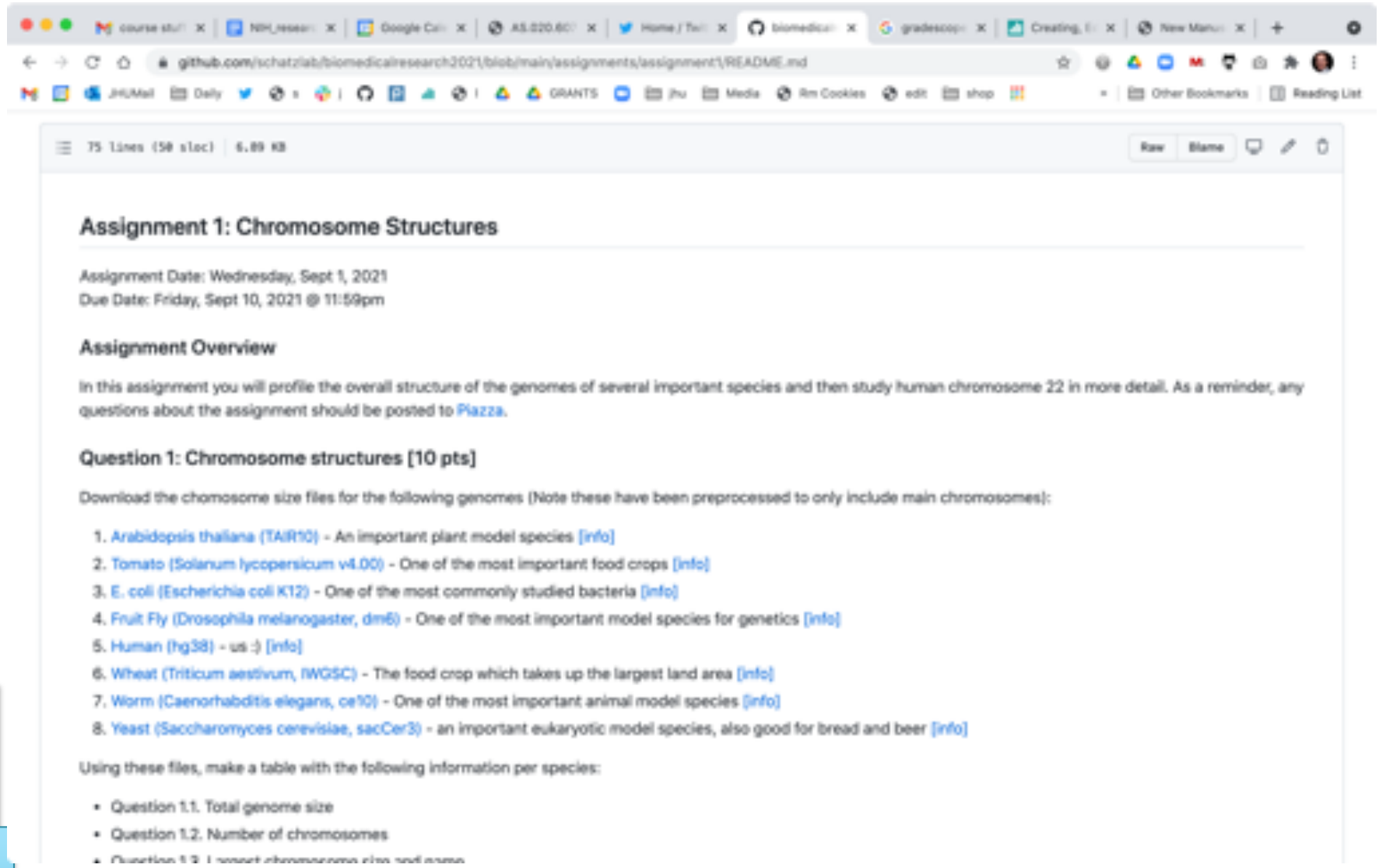
Genome Assembly

Michael Schatz

Sept 8, 2021

Lecture 3: Biomedical Research

Assignment 1: Chromosome Structures
Due Friday Sept 10 @ 11:59pm



github.com/schatzlab/biomedicalresearch2021/blob/main/assignments/assignment1/README.md

75 lines (50 sloc) | 6.89 KB

Raw | Blame

Assignment 1: Chromosome Structures

Assignment Date: Wednesday, Sept 1, 2021
Due Date: Friday, Sept 10, 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 [10 pts]

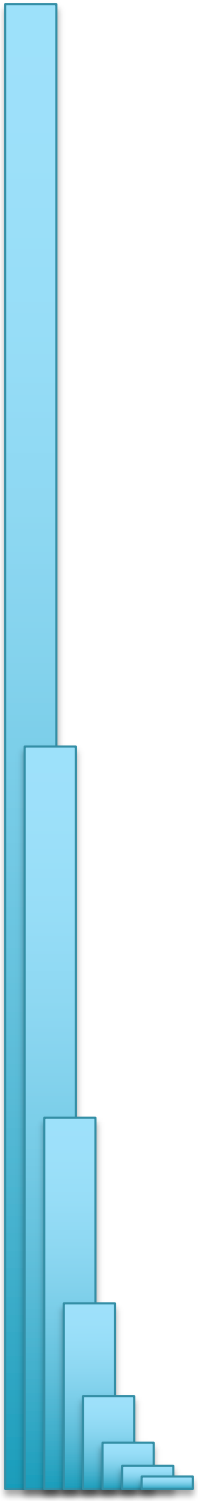
Download the chomosome size files for the following genomes (Note these have been preprocessed to only include main chromosomes):

