Assembling crop genomes with 2nd and 3rd generation sequencing

Michael Schatz

Oct 8, 2012
Strategies for de novo assemblies of complex crop genomes
The Genome Analysis Center, Norwich Research Park

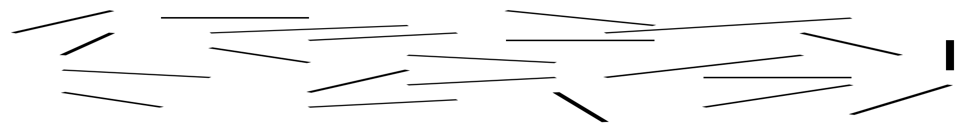
#ESFCrops / @mike_schatz

Outline

1. Ingredients for a good assembly
2. 2nd Generation Sequencing & Assembly
    1. Sacred Lotus
    2. Raspberry
    3. Wheat
3. 3rd Generation Sequence & Assembly
    1. Parrot
    2. Rice

Assembling a Genome

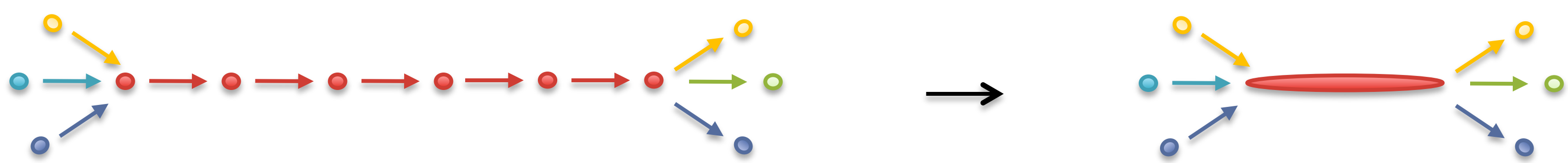
1. Shear & Sequence DNA



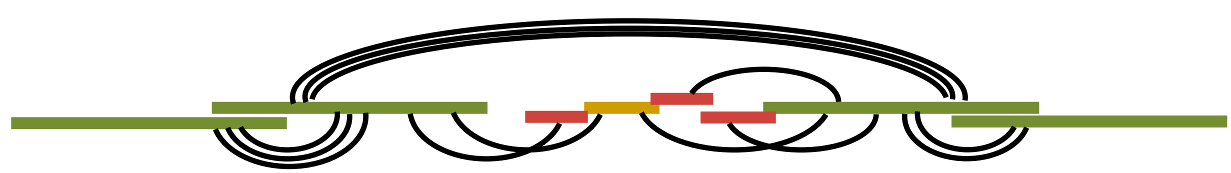
2. Construct assembly graph from overlapping reads

...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

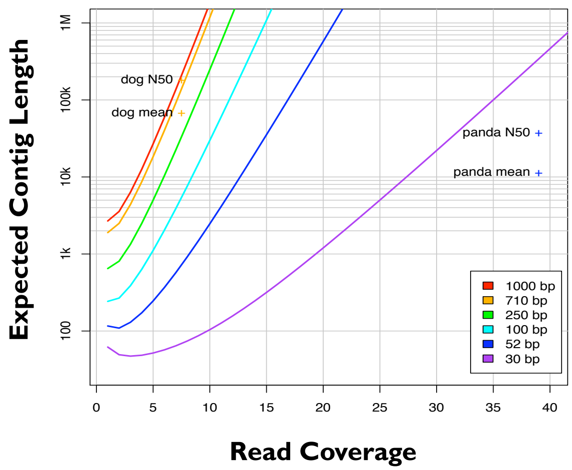
2. ***Sequencing***:
   – (Very) large genomes, imperfect sequencing

