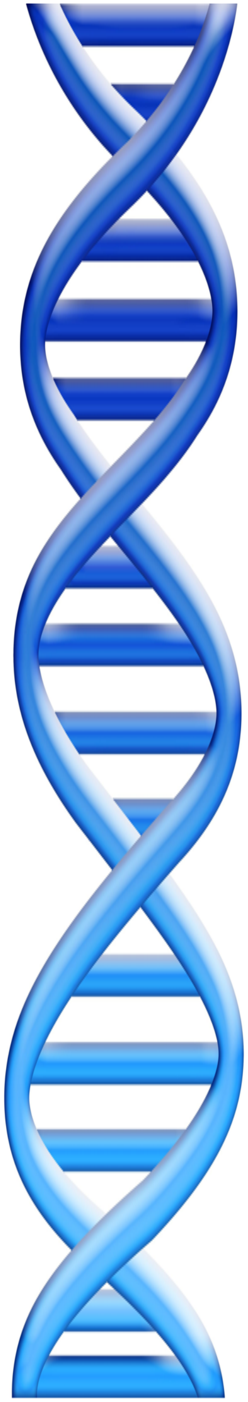
Genome Sequencing & Assembly

Michael Schatz

March 31, 2014

CSHL Genome Access

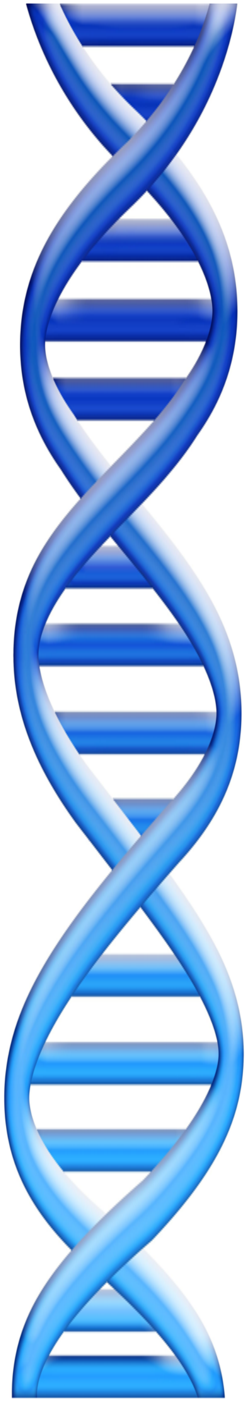


Outline

1. 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

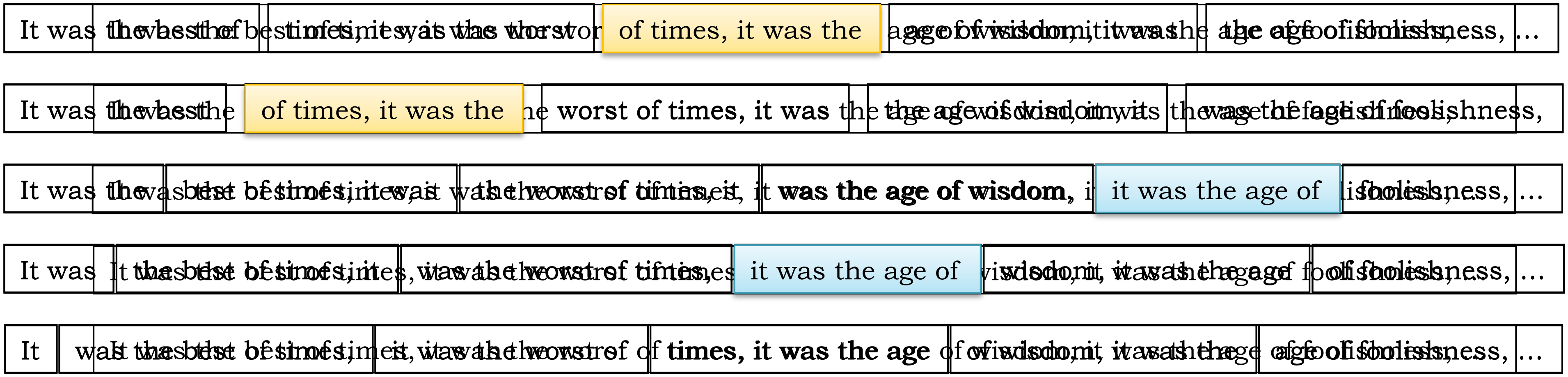


Outline

1. 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



- 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,

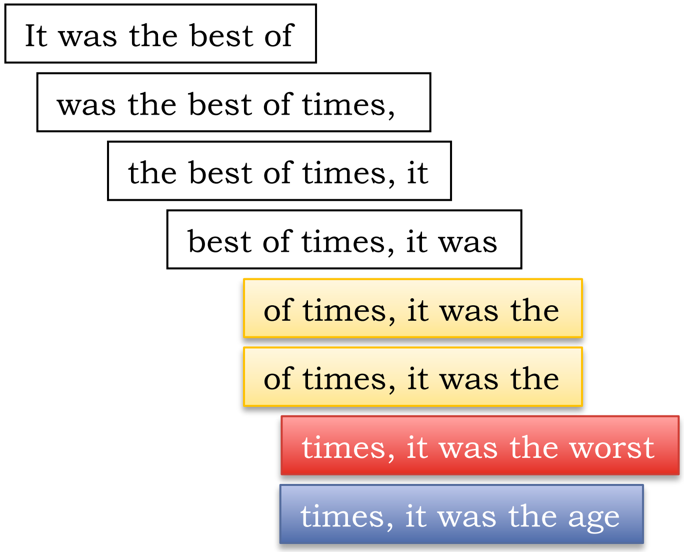
was the age of foolishness,

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

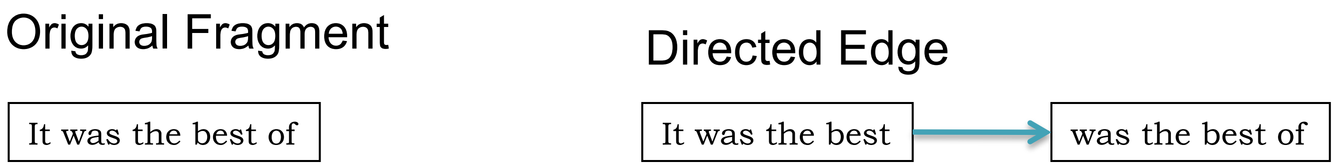
The repeated sequence make the correct reconstruction ambiguous

- 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

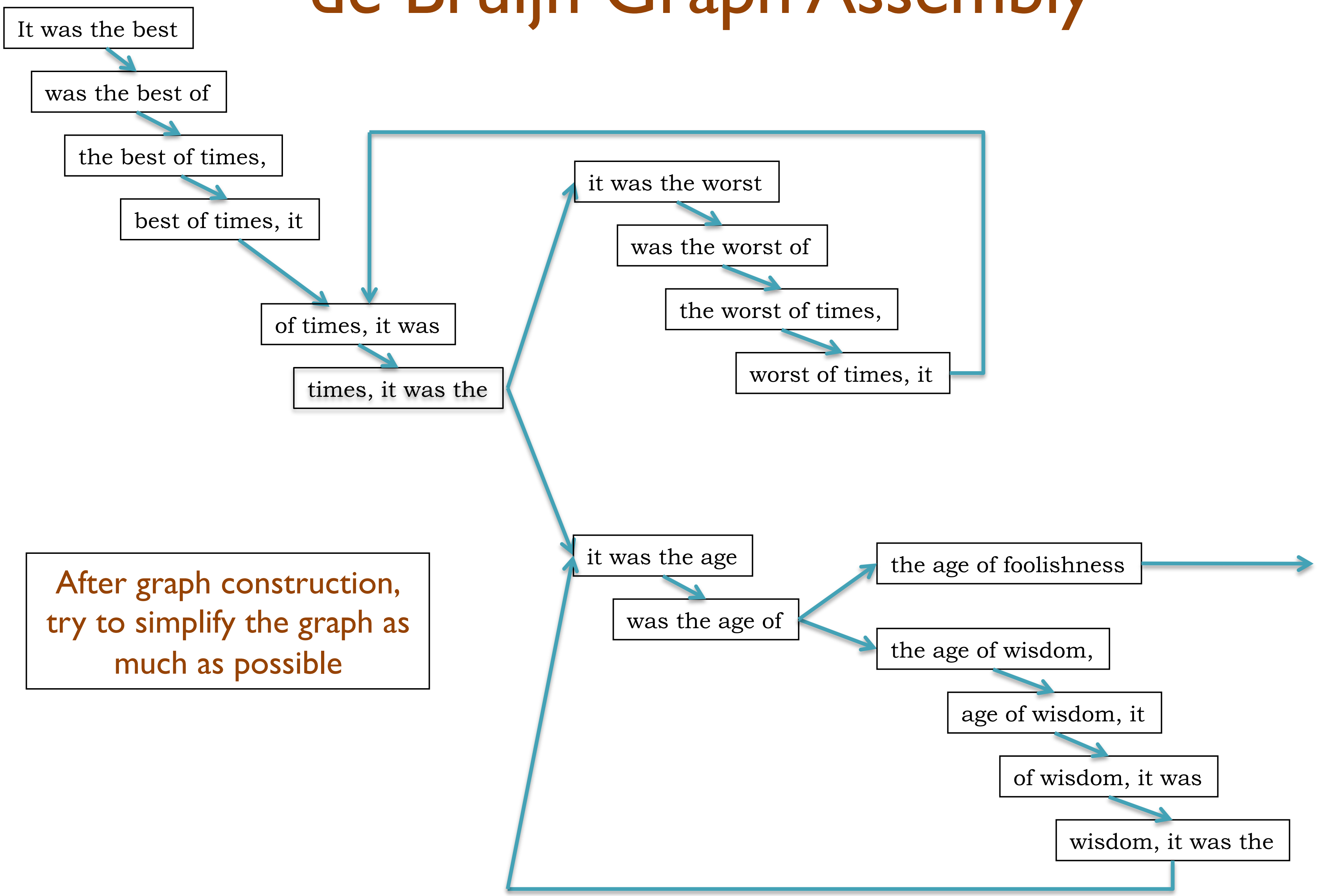
Directed Edge

It was the best → was the best of

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

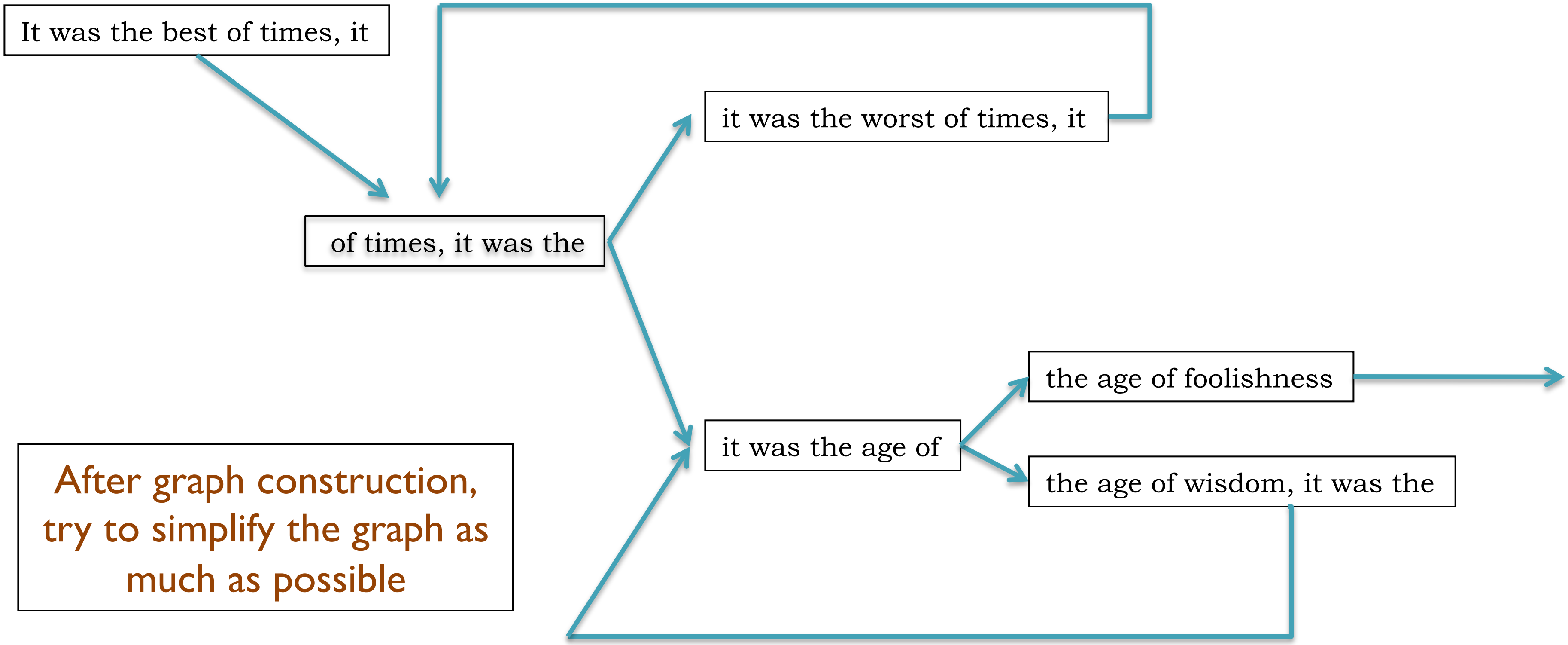
de Bruijn, 1946
Idury and Waterman, 1995
Pevzner, Tang, Waterman, 2001

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

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

The full tale

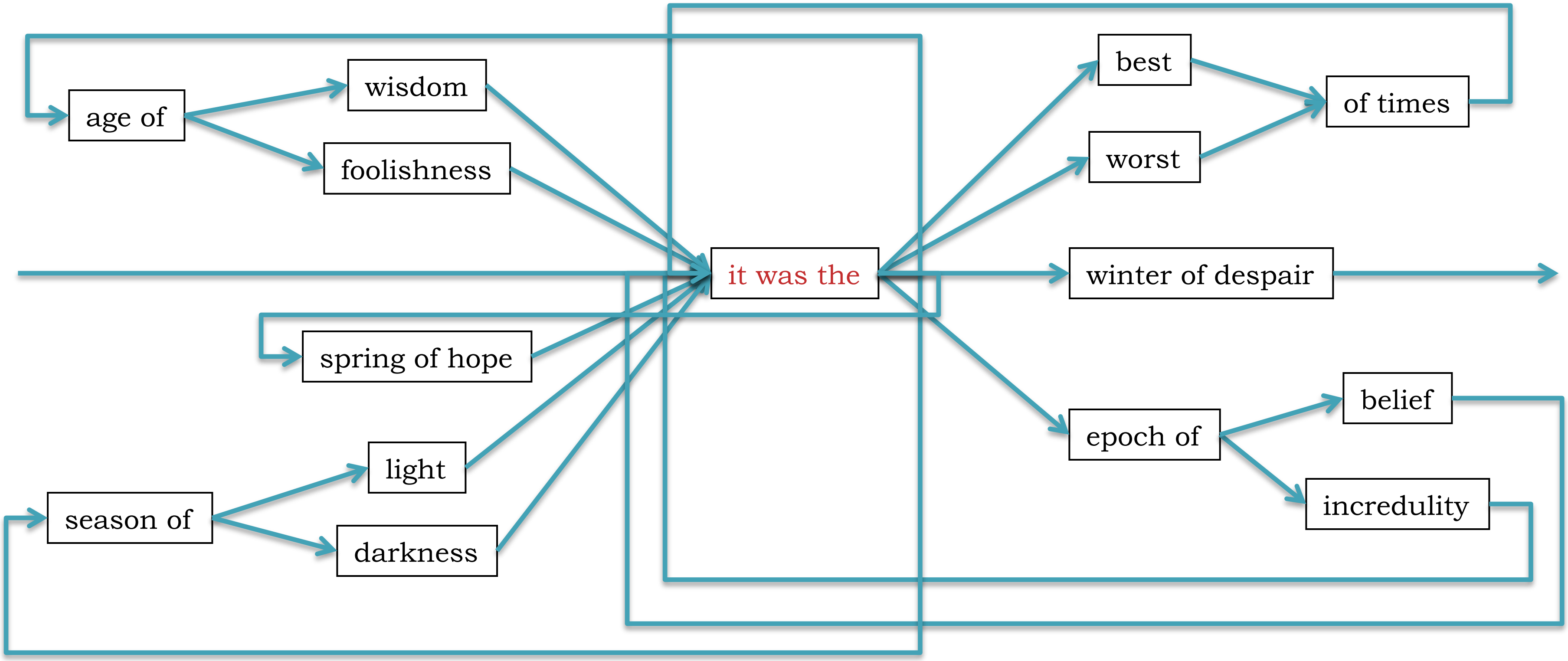
… it was the best of times it was the worst of times …

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

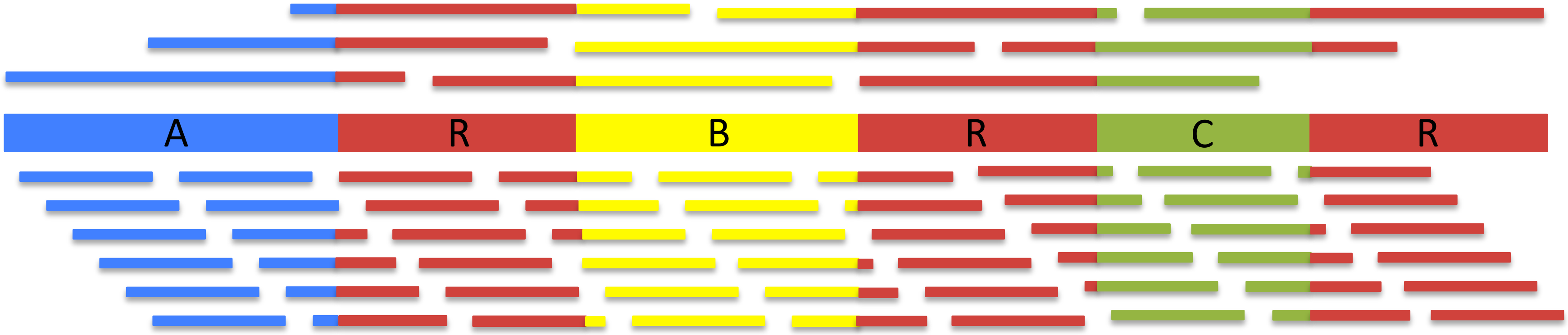
… it was the epoch of belief it was the epoch of incredulity …

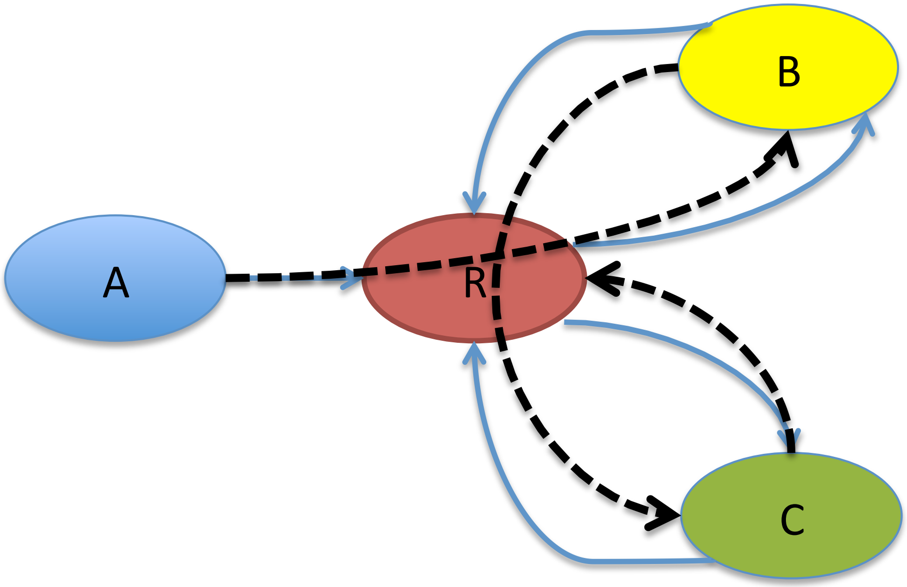
… it was the season of light it was the season of darkness …

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

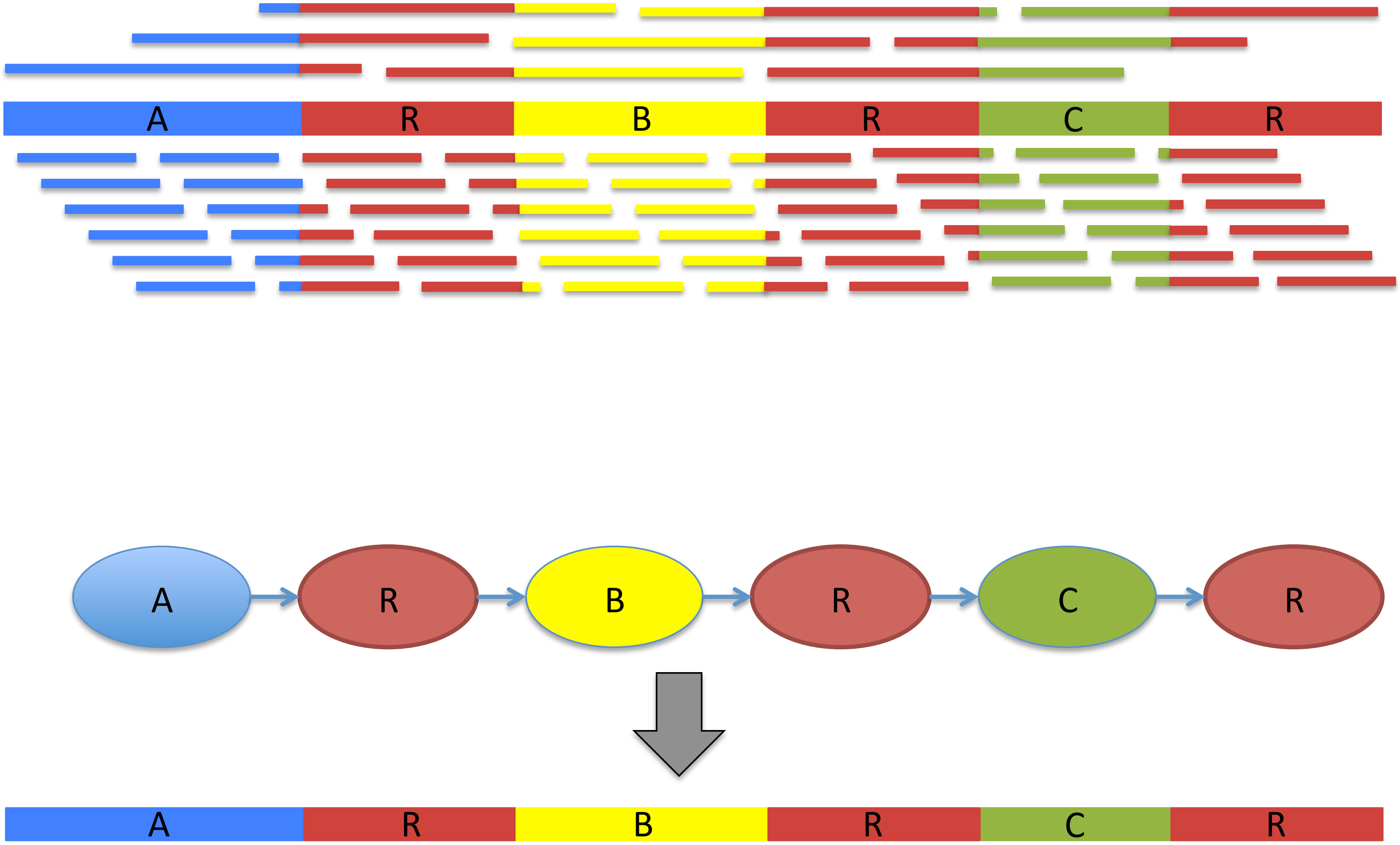


Assembly Complexity

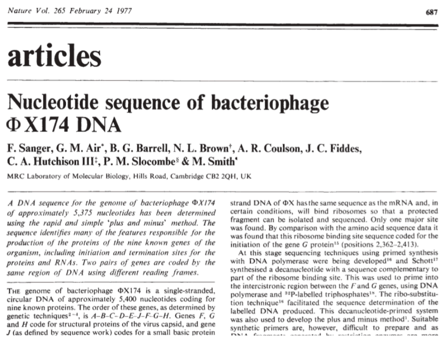




Assembly Complexity



Milestones in Genome Assembly



Nature Vol. 265 February 24 1977 **687**

articles

Nucleotide sequence of bacteriophage ΦX174 DNA

F. Sanger, G. M. Air[*], B. G. Barrell, N. L. Brown[†], A. R. Coulson, J. C. Fiddes, C. A. Hutchison III[‡], P. M. Slocombe[§] & M. Smith[*]

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

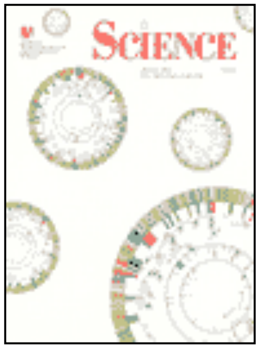
A DNA sequence for the genome of bacteriophage ΦX174 of approximately 5,375 nucleotides has been determined using the rapid and simple 'plus and minus' method. The sequence identifies many of the features responsible for the production of the proteins of the nine known genes of the organism, including 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 ΦX174 is a single-stranded, circular DNA of approximately 5,400 nucleotides coding for nine known proteins. The order of these genes, as determined by genetic techniques[2–4], is *A–B–C–D–E–J–F–G–H*. Genes *F*, *G* and *H* code for structural proteins of the virus capsid, and gene *J* (as defined by sequence work) codes for a small basic protein

strand DNA of ΦX has the same sequence as the mRNA and, in certain conditions, will bind ribosomes so that a protected fragment can be isolated and sequenced. Only one major site was found. By comparison with the amino acid sequence data it was found that this ribosome binding site sequence coded for the initiation of the gene *G* protein[15] (positions 2,362–2,413).

