Assembly and Validation of Large Genomes from Short Reads

Michael Schatz

March 16, 2011

Genome Assembly Workshop / Genome 10k

CSH

A Brief Aside

4.7GB / disc
~20 discs / 1G Genome

X

10,000 Genomes

=

1PB Data
200,000 DVDs

Hadoop MapReduce

http://hadoop.apache.org

- MapReduce is Google's framework for large data computations
  - Data and computations are spread over thousands of computers
    - Indexing the Internet, PageRank, Machine Learning, etc… (Dean and Ghemawat, 2004)
    - 946,460 TB processed in May 2010 (Jeff Dean at Stanford, 11.10.2010)
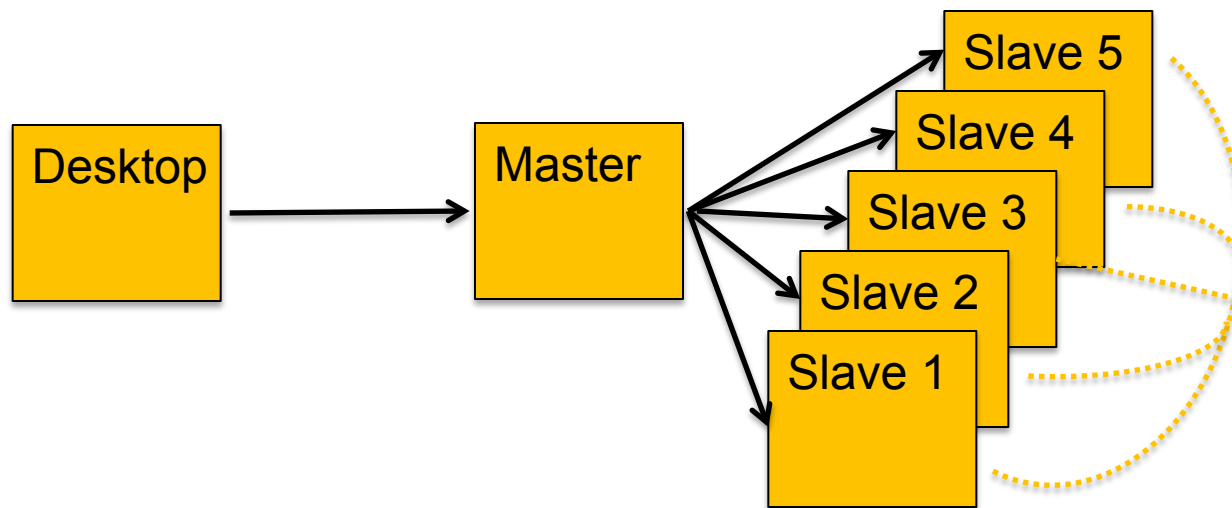
Hadoop for NGS Analysis

CloudBurst

Highly Sensitive Short Read
Mapping with MapReduce

*100x speedup mapping
on 96 cores @ Amazon*

http://cloudburst-bio.sf.net

(Schatz, 2009)

Crossbow
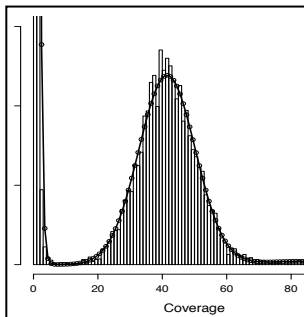
Searching for SNPs
with Cloud Computing

*Identify 3M SNPs from 38x coverage
in 3 hours on 320 cores*

(Langmead, Schatz,
Lin, Pop, Salzberg, 2010)

http://bowtie-bio.sf.net/crossbow/

Quake

Quality-aware error
correction of short reads

*Correct 97.9% of errors
with 99.9% accuracy*

http://www.cbcb.umd.edu/software/quake/

(Kelley, Schatz,
Salzberg, 2010)

Genome Indexing

Rapid Parallel Construction
of the Genome Index

*Construct the BWT of
the human genome in 9 minutes*

```
$GATTACA
A$GATTAC
ACA$GATT
ATTACA$G
CA$GATTA
GATTACA£
TACA$GAT
TTACA$GA
```

(Menon,
Bhat, Schatz, 2011*)

http://code.google.com/p/
genome-indexing/

Assembly of Large Genomes with Cloud Computing.
Schatz MC, Sommer D, Kelley D, Pop M, *et al. In Preparation.*

Outline & Acknowledgements

Quake

Quality-aware detection and correction of sequencing errors
Kelley, DR, Schatz, MC, Salzberg, SL (2010) *Genome Biology 11:R116*
http://www.cbcb.umd.edu/software/quake/

Celera Assembler

Aggressive Assembly of Pyrosequencing Reads with Mates.
Miller, J. *et al.* (2008) *Bioinformatics 24(24):2818-2824*
http://wgs-assembler.sf.net

Bambus 2

A Scaffolder for Polymorphic and Metagenomic Data
Koren, S, Pop, M (2011) *In Preparation.*
http://amos.sf.net/bambus2

Forensics

Assembly Forensics: Finding the ellusive mis-assembly
Phillippy, A, Schatz, MC, Pop, M. (2008) *Genome Biology 9:R55*
http://amos.sf.net/forensics

GAGE

Genome Assembly Gold Standard Evaluations
Salzberg, SL *et al.* (2011) *In Preparation*
http://gage.cbcb.umd.edu/

