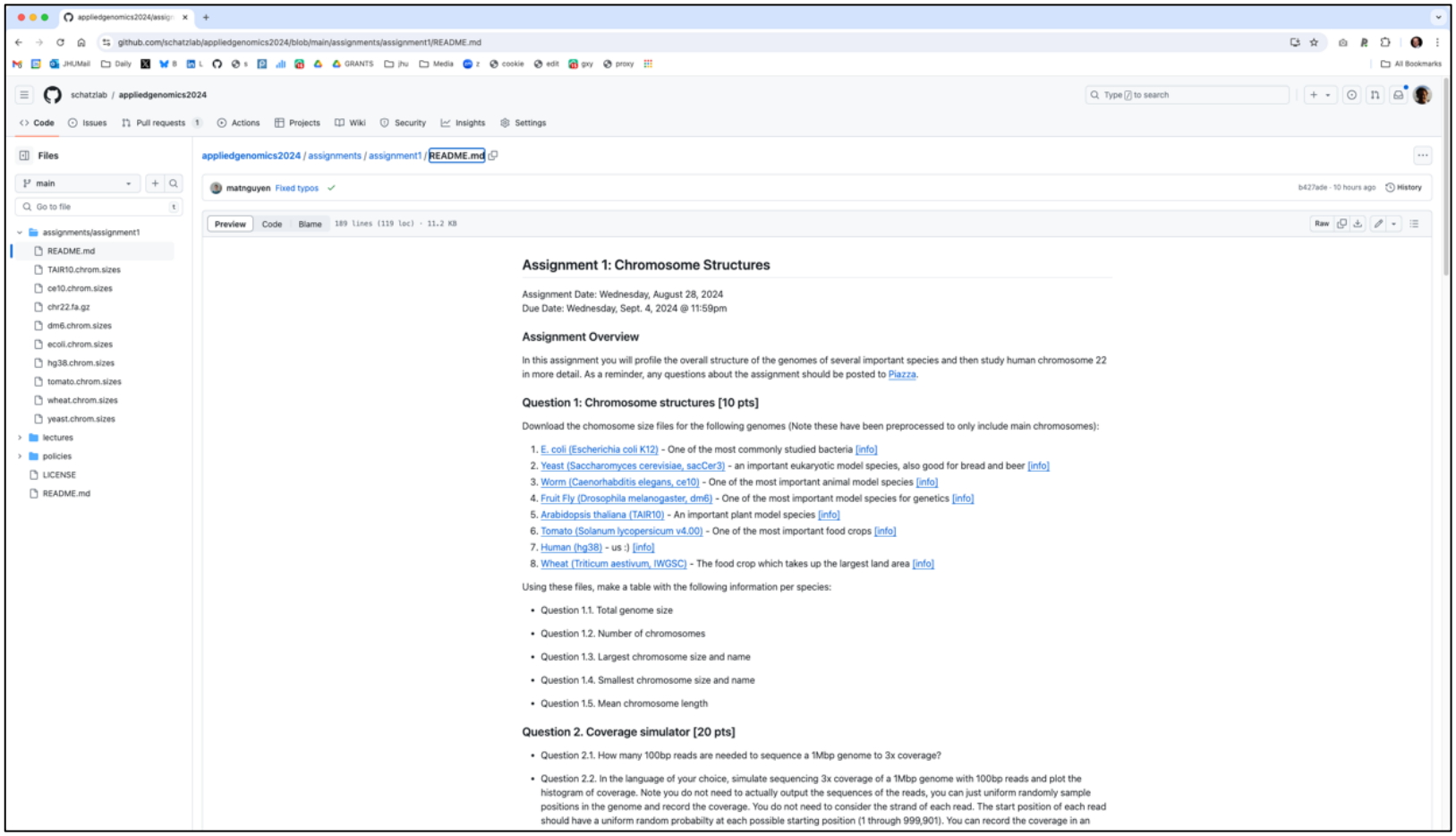
Genome Assembly

Michael Schatz

Sept 4, 2024

Lecture 3: Applied Comparative Genomics

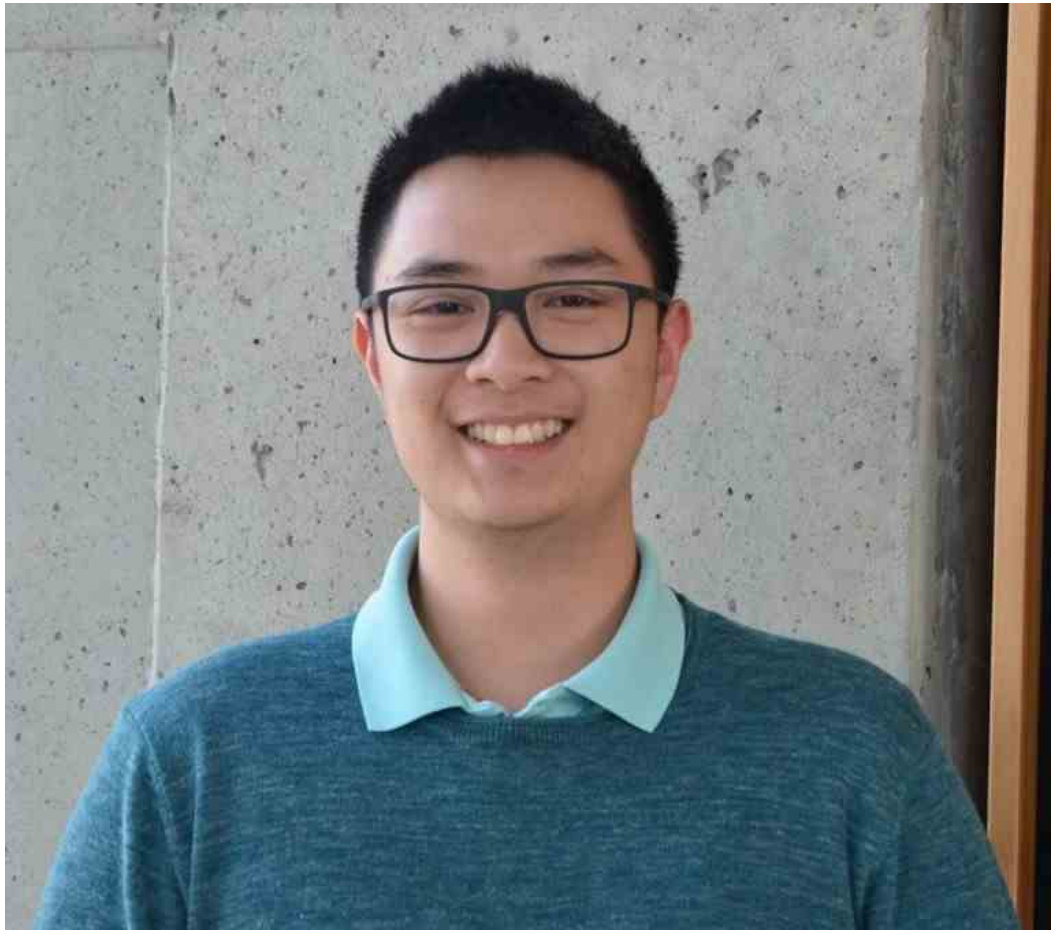
Assignment 1



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

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

TA: Matthew Nguyen



Starting next week:
Office hours Wednesdays 2pm-3pm in Malone 216.

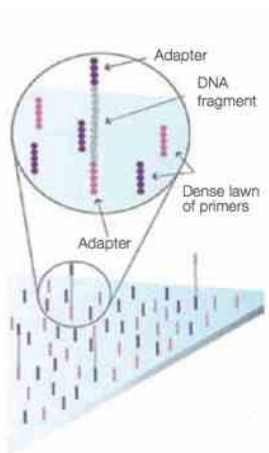
If this time doesn't work for you, you can always DM/email and Matthew will be happy to accommodate!

Part 1: Recap and Illumina Sequencing

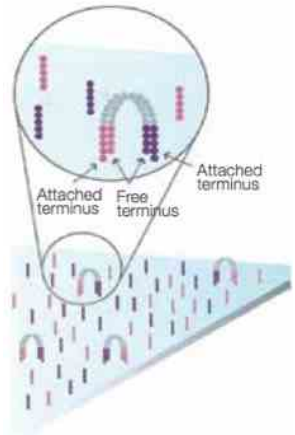
Second Generation Sequencing

Illumina NovaSeq 6000
Sequencing by Synthesis

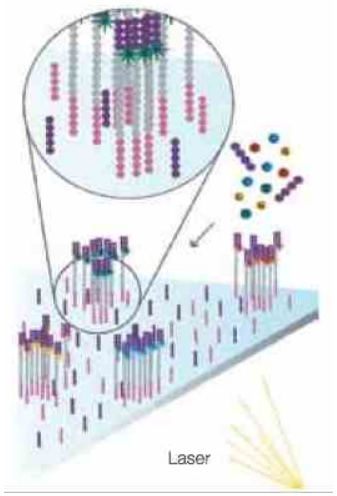
>3Tbp / day
(JHU has 4 of these!)