At this stage sequencing techniques using primed synthesis with DNA polymerase were being developed[16] and Schott[17] synthesised a decanucleotide with a sequence complementary to part of the ribosome binding site. This was used to prime into the intercistronic region between the *F* and *G* genes, using DNA polymerase and ^{32}P-labelled triphosphates[18]. The ribo-substitution technique[16] facilitated the sequence determination of the labelled DNA produced. This decanucleotide-primed system was also used to develop the plus and minus method[1]. Suitable synthetic primers are, however, difficult to prepare and as

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



1995. Fleischmann *et al.*
1st Free Living Organism
TIGR Assembler. 1.8Mbp

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

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

nature
the human genome
Nuclear fission
Five-dimensional energy landscapes
Seafloor spreading
The view from under the Arctic ice
Career prospects
Sequence creates new opportunities
naturejobs
genomics special

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

2010. Li *et al.*
1st Large SGS Assembly.
SOAPdenovo 2.2 Gbp

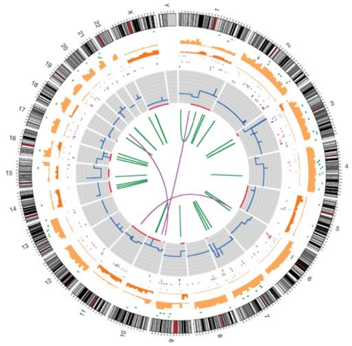
Like Dickens, we must computationally reconstruct a genome from short fragments

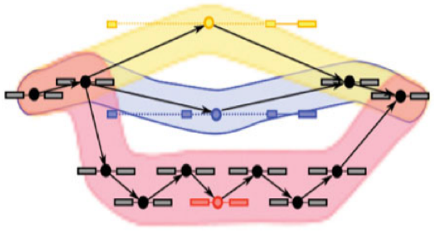
Assembly Applications

- Novel genomes

- Metagenomes

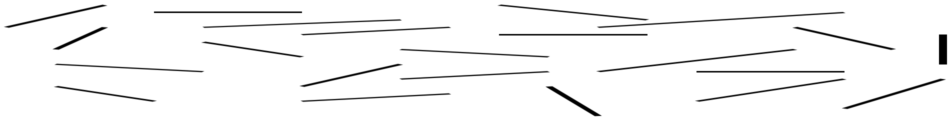
- Sequencing assays
  - Structural variations
  - Transcript assembly
  - …





Assembling a Genome

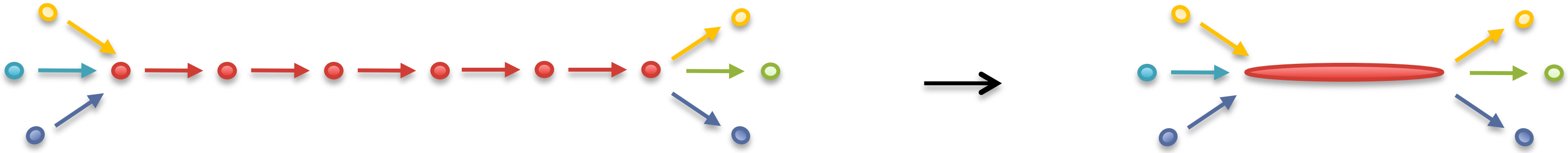
1. Shear & Sequence DNA



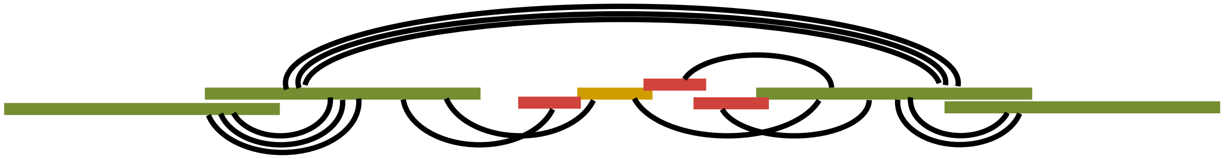
2. Construct assembly graph from overlapping reads

...AGCCTAGGGATGCGCGACACGT
GGATGCGCGACACGTCGCATATCCGGTTTGGTCAACCTCGGACGGAC
CAACCTCGGACGGACCTCAGCGAA...

3. Simplify assembly graph



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

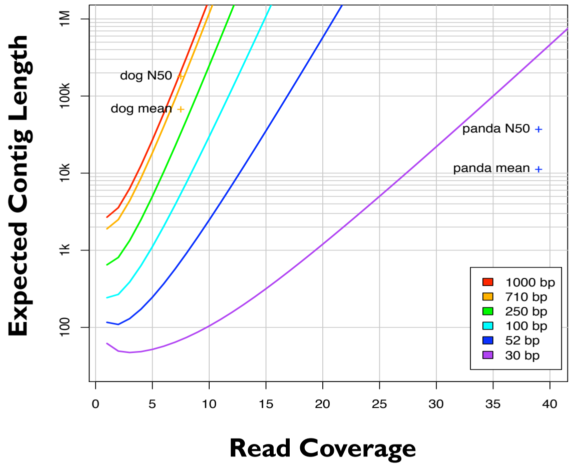


Why are genomes hard to assemble?

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

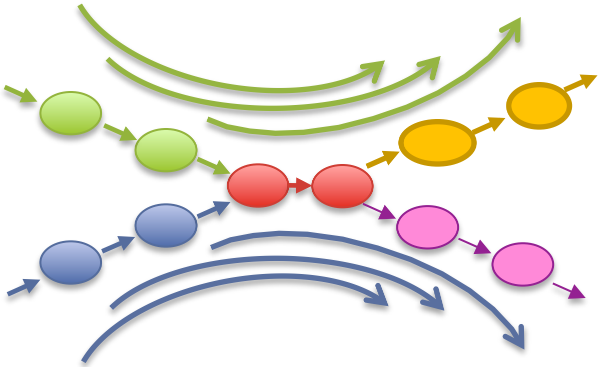
Coverage



High coverage is required

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

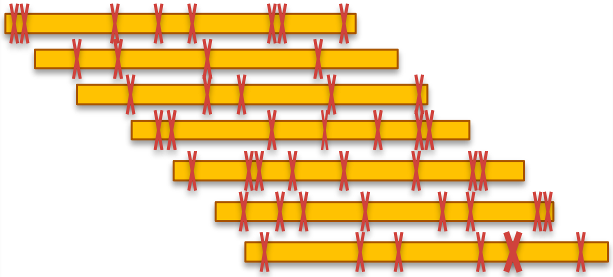
Read Length



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

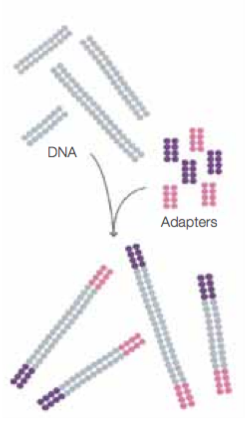


Errors obscure overlaps

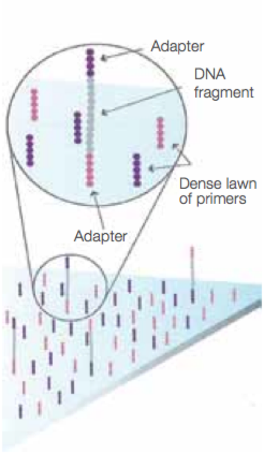
- Reads are assembled by finding kmers shared in pair of reads
- High error rate requires very short seeds, increasing complexity and forming assembly hairballs

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

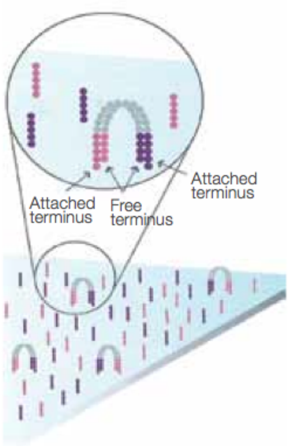
Illumina Sequencing by Synthesis



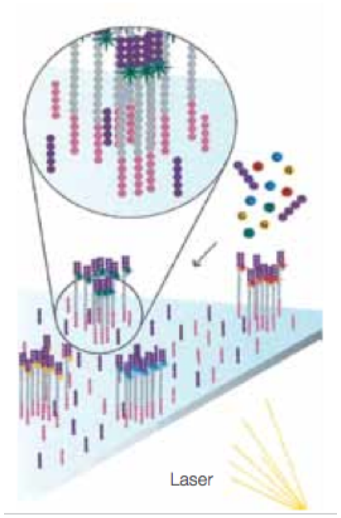
1. Prepare



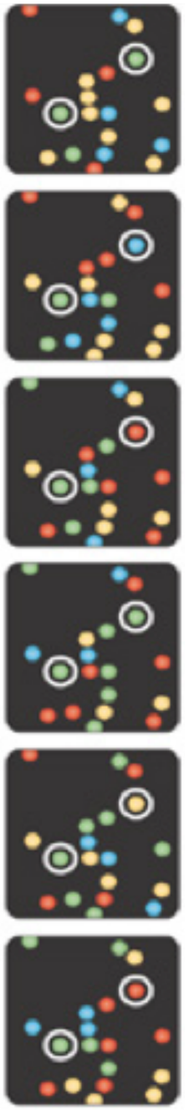
2. Attach



3. Amplify



4. Image

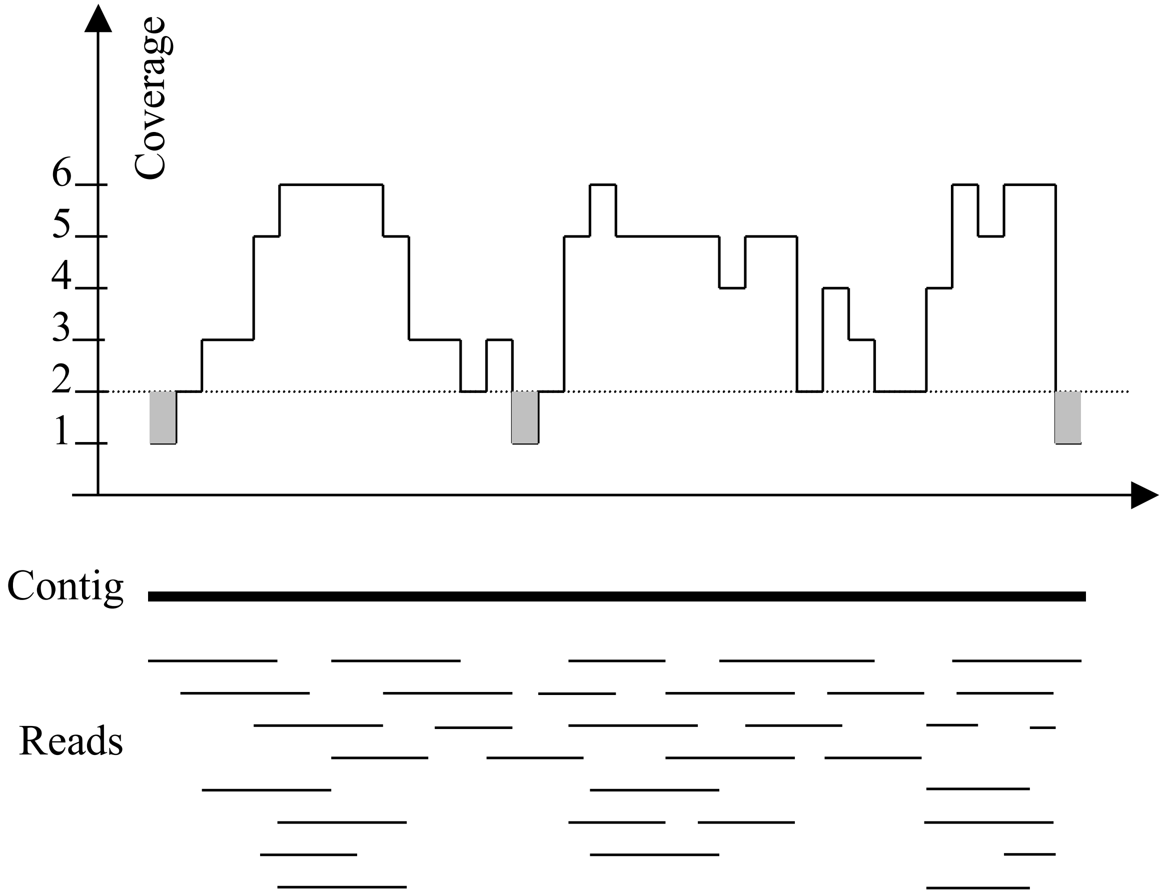


5. Basecall

Metzker (2010) Nature Reviews Genetics 11:31-46

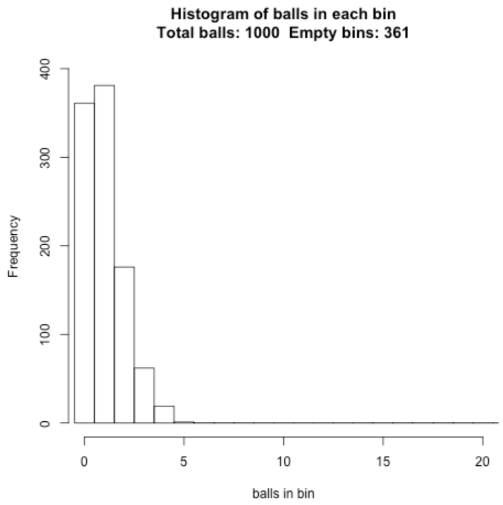
http://www.youtube.com/watch?v=l99aKKHcxC4

