Whole Genome Assembly and Alignment

Michael Schatz

Oct 25, 2012

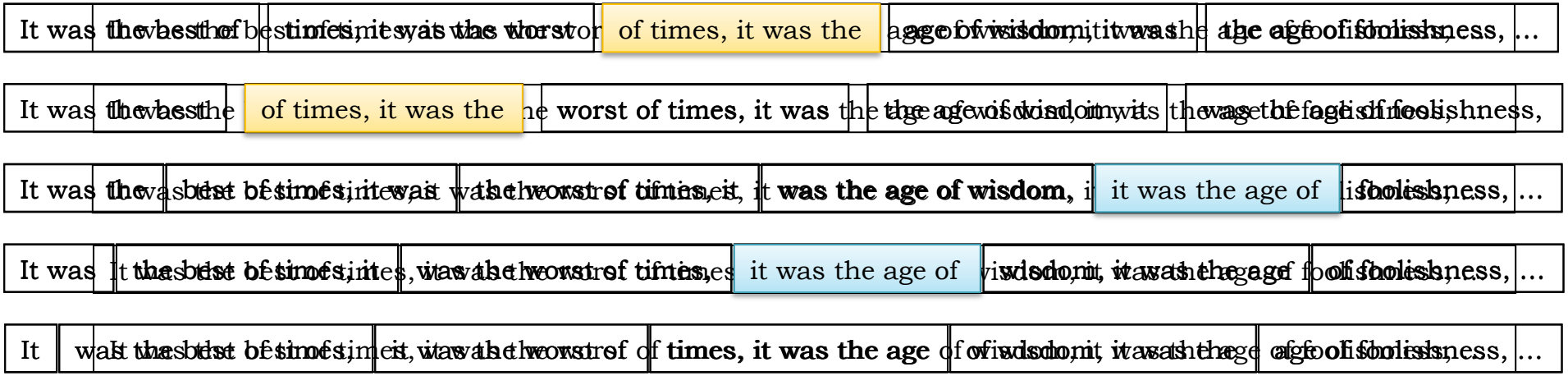
CSHL Sequencing Course

Outline

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

Shredded Book Reconstruction

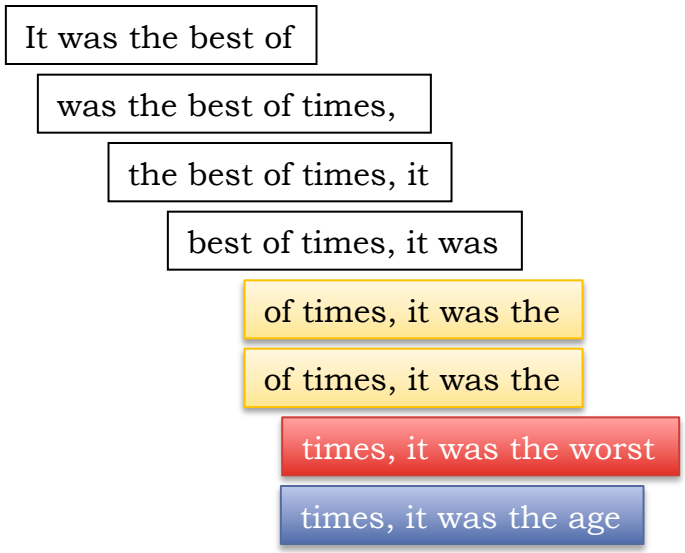
- Dickens accidentally shreds the first printing of A Tale of Two Cities
  - Text printed on 5 long spools



- 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



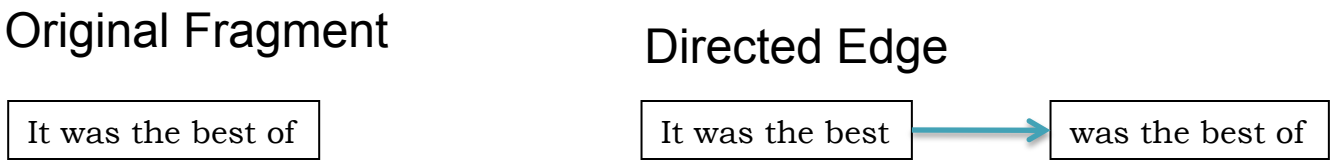
The repeated sequence make the correct reconstruction ambiguous

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

Model the assembly problem as a graph problem

de Bruijn Graph Construction

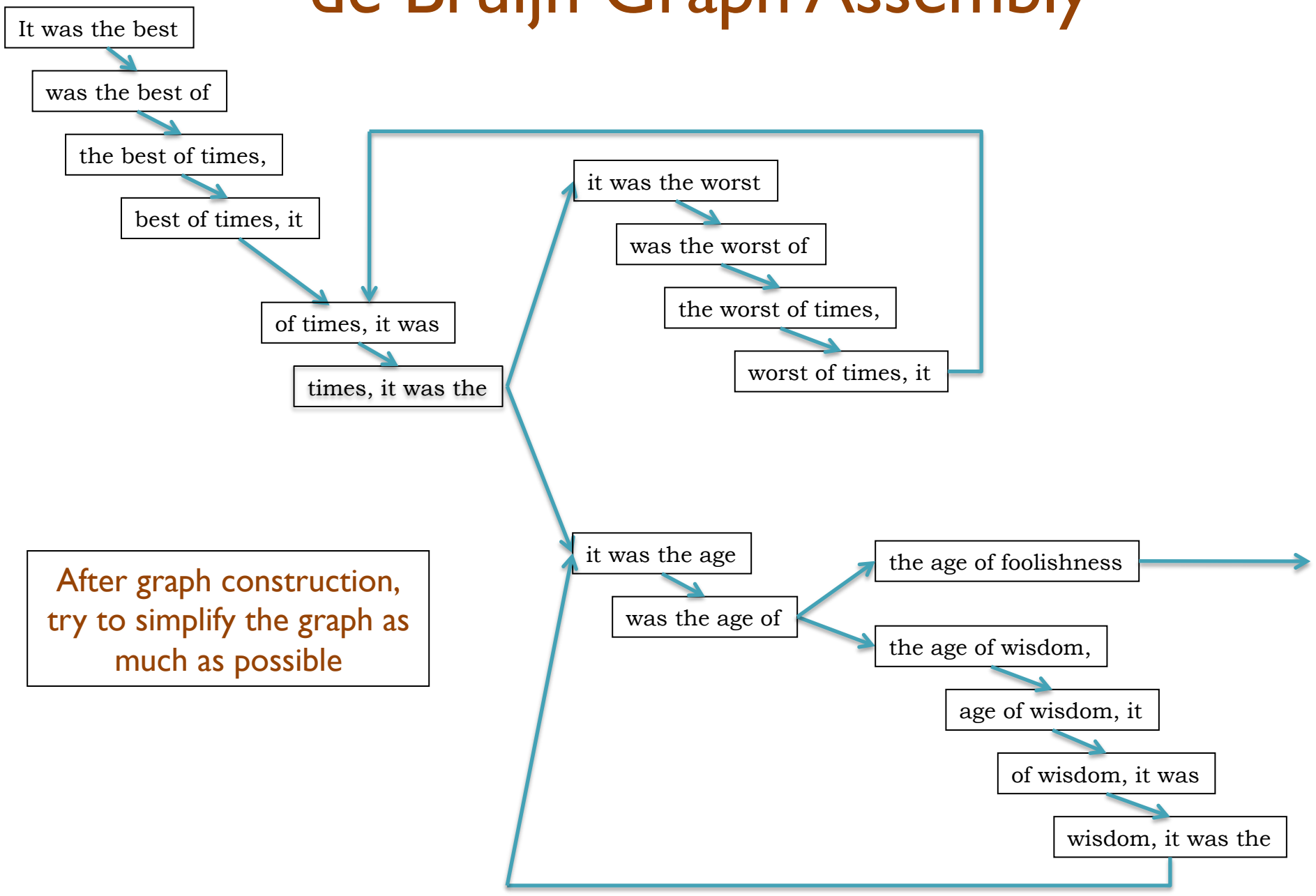
- $D_k = (V,E)$
  - V = All length-k subfragments (k < l)
  - E = Directed edges between consecutive subfragments
    - Nodes overlap by k-1 words



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

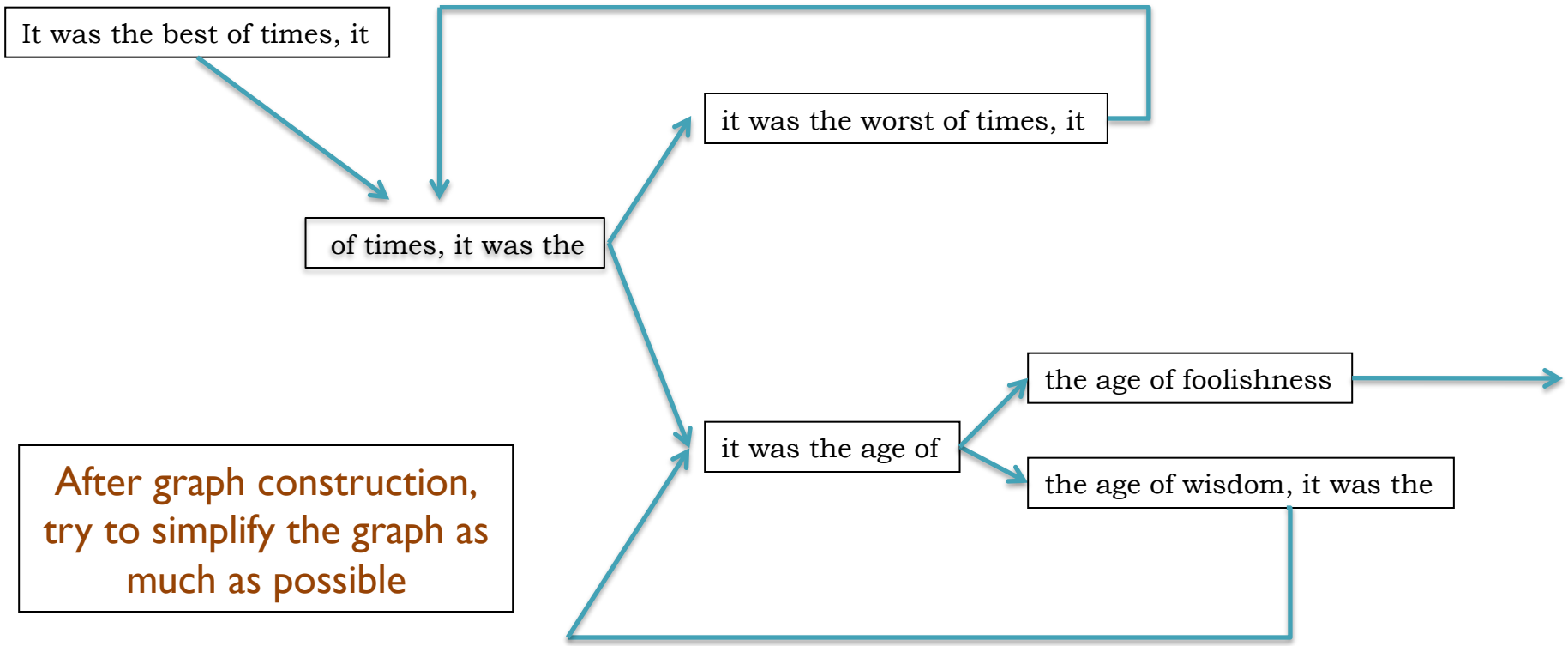
de Bruijn, 1946
Idury and Waterman, 1995
Pevzner, Tang, Waterman, 2001

de Bruijn Graph Assembly



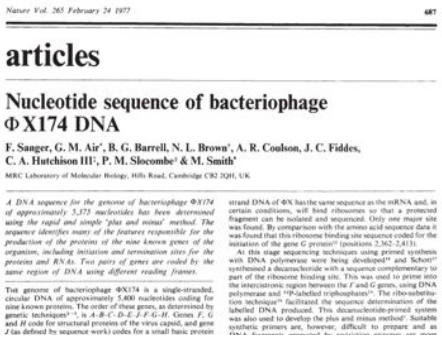
After graph construction, try to simplify the graph as much as possible

de Bruijn Graph Assembly



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

Milestones in Genome Assembly



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

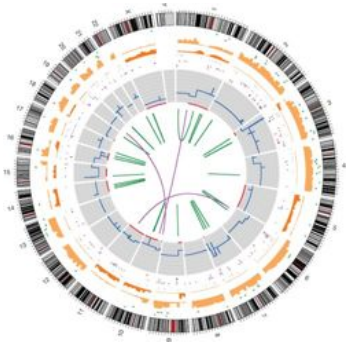
Like Dickens, we must computationally reconstruct a genome from short fragments

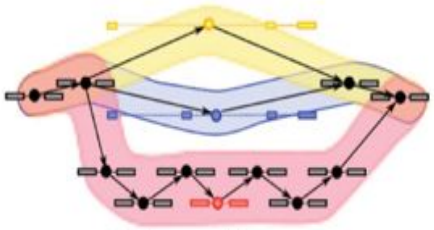
Assembly Applications

- Novel genomes

- Metagenomes

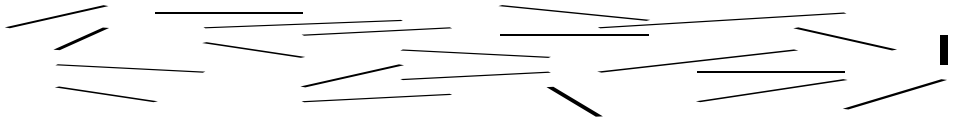
- Sequencing assays
  - Structural variations
  - Transcript assembly
  - …





Assembling a Genome

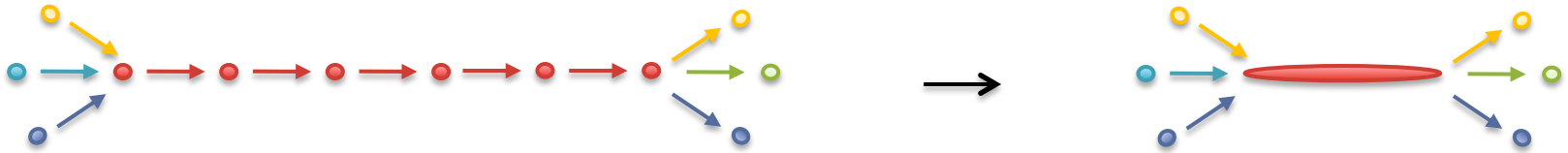
1. Shear & Sequence DNA



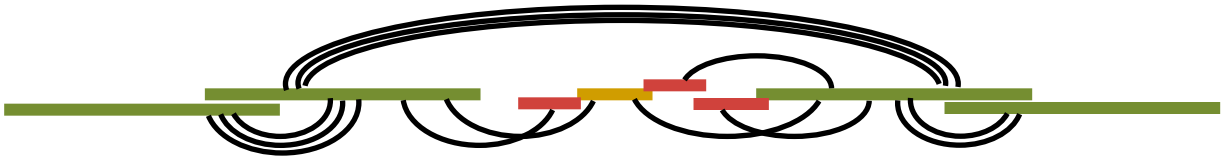
2. Construct assembly graph from overlapping reads

...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

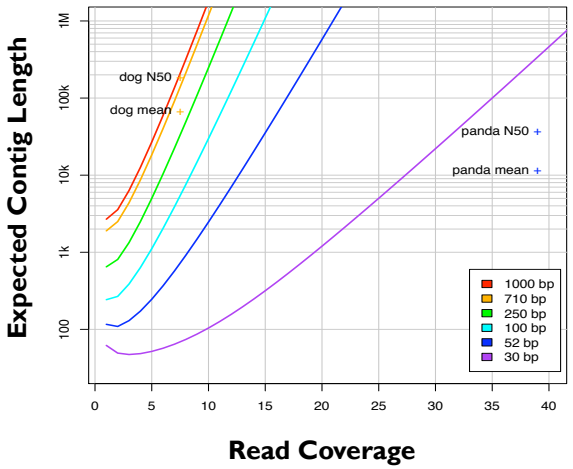
2. ***Sequencing***:
   - (Very) large genomes, imperfect sequencing

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

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

Ingredients for a good assembly

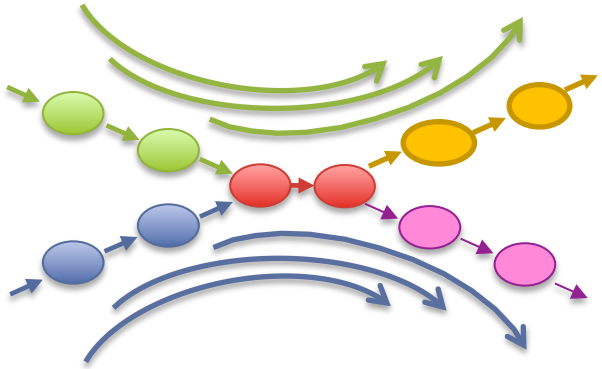
Coverage



High coverage is required

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

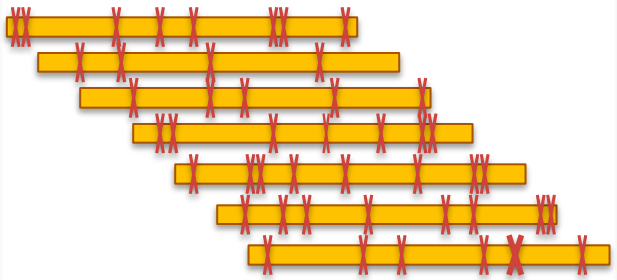
Read Length



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

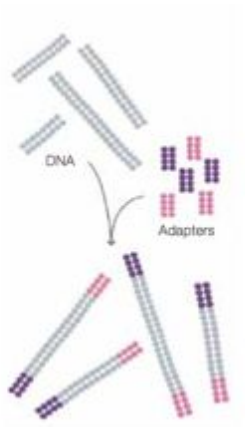


Errors obscure overlaps

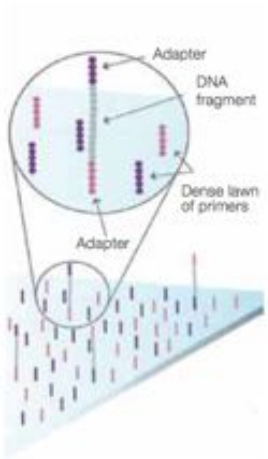
- Reads are assembled by finding kmers shared in pair of reads
- High error rate requires very short seeds, increasing complexity and forming assembly hairballs

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

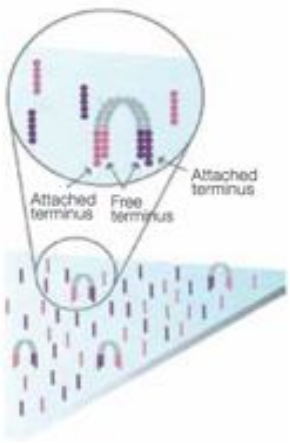
Illumina Sequencing by Synthesis



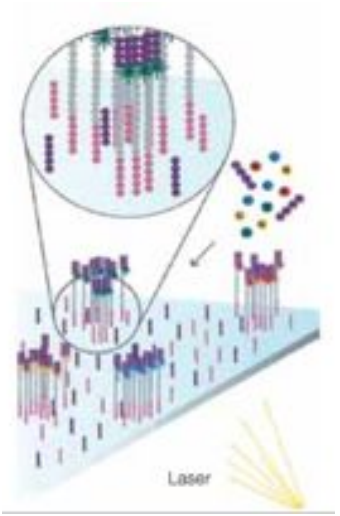
1. Prepare



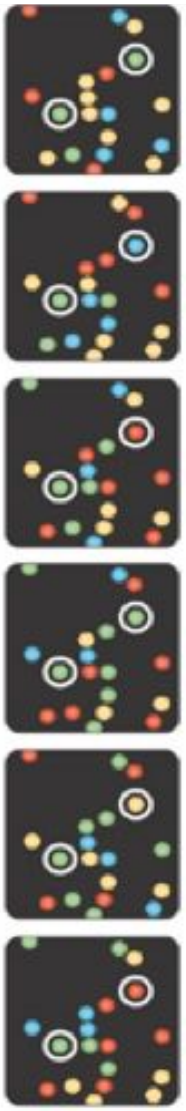
2. Attach



3. Amplify



4. Image



5. Basecall

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

http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

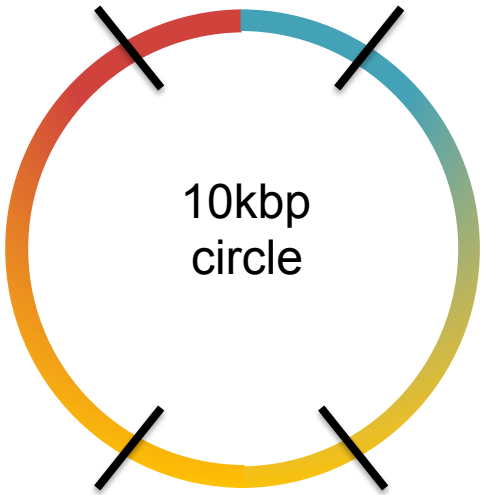
Paired-end and Mate-pairs

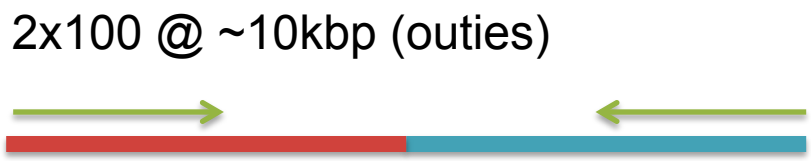
Paired-end sequencing

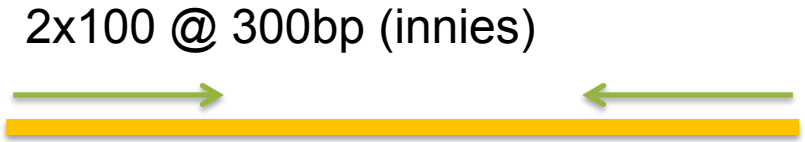
- Read one end of the molecule, flip, and read the other end
- Generate pair of reads separated by up to 500bp with inward orientation

