

# Genome Sequencing & Assembly

Michael Schatz

March 31, 2014

CSHL Genome Access





# Outline

## I. Assembly theory

1. Assembly by analogy
2. De Bruijn and Overlap graph
3. Coverage, read length, errors, and repeats

## 2. Whole Genome Alignment

1. Aligning & visualizing with MUMmer

## 3. Genome assemblers

1. ALLPATHS-LG: recommended for Illumina-only projects
2. Celera Assembler: recommended for PacBio projects



# Outline

## I. Assembly theory

1. Assembly by analogy
2. De Bruijn and Overlap graph
3. Coverage, read length, errors, and repeats

## 2. Whole Genome Alignment

1. Aligning & visualizing with MUMmer

## 3. Genome assemblers

1. ALLPATHS-LG: recommended for Illumina-only projects
2. Celera Assembler: recommended for PacBio projects

# Shredded Book Reconstruction

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

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

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

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

- How can he reconstruct the text?
  - $5 \text{ copies} \times 138,656 \text{ words} / 5 \text{ words per fragment} = 138k \text{ fragments}$
  - The short fragments from every copy are mixed together
  - Some fragments are identical

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

# Greedy Reconstruction

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

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

Original Fragment

It was the best of

Directed Edge

It was the best → was the best of

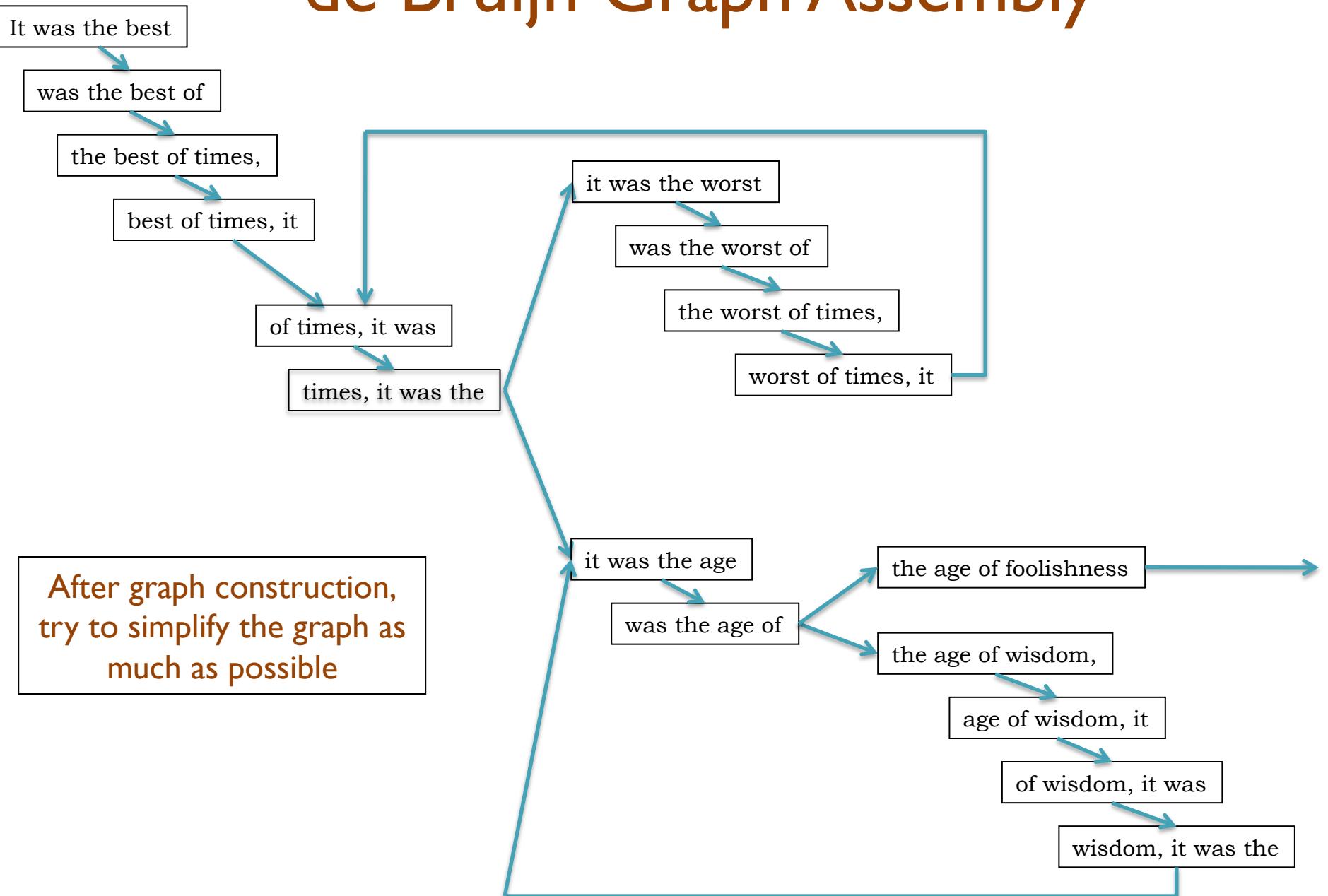
- Locally constructed graph reveals the global sequence structure
  - Overlaps between sequences implicitly computed

de Bruijn, 1946

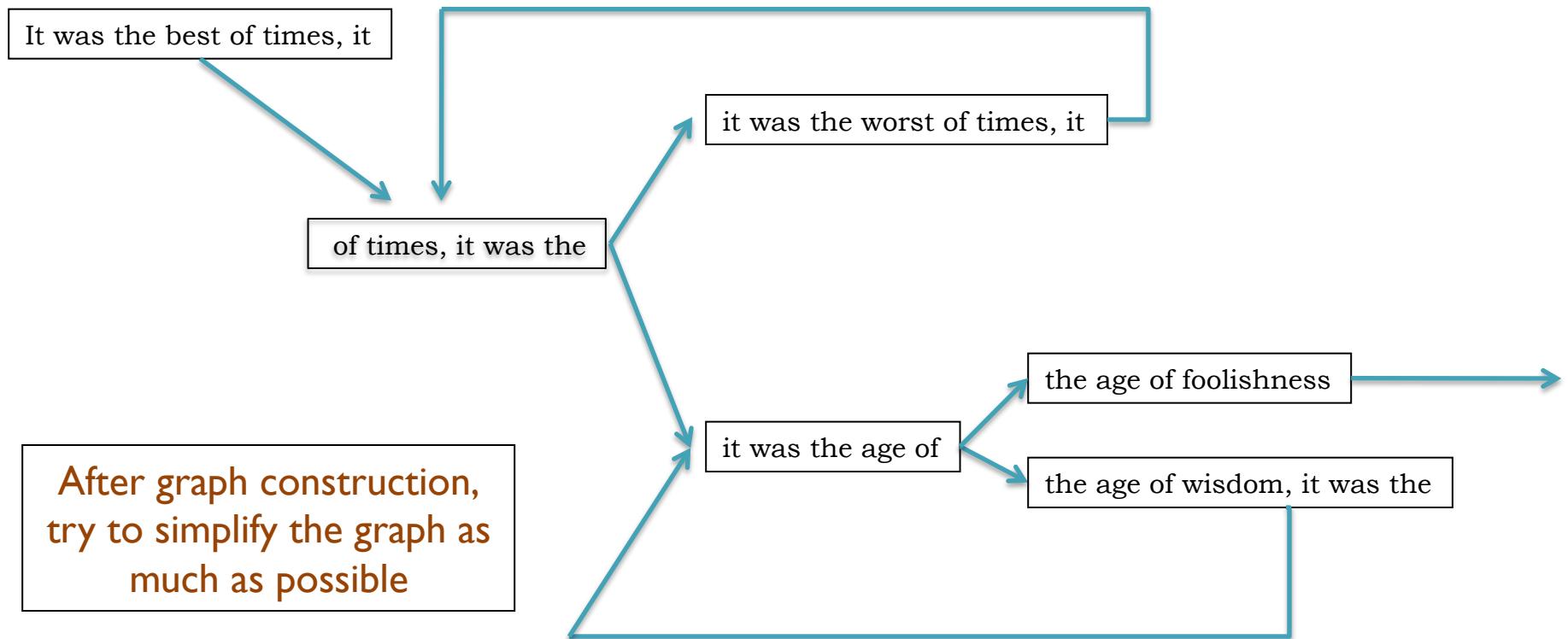
Idury and Waterman, 1995

Pevzner, Tang, Waterman, 2001

# de Bruijn Graph Assembly



# de Bruijn Graph Assembly



# The full tale

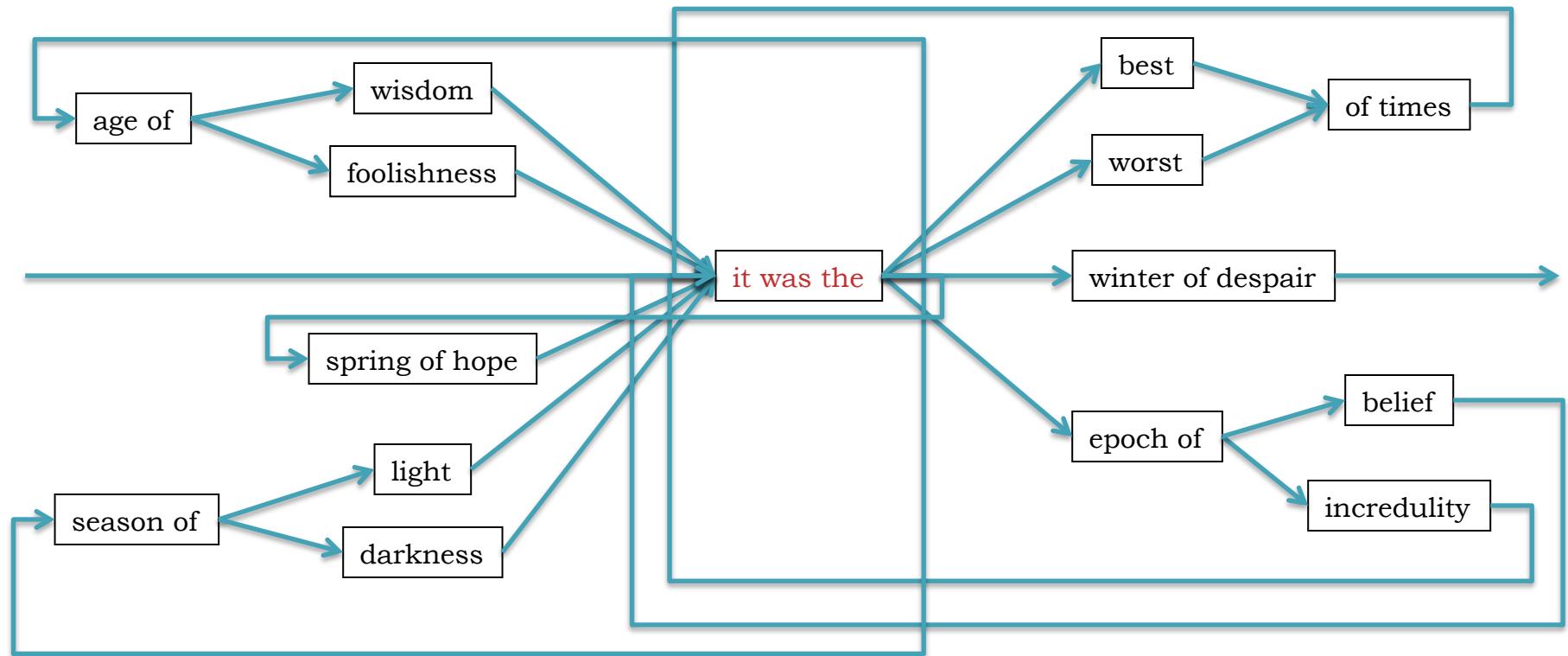
... it was the best of times it was the worst of times ...

... it was the age of wisdom it was the age of foolishness ...

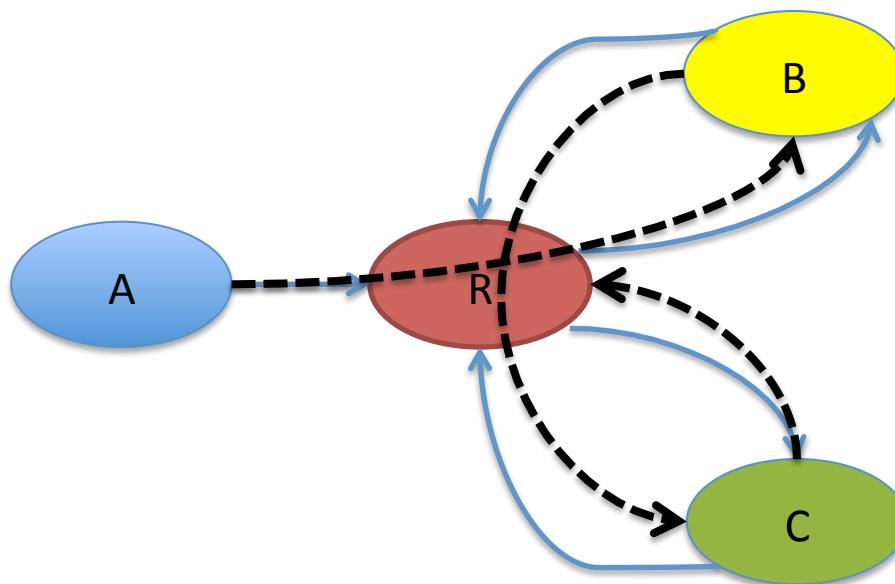
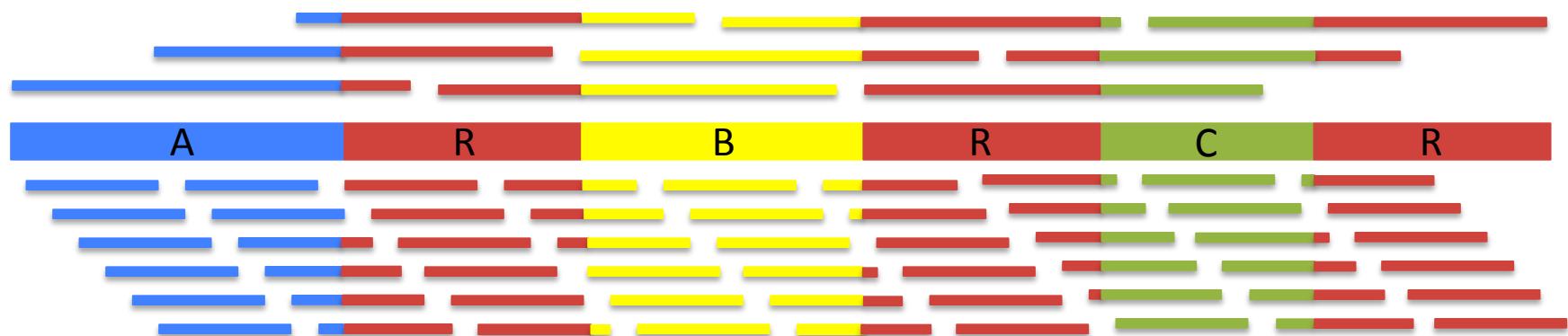
... it was the epoch of belief it was the epoch of incredulity ...

... it was the season of light it was the season of darkness ...

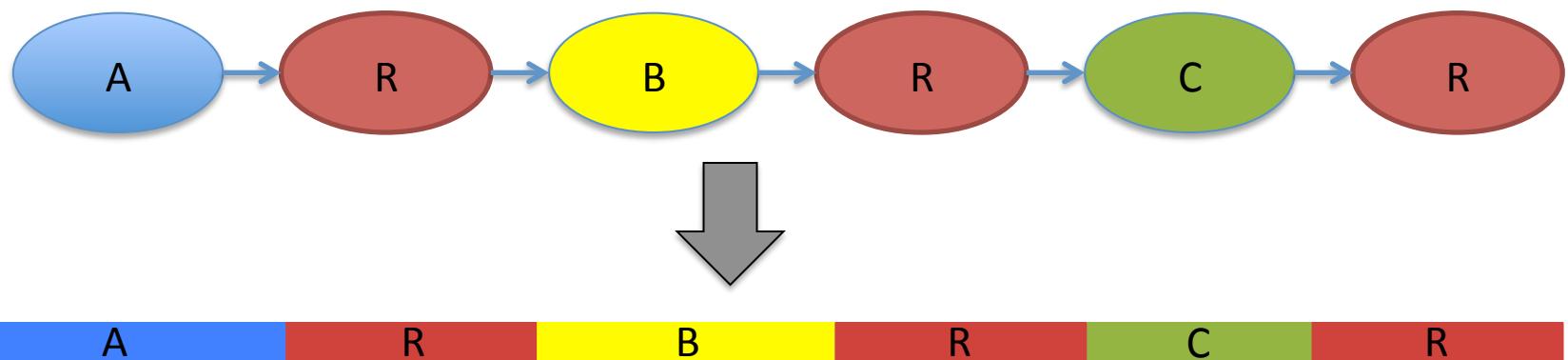
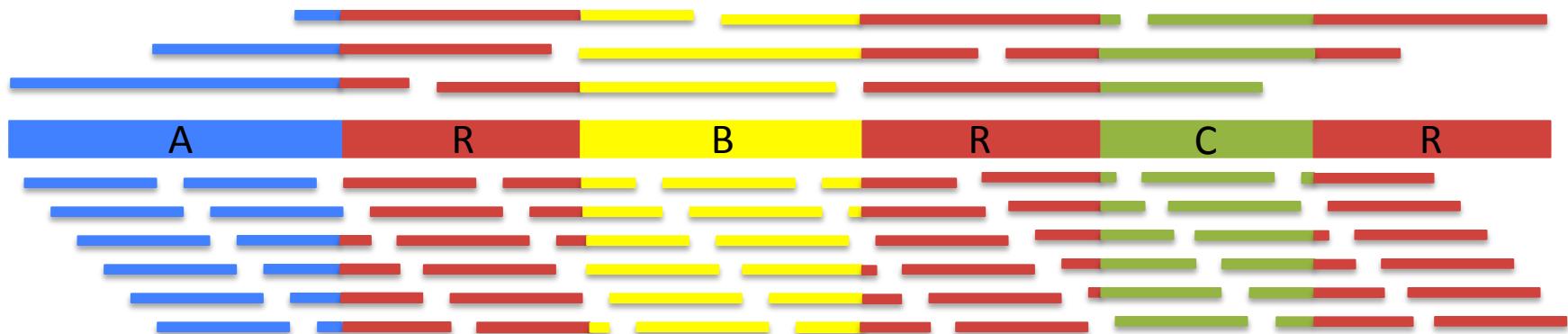
... it was the spring of hope it was the winter of despair ...



# Assembly Complexity



# Assembly Complexity



# Milestones in Genome Assembly

Nature Vol. 265 February 24 1977

487

## articles

### Nucleotide sequence of bacteriophage $\Phi$ X174 DNA

F. Sanger, G. M. Air<sup>1</sup>, B. G. Barrell, N. L. Brown<sup>1</sup>, A. R. Coulson, J. C. Fiddes,  
C. A. Hutchison III<sup>1</sup>, P. M. Slocombe<sup>2</sup> & M. Smith<sup>2</sup>

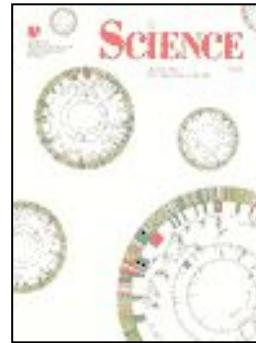
MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

A DNA sequence for the genome of bacteriophage  $\Phi$ X174 of approximately 5,375 nucleotides has been determined using the rapid and simple "plus and minus" method. The sequence identifies many of the features responsible for the production of the various proteins known to be produced by the organism, including initiation and termination sites for the proteins and RNAs. Two pairs of genes are coded by the same region of DNA using different reading frames.

The genome of bacteriophage  $\Phi$ X174 is a single-stranded, circular molecule of DNA. It contains 5,375 nucleotides and nine known proteins. The order of genes, as determined by genetic techniques<sup>1-3</sup>, is A-B-C-D-E-F-F-G-H. Genes F, G and H code for structural proteins, while genes A, B, C and D code for enzymes involved in the metabolism of host bacteria.

Two sets of synthetic primers were used to prime the DNA with DNA polymerase were being developed<sup>4</sup> and Scheraga<sup>5</sup> found that one of these primers was complementary to part of the ribosome binding site. This was used to prime into the intervening regions between the T and G genes using DNA, ribonuclease and protein sequencing. The ribonuclease technique<sup>6</sup> facilitated the sequence determination of the intervening regions. The plus and minus method<sup>7</sup> described was also used to develop the plus and minus method<sup>8</sup>. Suitable synthetic primers are, however, difficult to prepare and at

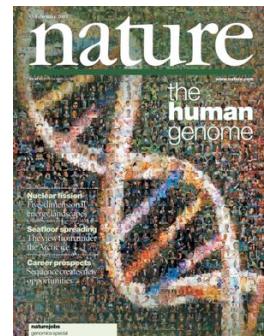
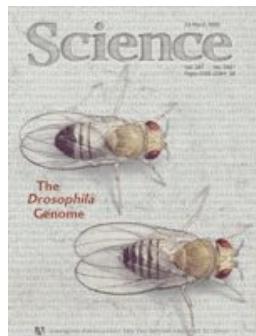
1977. Sanger et al.  
1<sup>st</sup> Complete Organism  
5375 bp



1995. Fleischmann et al.  
1<sup>st</sup> Free Living Organism  
TIGR Assembler. 1.8Mbp



1998. C.elegans SC  
1<sup>st</sup> Multicellular Organism  
BAC-by-BAC Phrap. 97Mbp



2000. Myers et al.  
1<sup>st</sup> Large WGS Assembly.  
Celera Assembler. 116 Mbp

2001. Venter et al., IHGSC  
Human Genome  
Celera Assembler/GigaAssembler. 2.9 Gbp



2010. Li et al.  
1<sup>st</sup> Large SGS Assembly.  
SOAPdenovo 2.2 Gbp

Like Dickens, we must computationally reconstruct a genome from short fragments

# Assembly Applications

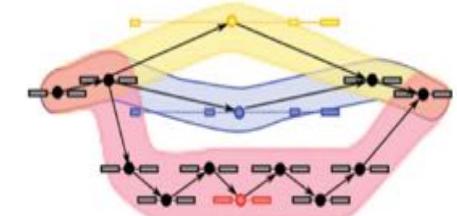
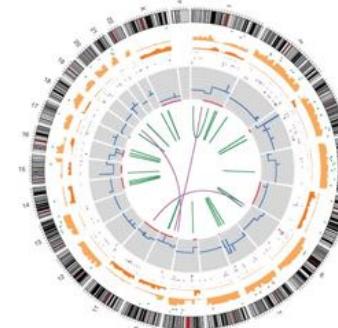
- Novel genomes



- Metagenomes

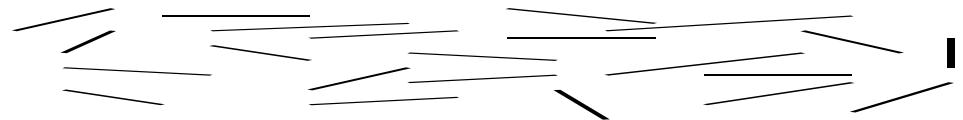


- Sequencing assays
  - Structural variations
  - Transcript assembly
  - ...



