# Applications of micro-, mega-, and meta- assembly

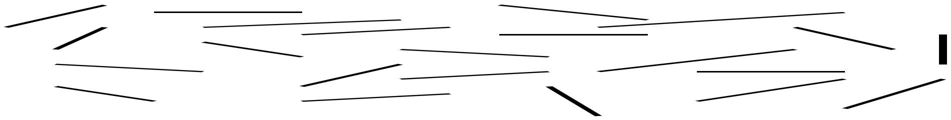
Michael Schatz

Dec. 9, 2011

CSHL In house

# Assembling a Genome

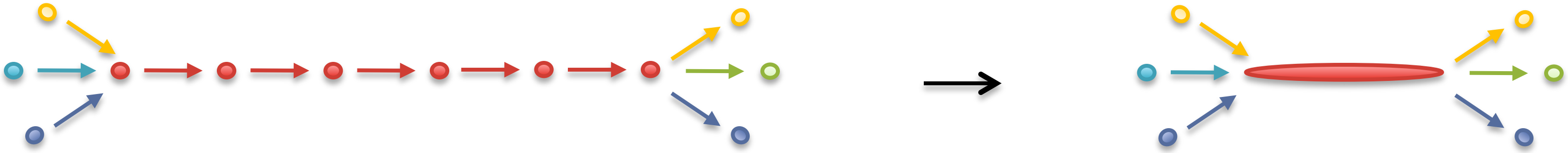
1. Shear & Sequence DNA



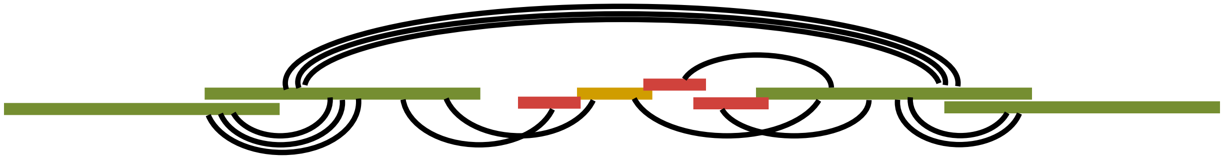
2. Construct assembly graph from overlapping reads

...AGCCTAGACCTACAGGATGCGCGACACGT
GGATGCGCGACACGTCGCATATCCGGT...

3. Simplify assembly graph



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



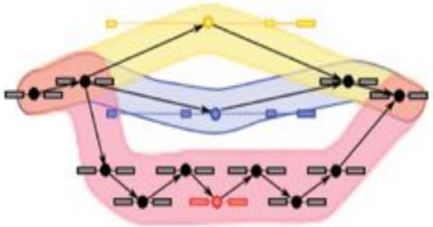
# Assembly Applications

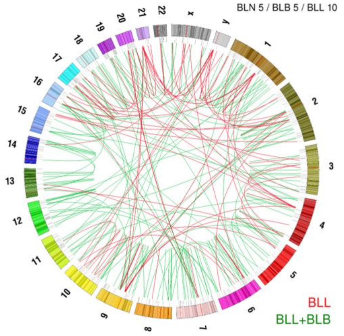
Novel genomes

Metagenomes

Sequencing assays

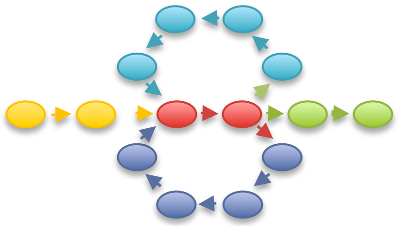
- Transcript assembly
- Structural variations
- Haplotype analysis
- …



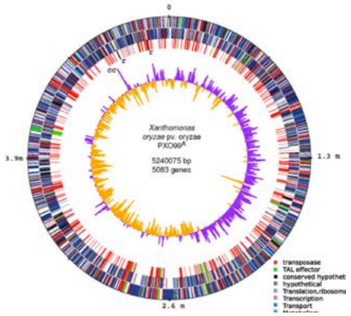


# Algorithms Overview

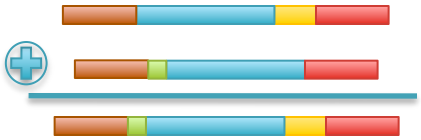
1. micro-
   - Microsatellite mutations
   - Haplotype Microassembly

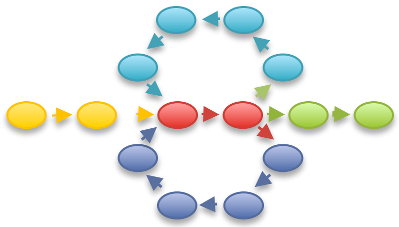


2. mega-
   - Genome Dark Matter
   - Cloud-scale Genome analysis
   - Single Molecule Sequencing & Assembly



