Whole Genome Assembly and Alignment

Michael Schatz

Nov 5, 2013

SBU Graduate Genetics

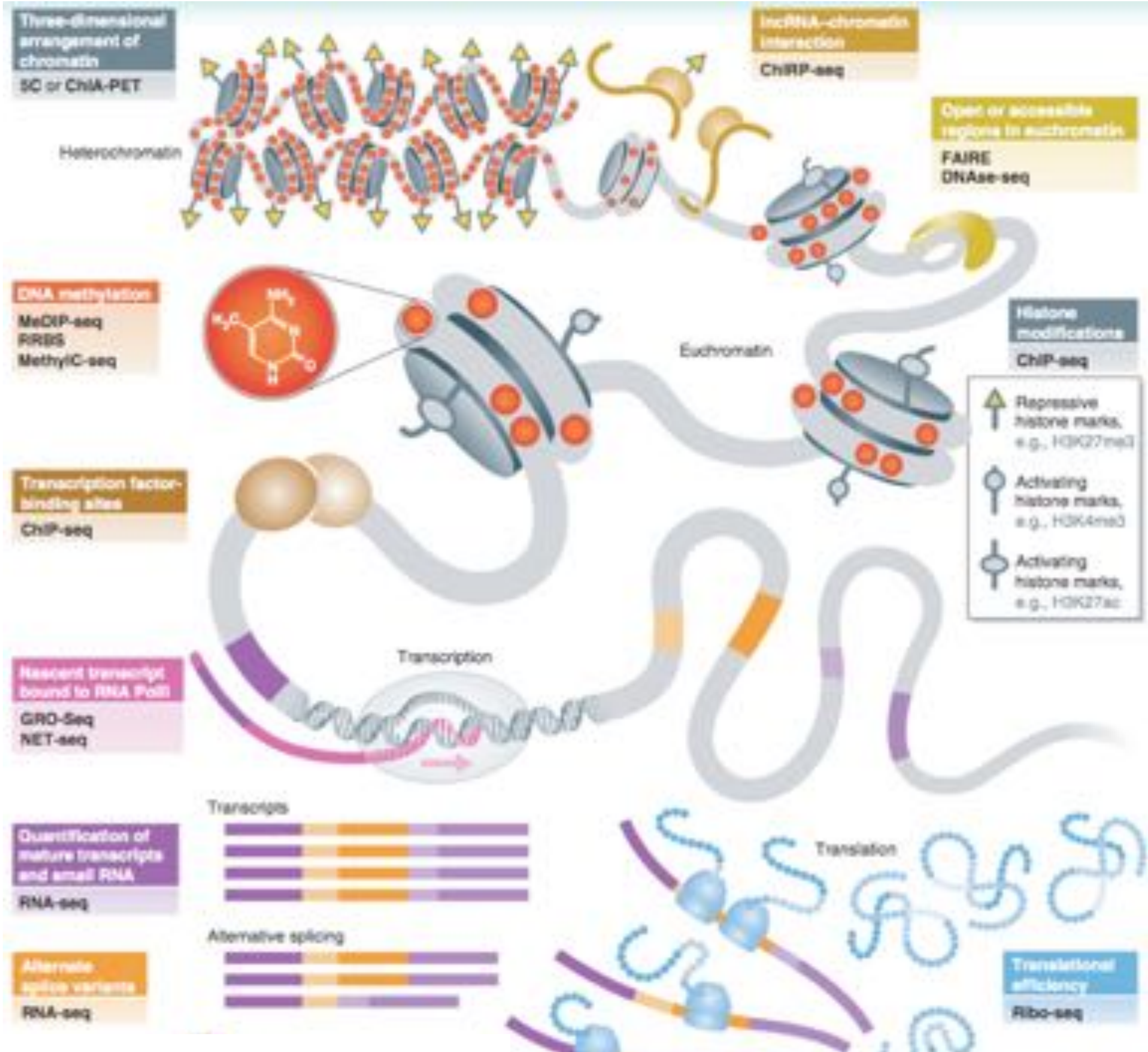
Outline

1. *-seq review
2. Assembly theory
    1. Assembly by analogy
    2. De Bruijn and Overlap graph
    3. Coverage, read length, errors, and repeats
3. Genome assemblers
    1. ALLPATHS-LG
    2. Celera Assembler
4. Whole Genome Alignment with MUMmer

Stories from the Supplement

Lior Pachter
Department of Mathematics and Molecular & Cell Biology
UC Berkeley

November 1, 2013
Genome Informatics, CSHL

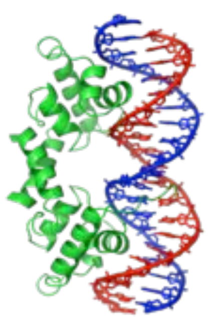


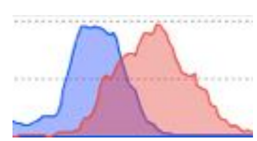
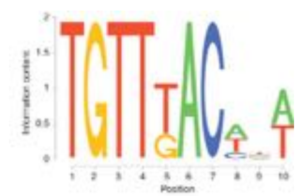
Soon et al., Molecular Systems Biology, 2013

First *Seq assay: ChIP-Seq

Desired measurement → **reduce to sequencing** → Sequence → Solve inverse problem → Analyze

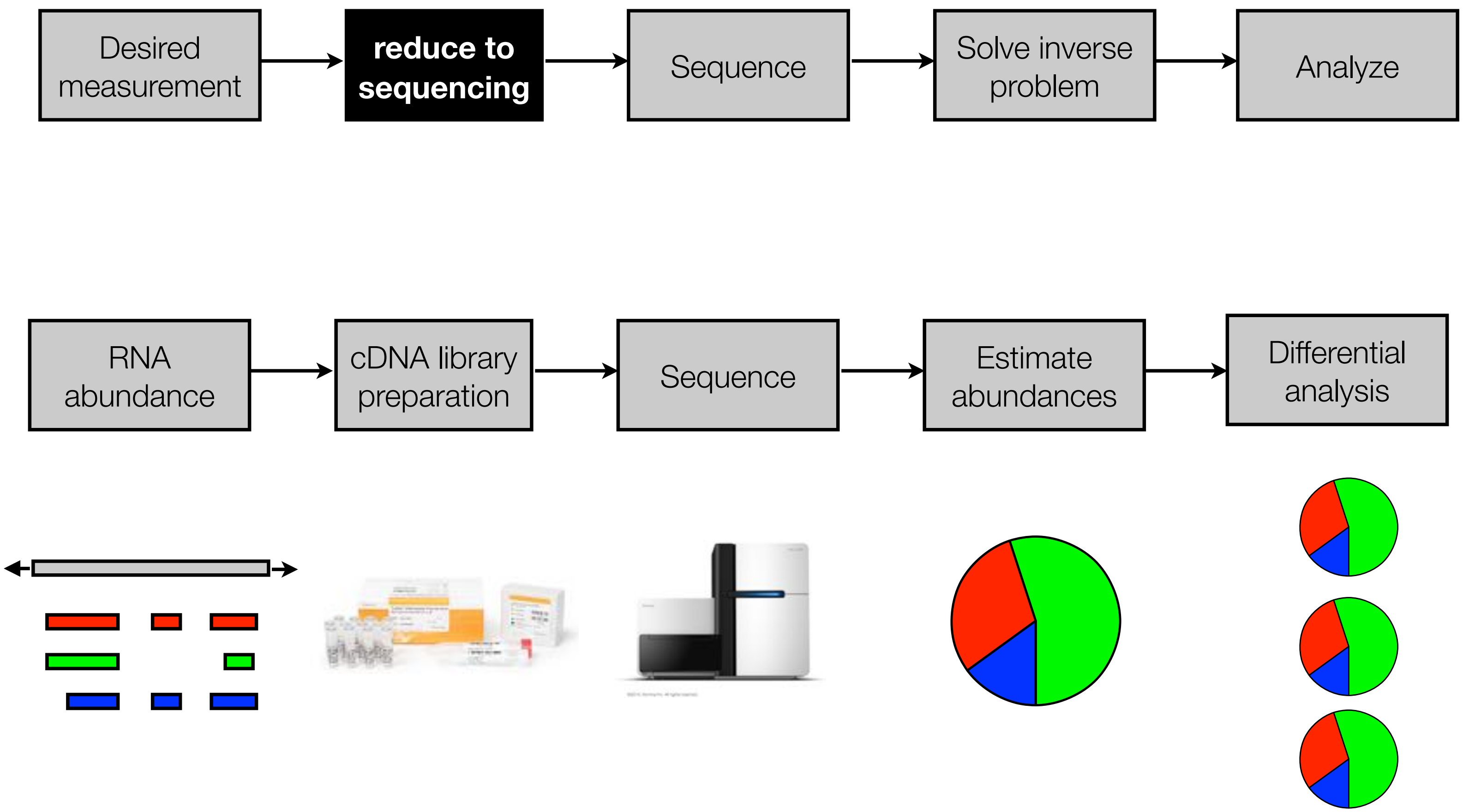
Protein-DNA binding → Cross link, shear, IP → Sequence → Infer binding → Binding affinity analysis



Johnson et al., Science, 2007

Most popular *Seq assay: RNA-Seq



Mortazavi et al., Nature Methods, 2008

What is a *Seq assay?

Molecular Biology

Mathematics & Computer Science

Computational Biology

Desired measurement → **reduce to sequencing** → Sequence → Solve inverse problem → Analyze

Creativity

Chemistry & Physics

Statistics

Biology

$$\mathbb{P}(f = (p,t,l)) \approx \frac{\lambda_l \cdot \frac{\tau_t}{\tilde{l}(t)} \cdot w_{p|t,l} \cdot \phi_{f|p,t,l}}{\sum_{(q,r,m)\in\mathcal{A}(f)} \lambda_m \cdot \frac{\tau_r}{\tilde{l}(r)} \cdot w_{q|r,m} \cdot \phi_{f|q,r,m}}$$

Sequencing Assays

1. Gregory E. Crawford et al., "Genome-wide Mapping of DNase Hypersensitive Sites Using Massively Parallel Signature Sequencing (MPSS)," Genome Research 16, no. 1 (January 1, 2006): 123–131, doi:10.1101/gr.4074106.
2. David S. Johnson et al., "Genome-Wide Mapping of in Vivo Protein-DNA Interactions," Science 316, no. 5830 (June 8, 2007): 1497–1502, doi:10.1126/science.1141319.
3. Tarjei S. Mikkelsen et al., "Genome-wide Maps of Chromatin State in Pluripotent and Lineage-committed Cells," Nature 448, no. 7153 (August 2, 2007): 553–560, doi:10.1038/nature06008.
4. Nathan A. Baird et al., "Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers," PLoS ONE 3, no. 10 (October 13, 2008): e3376, doi:10.1371/journal.pone.0003376.
5. Leighton J. Core, Joshua J. Waterfall, and John T. Lis, "Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters," Science 322, no. 5909 (December 19, 2008): 1845–1848, doi:10.1126/science.1162228.
6. Thomas A. Down et al., "A Bayesian Deconvolution Strategy for Immunoprecipitation-based DNA Methylome Analysis," Nature Biotechnology 26, no. 7 (July 2008): 779–785, doi:10.1038/nbt1414.
7. Ali Mortazavi et al., "Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq," Nature Methods 5, no. 7 (July 2008): 621–628, doi:10.1038/nmeth.1226.
8. Alayne L. Brunner et al., "Distinct DNA Methylation Patterns Characterize Differentiated Human Embryonic Stem Cells and Developing Human Fetal Liver," Genome Research 19, no. 6 (June 1, 2009): 1044–1056, doi:10.1101/gr.088773.108.
9. Melissa J. Fullwood et al., "An Oestrogen-receptor-α-bound Human Chromatin Interactome," Nature 462, no. 7269 (November 5, 2009): 58–64, doi:10.1038/nature08497.
10. Jay R. Hesselberth et al., "Global Mapping of protein-DNA Interactions in Vivo by Digital Genomic Footprinting," Nature Methods 6, no. 4 (April 2009): 283–289, doi:10.1038/nmeth.1313.
11. Nicholas T. Ingolia et al., "Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling," Science 324, no. 5924 (April 10, 2009): 218–223, doi:10.1126/science.1168978.
12. Gemma C. Langridge et al., "Simultaneous Assay of Every Salmonella Typhi Gene Using One Million Transposon Mutants," Genome Research (October 13, 2009), doi:10.1101/gr.097097.109.
13. Erez Lieberman-Aiden et al., "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome," Science 326, no. 5950 (October 9, 2009): 289–293, doi:10.1126/science.1181369.
14. Ryan Lister et al., "Human DNA Methylomes at Base Resolution Show Widespread Epigenomic Differences," Nature 462, no. 7271 (November 19, 2009): 315–322, doi:10.1038/nature08514.
15. Andrew M. Smith et al., "Quantitative Phenotyping via Deep Barcode Sequencing," Genome Research (July 21, 2009), doi:10.1101/

Outline

1. *-seq review

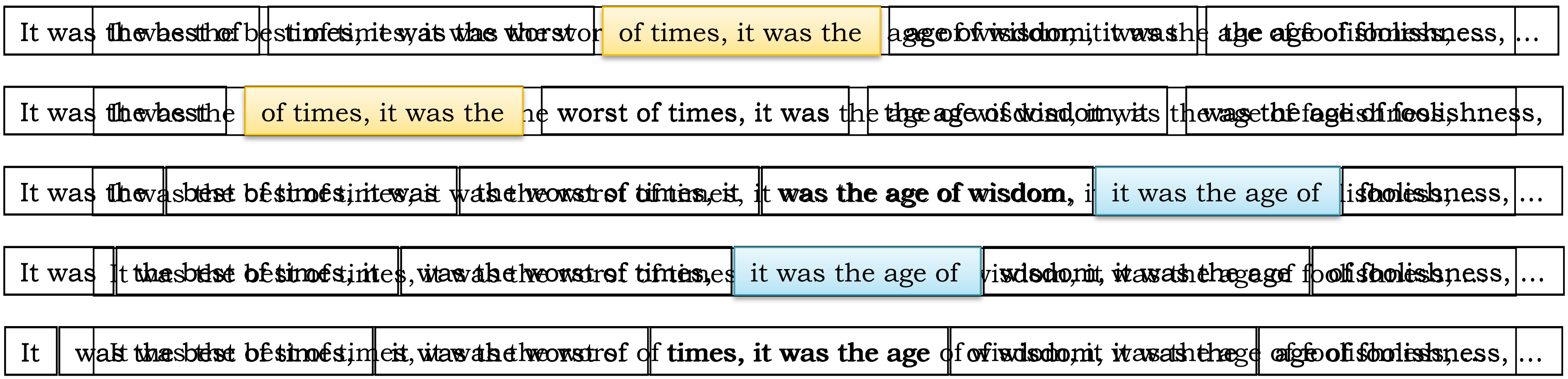
2. Assembly theory
    1. Assembly by analogy
    2. De Bruijn and Overlap graph
    3. Coverage, read length, errors, and repeats

3. Genome assemblers
    1. ALLPATHS-LG
    2. Celera Assembler

4. Whole Genome Alignment with MUMmer

Shredded Book Reconstruction

- Dickens accidentally shreds the first printing of <u>A Tale of Two Cities</u>
  - 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