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

https://github.com/schatzlab/biomedicalresearch2021

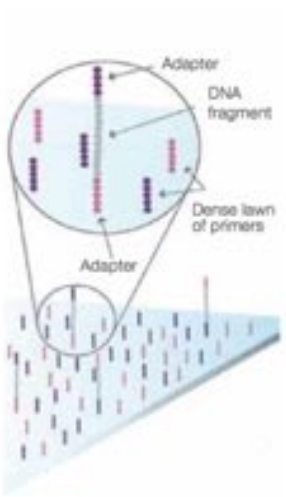


Recap: Sequencing

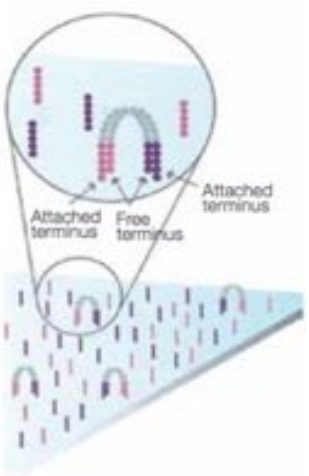
Second Generation Sequencing

Illumina NovaSeq 6000
Sequencing by Synthesis

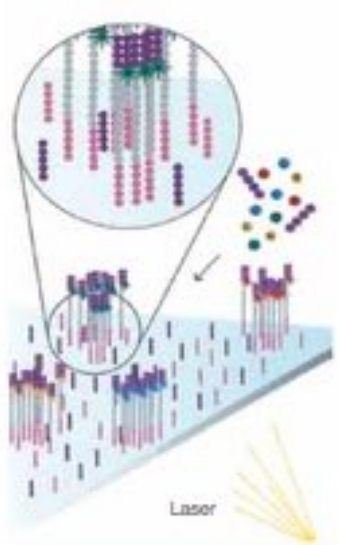
>3Tbp / day



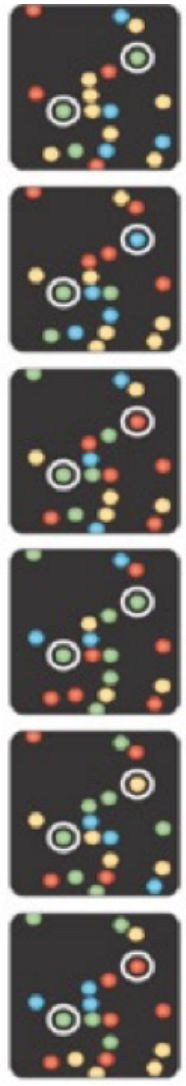
1. Attach



2. Amplify

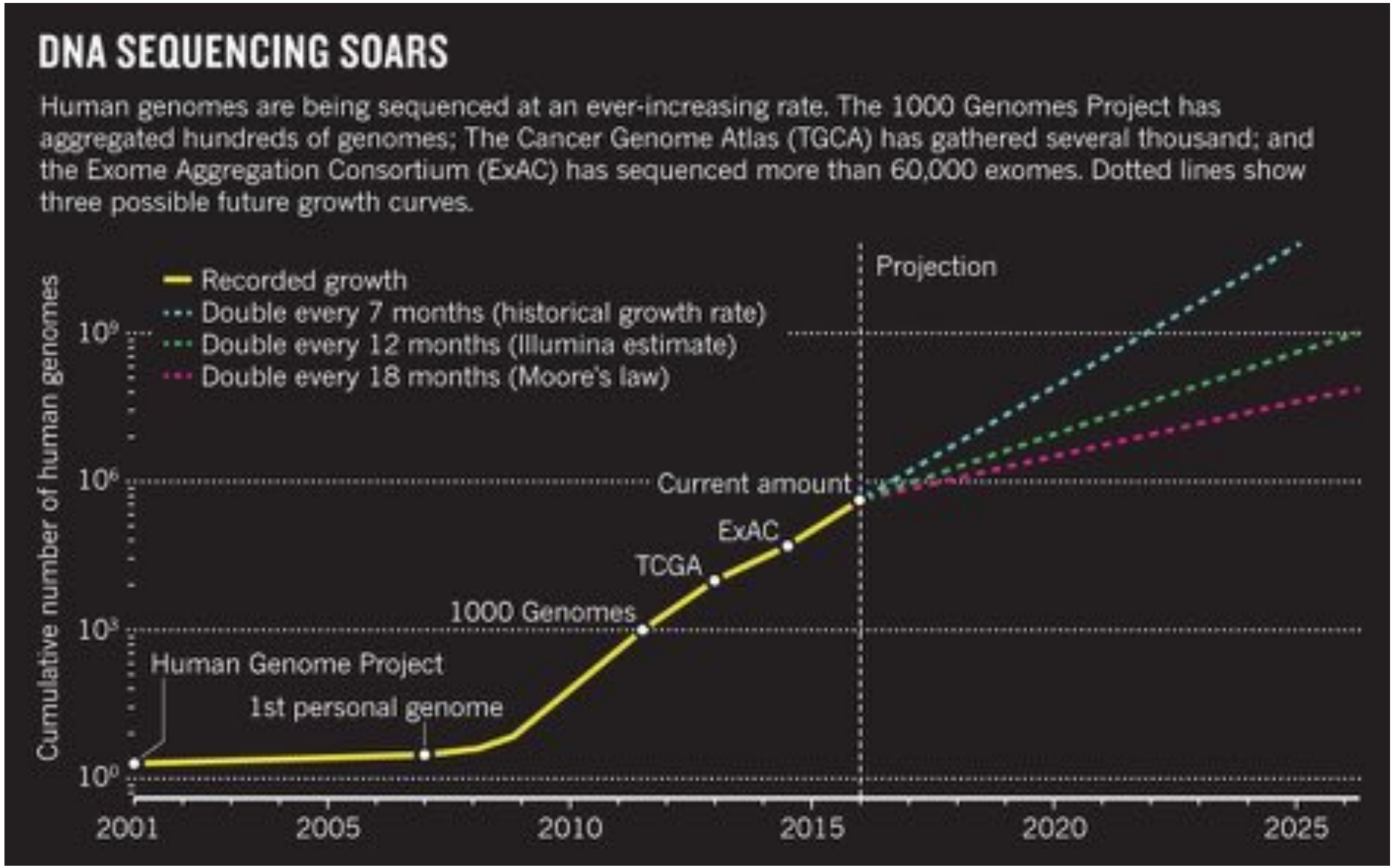


3. Image



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

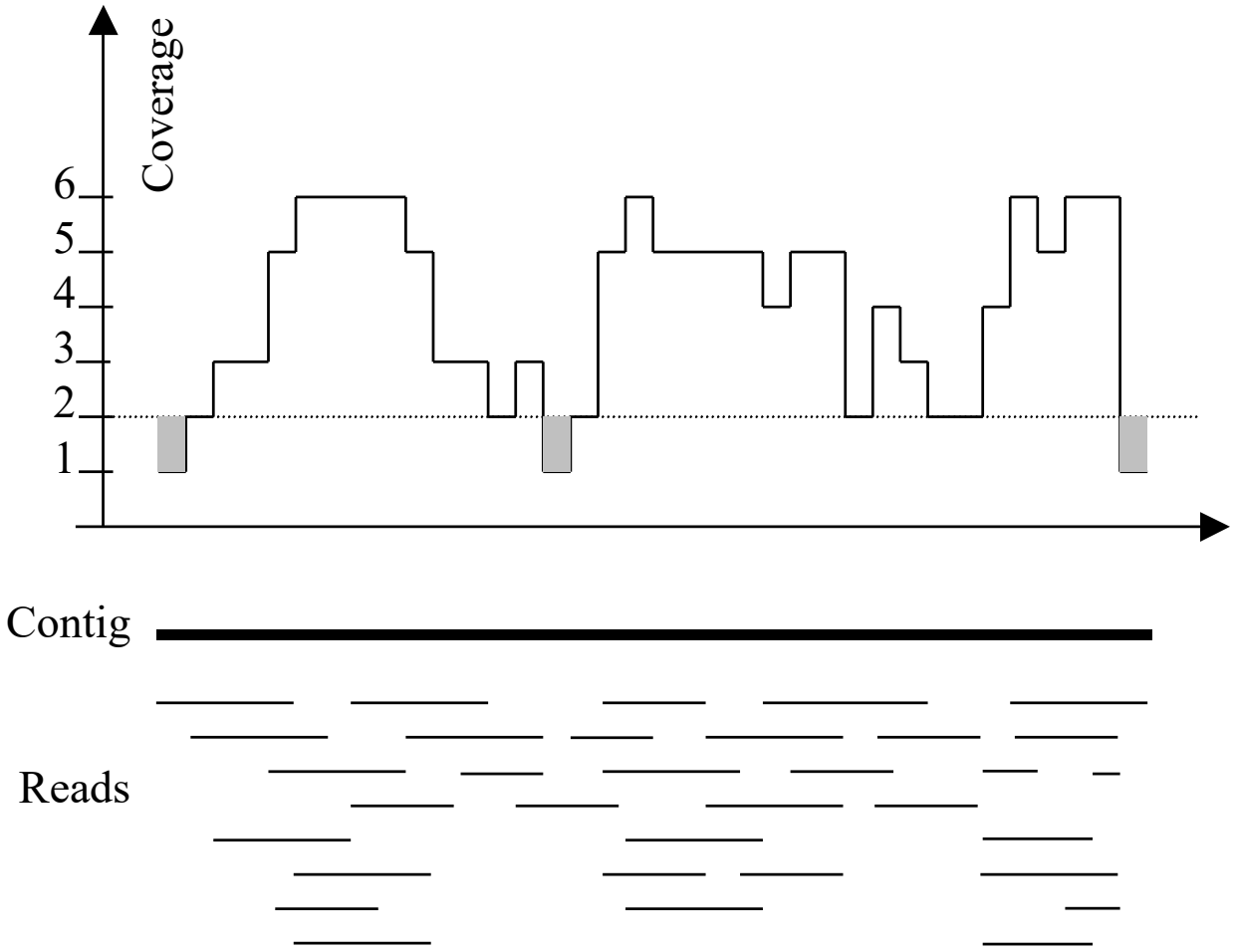
Sequencing Capacity



Big Data: Astronomical or Genomical?

Stephens, Z, et al. (2015) PLOS Biology DOI: 10.1371/journal.pbio.1002195

Typical sequencing coverage

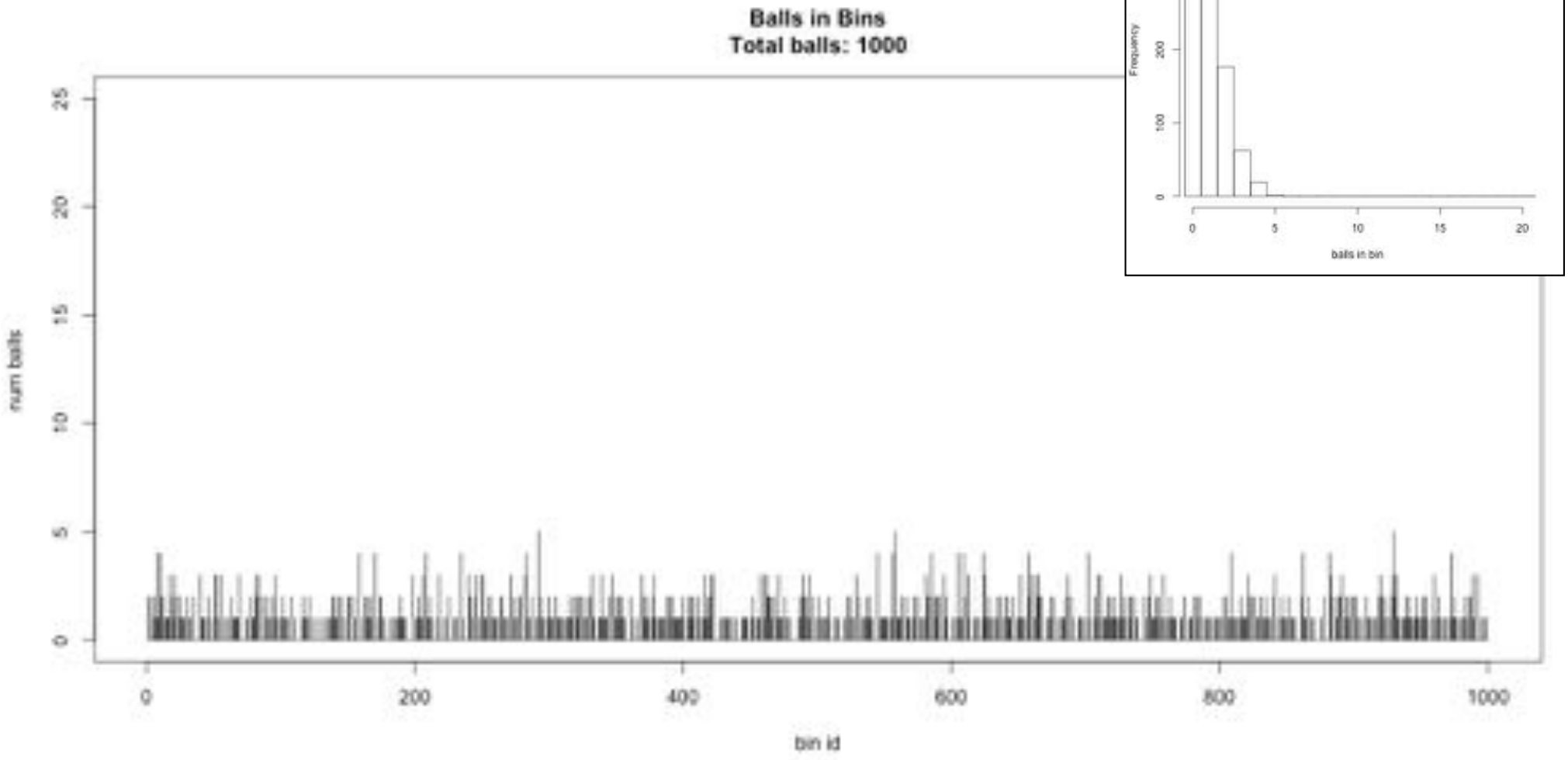


Imagine raindrops on a sidewalk

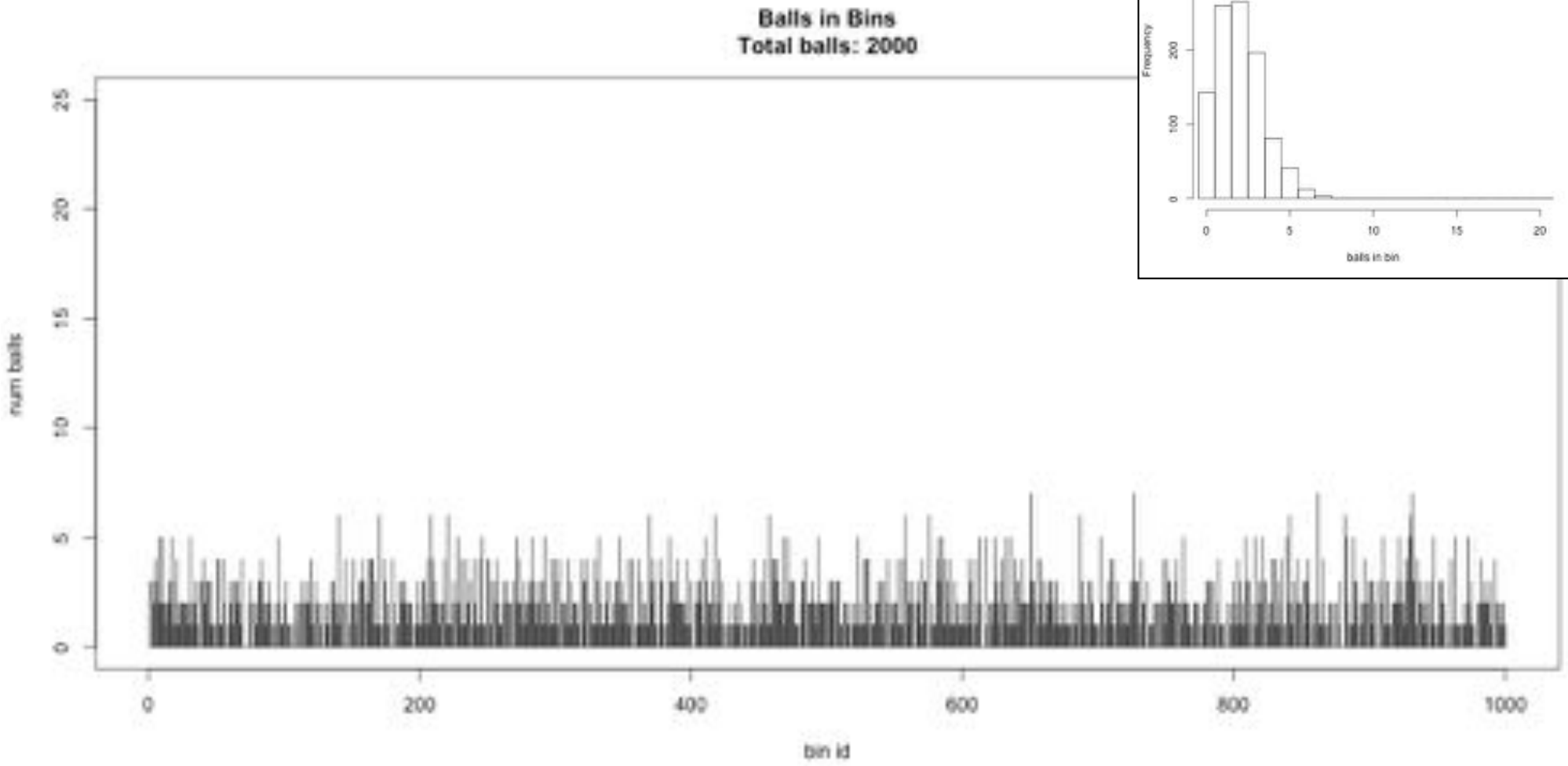
We want to cover the entire sidewalk but each drop costs $1

