Evaluating Assembly Quality

Michael Schatz

Dec 8, 2014

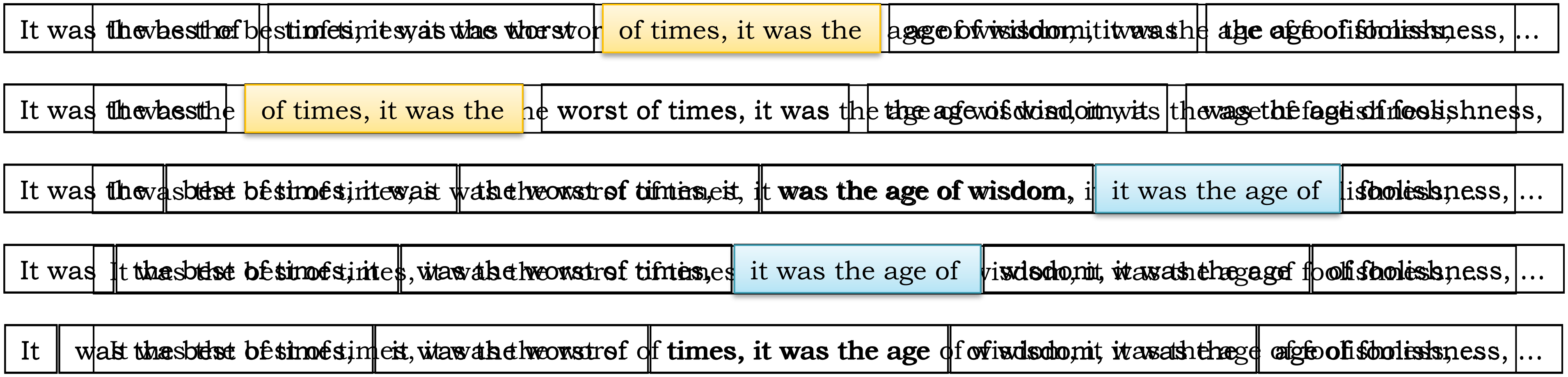
USDA/ARS Workshop

Outline

1. Assembly review
   1. Assembly by analogy
   2. Causes of Mis-assemblies
2. Evaluating Assembly Quality
   1. Assemblathon
   2. Size Statistics
   3. Mate-pair Happiness
   4. CEGMA
3. RNA-seq specific challenges

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,

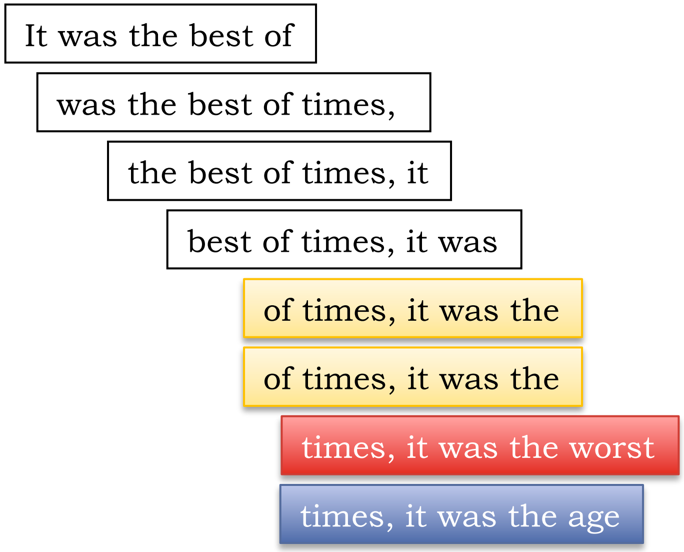
was the age of foolishness,

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

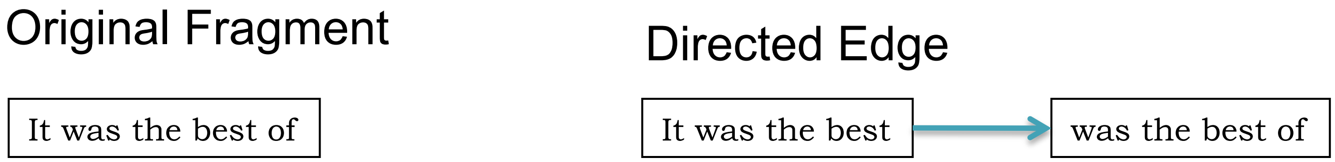
The repeated sequence make the correct reconstruction ambiguous

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

Model the assembly problem as a graph problem

de Bruijn Graph Construction

- $D_k = (V,E)$
  - V = All length-k subfragments ($k < l$)
  - E = Directed edges between consecutive subfragments
    - Nodes overlap by k-1 words



Original Fragment

It was the best of

Directed Edge

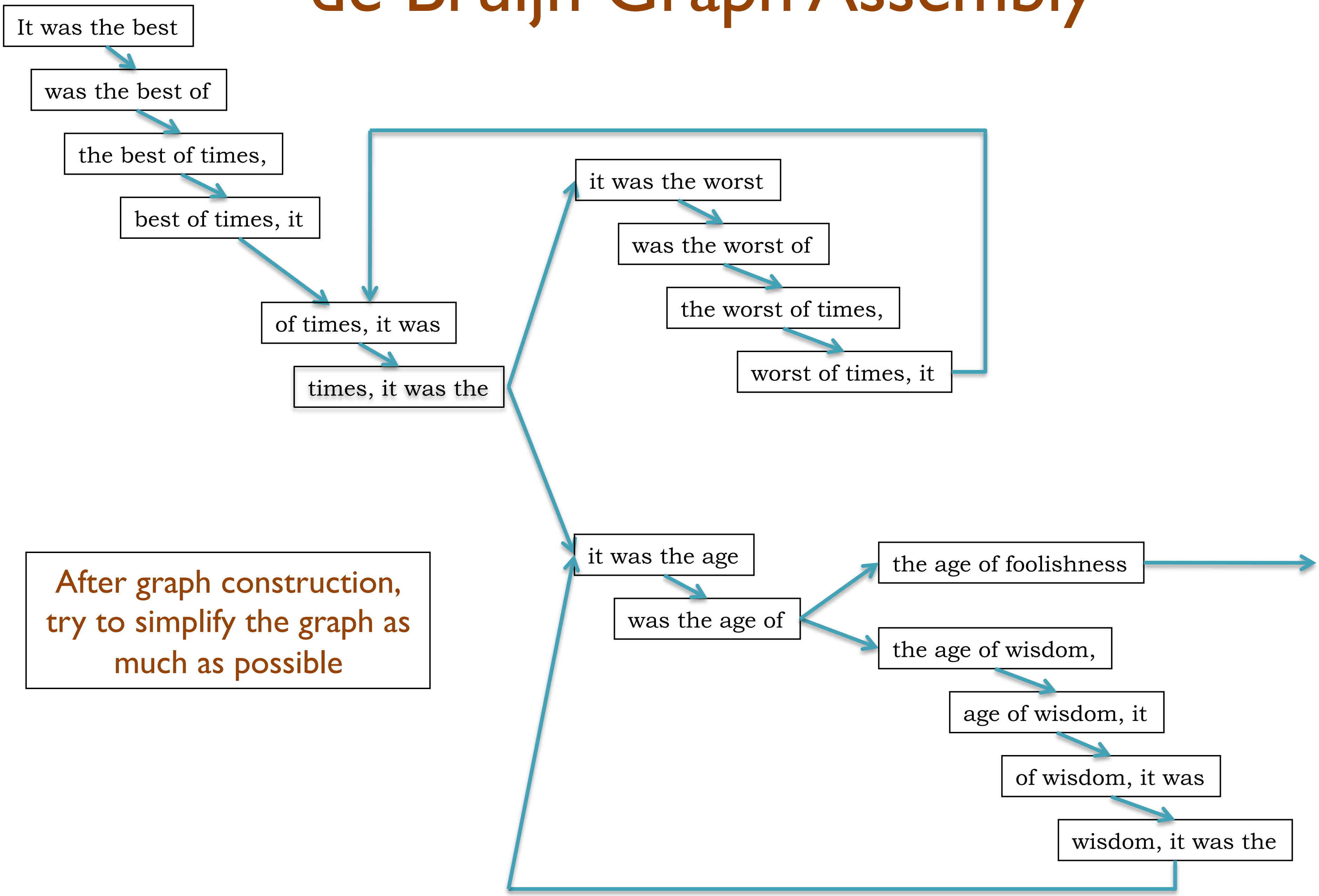
It was the best

was the best of

- Locally constructed graph reveals the global sequence structure
  - Overlaps between sequences implicitly computed

de Bruijn, 1946
Idury and Waterman, 1995
Pevzner, Tang, Waterman, 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,

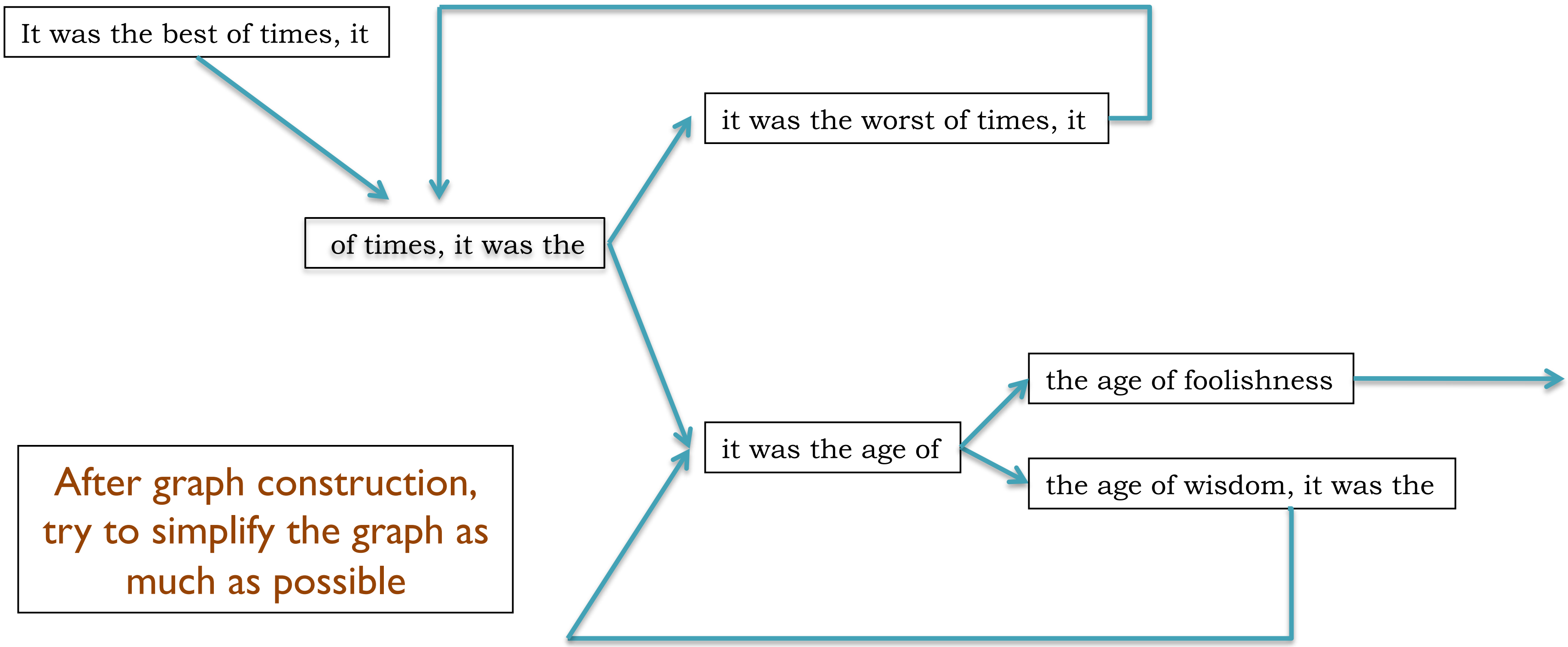
age of wisdom, it

of wisdom, it was

wisdom, it was the

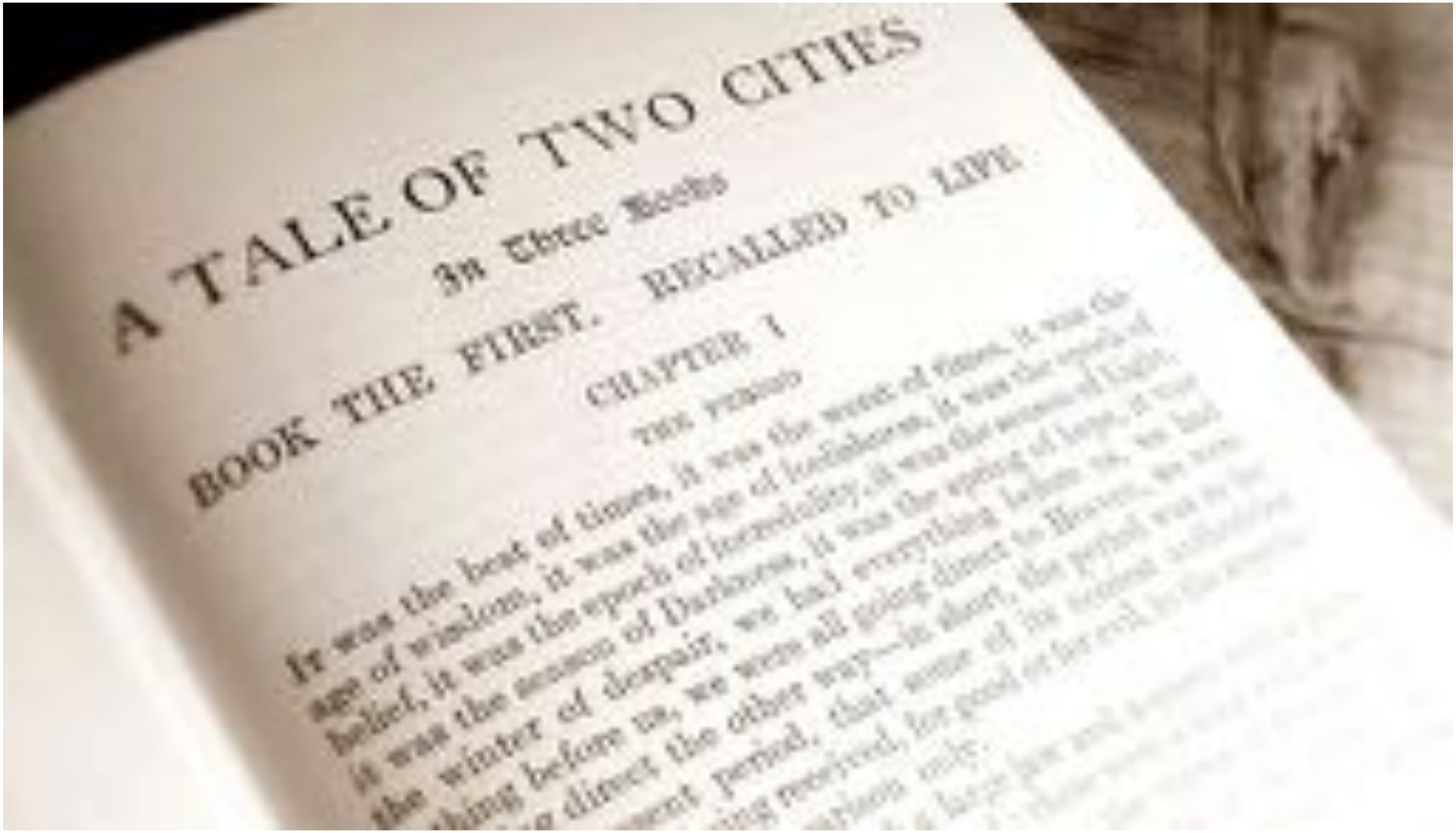
After graph construction, try to simplify the graph as much as possible

de Bruijn Graph Assembly


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

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

de Bruijn Graph Assembly



A TALE OF TWO CITIES

In Three Books

BOOK THE FIRST. RECALLED TO LIFE

CHAPTER I

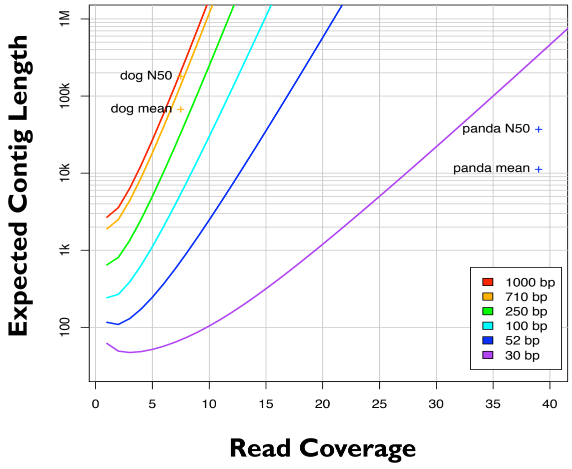
THE PERIOD

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 winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way—in short, the period was so far like the present period, that some of its noisiest authorities insisted on its being received, for good or for evil, in the superlative degree of comparison only.