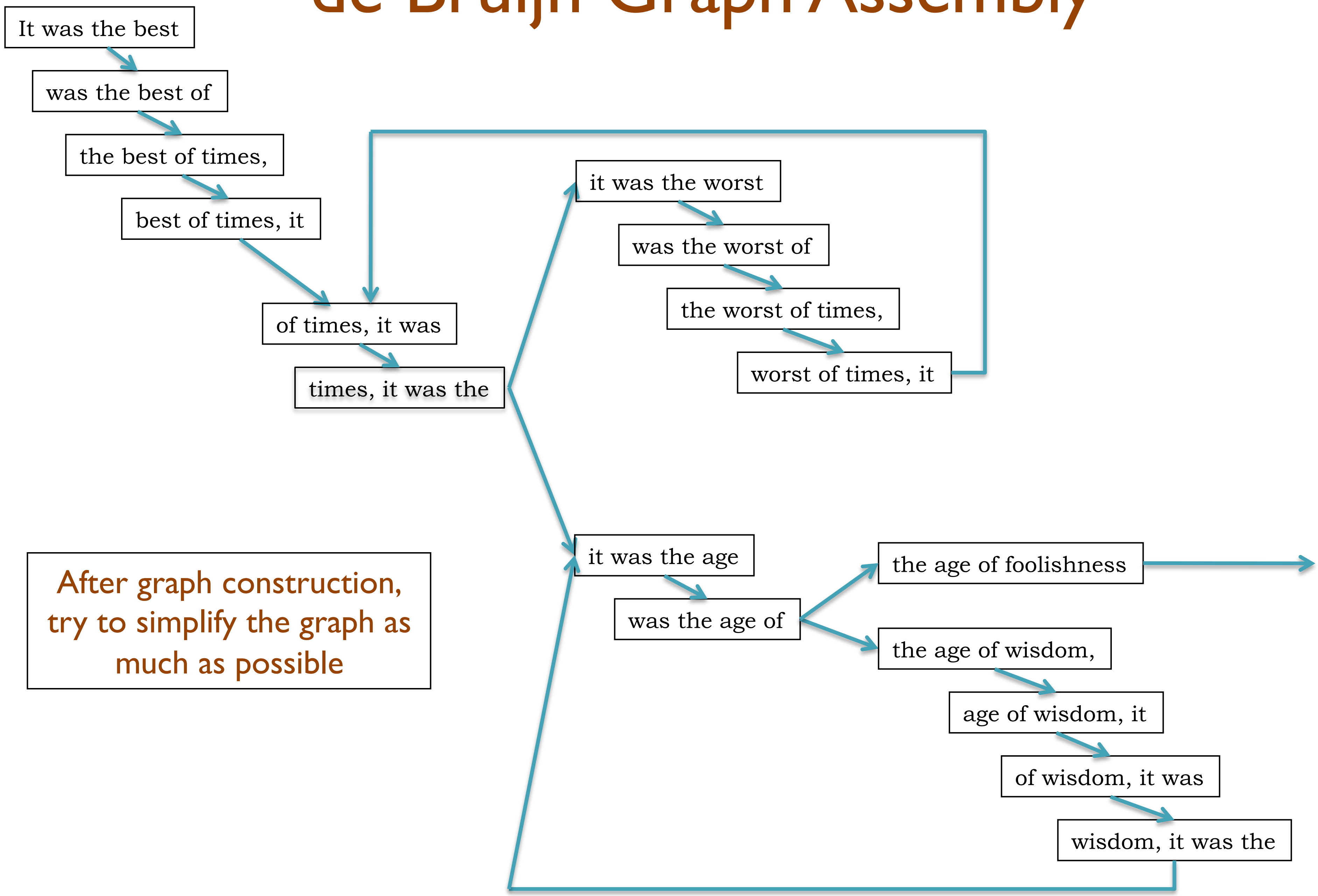
Directed Edge

It was the best → was the best of

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

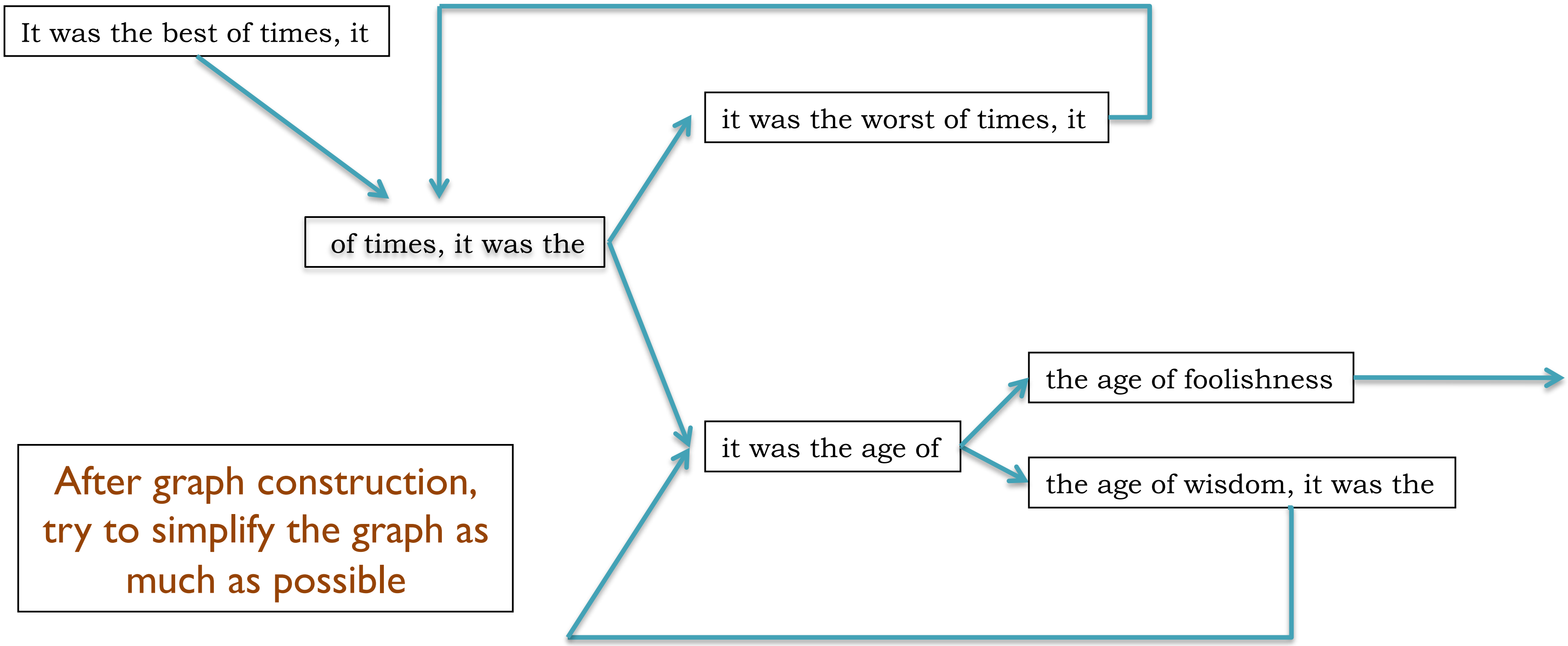
de Bruijn, 1946
Idury and Waterman, 1995
Pevzner, Tang, Waterman, 2001

de Bruijn Graph Assembly



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

de Bruijn Graph Assembly



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

The full tale

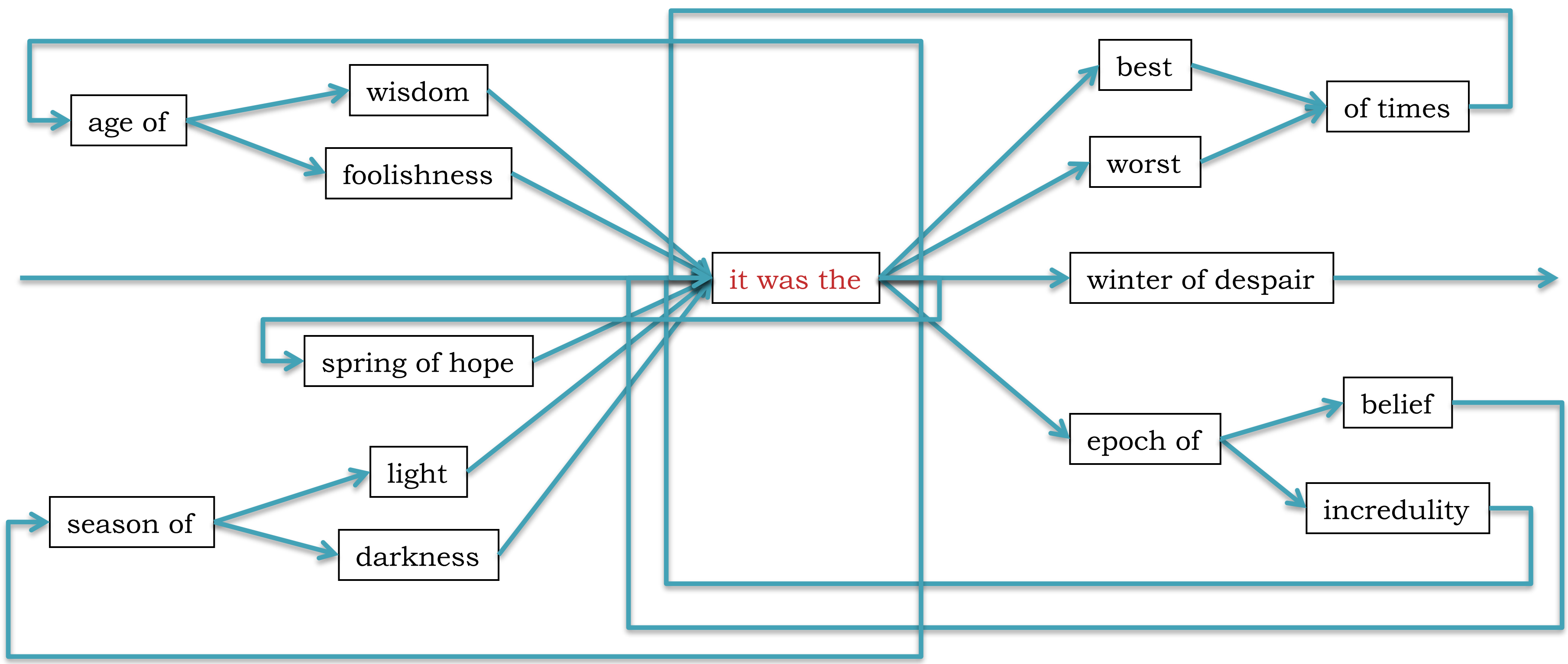
… it was the best of times it was the worst of times …

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

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

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

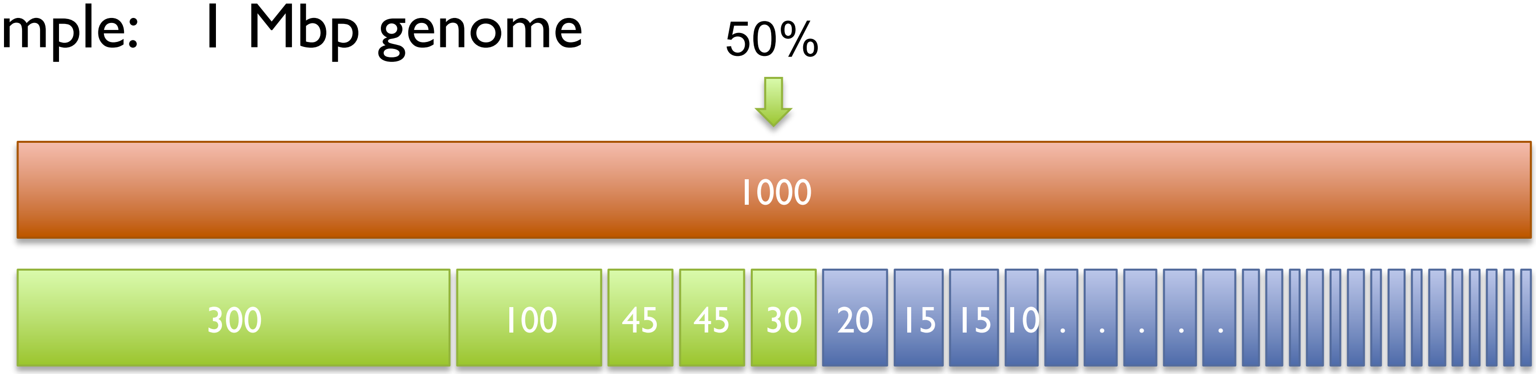
… it was the spring of hope it was the winder of despair …



N50 size

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

Example: 1 Mbp genome



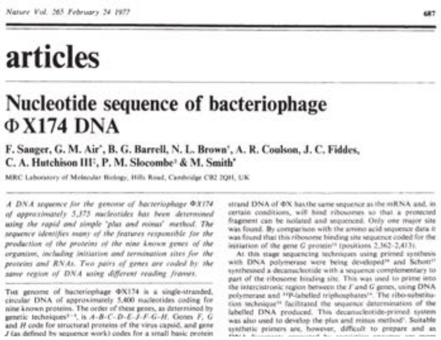
N50 size = 30 kbp

(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

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

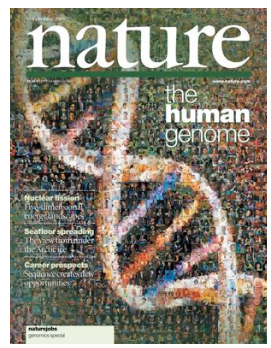
MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

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



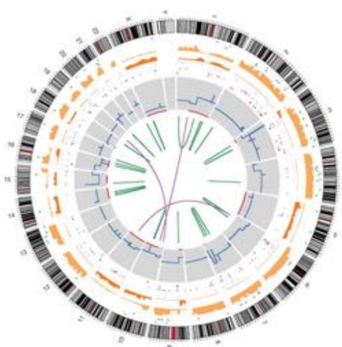
2001. Venter *et al.*, IHGSC
Human Genome
Celera Assembler/GigaAssembler. 2.9 Gbp

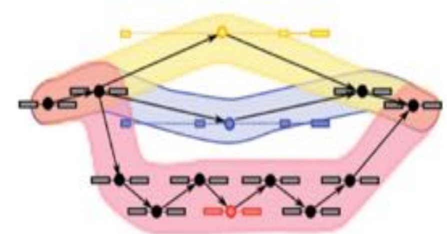
2010. Li *et al.*
1st Large SGS Assembly.
SOAPdenovo 2.2 Gbp

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

Assembly Applications

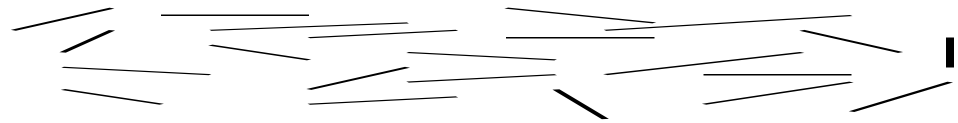
- Novel genomes
- Metagenomes
- Sequencing assays
  - Structural variations
  - Transcript assembly
  - …





Assembling a Genome

1. Shear & Sequence DNA

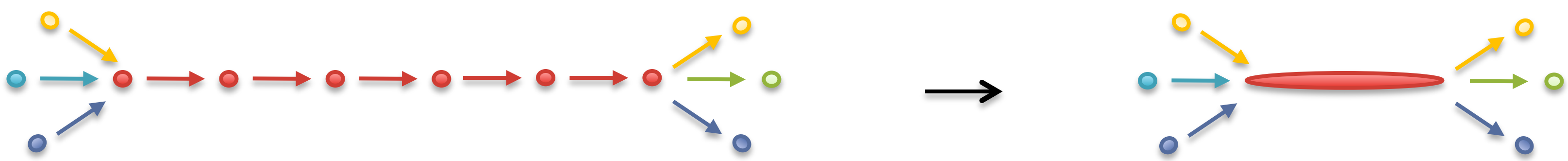


2. Construct assembly graph from overlapping reads

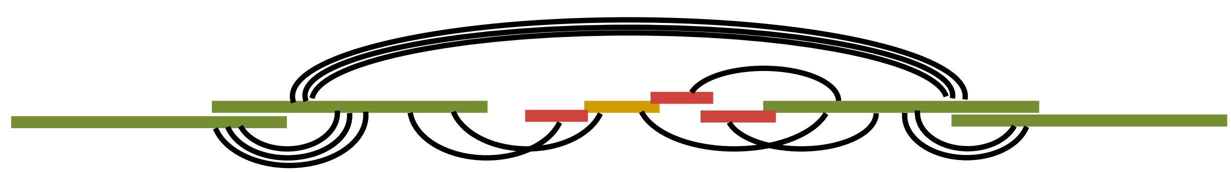
...AGCCTAGACCTACAGGATGCGCGACACGT

GGATGCGCGACACGTCGCATATCCGGT...