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

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

Ingredients for a good assembly

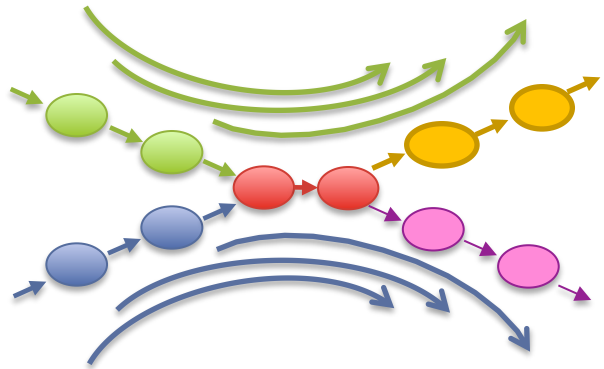
Coverage



High coverage is required

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

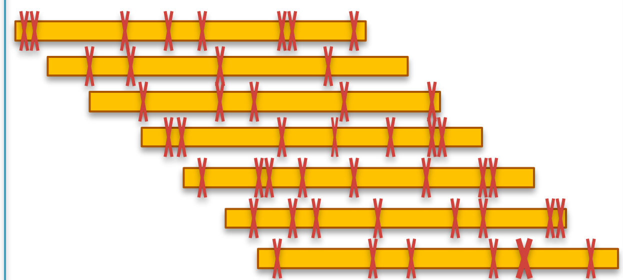
Read Length



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



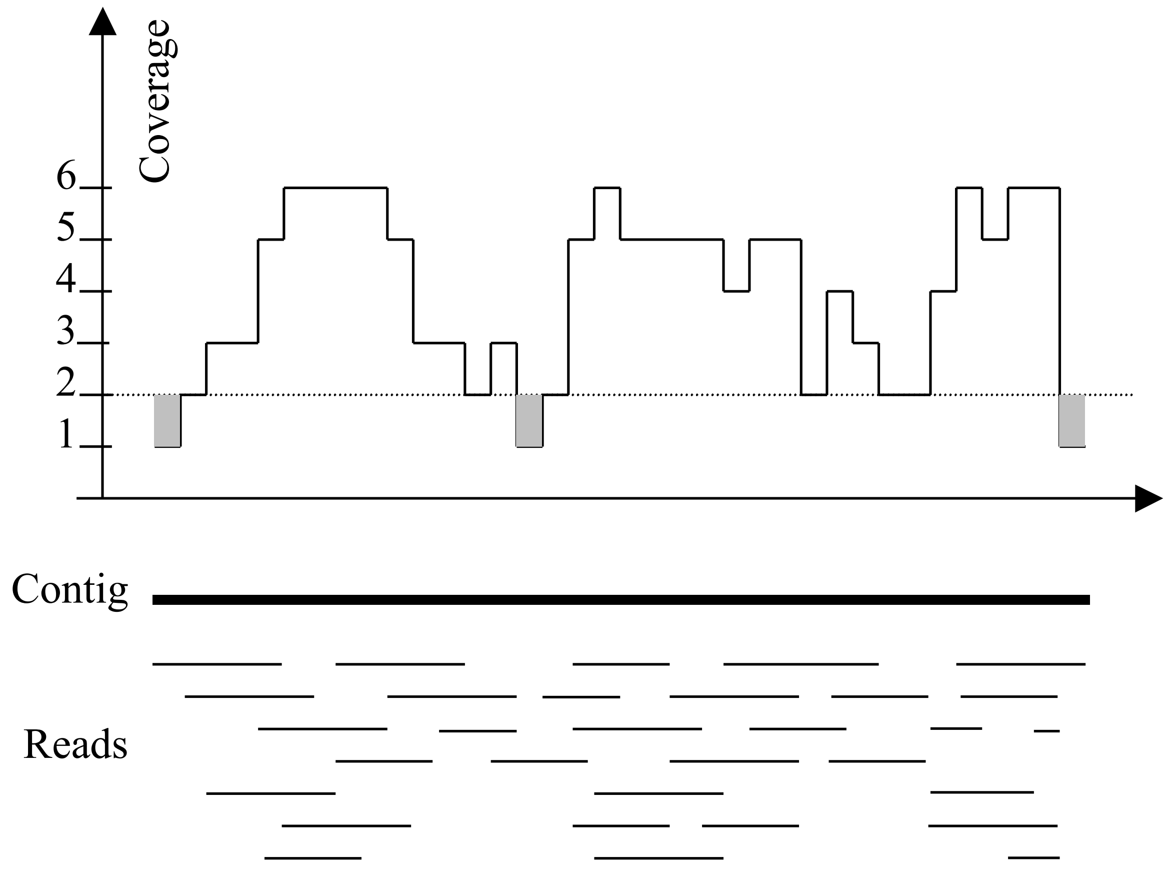
Errors obscure overlaps

- 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

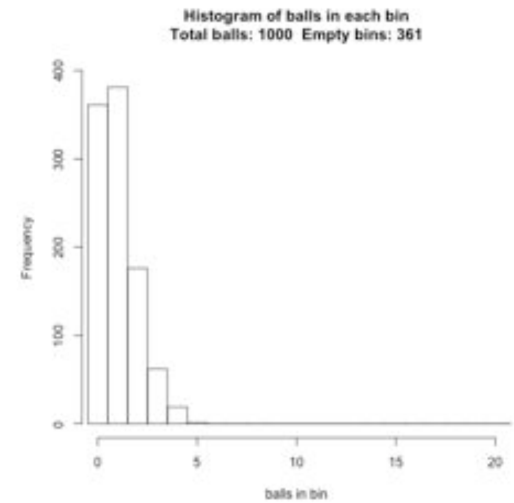
Coverage

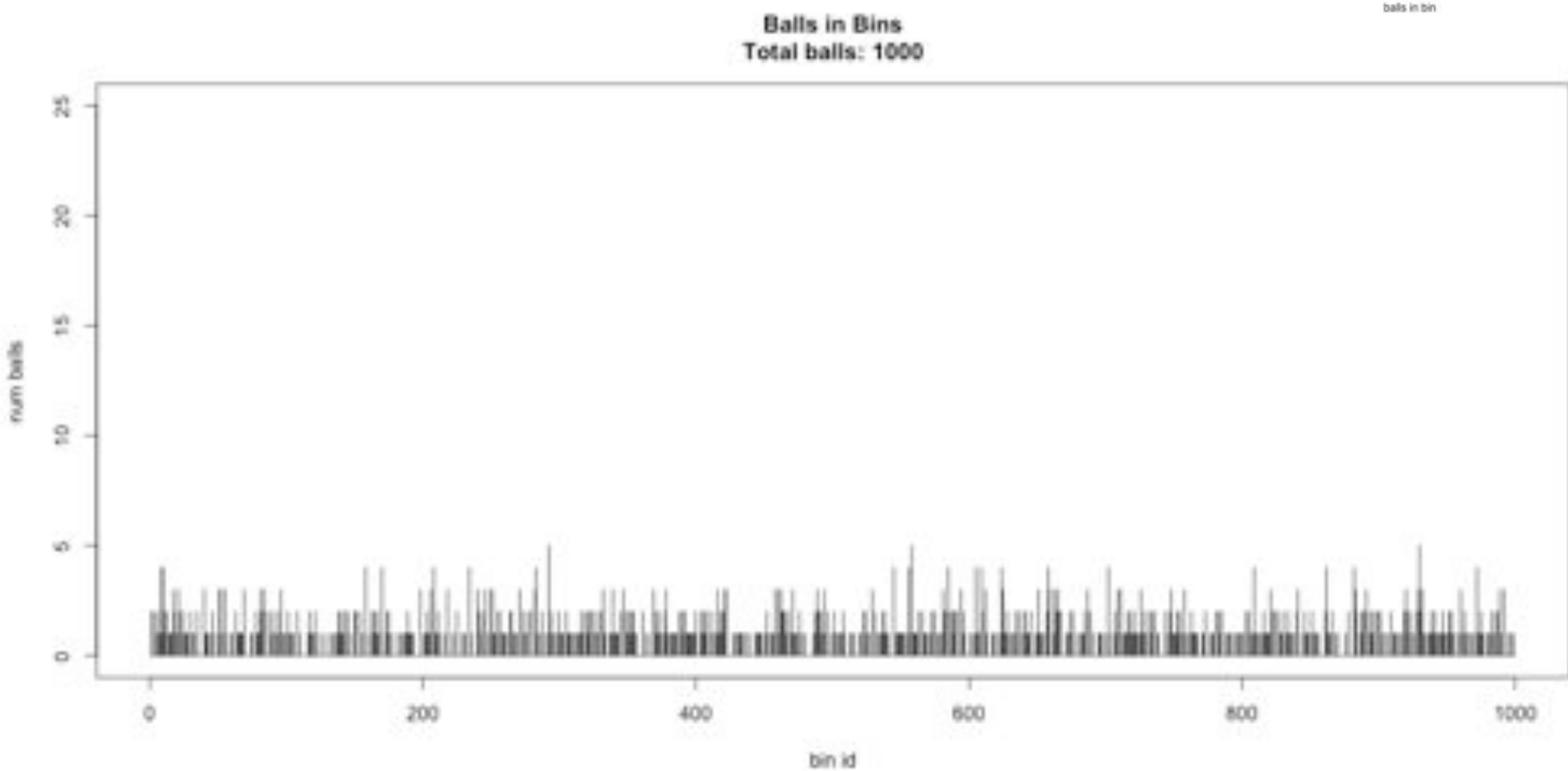
Typical contig coverage



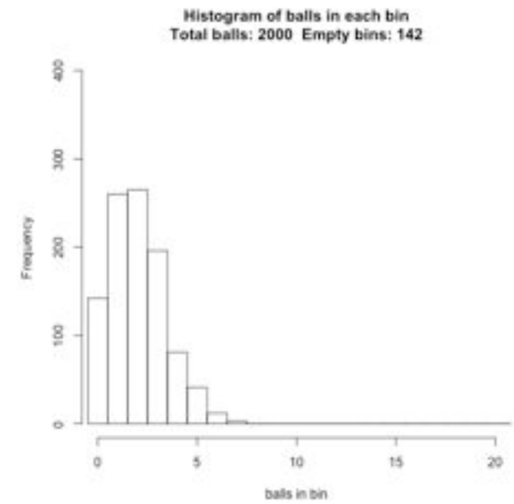
Imagine raindrops on a sidewalk

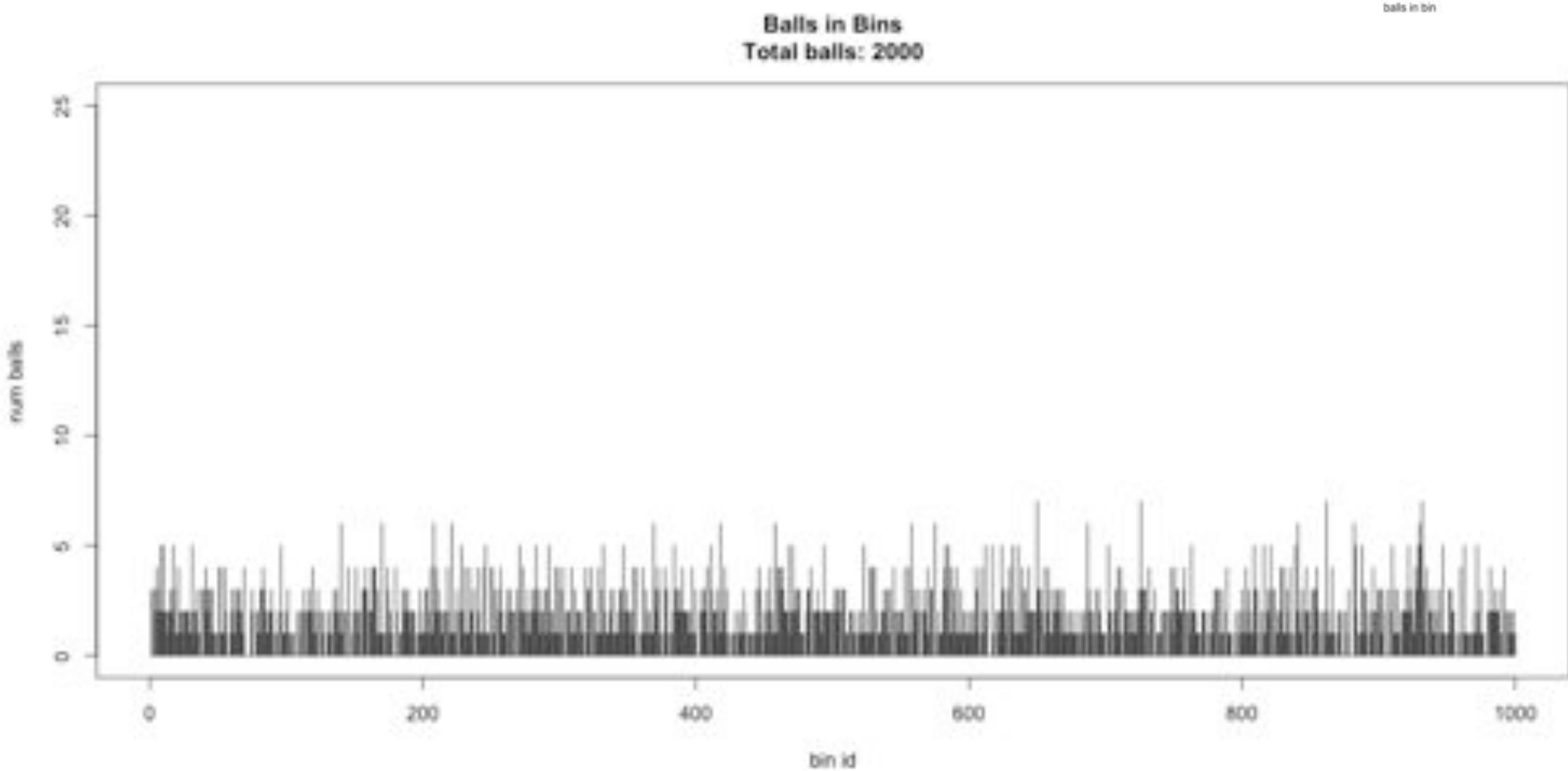
Balls in Bins 1x



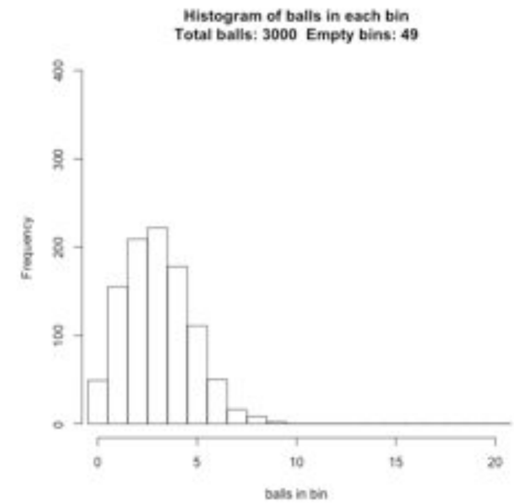


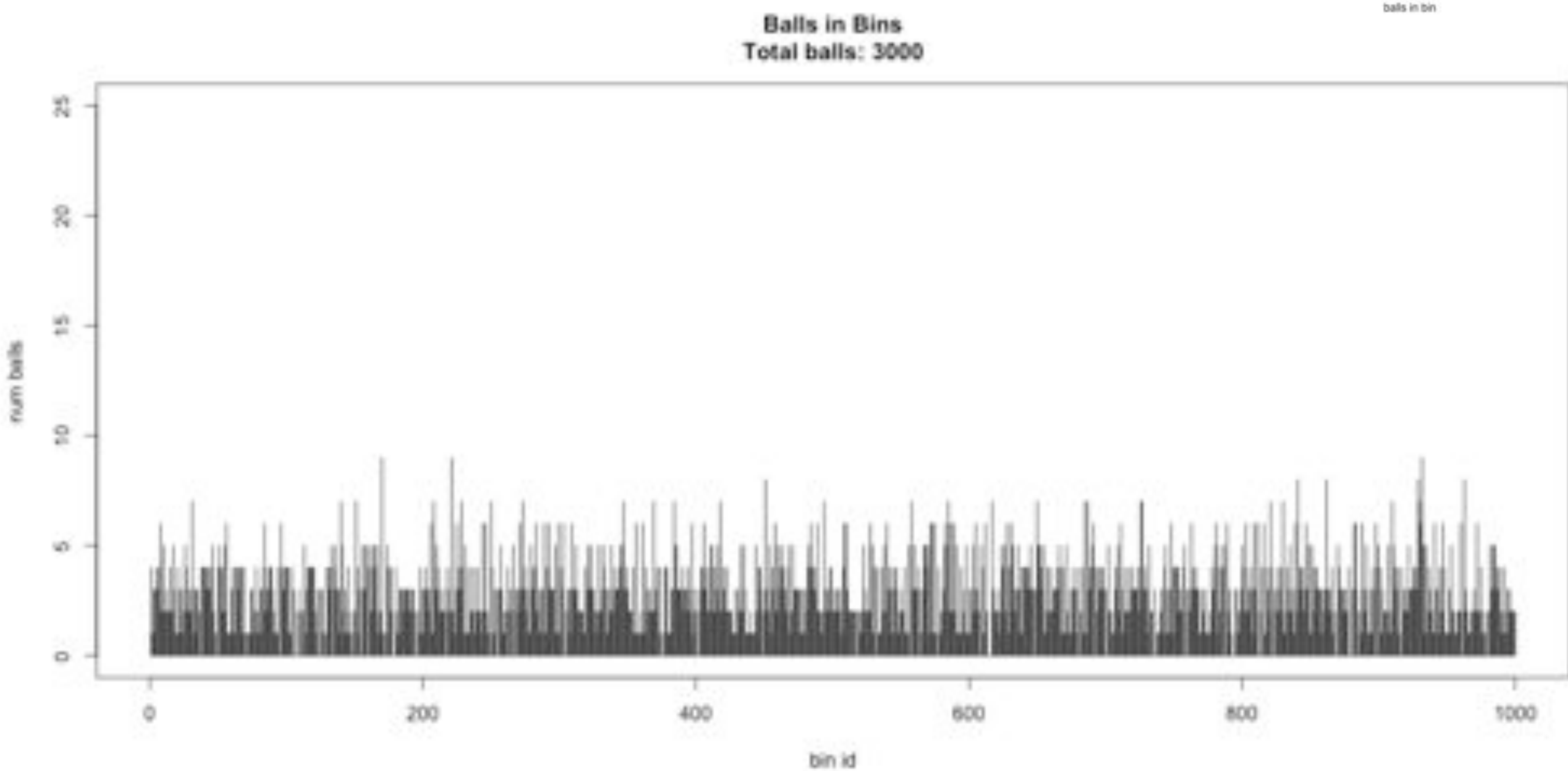
Balls in Bins 2x



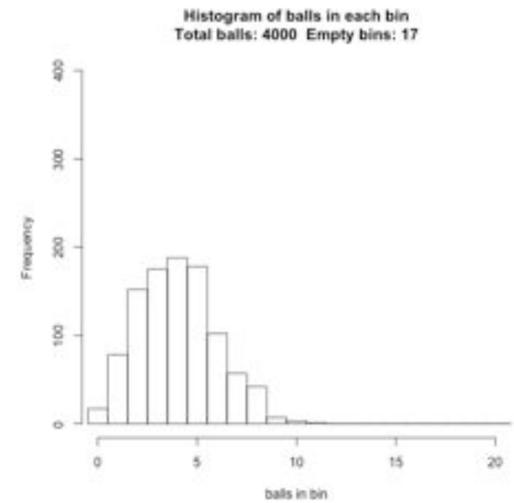


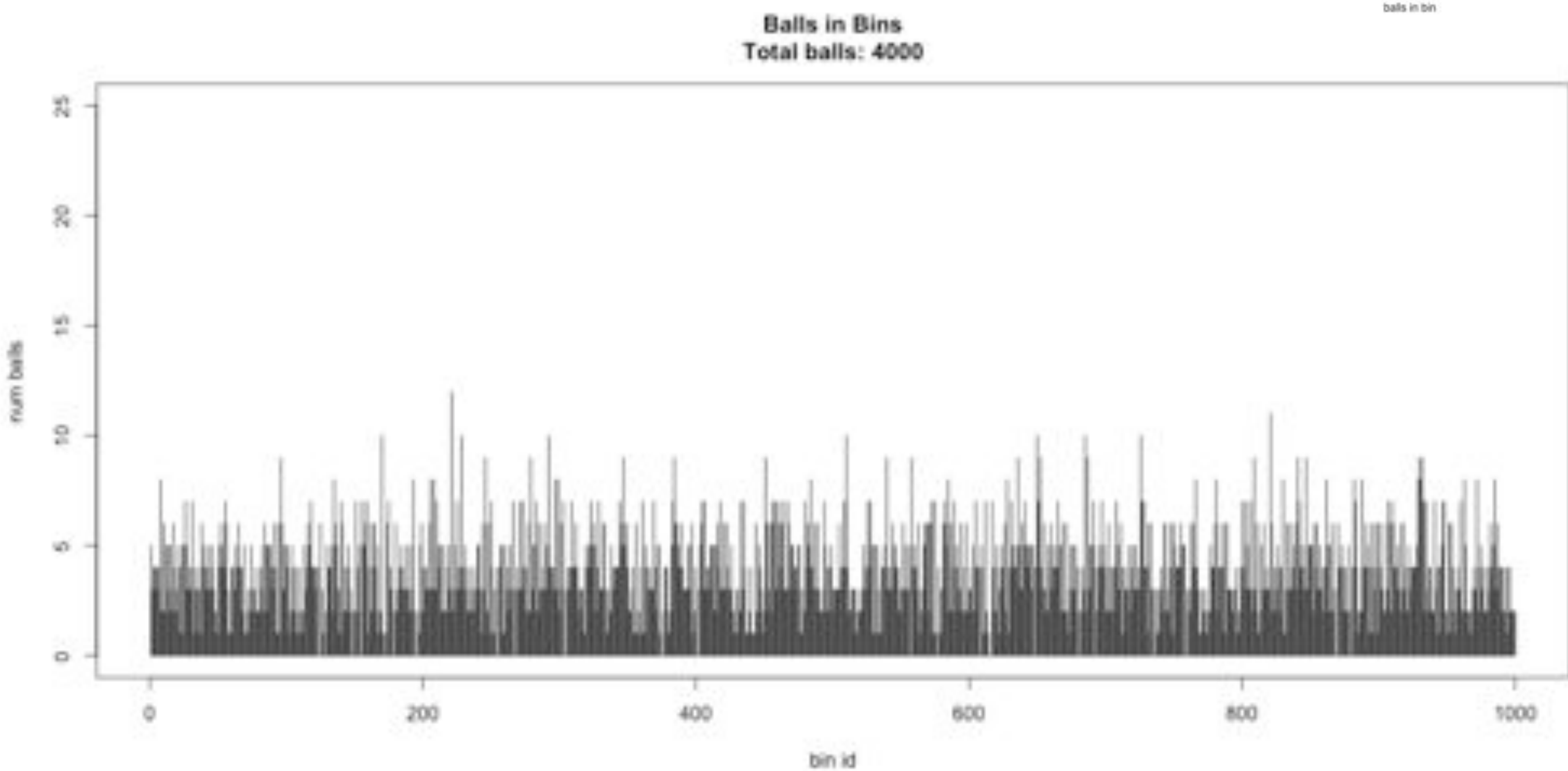
Balls in Bins 3x