Mate-pair sequencing

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

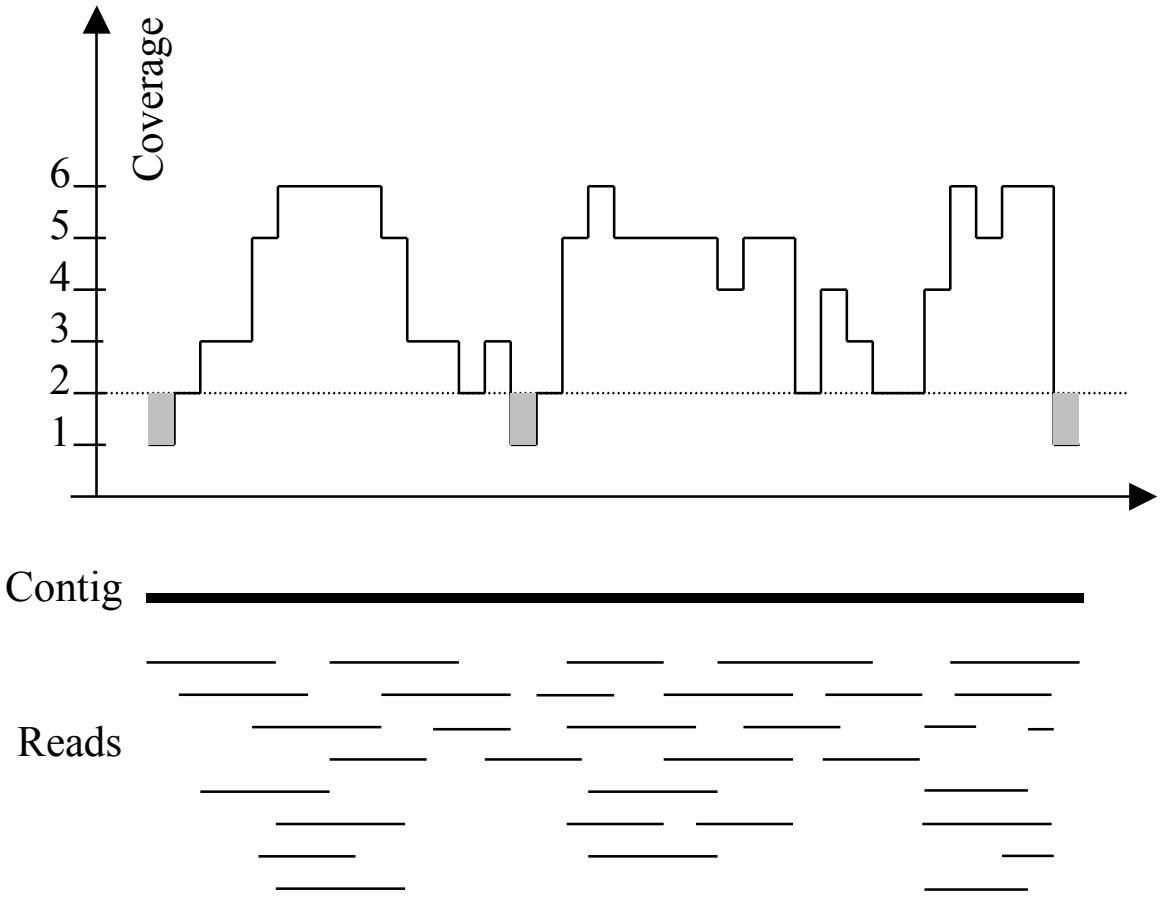






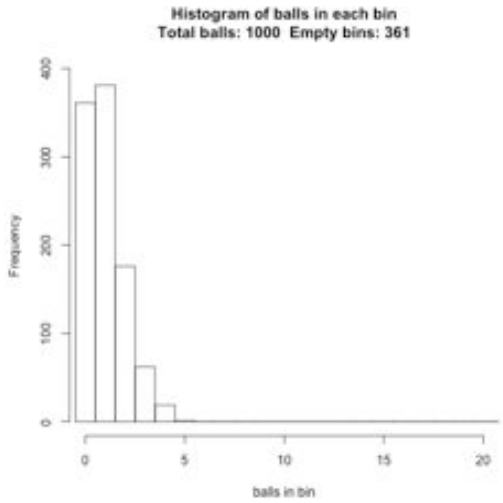
Coverage

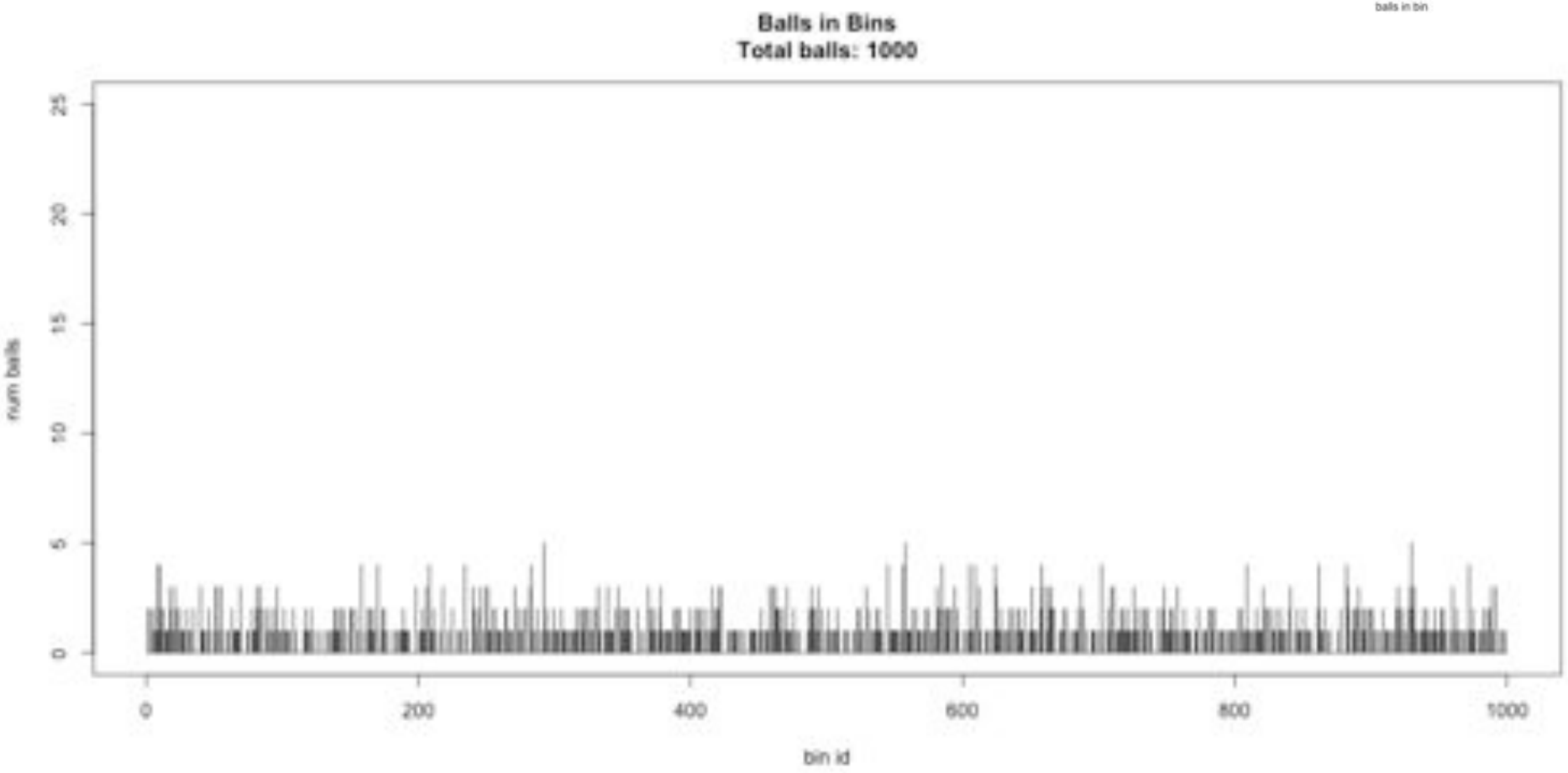
Typical contig coverage



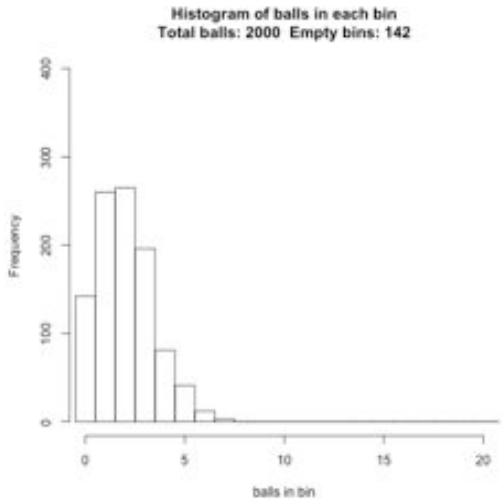
Imagine raindrops on a sidewalk

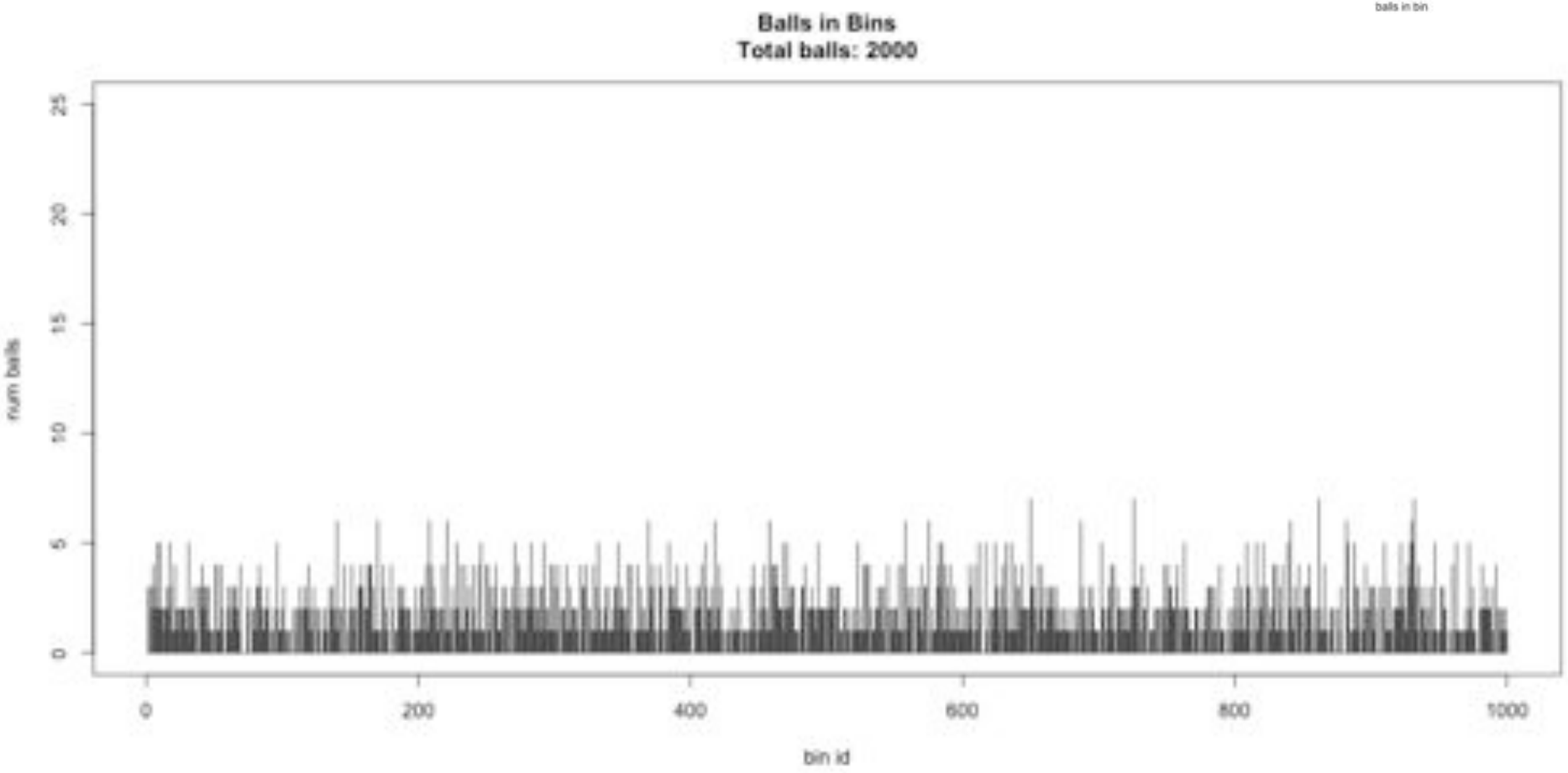
Balls in Bins 1x



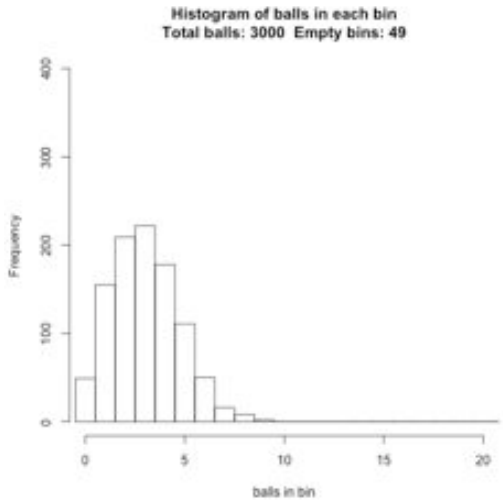


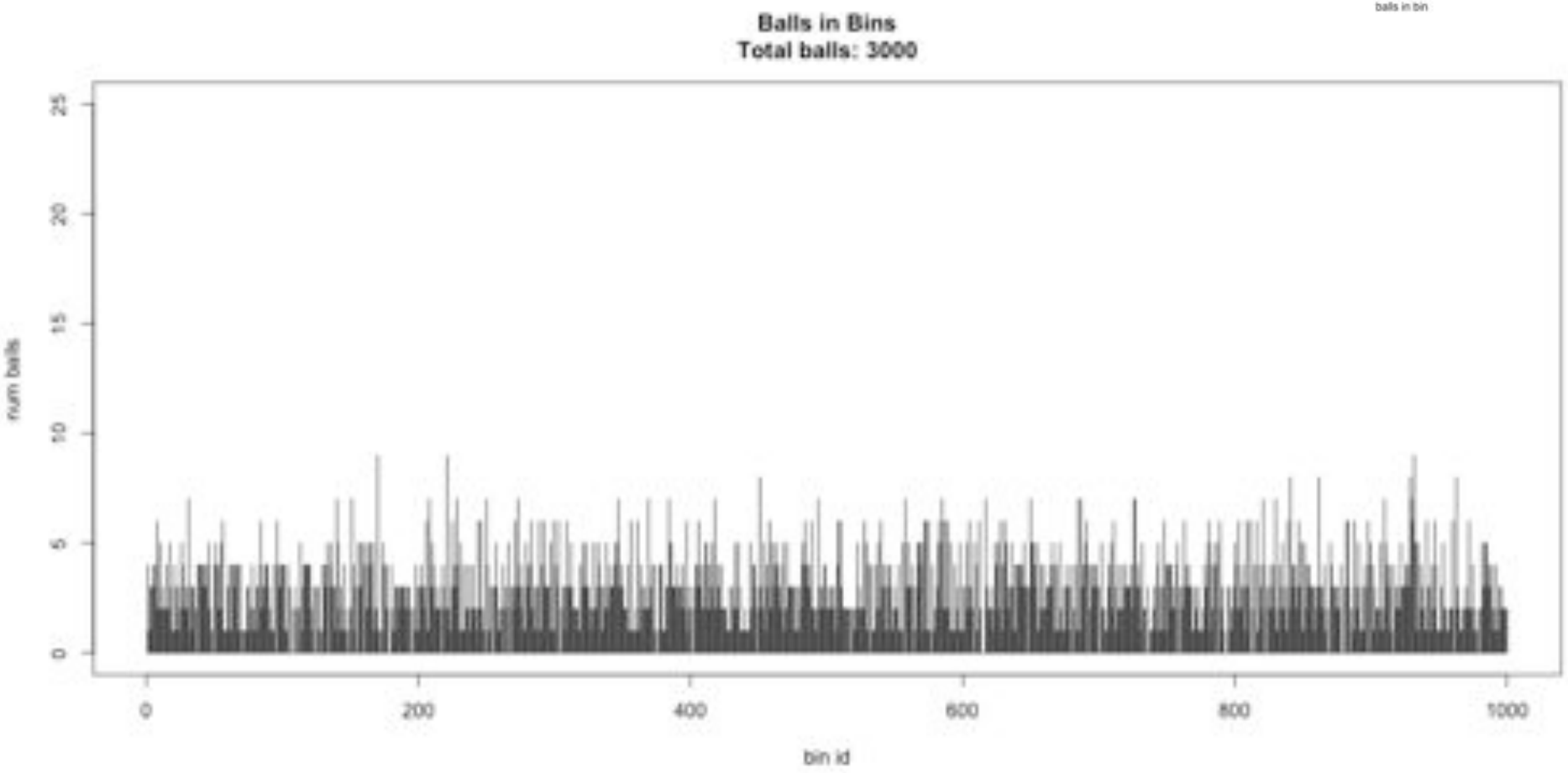
Balls in Bins 2x



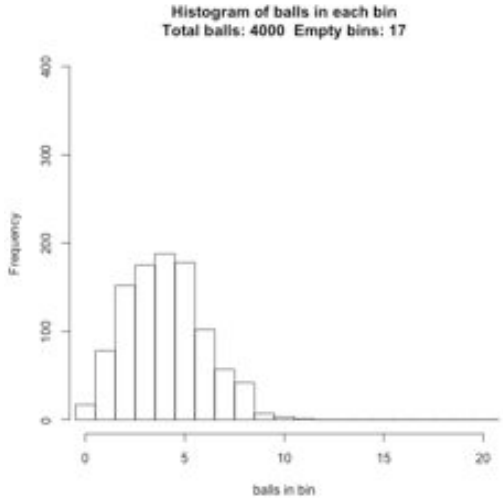


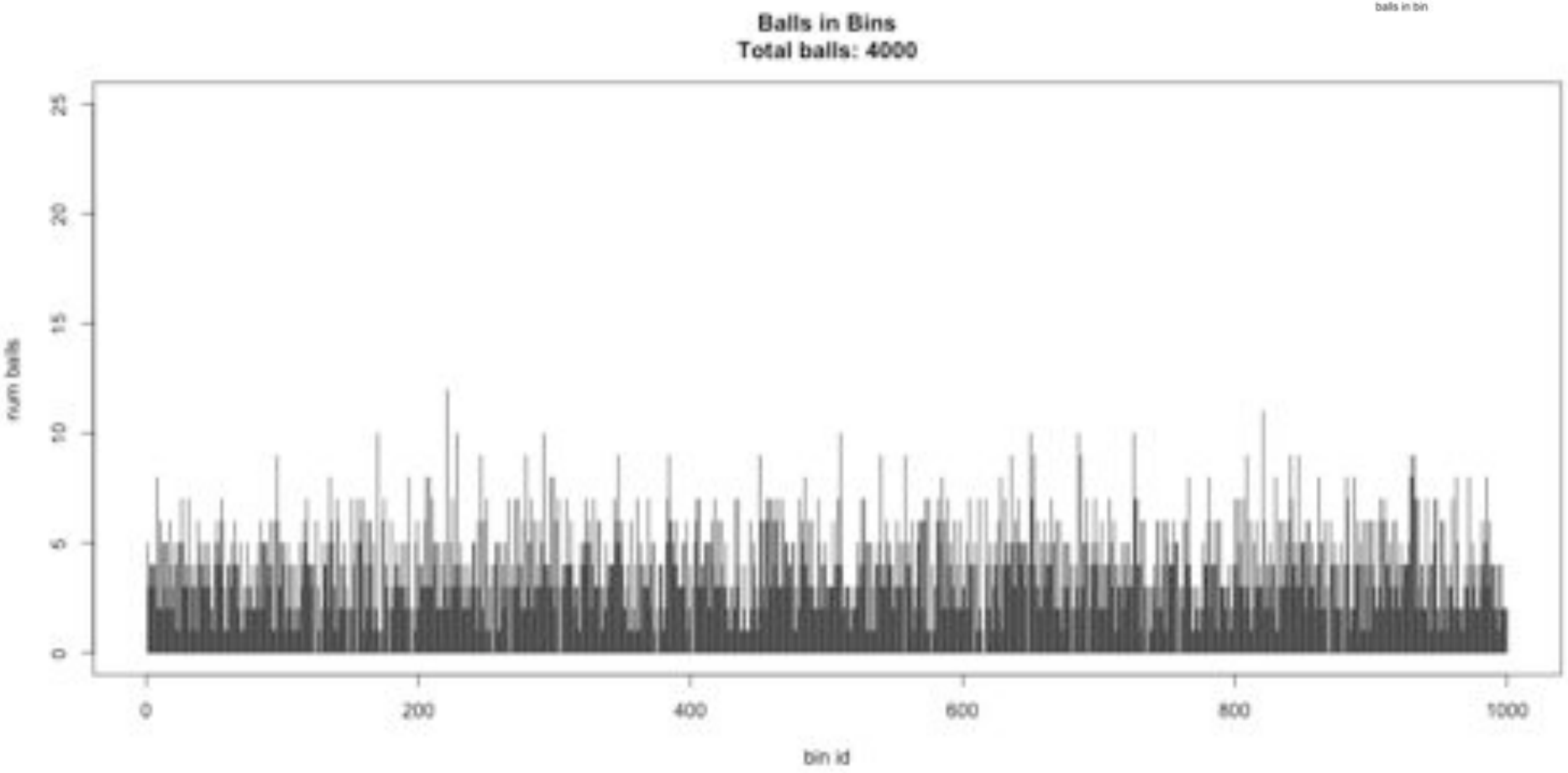
Balls in Bins 3x



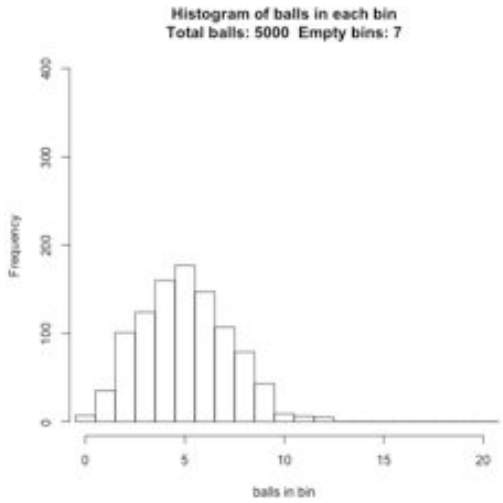


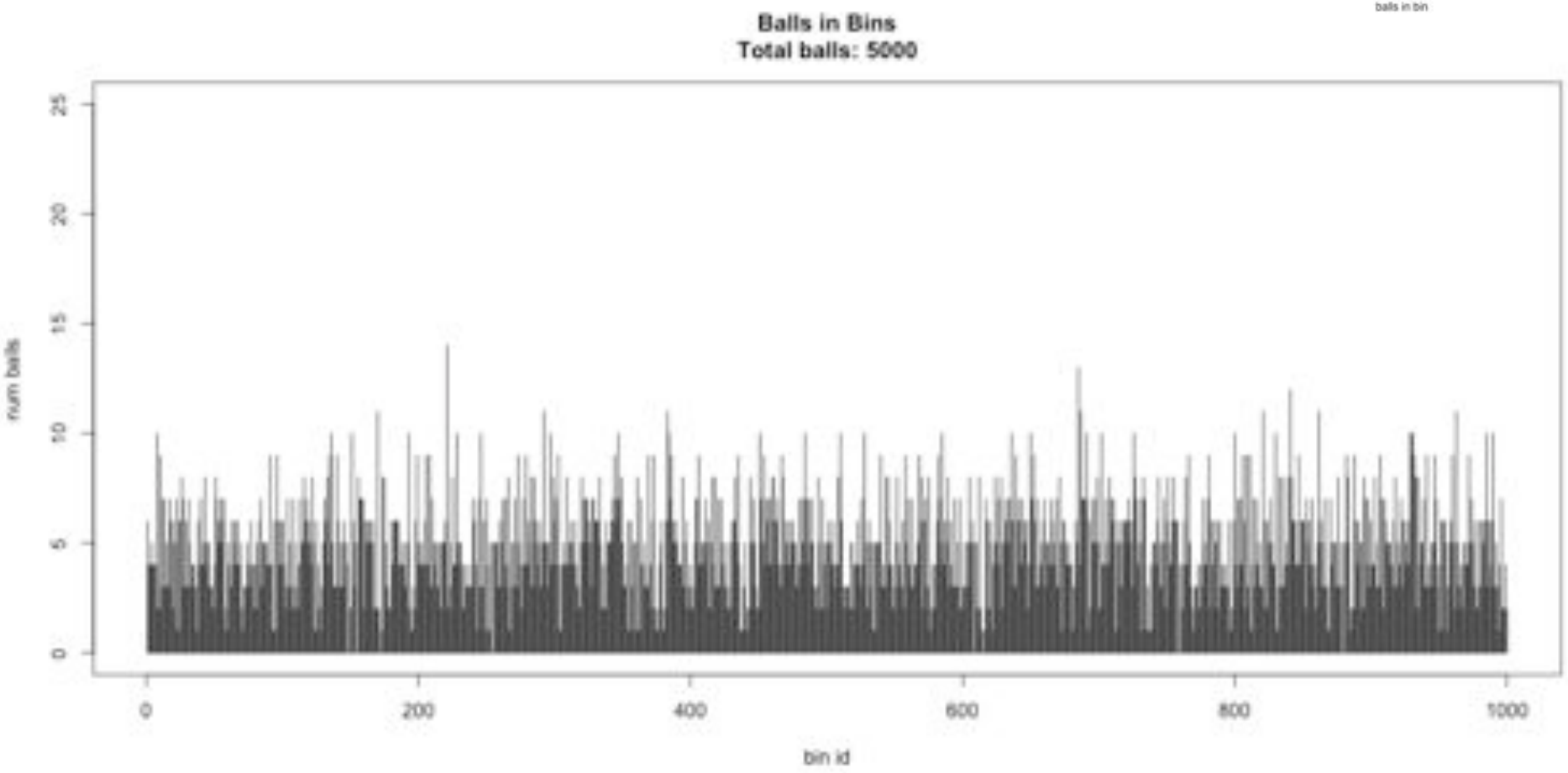
Balls in Bins 4x



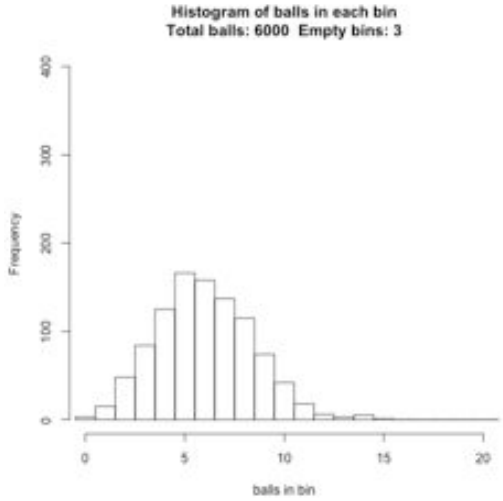


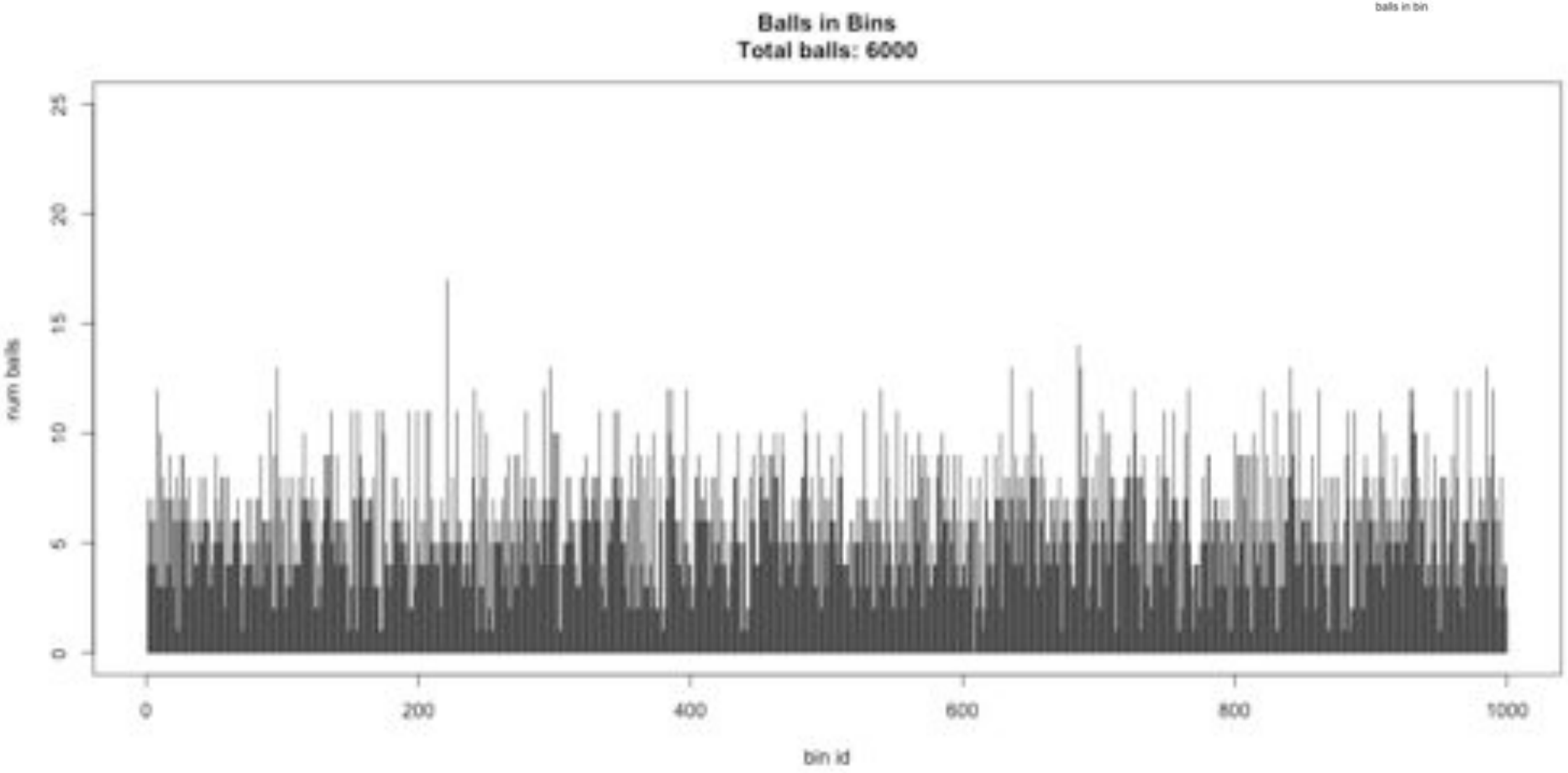
Balls in Bins 5x



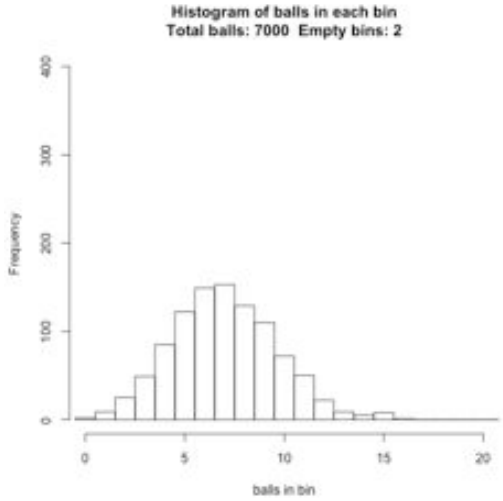


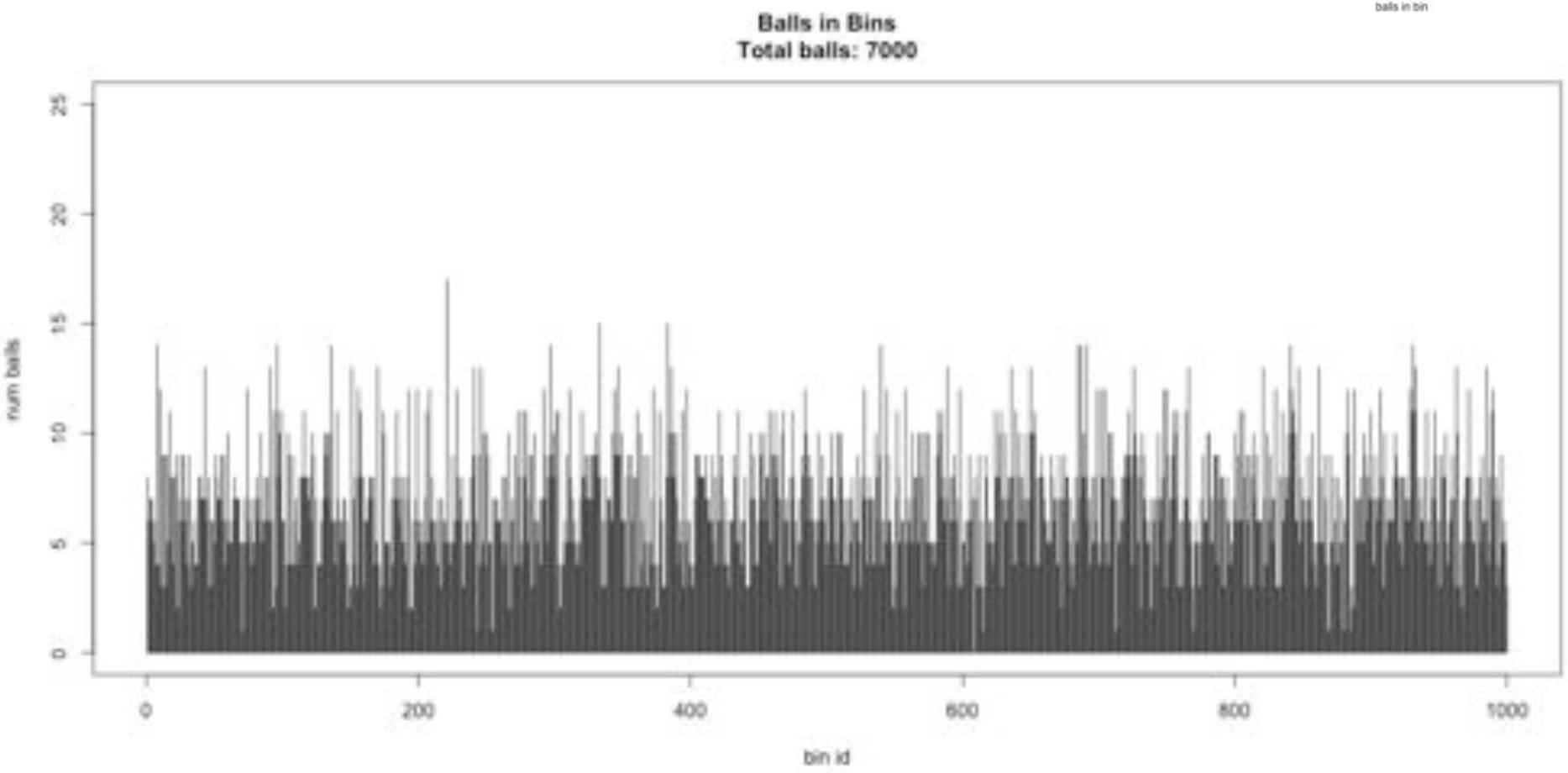
