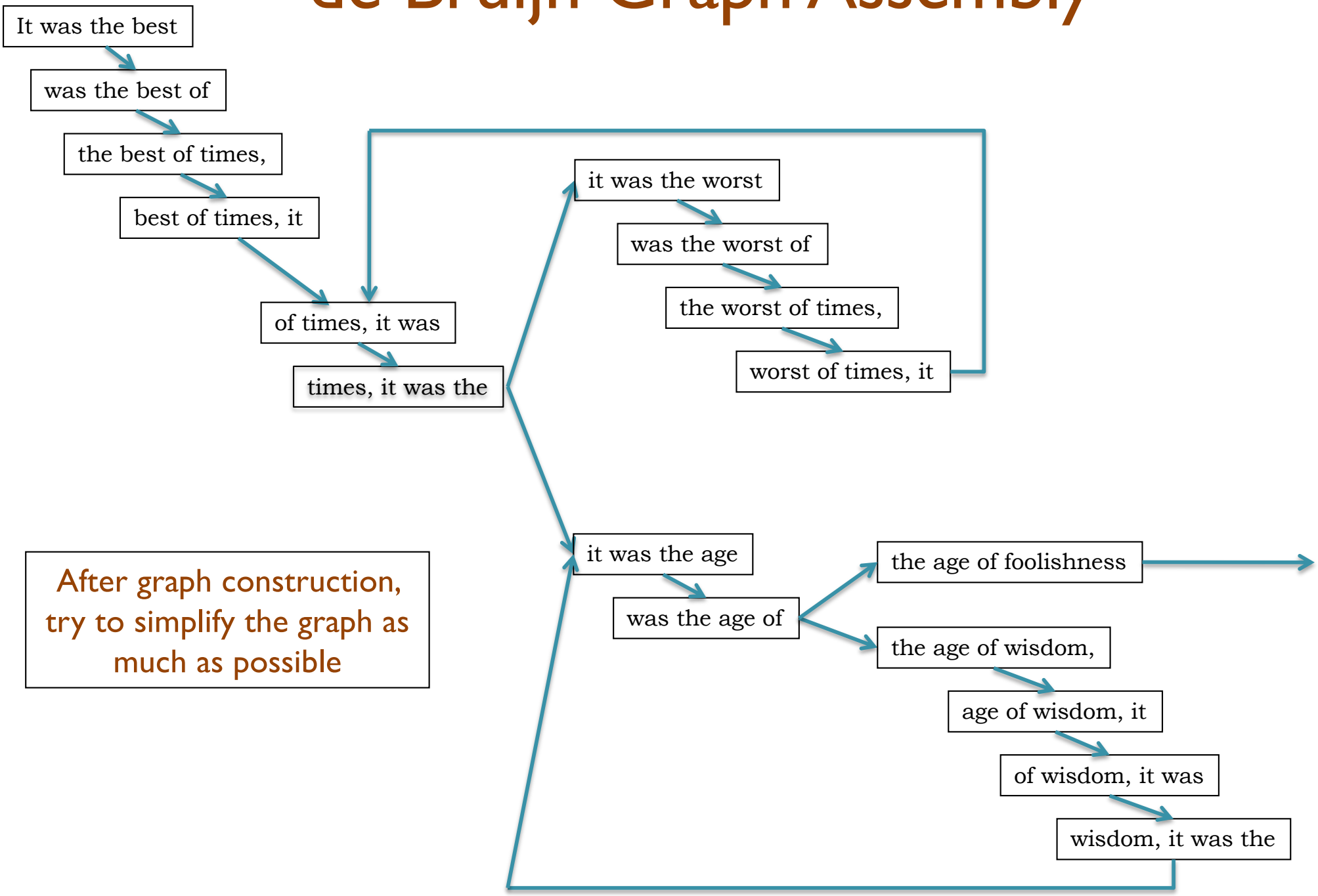


– Overlaps between fragments are implicitly computed

How to pronounce:
https://forvo.com/word/de_bruijn/

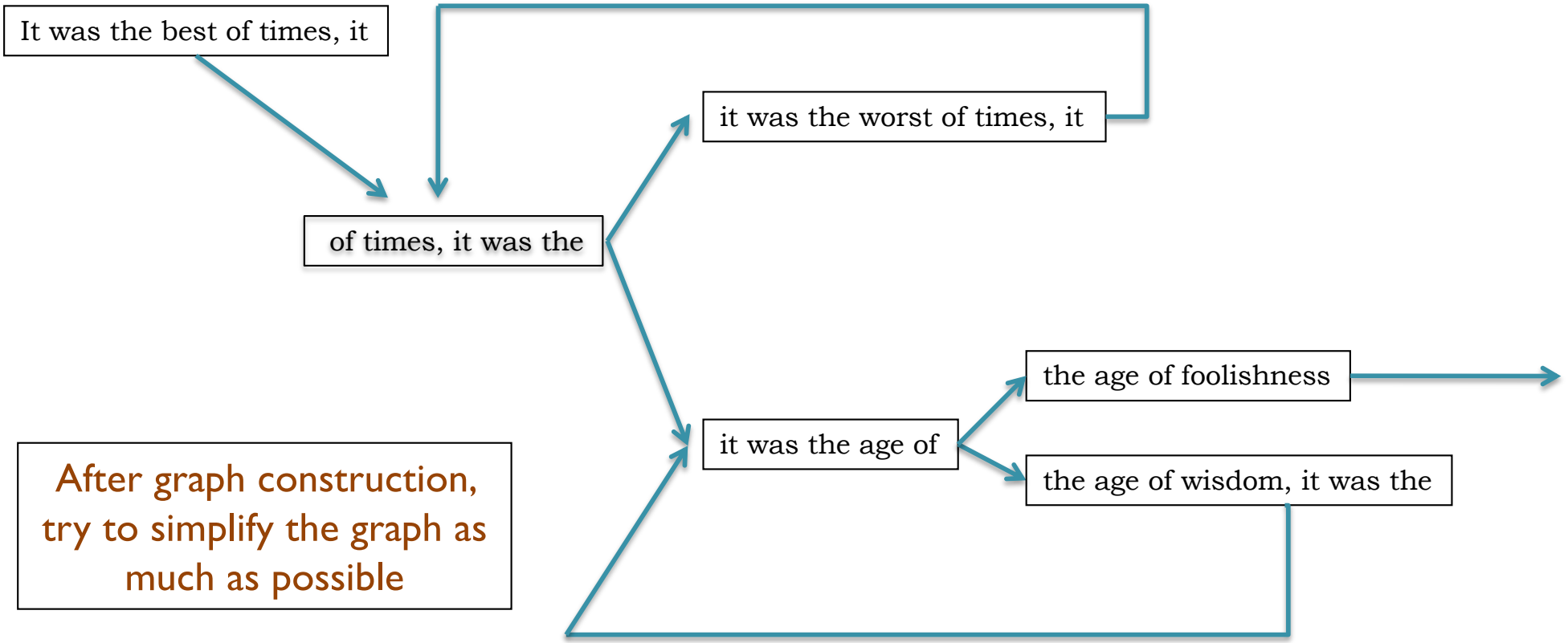
de Bruijn, 1946
Idury et al., 1995
Pevzner et al., 2001

de Bruijn Graph Assembly



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

de Bruijn Graph Assembly

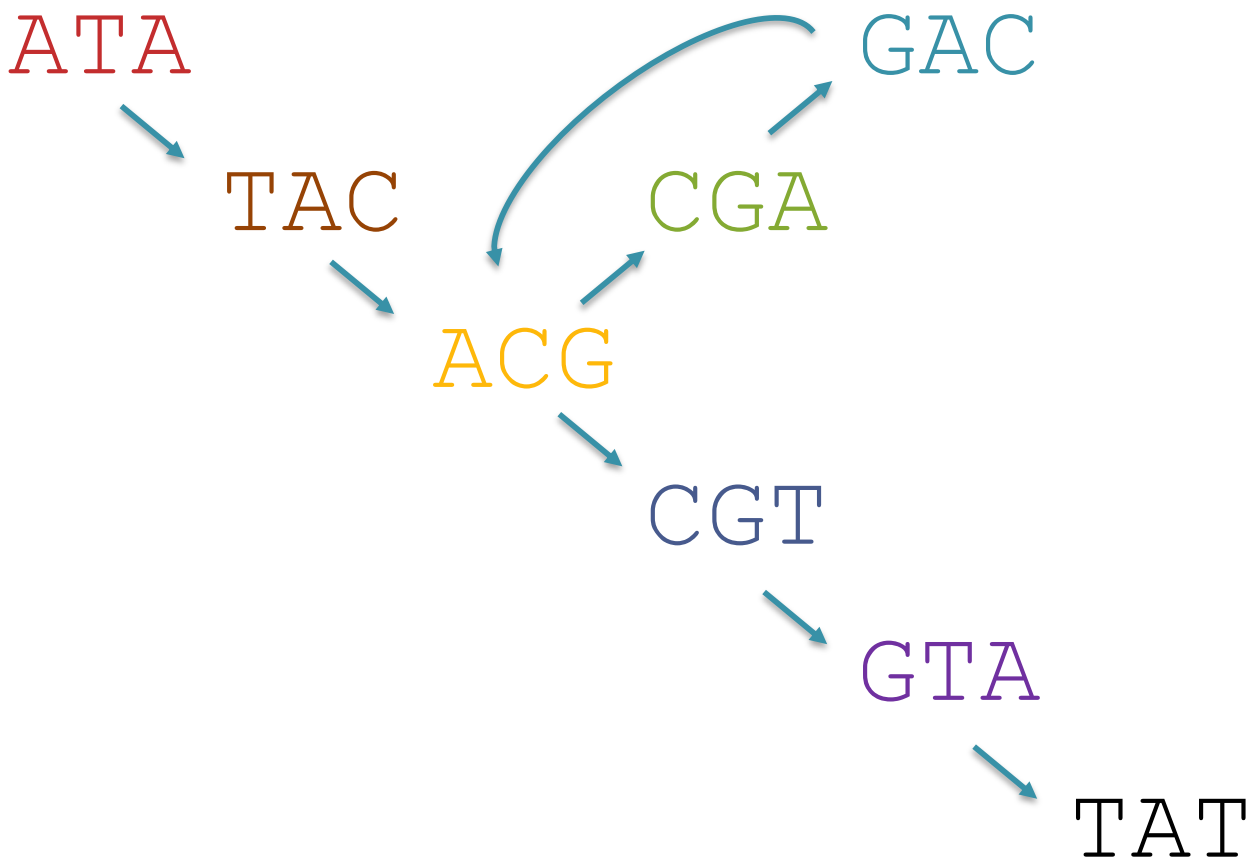


After graph construction, try to simplify the graph as much as possible

Pop Quiz 2

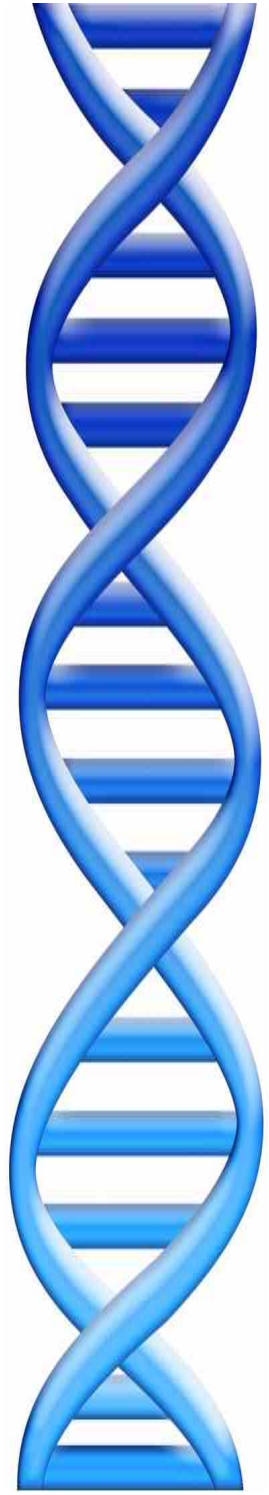
Assemble these reads using a de Bruijn graph approach (k=3):

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



Should we add the edge TAT -> ATA?

ATACGACGTAT



Outline

1. ***Assembly theory***
   - Assembly by analogy

2. ***Practical Issues***
   - Coverage, read length, errors, and repeats

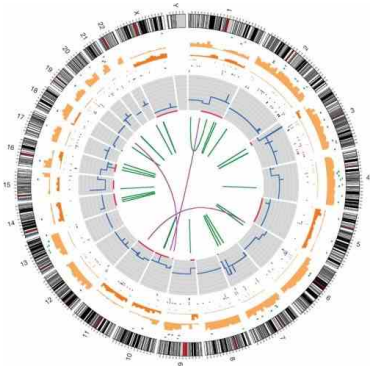
3. Whole Genome Alignment
   - MUMmer recommended

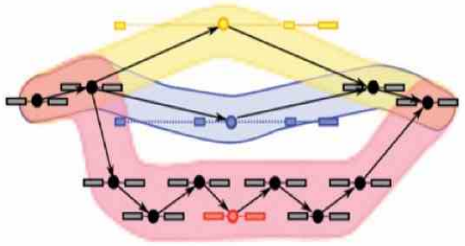
Assembly Applications

- Novel genomes

- Metagenomes

- Sequencing assays
  - Structural variations
  - Transcript assembly
  - …





Why are genomes hard to assemble?