3. meta-
   - Assembly Forensics & Metassembly





# micro-

# *MicroSeq*: Microsatellite Analysis

M. Bekritsky, J. Troge, D. Levy, M. Wigler, M. Schatz

- Highly variable simple sequence repeats
  - …GCACACACACAT… = …G(CA)$_5$T…
  - Mutate by slippage during replication, creating indels
  - High mutation rate makes it a useful marker for inferring phylogeny, associated with many diseases

- Genotyping with MicroSeq:
  1. Rapidly detect MS in short reads
  2. Map reads using a new MS-mapper
  3. Analyze profiles across populations

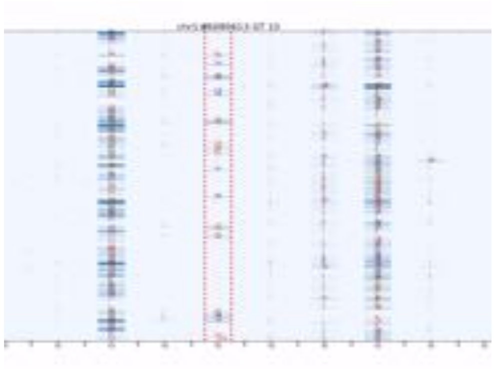
- Currently looking at de novo mutations associated with autism

```
          CTAGCCCCCTGTACG
          TAGCCCCCCCCCCTG
          AGCCCCCTGTACGAT
          GCCCCCCCCCCTGTA
```

↓ Map

```
REF: ATGACTAGCCCCCCCCCCTGTACGATTTCG
         CTAGCCCCC-----TGTACG
          TAGCCCCCCCCCCTG
           AGCCCCC-----TGTACGAT
            GCCCCCCCCCCTGTA
```

↓ Profile



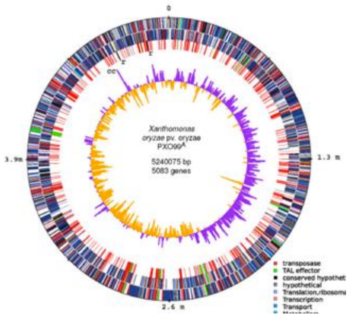
# *Scalpel*: Haplotype Microassembly

G. Narzisi, D. Levy, I. Iossifov, J. Kendall, M. Wigler, M. Schatz

- Use assembly techniques to identify complex variations from short reads
  - Improved power to find indels
  - Trace candidate haplotypes sequences as paths through assembly graphs

```
Ref:     ...CACAGGATCCACCTTTCTCAAAGACCCAGGATCCTCCTTCCTCGGTGACACTGTATACGTC...

Father:  ...CACAGGATCCACCTTTCTCAAAGACCCAGGATCCTCCTTCCTCGGTGACACTGTATACGTC... [cov:19.5]

Mother_1:...CACAGGATCCACCTTTCTCAAAGACCCAGGATCCTCCTTCCTCGGTGACACTGTATACGTC... [cov:19.4]
Mother_2:...CACAGGATCCACCTTT------------------------CTTGGTGACACTGTATACGTC... [cov:21.5]

Aut_2:   ...CACAGGATCCACCTTTCTCAAAGACCCAGGATCCTCCTTCCTCGGTGACACTGTATACGTC... [cov:28.2]
Aut_1:   ...CACAGGATCCACCTTT------------------------CTTGGTGACACTGTATACGTC... [cov:33.3]

Sib_1:   ...CACAGGATCCACCTTTCTCAAAGACCCAGGATCCTCCTTCCTCGGTGACACTGTATACGTC... [cov:19.4]
Sib_2:   ...CACAGGATCCACCTTT------------------------CTTGGTGACACTGTATACGTC... [cov:21.5]
```

24 bp heterozygous indel at chr5:176026122 GPRIN1


0
1.3 m
2.6 m
3.9m
Xanthomonas oryzae pv. oryzae PXO99A
5240075 bp
5083 genes
transposase
TAL effector
conserved hypothet
hypothetical
Translation,ribosoma
Transcription
Transport