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

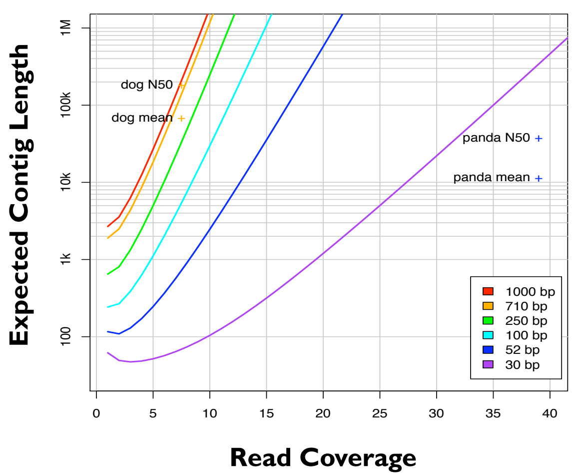
2. ***Sequencing***:
   – (Very) large genomes, imperfect sequencing

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

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

Ingredients for a good assembly

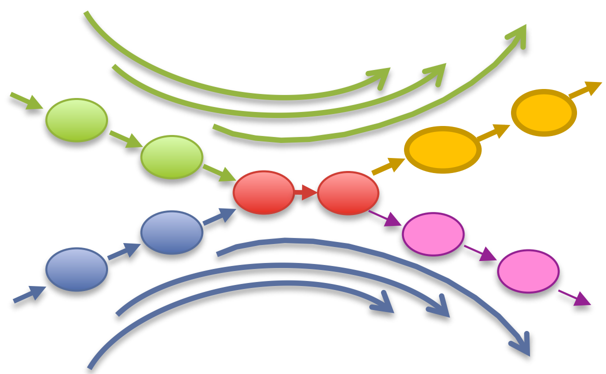
Coverage



High coverage is required

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

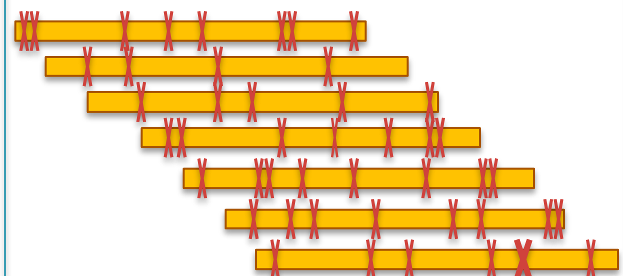
Read Length



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

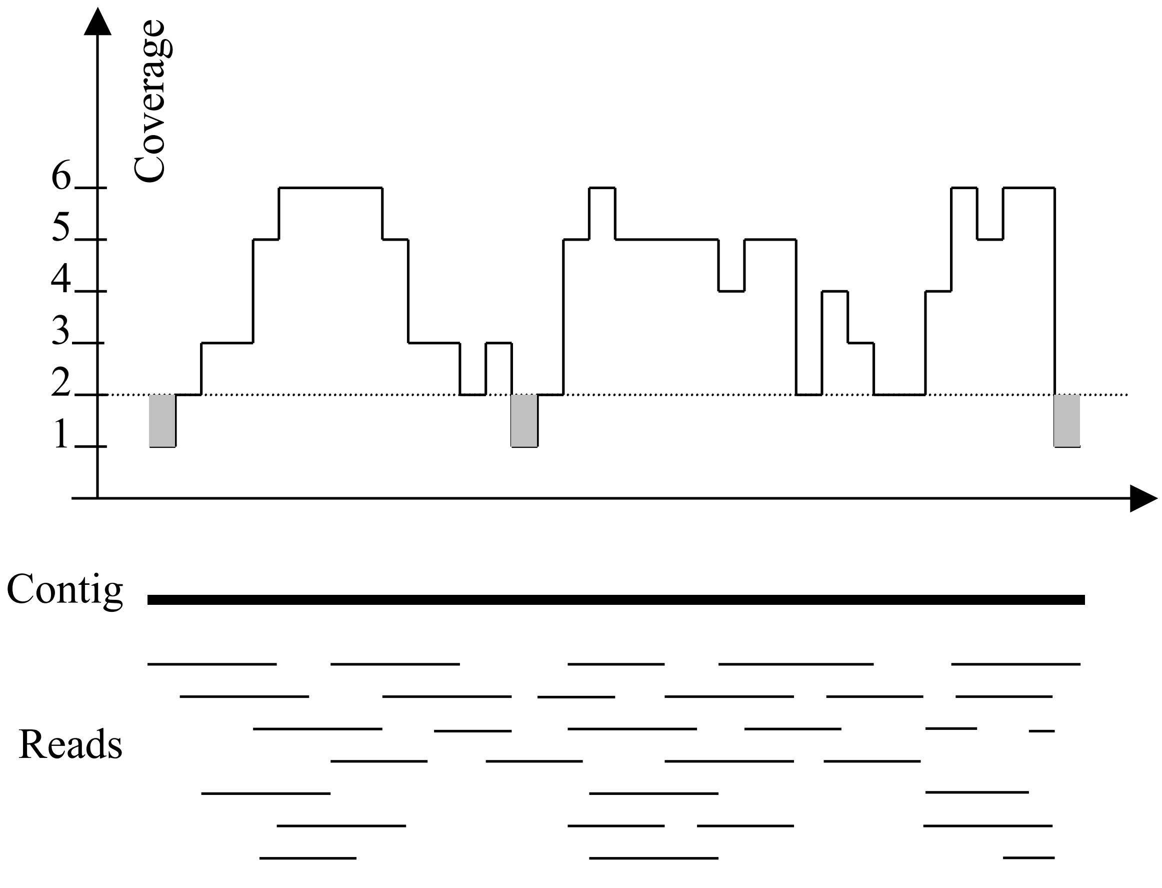


Errors obscure overlaps

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

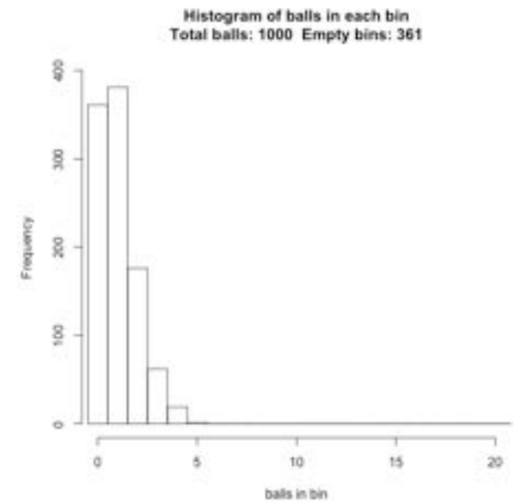
Current challenges in *de novo* plant genome sequencing and assembly
Schatz MC, Witkowski, McCombie, WR (2012) *Genome Biology*. 12:243