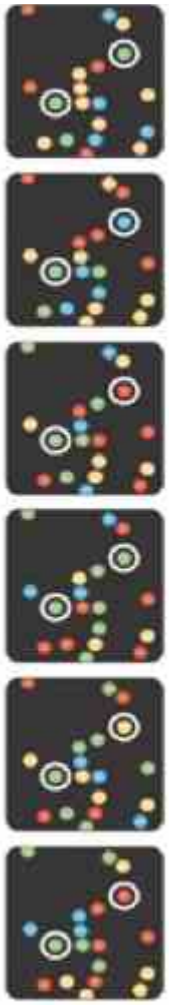
1. Attach



2. Amplify

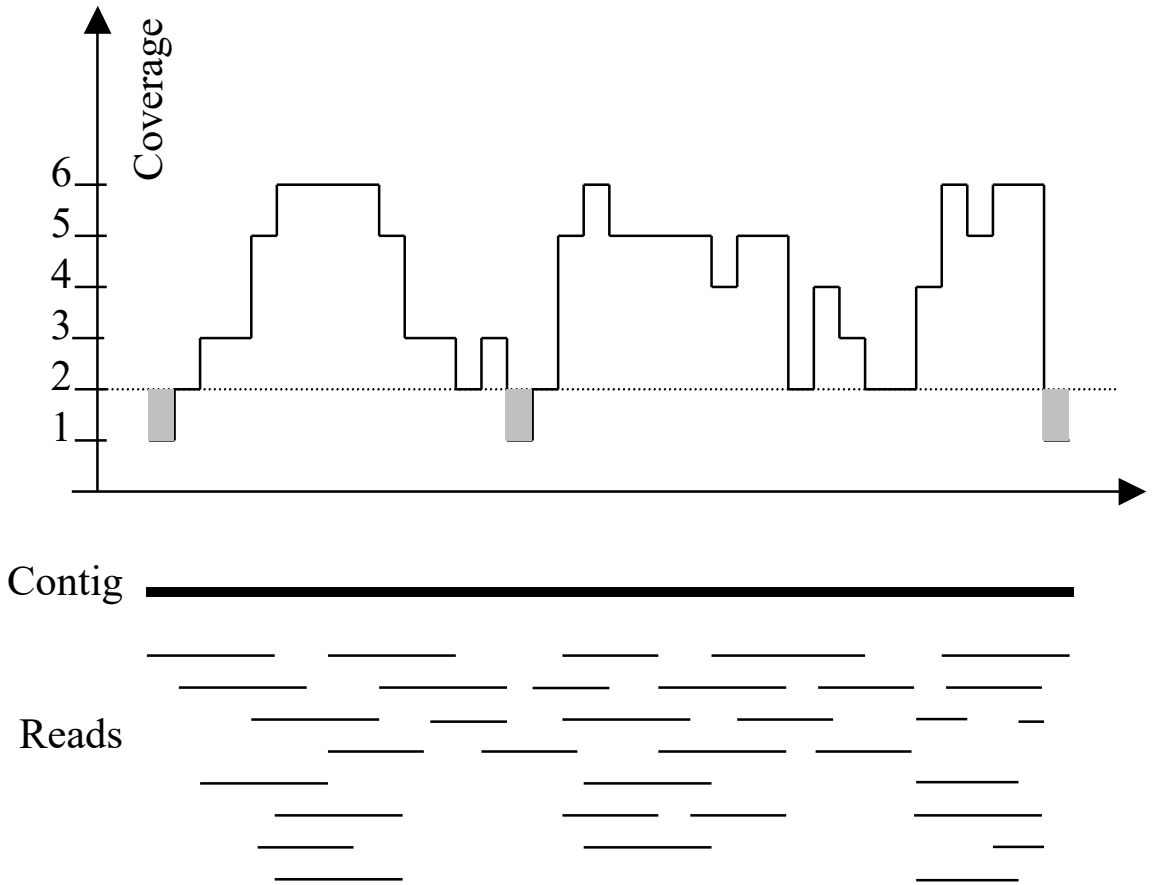


3. Image



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

Typical sequencing coverage

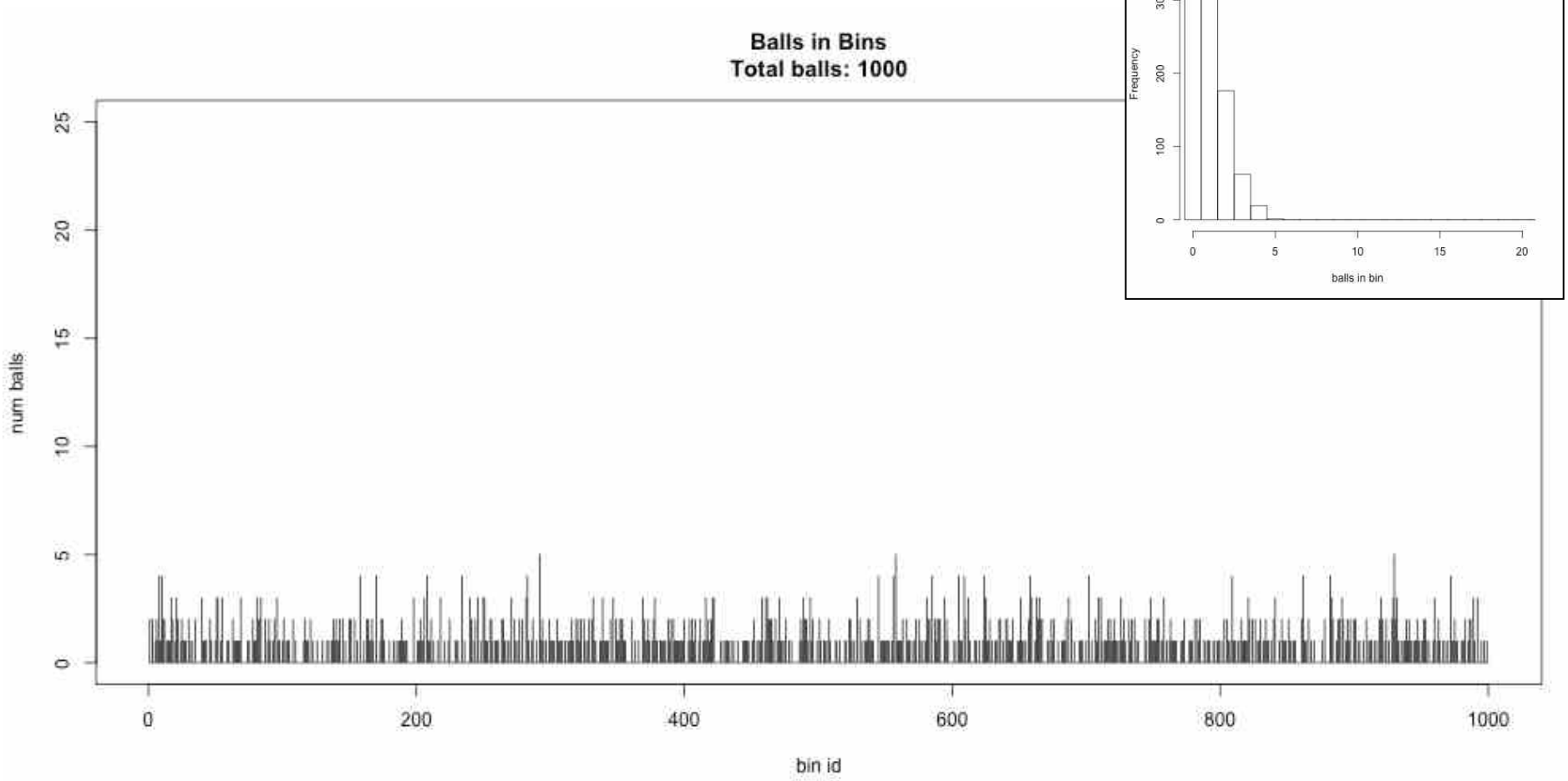


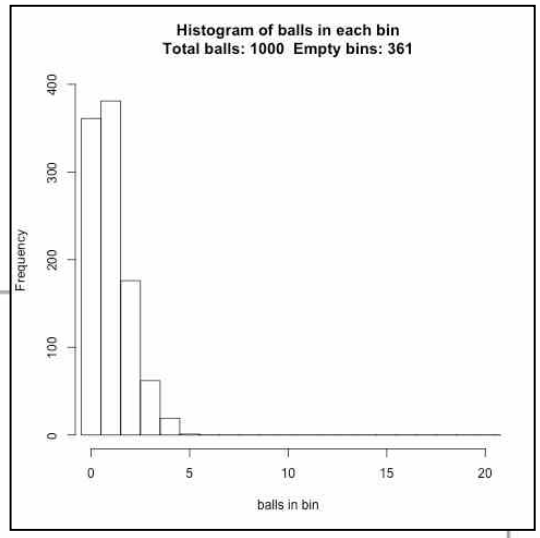
Imagine raindrops on a sidewalk

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

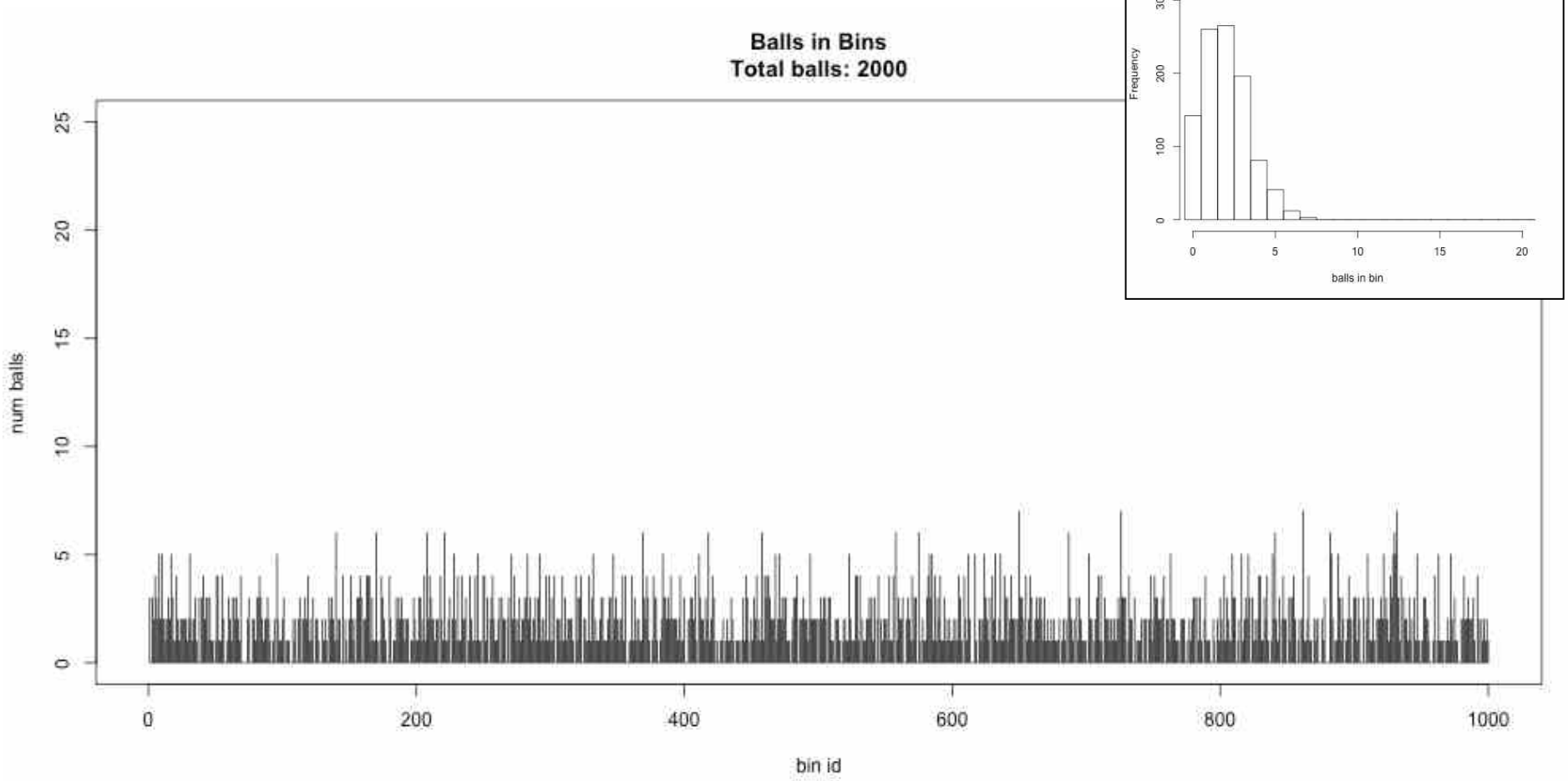
If the genome is 10 Mbp, should we sequence 100k 100bp reads?

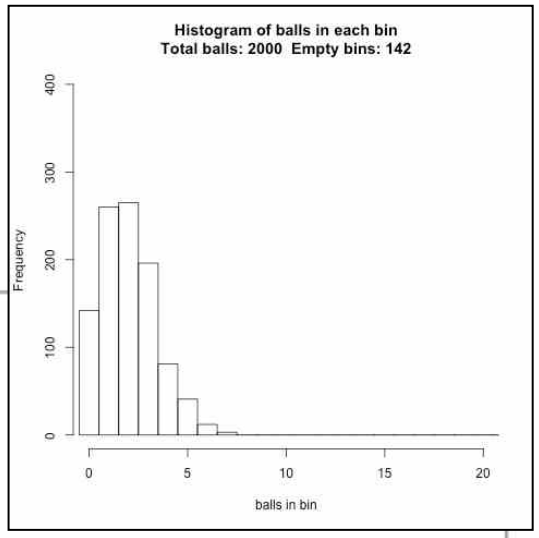
1x sequencing



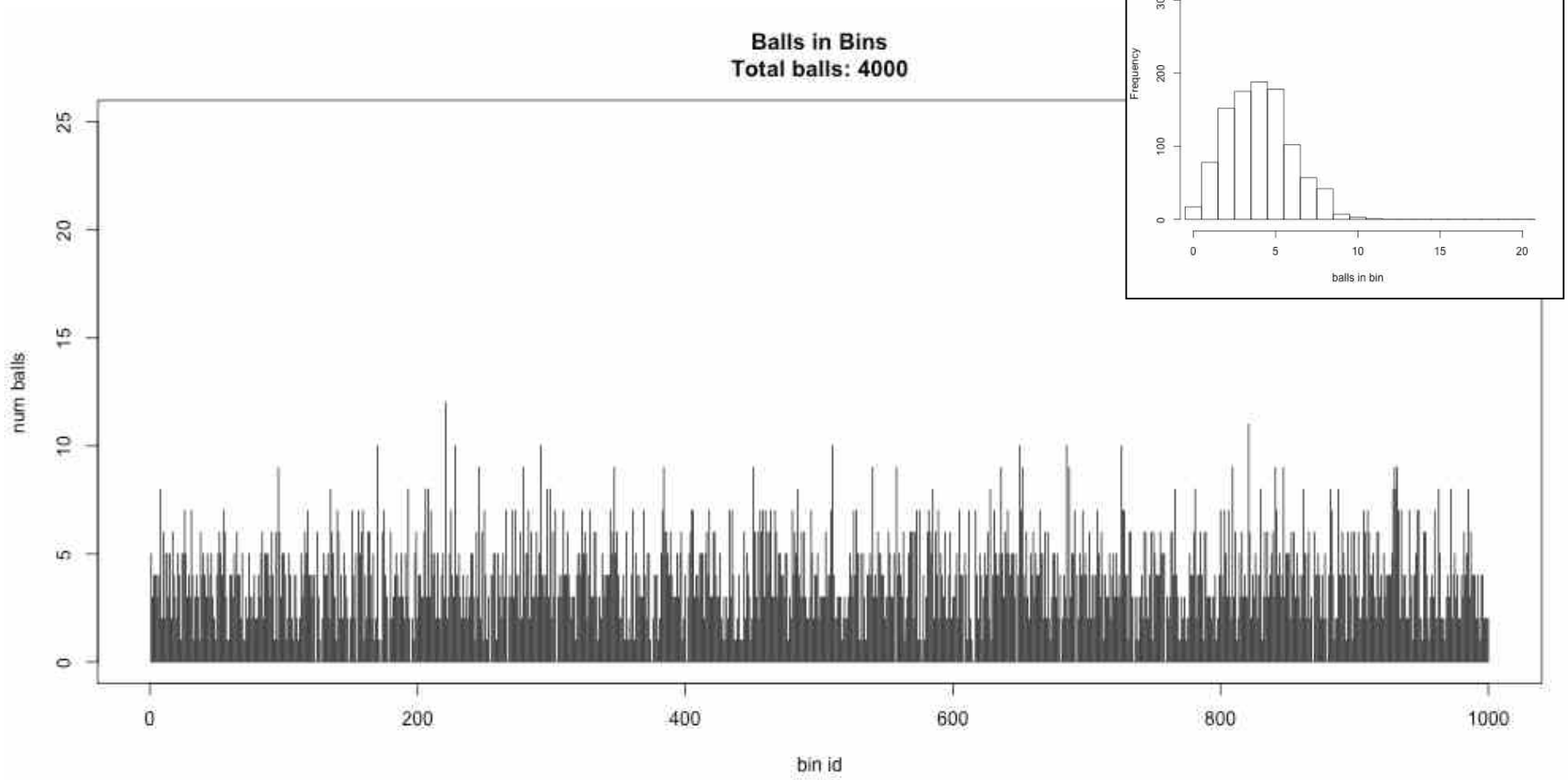


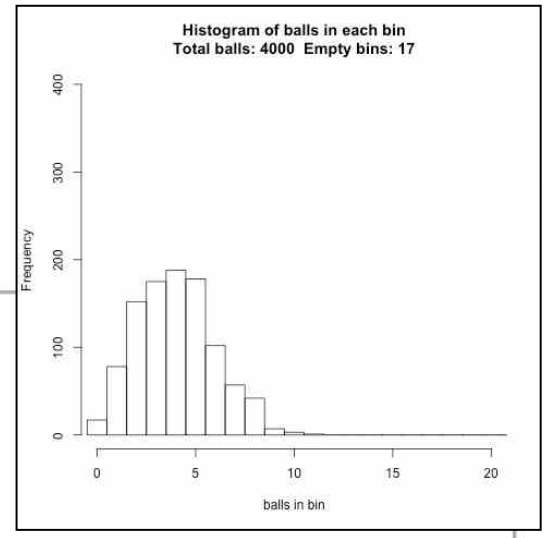
2x sequencing



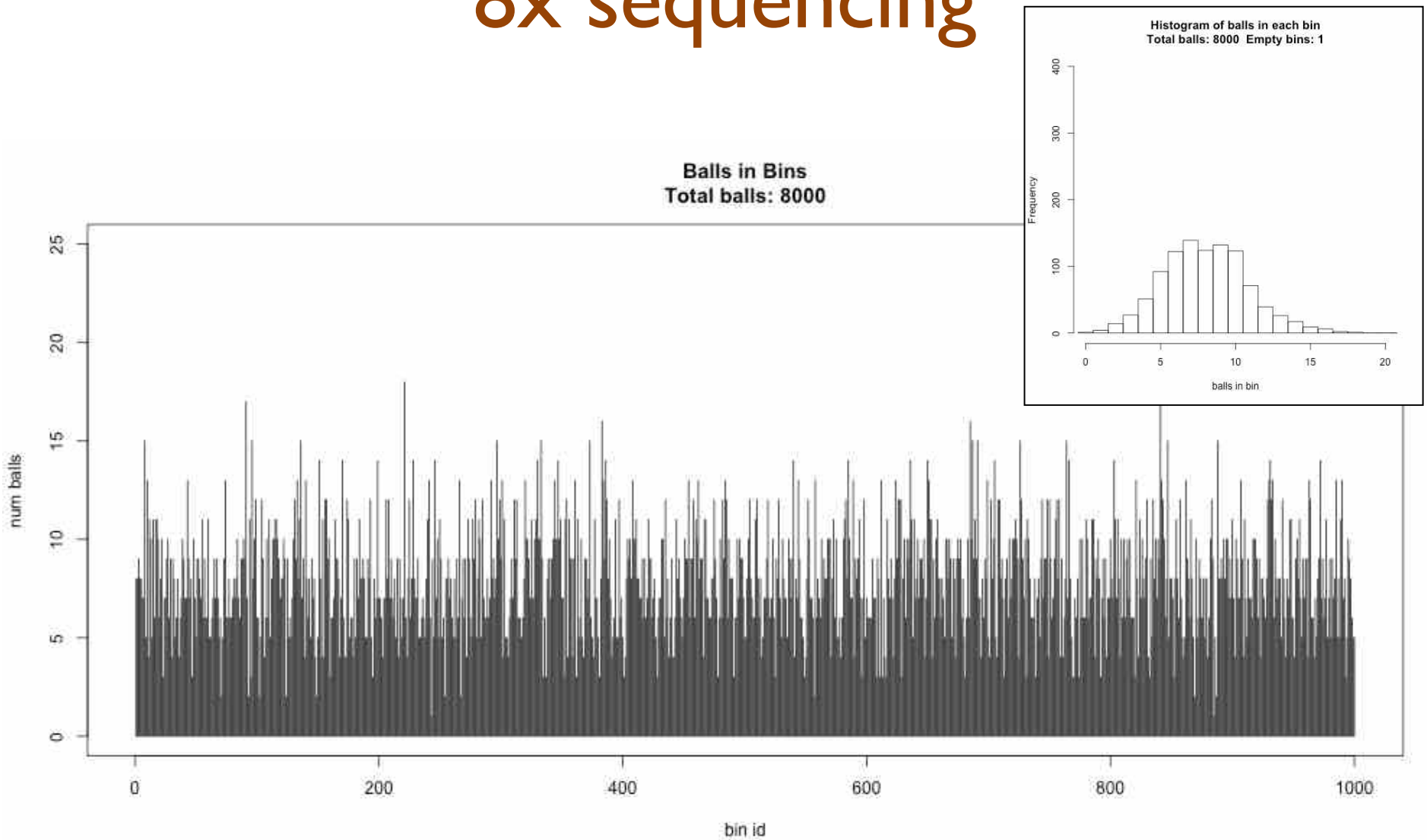


4x sequencing





8x sequencing



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.

