3. Genome Assembly

Michael Schatz

Feb 7, 2017

JHU 600.649: Applied Comparative Genomics

Welcome!

The primary goal of the course is for students to be grounded in theory and leave the course empowered to conduct independent genomic analyses.

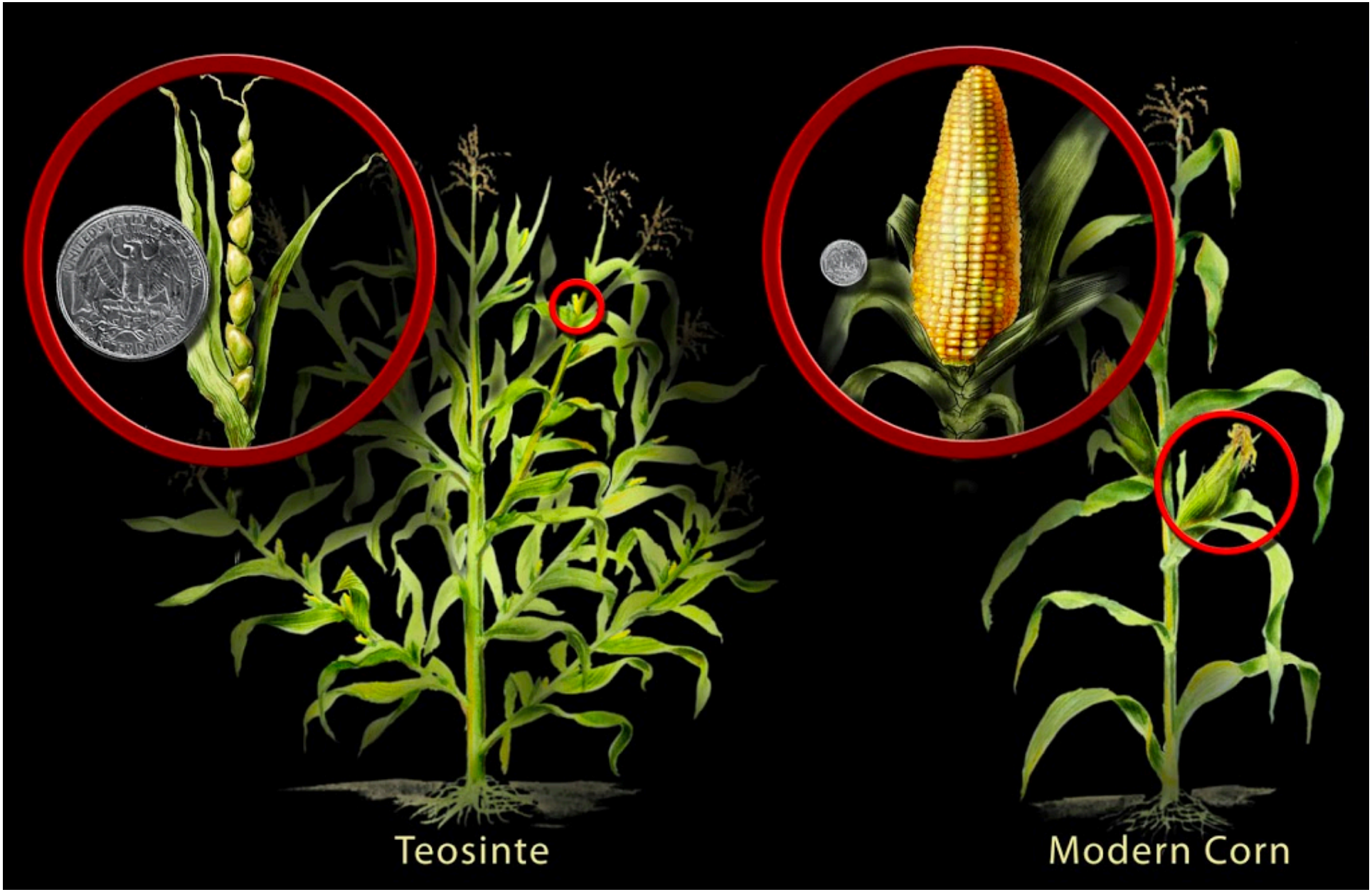
- We will study the leading computational and quantitative approaches for comparing and analyzing genomes starting from raw sequencing data.
- The course will focus on human genomics and human medical applications, but the techniques will be broadly applicable across the tree of life.
- The topics will include genome assembly & comparative genomics, variant identification & analysis, gene expression & regulation, personal genome analysis, and cancer genomics.

Course Webpage:	https://github.com/schatzlab/appliedgenomics
Course Discussions:	http://piazza.com
Class Hours:	Tues + Thurs @ 1:30p – 2:45p, Shaffer 304
Office Hours:	Tues + Thurs @ 3-4p and by appointment Please try Piazza first!

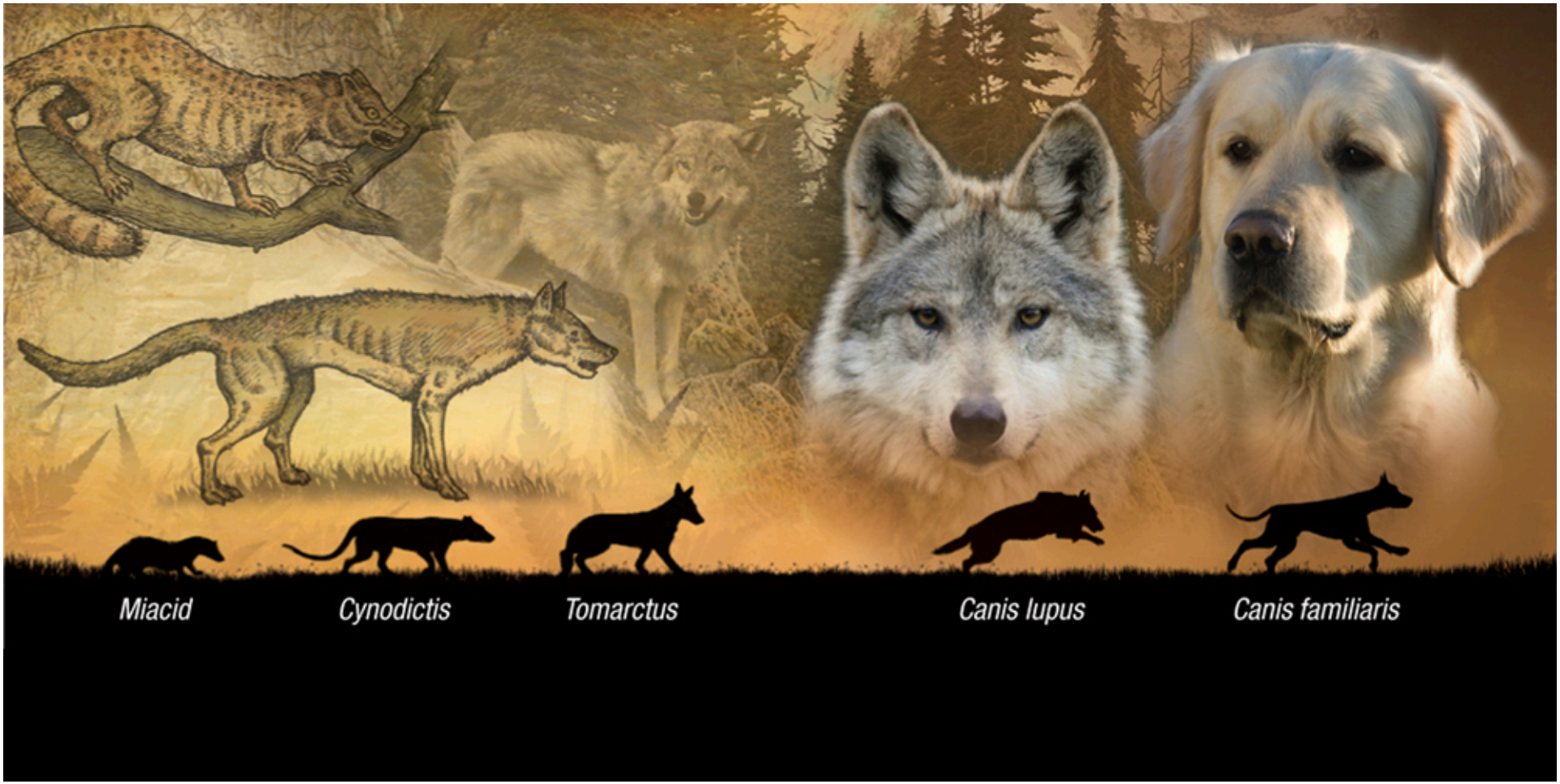
Earliest Genomics

Any Guesses?