Balls in Bins 6x



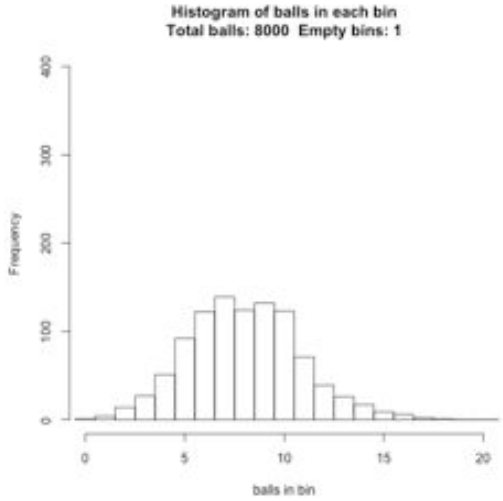


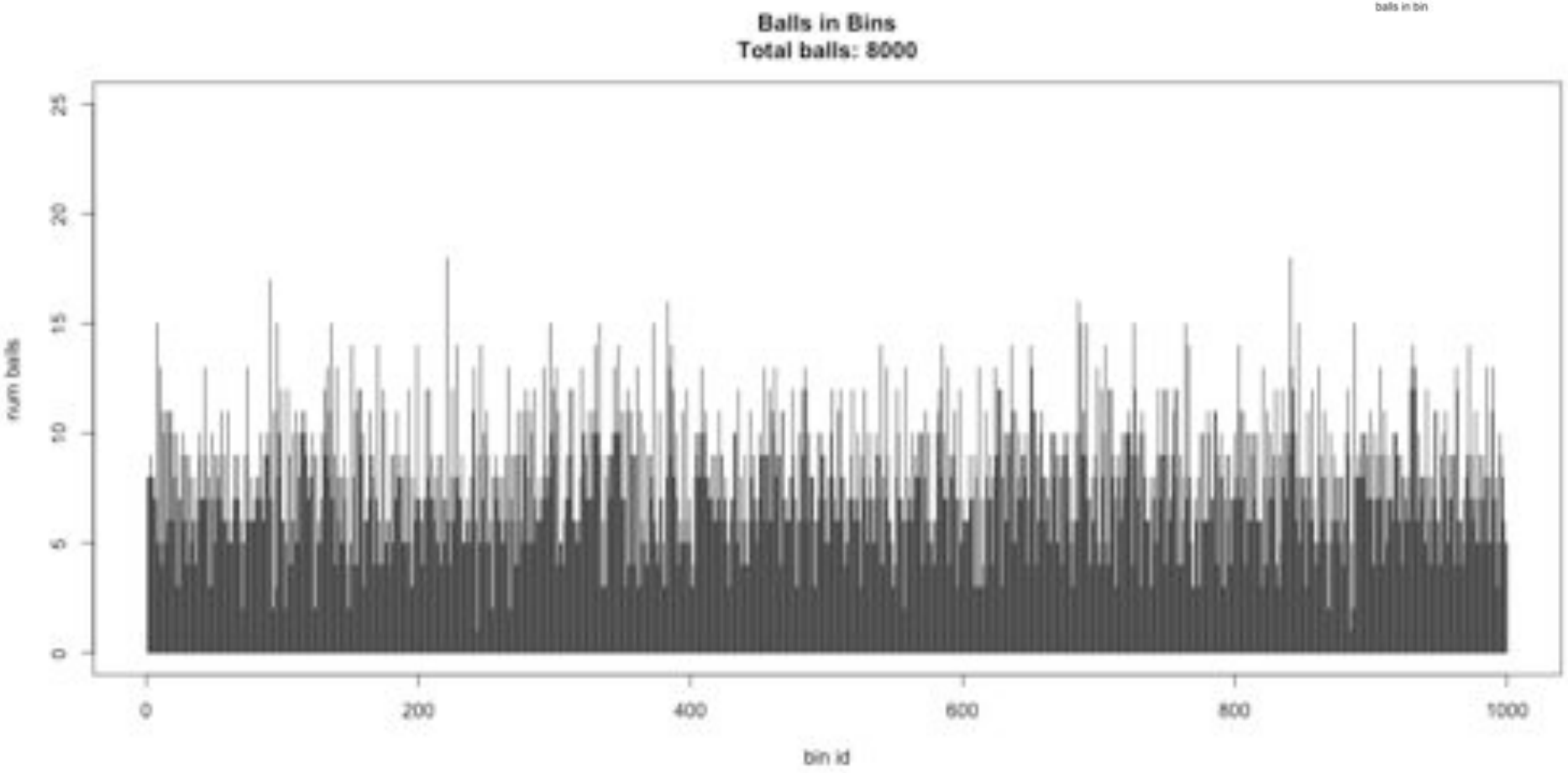
Balls in Bins 7x





Balls in Bins 8x

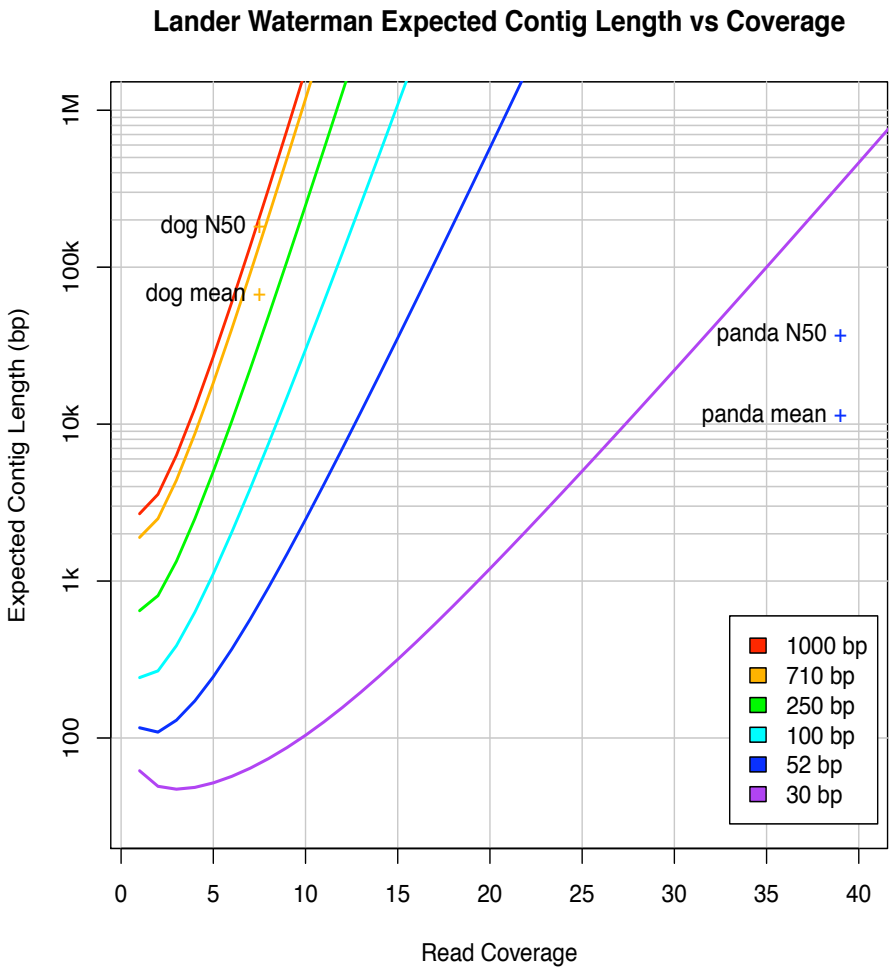




Coverage and Read Length

Idealized Lander-Waterman model

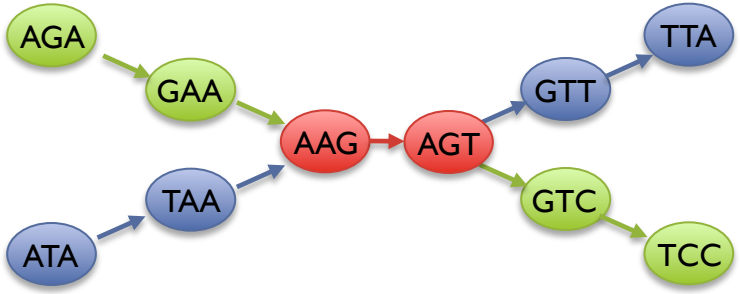
- Reads start at perfectly random positions
- Contig length is a function of coverage and read length
  - Short reads require much higher coverage to reach same expected contig length
- Need even high coverage for higher ploidy, sequencing errors, sequencing biases
  - Recommend 100x coverage



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

Two Paradigms for Assembly

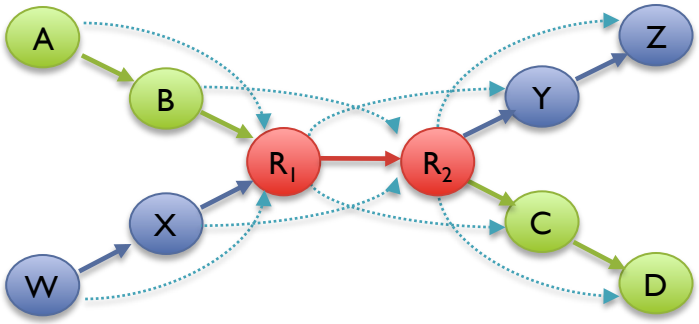
de Bruijn Graph



Short read assemblers

- Repeats depends on word length
- Read coherency, placements lost
- Robust to high coverage