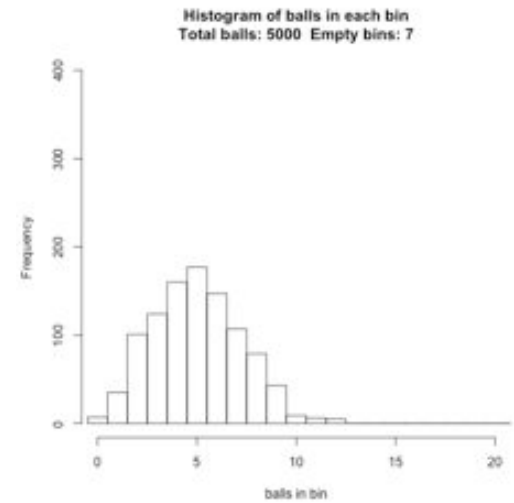


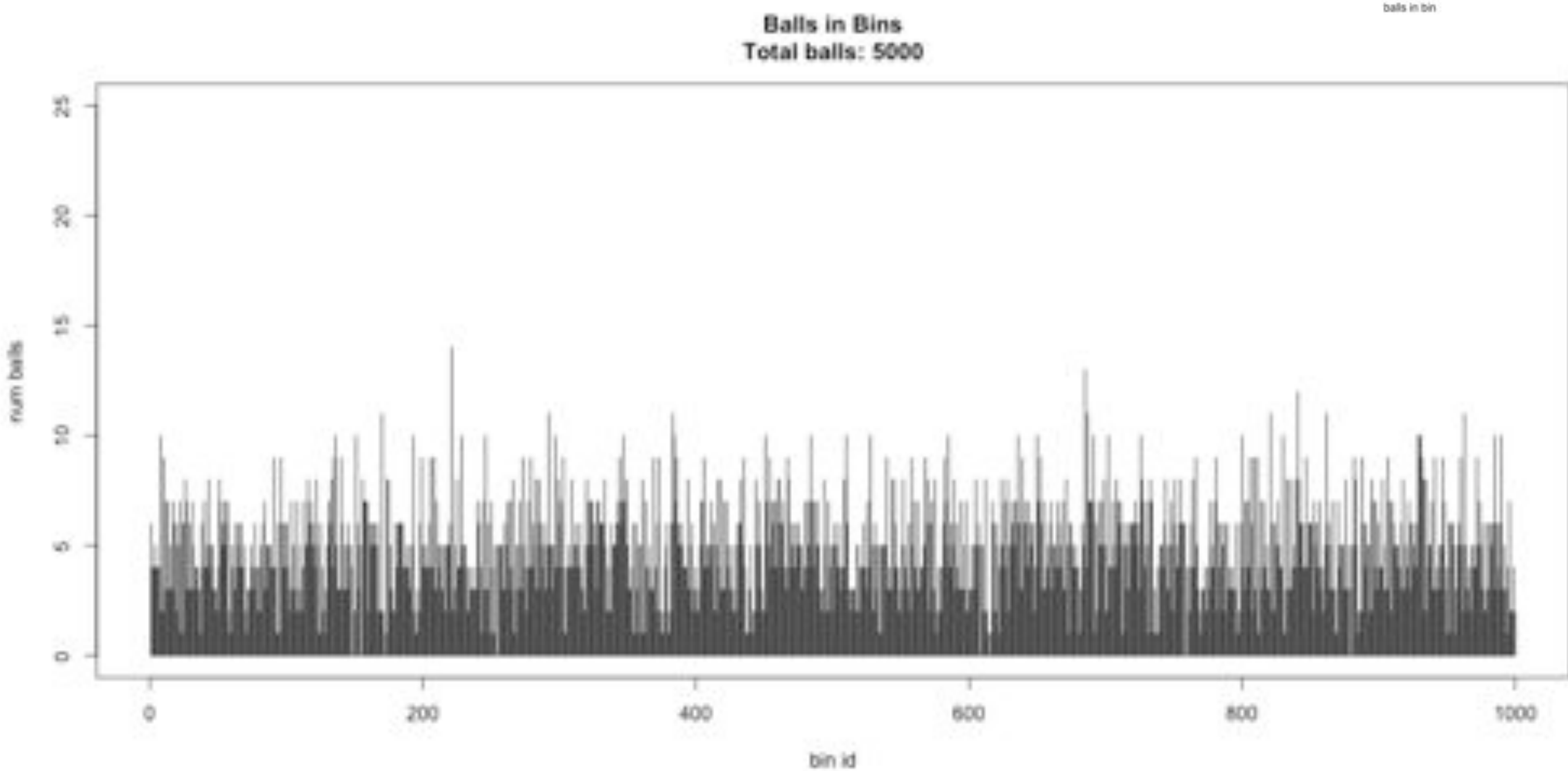
Balls in Bins 4x



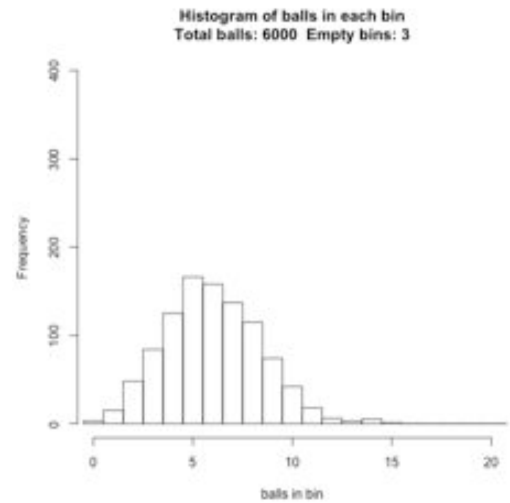


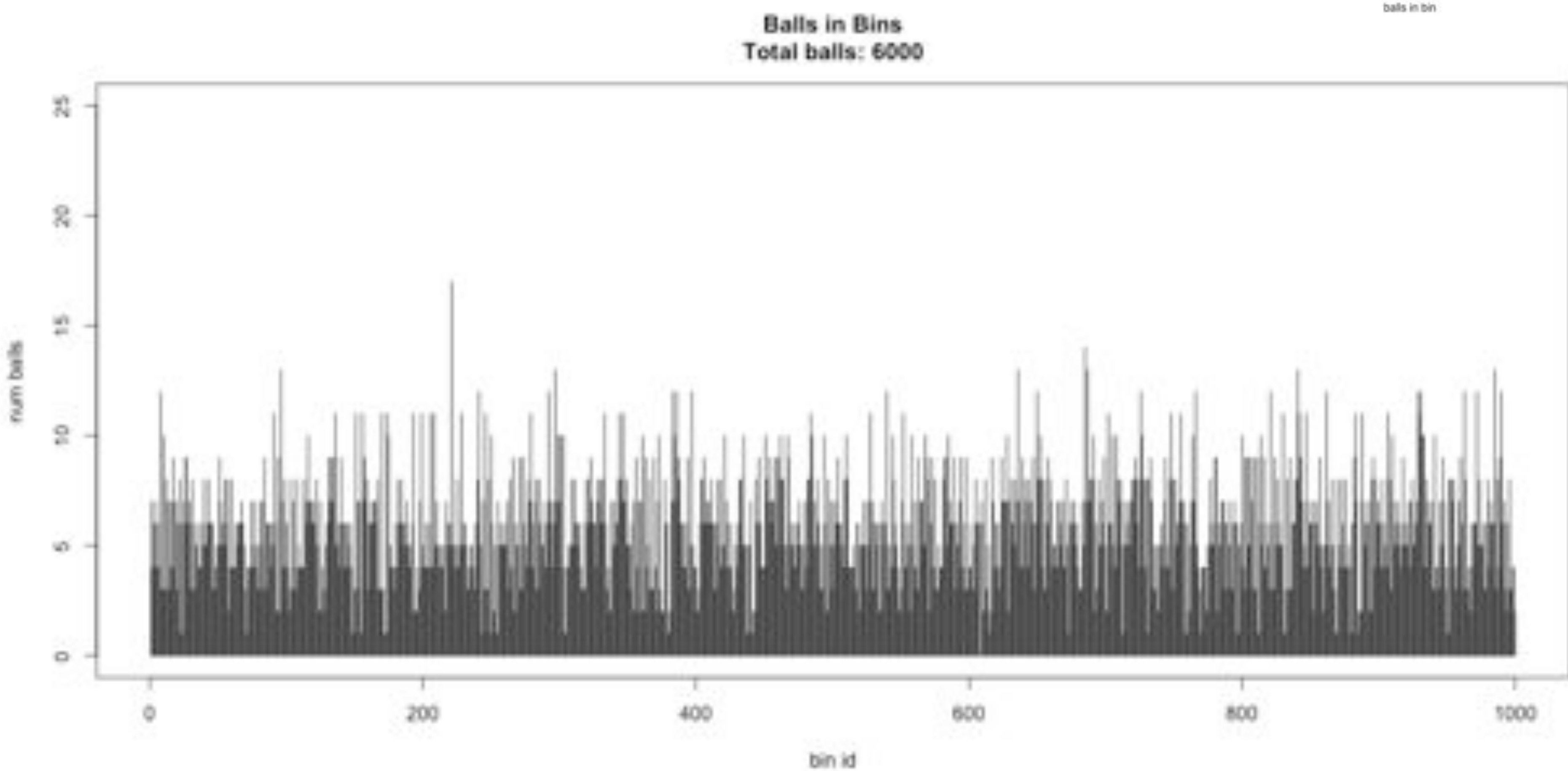
Balls in Bins 5x



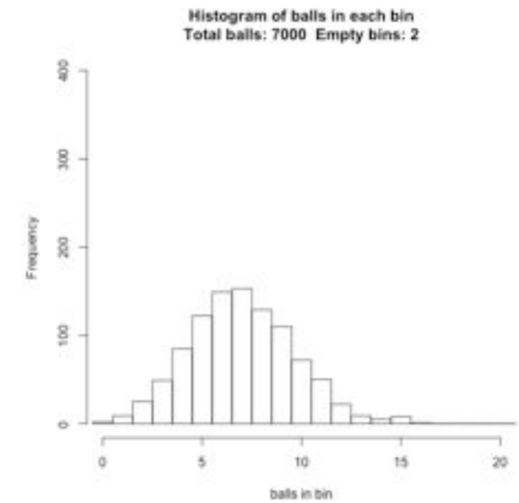


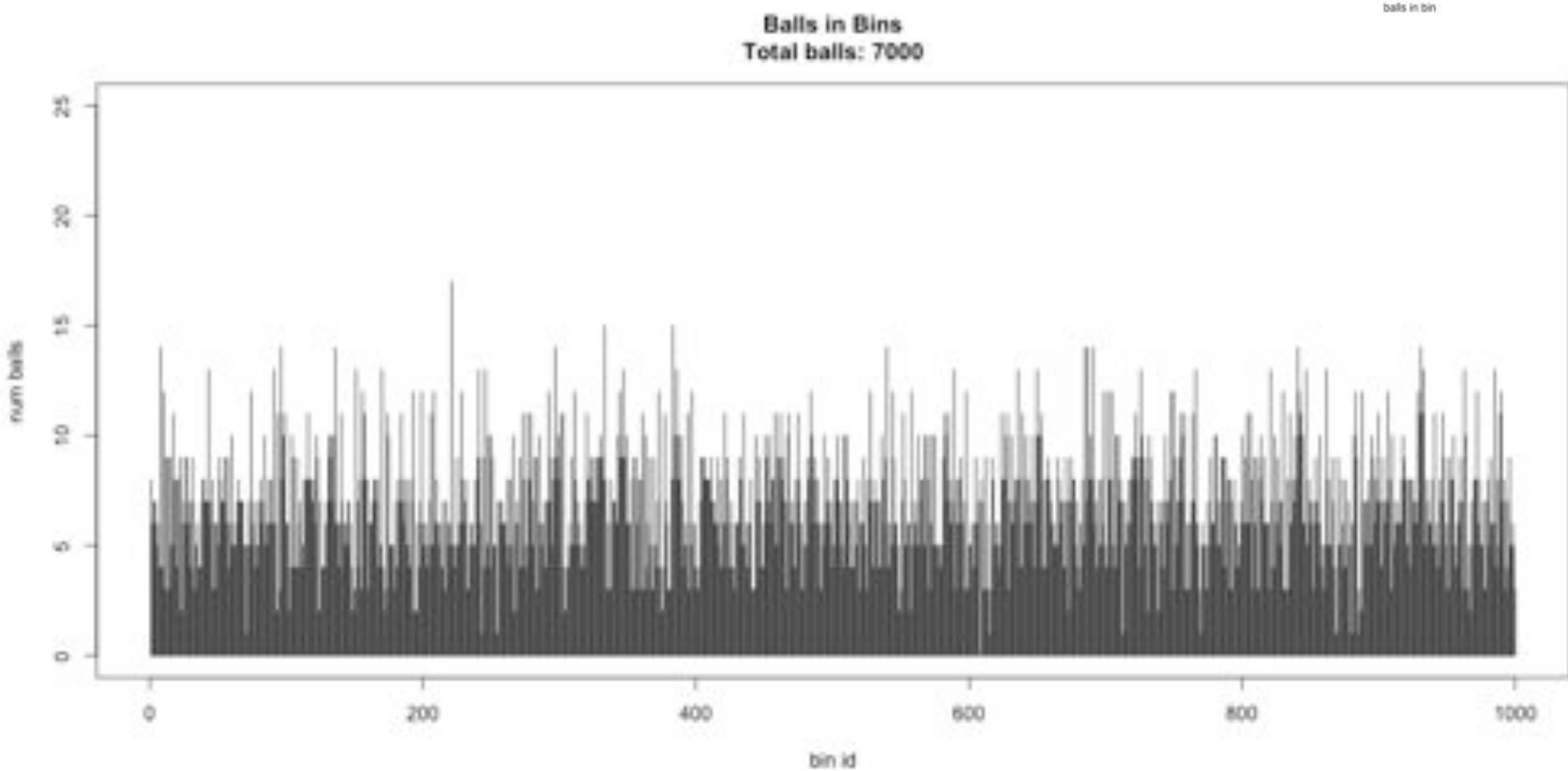
Balls in Bins 6x



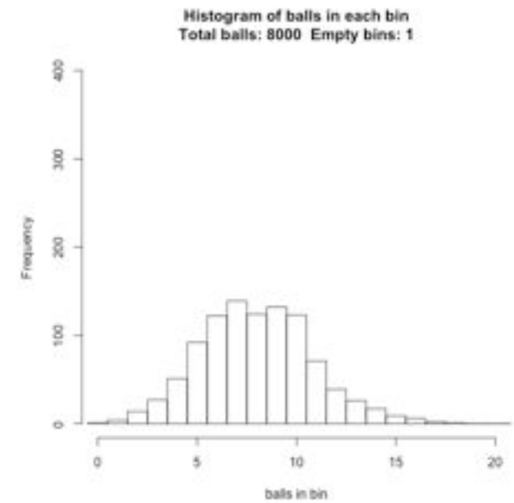


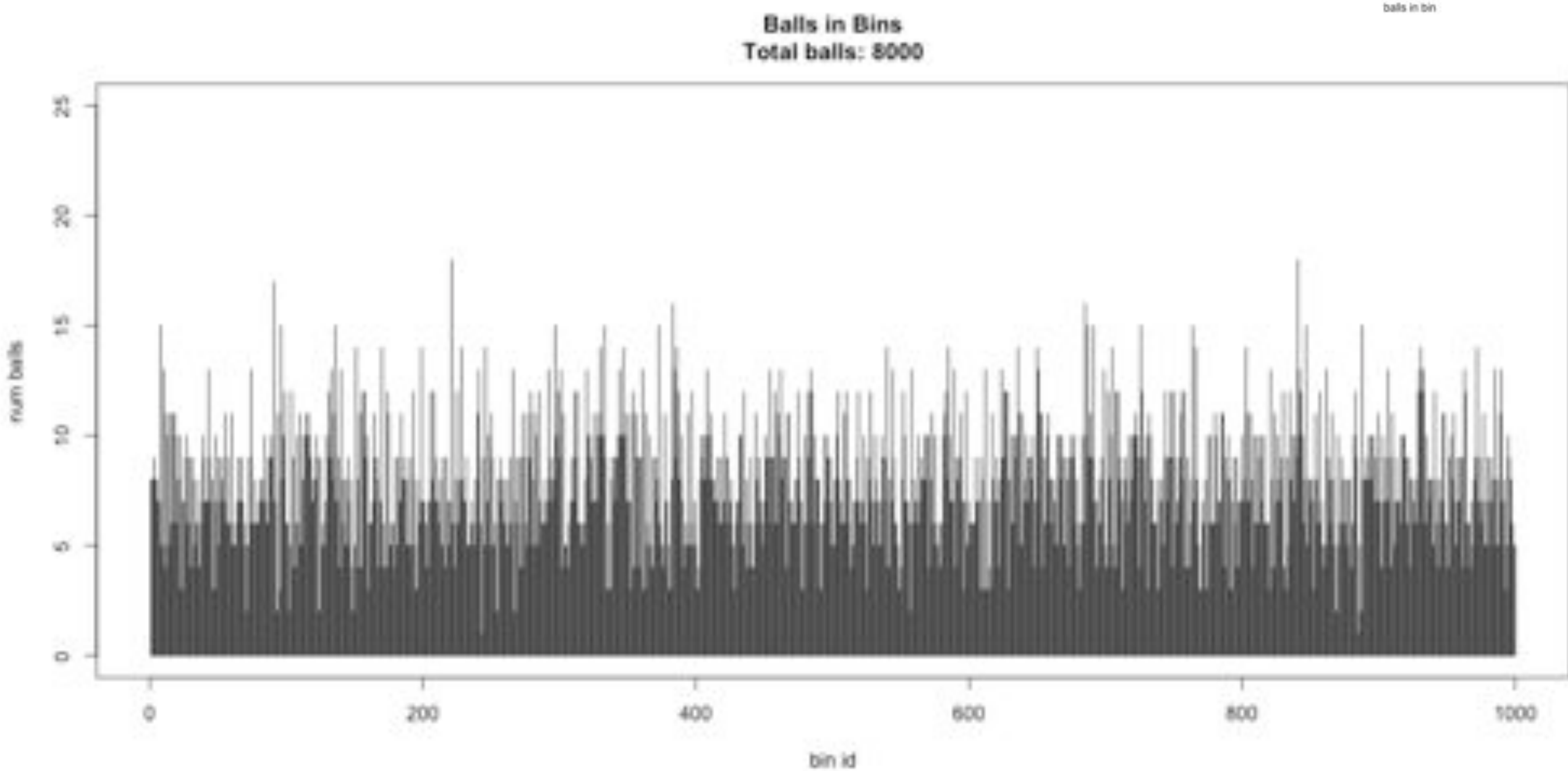
Balls in Bins 7x





Balls in Bins 8x

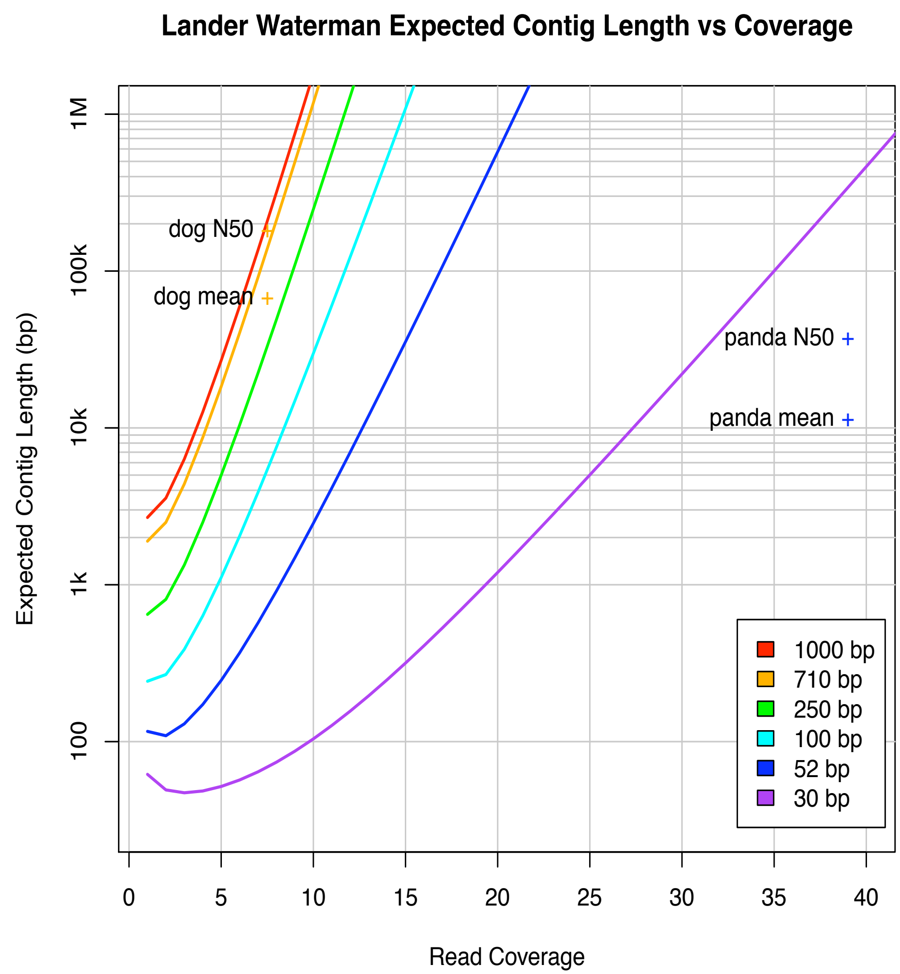




Coverage and Read Length

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

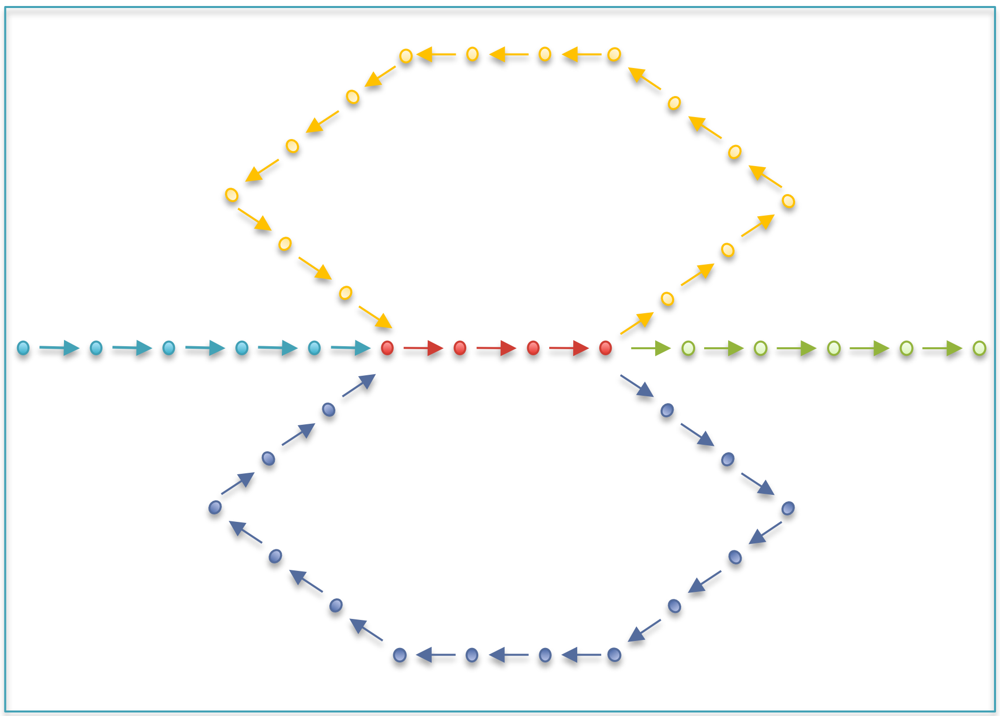


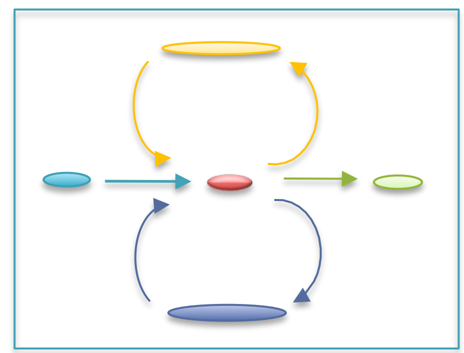
Assembly of Large Genomes using Second Generation Sequencing
Schatz MC, Delcher AL, Salzberg SL (2010) *Genome Research.* 20:1165-1173.

Repeats

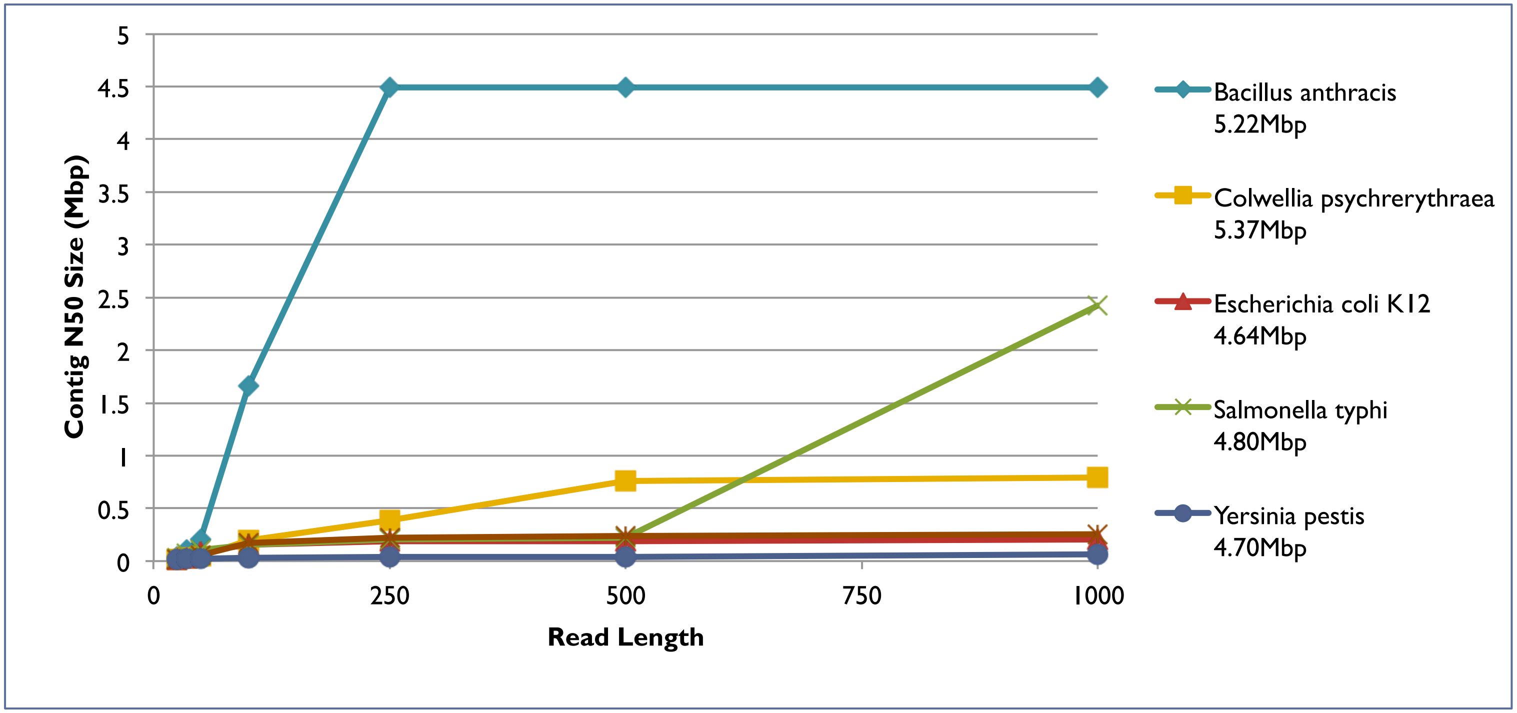
Unitigging / Unipathing

- After simplification and correction, compress graph down to its non-branching initial contigs