There were a king with a large jaw and a queen with a plain face,

Ingredients for a good assembly

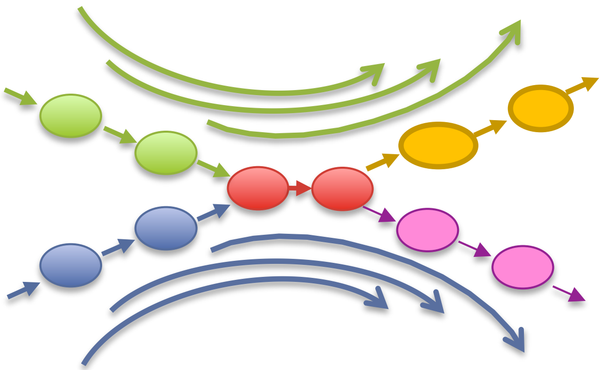
Coverage



High coverage is required

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

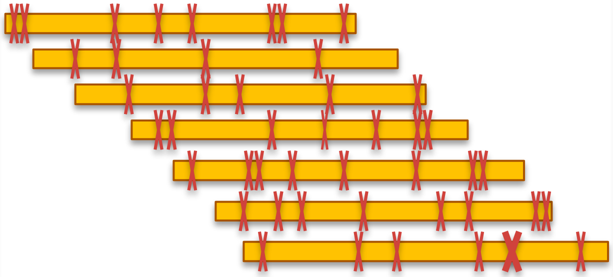
Read Length



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

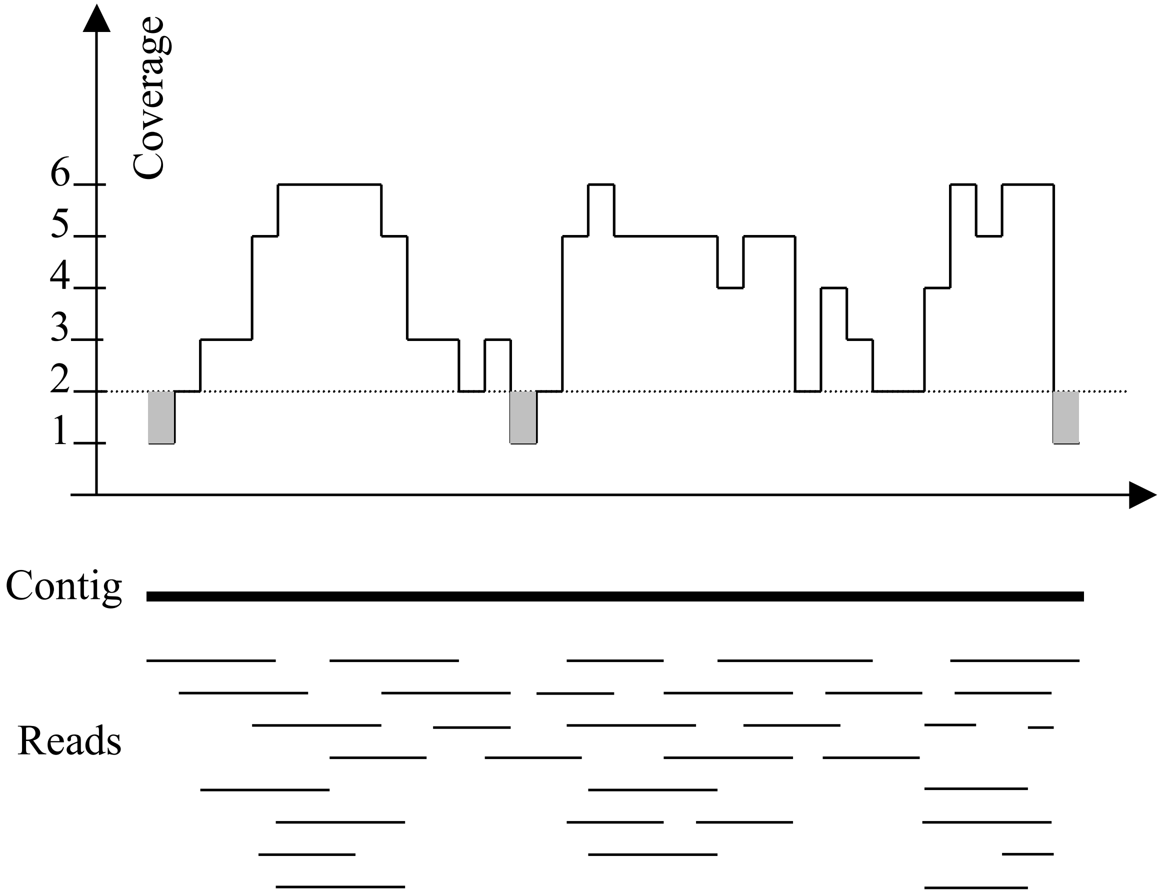


Errors obscure overlaps

- 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

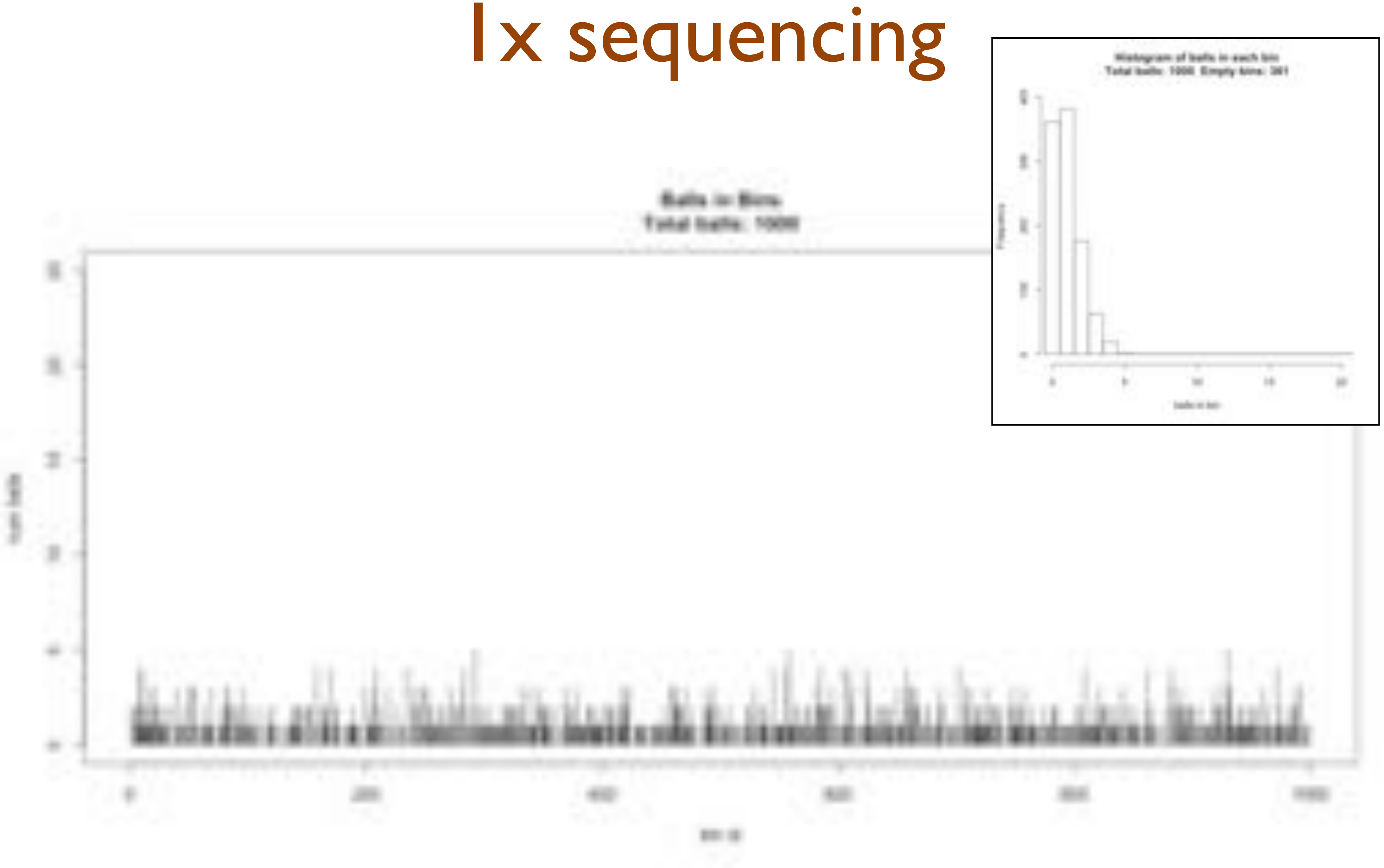
Typical sequencing coverage



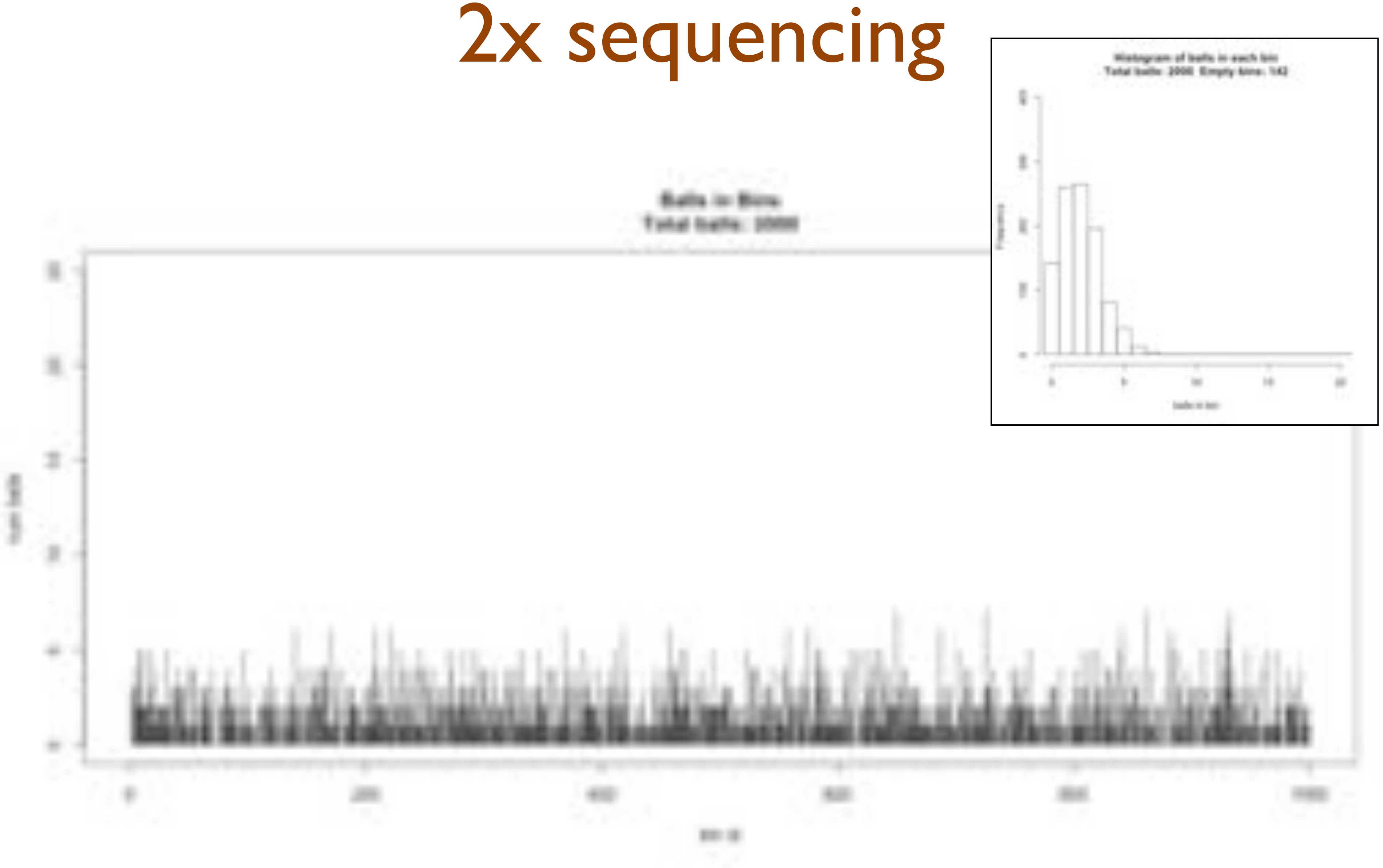
Imagine raindrops on a sidewalk

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

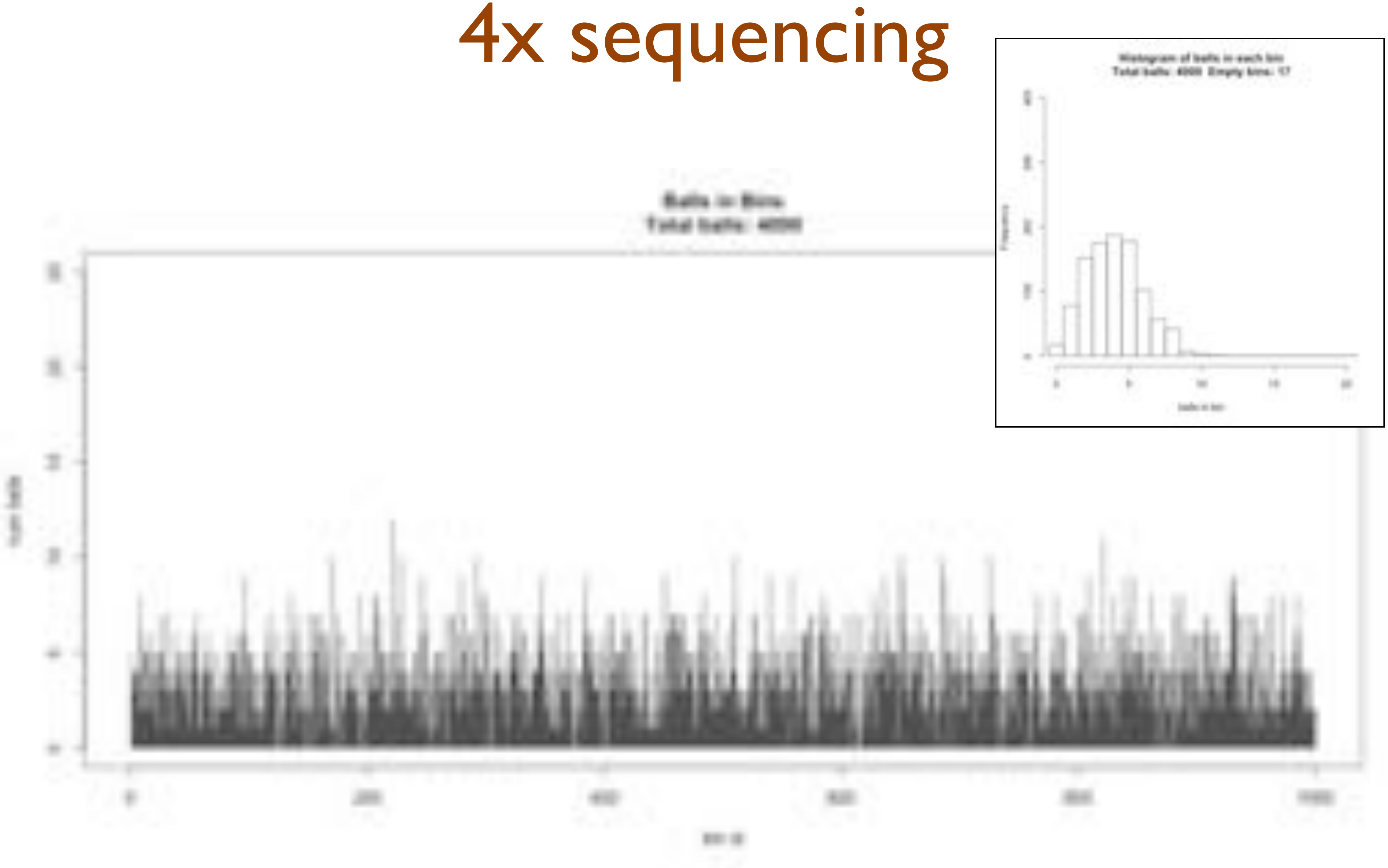
1x sequencing



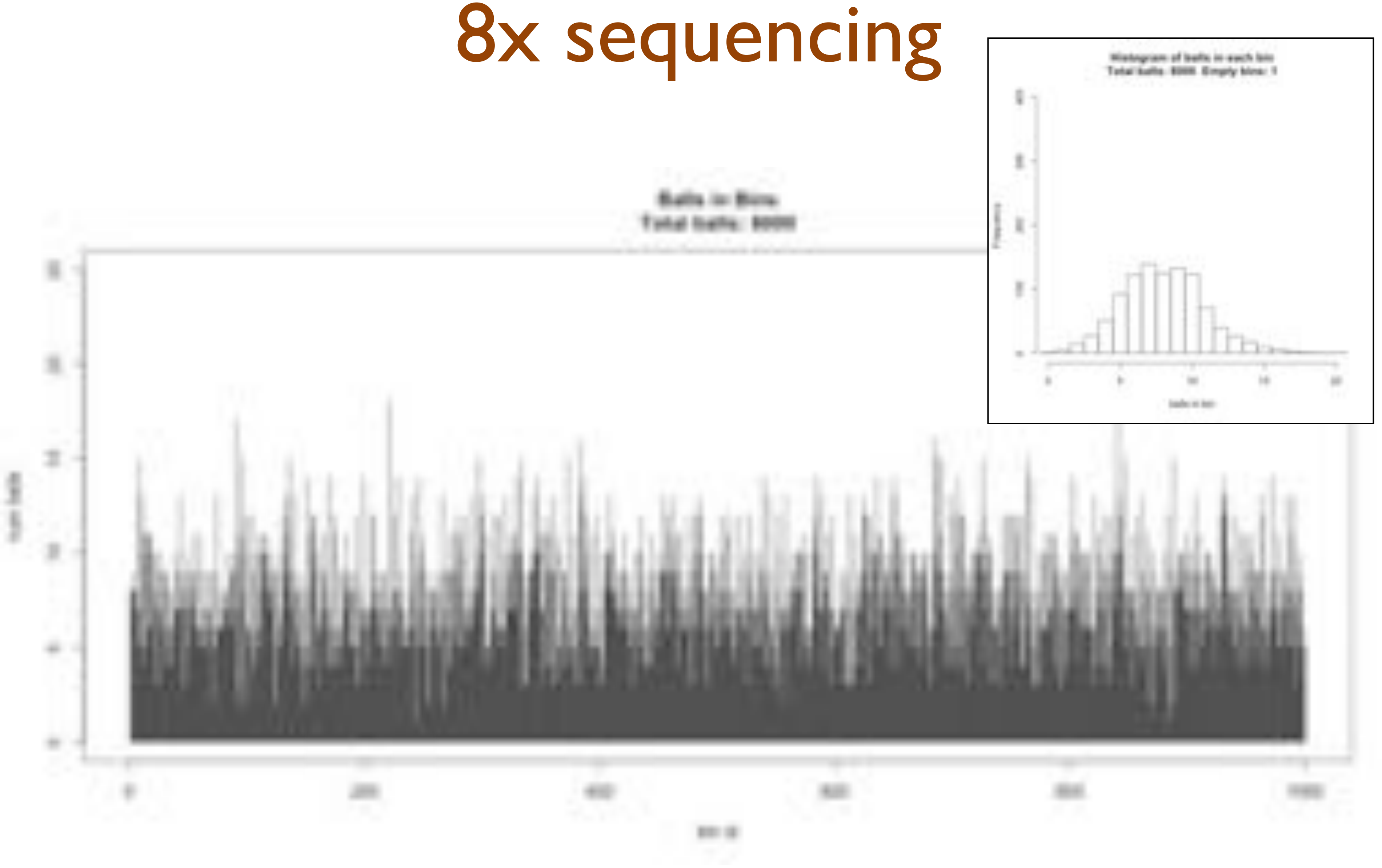
2x sequencing



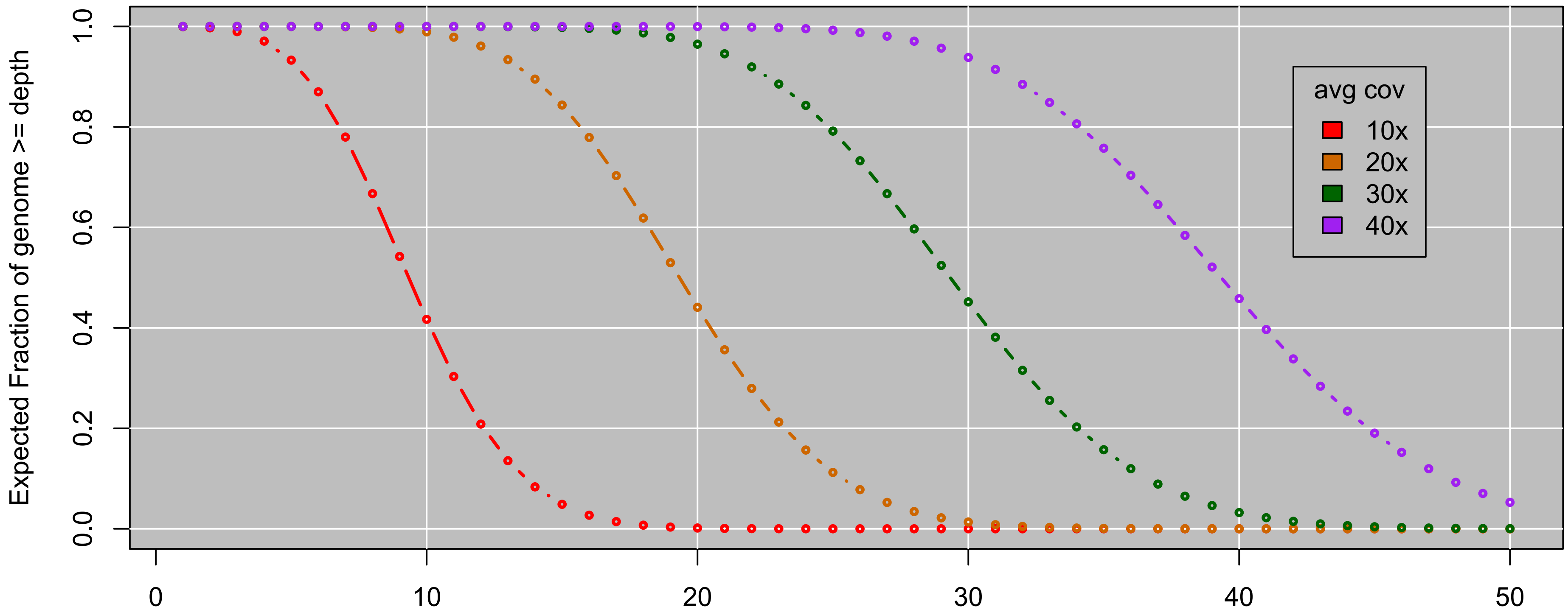
4x sequencing



8x sequencing



Genome Coverage Distribution



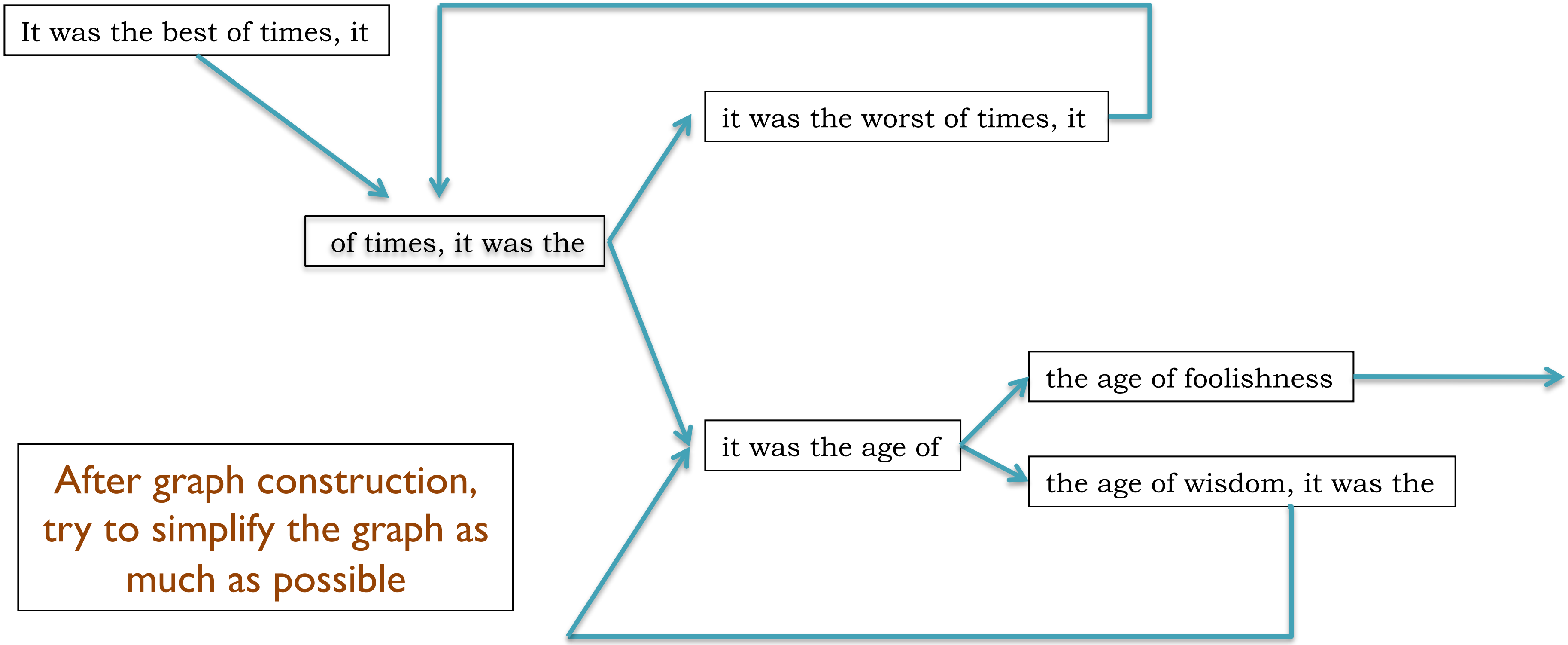
Expect Poisson distribution on depth

- Standard Deviation = sqrt(cov)

This is the mathematically model => reality may be much worse

- Double your coverage for diploid genomes
- Can use somewhat lower coverage in a population to find common variants

de Bruijn Graph Assembly



It was the best of times, it

of times, it was the

it was the worst of times, it

it was the age of

the age of foolishness

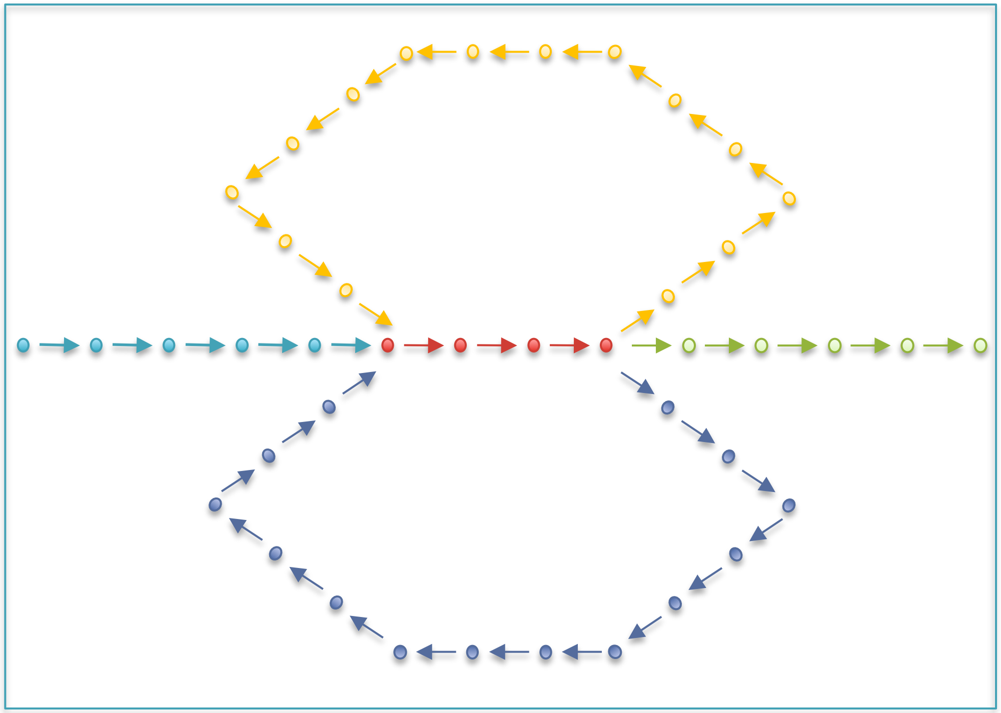
the age of wisdom, it was the

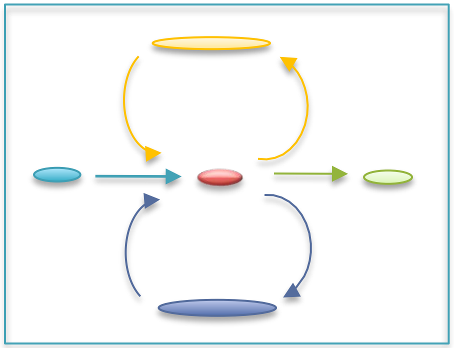