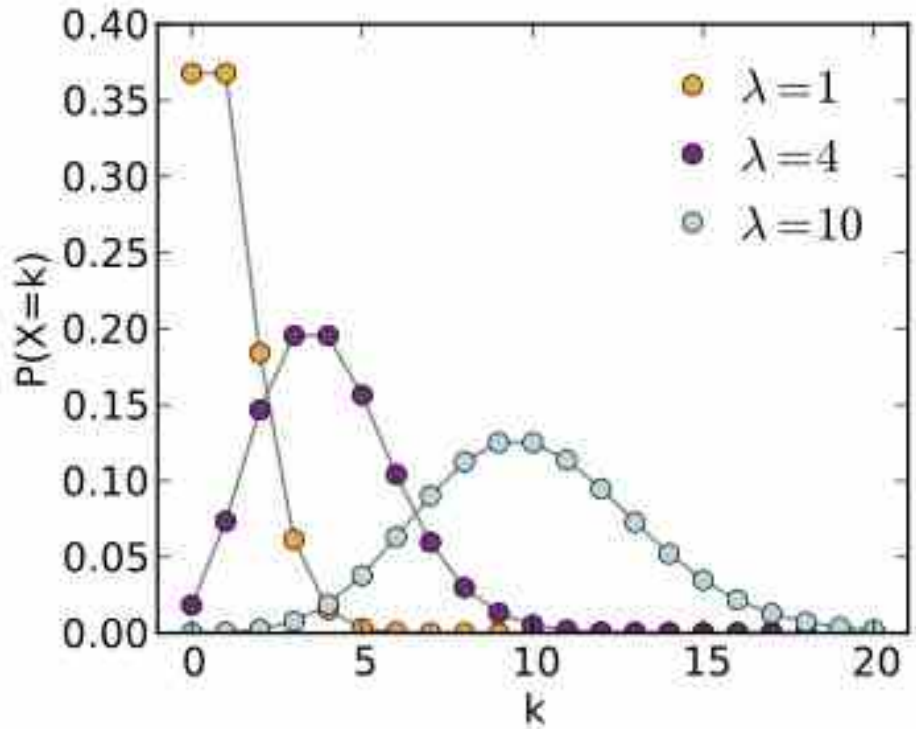
$$P(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Formulation comes from the limit of the binomial equation

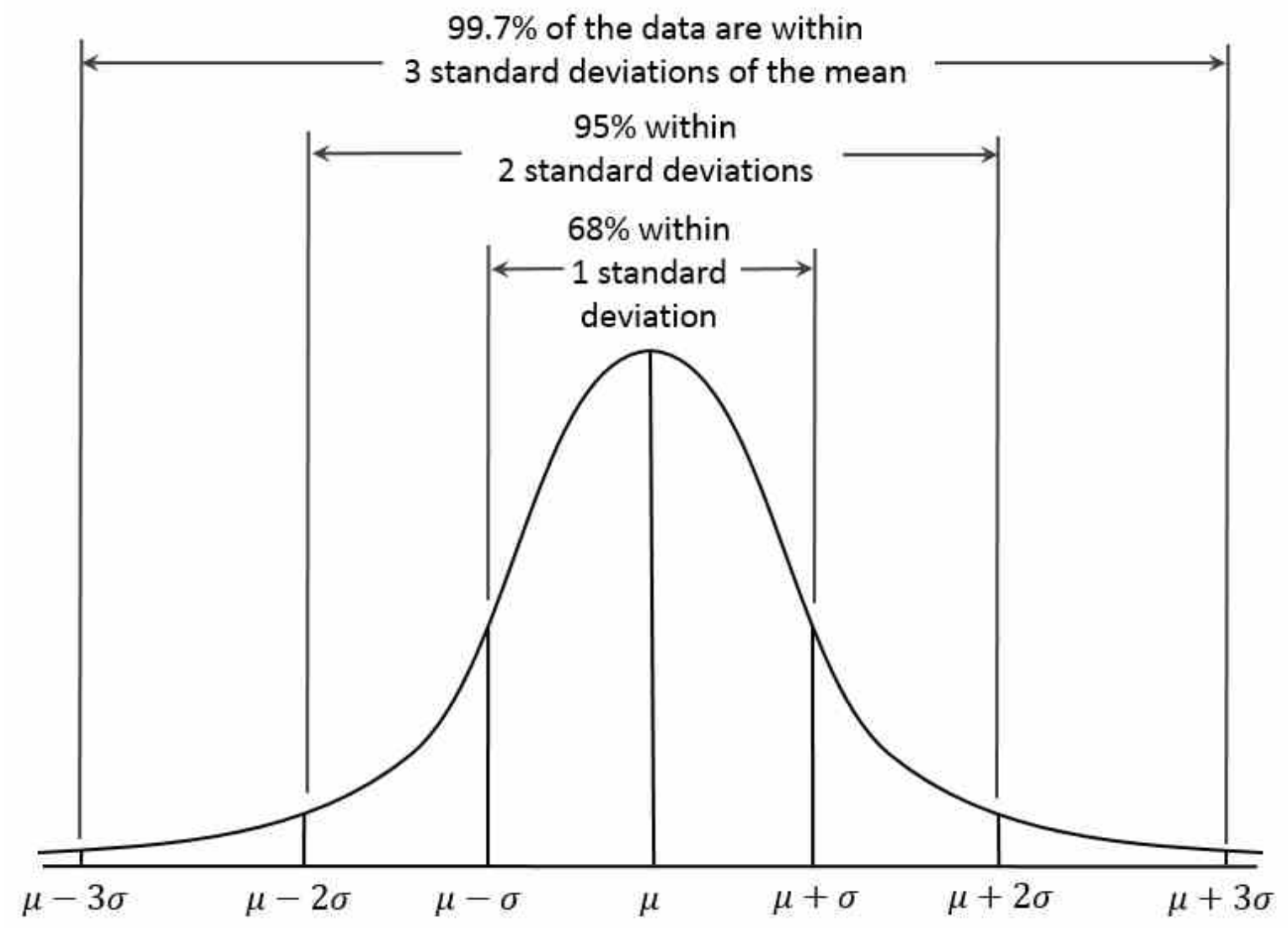
Resembles a normal distribution, but over the positive values, and with only a single parameter.

Key properties:

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



Normal Approximation



Can estimate Poisson distribution as a normal distribution when $\lambda > 10$

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 (50% more $$$)

K-mers and K-mer counting

GATTACATACACATTGGATG

K-mers and K-mer counting

GATTACATACACATTGGATG

GAT ACA ACA ATT GAT

ATT CAT CAC TTG ATG

TTA ATA ACA TGG

TAC TAC CAT GGA

Kmers:

- Divide a string into substrings of length k
- Notice every position is covered k times
- Notice there are G – k + 1 kmers from a string of length G

K-mers and K-mer counting

GATTACATACACATTGGATG

GAT ACA ACA ATT GAT

ATT CAT CAC TTG ATG

TTA ATA ACA TGG

TAC TAC CAT GGA

Kmers:
- Divide a string into substrings of length k
- Notice every position is covered k times
- Notice there are G – k + 1 kmers from a string of length G

Computation: Very easy to compute, exact matches, represent 32mers in 64 bits
Biological: The “atomic unit” of a sequence, creates a fingerprint of a genome/read

K-mers and K-mer counting

GATTACATACACATTGGATG

GAT ACA ACA ATT GAT

ATT CAT CAC TTG ATG

TTA ATA ACA TGG

TAC TAC CAT GGA

GAT:2	CAT:2	ATG:1	TGG:1
ACA:3	CAC:1	TTA:1	TAC:2
ATT:2	TTG:1	ATA:1	GGA:1

K-mers and K-mer counting

GATTACATACACATTGGATG

GAT:2	CAT:2	ATG:1	TGG:1
ACA:3	CAC:1	TTA:1	TAC:2
ATT:2	TTG:1	ATA:1	GGA:1

```
1:  7  (ATG,  TGG,  …)
2:  4  (GAT,  CAT,  ATT,  TAC)
3:  1  (ACA)
```

See HW1

K-mers and K-mer counting

GATTACATACACATTGGATG

```
1:  7  (ATG, TGG, …)
2:  4  (GAT, CAT, ATT, TAC)
3:  1  (ACA)
```

How long should k be?

K-mers and K-mer counting

GATTACATACACATTGGATG

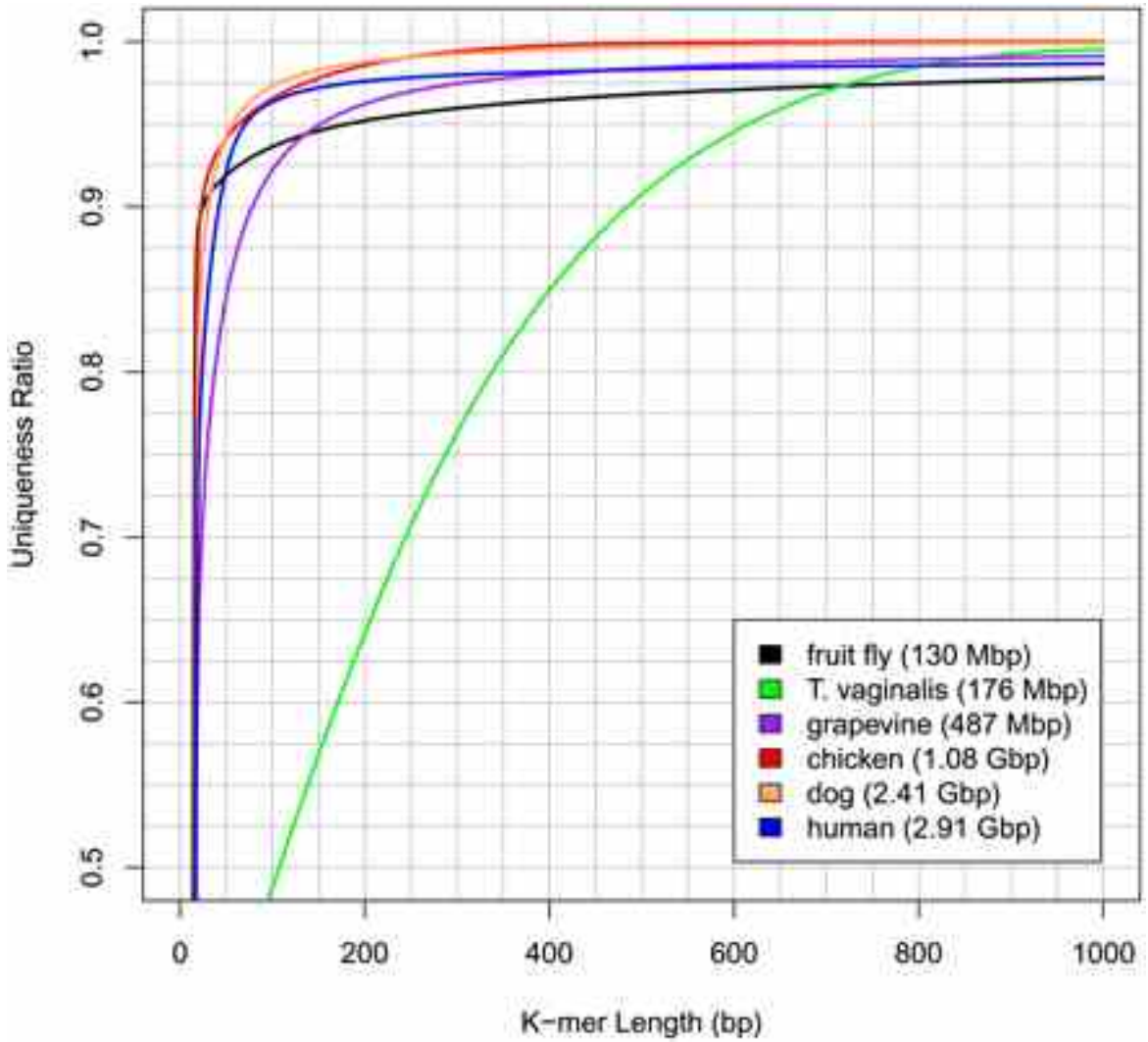
```
1:  7  (ATG, TGG, …)
2:  4  (GAT, CAT, ATT, TAC)
3:  1  (ACA)
```

How long should k be?

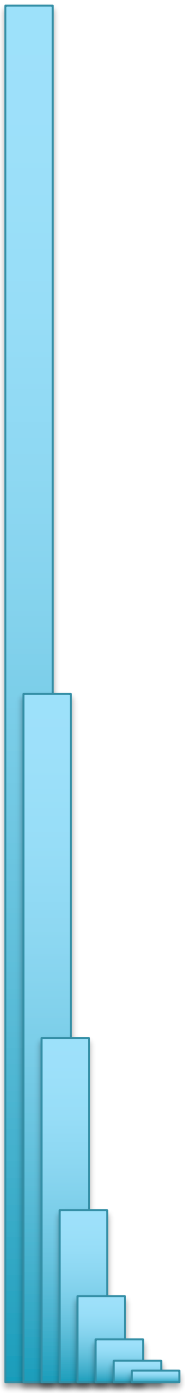
K=1 : Too short, every base is present
K=2 : Too short, every pair of bases will be present

Pick k so that $G/(4^k) << 1$
$k = \log_4 (G)$
At least 15 for human, often a bit longer
But not too long or could loose resolution

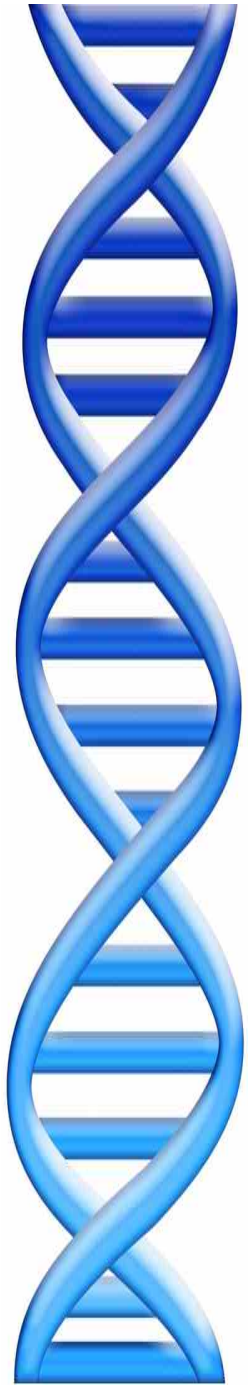
K-mer Uniqueness



Assembly of large genomes using second-generation sequencing
Schatz et al. (2010) Genome Research. doi: 10.1101/gr.101360.109



Part 2: De novo genome assembly

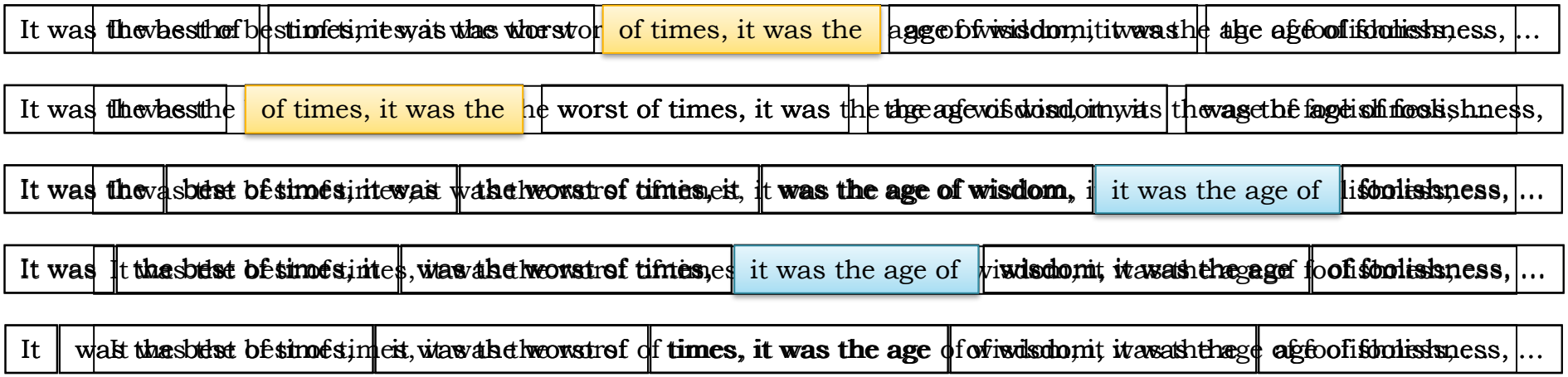


Outline

1. ***Assembly theory***
   - Assembly by analogy
2. ***Practical Issues***
   - Coverage, read length, errors, and repeats
3. ***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