1. ***Biological***:
   - (Very) High ploidy, heterozygosity, repeat content

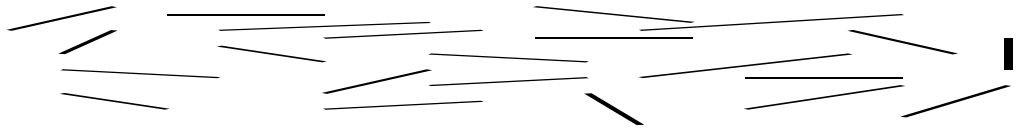
2. ***Sequencing***:
   - (Very) large genomes, imperfect sequencing

3. ***Computational***:
   - (Very) Large genomes, complex structure

4. ***Accuracy***:
   - (Very) Hard to assess correctness

Assembling a Genome

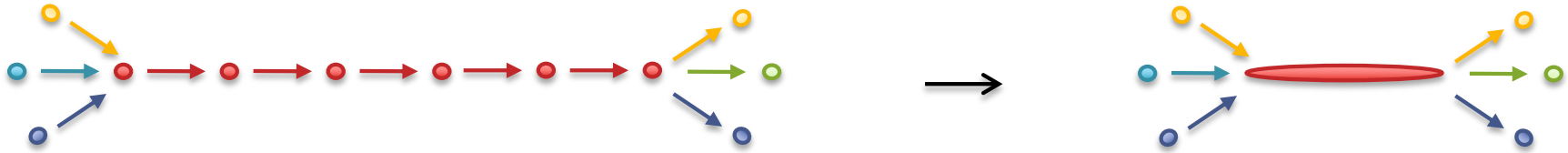
1. Shear & Sequence DNA



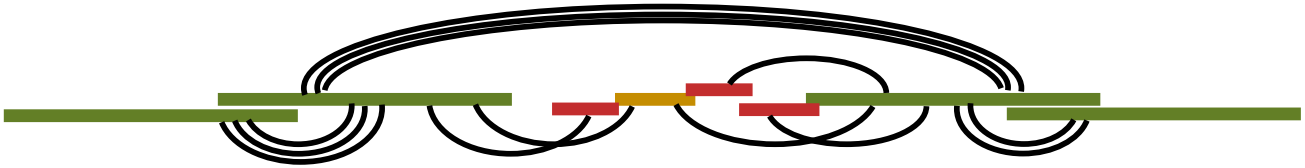
2. Construct assembly graph from reads (de Bruijn / overlap graph)

...AGCCTAGGGATGCGCGACACGT
GGATGCGCGACACGTCGCATATCCGGTTTGGTCAACCTCGGACGGAC
CAACCTCGGACGGACCTCAGCGAA...

3. Simplify assembly graph

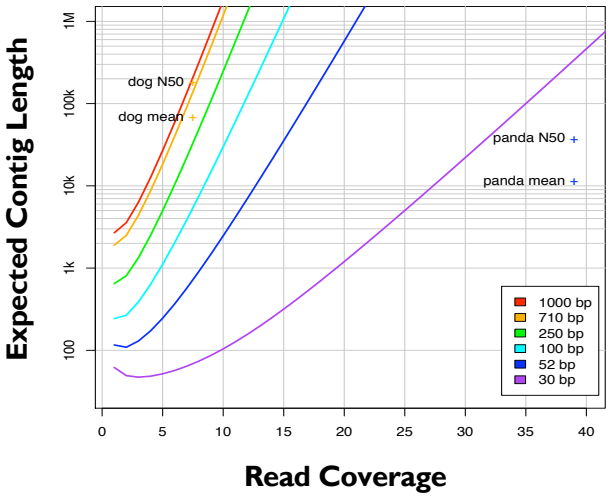


4. Detangle graph with long reads, mates, and other links



Ingredients for a good assembly

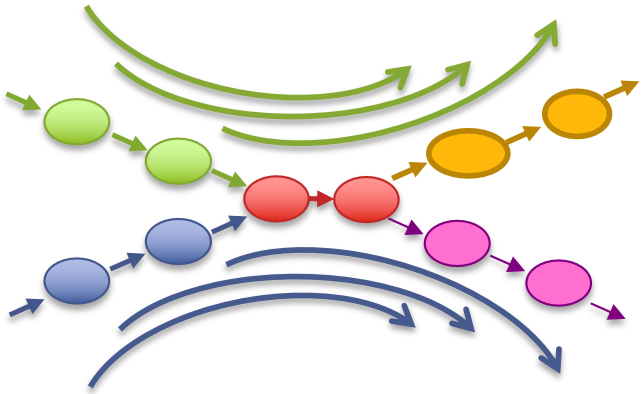
Coverage



High coverage is required

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

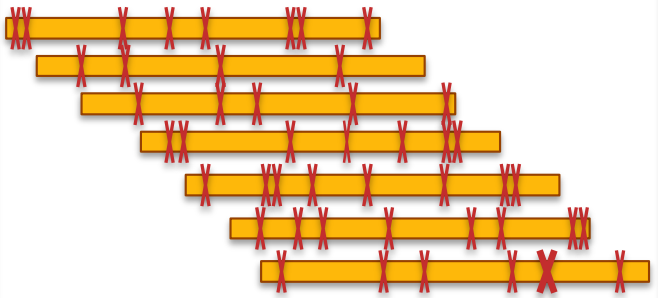
Read Length



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

$$\text{sequencing\_coverage} = \frac{\text{total\_bases\_sequenced}}{\text{genome\_size}}$$

$$\text{genome\_size} = \frac{\text{total\_bases\_sequenced}}{\text{sequencing\_coverage}}$$

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

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
```

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
```

sort count

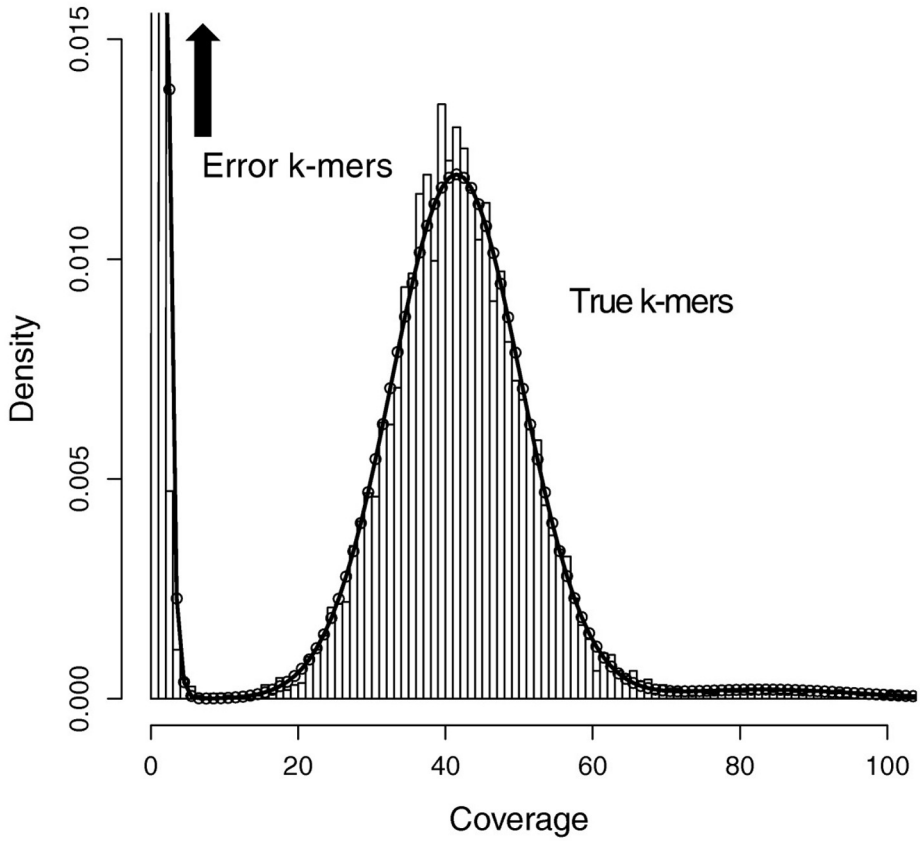
tally

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

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

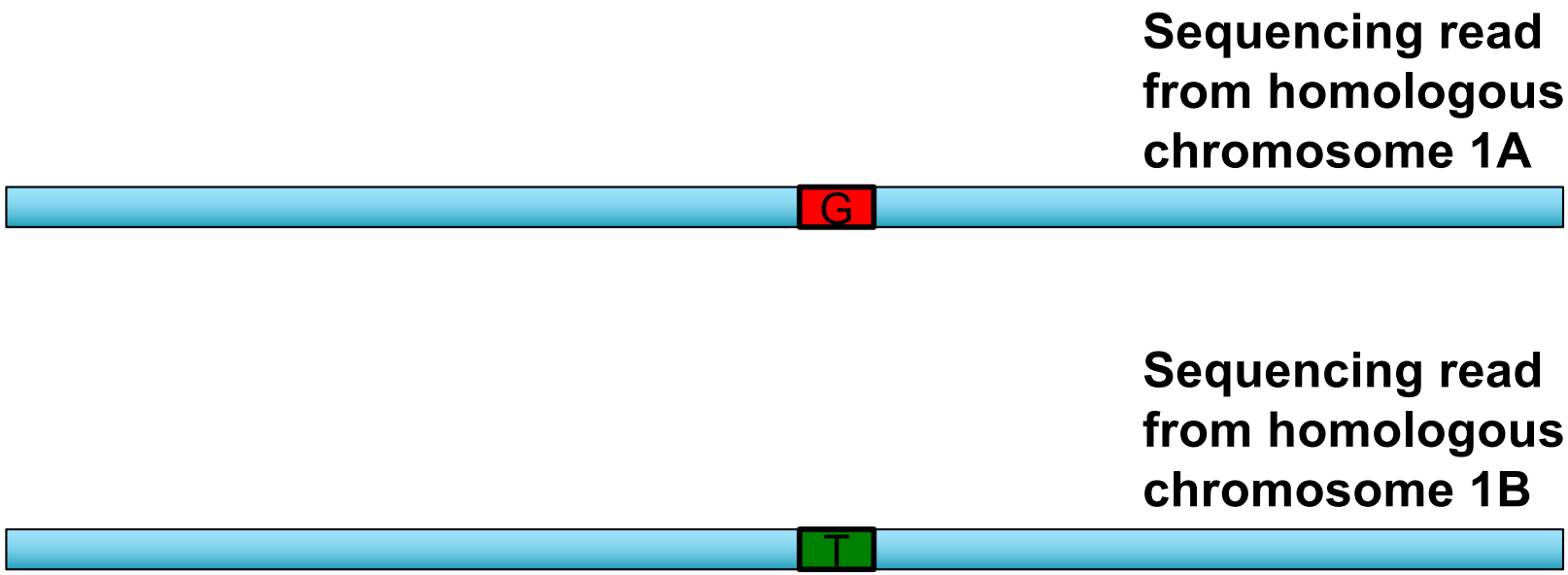
Fast to compute even over huge datasets

K-mer counting in real genomes

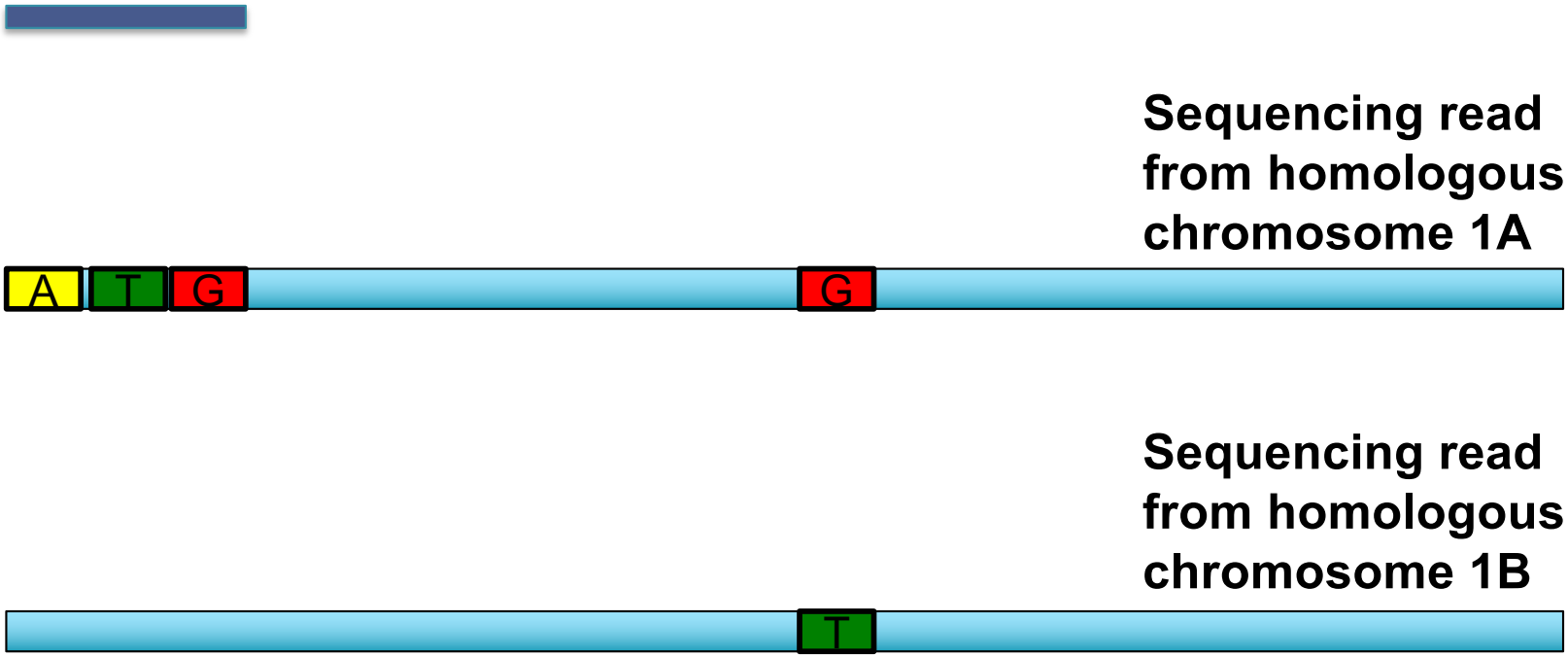


- The tally of k-mer counts in real genomes reveals the coverage distribution.
- Here we sequenced 120Gb of reads from a female human (haploid human genome size is 3Gb), and indeed we see a clear peak centered at 40x coverage
- There are also many kmers that only occur <5 times. These are from errors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