Detection and Correction with Quake
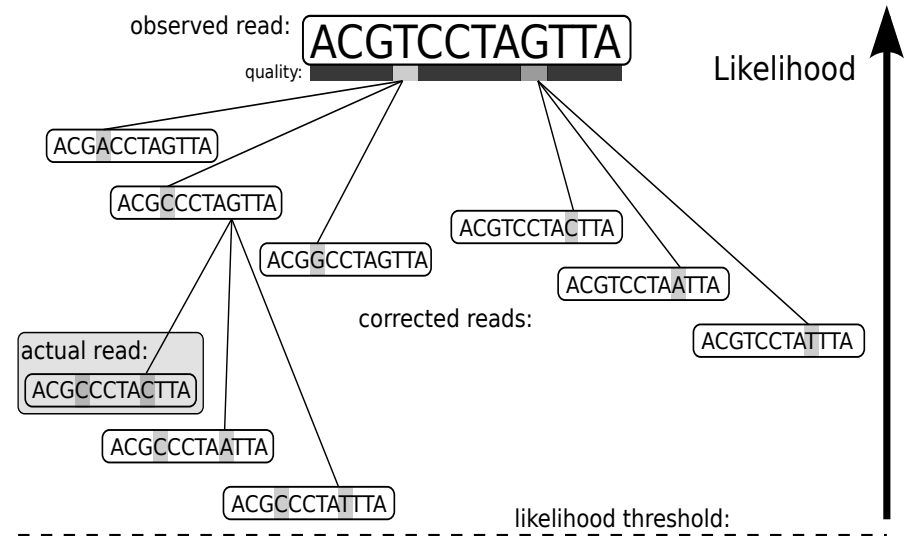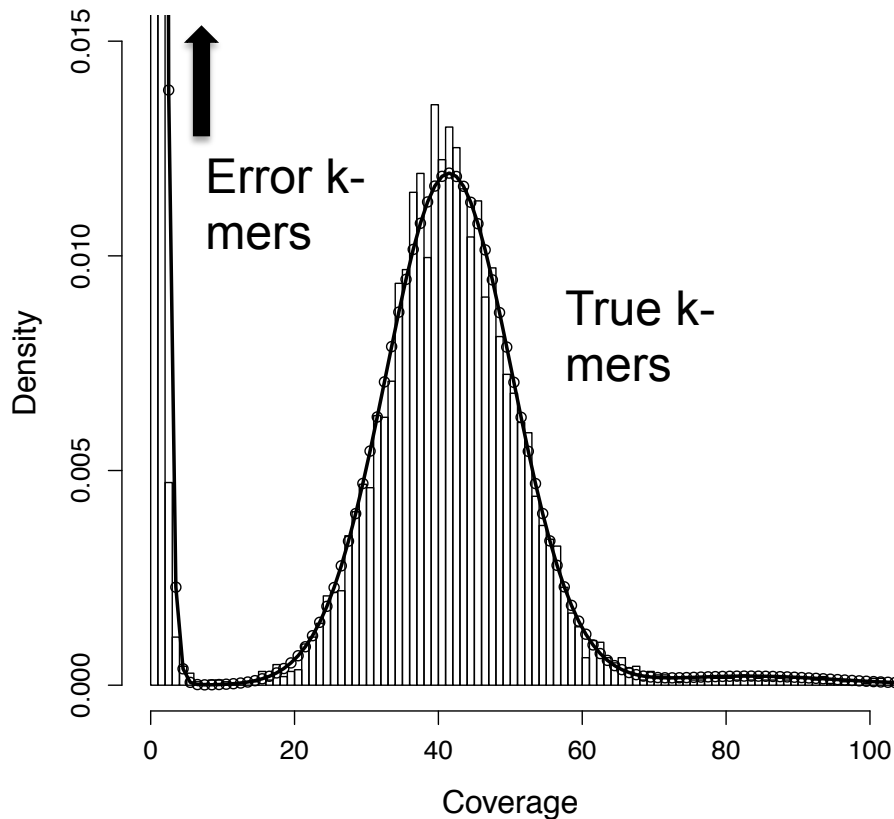
http://www.cbcb.umd.edu/software/quake/

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
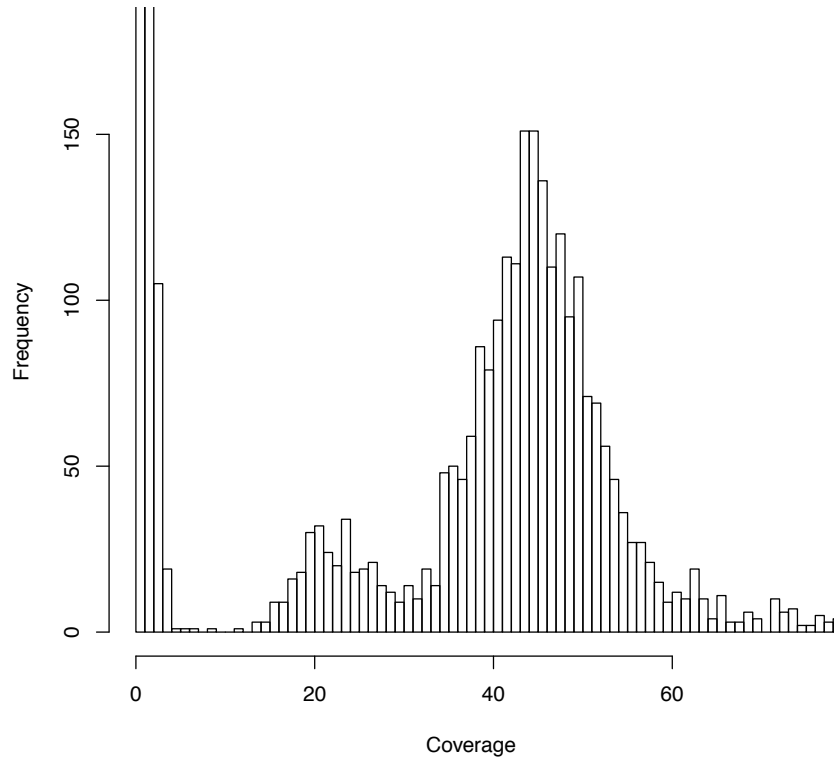
2. Correction Algorithm

- Consider editing erroneous kmers into trusted kmers in decreasing likelihood

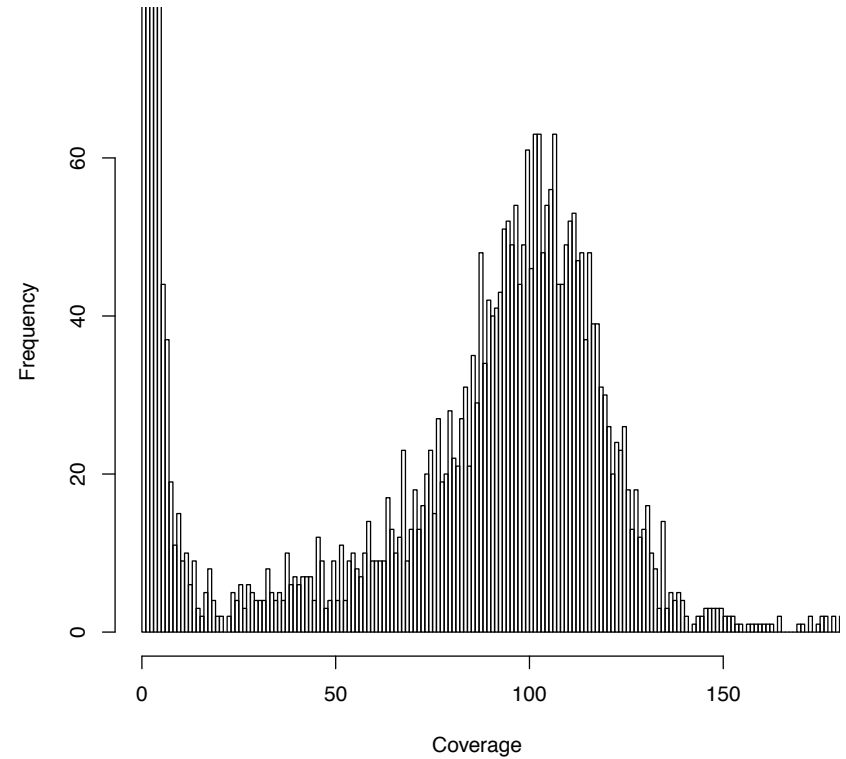- Includes quality values, nucleotide/ nucleotide substitution rate

Assemblathon Results

Species A

Validated	35996138	32.0%
Corrected	62502345	55.5%
Trim Only	7923360	7.0%
Removed	6076811	5.4%

Rice

| Validated	304488985	52.1%
Corrected	86383318	14.8%
Trim Only	190890445	27.5%
Removed	32648755	5.6%

Heterozygous Genomes

- Raspberry effectively has **3** genomes
  - 70% at full coverage
  - 2x30% at half coverage