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

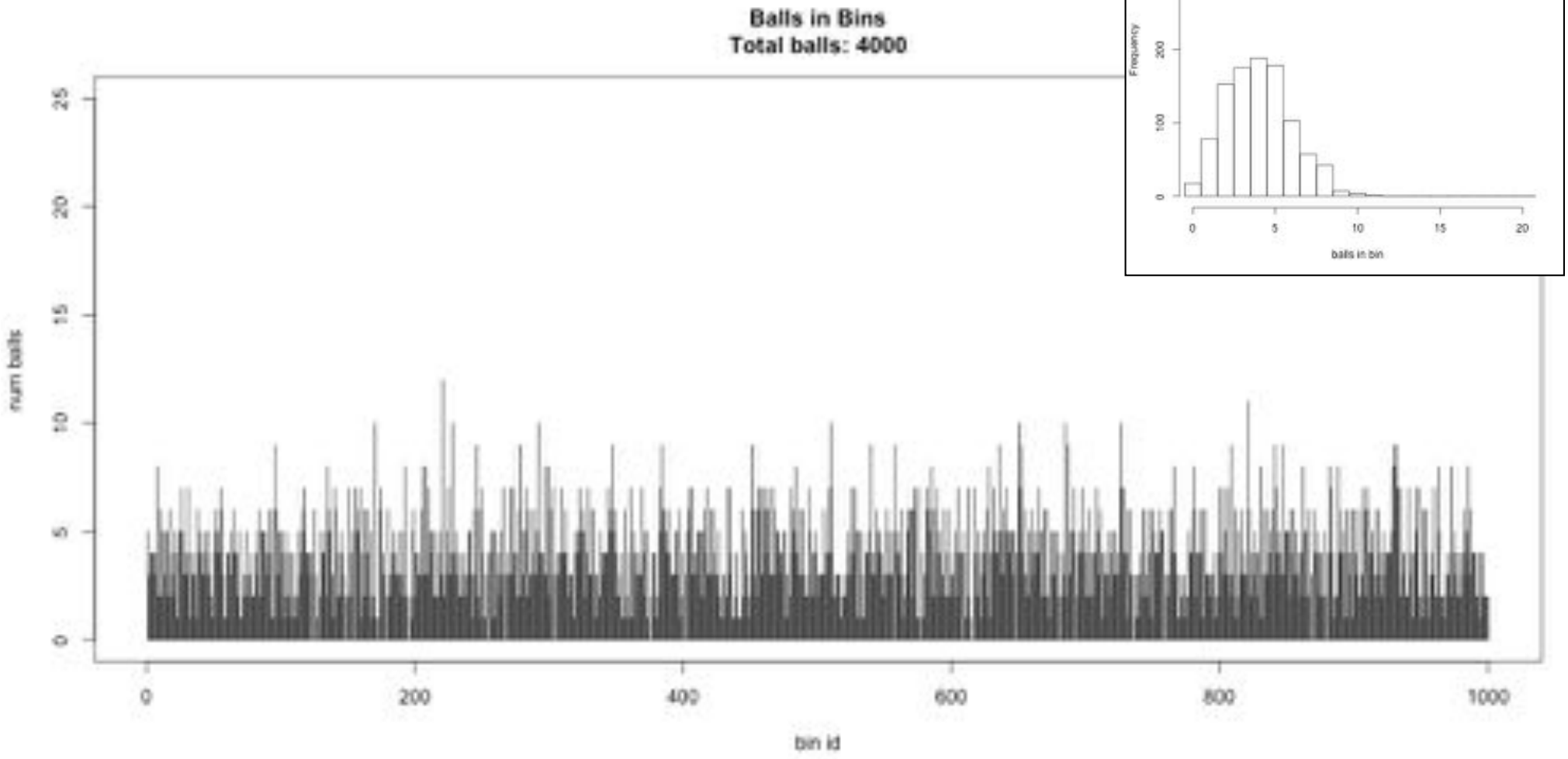
1x sequencing



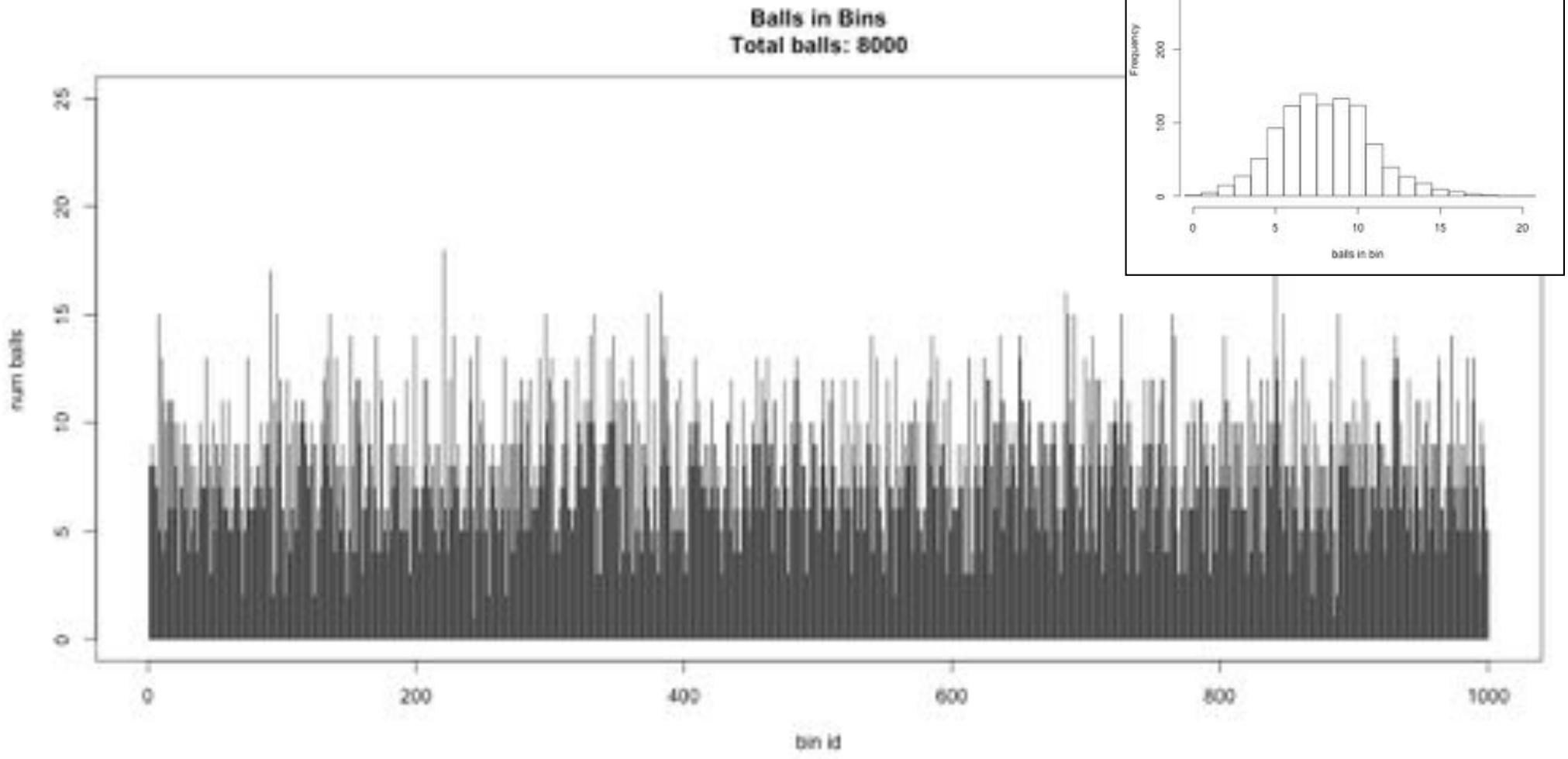
2x sequencing



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.

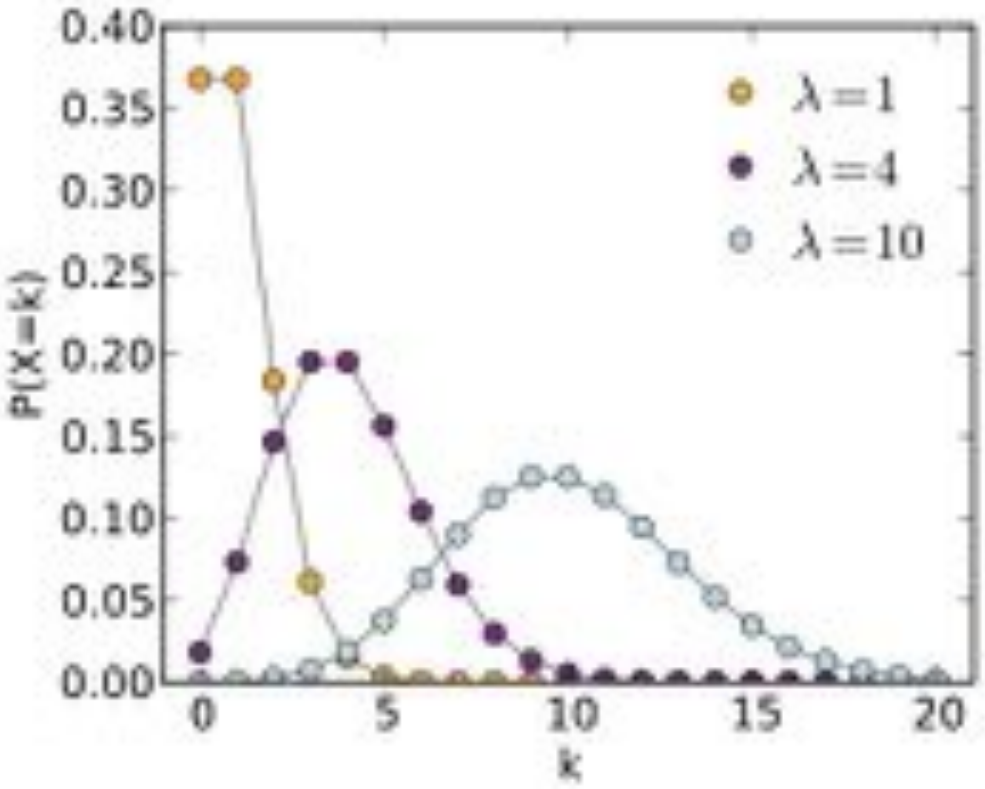
$$P(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Formulation comes from the limit of the binomial equation

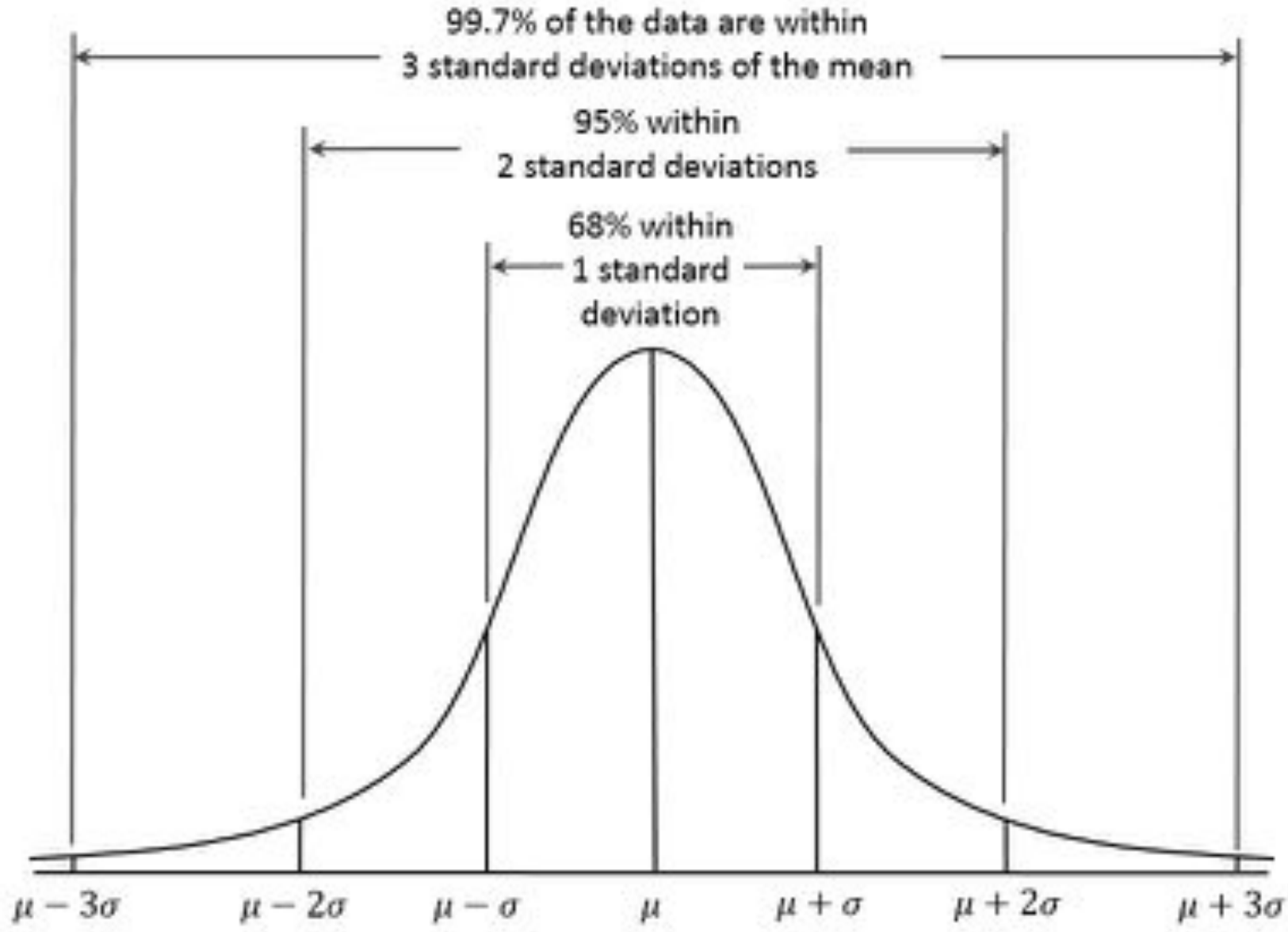
Resembles a normal distribution, but over the positive values, and with only a single parameter.

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?

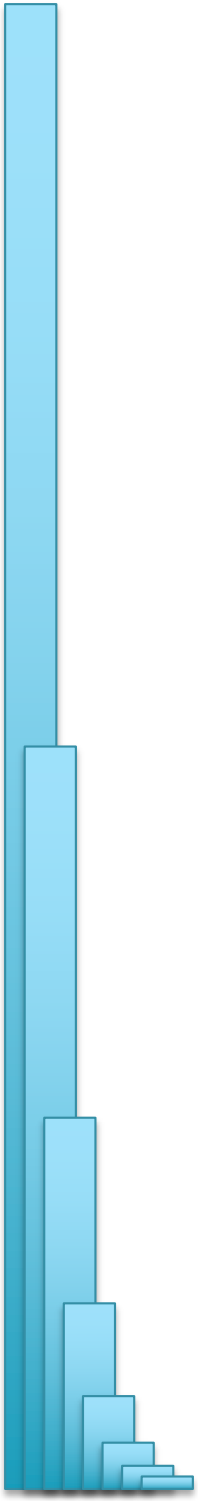
I need 10Mbp x 24x = 240Mbp of data
240Mbp / 120bp / read = 2M reads

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



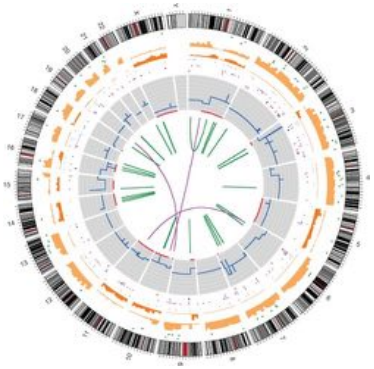
Genome Assembly

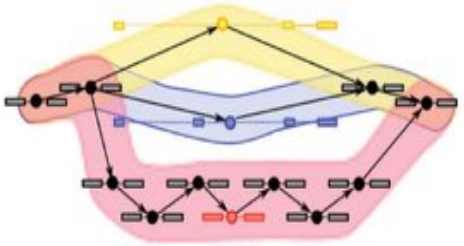
Assembly Applications

- Novel genomes

- Metagenomes

- Sequencing assays
  - Structural variations
  - Transcript assembly
  - …





Why are genomes hard to assemble?