Earliest Genomics



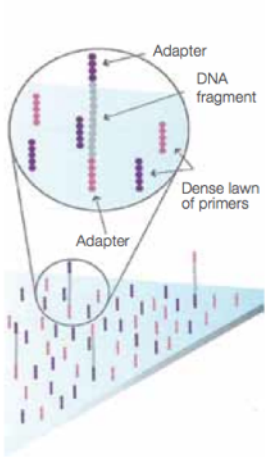
Earliest Genomics



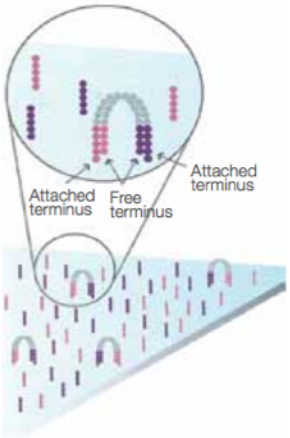
Second Generation Sequencing

Illumina HiSeq 2000
Sequencing by Synthesis

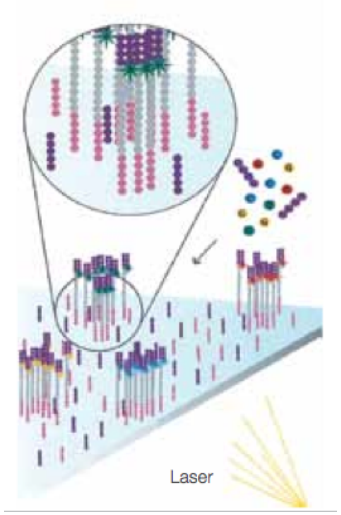
>60Gbp / day



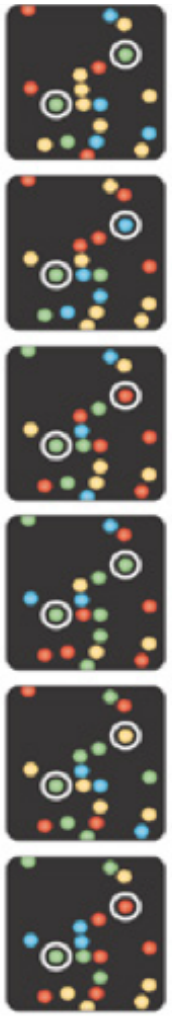
1. Attach



2. Amplify

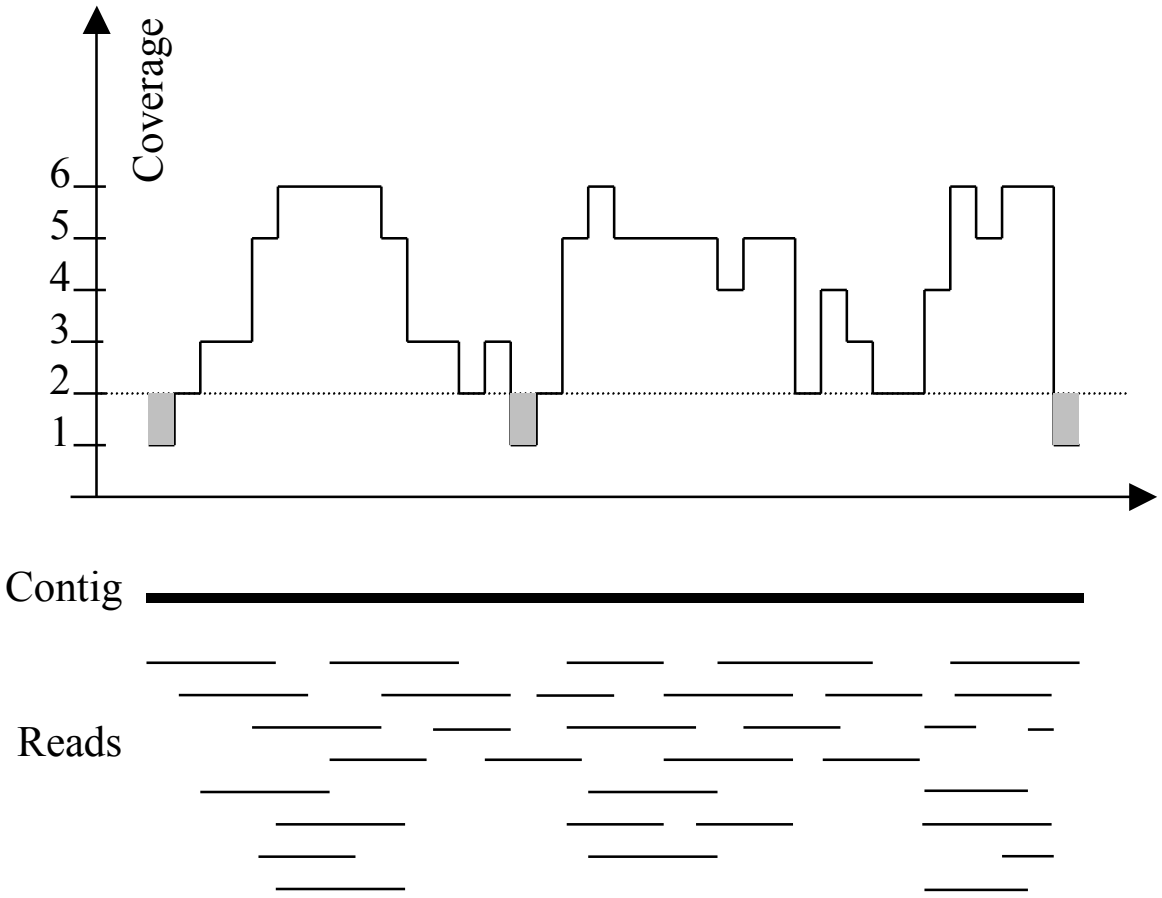


3. Image



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

Typical sequencing coverage

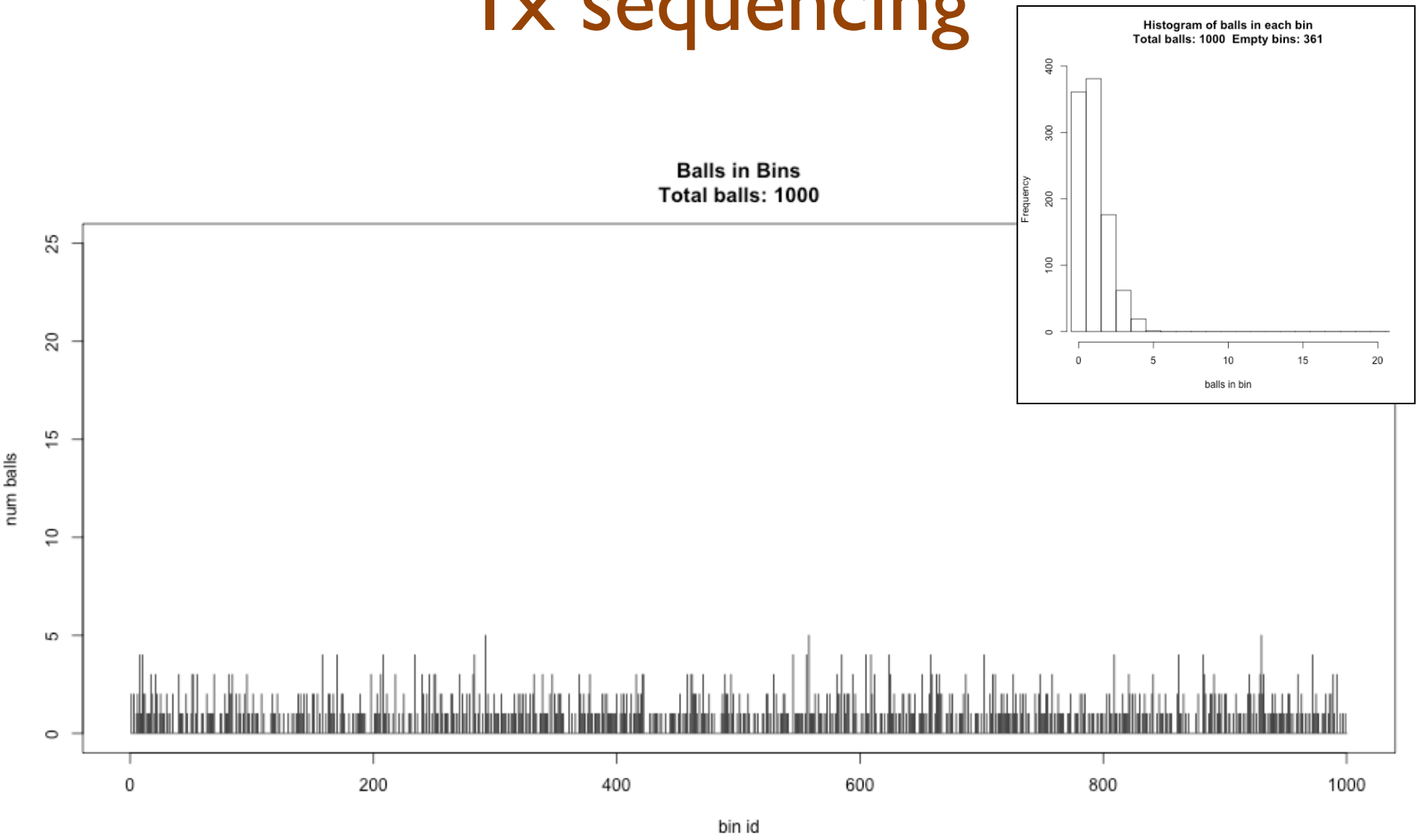


Imagine raindrops on a sidewalk

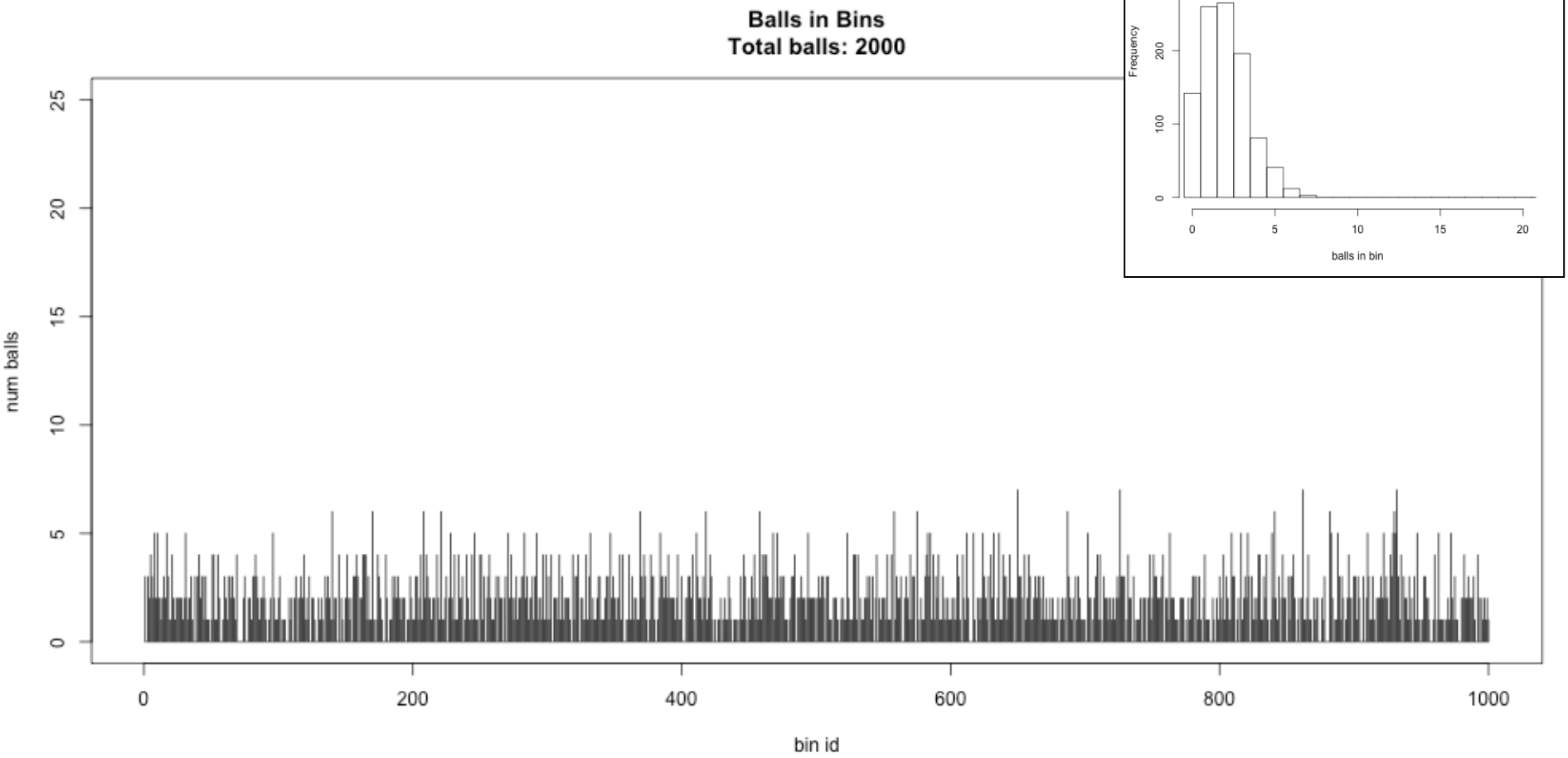
We want to cover the entire sidewalk but each drop costs $1

If the genome is 100 Mbp, should we sequence 1M 100bp reads?

