

Genome Sequencing & Assembly

Michael Schatz

July 6, 2014

Frontiers of techniques in plant sciences





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

4. Summary & Recommendations



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

4. Summary and Recommendations

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 = 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

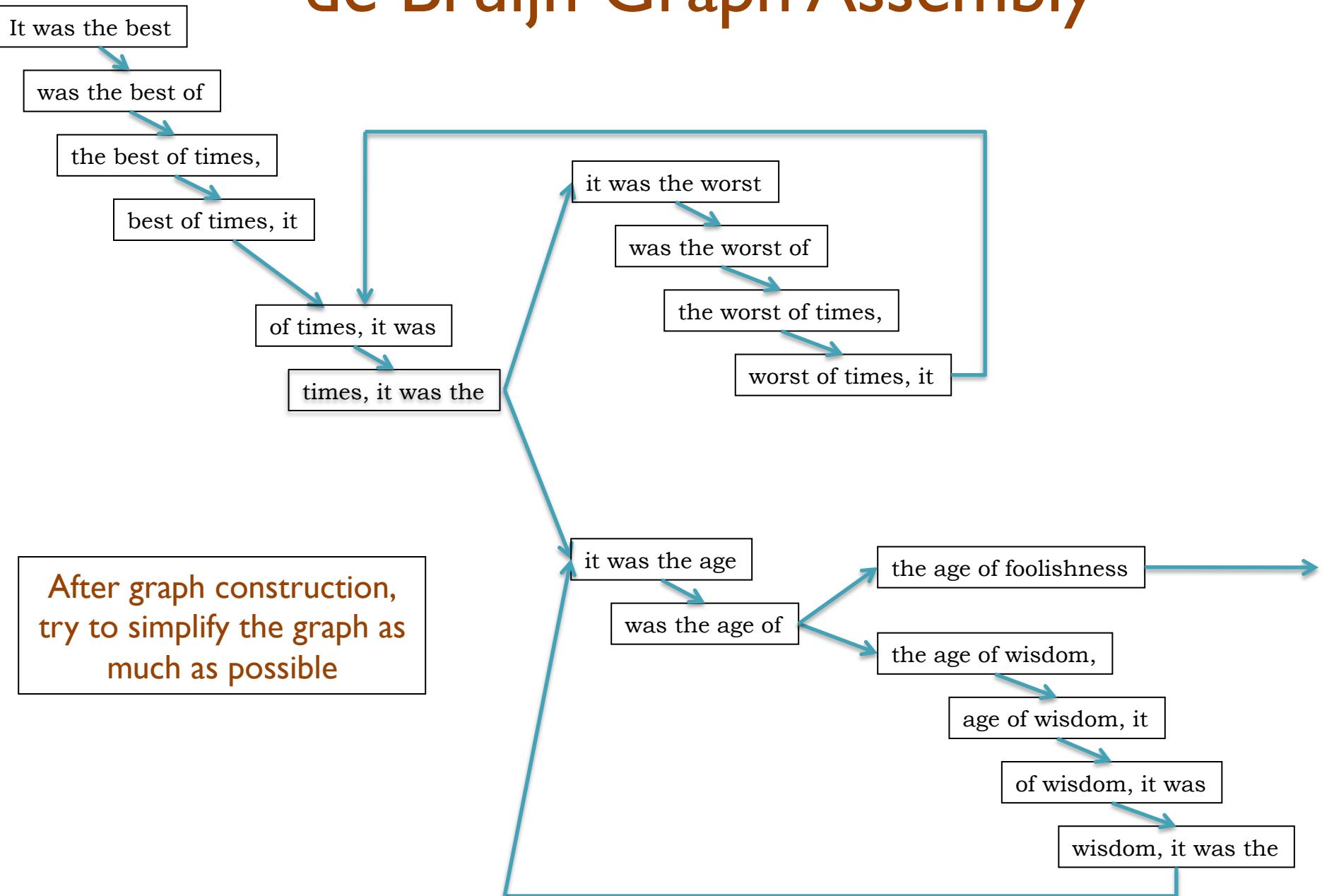
- Locally constructed graph reveals the global sequence structure
 - Overlaps between sequences implicitly computed

de Bruijn, 1946

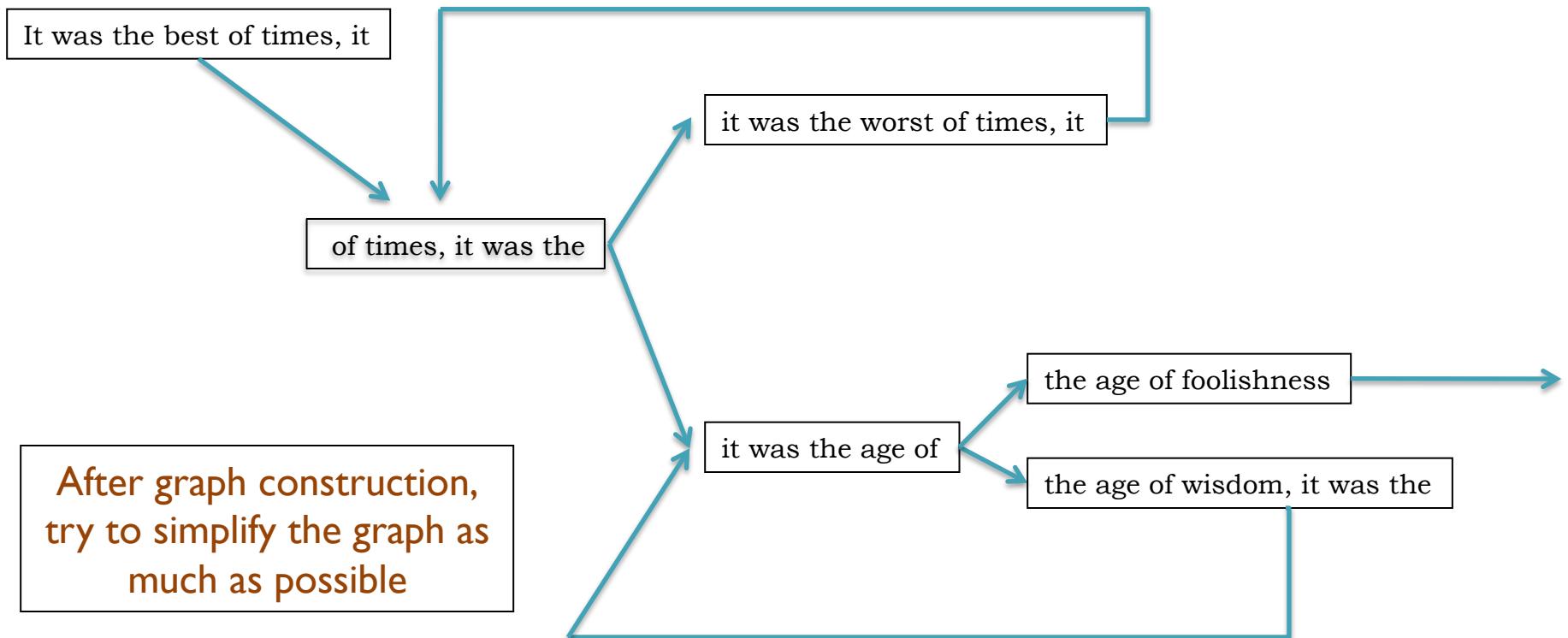
Idury and Waterman, 1995

Pevzner, Tang, Waterman, 2001

de Bruijn Graph Assembly



de Bruijn Graph Assembly



The full tale

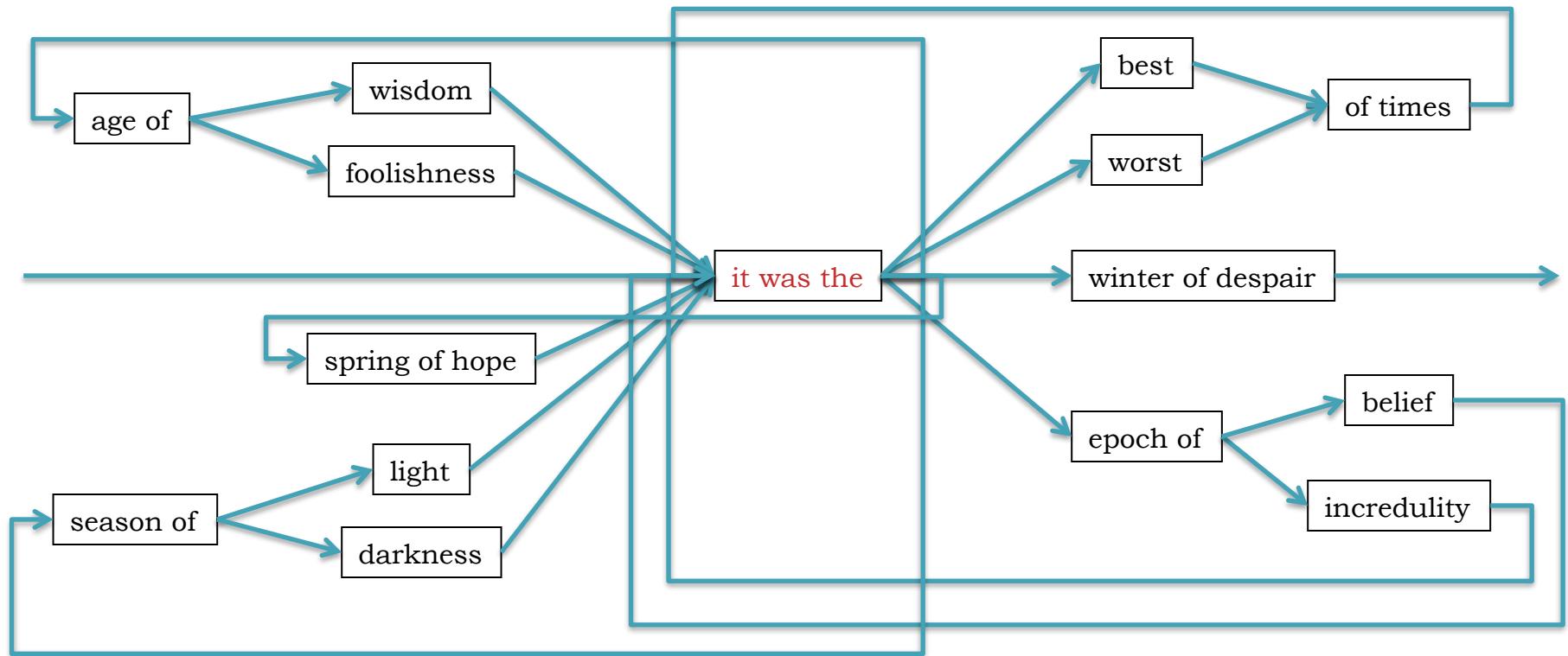
... it was the best of times it was the worst of times ...

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

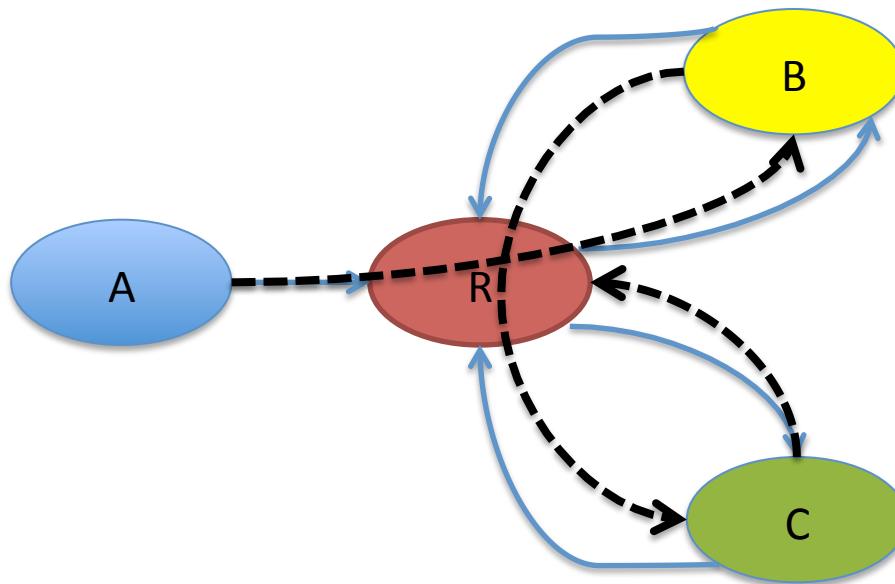
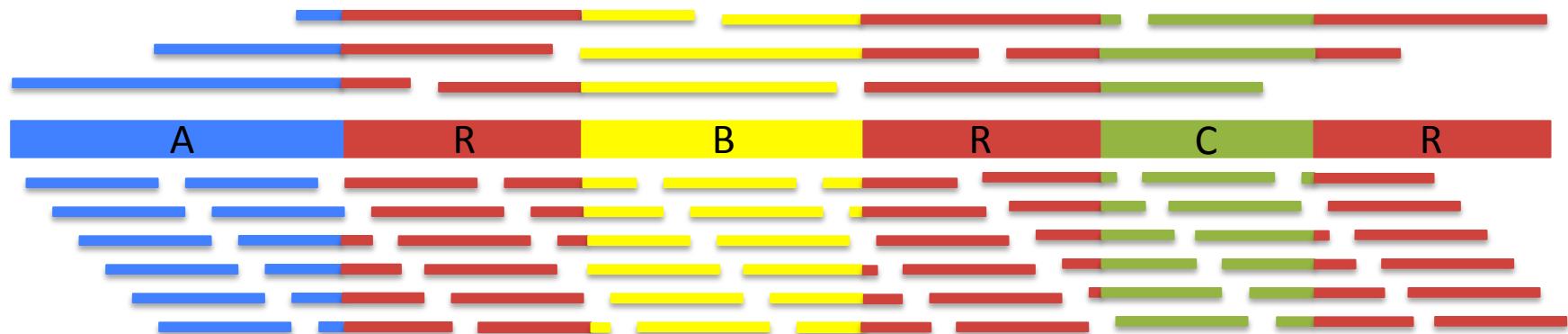
... it was the epoch of belief it was the epoch of incredulity ...