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

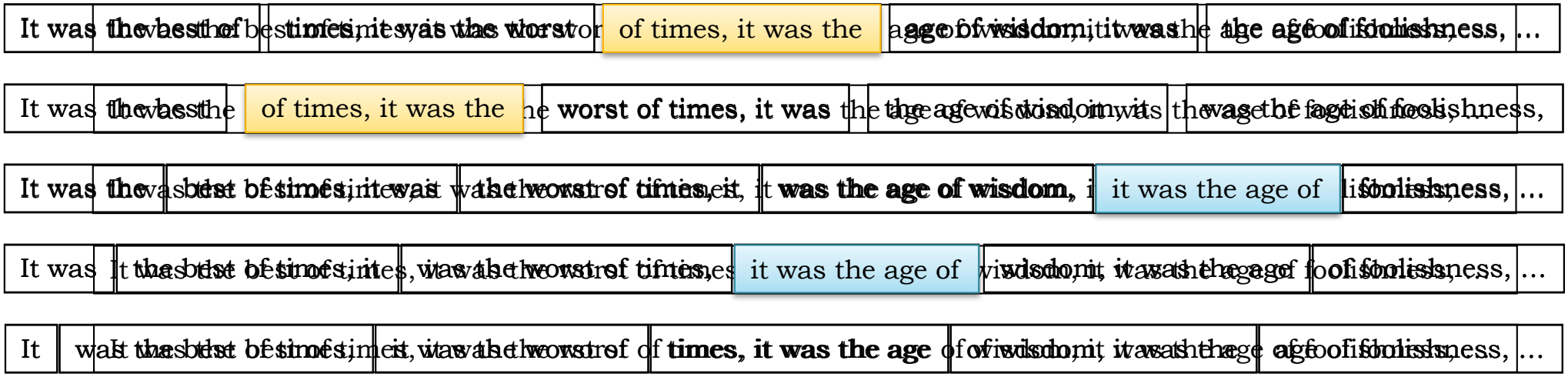
2. ***Sequencing***:
   - (Very) large genomes, imperfect sequencing

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

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

Shredded Book Reconstruction

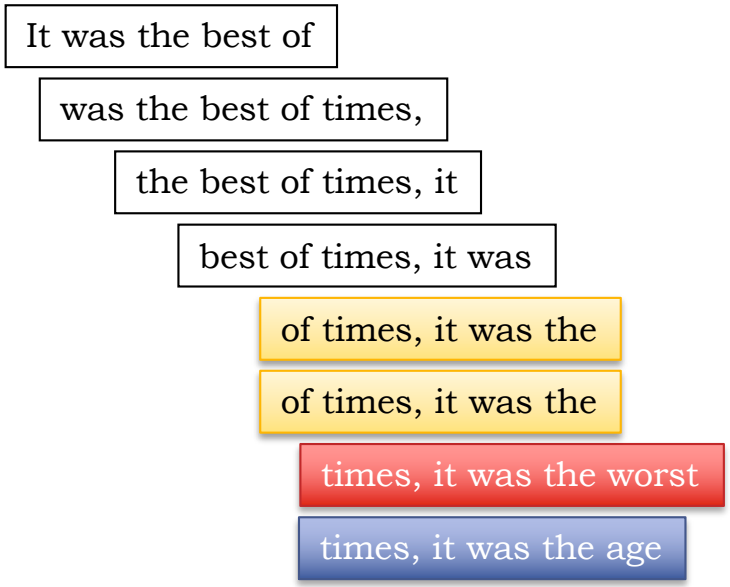
- Dickens accidentally shreds the first printing of A Tale of Two Cities
  - Text printed on 5 long spools



- 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



The repeated sequence make the correct reconstruction ambiguous

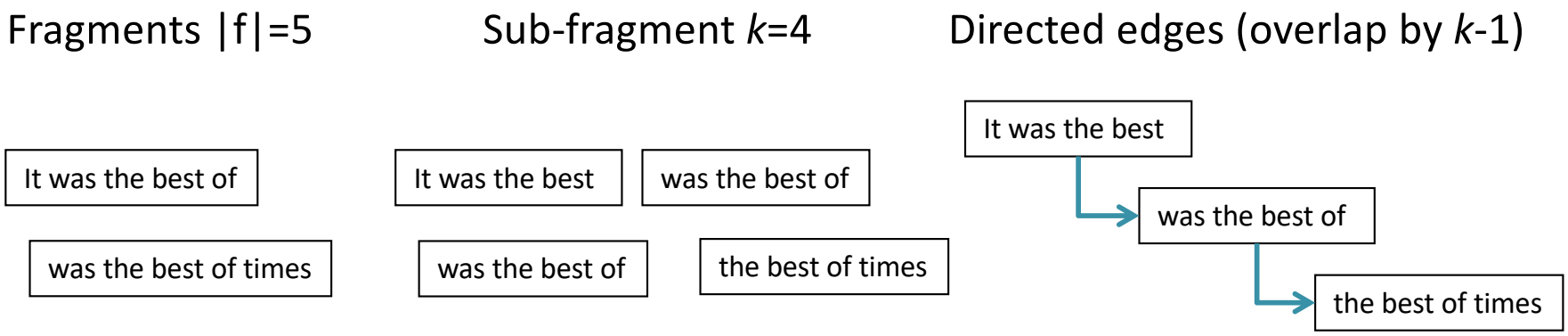
- It was the best of times, it was the [worst/age]

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

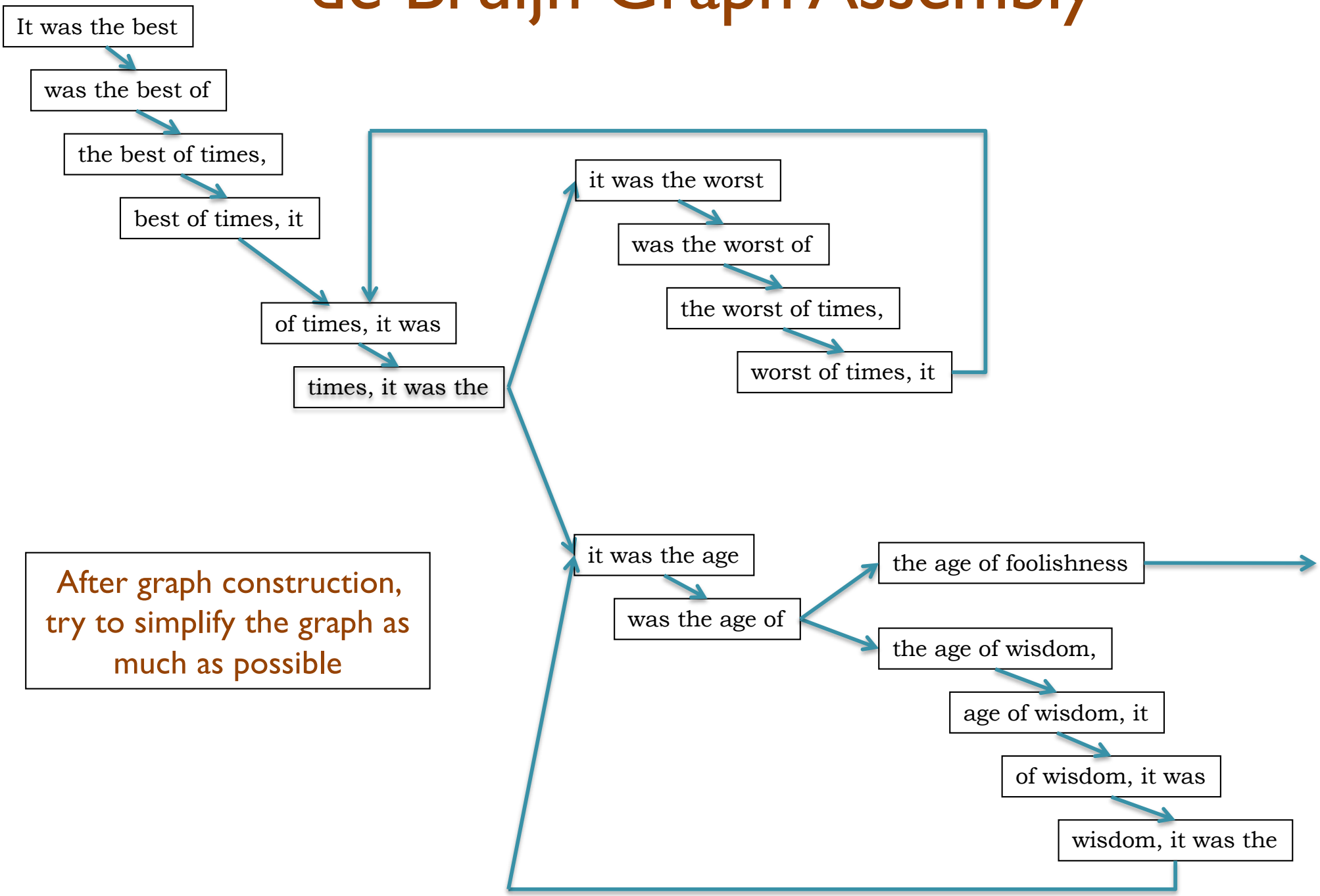


– Overlaps between fragments are implicitly computed

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

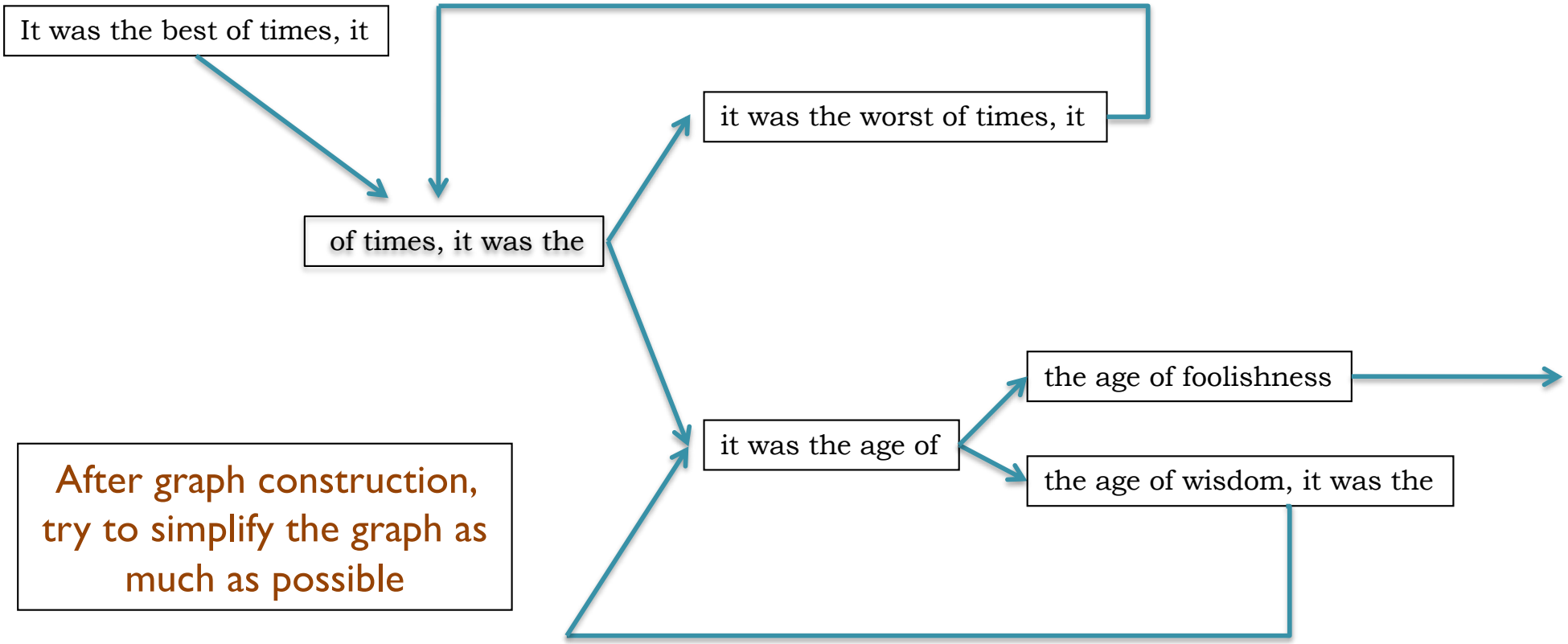
de Bruijn, 1946
Idury et al., 1995
Pevzner et al., 2001