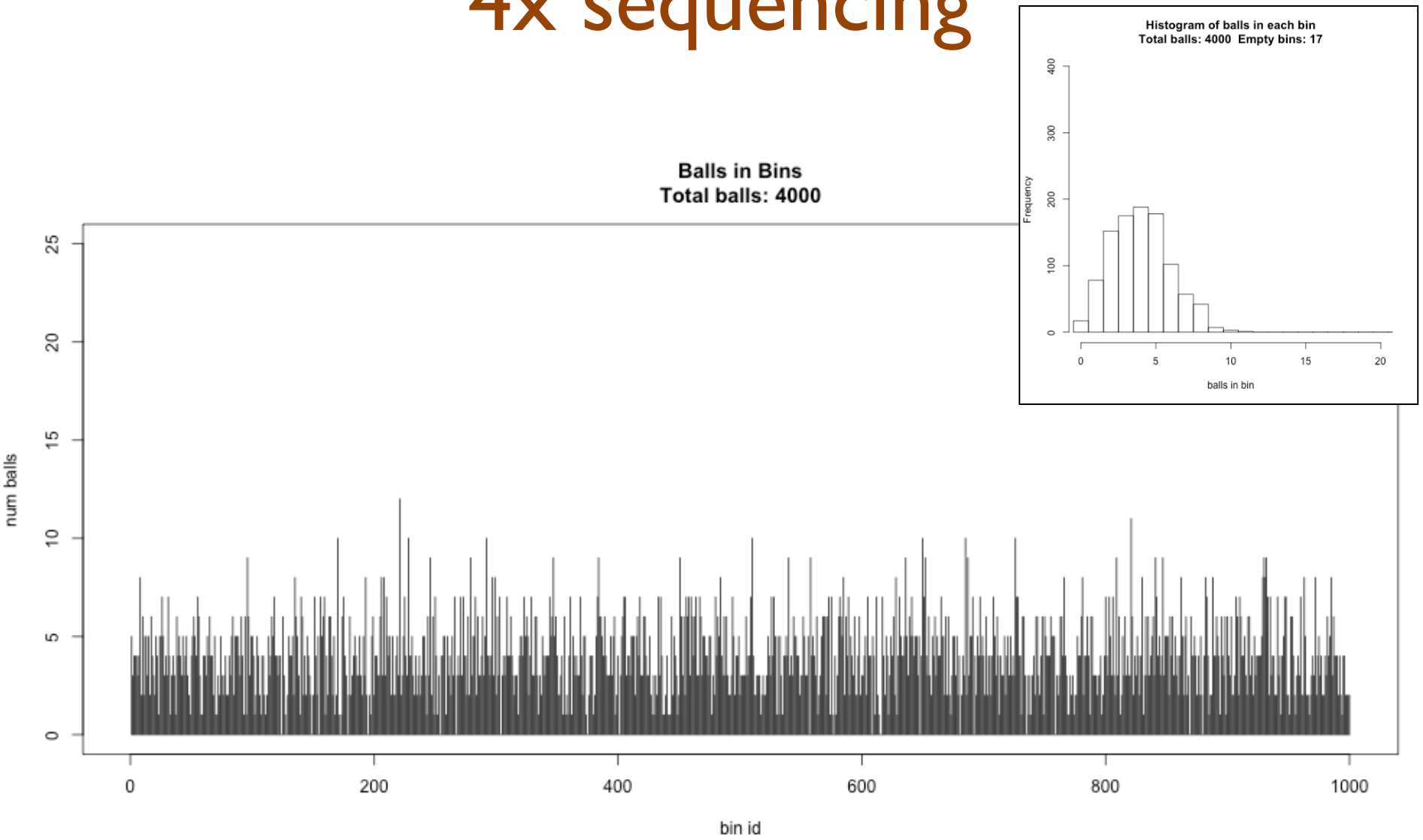
1x sequencing



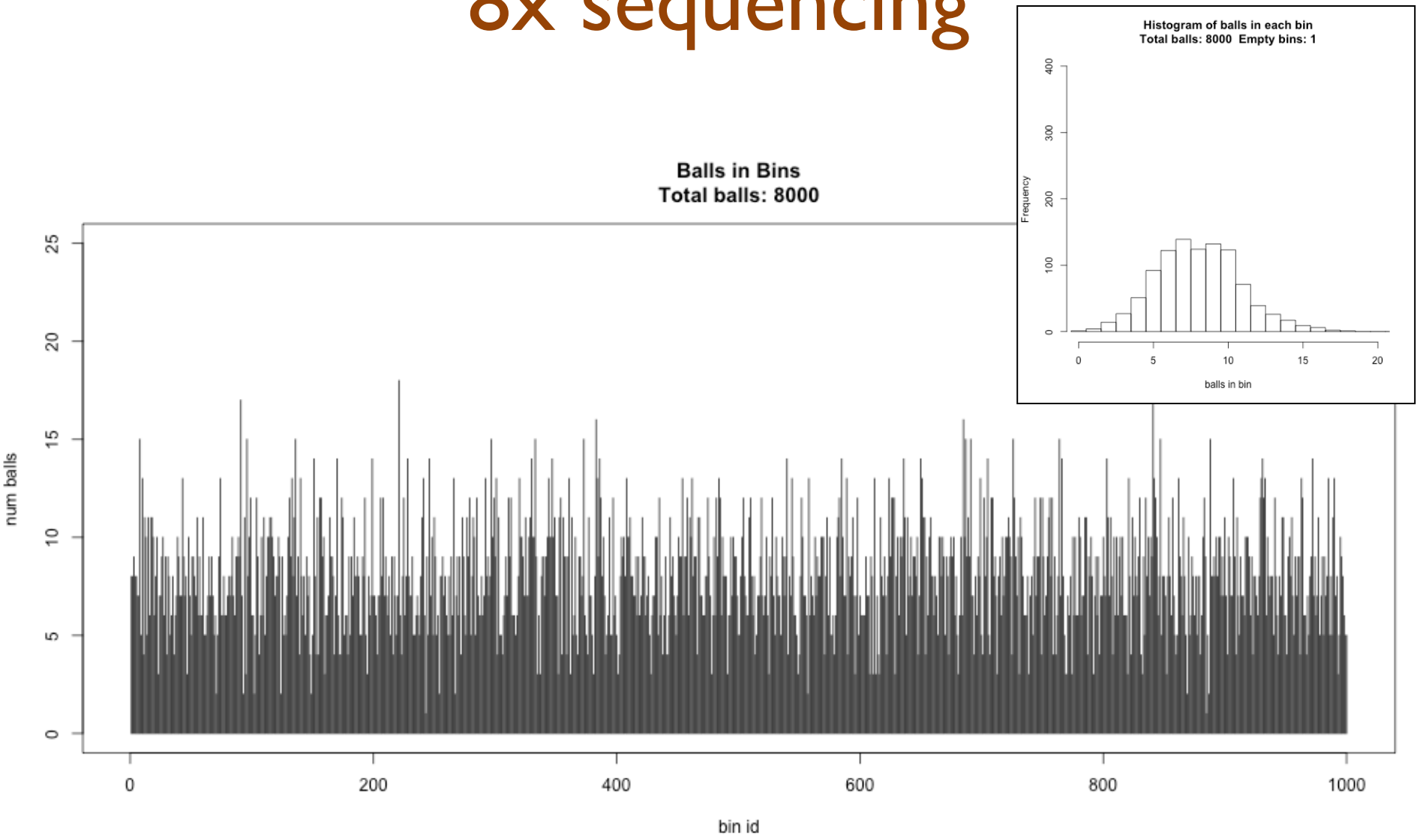
2x sequencing



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.

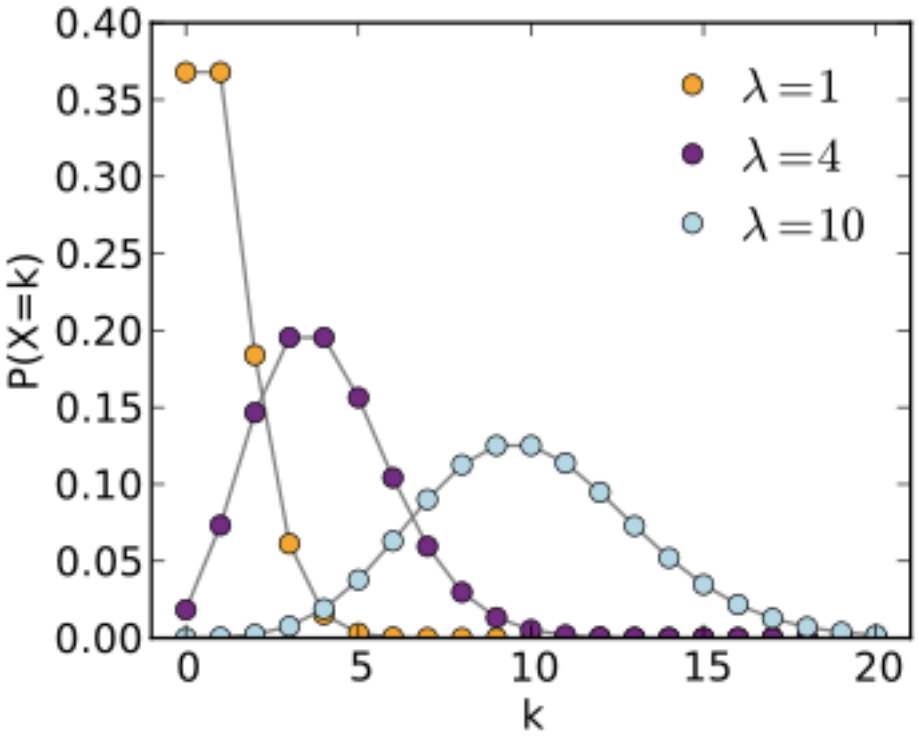
$$P(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

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***



Pop Quiz!

I want to sequence a 10Mbp genome to 24x coverage.
How many 150bp reads do I need?

I need 10Mbp x 24x = 240Mbp of data
240Mbp / 150bp / read = 1.6M reads

I want to sequence a 10Mbp genome so that
95% of the genome has at least 24x coverage.
How many 150bp reads do I need?

Find X such that X-2sqrt(X) = 24

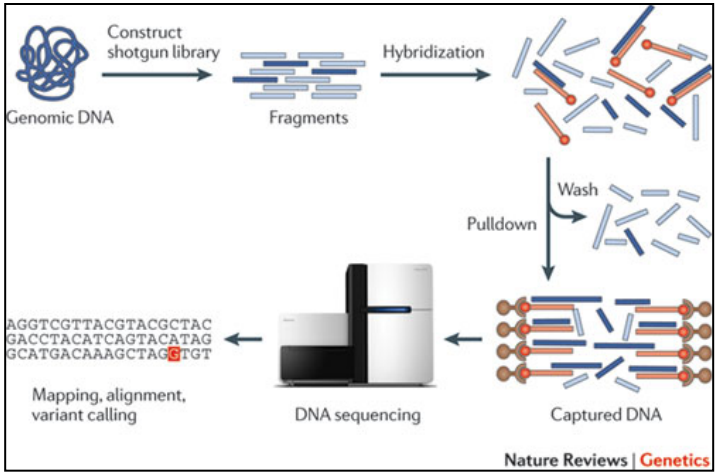
36-2*sqrt(36) = 24

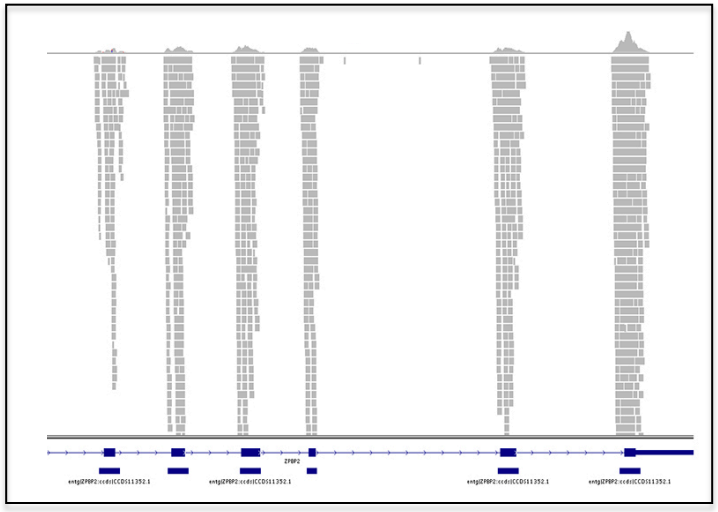
I need 10Mbp x 36x = 360Mbp of data
360Mbp / 150bp / read = 2.4M reads

Exome-Capture Sequencing

Exome-capture reduces the costs of sequencing

- Currently targets around 50Mbp of sequence: all exons plus flanking regions
- WGS currently costs ~$1500 per sample, while WES currently costs ~$300 per sample
- Coverage is highly localized around genes, although will get sparse coverage throughout rest of genome





Exome sequencing as a tool for Mendelian disease gene discovery
Bamshad et al. (2011) *Nature Reviews Genetics*. 12, 745-755

Illumina Sequencing Summary

Advantages:

- Best throughput, accuracy and read length for any 2nd gen. sequencer
- Fast & robust library preparation

Disadvantages:

- Inherent limits to read length (practically, 150bp)
- Some runs are error prone
- Requires amplification, sequences a population of molecules