# Assembling a Genome

## 1. Shear & Sequence DNA



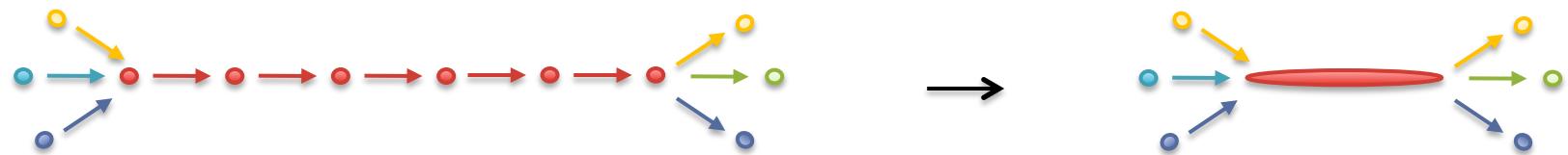
## 2. Construct assembly graph from overlapping reads

...AGCCTAGGGATGCGCGACACGT

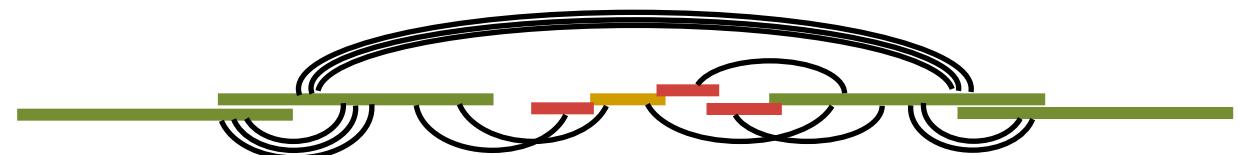
GGATGCGCGACACGT CGCATATCCGGTTGGT CAACCTCGGACGGAC

CAACCTCGGACGGAC CTCAGCGAA...

## 3. Simplify assembly graph



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



# Why are genomes hard to assemble?

## 1. **Biological:**

- (Very) High ploidy, heterozygosity, repeat content



## 2. **Sequencing:**

- (Very) large genomes, imperfect sequencing

## 3. **Computational:**

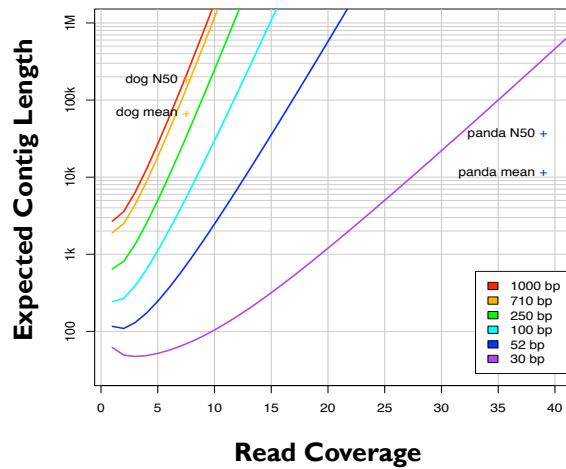
- (Very) Large genomes, complex structure

## 4. **Accuracy:**

- (Very) Hard to assess correctness

# Ingredients for a good assembly

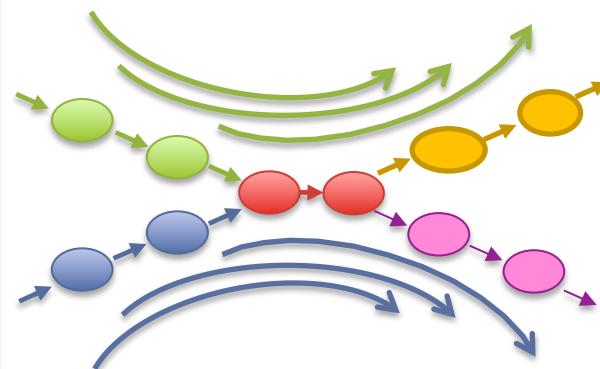
## Coverage



### High coverage is required

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

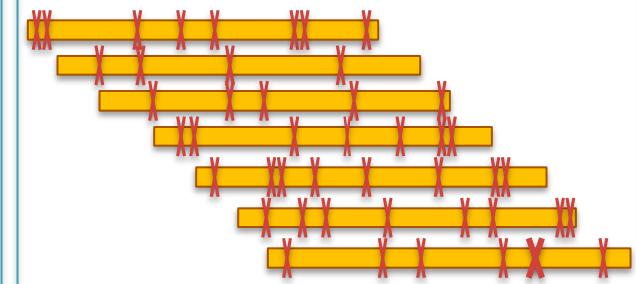
## Read Length



### 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

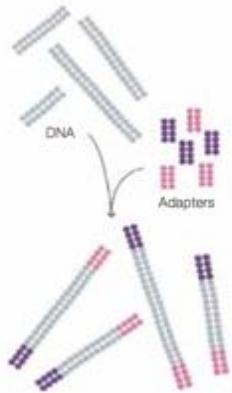


### Errors obscure overlaps

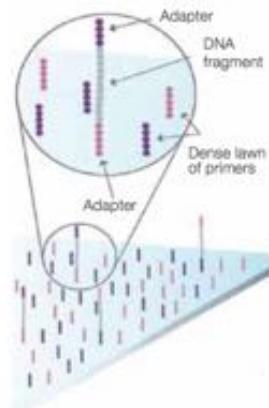
- Reads are assembled by finding kmers shared in pair of reads
- High error rate requires very short seeds, increasing complexity and forming assembly hairballs

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

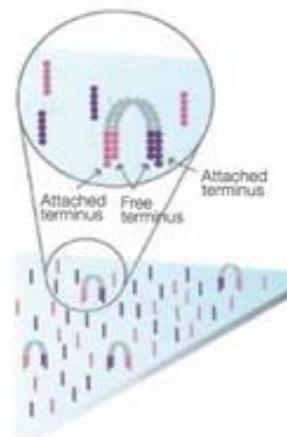
# Illumina Sequencing by Synthesis



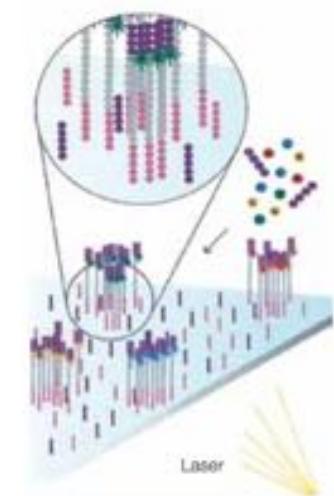
1. Prepare



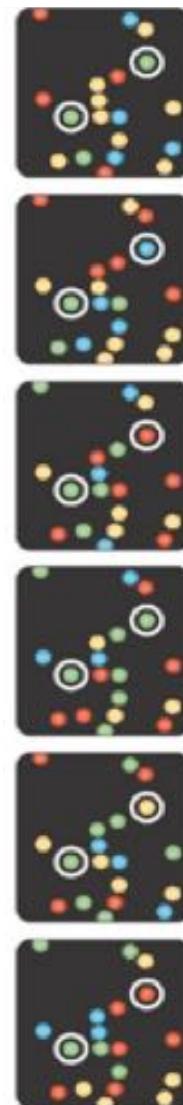
2. Attach



3. Amplify



4. Image

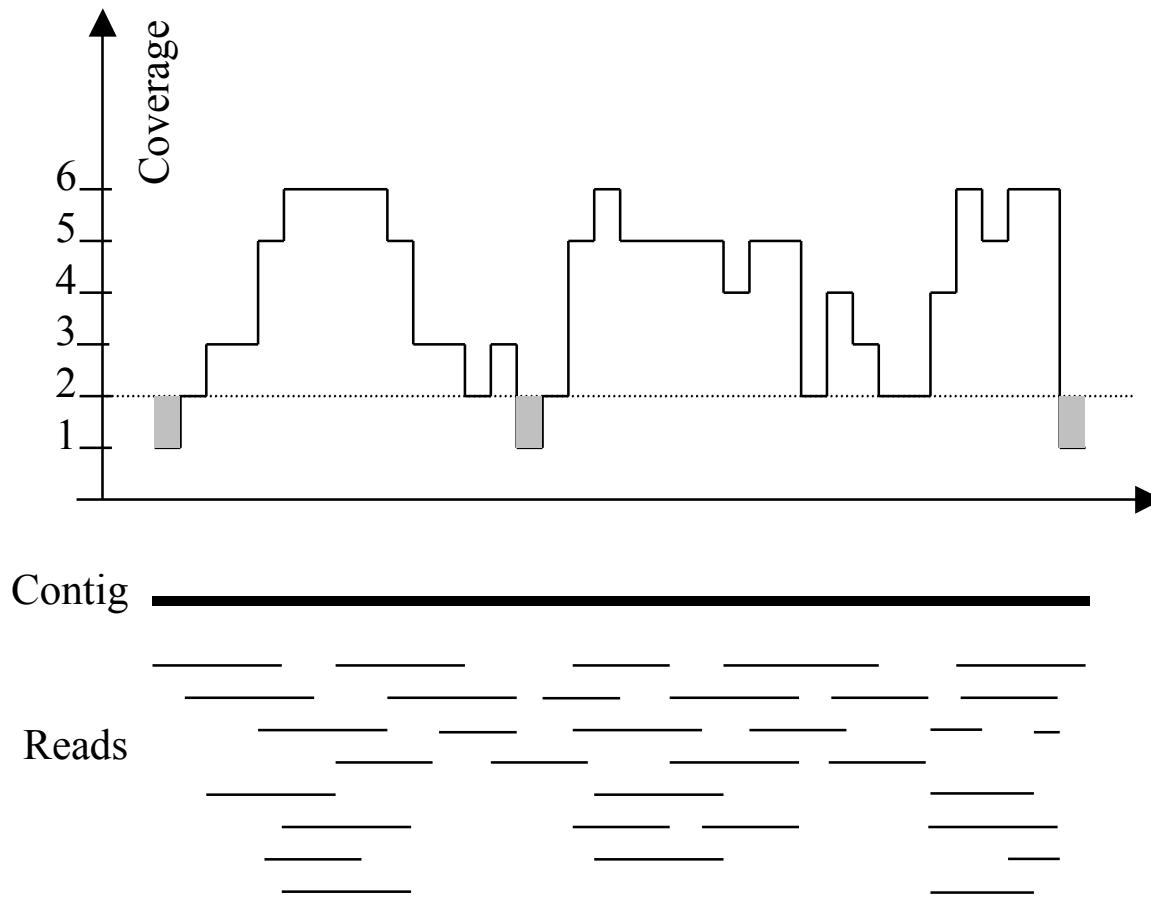


5. Basecall

Metzker (2010) Nature Reviews Genetics 11:31-46  
<http://www.youtube.com/watch?v=l99aKKHcxC4>