- 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

Greedy Reconstruction

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

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

It was the best of

was the best of times

Sub-fragment *k*=4

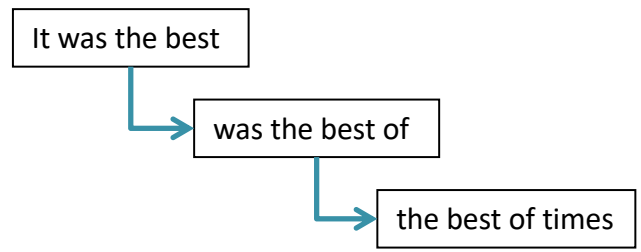
It was the best

was the best of

was the best of

the best of times

Directed edges (overlap by *k*-1)

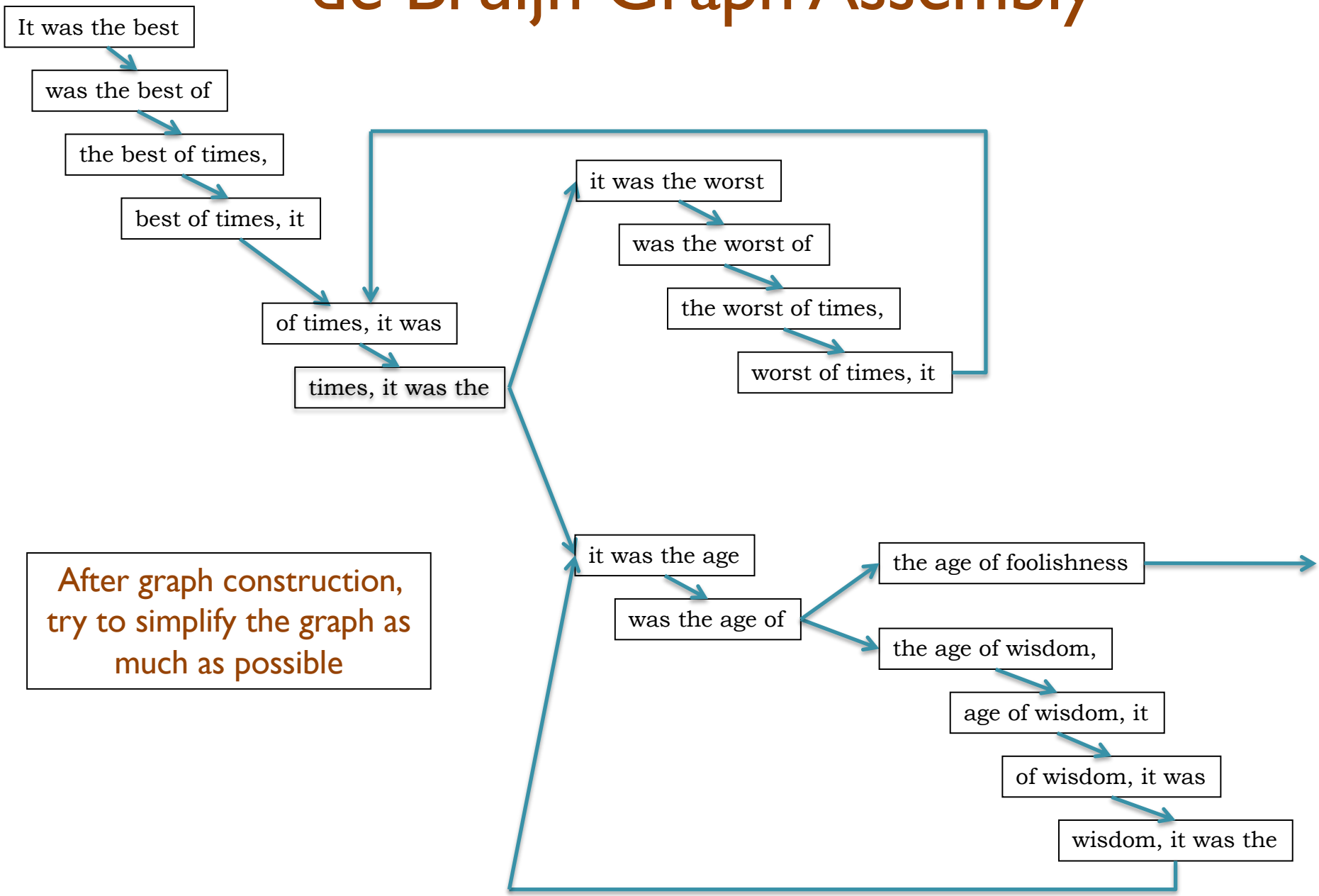


– Overlaps between fragments are implicitly computed

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

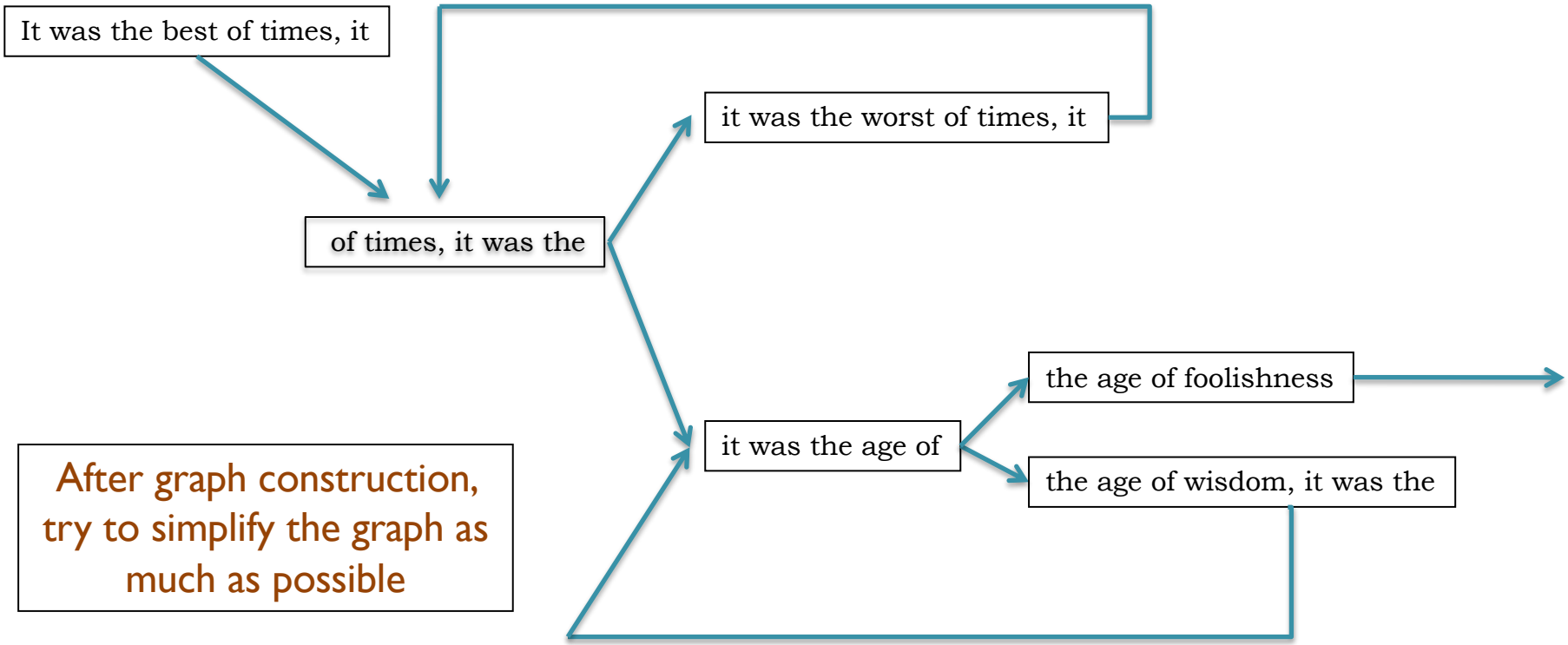
de Bruijn, 1946
Idury et al., 1995
Pevzner et al., 2001

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

The full tale

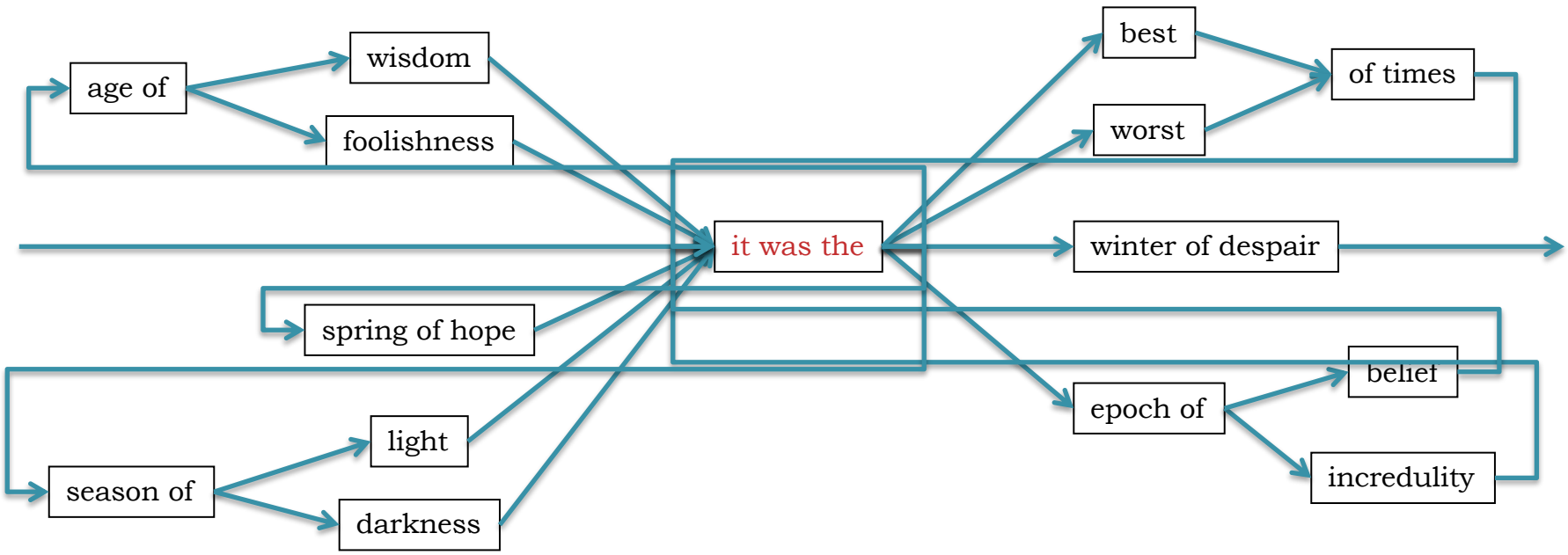
… it was the best of times it was the worst of times …

… it was the age of wisdom it was the age of foolishness …

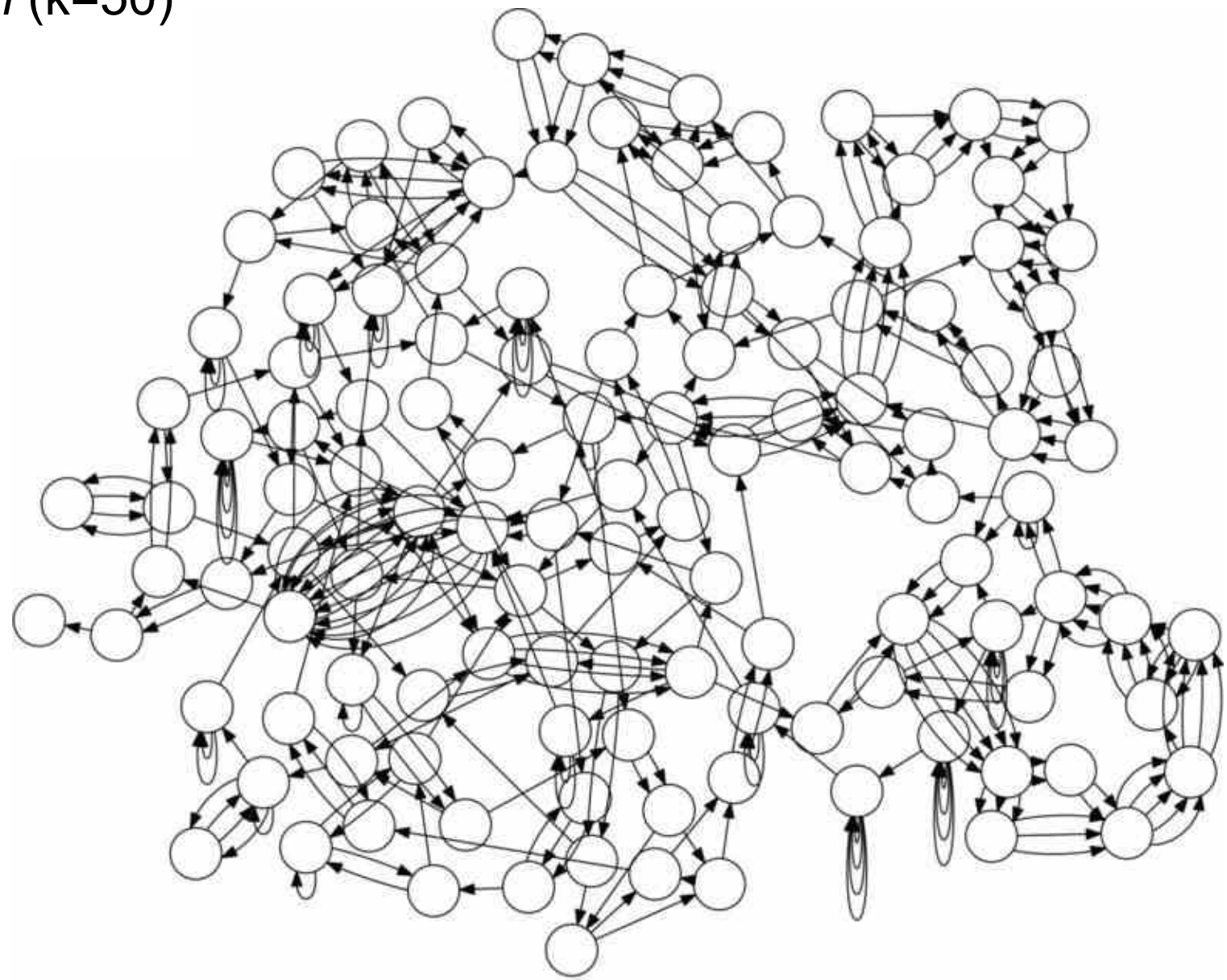
… it was the epoch of belief it was the epoch of incredulity …