Overlap Graph



Long read assemblers

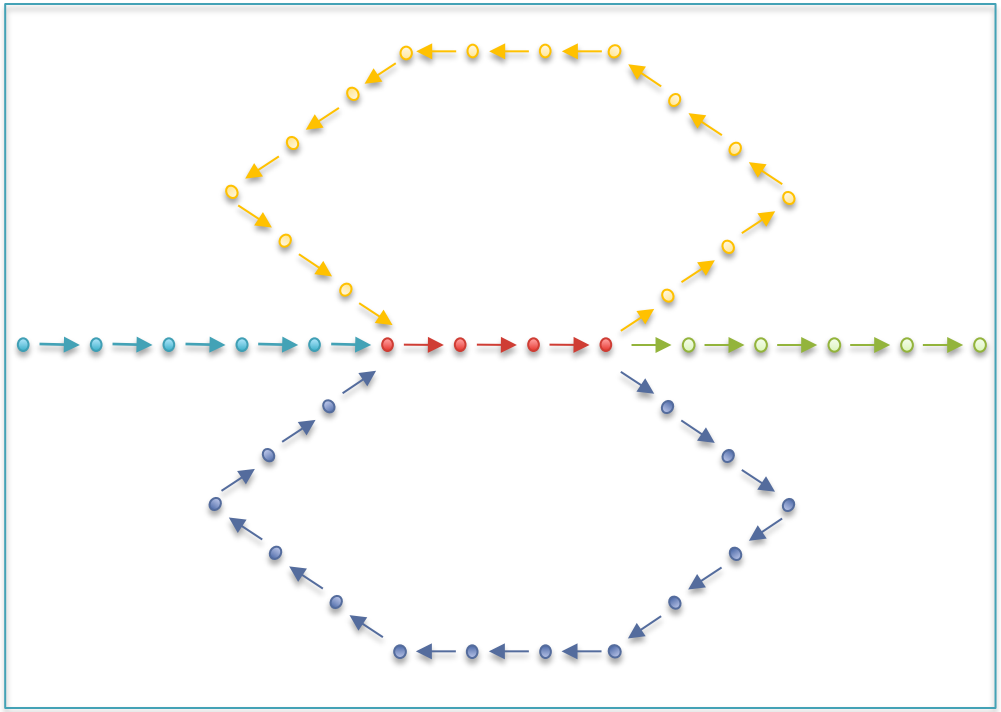
- Repeats depends on read length
- Read coherency, placements kept
- Tangled by high coverage

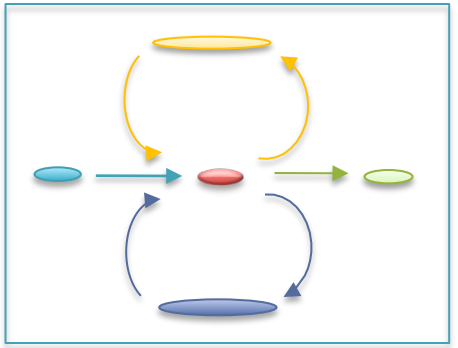
Assembly of Large Genomes using Second Generation Sequencing
Schatz MC, Delcher AL, Salzberg SL (2010) *Genome Research.* 20:1165-1173.

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





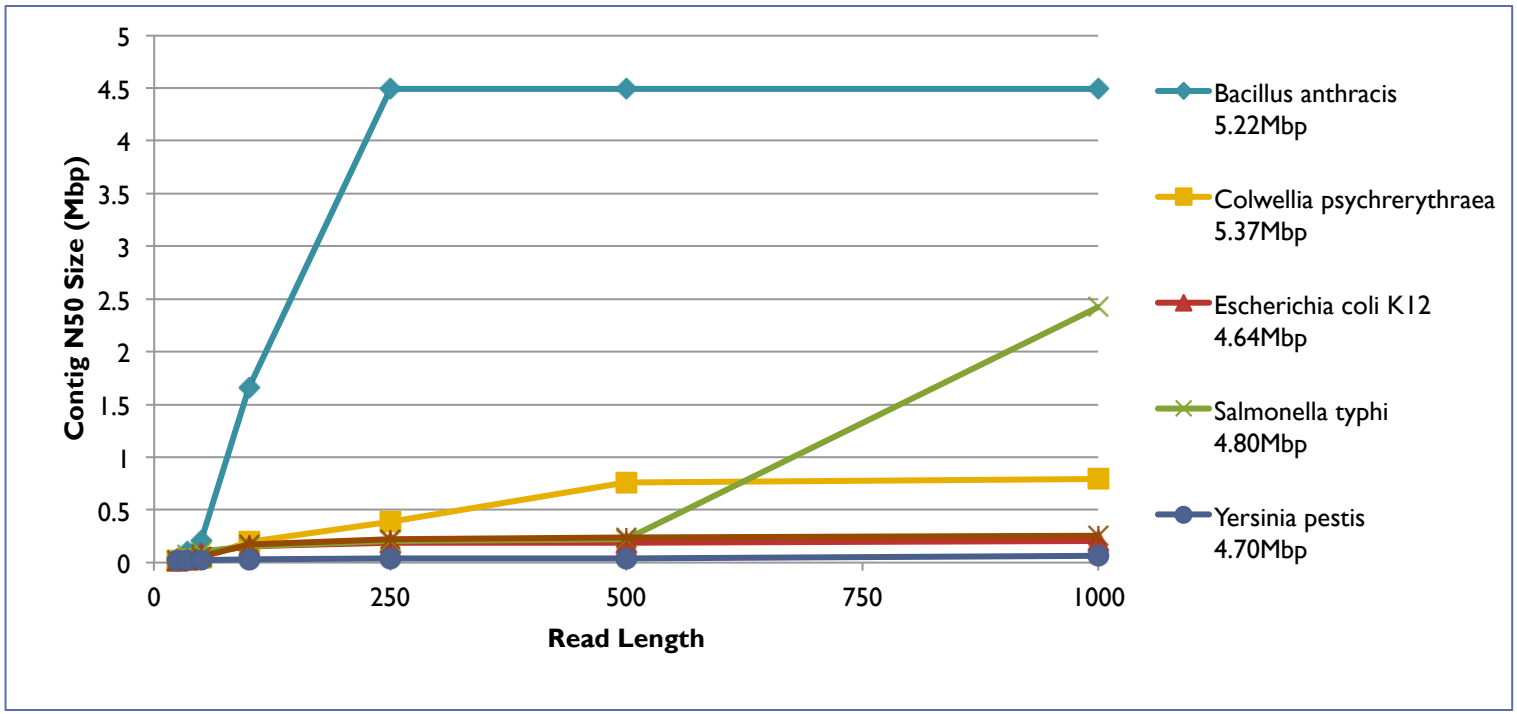
Errors in the graph

(Chaisson, 2009)

Clip Tips	Pop Bubbles
was the worst of times, was the worst of t**y**mes, the worst of times, it	was the worst of times, was the worst of t**y**mes, times, it was the age t**y**mes, it was the age
the worst of t**y**mes, was the worst of the worst of times, worst of times, it	t**y**mes, was the worst of it was the age times,

Repeats

Repeats and Read Length



- Explore the relationship between read length and contig N50 size
  - Idealized assembly of read lengths: 25, 35, 50, 100, 250, 500, 1000
  - Contig/Read length relationship depends on specific repeat composition

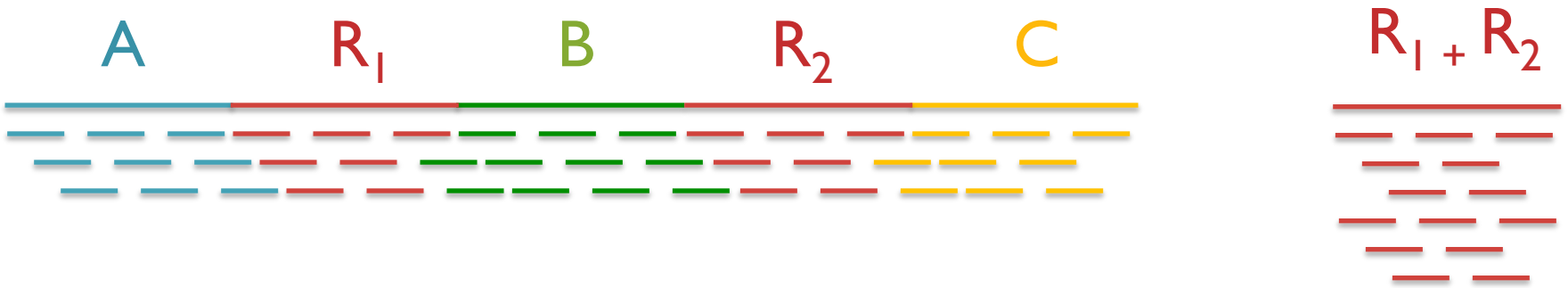
Assembly Complexity of Prokaryotic Genomes using Short Reads.
Kingsford C, Schatz MC, Pop M (2010) *BMC Bioinformatics*. 11:21.

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

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

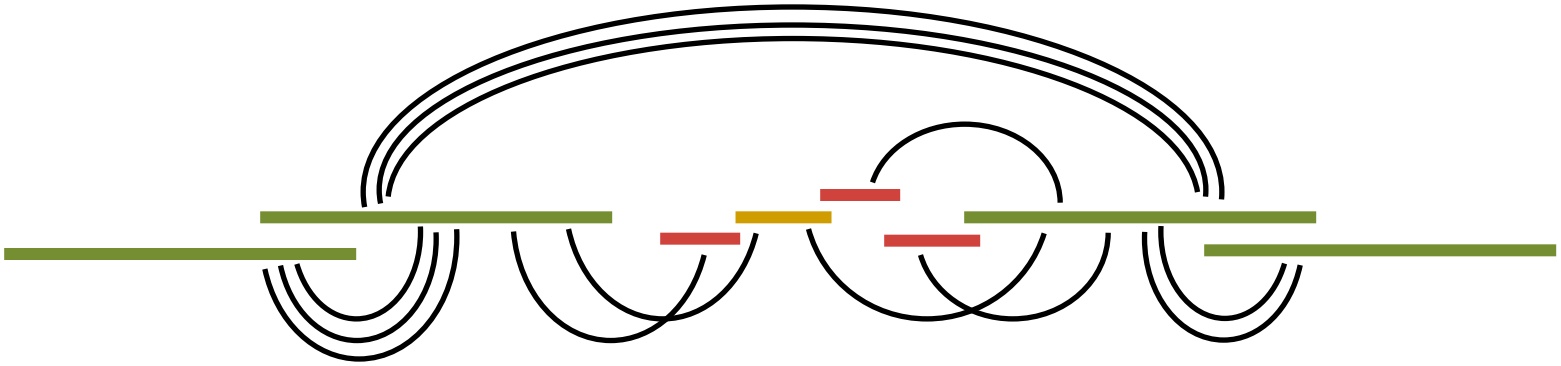
$$\Pr(X-copy)=\binom{n}{k}\left(\frac{X\Delta}{G}\right)^{k}\left(\frac{G-X\Delta}{G}\right)^{n-k}$$

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

Scaffolding

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

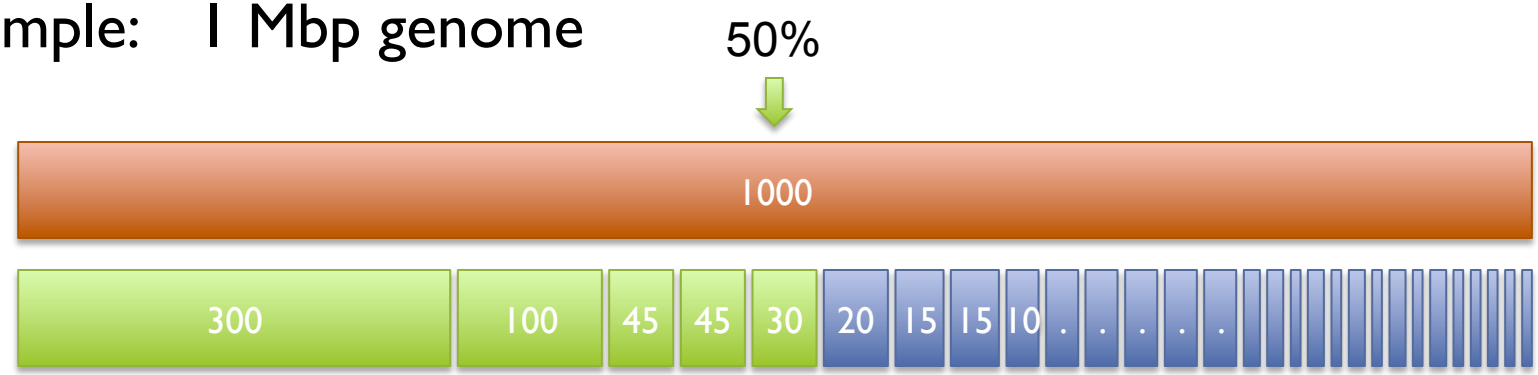
- Iteratively resolve longest, 'most unique' contigs
  - Both overlap graph and de Bruijn assemblers initially collapse repeats into single copies
  - Uniqueness measured by a statistical test on coverage



N50 size

Def: 50% of the genome is in contigs larger than N50

Example: 1 Mbp genome



N50 size = 30 kbp

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

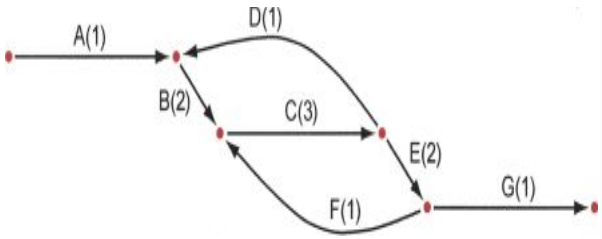
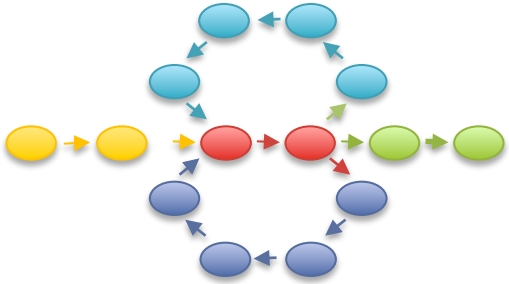
Note:

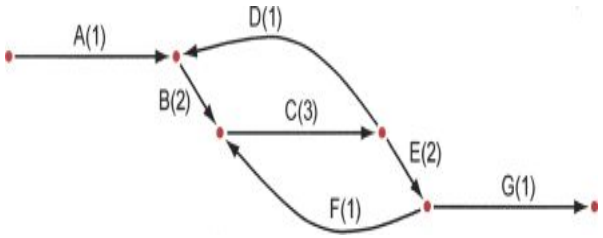
N50 values are only meaningful to compare when base genome size is the same in all cases

Break


PHASE TWO: INTERPRETATION
I THINK I FOUND A CORNER PIECE.
3 BILLION PIECES
GENOME

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

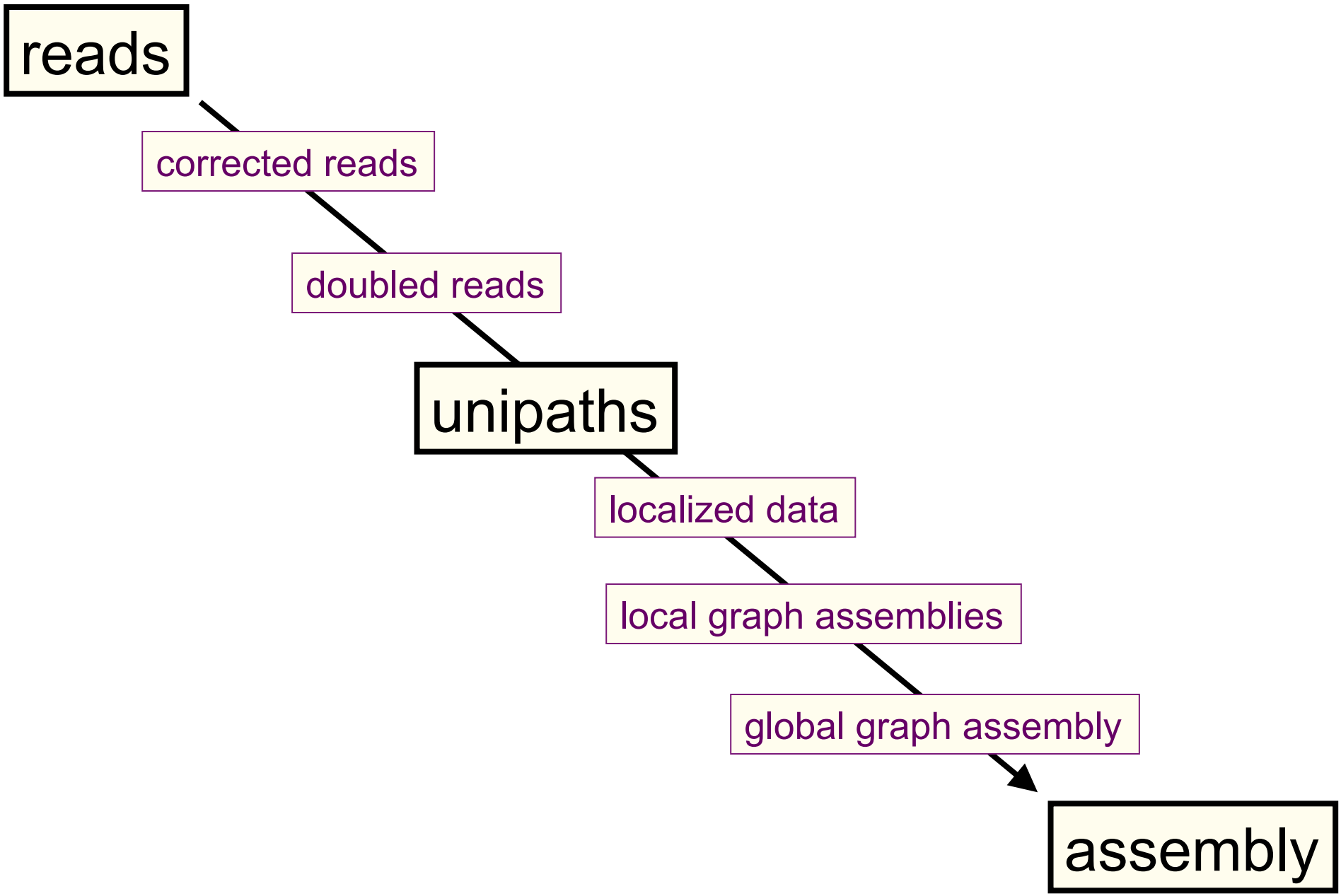


Genome assembly with ALLPATHS-LG

Iain MacCallum

BROAD INSTITUTE

How ALLPATHS-LG works



ALLPATHS-LG sequencing model

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

*See next slide.

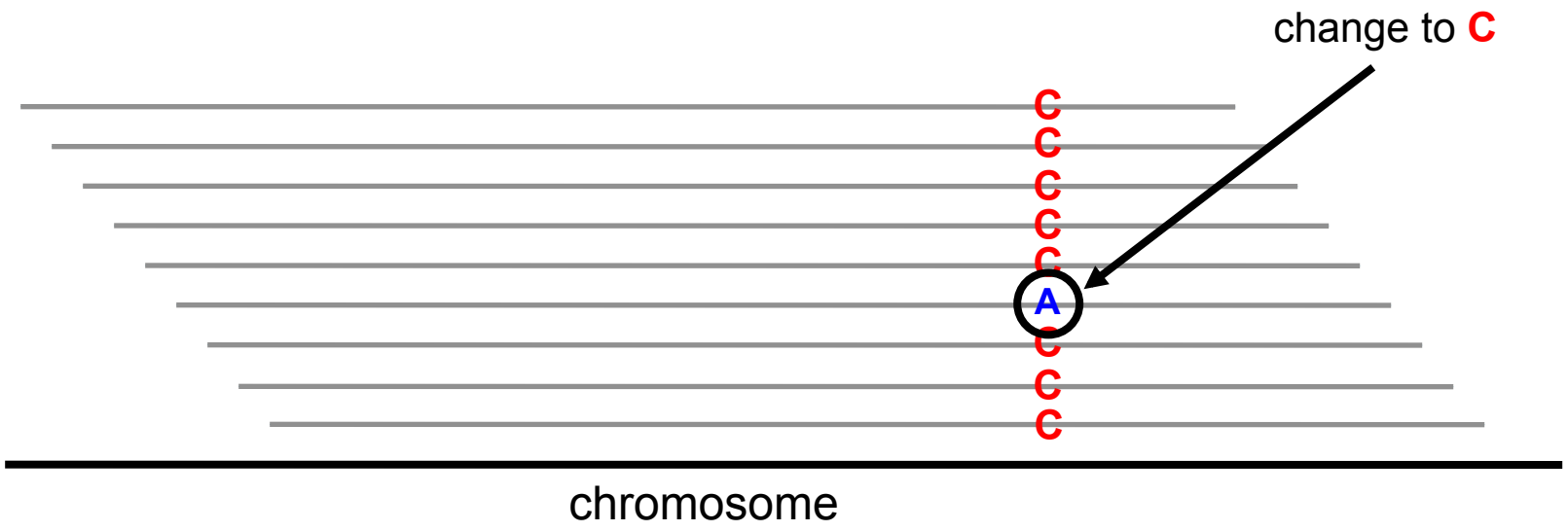
**For best results. Normally not used for small genomes.
However essential to assemble long repeats or duplications.

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

All: protocols are either available, or in progress.