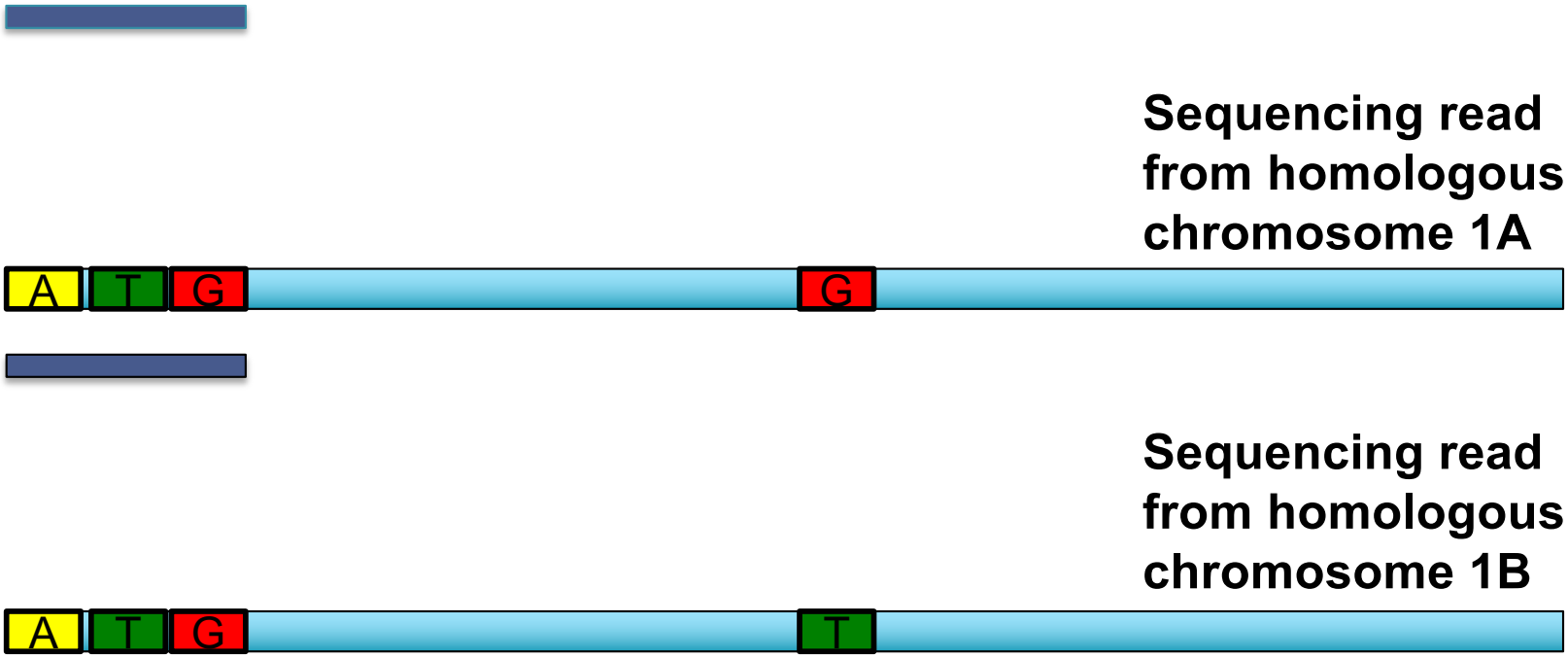
K-mer counting in heterozygous genomes



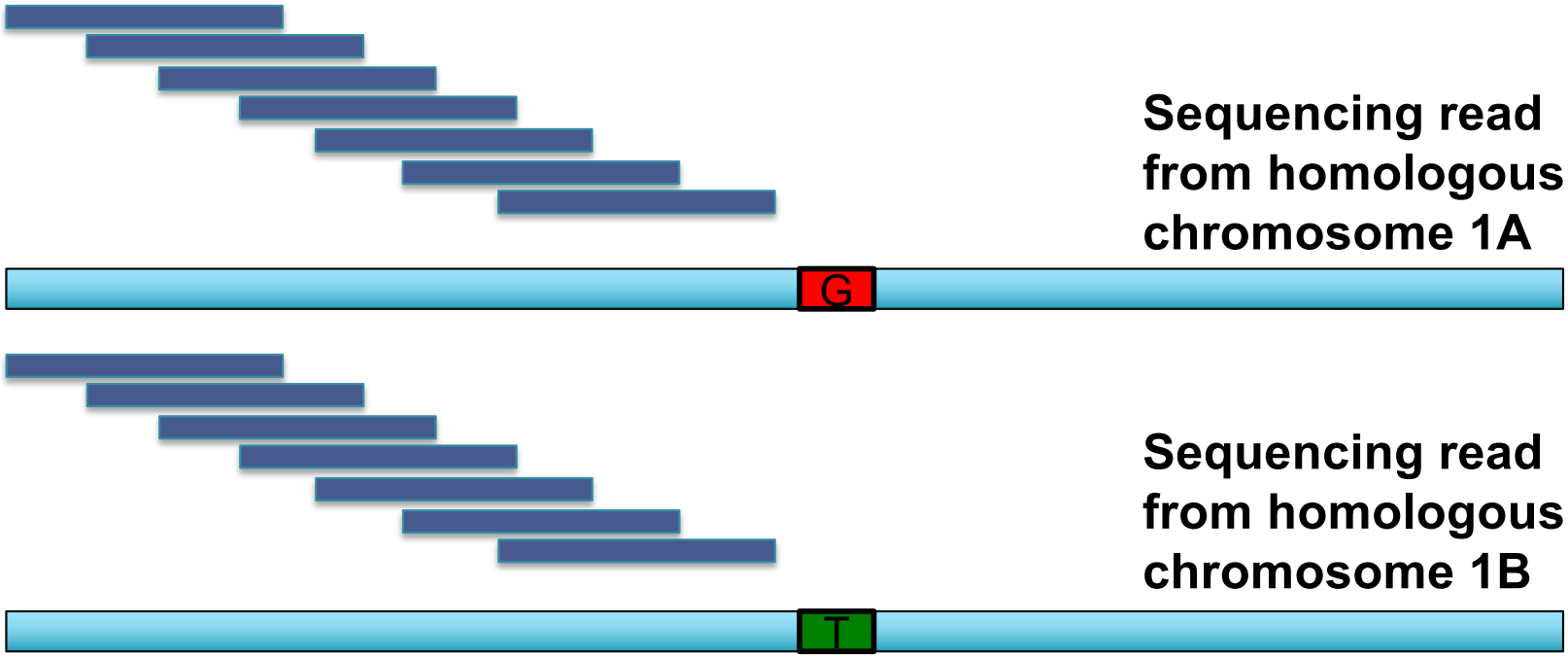
K-mer counting in heterozygous genomes



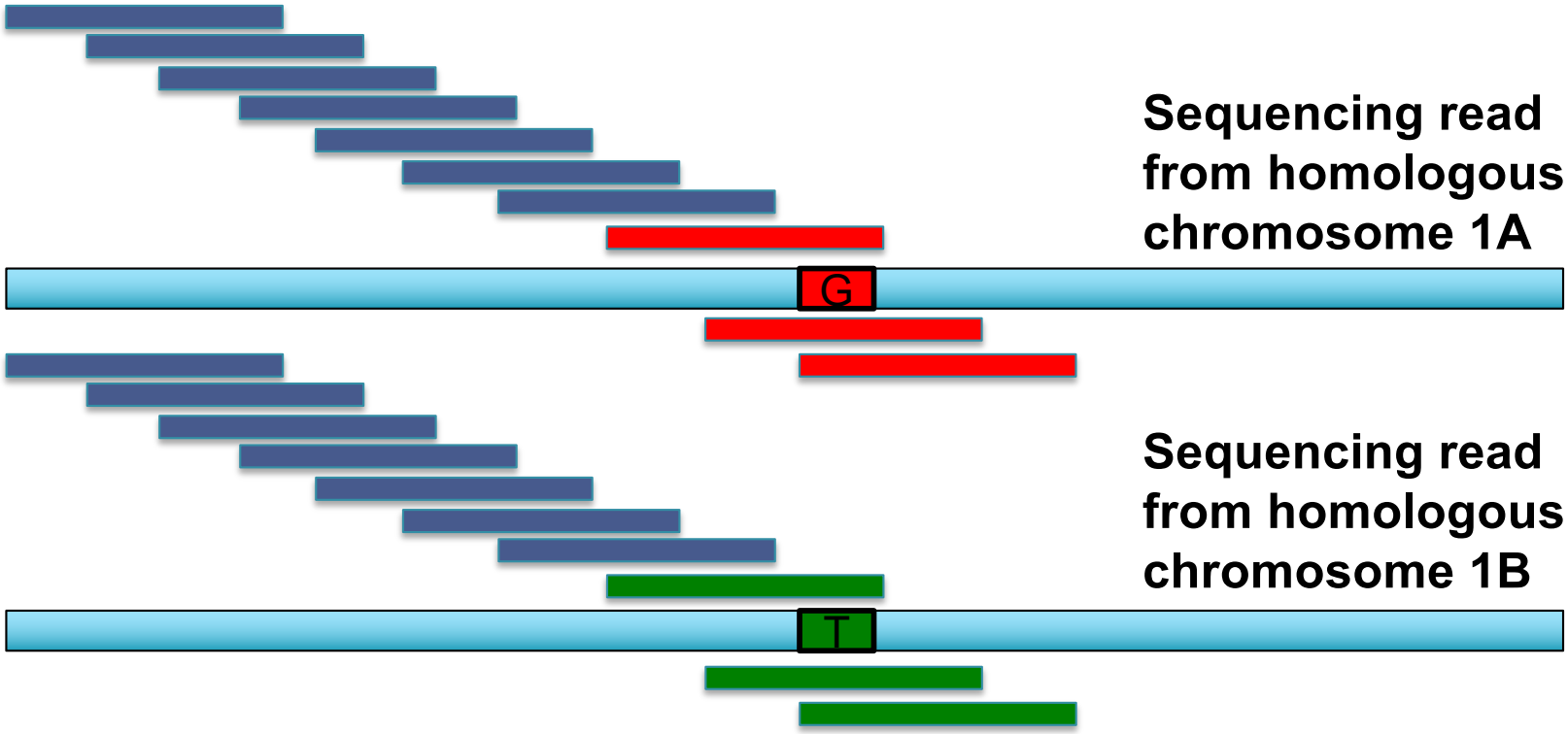
K-mer counting in heterozygous genomes



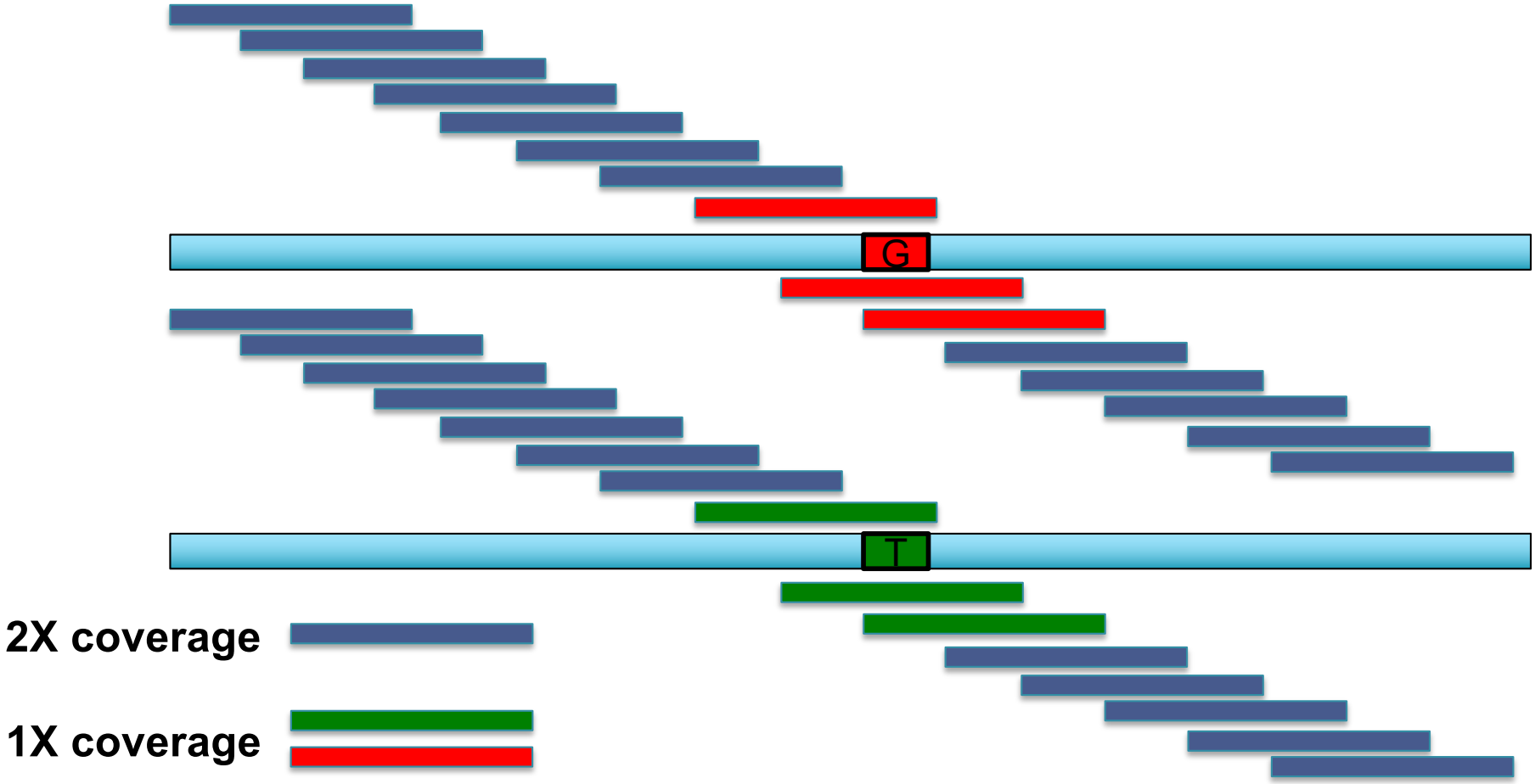
K-mer counting in heterozygous genomes



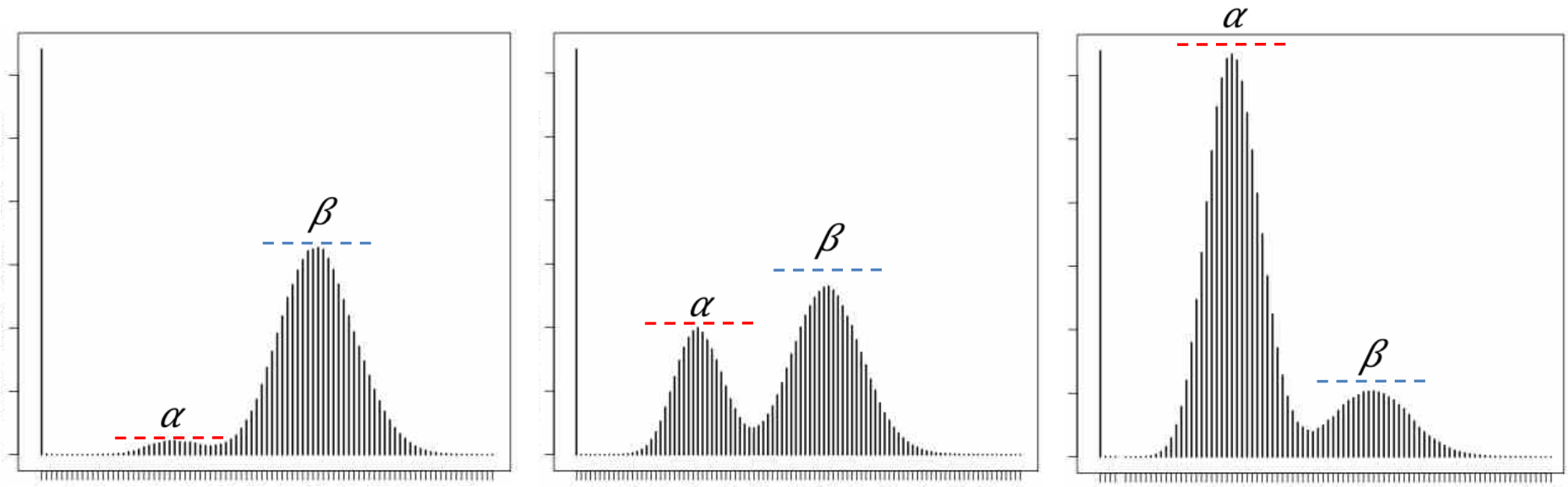
K-mer counting in heterozygous genomes



K-mer counting in heterozygous genomes



Heterozygous Kmer Profiles

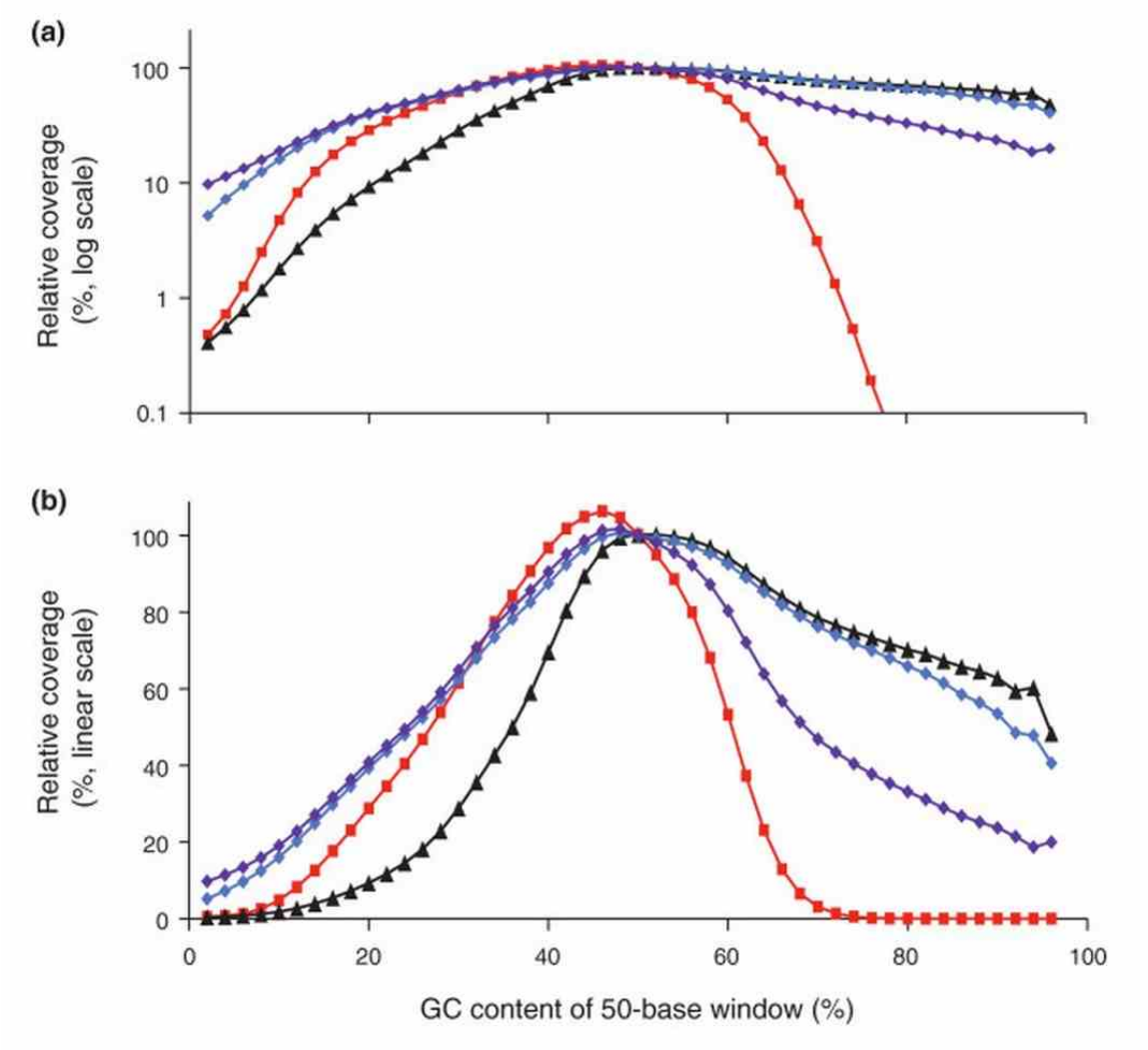


0.1% heterozygosity **1% heterozygosity** **5% heterozygosity**

- ***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)

Beware of GC Biases



Illumina sequencing does not produce uniform coverage over the genome

- Coverage of extremely high or extremely low GC content will have reduced coverage in Illumina sequencing
- Biases primarily introduced during PCR; lower temperatures, slower heating, and fewer rounds minimize biases