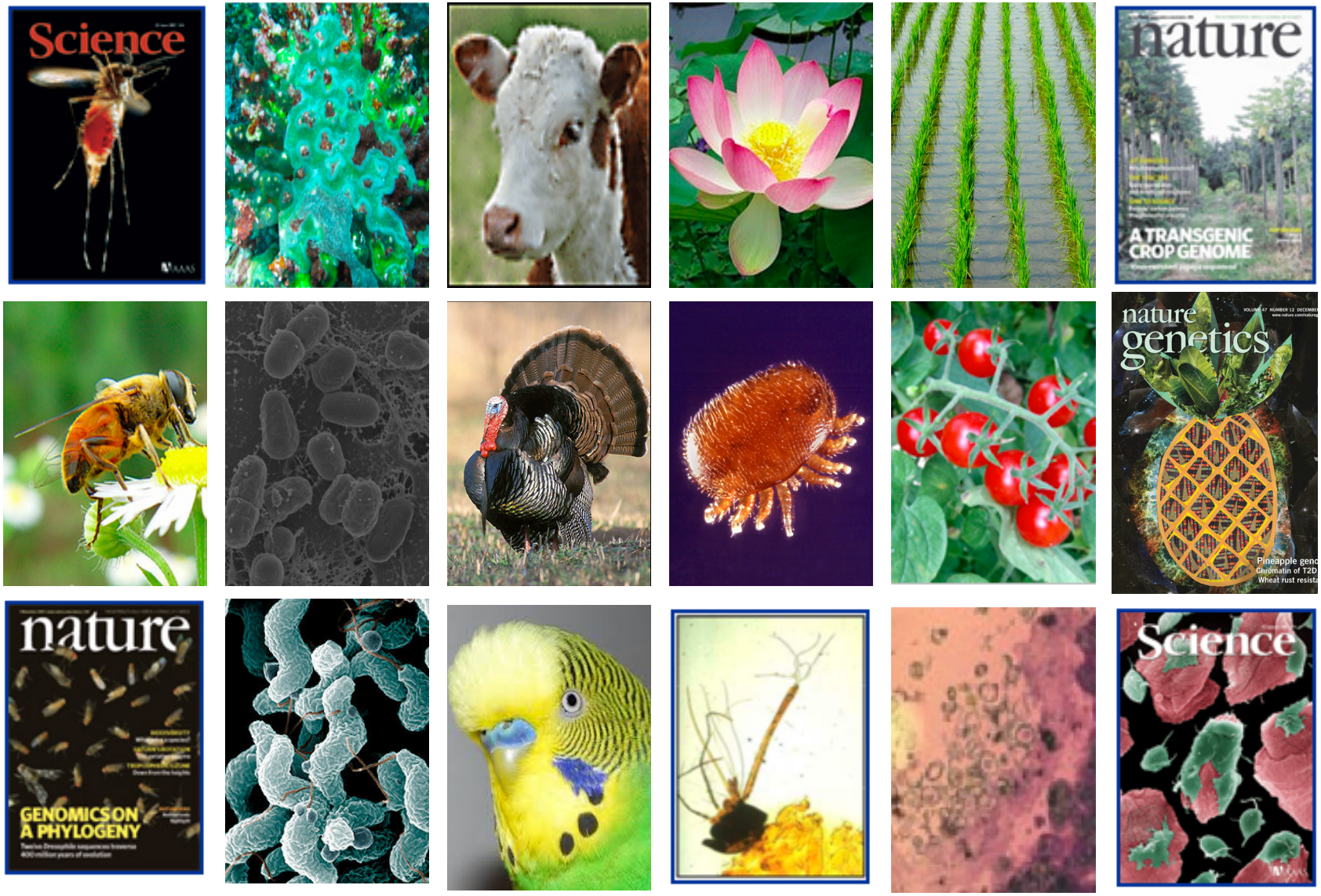
Illumina HiSeq
~3 billion paired 100bp reads
~600Gb, $10K, 8 days
(or “rapid run” ~90Gb in 1-2 days)

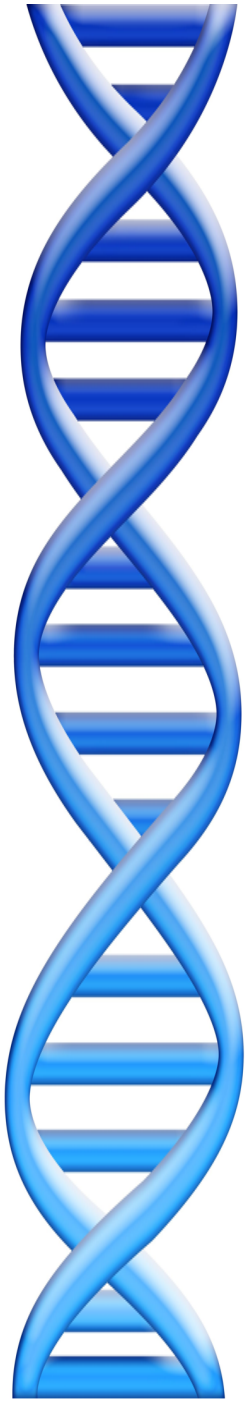
Illumina X Ten
~6 billion paired 150bp reads
1.8Tb, <3 days, ~1000 / genome($$)
(or “rapid run” ~90Gb in 1-2 days)

Illumina NextSeq
One human genome in **<30 hours**

Ira Hall

Genome Assembly





Outline

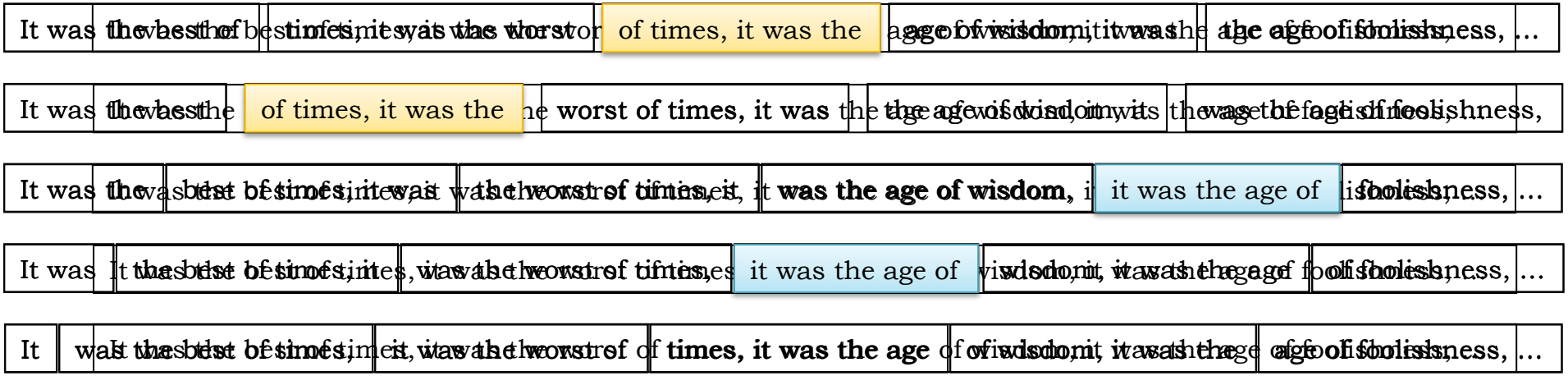
1. ***Assembly theory***
   - Assembly by analogy

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

3. ***Next-next-gen Assembly***
   - Canu: recommended for PacBio/ONT project

Shredded Book Reconstruction

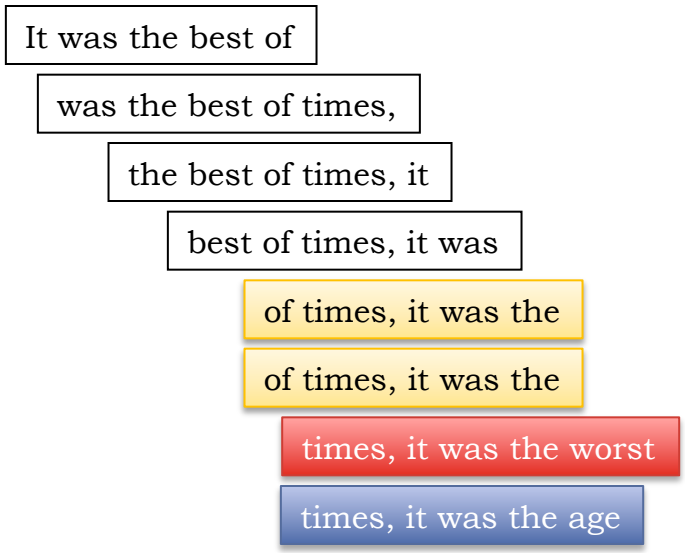
- Dickens accidentally shreds the first printing of A Tale of Two Cities
  - Text printed on 5 long spools



- 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

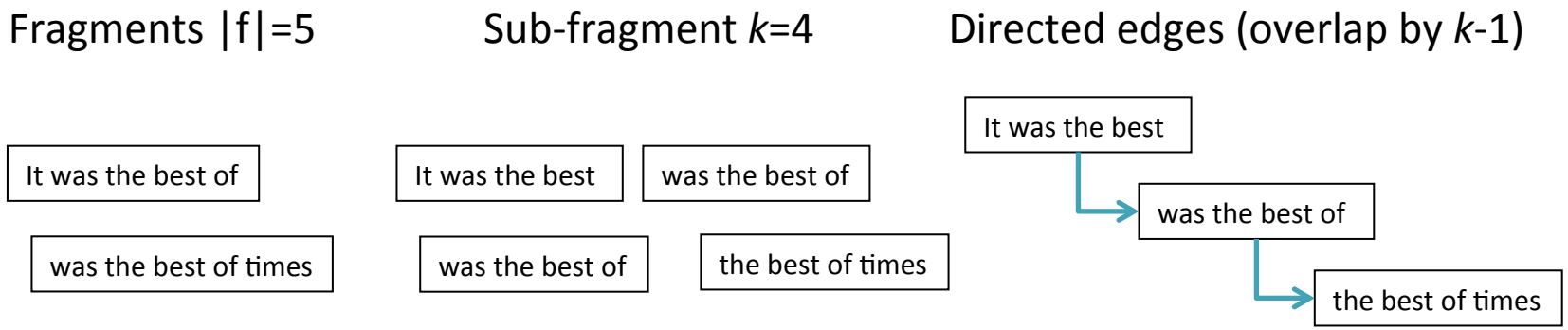
- It was the best of times, it was the [worst/age]

Model the assembly problem as a graph problem

How long will it take to compute the overlaps?

de Bruijn Graph Construction

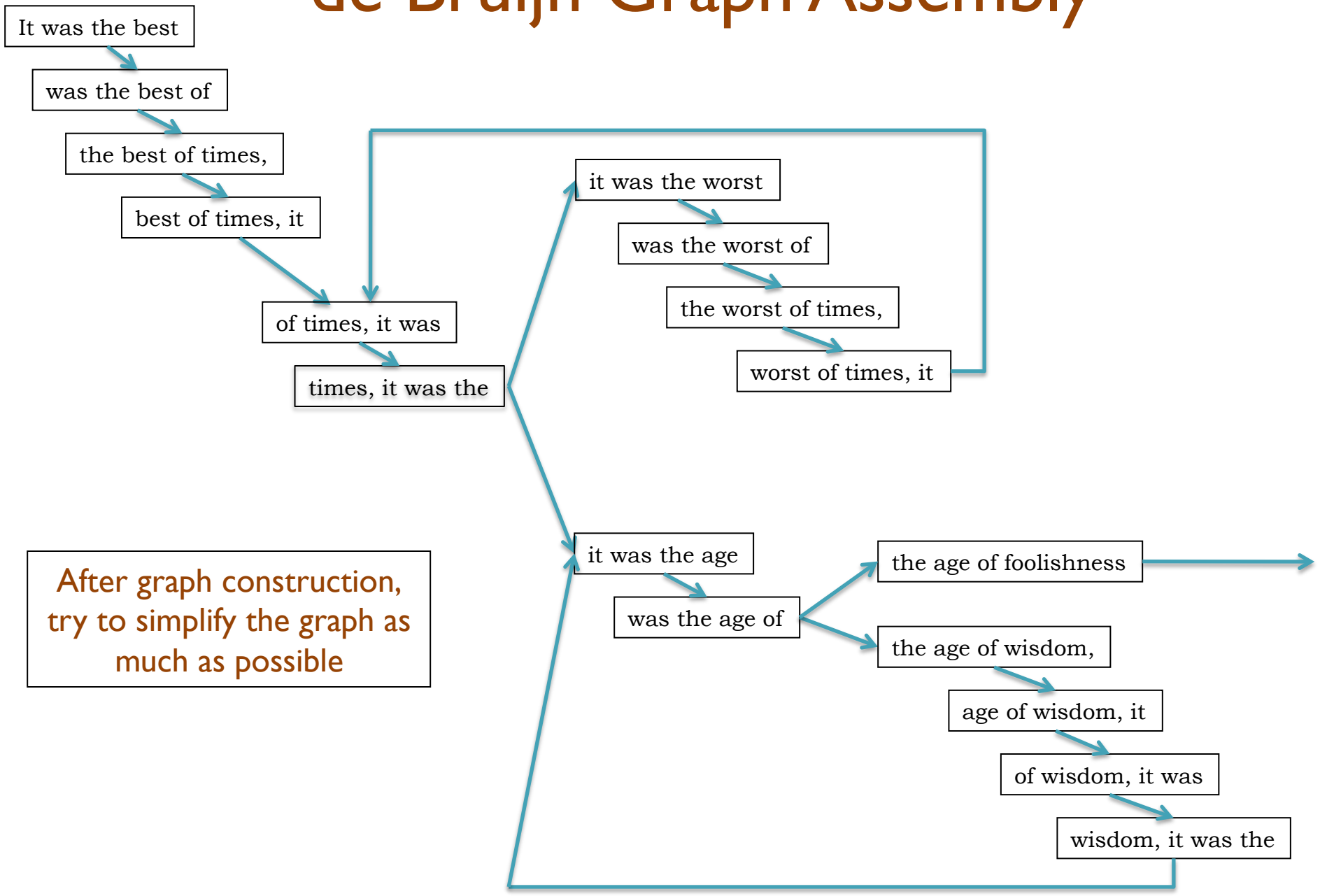
- $G_k = (V,E)$
  - V = Length-*k* sub-fragments
  - E = Directed edges between consecutive sub-fragments
    - Sub-fragments overlap by *k*-1 words