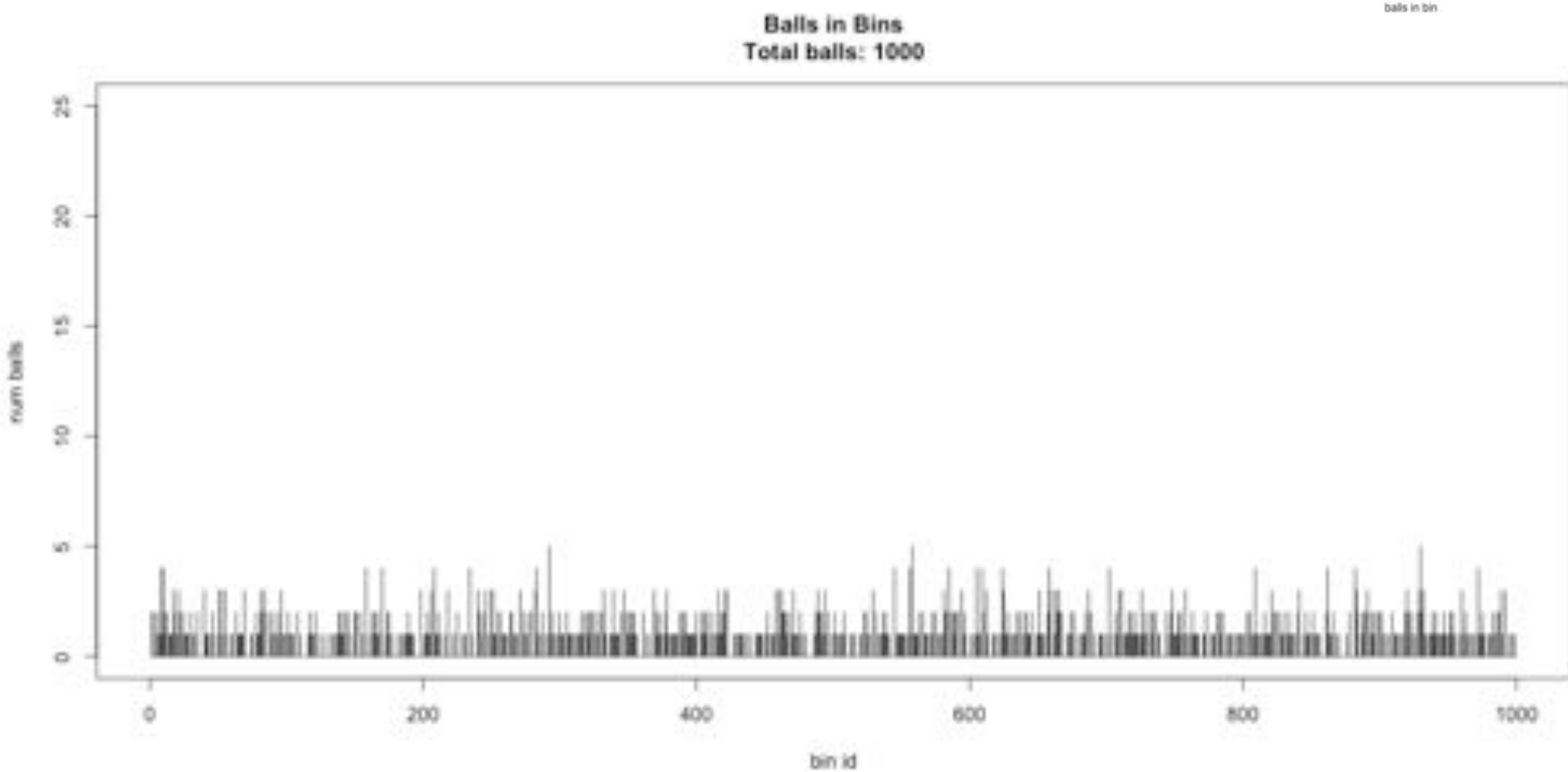
Typical contig coverage



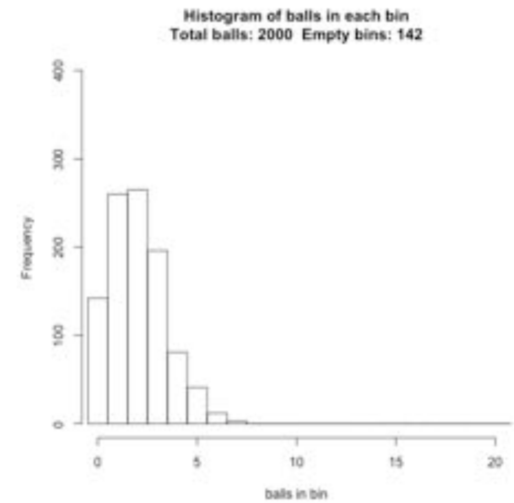
Imagine raindrops on a sidewalk

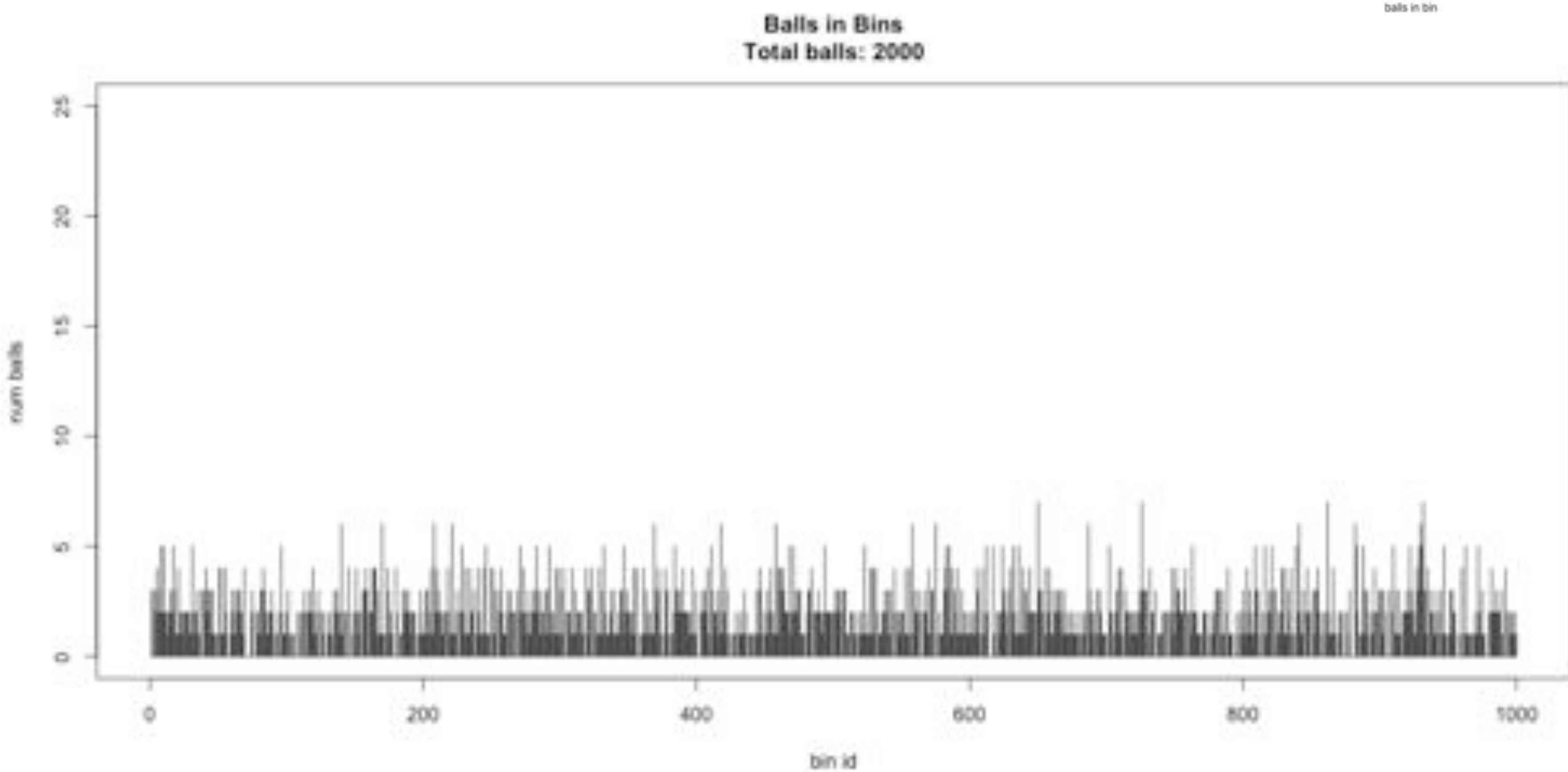
Balls in Bins 1x



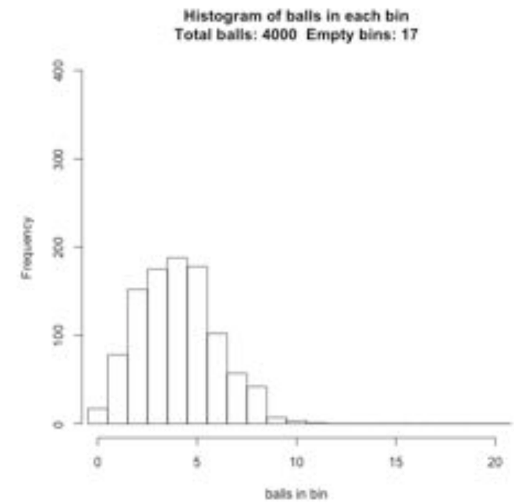


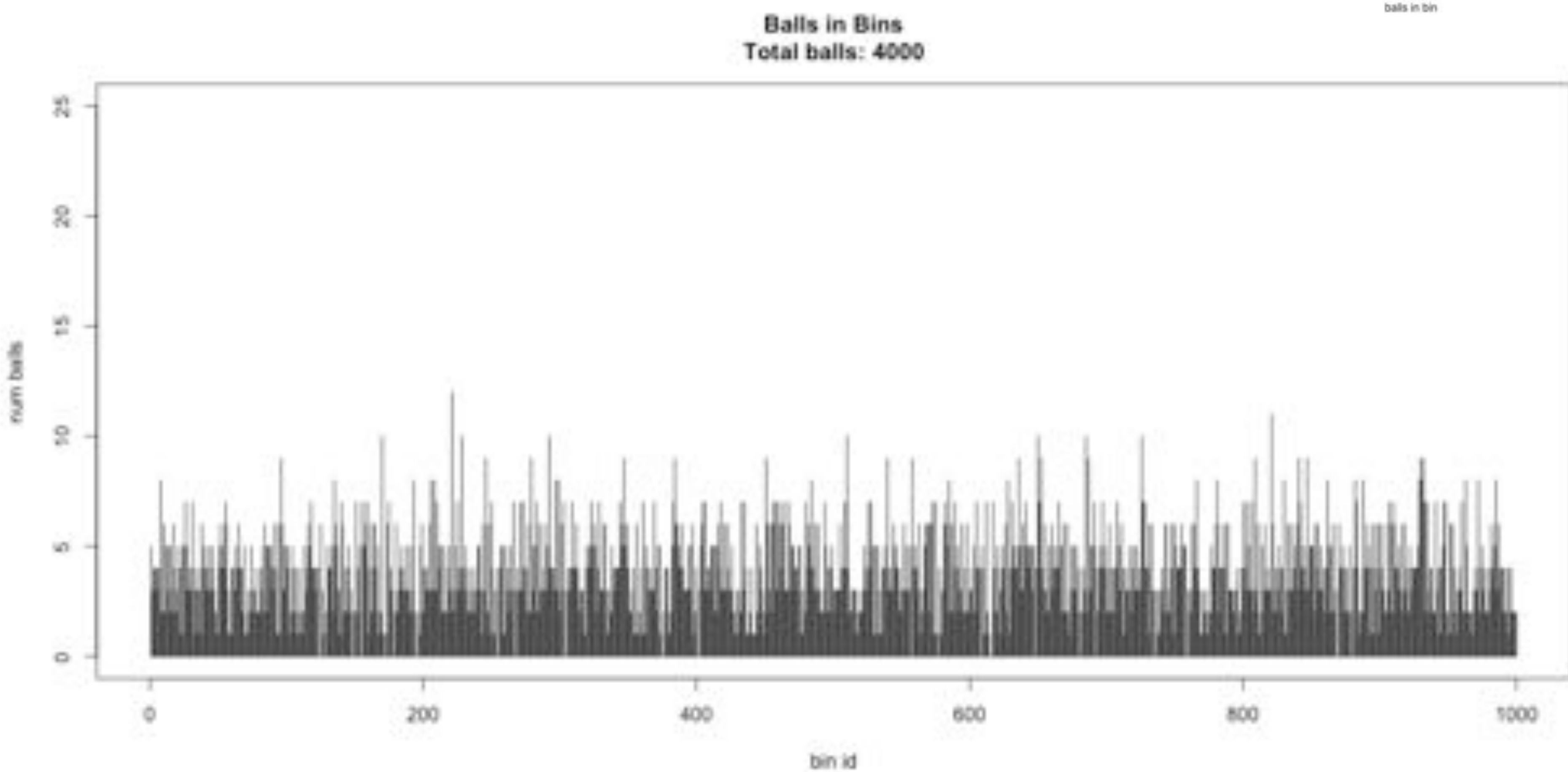
Balls in Bins 2x



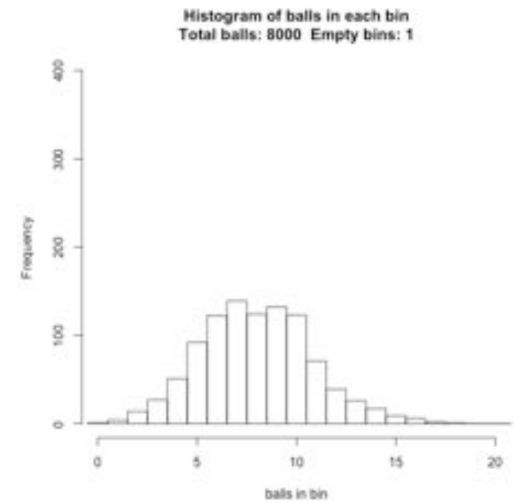


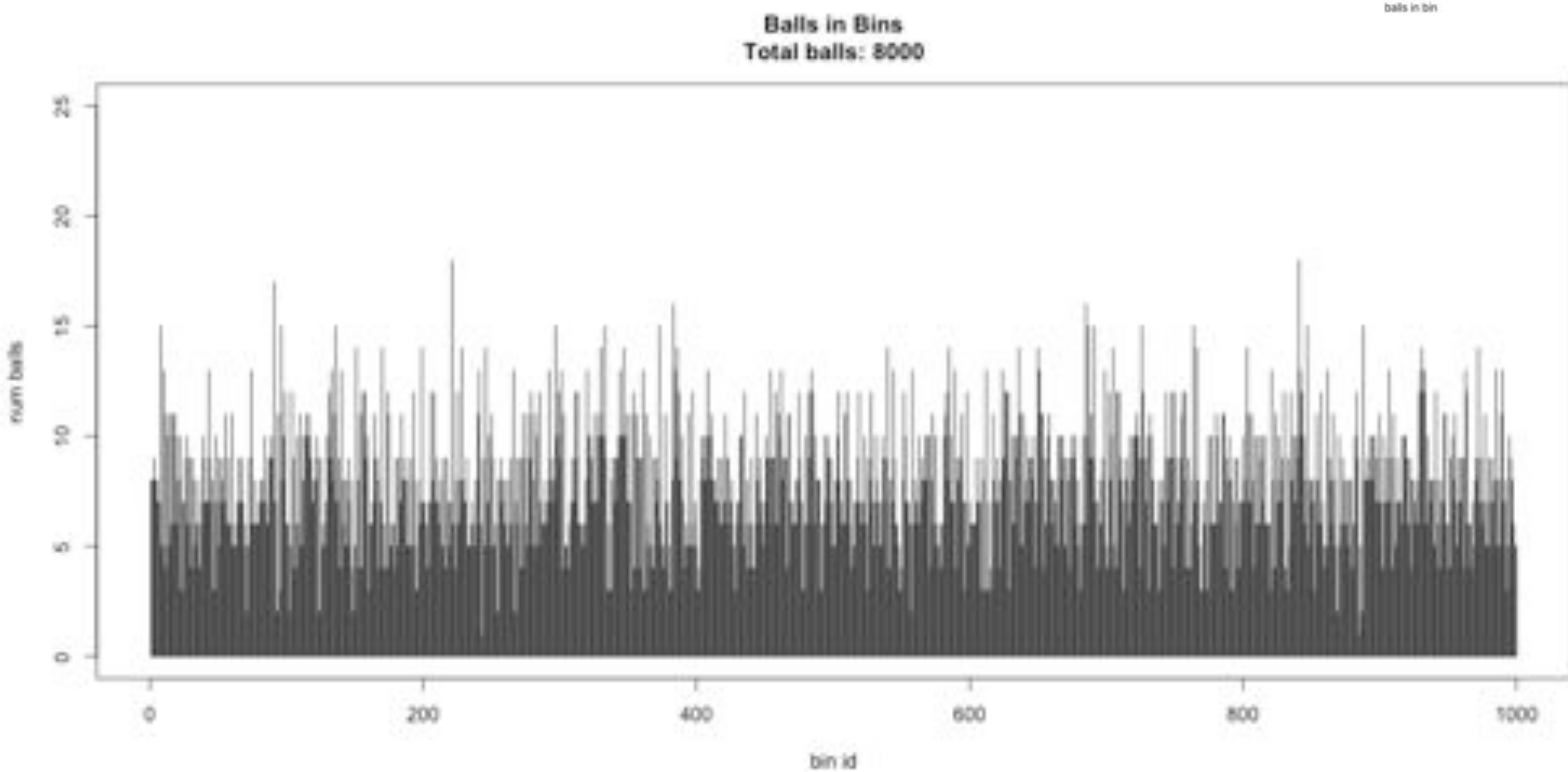
Balls in Bins 4x





Balls in Bins 8x

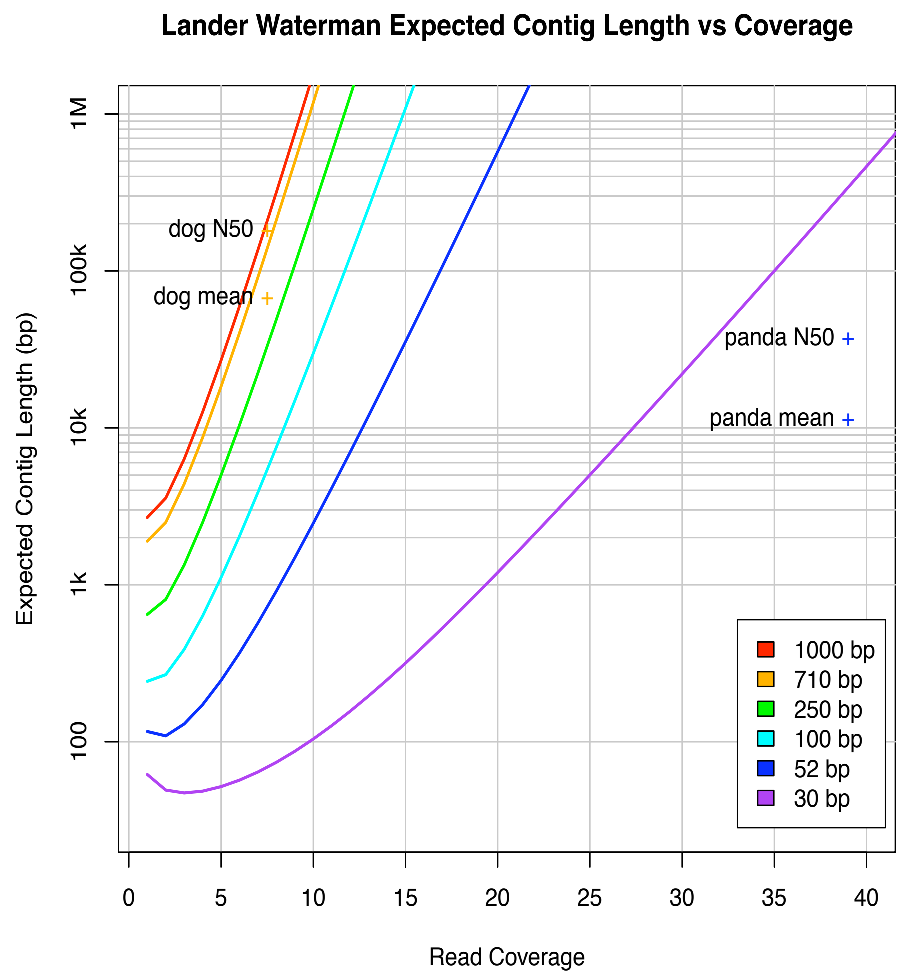




Coverage and Read Length

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

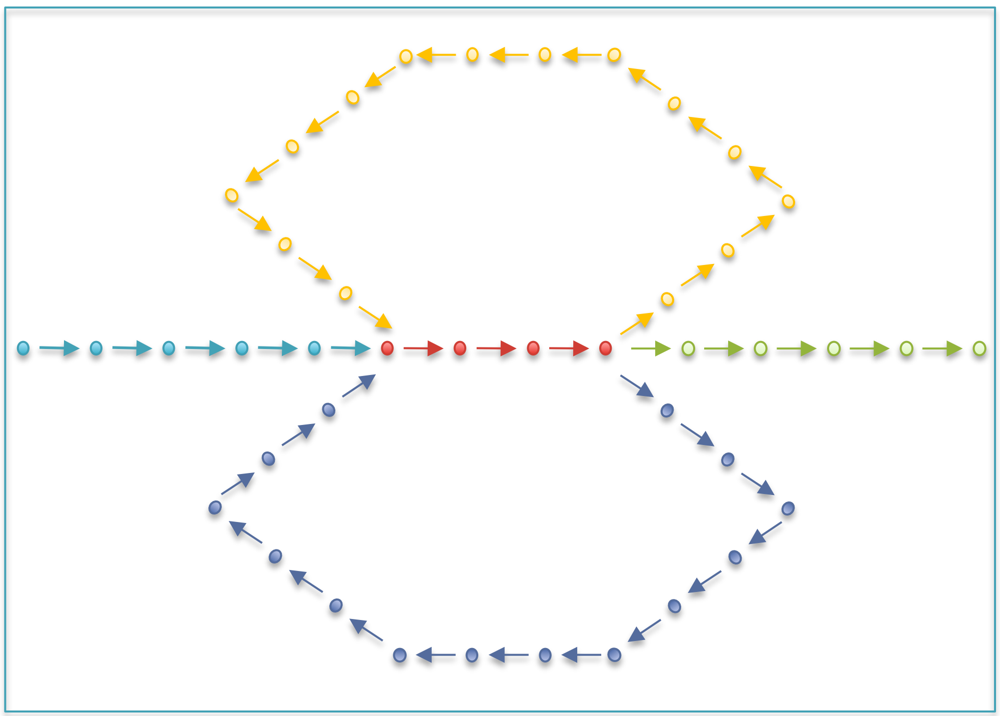


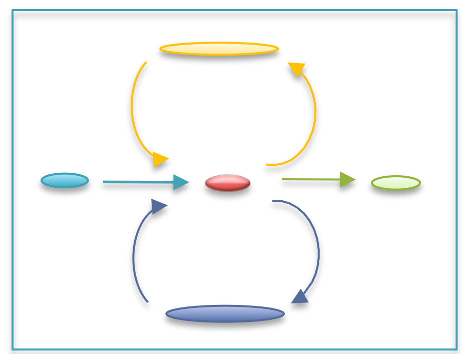
Assembly of Large Genomes using Second Generation Sequencing
Schatz MC, Delcher AL, Salzberg SL (2010) *Genome Research*. 20:1165-1173.

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





Repetitive regions

Repeat Type	Definition / Example	Prevalence
Low-complexity DNA / Microsatellites	$(b_1 b_2 \ldots b_k)^N$ where $1 \leq k \leq 6$ CACACACACACACACACACA	2%
SINEs (Short Interspersed Nuclear Elements)	*Alu* sequence (~280 bp) Mariner elements (~80 bp)	13%
LINEs (Long Interspersed Nuclear Elements)	~500 – 5,000 bp	21%
LTR (long terminal repeat) retrotransposons	Ty1-copia, Ty3-gypsy, Pao-BEL (~100 – 5,000 bp)	8%
Other DNA transposons		3%
Gene families & segmental duplications		4%

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

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

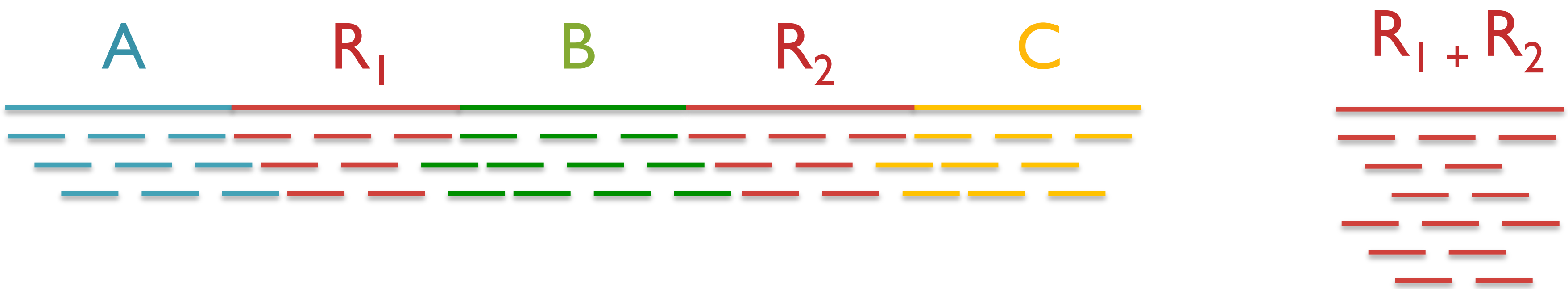
300bp

Mate-pair sequencing

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

10kbp

Repeats and Coverage Statistics



- If *n* reads are a uniform random sample of the genome of length *G*, we expect $k=n\Delta/G$ reads to start in a region of length Δ.
  - If we see many more reads than k (if the arrival rate is > A) , it is likely to be a collapsed repeat

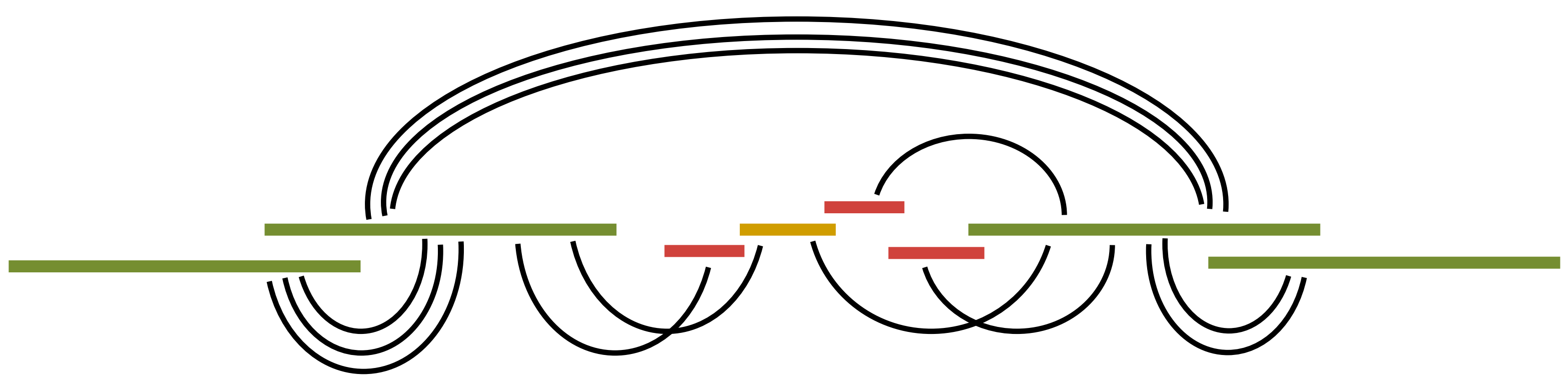
$$\Pr(X-copy) = \binom{n}{k}\left(\frac{X\Delta}{G}\right)^{k}\left(\frac{G-X\Delta}{G}\right)^{n-k} \qquad A(\Delta,k) = \ln\left(\frac{\Pr(1-copy)}{\Pr(2-copy)}\right) = \ln\left(\frac{\frac{(\Delta n/G)^k}{k!}e^{\frac{-\Delta n}{G}}}{\frac{(2\Delta n/G)^k}{k!}e^{\frac{-2\Delta n}{G}}}\right) = \frac{n\Delta}{G}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

The fragment assembly string graph
Myers, EW (2005) Bioinformatics. 21(suppl 2): ii79-85.