- This makes it very difficult to identify variants (SNPs, CNVs, etc) in certain regions of the genome

Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries.
Aird et al. (2011) *Genome Biology.* 12:R18.

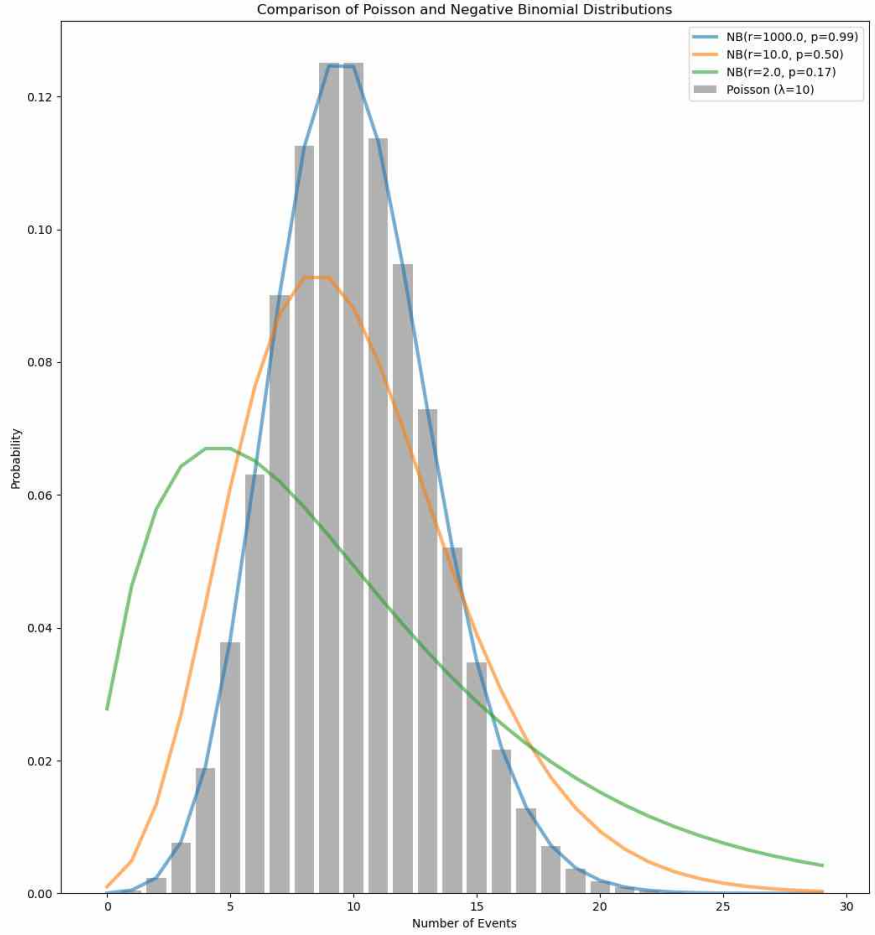
Negative Binomial Distribution

Models the number of failures in a sequence of independent and identically distributed Bernoulli trials before a specified number of successes (r) occurs

- Commonly used to model over-dispersed count data

Also arises as a continuous mixture of Poisson distributions where the mixing distribution of the Poisson rate is a gamma distribution.

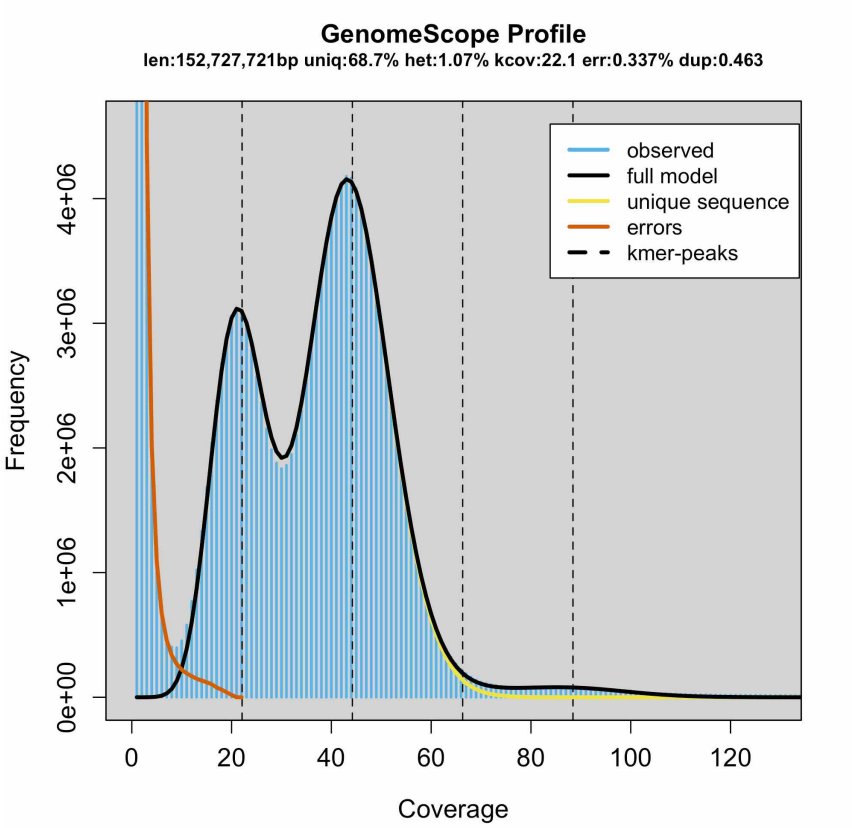
- Sequence coverage distribution where there is a non-uniform probability of a read starting at each position

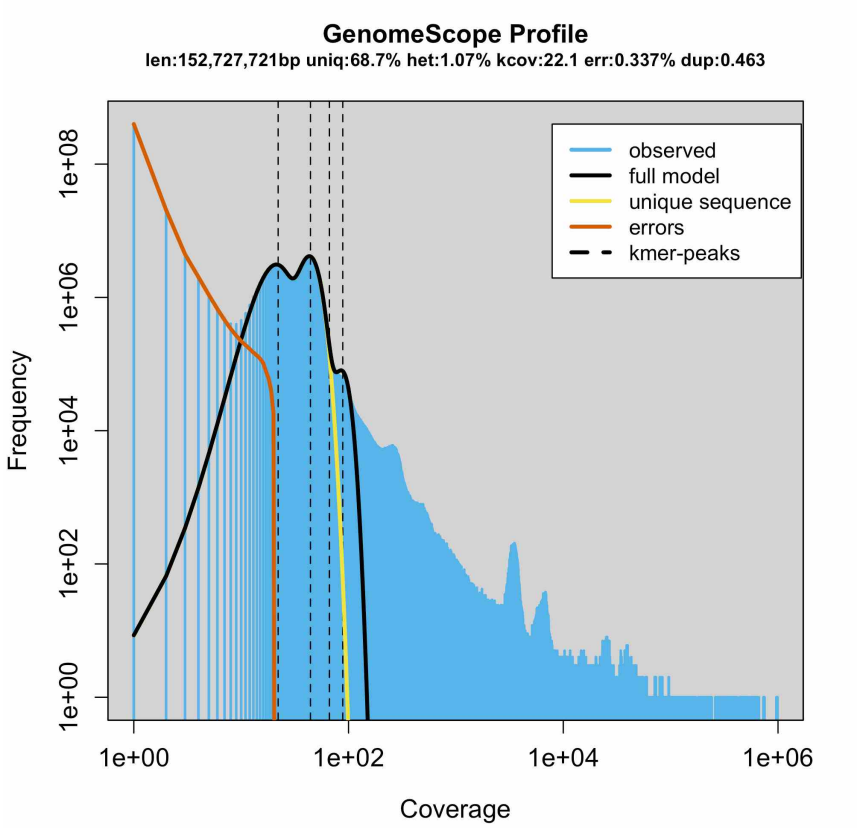


$$f(k;r,p) \equiv \Pr(X = k) = \binom{k+r-1}{k}(1-p)^k p^r$$

GenomeScope: Fast genome analysis from short reads

http://genomescope.org





$$f(k) = \alpha \cdot \mathrm{NB}(k; r_1, p_1) + \beta \cdot \mathrm{NB}(k; r_2, p_2)$$