Error correction

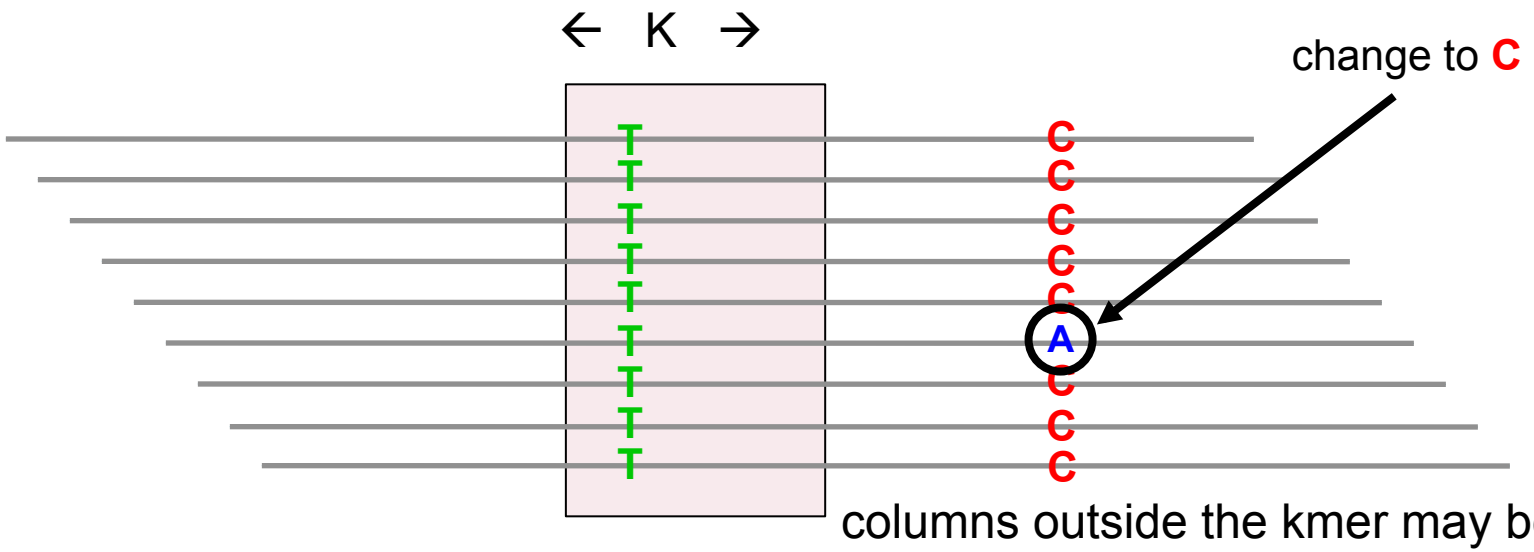
Given a crystal ball, we could stack reads on the chromosomes they came from (with homologous chromosomes separate), then let each column ‘vote’:



But we don’t have a crystal ball....

Error correction

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



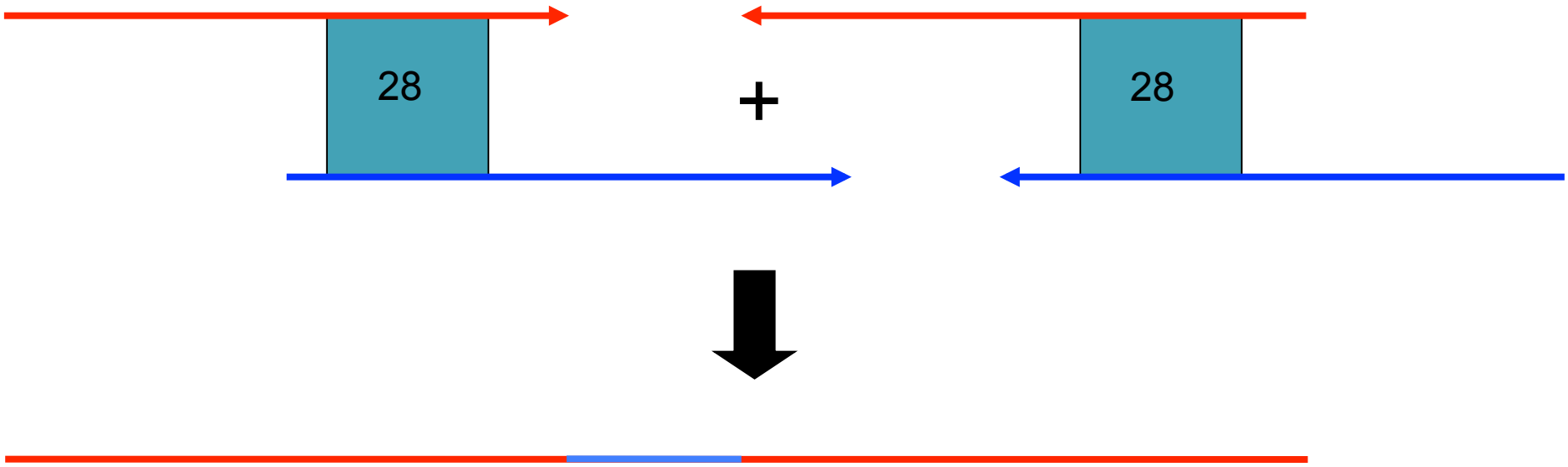
columns outside the kmer may be mixed

columns inside the kmer are homogeneous

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

Read doubling

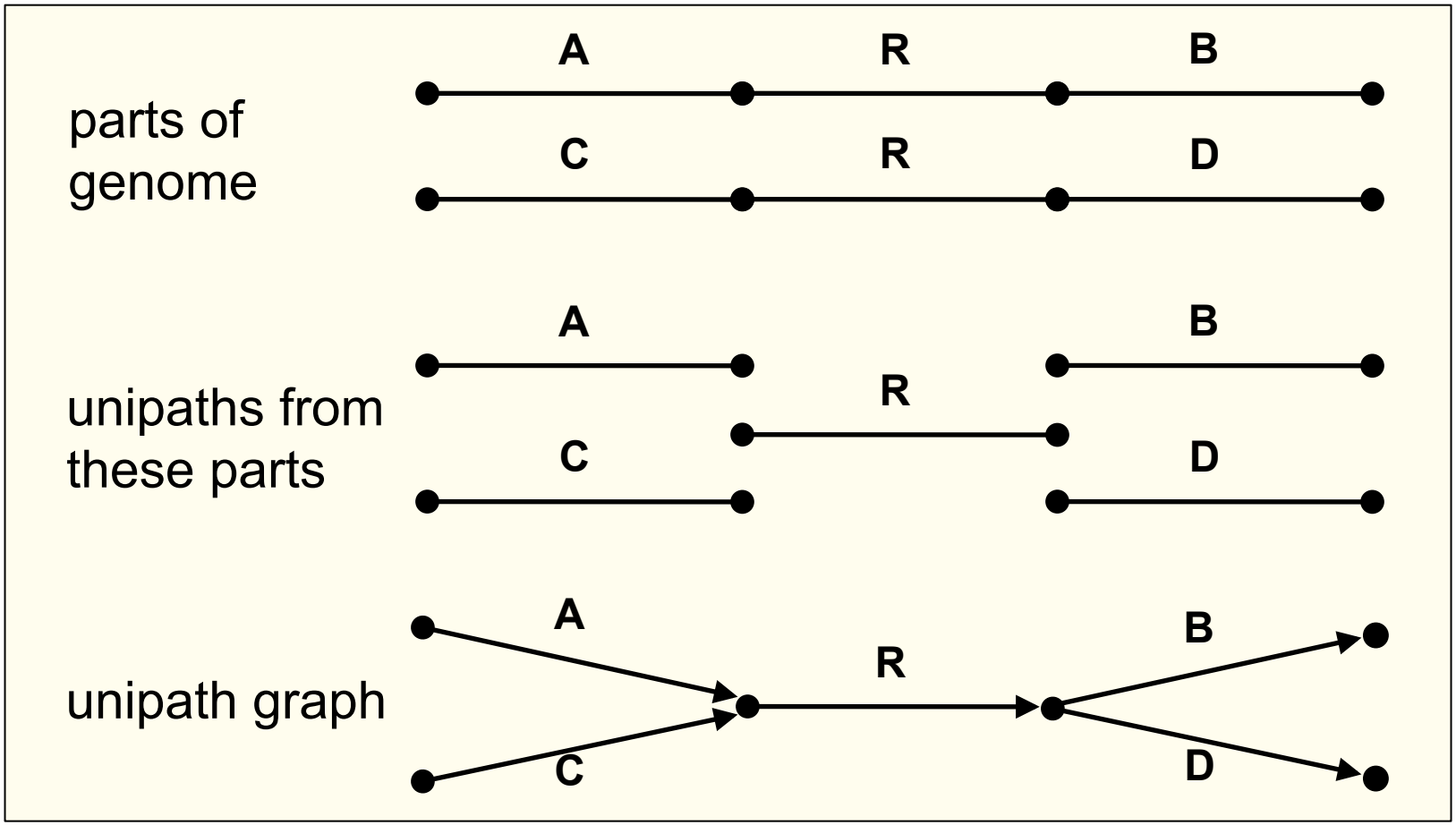
To close a read pair (red), we require the existence of another read pair (blue), overlapping perfectly like this:



More than one closure allowed (but rare).

Unipaths

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

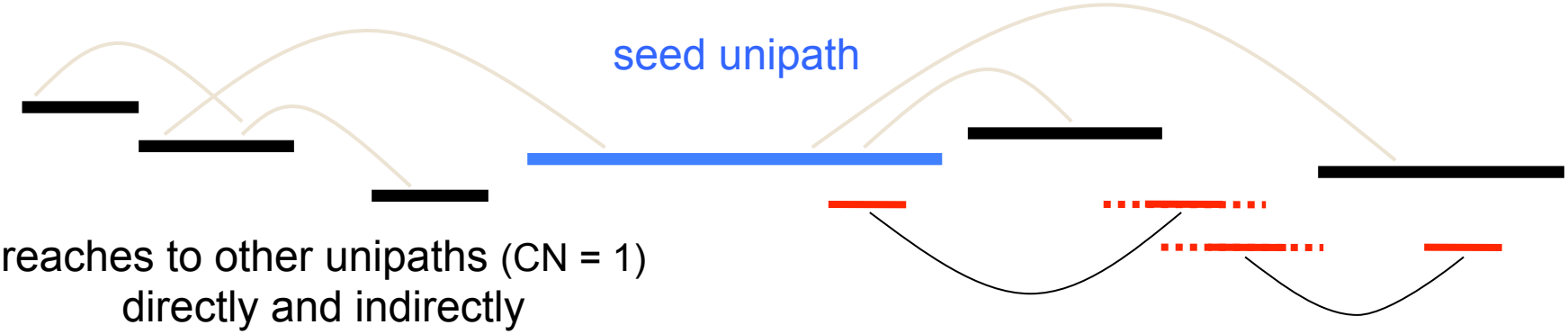


Adjacent unipaths overlap by K-1 bases

Localization

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

II. Form neighborhood around each seed



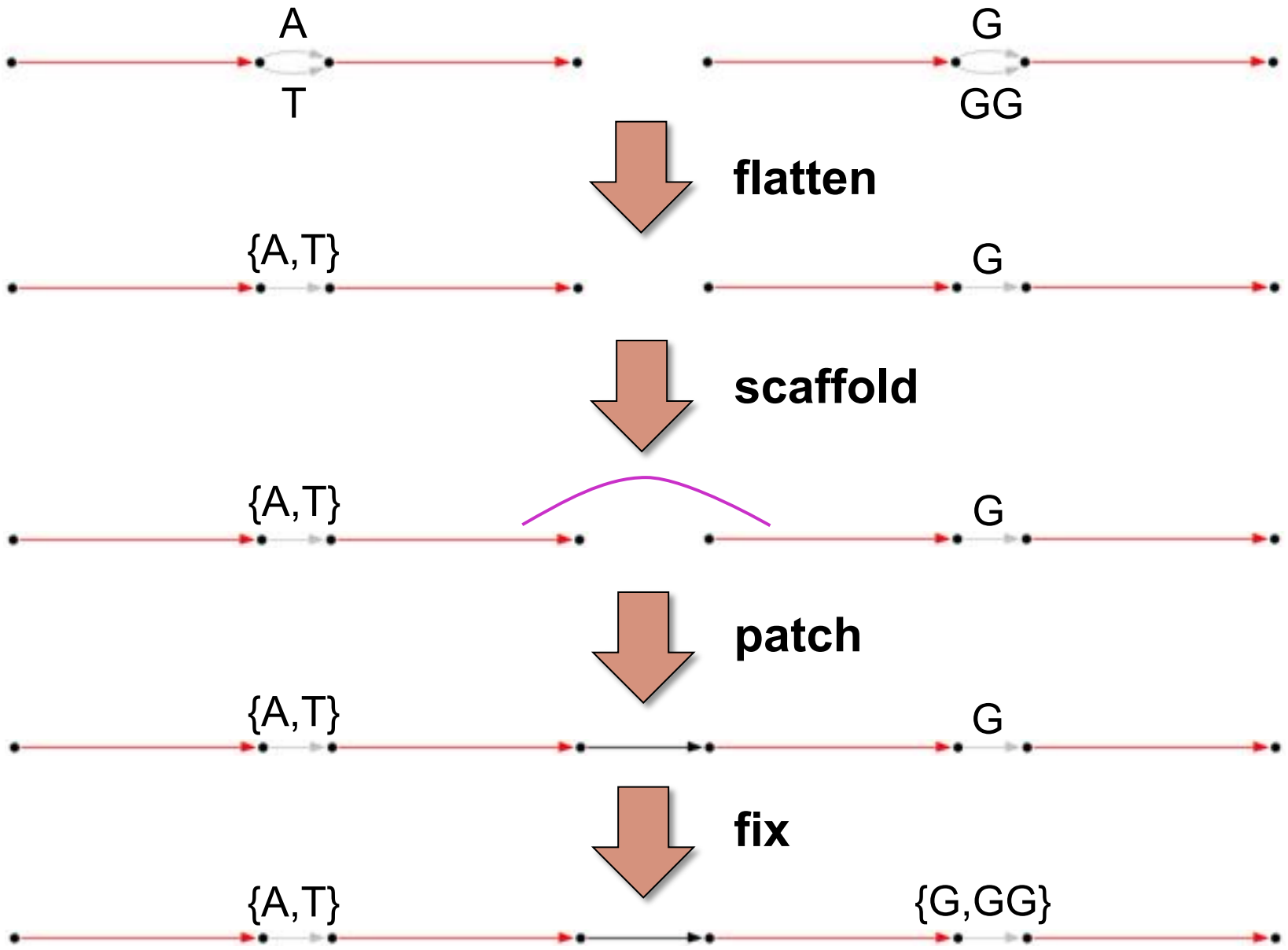
reaches to other unipaths (CN = 1)
directly and indirectly

read pairs reach into repeats

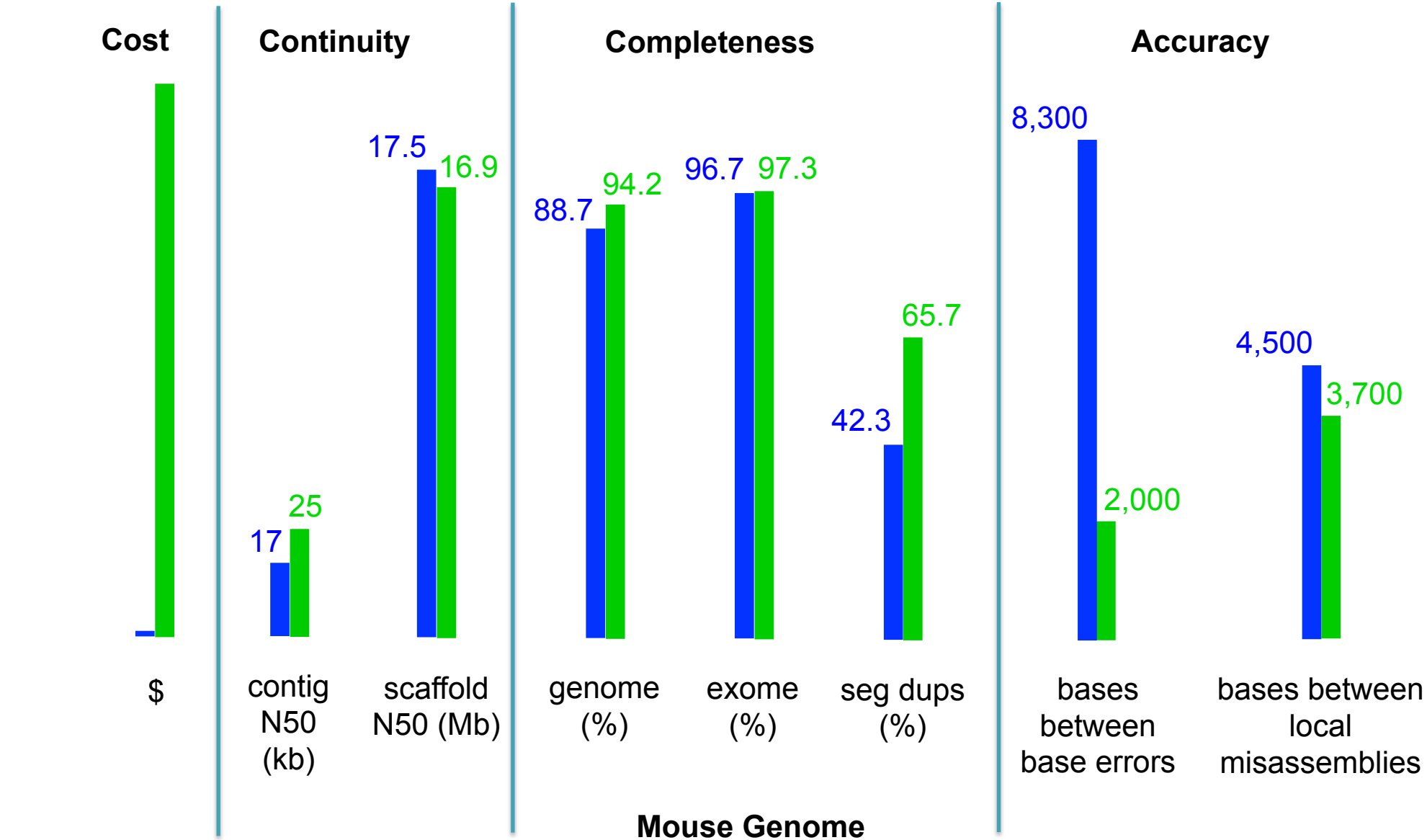
and are extended by other unipaths

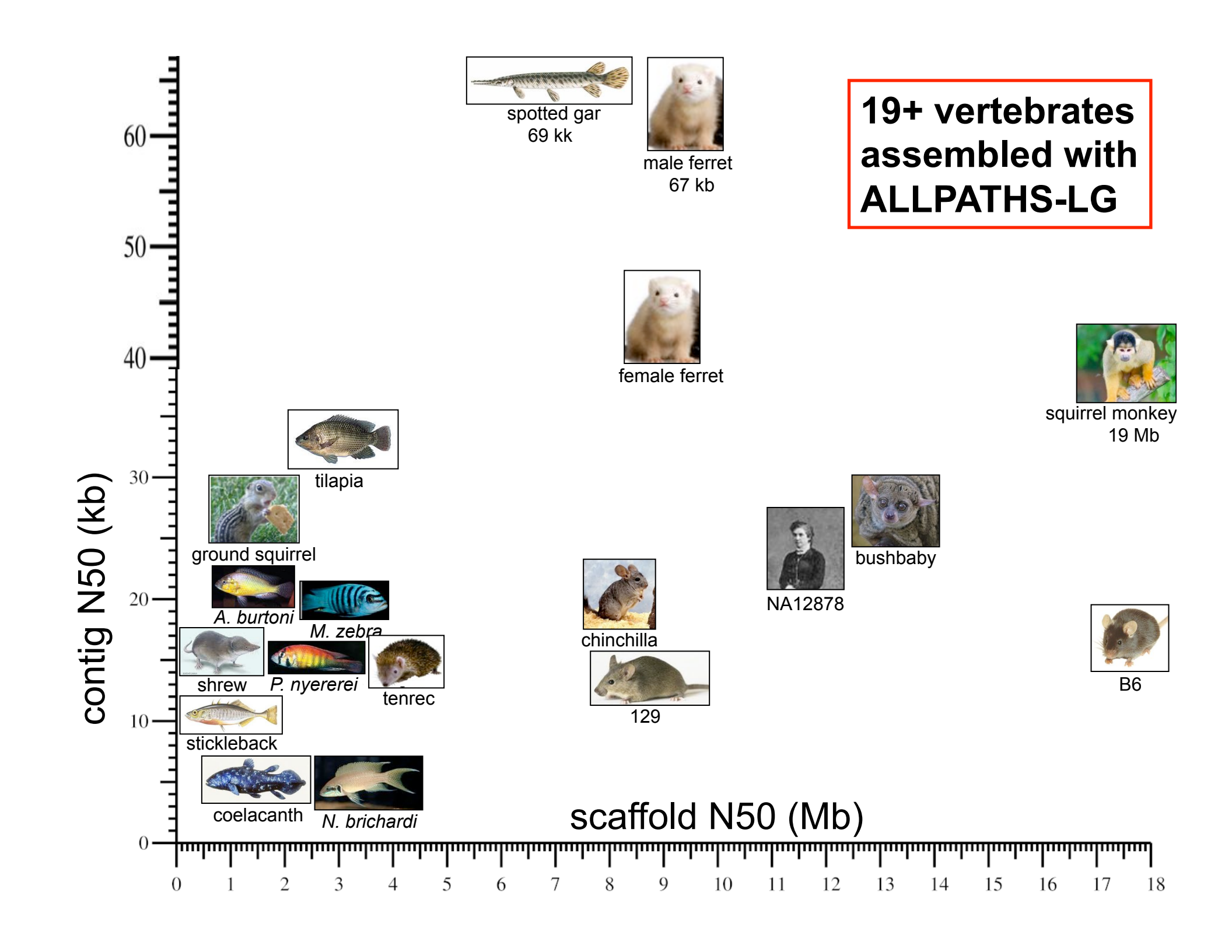
.....