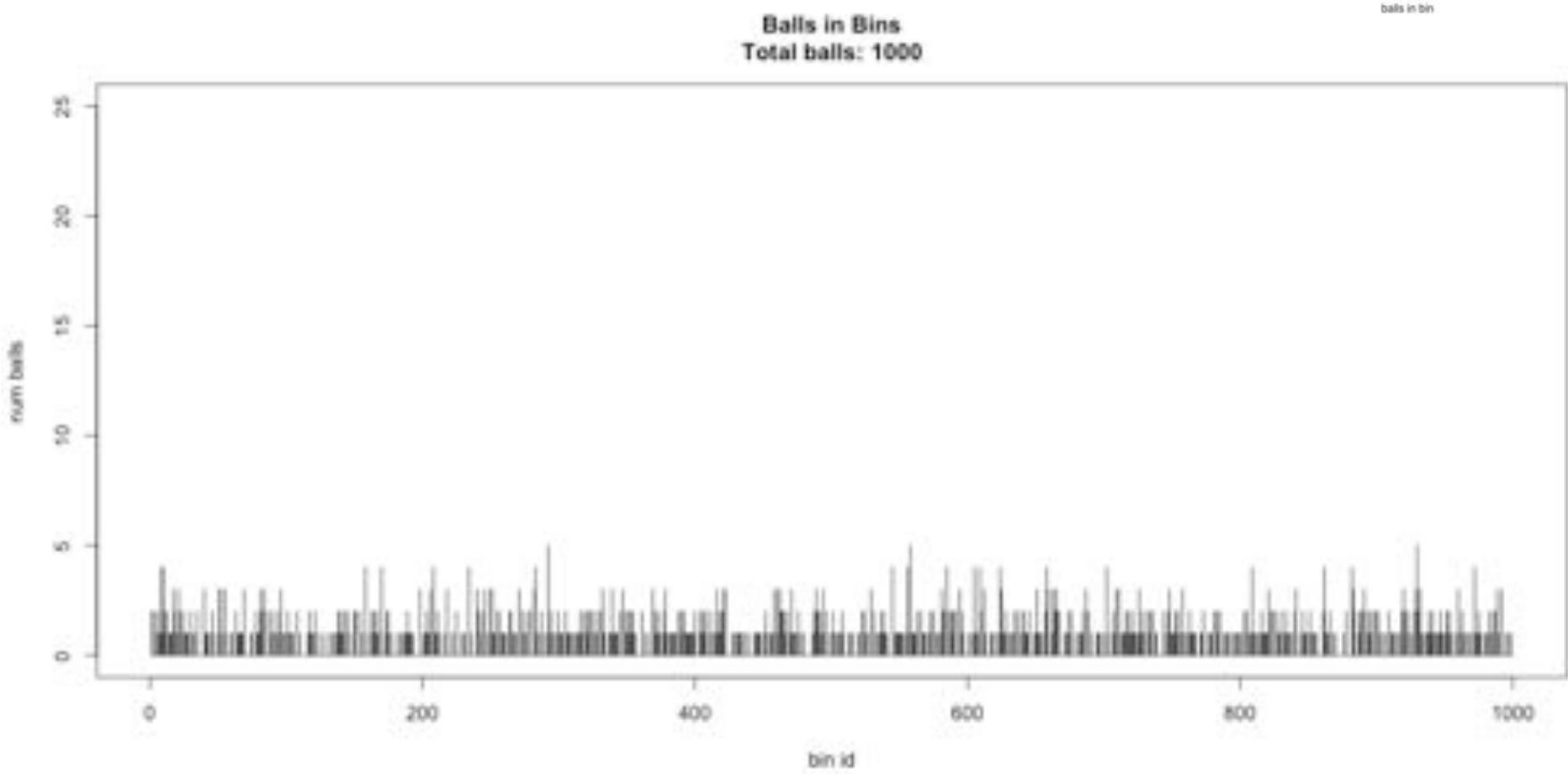
## Coverage

# Typical contig coverage



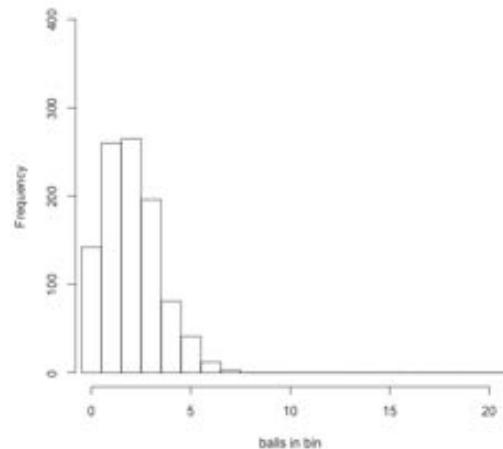
Imagine raindrops on a sidewalk

# Balls in Bins IX

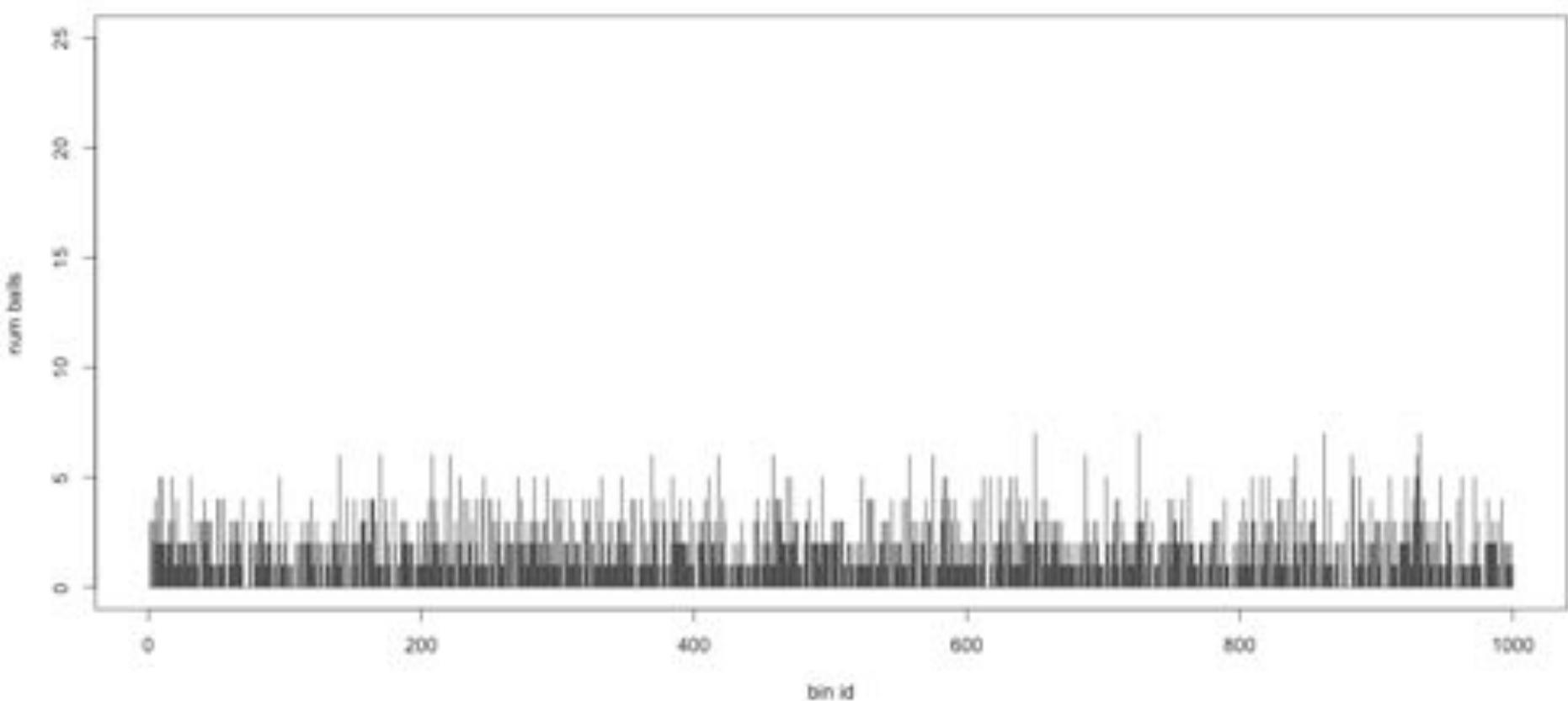


# Balls in Bins 2x

Histogram of balls in each bin  
Total balls: 2000 Empty bins: 142

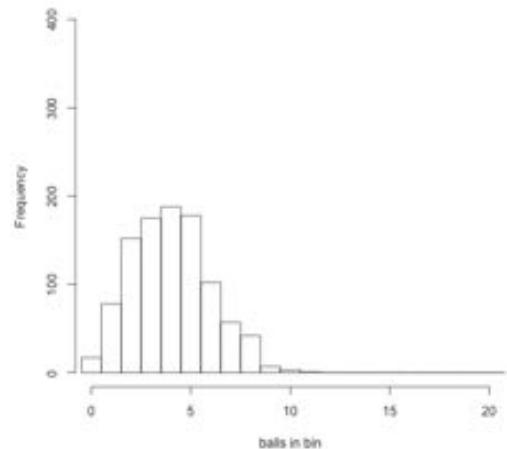


Balls in Bins  
Total balls: 2000

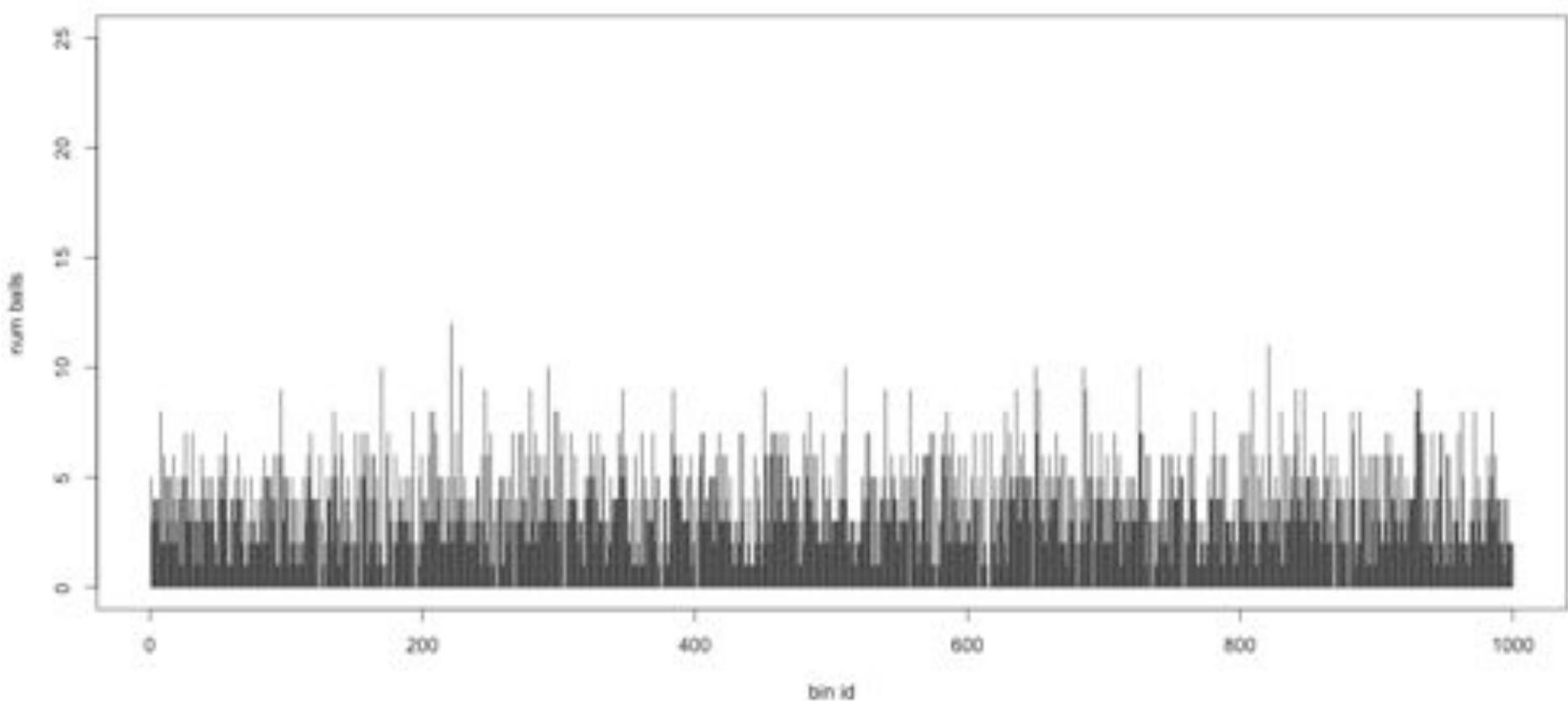


# Balls in Bins 4x

Histogram of balls in each bin  
Total balls: 4000 Empty bins: 17

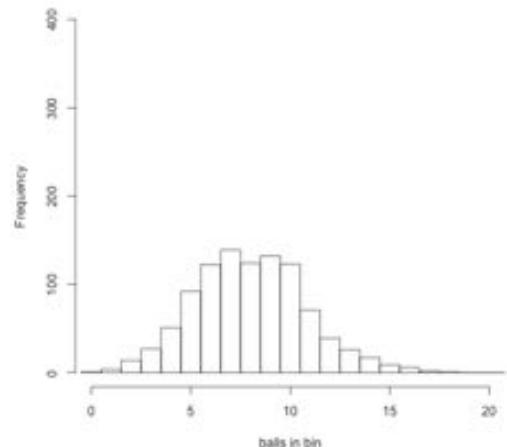


Balls in Bins  
Total balls: 4000

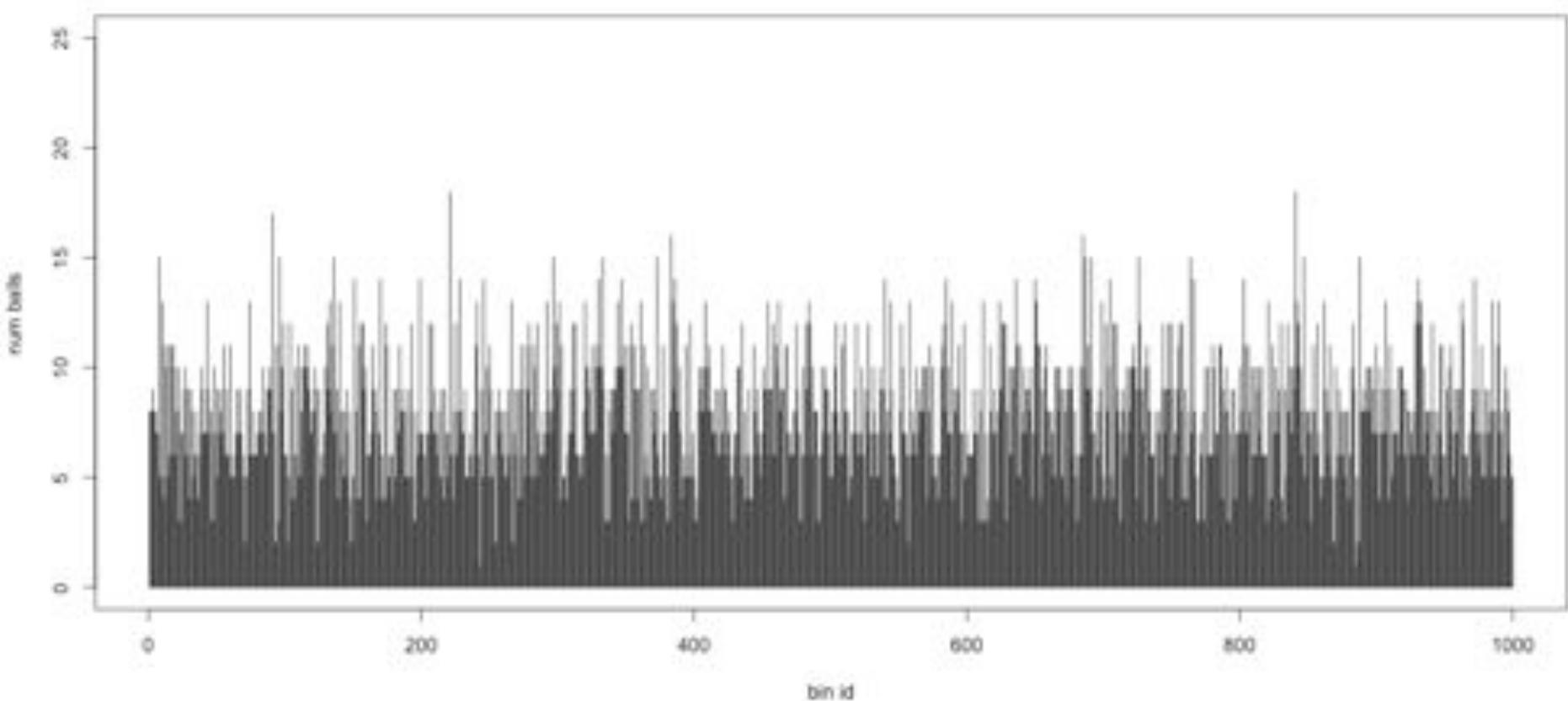


# Balls in Bins 8x

Histogram of balls in each bin  
Total balls: 8000 Empty bins: 1



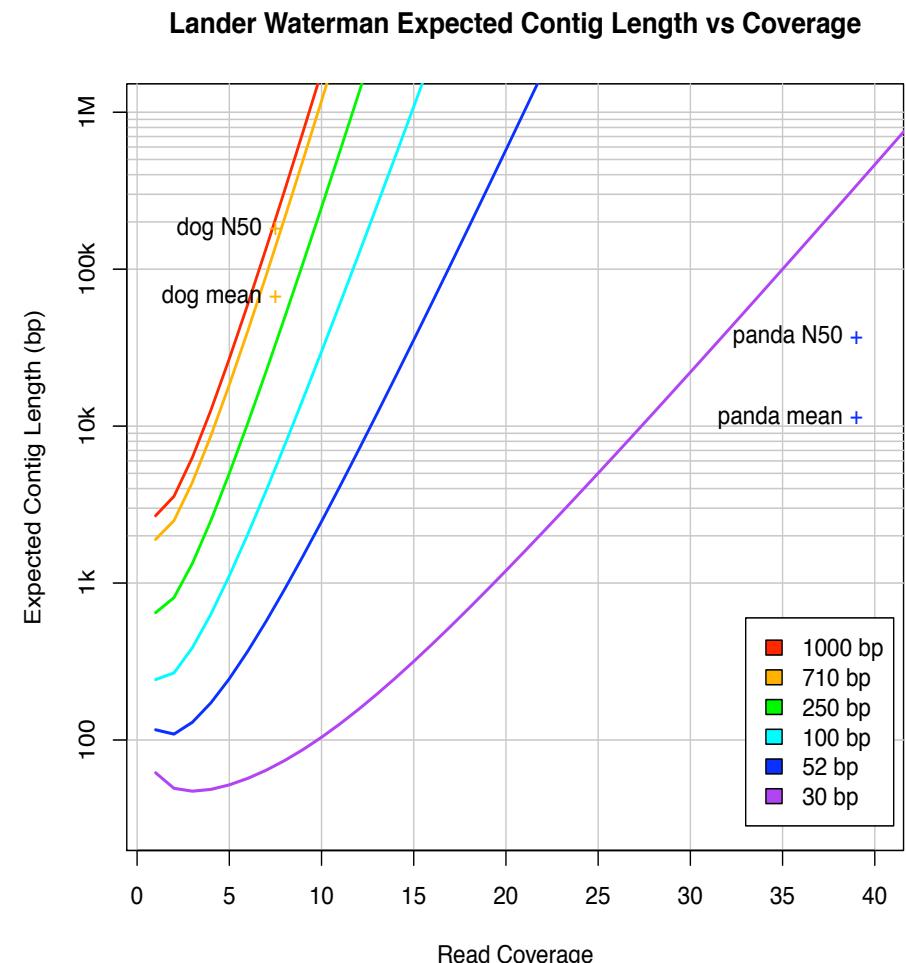
Balls in Bins  
Total balls: 8000



# Coverage and Read Length

## Idealized Lander-Waterman model

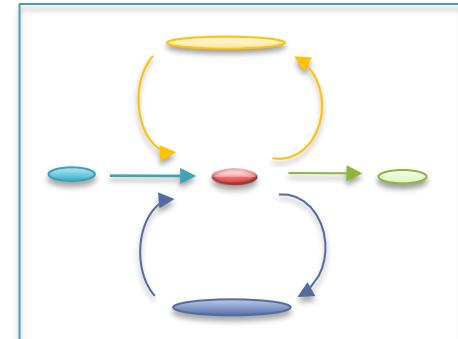
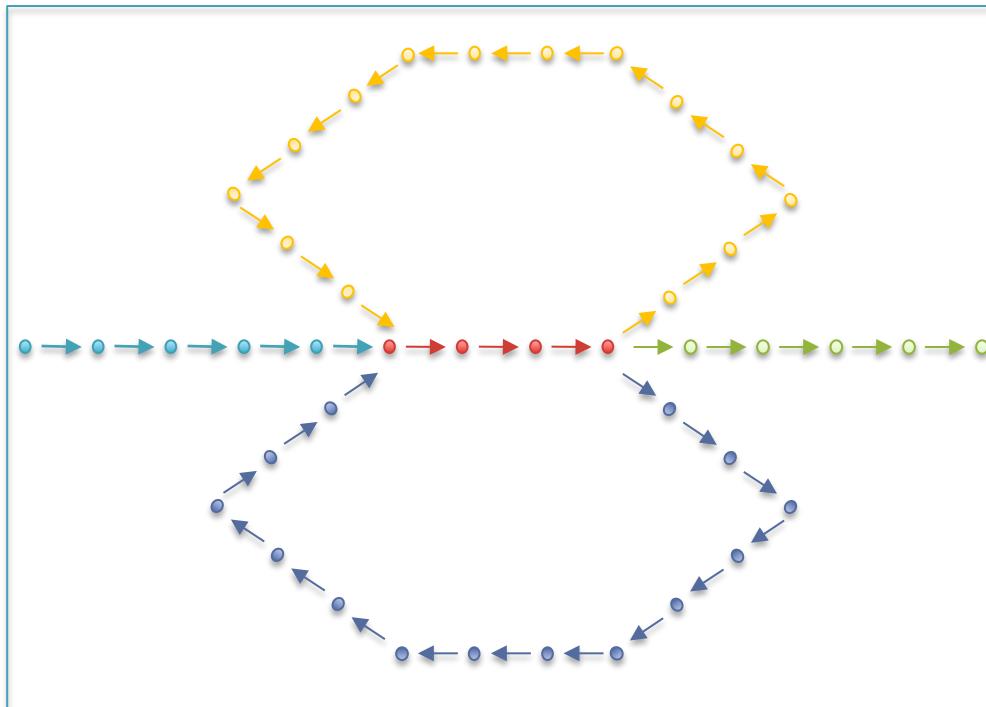
- Reads start at perfectly random positions
- Contig length is a function of coverage and read length
  - Short reads require much higher coverage to reach same expected contig length
- Need even high coverage for higher ploidy, sequencing errors, sequencing biases
  - Recommend 100x coverage



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

# 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, and (3) repeats



# Errors in the graph



(Chaisson, 2009)

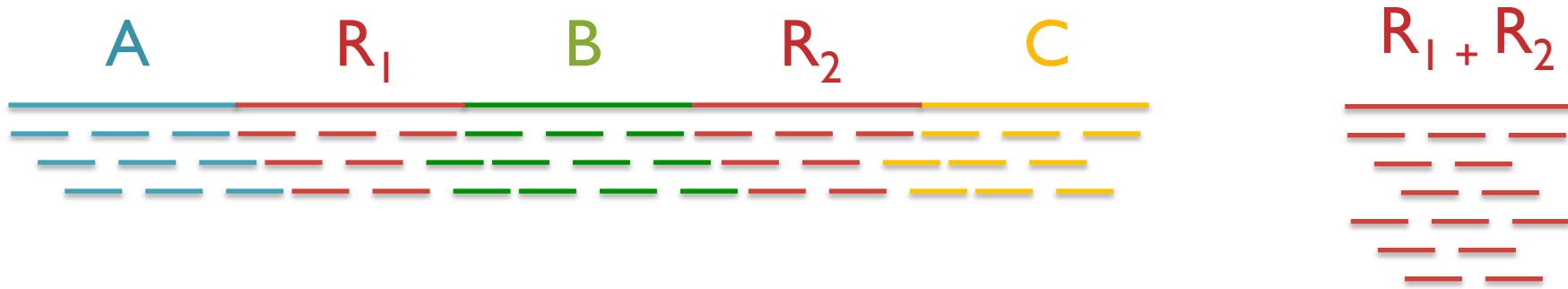
Clip Tips	Pop Bubbles
<p>was the worst of times,</p> <p>was the worst of <b>tymes</b>,</p> <p>the worst of times, it</p>	<p>was the worst of times,</p> <p>was the worst of <b>tymes</b>,</p> <p>times, it was the age</p> <p><b>tymes</b>, it was the age</p>
<p>the worst of <b>tymes</b>,</p> <p>was the worst of</p> <p>the worst of times,</p> <p>worst of times, it</p>	<p><b>tymes</b>,</p> <p>was the worst of</p> <p>it was the age</p> <p>times,</p>

# Repetitive regions

Repeat Type	Definition / Example	Prevalence
Low-complexity DNA / Microsatellites	$(b_1 b_2 \dots b_k)^N$ where $1 \leq k \leq 6$ CACACACACACACACACACA	2%
SINEs (Short Interspersed Nuclear Elements)	<i>Alu</i> 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  $\Delta$ .
  - If we see many more reads than  $k$  (if the arrival rate is  $> A$ ) , it is likely to be a collapsed repeat
  - Requires an accurate genome size estimate

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

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

# Paired-end and Mate-pairs

## Paired-end sequencing

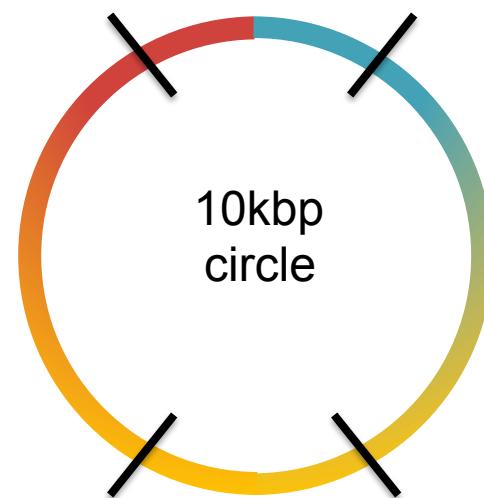
- Read one end of the molecule, flip, and read the other end
- Generate pair of reads separated by up to 500bp with inward orientation



## Mate-pair sequencing

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

10kbp



2x100 @ ~10kbp (outies)

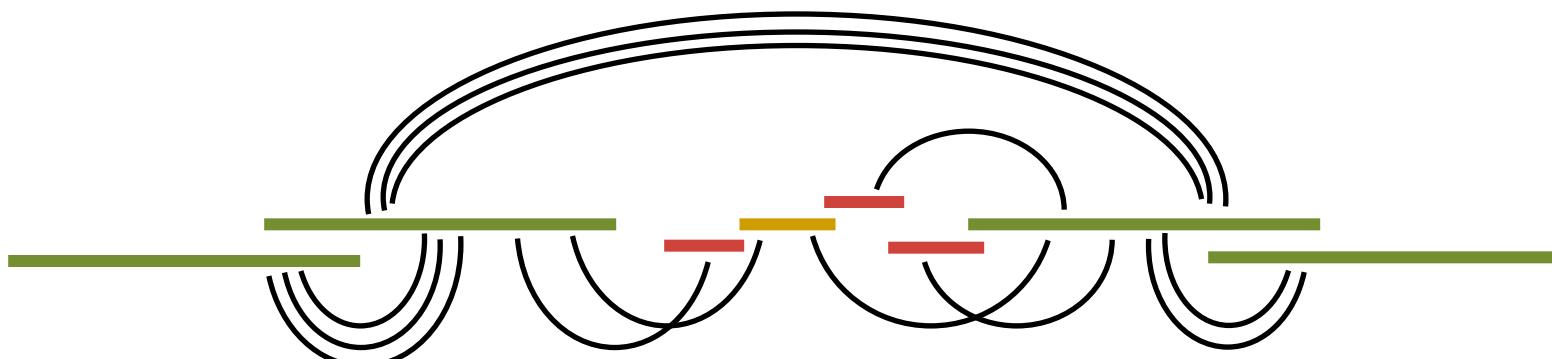


2x100 @ 300bp (innies)



# Scaffolding

- Initial contigs (aka unipaths, unitigs) terminate at
  - Coverage gaps: especially extreme GC regions
  - Conflicts: sequencing errors, repeat boundaries
- Iteratively resolve longest, ‘most unique’ contigs
  - Both overlap graph and de Bruijn assemblers initially collapse repeats into single copies
  - Uniqueness measured by a statistical test on coverage



# N50 size

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

Example: 1 Mbp genome



N50 size = 30 kbp

$$(300k+100k+45k+45k+30k = 520k \geq 500\text{ kbp})$$

Note:

N50 values are only meaningful to compare when base genome size is the same in all cases

# Break





# Outline

## I. Assembly theory

1. Assembly by analogy
2. De Bruijn and Overlap graph
3. Coverage, read length, errors, and repeats

## 2. Whole Genome Alignment

1. Aligning & visualizing with MUMmer

## 3. Genome assemblers

1. ALLPATHS-LG: recommended for Illumina-only projects
2. Celera Assembler: recommended for PacBio projects

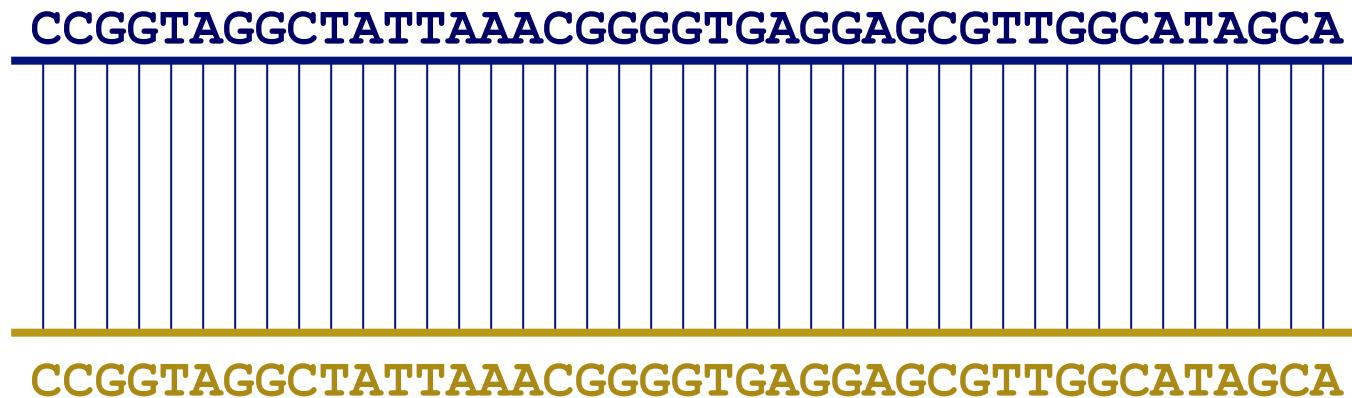


# Whole Genome Alignment with MUMmer

Slides Courtesy of Adam M. Phillippy  
University of Maryland

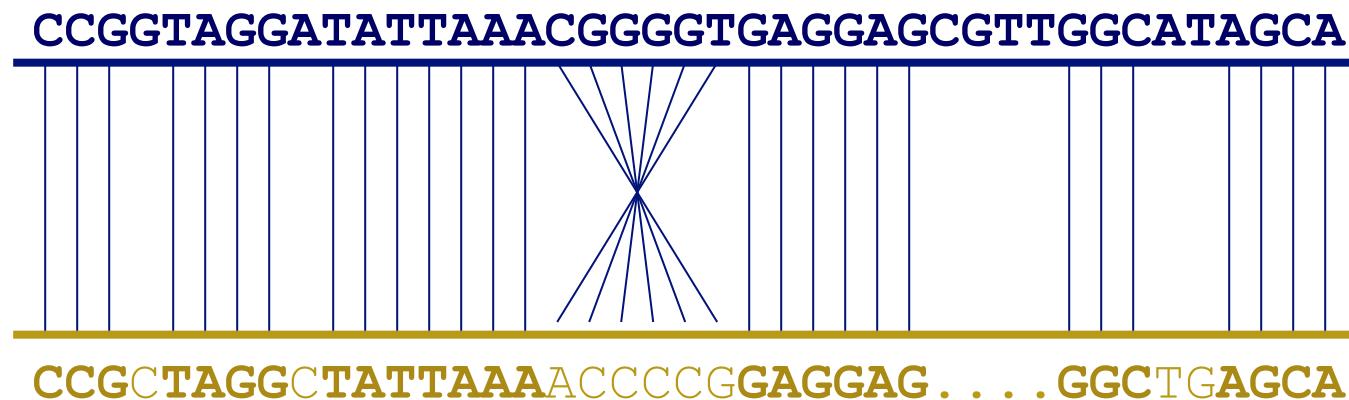
# Goal of WGA

- 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)



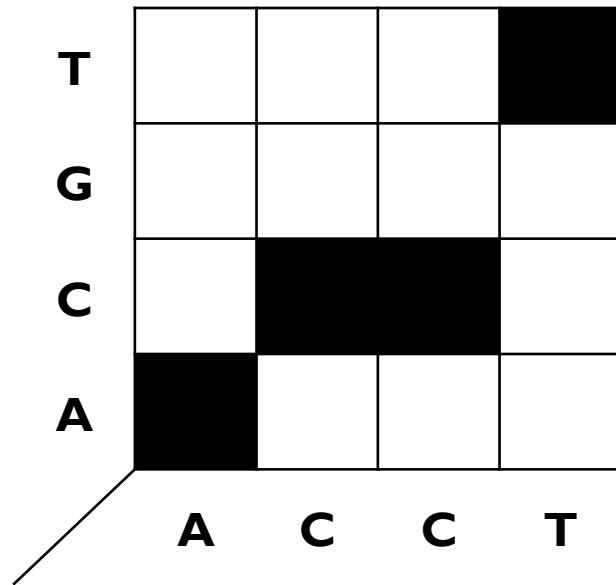
# WGA visualization

- How can we visualize *whole genome* alignments?