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

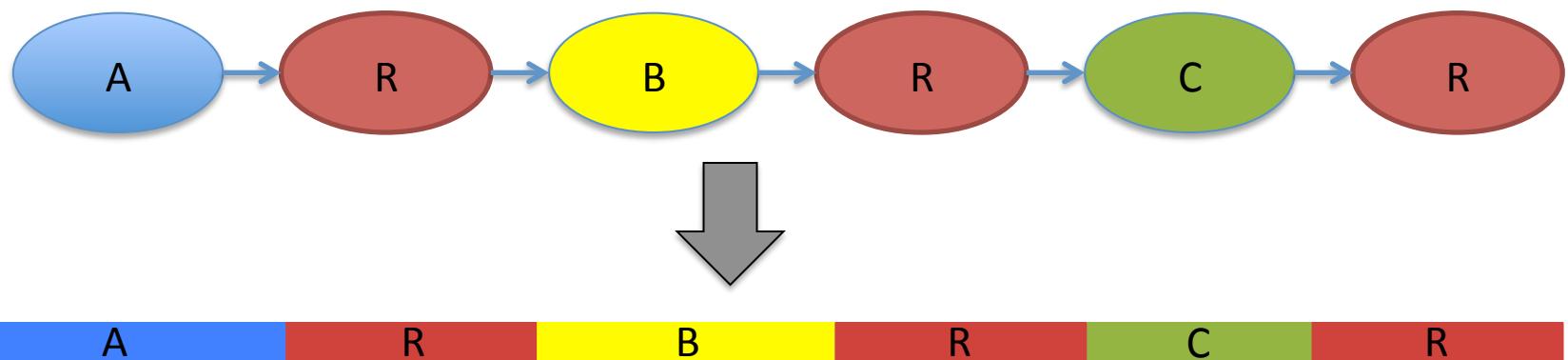
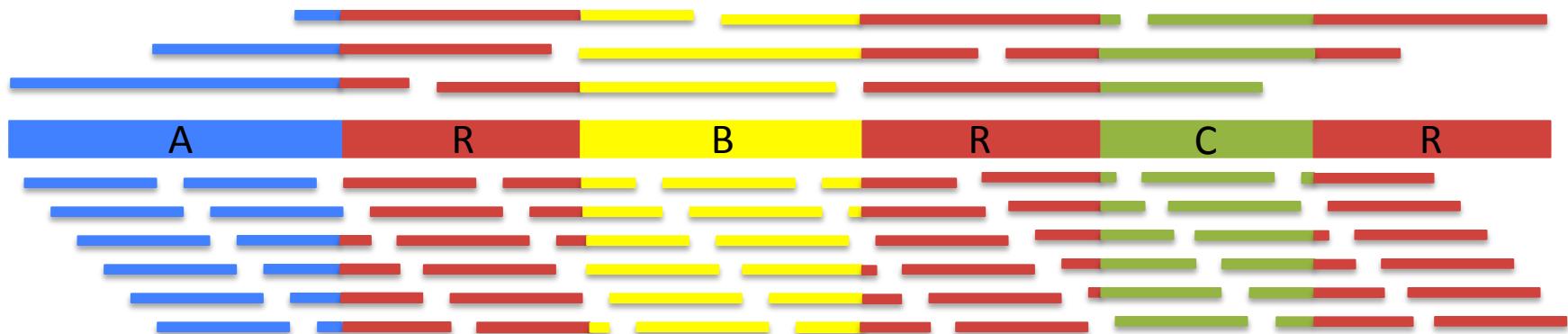
... it was the spring of hope it was the winter of despair ...



Assembly Complexity



Assembly Complexity



Milestones in Genome Assembly

Science Vol. 207 February 20, 1980
articles

Nucleotide sequence of bacteriophage Φ X174 DNA

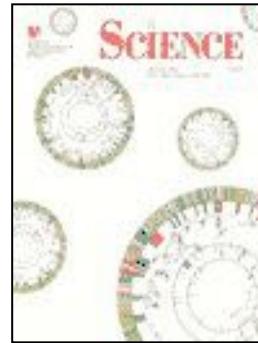
E.Sanger, G.M.Air, R.G.Bonell, N.E.Browne, A.R.Coulson, J.C.Fiddes,
C.A.Hinchliffe, B.P., P.M.Sherrard & M.Smith

MBI: Laboratory of Molecular Biology, 305 Brook Cambridge, MA 02139 USA

A Φ X174 sequence is the genome of bacteriophage Φ X174. It is a single molecule of DNA with a molecular weight of about 2.6 x 10⁶. It contains 5375 base pairs. The genome is composed of two segments, one coding for structural proteins and one coding for enzymes. The enzymes include restriction endonucleases, ligases, and polymerases. The genome is organized into three main regions: the structural genes, the enzyme genes, and the regulatory genes.

The genome of bacteriophage Φ X174 is a single molecule, containing 5375 base pairs. The code of these genes is determined by genetic recombination. The Φ X174 genome is composed of two segments: one coding for structural proteins and one coding for enzymes. The enzymes include restriction endonucleases, ligases, and polymerases. The genome is organized into three main regions: the structural genes, the enzyme genes, and the regulatory genes.

1977. Sanger et al.
1st Complete Organism
5375 bp



2000. Myers et al.
1st Large WGS Assembly.
Celera Assembler. 116 Mbp

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

1998. C.elegans SC
1st Multicellular Organism
BAC-by-BAC Phrap. 97Mbp

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

Assembly Applications

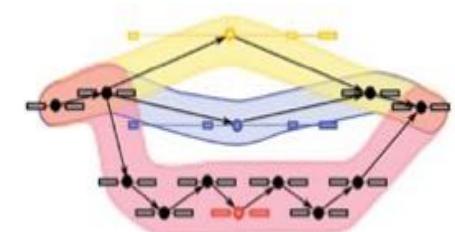
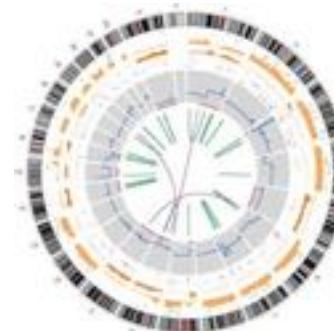
- Novel genomes



- Metagenomes

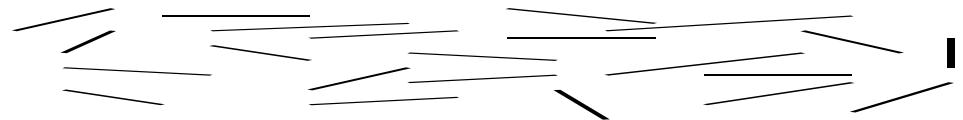


- Sequencing assays
 - Structural variations
 - Transcript assembly
 - ...



Assembling a Genome

1. Shear & Sequence DNA



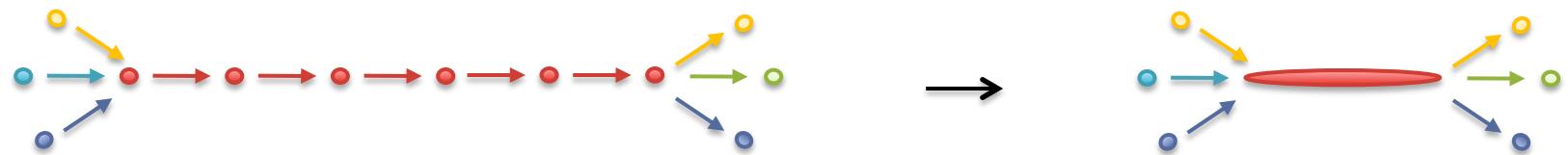
2. Construct assembly graph from overlapping reads

...AGCCTAGGGATGCGCGACACGT

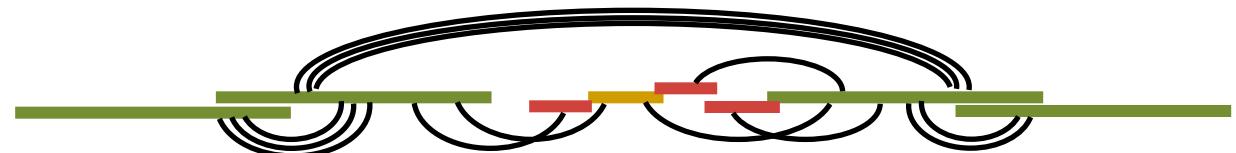
GGATGCGCGACACGT CGCATATCCGGTTGGT CAACCTCGGACGGAC

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:**

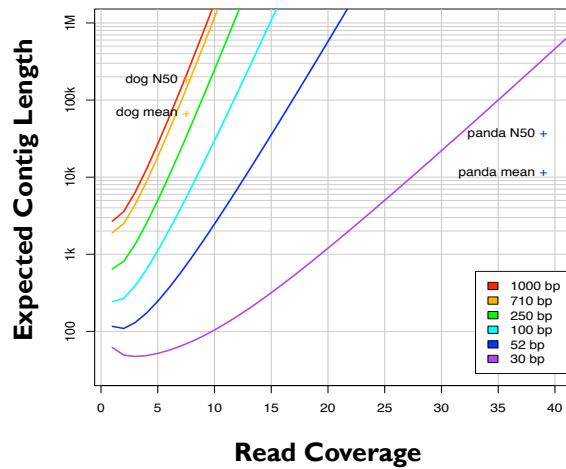
- (Very) Large genomes, complex structure

4. **Accuracy:**

- (Very) Hard to assess correctness

Ingredients for a good assembly

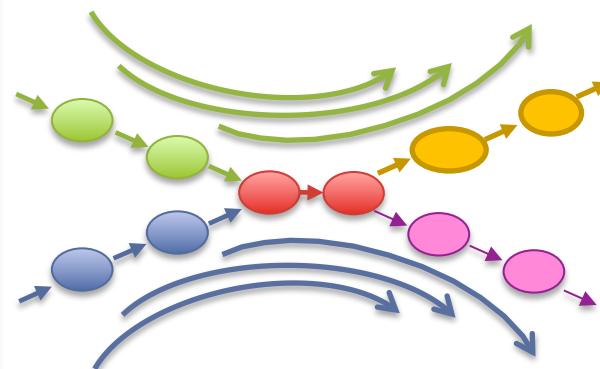
Coverage



High coverage is required

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

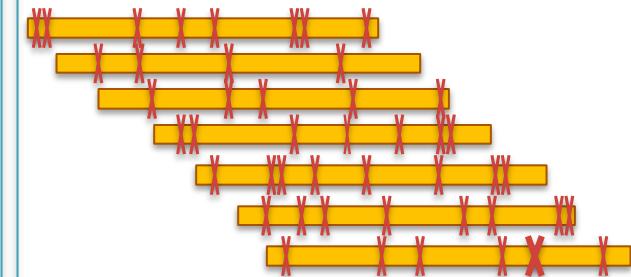
Read Length



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

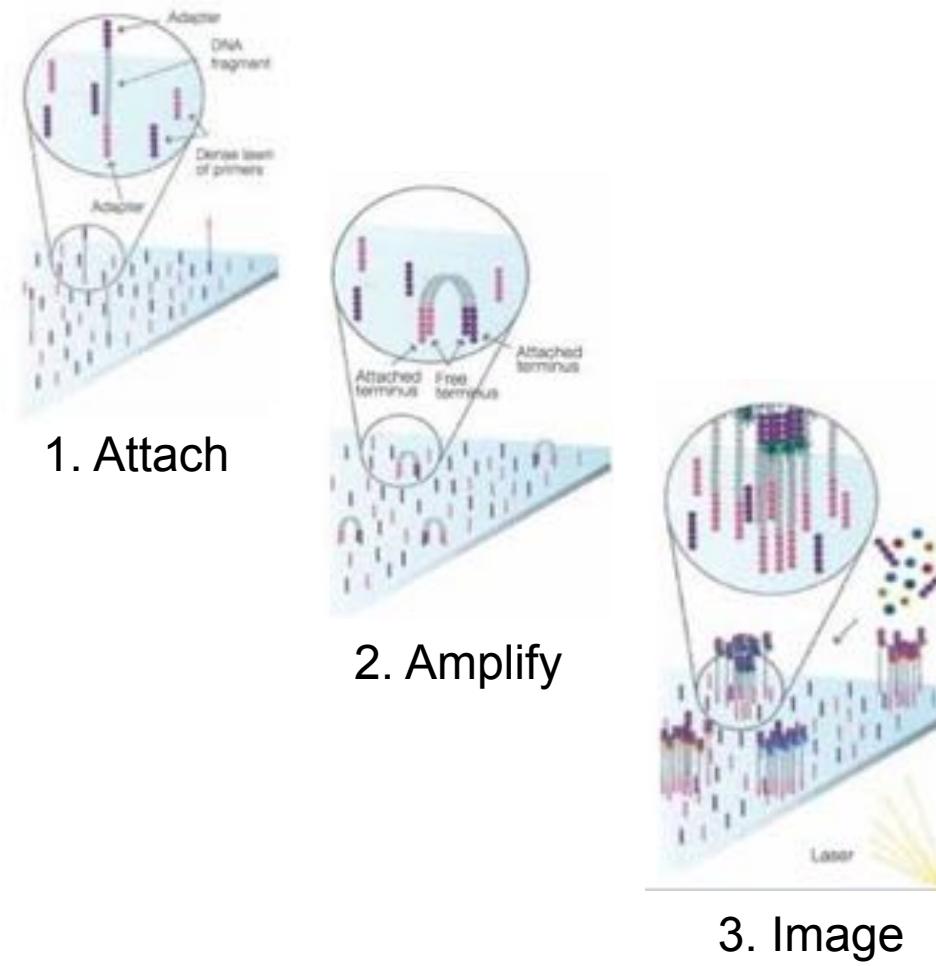
Current challenges in *de novo* plant genome sequencing and assembly
Schatz MC, Witkowski, McCombie, WR (2012) *Genome Biology*. 12:243

Massively Parallel Sequencing



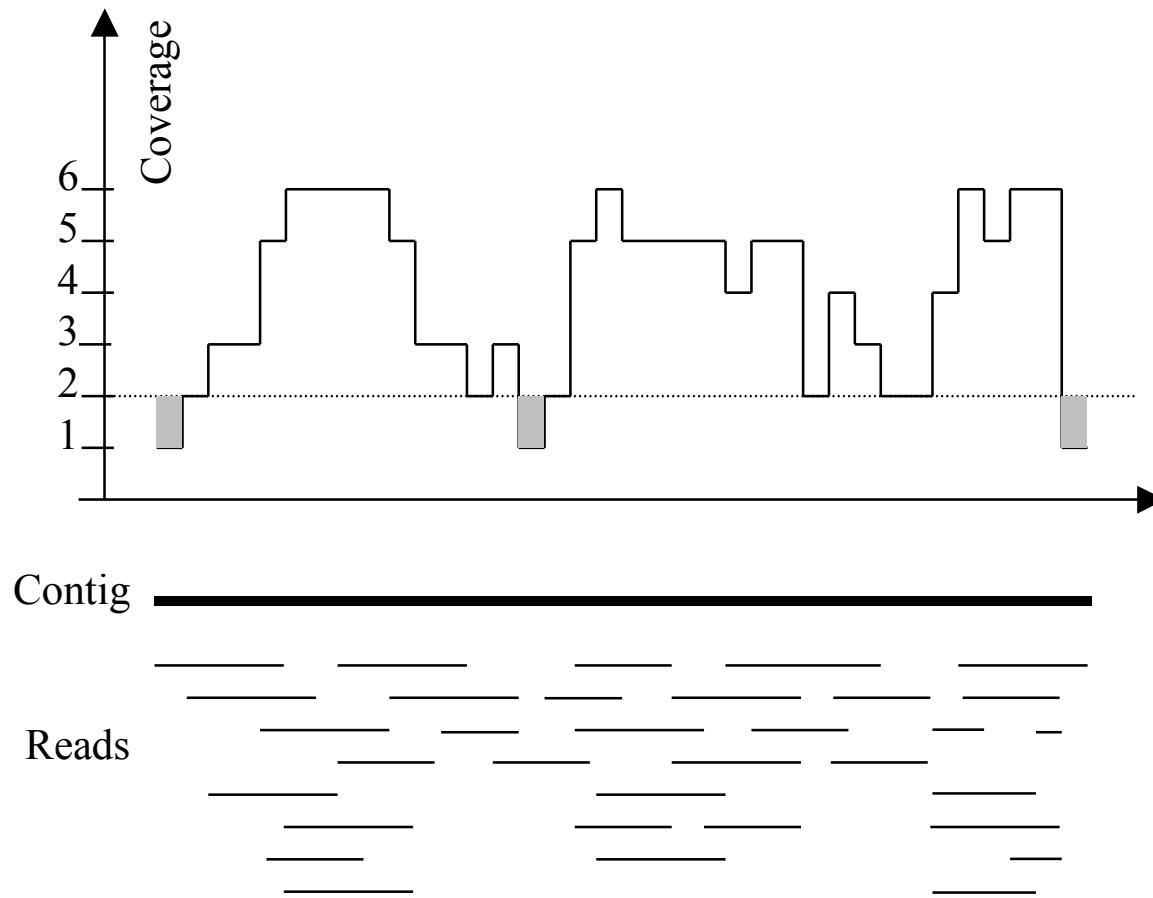
Illumina HiSeq 2000
Sequencing by Synthesis

