Practical Assembly

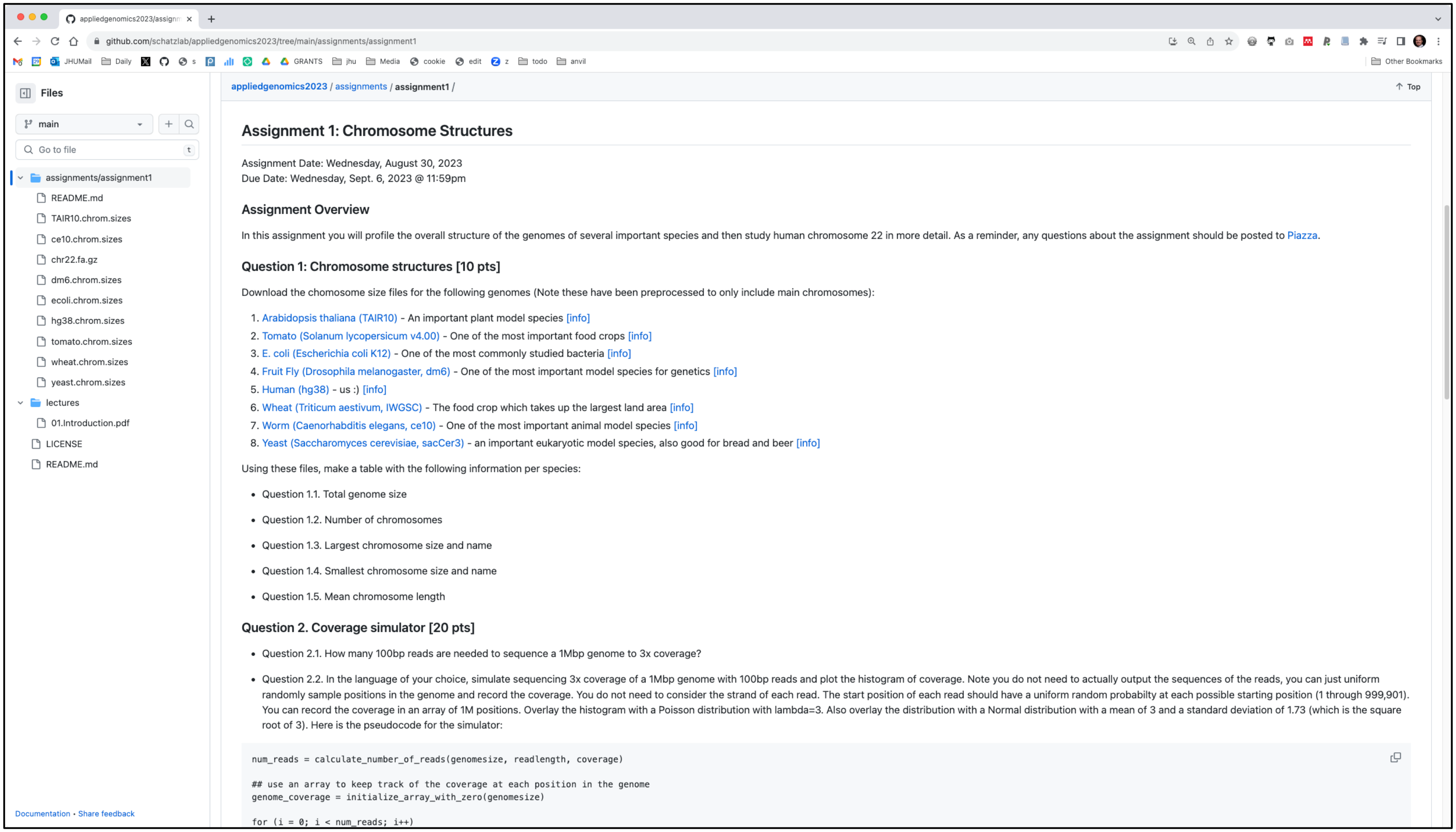
Michael Schatz

Sept 11, 2023

Lecture 4: Applied Comparative Genomics

Assignment 1

Due end of day on Sept 6 (right before midnight)



Assignment 1: Chromosome Structures

Assignment Date: Wednesday, August 30, 2023
Due Date: Wednesday, Sept. 6, 2023 @ 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. 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, dm6) - 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

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 array of 1M positions. Overlay the histogram with a Poisson distribution with lambda=3. Also overlay the distribution with a Normal distribution with a mean of 3 and a standard deviation of 1.73 (which is the square root of 3). Here is the pseudocode for the simulator:

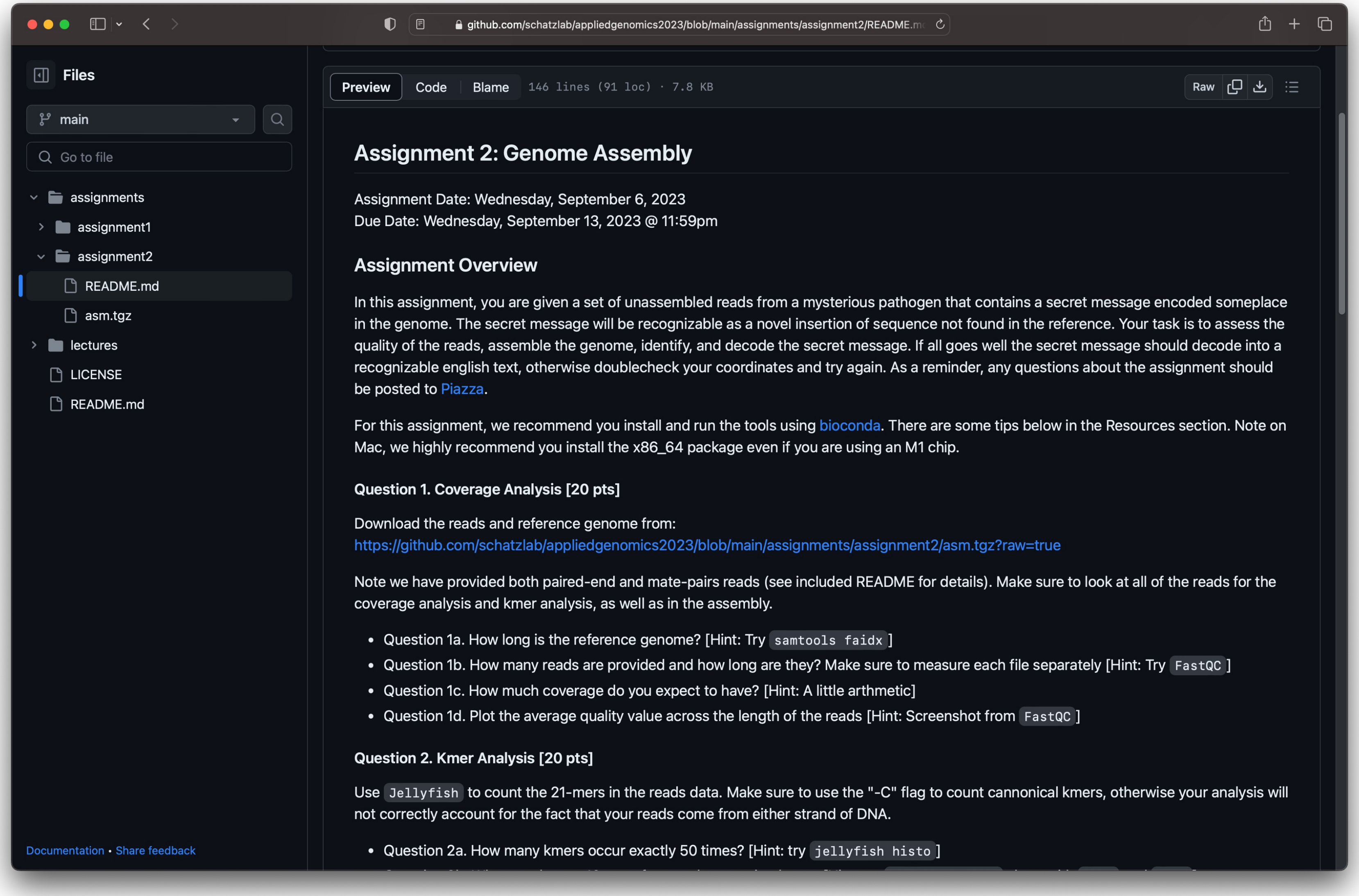
```
num_reads = calculate_number_of_reads(genomesize, readlength, coverage)

## use an array to keep track of the coverage at each position in the genome
genome_coverage = initialize_array_with_zero(genomesize)

for (i = 0; i < num_reads; i++)
```

https://github.com/schatzlab/appliedgenomics2023/tree/main/assignments/assignment1

Assignment 2: Genome Assembly
Due Wednesday Sept 13 by 11:59pm


github.com/schatzlab/appliedgenomics2023/blob/main/assignments/assignment2/README.m
Files
main
Go to file
assignments
assignment1
assignment2
README.md
asm.tgz
lectures
LICENSE
README.md
Documentation • Share feedback
Preview
Code
Blame
146 lines (91 loc) · 7.8 KB
Raw

Assignment 2: Genome Assembly

Assignment Date: Wednesday, September 6, 2023
Due Date: Wednesday, September 13, 2023 @ 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.

For this assignment, we recommend you install and run the tools using bioconda. There are some tips below in the Resources section. Note on Mac, we highly recommend you install the x86_64 package even if you are using an M1 chip.