– Overlaps between fragments are implicitly computed

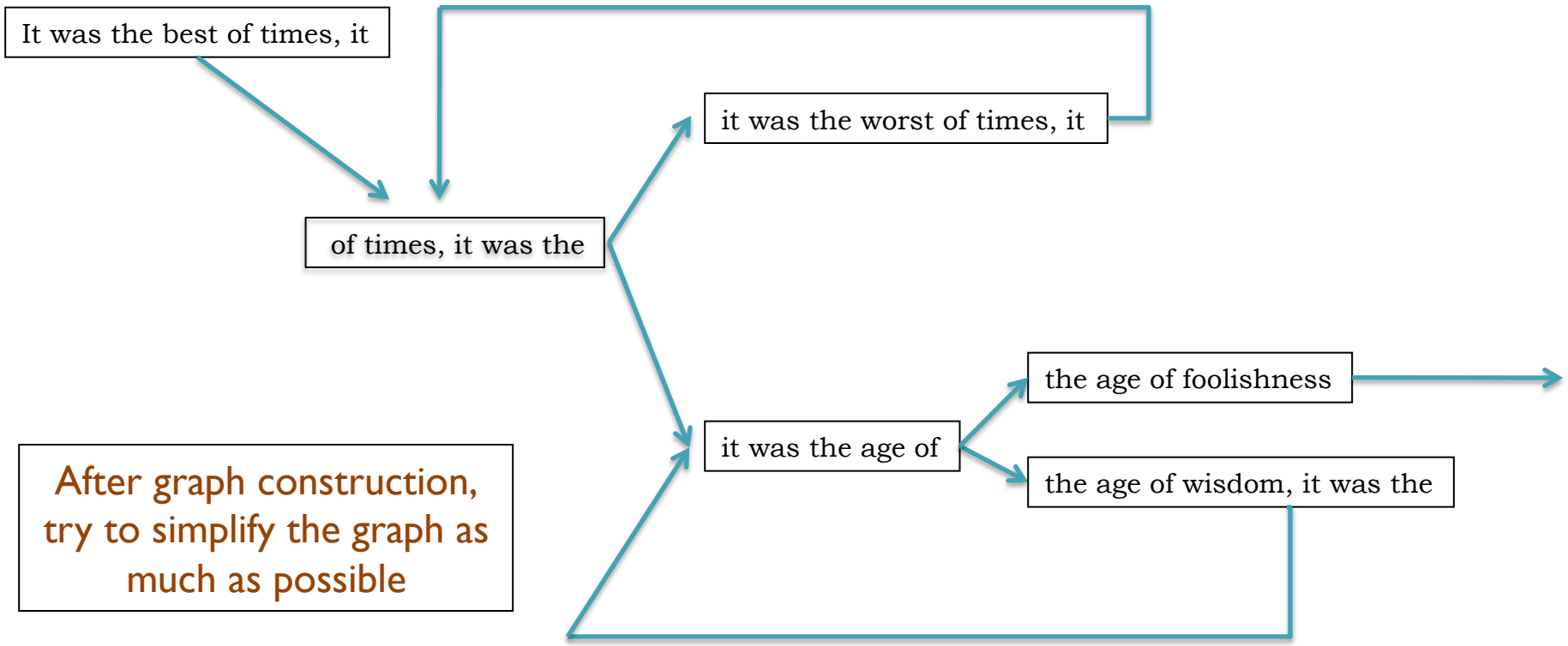
de Bruijn, 1946
Idury et al., 1995
Pevzner et al., 2001

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

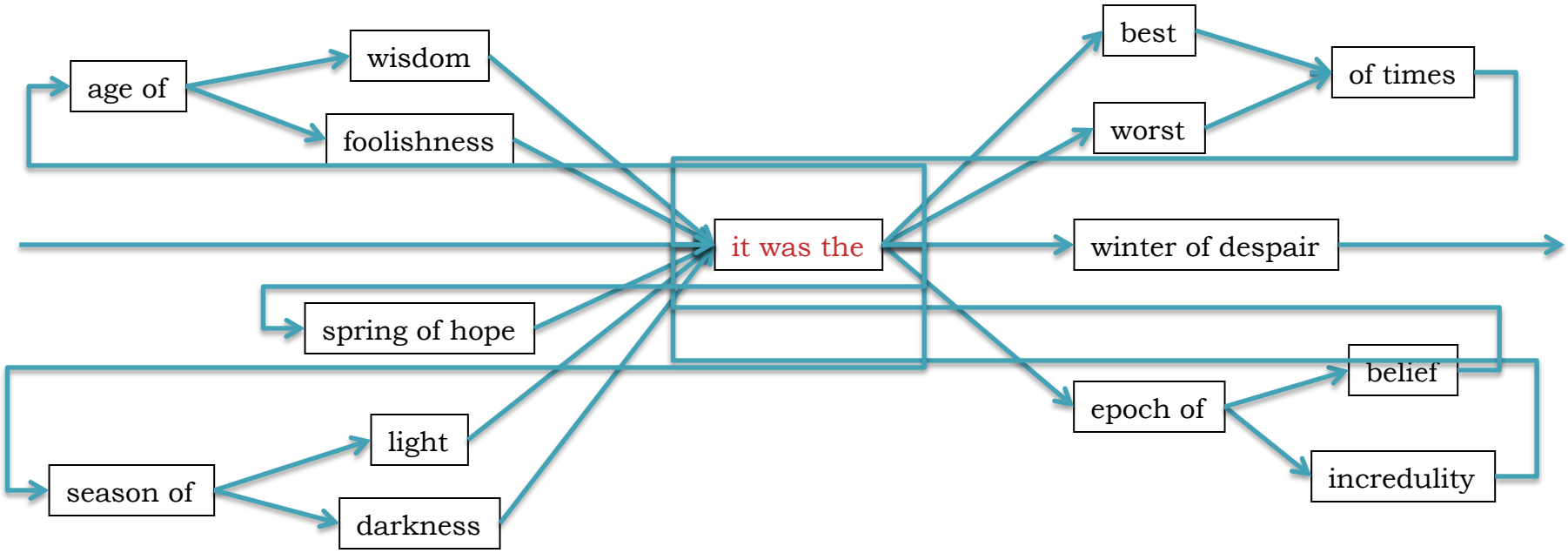
… it was the best of times it was the worst of times …

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

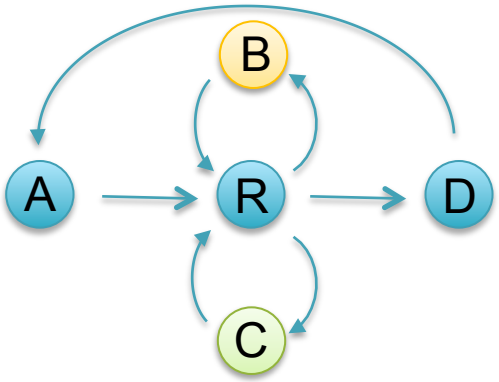
… it was the epoch of belief it was the epoch of incredulity …

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

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



Counting Eulerian Cycles



ARBRCRD
or
ARCRBRD

Generally an exponential number of compatible sequences

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

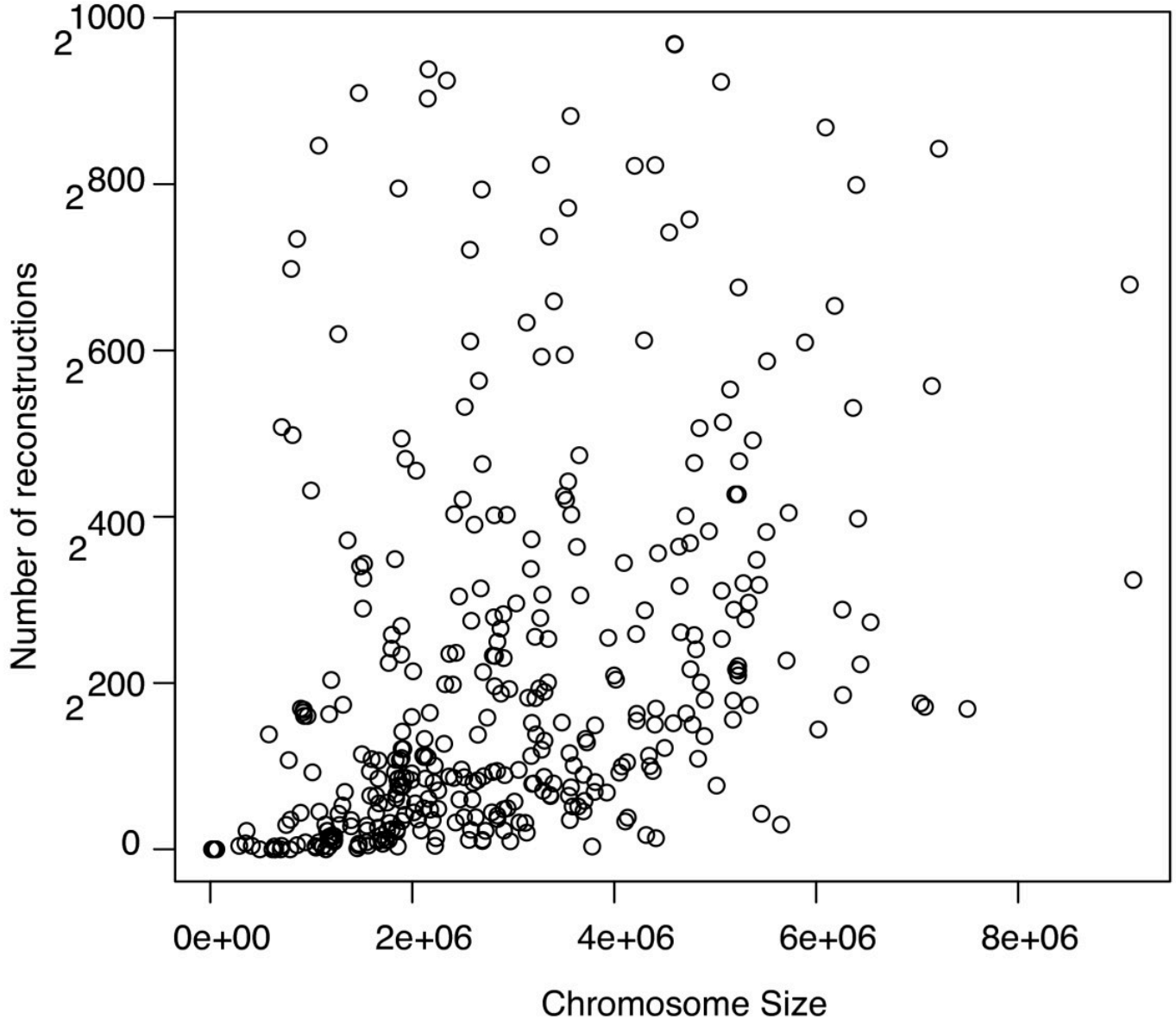
$$\mathcal{W}(G,t) = (\det L)\left\{\prod_{u\in V}(r_u-1)!\right\}\left\{\prod_{(u,v)\in E} a_{uv}!\right\}^{-1}$$