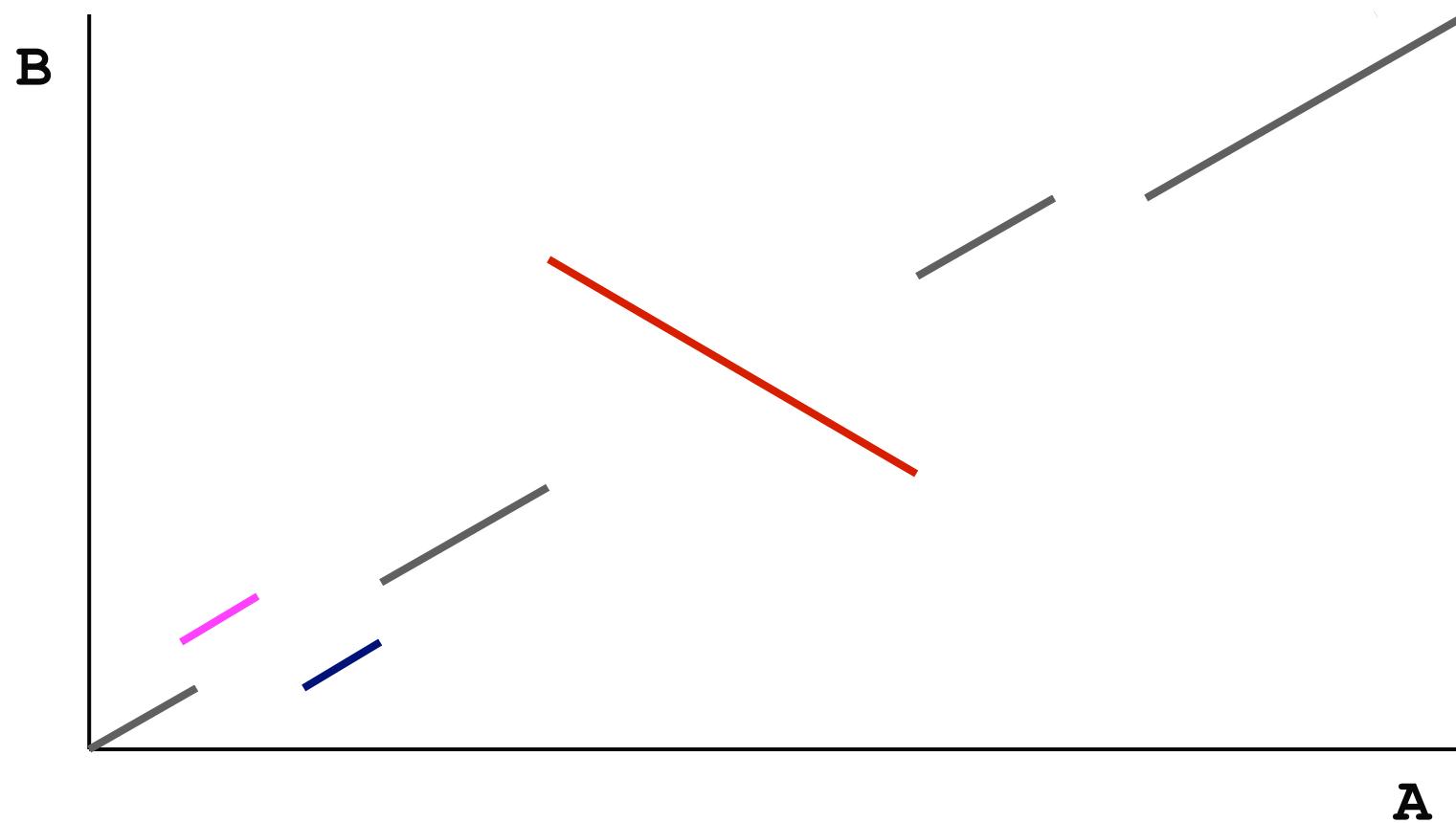
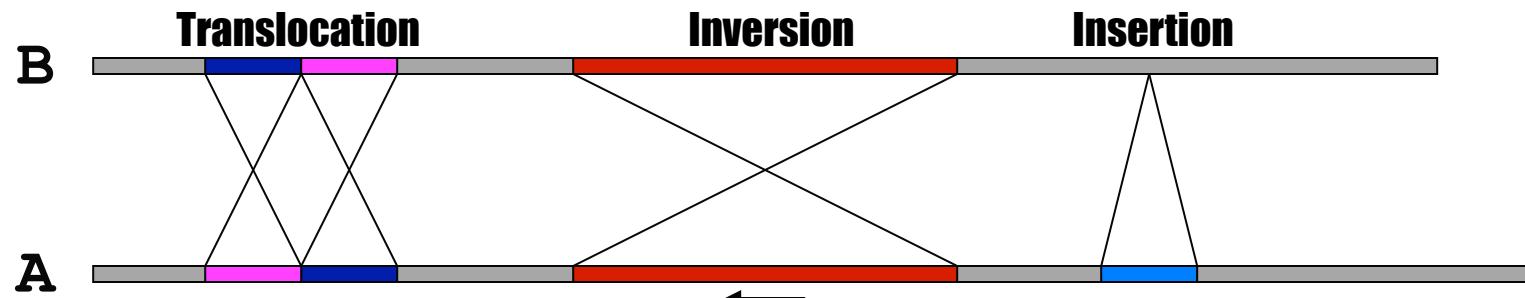
- With an alignment dot plot

- $N \times M$  matrix

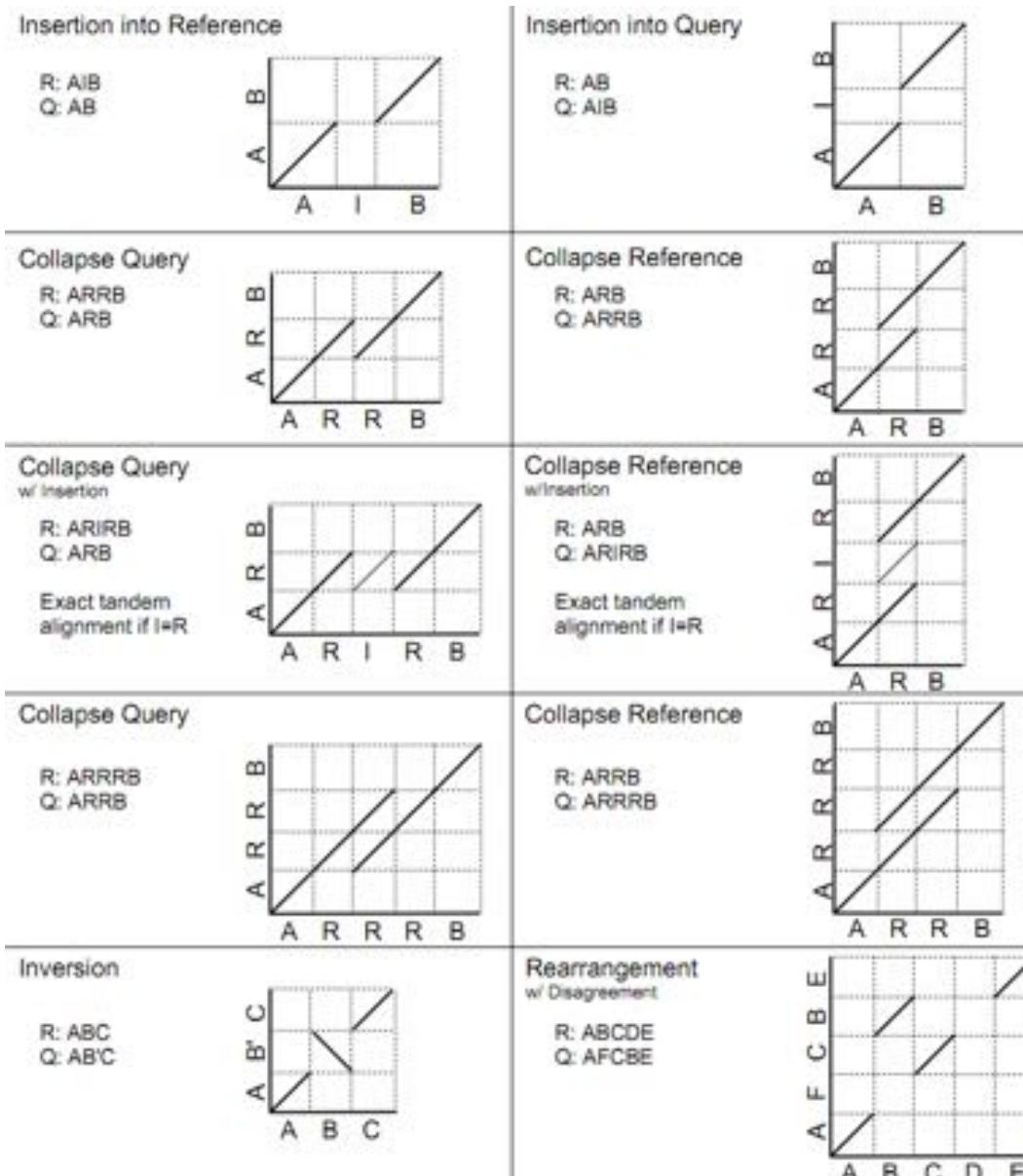
- Let  $i$  = position in genome A
    - Let  $j$  = position in genome B
    - Fill cell  $(i,j)$  if  $A_i$  shows similarity to  $B_j$



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



# 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](http://mummer.sf.net/manual/AlignmentTypes.pdf)

# Seed-and-extend with MUMmer

How can quickly align two genomes?

## I. Find maximal-unique-matches (MUMs)

- ◆ Match: exact match of a minimum length
- ◆ Maximal: cannot be extended in either direction without a mismatch
- ◆ Unique
  - ◆ occurs only once in both sequences (MUM)
  - ◆ occurs only once in a single sequence (MAM)
  - ◆ occurs one or more times in either sequence (MEM)

## 2. Cluster MUMs

- ◆ using size, gap and distance parameters

## 3. Extend clusters

- ◆ using modified Smith-Waterman algorithm

# WGA Alignment

**nucmer -maxmatch CO92.fasta KIM.fasta**

-maxmatch Find maximal exact matches (MEMs)

**delta-filter -m out.delta > out.filter.m**

-m Many-to-many mapping

**show-coords -r out.delta.m > out.coords**

-r Sort alignments by reference position

**dnadiff out.delta.m**

Construct catalog of sequence variations

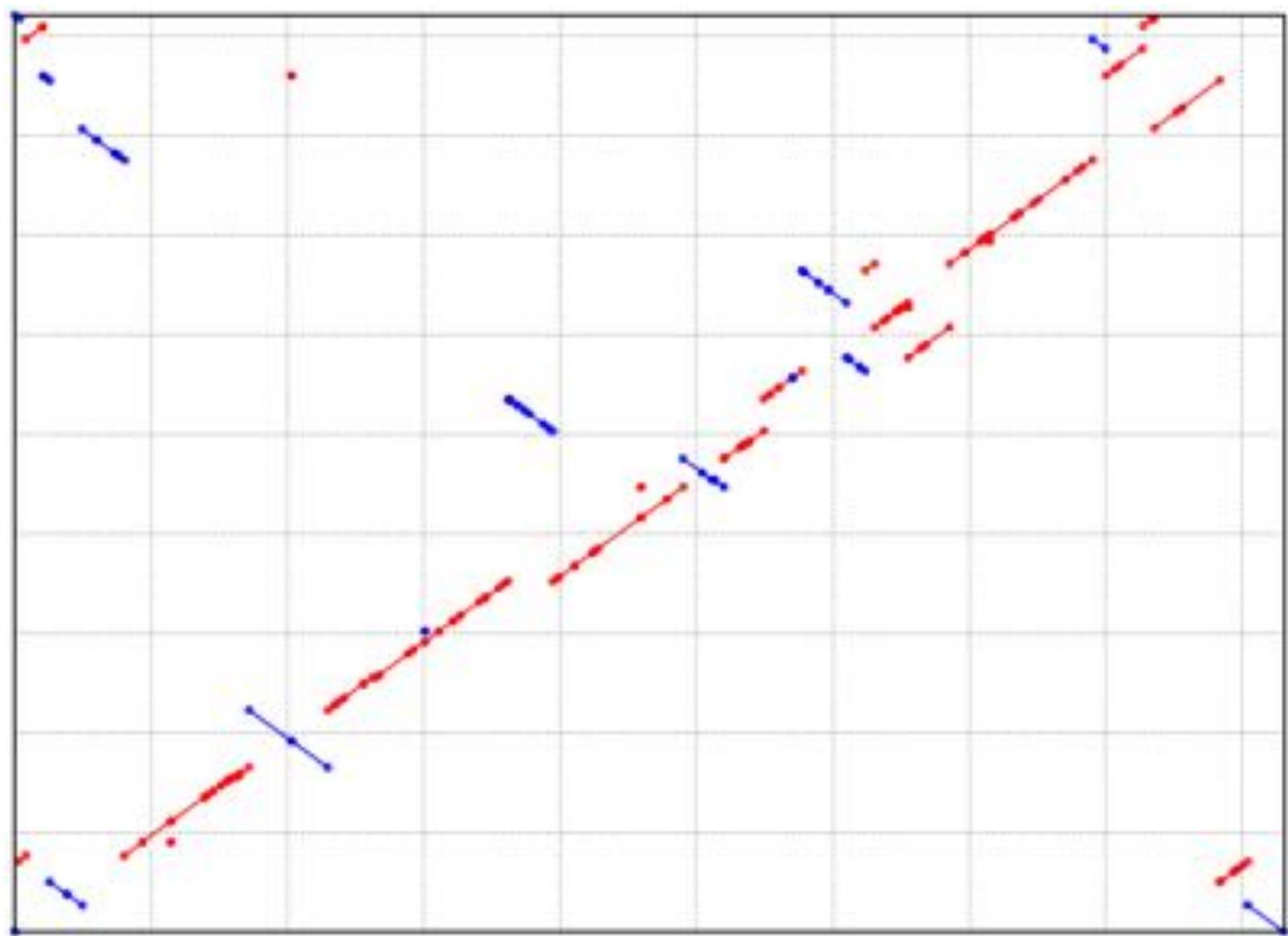
**mummerplot --large --layout out.delta.m**

--large Large plot

--layout Nice layout for multi-fasta files

--x11 Default, draw using x11 (--postscript, --png)

\*requires gnuplot





# Outline

## I. Assembly theory

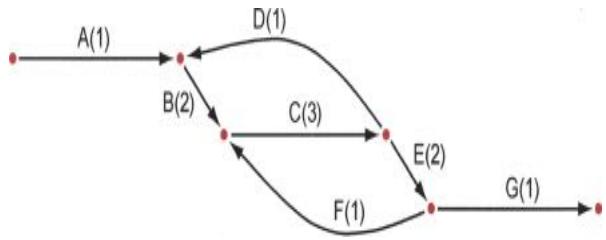
1. Assembly by analogy
2. De Bruijn and Overlap graph
3. Coverage, read length, errors, and repeats

## 2. Whole Genome Alignment

1. Aligning & visualizing with MUMmer

## 3. Genome assemblers

1. ALLPATHS-LG: recommended for Illumina-only projects
2. Celera Assembler: recommended for PacBio projects

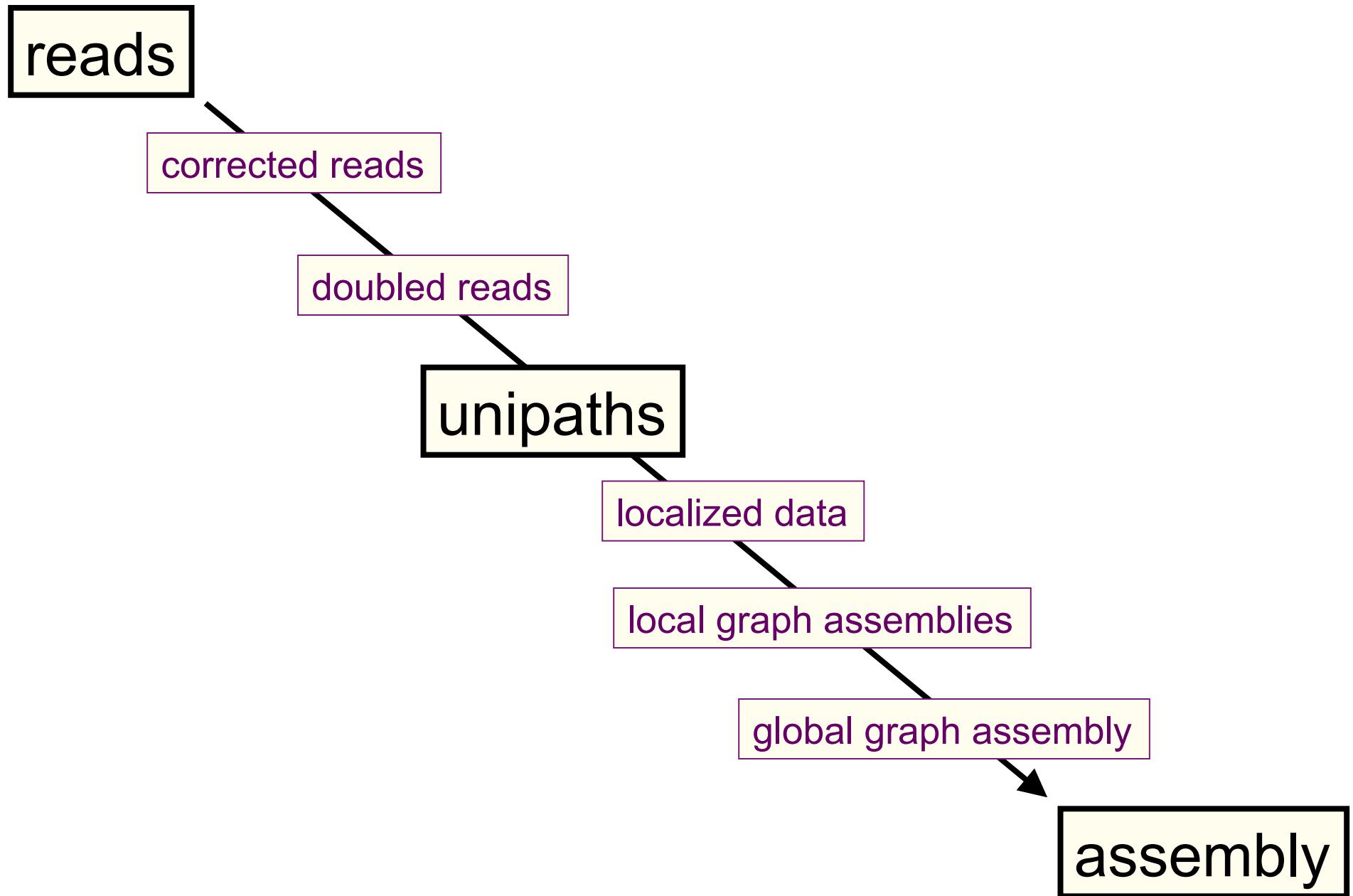


# Genome assembly with ALLPATHS-LG

## Iain MacCallum

## How ALLPATHS-LG works

---



## ALLPATHS-LG sequencing model

---

Libraries (insert types)	Fragment size (bp)	Read length (bases)	Sequence coverage (x)	Required
Fragment	180*	≥ 100	45	yes
Short jump	3,000	≥ 100 preferable	45	yes
Long jump	6,000	≥ 100 preferable	5	no**
Fosmid jump	40,000	≥ 26	1	no**

\*See next slide.

\*\*For best results. Normally not used for small genomes.  
However essential to assemble long repeats or duplications.

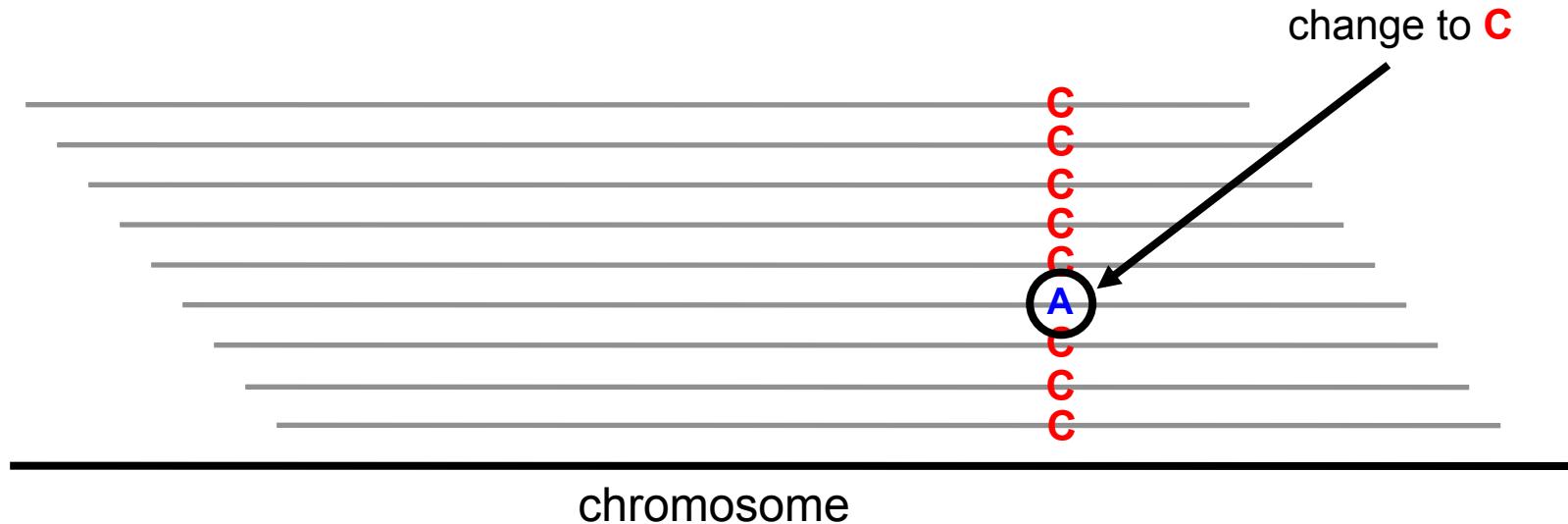
Cutting coverage in half still works, with some reduction in quality of results.