After graph construction, try to simplify the graph as much as possible

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/isoform differences, and (4) repeats





Repetitive regions

Repeat Type	Definition / Example	Prevalence
Low-complexity DNA / Microsatellites	$(b_1b_2\ldots b_k)^N$ where $1 \le k \le 6$ 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

The full tale

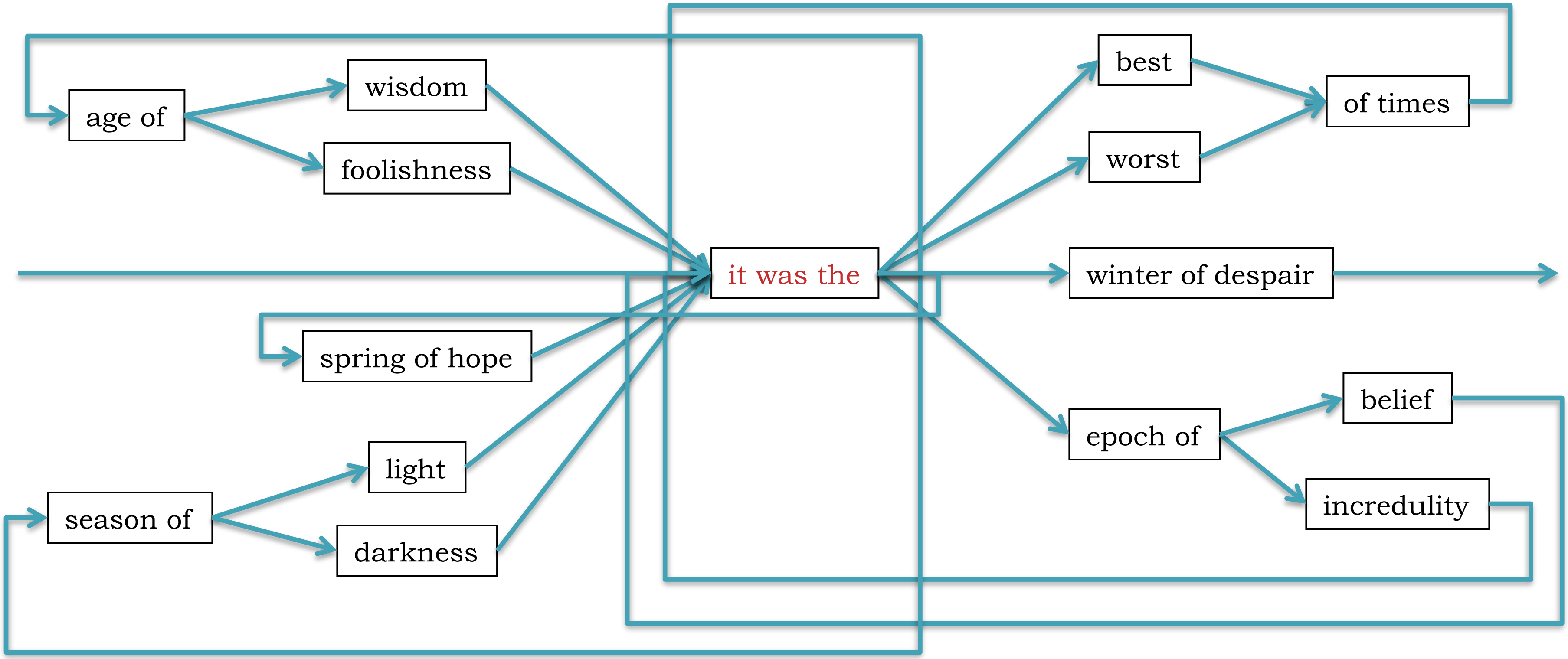
… it was the best of times it was the worst of times …

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

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

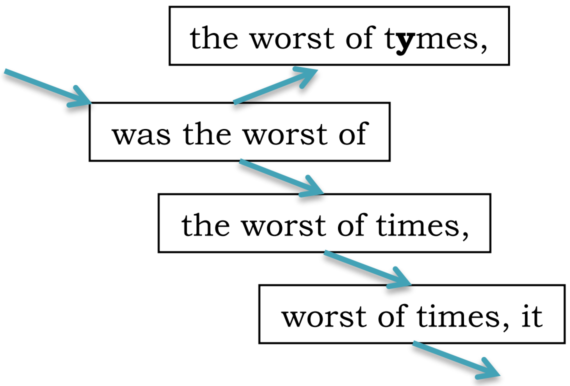
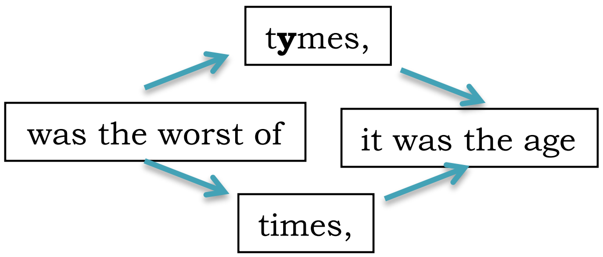
… it was the season of light it was the season of darkness …

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



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
 the worst of t**y**mes, was the worst of the worst of times, worst of times, it	 t**y**mes, was the worst of it was the age times,

Outline

1. Assembly review
    1. Assembly by analogy
    2. Causes of Mis-assemblies
2. Evaluating Assembly Quality
    1. Assemblathon
    2. Size Statistics
    3. Mate-pairs
    4. CEGMA
3. RNA-seq specific challenges

THE ASSEMBLATHON

- Attempt to answer the question:
  “What makes a good assembly?”

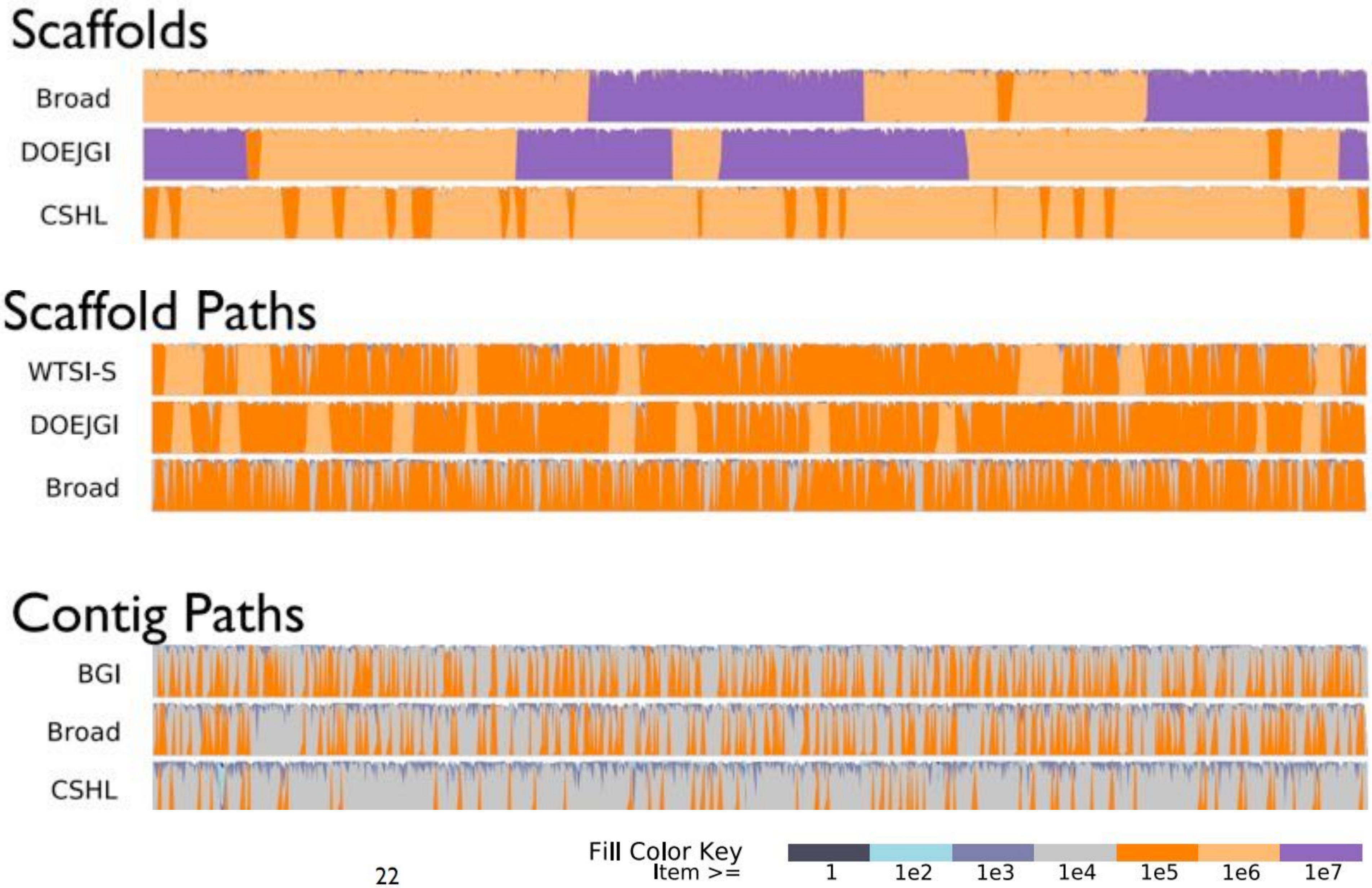
- Organizers provided sequence data to assembly experts around the world
  - Assemblathon 1: ~100Mbp simulated genome
  - Assemblathon 2: 3 vertebrate genomes each ~1GB