… it was the season of light it was the season of darkness …

… it was the spring of hope it was the winder of despair …



E. coli (k=50)

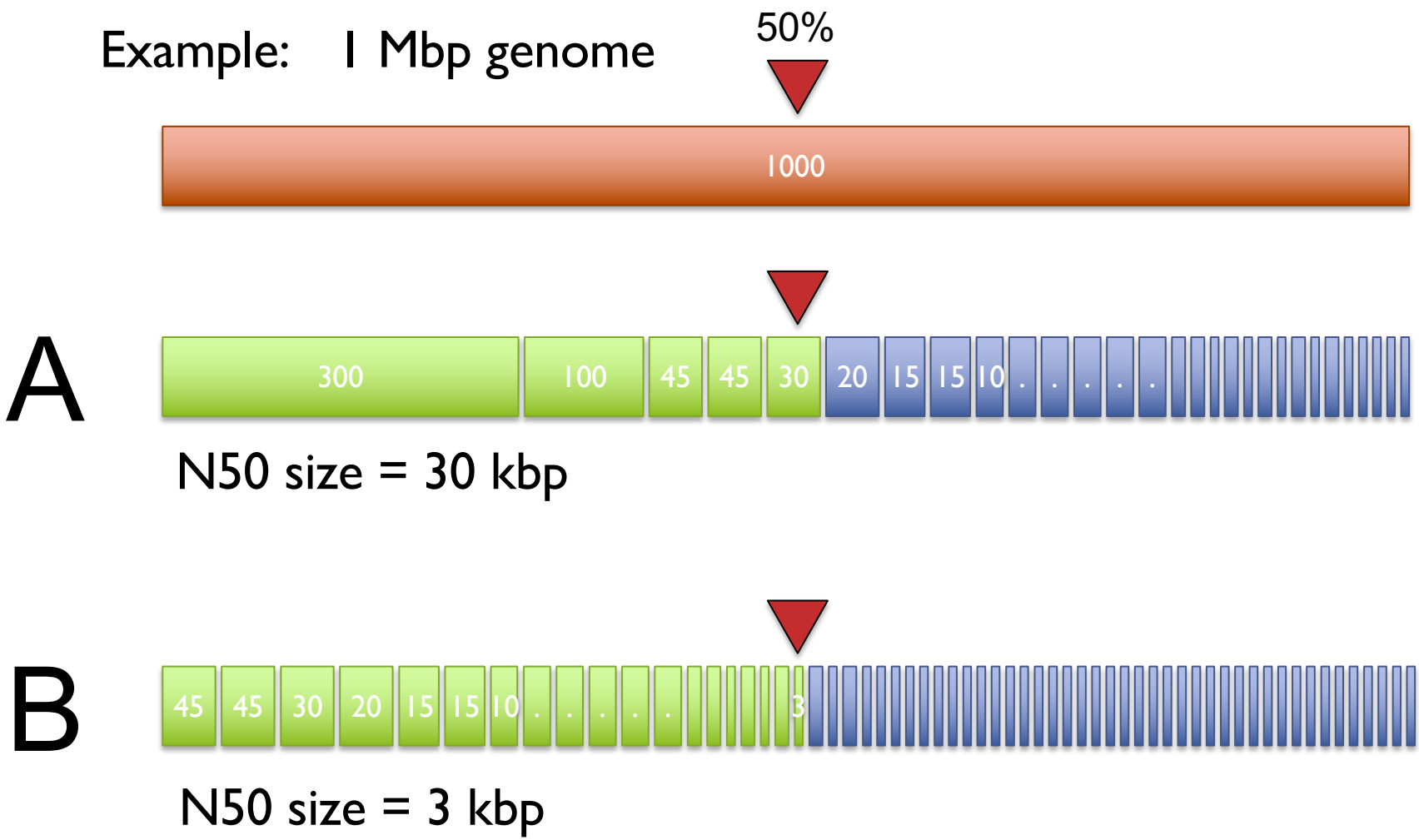


Reducing assembly complexity of microbial genomes with single-molecule sequencing
Koren et al (2013) Genome Biology. **14**:R101 https://doi.org/10.1186/gb-2013-14-9-r101

Contig N50

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

Example: 1 Mbp genome



Contig N50

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

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 1

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

ATTA

GATT

TACA

TTAC

Pop Quiz 1

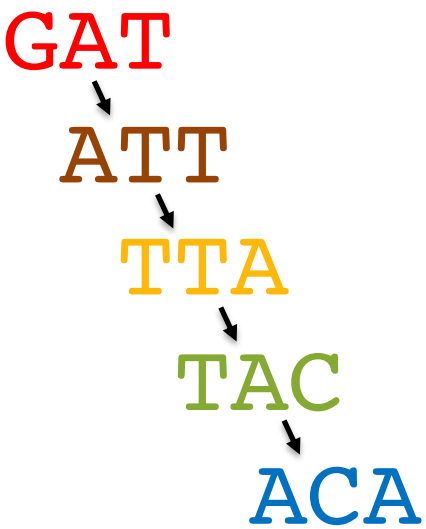
Assemble these reads using a de Bruijn graph approach (k=3):

```
ATTA: ATT -> TTA
GATT: GAT -> ATT
TACA: TAC -> ACA
TTAC: TTA -> TAC
```

Pop Quiz 1

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

```
ATTA:  ATT  ->  TTA
GATT:  GAT  ->  ATT
TACA:  TAC  ->  ACA
TTAC:  TTA  ->  TAC
```



GATTACA

Pop Quiz 2

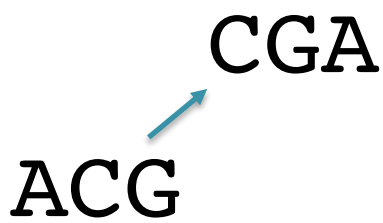
Assemble these reads using a de Bruijn graph approach (k=3):

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

Pop Quiz 2

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

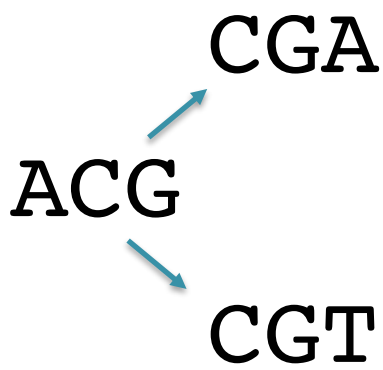
~~ACGA~~
ACGT
ATAC
CGAC
CGTA
GACG
GTAT
TACG



Pop Quiz 2

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

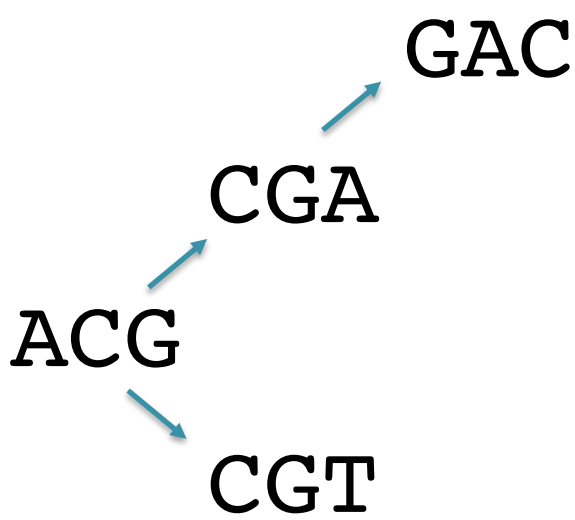
~~ACGA~~
~~ACGT~~
ATAC
CGAC
CGTA
GACG
GTAT
TACG



Pop Quiz 2

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

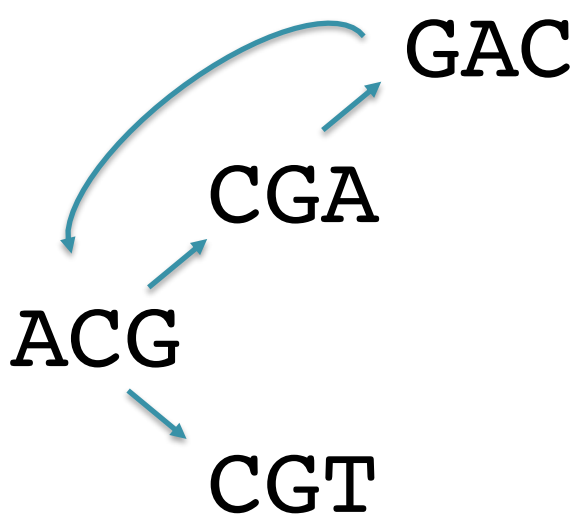
~~ACGA~~
~~ACGT~~
ATAC
~~CGAC~~
CGTA
GACG
GTAT
TACG



Pop Quiz 2

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

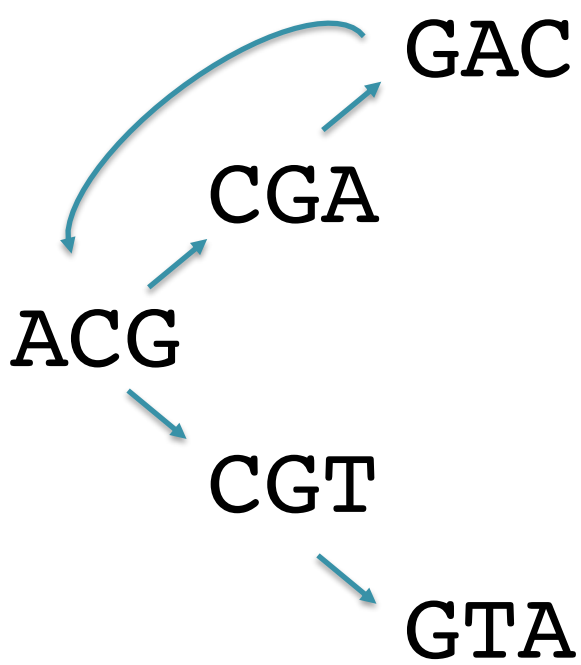
~~ACGA~~
~~ACGT~~
ATAC
~~CGAC~~
CGTA
~~GACG~~
GTAT
TACG



Pop Quiz 2

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

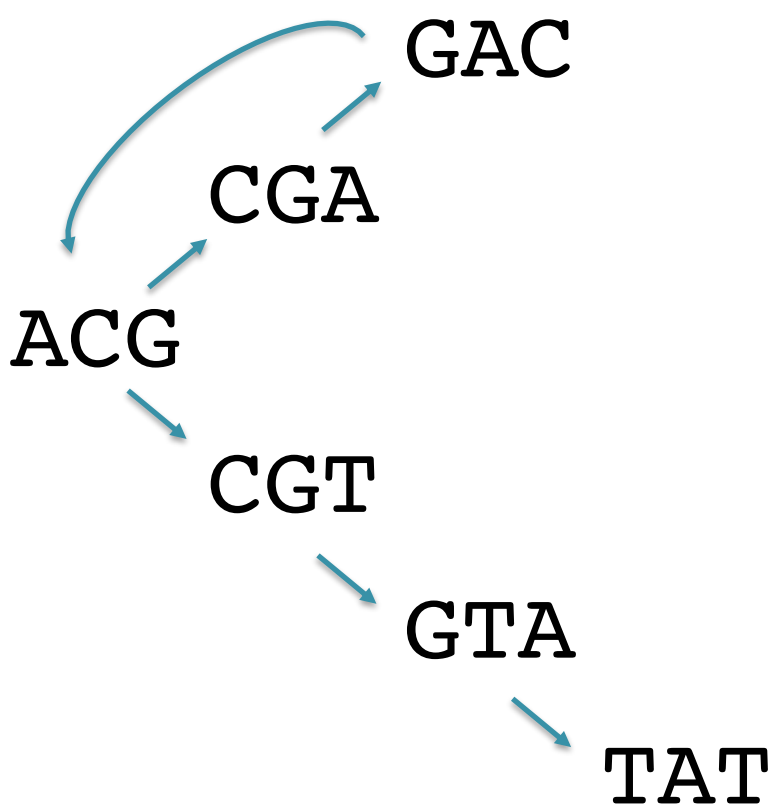
~~ACGA~~
~~ACGT~~
ATAC
~~CGAC~~
~~CGTA~~
~~GACG~~
GTAT
TACG



Pop Quiz 2

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

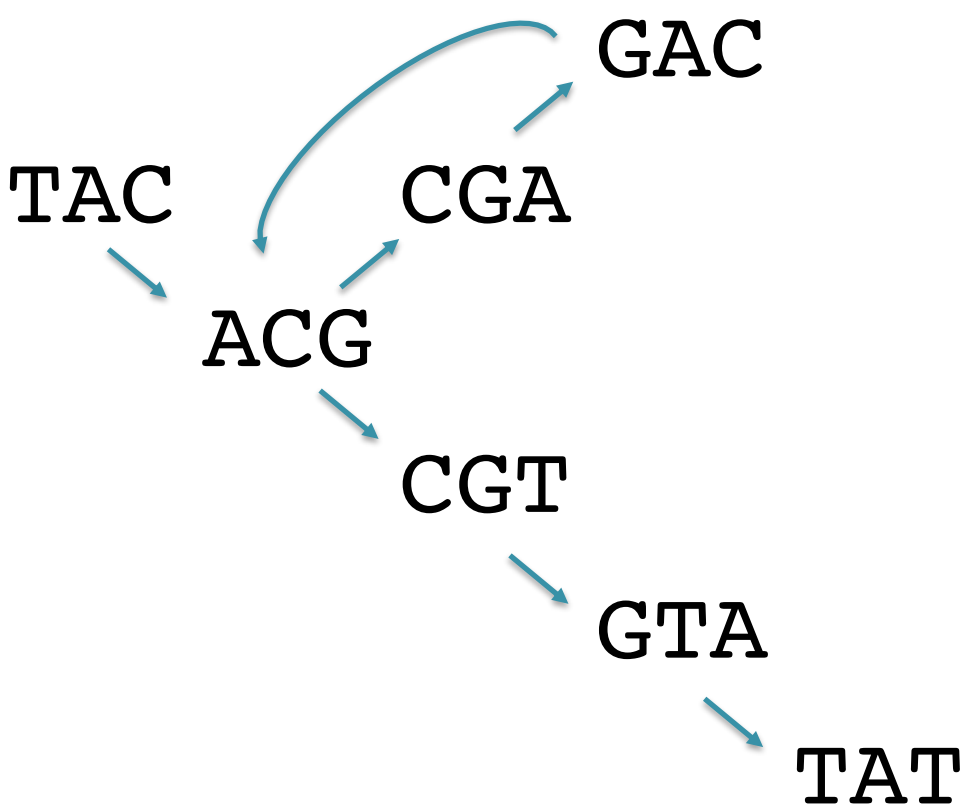
~~ACGA~~
~~ACGT~~
ATAC
~~CGAC~~
~~CGTA~~
~~GACG~~
~~GTAT~~
TACG