All: protocols are either available, or in progress.

## Error correction

---

Given a crystal ball, we could stack reads on the chromosomes they came from (with homologous chromosomes separate), then let each column ‘vote’:

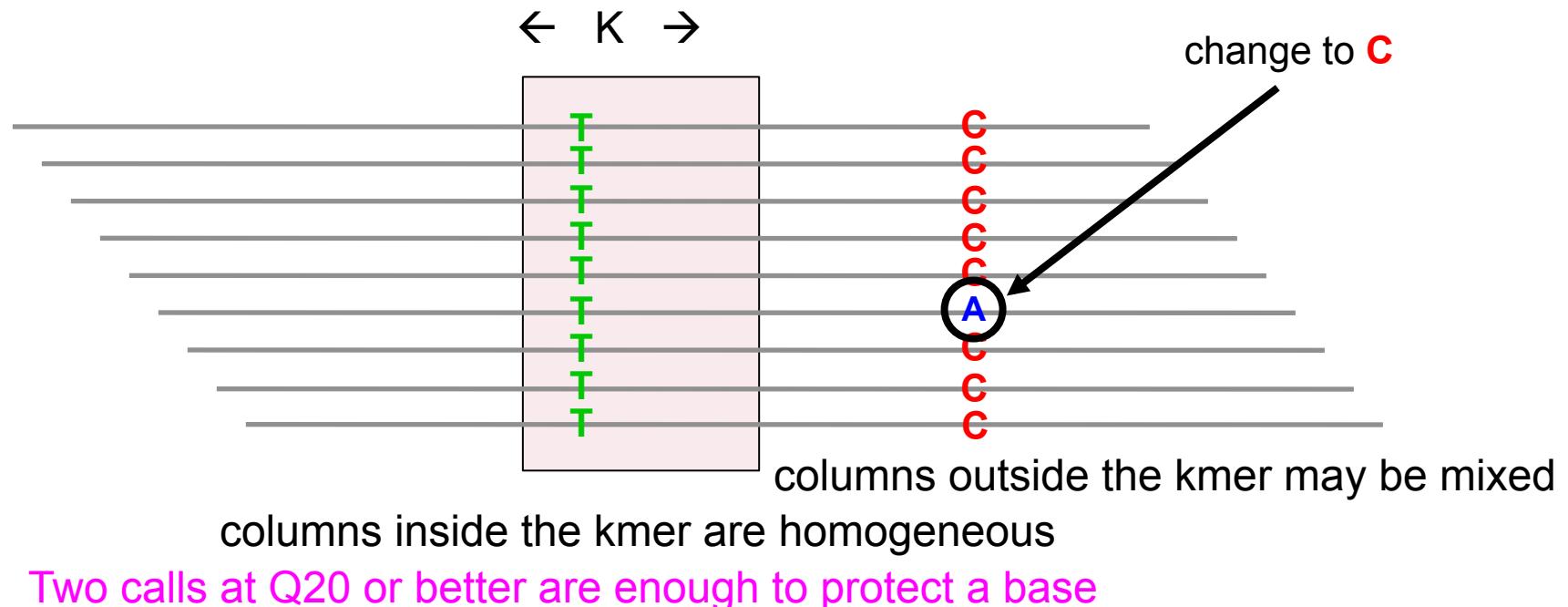


But we don't have a crystal ball....

## Error correction

---

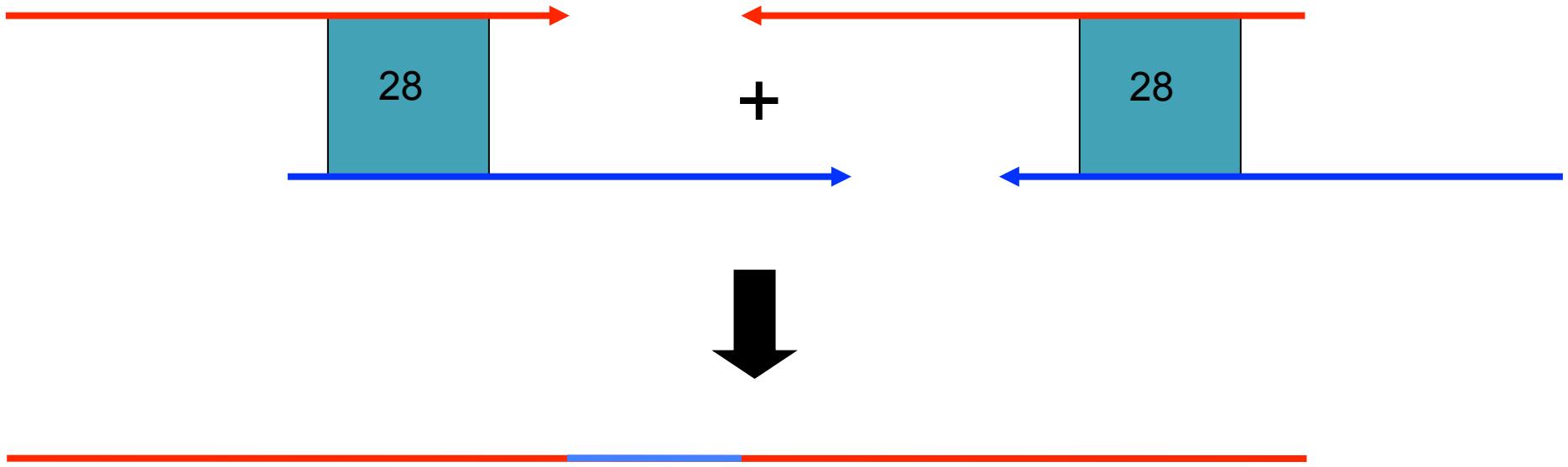
ALLPATHS-LG. For every K-mer, examine the stack of all reads containing the K-mer. Individual reads may be edited if they differ from the overwhelming consensus of the stack. If a given base on a read receives conflicting votes (arising from membership of the read in multiple stacks), it is not changed. (K=24)



## Read doubling

---

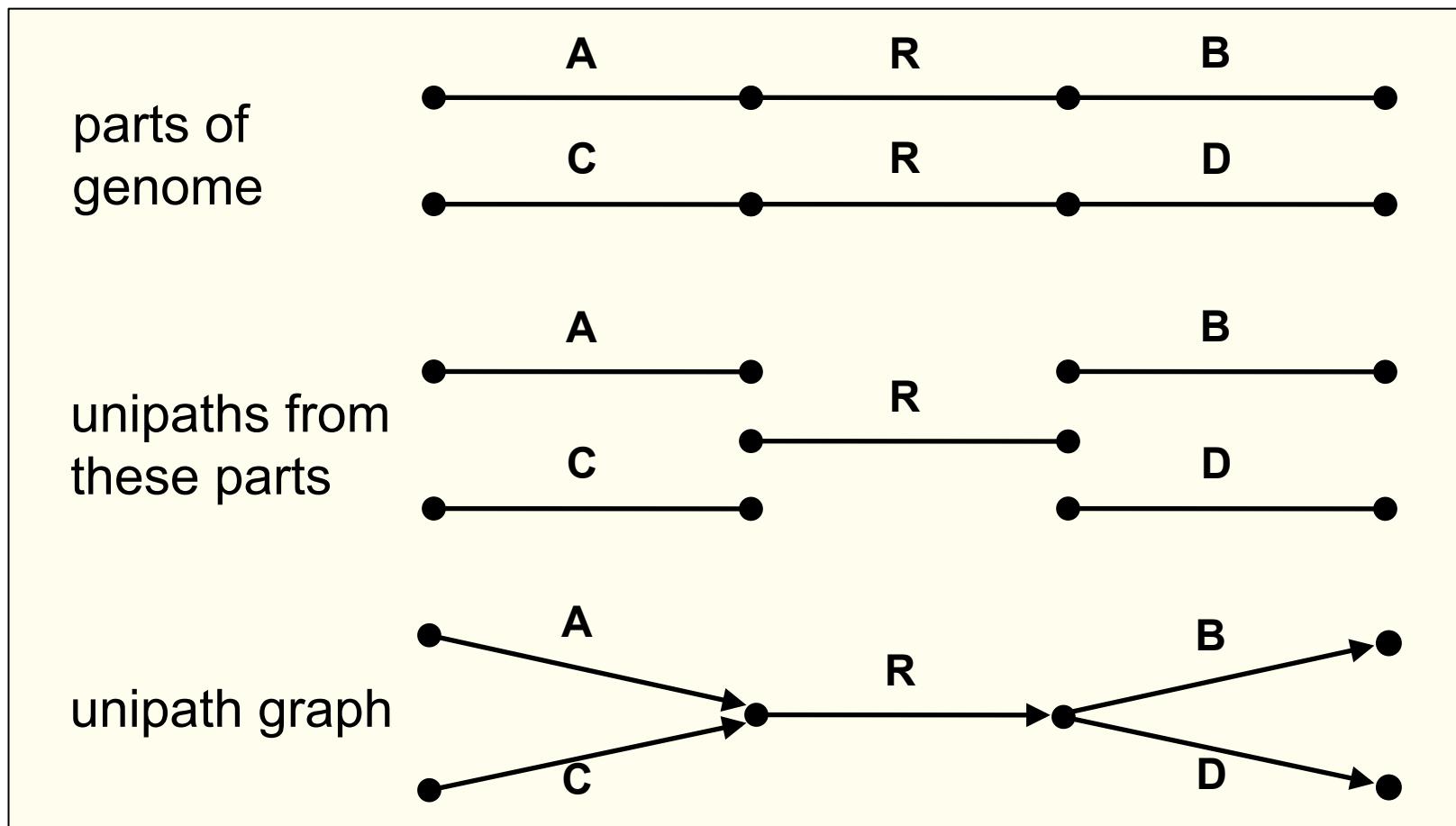
To close a read pair (red), we require the existence of another read pair (blue), overlapping perfectly like this:



More than one closure allowed (but rare).

## Unipaths

*Unipath*: unbranched part of genome – squeeze together perfect repeats of size  $\geq K$



Adjacent unipaths overlap by  $K-1$  bases

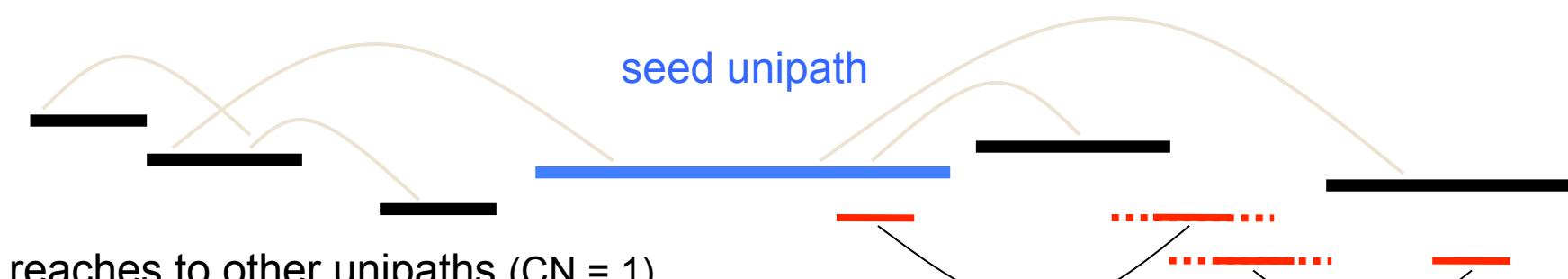
## Localization

---

- I. Find ‘seed’ unipaths, evenly spaced across genome**  
(ideally long, of copy number CN = 1)



- II. Form neighborhood around each seed**



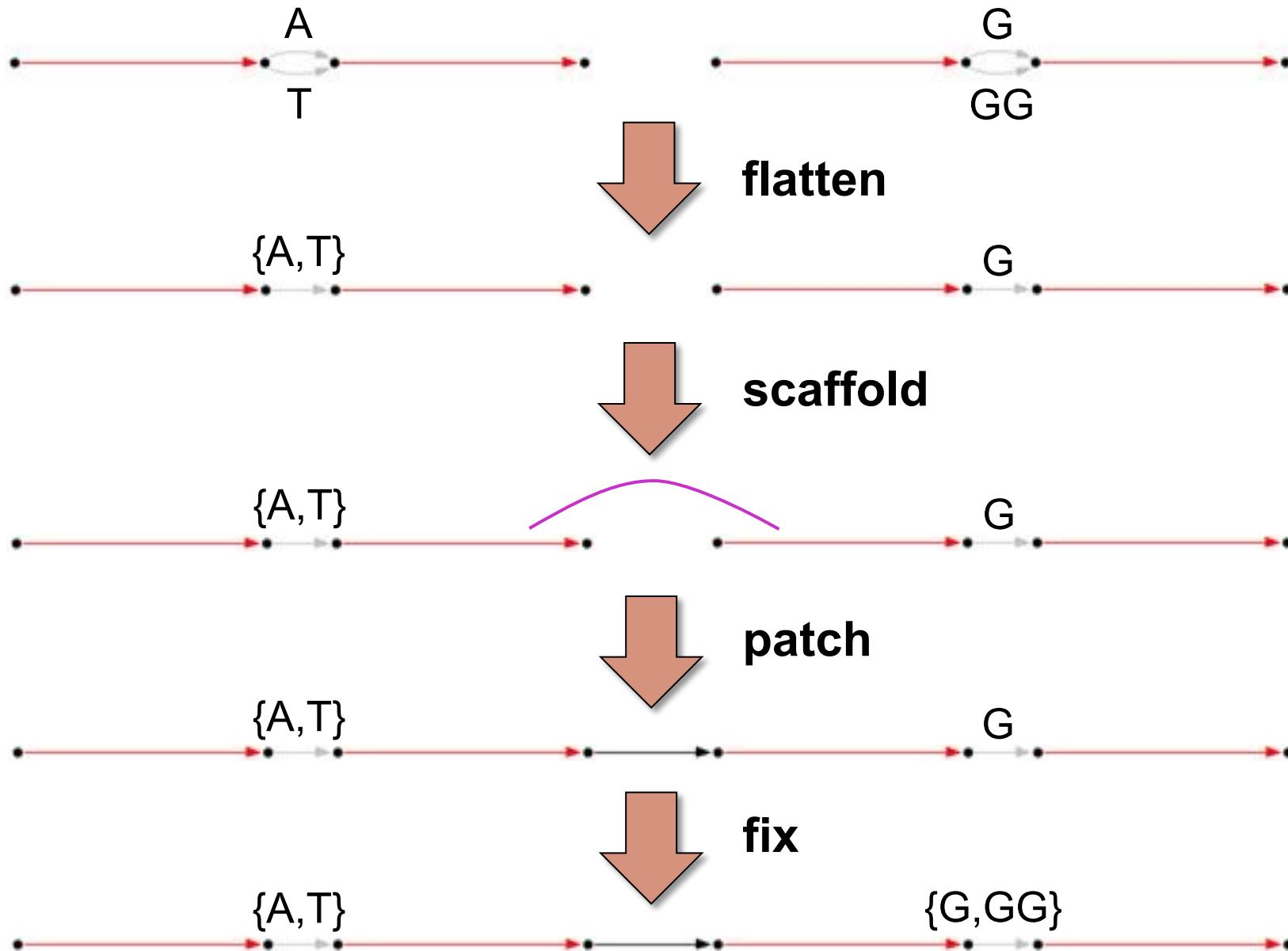
reaches to other unipaths (CN = 1)  
directly and indirectly

read pairs reach into repeats

and are extended by other  
unipaths

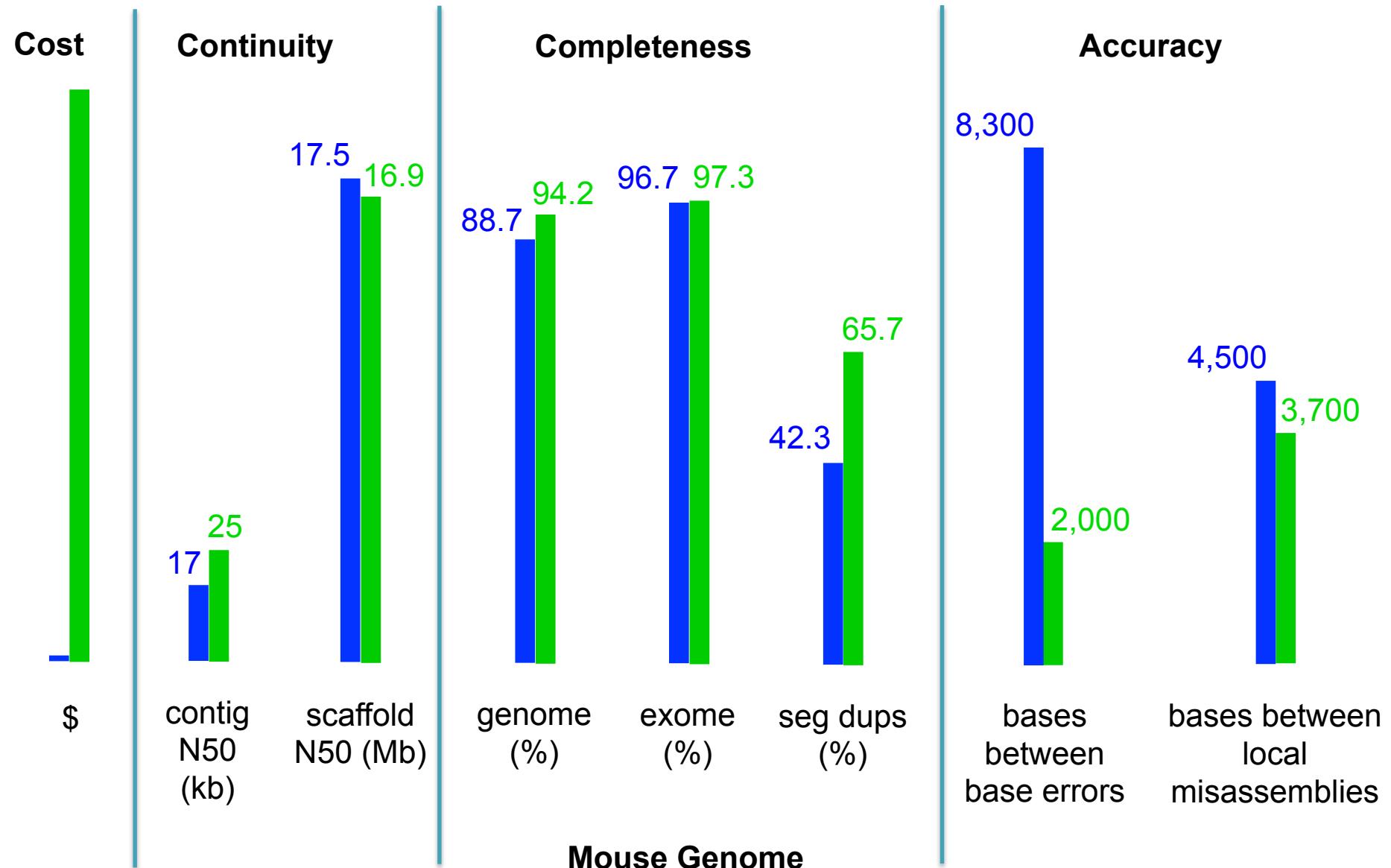
.....