>60Gbp / day



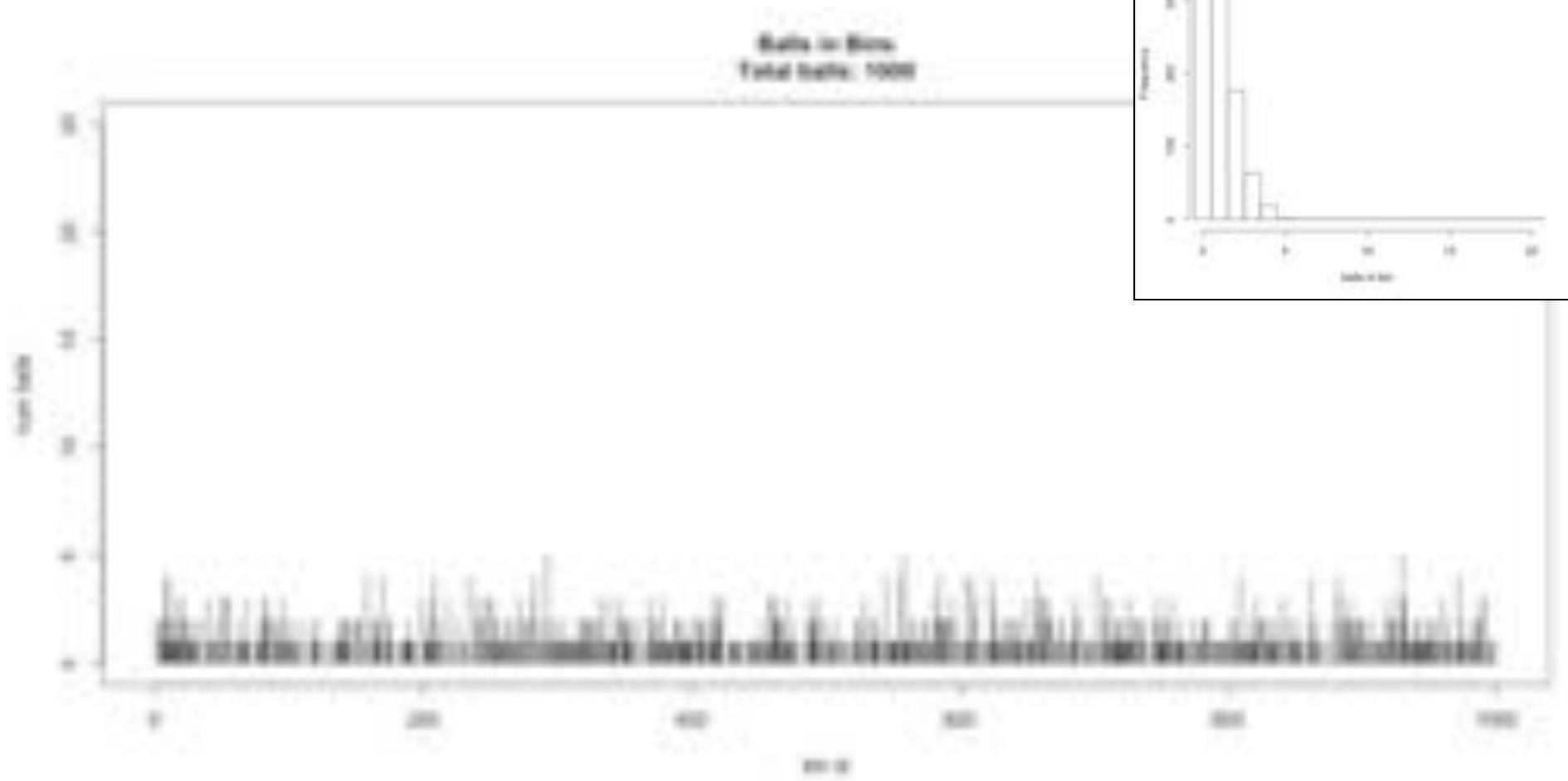
Metzker (2010) Nature Reviews Genetics 11:31-46
<http://www.youtube.com/watch?v=l99aKKHcxC4>

Typical contig coverage

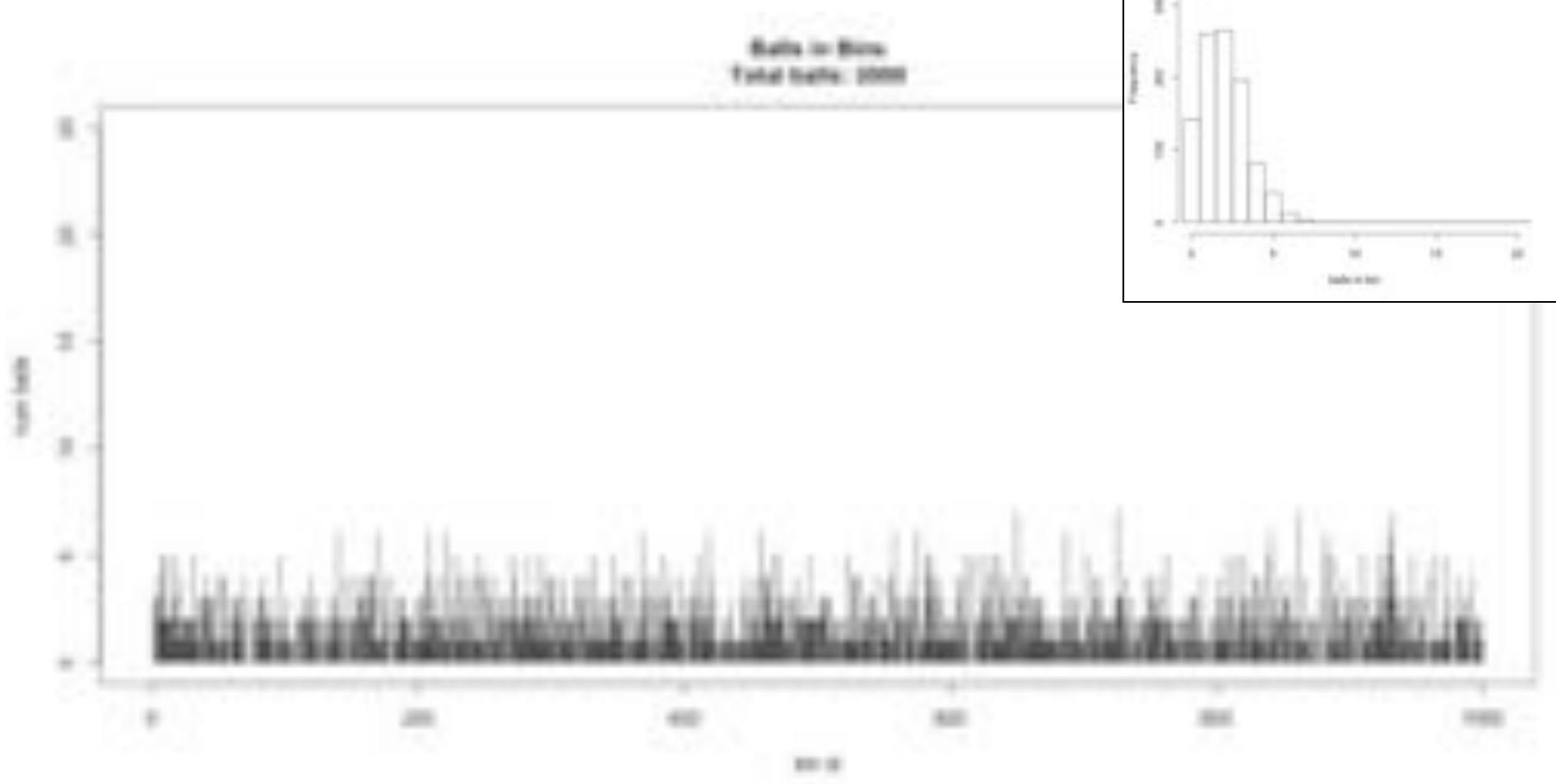


Imagine raindrops on a sidewalk
How many rain drops should we collect?