  - Aka "unitigs", "unipaths"





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

Quality

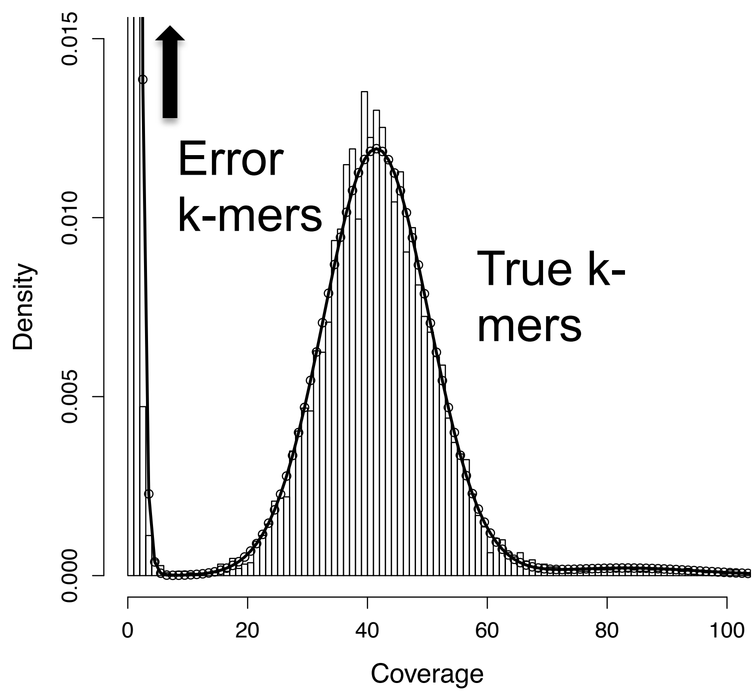
Error Correction with Quake

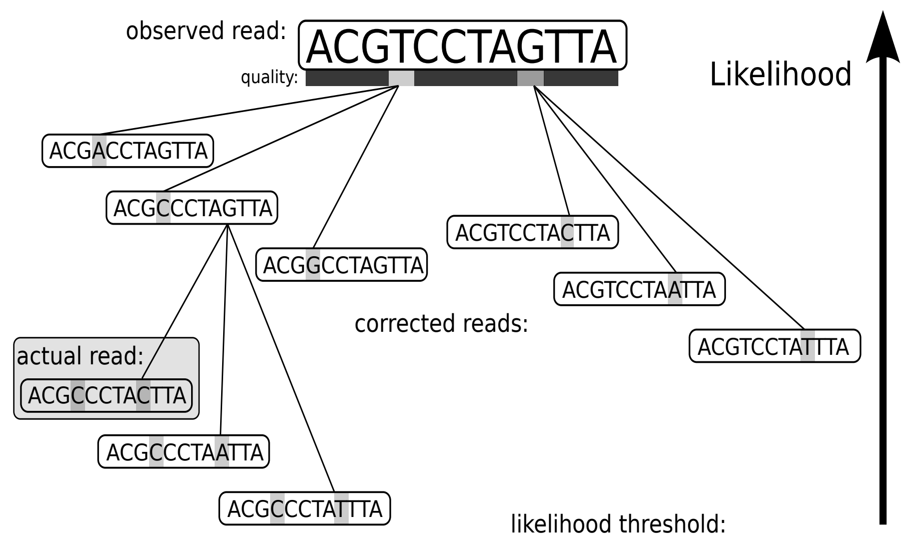
1. Count all "Q-mers" in reads

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

2. Correction Algorithm

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



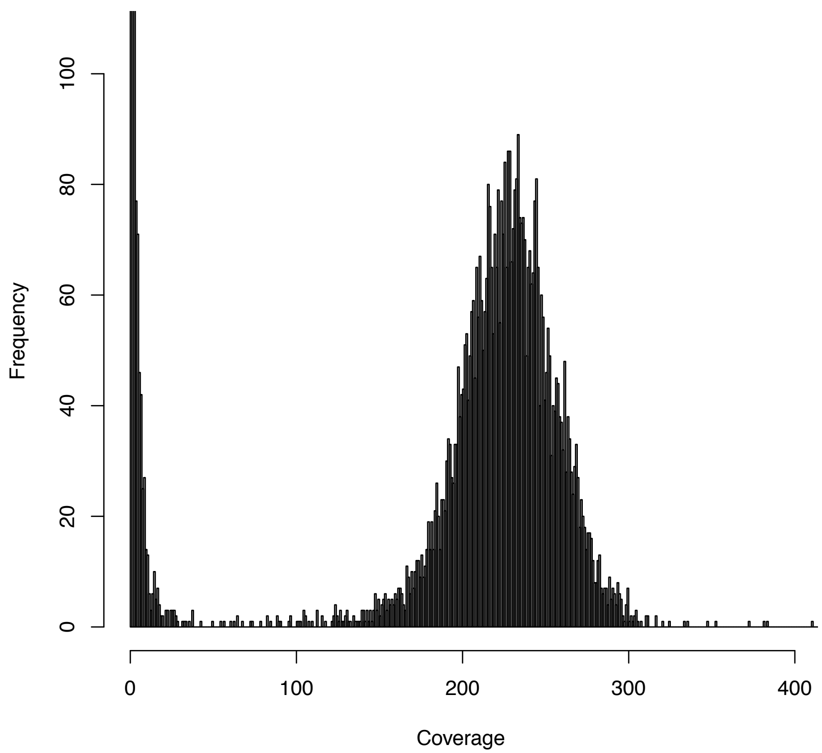


Quake: quality-aware detection and correction of sequencing reads.
Kelley, DR, Schatz, MC, Salzberg, SL (2010) *Genome Biology*. 11:R116

Illumina Sequencing & Assembly

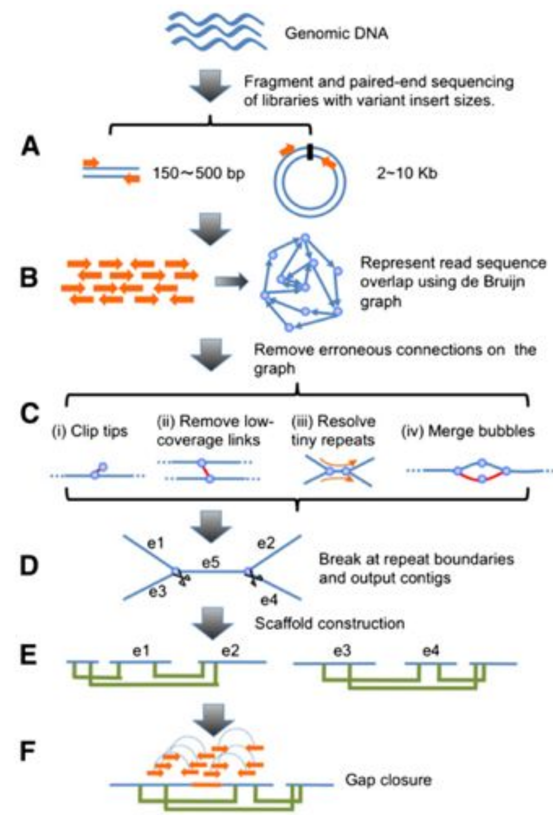
Quake Results

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

Outline

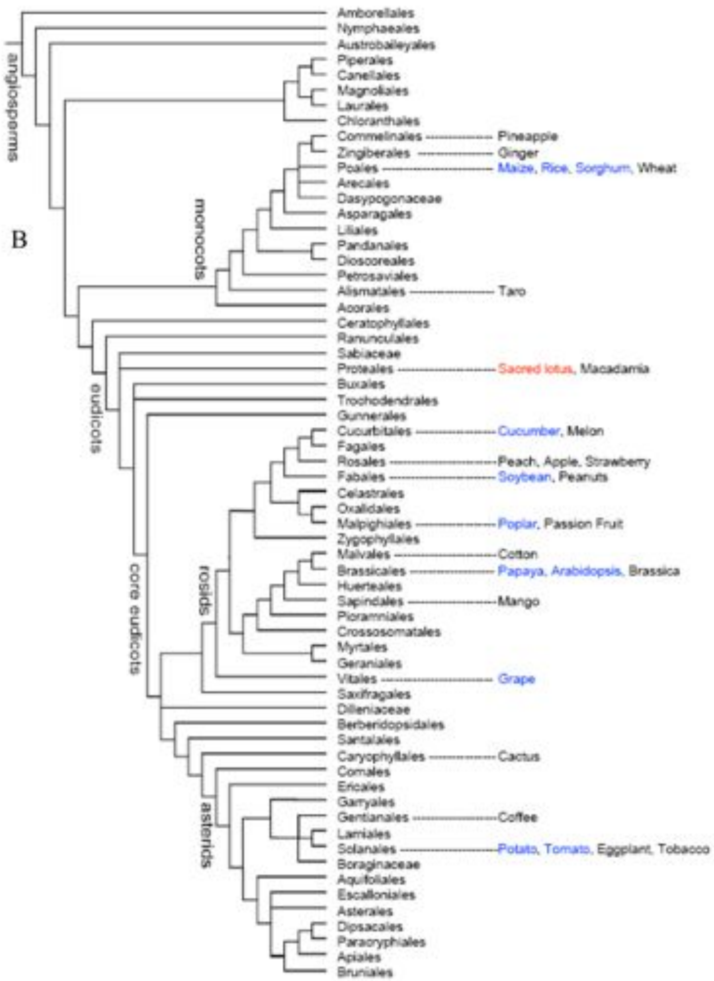
1. Ingredients for a good assembly

2. 2nd Generation Sequencing & Assembly
   1. Sacred Lotus
   2. Raspberry
   3. Wheat

3. 3rd Generation Sequence & Assembly
   1. Parrot
   2. Rice

Sacred Lotus Sequencing

Nelumbo nucifera Gaertn.