Celera Assembler

http://wgs-assembler.sf.net
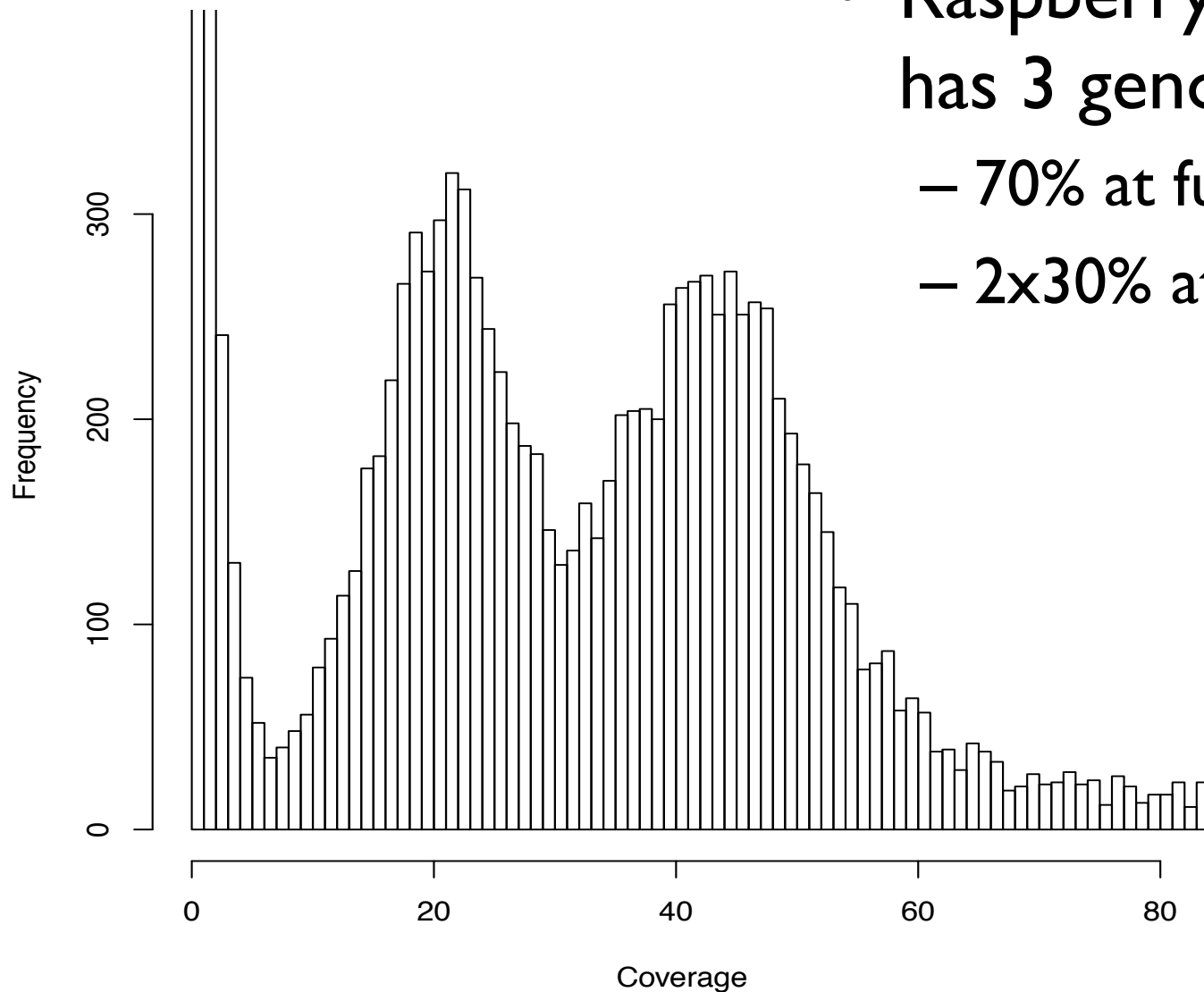
1. **Pre-overlap**
   – Consistency checks

2. **Trimming**
   – Vector 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

Recent CA Results

Celera Assembler
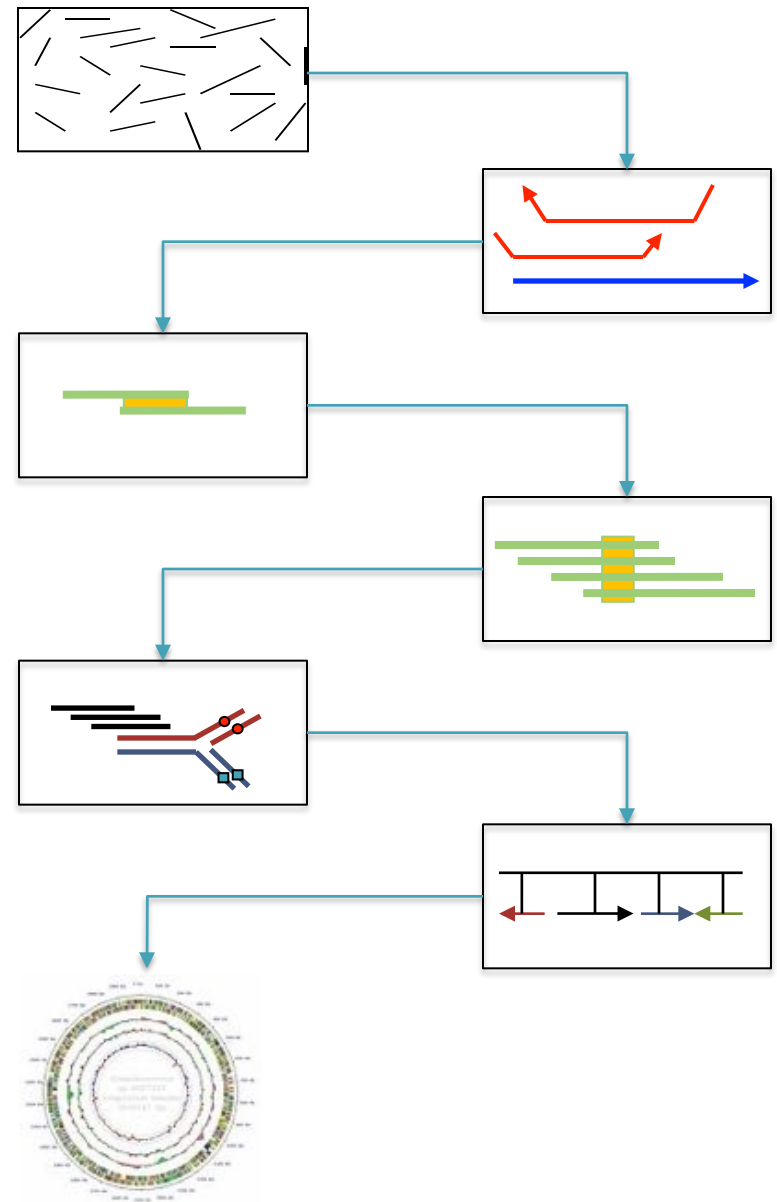
	Species A	Bumble Bee[1]	Argentine Ant[2]	Parrot[3]
Species		*Bombus impatiens*	*Linepithema humile*	*Melopsittacus undulatus*
Total Scaffolds	137	1,896	3,030	25,212
Scaffolds Bases	121,259,411	287,738,041	215,552,578	1,086,605,544
Scaffold N50	3,254,796	1,124,853	1,386,360	11,201,952
Max Scaffold	8,283,751	4,021,294	-	39,665,220
Total Contigs*	37,571	92,307	18,227	404,592
Contig N50	139,666	23,515	35,858	55,633
Max Contig	1,442,666	297,795	-	465,633

*Includes "degenerate contigs"