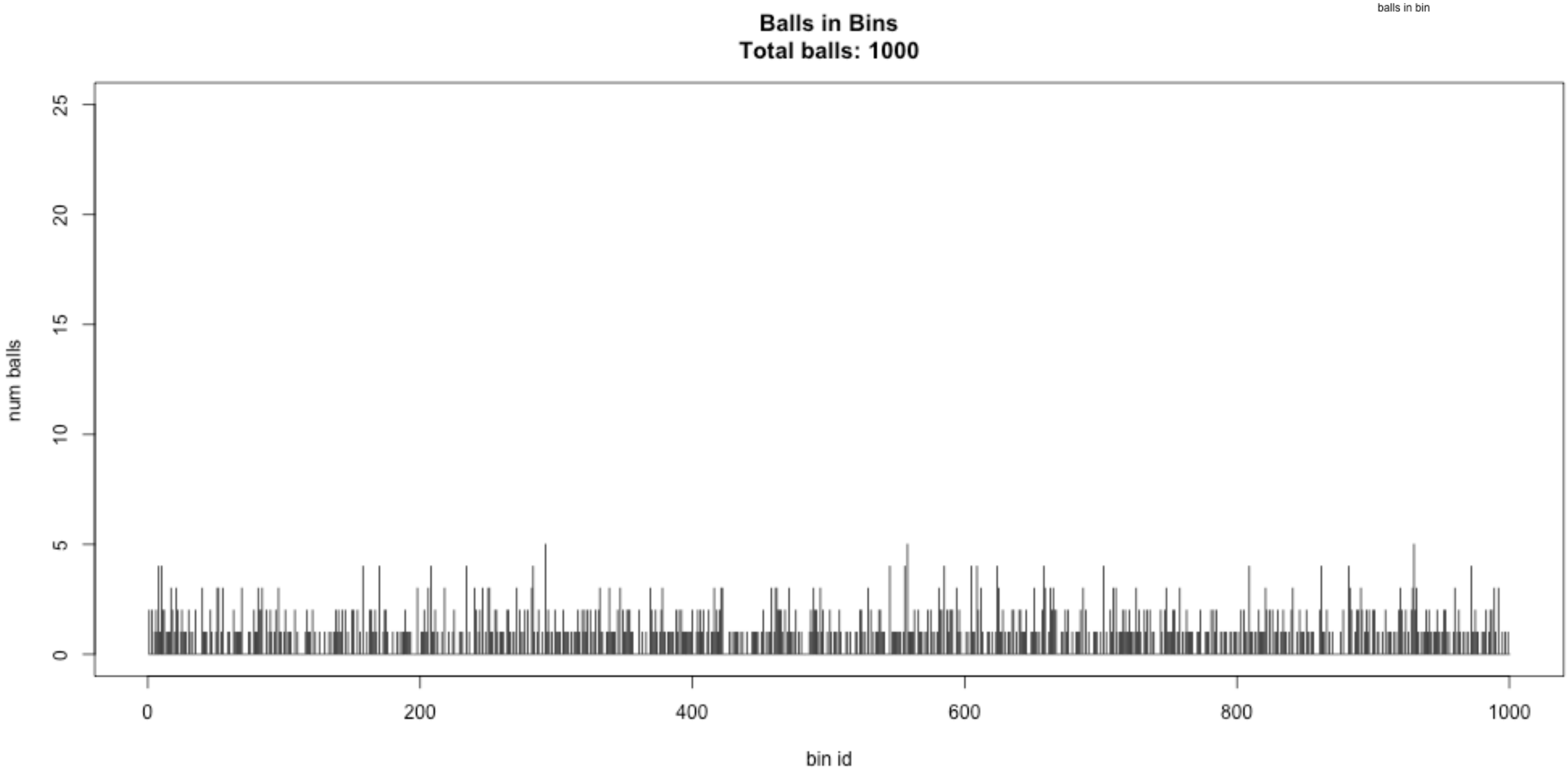
Typical contig coverage



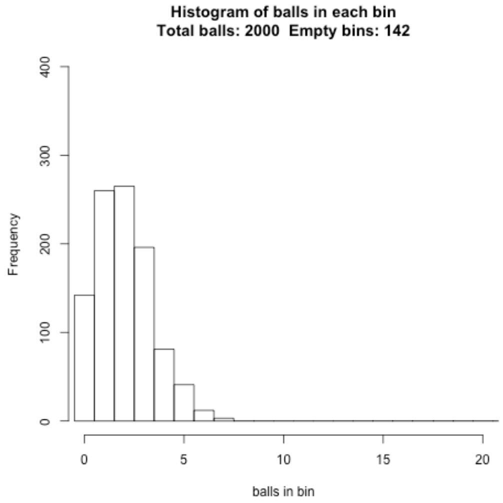
Imagine raindrops on a sidewalk

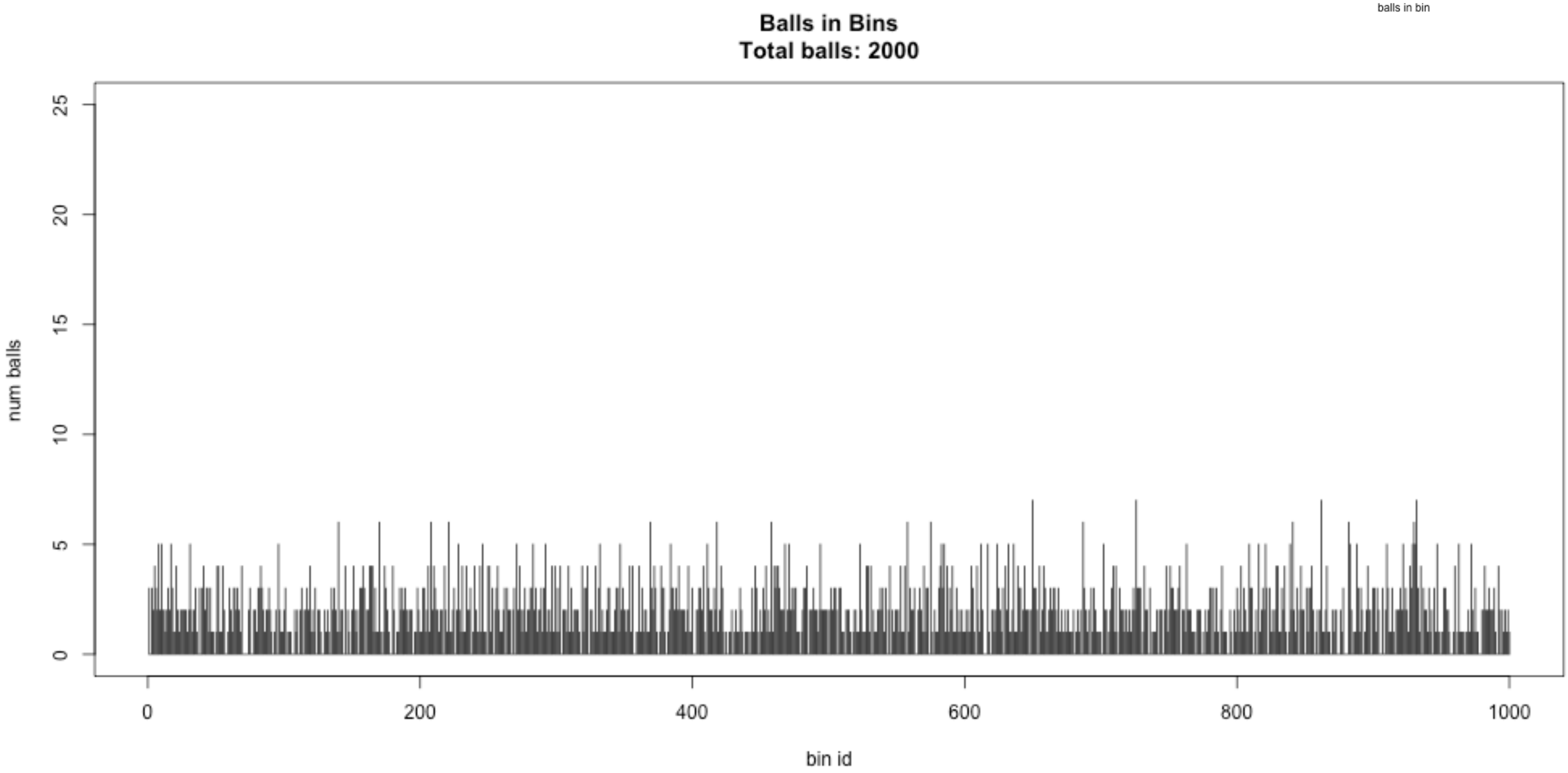
Balls in Bins 1x



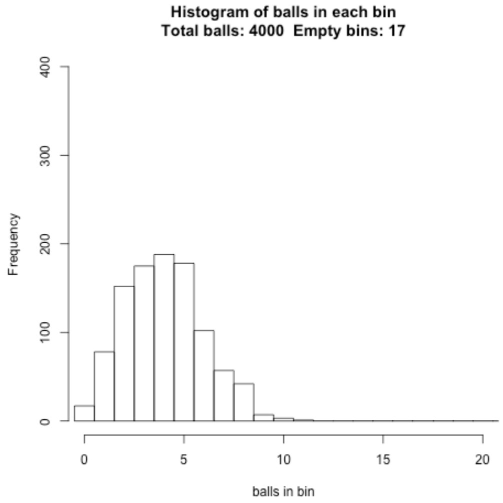


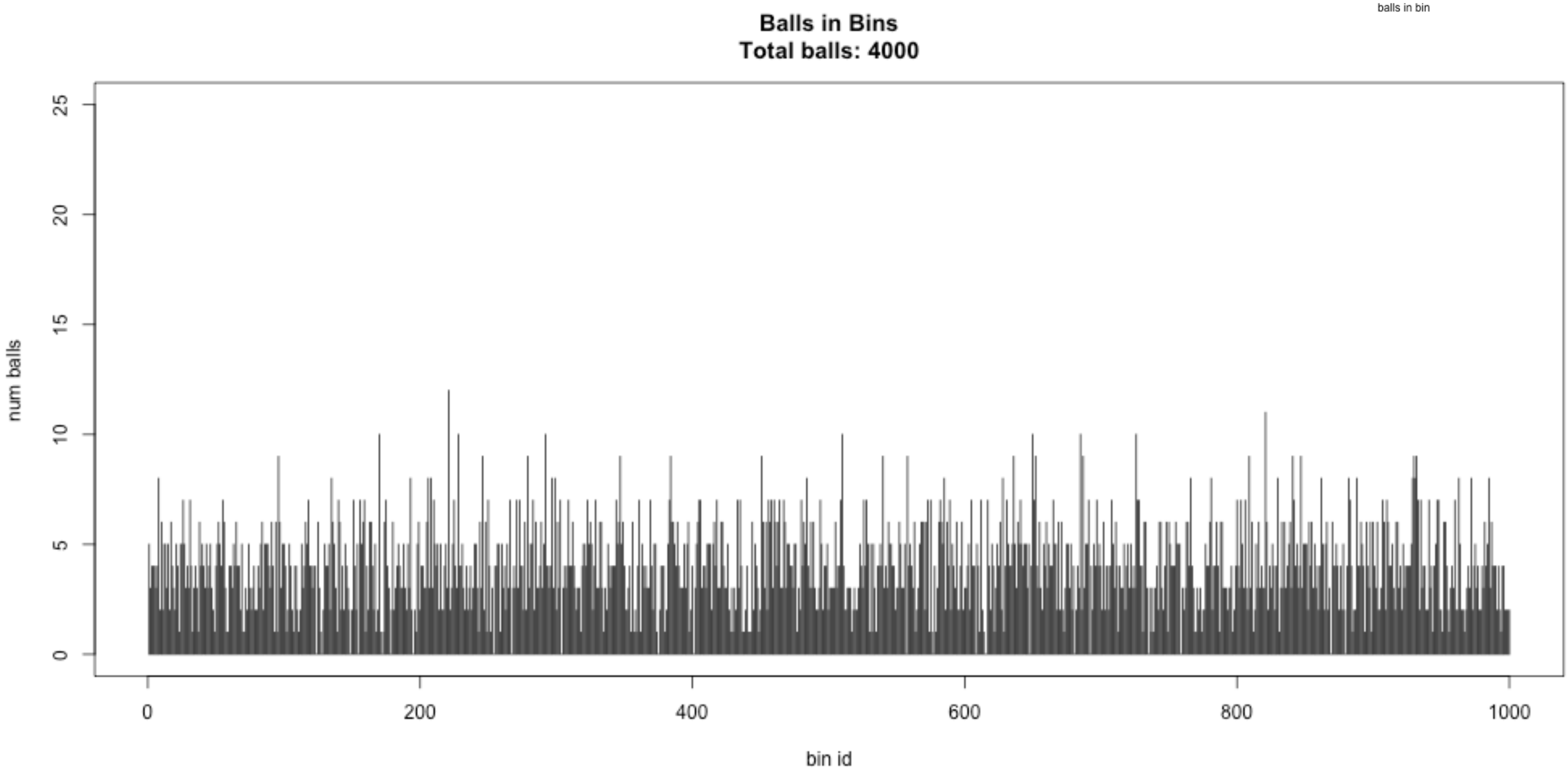
Balls in Bins 2x



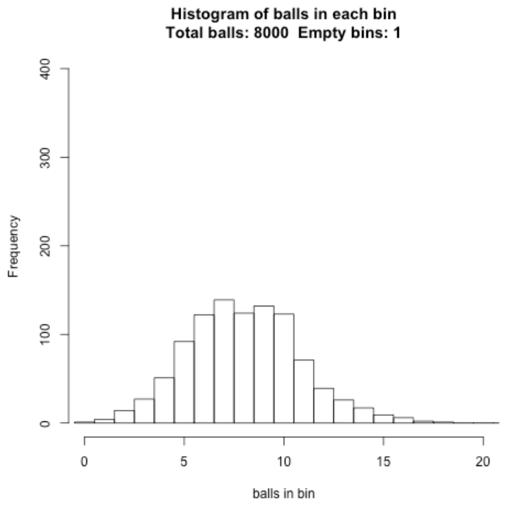


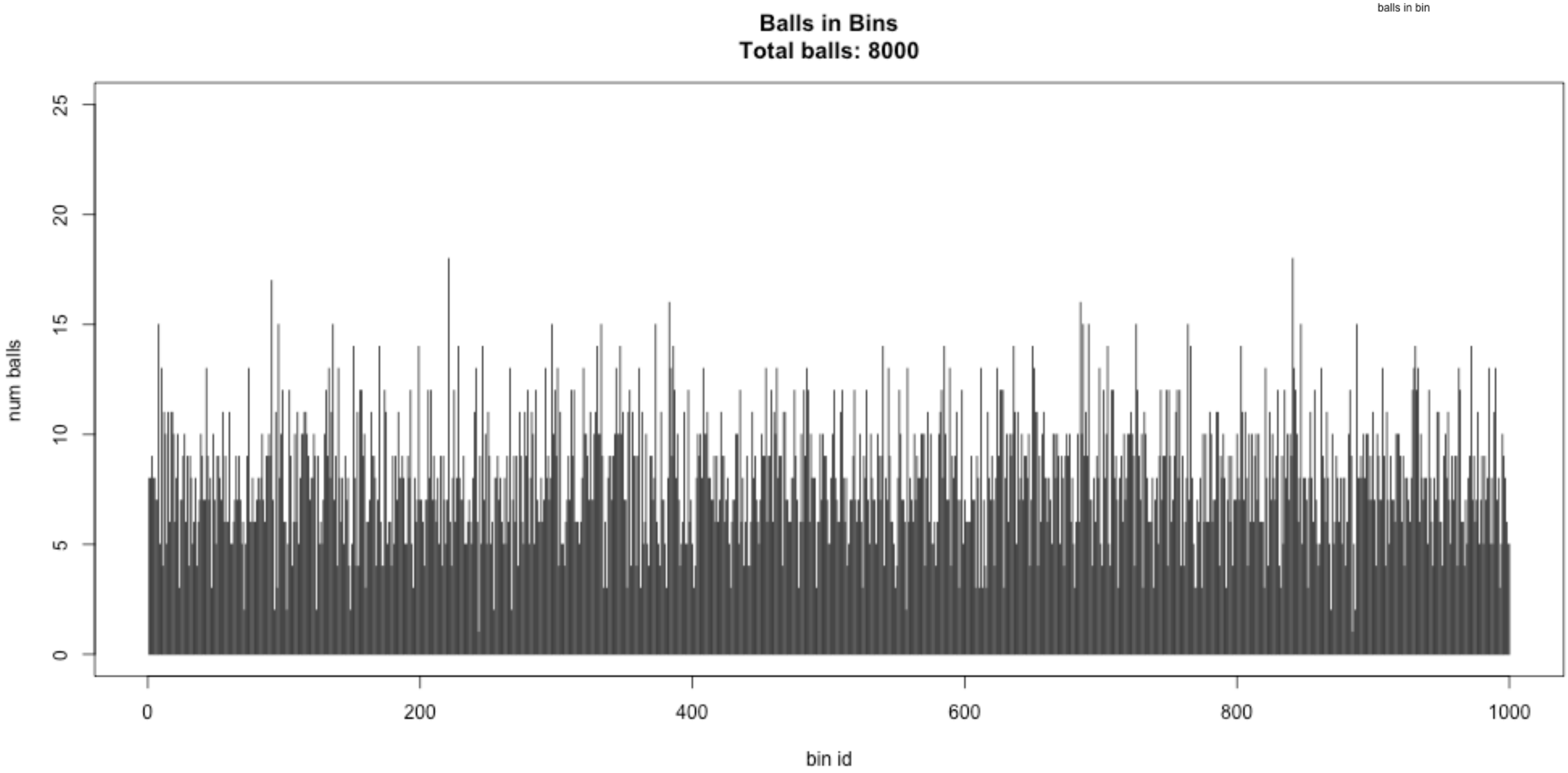
Balls in Bins 4x





Balls in Bins 8x





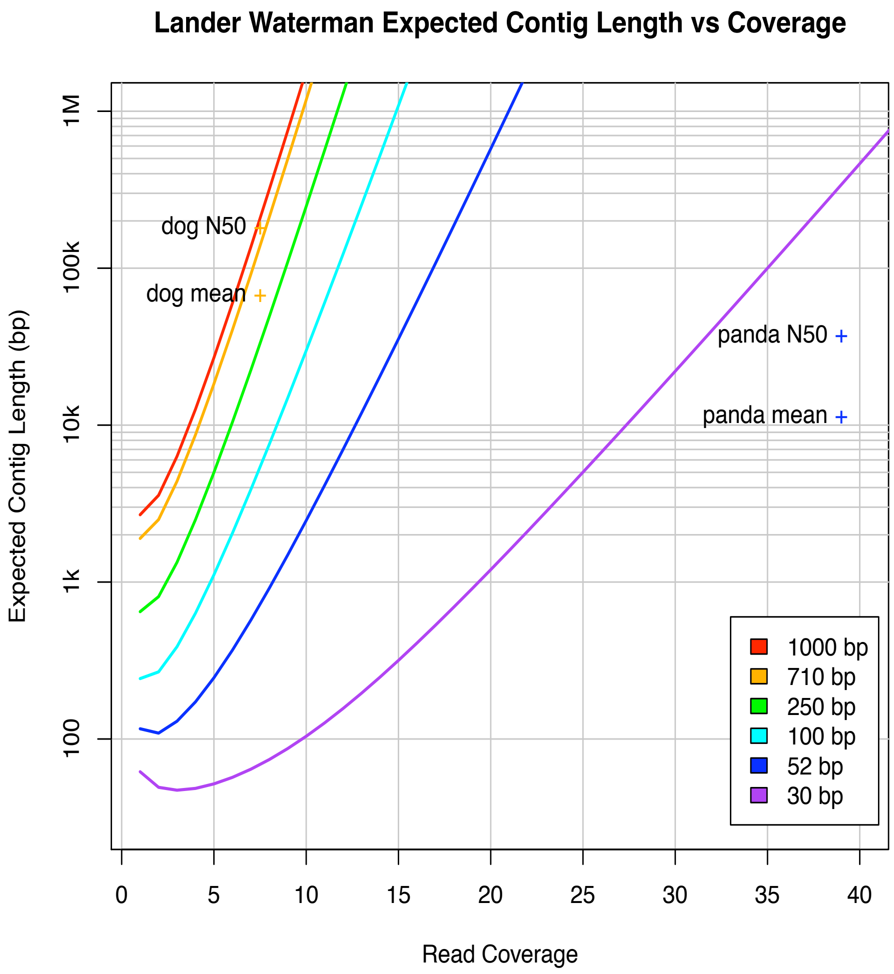
Coverage and Read Length

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

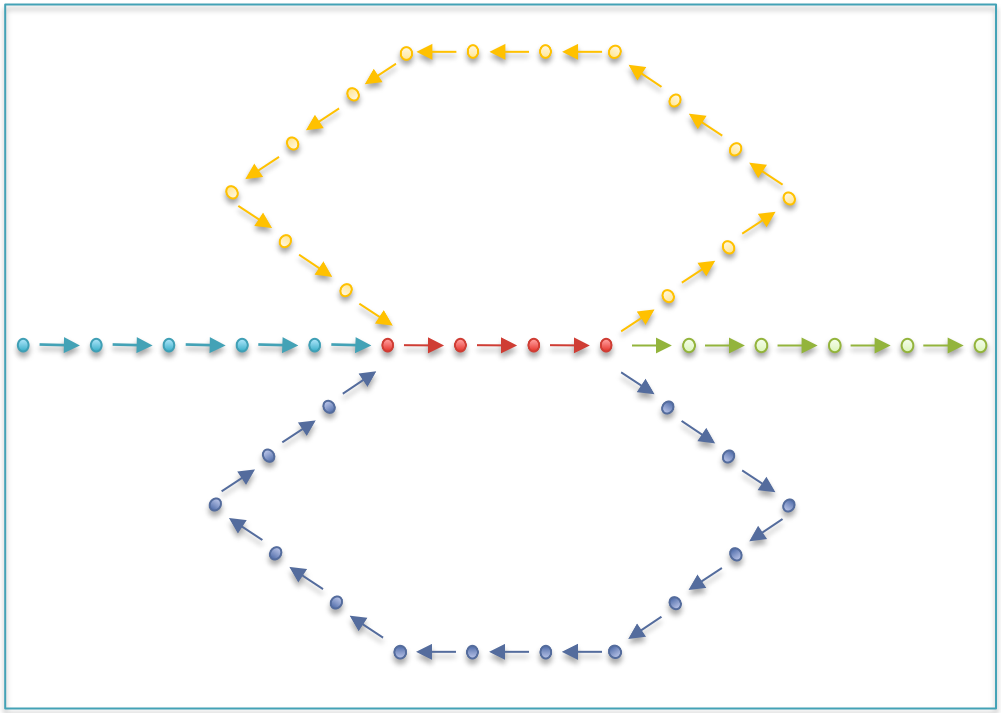


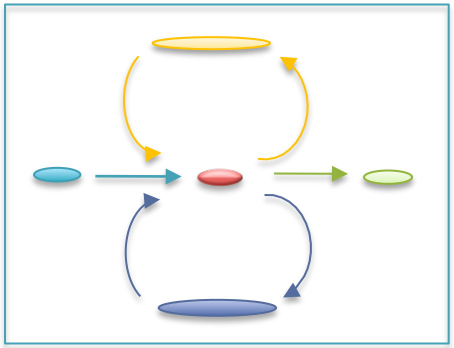
Assembly of Large Genomes using Second Generation Sequencing
Schatz MC, Delcher AL, Salzberg SL (2010) *Genome Research.* 20:1165-1173.

Errors

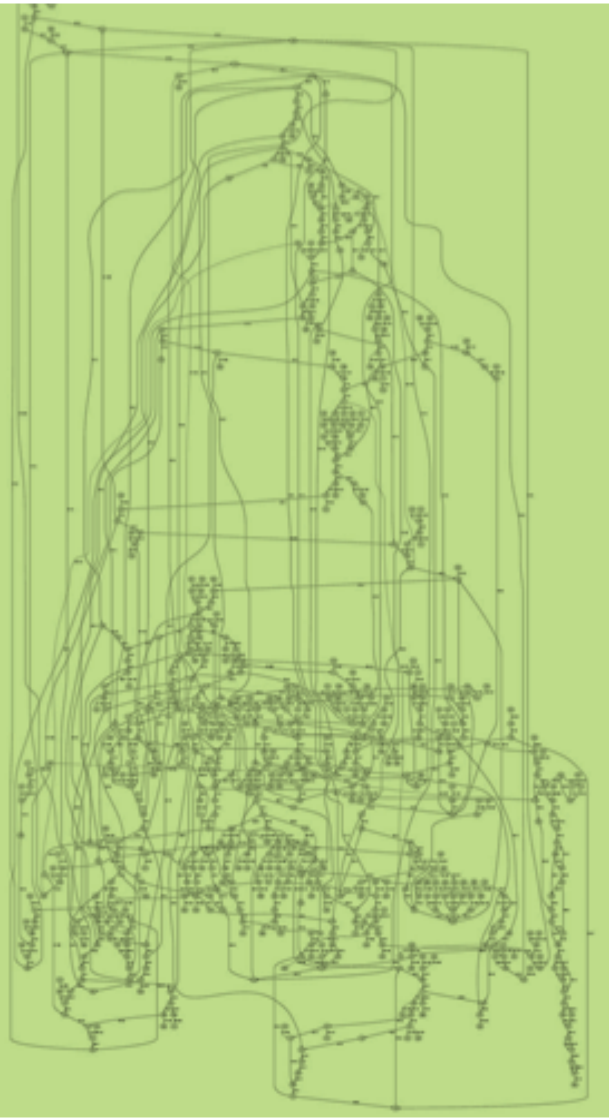
Unitigging / Unipathing

- 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, and (3) repeats





Errors in the graph



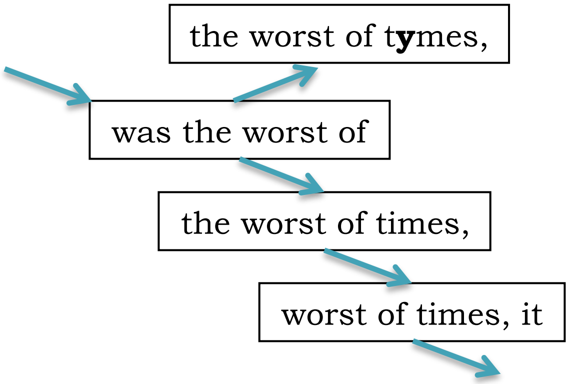
(Chaisson, 2009)

Clip Tips

was the worst of times,

was the worst of t**y**mes,

the worst of times, it



the worst of t**y**mes,

was the worst of

the worst of times,

worst of times, it

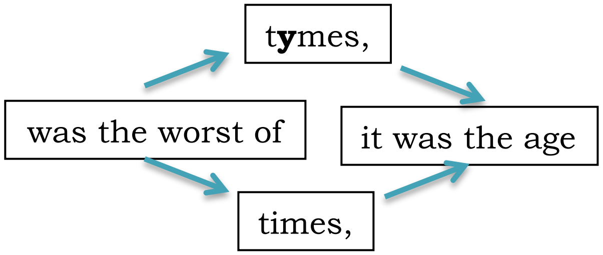
Pop Bubbles

was the worst of times,

was the worst of t**y**mes,

times, it was the age

t**y**mes, it was the age



t**y**mes,

was the worst of

it was the age

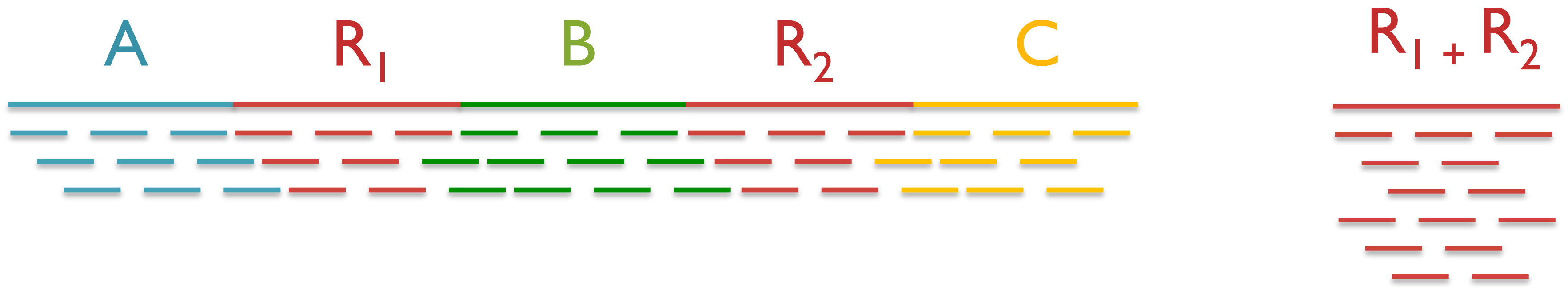
times,

Repetitive regions

Repeat Type	Definition / Example	Prevalence
Low-complexity DNA / Microsatellites	$(b_1b_2\ldots b_k)^N$ where $1 \leq k \leq 6$ CACACACACACACACACACA	2%
SINEs (Short Interspersed Nuclear Elements)	*Alu* sequence (~280 bp) Mariner elements (~80 bp)	13%
LINEs (Long Interspersed Nuclear Elements)	~500 – 5,000 bp	21%
LTR (long terminal repeat) retrotransposons	Ty1-copia, Ty3-gypsy, Pao-BEL (~100 – 5,000 bp)	8%
Other DNA transposons		3%
Gene families & segmental duplications		4%

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

Repeats and Coverage Statistics



- If *n* reads are a uniform random sample of the genome of length *G*, we expect *k*=*n Δ* /*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

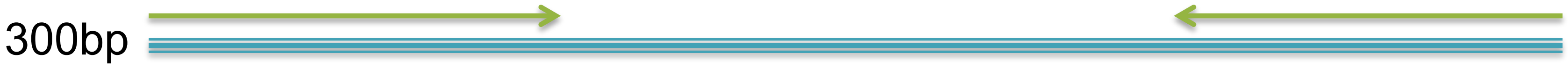
$$\Pr(X-copy) = \binom{n}{k}\left(\frac{X\Delta}{G}\right)^{k}\left(\frac{G-X\Delta}{G}\right)^{n-k}$$

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

Paired-end and Mate-pairs

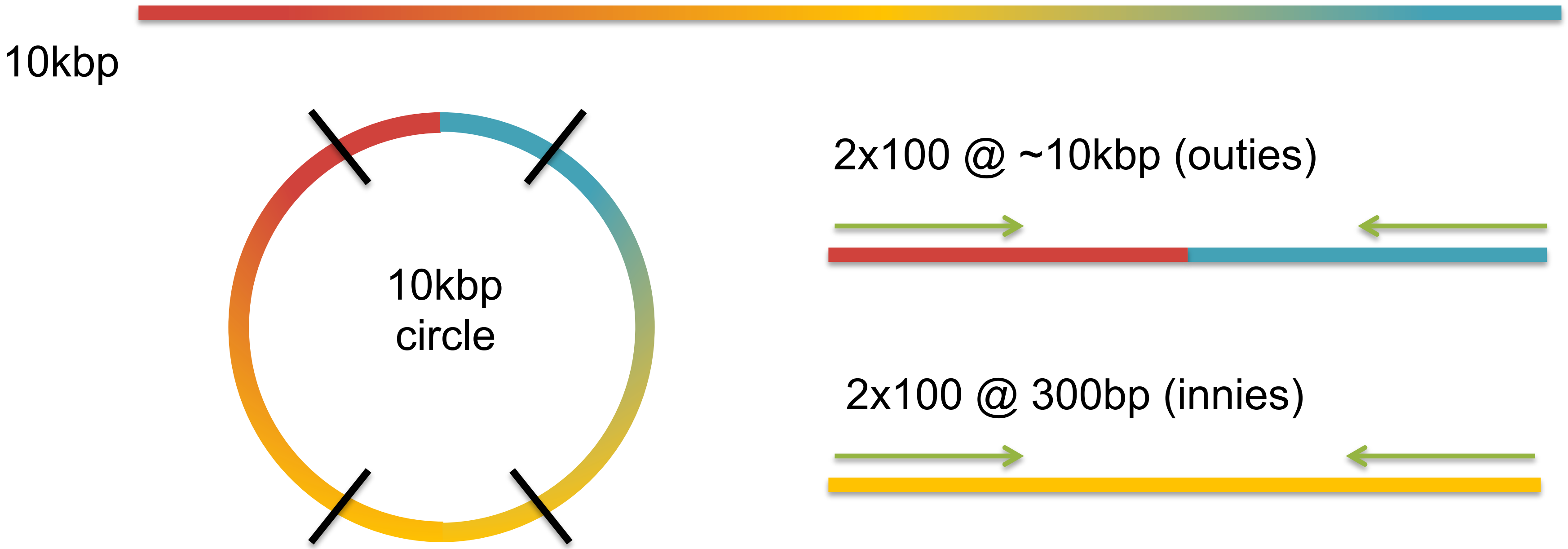
Paired-end sequencing

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



Mate-pair sequencing

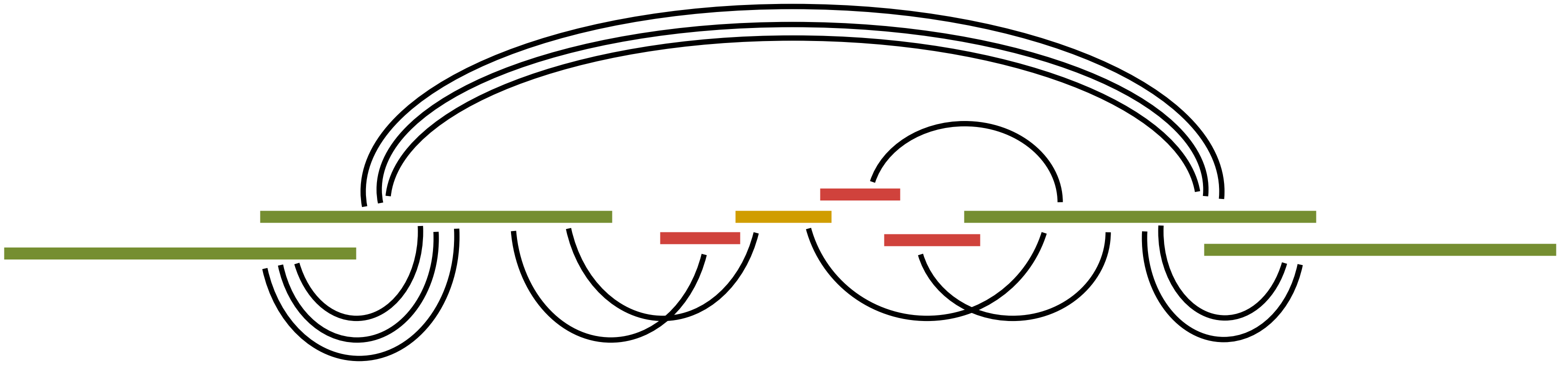
- Circularize long molecules (1-10kbp), shear into fragments, & sequence
- Mate failures create short paired-end reads