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

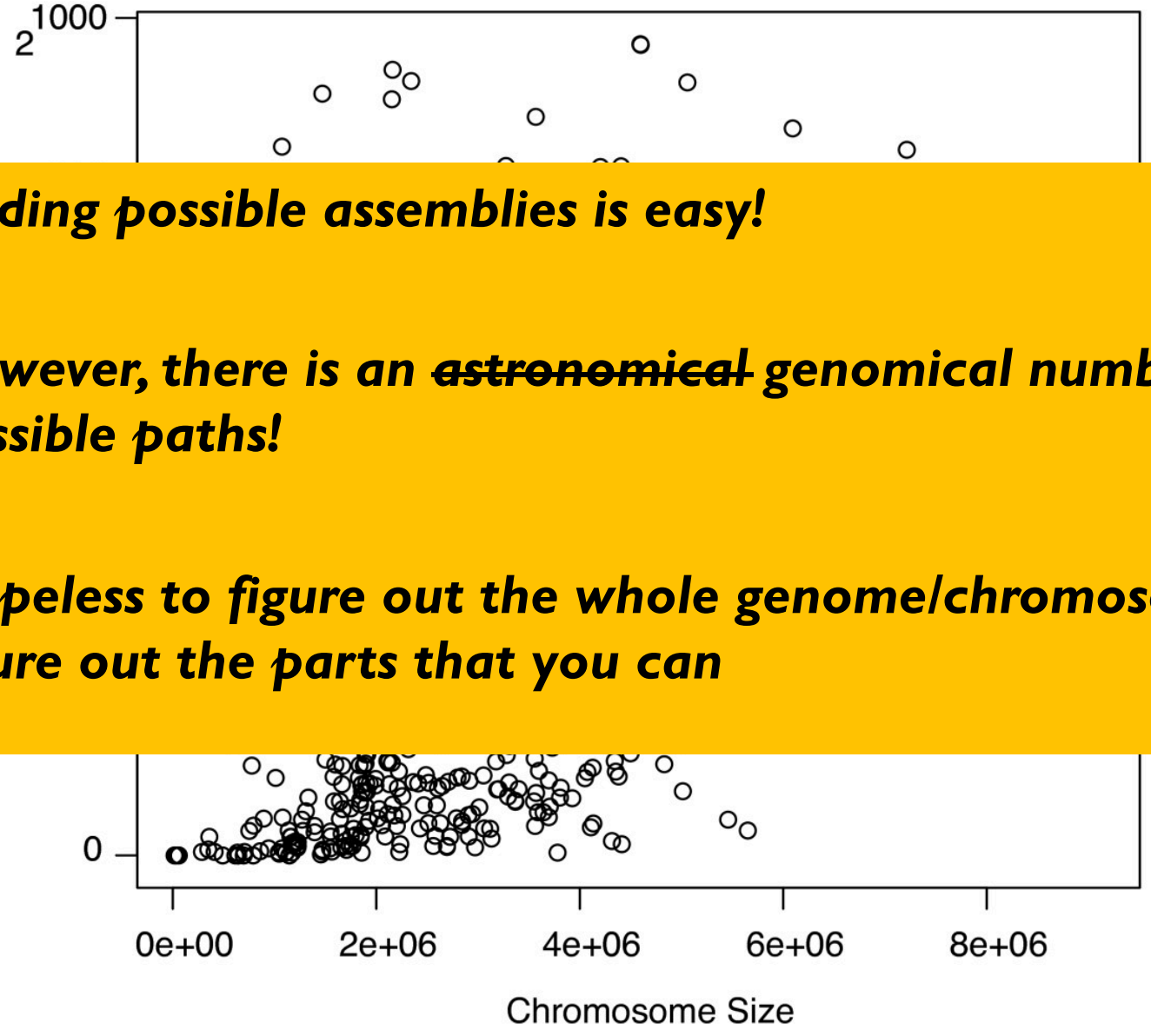
$r_u = d^+(u)+1$ if $u=t$, or $d^+(u)$ otherwise

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*.

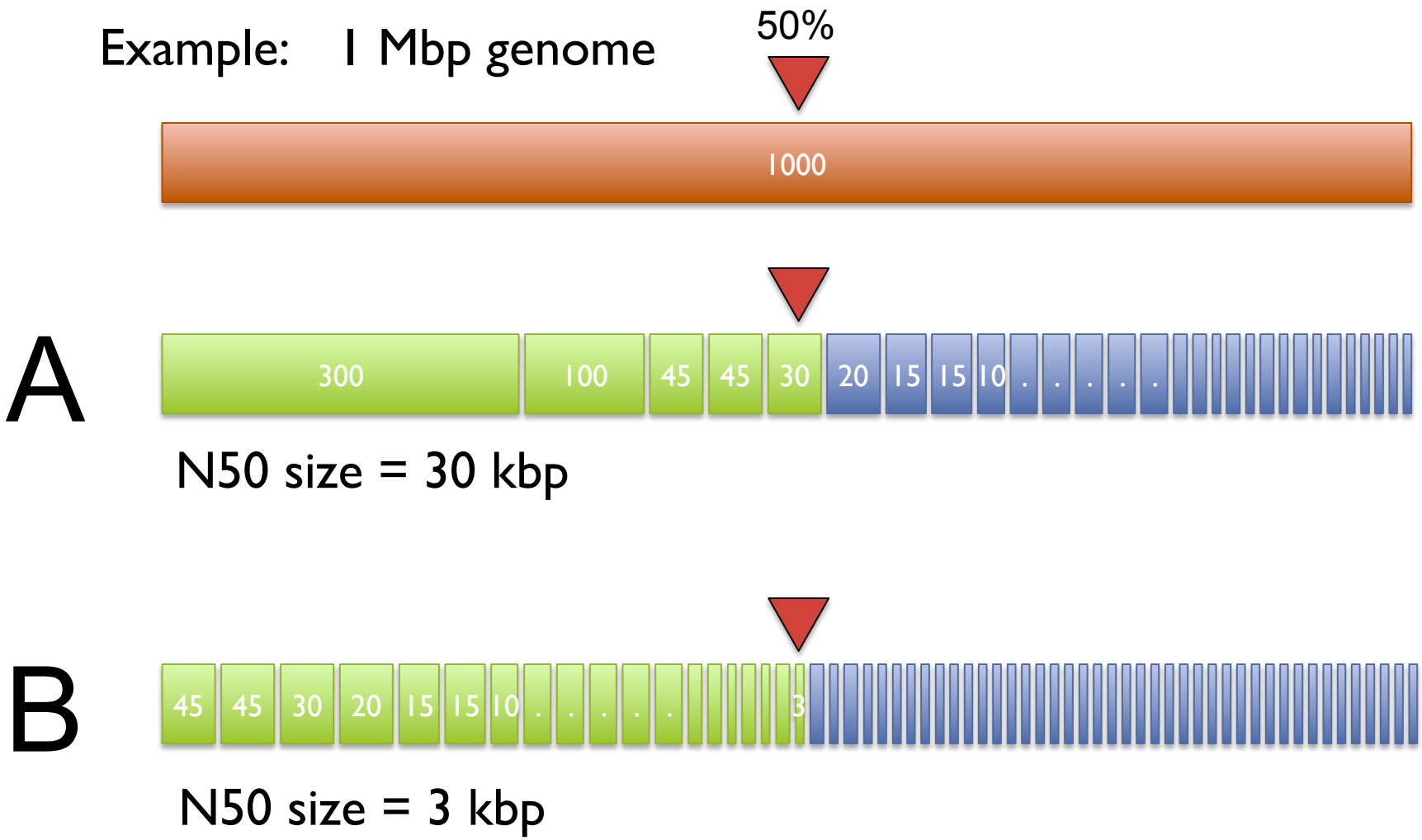


- ***Finding possible assemblies is easy!***
- ***However, there is an ~~astronomical~~ genomical number of possible paths!***
- ***Hopeless to figure out the whole genome/chromosome, figure out the parts that you can***

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

Contig N50

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



Contig N50

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

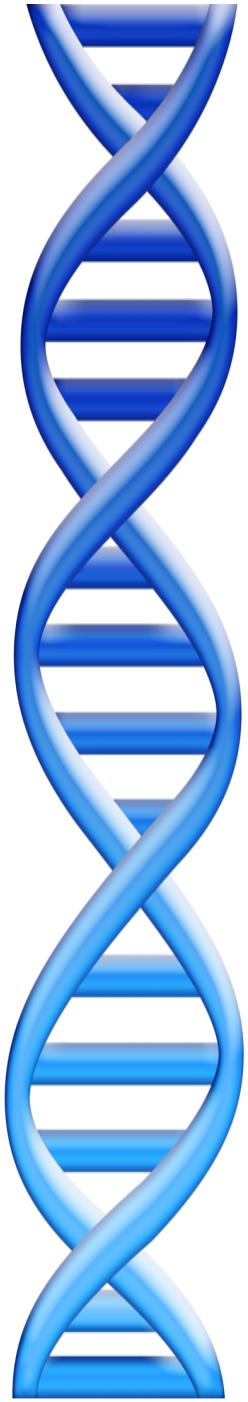
Better N50s improves the analysis in every dimension

- 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



Outline

1. *Assembly theory*
   - Assembly by analogy

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

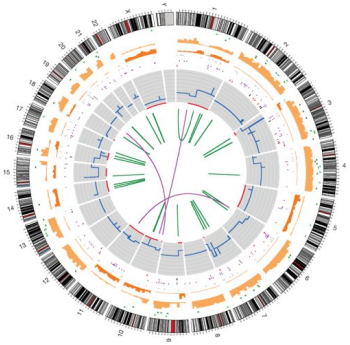
3. *Next-next-gen Assembly*
   - Canu: recommended for PacBio/ONT project

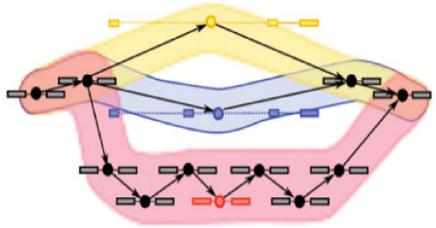
Assembly Applications

- Novel genomes

- Metagenomes

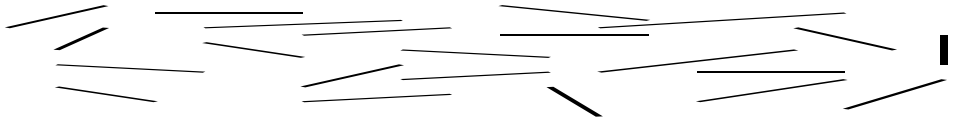
- Sequencing assays
  - Structural variations
  - Transcript assembly
  - …





Assembling a Genome

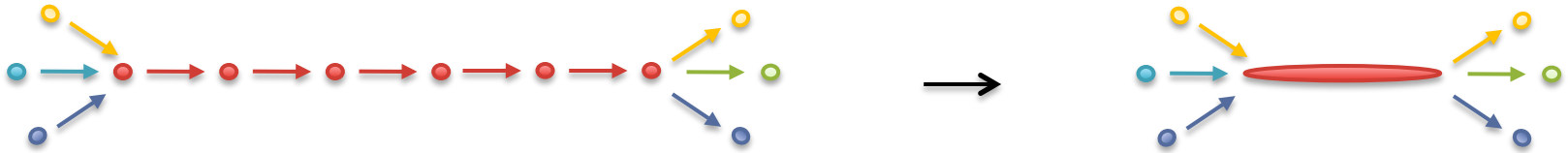
1. Shear & Sequence DNA



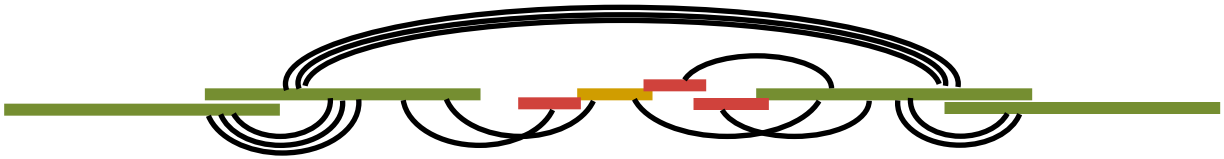
2. Construct assembly graph from reads (de Bruijn / overlap graph)

...AGCCTAGGGATGCGCGACACGT
GGATGCGCGACACGTCGCATATCCGGTTTGGTCAACCTCGGACGGAC
CAACCTCGGACGGACCTCAGCGAA...