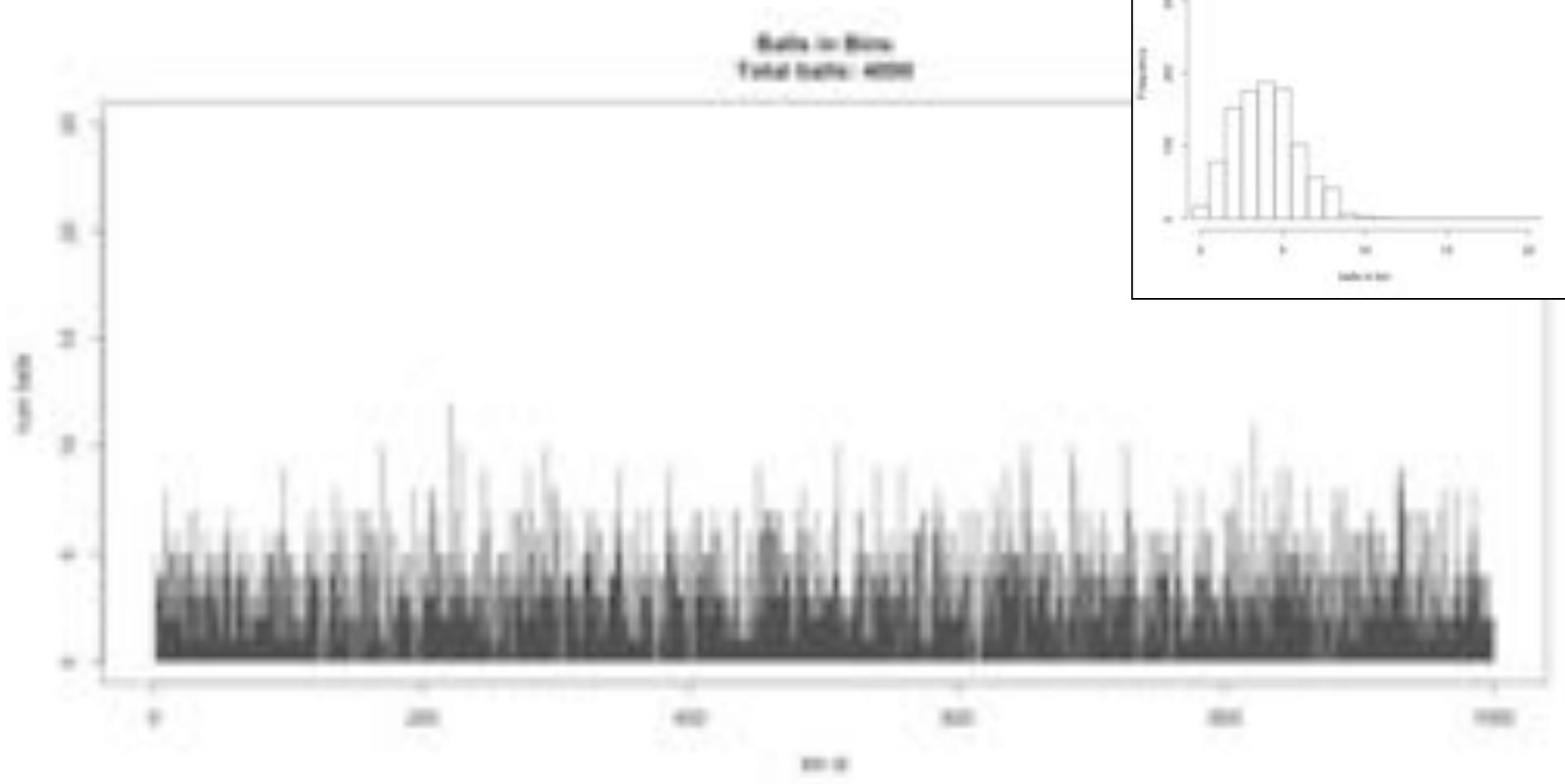
Ix sequencing



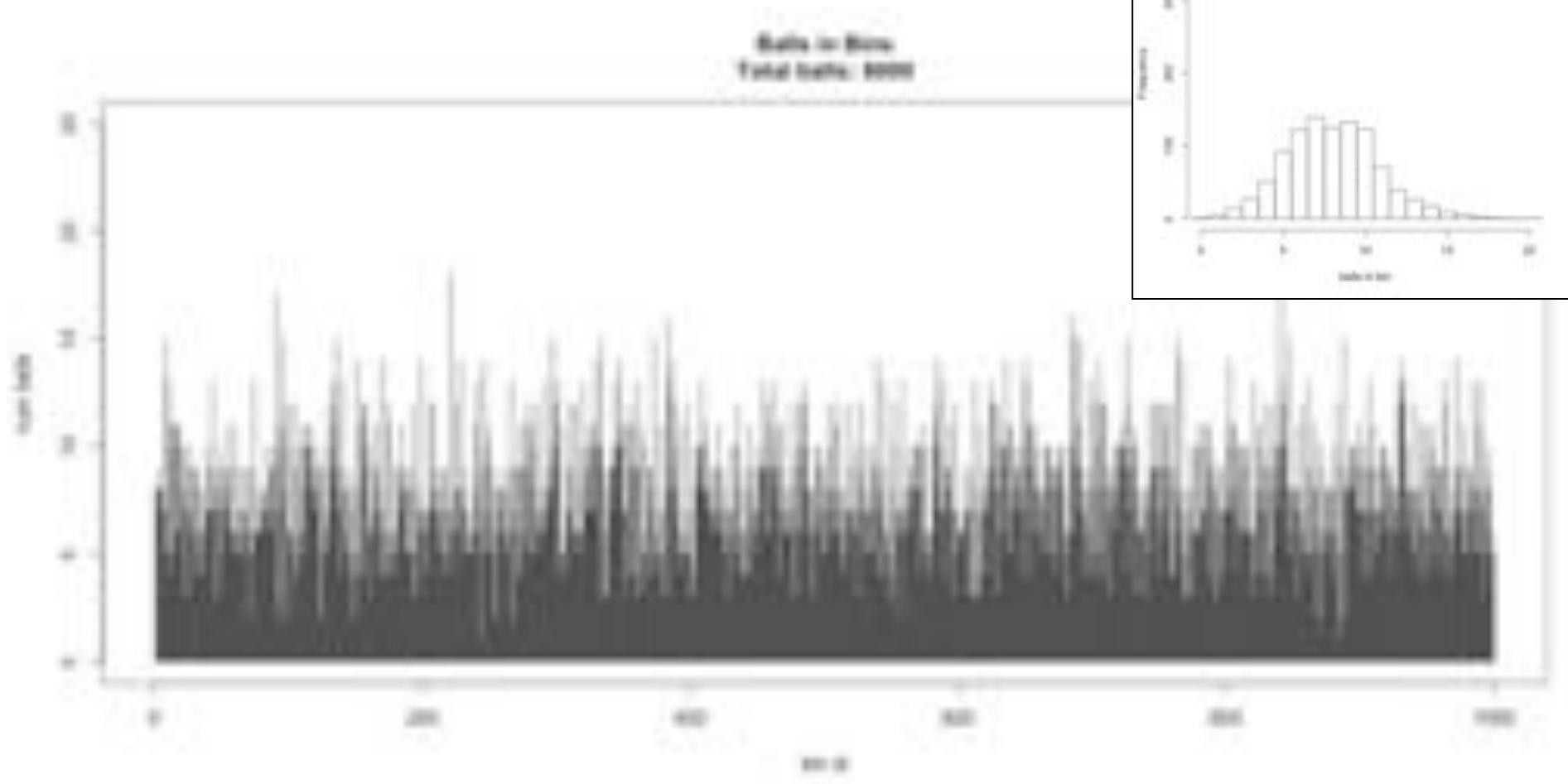
2x sequencing



4x sequencing



8x sequencing



Poisson Distribution

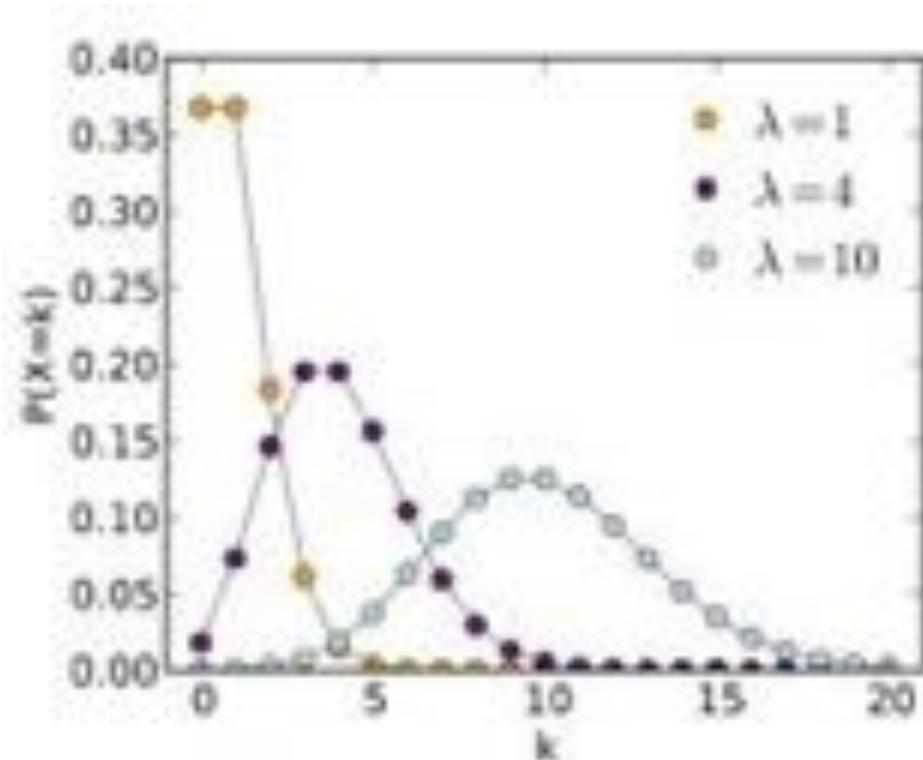
The probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and independently of the time since the last event.

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 property: The standard deviation is the square root of the mean.

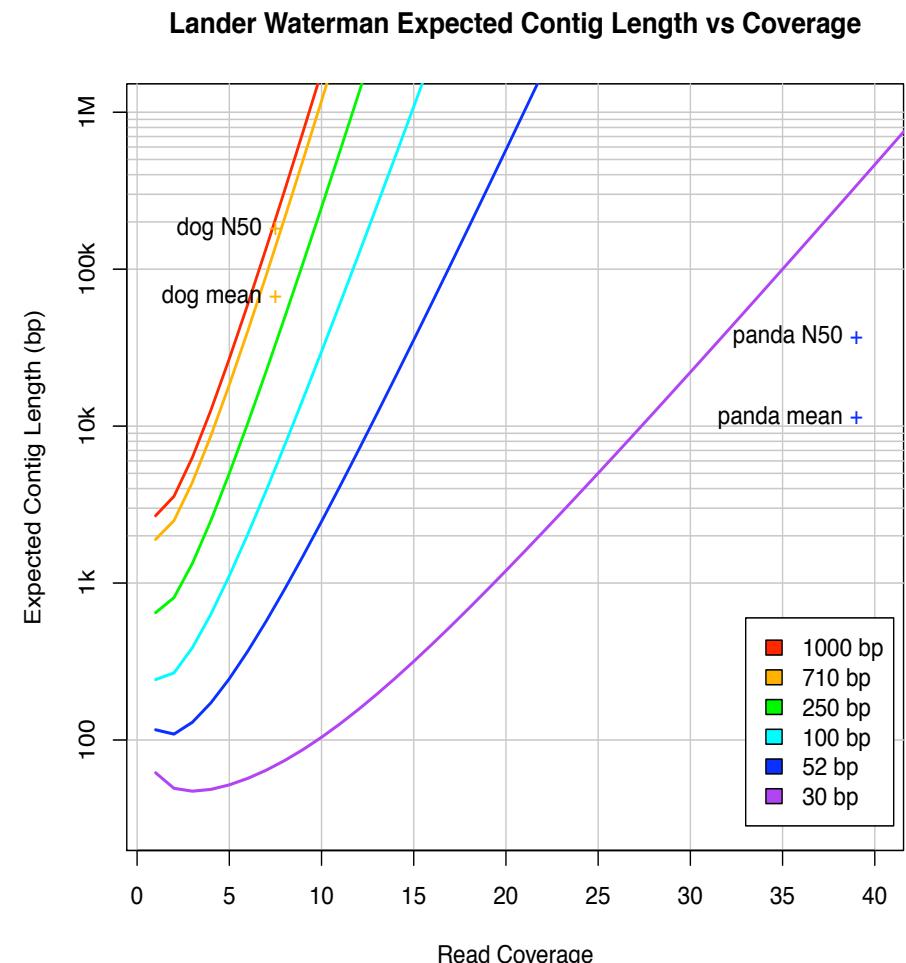
$$P(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$



Coverage and Read Length

Idealized Lander-Waterman model

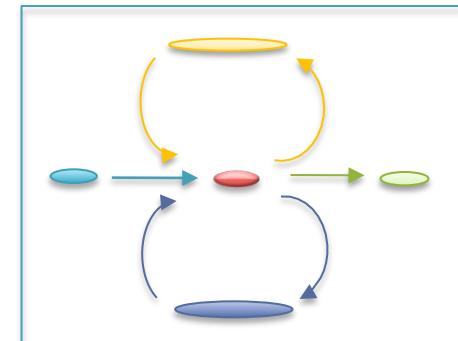
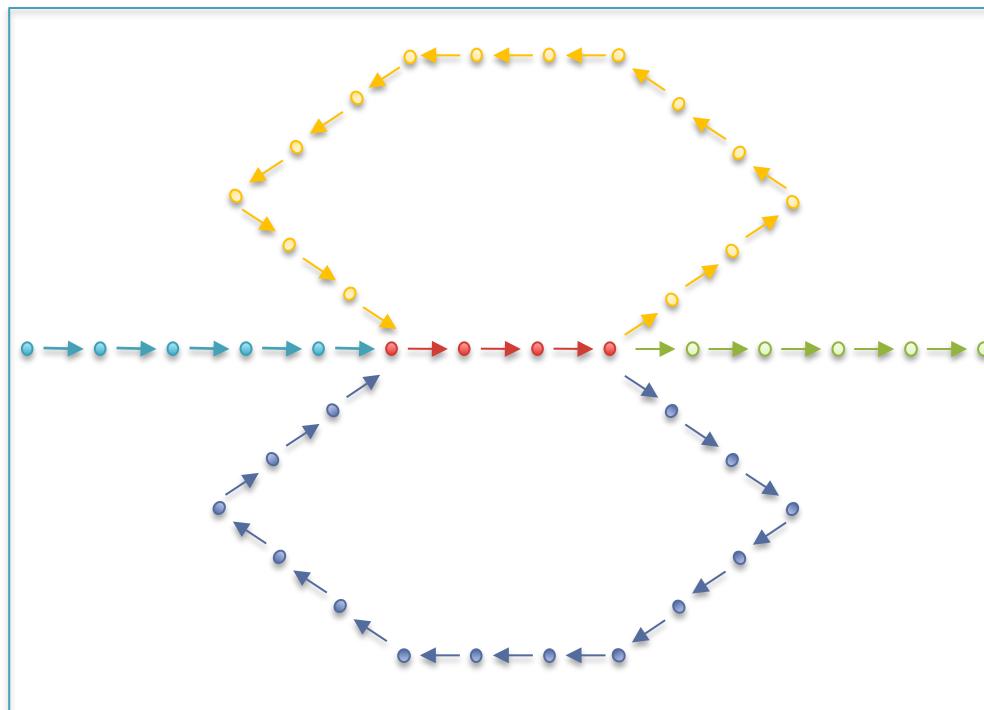
- Reads start at perfectly random positions
- Contig length is a function of coverage and read length
 - Short reads require much higher coverage to reach same expected contig length
- Need even high coverage for higher ploidy, sequencing errors, sequencing biases
 - Recommend 100x coverage



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, (3) heterozygosity, and (3) repeats



Errors in the graph



(Chaisson, 2009)

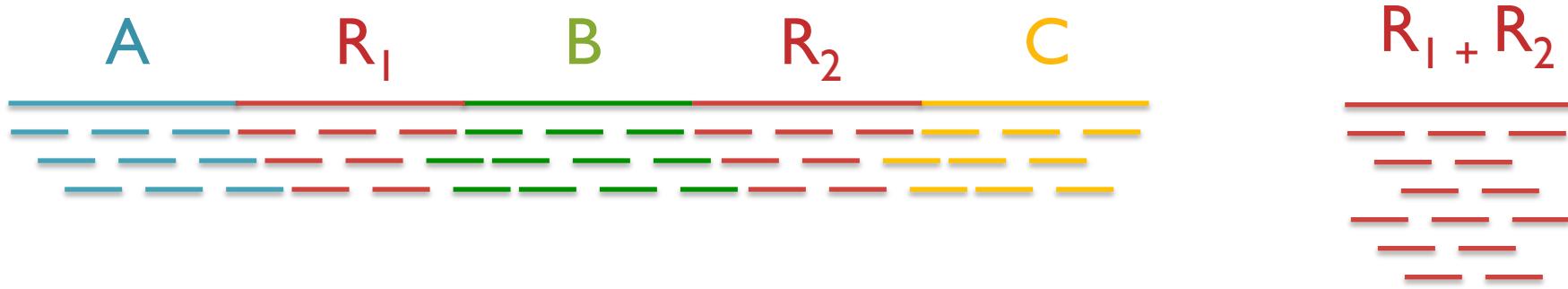
Clip Tips	Pop Bubbles
<p>was the worst of times,</p> <p>was the worst of tymes,</p> <p>the worst of times, it</p>	<p>was the worst of times,</p> <p>was the worst of tymes,</p> <p>times, it was the age</p> <p>tymes, it was the age</p>
<p>the worst of tymes,</p> <p>was the worst of</p> <p>the worst of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

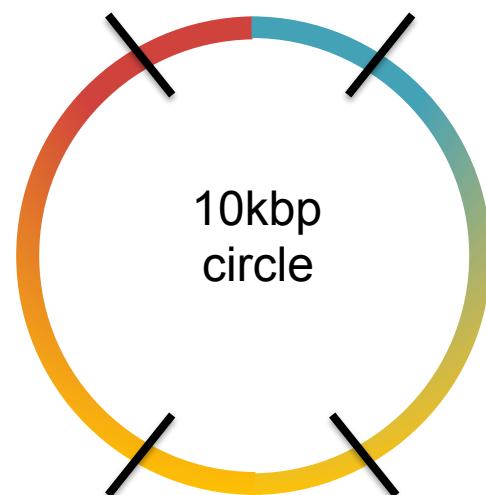
- Read one end of the molecule, flip, and read the other end
- Generate pair of reads separated by up to 500bp with inward orientation



Mate-pair sequencing

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

10kbp



2x100 @ ~10kbp (outies)

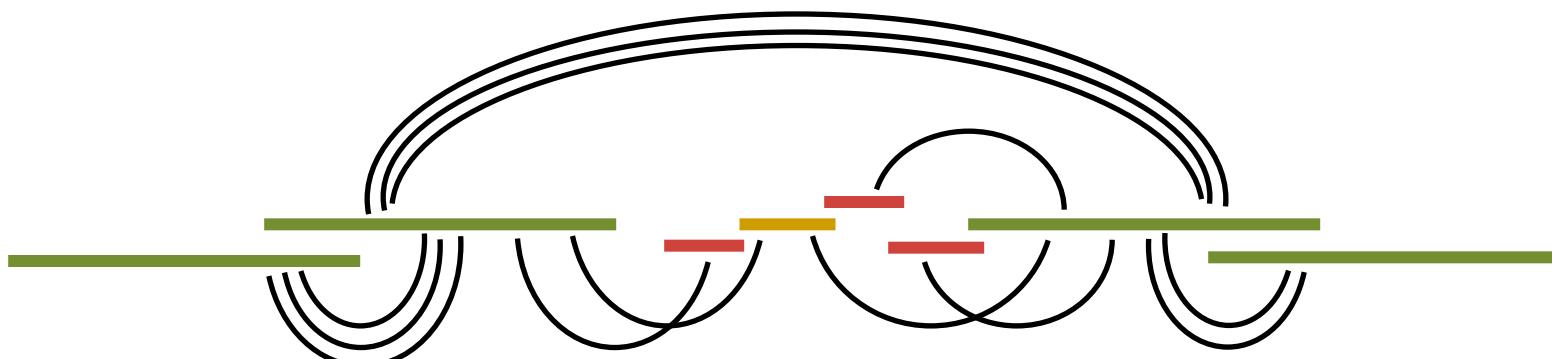


2x100 @ 300bp (innies)



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



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
 - I. Aligning & visualizing with MUMmer
3. Genome assemblers
 1. ALLPATHS-LG: recommended for Illumina-only projects
 2. Celera Assembler: recommended for PacBio projects
4. Summary and Recommendations



Whole Genome Alignment with MUMmer

Slides Courtesy of Adam M. Phillippy
University of Maryland

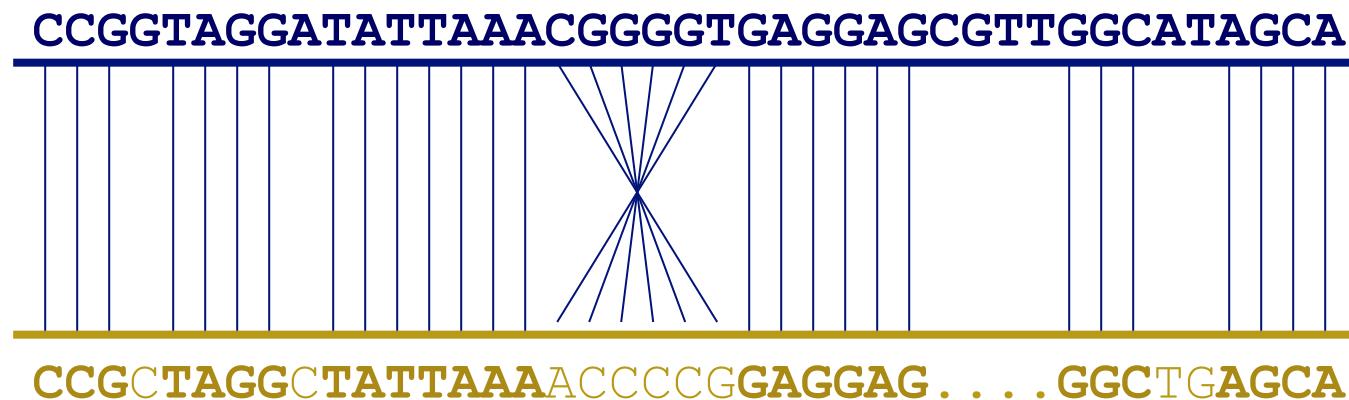
Goal of WGA

- 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)



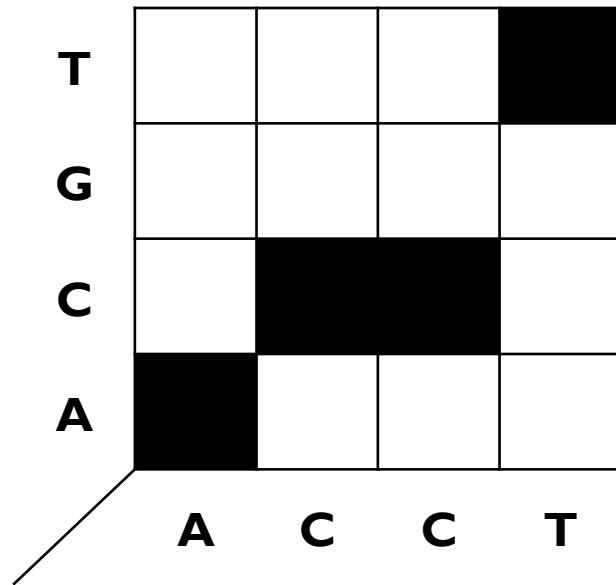
WGA visualization

- How can we visualize *whole genome* alignments?