Scaffolding

- Initial contigs (*aka* unipaths, unitigs) terminate at
  - *Coverage gaps*: especially extreme GC regions
  - *Conflicts*: sequencing errors, repeat boundaries

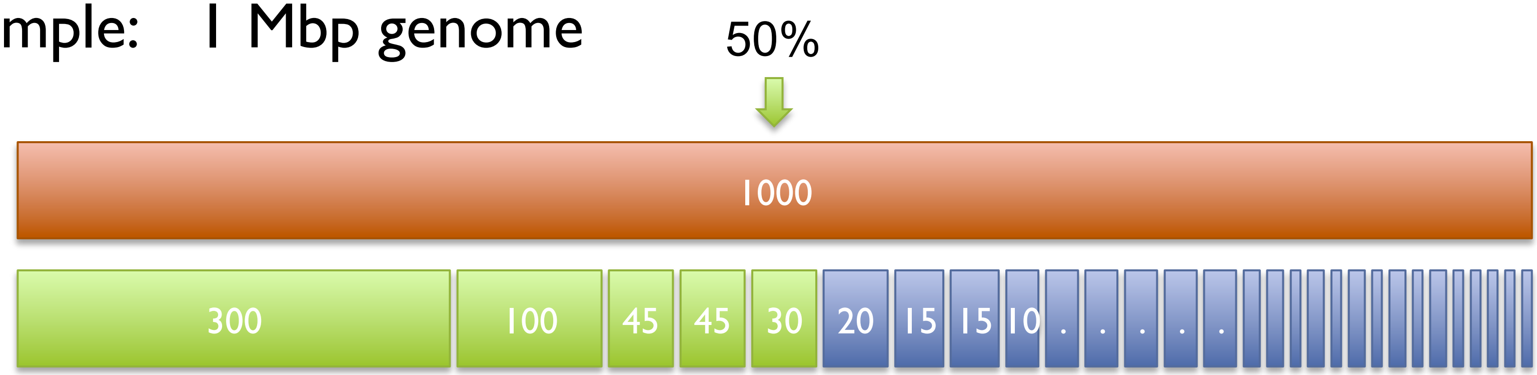
- Iteratively resolve longest, 'most unique' contigs
  - Both overlap graph and de Bruijn assemblers initially collapse repeats into single copies
  - Uniqueness measured by a statistical test on coverage



N50 size

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

Example: 1 Mbp genome



N50 size = 30 kbp

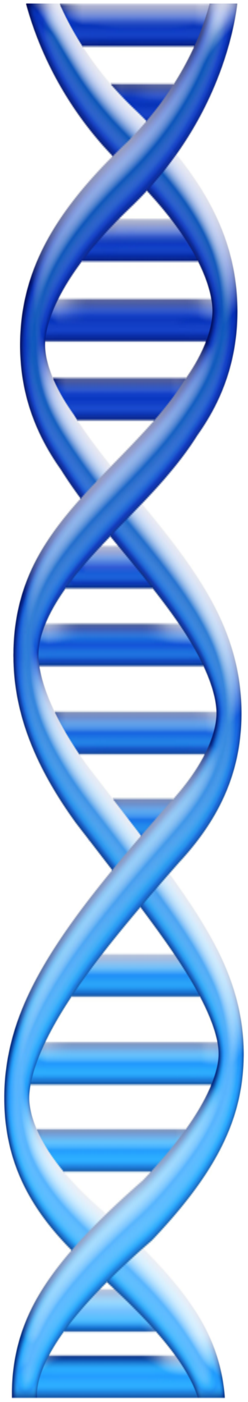
(300k+100k+45k+45k+30k = 520k >= 500kbp)

Note:

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

Break

PHASE TWO : INTERPRETATION
SHENEMAN The Star-Ledger
I THINK I FOUND A CORNER PIECE.
3 BILLION PIECES
GENOME



Outline

1. 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

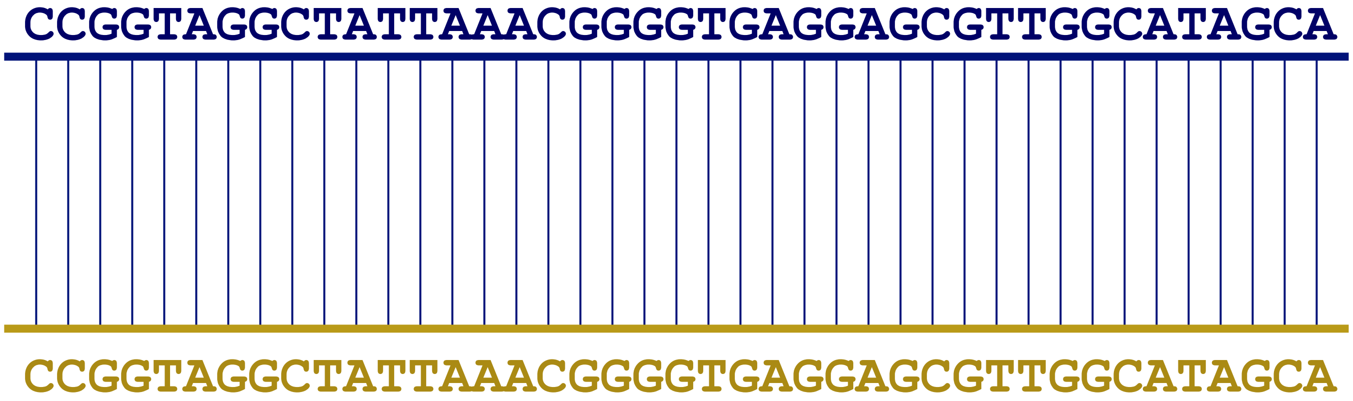
3. Genome assemblers
   1. ALLPATHS-LG: recommended for Illumina-only projects
   2. Celera Assembler: recommended for PacBio projects

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*

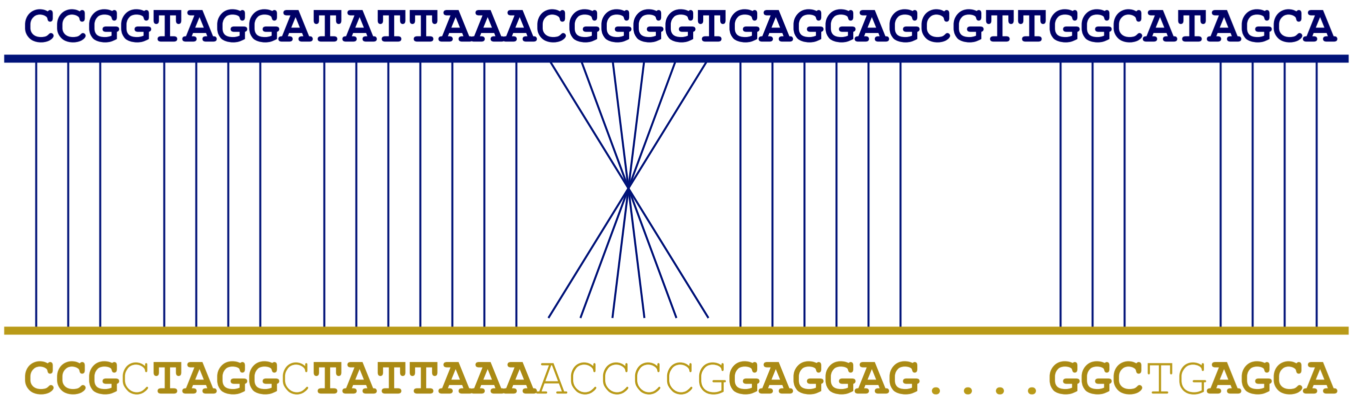


CCGGTAGGCTATTAAACGGGGTGAGGAGCGTTGGCATAGCA

CCGGTAGGCTATTAAACGGGGTGAGGAGCGTTGGCATAGCA

Not so fast...

- Genome *A* may have insertions, deletions, translocations, inversions, duplications or SNPs with respect to *B* (sometimes all of the above)



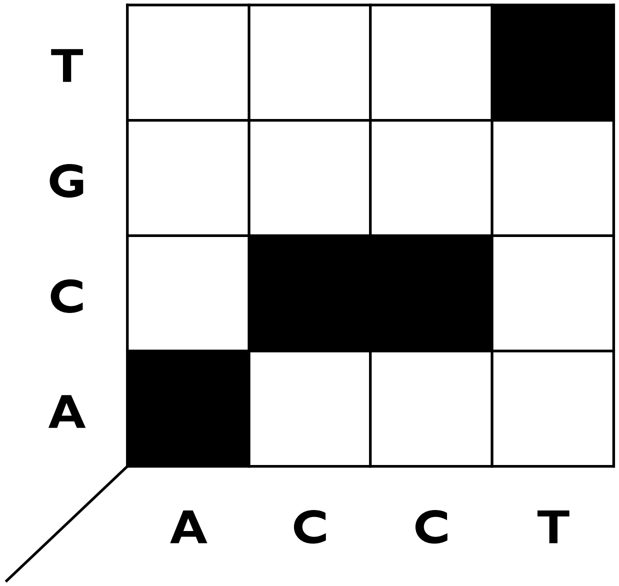
CCGGTAGGATATTAAACGGGGTGAGGAGCGTTGGCATAGCA

CCGCTAGGCTATTAAAACCCCGGAGGAG....GGCTGAGCA

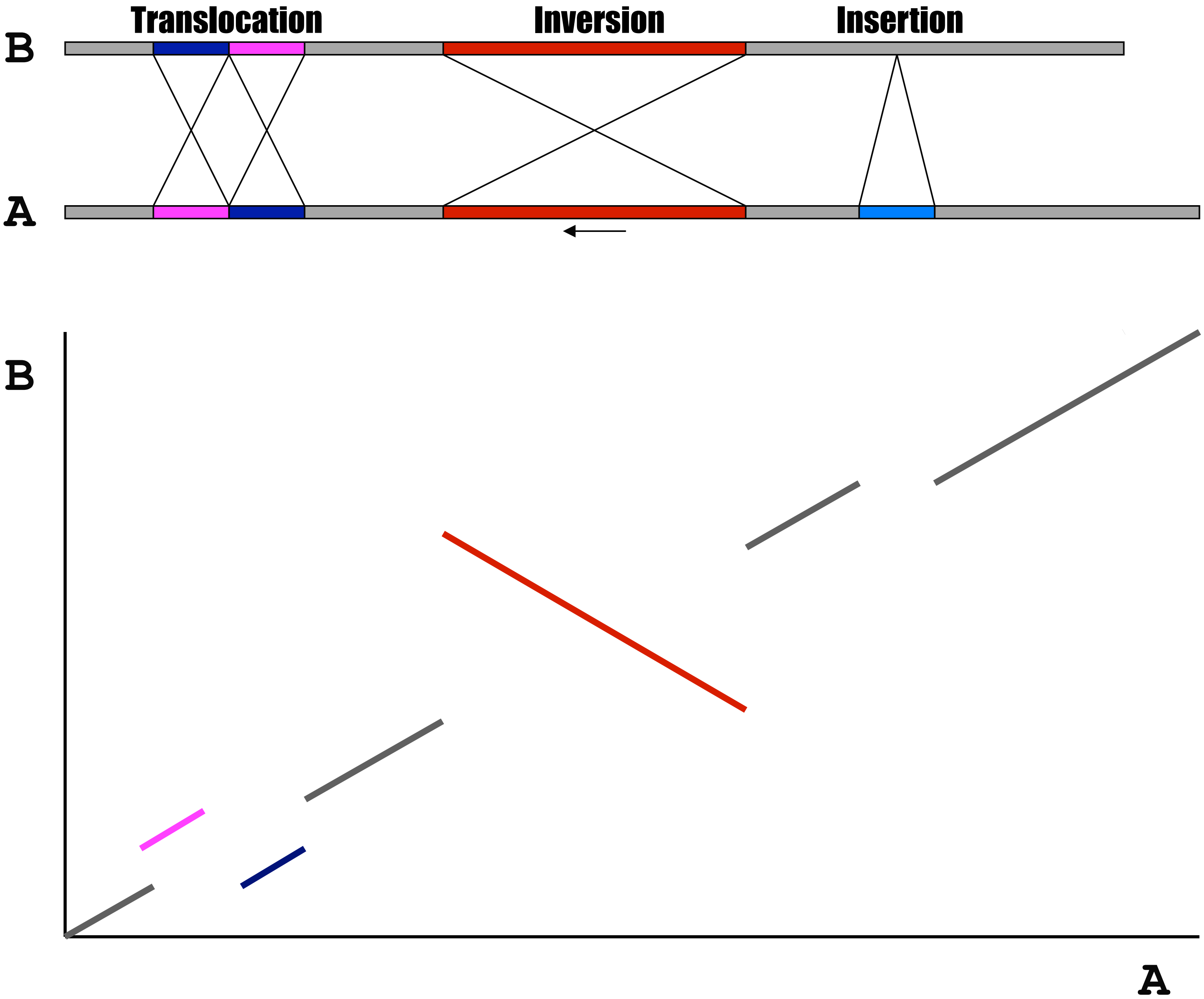
WGA visualization

- How can we visualize *whole* genome alignments?

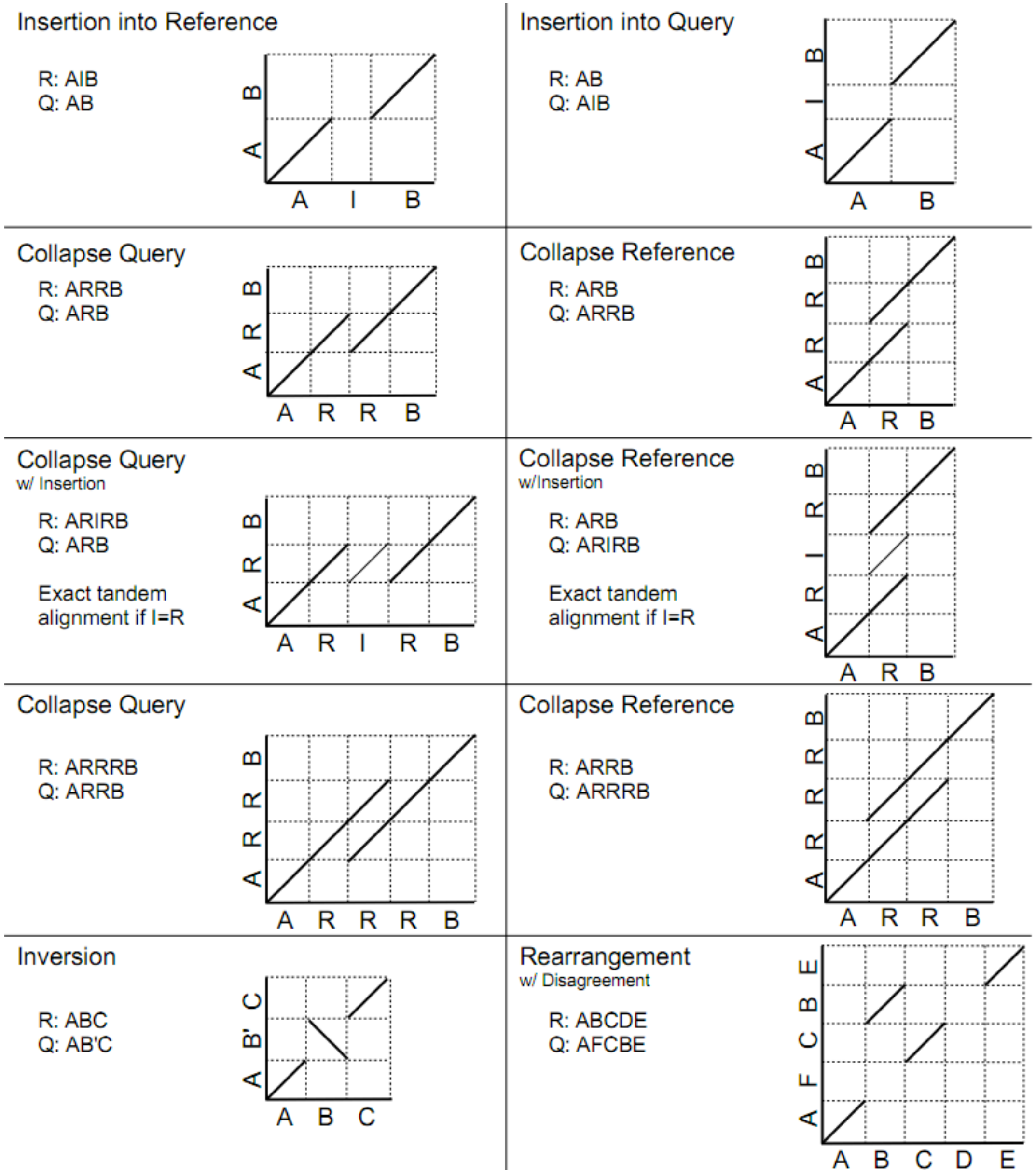
- With an alignment dot plot
  - *N* x *M* matrix
    - Let *i* = position in genome *A*
    - Let *j* = position in genome *B*
    - Fill cell *(i,j)* if A_i shows similarity to B_j



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


Translocation
Inversion
Insertion
B
A
B
A

SV Types



Insertion into Reference

R: AIB
Q: AB

Insertion into Query

R: AB
Q: AIB

Collapse Query

R: ARRB
Q: ARB

Collapse Reference

R: ARB
Q: ARRB

Collapse Query
w/ Insertion

R: ARIRB
Q: ARB

Exact tandem alignment if I=R

Collapse Reference
w/Insertion

R: ARB
Q: ARIRB

Exact tandem alignment if I=R

Collapse Query

R: ARRRB
Q: ARRB

Collapse Reference

R: ARRB
Q: ARRRB

Inversion

R: ABC
Q: AB'C

Rearrangement
w/ Disagreement

R: ABCDE
Q: AFCBE

- 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

Seed-and-extend with MUMmer

How can quickly align two genomes?

1. 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)

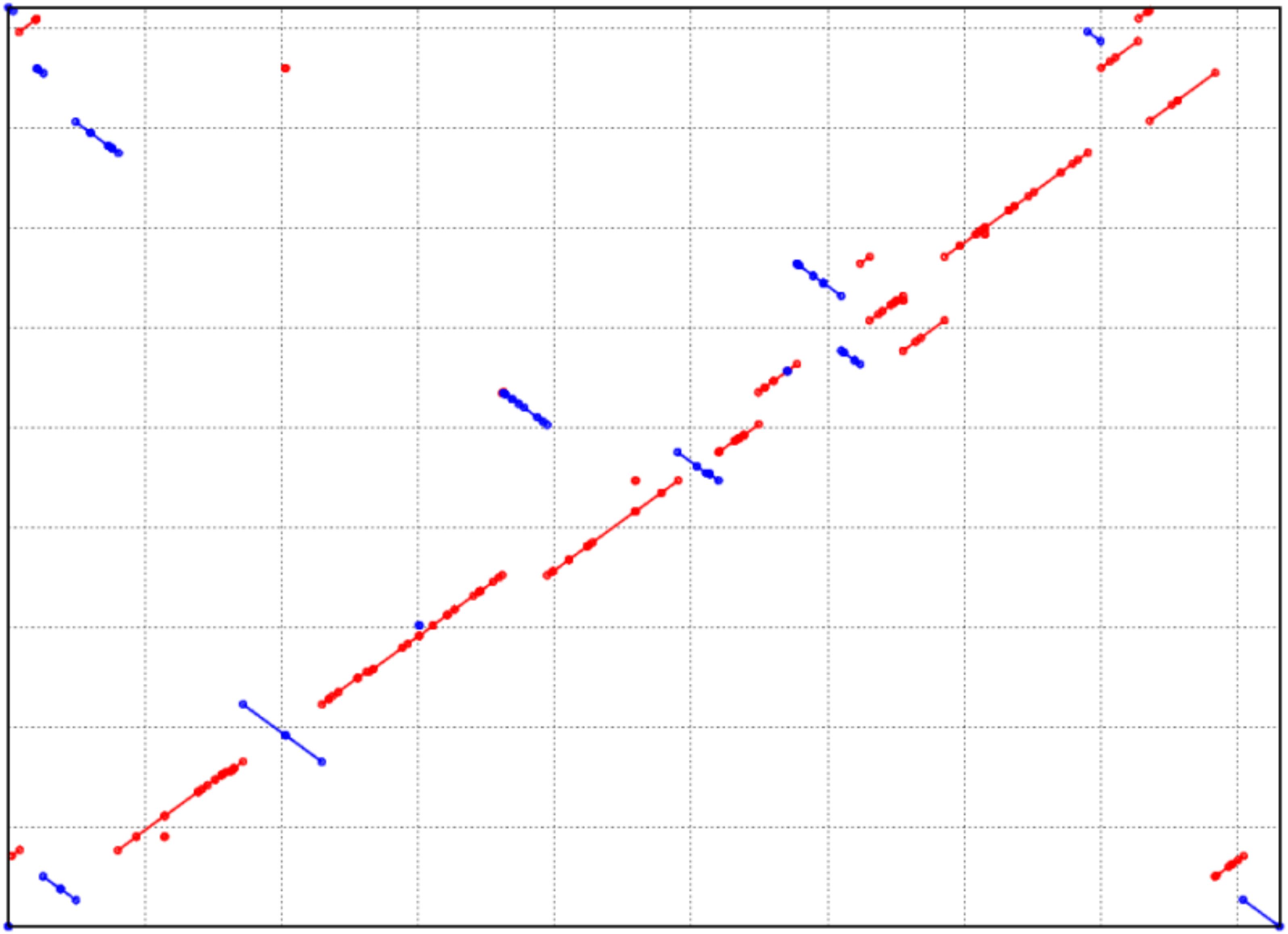
2. Cluster MUMs
   - using size, gap and distance parameters