3. Simplify assembly graph

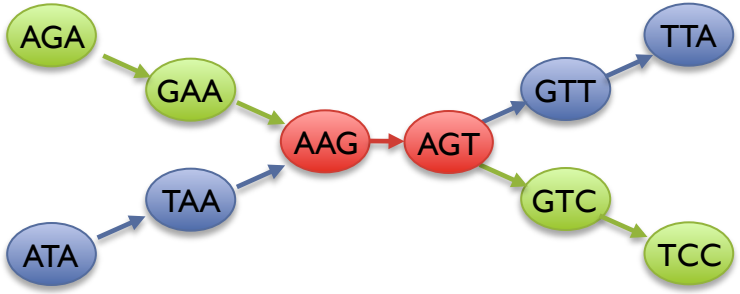


4. Detangle graph with long reads, mates, and other links



Two Paradigms for Assembly

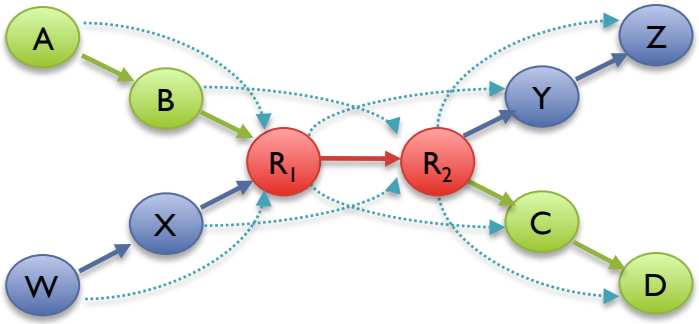
de Bruijn Graph



Short read assemblers

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

Overlap Graph



Long read assemblers

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

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

Why are genomes hard to assemble?

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

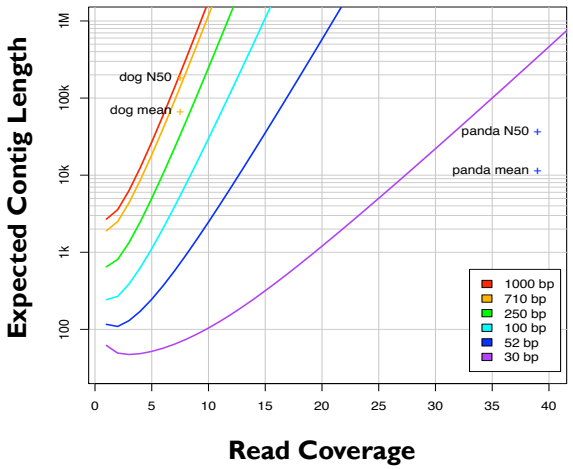
2. ***Sequencing***:
   - (Very) large genomes, imperfect sequencing

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

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

Ingredients for a good assembly

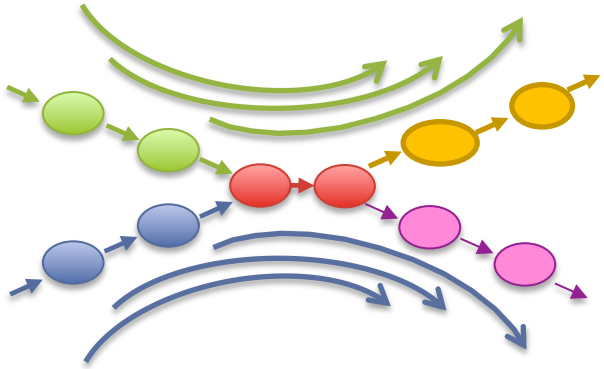
Coverage



High coverage is required

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

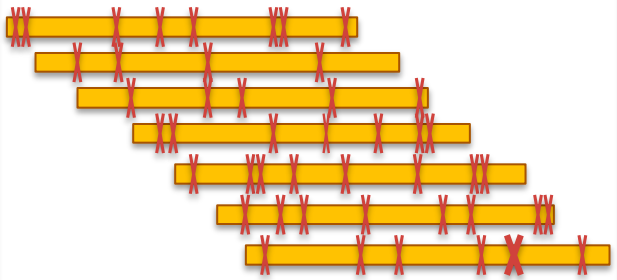
Read Length



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

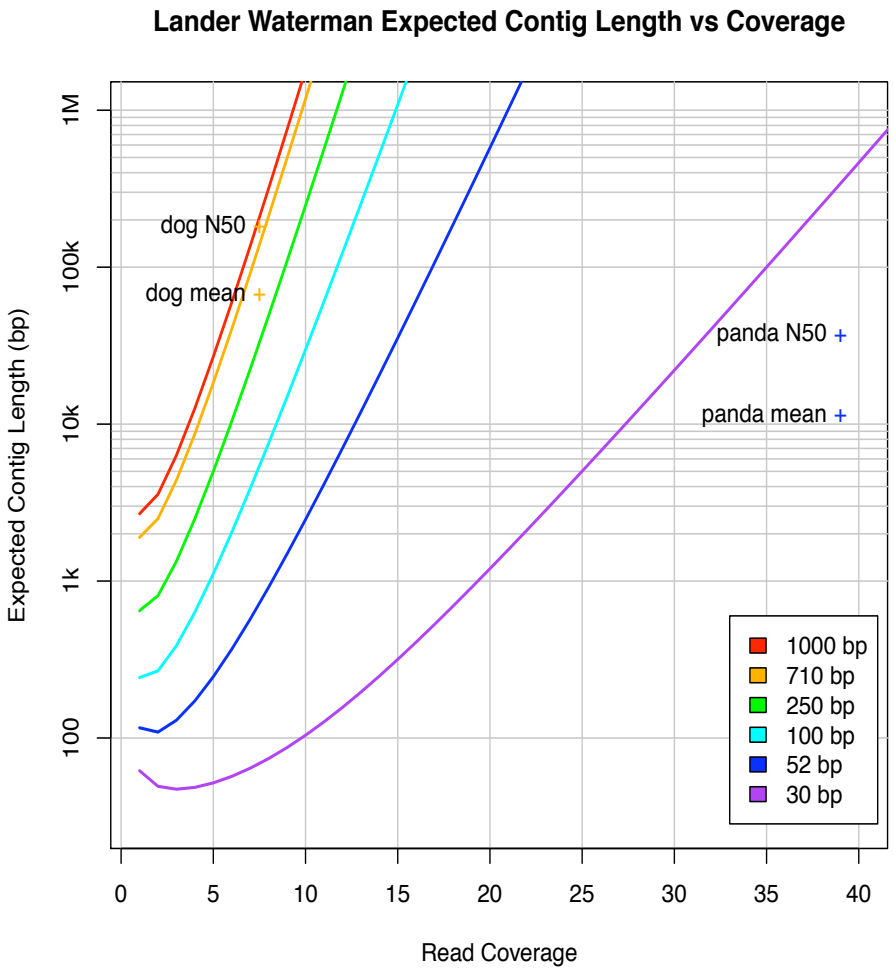
- Reads are assembled by finding kmers shared in pair of reads
- High error rate requires very short seeds, increasing complexity and forming assembly hairballs

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

Coverage and Read Length

Idealized Lander-Waterman model

- Reads start at perfectly random positions
- Contig length is a function of coverage and read length
  - Short reads require much higher coverage to reach same expected contig length
- Need even high coverage for higher ploidy, sequencing errors, sequencing biases
  - Recommend 100x coverage



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

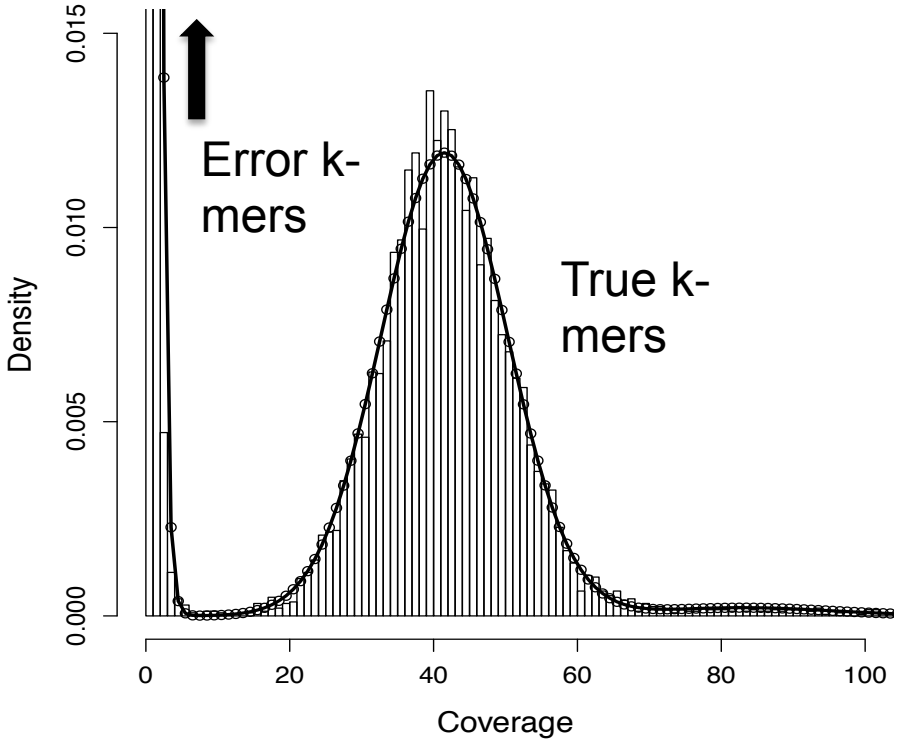
Error Correction with Quake

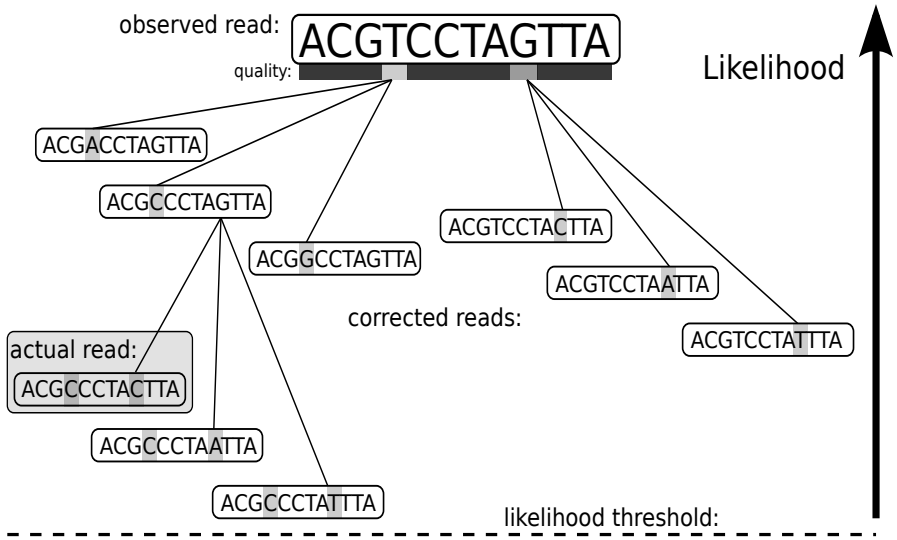
1. Count all "Q-mers" in reads

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

2. Correction Algorithm

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



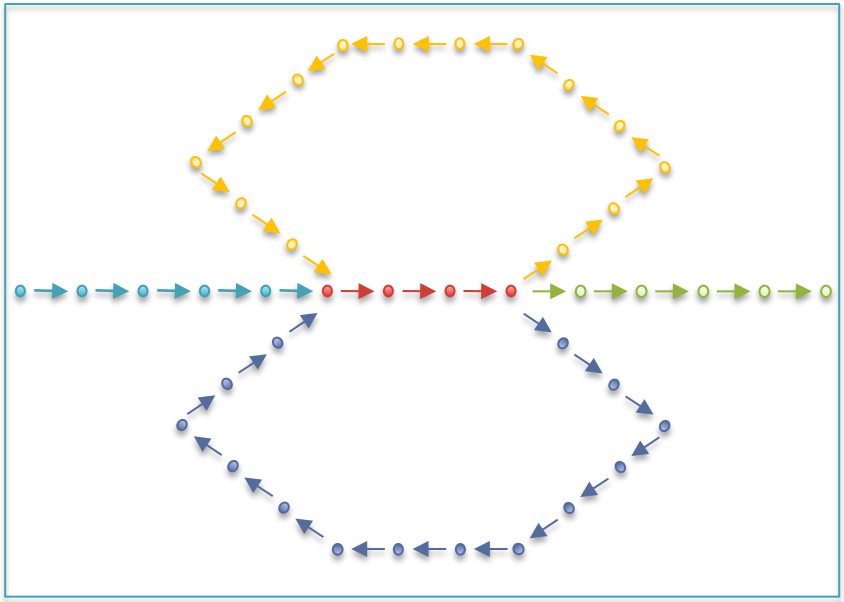


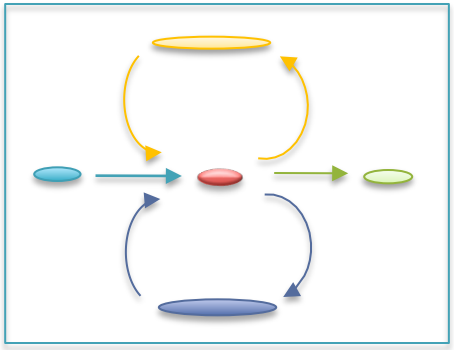
Quake: quality-aware detection and correction of sequencing reads.
Kelley, DR, Schatz, MC, Salzberg SL (2010) *Genome Biology*. 11:R116