Question 1. Coverage Analysis [20 pts]

Download the reads and reference genome from:
https://github.com/schatzlab/appliedgenomics2023/blob/main/assignments/assignment2/asm.tgz?raw=true

Note we 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 arthmetic]
- Question 1d. Plot the average quality value across the length of the reads [Hint: Screenshot from `FastQC`]

Question 2. Kmer Analysis [20 pts]

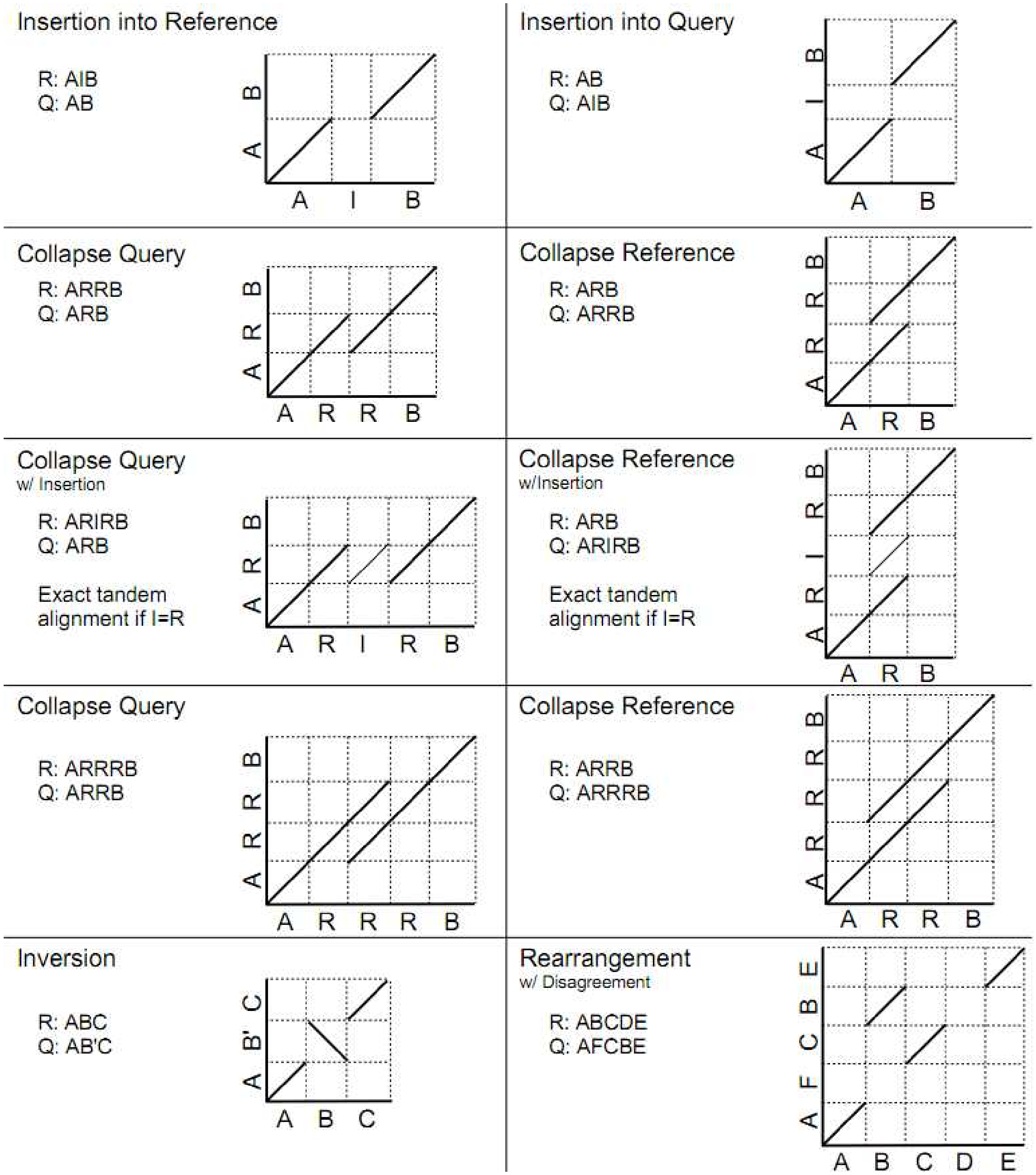
Use `Jellyfish` to count the 21-mers in the reads data. Make sure to use the "-C" flag to count cannonical kmers, otherwise your analysis will not correctly account for the fact that your reads come from either strand of DNA.

- Question 2a. How many kmers occur exactly 50 times? [Hint: try `jellyfish histo`]

https://github.com/schatzlab/appliedgenomics2023/tree/main/assignments/assignment2

Check Piazza for questions!

SV Types



Insertion into Reference

R: AIB
Q: AB

Insertion into Query

R: AB
Q: AIB

Collapse Query

R: ARRB
Q: ARB

Collapse Reference

R: ARB
Q: ARRB

Collapse Query
w/ Insertion

R: ARIRB
Q: ARB

Exact tandem
alignment if I=R

Collapse Reference
w/Insertion

R: ARB
Q: ARIRB

Exact tandem
alignment if I=R

Collapse Query

R: ARRRB
Q: ARRB

Collapse Reference

R: ARRB
Q: ARRRB

Inversion

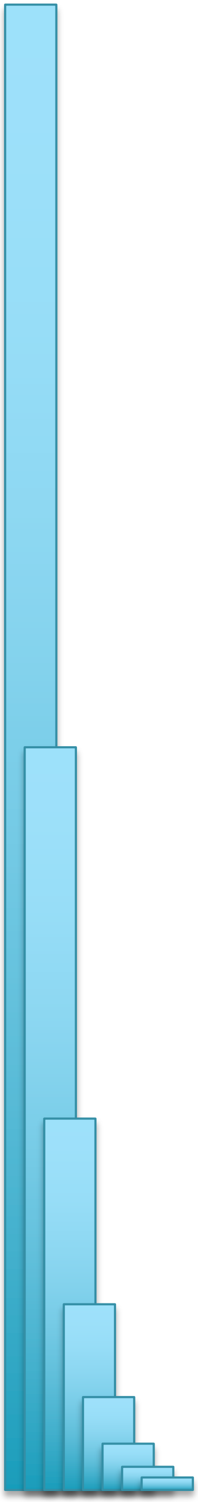
R: ABC
Q: AB'C

Rearrangement
w/ Disagreement

R: ABCDE
Q: AFCBE

- 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

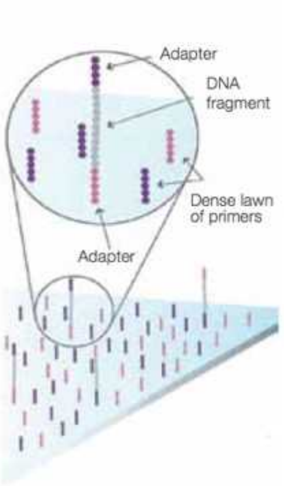


Part 1: Recap

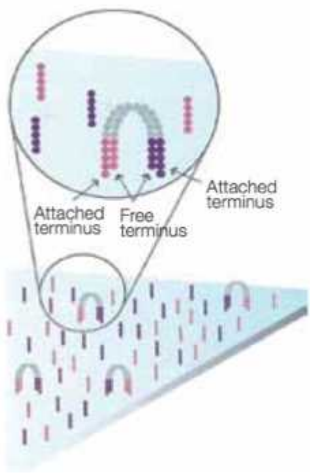
Second Generation Sequencing

Illumina NovaSeq 6000
Sequencing by Synthesis

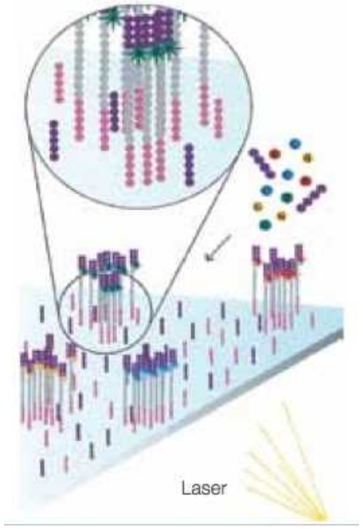
>3Tbp / day
(JHU has 4 of these!)