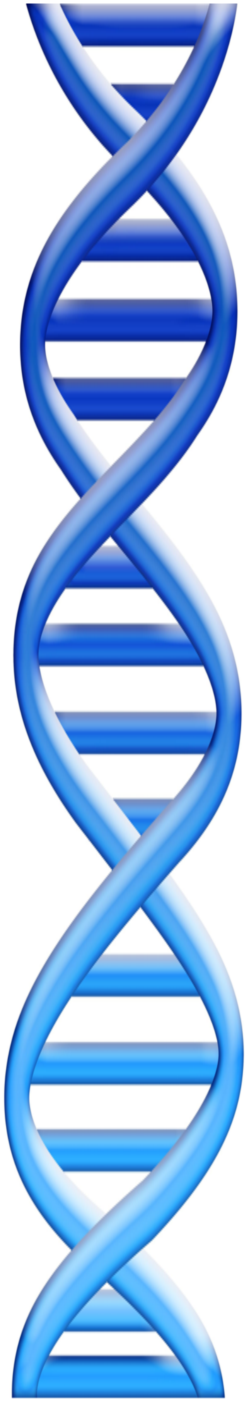
3. Extend clusters
   - using modified Smith-Waterman algorithm

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
```

See manual at http://mummer.sourceforge.net/manual



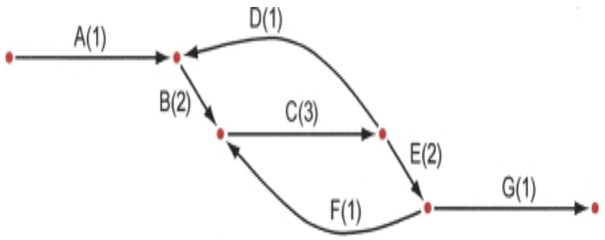


Outline

1. 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

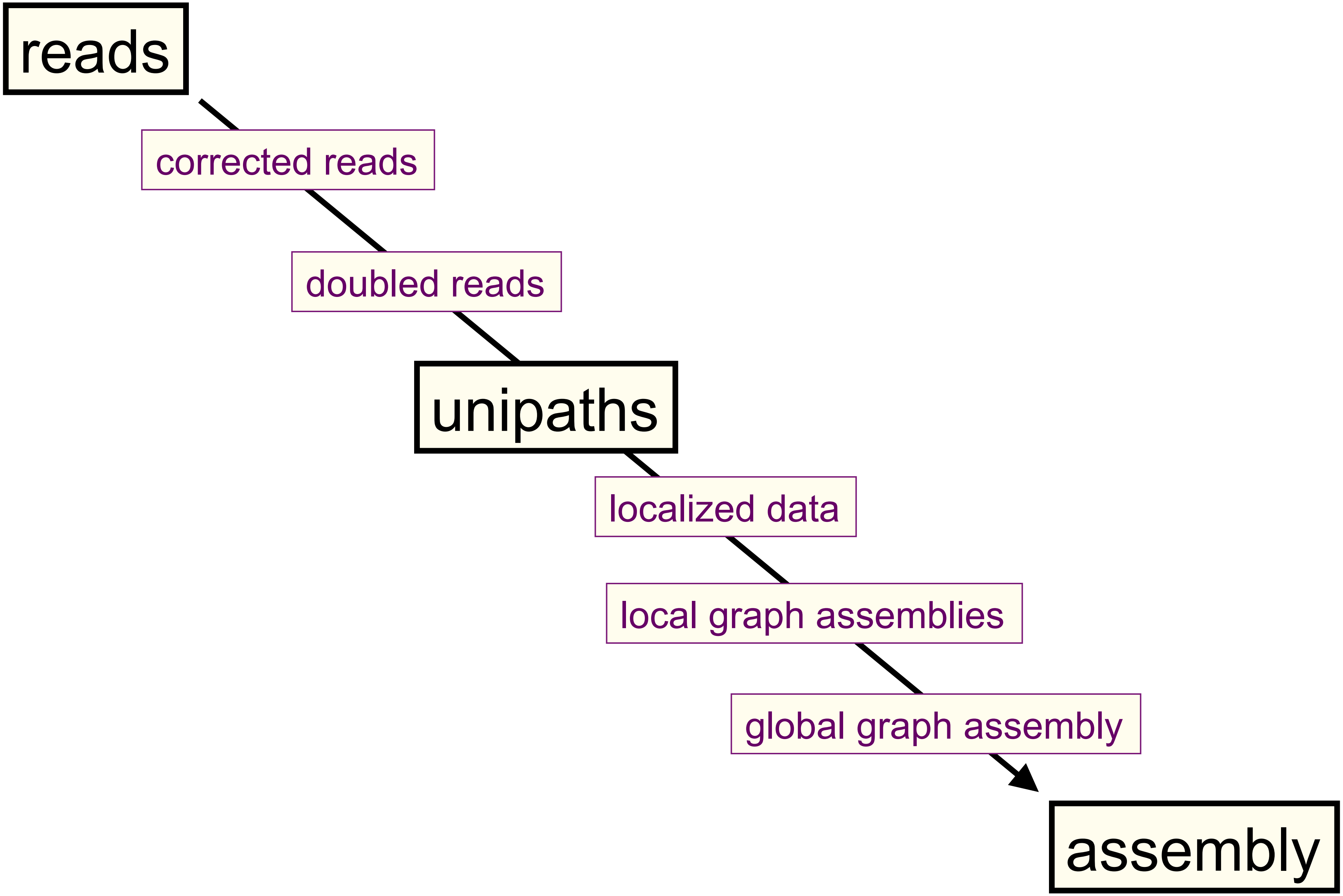


Genome assembly with ALLPATHS-LG

Iain MacCallum

BROAD INSTITUTE

How ALLPATHS-LG works



reads

corrected reads

doubled reads

unipaths

localized data

local graph assemblies

global graph assembly

assembly

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.

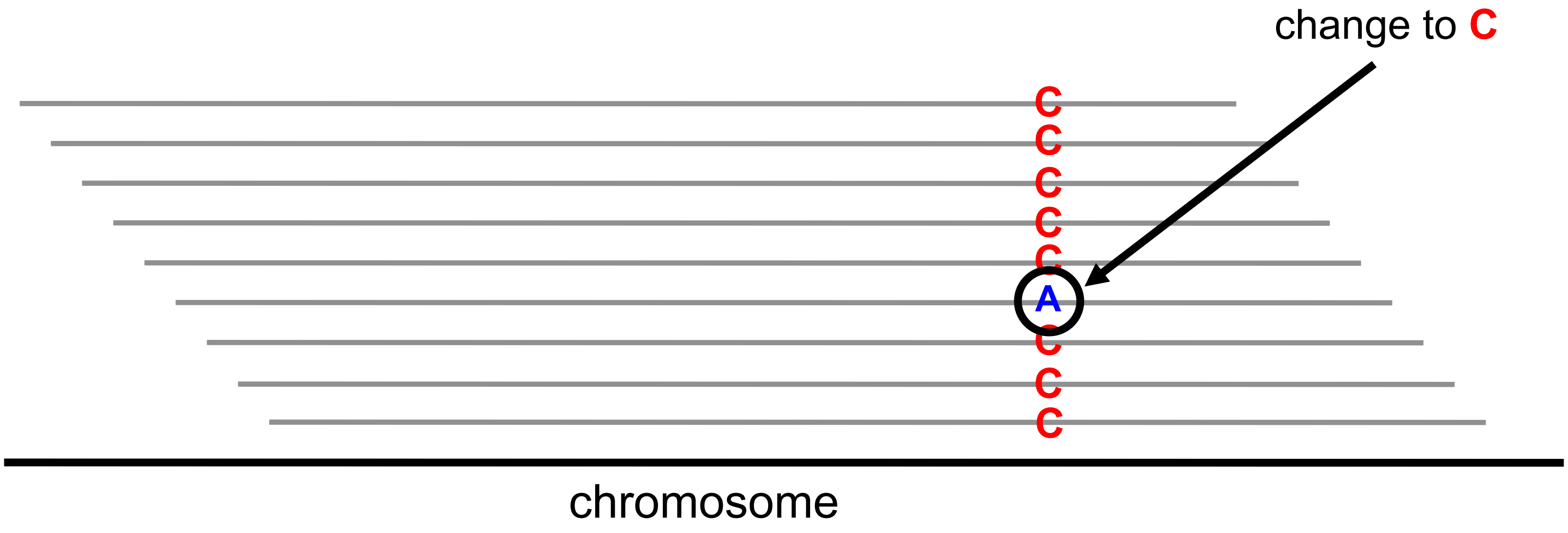
**For best results. Normally not used for small genomes.
However essential to assemble long repeats or duplications.

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

All: protocols are either available, or in progress.

Error correction

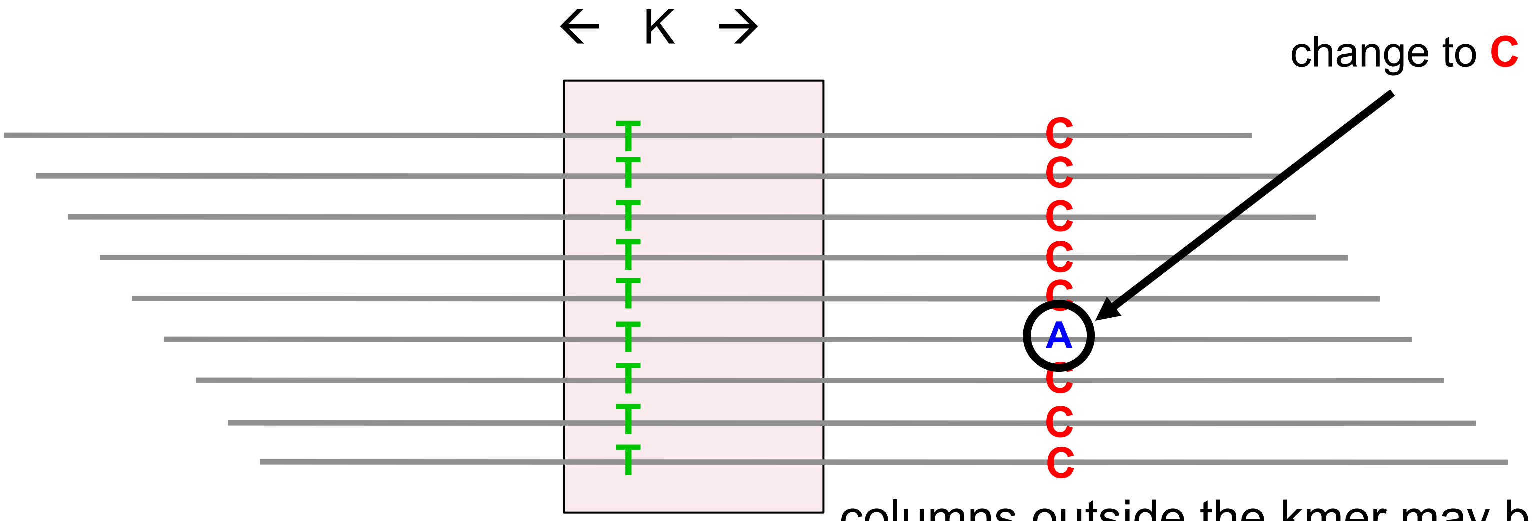
Given a crystal ball, we could stack reads on the chromosomes they came from (with homologous chromosomes separate), then let each column 'vote':



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)



change to C

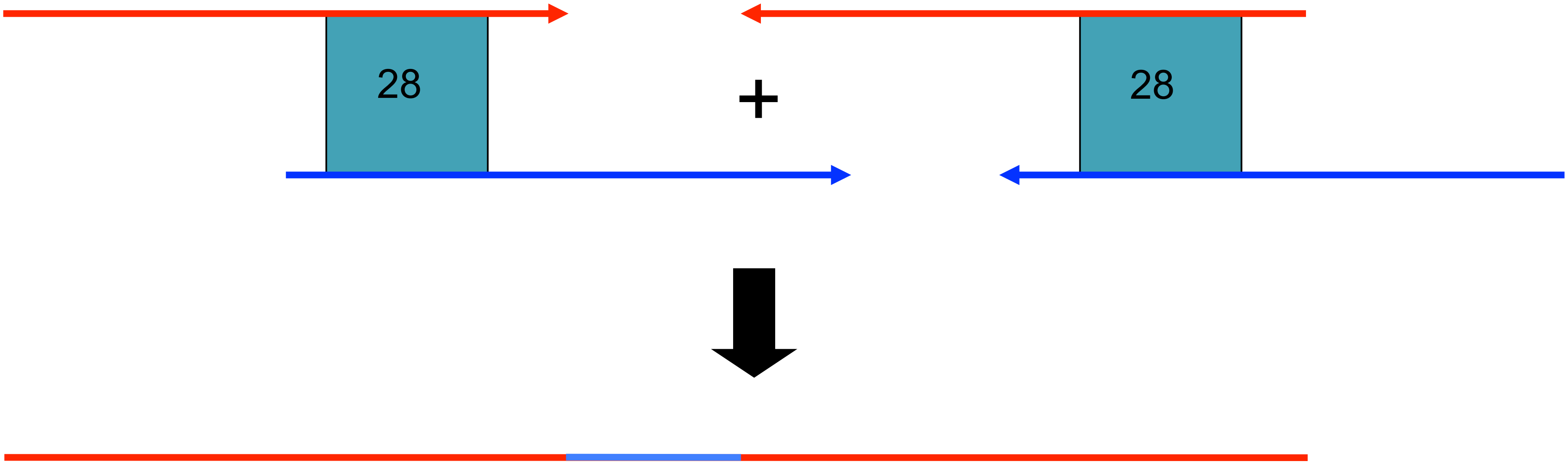
columns outside the kmer may be mixed

columns inside the kmer are homogeneous

Two calls at Q20 or better are enough to protect a base

Read doubling

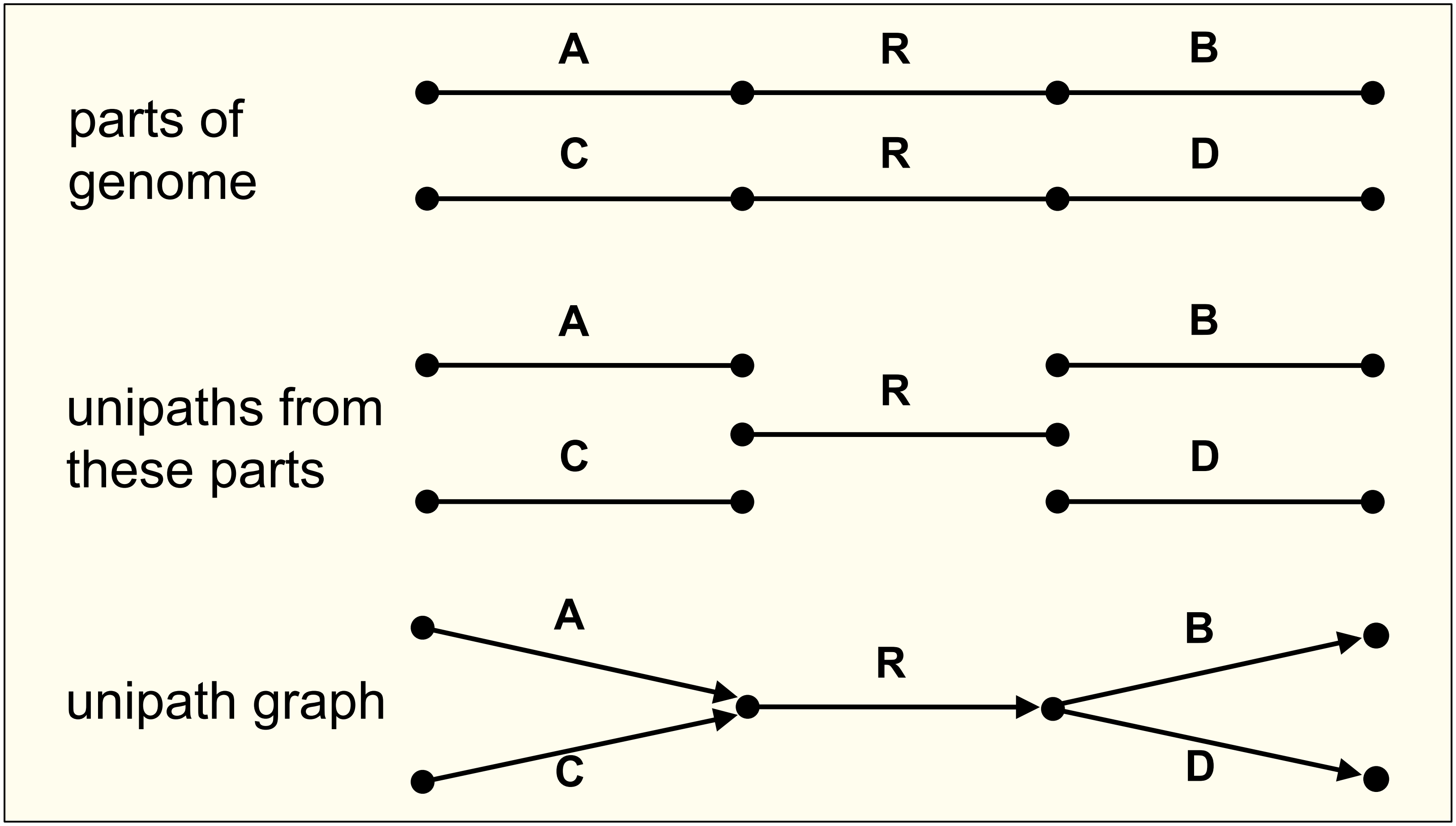
To close a read pair (red), we require the existence of another read pair (blue), overlapping perfectly like this:



More than one closure allowed (but rare).

Unipaths

Unipath: unbranched part of genome – squeeze together perfect repeats of size ≥ K



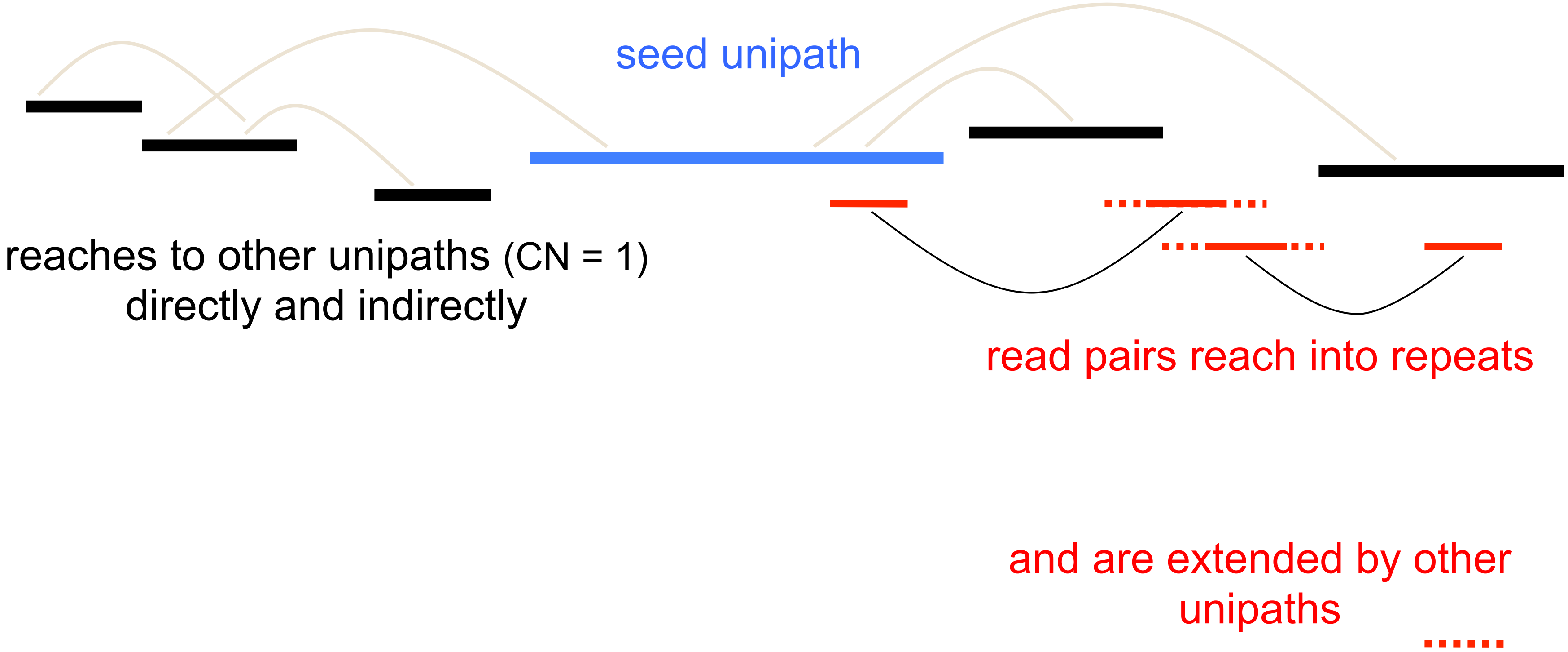
Adjacent unipaths overlap by K-1 bases

Localization

I. Find ‘seed’ unipaths, evenly spaced across genome

(ideally long, of copy number CN = 1)

II. Form neighborhood around each seed



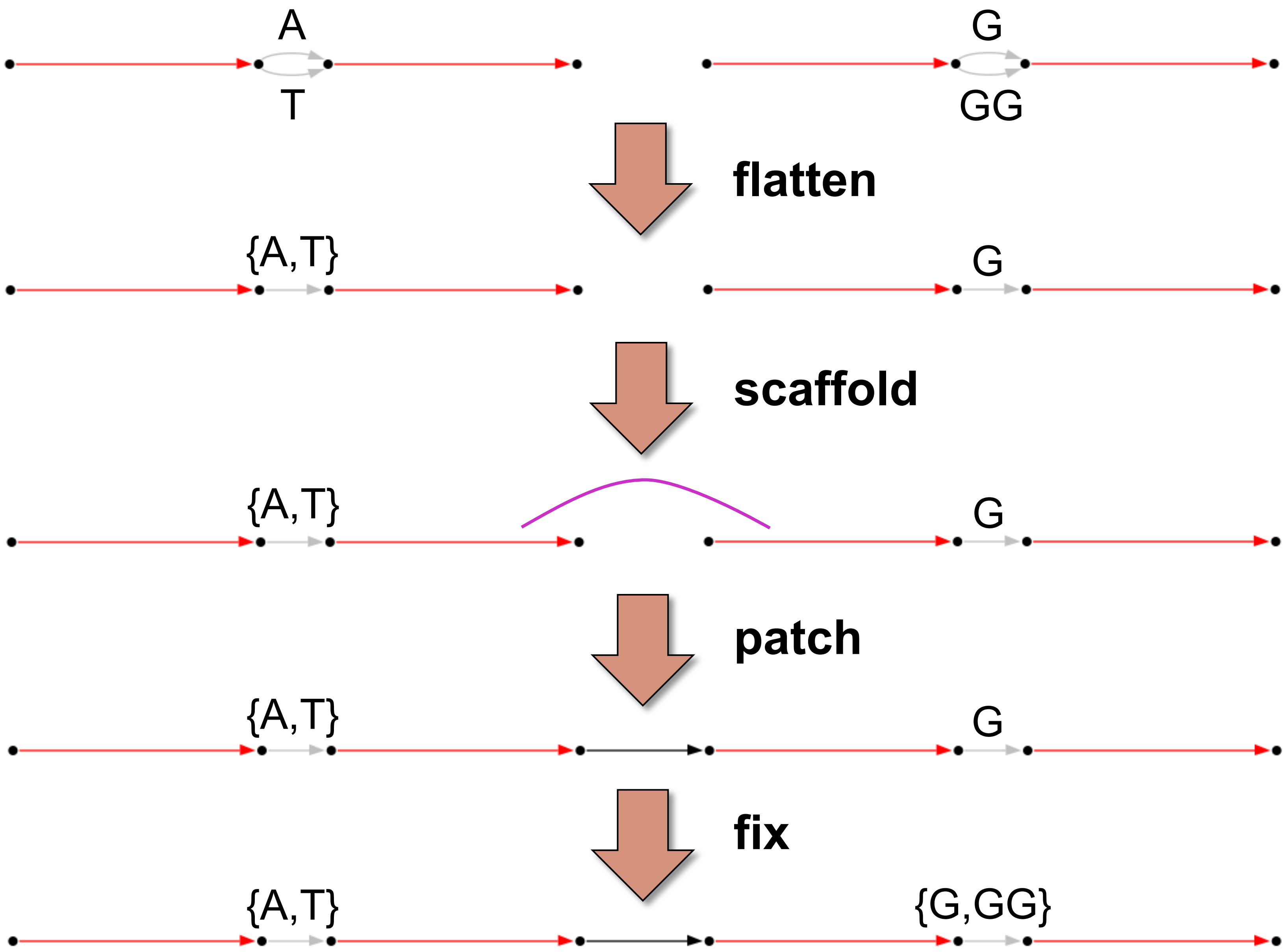
seed unipath

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

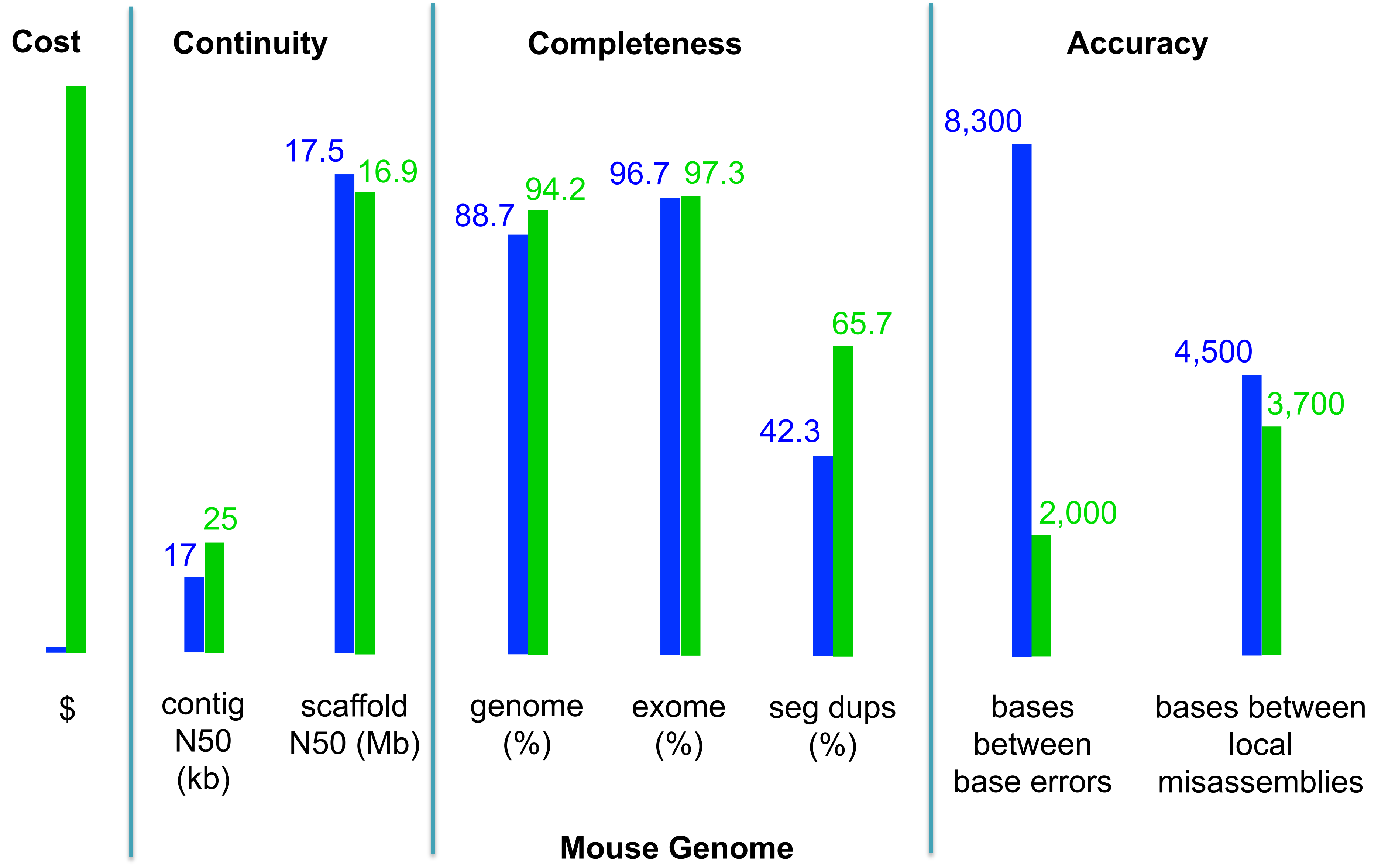
read pairs reach into repeats

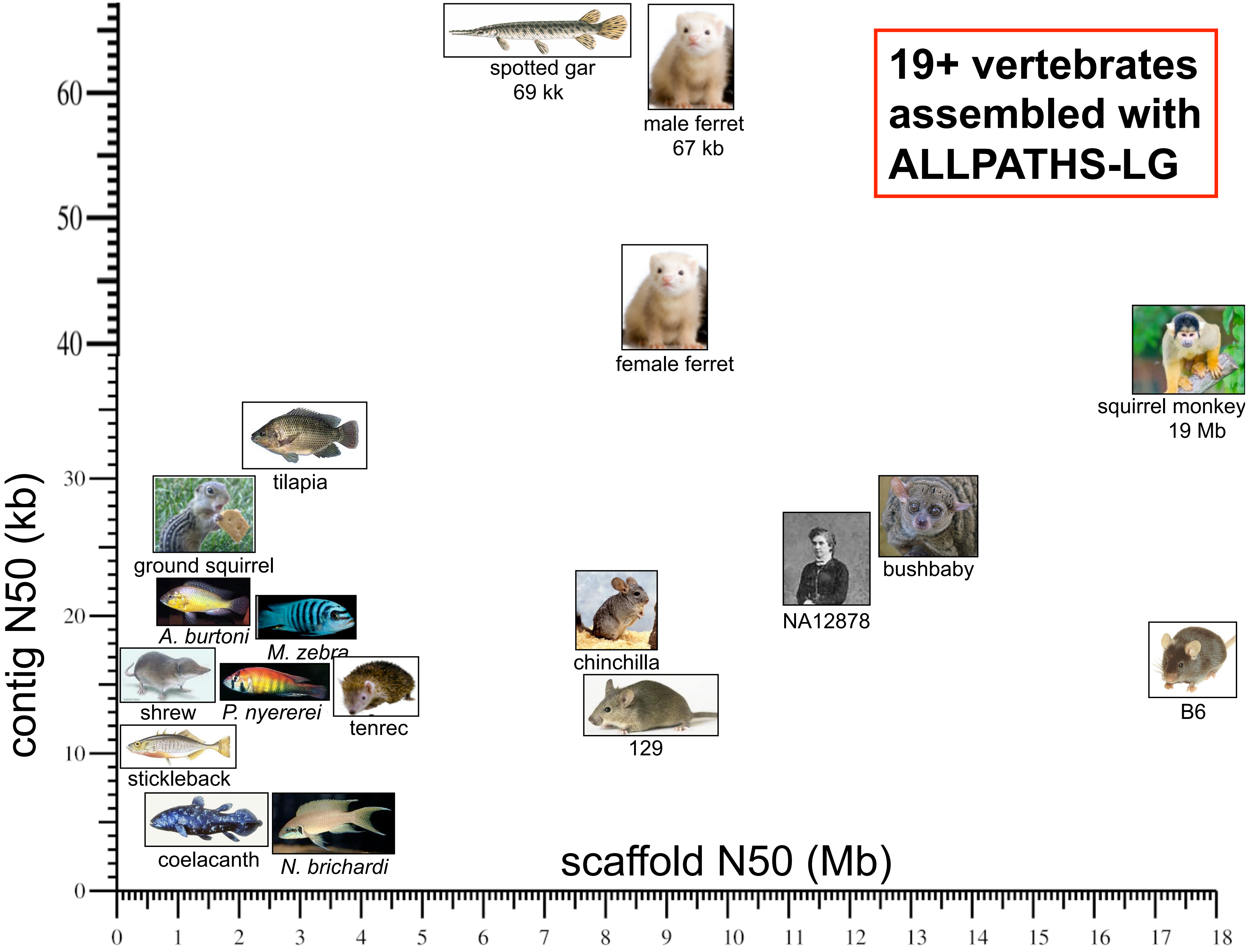
and are extended by other unipaths

Create assembly from global assembly graph

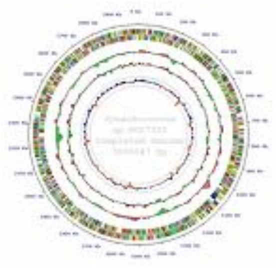


Large genome recipe: ALLPATHS-LG vs capillary





19+ vertebrates assembled with ALLPATHS-LG

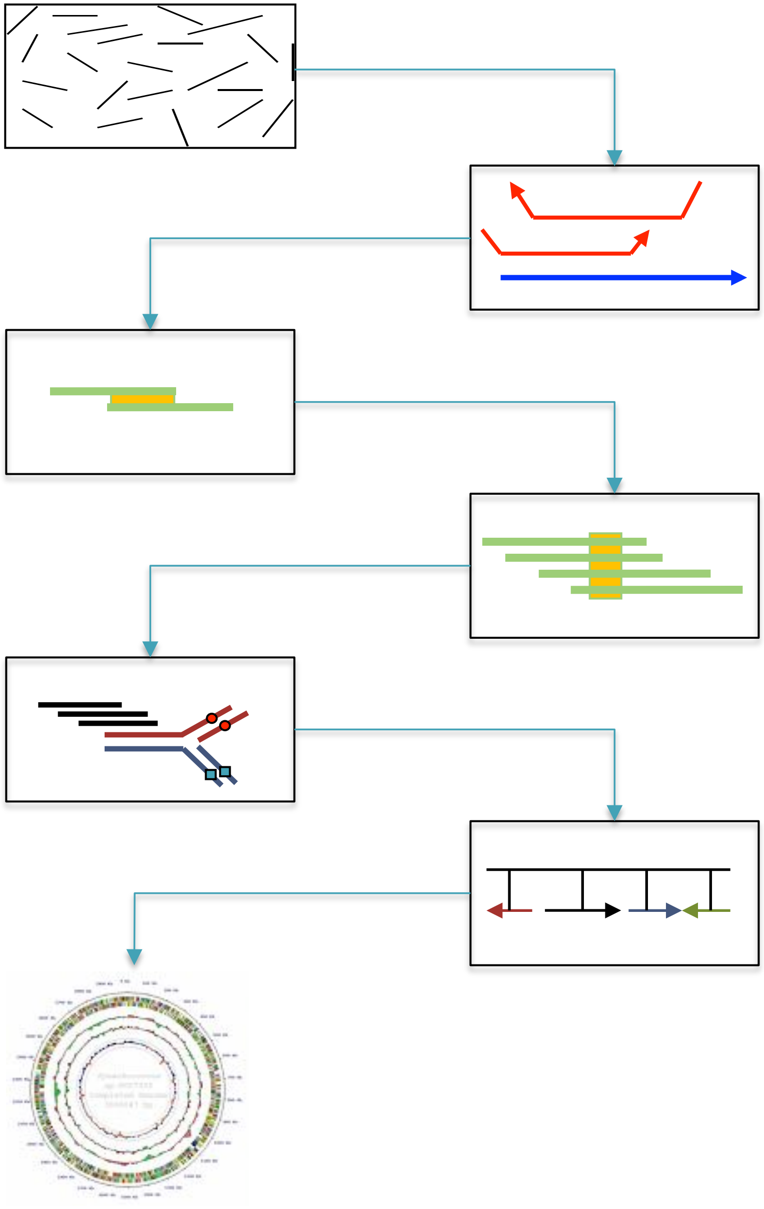


Genome assembly with the Celera Assembler

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

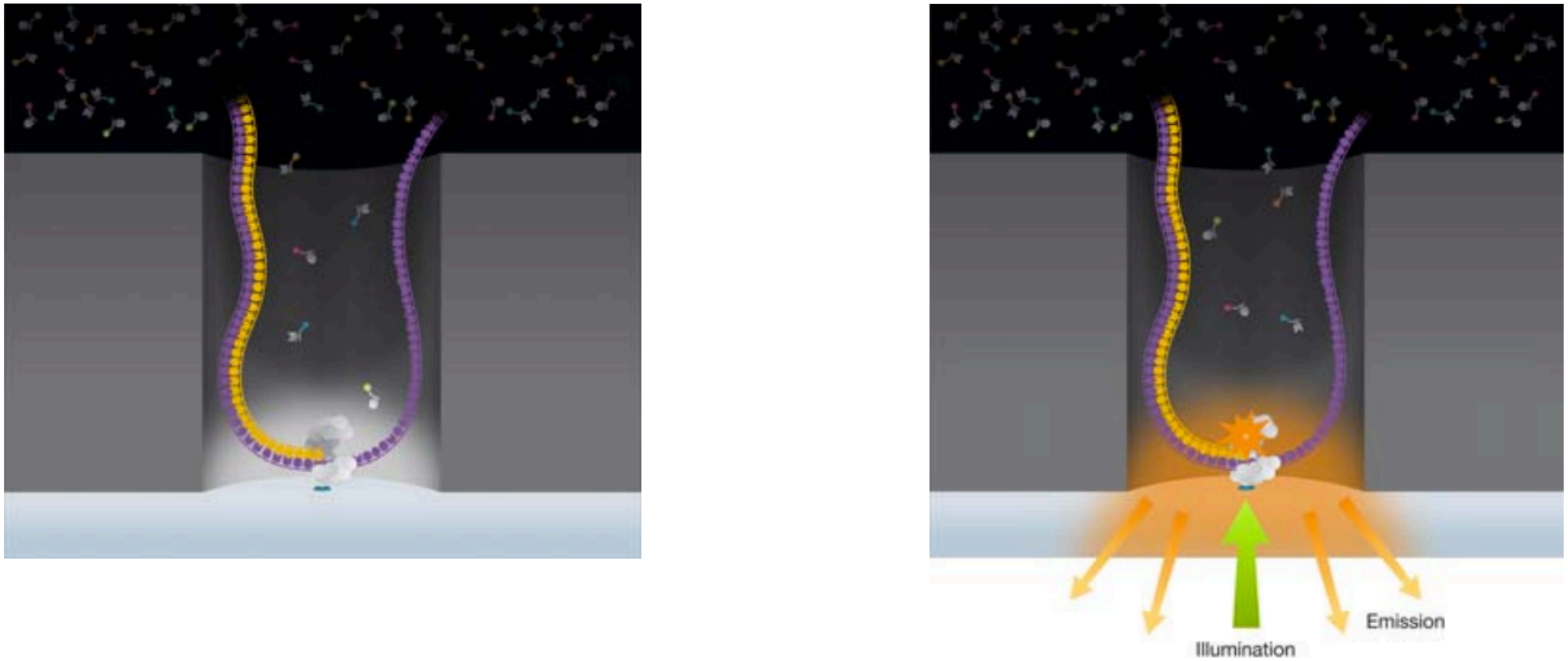
High throughput (60Gbp/day)
High accuracy (~99%)
Short reads (~100bp)

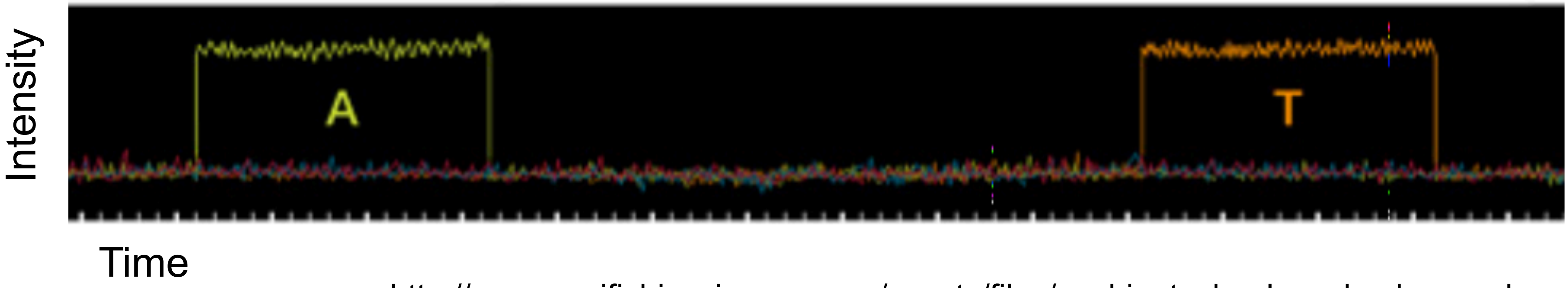
Pacific Biosciences
SMRT Sequencing

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

SMRT Sequencing

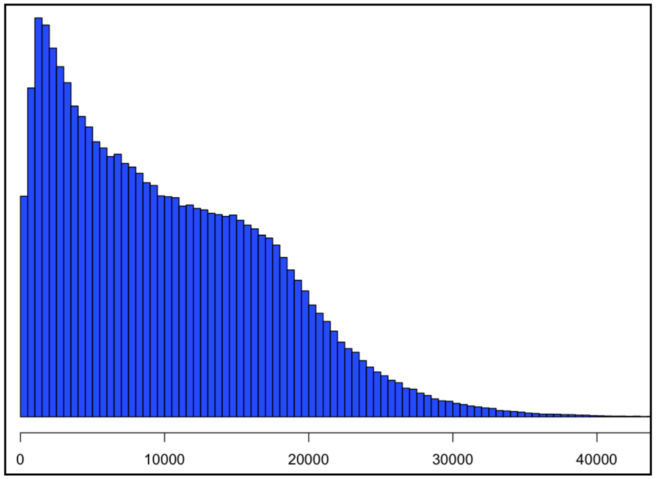
Imaging of fluorescently phospholinked labeled nucleotides as they are incorporated by a polymerase anchored to a Zero-Mode Waveguide (ZMW).





http://www.pacificbiosciences.com/assets/files/pacbio_technology_backgrounder.pdf

SMRT Sequencing Data

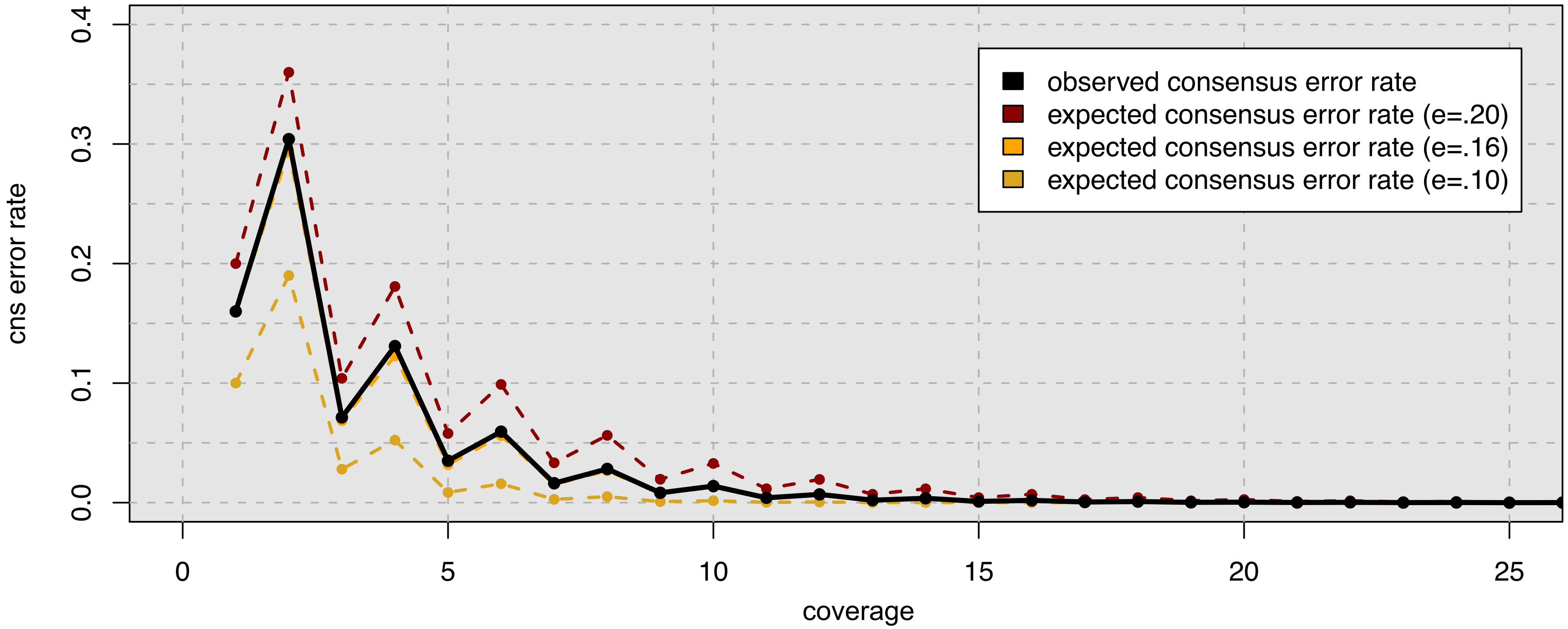


Match	83.7%
Insertions	11.5%
Deletions	3.4%
Mismatch	1.4%