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

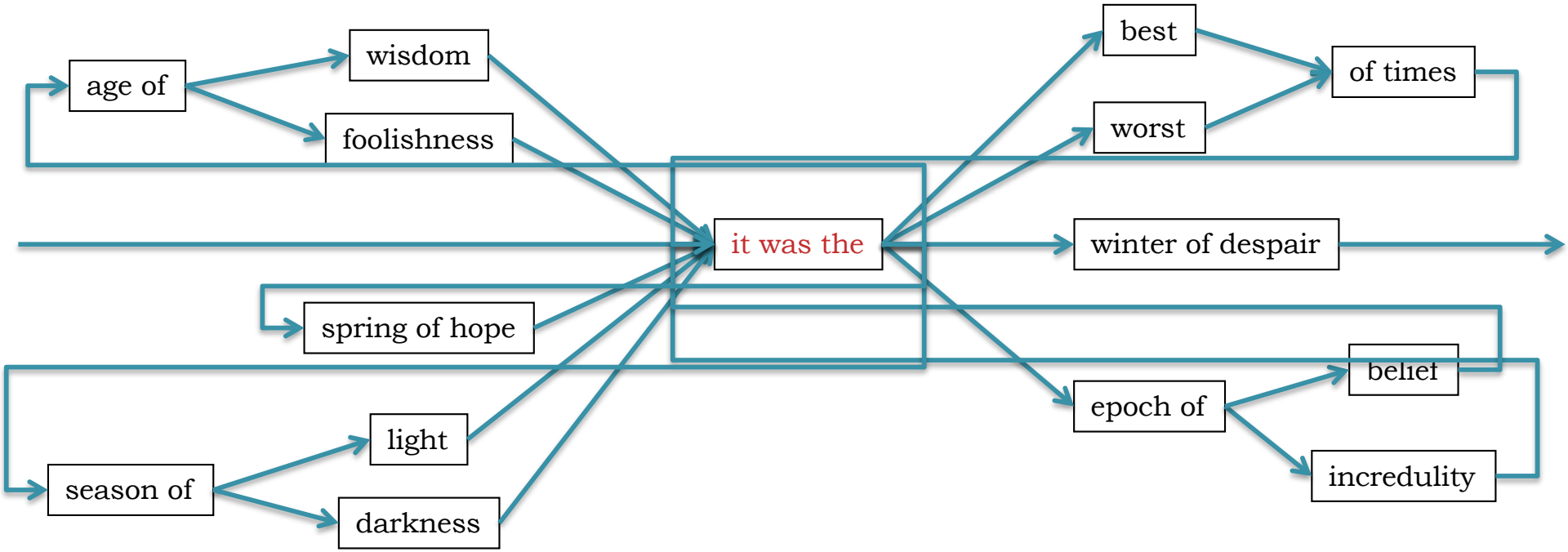
… it was the best of times it was the worst of times …

… 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 …

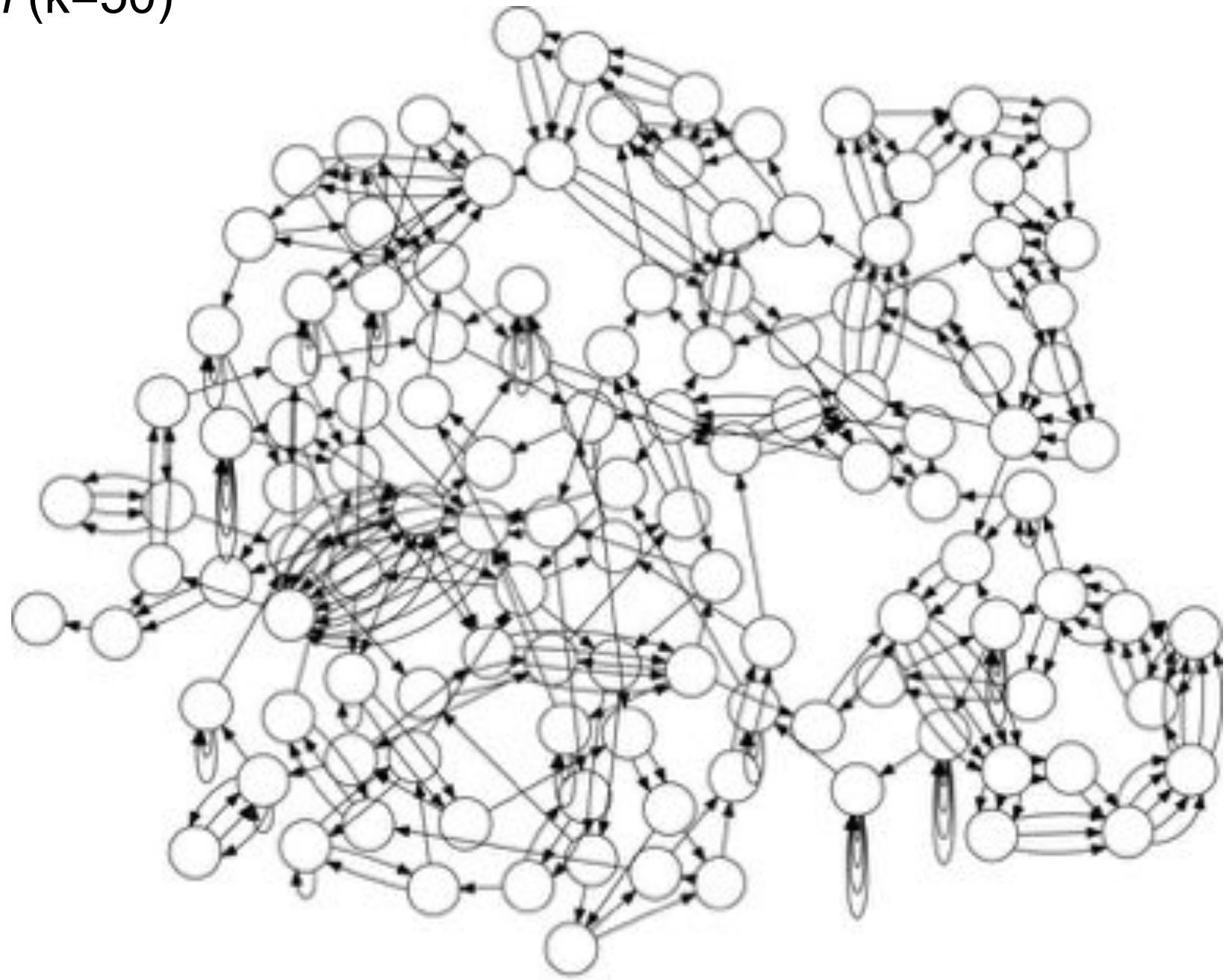


Repetitive regions

Repeat Type	Definition / Example	Prevalence
Low-complexity DNA / Microsatellites	$(b_1b_2\ldots b_k)^N$ 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%

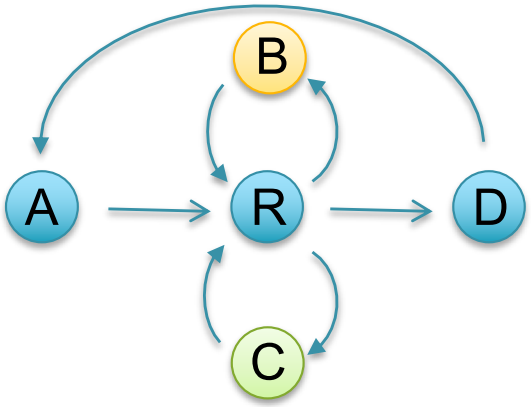
- Over 50% of mammalian genomes are repetitive
  - Large plant genomes tend to be even worse
  - Wheat: 16 Gbp; Pine: 24 Gbp

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)

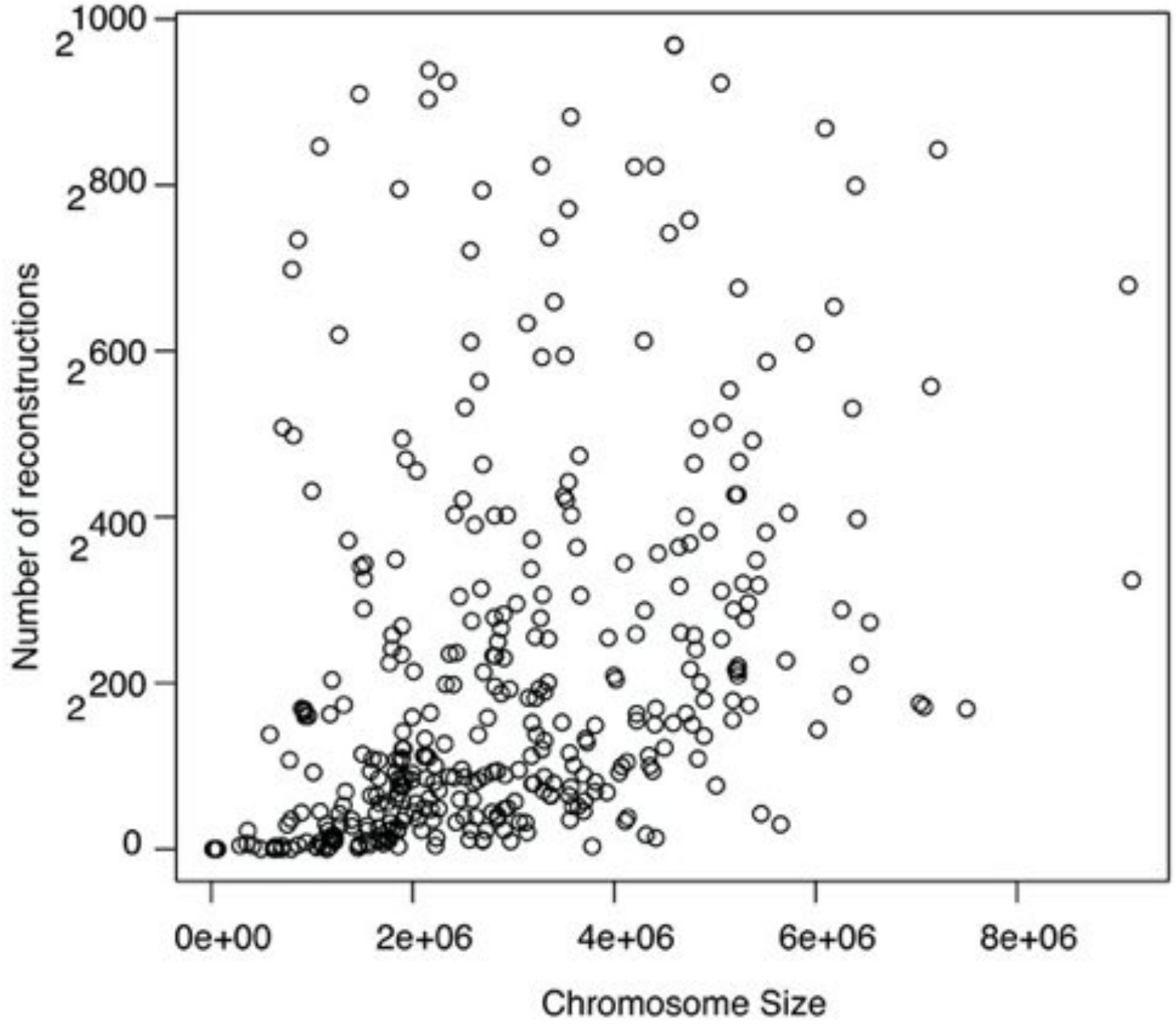
$$\mathcal{W}(G,t) = (\det L)\left\{\prod_{u\in V}(r_u-1)!\right\}\left\{\prod_{(u,v)\in E} a_{uv}!\right\}^{-1}$$