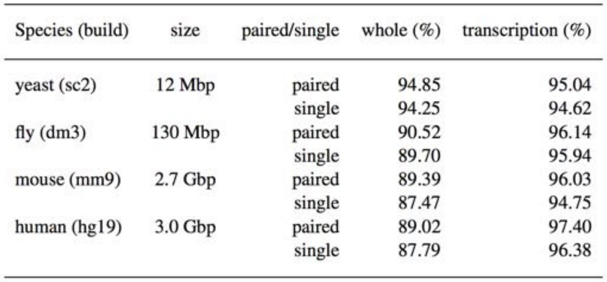
# mega-

# Genomic Dark Matter

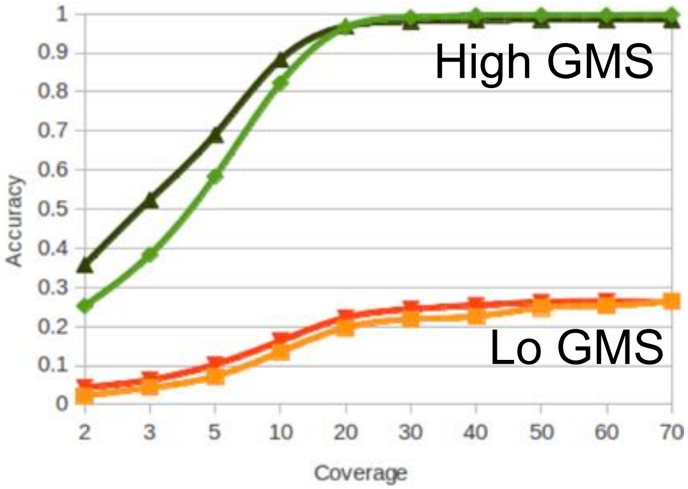
Hayan Lee, Michael Schatz

- Short read mapping is a essential for identifying mutations in the genome
  - Not every base of the genome can mapped equally well, especially because of repeats

- Introduced a new probabilistic metric - the Genome Mappability Score - that quantifies how reliably reads can be mapped to every position in the genome
  - We have little power to measure 11-13% of the human genome, including of known clinically relevant variations
  - Errors in variation discovery are dominated by false positives, especially in low GMS regions



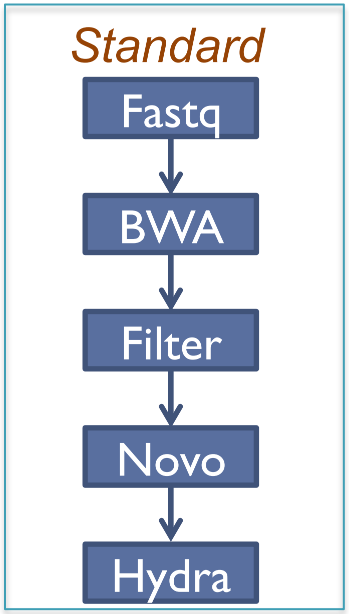
| Species (build) | size | paired/single | whole (%) | transcription (%) |
|---|---|---|---|---|
| yeast (sc2) | 12 Mbp | paired | 94.85 | 95.04 |
| | | single | 94.25 | 94.62 |
| fly (dm3) | 130 Mbp | paired | 90.52 | 96.14 |
| | | single | 89.70 | 95.94 |
| mouse (mm9) | 2.7 Gbp | paired | 89.39 | 96.03 |
| | | single | 87.47 | 94.75 |
| human (hg19) | 3.0 Gbp | paired | 89.02 | 97.40 |
| | | single | 87.79 | 96.38 |