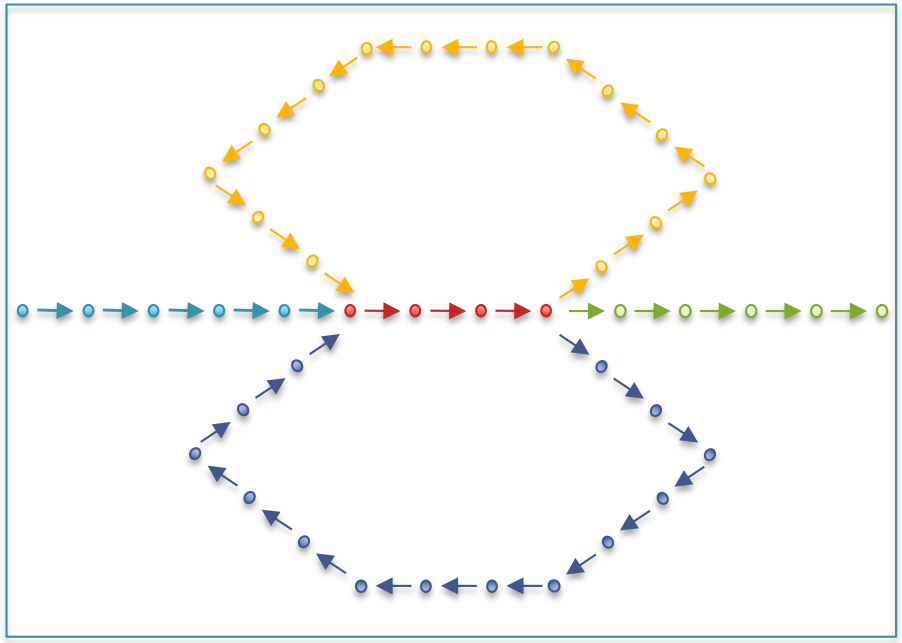
- Theoretical model agrees well with published results:
  - Quickly estimate genome size, rate of heterozygosity, and other genome properties
  - Generalized to higher ploidies by introducing additional terms
  - “Reference-free analysis” does not require the use of an assembled genome

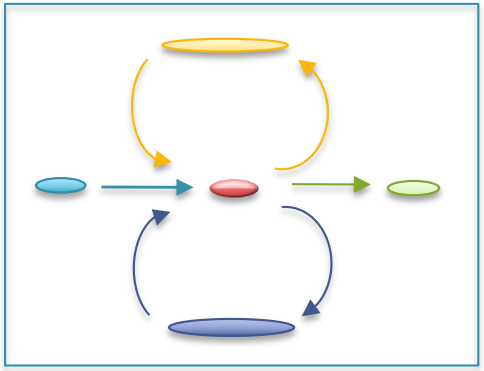
Vurture, GW*, Sedlazeck FJ*, et al. (2017) *Bioinformatics*
Ranallo-Benavidez, TR. *et al.* (2020) *Nature Communication*

Errors

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

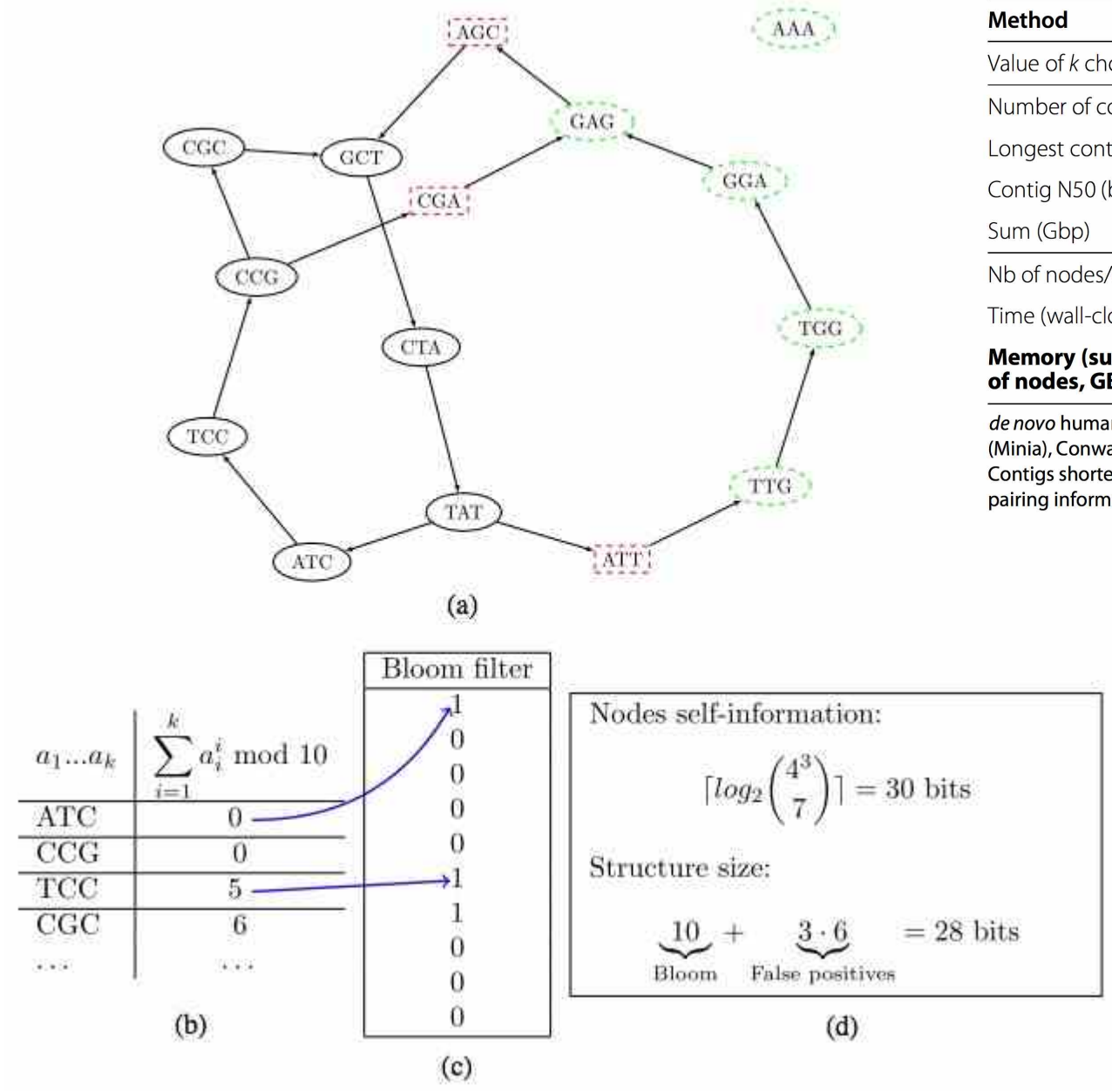


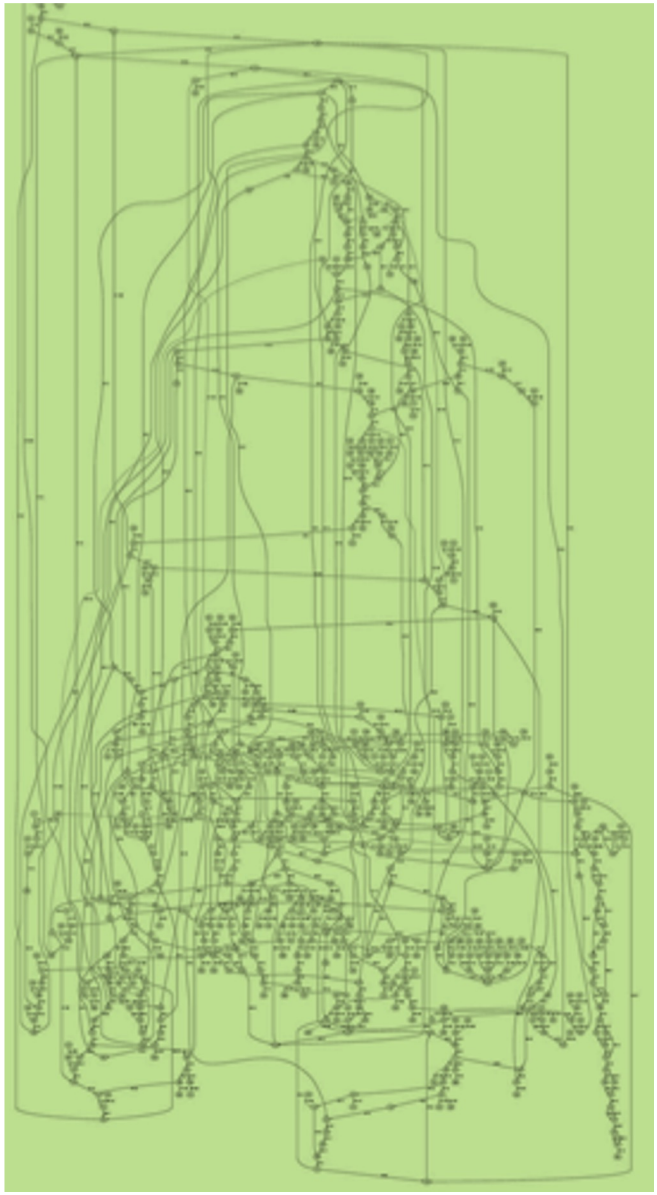
Table 2 de novo human genome (NA18507) assemblies

Method	Minia	C. & B.	ABySS	SOAPdenovo
Value of *k* chosen	27	27	27	25
Number of contigs (M)	3.49	7.69	4.35	-
Longest contig (kbp)	18.6	22.0	15.9	-
Contig N50 (bp)	1156	250	870	886
Sum (Gbp)	2.09	1.72	2.10	2.08
Nb of nodes/cores	1/1	1/8	21/168	1/16
Time (wall-clock, h)	23	50	15	33
Memory (sum of nodes, GB)	**5.7**	**32**	**336**	**140**

de novo human genome (NA18507) assemblies reported by our assembler (Minia), Conway and Bromage assembler [9], ABySS [8], and SOAPdenovo [7]. Contigs shorter than 100 bp were discarded. Assemblies were made without any pairing information.

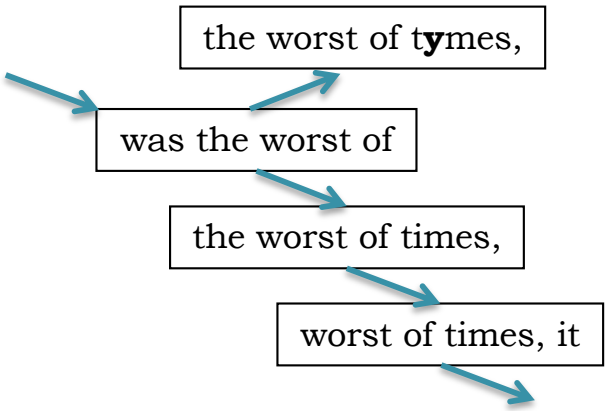
Space-efficient and exact de Bruijn graph representation based on a Bloom filter
Chikhi and Rizk (2013) *Algorithms for Molecular Biology*. 8:22

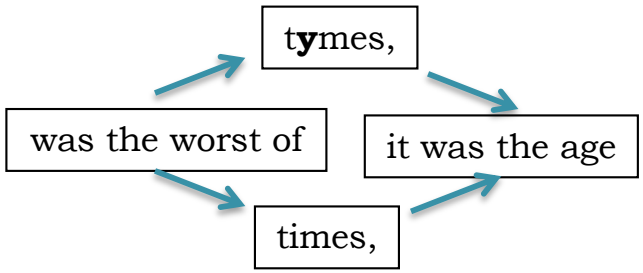
Errors in the graph



(Chaisson, 2009)

Clip Tips	Pop Bubbles
was the worst of times,	was the worst of times,
was the worst of t**y**mes,	was the worst of t**y**mes,
the worst of times, it	times, it was the age
	t**y**mes, it was the age



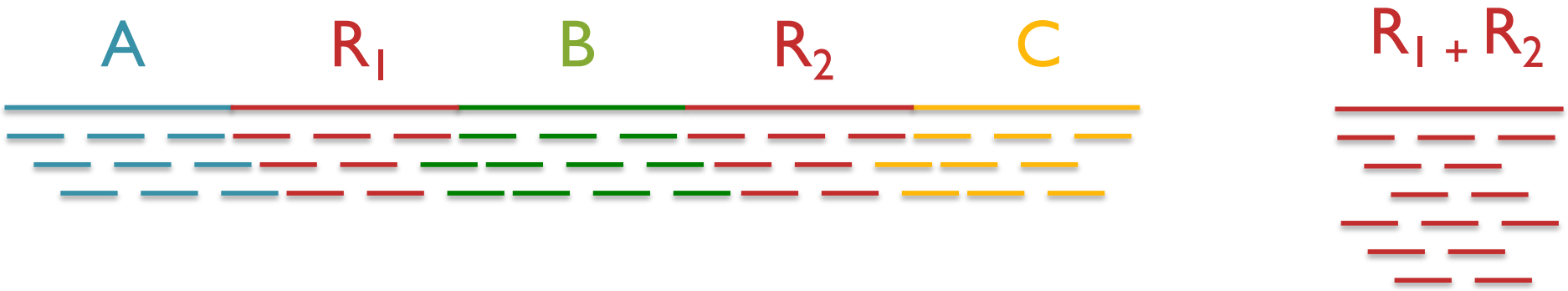


Repetitive regions

Repeat Type	Definition / Example	Prevalence
Low-complexity DNA / Microsatellites	$(b_1 b_2 \ldots b_k)^N$ where $1 \leq k \leq 6$ CACACACACACACACACACA	2%
SINEs (Short Interspersed Nuclear Elements)	*Alu* sequence (~280 bp) Mariner 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

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

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

The fragment assembly string graph
Myers, EW (2005) Bioinformatics. 21(suppl 2): ii79-85.