[1]Robertson, H. *et al.* (2011) *Under Review.*   Illumina: 75x 400bp, 14x 4kbp, 13x 8kbp

[2]Smith, C.D. *et al.* (2011) *PNAS.*   Illumina: 8x unpaired, 4x 3kbp, 1x 8kbp
        454:        8x unpaired, 1x 3kbp, .3x 8kbp

[3]Jarvis, E. *et al.* (2011) *Details Friday.*   Illumina 12x, 454: 6x

Scaffolding with Bambus

http://amos.sf.net/bambus2

- Algorithm Overview
  - Hierarchical scaffolding of the most "strongly" connected contigs

- Design
  - Identify consistent bundles of "links"
    - Mate-pairs, but also any other relationships
  - Prioritize link types, link requirements
    - Prefer mate-links to distant synteny
  - Standalone module that can be used with any assembler
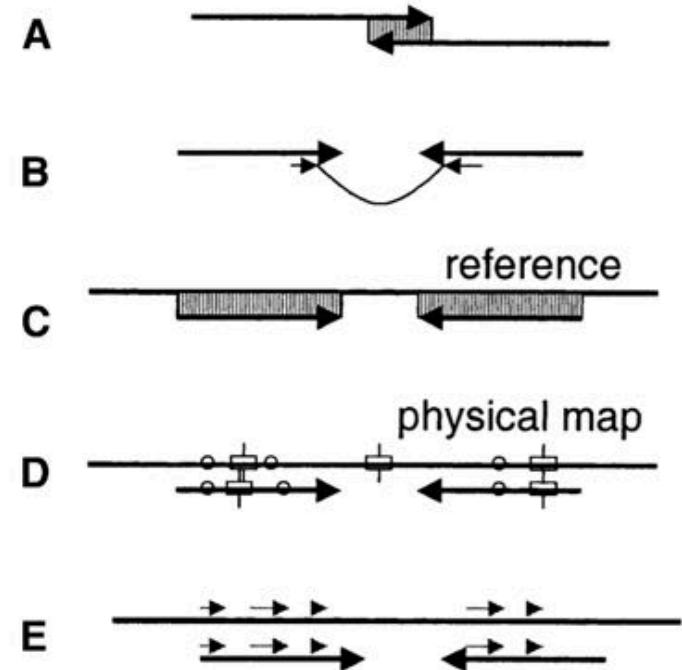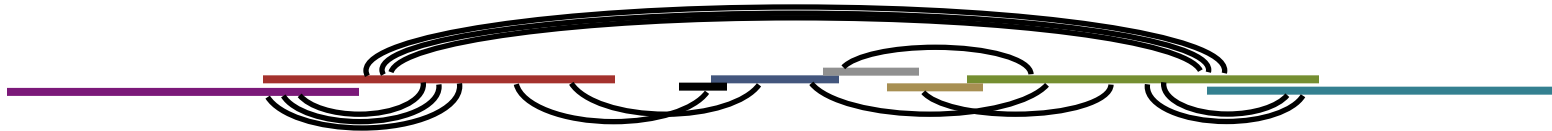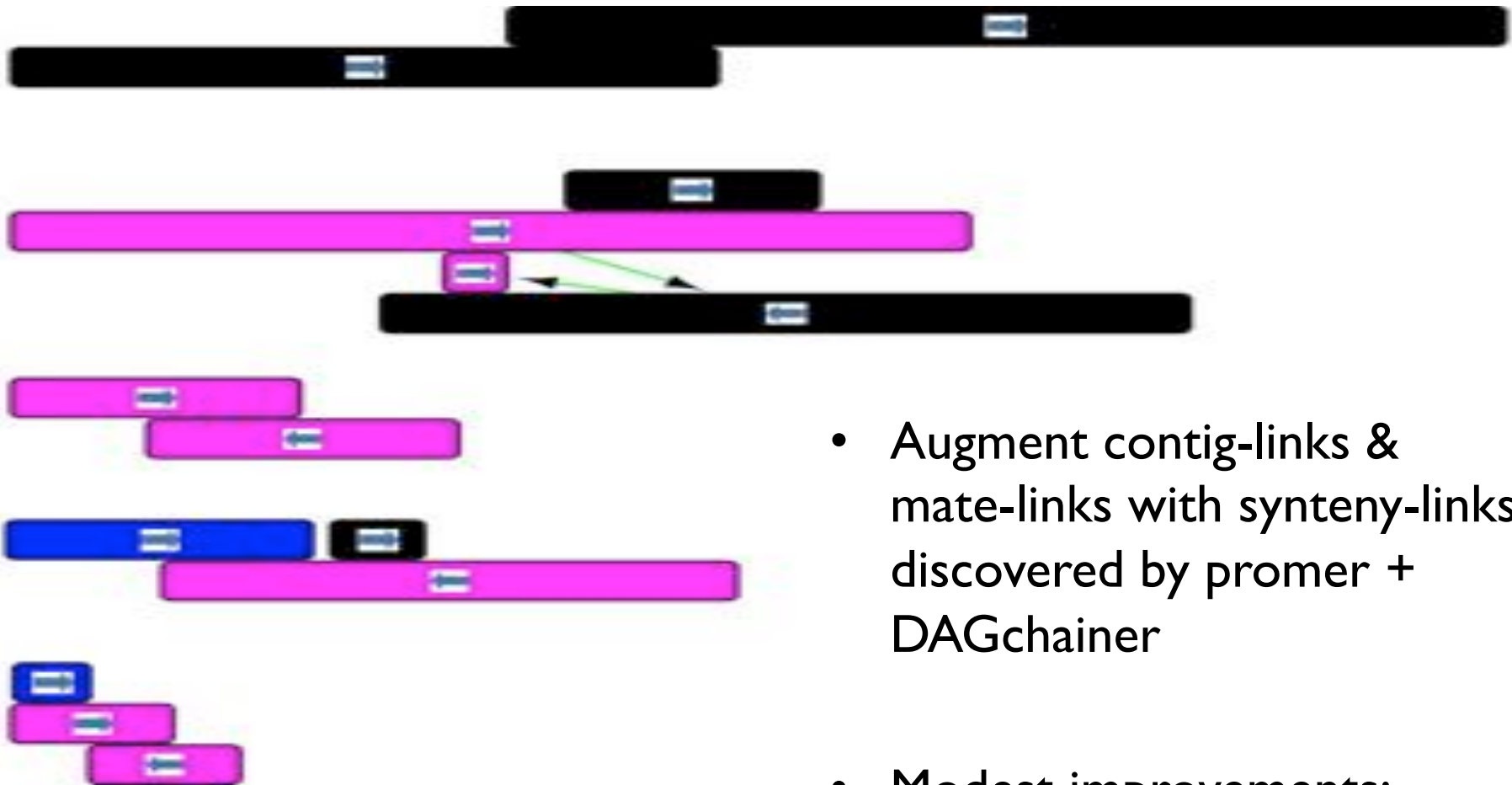   - Support for strobed-reads in development

Figure 3 Sources of linking information between contigs. (*A*) overlaps, (*B*) clone mates, (*C*) alignments to reference genome, (*D*) alignments to physical maps, (*E*) conservation of gene synteny.