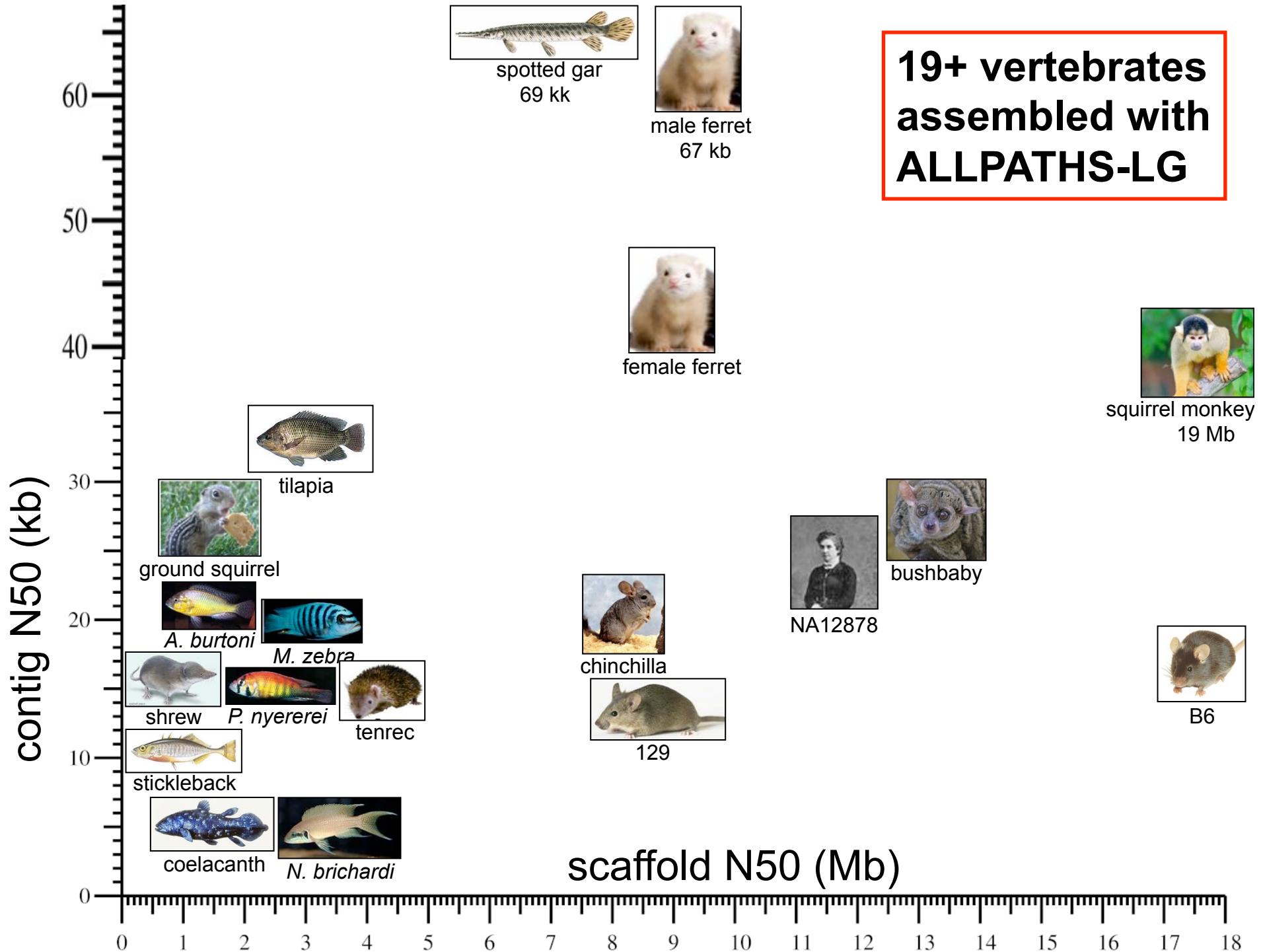
## Create assembly from global assembly graph

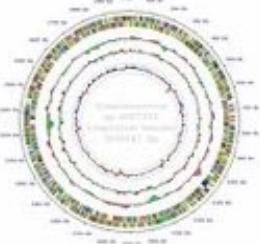




# Large genome recipe: ALLPATHS-LG vs capillary





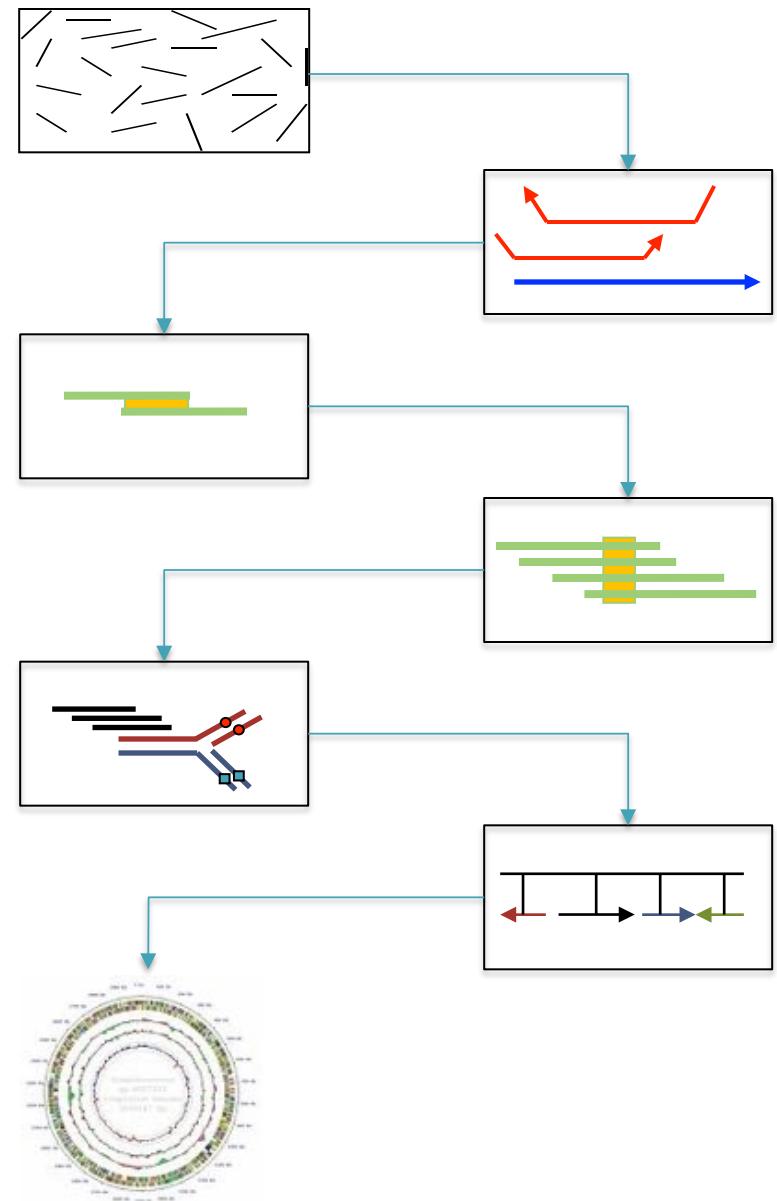


## **Genome assembly with the Celera Assembler**

# Celera Assembler

<http://wgs-assembler.sf.net>

1. Pre-overlap
  - Consistency checks
2. Trimming
  - Quality trimming & partial overlaps
3. Compute Overlaps
  - Find high quality overlaps
4. Error Correction
  - Evaluate difference in context of overlapping reads
5. Unitigging
  - Merge consistent reads
6. Scaffolding
  - Bundle mates, Order & Orient
7. Finalize Data
  - Build final consensus sequences



# Hybrid Sequencing



**Illumina**  
*Sequencing by Synthesis*

High throughput (60Gbp/day)  
High accuracy (~99%)  
Short reads (~100bp)

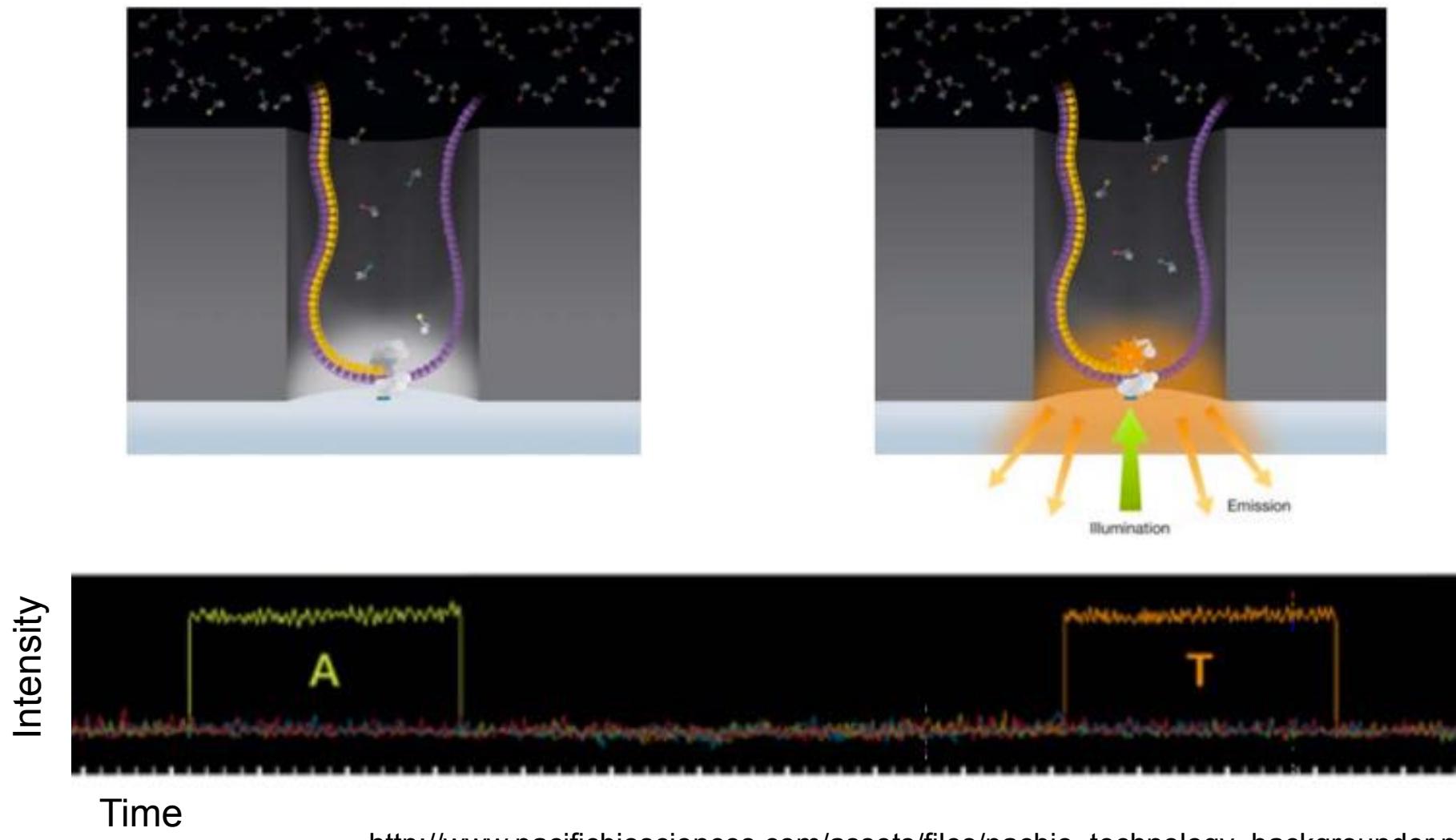


**Pacific Biosciences**  
*SMRT Sequencing*

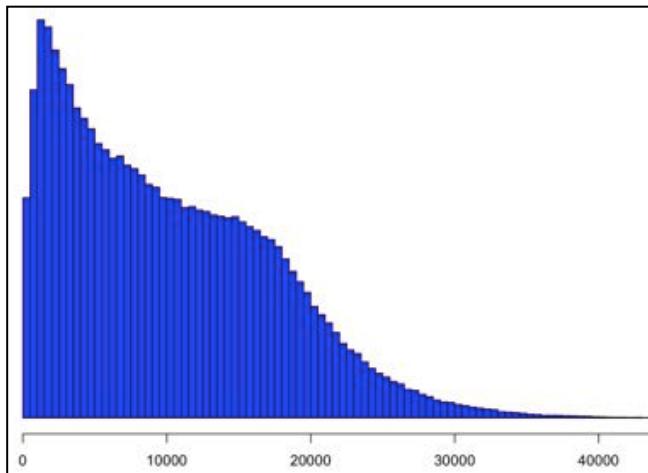
Lower throughput (600Mbp/day)  
Lower accuracy (~85%)  
Long reads (2-5kbp+)

# SMRT Sequencing

Imaging of fluorescently phospholinked labeled nucleotides as they are incorporated by a polymerase anchored to a Zero-Mode Waveguide (ZMW).



# SMRT Sequencing Data

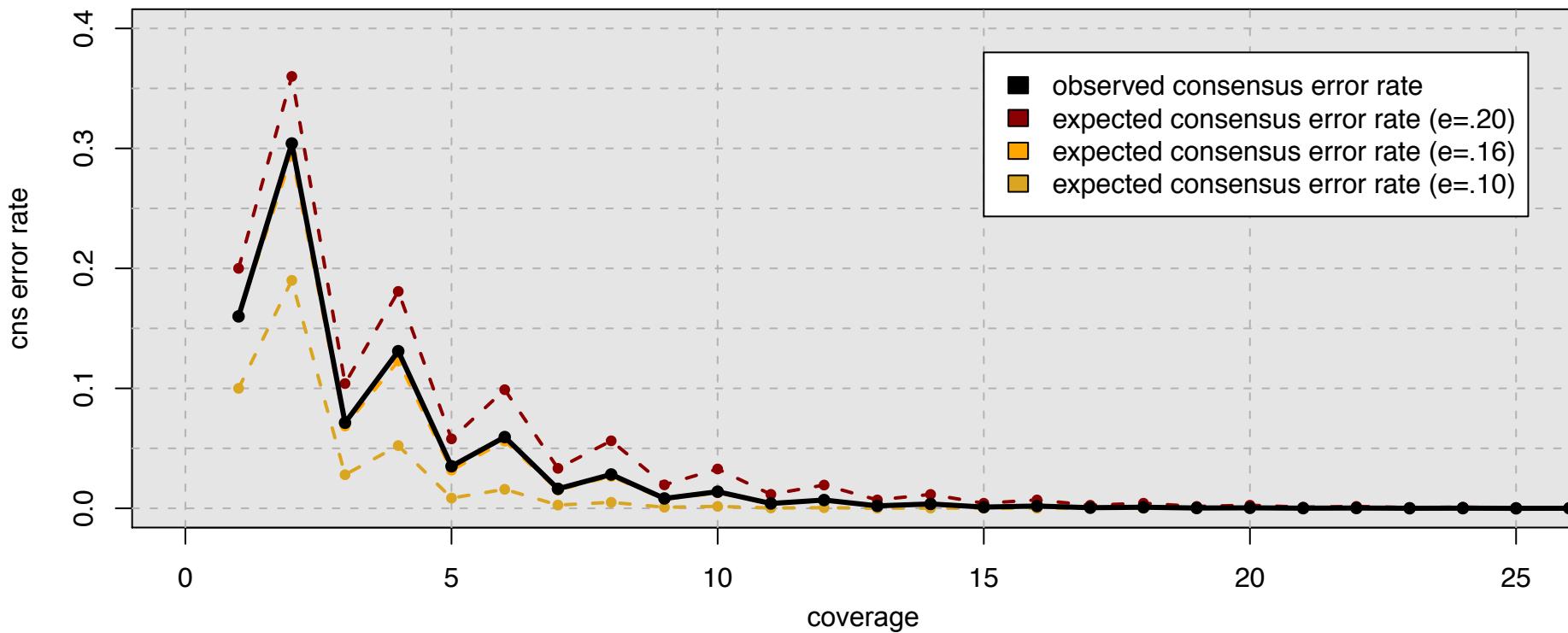


Match	83.7%
Insertions	11.5%
Deletions	3.4%
Mismatch	1.4%

TTGTAAGCAGTTGAAAACATATGTGTGGATTAGATAAAGAACATGAAAG  
||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| |||||  
TTGTAAGCAGTTGAAAACATATGTGT-GATTAG-ATAAAGAACATGGAAG  
  
ATTATAAA-CAGTTGATCCATT-AGAAGA-AAACGCAGGCGCTAGG  
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |  
A-TATAAAATCAGTTGATCCATTAGAA-AGAACGC-AAAGGC-GCTAGG  
  
CAACCTTGAAATGTAATCGCACTGAAAGAACAGATTTATTCCGCGCCCG  
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |  
C-ACCTTG-ATGT-AT--CACTGAAAGAACAGATTTATTCCGCGCCCG  
  
TAACGAATCAAGATTCTGAAAACACAT-ATAACAACCTCCAAAA-CACAA  
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |  
T-ACGAATC-AGATTCTGAAAACA-ATGAT---ACCTCCAAAAAGCACAA  
  
-AGGAGGGGAAAGGGGGGAATATCT-ATAAAAGATTACAAATTAGA-TGA  
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |  
GAGGAGG-AA-----GAATATCTGAT-AAAGATTACAAATT-GAGTGA  
  
ACT-AATTCACAA TA-AATAACACTTTA-ACAGAATTGAT-GGAA-GTT  
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |  
ACTAATTCACAA-ATAATAACACTTTAGACAAATTGATGGGAAGGTT  
  
TCGGAGAGATCCAACAAATGGGC-ATCGCCTTGAGTTAC-AATCAA  
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |  
TC-GAGAGATCC-AAACAAAT-GGC GATCG-CTTGACGTTACAATCAA  
  
ATCCAGTGGAAAATATAATTTATGCAATCCAGGAACCTATTACAATTAG  
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |  
ATCCAGT-GAAAATATA- TTATGC-ATCCA-GAACTTATTACAATTAG

Sample of 100k reads aligned with BLASR requiring >100bp alignment

# Consensus Accuracy and Coverage



Coverage can overcome random errors

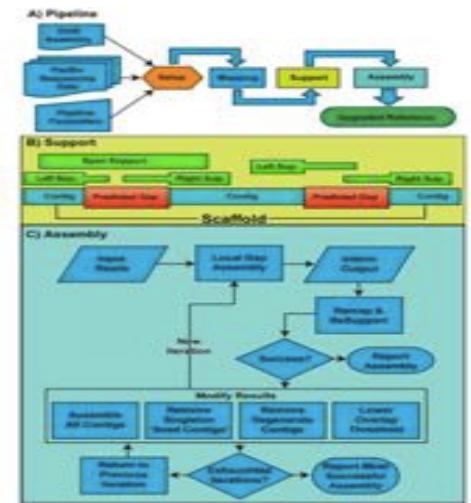
- Dashed: error model from binomial sampling
- Solid: observed accuracy

Koren, Schatz, et al (2012)  
Nature Biotechnology. 30:693–700

$$CNS\ Error = \sum_{i=\lceil c/2 \rceil}^c \binom{c}{i} (e)^i (1-e)^{n-i}$$

# PacBio Assembly Algorithms

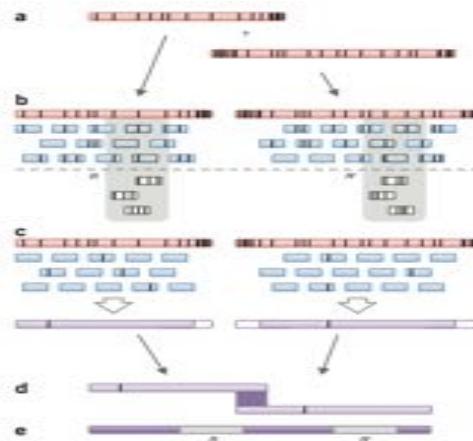
## PBJelly



### Gap Filling and Assembly Upgrade

English et al (2012)  
PLOS One. 7(11): e47768

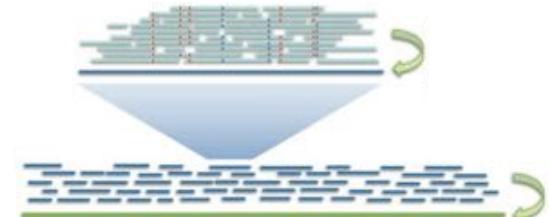
## PacBioToCA & ECTools



### Hybrid/PB-only Error Correction

Koren, Schatz, et al (2012)  
Nature Biotechnology. 30:693–700

## HGAP & Quiver



$$\Pr(R | T) = \prod_k \Pr(R_k | T)$$

Pr(R | T)

Pr(R | T) =  $\prod_k \Pr(R_k | T)$

$T$

$R_1$ ,  $R_2$ , ...,  $R_n$

	Initial Assembly	Quiver Consensus
QV	43.4	54.5
Accuracy	99.99540%	99.99964%
Differences	141	11

### PB-only Correction & Polishing

Chin et al (2013)  
Nature Methods. 10:563–569

< 5x

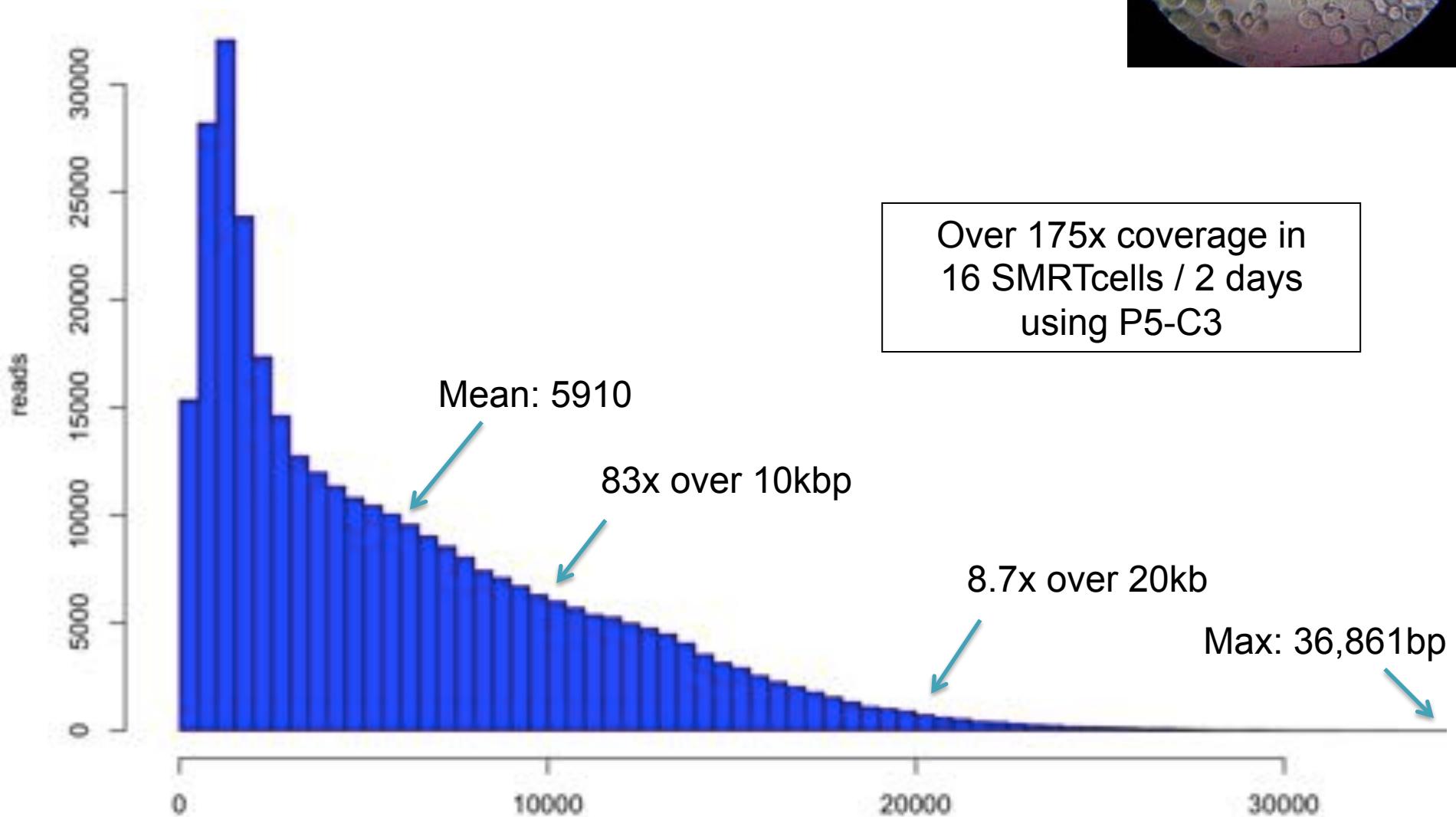
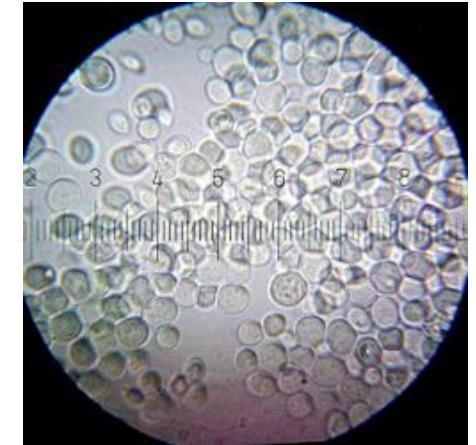
PacBio Coverage