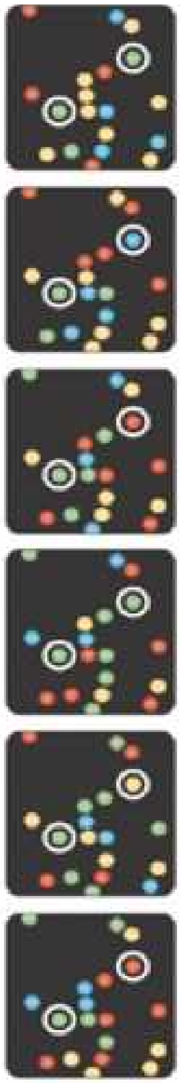
1. Attach



2. Amplify

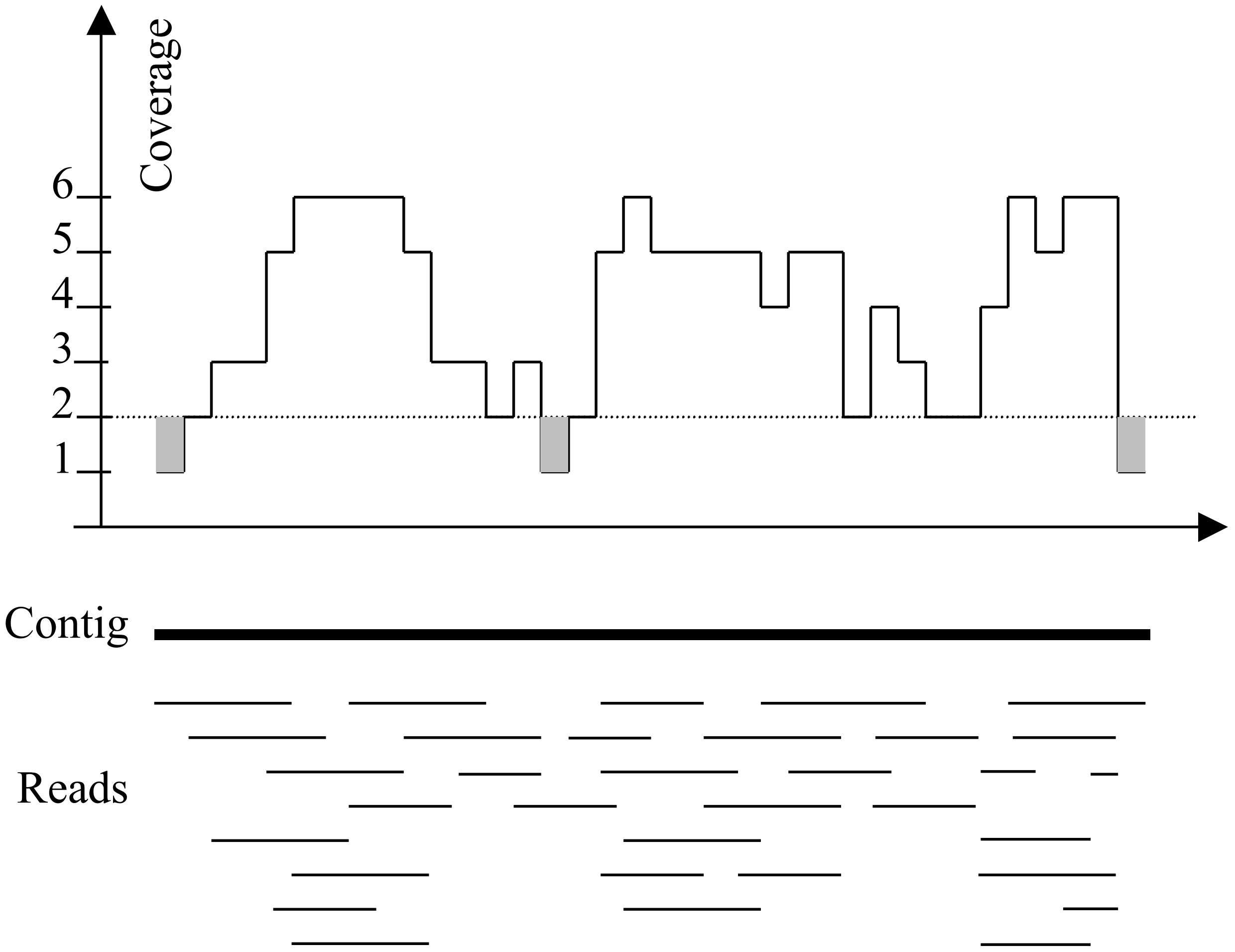


3. Image



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?

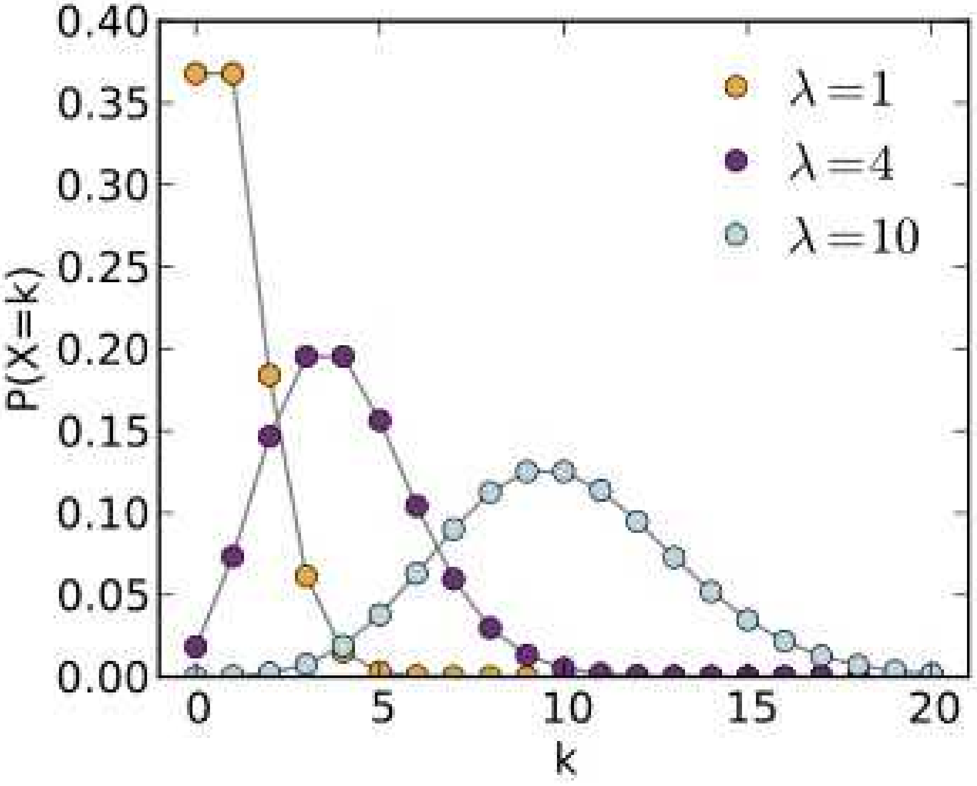
Poisson Distribution

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

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.

Formulation comes from the limit of the binomial equation

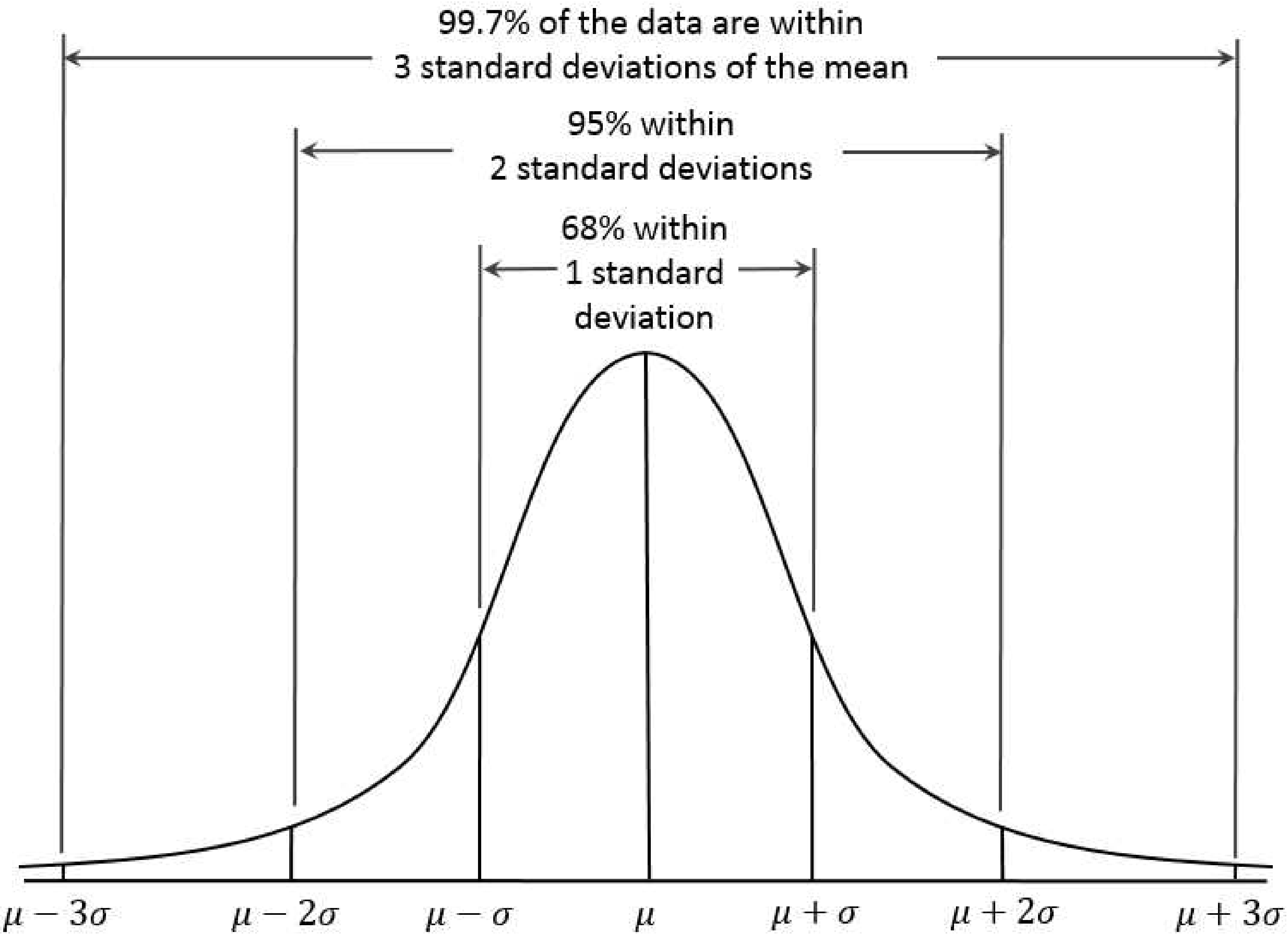
Resembles a normal distribution, but over the positive values, and with only a single parameter.



Key properties:

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

Normal Approximation



Can estimate Poisson distribution as a normal distribution when λ > 10

Pop Quiz!