Errors

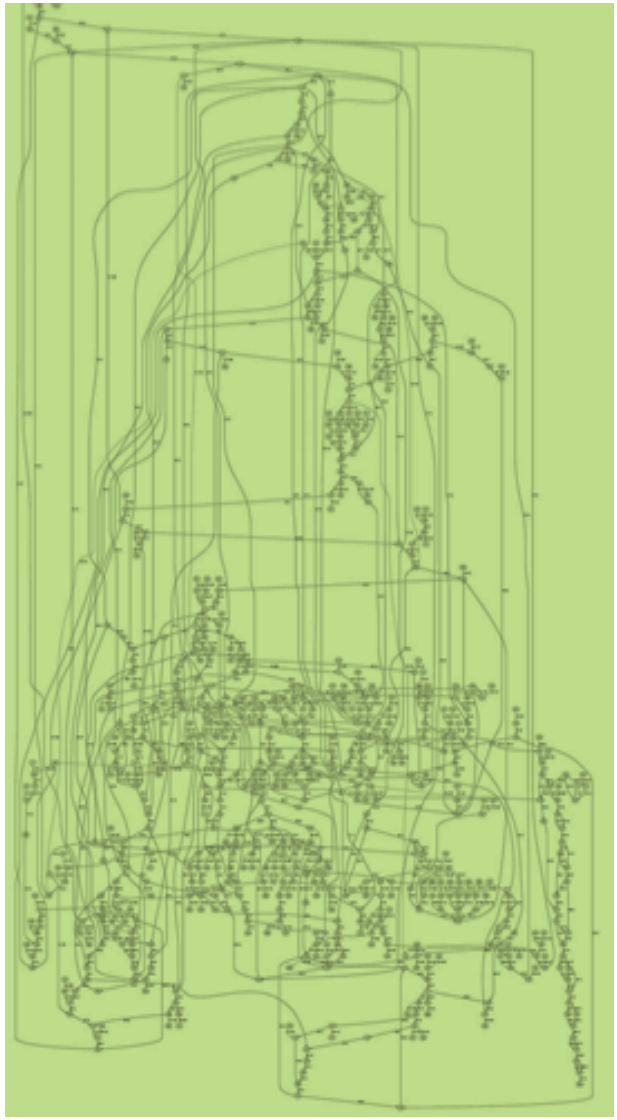
Unitigging / Unipathing

- After simplification and correction, compress graph down to its non-branching initial contigs
  - Aka "unitigs", "unipaths"
  - Unitigs end because of (1) lack of coverage, (2) errors, (3) heterozygosity and (4) repeats

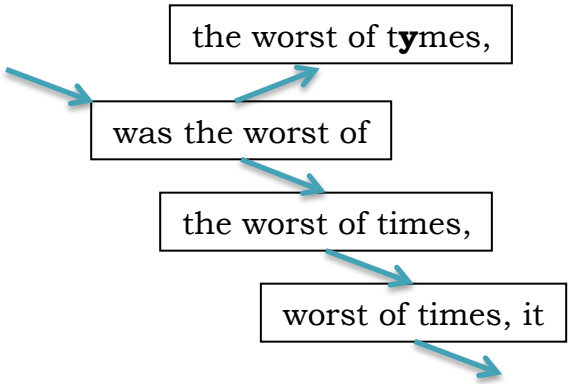
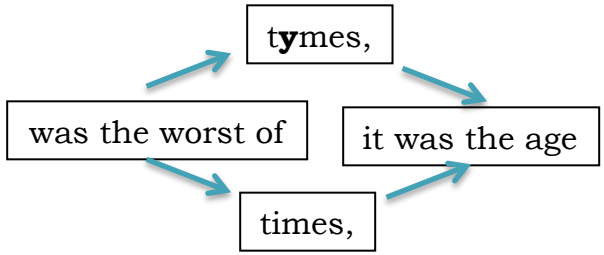




Errors in the graph



(Chaisson, 2009)

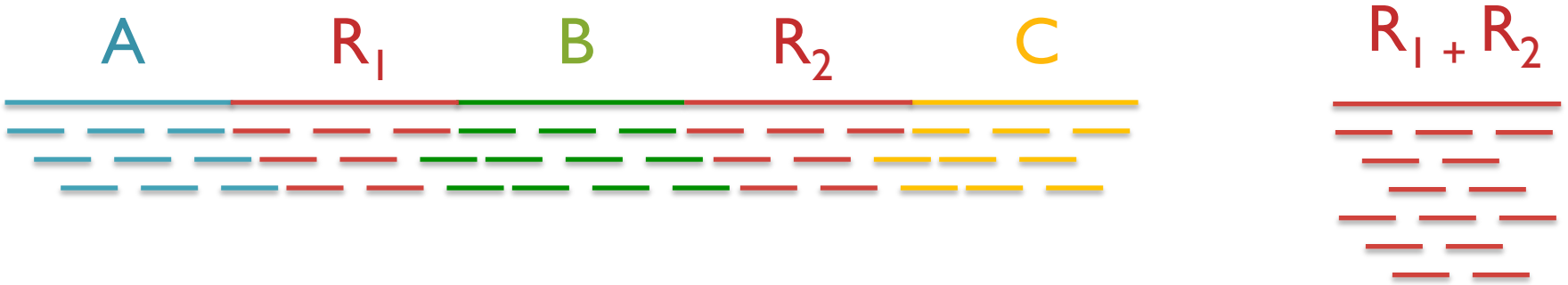
Clip Tips	Pop Bubbles
was the worst of times, was the worst of t**y**mes, the worst of times, it	was the worst of times, was the worst of t**y**mes, times, it was the age t**y**mes, it was the age
	

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.

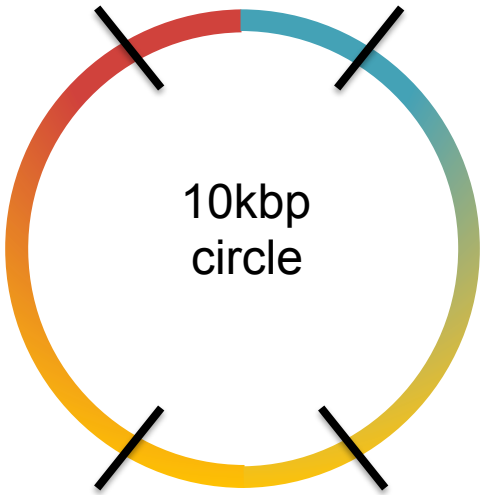
Paired-end and Mate-pairs

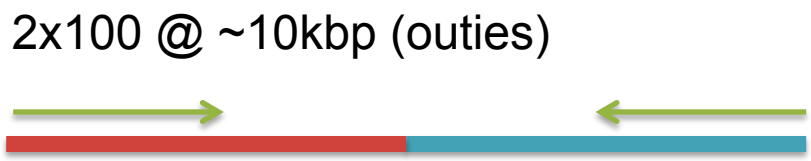
Paired-end sequencing

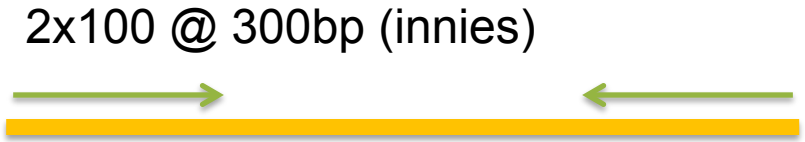
- Read one end of the molecule, flip, and read the other end
- Generate pair of reads separated by up to 500bp with inward orientation

Mate-pair sequencing

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

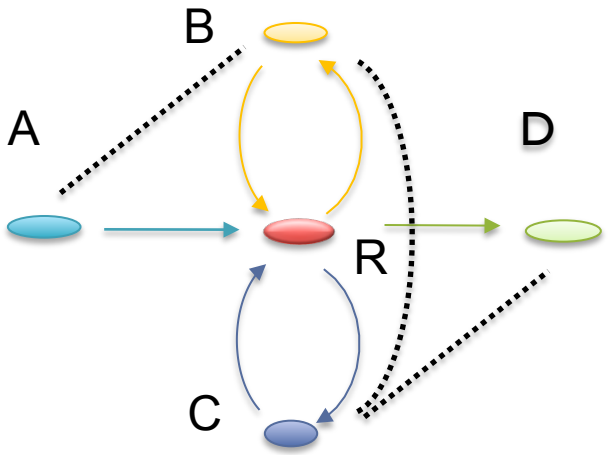




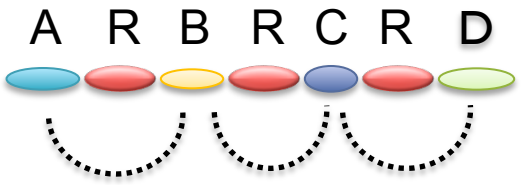


Scaffolding

- 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

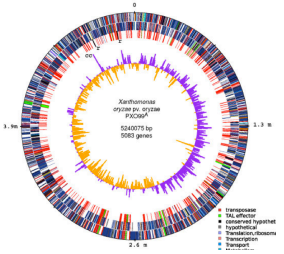
Assemblathon Results

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

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

Assemblathon 1: A competitive assessment of de novo short read assembly methods
Earl et al. (2011) Genome Research. 21: 2224-2241

Assembly Summary



Assembly quality depends on

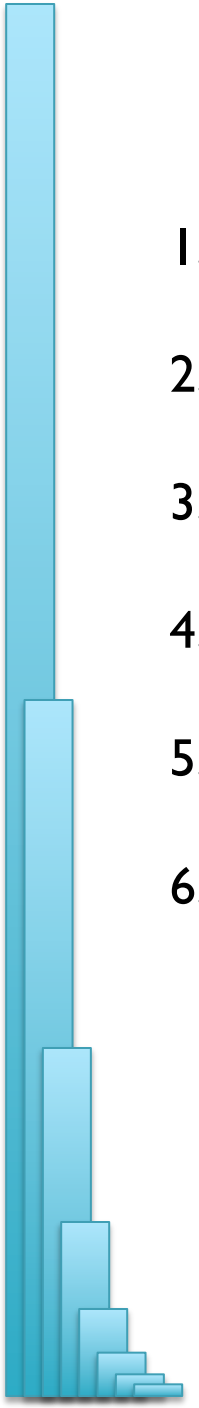
1. ***Coverage***: low coverage is mathematically hopeless
2. ***Repeat composition***: high repeat content is challenging
3. ***Read length***: longer reads help resolve repeats
4. ***Error rate***: errors reduce coverage, obscure true overlaps

- Assembly is a hierarchical, starting from individual reads, build high confidence contigs/unitigs, incorporate the mates to build scaffolds
  - Extensive error correction is the key to getting the best assembly possible from a given data set

- Watch out for collapsed repeats & other misassemblies
  - Globally/Locally reassemble data from scratch with better parameters & stitch the 2 assemblies together

Next Steps

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




Welcome to Applied Comparative Genomics
https://github.com/schatzlab/appliedgenomics

Questions?