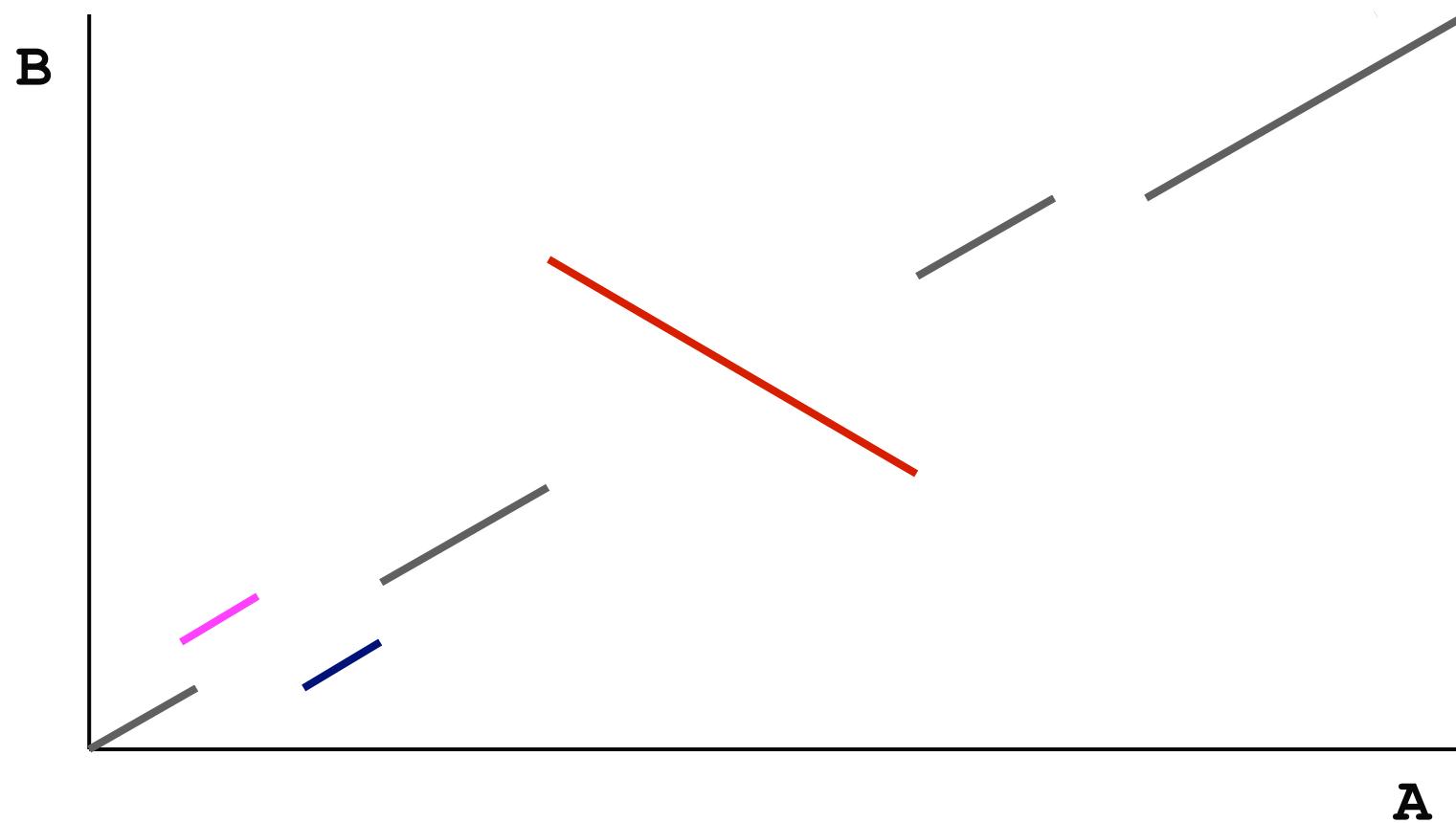
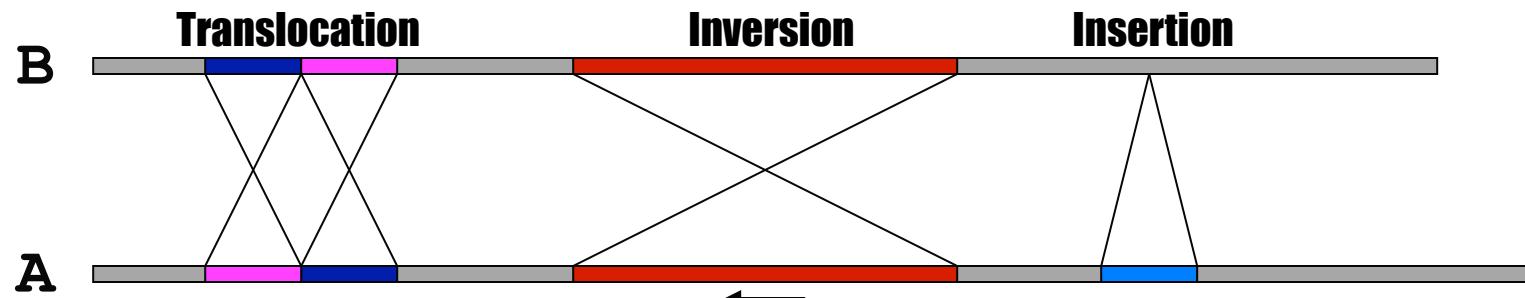
- With an alignment dot plot

- $N \times M$ matrix

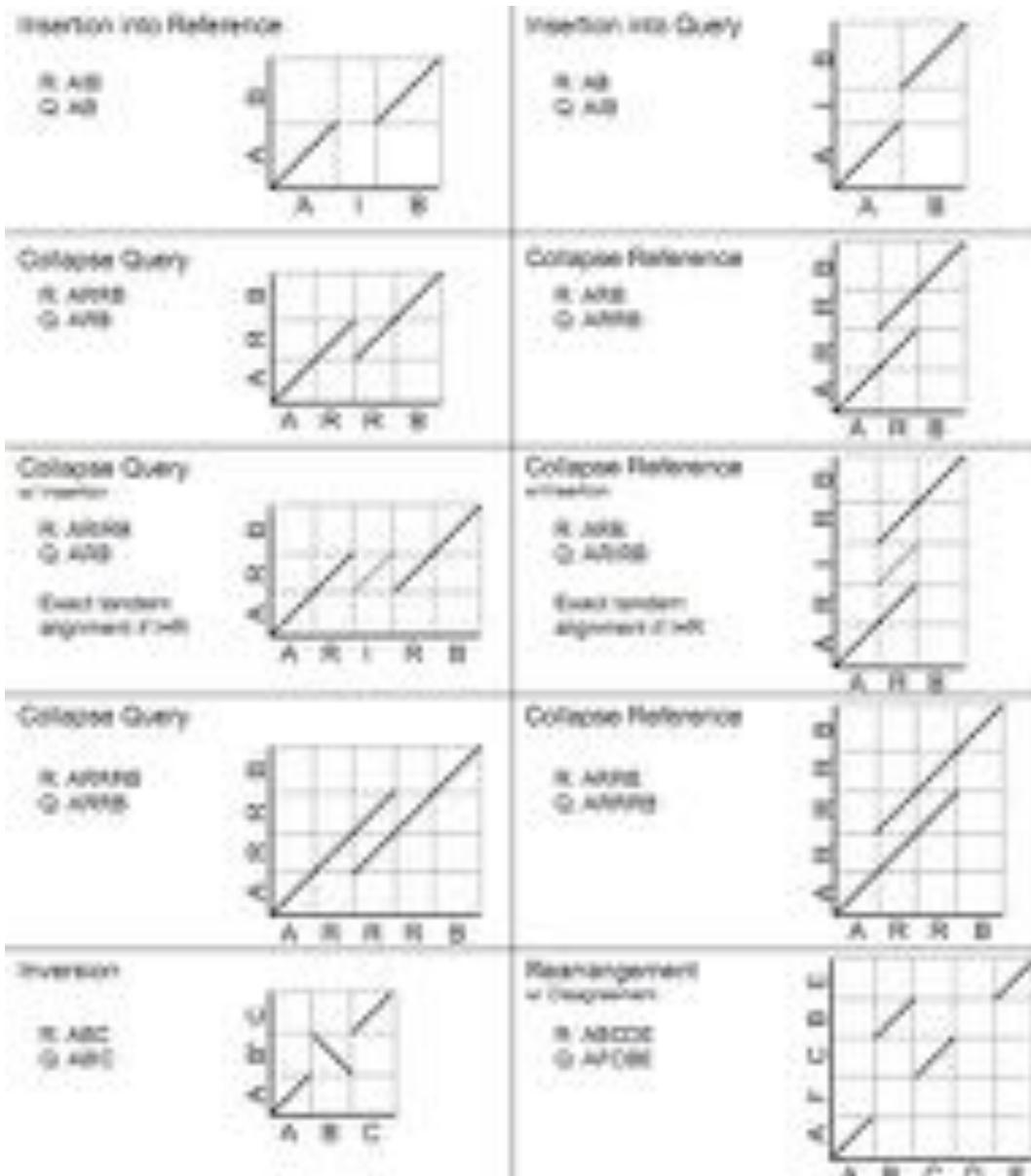
- Let i = position in genome A
 - Let j = position in genome B
 - Fill cell (i,j) if A_i shows similarity to B_j



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

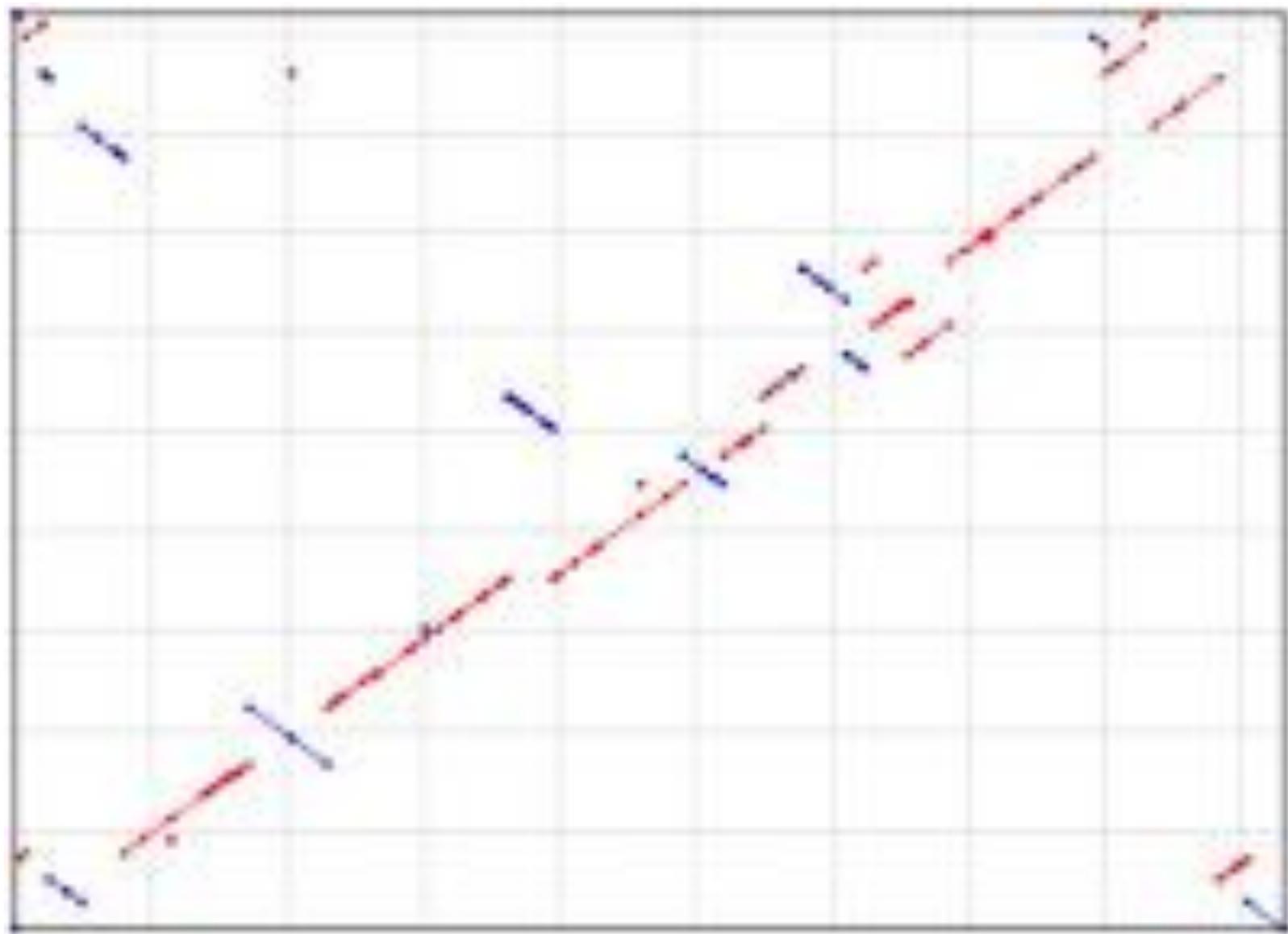


SV Types



- 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)