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

I need 10Mbp x 24x = 240Mbp of data
240Mbp / 120bp / read = 2M 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

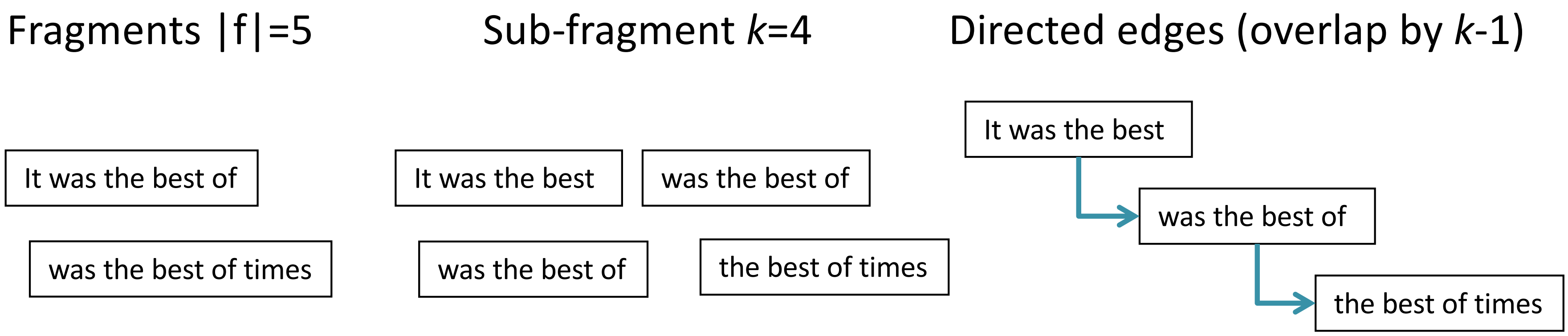
36-2*sqrt(36) = 24

I need 10Mbp x 36x = 360Mbp of data
360Mbp / 120bp / read = 3M reads

Part 2: De novo genome assembly

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

It was the best of

was the best of times

Sub-fragment *k*=4

It was the best

was the best of

was the best of

the best of times

Directed edges (overlap by *k*-1)

It was the best

was the best of

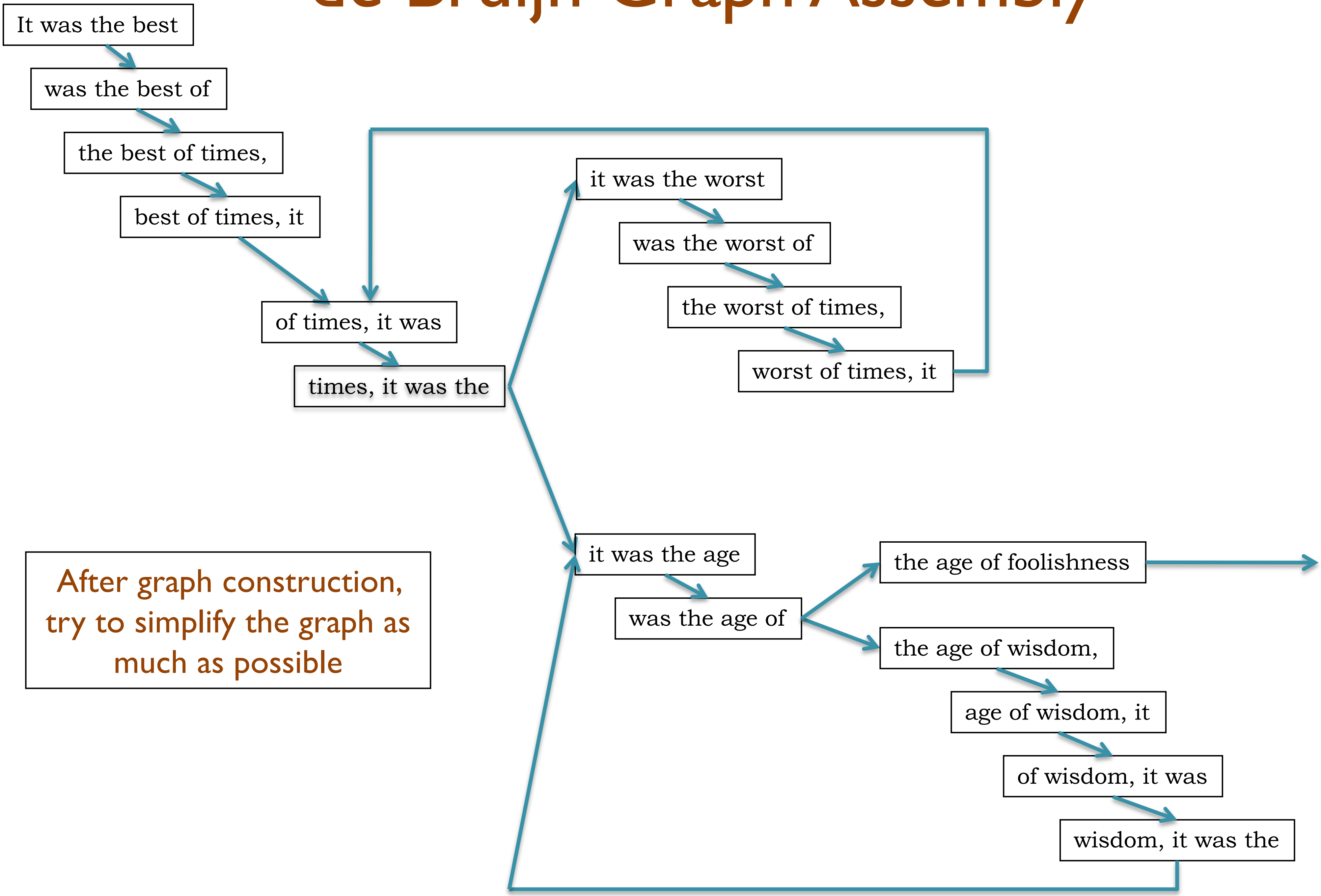
the best of times

– 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



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,

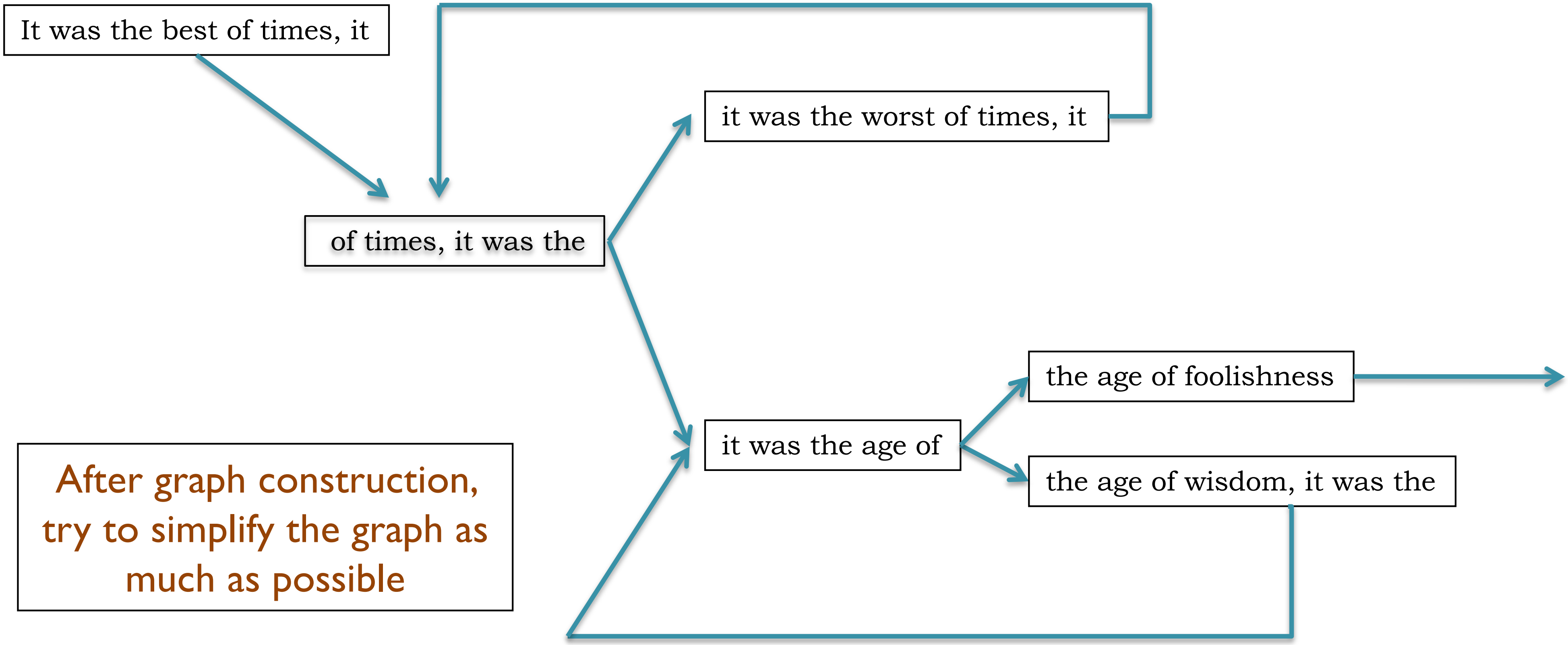
age of wisdom, it

of wisdom, it was

wisdom, it was the

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

de Bruijn Graph Assembly


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

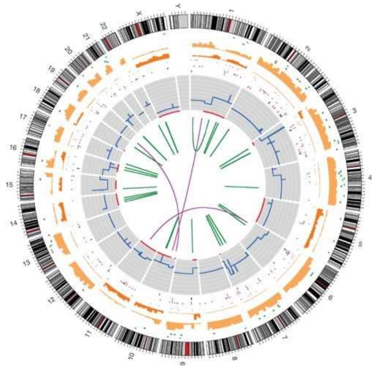
After graph construction, try to simplify the graph as much as possible

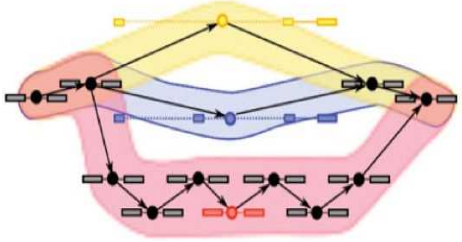
Assembly Applications

- Novel genomes

- Metagenomes

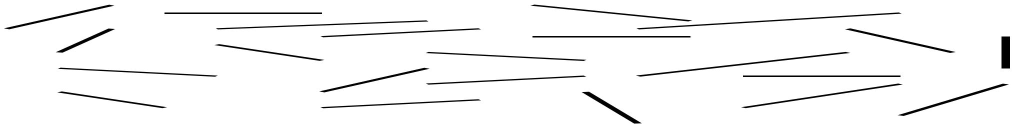
- Sequencing assays
  - Structural variations
  - Transcript assembly
  - …





Assembling a Genome

1. Shear & Sequence DNA



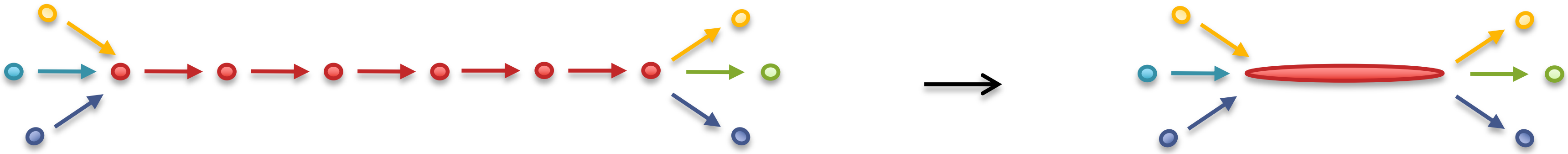
2. Construct assembly graph from reads (de Bruijn / overlap graph)

...AGCCTAGGGATGCGCGACACGT

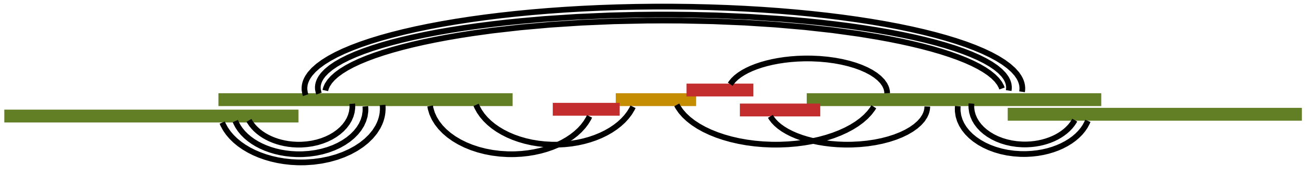
GGATGCGCGACACGTCGCATATCCGGTTTGGTCAACCTCGGACGGAC

CAACCTCGGACGGACCTCAGCGAA...