Assemblathon Results

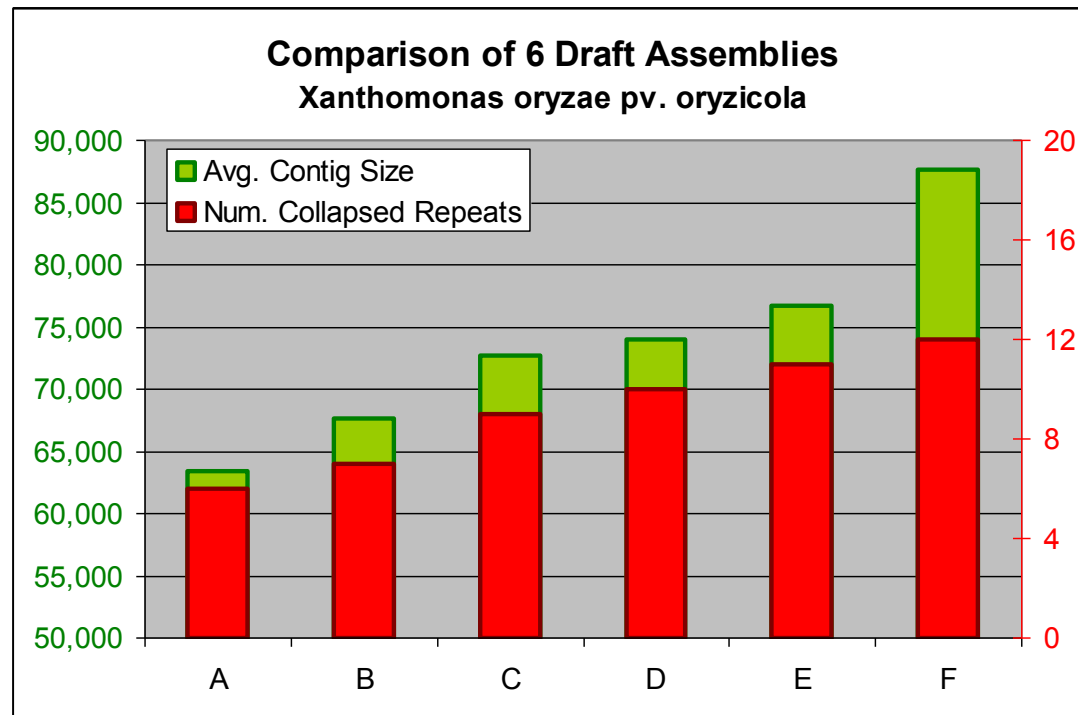- Augment contig-links & mate-links with synteny-links discovered by promer + DAGchainer

- Modest improvements:
  - Max: 10,924,052 (+30%)
  - N50: Unchanged

Assembly Forensics

http://amos.sf.net/forensics

- Assembly is often a balancing act
  - Tension between sequencing errors, repeats, coverage, other factors
  - Size statistics alone can be misleading

Comparison of 6 Draft Assemblies
Xanthomonas oryzae pv. oryzicola

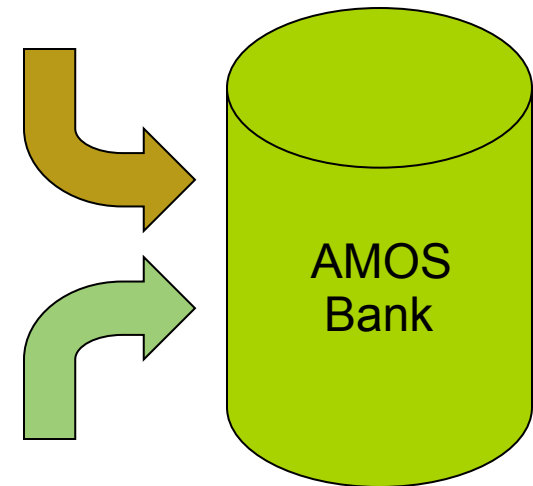- Avg. Contig Size
- Num. Collapsed Repeats

Forensics Pipeline

Computationally scan an assembly for mis-assemblies.

- Data inconsistencies are indicators for mis-assembly
- Some inconsistencies are merely statistical variations

AMOSvalidate

1. Load Assembly Data into Bank
2. Analyze Mate Pairs & Libraries
3. Analyze Depth of Coverage
4. Analyze Normalized K-mers
5. Analyze Read Alignments
6. Analyze Read Breakpoints
7. Load Mis-assembly Signatures into Bank

AMOS Bank

Compression/Expansion Statistic

Forensics

Normal Library
Count=10000, Mean=4000, SD=400

0kb   2kb   4kb   6kb
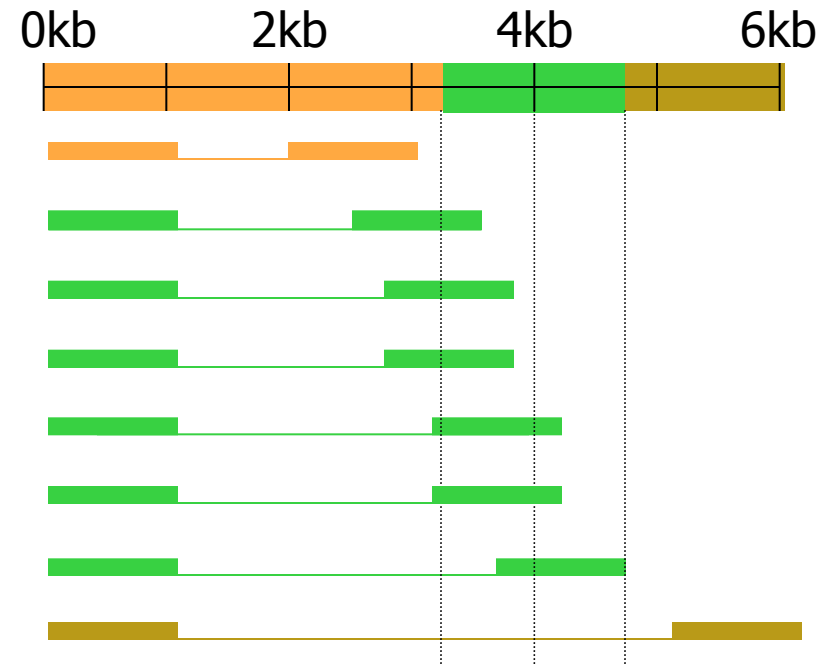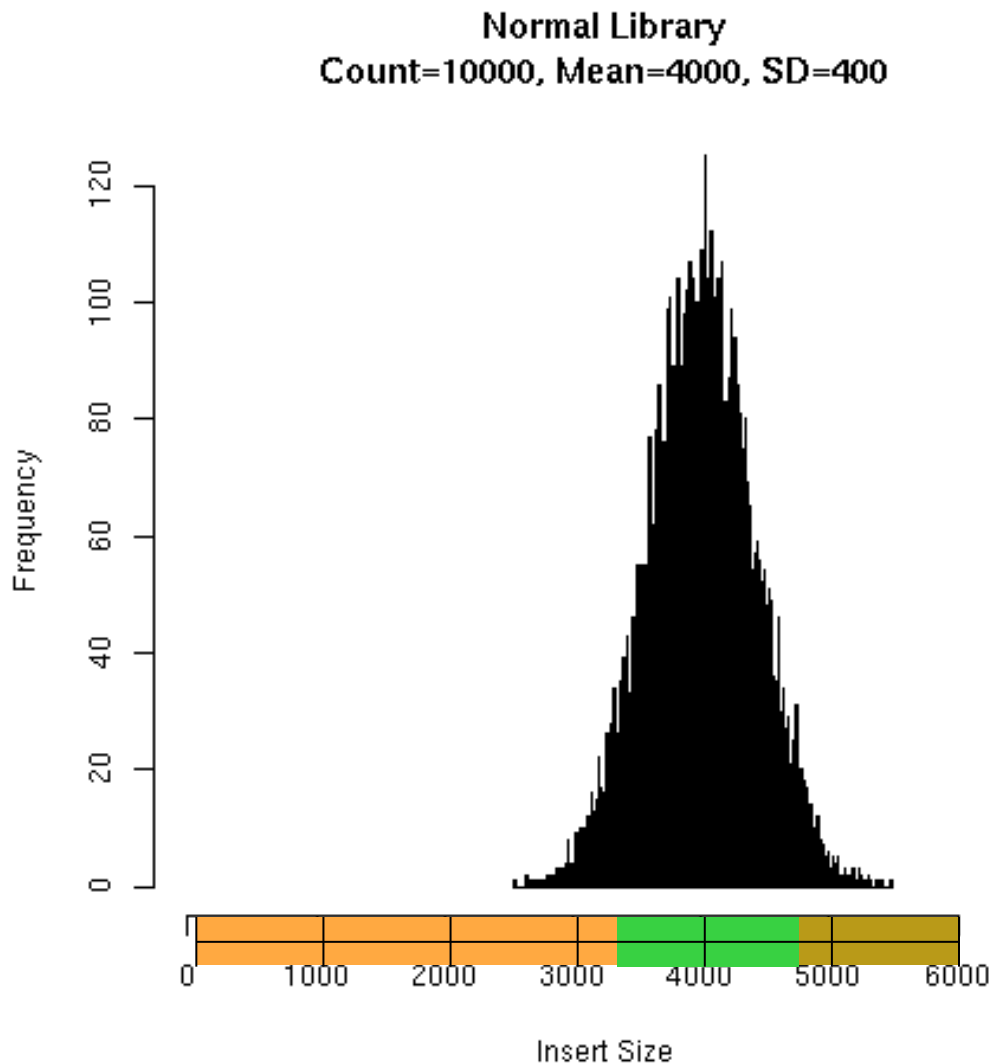
8 inserts: 3kb-6kb