Alignment of 2 strains of *Y. pestis*
<http://mummer.sourceforge.net/manual/>



Outline

I. Assembly theory

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

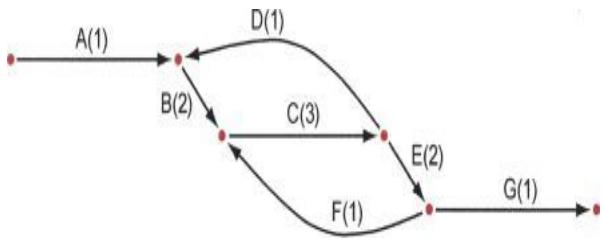
2. Whole Genome Alignment

- 1. Aligning & visualizing with MUMmer

3. Genome assemblers

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

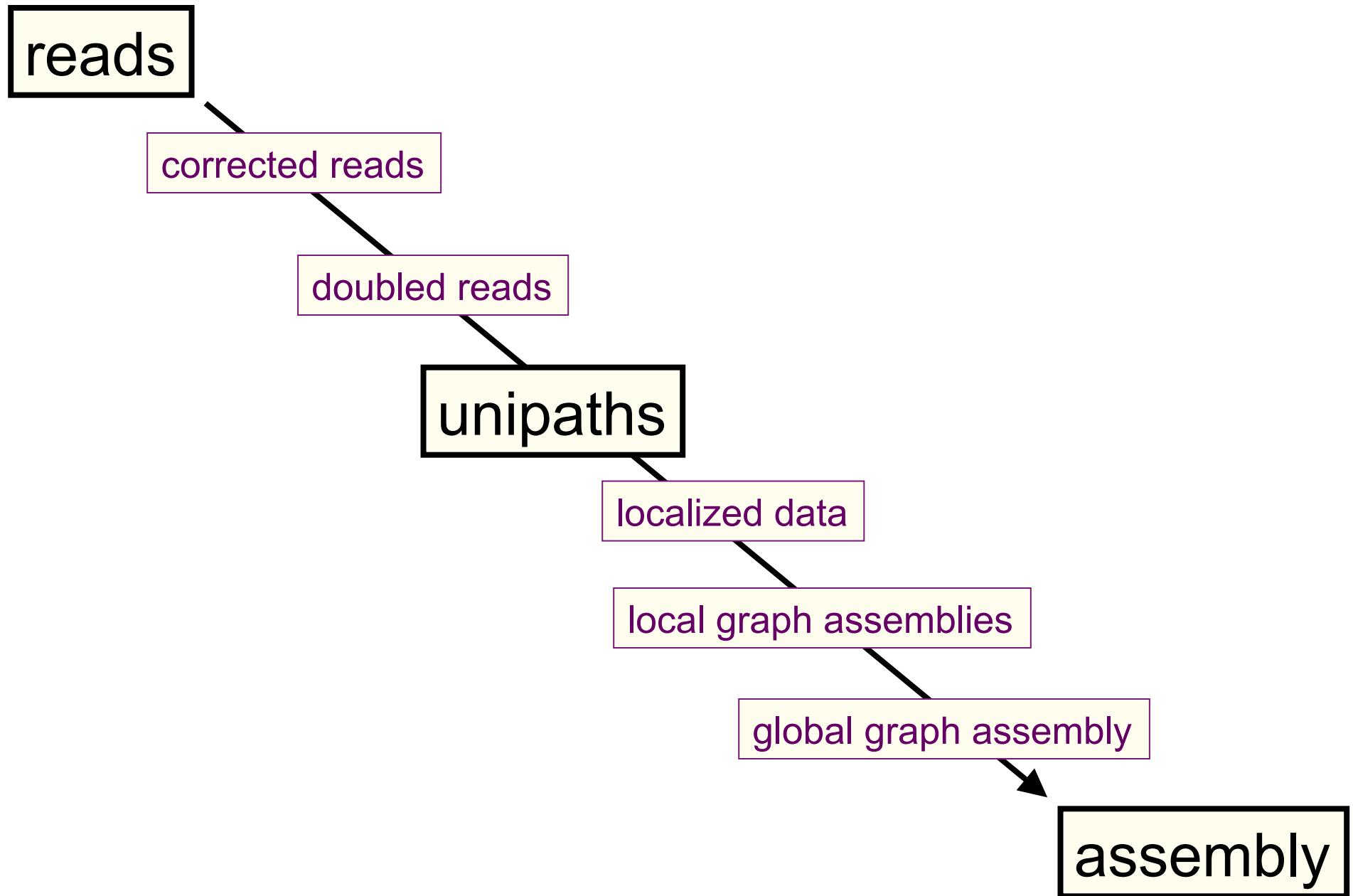
4. Summary and Recommendations



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.

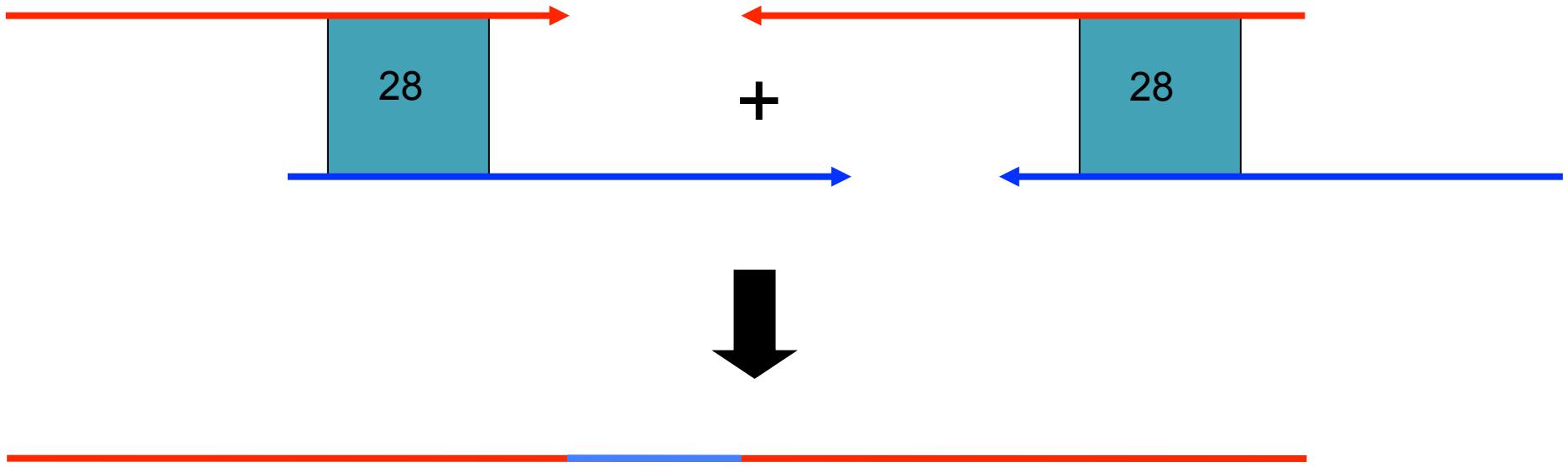
**For best results. Normally not used for small genomes.
However essential to assemble long repeats or duplications.

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

All: protocols are either available, or in progress.

Read doubling

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



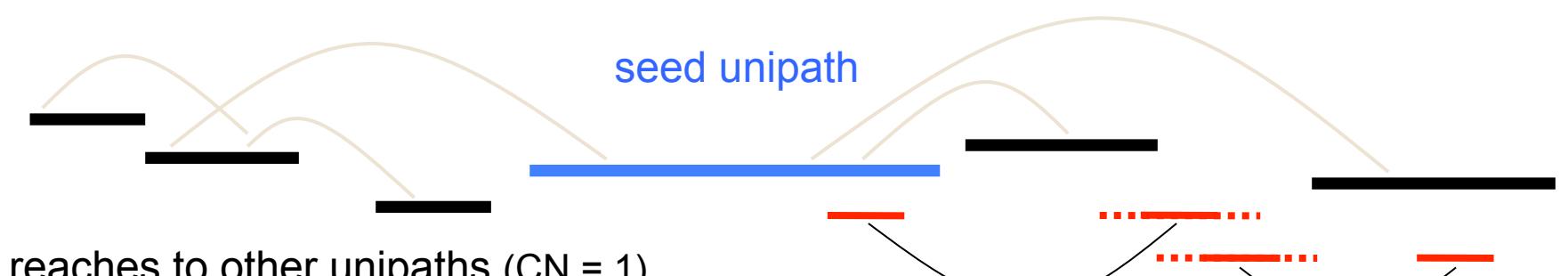
More than one closure allowed (but rare).

Localization

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



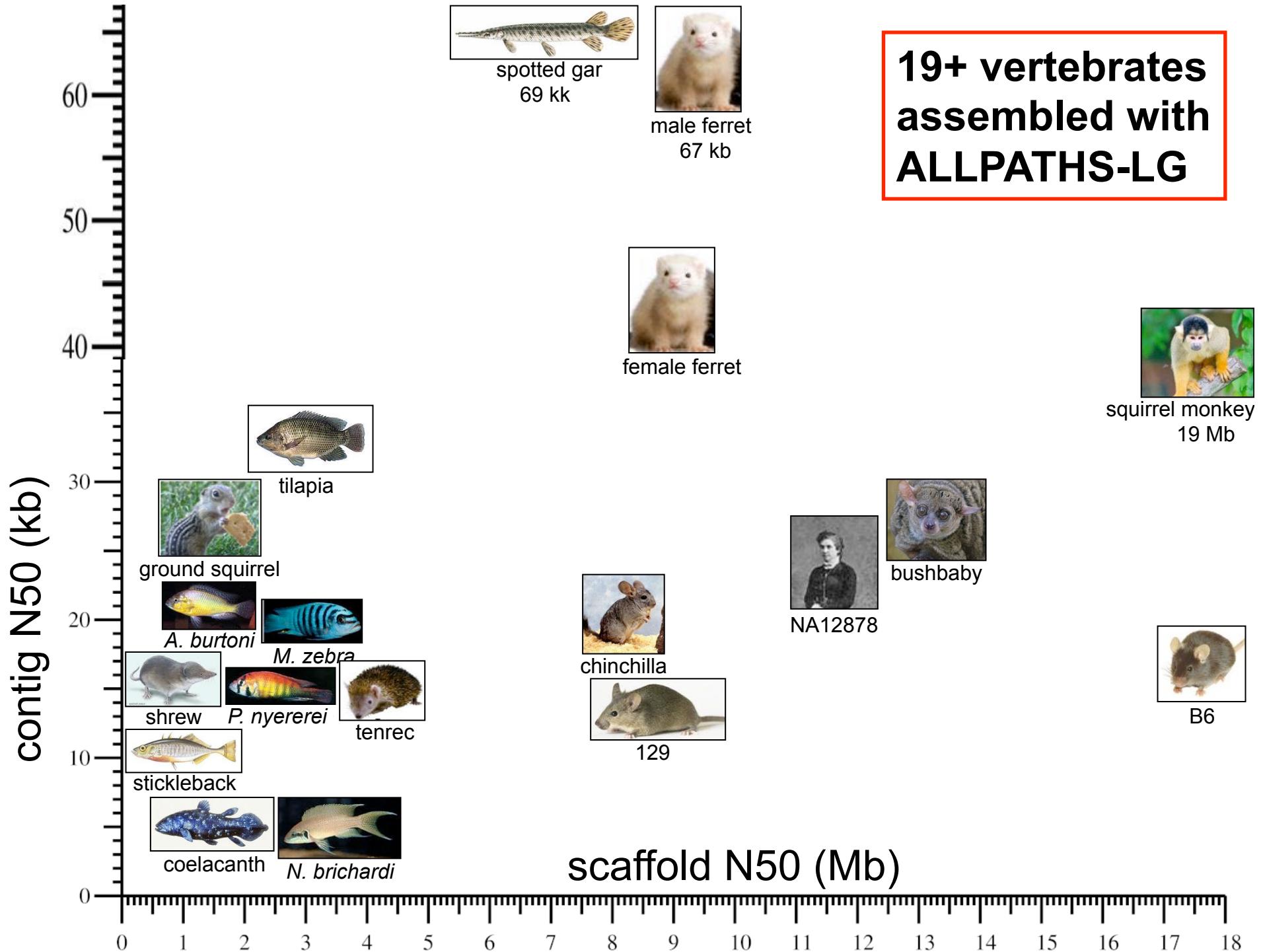
II. Form neighborhood around each seed



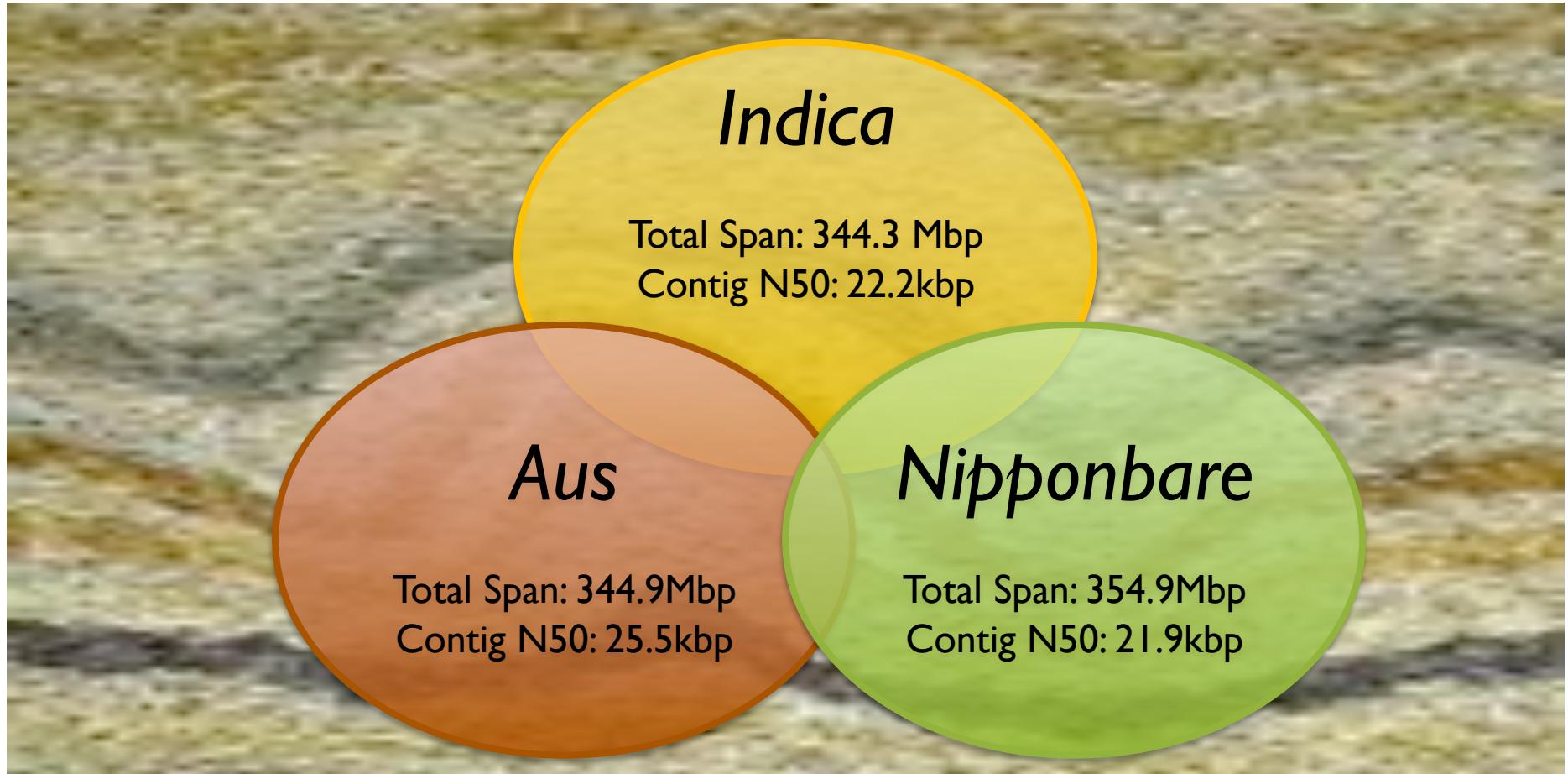
reaches to other unipaths ($CN = 1$)
directly and indirectly

read pairs reach into repeats

and are extended by other
unipaths



Population structure of *Oryza sativa*



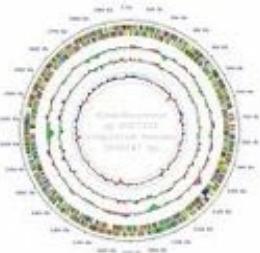
New whole genome de novo assemblies of three divergent strains of rice (*O. sativa*) documents novel gene space of aus and indica
Schatz, MC, Maron, L, Stein, et al (2014) Under Review.