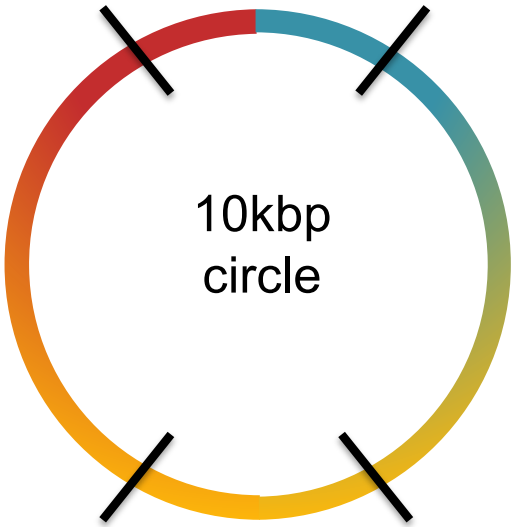
Paired-end and Mate-pairs

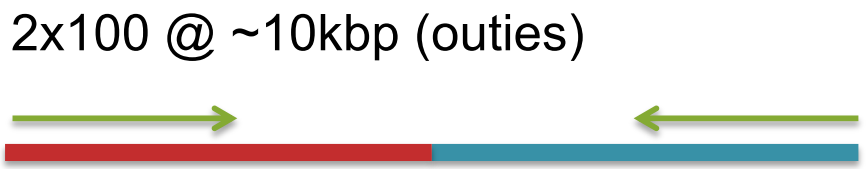
Paired-end sequencing

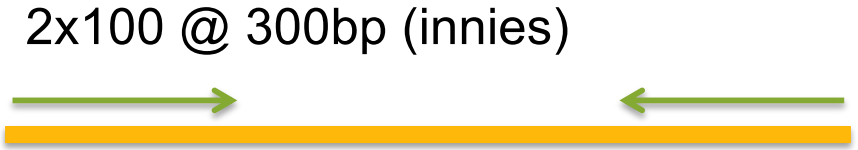
- Read one end of the molecule, flip, and read the other end
- Generate pair of reads separated by up to 500bp with inward orientation

Mate-pair sequencing

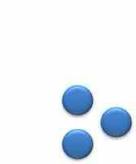
- Circularize long molecules (1-10kbp), shear into fragments, & sequence
- Mate failures create short paired-end reads



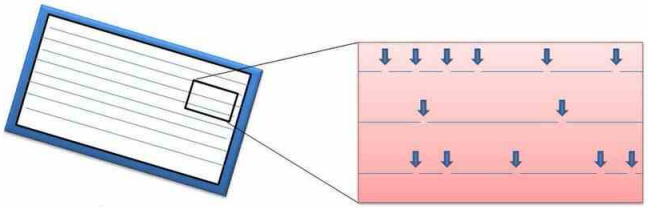




Optical Mapping

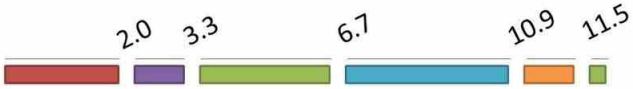


1. Cells are lysed to retrieve genomic DNA

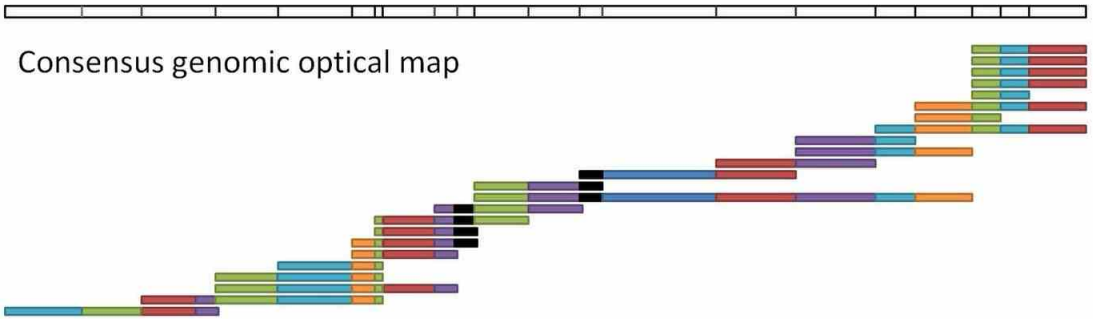


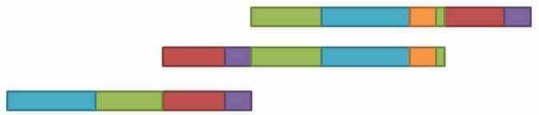
2. Single genomic DNA molecules are placed onto a microfluidic device

3. Restriction enzymes are added to cut the DNA molecules at specific positions



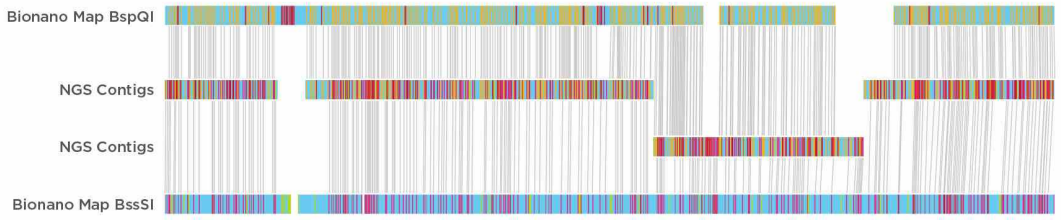
4. Each DNA molecule is stained with a fluorescent dye. An optical map of single-molecules are derived by measuring the fluorescent intensity.





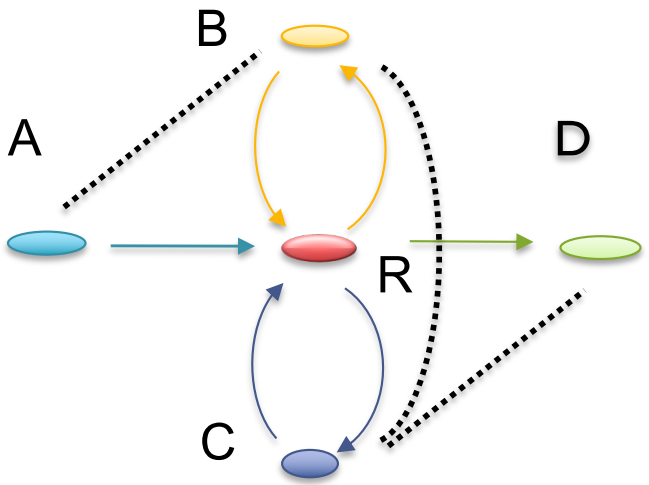
5. Overlapping of the multiple single-molecule maps gives us the consensus genomic optical map

6. Align assembled optical map to in-silico digestion of sequence contigs to estimate their relative position and orientation, “hybrid scaffolding”

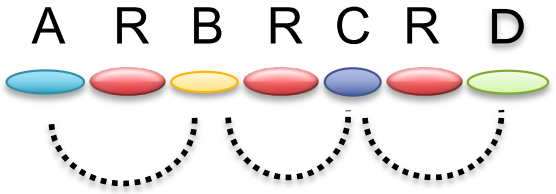


Scaffolding

- Initial contigs (*aka* unipaths, unitigs) terminate at
  - *Coverage gaps*: especially extreme GC
  - *Conflicts*: errors, repeat boundaries



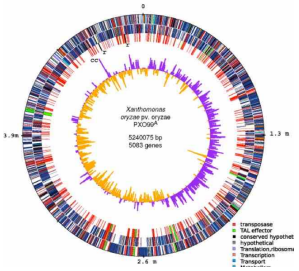
- Use mate-pairs to resolve correct order through assembly graph
  - Place sequence to satisfy the mate constraints
  - Mates through repeat nodes are tangled



- 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?

Assembly Summary

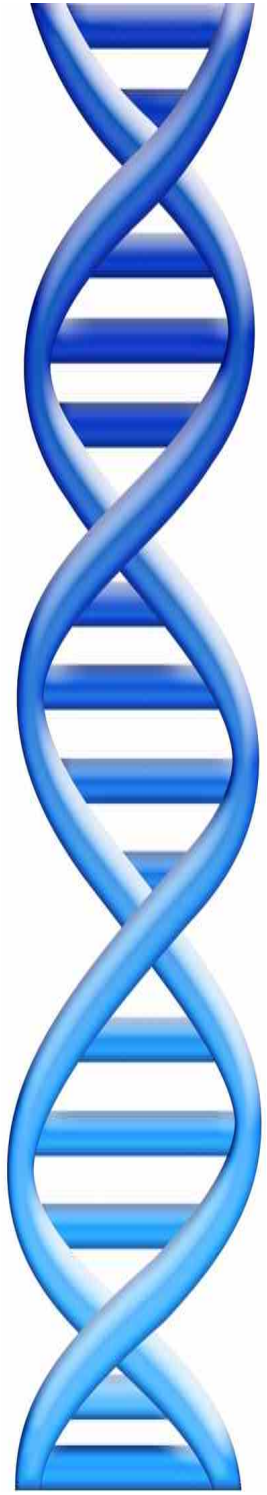


Assembly quality depends on

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

- Recommend spades for short read assembly
  - Integrates error correction and scaffolding



Outline

1. *Assembly theory*
   - Assembly by analogy

2. *Practical Issues*
   - Coverage, read length, errors, and repeats

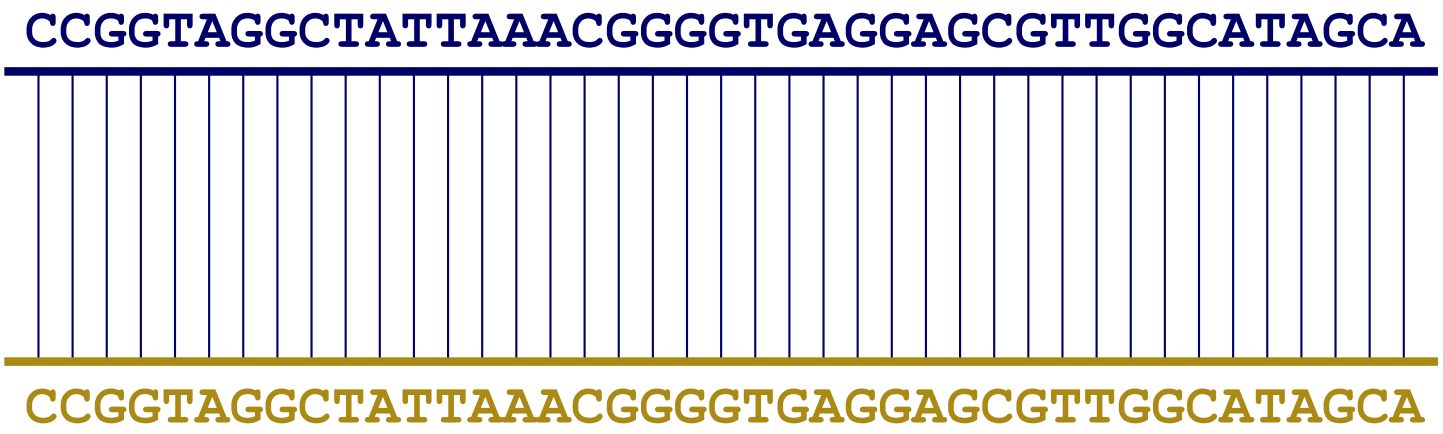
3. **Whole Genome Alignment**
   - **MUMmer recommended**

Whole Genome Alignment with MUMmer

Slides Courtesy of Adam M. Phillippy
NHGRI

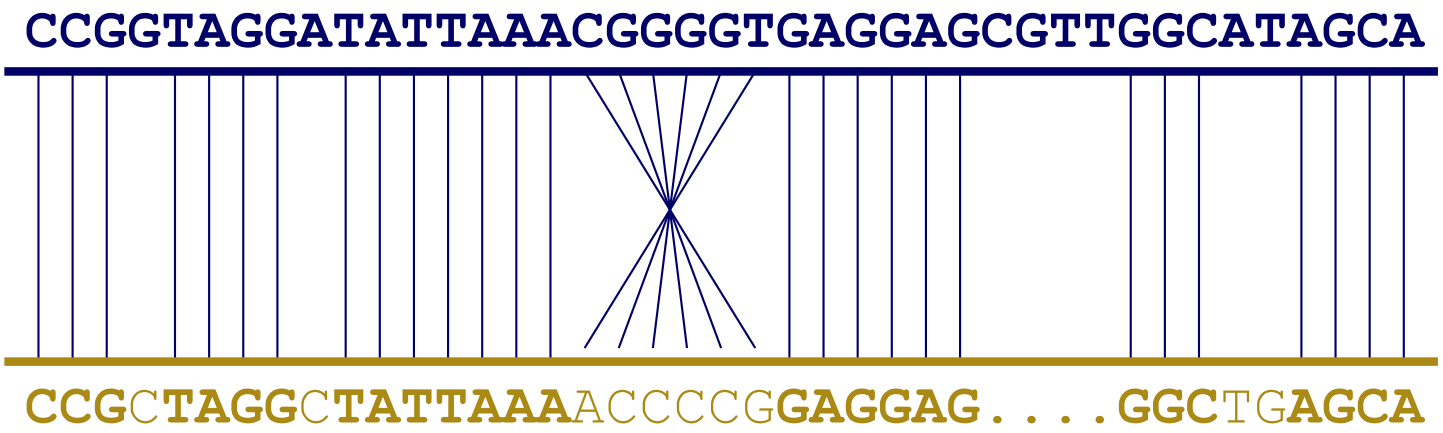
Goal of WGA

- For two genomes, *A* and *B*, find a mapping from each position in *A* to its corresponding position in *B*



Not so fast...