> 50x

# *S. cerevisiae* W303

PacBio RS II sequencing at CSHL by Dick McCombie

- Size selection using an 7 Kb elution window on a BluePippin™ device from Sage Science



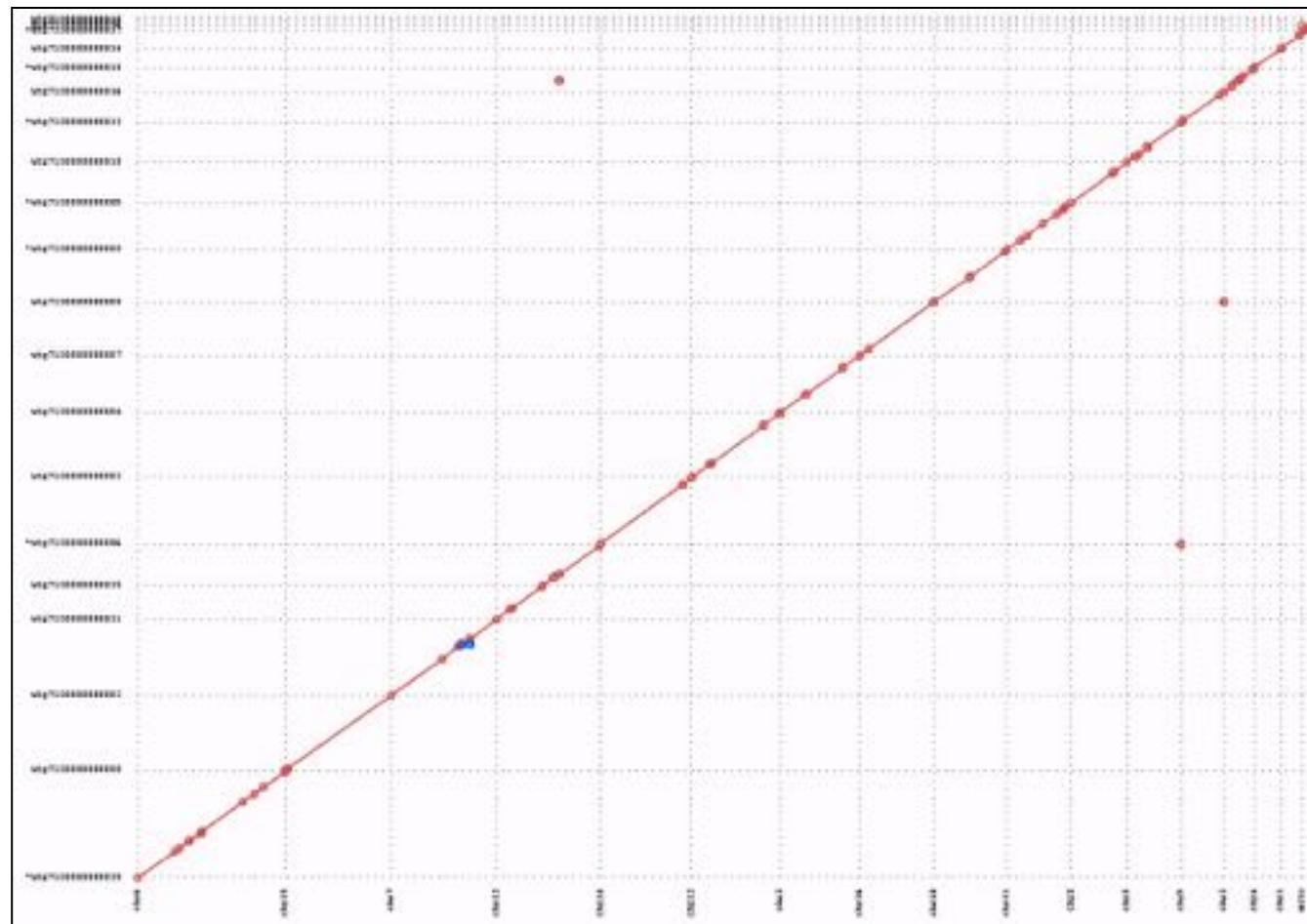
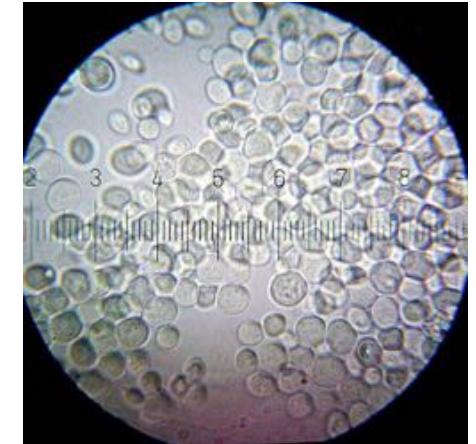
# *S. cerevisiae* W303

S288C Reference sequence

- 12.1Mbp; 16 chromo + mitochondria; N50: 924kbp

PacBio assembly using HGAP + Celera Assembler

- 12.4Mbp; 21 non-redundant contigs; N50: 811kbp; >99.8% id



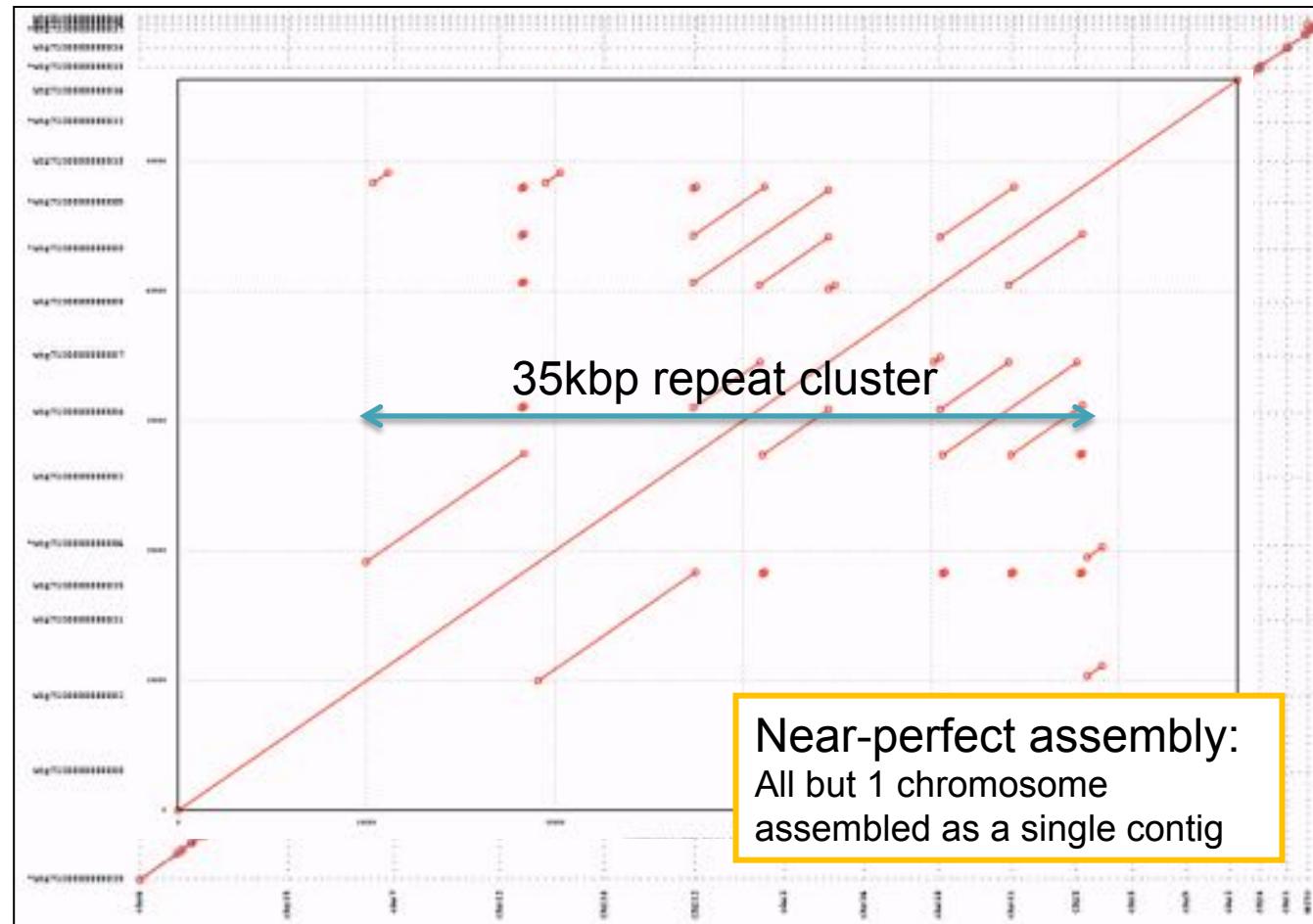
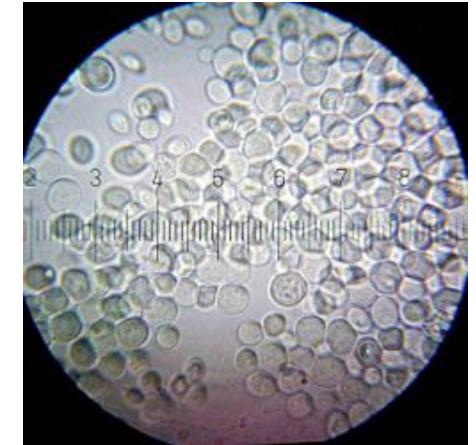
# *S. cerevisiae* W303

S288C Reference sequence

- 12.1Mbp; 16 chromo + mitochondria; N50: 924kbp

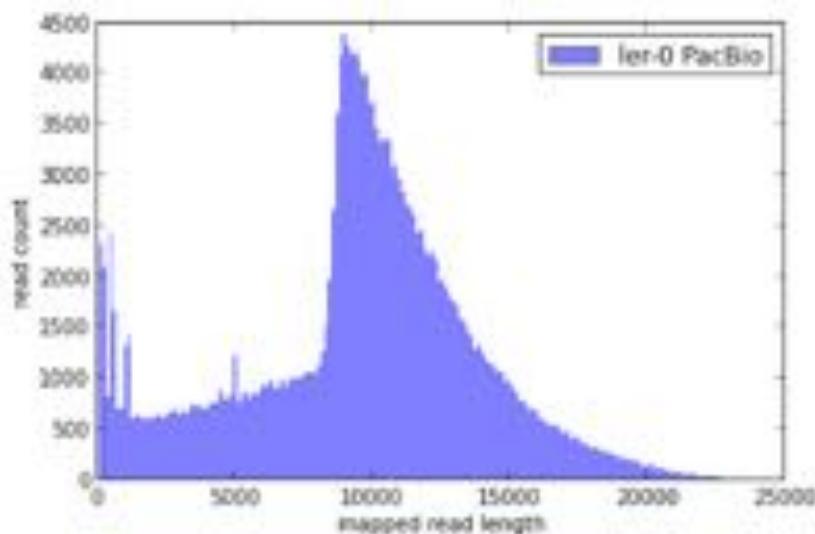
PacBio assembly using HGAP + Celera Assembler

- 12.4Mbp; 21 non-redundant contigs; N50: 811kbp; >99.8% id



# *A. thaliana* Ler-0

<http://blog.pacificbiosciences.com/2013/08/new-data-release-arabidopsis-assembly.html>



Genome size: 124.6 Mbp  
Chromosome N50: 23.0 Mbp  
Corrected coverage: 20x over 10kb

## *A. thaliana* Ler-0 sequenced at PacBio

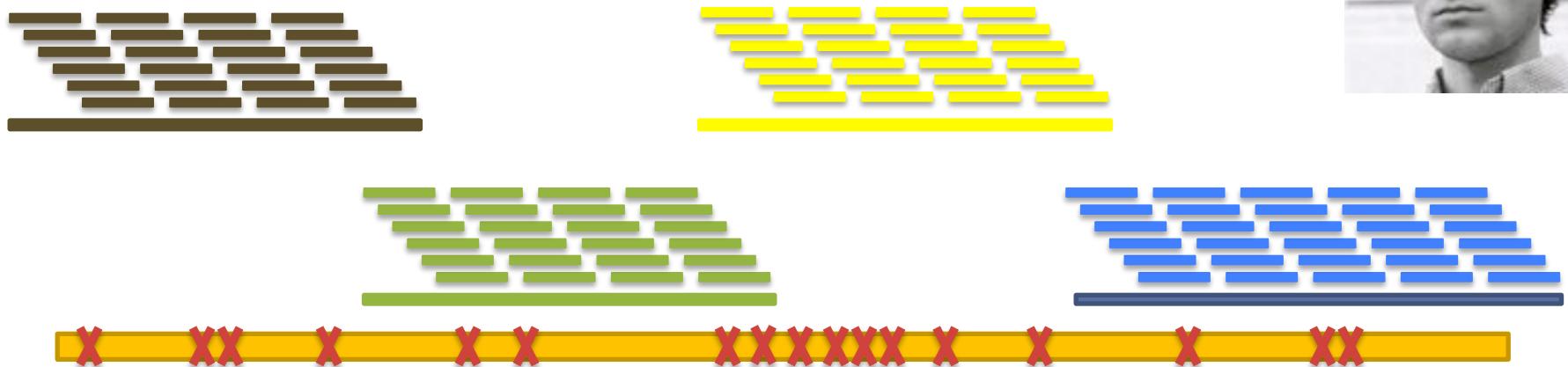
- Sequenced using the previous P4 enzyme and C2 chemistry
- Size selection using an 8 Kb to 50 Kb elution window on a BluePippin™ device from Sage Science
- Total coverage >119x

Sum of Contig Lengths: 149.5Mb  
N50 Contig Length: 8.4 Mb  
Number of Contigs: 1788

High quality assembly of chromosome arms  
Assembly Performance: 8.4Mbp/23Mbp = 36%  
MiSeq assembly: 63kbp/23Mbp = .2%

# ECTools: Error Correction with pre-assembled reads

<https://github.com/jgurtowski/ectools>



**Short Reads -> Assemble Unitigs -> Align & Select -> Error Correct**

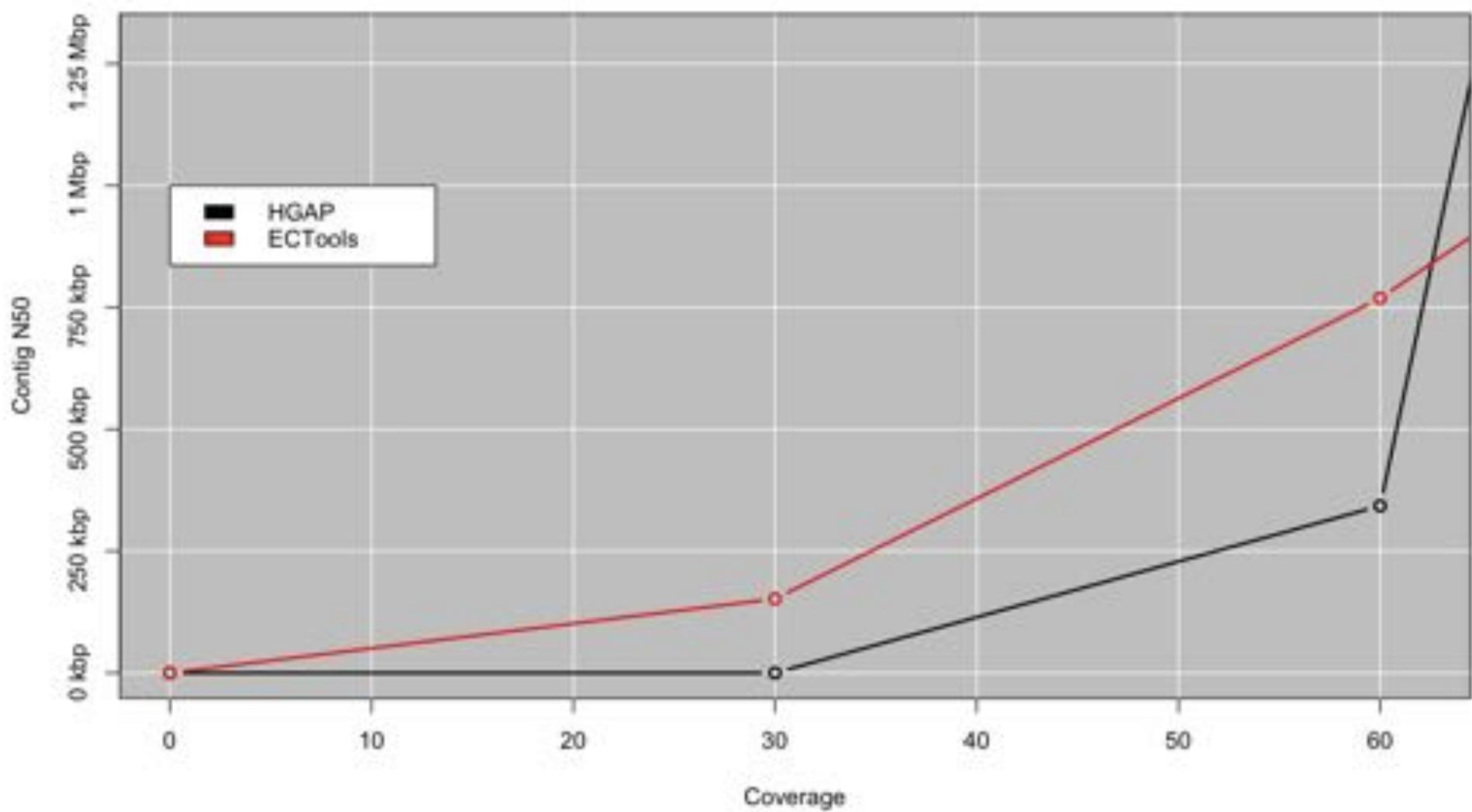
Can Help us overcome:

1. Error Dense Regions – Longer sequences have more seeds to match
2. Simple Repeats – Longer sequences easier to resolve

**However, cannot overcome Illumina coverage gaps & other biases**

# *A. thaliana* Ler-0

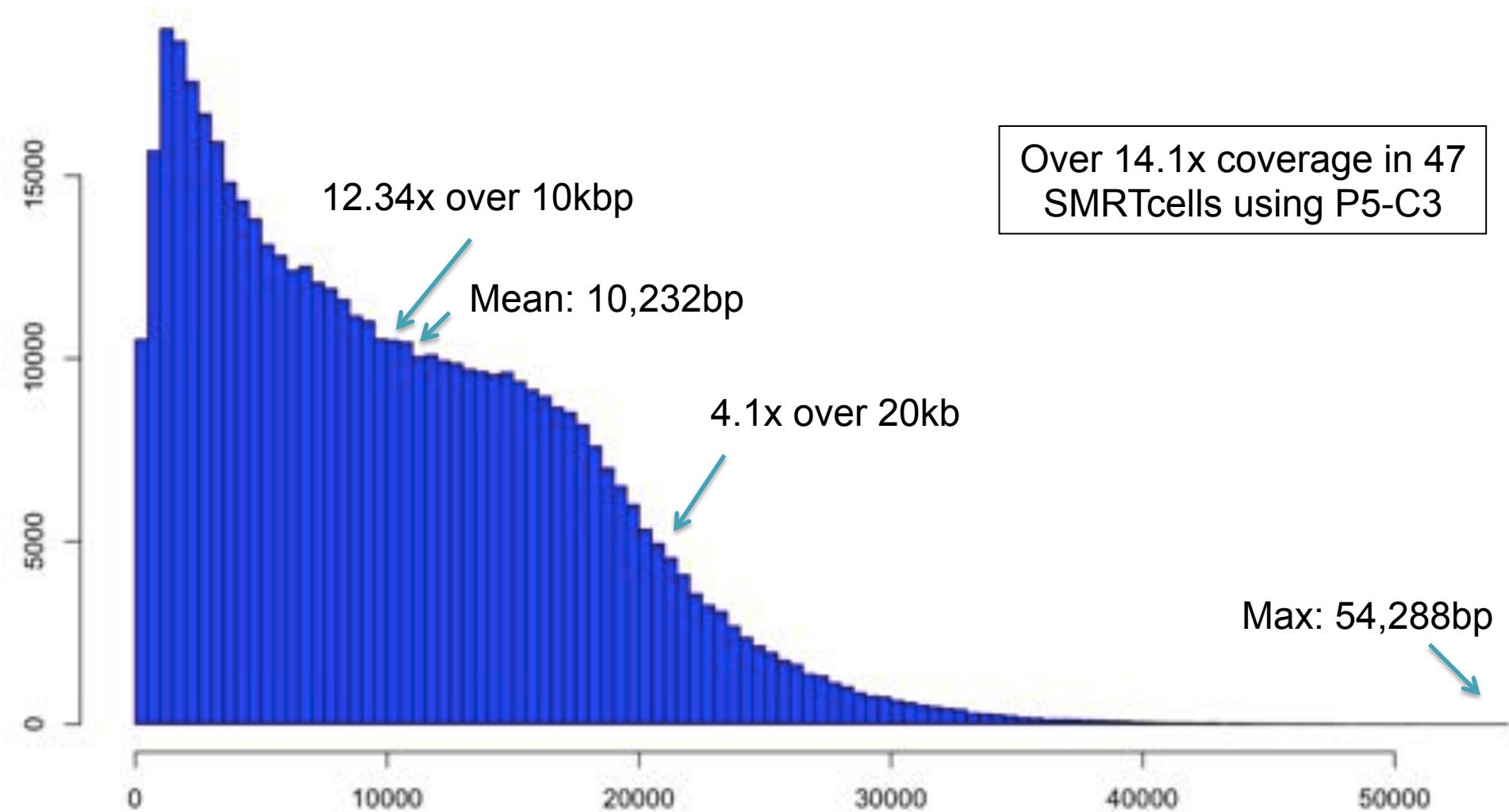
<http://blog.pacificbiosciences.com/2013/08/new-data-release-arabidopsis-assembly.html>