L = *n x n* matrix with r_u-a_{uu} along the diagonal and -a_{uv} in entry uv

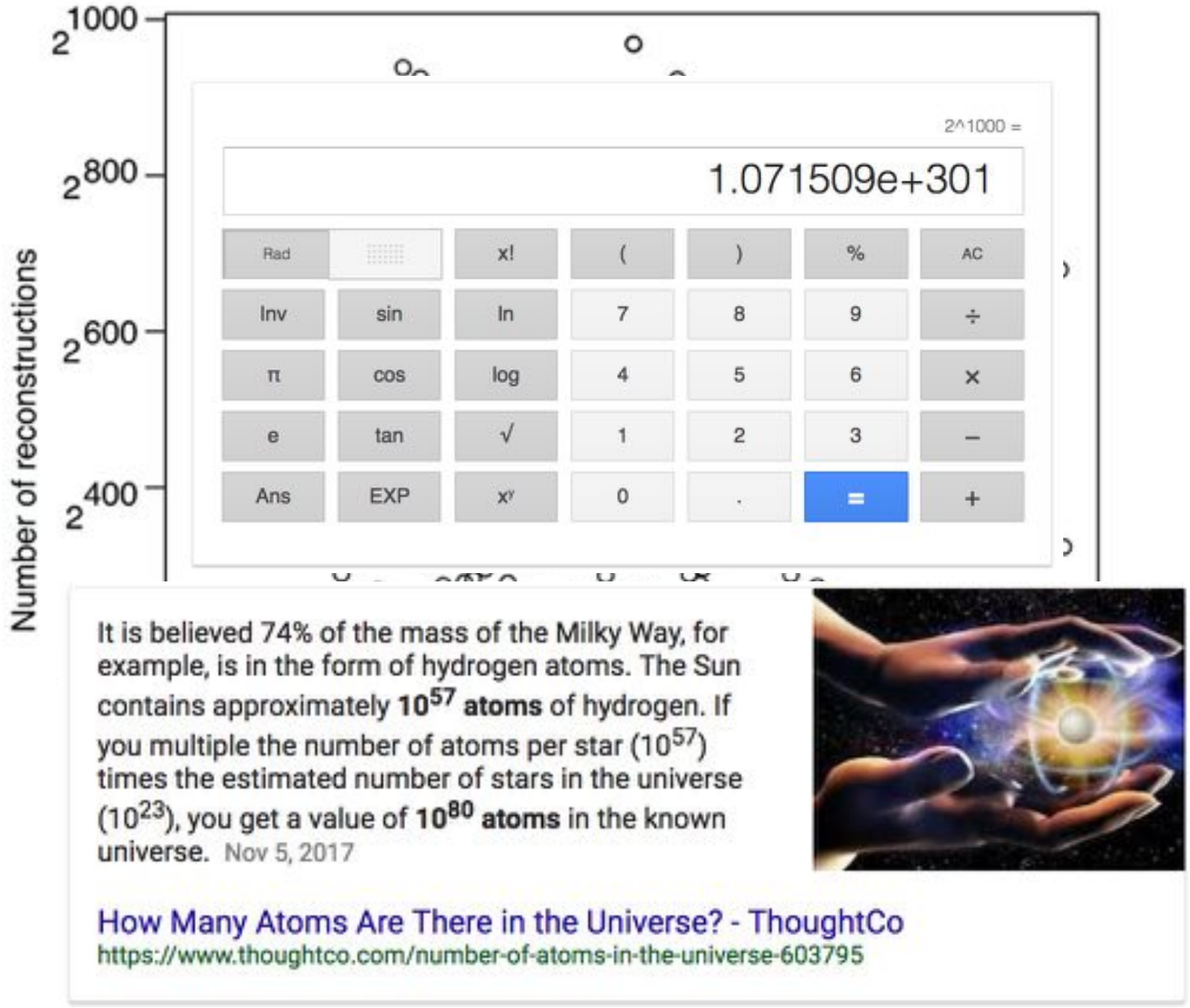
$r_u = d^+(u)+1$ if $u=t$, or $d^+(u)$ otherwise

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

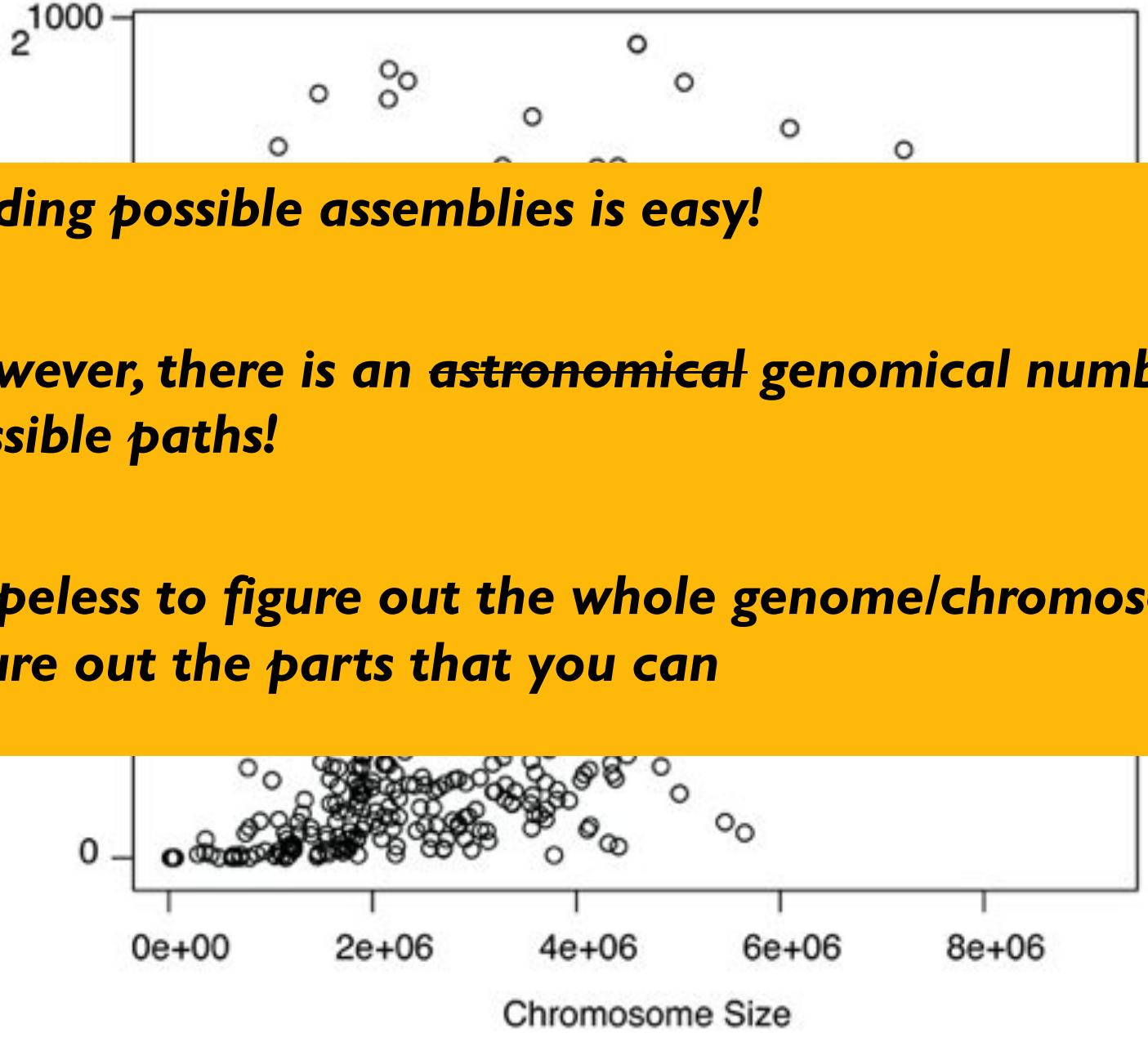
Assembly Complexity of Prokaryotic Genomes using Short Reads.
Kingsford C, Schatz MC, Pop M (2010) *BMC Bioinformatics*.



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



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~~ genomical 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*.

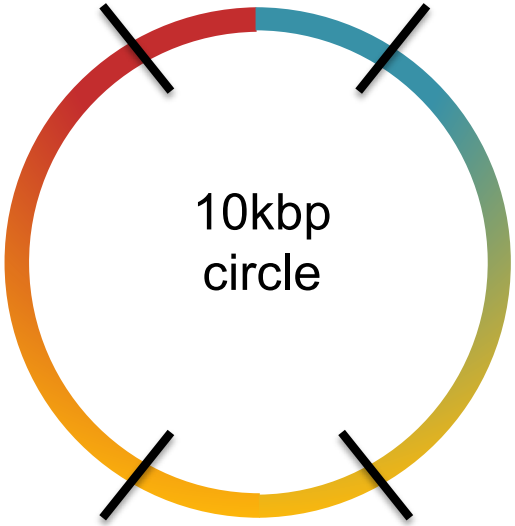
Paired-end and Mate-pairs

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

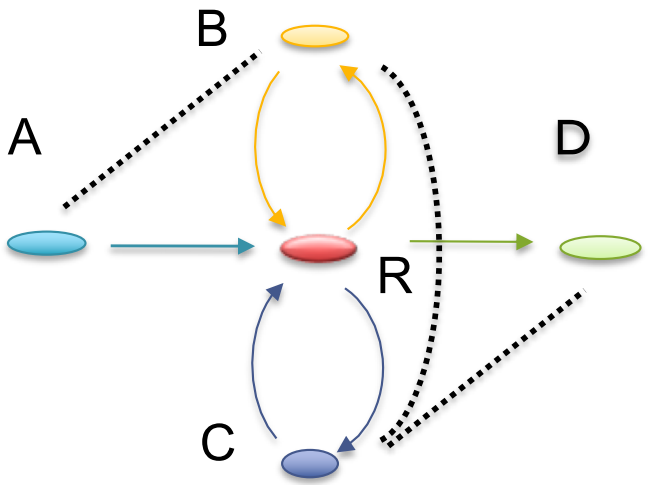
Mate-pair sequencing

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

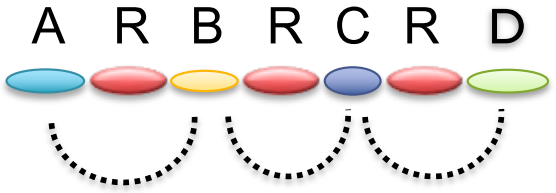


Scaffolding

- 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



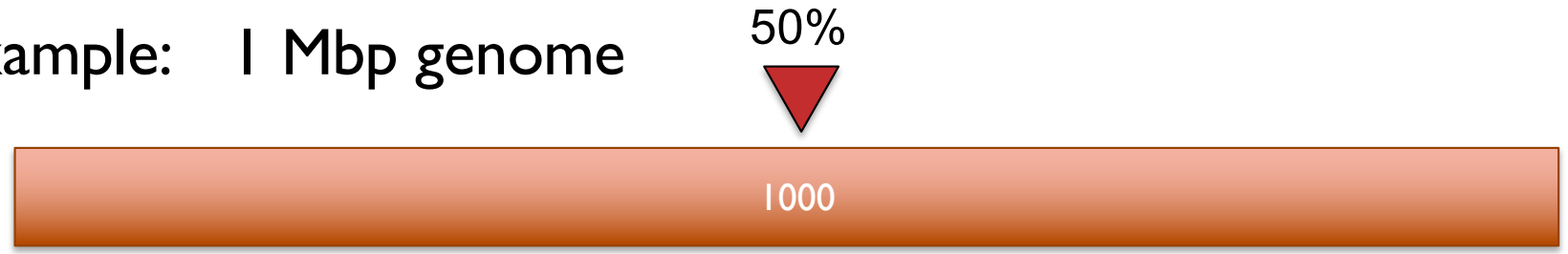
- Final scaffold may have internal gaps called sequencing gaps
  - We know the order, orientation, and spacing, but just not the bases. Fill with Ns instead

Why do scaffolds end?

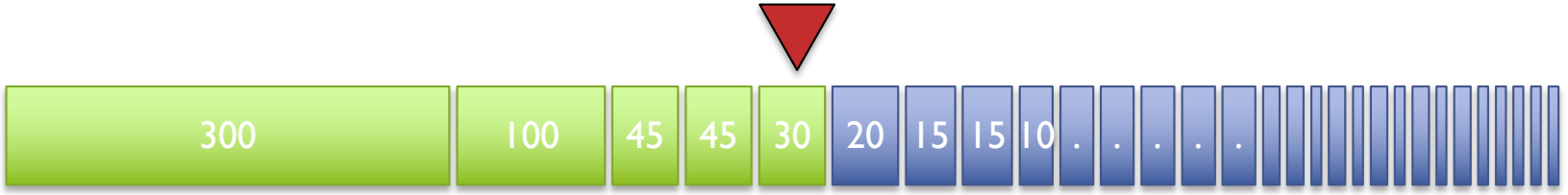
Contig/Scaffold N50

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

Example: 1 Mbp genome

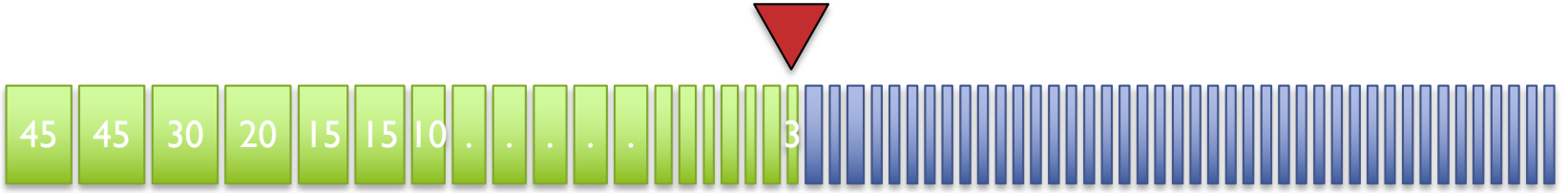


A



N50 size = 30 kbp

B



N50 size = 3 kbp

Contig/Scaffold N50

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

Better N50s improves the analysis in every dimension

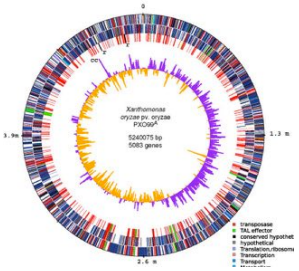
- Better resolution of genes and flanking regulatory regions
- Better resolution of transposons and other complex sequences
- Better resolution of chromosome organization
- Better sequence for all downstream analysis