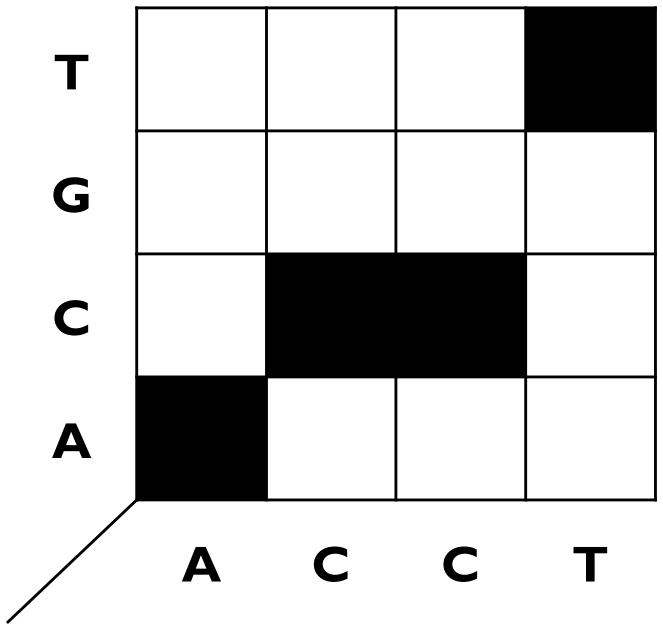
- Genome *A* may have insertions, deletions, translocations, inversions, duplications or SNPs with respect to *B* (sometimes all of the above)



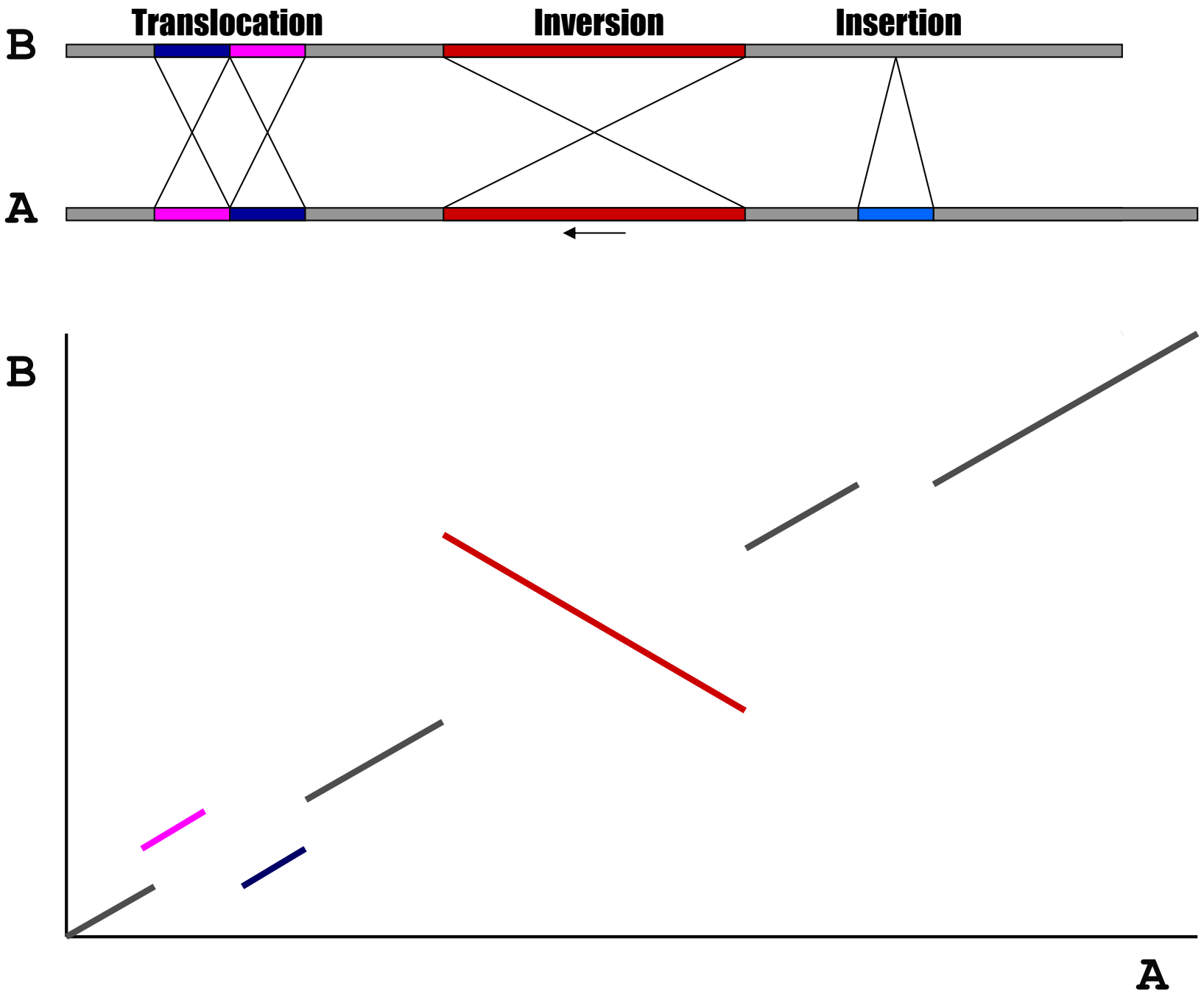
WGA visualization

- How can we visualize *whole* genome alignments?

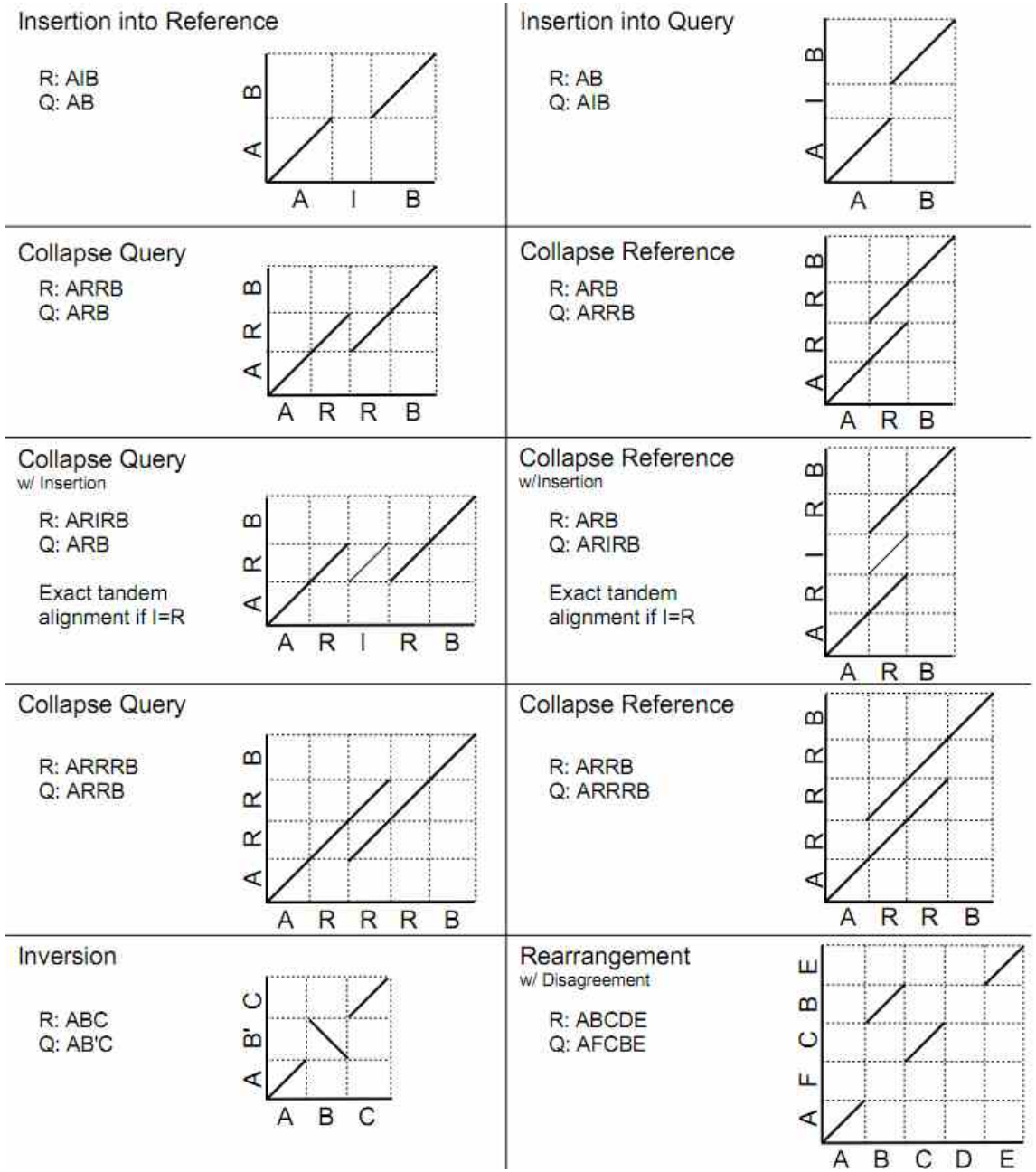
- With an alignment dot plot
  - *N* x *M* matrix
    - Let *i* = position in genome *A*
    - Let *j* = position in genome *B*
    - Fill cell *(i,j)* if A_i shows similarity to B_j



  - A perfect alignment between *A* and *B* would completely fill the positive diagonal

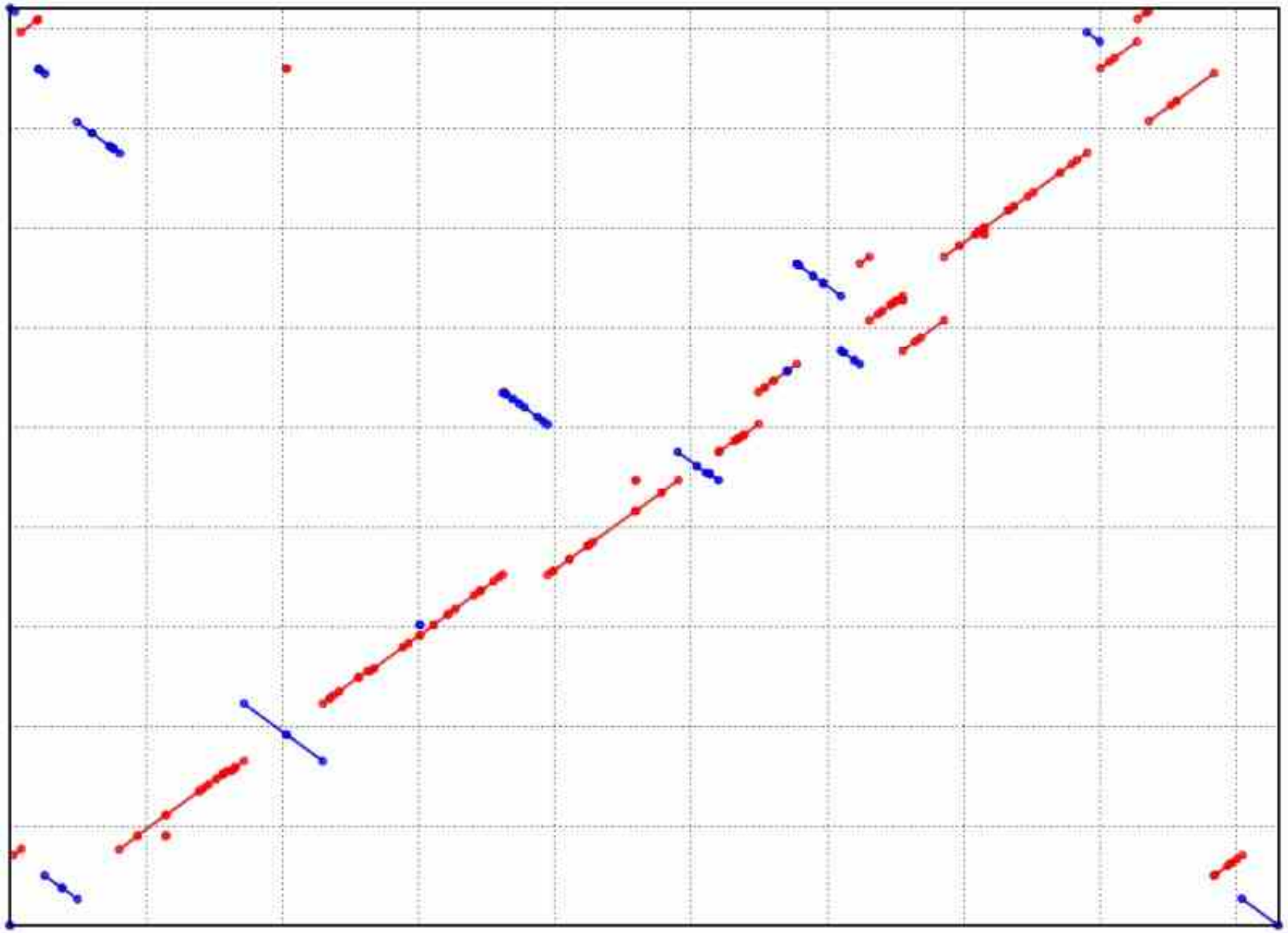

Translocation
Inversion
Insertion
B
A
B
A

SV Types



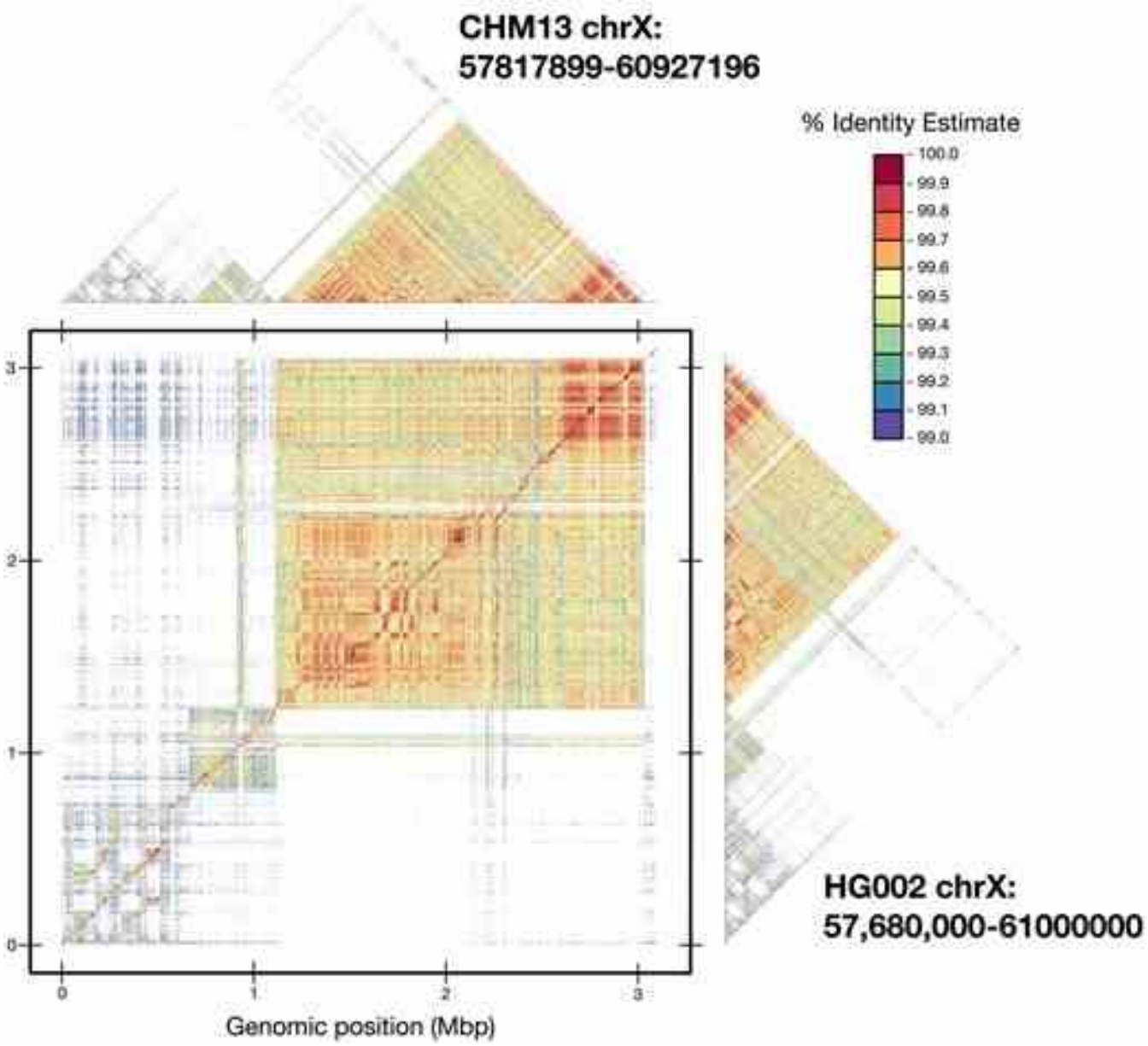
- Different structural variation types / misassemblies will be apparent by their pattern of breakpoints
- Most breakpoints will be at or near repeats
- Things quickly get complicated in real genomes