3. Simplify assembly graph

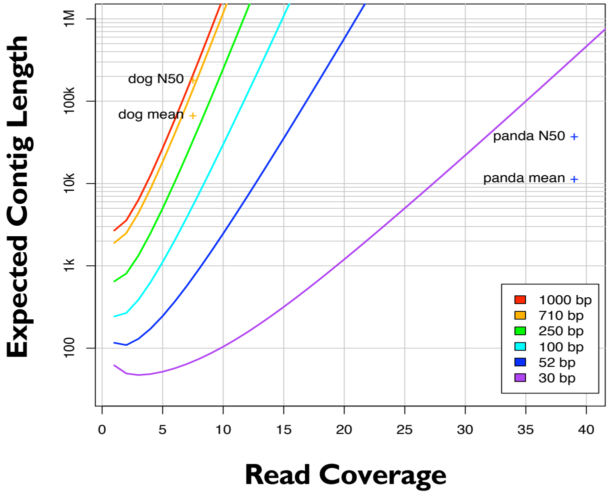


4. Detangle graph with long reads, mates, and other links



Ingredients for a good assembly

Coverage

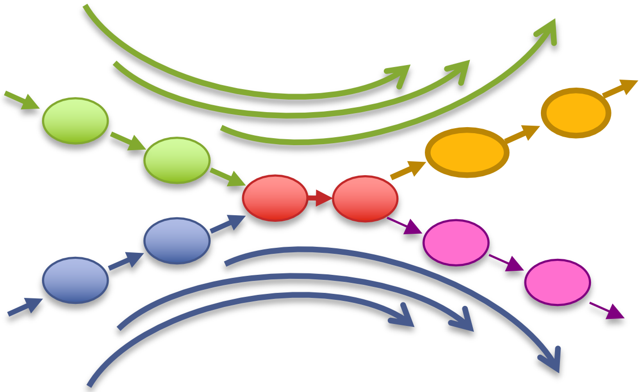


High coverage is required

– Oversample the genome to ensure every base is sequenced with long overlaps between reads

– Biased coverage will also fragment assembly

Read Length

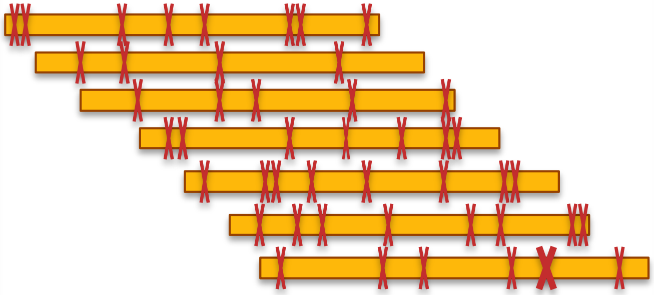


Reads & mates must be longer than the repeats

– 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

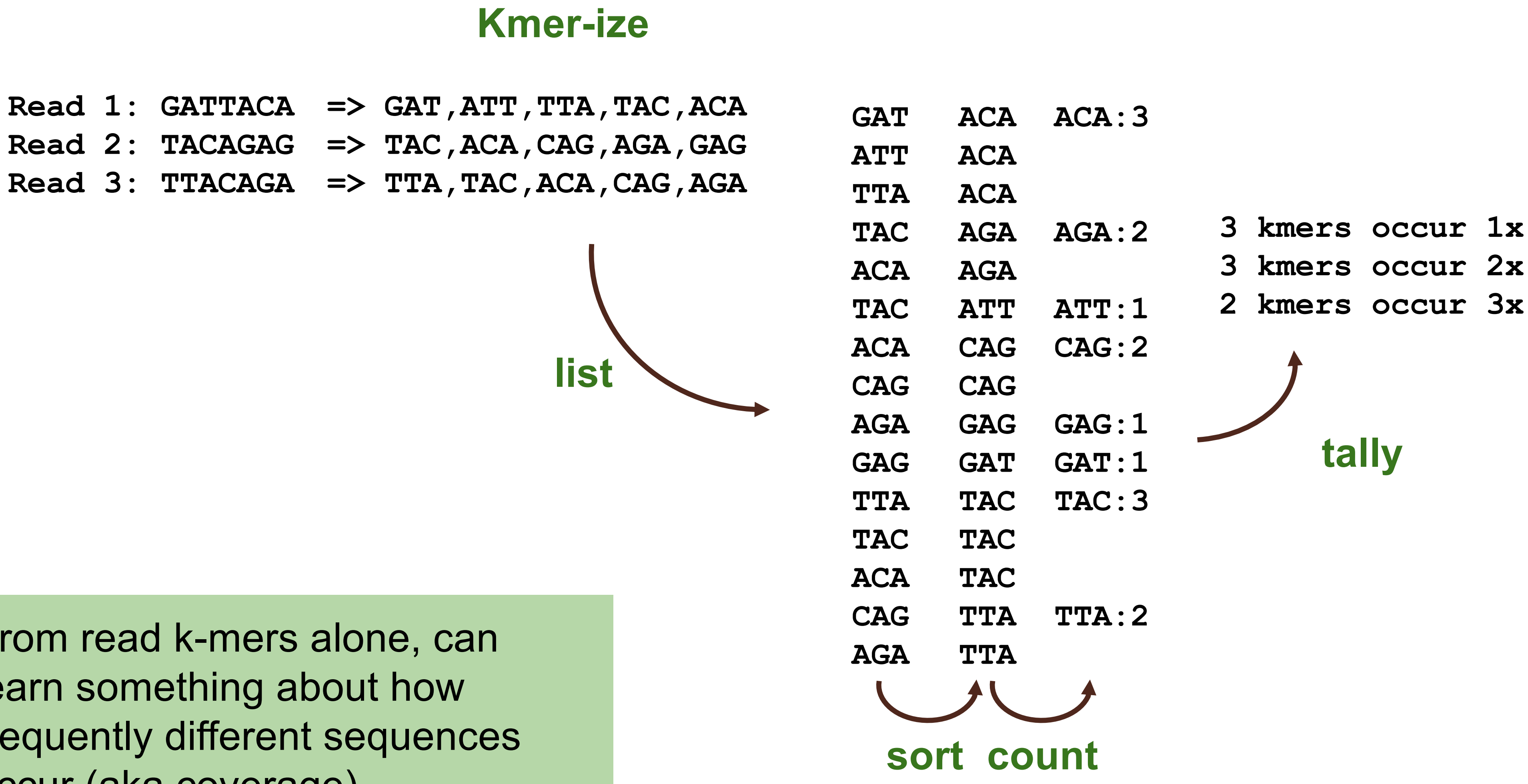
$$\text{sequencing\_coverage} = \frac{\text{total\_bases\_sequenced}}{\text{genome\_size}}$$