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



Outline

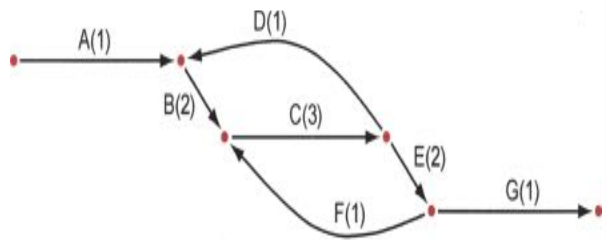
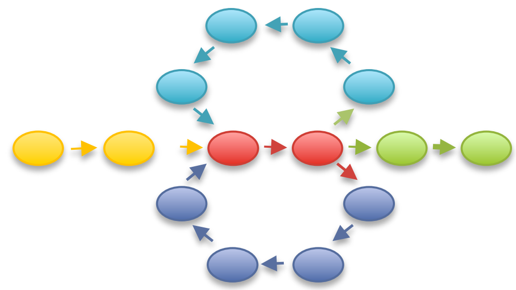
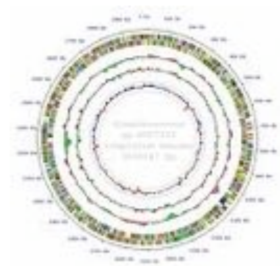
1. *-seq review

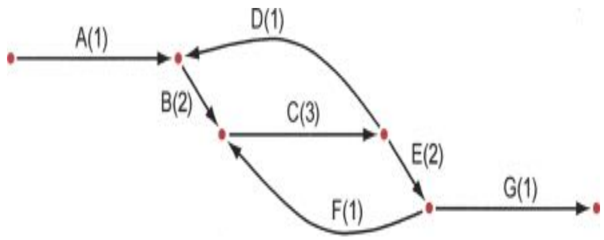
2. Assembly theory
    1. Assembly by analogy
    2. De Bruijn and Overlap graph
    3. Coverage, read length, errors, and repeats

3. Genome assemblers
    1. ALLPATHS-LG
    2. Celera Assembler

4. Whole Genome Alignment with MUMmer

Assembly Algorithms

ALLPATHS-LG	SOAPdenovo	Celera Assembler
		
Broad's assembler (Gnerre et al. 2011)	BGI's assembler (Li et al. 2010)	JCVI's assembler (Miller et al. 2008)
De bruijn graph Short + PacBio (patching)	De bruijn graph Short reads	Overlap graph Medium + Long reads
Easy to run if you have compatible libraries	Most flexible, but requires a lot of tuning	Supports Illumina/454/PacBio Hybrid assemblies
http://www.broadinstitute.org/software/allpaths-lg/blog/	http://soap.genomics.org.cn/soapdenovo.html	http://wgs-assembler.sf.net



Genome assembly with ALLPATHS-LG
Iain MacCallum

BROAD INSTITUTE

ALLPATHS-LG sequencing model

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

*See next slide.

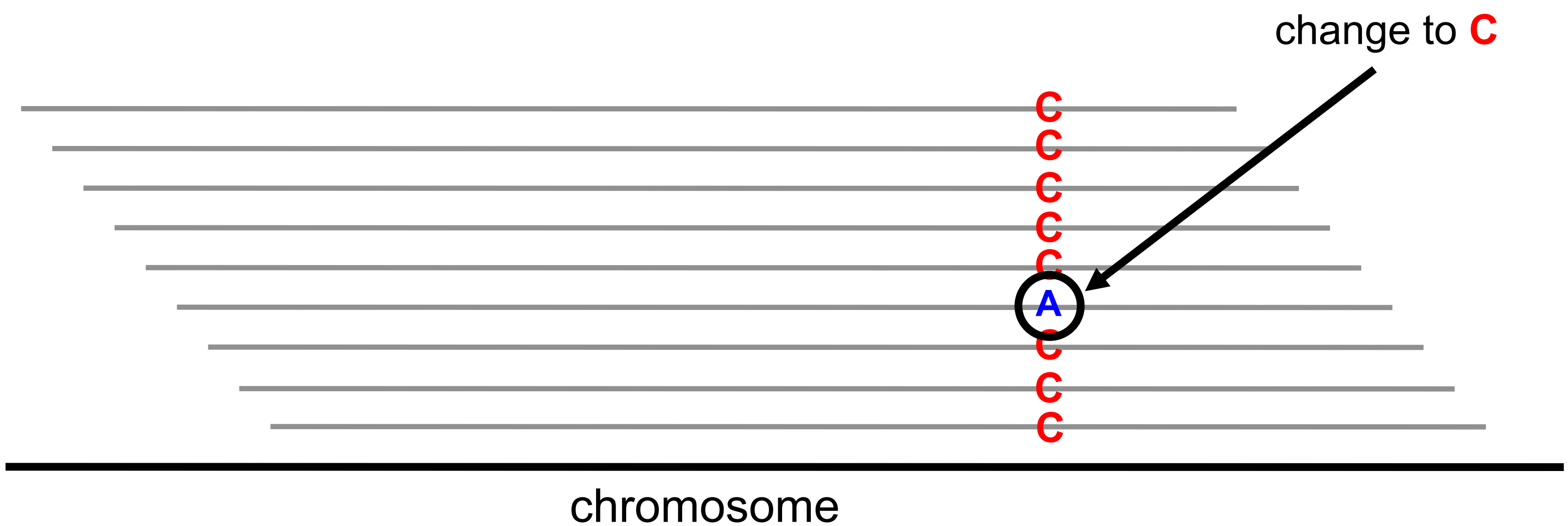
**For best results. Normally not used for small genomes. However essential to assemble long repeats or duplications.

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

All: protocols are either available, or in progress.

Error correction

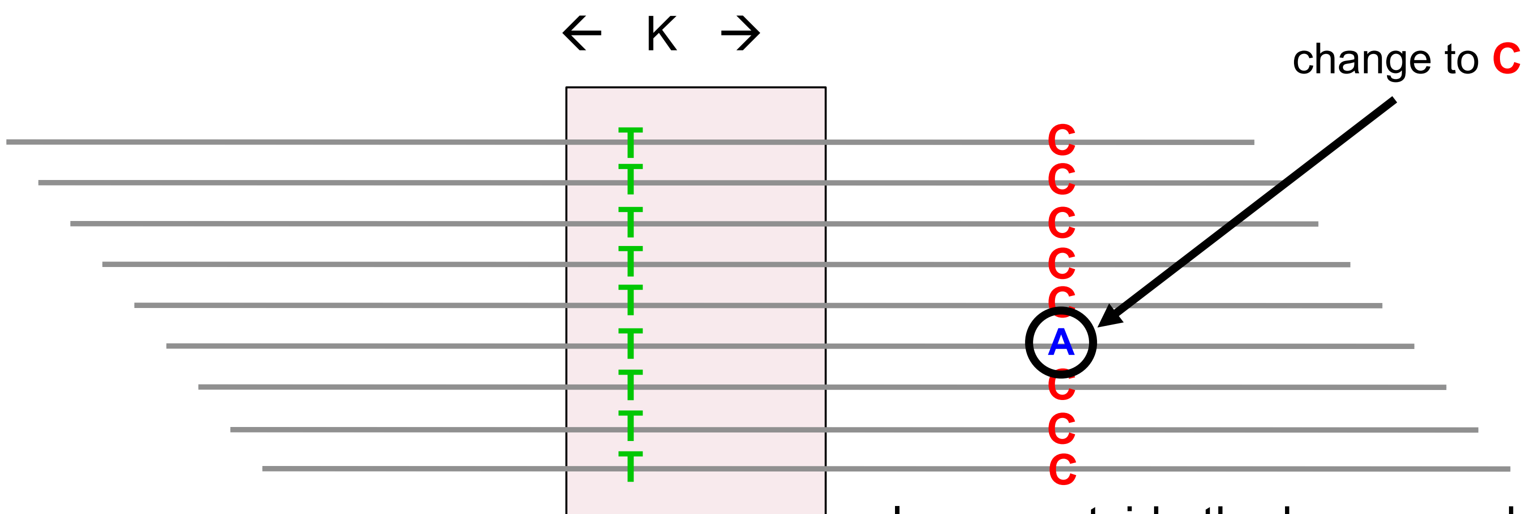
Given a crystal ball, we could stack reads on the chromosomes they came from (with homologous chromosomes separate), then let each column 'vote':



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

Error correction

<u>ALLPATHS-LG.</u> 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)



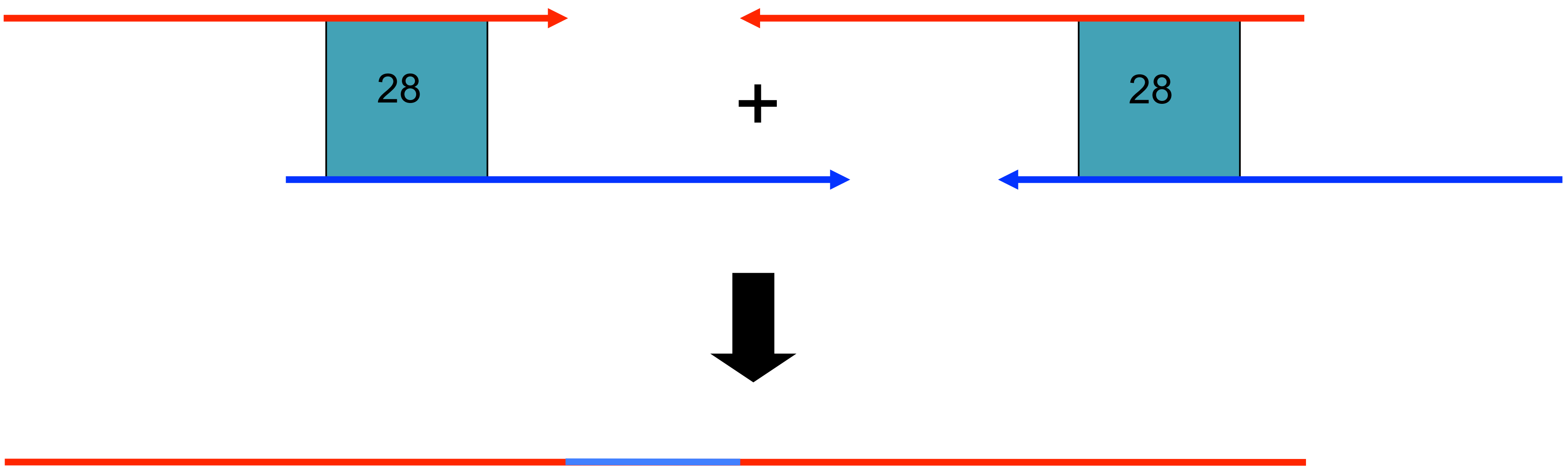
columns outside the kmer may be mixed

columns inside the kmer are homogeneous

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

Read doubling

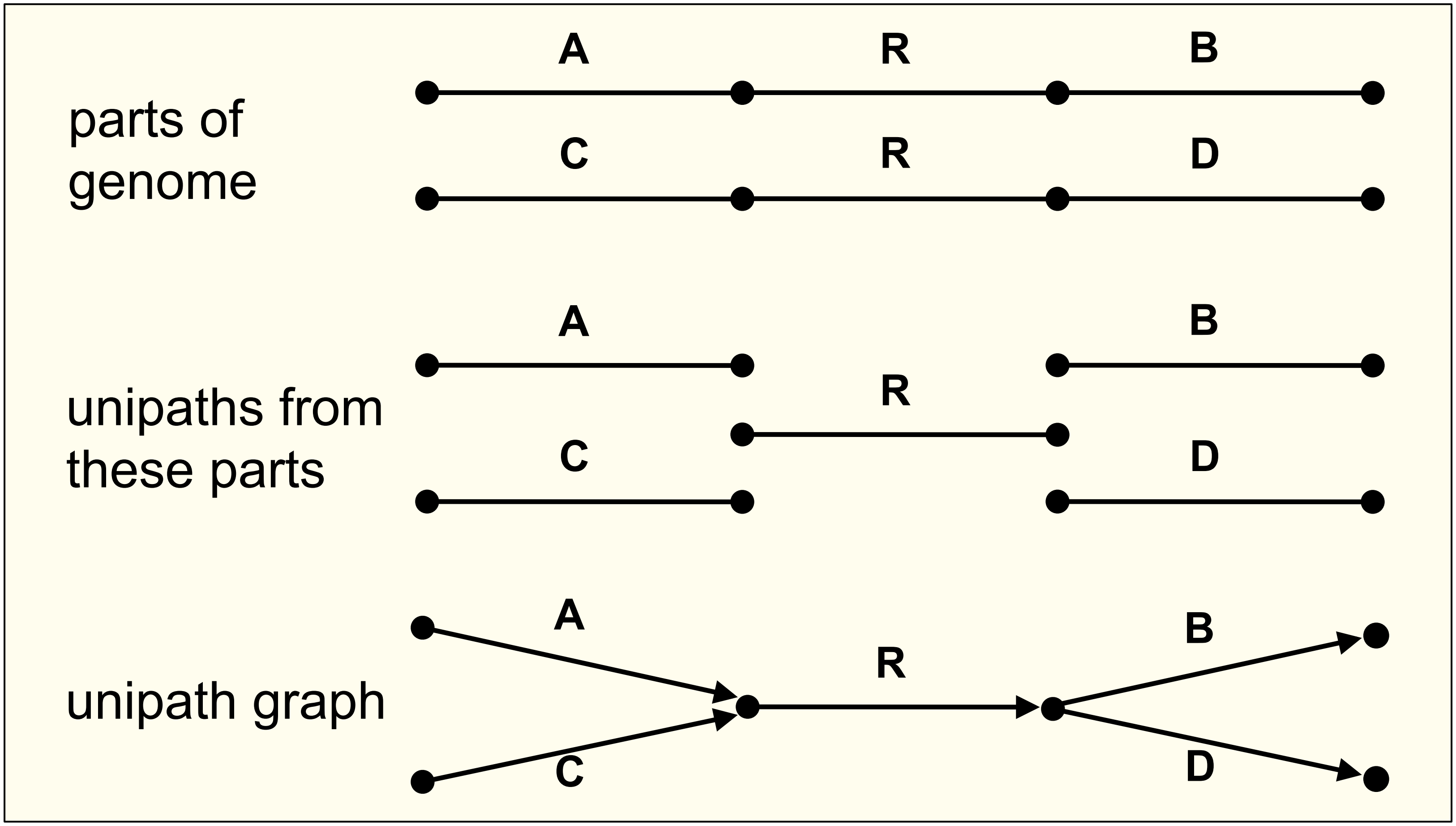
To close a read pair (red), we require the existence of another read pair (blue), overlapping perfectly like this:



More than one closure allowed (but rare).