http://mummer.sf.net/manual/AlignmentTypes.pdf



Alignment of 2 strains of *Y. pestis*

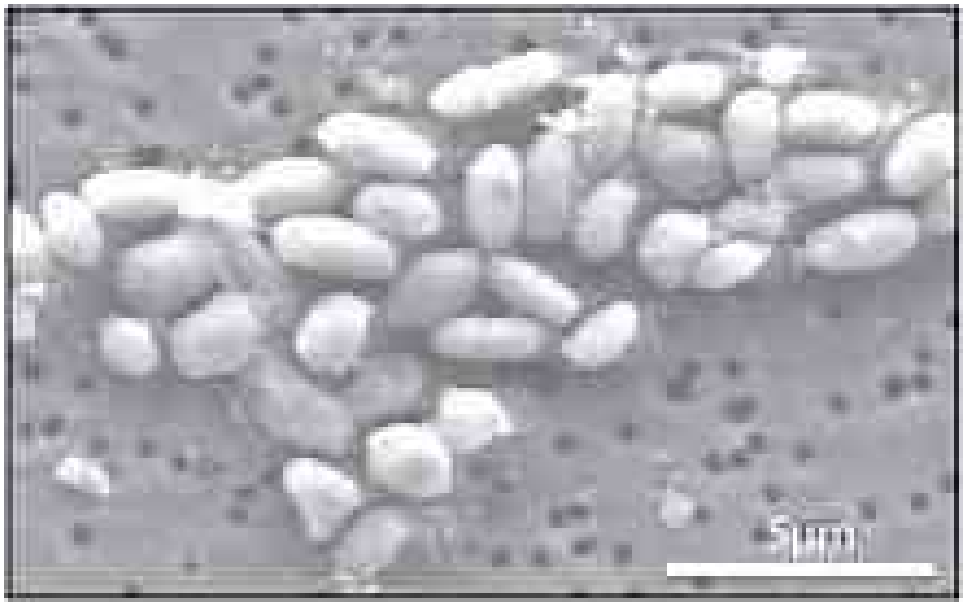
http://mummer.sourceforge.net/manual/



ModDotPlot—rapid and interactive visualization of tandem repeats

Sweeten, Schatz, Phillippy (2024) Bioinformatics. https://doi.org/10.1093/bioinformatics/btae493

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

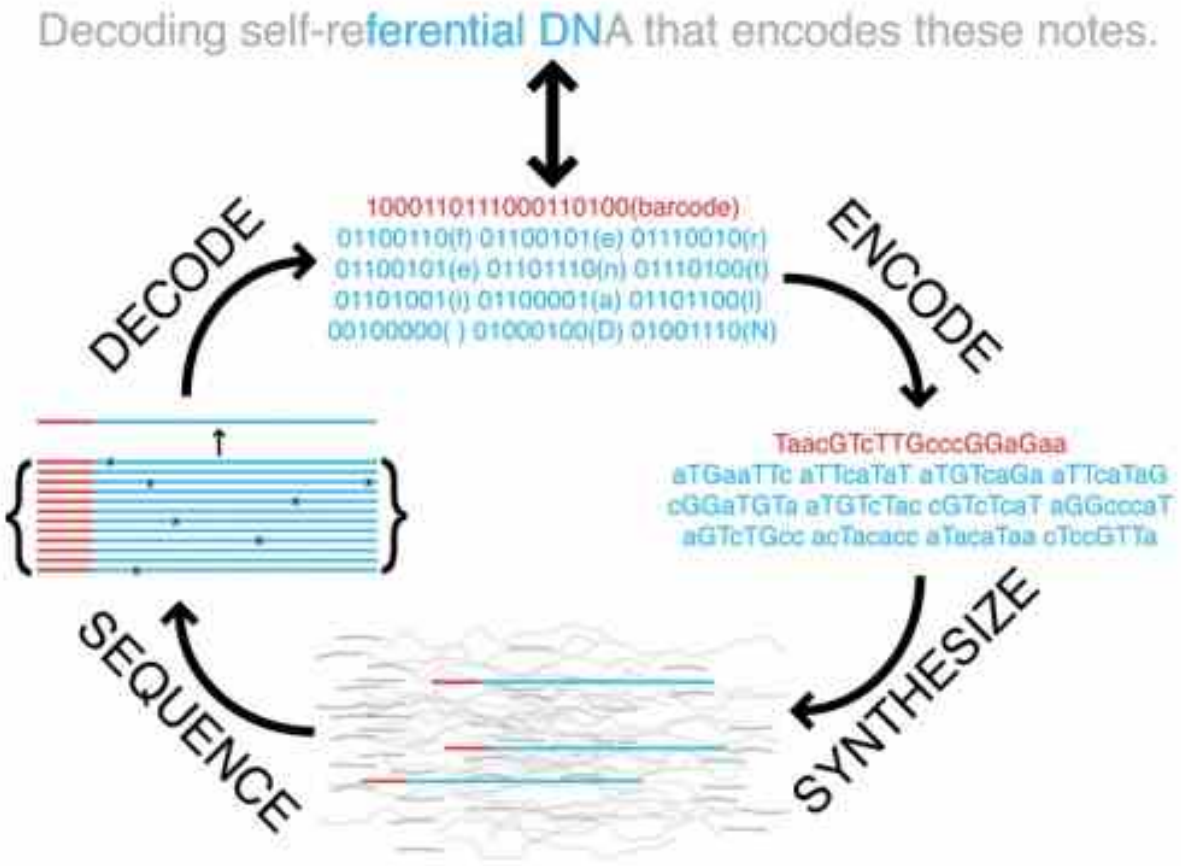


Fig. S1. Schematic of DNA information storage.

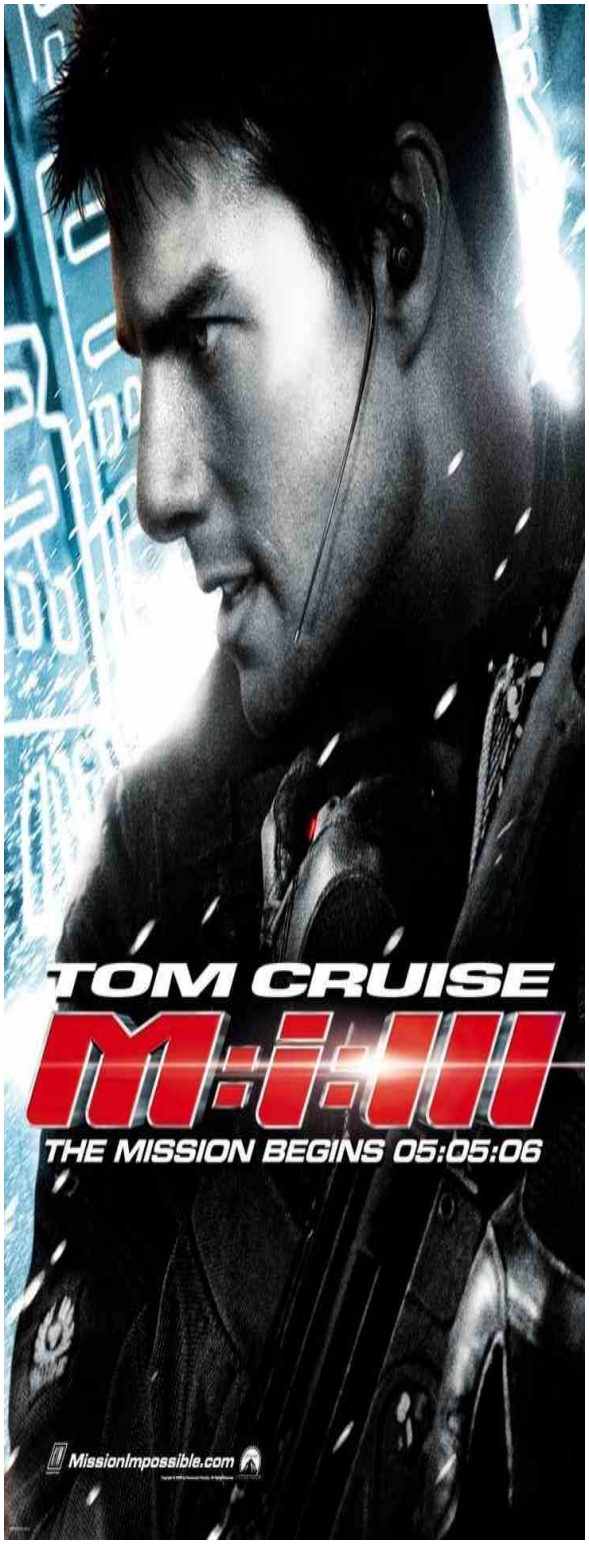
Encoding/decoding algorithm implemented in dna-encode.pl from David Dooling.

Next-generation Digital Information Storage in DNA
Church et al (2010) *Science*. 337(6102)1628

Assignment 2: Genome Assembly
Due Monday Sept 16 @ 11:59pm

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

https://github.com/schatzlab/appliedgenomics2024/blob/master/assignments/assignment4/README.md



Find and decode

nucmer --maxmatch ref.fasta 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 contigs.fasta

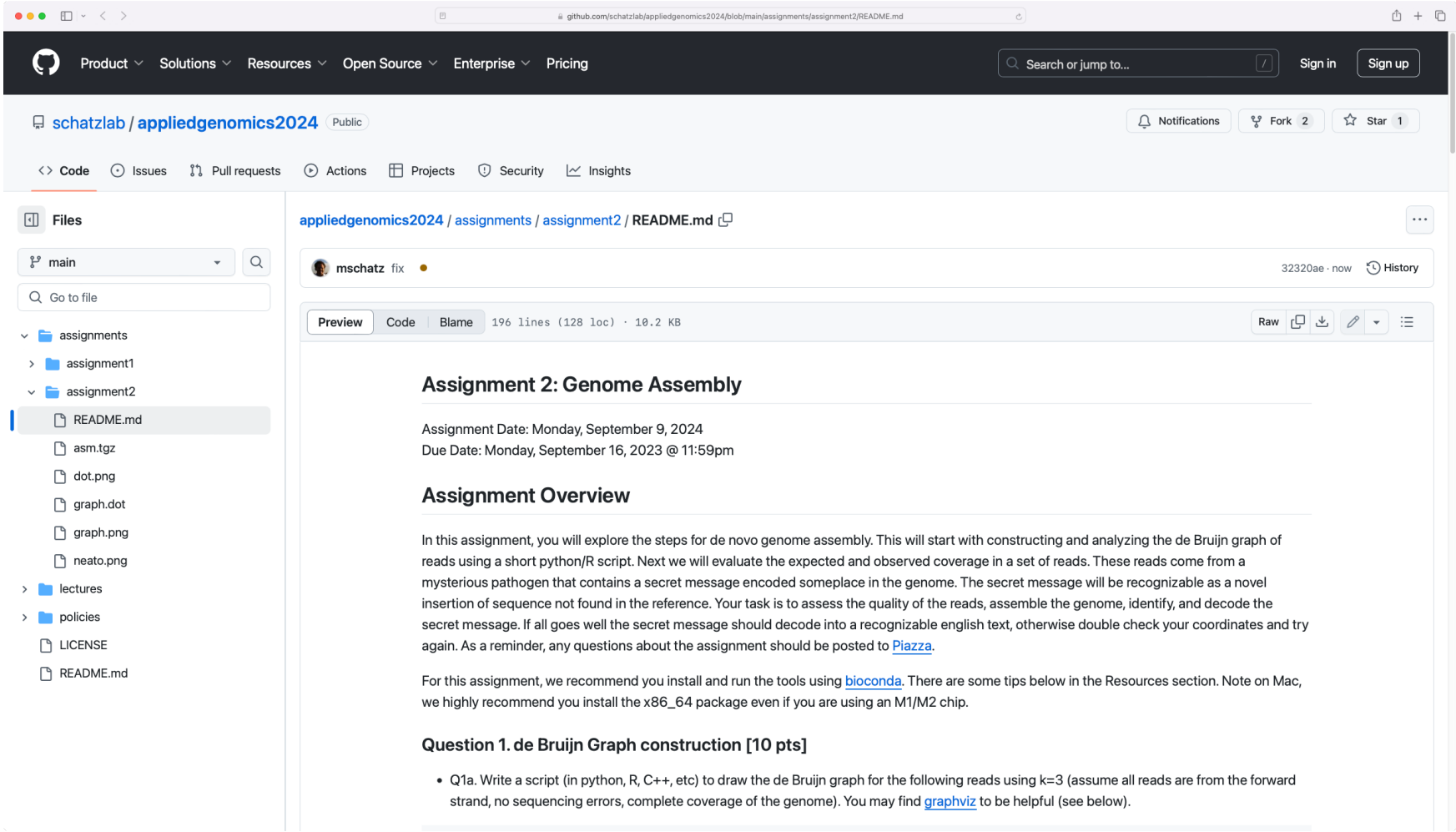
```
   Index the fasta file
```

samtools faidx contigs.fasta contig_XXX:YYY-ZZZ > msg.fa

dna-decode.py -d -input msg.fa

See manual at http://mummer.sourceforge.net/manual

Assignment 2: Genome Assembly
Due Monday Sept 16 by 11:59pm



https://github.com/schatzlab/appliedgenomics2024/tree/main/assignments/assignment2

Check Piazza for questions!