Strain specific regions

(A) Nipponbare

Conclusions

- Very high quality representation of the “gene-space”
 - Overall identity ~99.9%
 - Less than 1% of exonic bases missing
- Genome-specific genes enriched for disease resistance
 - Reflects their geographic and environmental diversity
 - Detailed analysis of agriculturally important loci
- Assemblies fragmented at (high copy) repeats
 - Missing regions have mean k-mer coverage >10,000x
 - Difficult to identify full length gene models and regulatory features

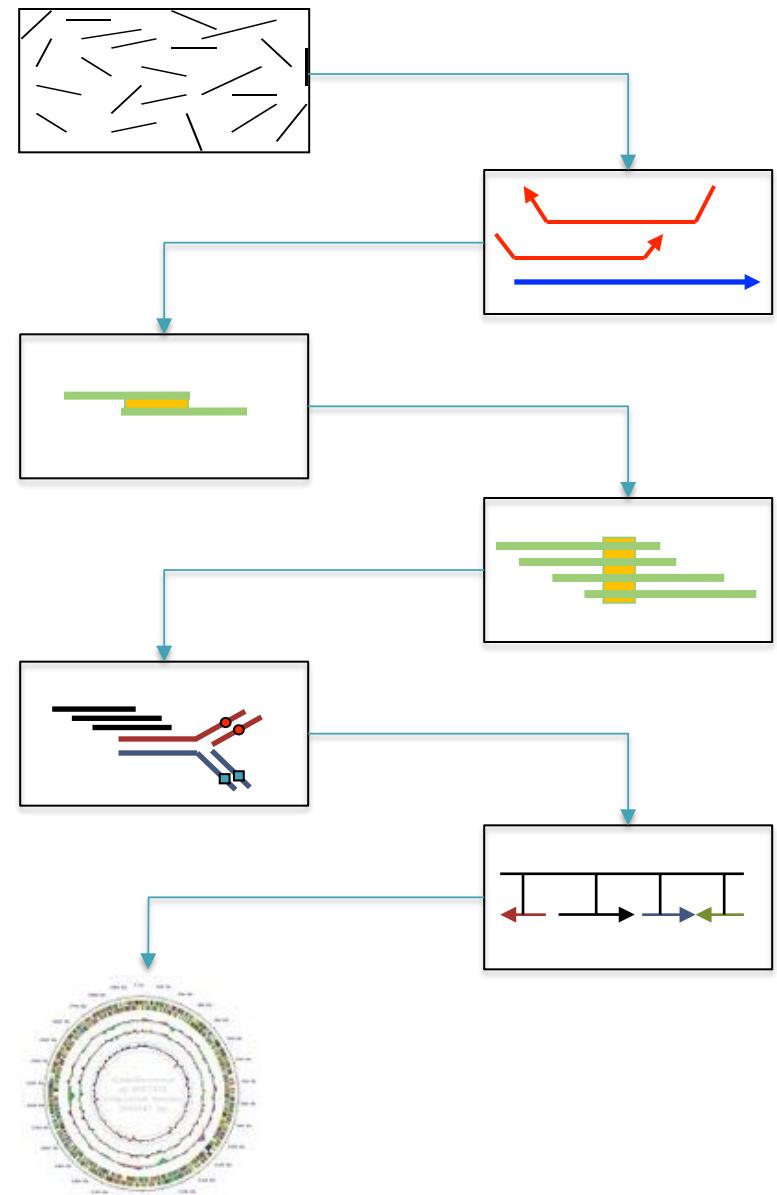


Genome assembly with the Celera Assembler

Celera Assembler

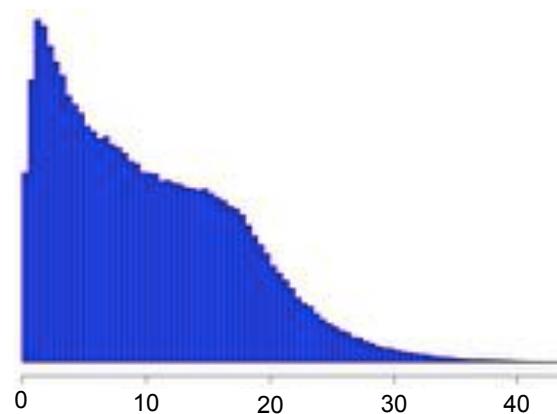
<http://wgs-assembler.sf.net>

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

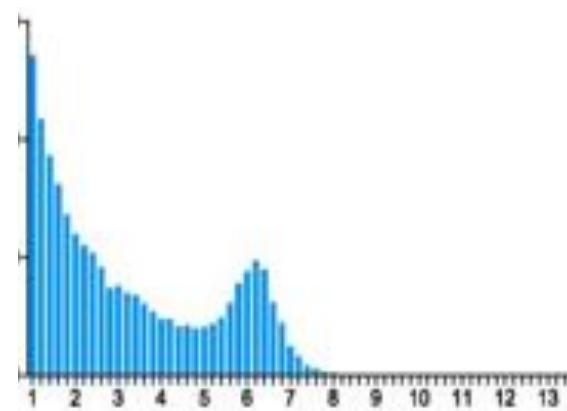


Long Read Sequencing Technology

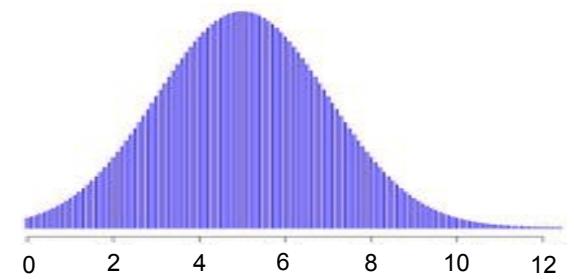
PacBio RS II



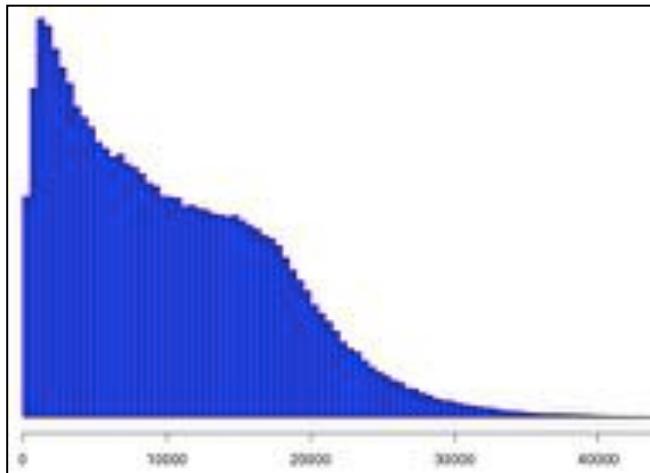
Moleculo



Oxford Nanopore



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

CAACCTTGAAATGTAATCGCACTGAGAACACAAGATTTATTCCGCGCCCG
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
C-ACCTTG-ATGT-AT--CACTGAGAACACAAGATTTATTCCGCGCCCG

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

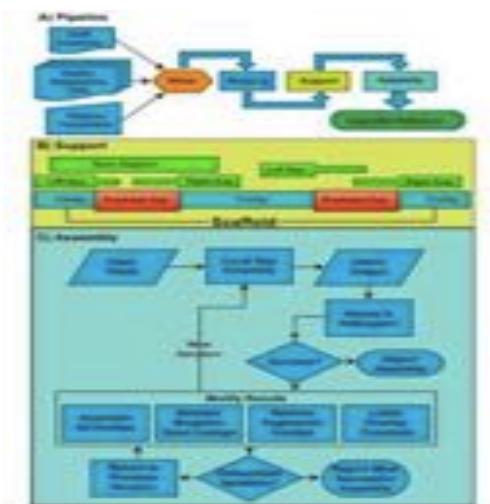
TCGGAGAGATCCAACAAATGGGC-ATCGCCTTGAGTTAC-AATCAA
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
TC-GAGAGATCC-AAACAAAT-GGC GATCG-CTTGACGTTACAATCAA

ATCCAGTGGAAAATATAATTTATGCAATCCAGGAACCTATTACAATTAG
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
ATCCAGT-GAAAATATA- TTATGC-ATCCA-GAACTTATTACAATTAG

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

PacBio Assembly Algorithms

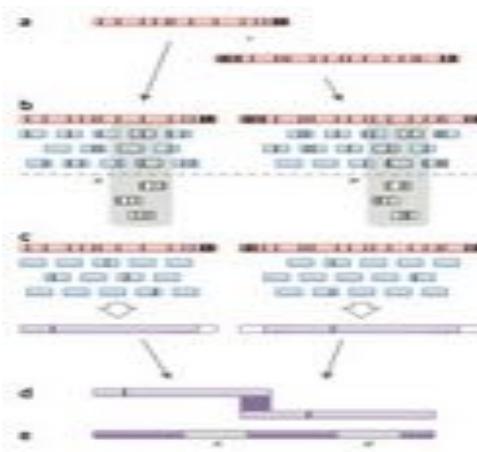
PBJelly



Gap Filling and Assembly Upgrade

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

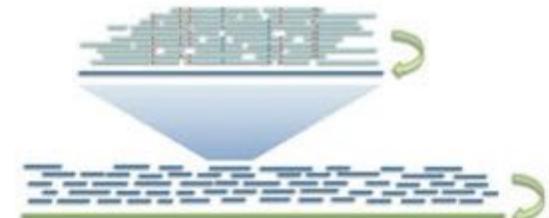
PacBioToCA & ECTools



Hybrid/PB-only Error Correction

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

HGAP & Quiver



$$\Pr(\mathbf{R} \mid T) = \prod_k \Pr(R_k \mid T)$$

Quiver Performance Results Comparison to Reference Genome (<i>M. ruber</i> ; 3.1 MB; SMRT® Cells)		
	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