Pop Quiz 2

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

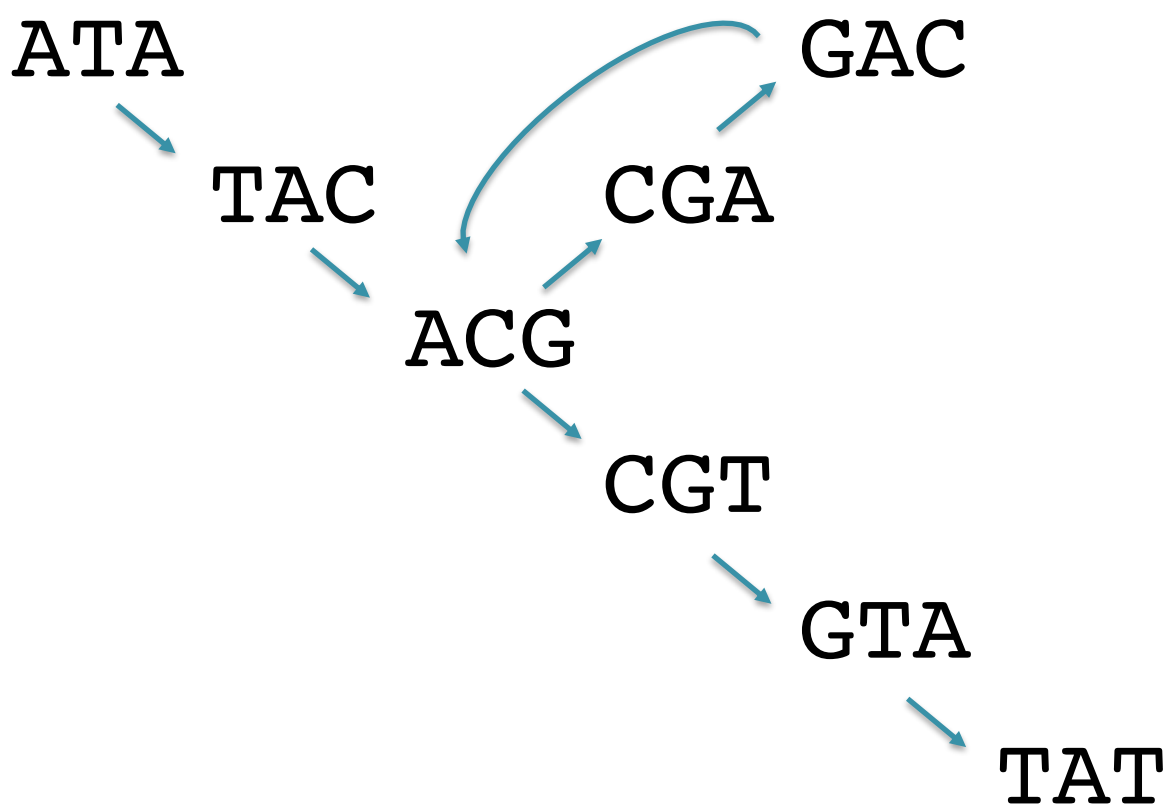
~~ACGA~~
~~ACGT~~
ATAC
~~CGAC~~
~~CGTA~~
~~GACG~~
~~GTAT~~
~~TACG~~



Pop Quiz 2

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

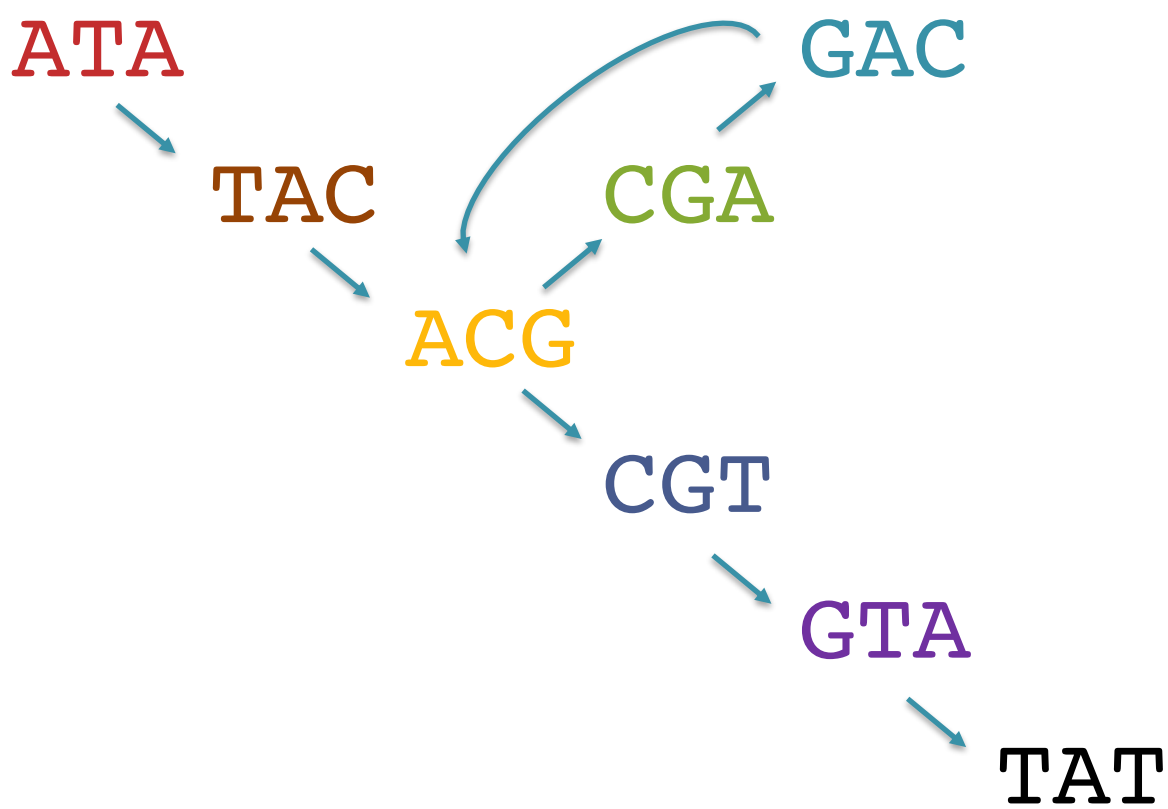
~~ACGA~~
~~ACGT~~
~~ATAC~~
~~CGAC~~
~~CGTA~~
~~GACG~~
~~GTAT~~
~~TACG~~



Pop Quiz 2

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

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

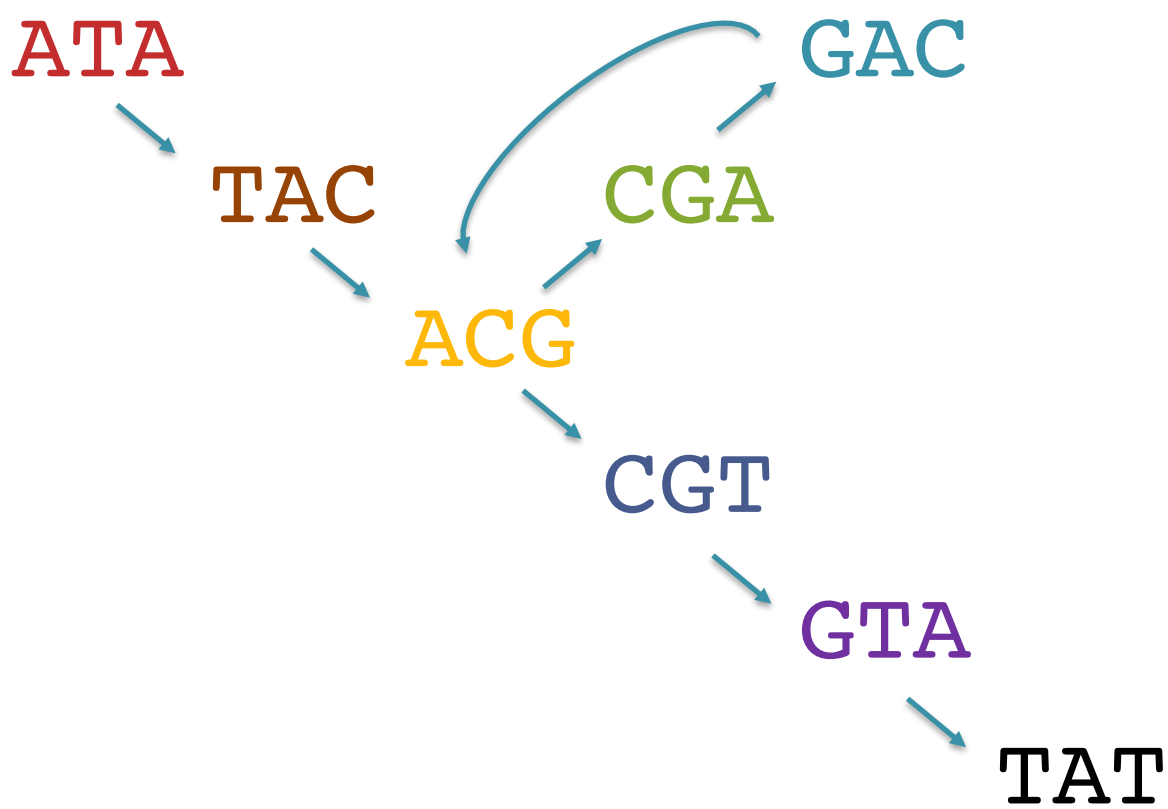


ATACGACGTAT

Pop Quiz 2

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

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



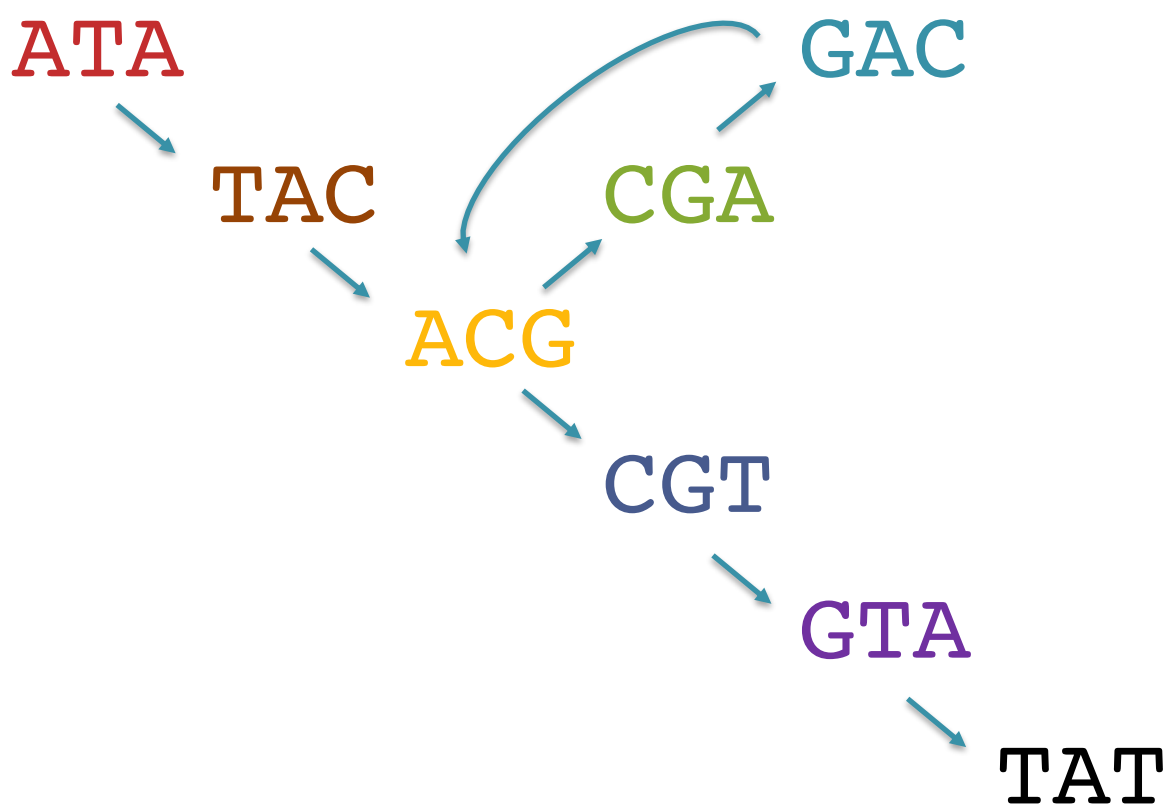
Whats another possible genome?

ATACGACGTAT

Pop Quiz 2

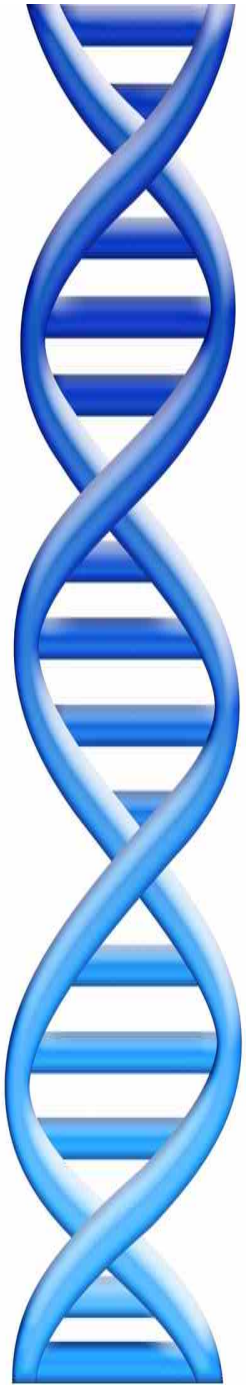
Assemble these reads using a de Bruijn graph approach (k=3):

~~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

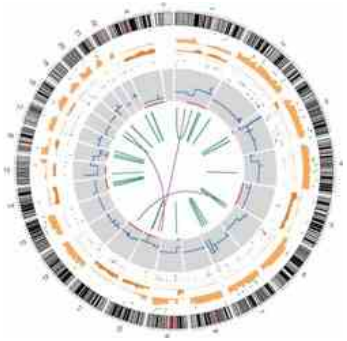
3. Whole Genome Alignment
   - MUMmer recommended

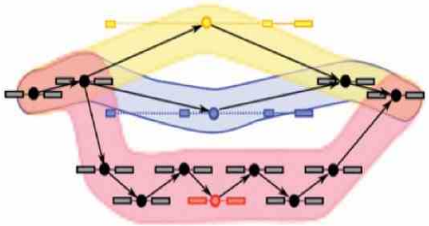
Assembly Applications

- Novel genomes

- Metagenomes

- Sequencing assays
  - Structural variations
  - Transcript assembly
  - …



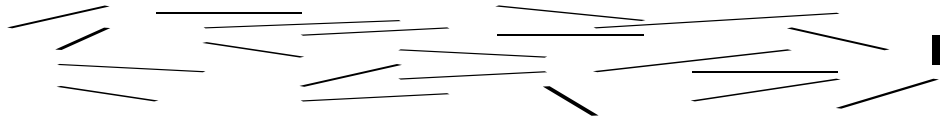


Why are genomes hard to assemble?

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

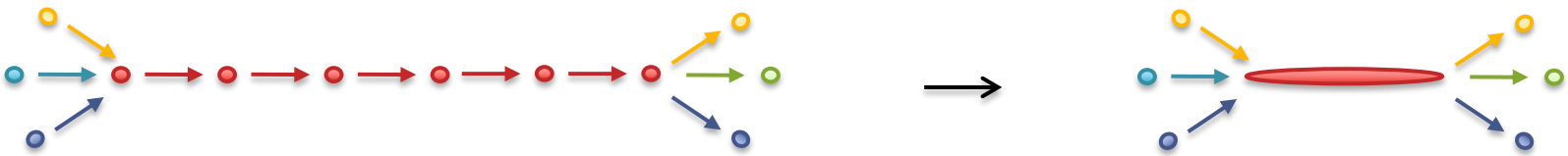
1. Shear & Sequence DNA



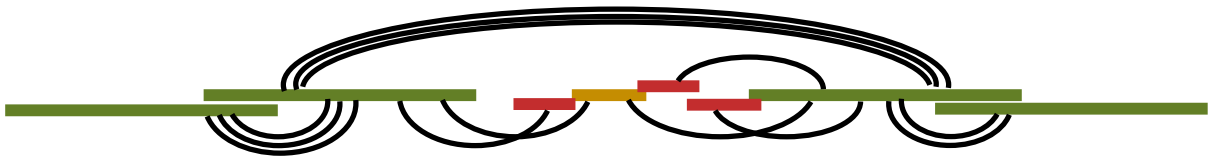
2. Construct assembly graph from reads (de Bruijn / overlap graph)

...AGCCTAGGGATGCGCGACACGT
GGATGCGCGACACGTCGCATATCCGGTTTGGTCAACCTCGGACGGAC
CAACCTCGGACGGACCTCAGCGAA...