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

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

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      3 kmers occur 1x
ACA   AGA              3 kmers occur 2x
TAC   ATT   ATT:1      2 kmers occur 3x
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
```

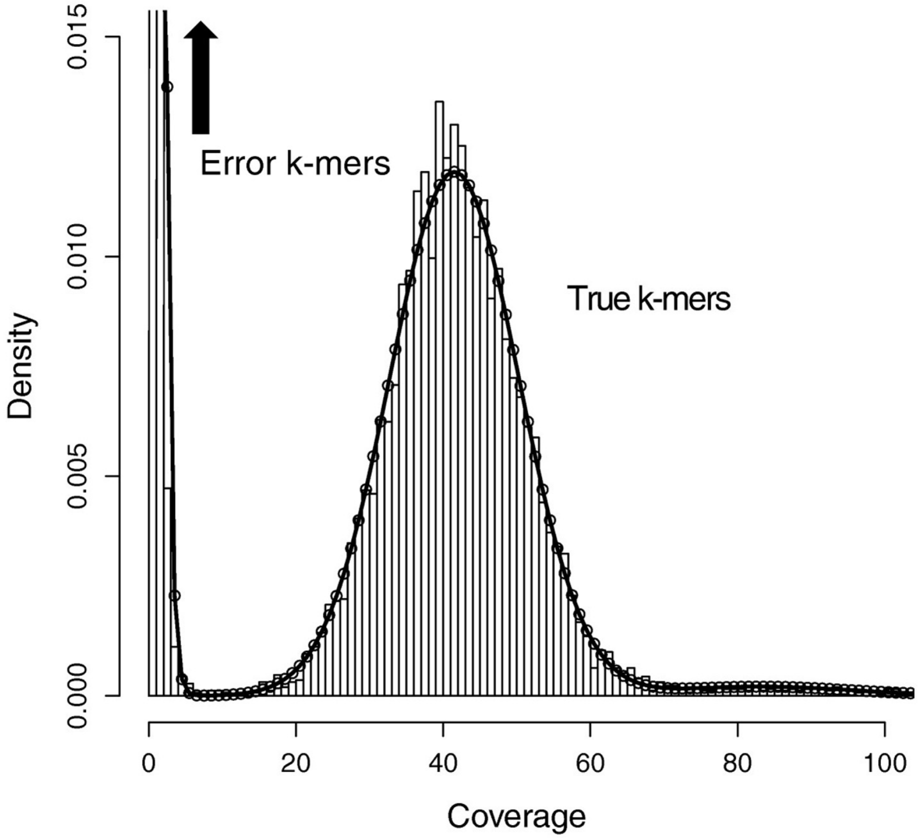
tally

sort 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



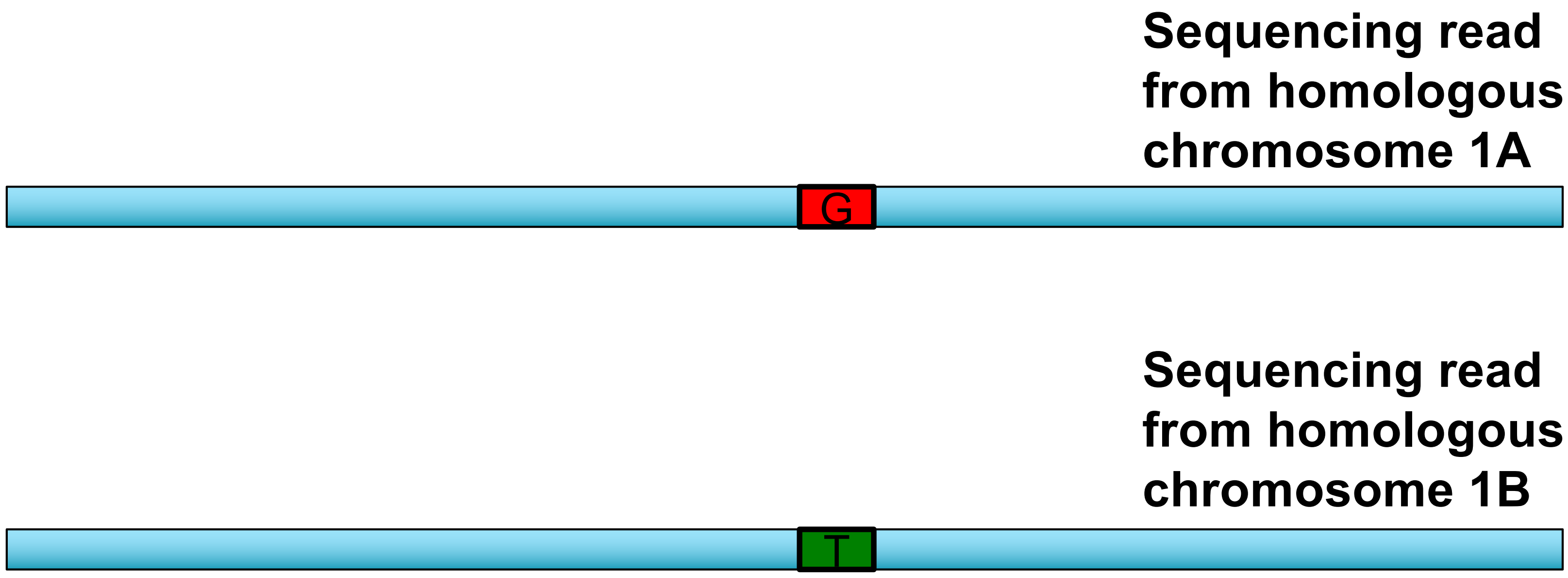
- The tally of k-mer counts in real genomes reveals the coverage distribution.

- 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

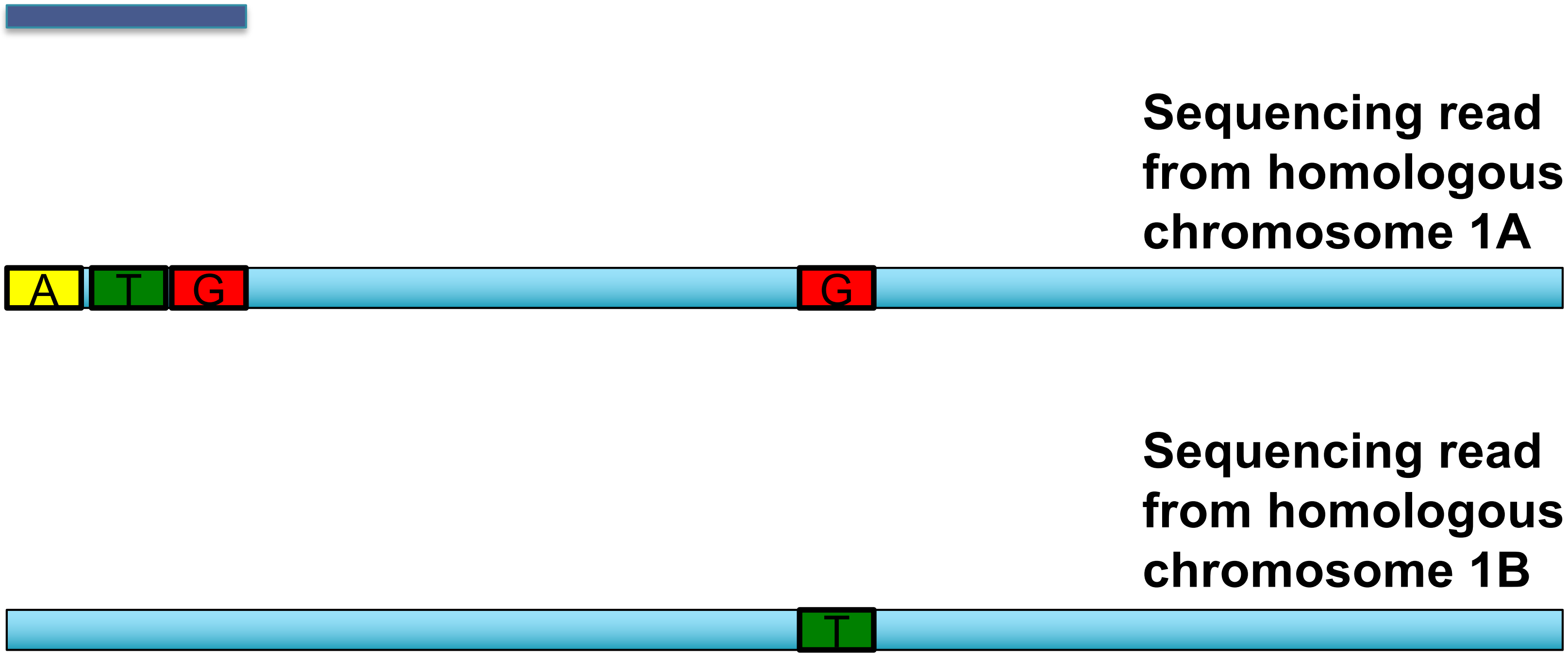
K-mer counting in heterozygous genomes



Sequencing read from homologous chromosome 1A