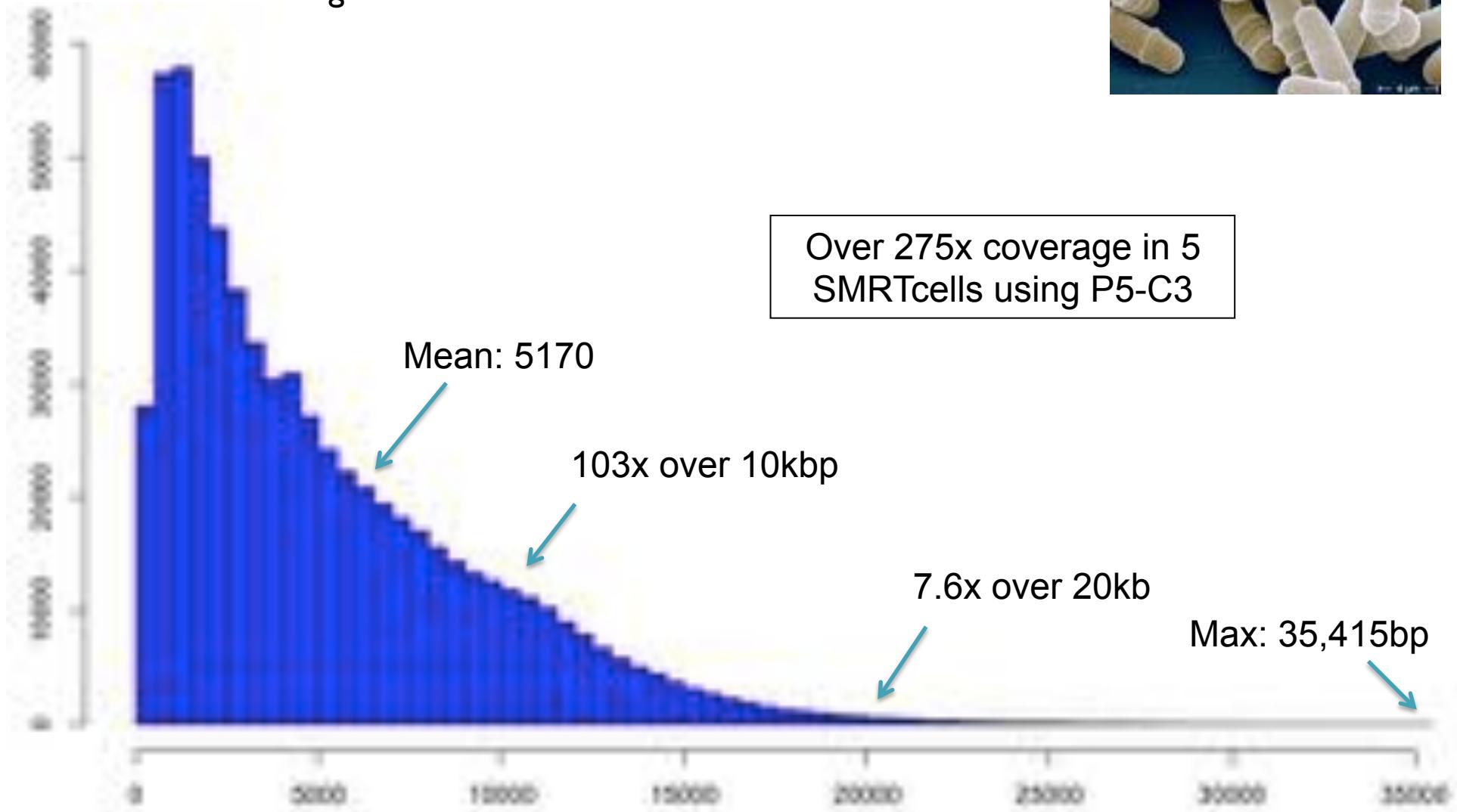
PacBio Coverage

> 50x

S. pombe dg21

PacBio RS II sequencing at CSHL

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



S. pombe dg21

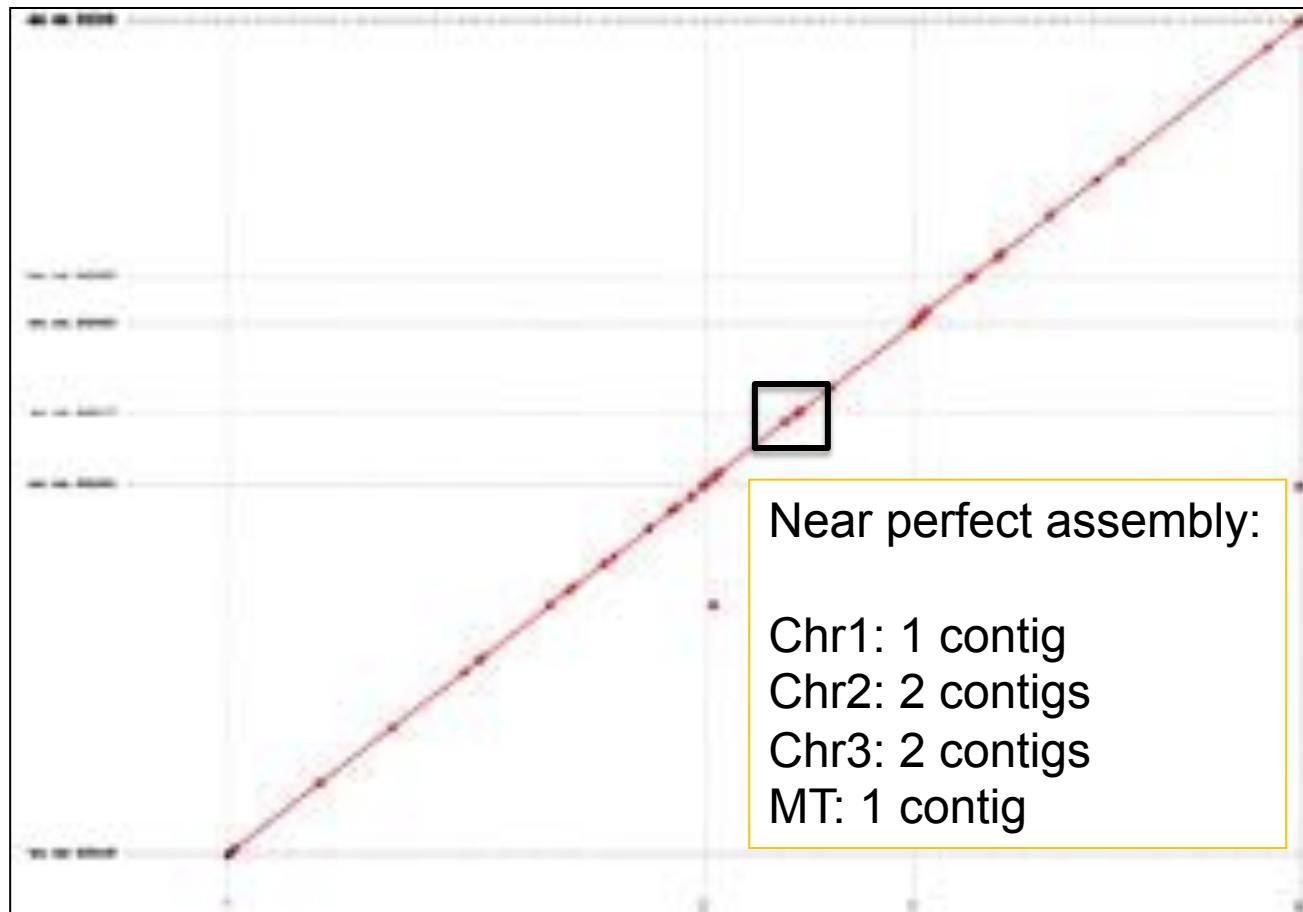
ASM294 Reference sequence

- 12.6Mbp; 3 chromo + mitochondria; N50: 4.53Mbp



PacBio assembly using HGAP + Celera Assembler

- 12.7Mbp; 13 non-redundant contigs; N50: 3.83Mbp; >99.98% id



S. pombe dg21

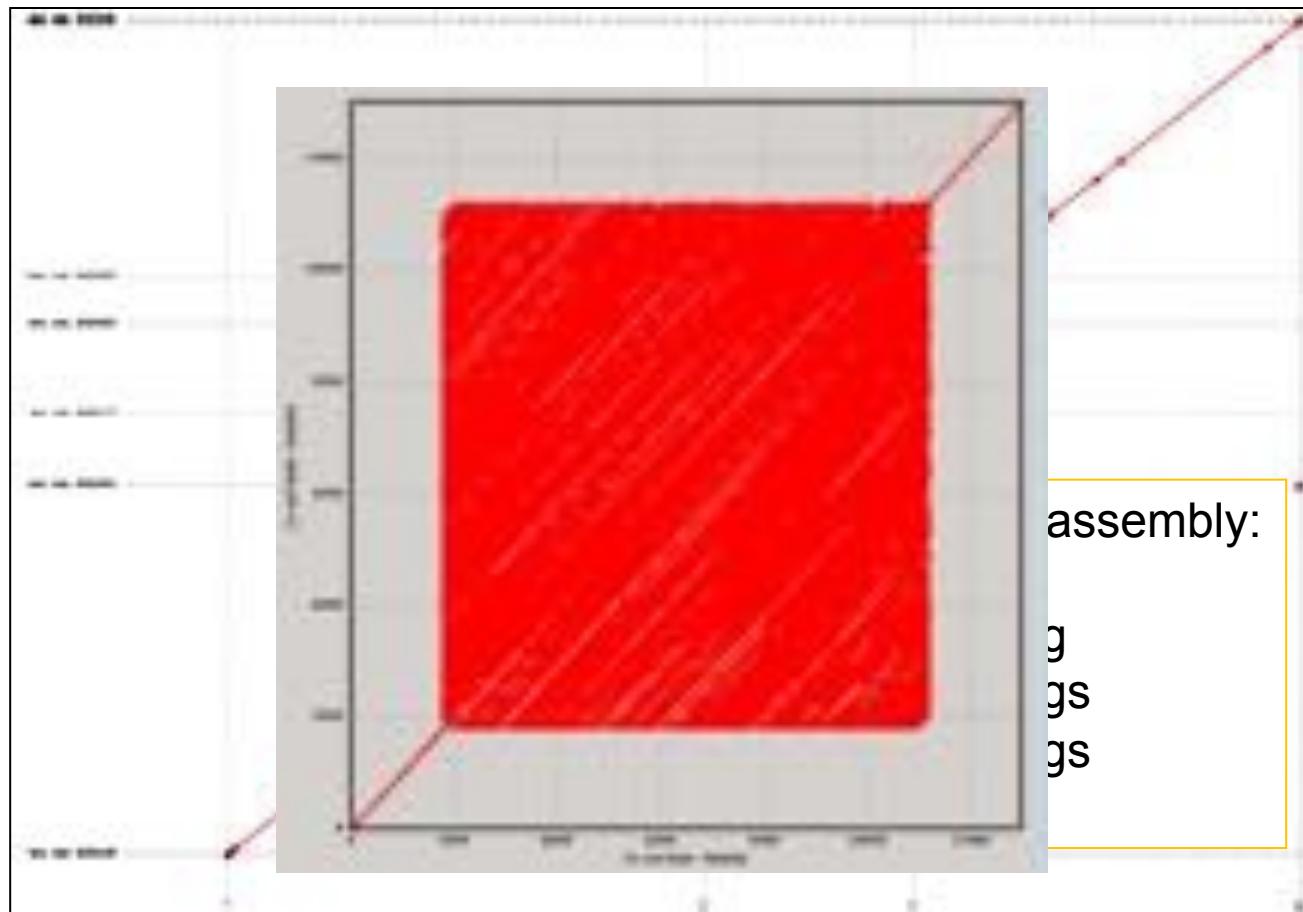
ASM294 Reference sequence

- 12.6Mbp; 3 chromo + mitochondria; N50: 4.53Mbp



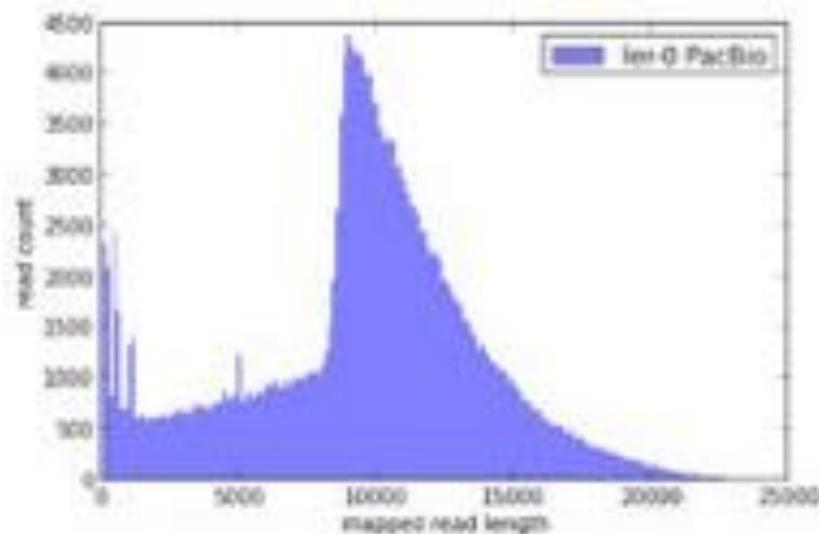
PacBio assembly using HGAP + Celera Assembler

- 12.7Mbp; 13 non-redundant contigs; N50: 3.83Mbp; >99.98% id



A. thaliana Ler-0

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



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

Genome size: 124.6 Mbp

Chromosome N50: 23.0 Mbp

Corrected coverage: 20x over 10kb

Sum of Contig Lengths: 149.5Mb

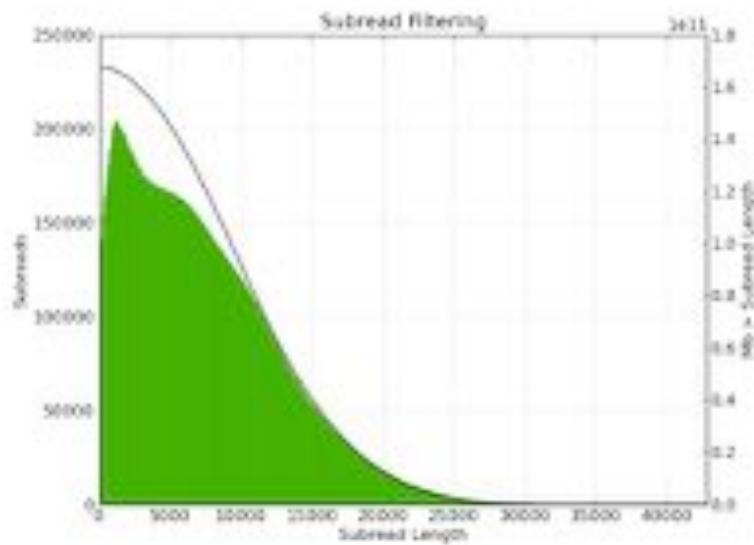
N50 Contig Length: 8.4 Mb

Number of Contigs: 1788

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

Human CHM I

<http://blog.pacificbiosciences.com/2014/02/data-release-54x-long-read-coverage-for.html>



CHM I *hert* sequenced at PacBio

- Sequenced using the P5 enzyme and C3 chemistry
- Size selection using an 20kb elution window on a BluePippin™ device from Sage Science
- Total coverage: 54x

Genome size: 3.0 Gb

Chromosome N50: 90.5 Mbp

Average read length: 7,680 bp

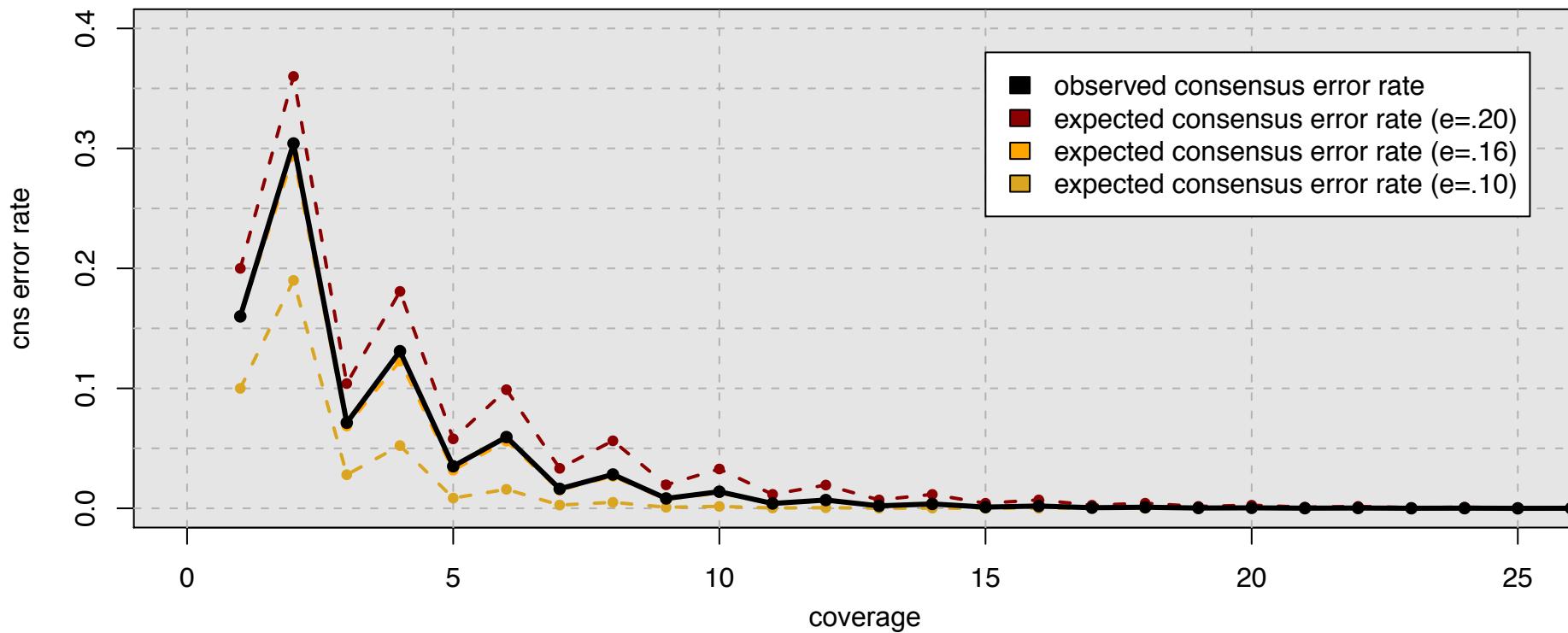
Sum of Contig Lengths: 3.2 Gb

N50 Contig Length: 4.38 Mbp

Max Contig: 44 Mbp

High quality draft assembly
Assembly Performance: 4.38Mbp/90.5Mbp = 4.5%
Sanger HuRef assembly: 107kbp / 90.5Mbp = .1%

Consensus Accuracy and Coverage



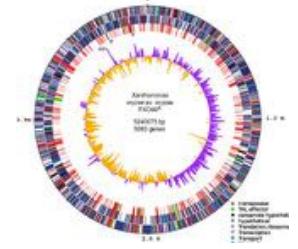
Coverage can overcome random errors

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

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

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

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

What should we expect from an assembly?

Analysis of dozens of genomes from across the tree of life with real and simulated data

Summary & Recommendations

- < 100 Mbp: HGAP/PacBio2CA @ 100x PB C3-P5
expect near perfect chromosome arms
- < 1GB: HGAP/PacBio2CA @ 100x PB C3-P5
high quality assembly: contig N50 over 1Mbp
- > 1GB: hybrid/gap filling
expect contig N50 to be 100kbp – 1Mbp
- > 5GB: Email mschatz@cshl.edu



Error correction and assembly complexity of single molecule sequencing reads.

Lee, H*, Gurtowski, J*, Yoo, S, Marcus, S, McCombie, WR, Schatz, MC

<http://www.biorxiv.org/content/early/2014/06/18/006395>

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