Local Mean: 4048

C/E Stat: $\dfrac{(4048-4000)}{(400 / \sqrt{8})} = +0.33$

Near 0 indicates overall happiness

CE Expansion

Forensics

Normal Library
Count=10000, Mean=4000, SD=400
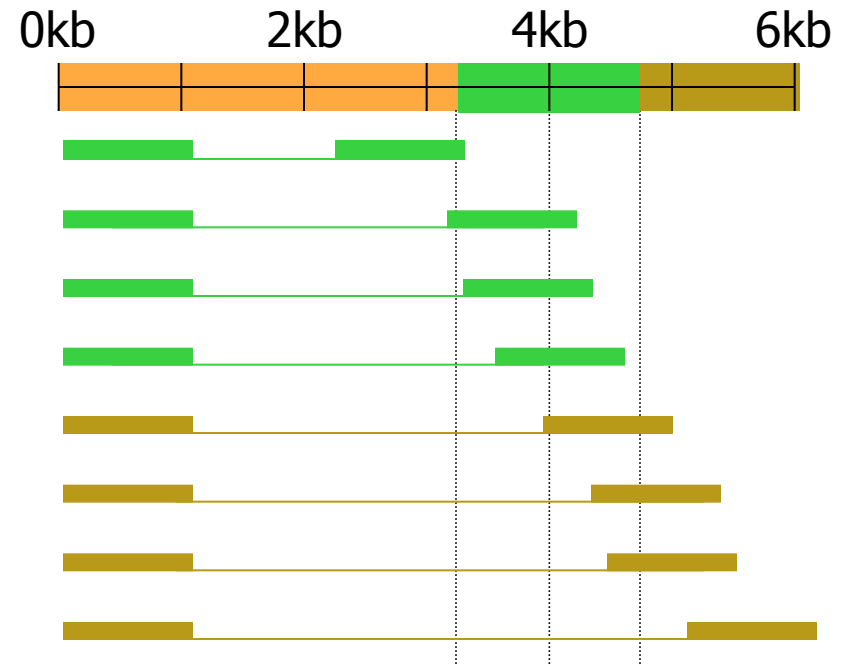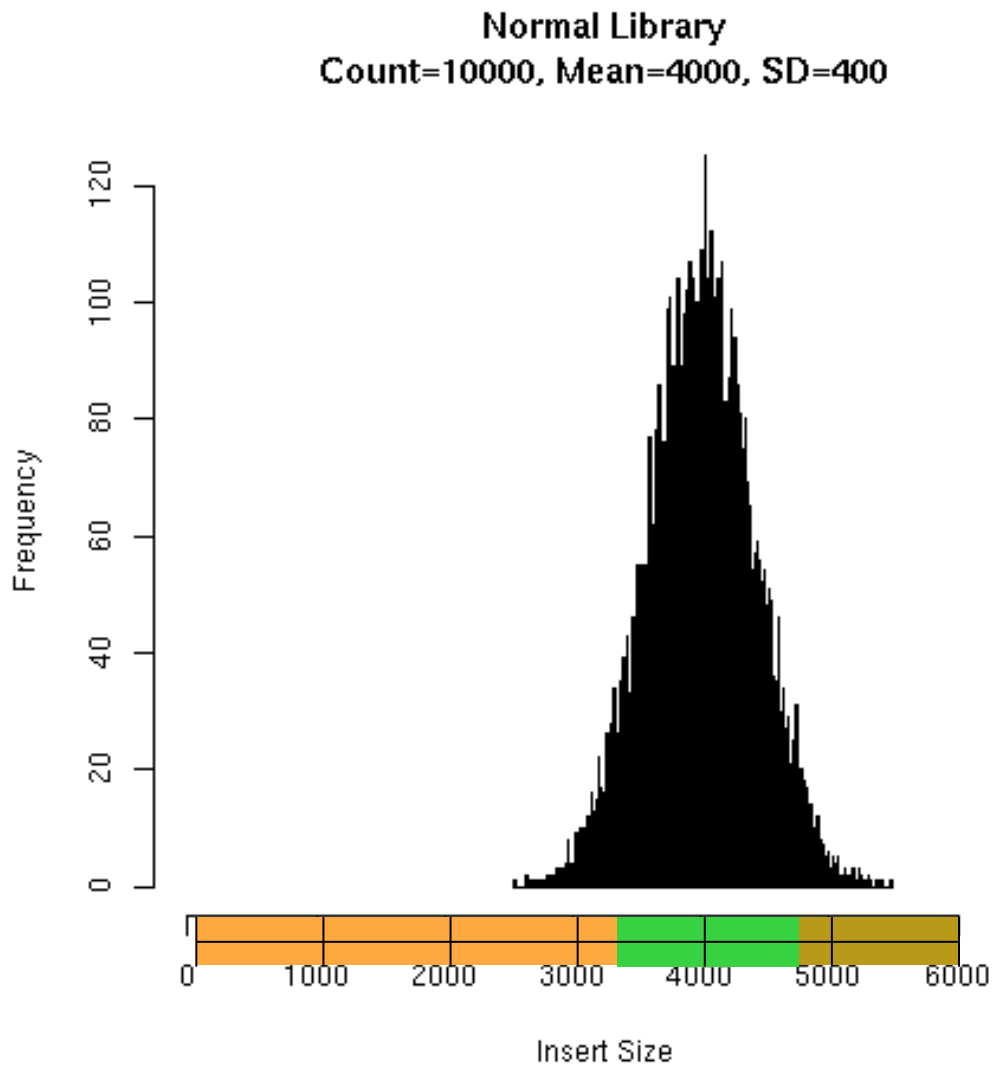
8 inserts: 3.2kb-6kb