- Results demonstrate trade-offs assemblers must make

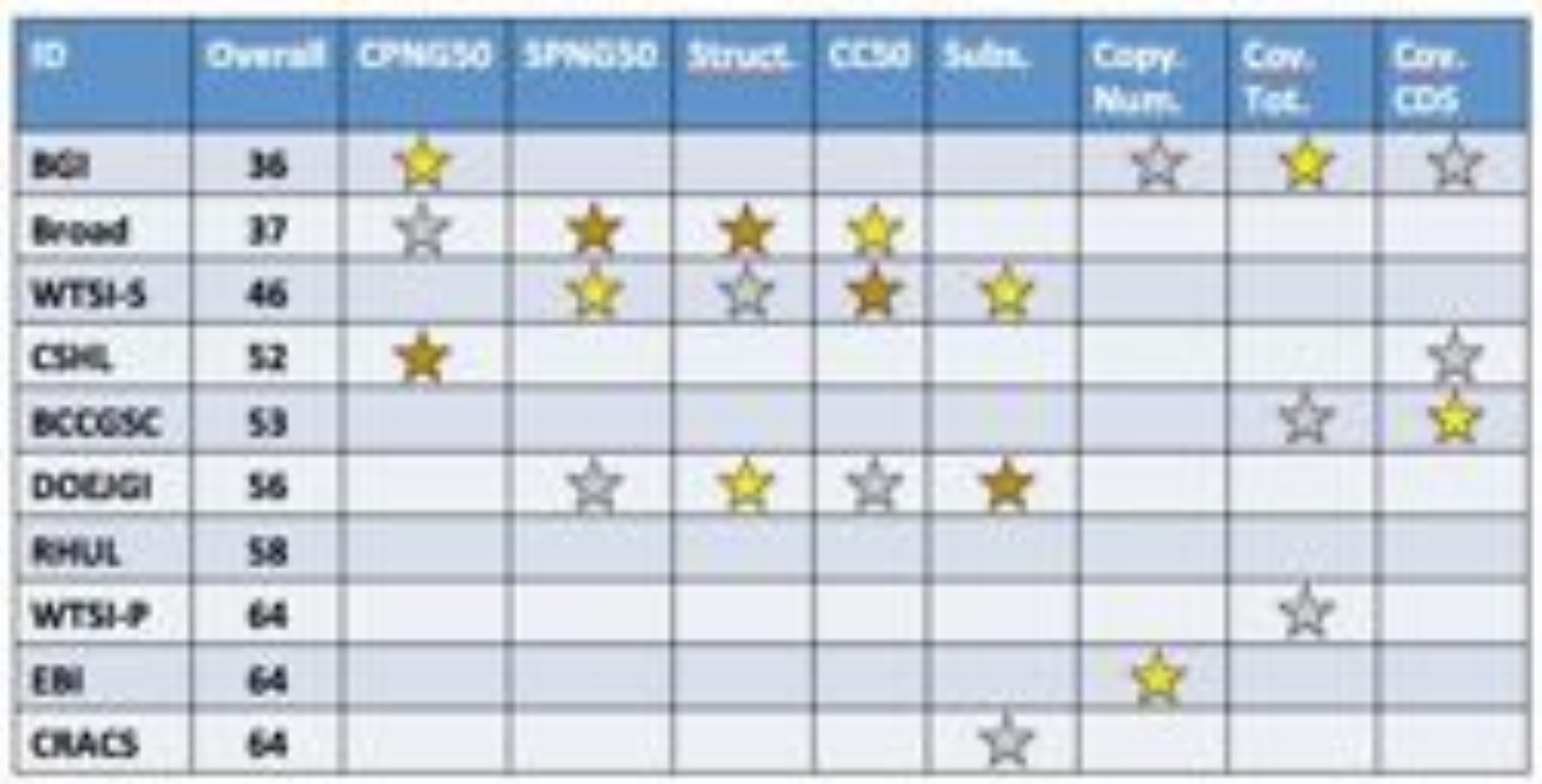
Assemblathon 1: A competitive assessment of de novo short read assembly methods.
Earl, DA, et al. (2011) Genome Research. doi: 10.1101/gr.126599.111

Assemblathon 2: Evaluating de novo methods of genome assembly in three vertebrate species
Bradnam, KR. et al (2013) GigaScience 2:10 doi:10.1186/2047-217X-2-10

Assembly Results



Final Rankings



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								
WTSI-P	64							🥈	
EBI	64						🥇		
CRACS	64					🥈			

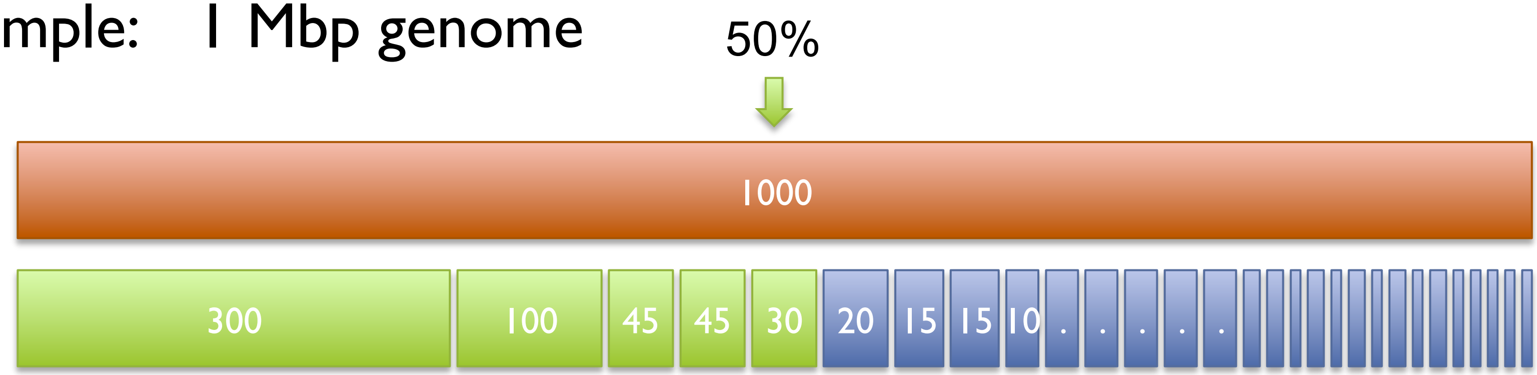
- ALLPATHS and SOAPdenovo came out neck-and-neck followed closely behind by Celera Assembler, SGA, and ABySS

- My recommendation for "typical" short read assembly is to use ALLPATHS
- Single molecule sequencing becoming extremely attractive if you have access

N50 size

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

Example: 1 Mbp genome



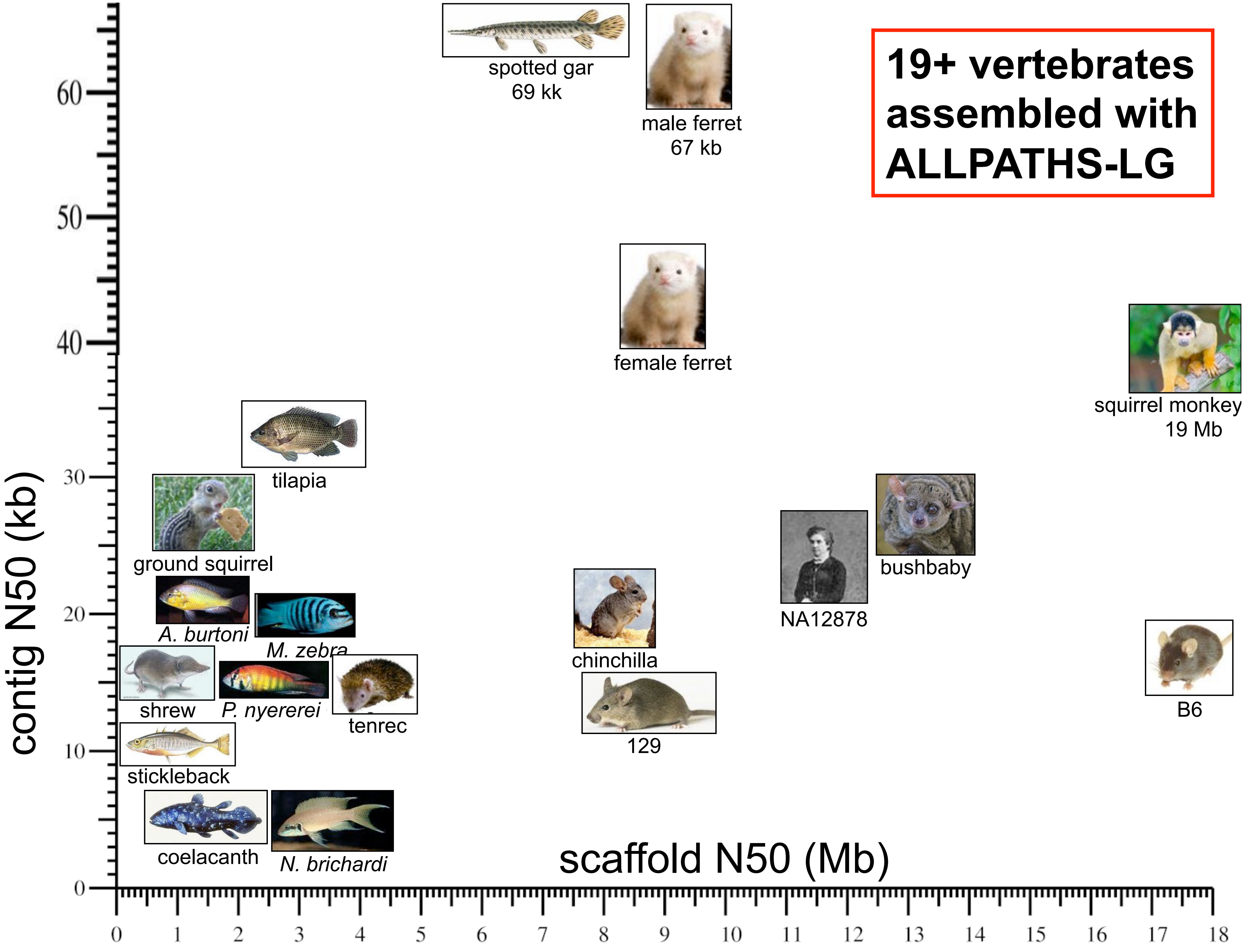
N50 size = 30 kbp

(300k+100k+45k+45k+30k = 520k >= 500kbp)

A greater N50 is indicative of improvement in every dimension:

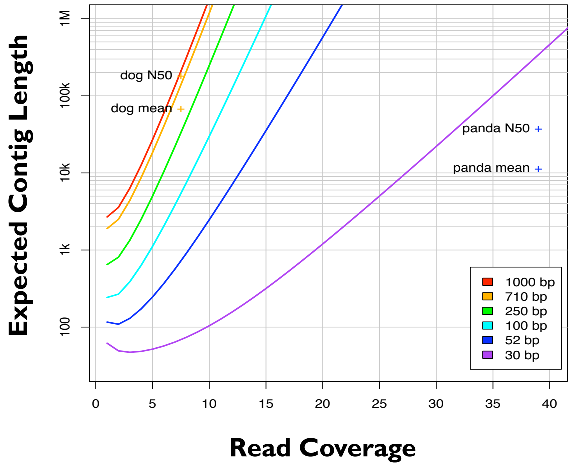
- Better resolution of genes and flanking regulatory regions
- Better resolution of transposons and other complex sequences
- Better resolution of chromosome organization
- Better sequence for all downstream analysis

19+ vertebrates assembled with ALLPATHS-LG



Ingredients for a good assembly

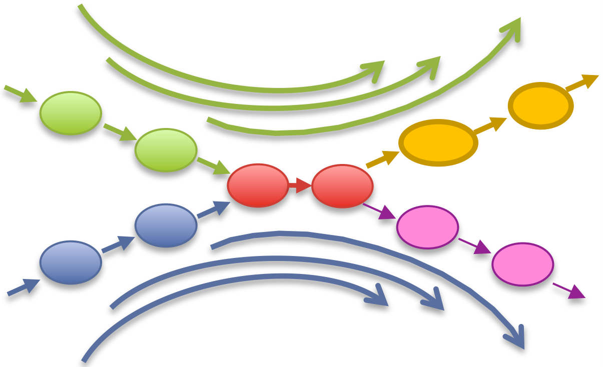
Coverage



High coverage is required

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

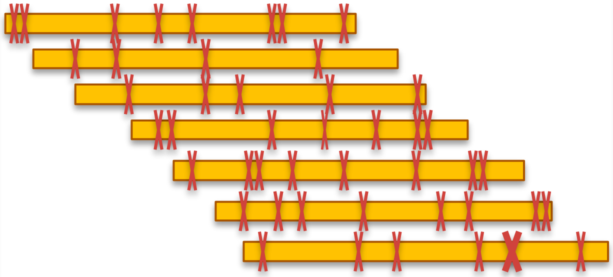
Read Length



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

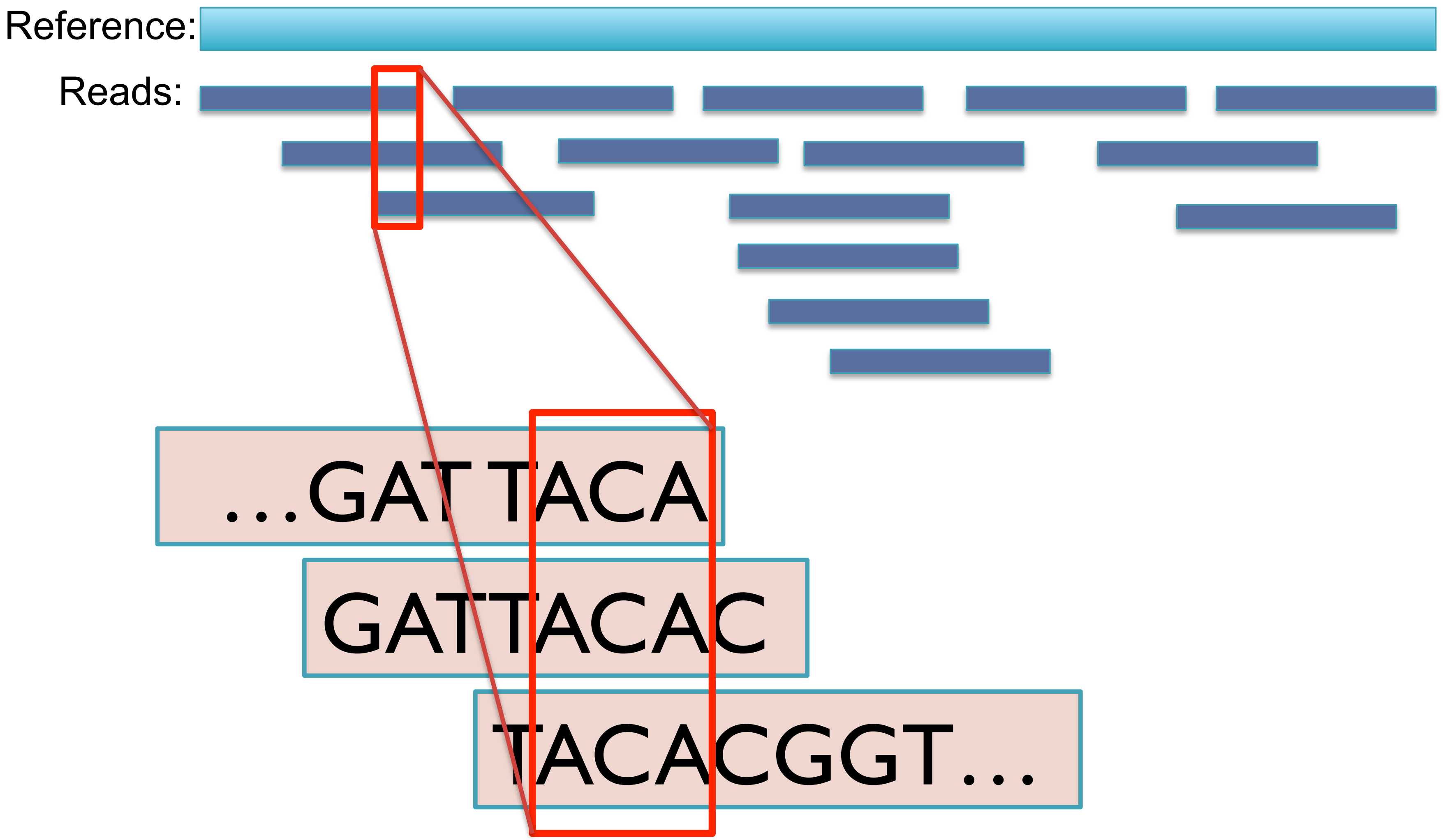


Errors obscure overlaps

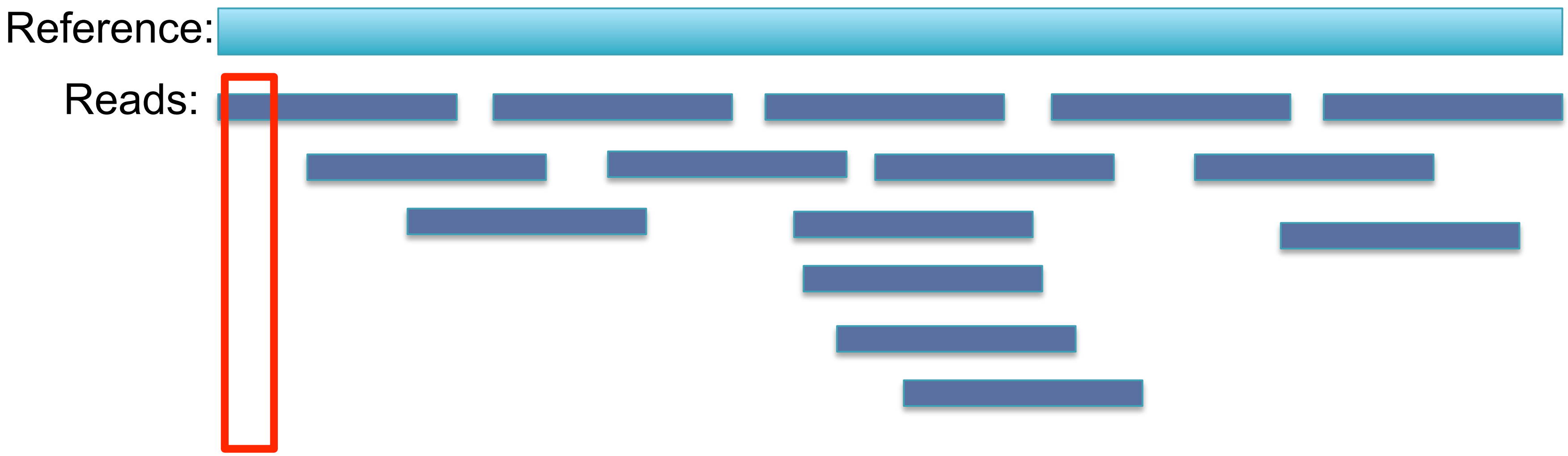
- Reads are assembled by finding kmers shared in pair of reads
- High error rate requires very short seeds, increasing complexity and forming assembly hairballs