```
TTGTAAGCAGTTGAAAACTATGTGTGGATTTAGAATAAAGAACATGAAAG
||||||||||||||||||||||||| ||||||| |||||||||||| |||
TTGTAAGCAGTTGAAAACTATGTGT-GATTTAG-ATAAAGAACATGGAAG

ATTATAAA-CAGTTGATCCATT-AGAAGA-AAACGCAAAAGGCGGCTAGG
| |||||| ||||||||||||| |||| | |||||| |||||| ||||||
A-TATAAATCAGTTGATCCATTAAGAA-AGAAACGC-AAAGGC-GCTAGG

CAACCTTGAATGTAATCGCACTTGAAGAACAAGATTTTATTCCGCGCCCG
| |||||| |||| ||  ||||||||||||||||||||||||||||||||
C-ACCTTG-ATGT-AT--CACTTGAAGAACAAGATTTTATTCCGCGCCCG

TAACGAATCAAGATTCTGAAAACACAT-ATAACAACCTCCAAAA-CACAA
| ||||||| |||||||||||||| || ||    |||||||||| |||||
T-ACGAATC-AGATTCTGAAAACA-ATGAT----ACCTCCAAAAGCACAA

-AGGAGGGGAAAGGGGGGAATATCT-ATAAAAGATTACAAATTAGA-TGA
 ||||||   ||     |||||||| || |||||||||||||| || |||
GAGGAGG---AA-----GAATATCTGAT-AAAGATTACAAATT-GAGTGA

ACT-AATTCACAATA-AATAACACTTTTA-ACAGAATTGAT-GGAA-GTT
||| ||||||||| | ||||||||||||| ||| ||||||| |||| |||
ACTAAATTCACAA-ATAATAACACTTTTAGACAAAATTGATGGGAAGGTT

TCGGAGAGATCCAAAACAATGGGC-ATCGCCTTTGA-GTTAC-AATCAAA
|| ||||||||| ||||||| ||| |||| |||||| ||||| |||||||
TC-GAGAGATCC-AAACAAT-GGCGATCG-CTTTGACGTTACAAATCAAA