3. Simplify assembly graph

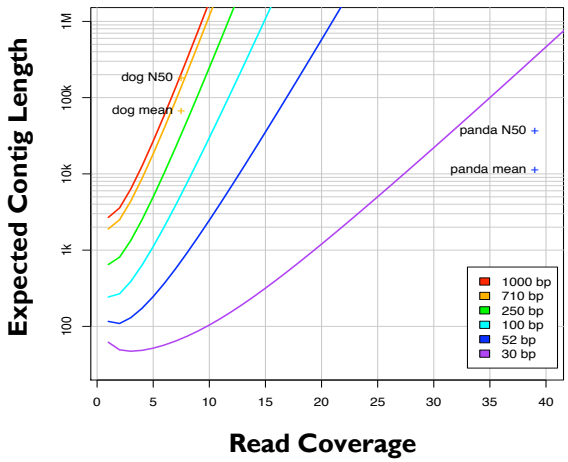


4. Detangle graph with long reads, mates, and other links



Ingredients for a good assembly

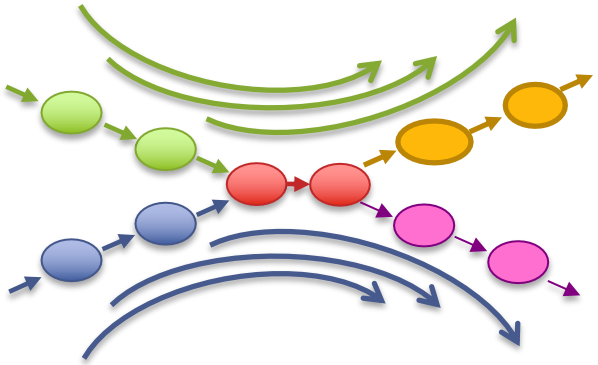
Coverage



High coverage is required

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

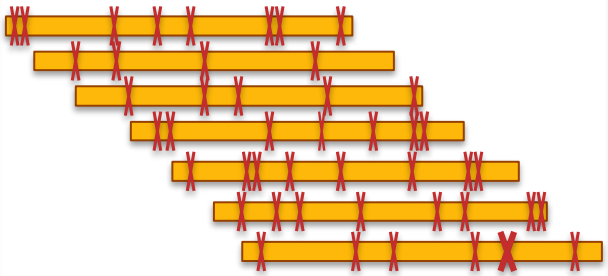
Read Length



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

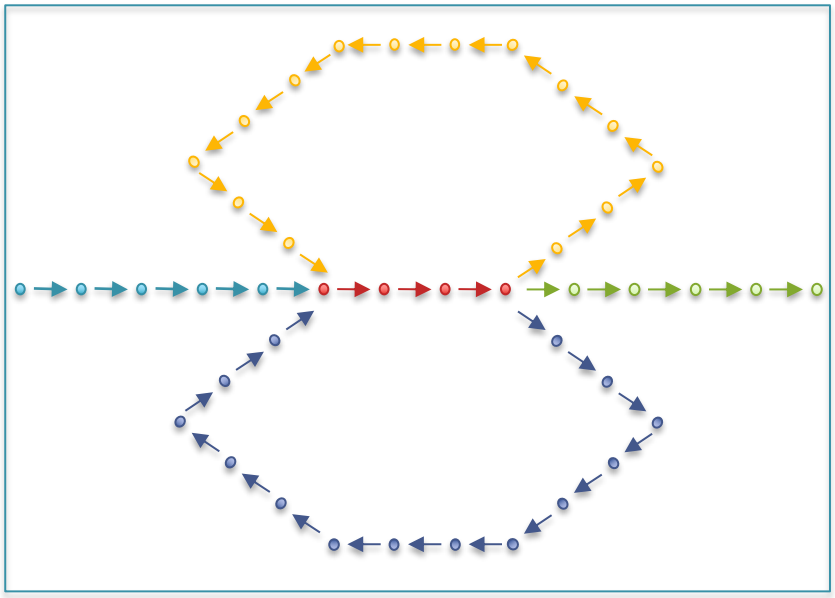
- Reads are assembled by finding kmers shared in pair of reads
- High error rate requires very short seeds, increasing complexity and forming assembly hairballs

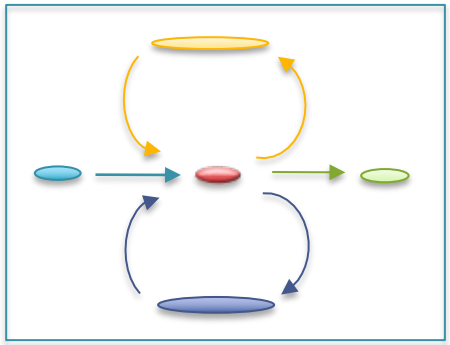
Current challenges in *de novo* plant genome sequencing and assembly
Schatz MC, Witkowski, McCombie, WR (2012) *Genome Biology*. 12:243

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

Errors in the graph

(Chaisson, 2009)

Clip Tips

was the worst of times,

was the worst of t**y**mes,

the worst of times, it

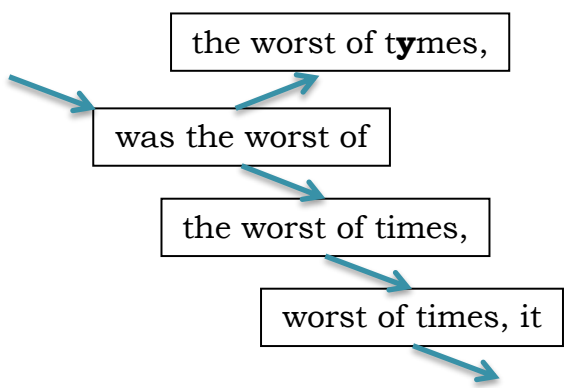
Pop Bubbles

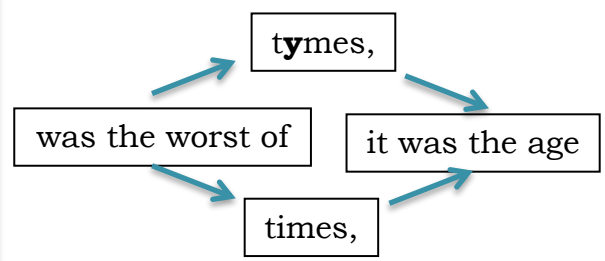
was the worst of times,

was the worst of t**y**mes,

times, it was the age

t**y**mes, it was the age



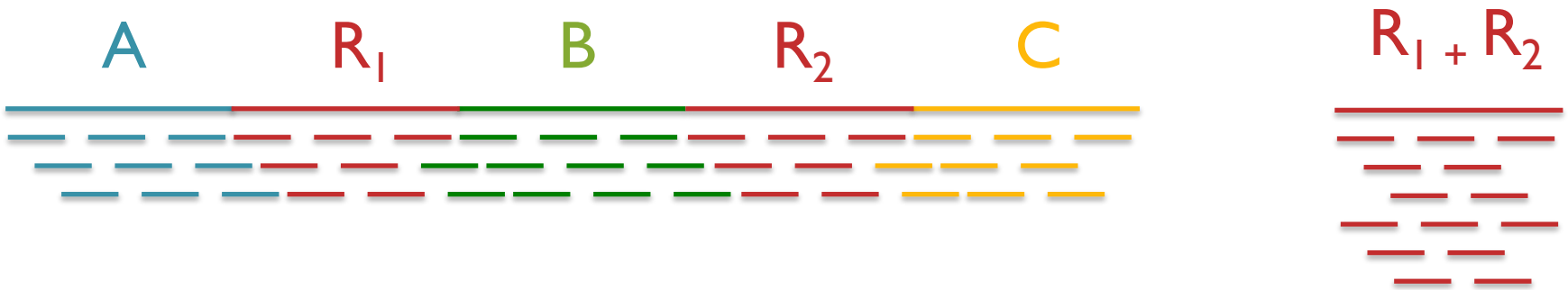


Repetitive regions

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

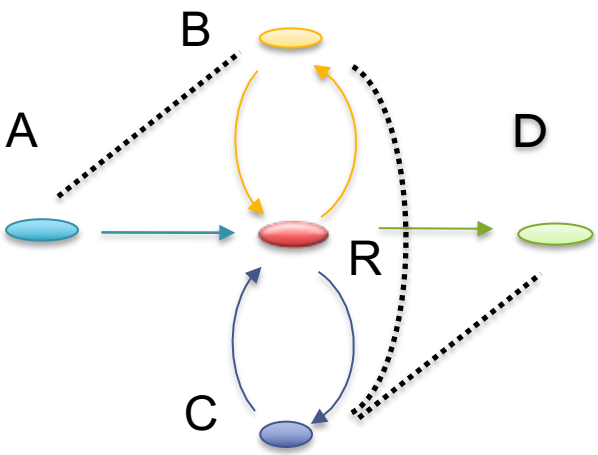
$$\Pr(X-copy)=\binom{n}{k}\left(\frac{X\Delta}{G}\right)^{k}\left(\frac{G-X\Delta}{G}\right)^{n-k}$$

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

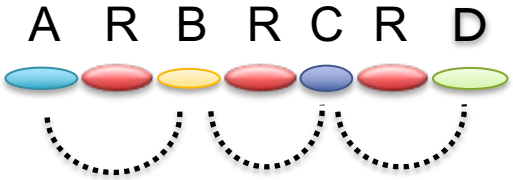
The fragment assembly string graph
Myers, EW (2005) Bioinformatics. 21(suppl 2): ii79-85.

Scaffolding

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



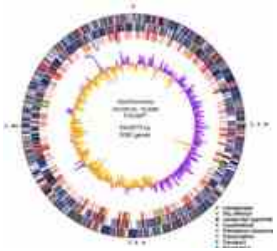
- Use mate-pairs to resolve correct order through assembly graph
  - Place sequence to satisfy the mate constraints
  - Mates through repeat nodes are tangled



- 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

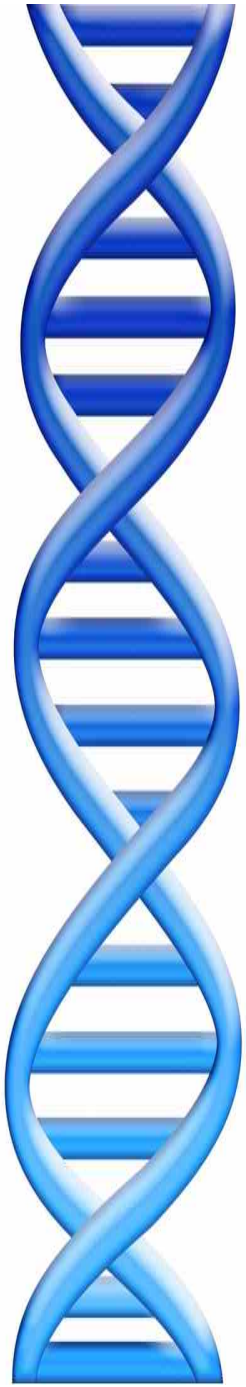


Assembly quality depends on

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 assembl
  - Integrates error correction and scaffolding



Outline

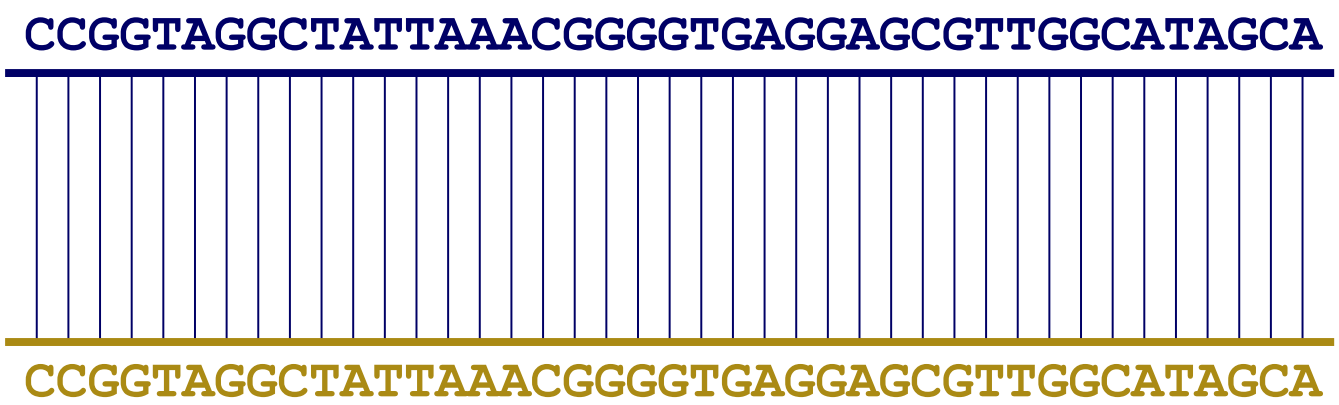
1. ***Assembly theory***
   - Assembly by analogy
2. ***Practical Issues***
   - Coverage, read length, errors, and repeats
3. **Whole Genome Alignment**
   - **MUMmer recommended**

Whole Genome Alignment with MUMmer

Slides Courtesy of Adam M. Phillippy
NHGRI

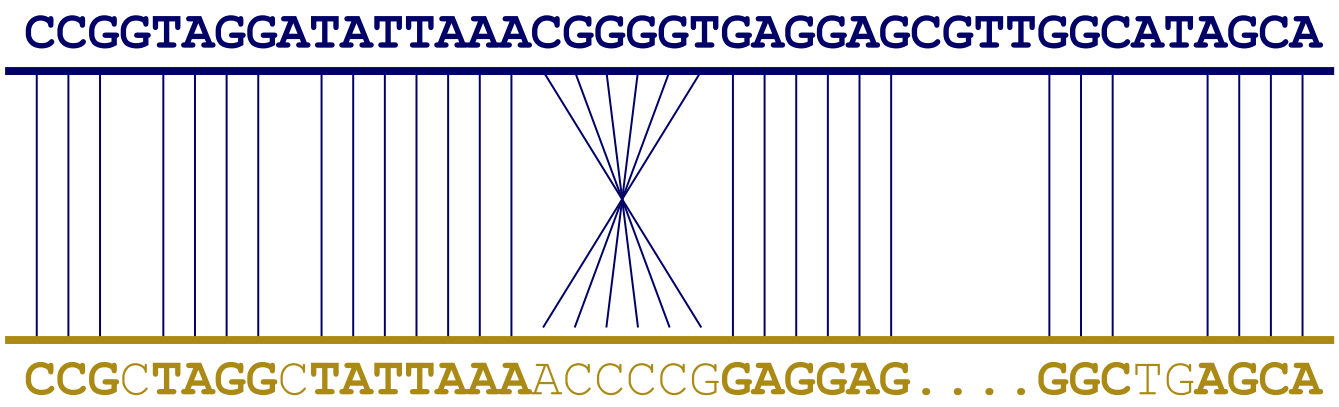
Goal of WGA

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



Not so fast...