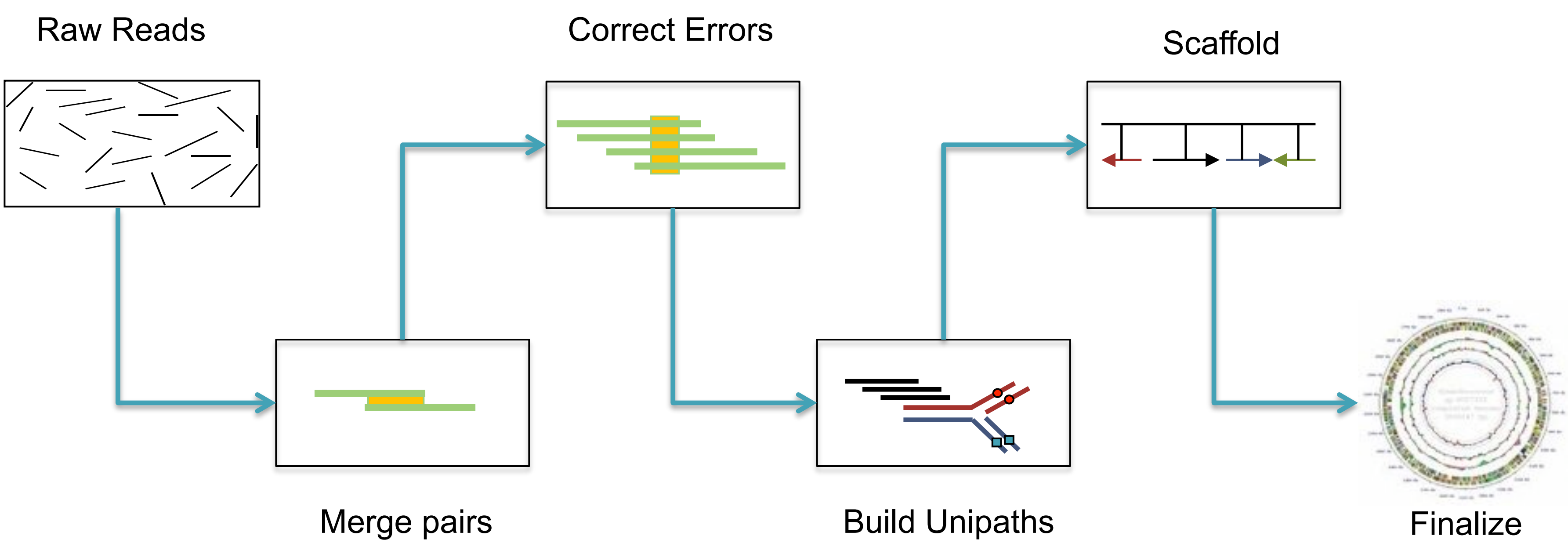


- Known for religious significance, herbal medicines, seed longevity, and water repellency
- Member of the Proteales, which lies outside of the core eudicots
  - Closest relatives are shrubs and trees belonging to the Proteaceae and Platanaceae
  - ~929Mbp Genome Size

Genome of the long-living sacred lotus (*Nelumbo nucifera* Gaertn.)
Ming, R, *et al.* (2012) *Under Review*

Sacred Lotus Sequencing Approach

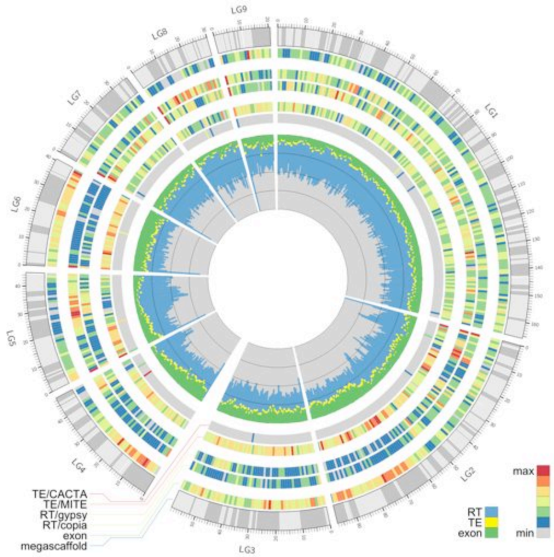
Technology	Read Length	Fragment Length	Coverage
Illumina	100 bp	180 bp	33x
	100 bp	500 bp	35x
	35 bp	3,800 bp	6.4x
	35 bp	8,000 bp	6.1x
454	*** 35 bp	20,000 bp	0.2x



Sacred Lotus Assembly

Adding 20kbp mates improved scaffold N50 from 600kbp to 3.4Mbp

- Align 454 mates to draft assembly, extract the 35bp sequence from consensus
- Error corrects, remove duplicates

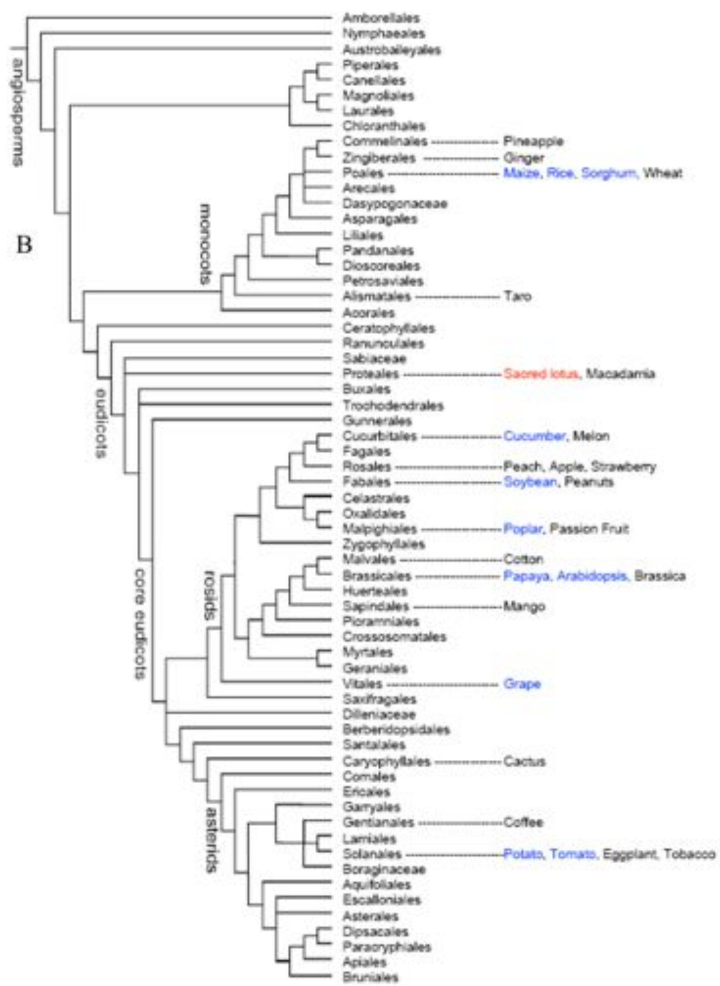


Assembly	Status	Number	N50 (kb)	Longest (kb)	size (Mb)	% cov
Contigs	All	58409	38.8	286	707	76.1
Scaffold	All	3605	3,435	14,300	804	86.5

Annotation	number	Mean (bp)	Median (bp)	Length (Mb)	% genome	% GC
Gene	26,685	6562	3917	175	21.7	36
Exons	132,653	294	153	39	4.8	43
Introns	108,887	1249	283	136	16.9	34
TE	396,000	1111		440	47	
Repeats	232,000	370		86	8.9	

Raspberry Sequencing

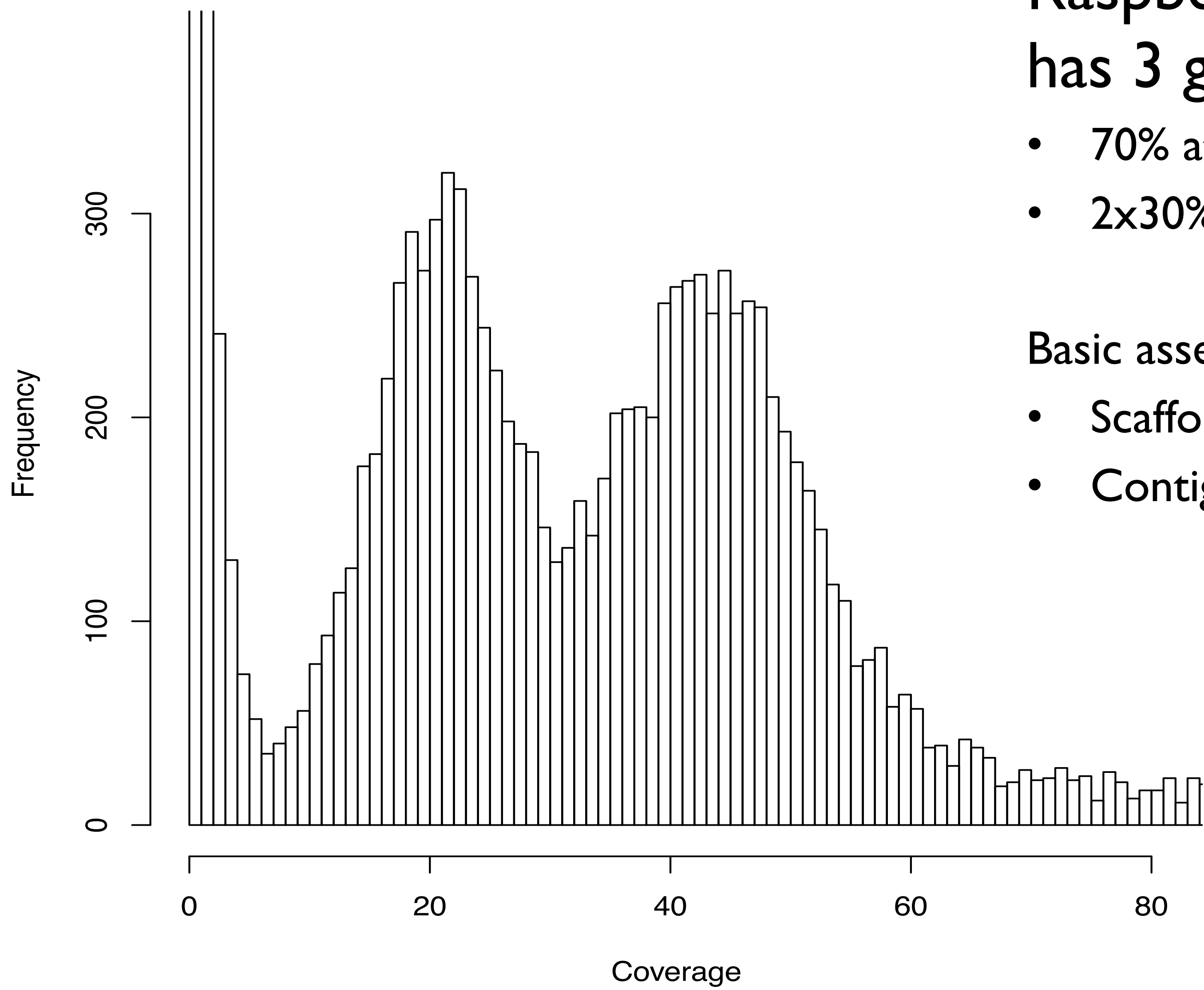
Rubus idaeus



- Important food crop (~$1B / year in production). High amounts of fiber, vitamin C, manganese, and other nutrients
- Member of the Rosaceae family, along with other common fruits
  - Including apple, peach, and strawberry
  - ~350Mbp Genome Size

The genome of the red raspberry (Rubus idaeus L.)
Price J, Ward JA *et al.* (2012) *In preparation*

Heterozygous Genomes



Raspberry effectively has 3 genomes

- 70% at full coverage
- 2x30% at half coverage

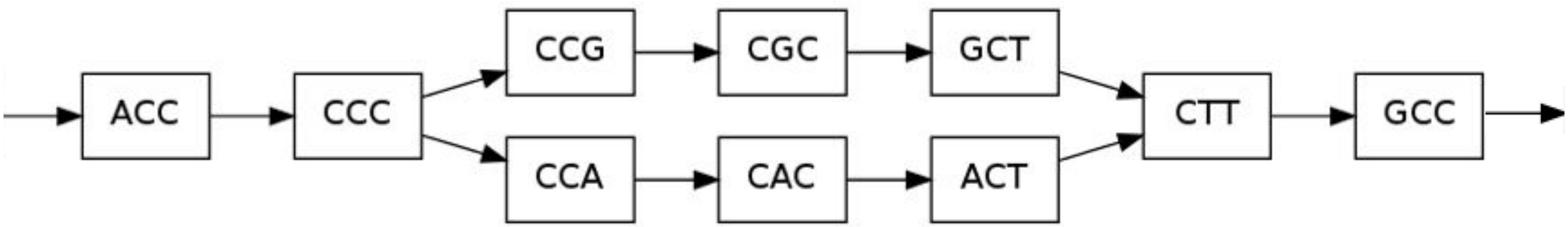
Basic assembly stats

- Scaffold N50: 17kbp
- Contig N50: 12kbp

Resolving the Heterozygosity

Chromosome 1 **TATAATCAACCCGCTTGCCGATCTGATG**

Chromosome 2 **TATAATCAACCCACTTGCCGATCTGATG**

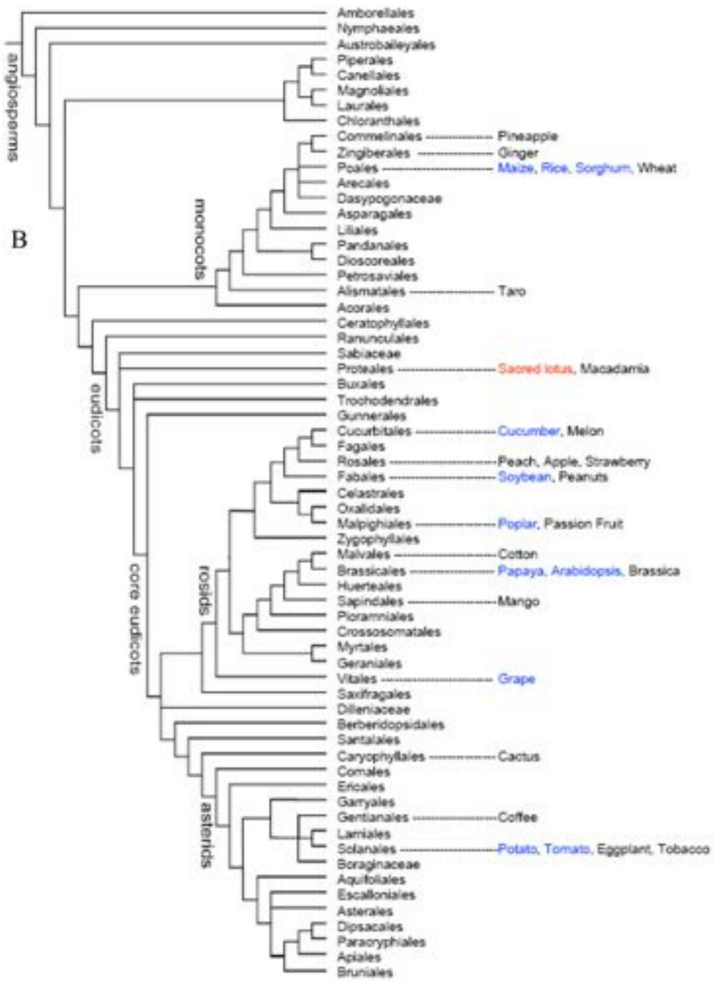


- Exploring various approaches to identify and resolve the heterozygosity.
  - Improved scaffold N50 to more than 250kbp
  - Currently using genetic map to form larger linkage groups

De novo identification of "heterotigs" towards accurate and in-phase assembly of complex plant genomes
Price J, *et al.* (2012) *Proceedings of BIOCOMP'12.* Las Vegas, NV