Sequencing read from homologous chromosome 1B

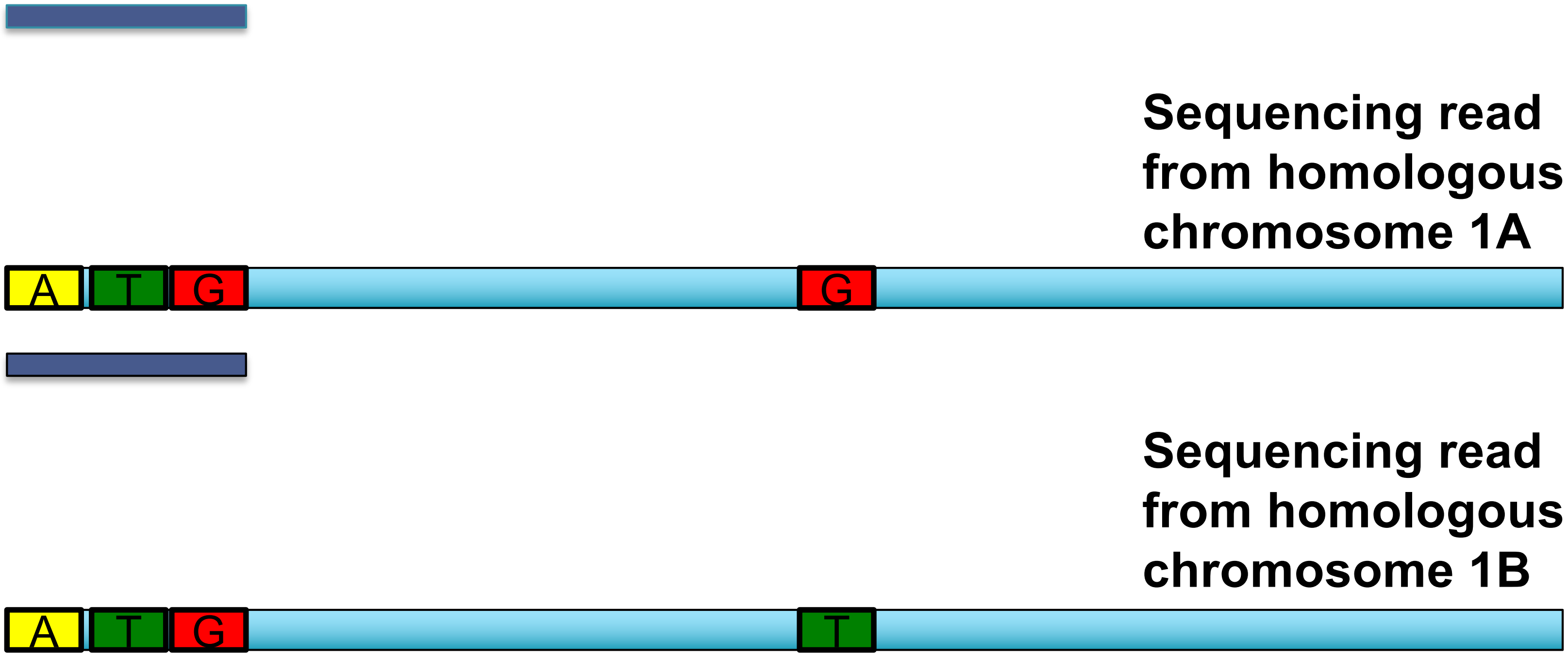
K-mer counting in heterozygous genomes



Sequencing read from homologous chromosome 1A

Sequencing read from homologous chromosome 1B

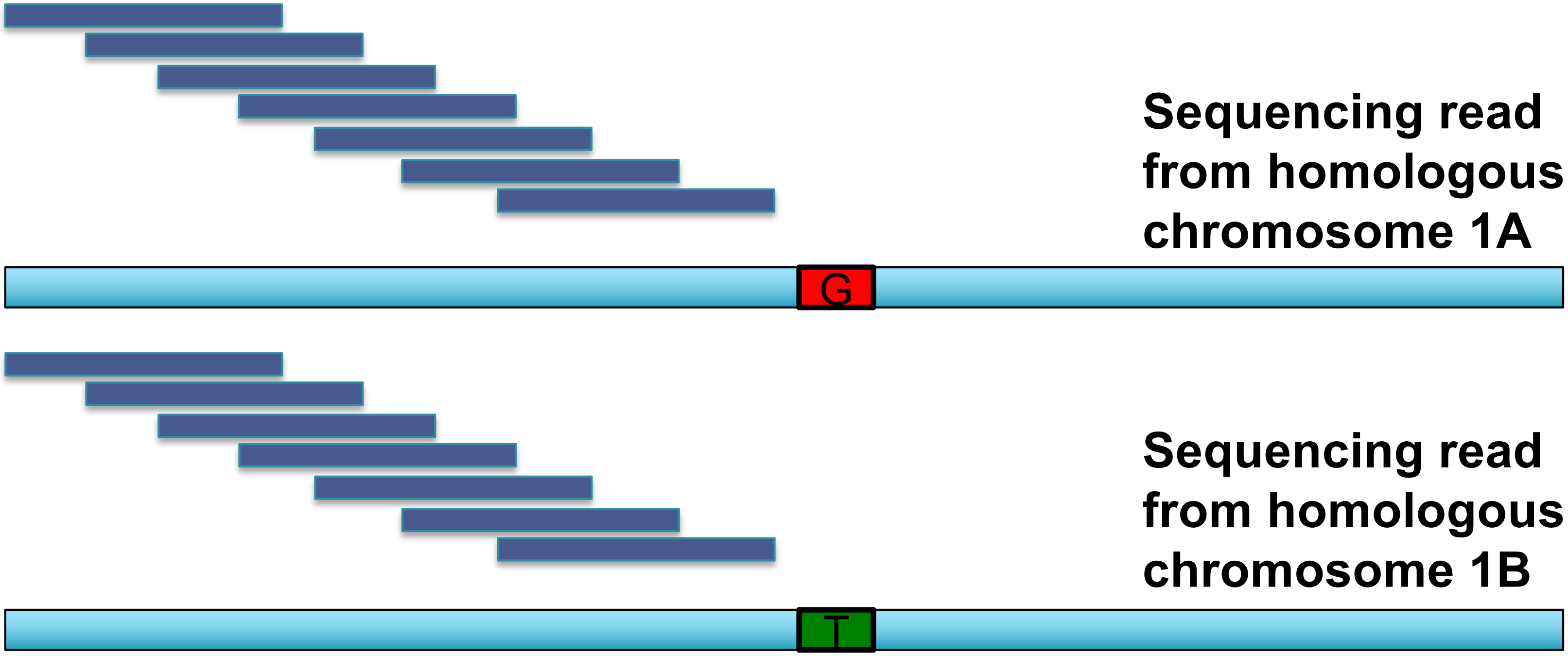
K-mer counting in heterozygous genomes



Sequencing read from homologous chromosome 1A

Sequencing read from homologous chromosome 1B

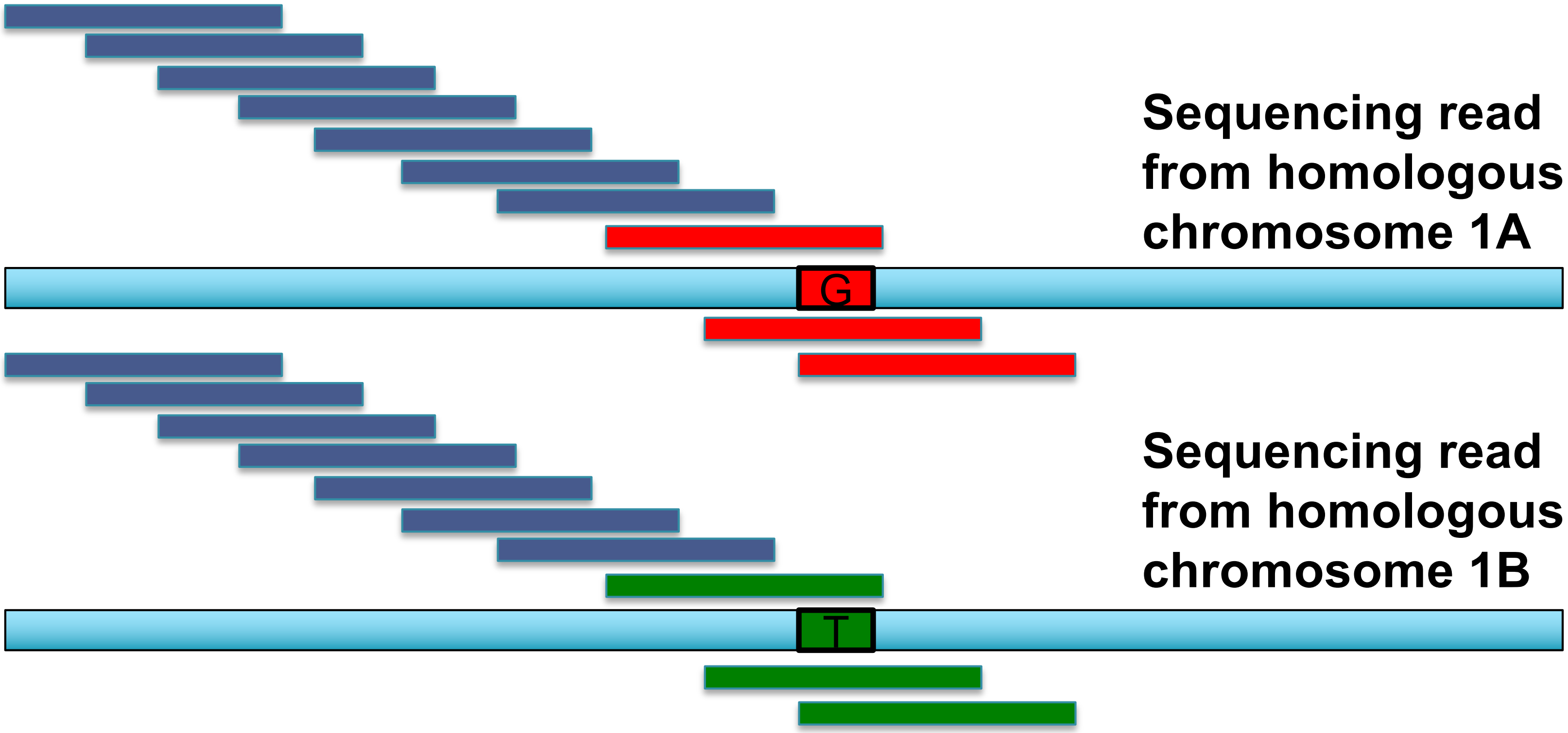
K-mer counting in heterozygous genomes



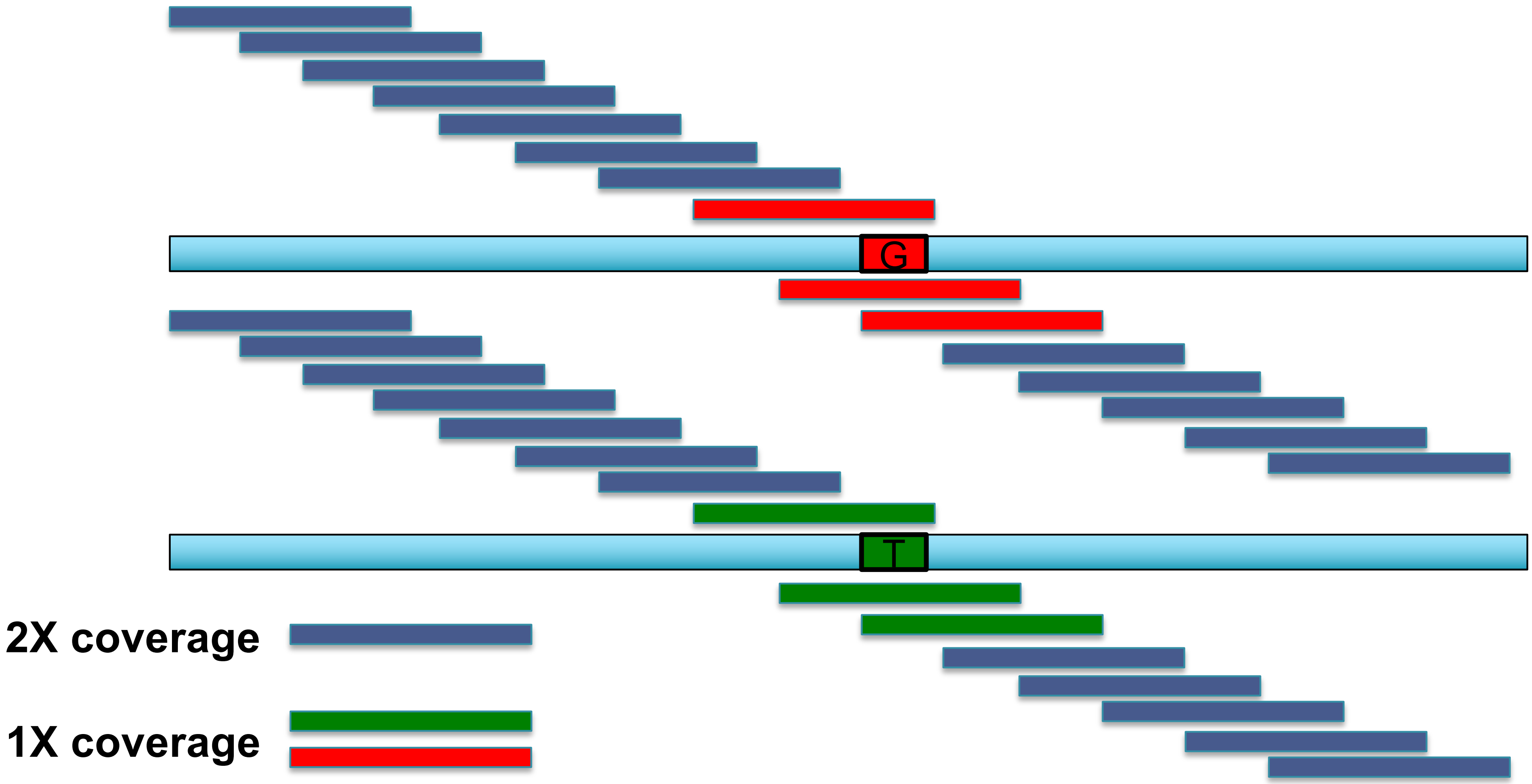
Sequencing read from homologous chromosome 1A

Sequencing read from homologous chromosome 1B

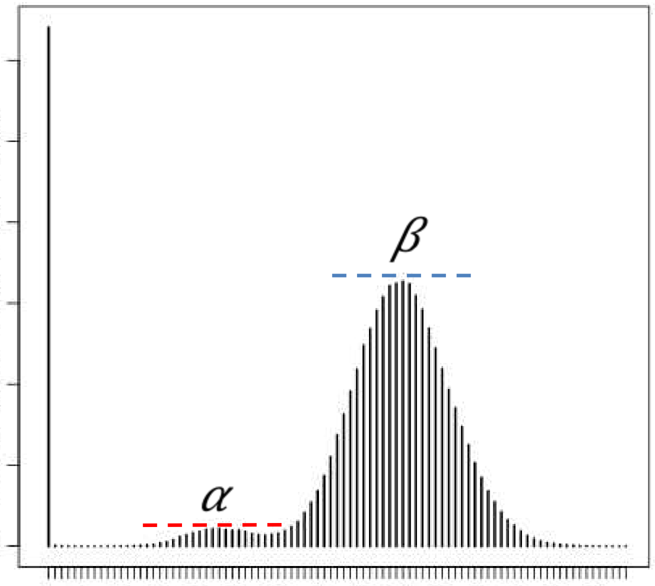
K-mer counting in heterozygous genomes


Sequencing read from homologous chromosome 1A