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

Estimating coverage with Kmers

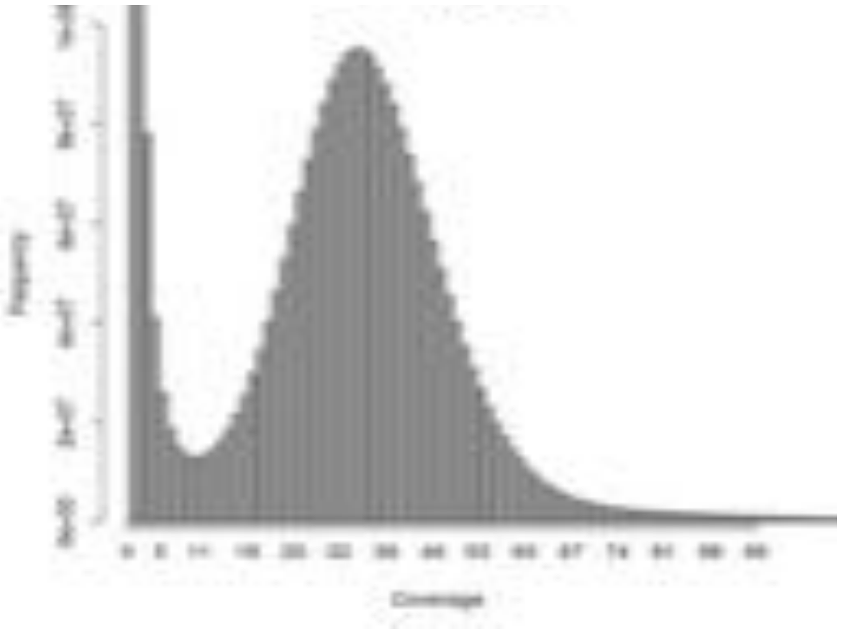


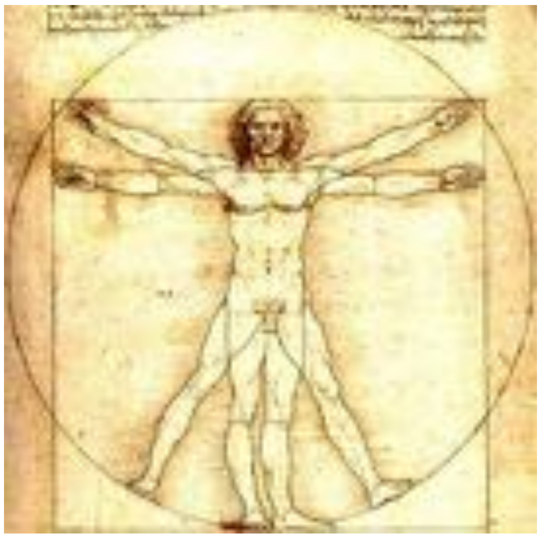
Estimating coverage with Kmers



Reference:

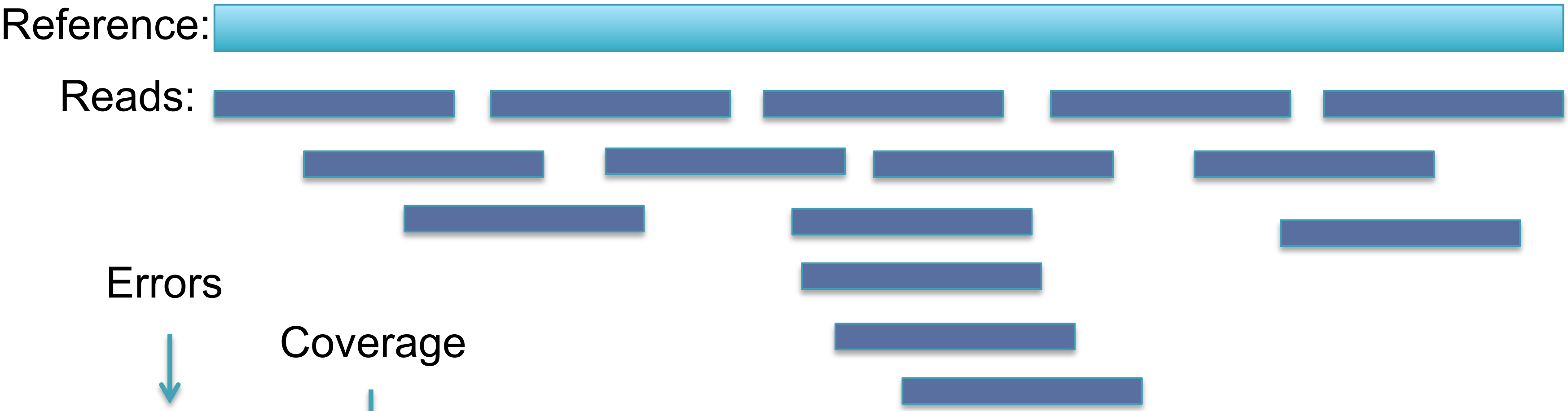
Reads:





NA12878

QC: Read Coverage

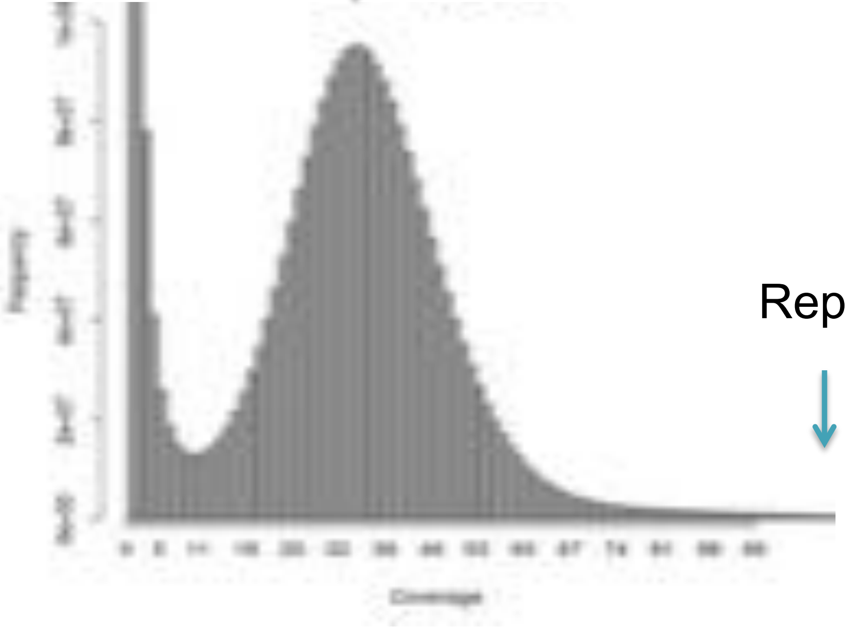

Reference:

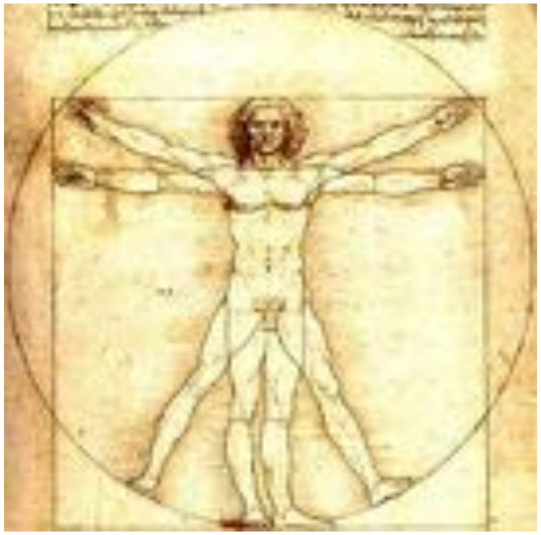
Reads:

Errors

Coverage

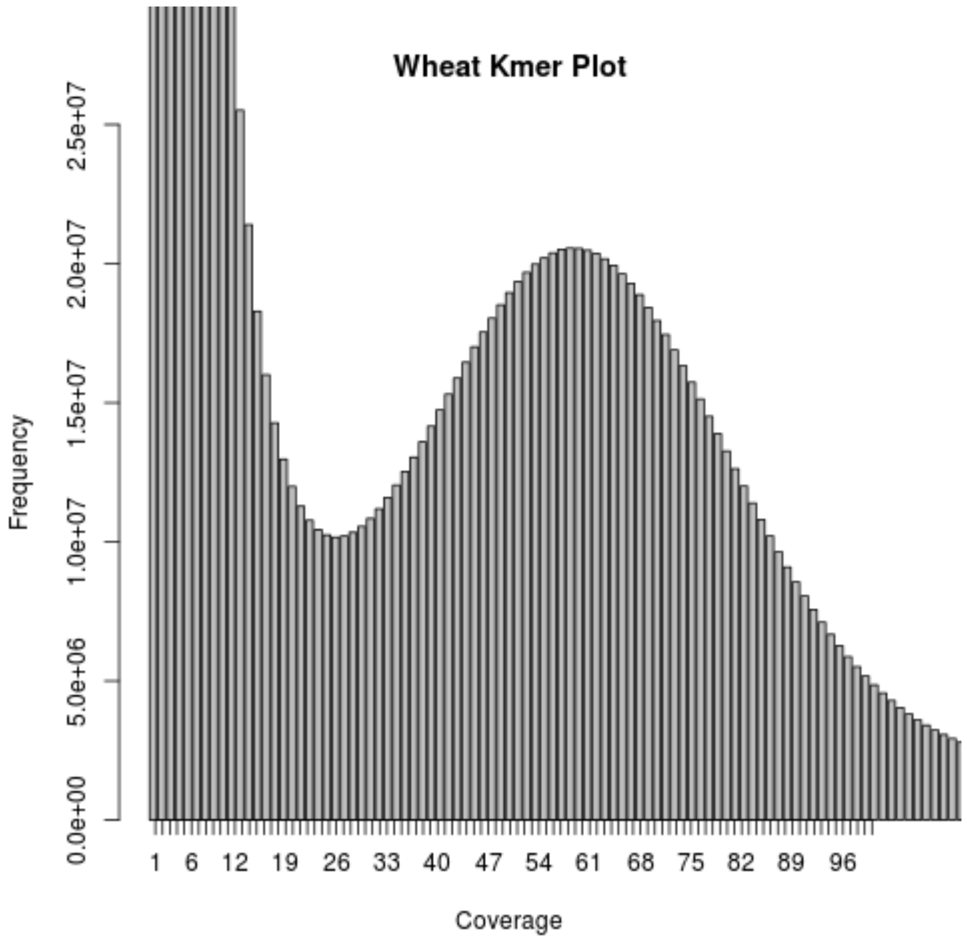
Repeats



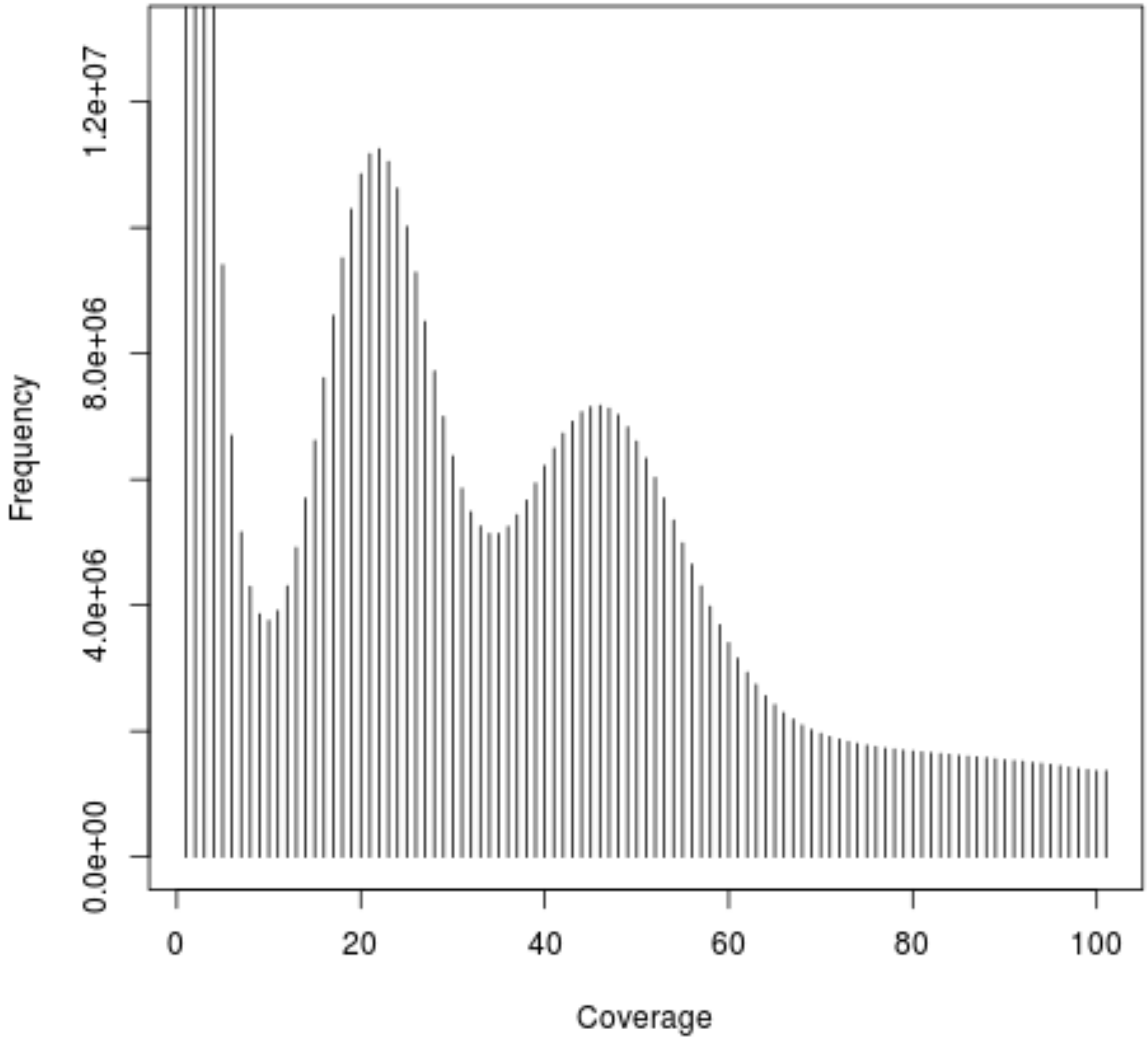


Wheat Genome

(A. tauschi / CSHL)