Unipaths

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

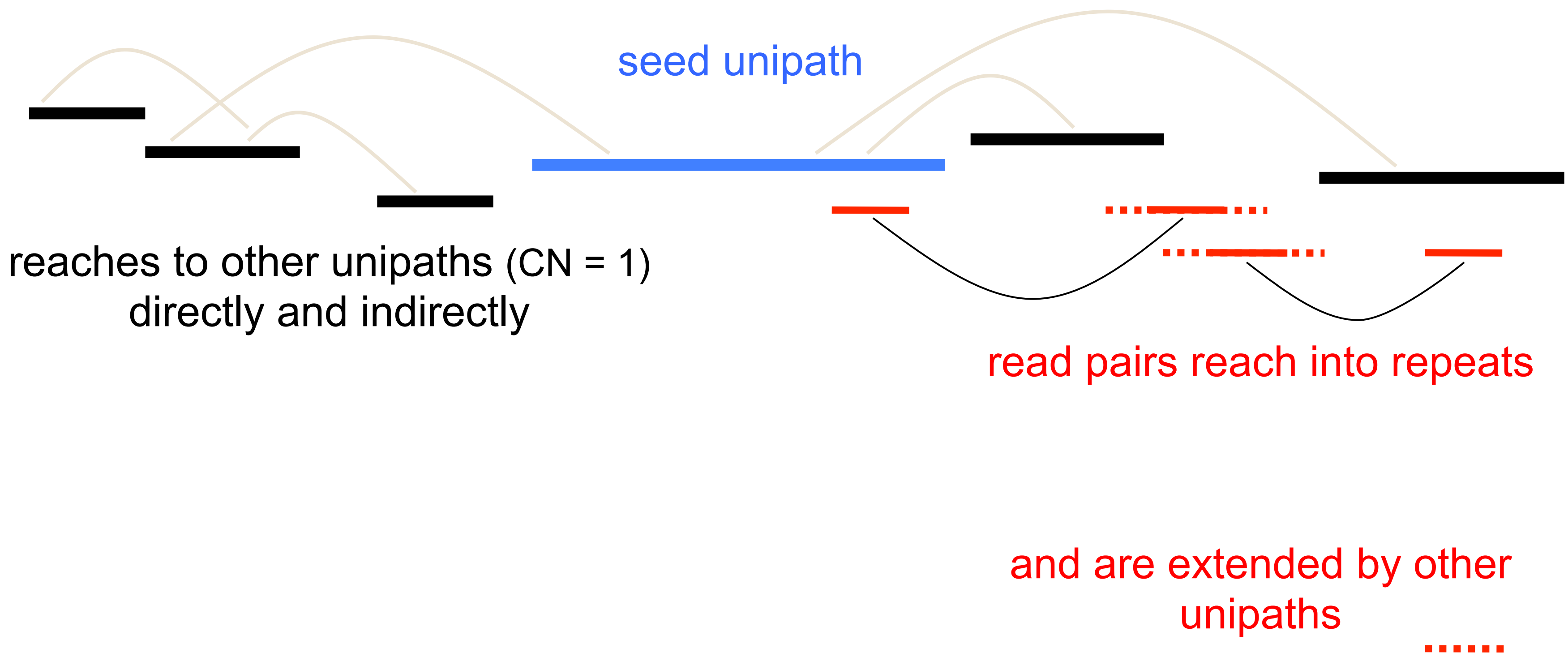


Adjacent unipaths overlap by K-1 bases

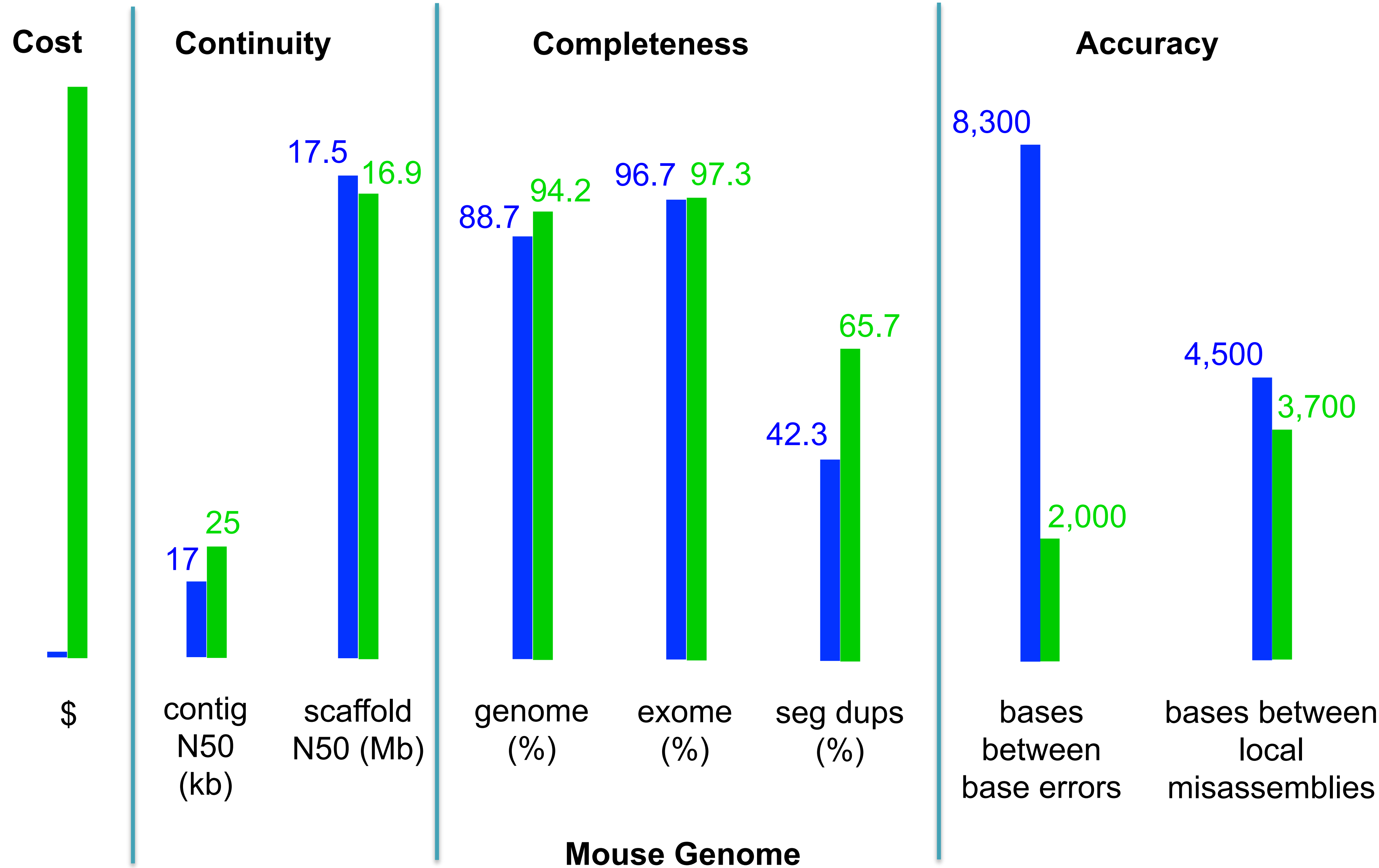
Localization

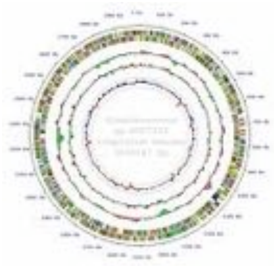
I. Find ‘seed’ unipaths, evenly spaced across genome
(ideally long, of copy number CN = 1)

II. Form neighborhood around each seed



Large genome recipe: ALLPATHS-LG vs capillary



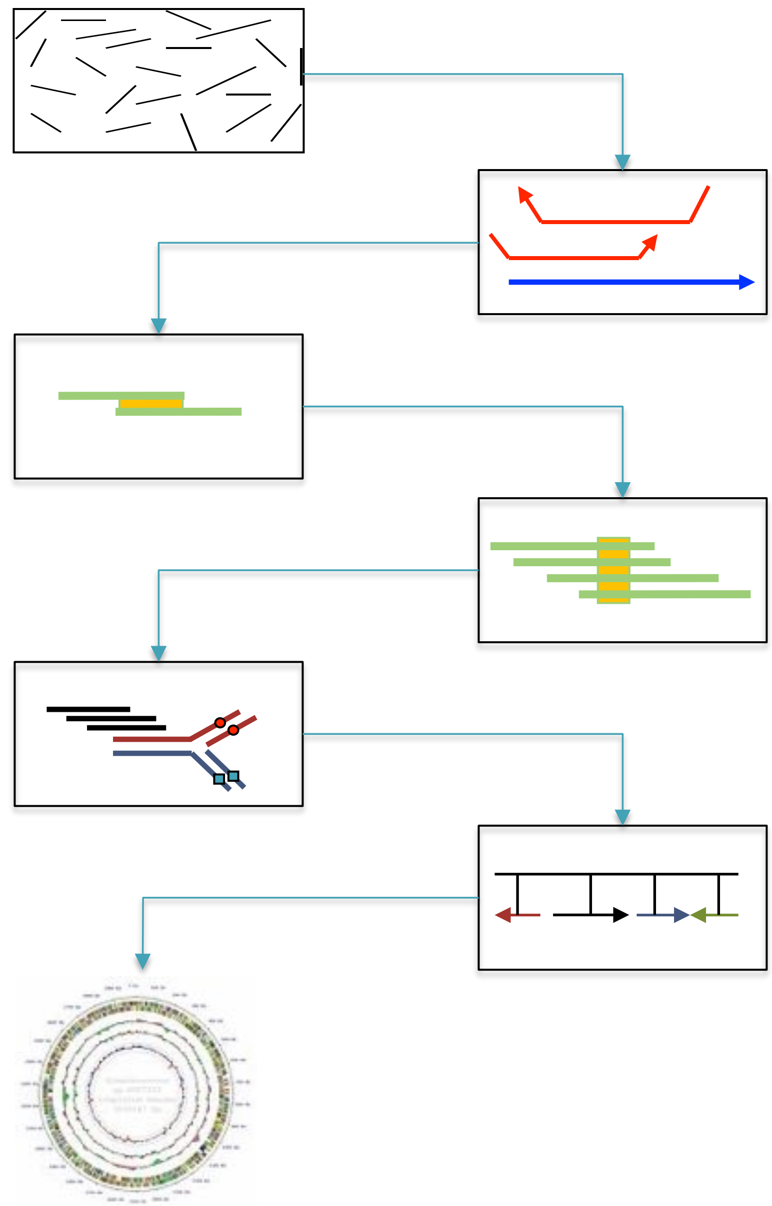


Genome assembly with the Celera Assembler

Celera Assembler

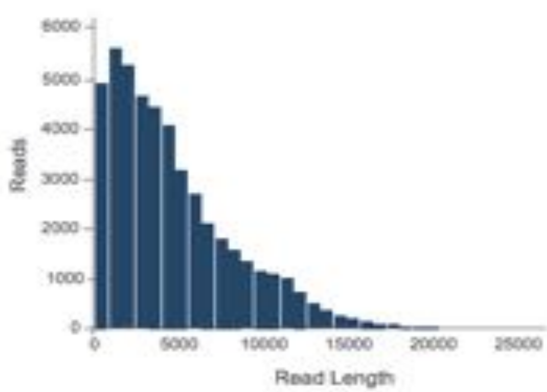
http://wgs-assembler.sf.net

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

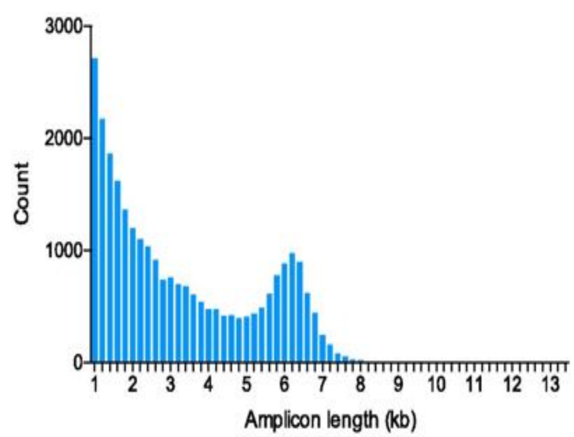


Single Molecule Sequencing Technology

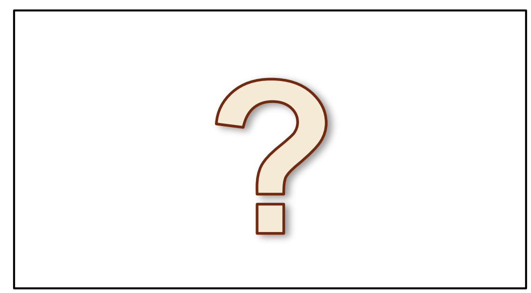
PacBio RS II



Moleculo

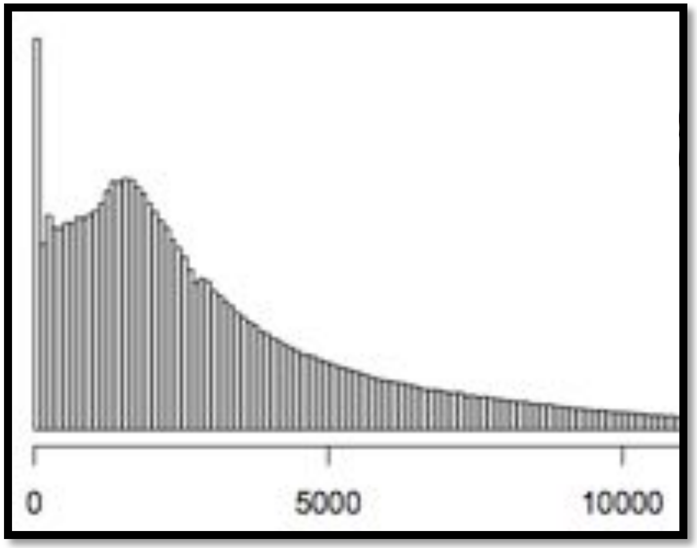


Oxford Nanopore



Clive G. Brown @Clive_G_Brown 9 Oct
I've reluctantly rejoined twitter purely so that I can make one tweet - when the appropriate time arises ...
Expand

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

PacBio Error Correction: HGAP

- With 50-100x of Pacbio coverage, virtually all of the errors can be eliminated
  - Works well for Microbial genomes: single contig per chromosome routinely achieved
  - Difficult to scale up for use with eukaryotic genomes

Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data
Chin, CS *et al*. (2013) *Nature Methods*. 10: 563-569

Hybrid Sequencing

Illumina
Sequencing by Synthesis

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

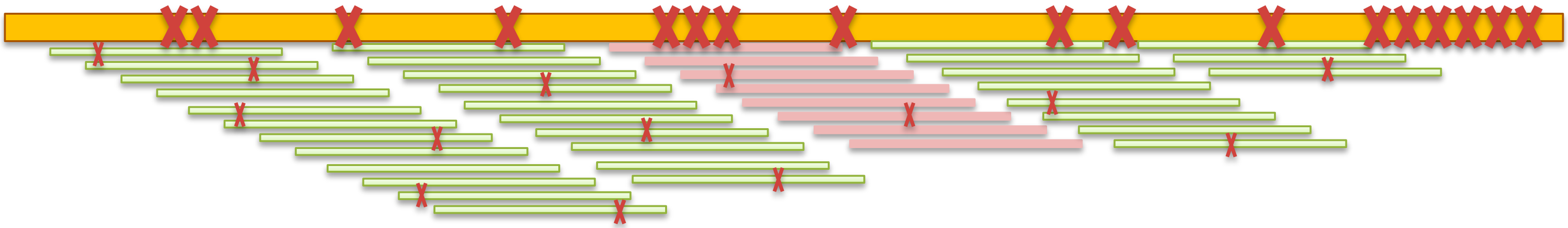
Pacific Biosciences
SMRT Sequencing

Lower throughput (1Gbp/day)
Lower accuracy (~85%)
Long reads (5kbp+)

Hybrid Error Correction: PacBioToCA

http://wgs-assembler.sf.net

1. Correction Pipeline
    1. Map short reads to long reads
    2. Trim long reads at coverage gaps
    3. Compute consensus for each long read
2. Error corrected reads can be easily assembled, aligned

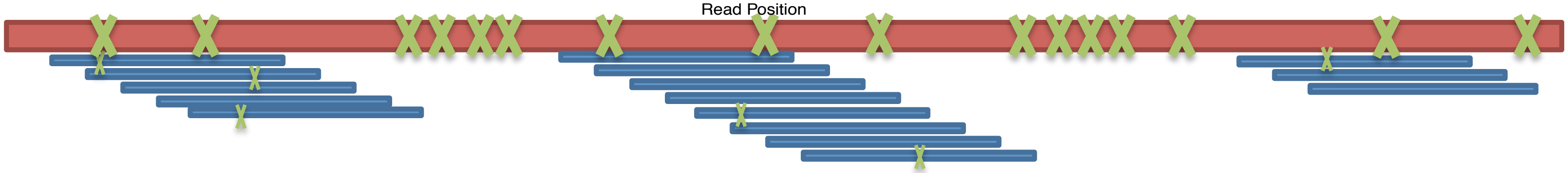


Hybrid error correction and de novo assembly of single-molecule sequencing reads.
Koren, S, Schatz, MC, *et al.* (2012) *Nature Biotechnology*. doi:10.1038/nbt.2280

Enhanced PacBio Error Correction

PacBioToCA fails in complex regions

1. Simple Repeats – Kmer Frequency Too High to Seed Overlaps
2. Error Dense Regions – Difficult to compute overlaps with many errors
3. Extreme GC – Lacks Illumina Coverage

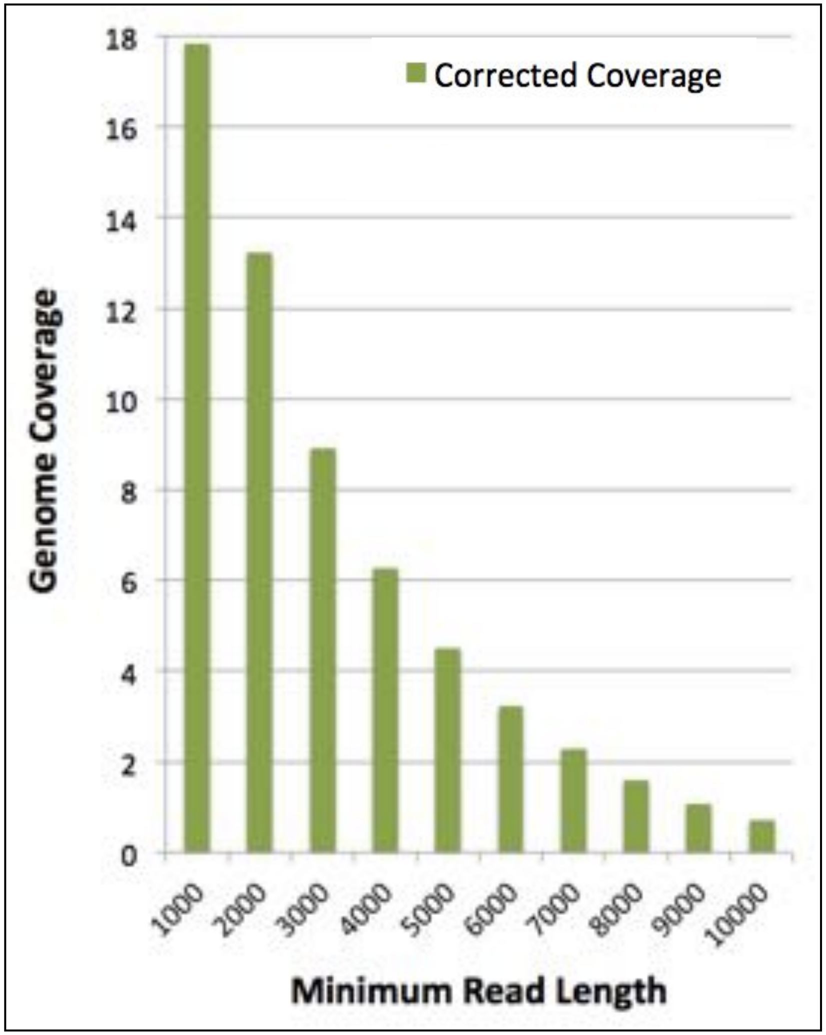


Assembly complexity of long read sequencing

Lee, H*, Gurtowski, J*, Yoo, S, Marcus, S, McCombie, WR, Schatz MC *et al.* (2013) *In preparation*