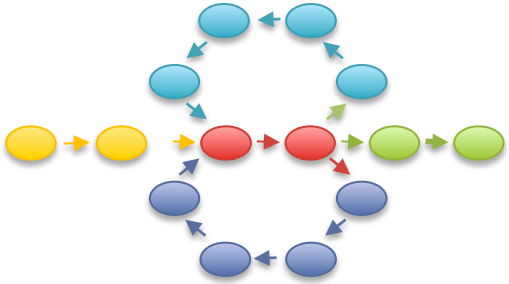
Create assembly from global assembly graph



Large genome recipe: ALLPATHS-LG vs capillary




19+ vertebrates assembled with ALLPATHS-LG
spotted gar
69 kk
male ferret
67 kb
female ferret
squirrel monkey
19 Mb
tilapia
ground squirrel
A. burtoni
M. zebra
shrew
P. nyererei
tenrec
stickleback
coelacanth
N. brichardi
chinchilla
129
NA12878
bushbaby
B6
contig N50 (kb)
scaffold N50 (Mb)
0
10
20
30
40
50
60
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18



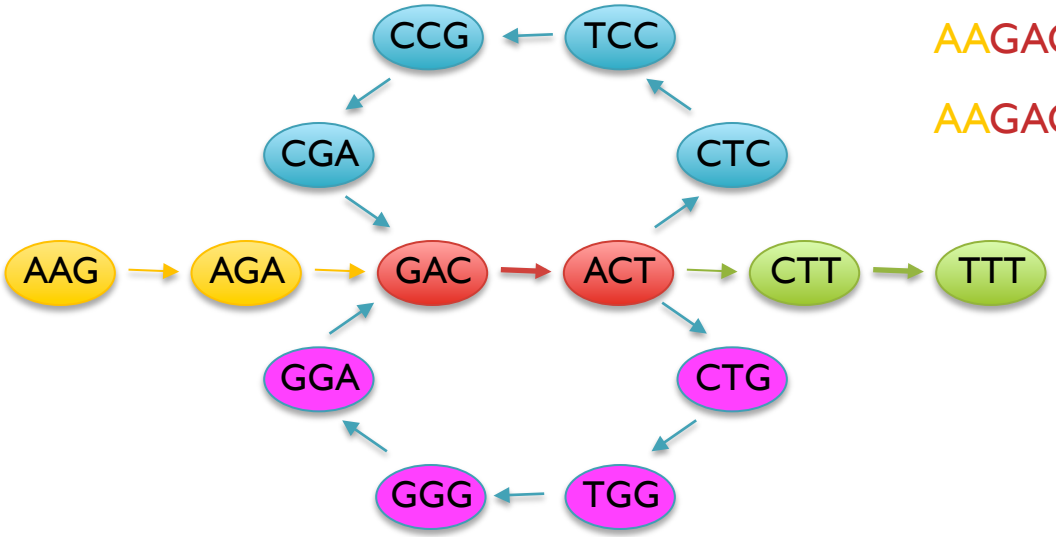
Genome assembly with SOAPdenovo

Short Read Assembly

Reads

AAGA
ACTT
ACTC
ACTG
AGAG
CCGA
CGAC
CTCC
CTGG
CTTT
…

de Bruijn Graph



Potential Genomes

AAGACTCCGACTGGGACTTT

AAGACTGGGACTCCGACTTT

- Genome assembly as finding an Eulerian tour of the de Bruijn graph
  - Human genome: >3B nodes, >10B edges
- The new short read assemblers require tremendous computation
  - Velvet (Zerbino & Birney, 2008) serial: > 2TB of RAM
  - ABySS (Simpson *et al.*, 2009) MPI: 168 cores x ~96 hours
  - SOAPdenovo (Li *et al.*, 2010) pthreads: 40 cores x 40 hours, >140 GB RAM

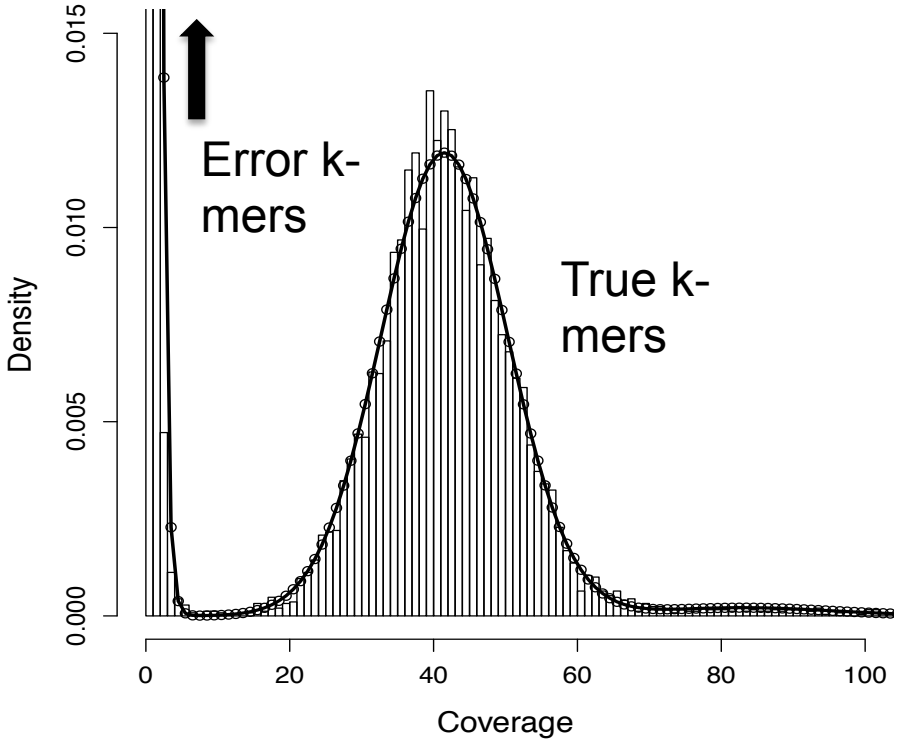
Error Correction with Quake

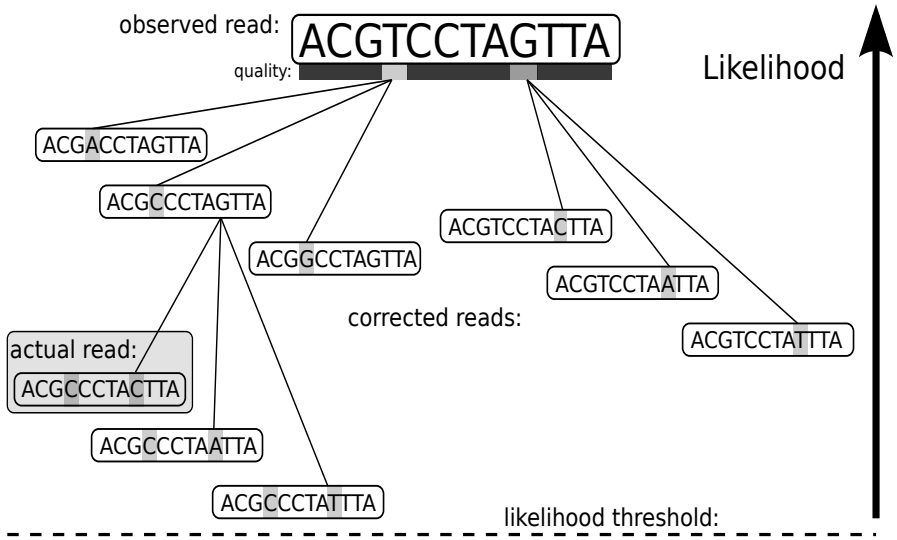
1. Count all "Q-mers" in reads

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

2. Correction Algorithm

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



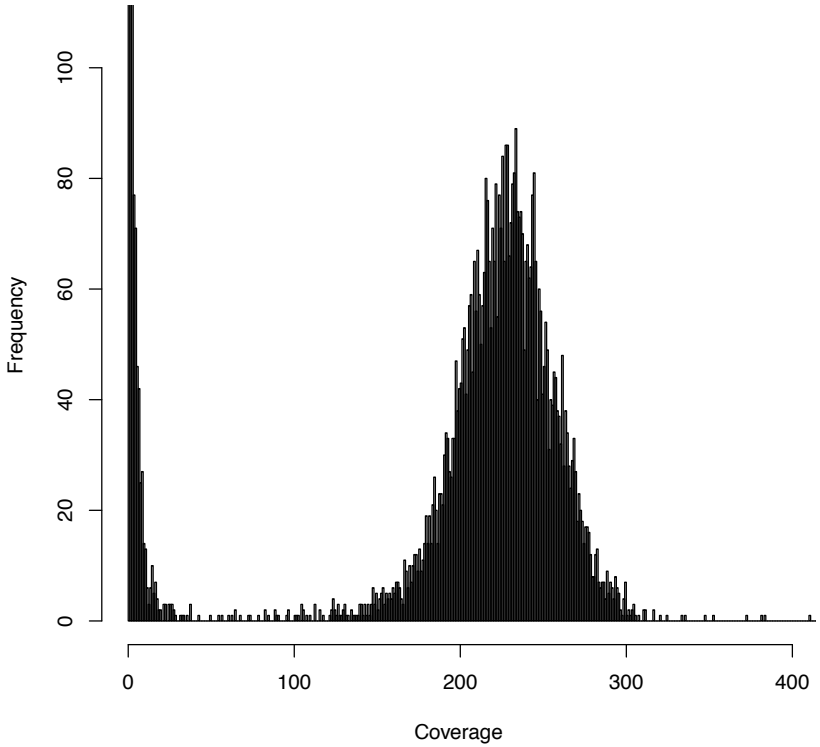


Quake: quality-aware detection and correction of sequencing reads.
Kelley, DR, Schatz, MC, Salzberg SL (2010) *Genome Biology*. 11:R116

Illumina Sequencing & Assembly

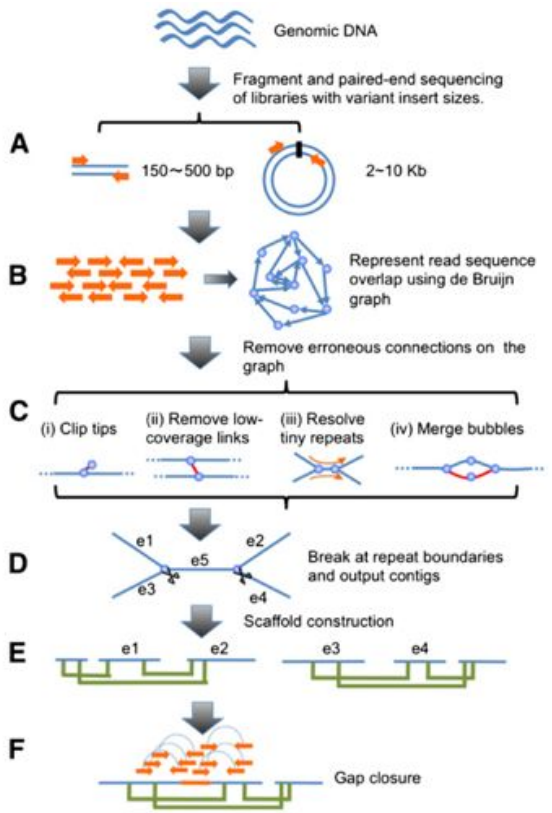
Quake Results

2x76bp @ 275bp
2x36bp @ 3400bp

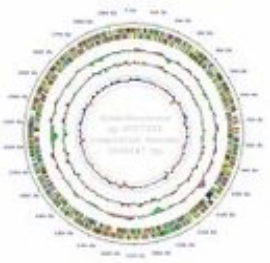


Validated	51,243,281	88.5%
Corrected	2,763,380	4.8%
Trim Only	3,273,428	5.6%
Removed	606,251	1.0%

SOAPdenovo Results



	# ≥ 100bp	N50 (bp)
Scaffolds	2,340	253,186
Contigs	2,782	56,374
Unitigs	4,151	20,772

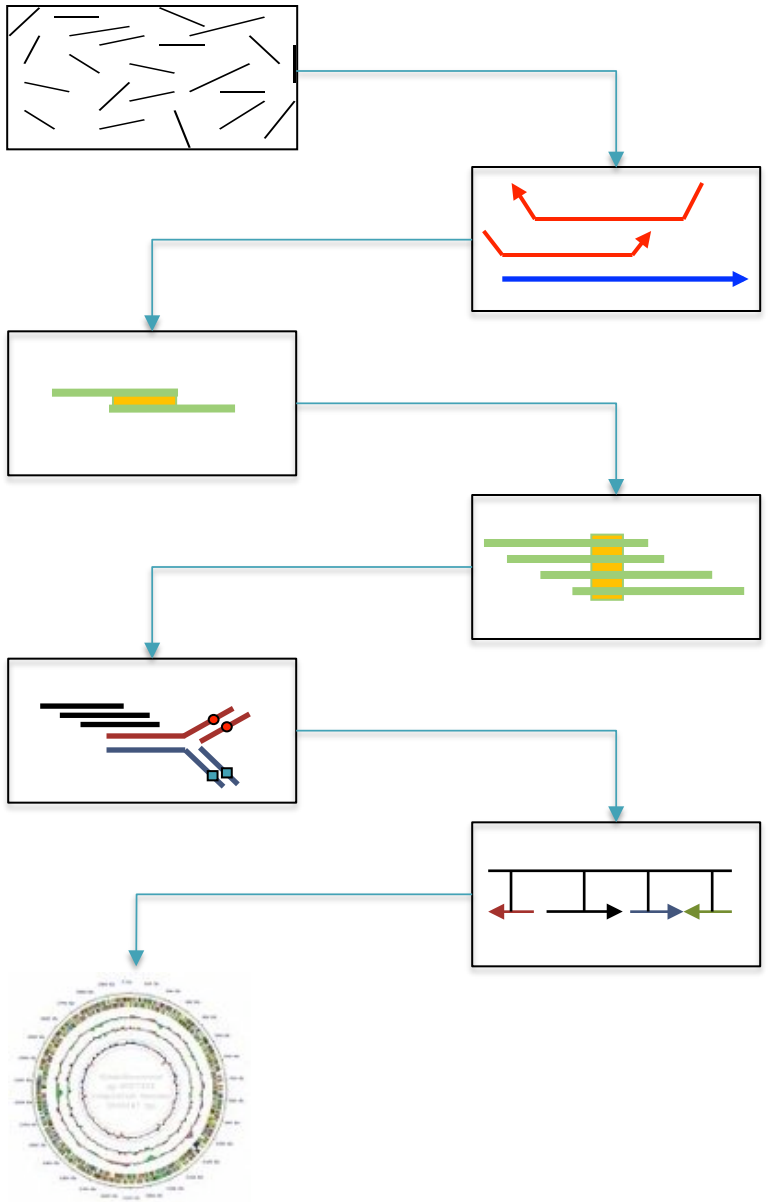


Genome assembly with the Celera Assembler

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

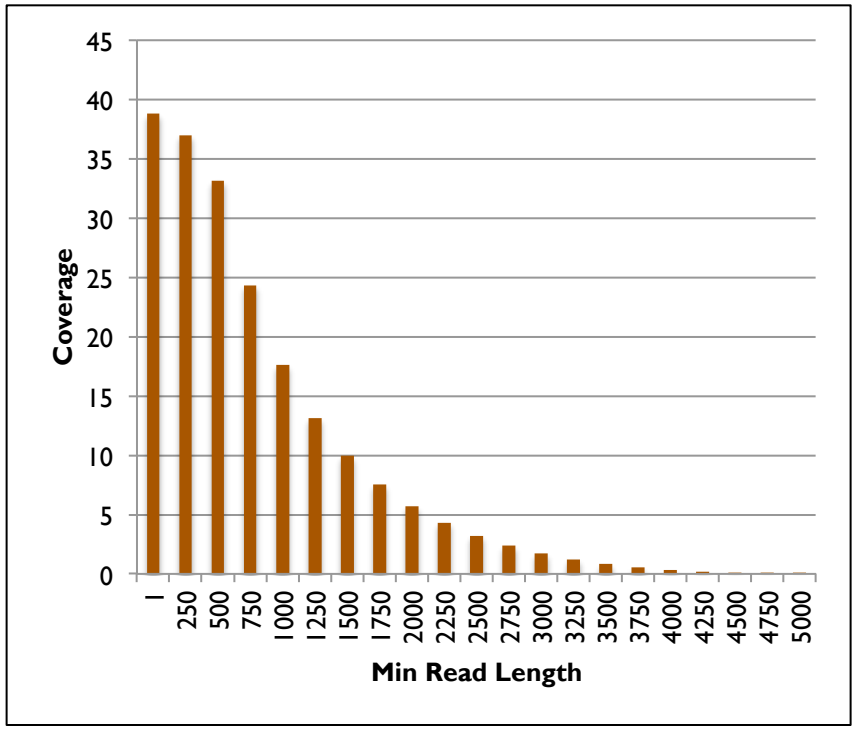
SMRT Sequencing

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

SMRT Sequencing Data

Yeast
(Pre-release Chemistry / 2010)

65 SMRT cells
734,151 reads after filtering
Mean: 642.3 +/- 587.3
Median: 553 Max: 8,495




```
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
```

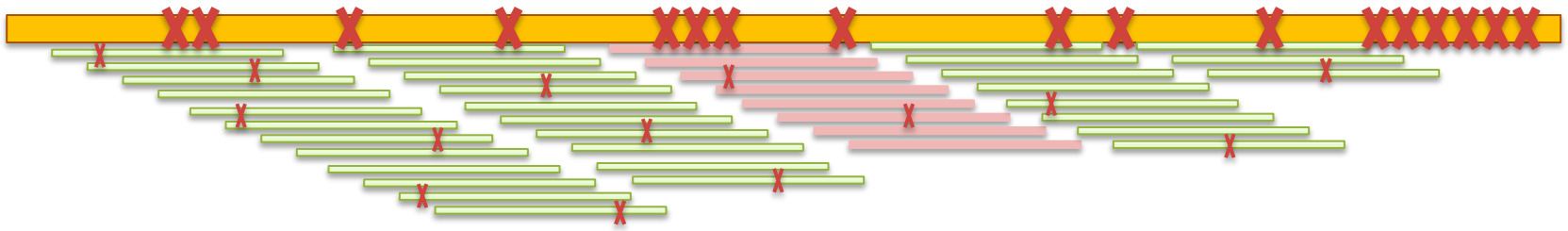
Sample of 100k reads aligned with BLASR requiring >100bp alignment
Average overall accuracy: 83.7%, 11.5% insertions, 3.4% deletions, 1.4% mismatch

PacBio Error Correction

http://wgs-assembler.sf.net

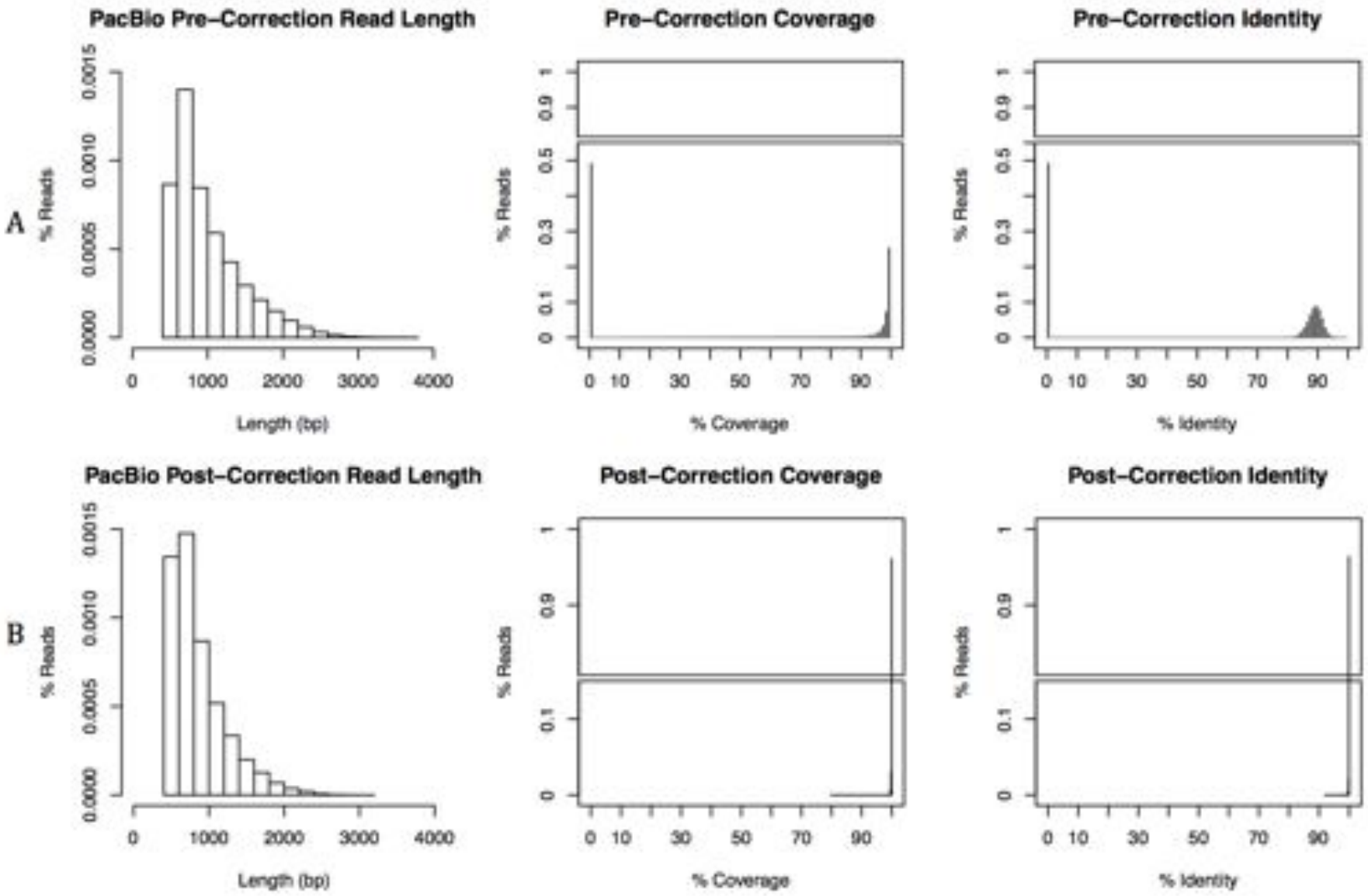
1. Correction Pipeline
    1. Map short reads to long reads
    2. Trim long reads at coverage gaps
    3. Compute consensus for each long read