G
Sequencing read from homologous chromosome 1B
T

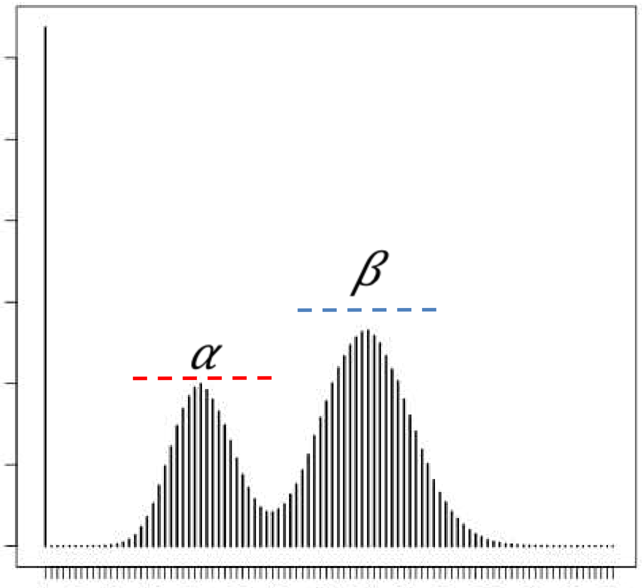
K-mer counting in heterozygous genomes



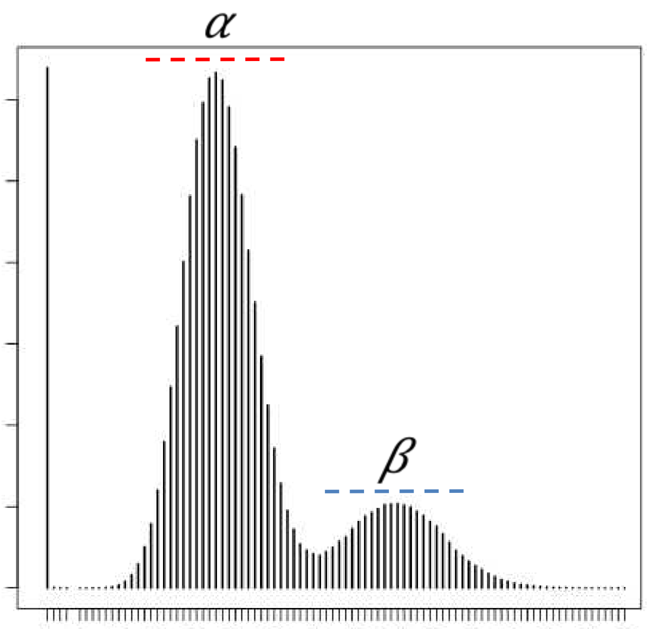
Heterozygous Kmer Profiles



0.1% heterozygosity



1% heterozygosity



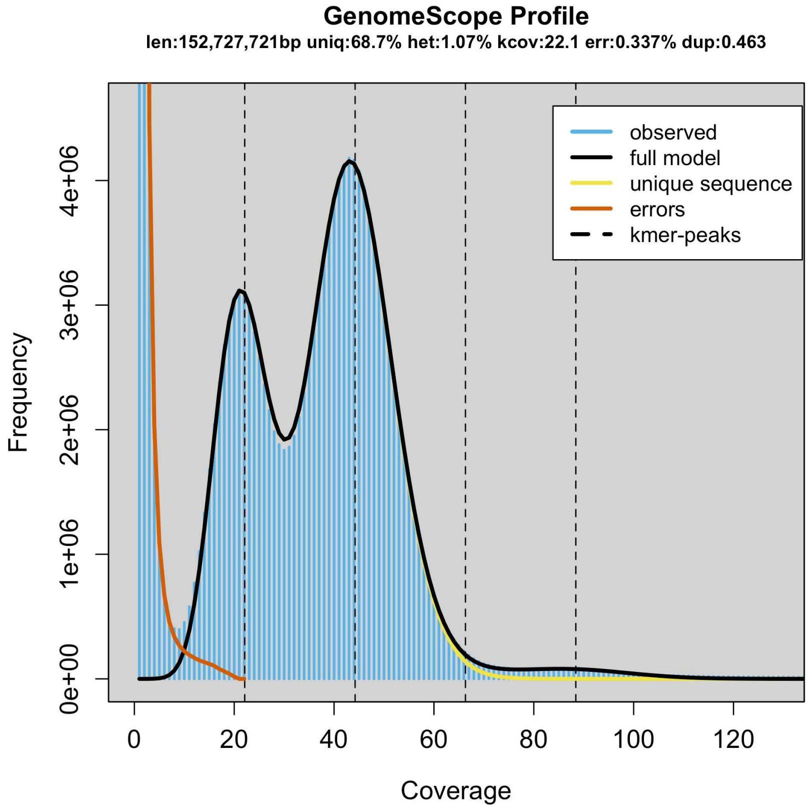
5% heterozygosity

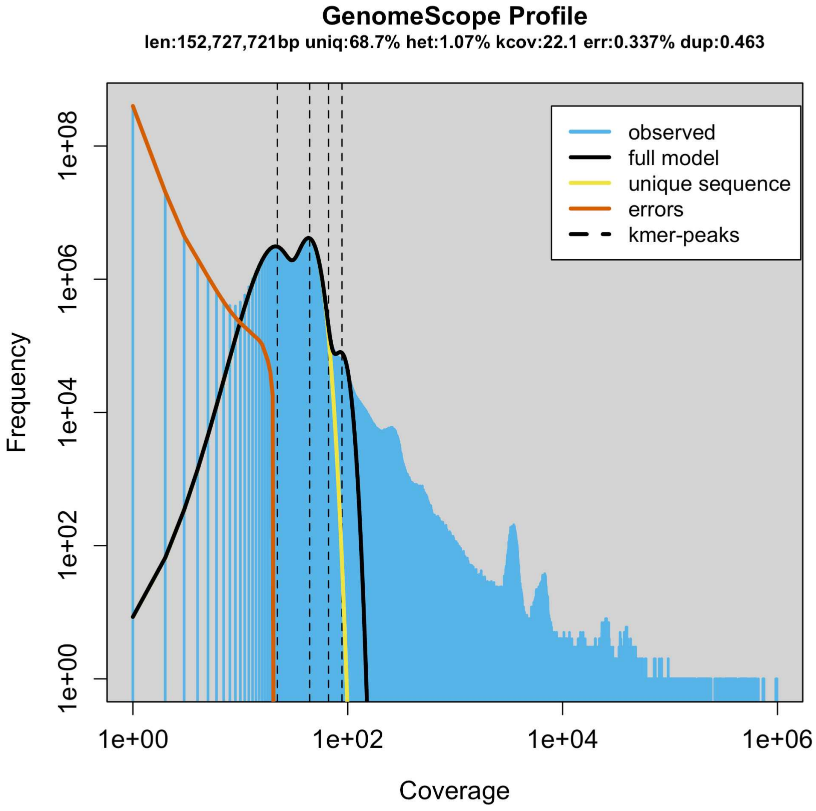
- ***Heterozygosity creates a characteristic "double-peak" in the Kmer profile***
  - Second peak at twice k-mer coverage as the first: heterozygous kmers average 50x coverage, homozygous kmers average 100x coverage

- ***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)

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)

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