Wheat Sequencing

Aegilops tauschii



- One of the most important cereal crops in the world
- A. tauschii is one of the three ancestral species (DD) in modern bread wheat (*Triticum aestivum*)
  - Also looking to sequence other 2 species, and bread wheat
  - ~4.5Gbp Genome Size

In Collaboration with McCombie and Ware labs

Wheat Sequencing & Assembly

Technology	Read Length	Fragment Length	Coverage
Illumina	100 bp	180 bp	69x
	100 bp	300 bp	50x
	35 bp	2,000 bp	6.6x
	35 bp	5,000 bp	6.5x

Assembly	Count	Max	N50	Sum
Scaffolds	97,313	2.76 Mbp	23,193	1.36 Gbp (30%)
Contigs	556,767	165 kbp	4,623	928 Mbp (20%)

- Poor coverage of the genome due to extreme repeat content
  - Had to downsample reads to fit into RAM
  - Randomly discard reads covered by kmers that occur more than 500 times
- Coverage may be sufficient for "***gene-space***"

Outline

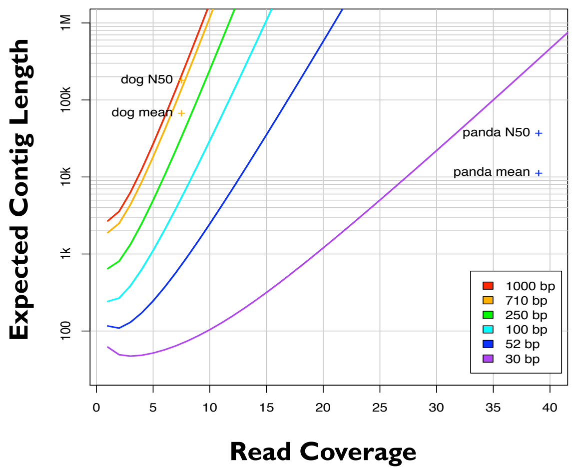
1. Ingredients for a good assembly

2. 2nd Generation Sequencing & Assembly
    1. Sacred Lotus
    2. Raspberry
    3. Wheat

3. 3rd Generation Sequence & Assembly
    1. Parrot
    2. Rice

Ingredients for a good assembly

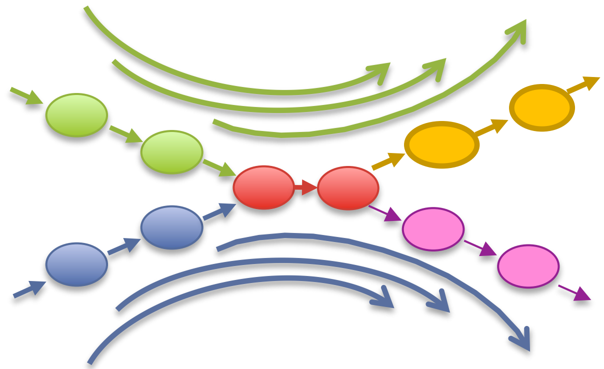
Coverage



High coverage is required

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

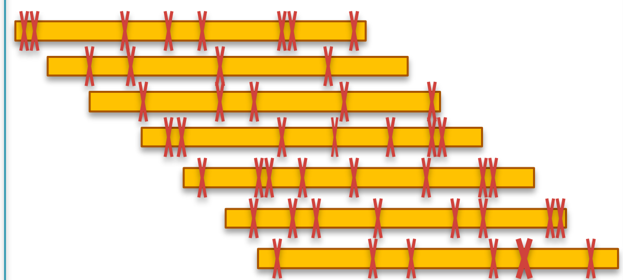
Read Length



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

Hybrid Sequencing

Illumina
Sequencing by Synthesis

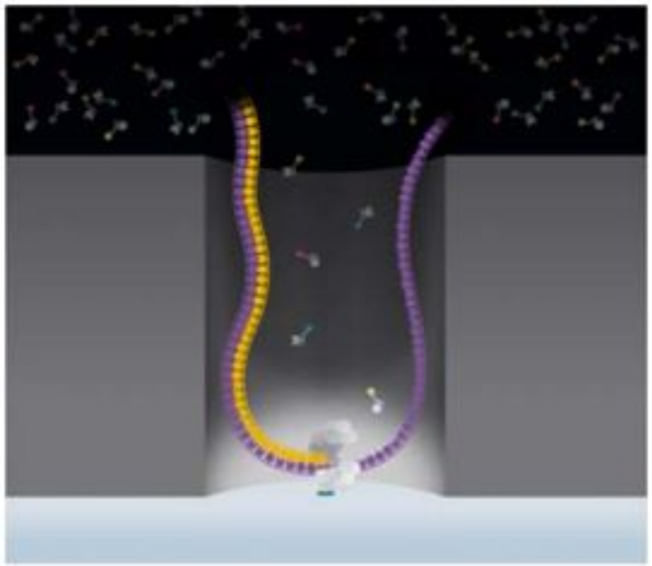
High throughput (60Gbp/day)
High accuracy (~99%)
Short reads (~100bp)

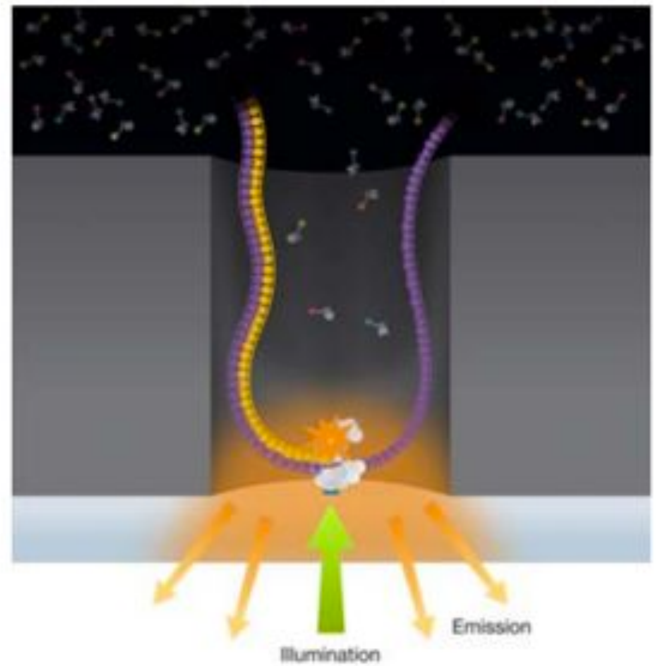
Pacific Biosciences
SMRT Sequencing

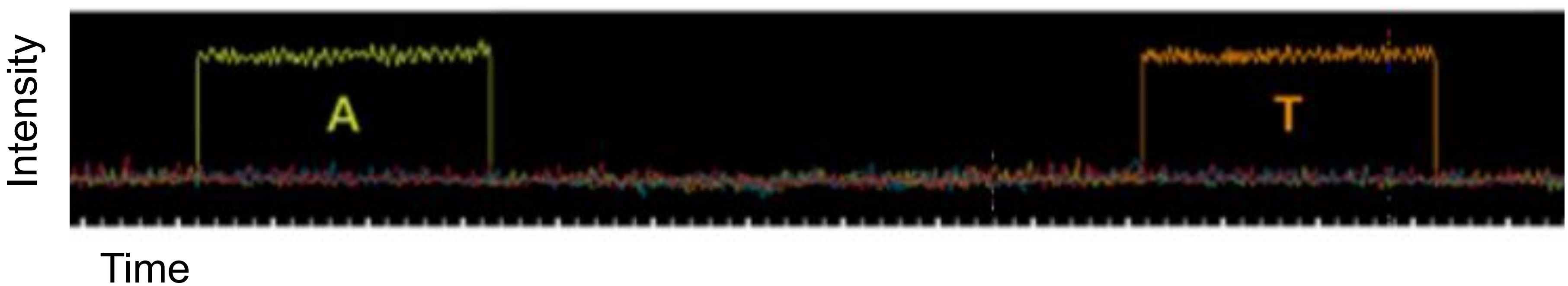
Lower throughput (600Mbp/day)
Lower accuracy (~85%)
Long reads (1-2kbp+)

SMRT Sequencing

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

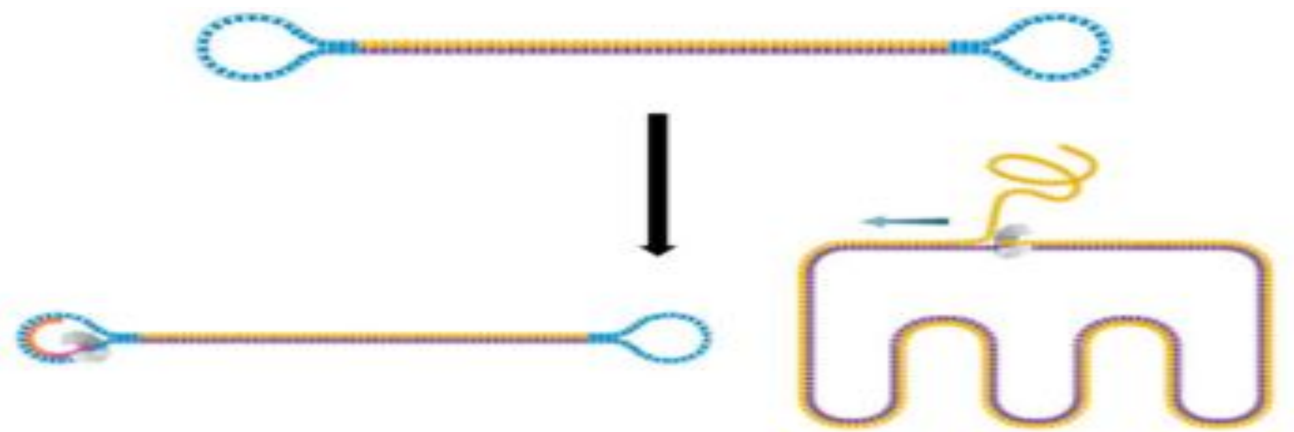






http://www.pacificbiosciences.com/assets/files/pacbio_technology_backgrounder.pdf

SMRT Read Types



- ***Standard sequencing***
  - Long inserts so that the polymerase can synthesize along a single strand

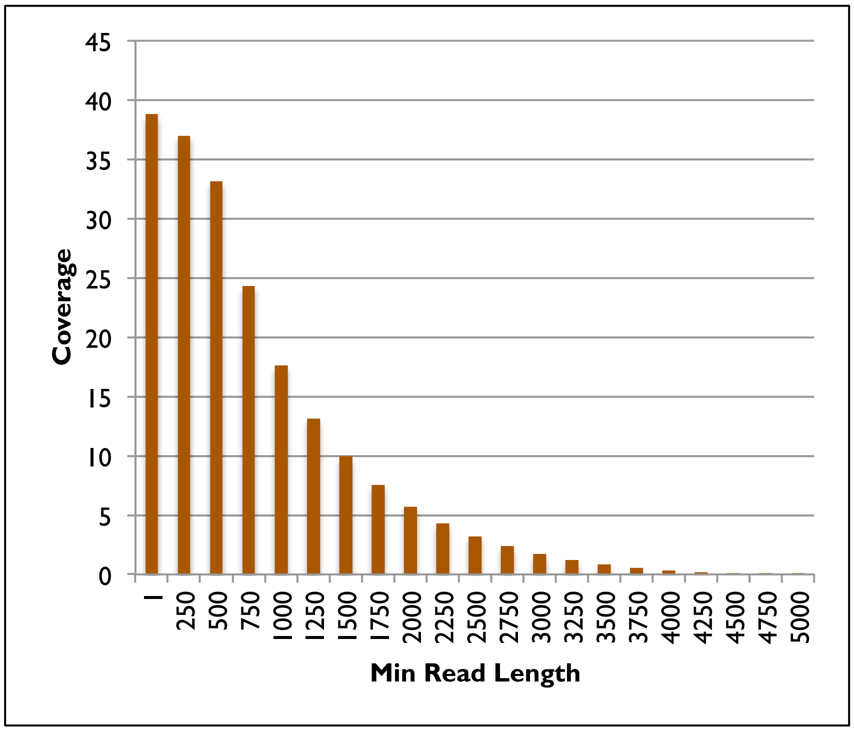
- ***Circular consensus sequencing***
  - Short inserts, so polymerase can continue around the entire SMRTbell multiple times and generate multiple sub-reads from the same single molecule.

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

SMRT Sequencing Data

Yeast
(Pre-release Chemistry / 2010)