Local Mean: 4461

C/E Stat: $\dfrac{(4461-4000)}{(400 / \sqrt{8})} = +3.26$

C/E Stat ≥ 3.0 indicates Expansion

CE Compression

Forensics

Normal Library
Count=10000, Mean=4000, SD=400
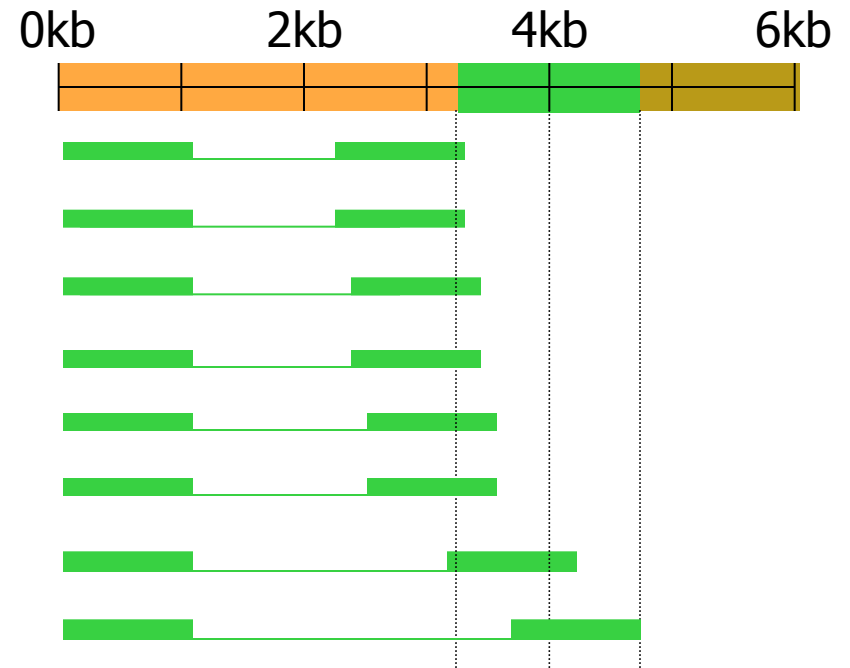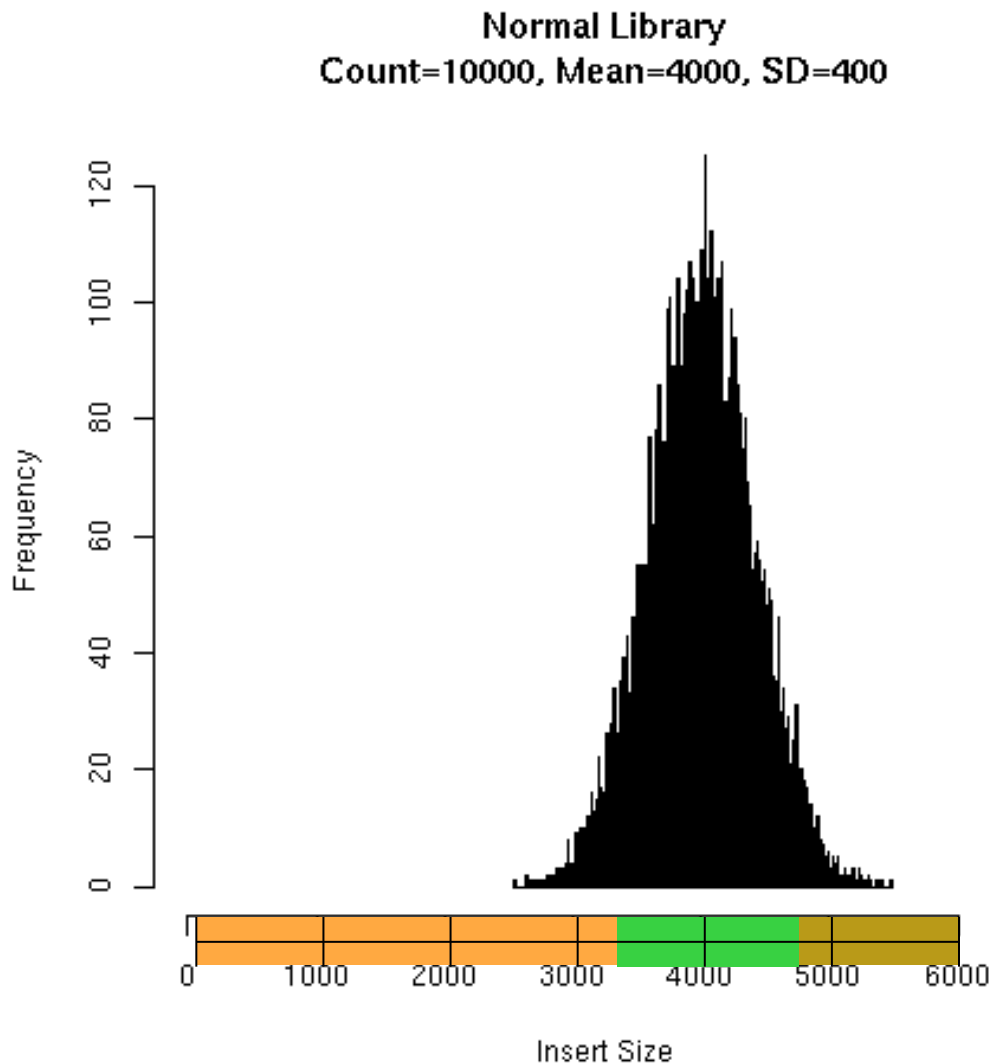
8 inserts: 3.2 kb-4.8kb
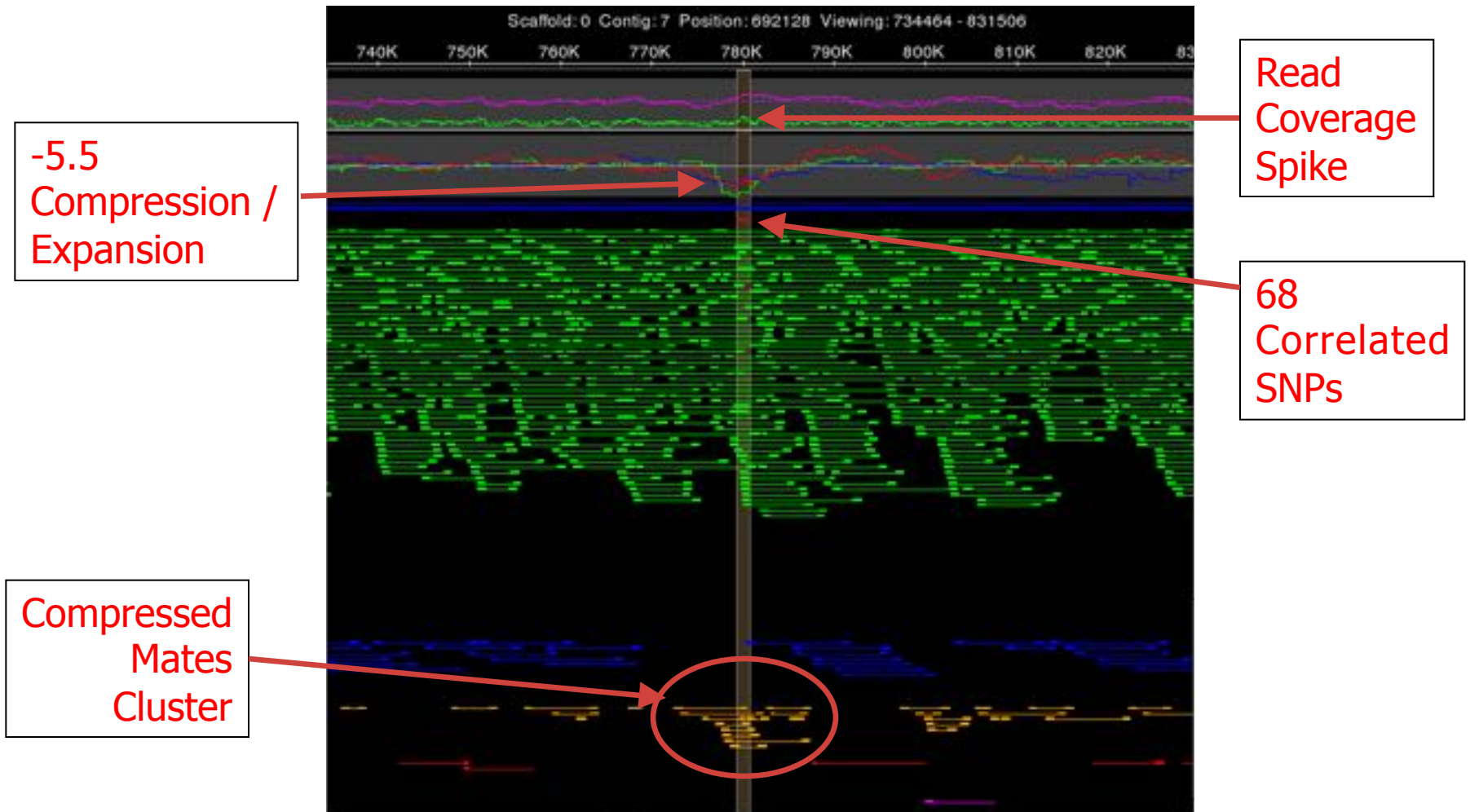
Local Mean: 3488