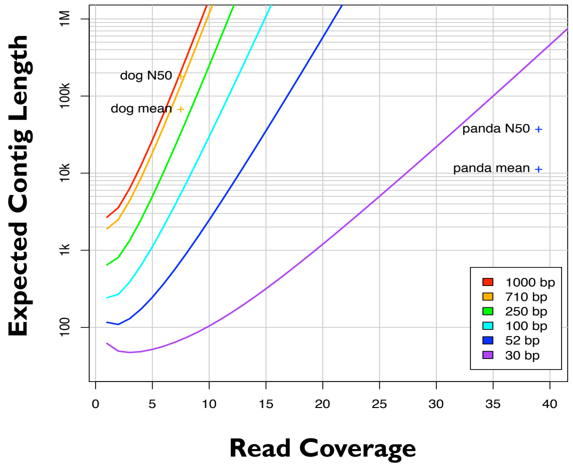
Heterozygous Genome



Contact: @mike_schatz

Ingredients for a good assembly

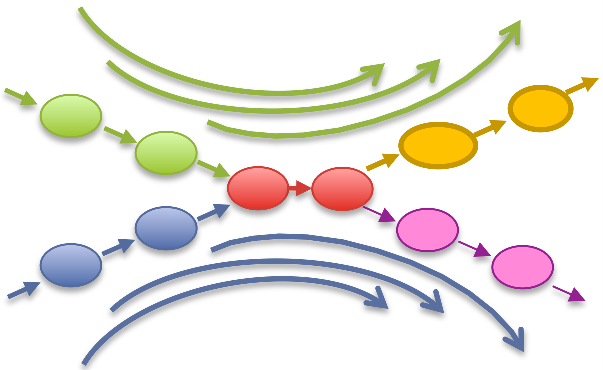
Coverage



High coverage is required

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

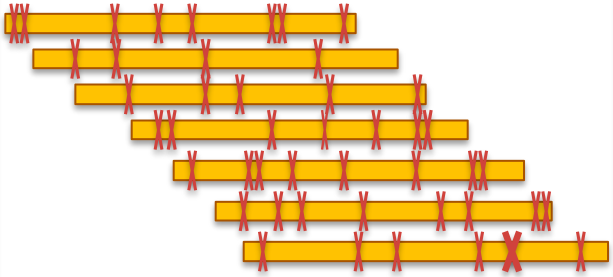
Read Length



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

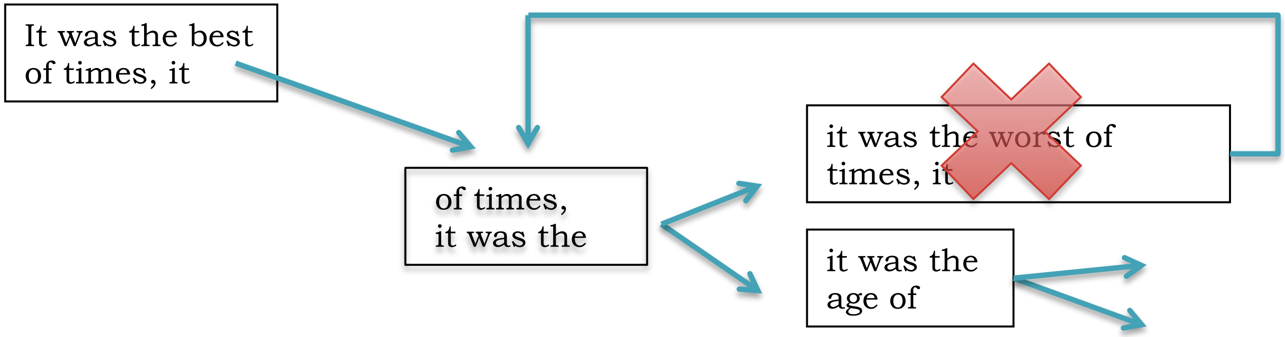
Current challenges in *de novo* plant genome sequencing and assembly
Schatz MC, Witkowski, McCombie, WR (2012) *Genome Biology*. 12:243

Assembly Validation

Automatically scan an assembly to locate misassembly signatures for further analysis and correction

Assembly-validation pipeline

1. Evaluate Mate Pairs & Libraries
2. Evaluate Read Alignments
3. Evaluate Read Breakpoints
4. Analyze Depth of Coverage



It was the best of times, it

of times, it was the

it was the worst of times, it

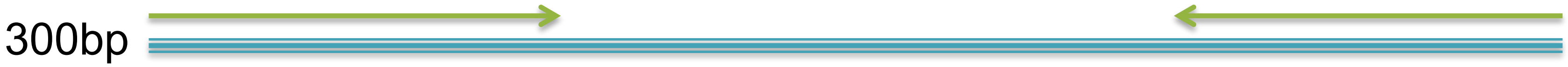
it was the age of

Genome Assembly forensics: finding the elusive mis-assembly.
Phillippy, AM, Schatz, MC, Pop, M. (2008) *Genome Biology* 9:R55.

Paired-end and Mate-pairs

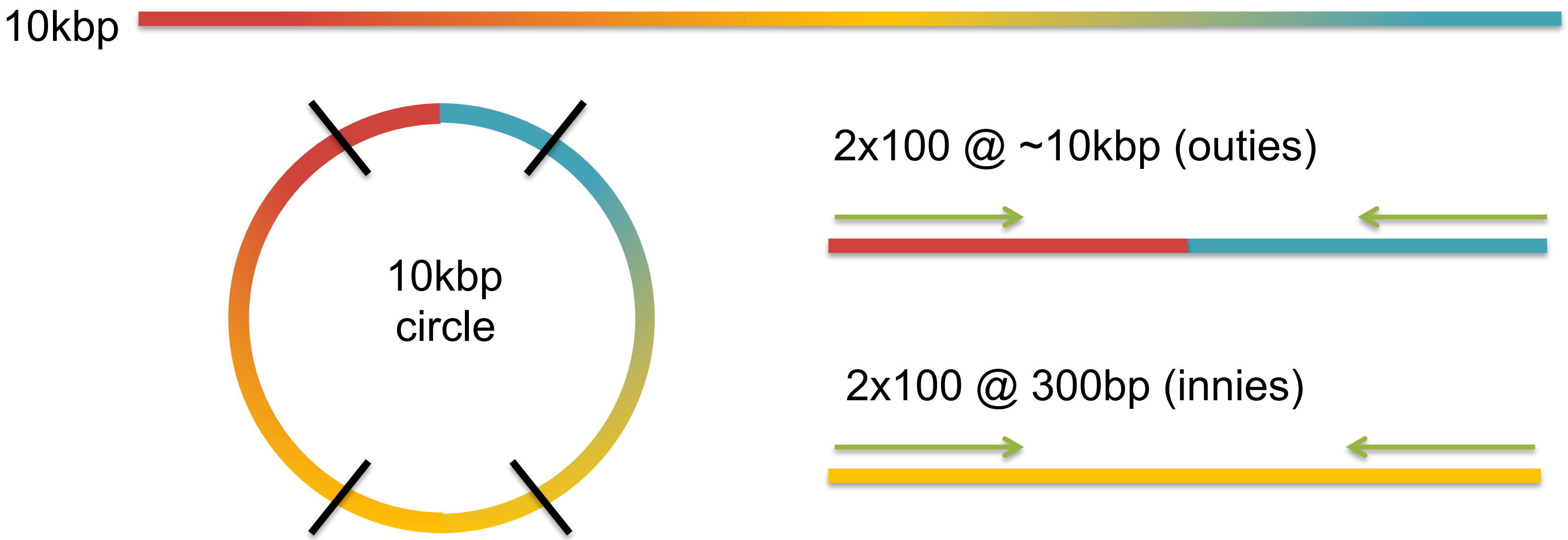
Paired-end sequencing

- 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

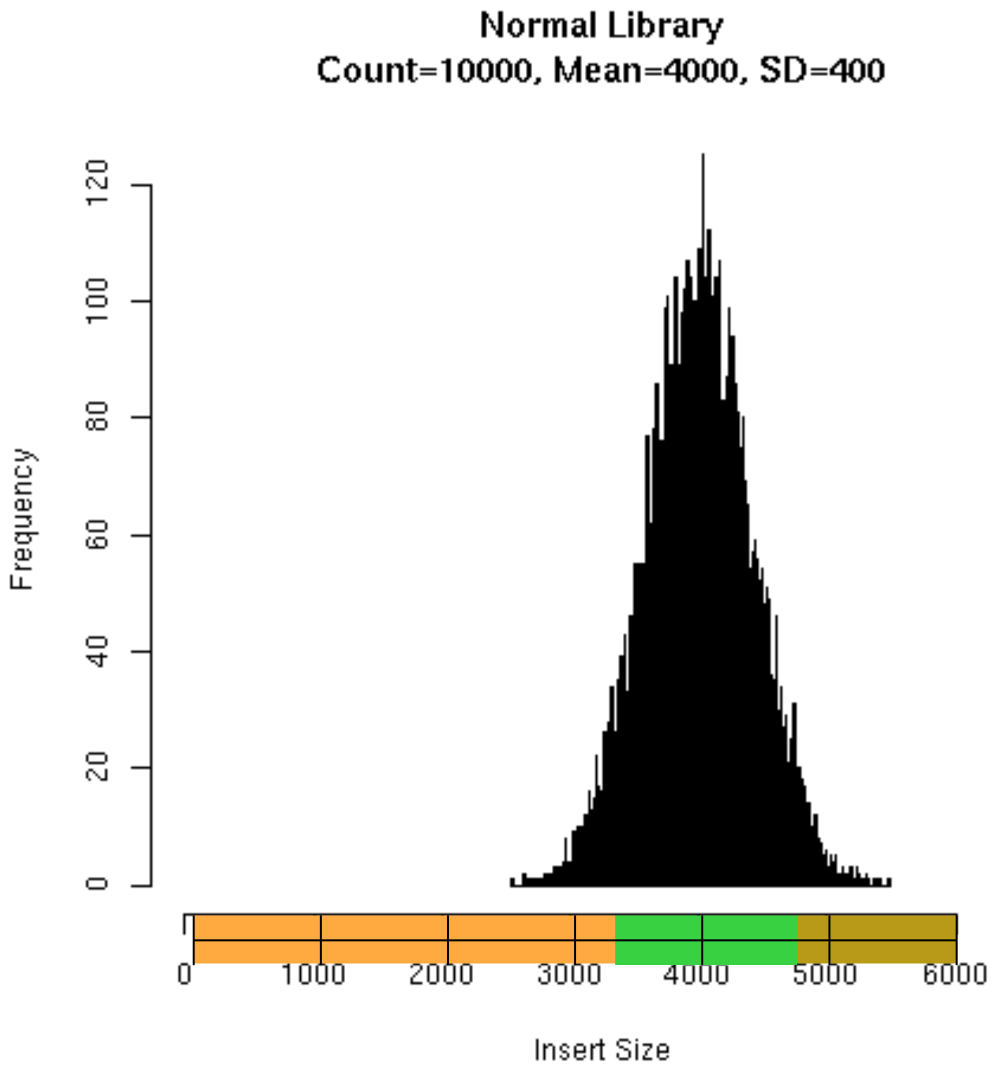


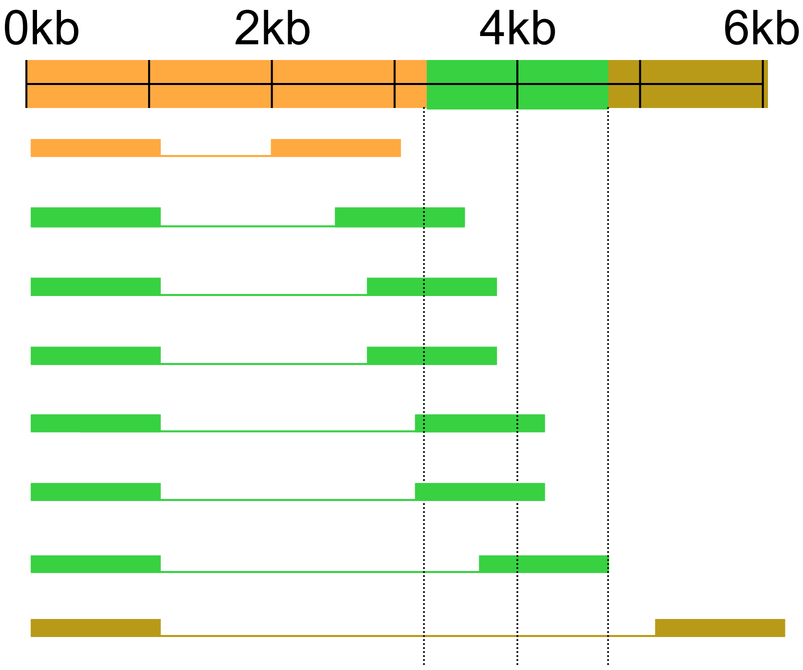
C/E Statistic

- The presence of individual compressed or expanded mates is rare but expected.
- Do the inserts spanning a given position differ from the rest of the library?
  - Flag large differences as potential misassemblies
  - Even if each individual mate is “happy”
- Compute the statistic at all positions
  - (Local Mean – Global Mean) / Scaling Factor
- Introduced by Jim Yorke’s group at UMD

Forensics

Sampling the Genome





8 inserts: 3kb-6kb

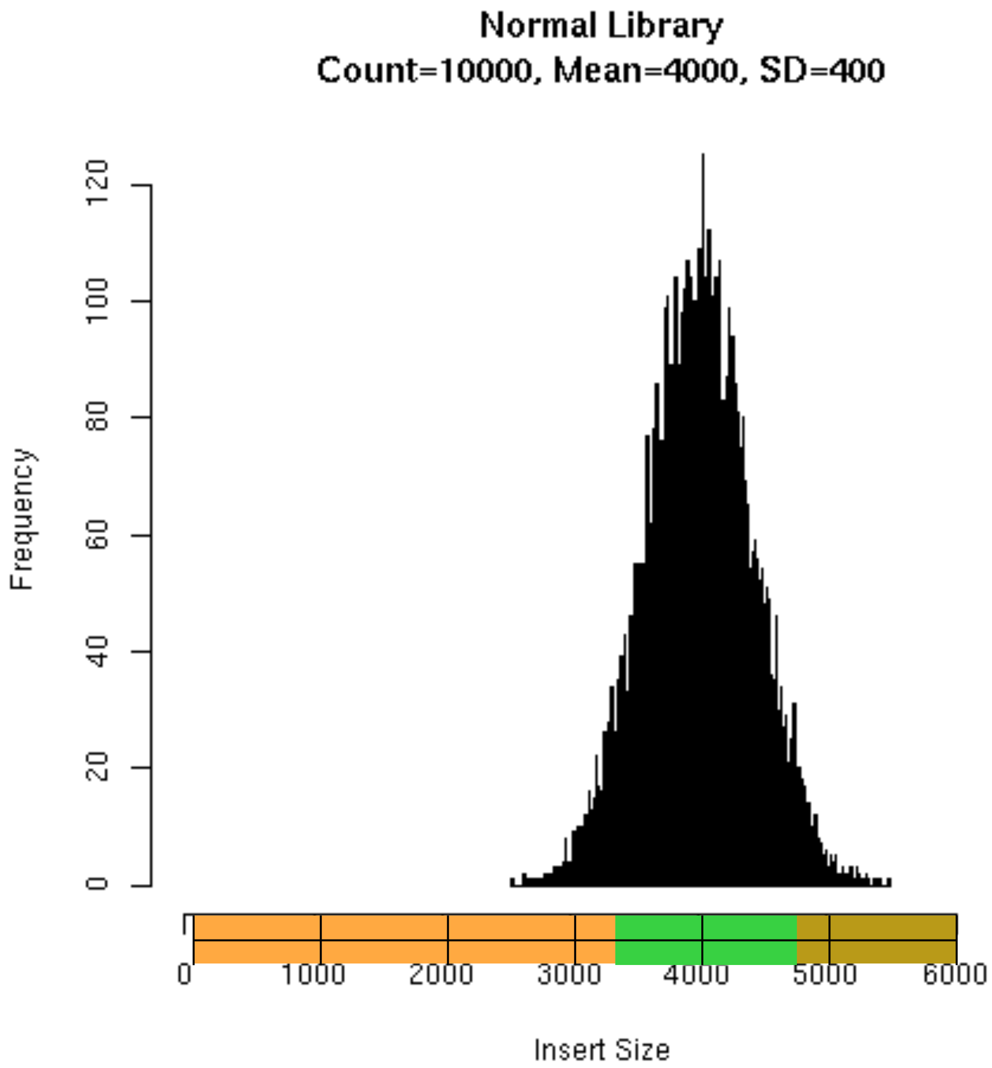
Local Mean: 4048

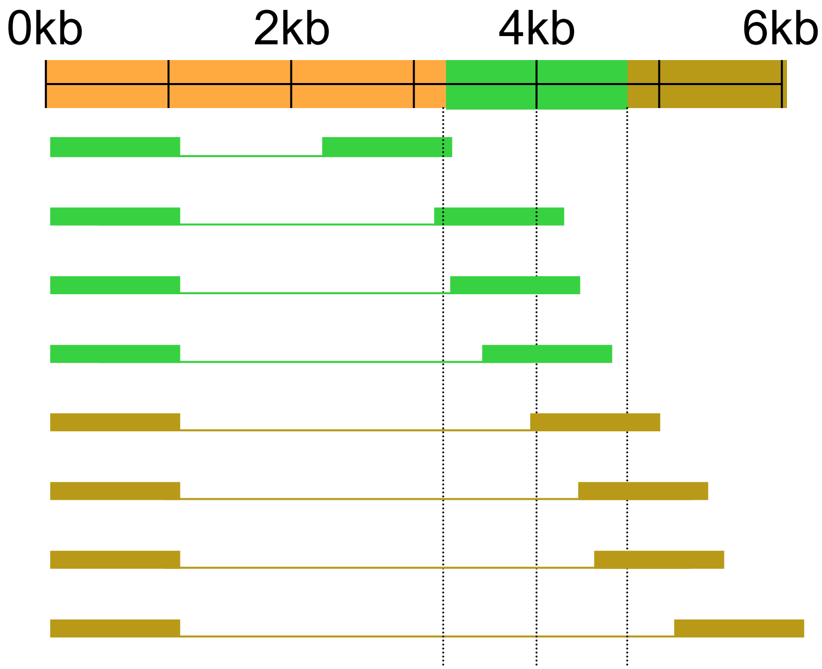
C/E Stat: $\frac{(4048-4000)}{(400 / \sqrt{8})} = +0.33$