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

Error Correction Results



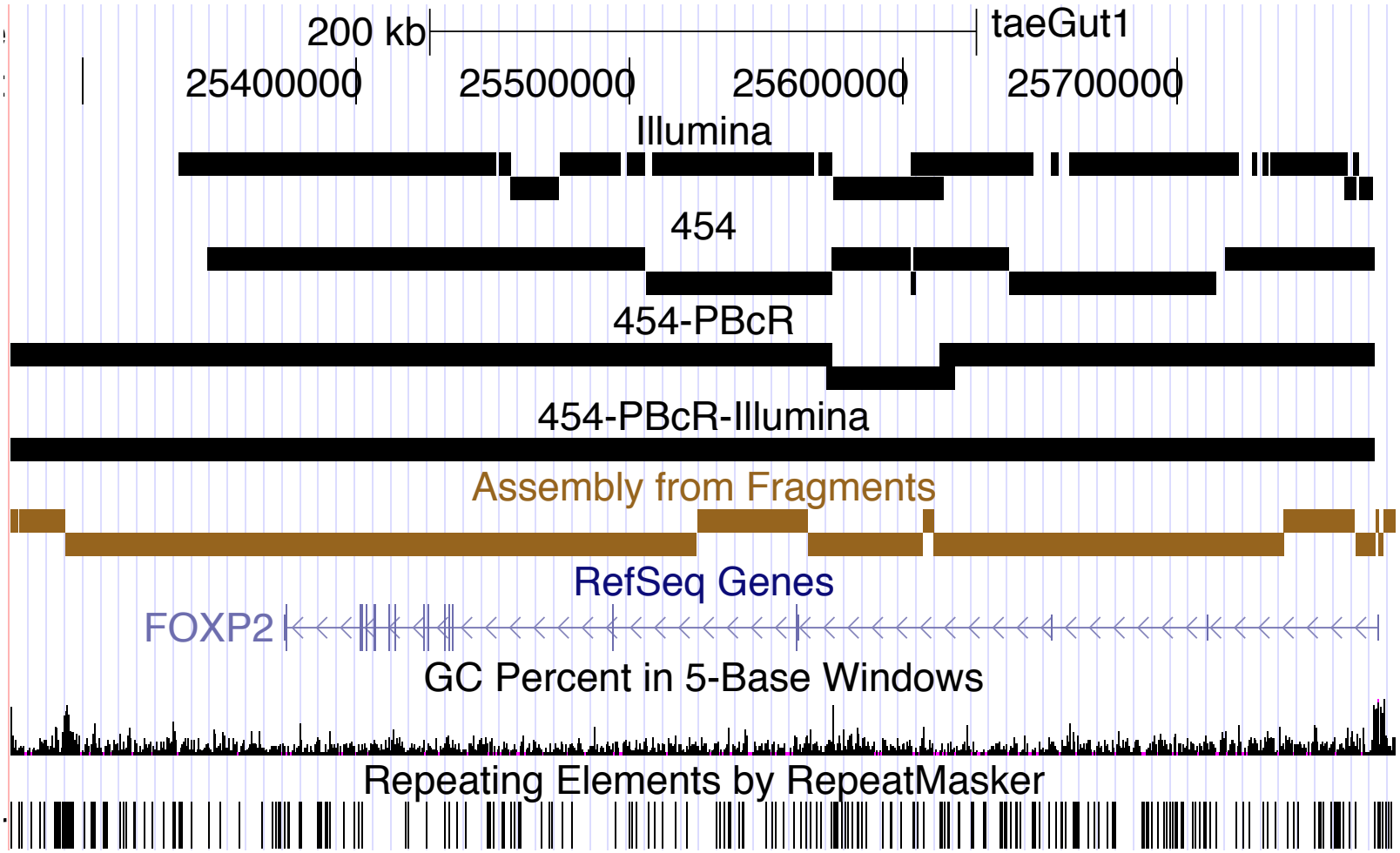
Correction results of 20x PacBio coverage of E. coli K12 corrected using 50x Illumina

SMRT-Assembly Results

Organism	Technology	Reference bp	Assembly bp	# Contigs	Max Contig Length	N50
Lambda NEB3011	Illumina 100X 200bp	48 502	48 492	1	48 492 / 48 492	48 492 / 48 492 (100%) *
(median: 727 max: 3 280)	PacBio PBcR 25X		48 440	1	48 444 / 48 444	48 444 / 48 440 (100%) *
E. coli K12	Illumina 100X 500bp	4 639 675	4 462 836	61	221 615 / 221 553	100 338 / 83 037 (82.76%) *
(median: 747 max: 3 068)	PacBio PBcR 18X		4 465 533	77	239 058 / 238 224	71 479 / 68 309 (95.57%) *
	Both 18X PacBio PBcR + Illumina 50X 500bp		4 576 046	65	238 272 / 238 224	93 048 / 89 431 (96.11%) *
E. coli C227-11	PacBio CCS 50X	5 504 407	4 917 717	76	249 515	100 322
(median: 1 217 max: 14 901)	PacBio 25X PBcR (corrected by 25X CCS)		5 207 946	80	357 234	98 774
	Both PacBio PBcR 25X + CCS 25X		5 269 158	39	647 362	227 302
	PacBio 50X PBcR (corrected by 50X CCS)		5 445 466	35	1 076 027	376 443
	Both PacBio PBcR 50X + CCS 25X		5 453 458	33	1 167 060	527 198
	Manually Corrected ALLORA Assembly[a]		5 452 251	23	653 382	402 041
S. cerevisiae S228c	Illumina 100X 300bp	12 157 105	11 034 156	192	266 528 / 227 714	73 871 / 49 254 (66.68%) *
(median: 674 max: 5 994)	PacBio PBcR 13X		11 110 420	224	224 478 / 217 704	62 898 / 54 633 (86.86%) *
	Both PacBio PBcR 13X + Illumina 50X 300bp		11 286 932	177	262 846 / 260 794	82 543 / 59 792 (72.44%) *
Melopsittacus undulatus	Illumina 194X (220/500/800 paired-end 2/5/10Kb mate-pairs)	1.23 Gbp	1 023 532 850	24 181	1 050 202	47 383
	454 15.4X (FLX + FLX Plus + 3/8/20Kbp paired-ends)		999 168 029	16 574	751 729	75 178
(median 997, max 13 079)	454 15.4X + PacBio PBcR 3.75X		1 071 356 415	15 081	1 238 843	99 573

Hybrid assembly results using error corrected PacBio reads
Meets or beats Illumina-only or 454-only assembly in every case

Improved Gene Reconstruction

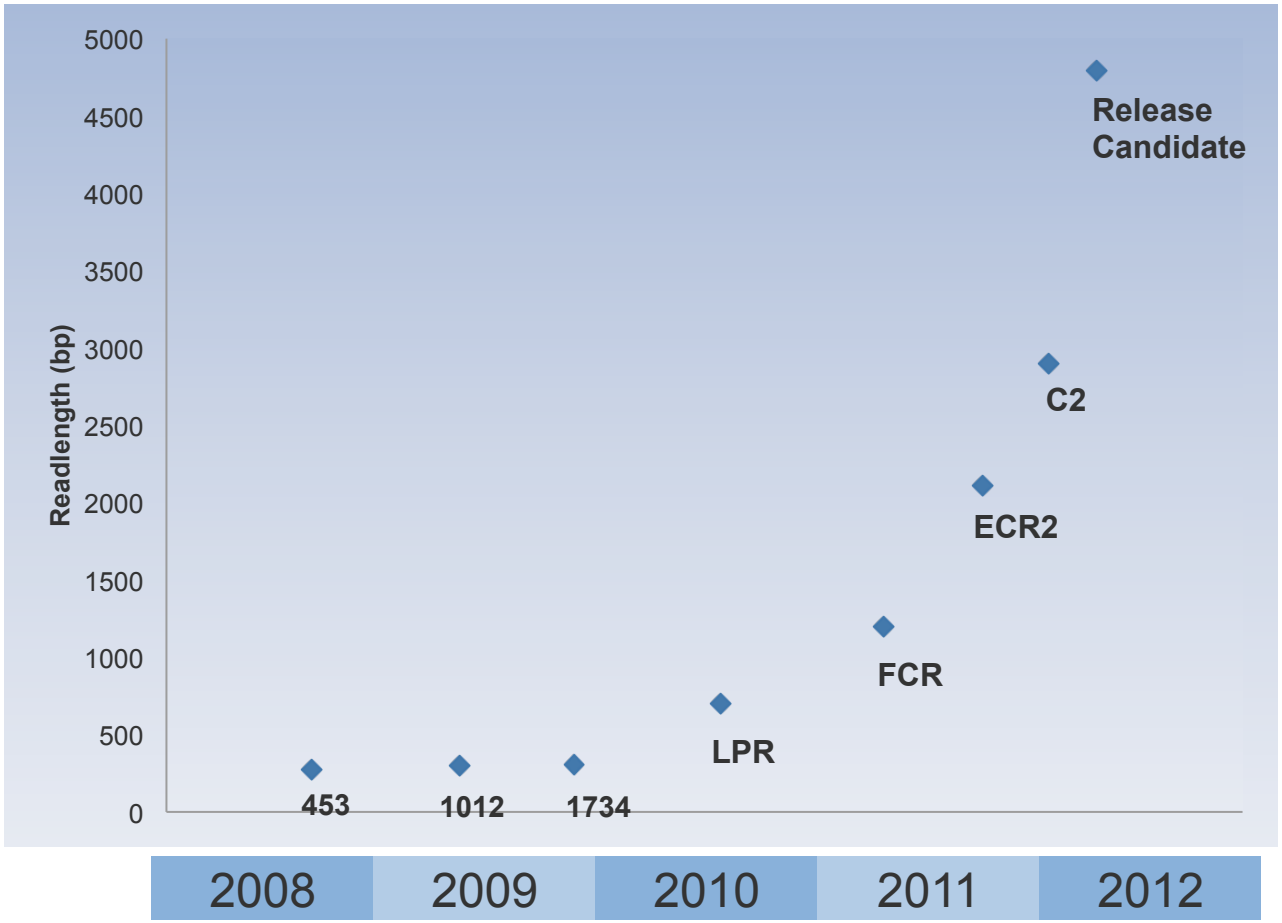


FOXP2 assembled on a single contig

Transcript Alignment

- Long-read single-molecule sequencing has potential to directly sequence full length transcripts
  - Raw reads and raw alignments (red) have many spurious indels inducing false frameshifts and other artifacts
  - Error corrected reads almost perfectly match the genome, pinpointing splice sites, identifying alternative splicing
- New collaboration with Gingeras Lab looking at splicing in human

PacBio Technology Roadmap



Internal Roadmap has made steady progress towards improving read length and throughput

Very recent improvements:

1. Improved enzyme:

   Maintains reactions longer

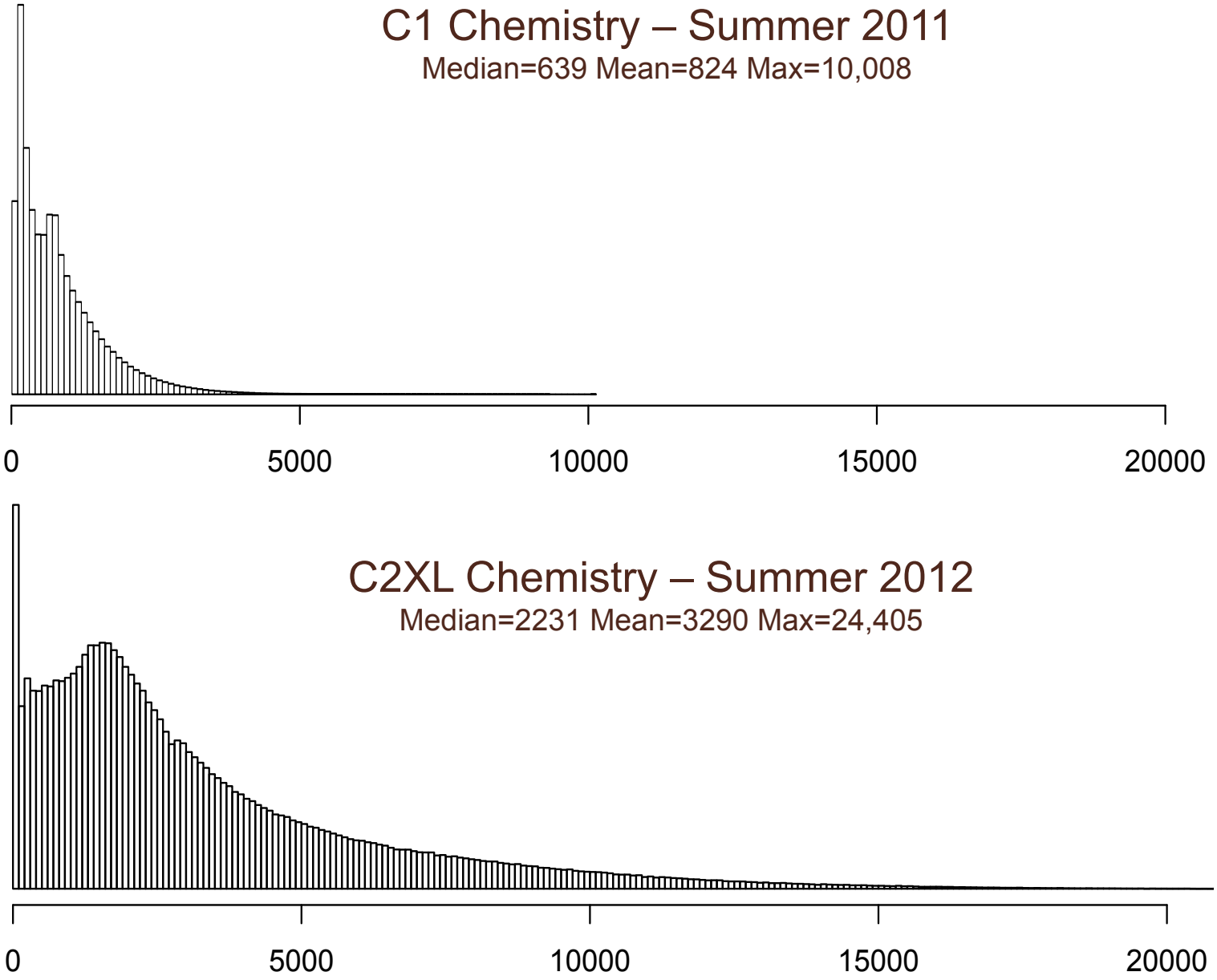
2. "Hot Start" technology:

   Maximize subreads

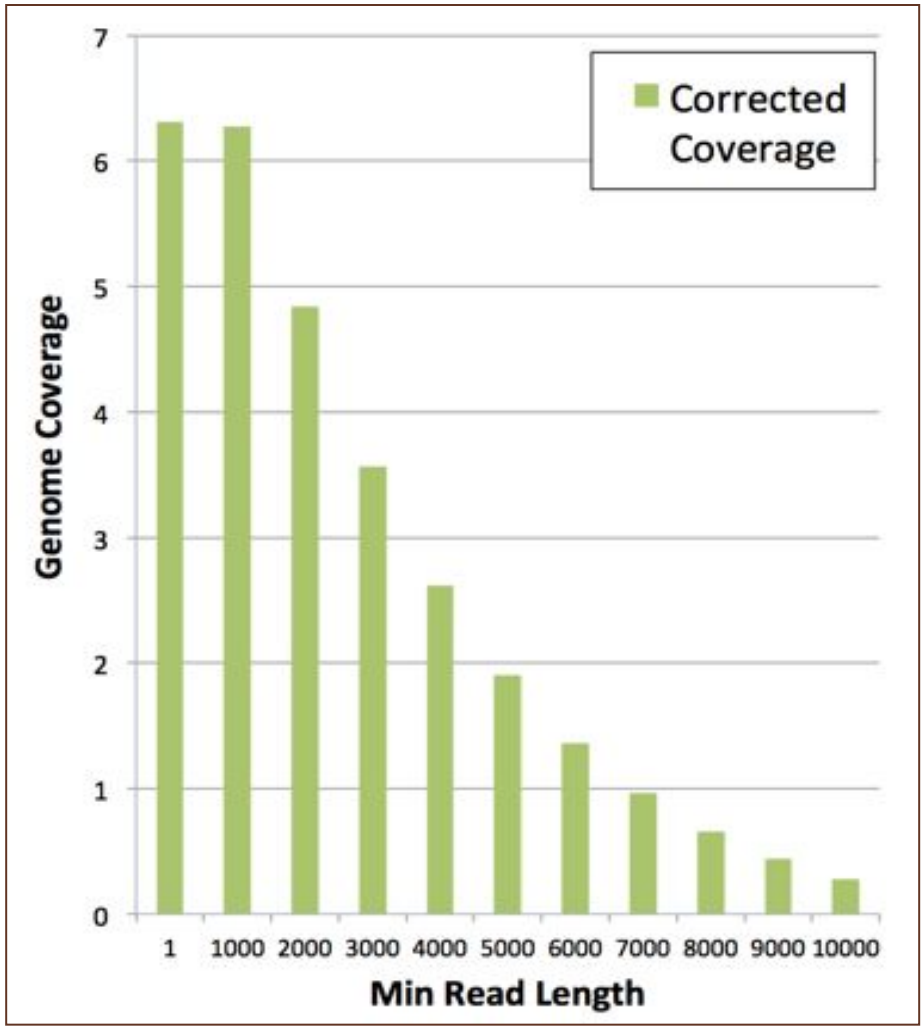
3. MagBead loading:

   Load longest fragments

PacBio Long Read Rice Sequencing



Preliminary Rice Assemblies



Assembly	Contig N50
Illumina Fragments 50x 2x100bp @ 180	3,925
MiSeq Fragments 23x 459bp 8x 2x251bp @ 450	6,444
PBeCR Reads 6.3x 2146bp ** MiSeq for correction	13,600
Illumina Mates 50x 2x100bp @ 180 36x 2x50bp @ 2100 51x 2x50bp @ 4800	13,696
PBeCR + Illumina Shred 6.3x 2146bp ** MiSeq for correction 51x 2x50bp @ 4800	25,108

In collaboration with McCombie & Ware labs @ CSHL

THE ASSEMBLATHON

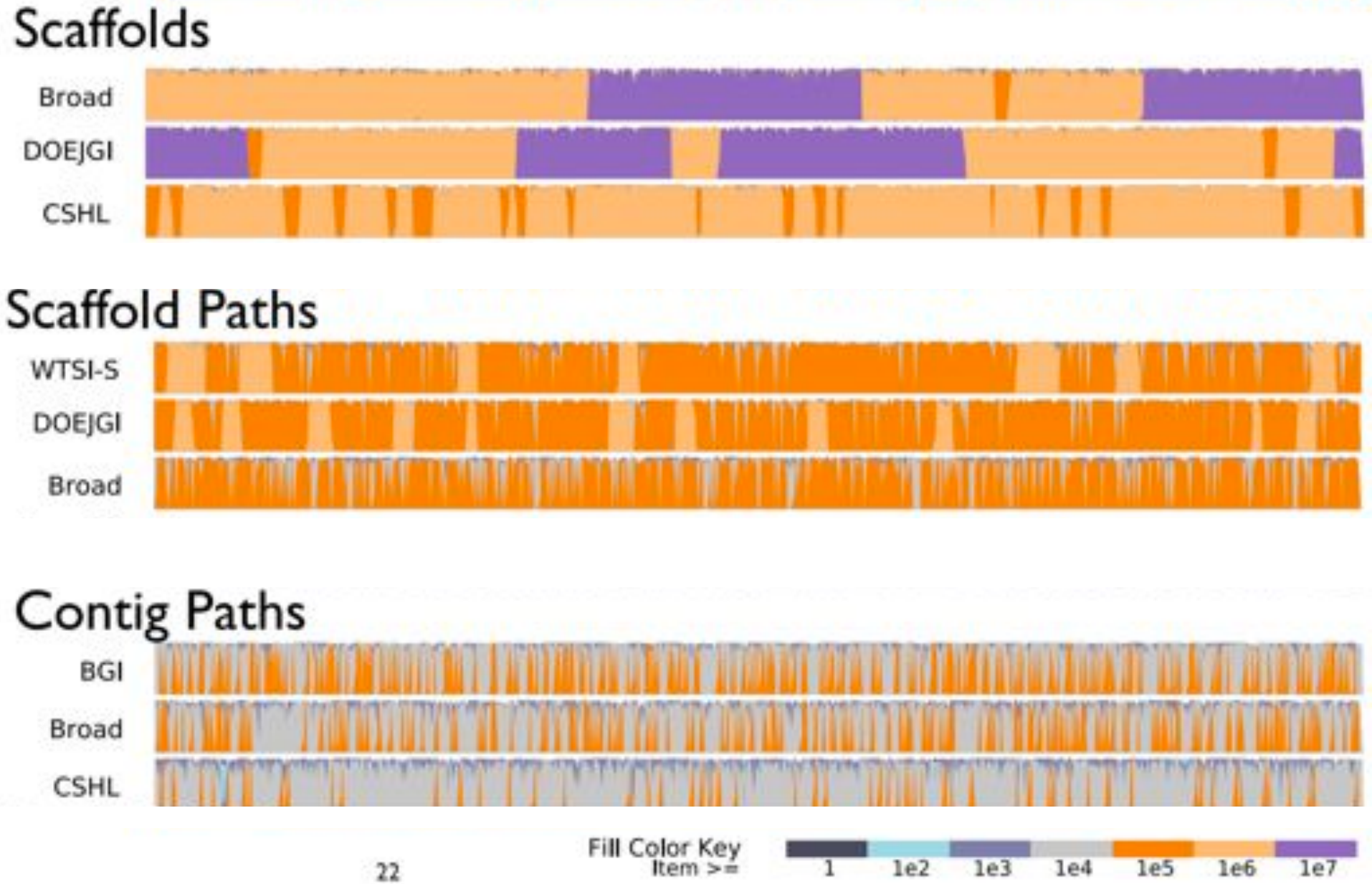
- Attempt to answer the question:
  "What makes a good assembly?"