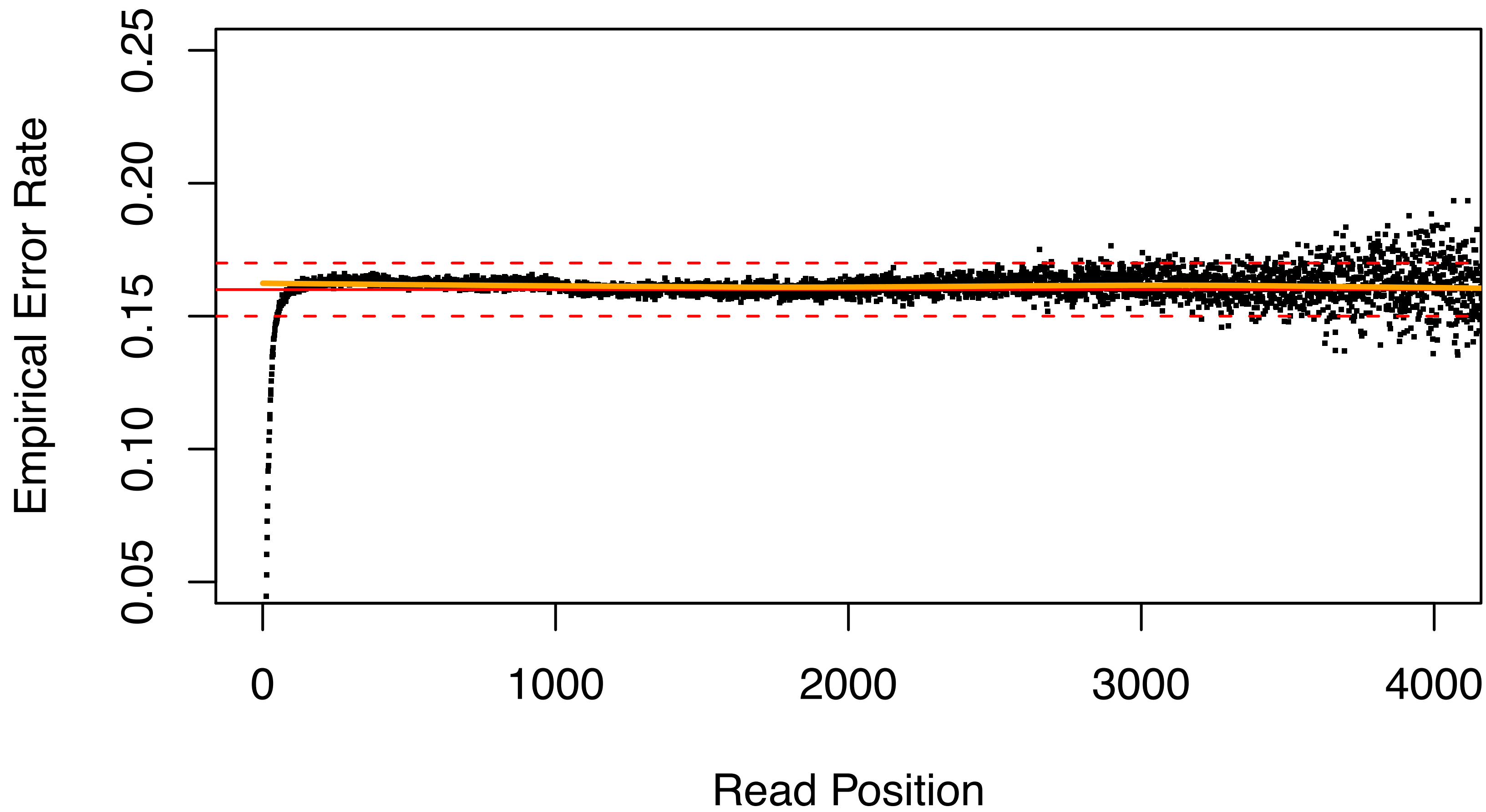
65 SMRT cells
734,151 reads after filtering
Mean: 642.3 +/- 587.3
Median: 553 Max: 8,495




```
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

Read Quality



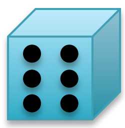
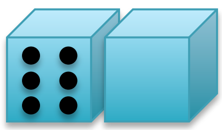
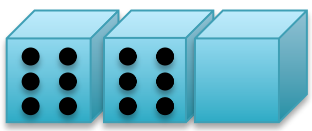
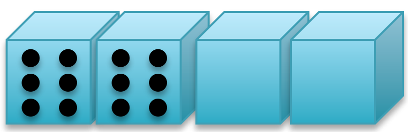
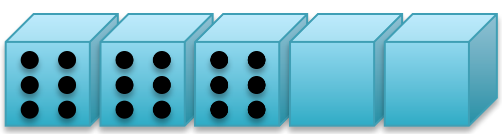
Consistent quality across the entire read

- Uniform error rate, no apparent biases for GC/motifs
- Sampling artifacts at beginning and ends of alignments

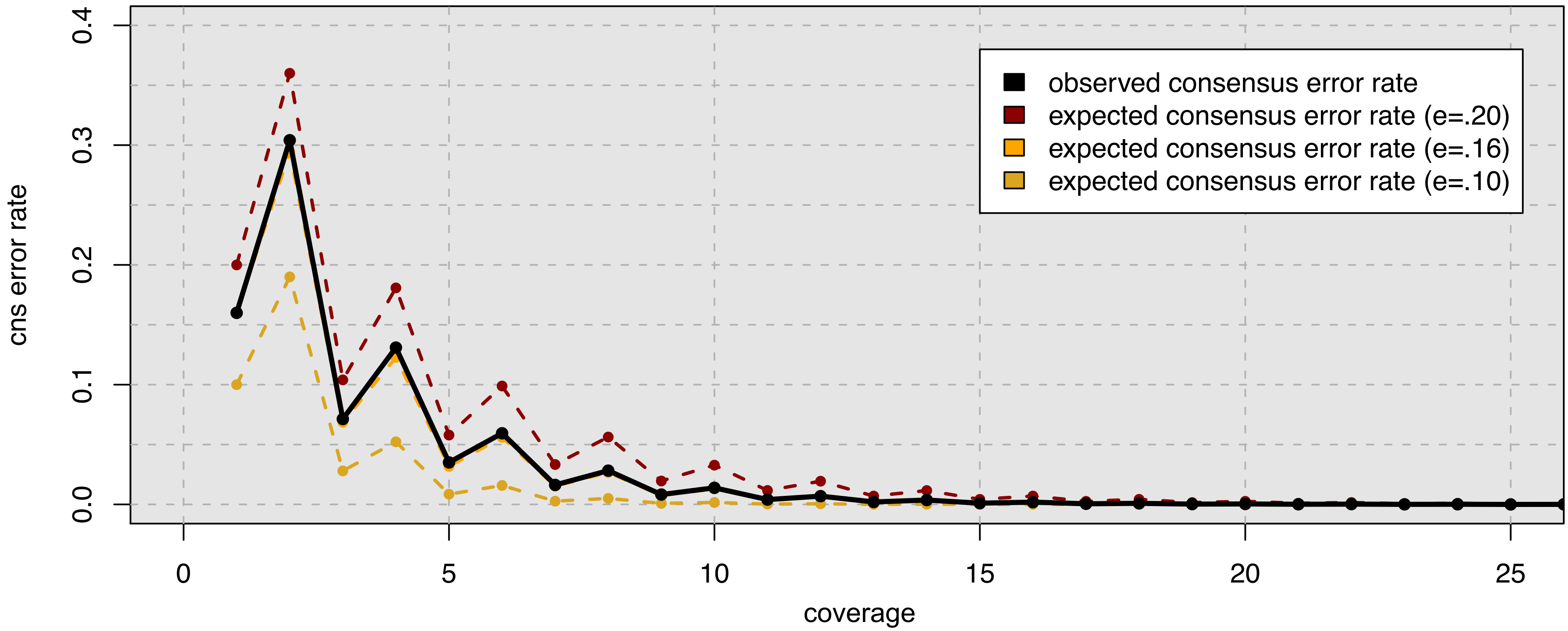
Consensus Quality: Probability Review

Roll *n* dice => What is the probability that at least half are 6's

(Consensus is wrong if at least half the bases are wrong)

n	*Min to Lose*	*Losing Events*	*P(Lose)*
1		*1/6*	*16.7%*
2		*P(1 of 2) + P(2 of 2)*	*30.5%*
3		*P(2 of 3) + P(3 of 3)*	*7.4%*
4		*P(2 of 4) + P(3 of 4) + P(4 of 4)*	*13.2%*
5		*P(3 of 5) + P(4 of 5) + P(5 of 5)*	*3.5%*
n	*ceil(n/2)*	$\sum_{i=\lceil n/2 \rceil}^{n} P(i \;\; of \;\; n) = \sum_{i=\lceil n/2 \rceil}^{n} \binom{n}{i} (p)^i (1-p)^{n-i}$	

Consensus Accuracy and Coverage



Coverage can overcome random errors

- Dashed: error model from binomial sampling; solid: observed accuracy
- For same reason, CCS is extremely accurate when using 5+ subreads

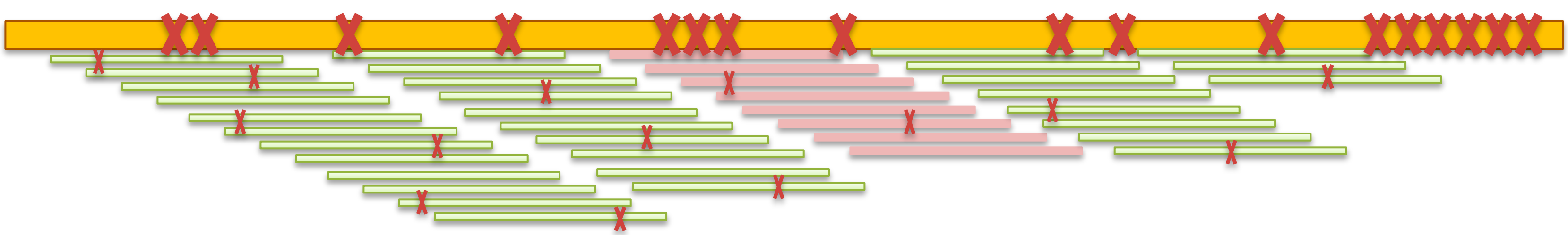
$$CNS\ Error = \sum_{i=\lceil c/2 \rceil}^{c} \binom{c}{i} (e)^{i} (1-e)^{n-i}$$

PacBio Error Correction

http://wgs-assembler.sf.net

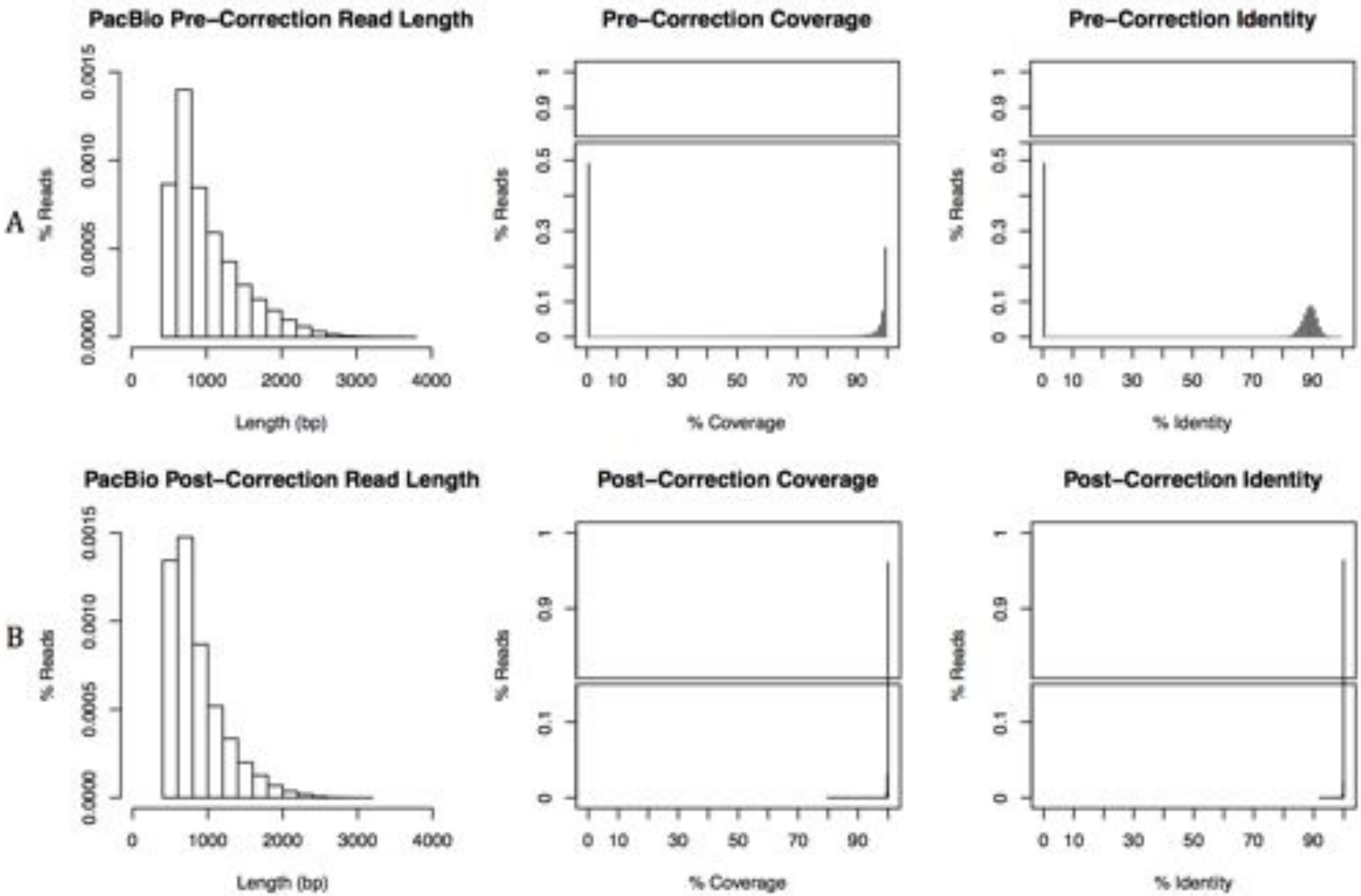
1. Correction Pipeline
   1. Map short reads (SR) to long reads (LR)
   2. Trim LRs at coverage gaps
   3. Compute consensus for each LR

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

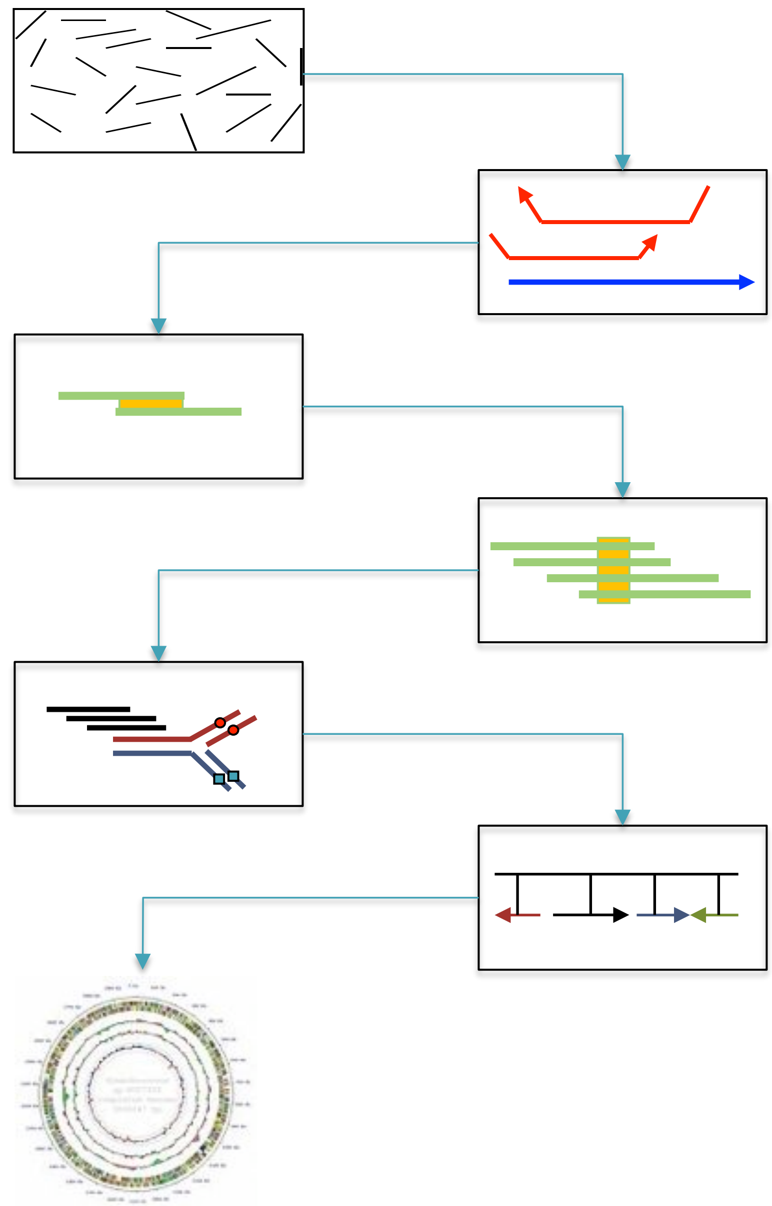


Correction results of 20x PacBio coverage of E. coli K12 corrected using 50x Illumina

Celera Assembler

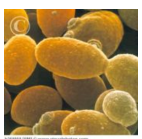
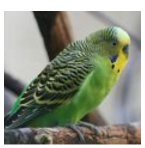
http://wgs-assembler.sf.net

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



SMRT-Assembly Results