# *O. sativa* pv Indica (IR64)

PacBio RS II sequencing at PacBio

- Size selection using an 10 Kb elution window on a BluePippin™ device from Sage Science



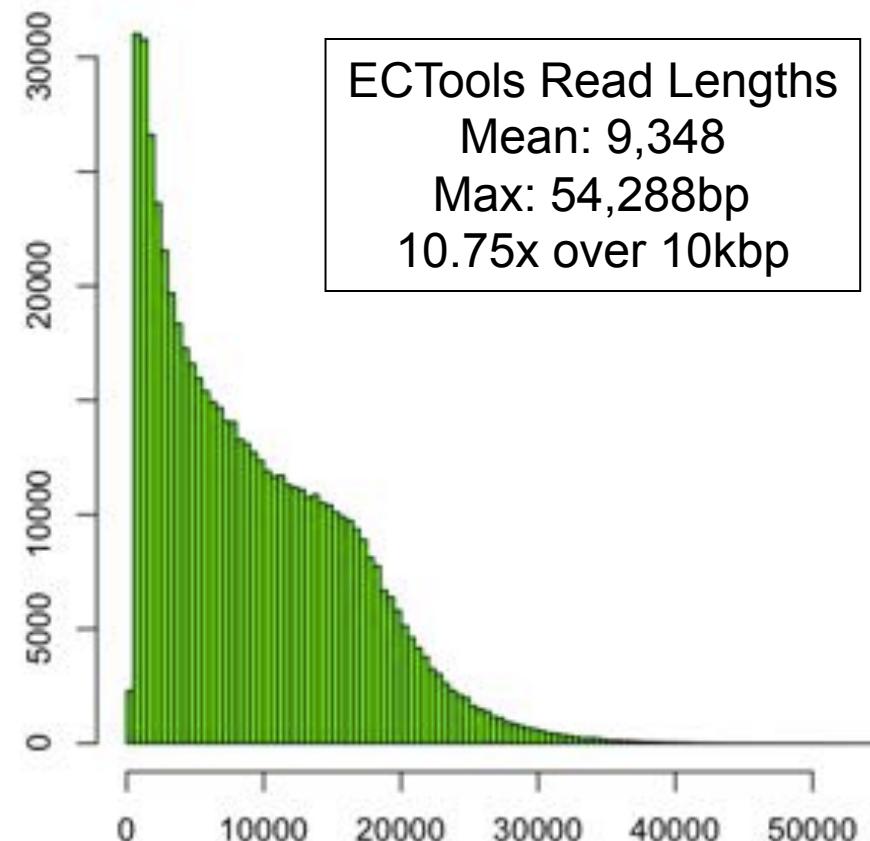
# *O. sativa* pv Indica (IR64)

Genome size: ~370 Mb

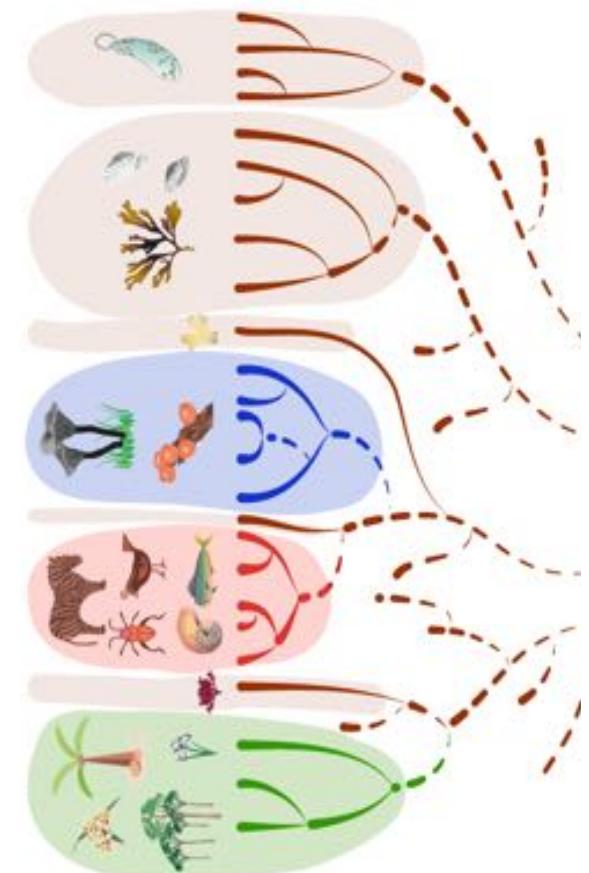
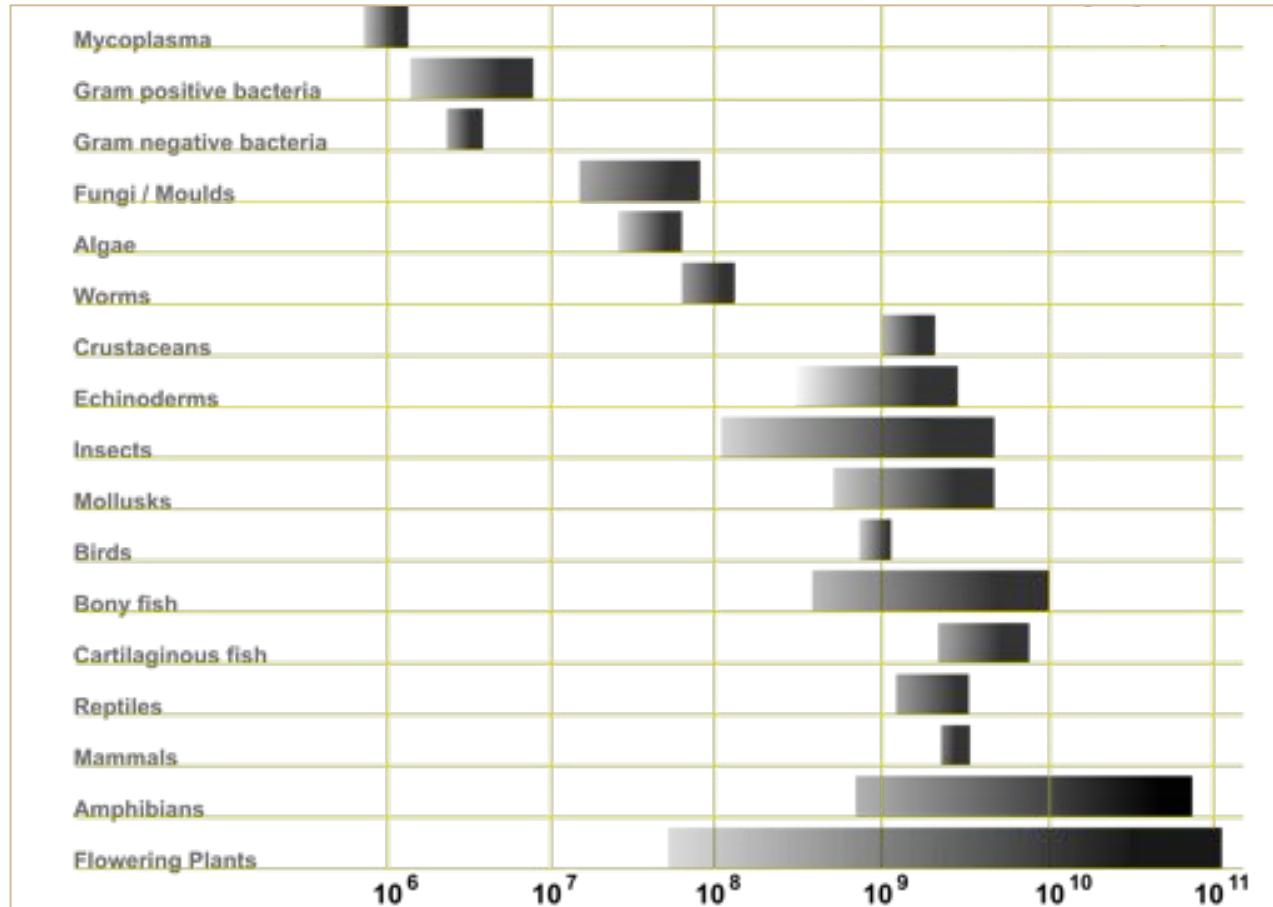
Chromosome N50: ~29.7 Mbp



Assembly	Contig NG50
MiSeq Fragments 25x 456bp (3 runs 2x300 @ 450 FLASH)	19,078
“ALLPATHS-recipe” 50x 2x100bp @ 180 36x 2x50bp @ 2100 51x 2x50bp @ 4800	18,450
ECTools 10.7x @ 10kbp	271,885

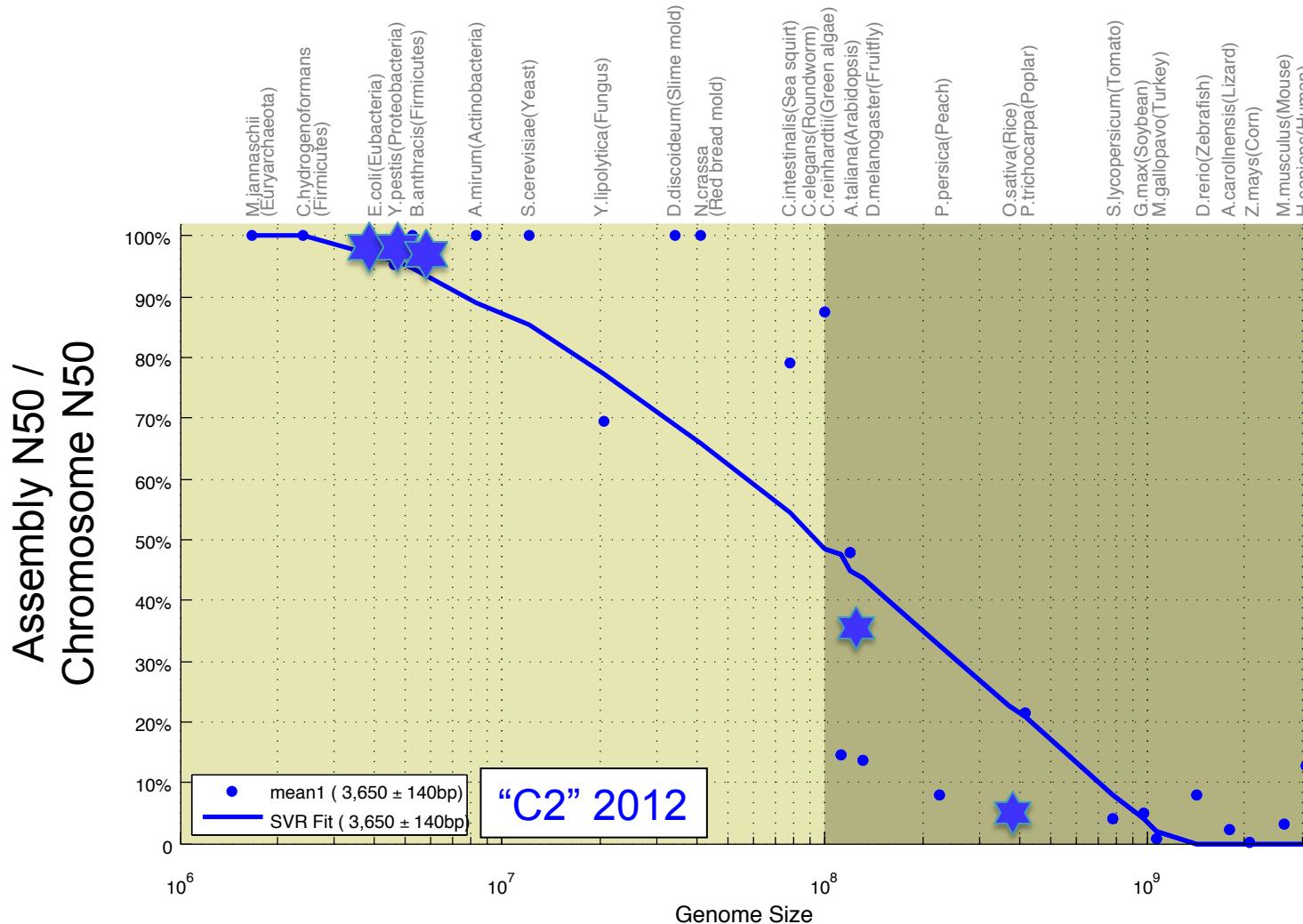


# What should we expect from an assembly?



[https://en.wikipedia.org/wiki/Genome\\_size](https://en.wikipedia.org/wiki/Genome_size)

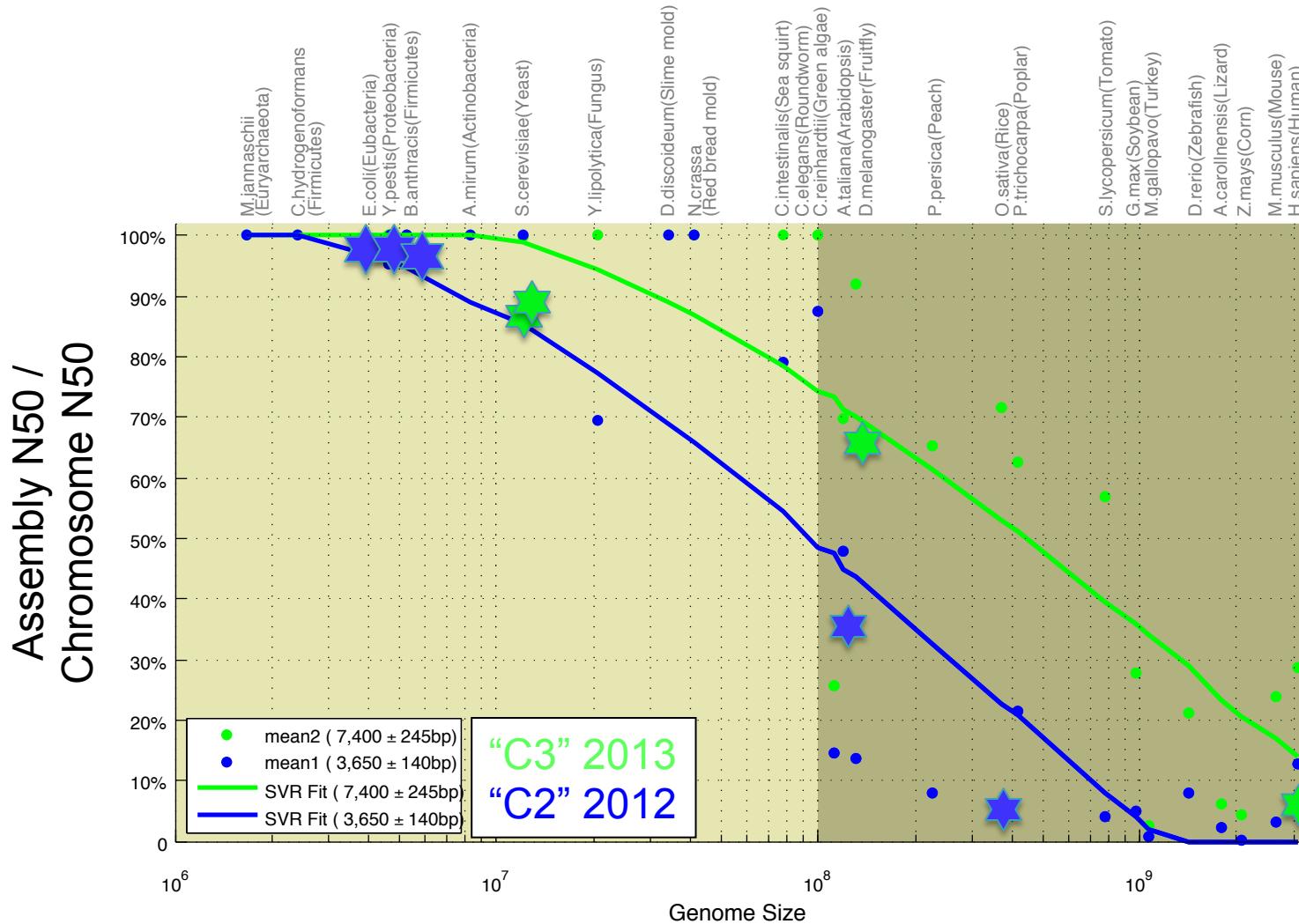
# Assembly Complexity of Long Reads



## Assembly complexity of long read sequencing

Lee, H\*, Gurtowski, J\*, Yoo, S, Marcus, S, McCombie, WR, Schatz MC (2014) *In preparation*

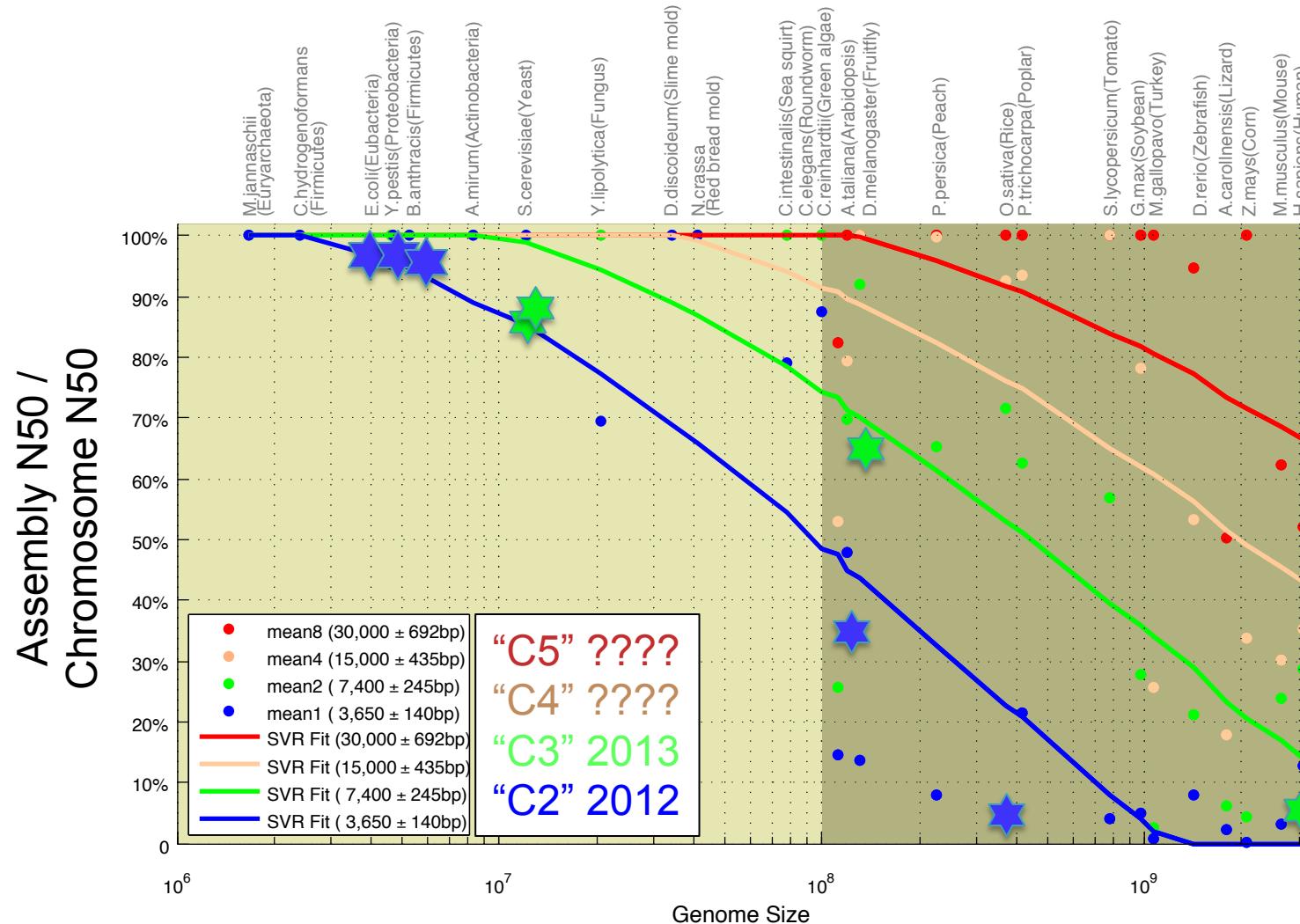
# Assembly Complexity of Long Reads



## Assembly complexity of long read sequencing

Lee, H\*, Gurtowski, J\*, Yoo, S, Marcus, S, McCombie, WR, Schatz MC (2014) *In preparation*

# Assembly Complexity of Long Reads



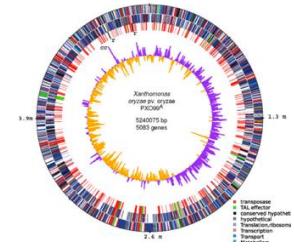
## Assembly complexity of long read sequencing

Lee, H\*, Gurtowski, J\*, Yoo, S, Marcus, S, McCombie, WR, Schatz MC. (2014) *In preparation*

# Assembly Recommendations

- **Long read sequencing of eukaryotic genomes is here**
- **Recommendations**
  - < 100 Mbp: HGAP/PacBio2CA @ 100x PB C3-P5  
expect near perfect chromosome arms
  - < 1GB: HGAP/PacBio2CA @ 100x PB C3-P5  
expect high quality assembly: contig N50 over 1Mbp
  - > 1GB: hybrid/gap filling  
expect contig N50 to be 100kbp – 1Mbp
  - > 5GB: Email [mschatz@cshl.edu](mailto:mschatz@cshl.edu)
- **Caveats**
  - Model only as good as the available references (esp. haploid sequences)
  - Technologies are quickly improving, exciting new scaffolding technologies

# 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

# Acknowledgements

## Schatz Lab

Giuseppe Narzisi  
Shoshana Marcus  
James Gurtowski  
Alejandro Wences  
Hayan Lee  
Rob Aboukhalil  
Mitch Bekritsky  
Charles Underwood  
Rushil Gupta  
Avijit Gupta  
Shishir Horane  
Deepak Nettem  
Varrun Ramani  
Piyush Kansal  
Eric Biggers  
Aspyn Palatnick

## CSHL

Hannon Lab  
Gingeras Lab  
Iossifov Lab  
Levy Lab  
Lippman Lab  
Lyon Lab  
Martienssen Lab  
McCombie Lab  
Ware Lab  
Wigler Lab

IT Department

## NBACC

Adam Phillippy  
Sergey Koren



# Thank You!

<http://schatzlab.cshl.edu>

@mike\_schatz