- Organizers provided simulated sequence data
  - Simulated 100 base pair Illumina reads from simulated diploid organism

- 41 submissions from 17 groups

- Results demonstrate trade-offs assemblers must make

Assembly Results

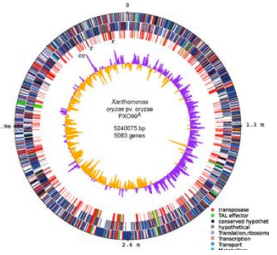


Final Rankings

ID	Overall	CPNG50	SPNG50	Struct.	CC50	Subs.	Copy. Num.	Cov. Tot.	Cov. CDS
BGI	36	★					☆	★	☆
Broad	37	☆	★	★	★				
WTSI-S	46		★	☆	★	★			
CSHL	52	★							☆
BCCGSC	53							☆	★
DOEJGI	56		☆	★	☆	★			
RHUL	58								
WTSI-P	64							☆	
EBI	64						★		
CRACS	64					☆			

- SOAPdenovo and ALLPATHS came out neck-and-neck followed closely behind by SGA, Celera Assembler, ABySS
- My recommendation for "typical" short read assembly is to use ALLPATHS

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

Break

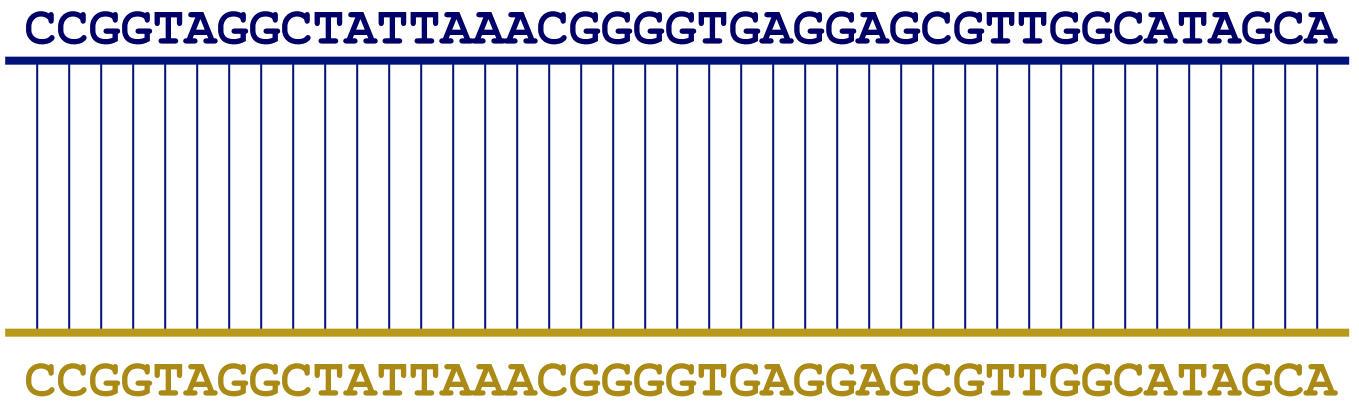

READING THE ENTIRE GENOME CODE...
Listen to this part: 'AGGCTAATC CGCATAACTG'
Wow!

Whole Genome Alignment with MUMmer

Slides Courtesy of Adam M. Phillippy
amp@umics.umd.edu

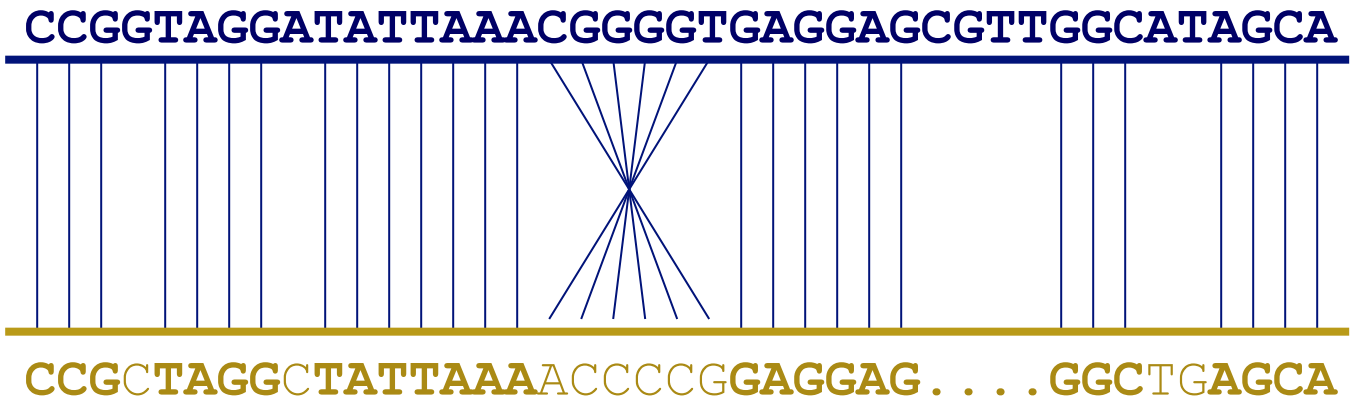
Goal of WGA

- For two genomes, *A* and *B*, find a mapping from each position in *A* to its corresponding position in *B*



Not so fast...

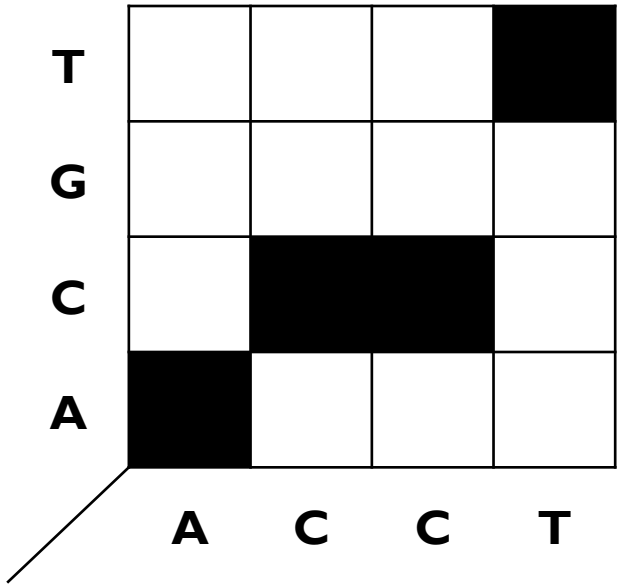
- Genome *A* may have insertions, deletions, translocations, inversions, duplications or SNPs with respect to *B* (sometimes all of the above)



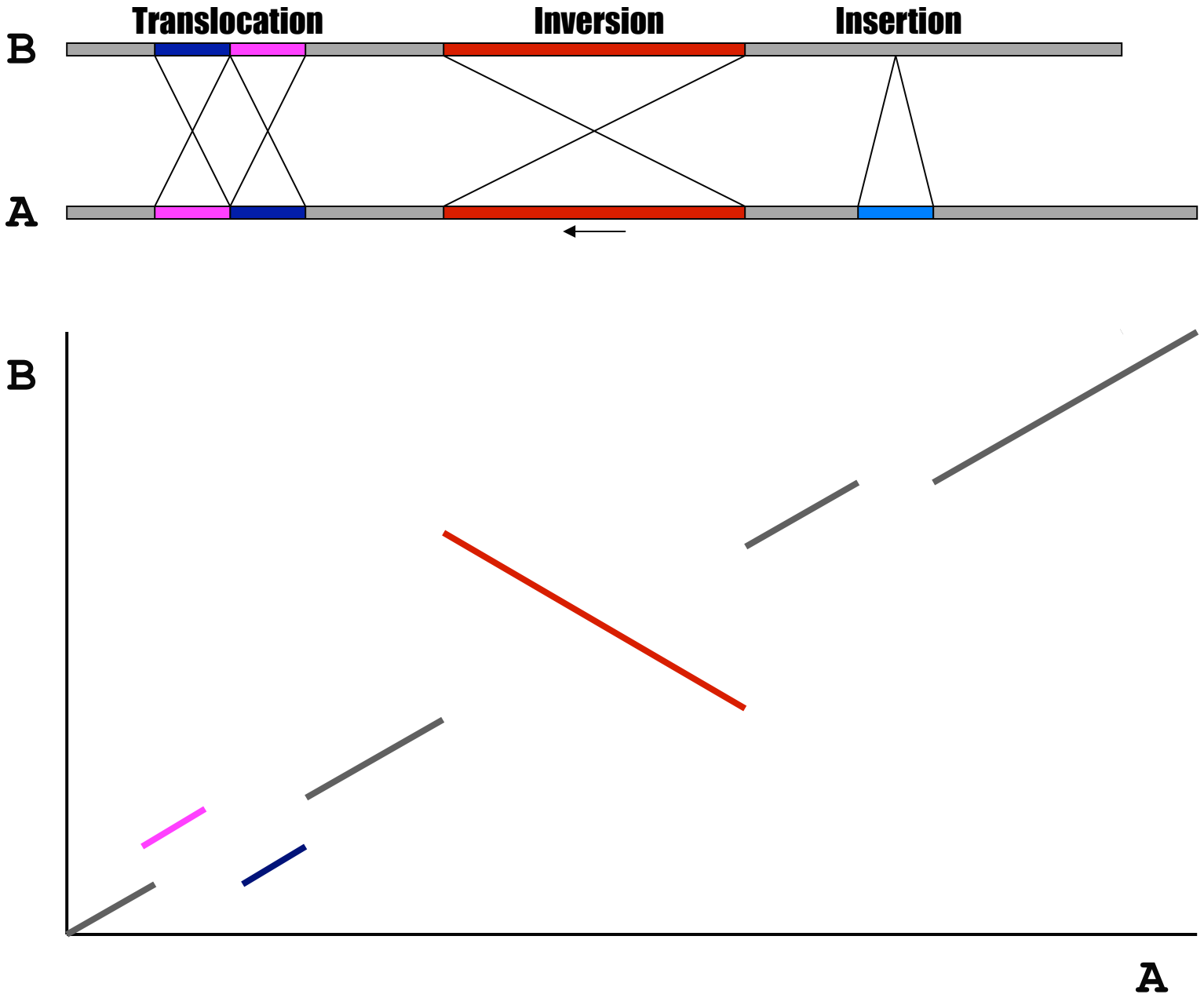
WGA visualization

- How can we visualize *whole* genome alignments?

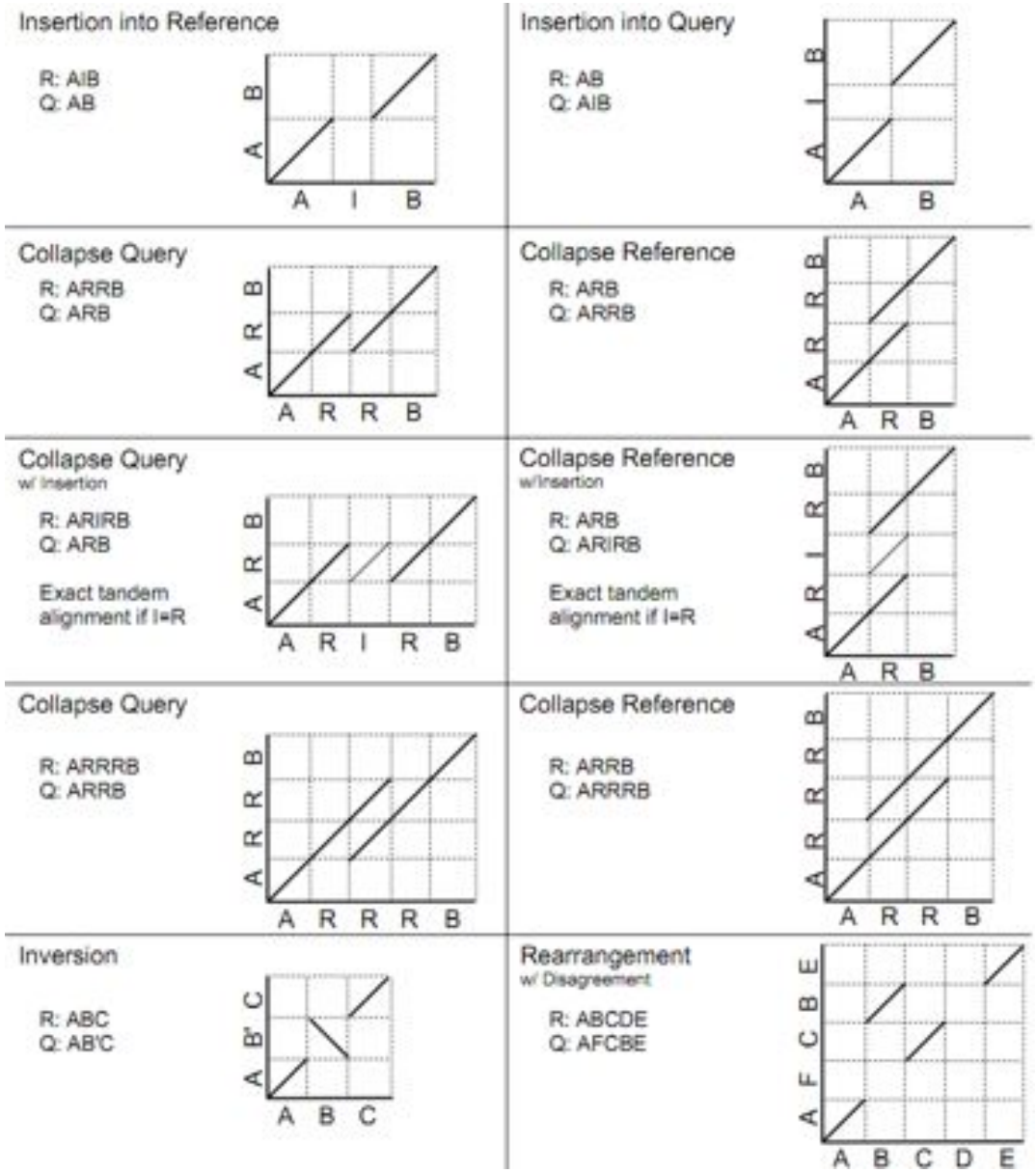
- With an alignment dot plot
  - *N* x *M* matrix
    - Let *i* = position in genome *A*
    - Let *j* = position in genome *B*
    - Fill cell *(i,j)* if A_i shows similarity to B_j



  - 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

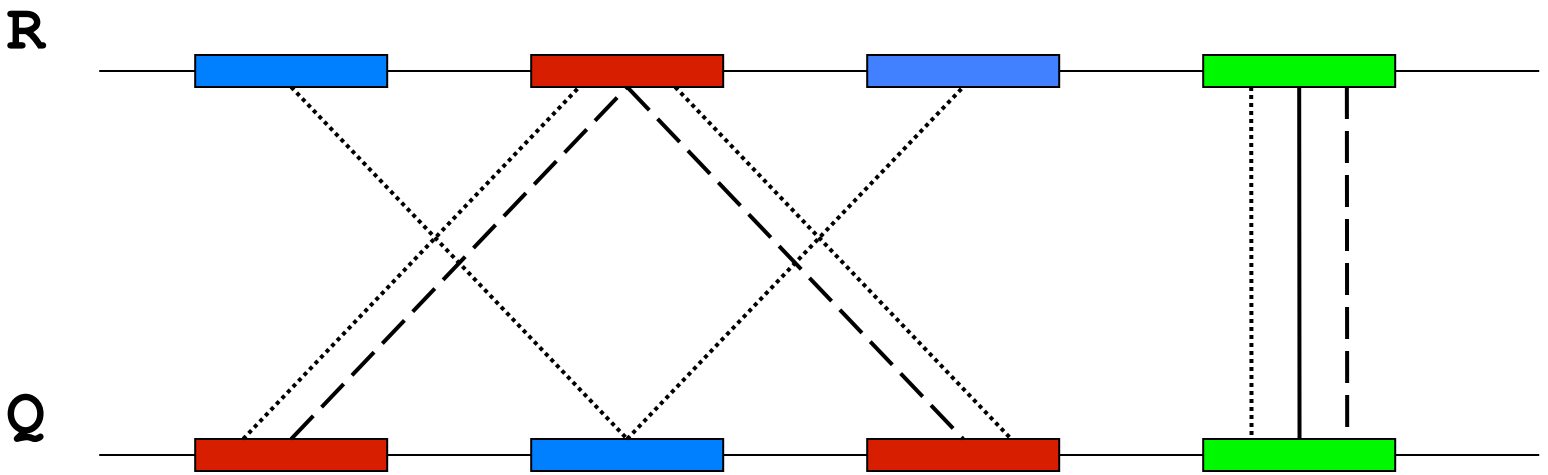
Fee Fi Fo Fum,

is it a MAM, MEM or MUM?

MUM : maximal unique match

MAM : maximal almost-unique match

MEM : maximal exact match



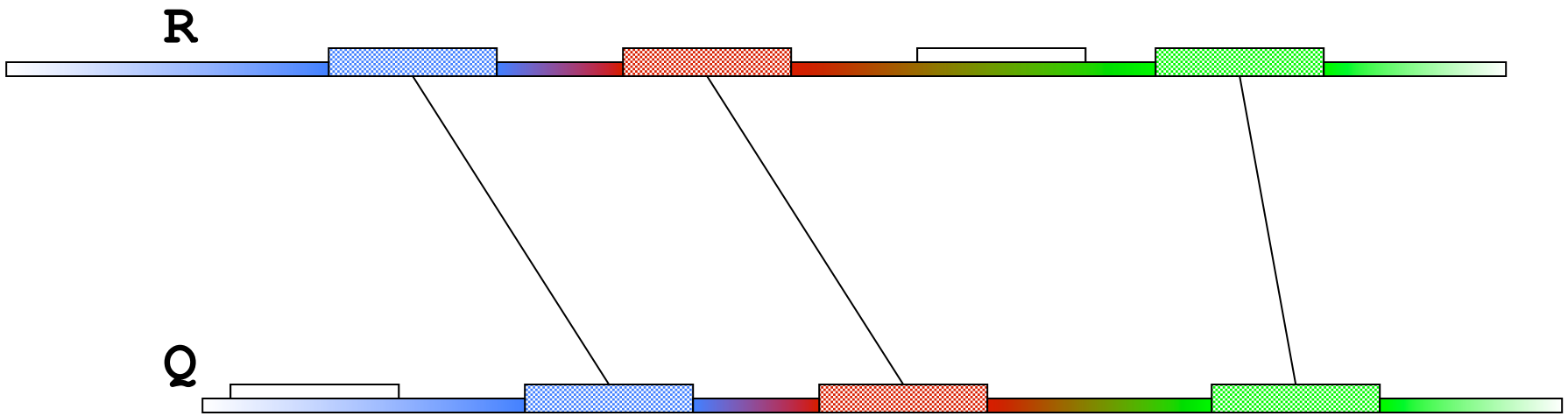
Seed and Extend

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
```

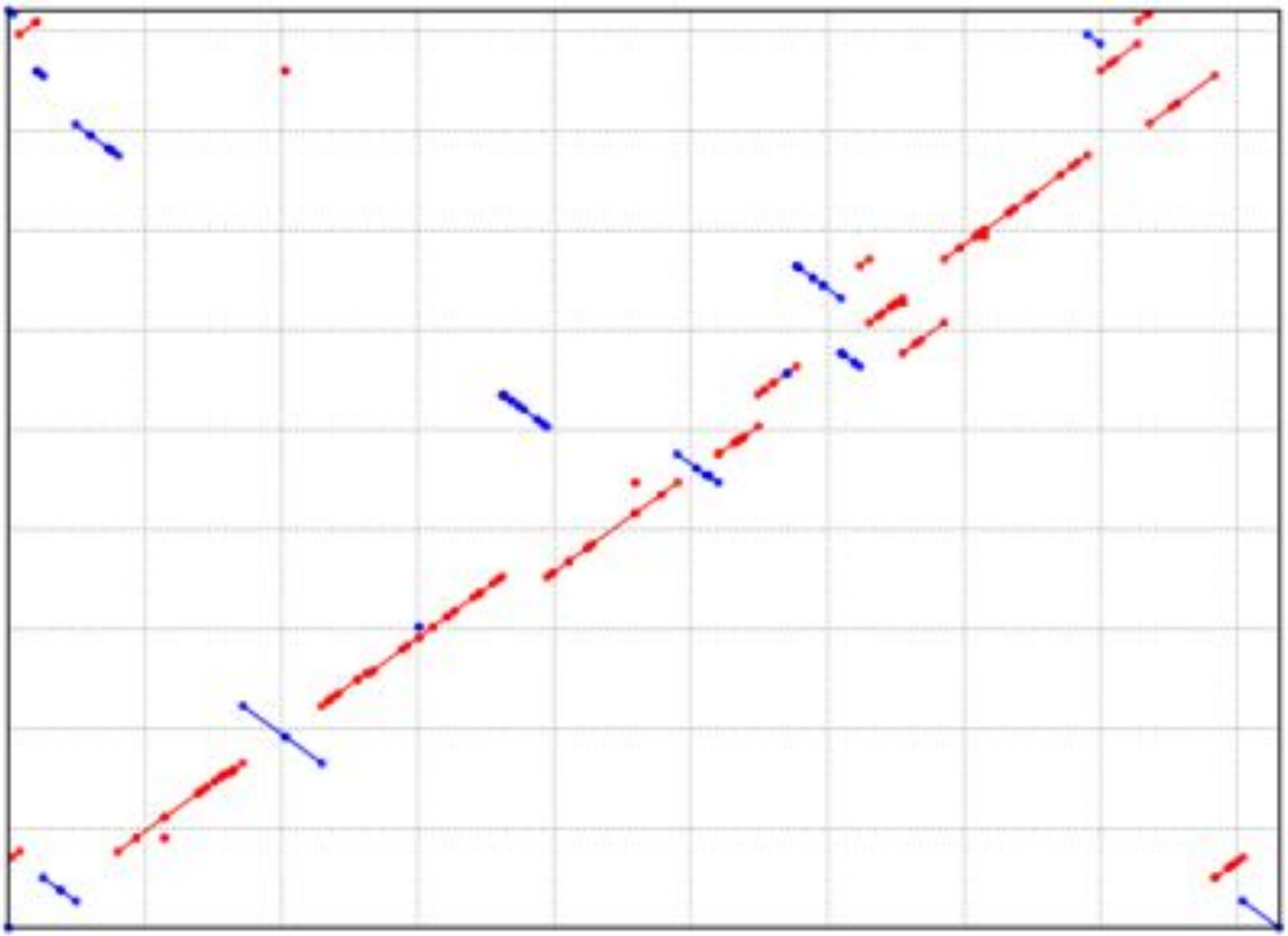
dnadiff out.delta.m

```
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



References

- Documentation
  - http://mummer.sourceforge.net
    - » publication listing
  - http://mummer.sourceforge.net/manual
    - » documentation
  - http://mummer.sourceforge.net/examples
    - » walkthroughs

- Email
  - mummer-help@lists.sourceforge.net

Acknowledgements

Schatz Lab

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

CSHL

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

NBACC

Adam Phillippy
Sergey Koren

JHU/UMD

Steven Salzberg
Mihai Pop
Ben Langmead
Cole Trapnell

Thank You!

http://schatzlab.cshl.edu/