Preliminary Rice Assemblies

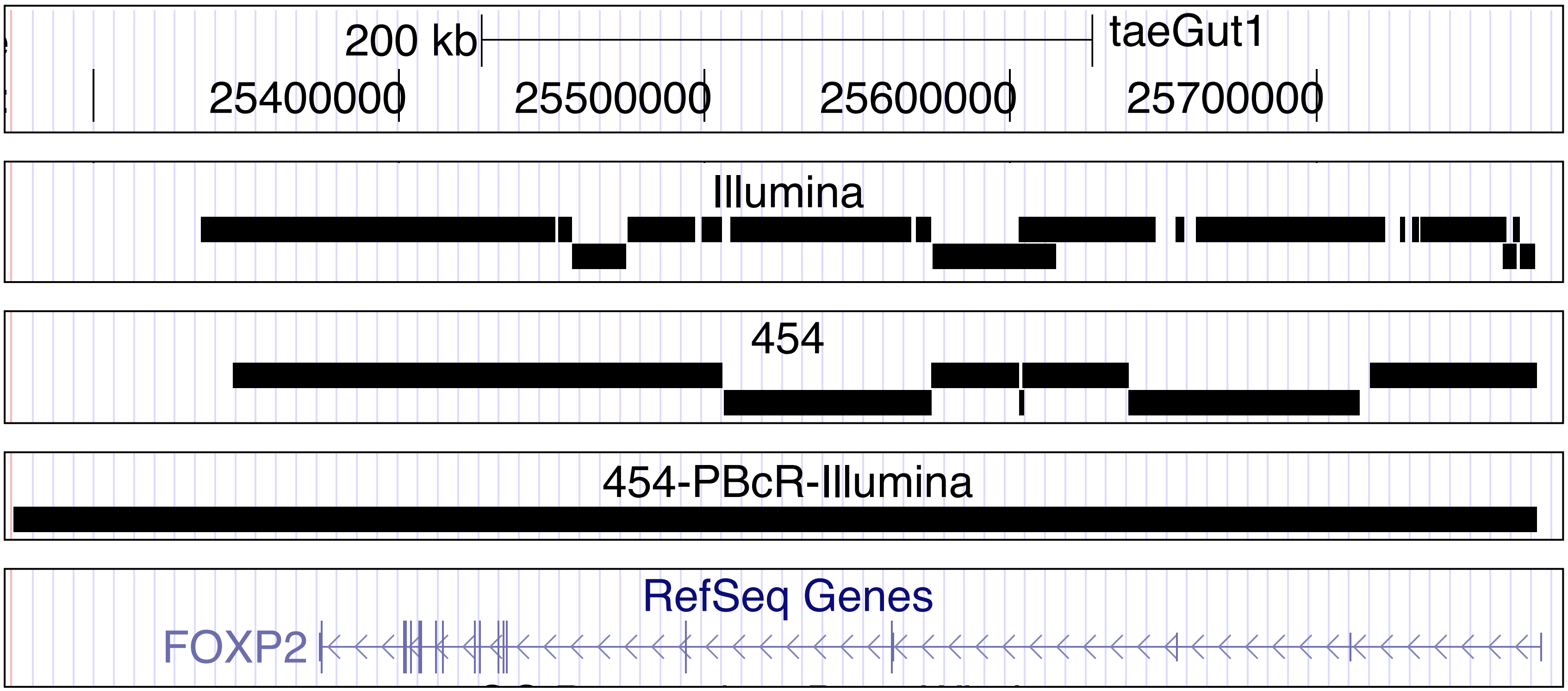
Assembly	Contig NG50
HiSeq Fragments 50x 2x100bp @ 180	3,925
MiSeq Fragments 23x 459bp 8x 2x251bp @ 450	6,332
"ALLPATHS-recipe" 50x 2x100bp @ 180 36x 2x50bp @ 2100 51x 2x50bp @ 4800	18,248



In collaboration with McCombie & Ware labs @ CSHL

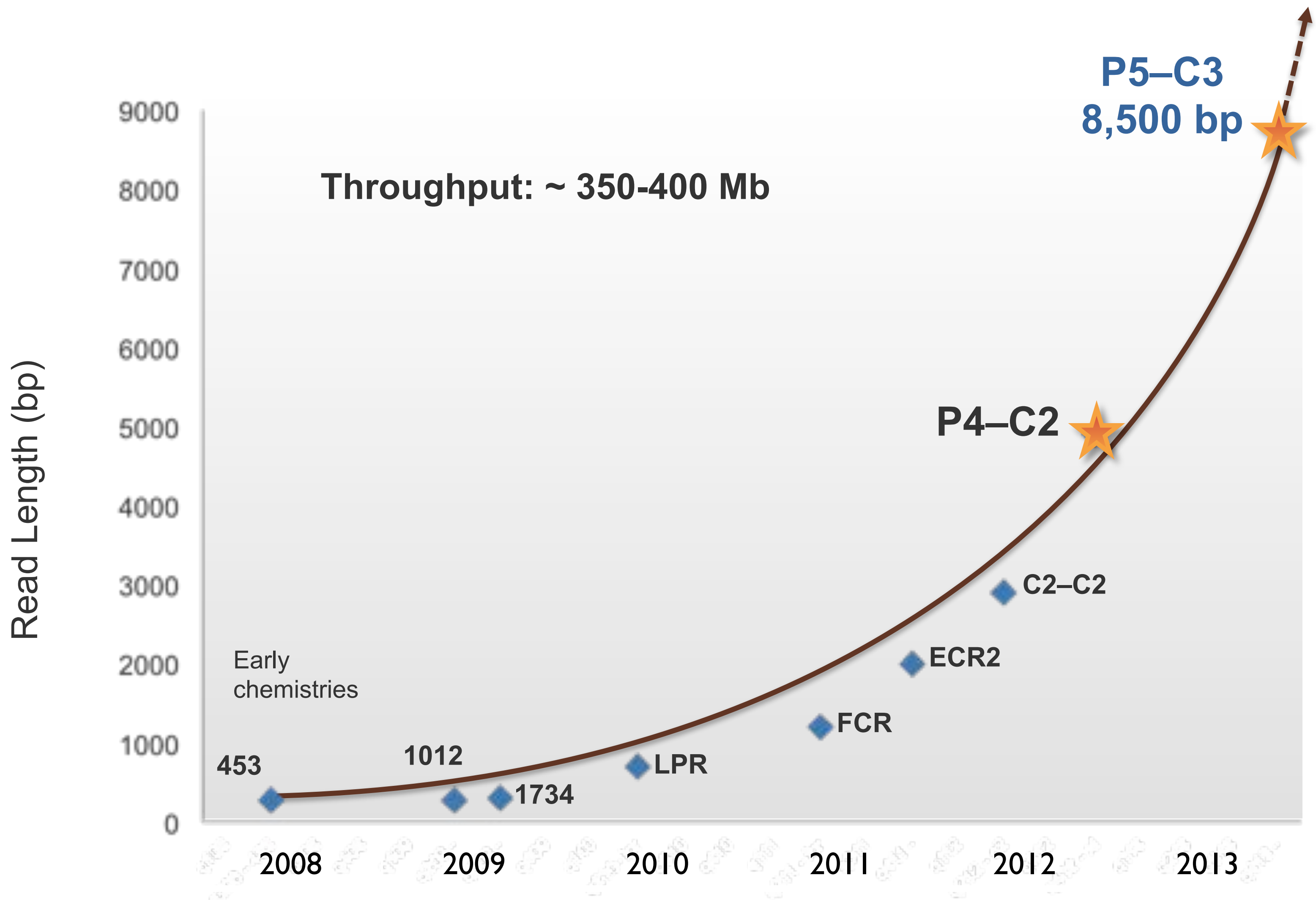
Improved Gene Reconstruction

FOXP2 assembled in a single contig in the PacBio parrot assembly



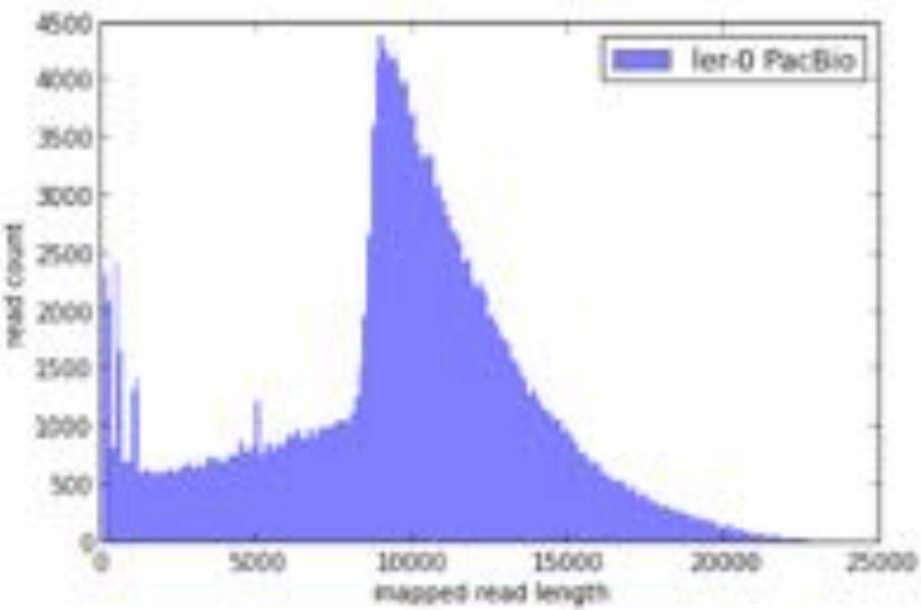
Hybrid error correction and de novo assembly of single-molecule sequencing reads.
Koren, S, Schatz, MC, *et al.* (2012) *Nature Biotechnology*. doi:10.1038/nbt.2280

P5-C3 Chemistry Read Lengths



De novo assembly of Arabidopsis

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

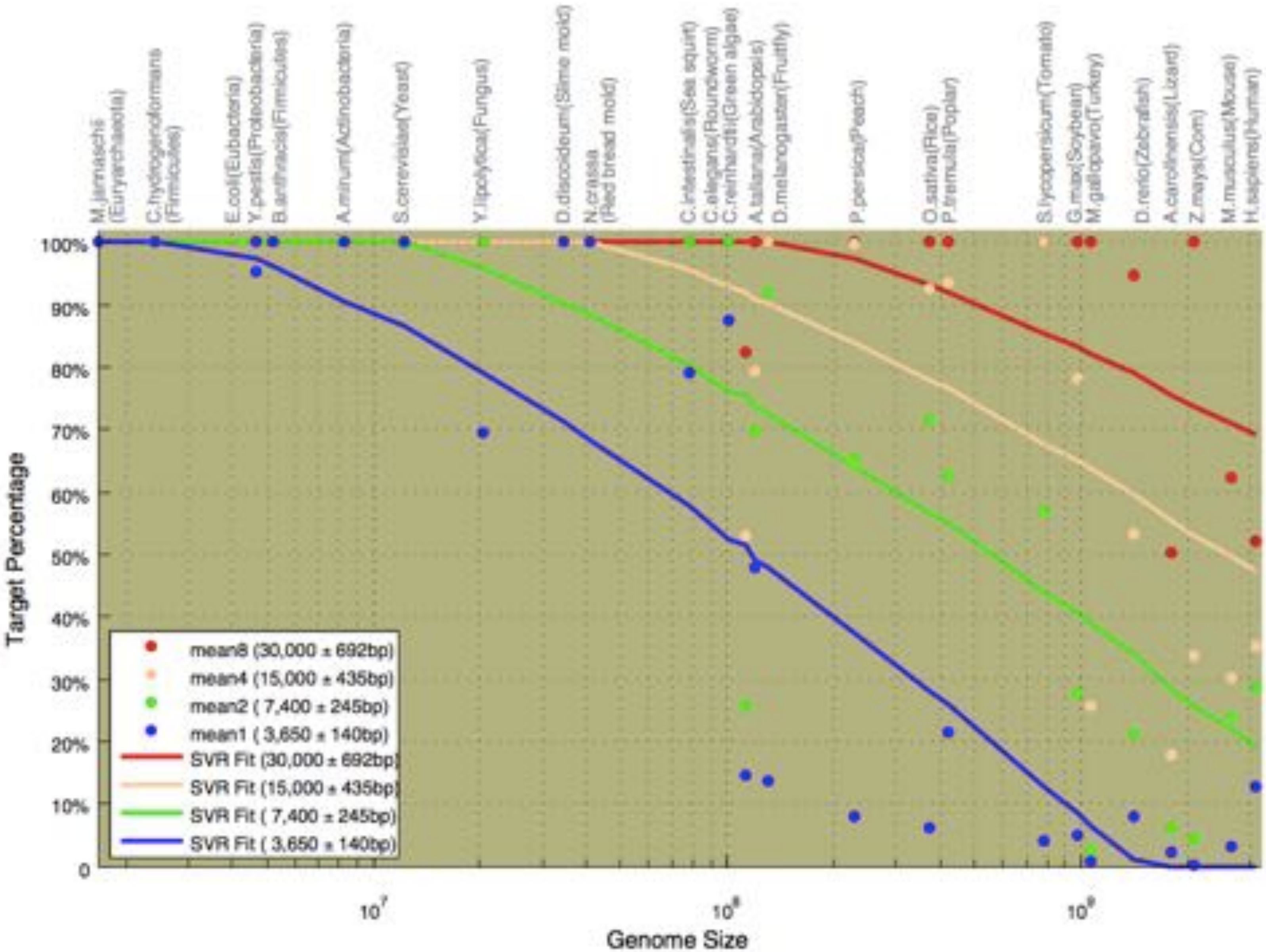


A. thaliana Ler-0 sequenced at PacBio

- Sequenced using the latest 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 >100x

Genome size:	124.6 Mb	Sum of Contig Lengths:	149.5Mb
GC content:	33.92%	Number of Contigs:	1788
Raw data:	11 Gb	Max Contig Length:	12.4 Mb
Assembly coverage:	15x over 9kbp	N50 Contig Length:	8.4 Mb

Assembly Complexity of Long Reads



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
  - Reads

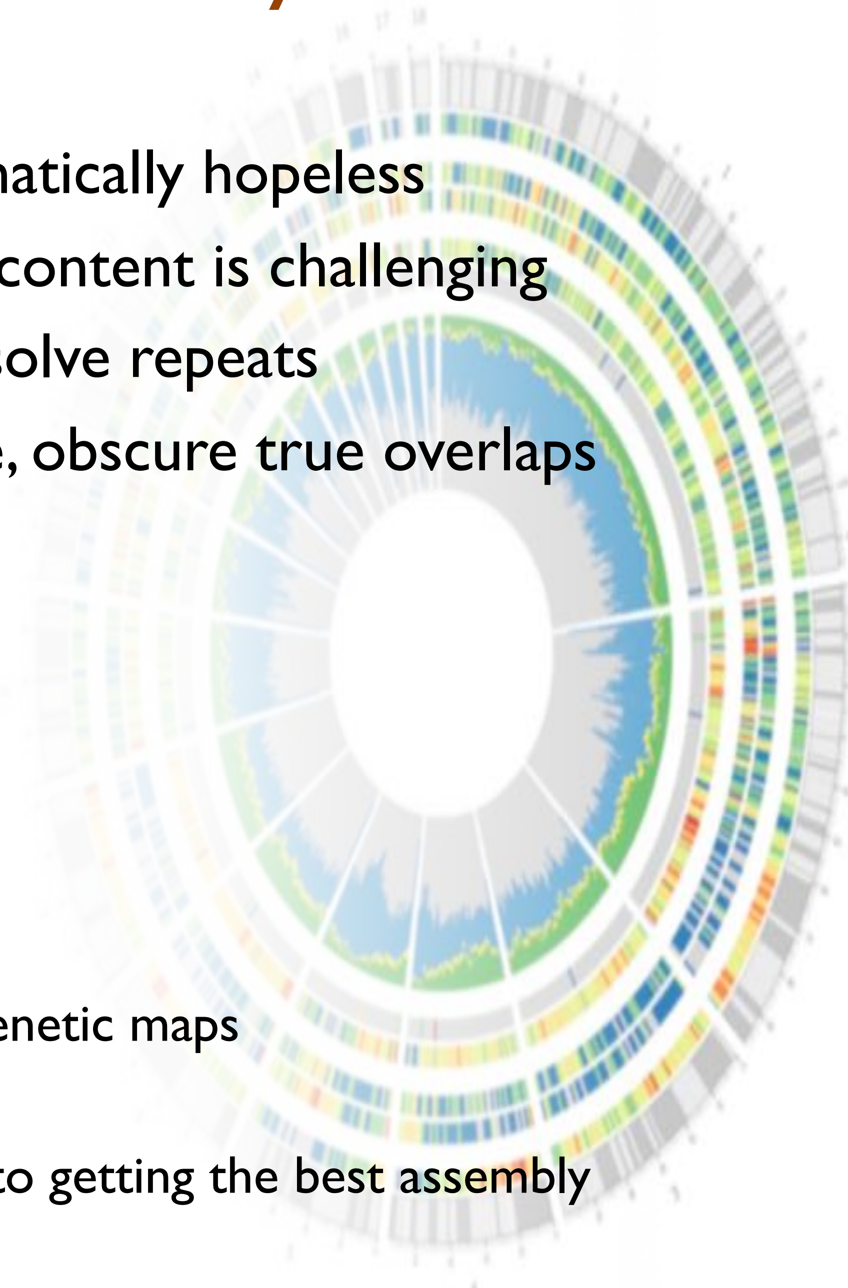
    -> unitigs

    -> mates

    -> scaffolds

    -> optical / physical / genetic maps

    -> chromosomes
  - Extensive error correction is the key to getting the best assembly possible from a given data set