Near 0 indicates overall happiness

Forensics

C/E-Statistic: Expansion





8 inserts: 3.2kb-6kb

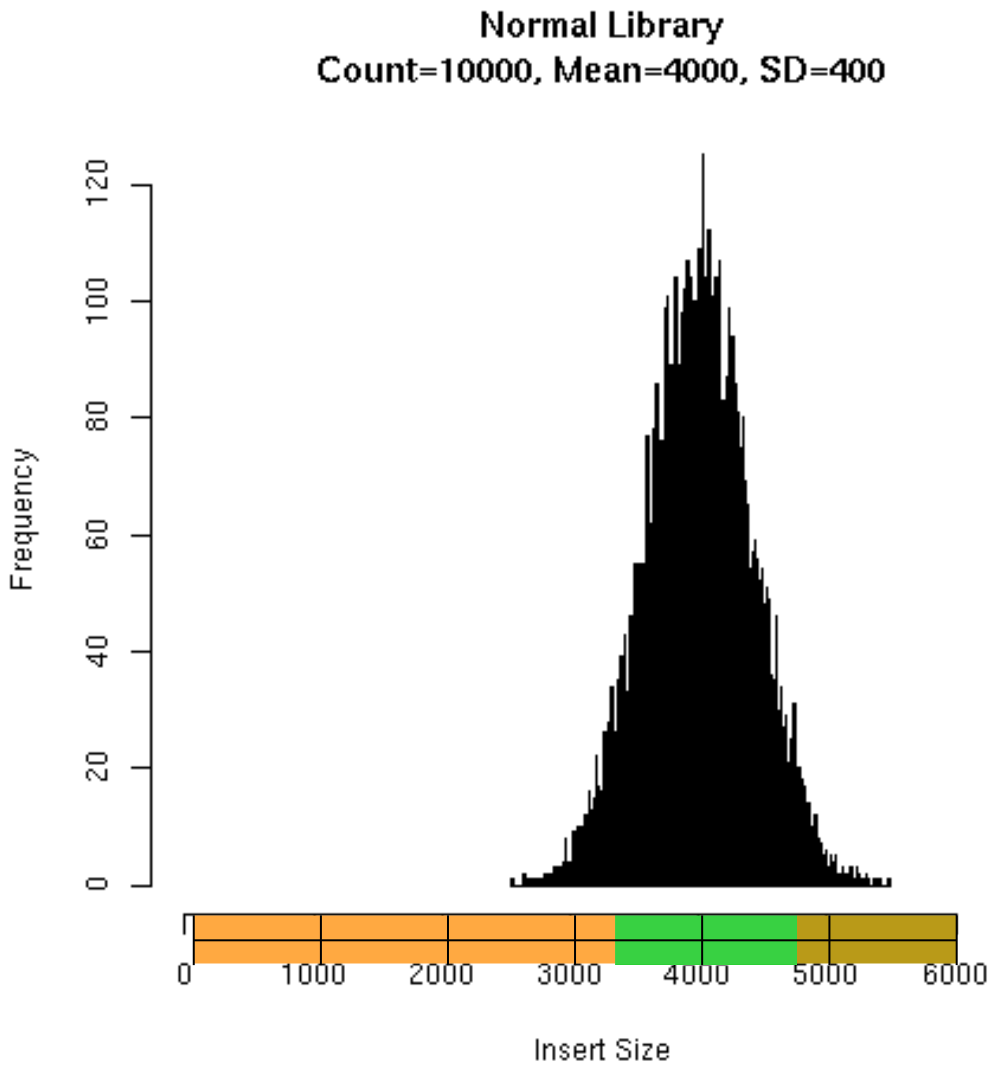
Local Mean: 4461

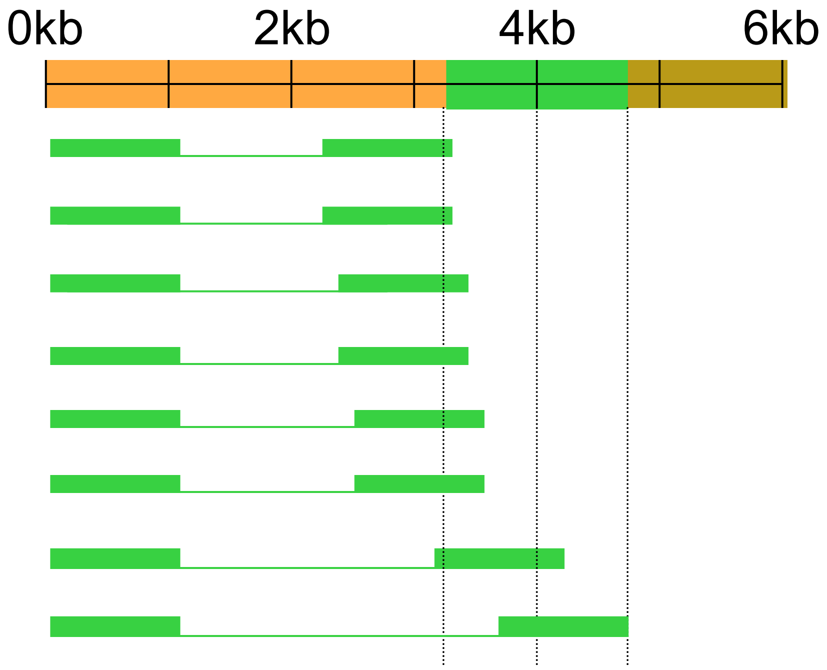
C/E Stat: $$\frac{(4461-4000)}{(400 / \sqrt{8})} = +3.26$$

C/E Stat ≥ 3.0 indicates Expansion

Forensics

C/E-Statistic: Compression





8 inserts: 3.2 kb-4.8kb

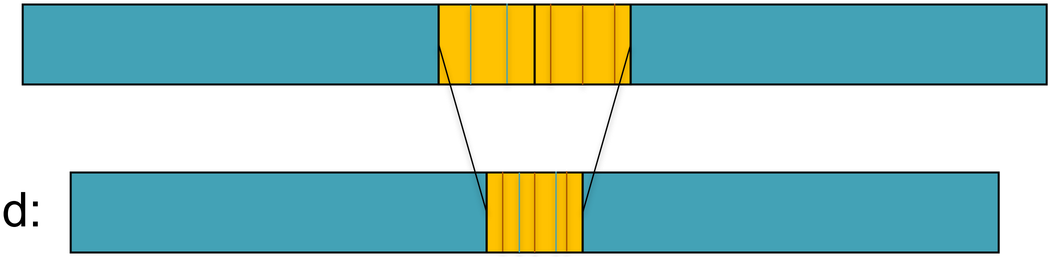
Local Mean: 3488

C/E Stat: $\frac{(3488-4000)}{(400 / \sqrt{8})} = -3.62$

C/E Stat ≤ -3.0 indicates Compression

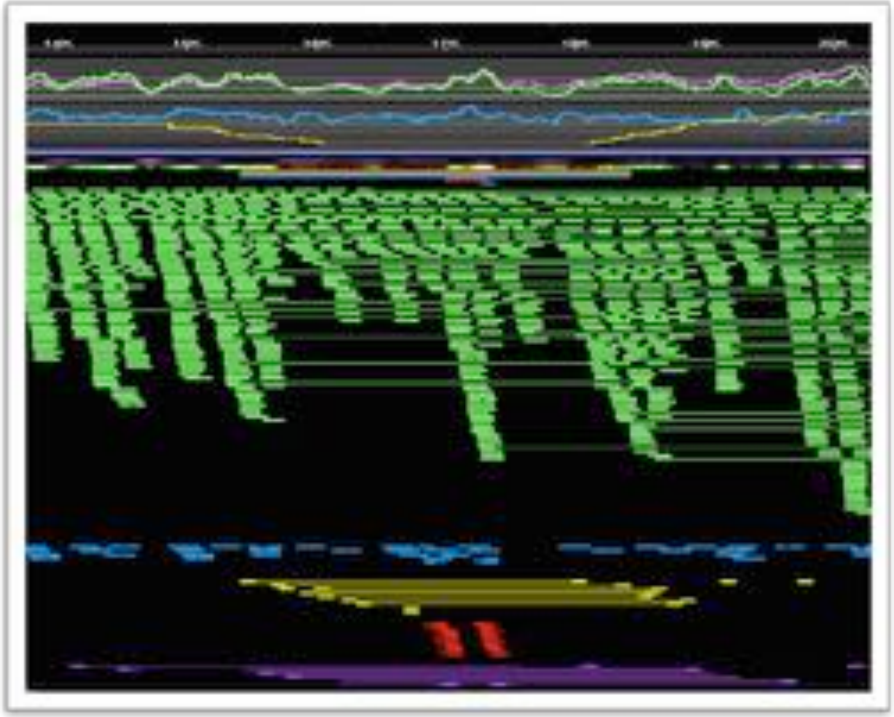
Forensics

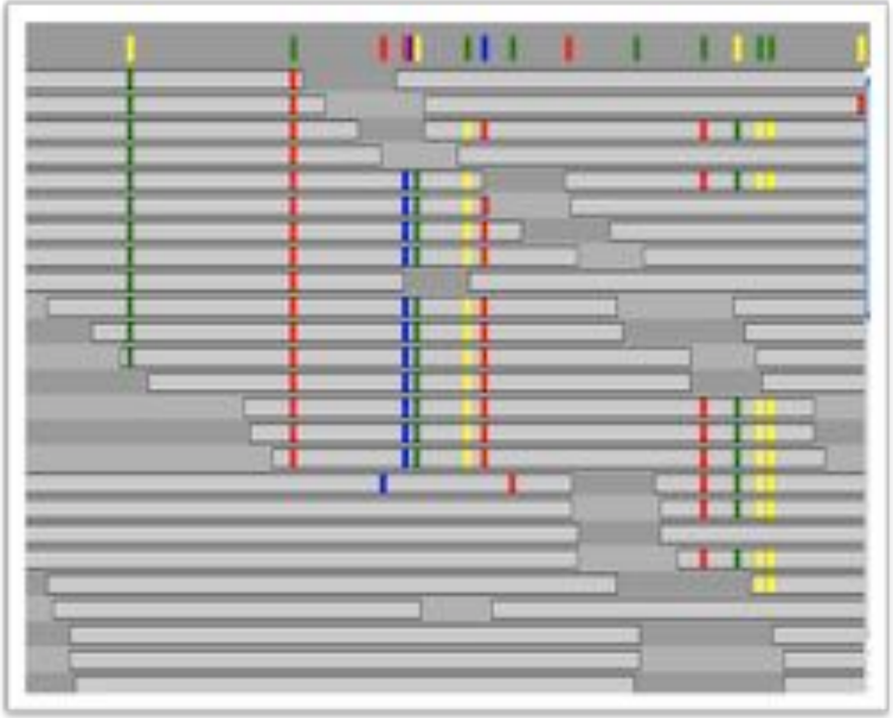
Assembly Forensics



Truth:

Mis-assembled:

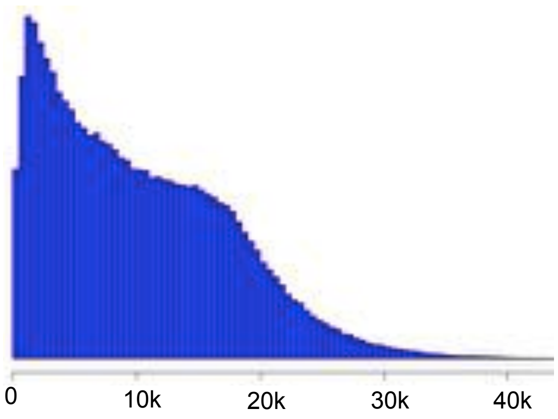




Hawkeye & AMOS: Visualizing and assessing the quality of genome assemblies
Schatz, M.C. *et al.* (2011) *Briefings in Bioinformatics.* doi: 10.1093/bib/bbr074

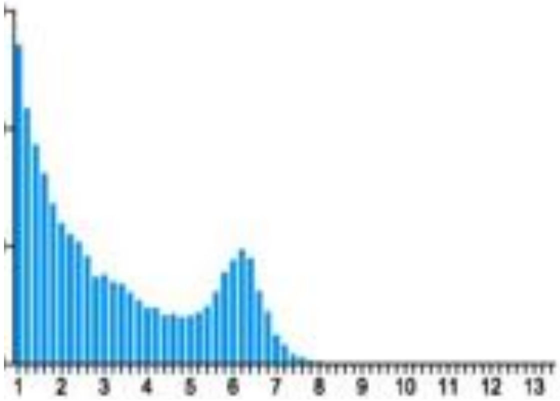
Long Read Sequencing Technology

PacBio RS II



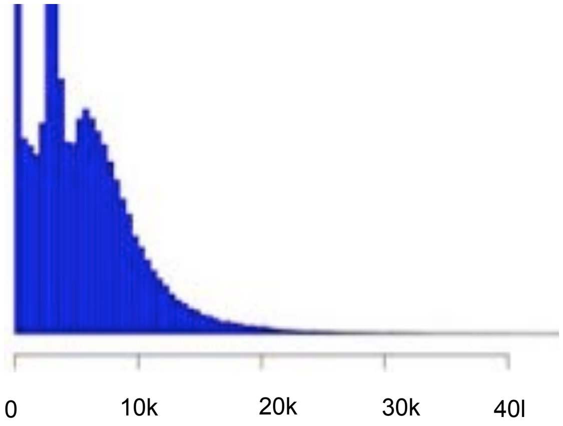
CSHL/PacBio

Moleculo



(Voskoboynik et al. 2013)

Oxford Nanopore

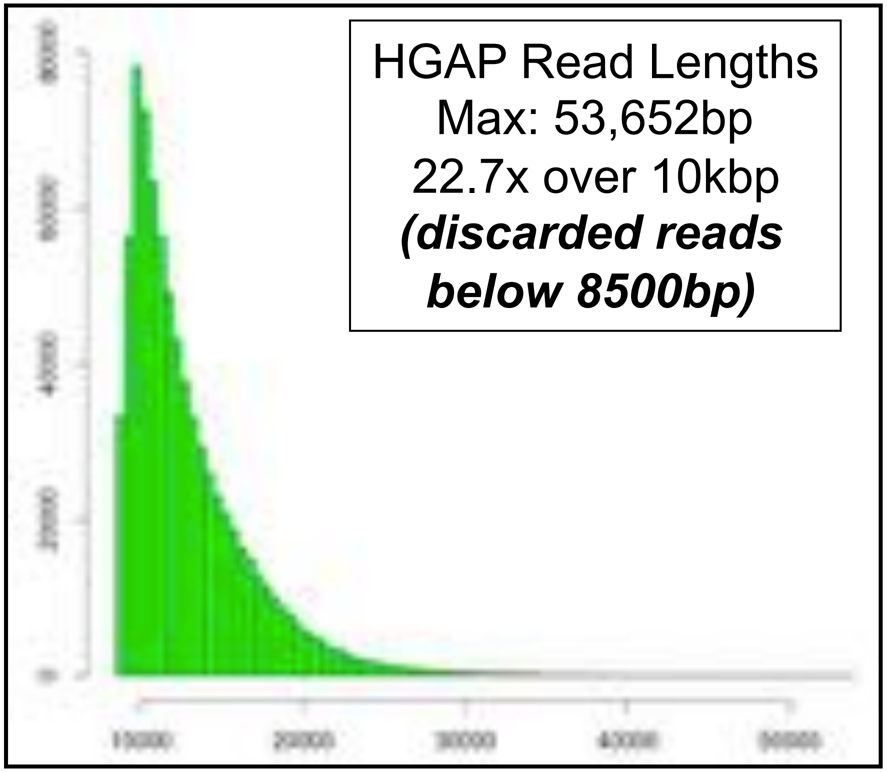


CSHL/ONT

O. sativa pv Indica (IR64)

Genome size:	~370 Mb
Chromosome N50:	~29.7 Mbp

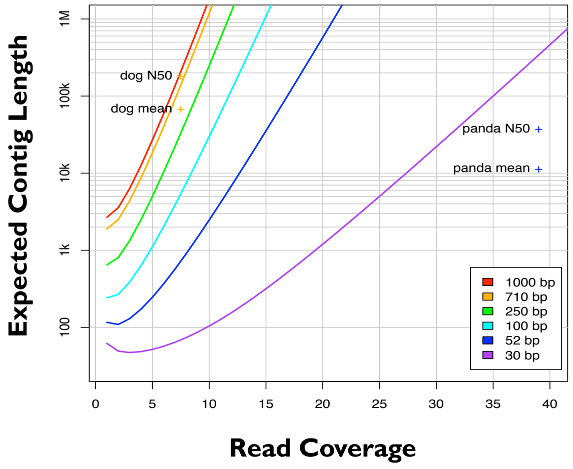
Assembly	Contig NG50
MiSeq Fragments 25x 456bp (3 runs 2x300 @ 450 FLASH)	19 kbp
"ALLPATHS-recipe" 50x 2x100bp @ 180 36x 2x50bp @ 2100 51x 2x50bp @ 4800	18 kbp
HGAP 22.7x @ 10kbp	4.0 Mbp
Nipponbare BAC-by-BAC Assembly	5.1 Mbp



HGAP Read Lengths
Max: 53,652bp
22.7x over 10kbp
(discarded reads below 8500bp)

Ingredients for a good assembly

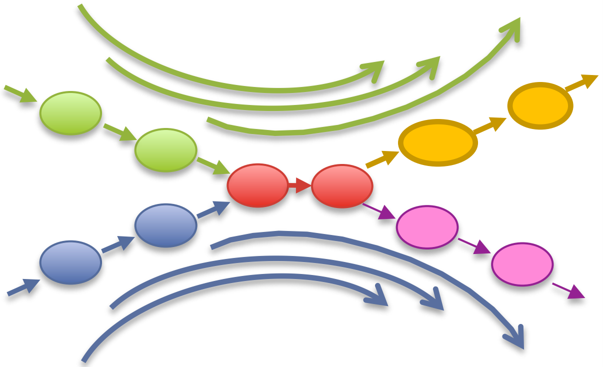
Coverage



High coverage is required

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

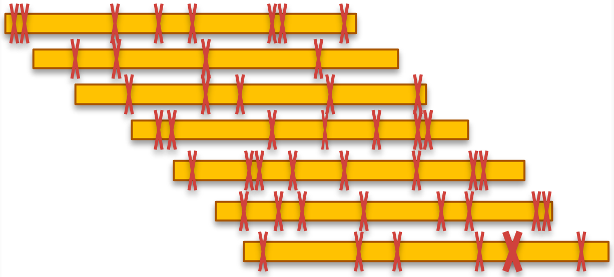
Read Length



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

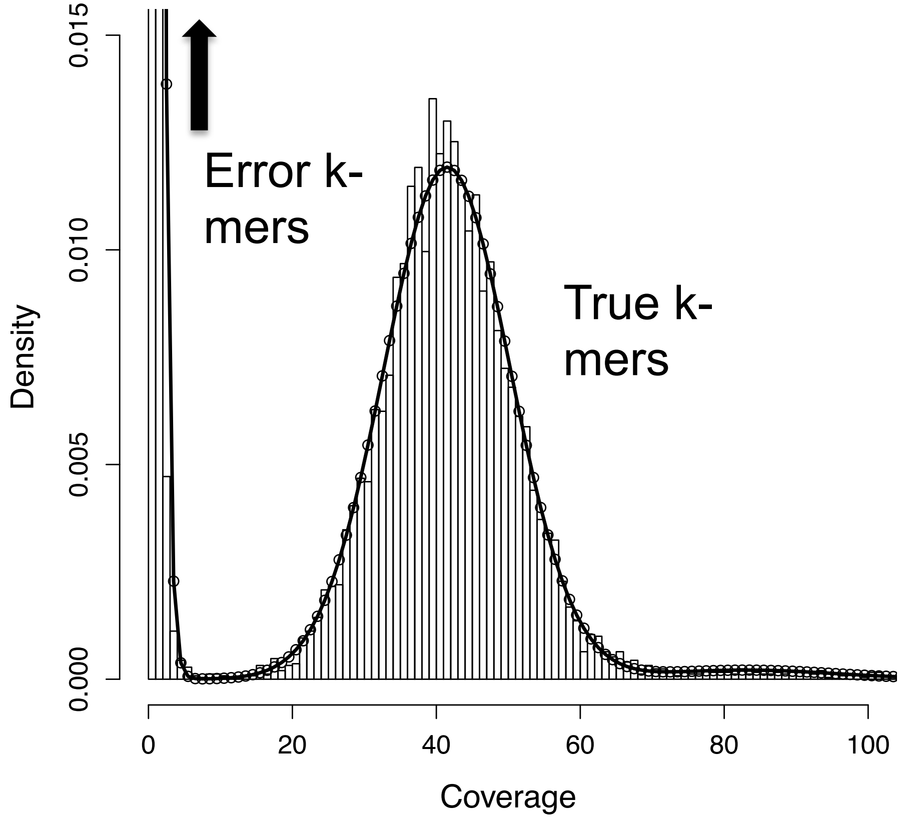
- Reads are assembled by finding kmers shared in pair of reads
- High error rate requires very short seeds, increasing complexity and forming assembly hairballs