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

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

Pop Quiz 1

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

```
ATTA
GATT
TACA
TTAC
```

Pop Quiz 1

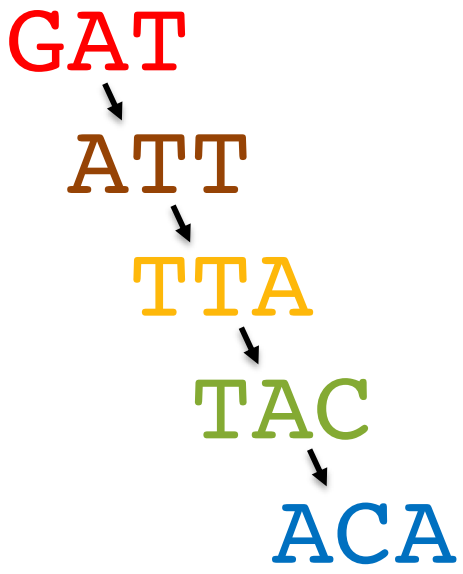
Assemble these reads using a de Bruijn graph approach (k=3):

```
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

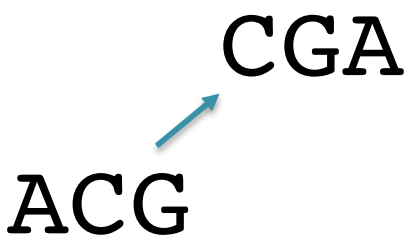
CGAC

CGTA

GACG

GTAT

TACG



Pop Quiz 2

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

~~ACGA~~

~~ACGT~~

ATAC

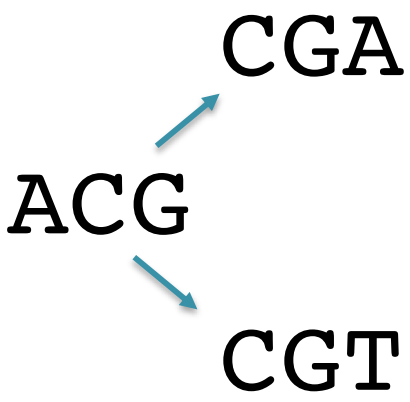
CGAC

CGTA

GACG

GTAT

TACG



Pop Quiz 2

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

~~ACGA~~

~~ACGT~~

ATAC

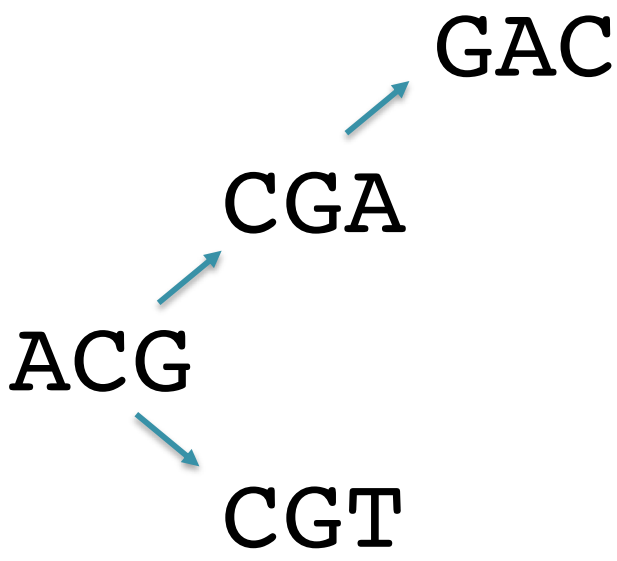
~~CGAC~~

CGTA

GACG

GTAT

TACG



Pop Quiz 2

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

~~ACGA~~

~~ACGT~~

ATAC

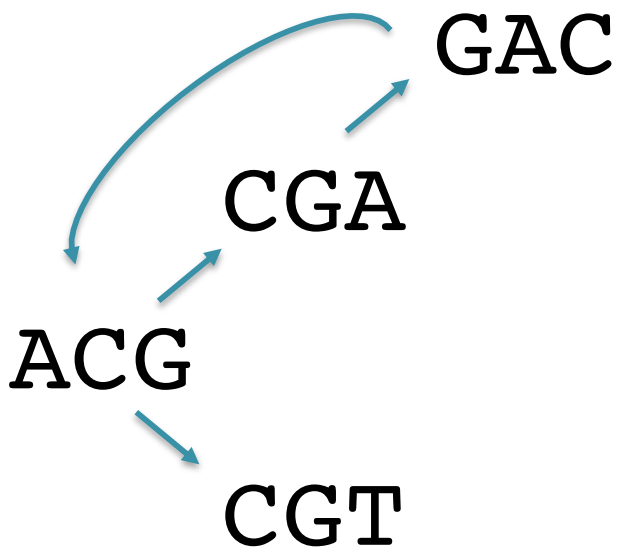
~~CGAC~~

CGTA

~~GACG~~

GTAT

TACG



Pop Quiz 2

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

~~ACGA~~

~~ACGT~~

ATAC

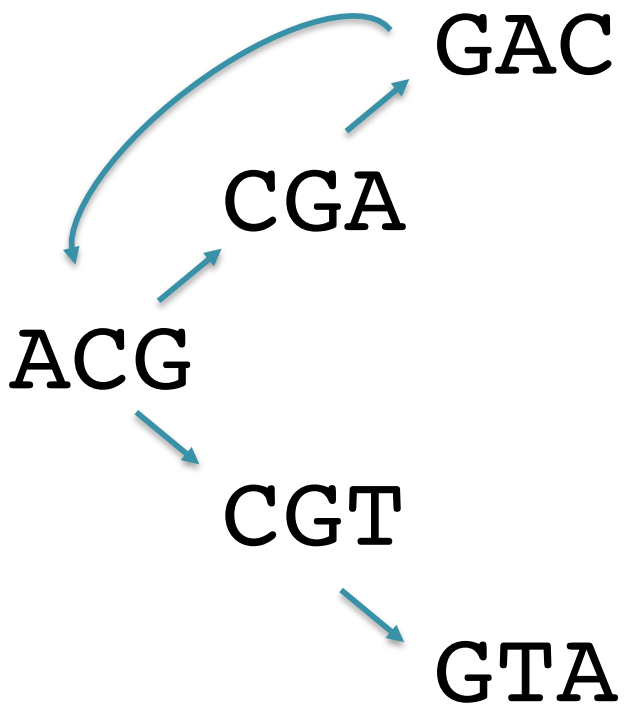
~~CGAC~~