Outline

1. *-seq review

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

3. Genome assemblers
   1. ALLPATHS-LG
   2. Celera Assembler

4. Whole Genome Alignment with MUMmer

Whole Genome Alignment with MUMmer

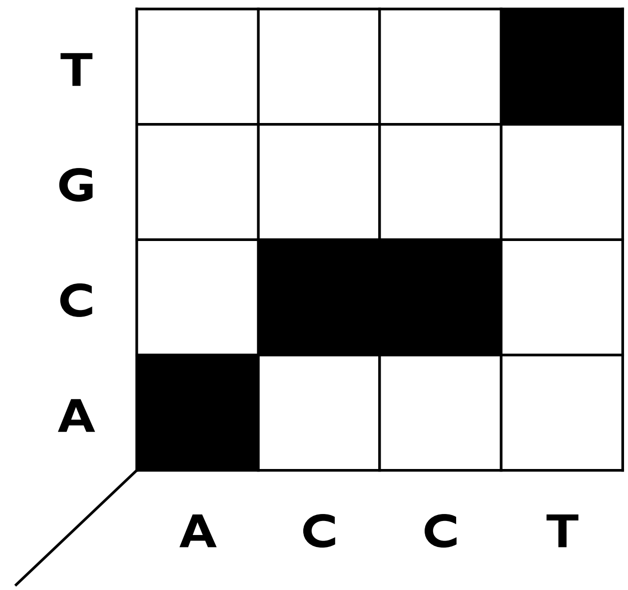
Slides Courtesy of Adam M. Phillippy

amp@umics.umd.edu

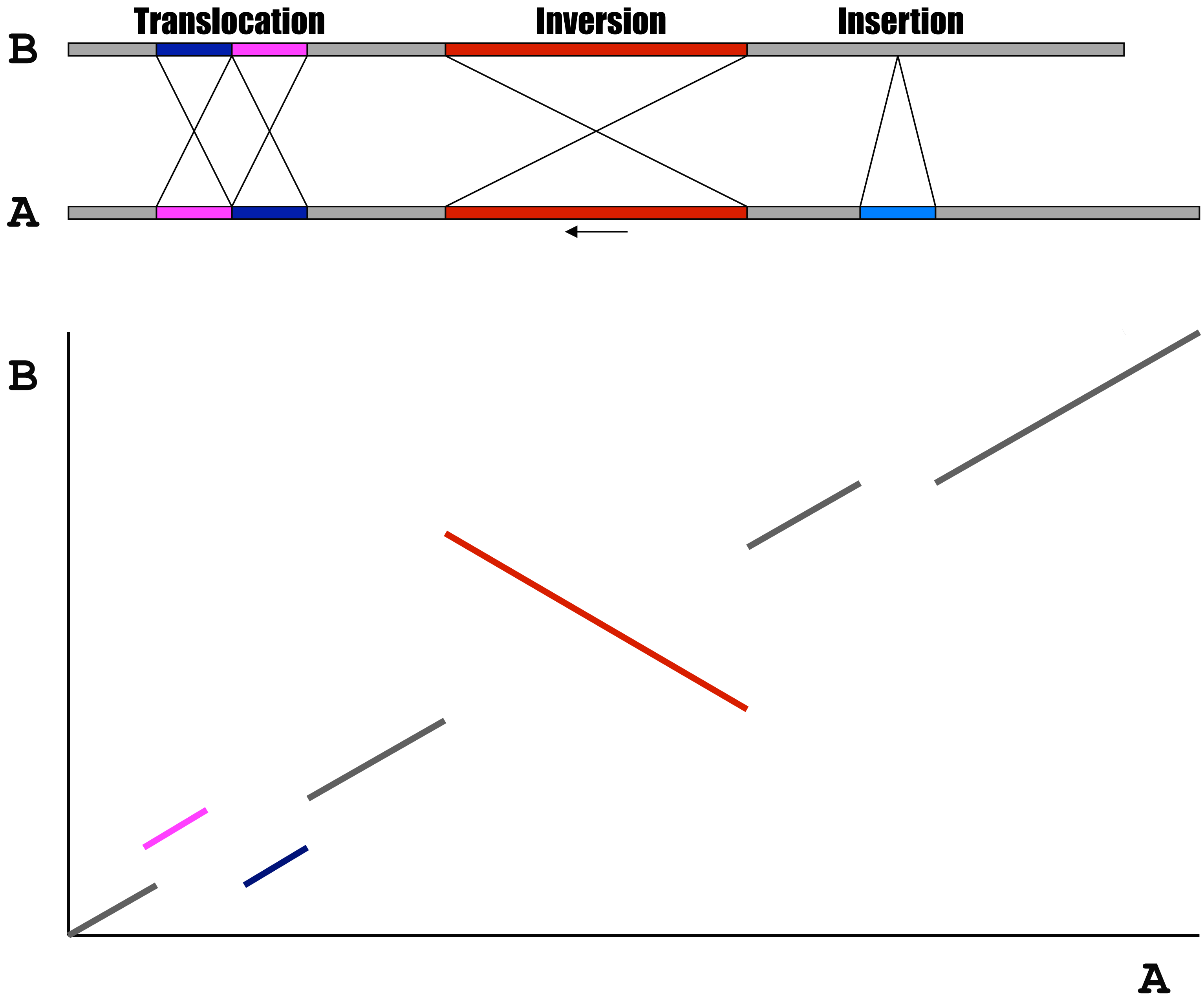
WGA visualization

- How can we visualize *whole* genome alignments?

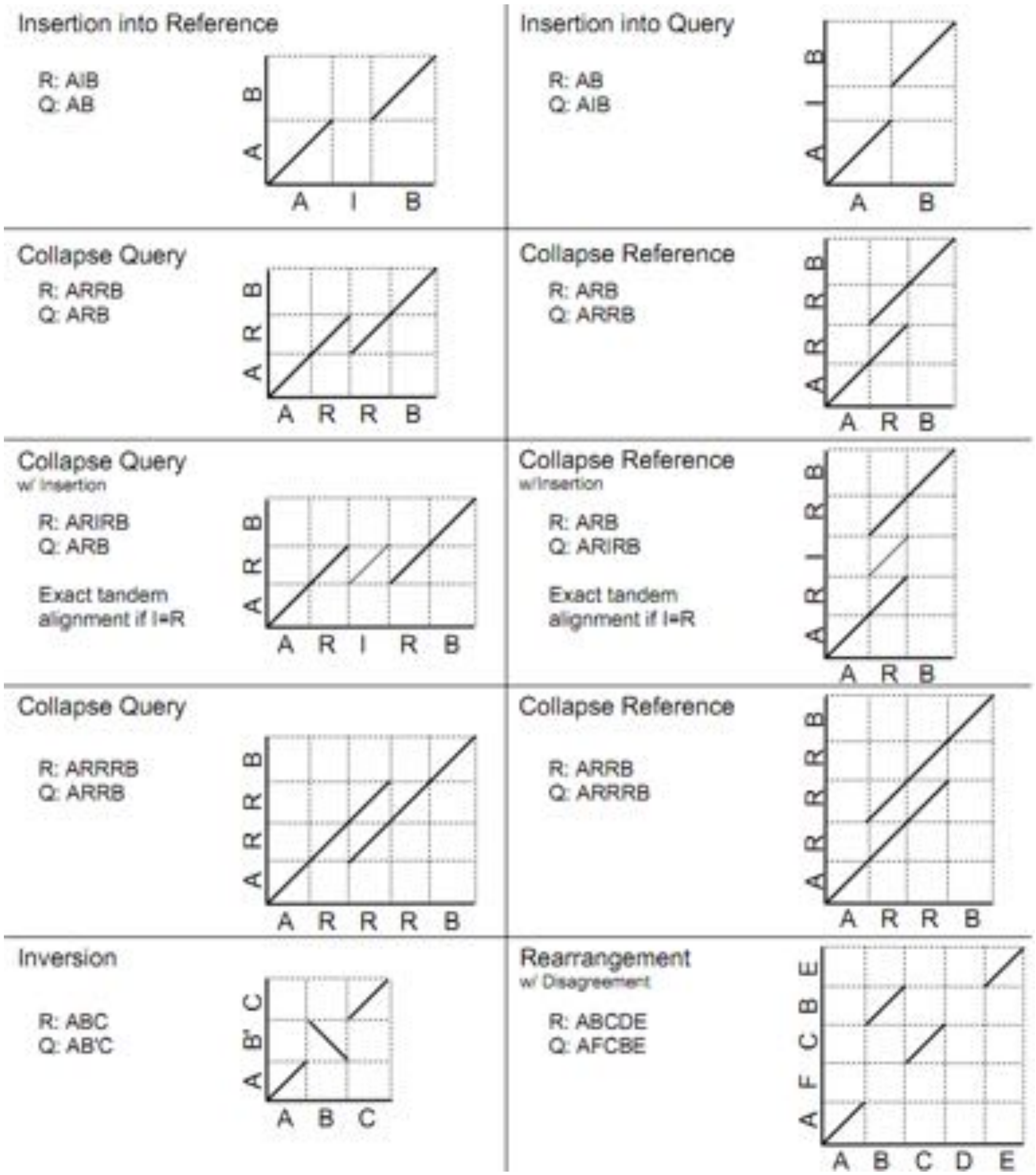
- With an alignment dot plot
  - *N* x *M* matrix
    - Let i = position in genome A
    - Let j = position in genome B
    - Fill cell (i,j) if A_i shows similarity to B_j



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


Translocation
Inversion
Insertion
B
A
B
A

SV Types



- 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

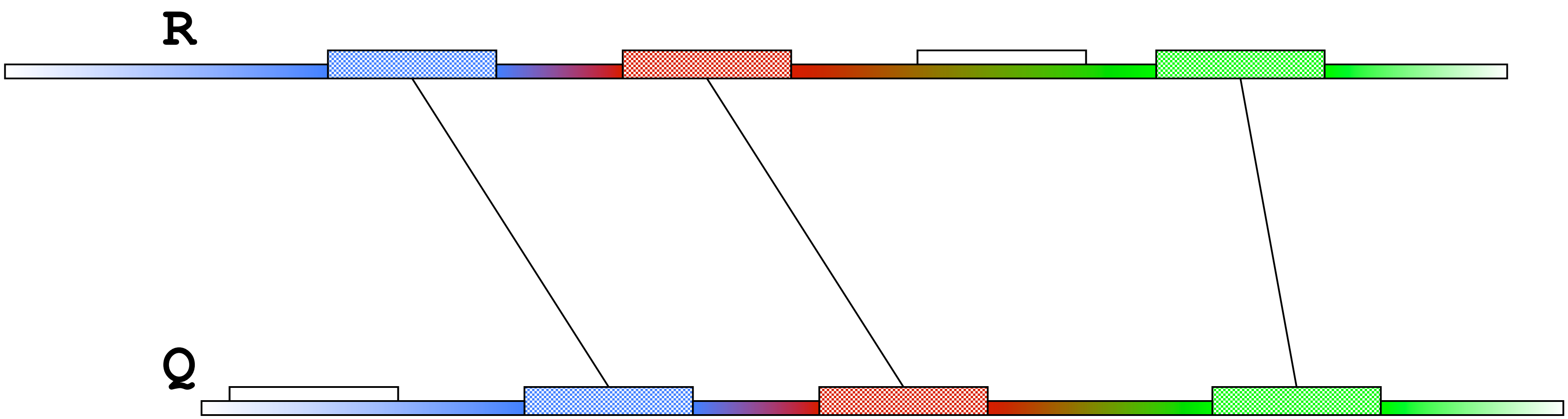
Seed and Extend

visualization

FIND all MUMs

CLUSTER consistent MUMs

EXTEND alignments



WGA example with **nucmer**

- *Yersina pestis CO92* vs. *Yersina pestis KIM*
  - High nucleotide similarity, 99.86%
    - Two strains of the same species
  - Extensive genome shuffling
    - Global alignment will not work
  - Highly repetitive
    - Many local alignments

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
```

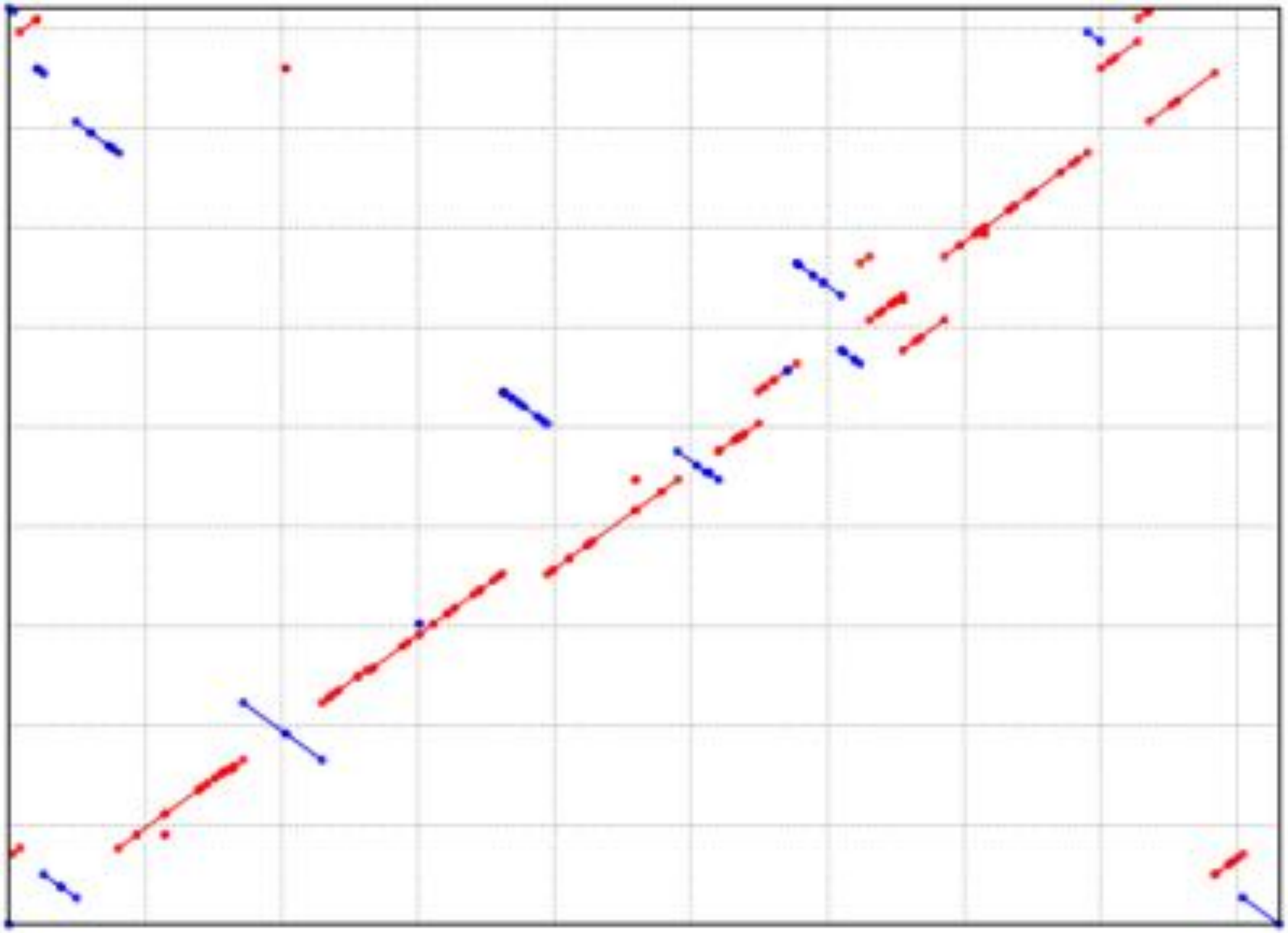
`dnadiff out.delta.m`

```
Construct catalog of sequence variations
```

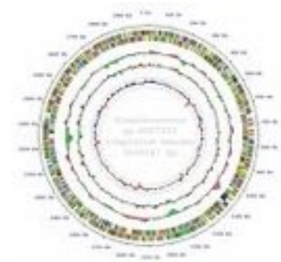
`mummerplot --large --layout out.delta.m`

```
--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



Resources



- Assembly Competitions
  - Assemblathon: http://assemblathon.org/
  - GAGE: http://gage.cbcb.umd.edu/

- Assembler Websites:
  - ALLPATHS-LG: http://www.broadinstitute.org/software/allpaths-lg/blog/
  - SOAPdenovo: http://soap.genomics.org.cn/soapdenovo.html
  - Celera Assembler: http://wgs-assembler.sf.net

- Tools:
  - MUMmer: http://mummer.sourceforge.net/
  - Quake: http://www.cbcb.umd.edu/software/quake/
  - AMOS: http://amos.sf.net

Acknowledgements

Schatz Lab
Giuseppe Narzisi
Shoshana Marcus
James Gurtowski
Srividya Ramakrishnan
Hayan Lee
Rob Aboukhalil
Mitch Bekritsky
Charles Underwood
Tyler Gavin
Alejandro Wences
Greg Vurture
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