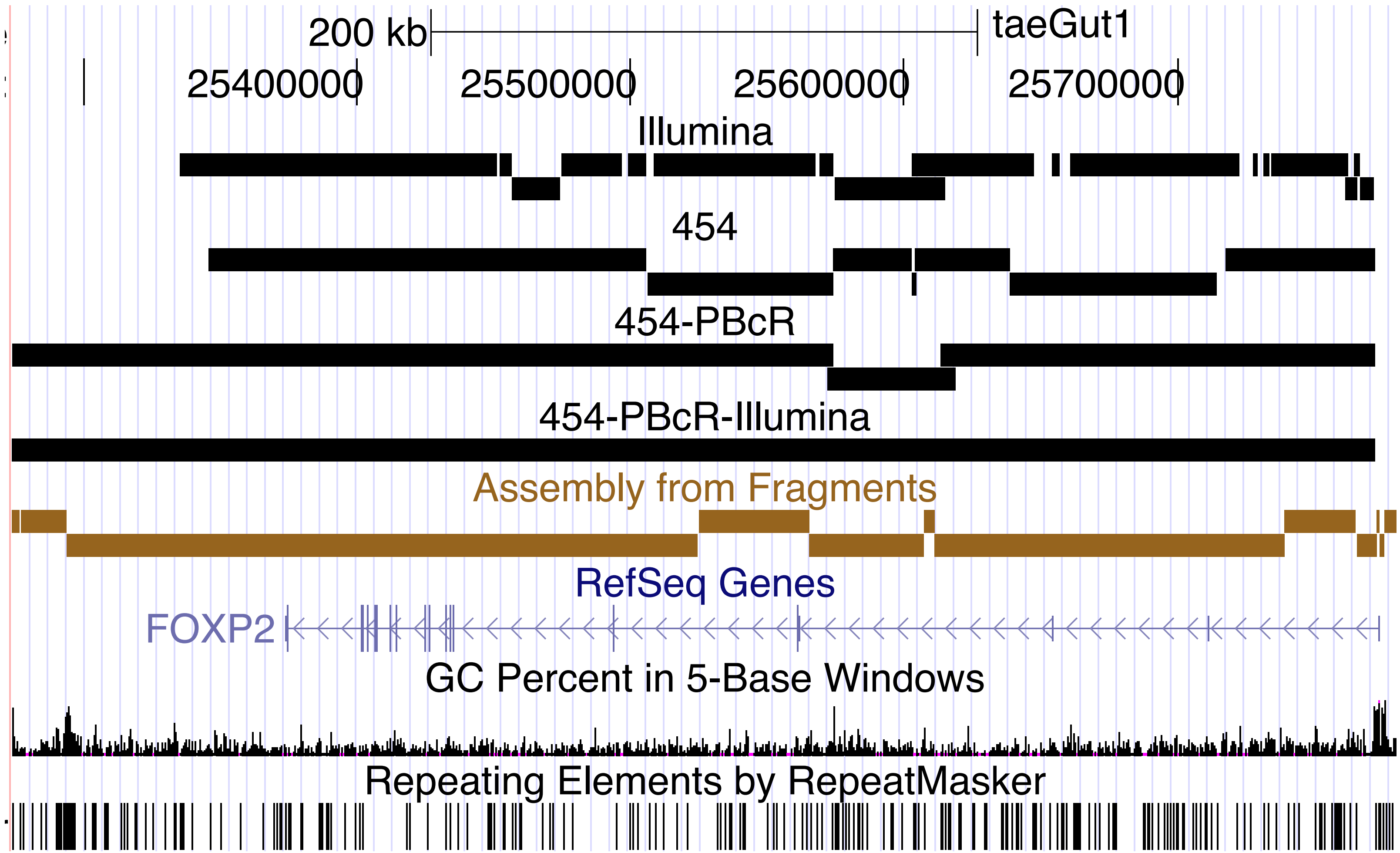


Organism	Technology	Reference bp	Assembly bp	# Contigs	Max Contig Length	N50
Lambda NEB3011	Illumina 100X 200bp	48 502	48 492	1	48 492 / 48 492	48 492 / 48 492 (100%) *
(median: 727 max: 3 280)	PacBio PBcR 25X		48 440	1	48 444 / 48 444	48 444 / 48 440 (100%) *
E. coli K12	Illumina 100X 500bp	4 639 675	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
*** Able to assemble entire microbial chromosomes into individual contigs ***

Improved Gene Reconstruction



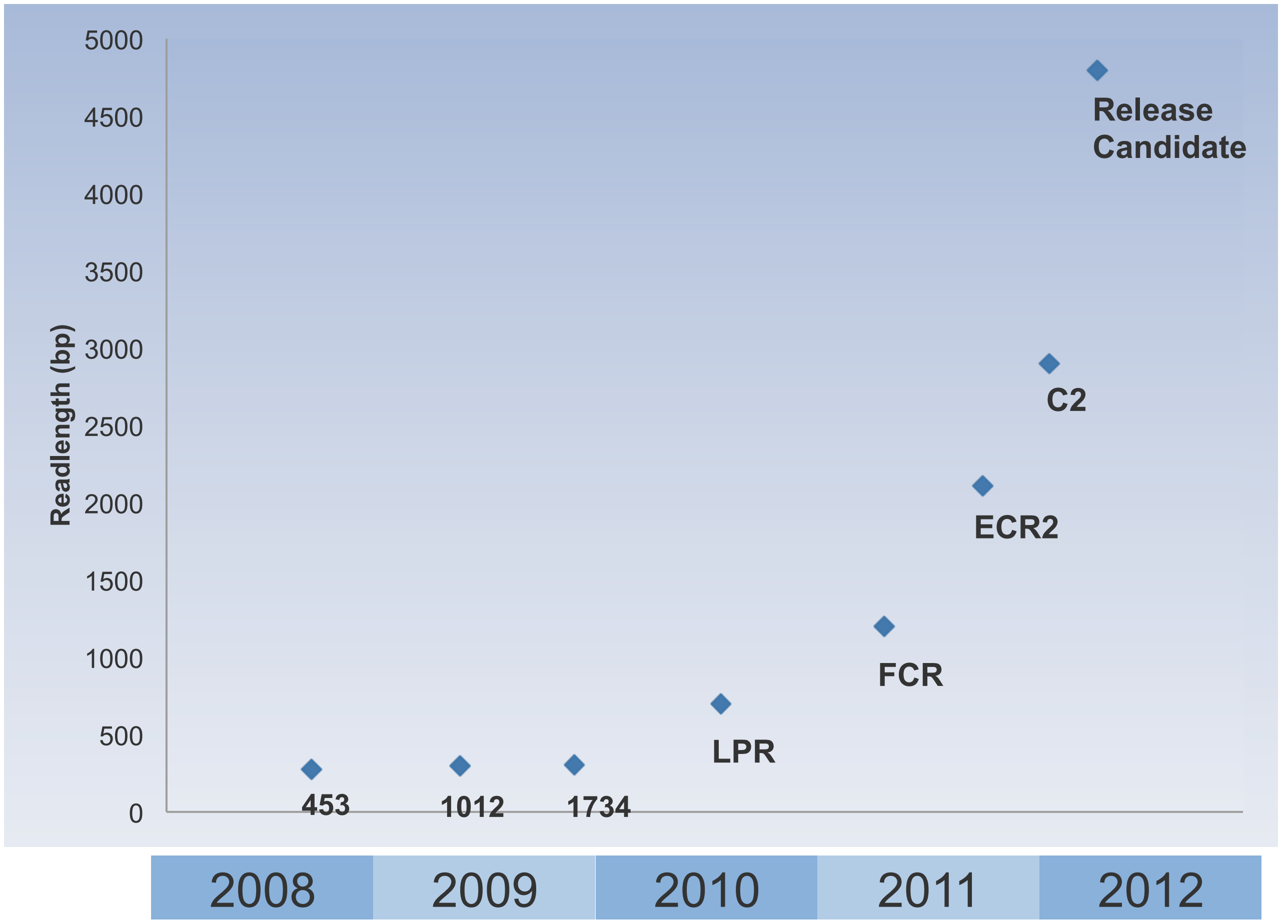
FOXP2 assembled on a single contig

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

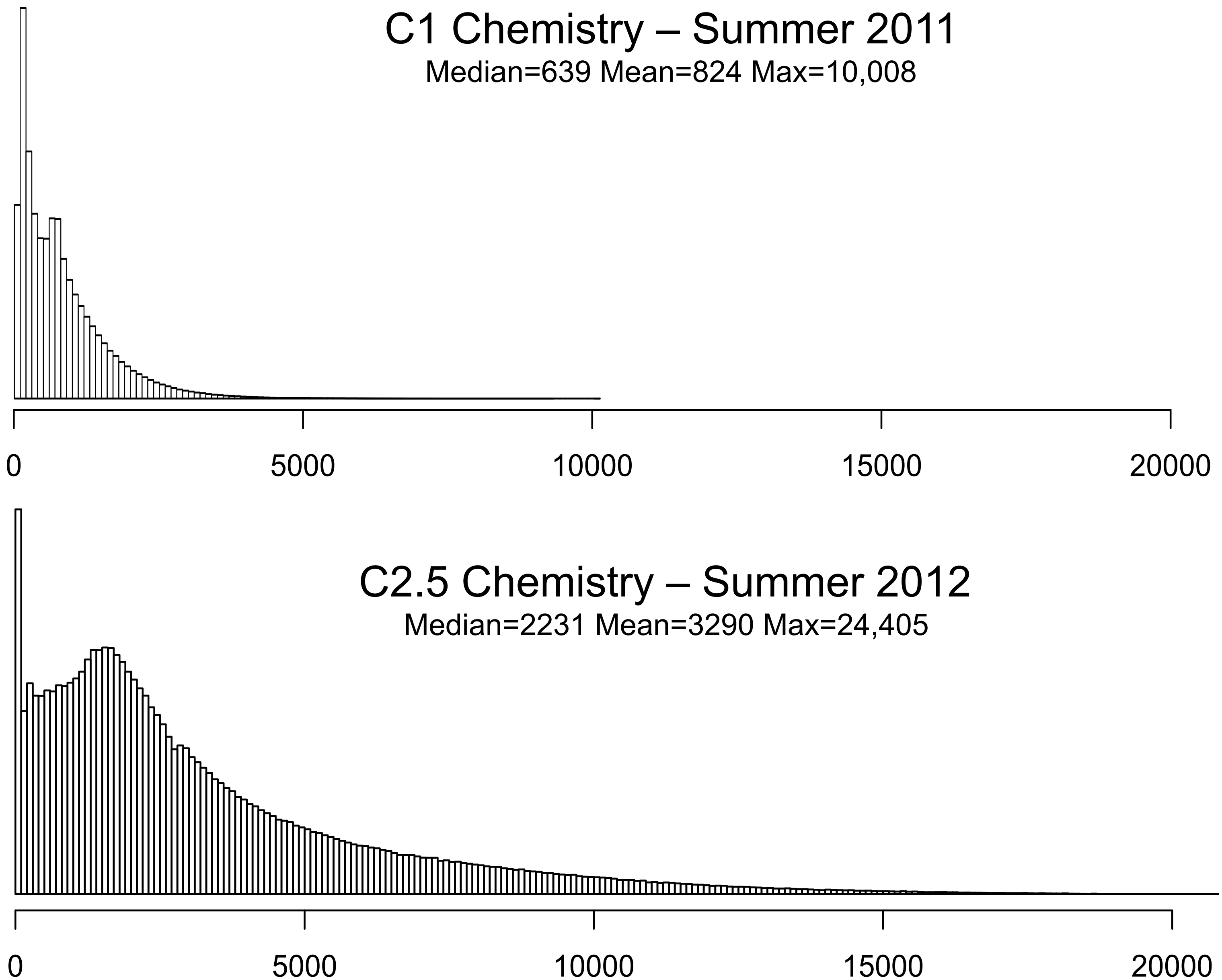
2. "Hot Start" technology:

   Maximize subreads

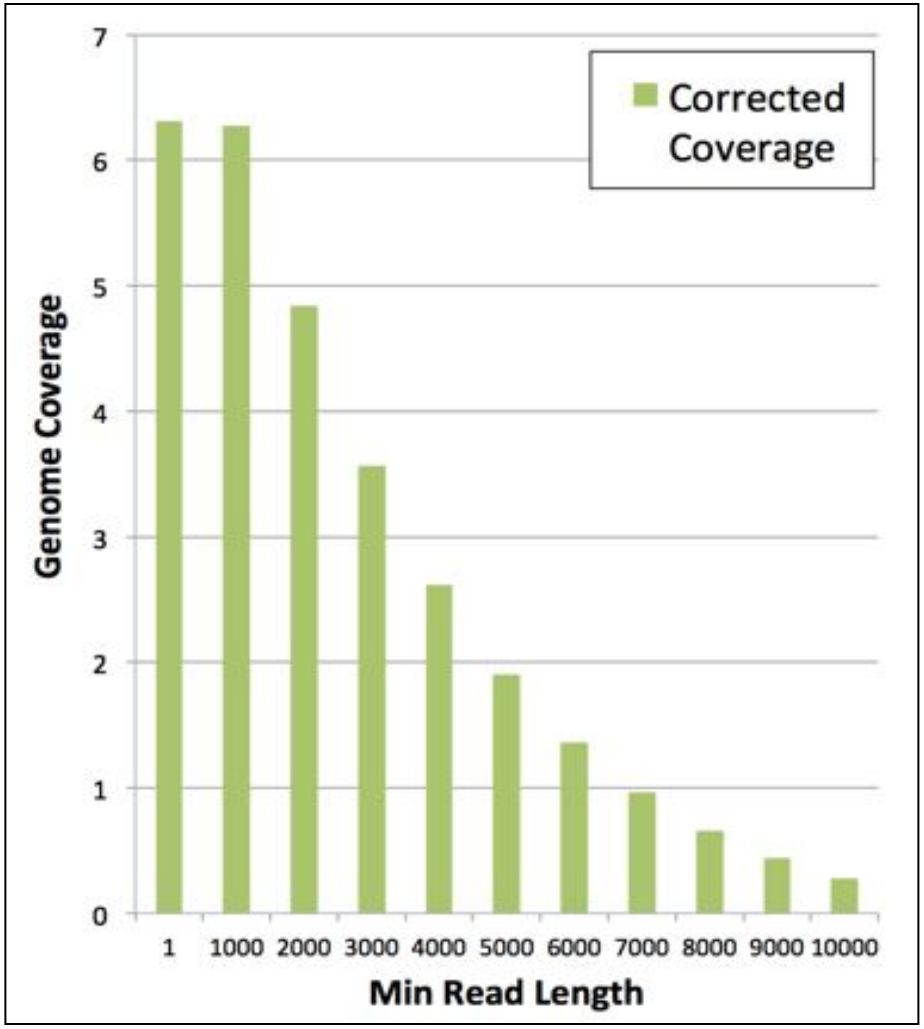
3. MagBead loading:

   Load longest fragments

PacBio Rice Sequencing



Preliminary Rice Assemblies



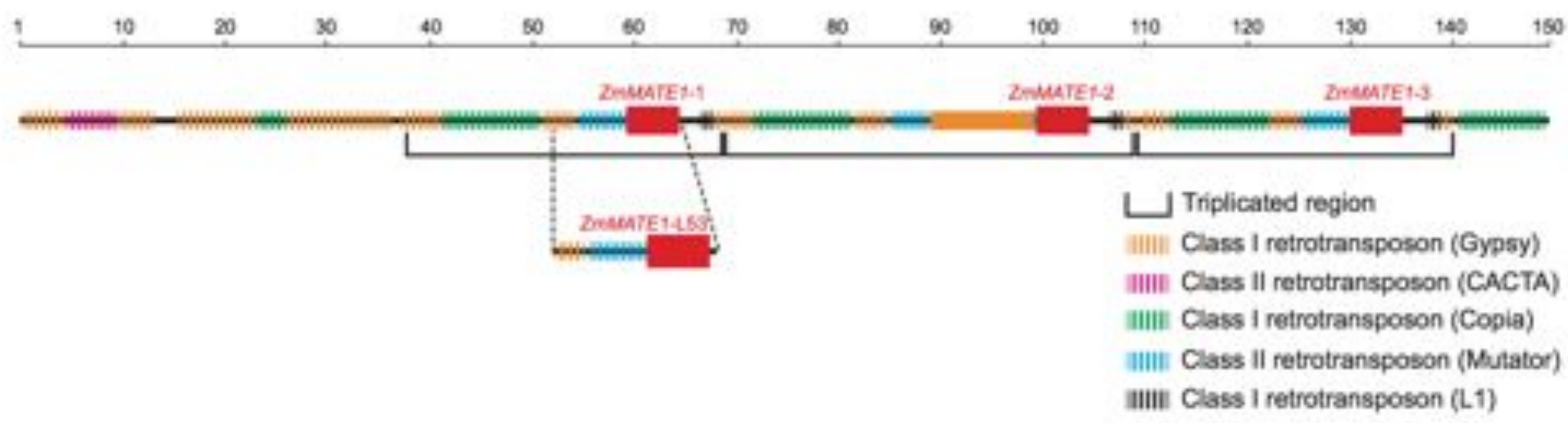
Assembly	Contig N50
Illumina Fragments 50x 2x100bp @ 180	3925
Illumina Mates 50x 2x100bp @ 180 36x 2x50bp @ 2100 51x 2x50bp @ 4800	13696
MiSeq Fragments 23x 459bp 8x 2x251bp @ 450	6444
PBeCR Reads 6.3x 2146bp ** MiSeq for correction	13600
PBeCR + Mates 6.3x 2146bp ** MiSeq for correction 51x 2x50bp @ 4800	In Progress

In collaboration with McCombie & Ware labs @ CSHL

Long Read CNV Analysis

Aluminum tolerance in maize is important for drought resistance and protecting against nutrient deficiencies

- Segregating population localized a QTL on a BAC, but unable to genotype with Illumina sequencing because of high repeat content
- Long read PacBio sequencing revealed an additional copy of the ZnMATE1 membrane transporter and enabled assembly of the entire gene cluster



A rare gene copy-number variant that contributes to maize aluminum tolerance and adaptation to acid soils

Maron, LG *et al.* (2012) *Under review.*

Why are crop 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

With new biotechnologies and improved algorithms we can address these challenges

=> Cautiously optimistic

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/