ATCCAGTGGAAAATATAATTTATGCAATCCAGGAACTTATTCACAATTAG
||||||| |||||||||  |||||| ||||| ||||||||||||||||||
ATCCAGT-GAAAATATA--TTATGC-ATCCA-GAACTTATTCACAATTAG
```

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

Consensus Accuracy and Coverage



Coverage can overcome random errors

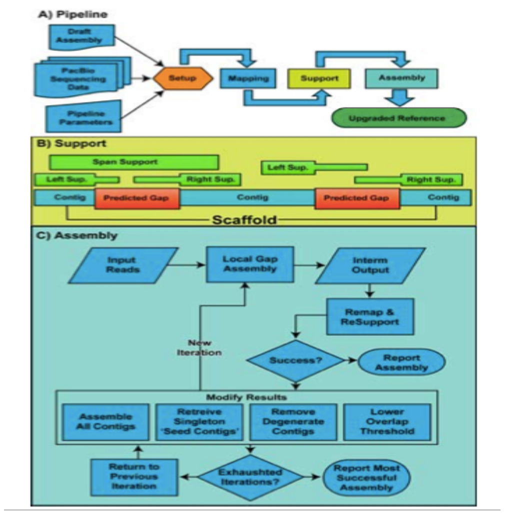
- Dashed: error model from binomial sampling
- Solid: observed accuracy

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

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

PacBio Assembly Algorithms

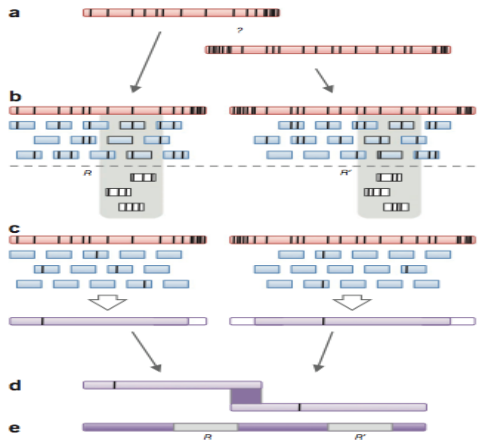
PBJelly



Gap Filling and Assembly Upgrade

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

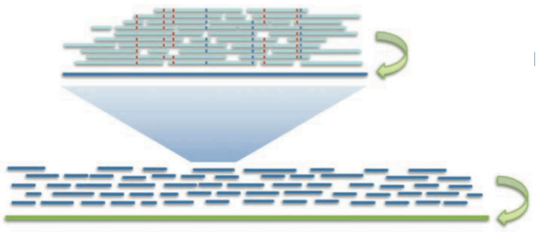
PacBioToCA & ECTools



Hybrid/PB-only Error Correction

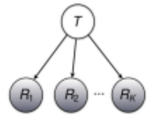
Koren, Schatz, *et al* (2012)
Nature Biotechnology. 30:693–700

HGAP & Quiver



$$\Pr(\mathbf{R} \mid T)$$

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



Quiver Performance Results
Comparison to Reference Genome
(*M. ruber* ; 3.1 MB ; SMRT® Cells)

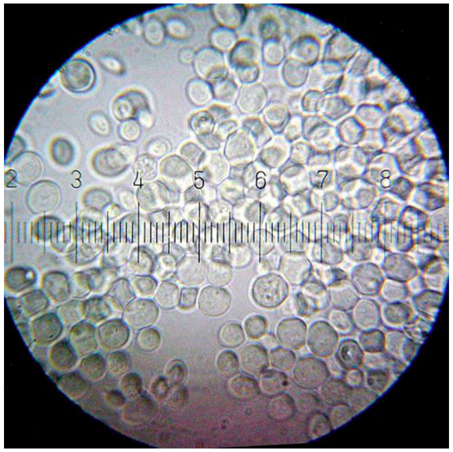
	Initial Assembly	Quiver Consensus
QV	43.4	54.5
Accuracy	99.99540%	99.99964%
Differences	141	11

PB-only Correction & Polishing

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

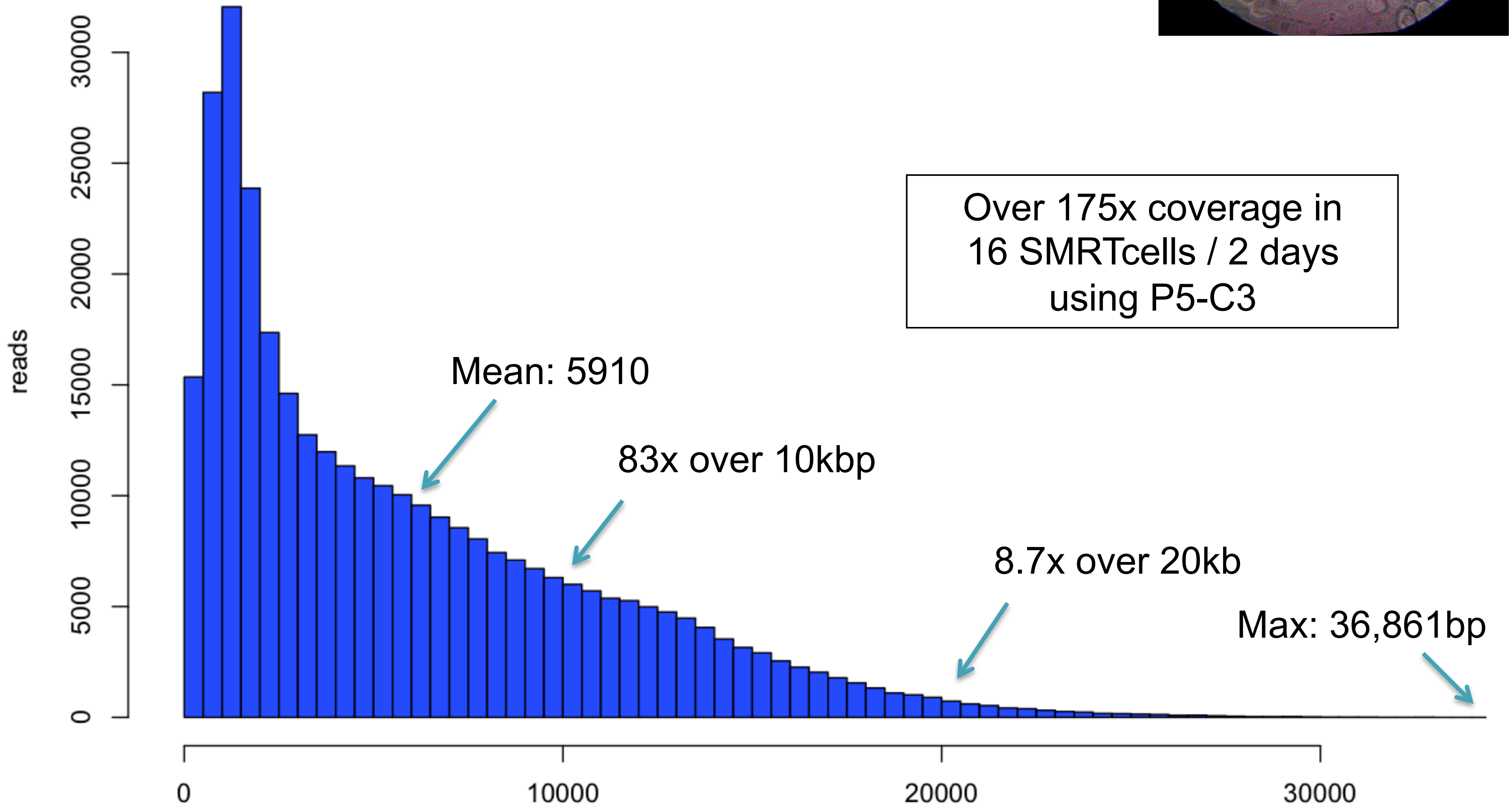
< 5x   PacBio Coverage   > 50x

S. 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



Over 175x coverage in 16 SMRTcells / 2 days using P5-C3

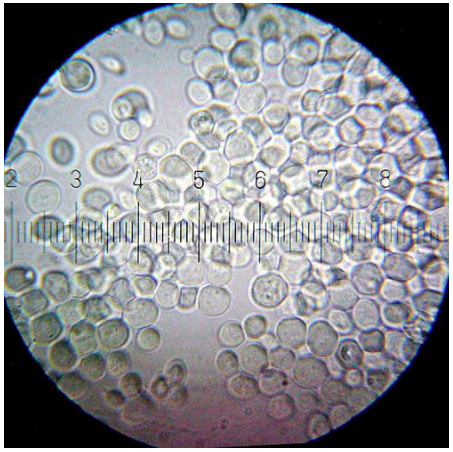
Mean: 5910

83x over 10kbp

8.7x over 20kb

Max: 36,861bp

S. cerevisiae W303

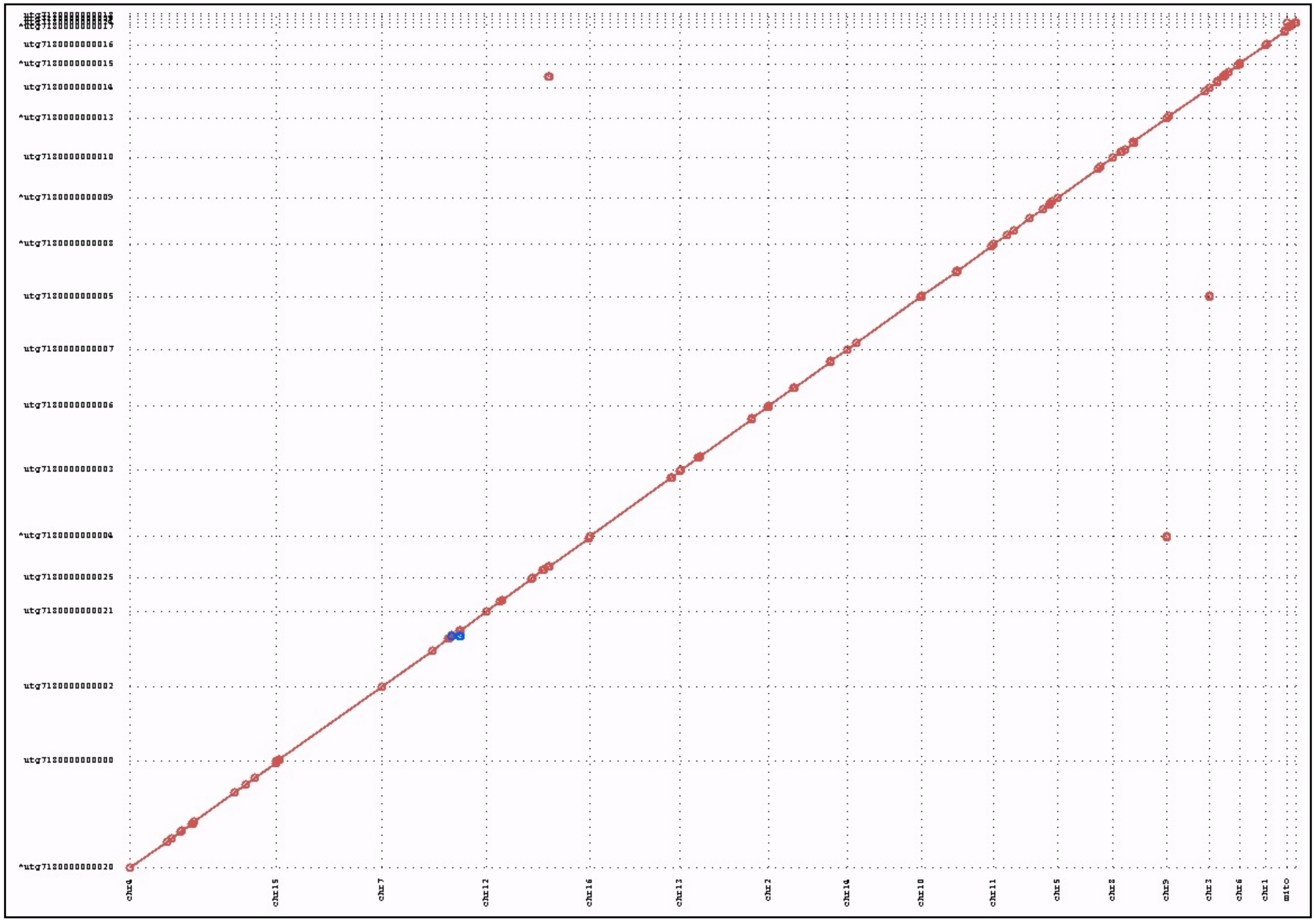


S288C Reference sequence

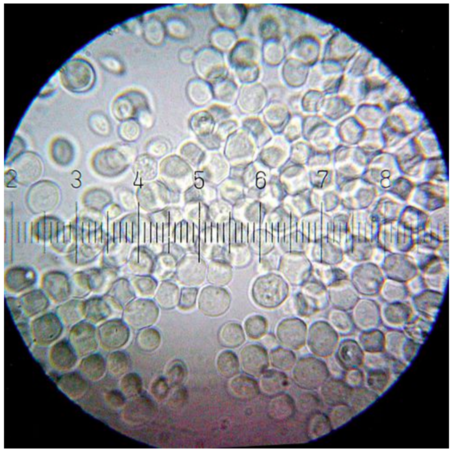
- 12.1Mbp; 16 chromo + mitochondria; N50: 924kbp

PacBio assembly using HGAP + Celera Assembler

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



S. cerevisiae W303

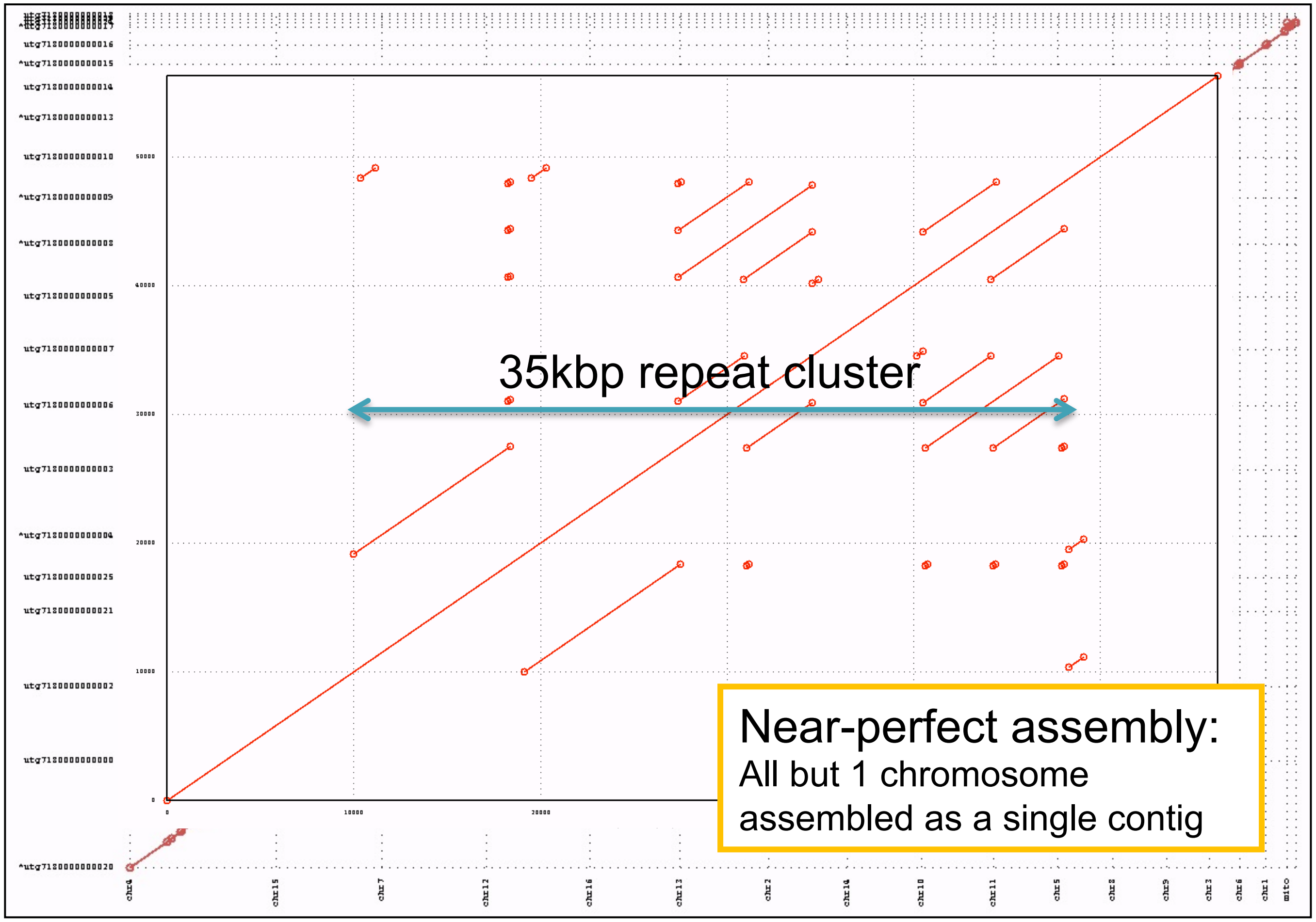


S288C Reference sequence

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

PacBio assembly using HGAP + Celera Assembler

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

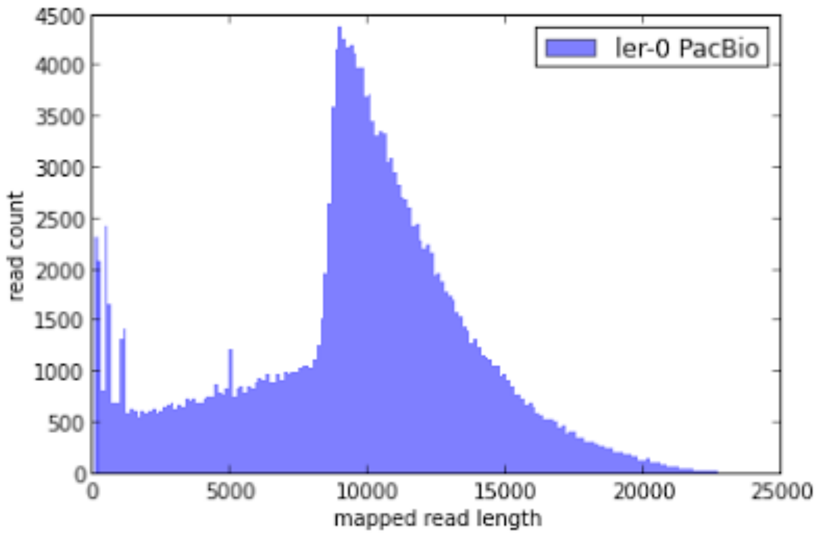


35kbp repeat cluster