~~CGTA~~

~~GACG~~

GTAT

TACG



Pop Quiz 2

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

~~ACGA~~

~~ACGT~~

ATAC

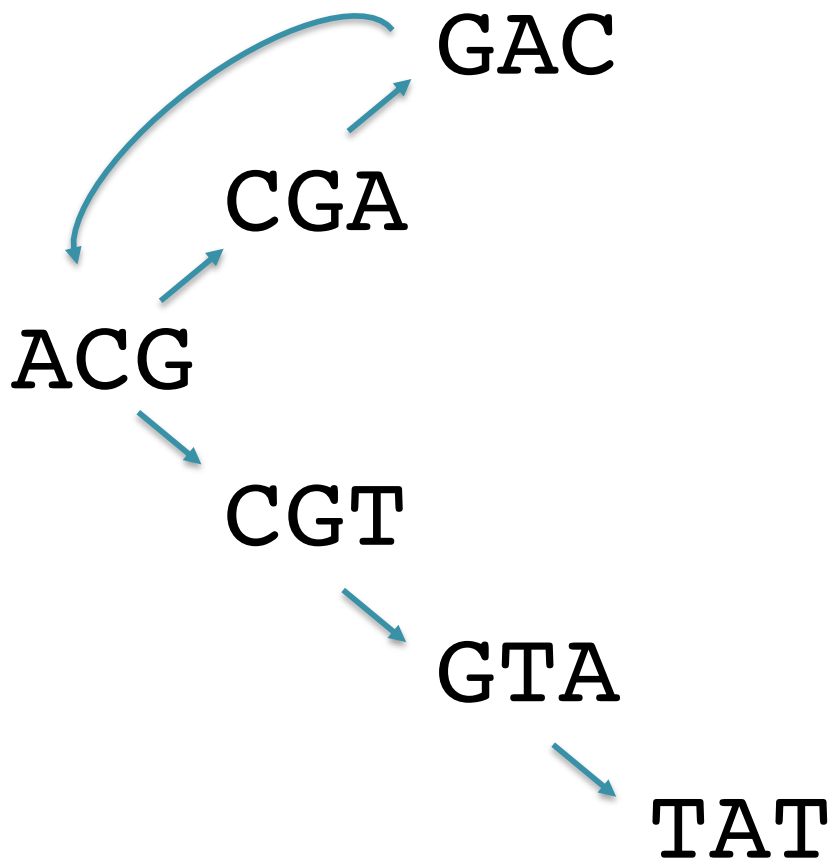
~~CGAC~~

~~CGTA~~

~~GACG~~

~~GTAT~~

TACG



Pop Quiz 2

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

~~ACGA~~

~~ACGT~~

ATAC

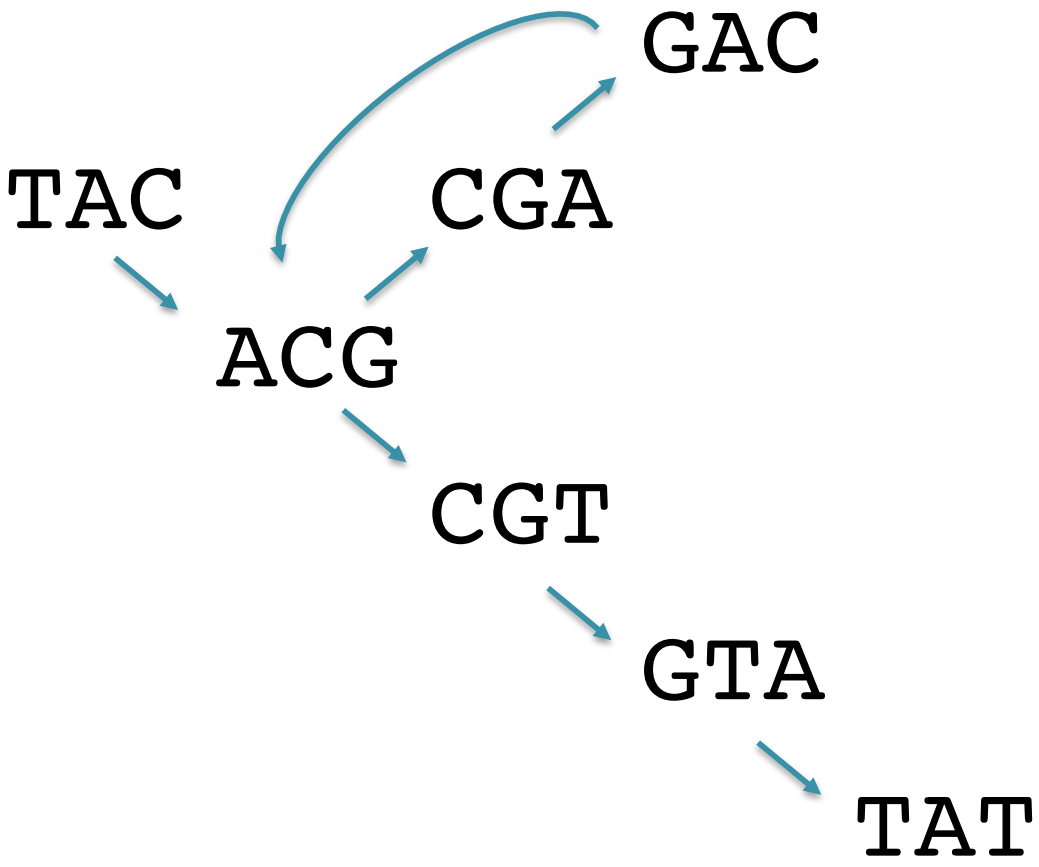
~~CGAC~~

~~CGTA~~

~~GACG~~

~~GTAT~~

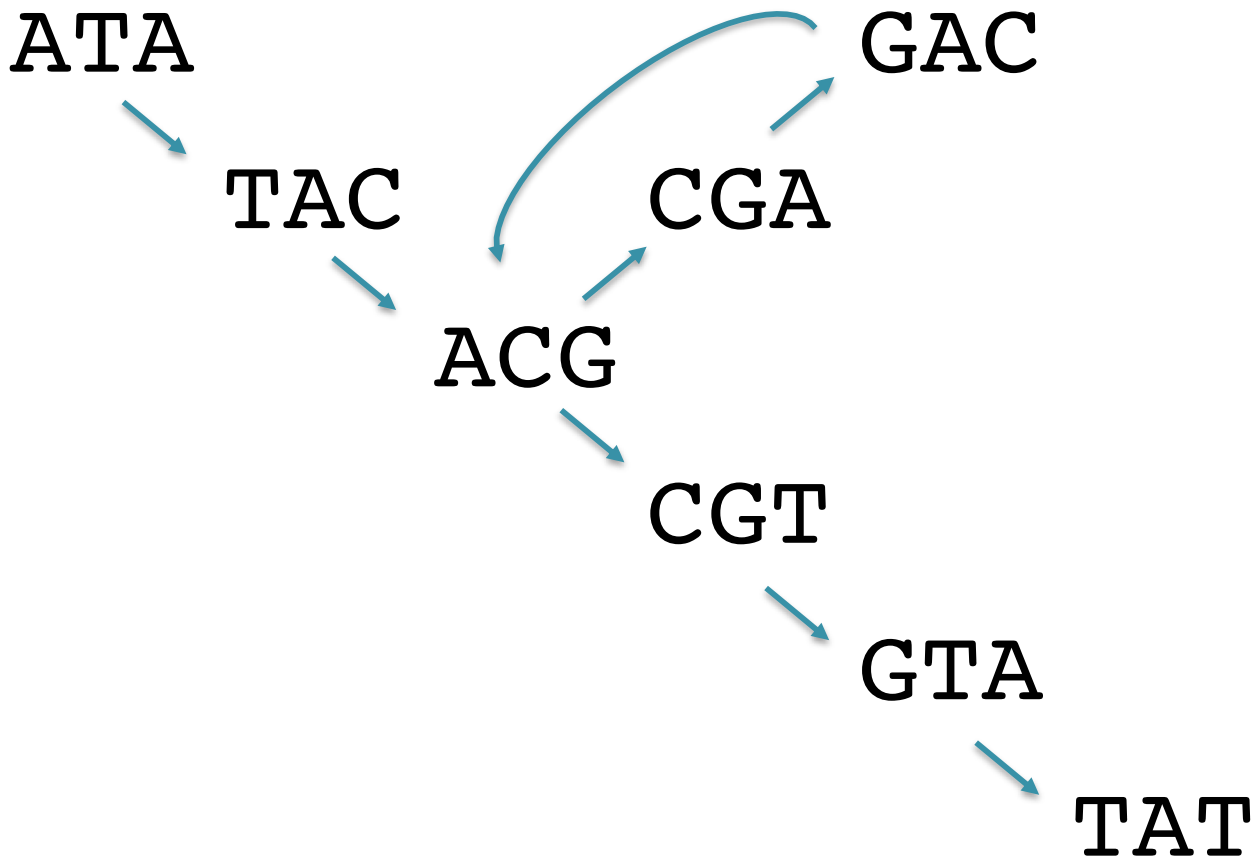
~~TACG~~



Pop Quiz 2

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

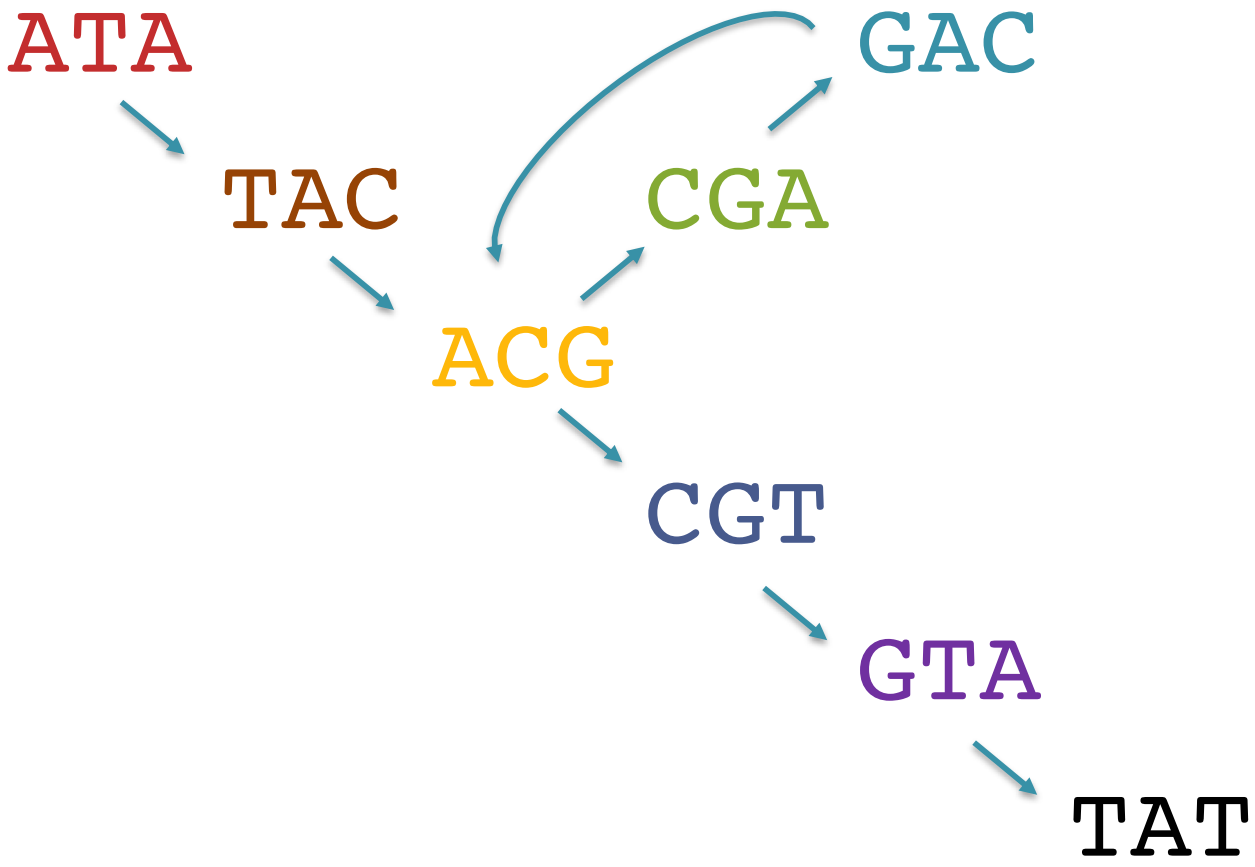
~~ACGA~~
~~ACGT~~
~~ATAC~~
~~CGAC~~
~~CGTA~~
~~GACG~~
~~GTAT~~
~~TACG~~



Pop Quiz 2

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

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

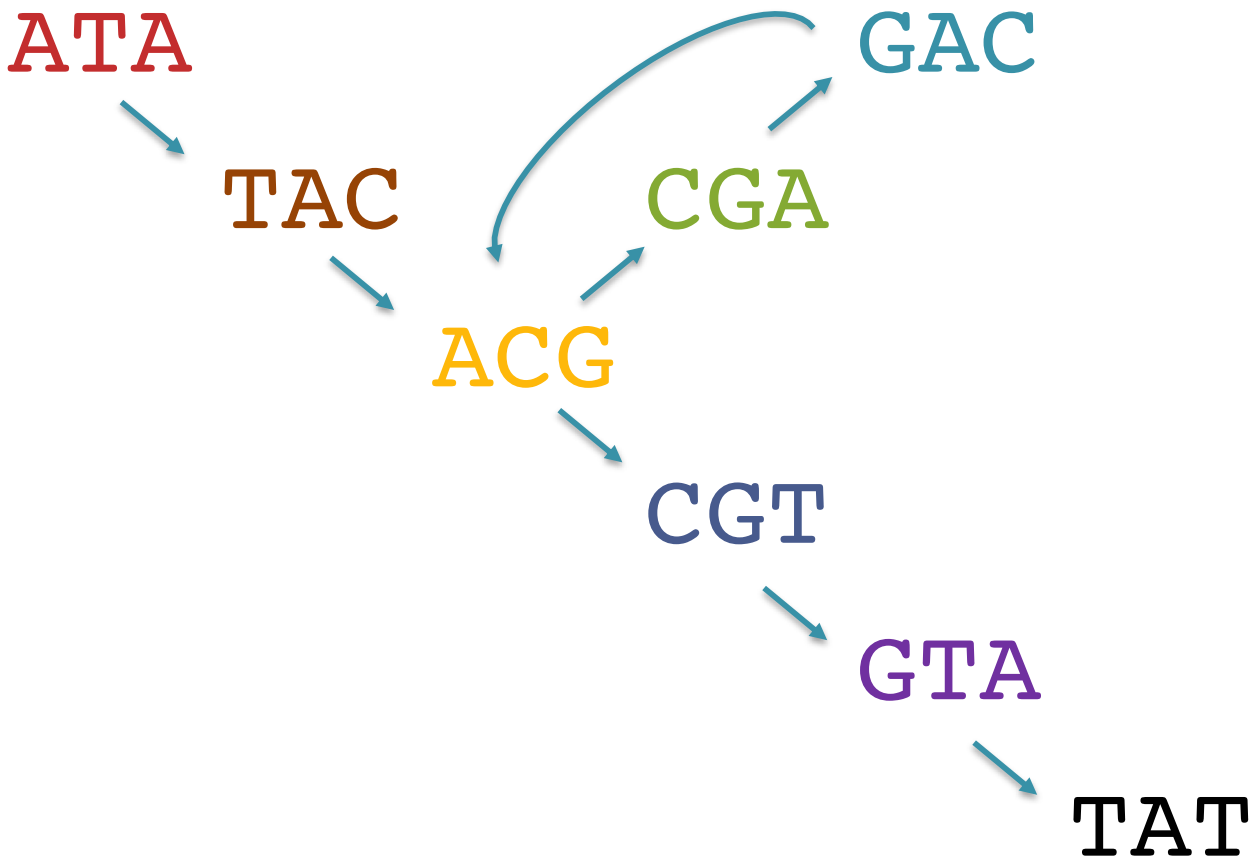


ATACGACGTAT

Pop Quiz 2

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

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



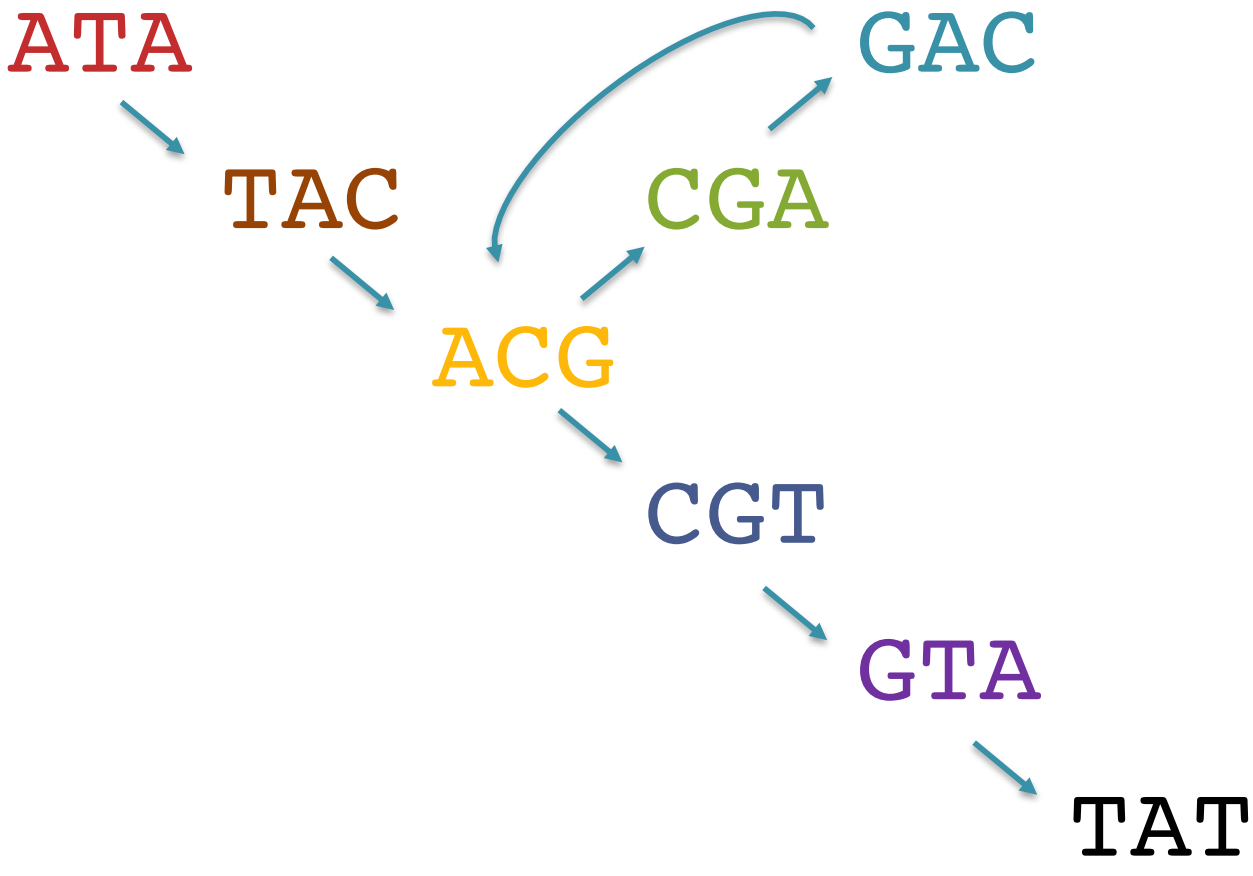
Whats another possible genome?

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

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