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

Detection and Correction with Quake

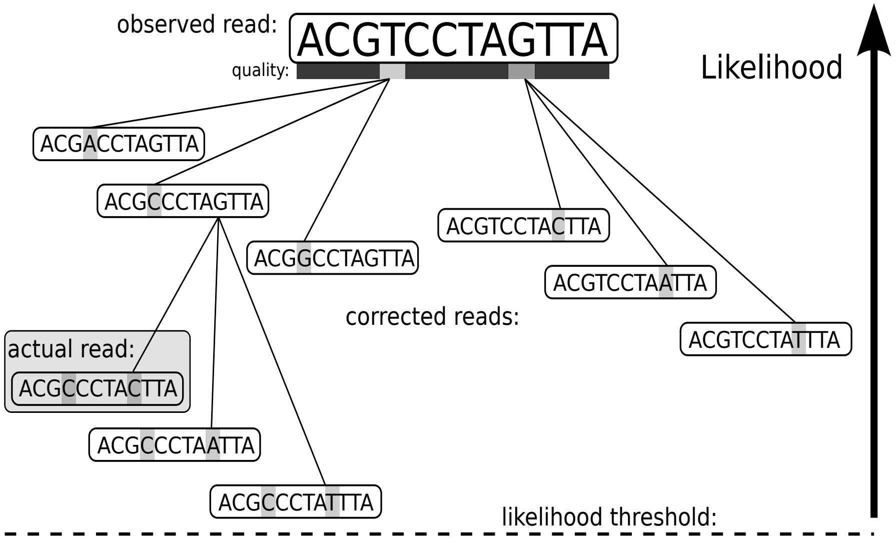
1. Count all "Q-mers" in reads

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



2. Correction Algorithm

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

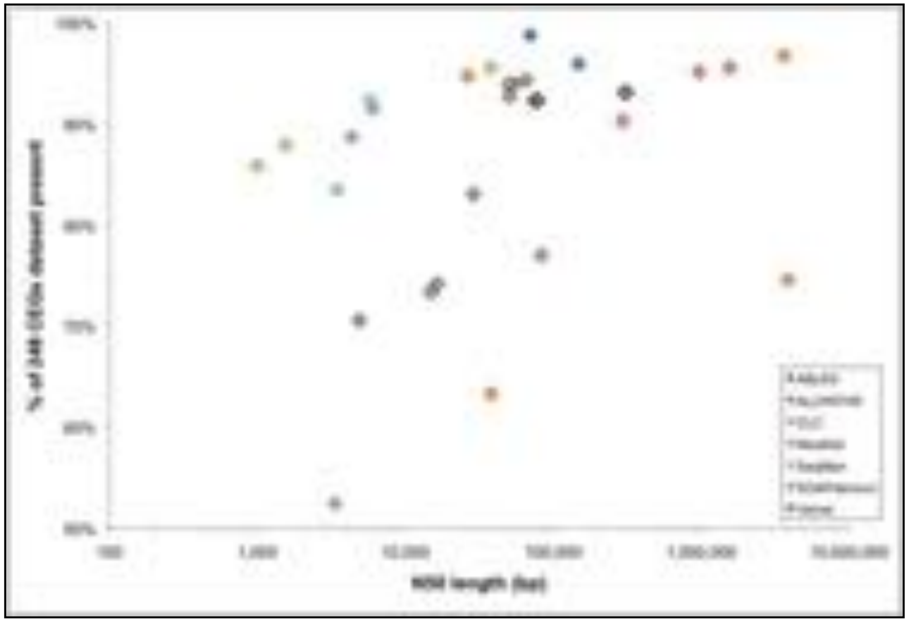


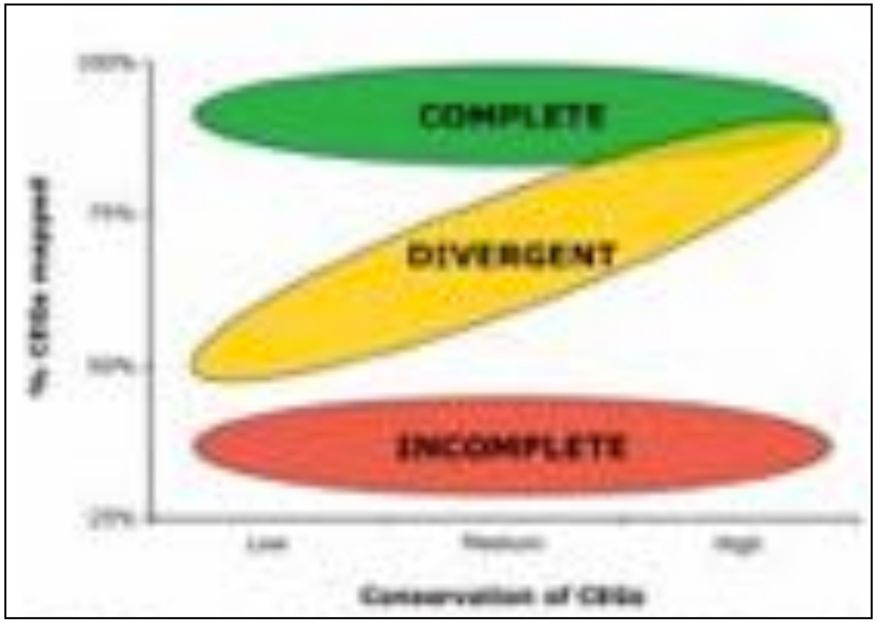
Quake: quality-aware detection and correction of sequencing reads.
Kelley, DR, Schatz, MC, Salzberg, SL (2010) Genome Biology. 11:R116

Gene Analysis with CEGMA

- Defined a set of 248 "core eukaryotic genes" (CEGs)
  - Highly conserved and in low copy numbers across all known eukaryotic species
  - House keeping genes and other basic functions

- Developed a robust alignment-based search tool to seek out those genes in your new assembly
  - Your ability to discover these 248 CEGs is highly correlated with finding the rest of the genes in the genome





Assessing the gene space in draft genomes

Parra, G, Bradnam, B, Ning, Z, Keane, T, Korf ,I (2009) 37(1) 289–297. doi:10.1093/nar/gkn916

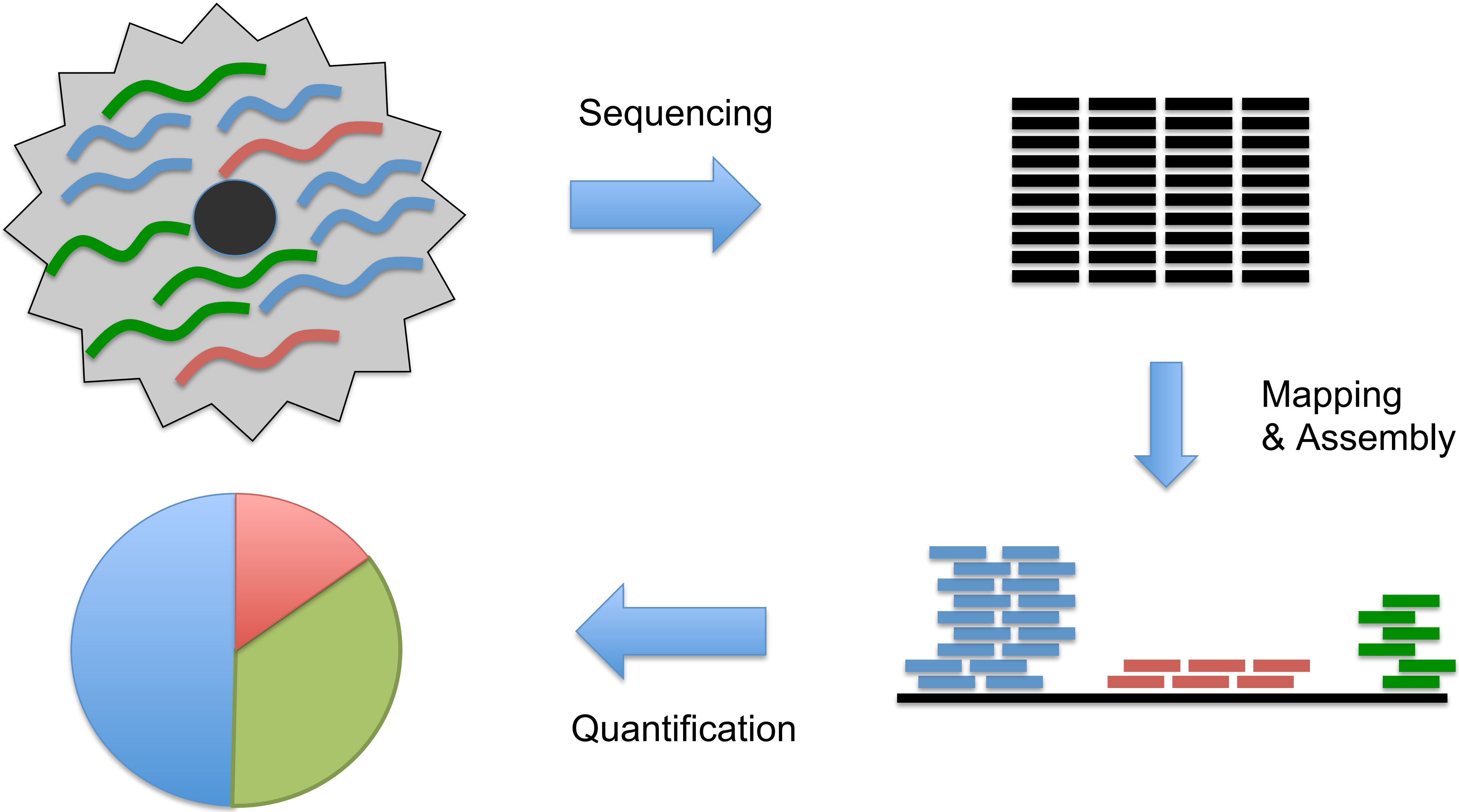
Outline

1. Assembly review
   1. Assembly by analogy
   2. Causes of Mis-assemblies

2. Evaluating Assembly Quality
   1. Assemblathon
   2. Size Statistics
   3. Mate-pairs
   4. CEGMA

3. RNA-seq specific challenges

RNA-seq Overview

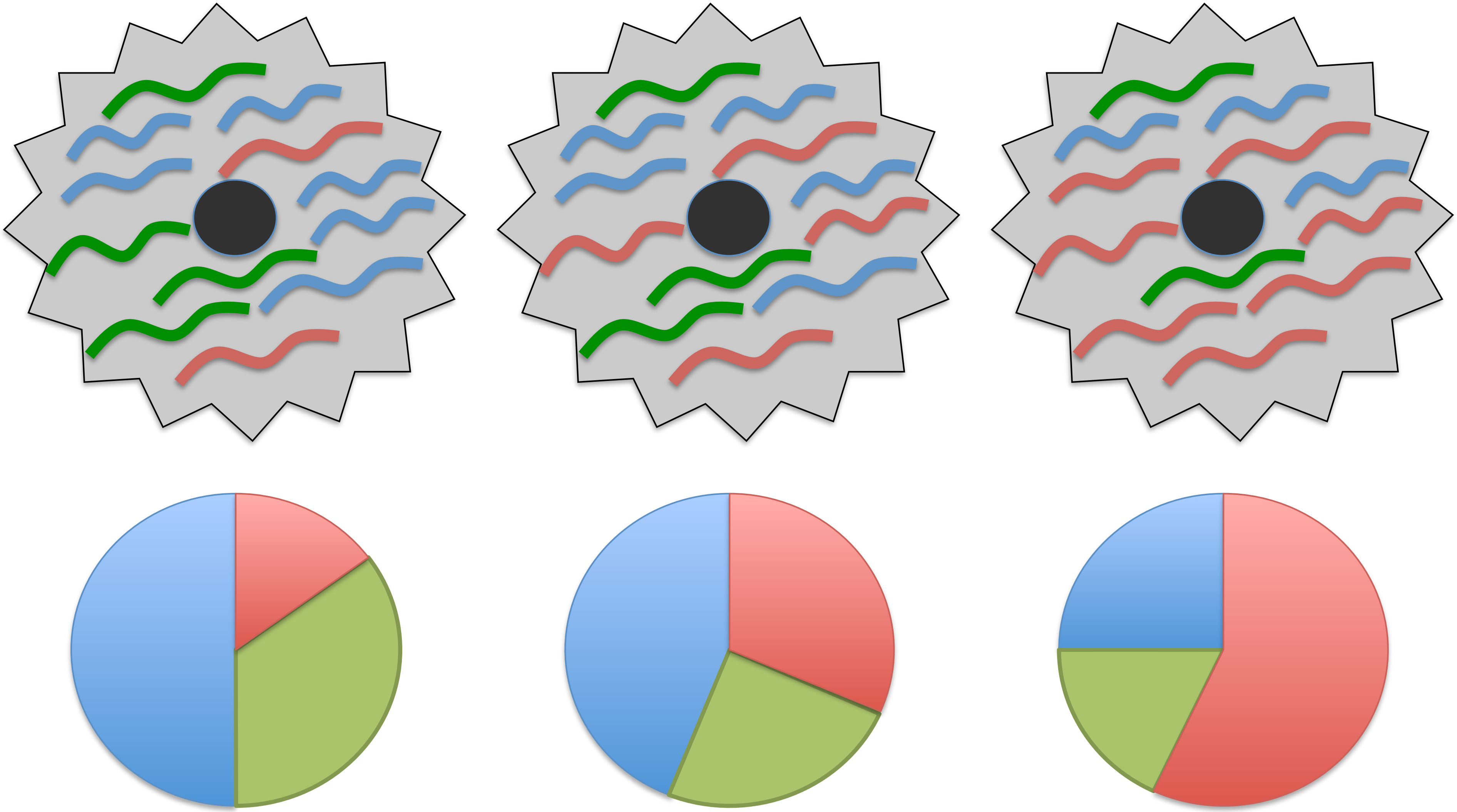


Sequencing

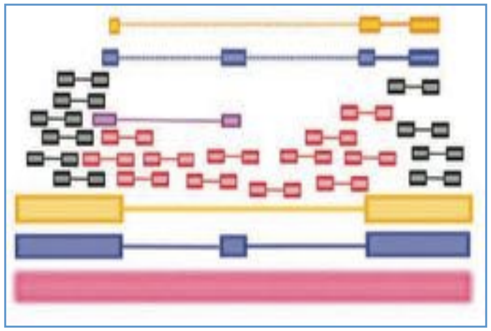
Mapping
& Assembly

Quantification

RNA-seq Overview



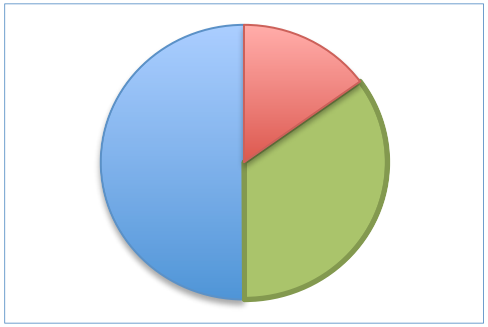
RNA-seq Challenges



Challenge 1: Eukaryotic genes are spliced

Solution: Use a spliced aligner, and assemble isoforms

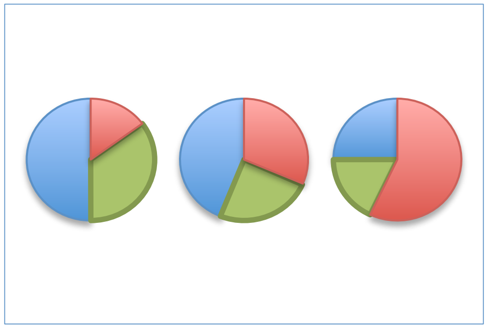
TopHat: discovering spliced junctions with RNA-Seq.
Trapnell et al (2009) *Bioinformatics*. 25:0 1105-1111



Challenge 2: Read Count != Transcript abundance

Solution: Infer underlying abundances (e.g. FPKM)

Transcript assembly and quantification by RNA-seq
Trapnell et al (2010) *Nat. Biotech*. 25(5): 511-515

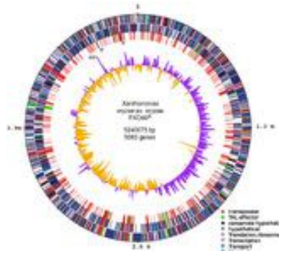


Challenge 3: Transcript abundances are stochastic

Solution: Replicates, replicates, and more replicates

RNA-seq differential expression studies: more sequence or more replication?
Liu et al (2013) *Bioinformatics*. doi:10.1093/bioinformatics/btt688

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

Questions?

http://schatzlab.cshl.edu/

@mike_schatz