Near-perfect assembly:
All but 1 chromosome
assembled as a single contig

A. thaliana Ler-0

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



A. thaliana Ler-0 sequenced at PacBio

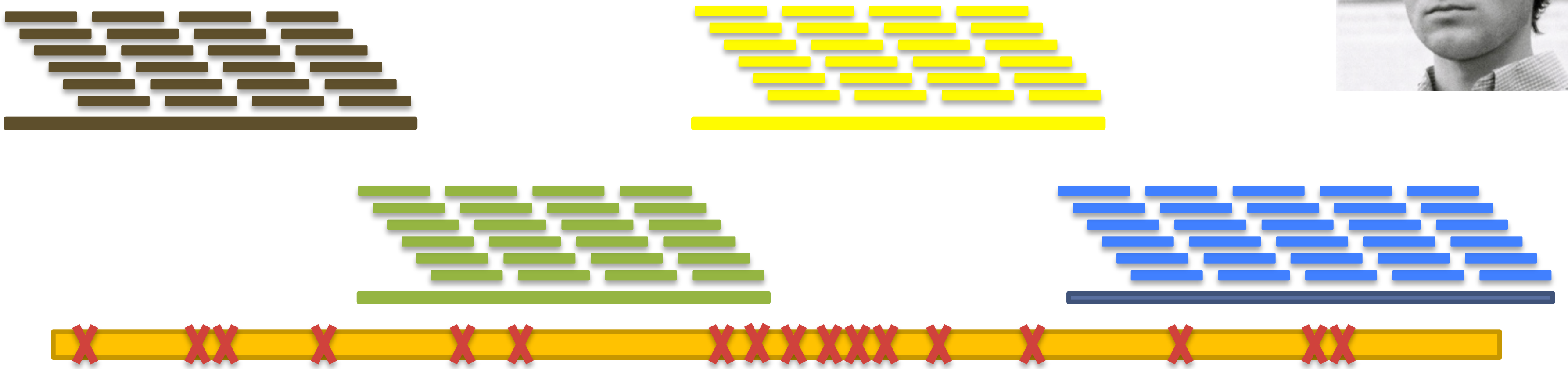
- Sequenced using the previous P4 enzyme and C2 chemistry
- Size selection using an 8 Kb to 50 Kb elution window on a BluePippin™ device from Sage Science
- Total coverage >119x

Genome size:	124.6 Mbp	Sum of Contig Lengths:	149.5Mb
Chromosome N50:	23.0 Mbp	N50 Contig Length:	8.4 Mb
Corrected coverage:	20x over 10kb	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

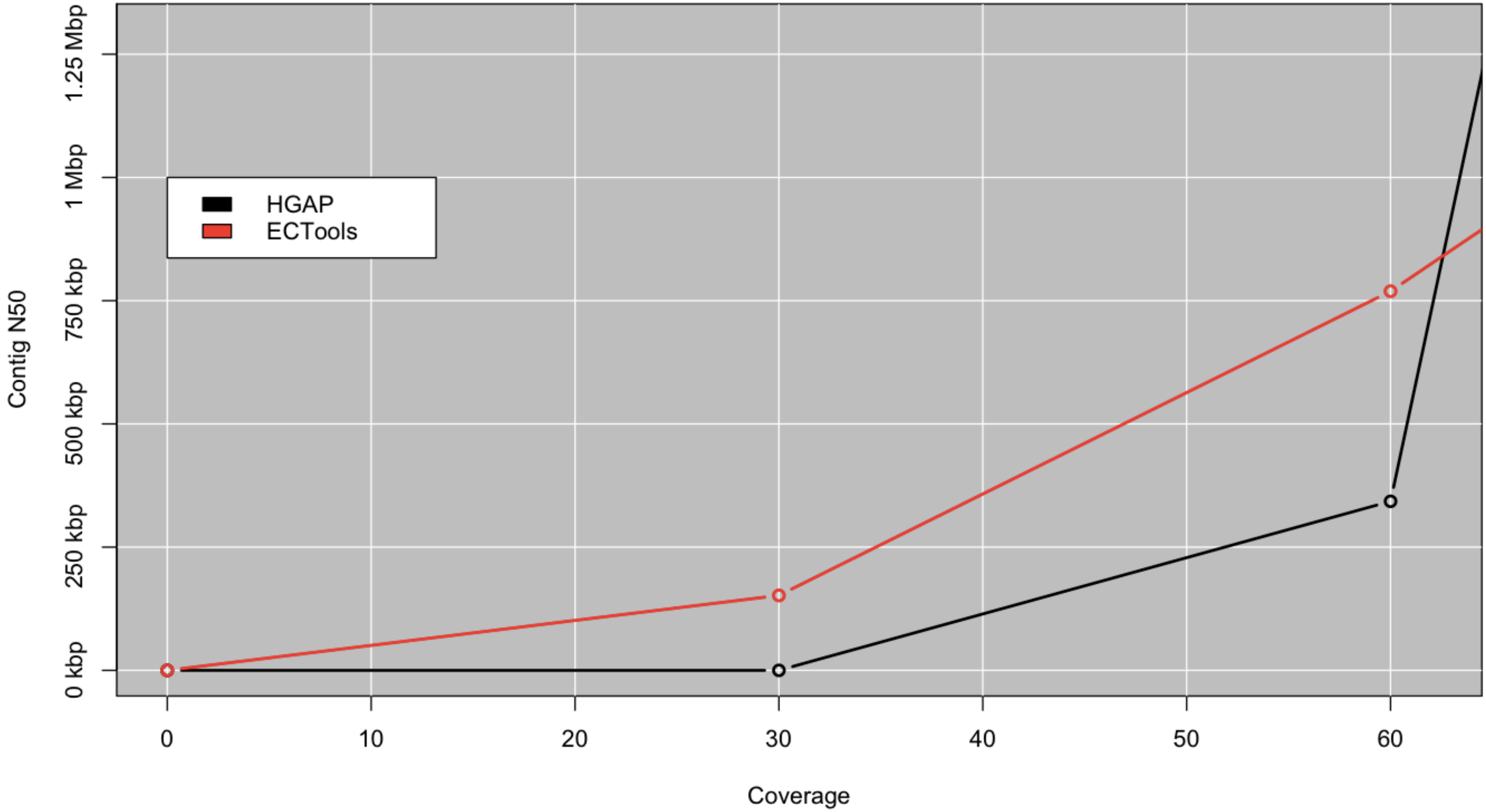
Can Help us overcome:

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

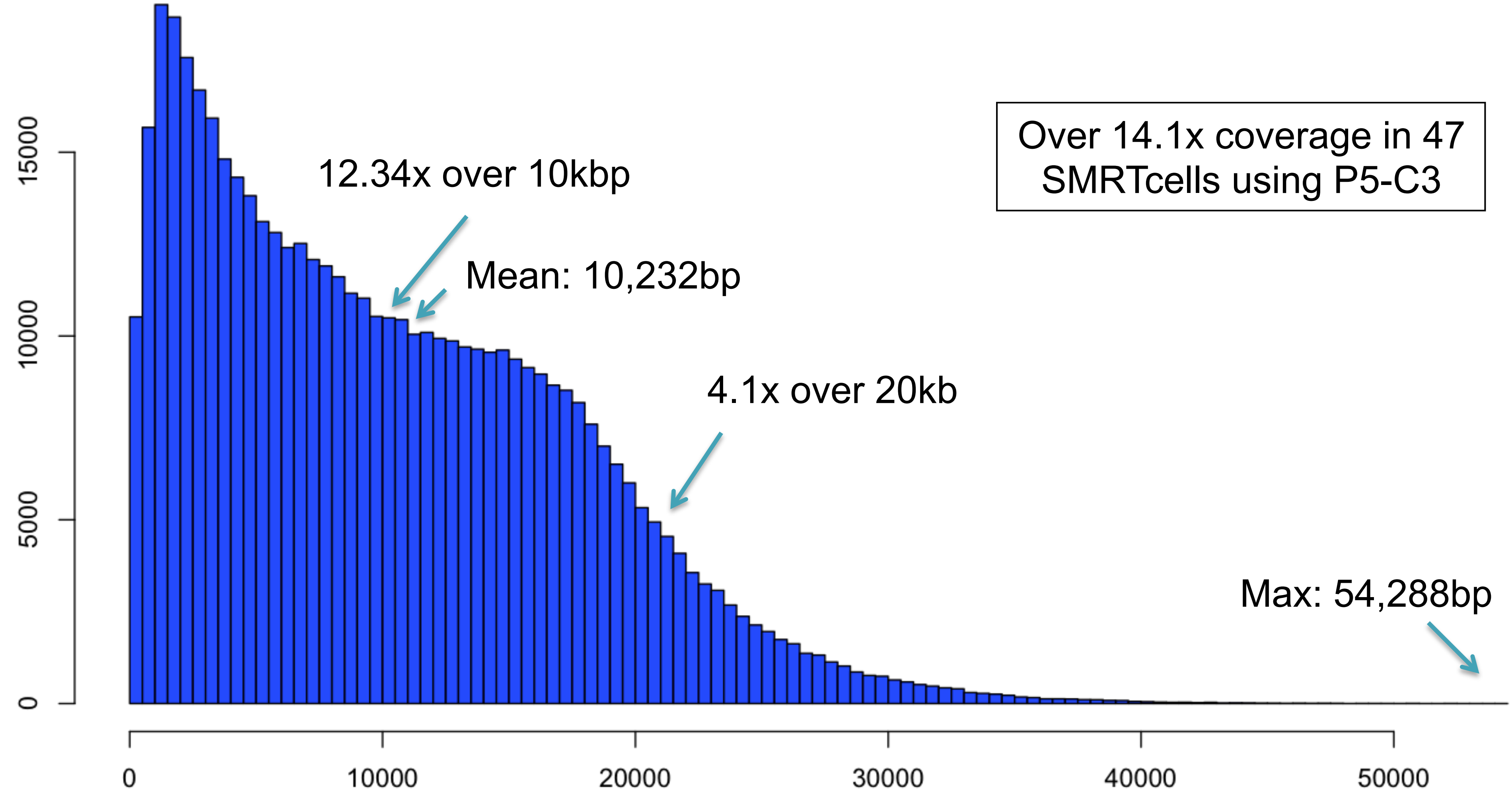
http://blog.pacificbiosciences.com/2013/08/new-data-release-arabidopsis-assembly.html



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



12.34x over 10kbp

Mean: 10,232bp

4.1x over 20kb

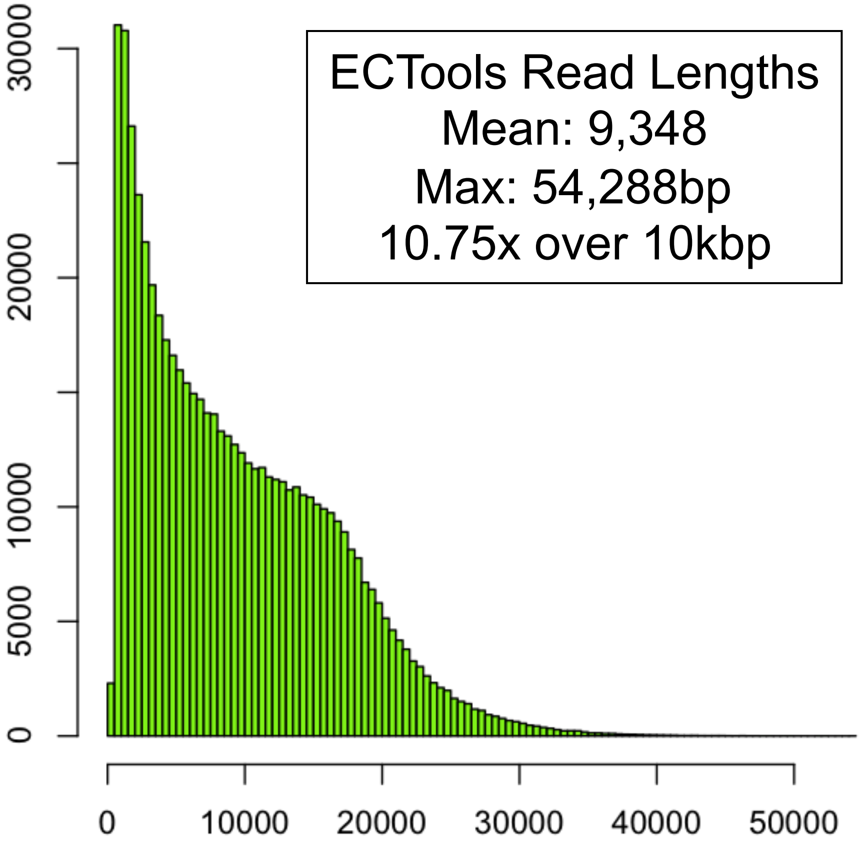
Max: 54,288bp

Over 14.1x coverage in 47 SMRTcells using P5-C3

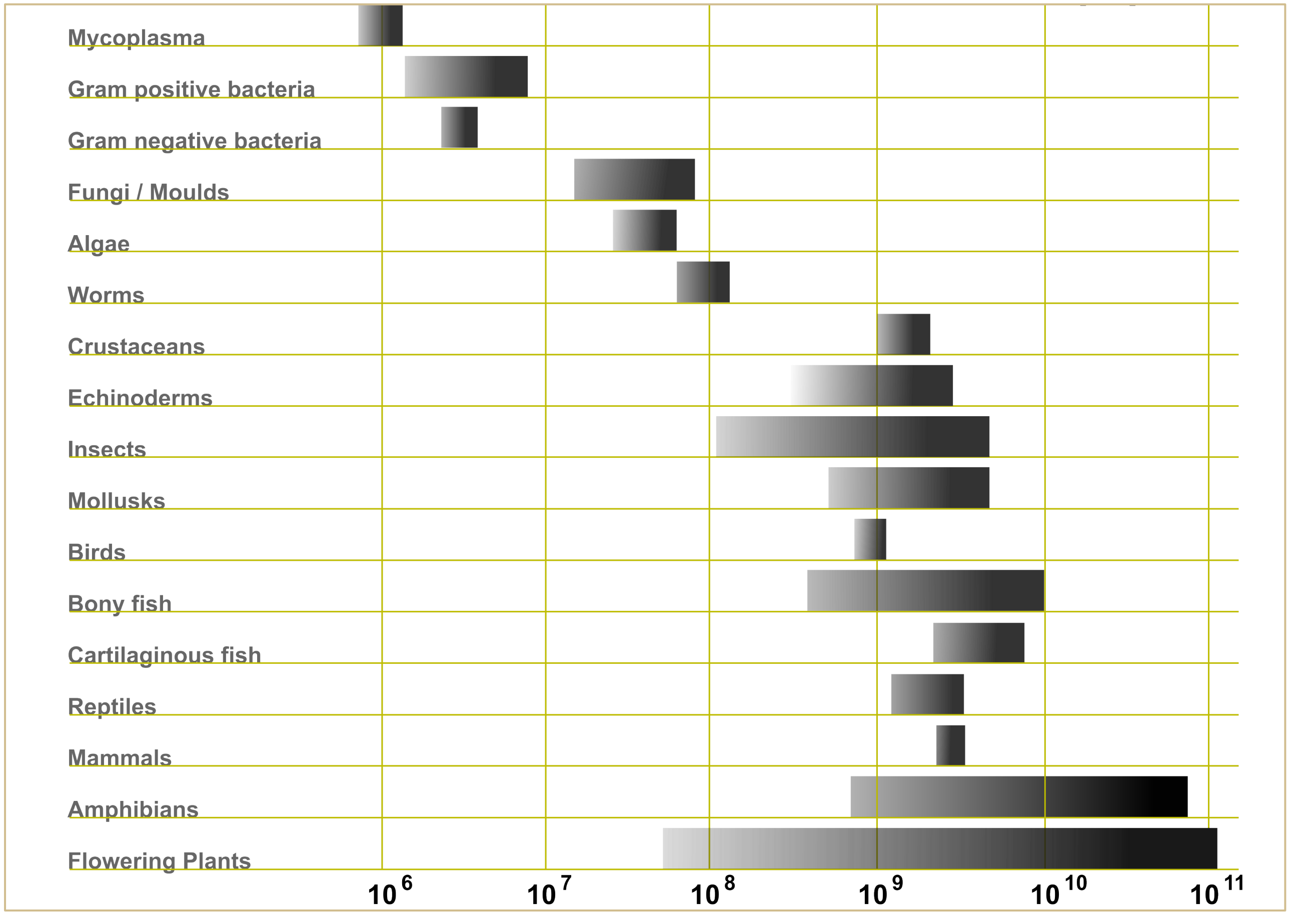
O. sativa pv Indica (IR64)

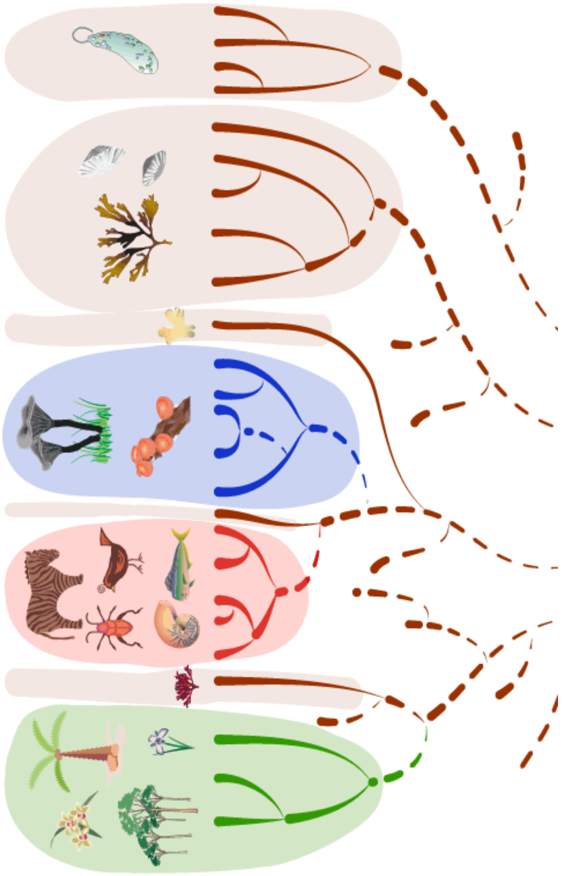
Genome size: ~370 Mb
Chromosome N50: ~29.7 Mbp

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



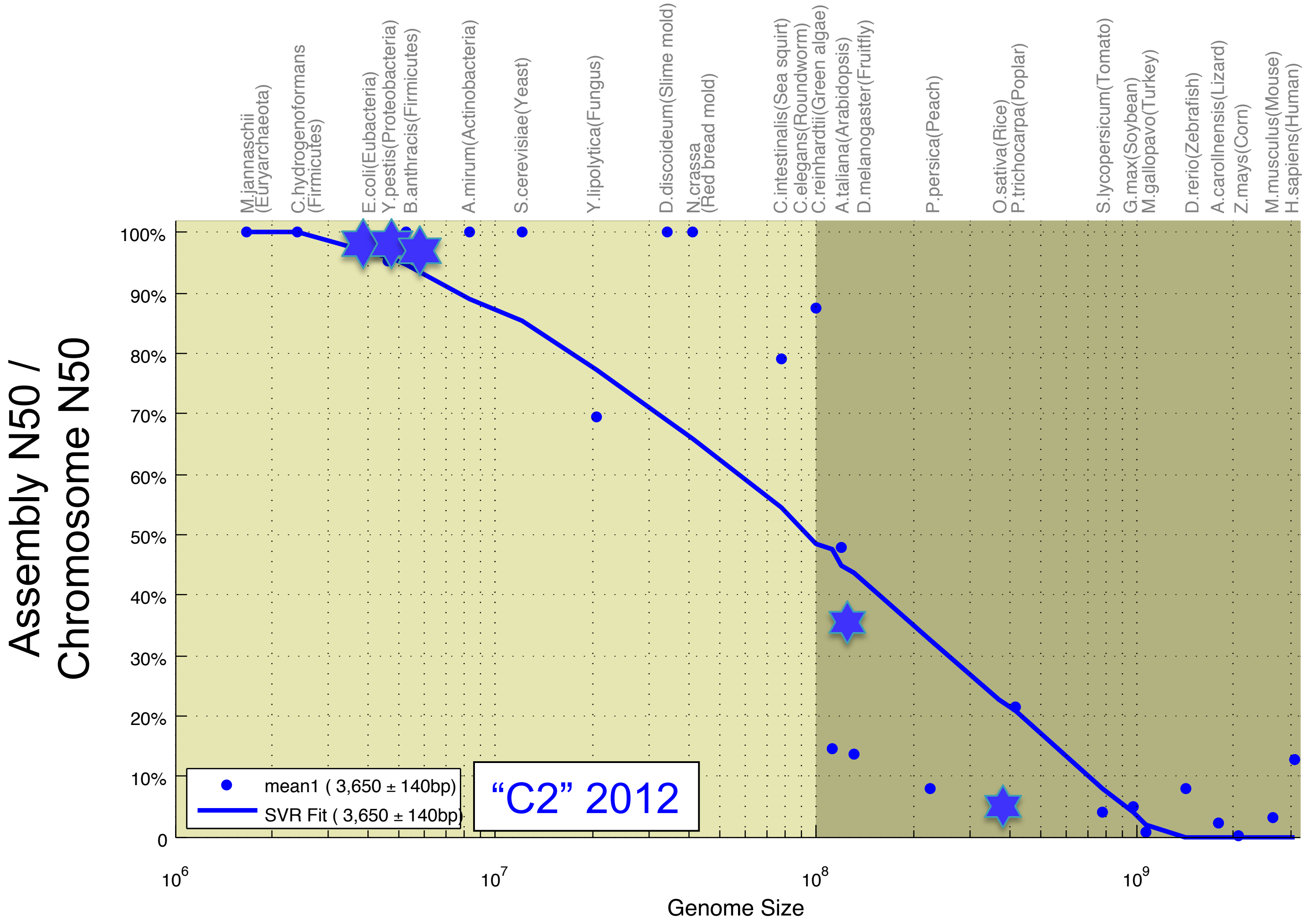
What should we expect from an assembly?





https://en.wikipedia.org/wiki/Genome_size

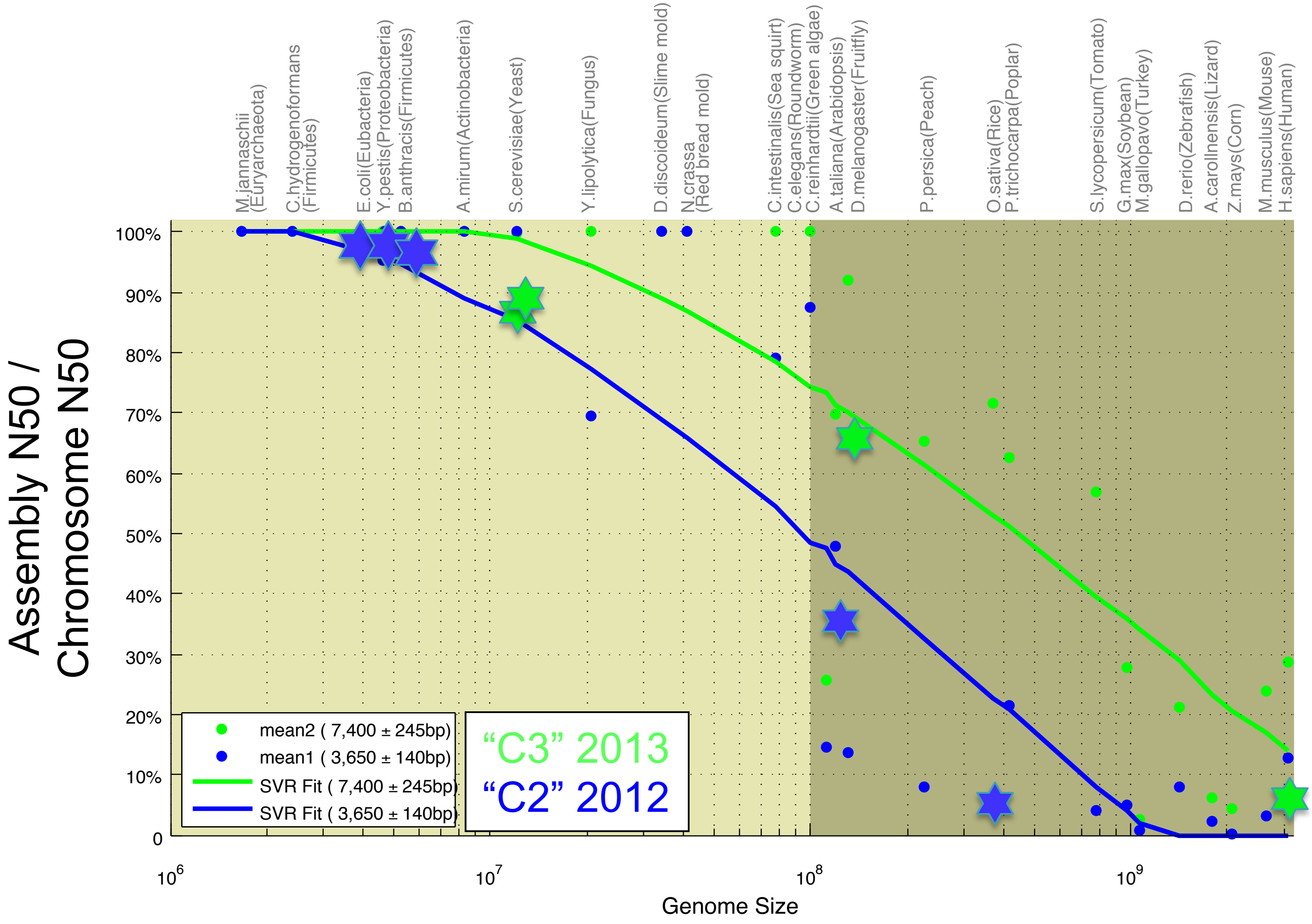
Assembly Complexity of Long Reads



Assembly complexity of long read sequencing

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

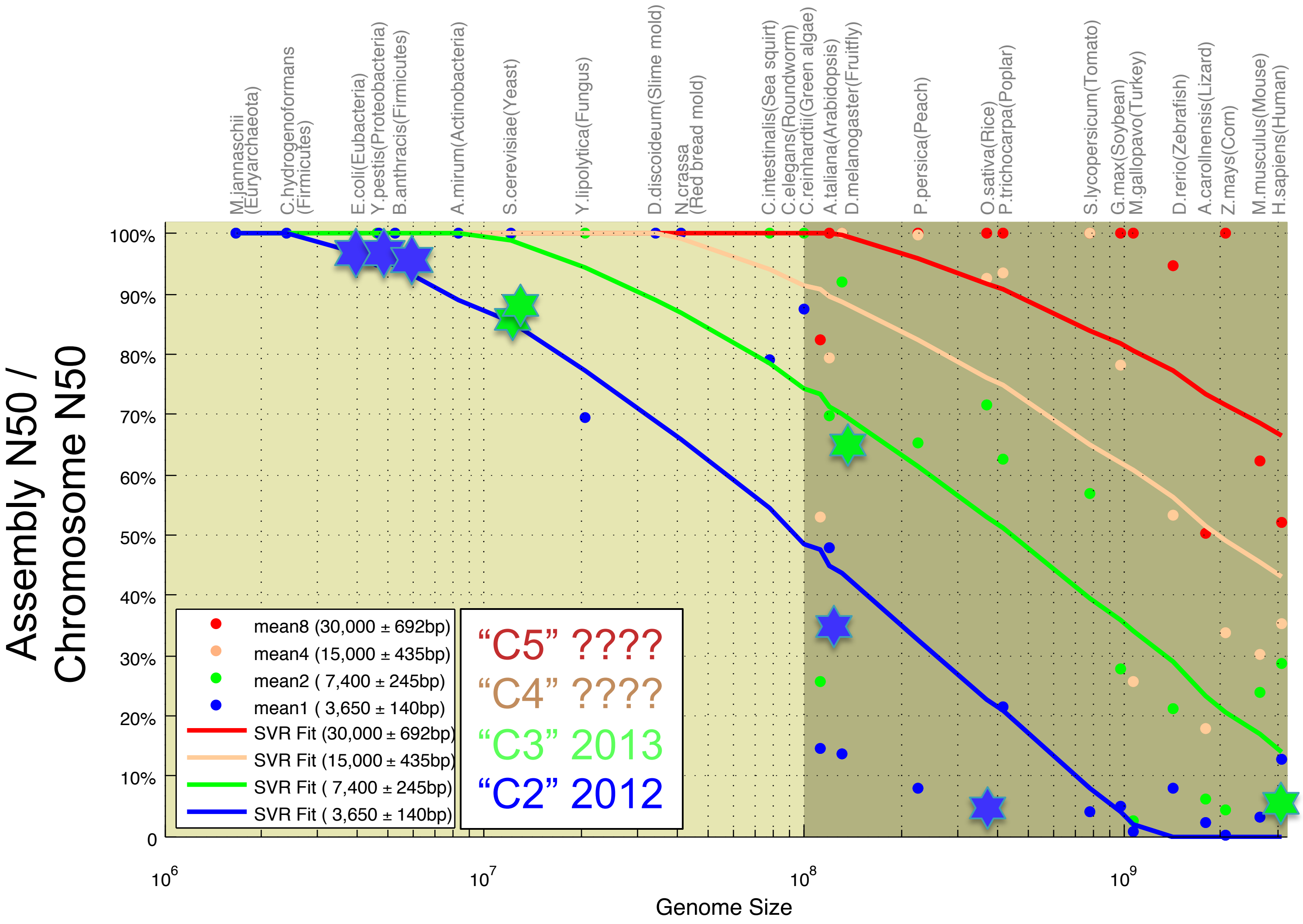
Assembly Complexity of Long Reads



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

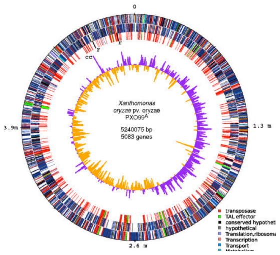
  < 1GB: HGAP/PacBio2CA @ 100x PB C3-P5
  expect high quality assembly: contig N50 over 1Mbp

  > 1GB: hybrid/gap filling
  expect contig N50 to be 100kbp – 1Mbp

  > 5GB: Email mschatz@cshl.edu

- **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

SFARI
SIMONS FOUNDATION
AUTISM RESEARCH INITIATIVE

National Human
Genome Research
Institute

U.S. DEPARTMENT OF
ENERGY

NSF

Thank You!

http://schatzlab.cshl.edu

@mike_schatz