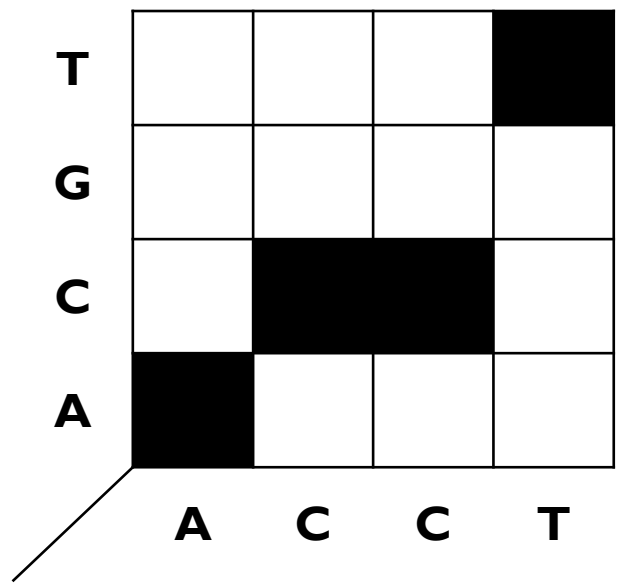
- Genome *A* may have insertions, deletions, translocations, inversions, duplications or SNPs with respect to *B* (sometimes all of the above)



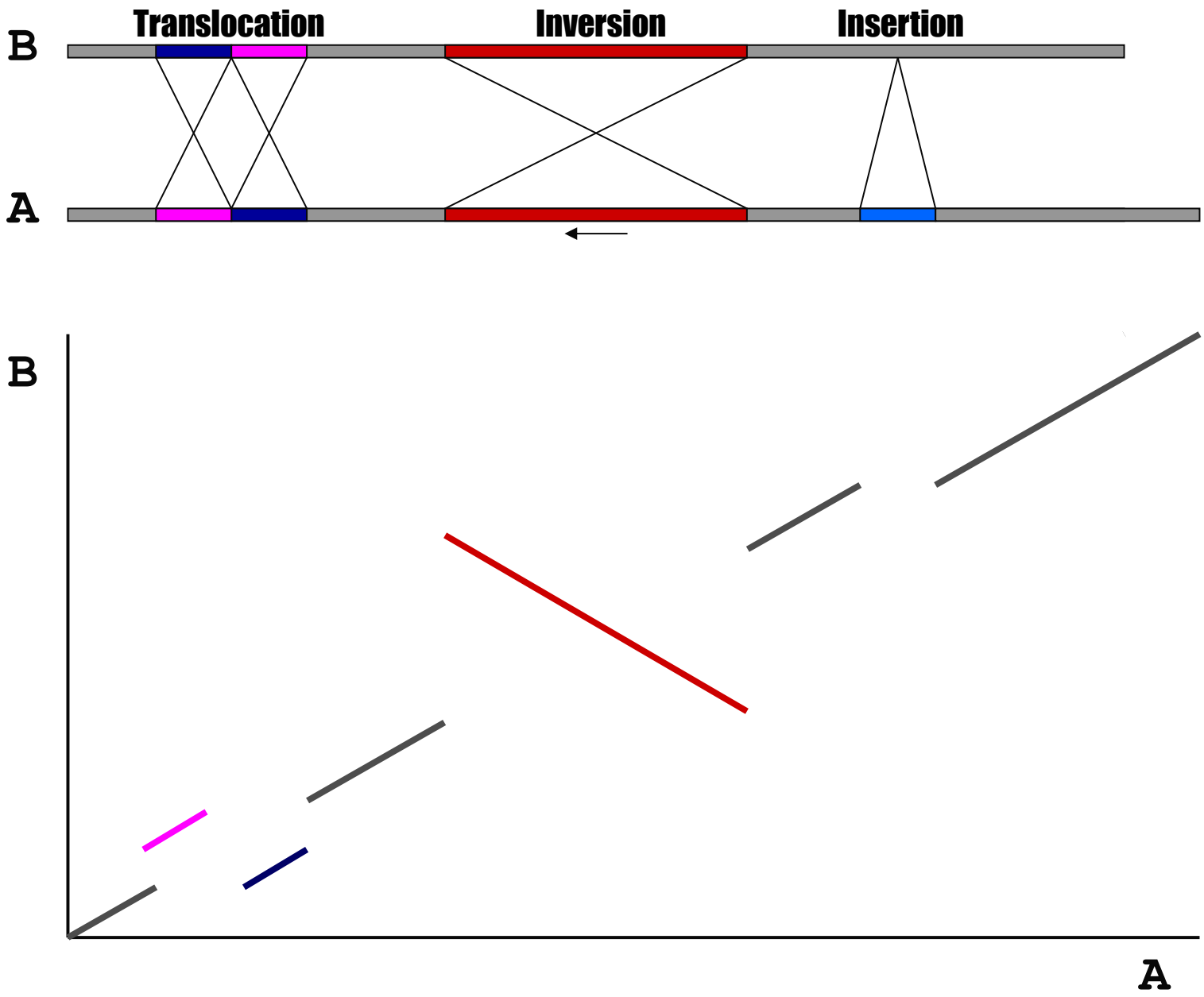
WGA visualization

- How can we visualize *whole* genome alignments?

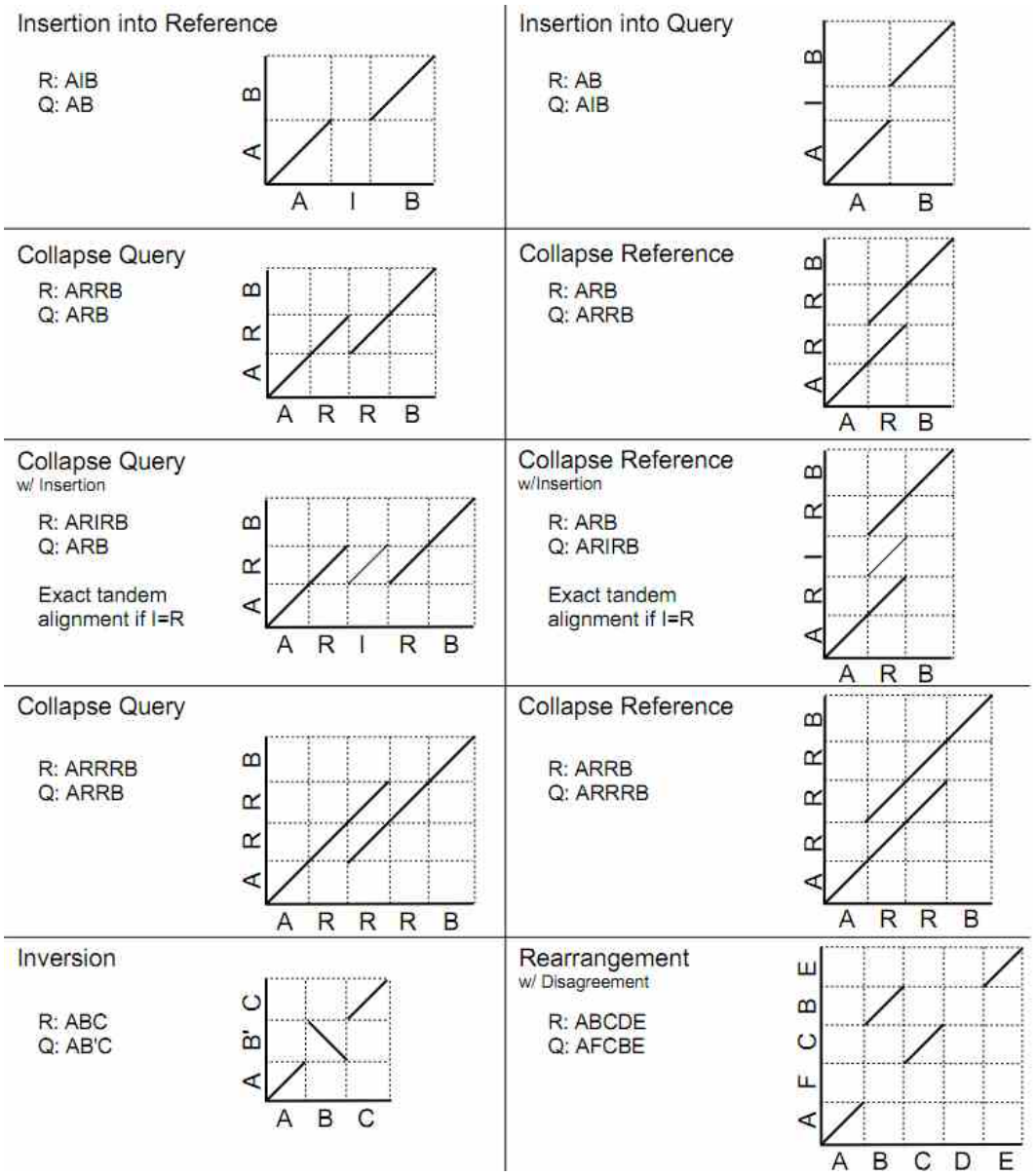
- With an alignment dot plot
  - *N* x *M* matrix
    - Let *i* = position in genome *A*
    - Let *j* = position in genome *B*
    - Fill cell *(i,j)* if A_i shows similarity to B_j



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

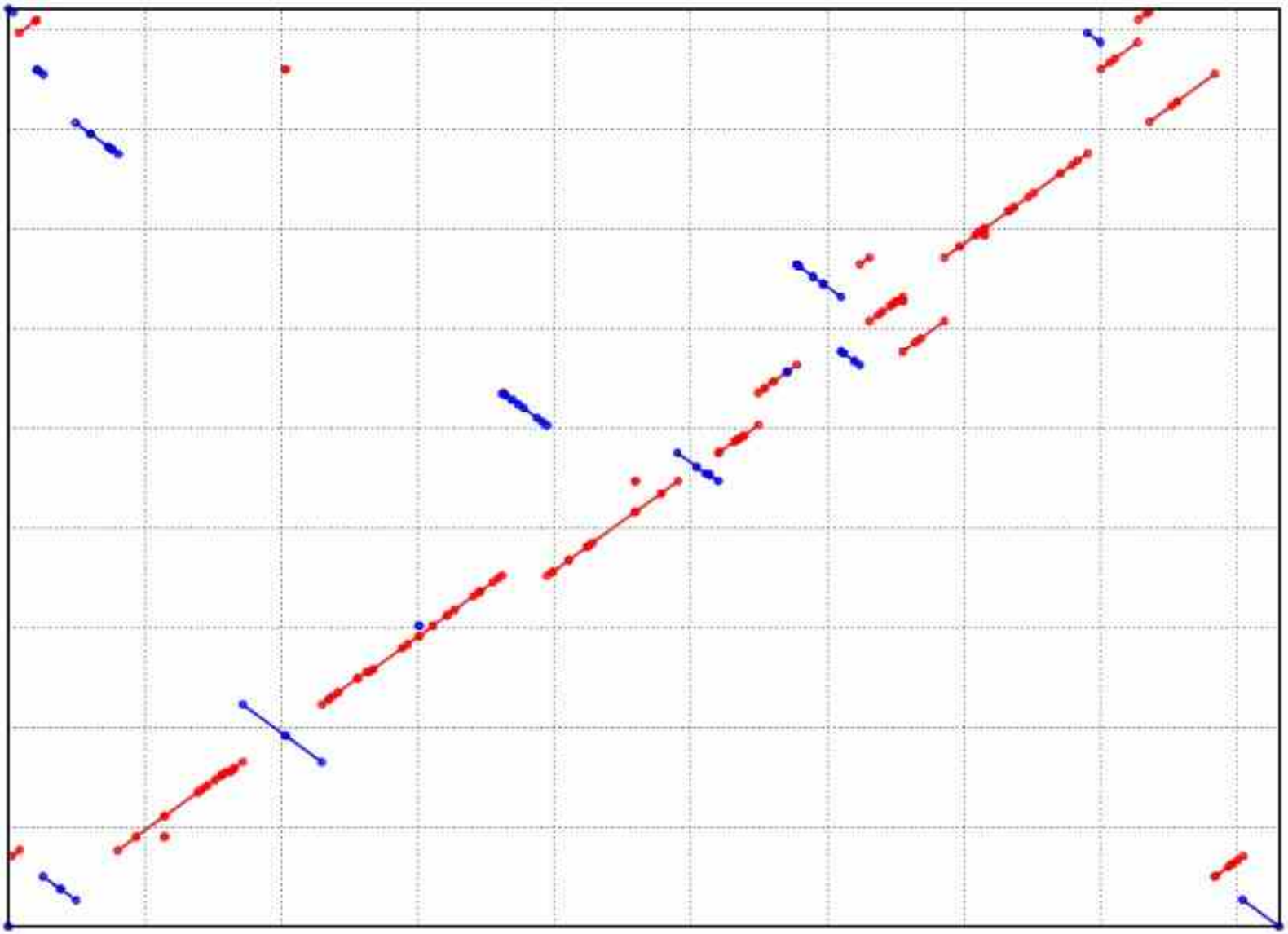

Translocation
Inversion
Insertion
B
A
B
A

SV Types

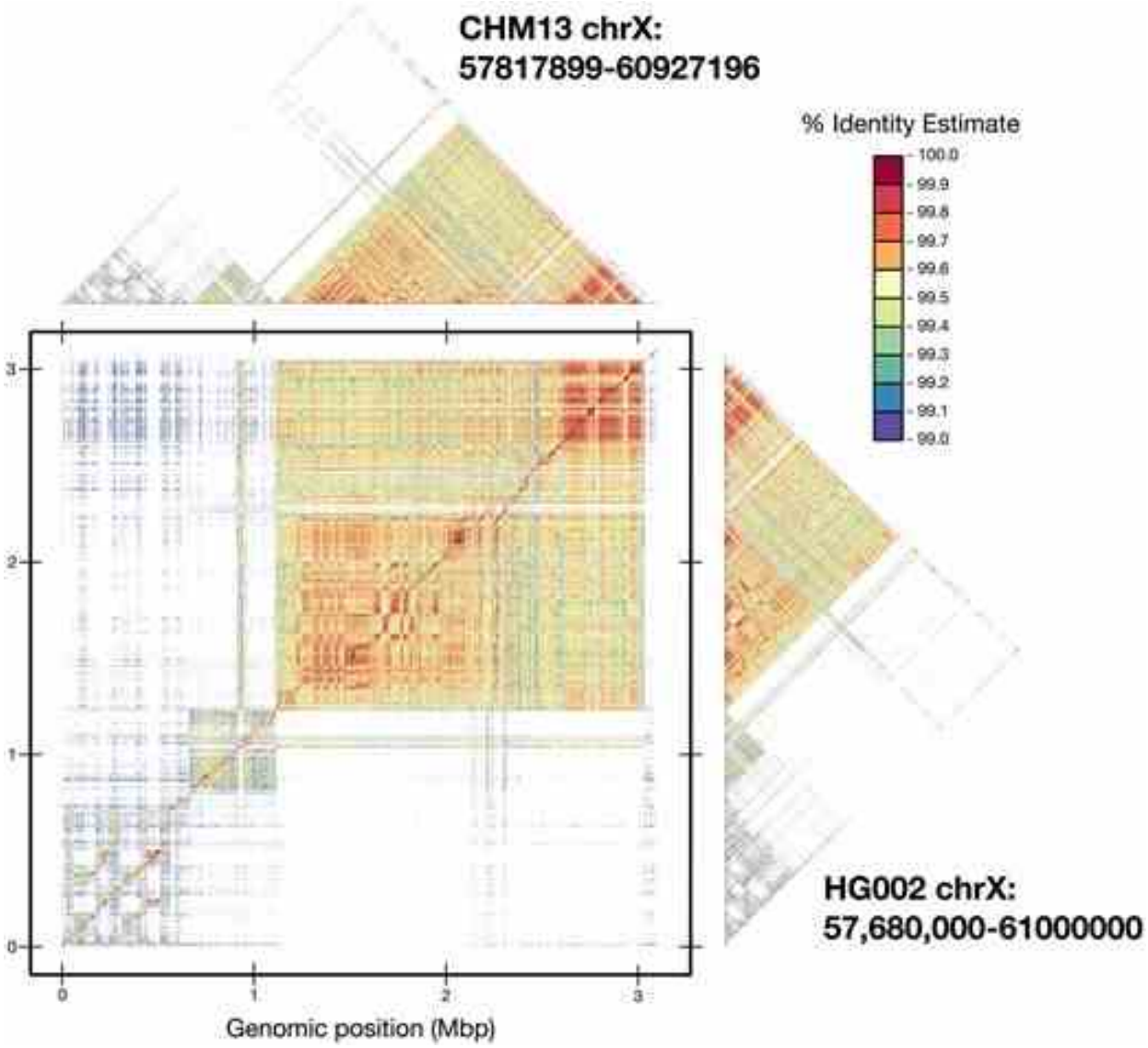


- Different structural variation types / misassemblies will be apparent by their pattern of breakpoints
- Most breakpoints will be at or near repeats
- Things quickly get complicated in real genomes

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