C/E Stat: $\dfrac{(3488-4000)}{(400 / \sqrt{8})}$ = -3.62

C/E Stat ≤ -3.0 indicates Compression

Collapsed Repeat Signature

Read Coverage Spike

-5.5 Compression / Expansion

68 Correlated SNPs

Compressed Mates Cluster

Hawkeye: a visual analytics tool for genome assemblies.
Schatz, MC, Phillippy, AM, Shneiderman, B, Salzberg, SL. (2007) Genome Biology 8:R34.

Forensics Performance

Table 1

Accuracy of amosvalidate mis-assembly signatures and suspicious regions summarized for 16 bacterial genomes assembled with Phrap

Species	Len	Ctgs	Errs	Mis-assembly signatures			Suspicious regions		
				Num	Valid	Sens	Num	Valid	Sens
B. anthracis	5.2	87	2	1,336	21	100.0	127	2	100.0
B. suis	3.4	120	10	1,047	30	80.0	158	9	90.0
C. burnetii	2.0	55	22	1,375	70	100.0	124	19	100.0
C. caviae	1.4	270	12	625	16	83.3	50	8	66.7
C. jejuni	1.8	53	5	290	11	80.0	61	3	60.0
D. ethenogenes	1.8	632	12	688	22	91.7	88	9	100.0
F. succinogenes	4.0	455	21	1,670	27	95.2	266	14	66.7
L. monocytogenes	2.9	172	1	1,381	5	100.0	201	1	100.0
M. capricolum	1.0	17	3	83	0	0.0	16	0	0.0
N. sennetsu	0.9	16	0	91	0	NA	13	0	NA
P. intermedia	2.7	243	21	1,655	57	100.0	201	20	100.0
P. syringae	6.4	274	64	2,841	200	98.4	366	55	98.4
S. agalactiae	2.1	127	21	687	53	95.2	112	18	85.7
S. aureus	2.8	824	41	1,850	69	97.6	227	18	75.6
W. pipientis	3.3	2017	31	761	92	100.0	132	30	100.0
X. oryzae	5.0	50	151	2,569	379	100.0	100	69	100.0
Totals	46.8	5412	417	18,949	1,052	96.9	2,242	275	92.6

Species name, genome length (Len), number of assembled contigs (Ctgs), and alignment inferred mis-assemblies (Errs) are given in the first four columns. Number of mis-assembly signatures output by amosvalidate (Num) is given in column 5, along with the number of signatures coinciding with a known mis-assembly in column 6 (Valid), and percentage of known mis-assemblies identified by one or more signatures in column 7 (Sens). The same values are given in columns 8-10 for the suspicious regions output by amosvalidate. The suspicious regions represent at least two different, coinciding lines of evidence, whereas the signatures represent a single line of evidence. A signature or region is deemed 'validated' if its location interval overlaps a mis-assembled region identified by dnadiff. Thus, a single signature or region can identify multiple mis-assemblies, and vice versa, a single mis-assembly can be identified by multiple signatures or regions.

Phillippy et al. Genome Biology 2008 **9**:R55   doi:10.1186/gb-2008-9-3-r55

96.9% sensitivity of mis-assemblies
Combining signatures into suspicious regions greatly improves specificity.

Genome Assembly
Gold-Standard Evaluation

http://gage.cbcb.umd.edu/

Ongoing Internal Evaluation Gone Public

- How much sequencing coverage do I need for my genome project?

- What can I expect the resulting assembly to look like?

- Which assembly software should I use?

- What parameters should I use when I run the software?

Genomes	***Assemblers***	***Evaluations***
Staphylococcus aureus	*ALLPATHS-LG*	*Connectivity*
Human chromosome 14	*Celera Assembler*	*Correctness*
Bombus impatiens	*Contrail*	*"Effort"*
Linepithema humile	*SOAPdenovo*	
	Velvet	

Final Thoughts

- Assembling 10,000 large vertebrate genomes requires substantial computational and human resources
  - Automate and parallelize as much as possible
  - Every genome seems to have its own challenges

- Any specific characteristics we focus on today will be hopelessly out of date tomorrow (or the next day)
  - Cost, read lengths, error model, pairs & strobes, bias
  - Software methods

- The consensus sequence is not sufficient
  - Where are the reads placed?
  - Where are the ambiguities?
  - How are the contigs related?

Acknowledgements

Univ. of Maryland
Steven Salzberg
Mihai Pop
Art Delcher
David Kelley
Daniella Puiu
James Yorke
Aleksey Zimin

NBACC
Adam Phillippy
Sergey Koren

JCVI
Granger Sutton
Jason Miller
Brain Wallenz

CSHL
Dick McCombie

Thank You

http://schatzlab.cshl.edu
@mike_schatz