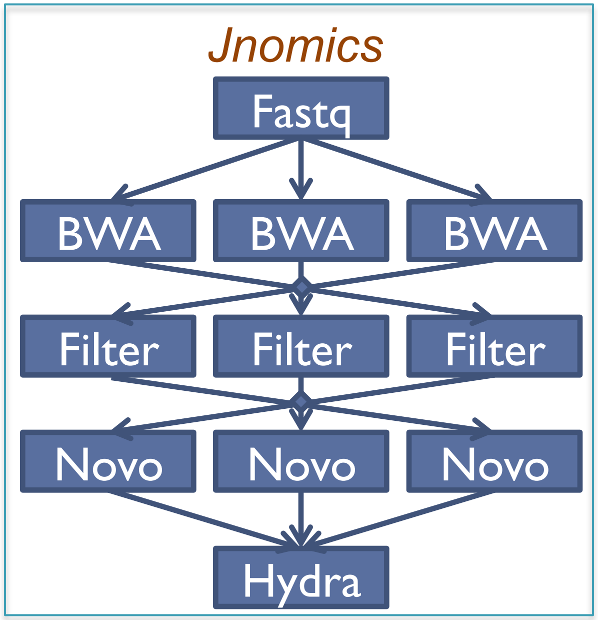


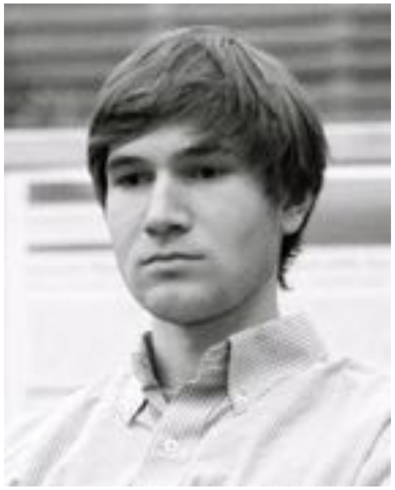
**Genomic Dark Matter: The reliability of short read mapping illustrated by the GMS.**
Lee, H., Schatz, M.C. (2011) *Under Review*

# *Jnomics*: Cloud-scale genomics

Matt Titmus, James Gurtowski, Michael Schatz







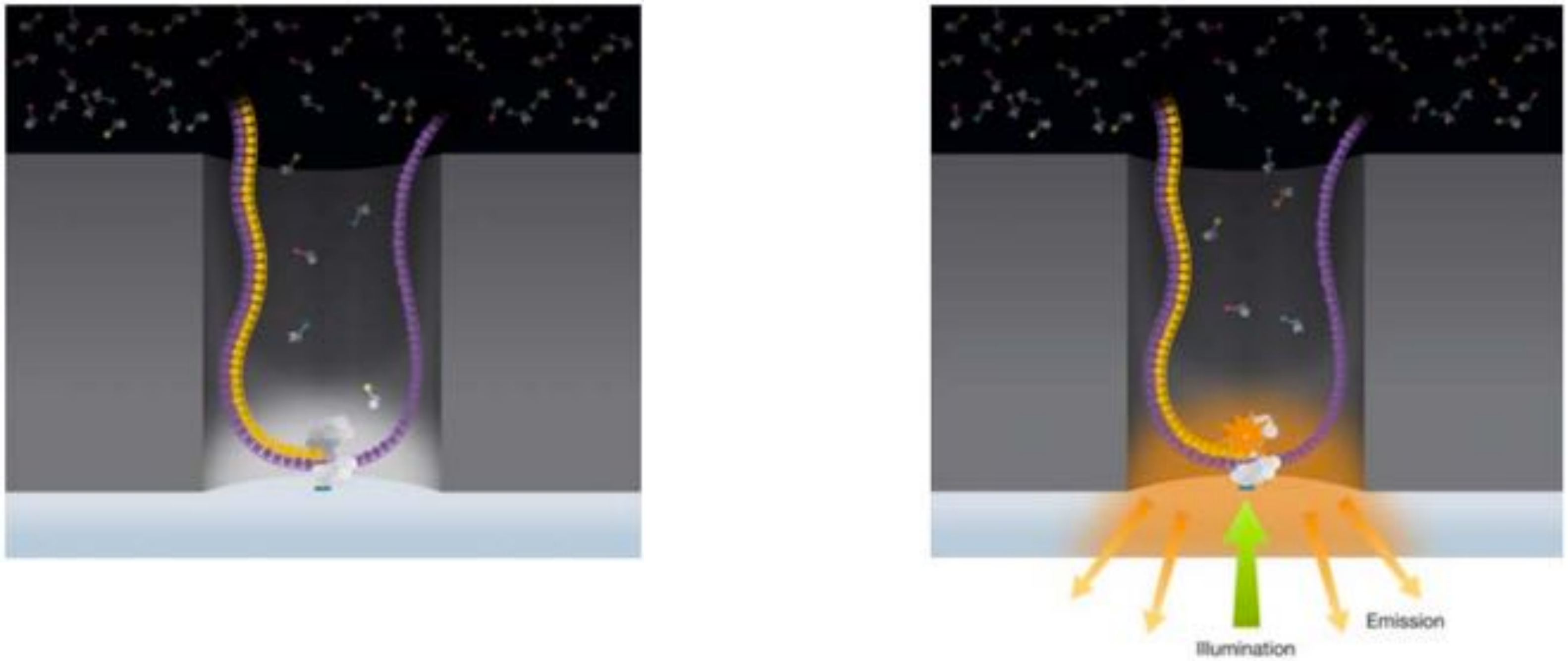
- Rapid parallel execution of NGS analysis pipelines
  - FASTX, BWA, Novoalign, SAMTools, Hydra
  - Sorting, merging, filtering, selection, of BAM, SAM, BED, fastq
- Case study: Structural variations in esophageal cancer

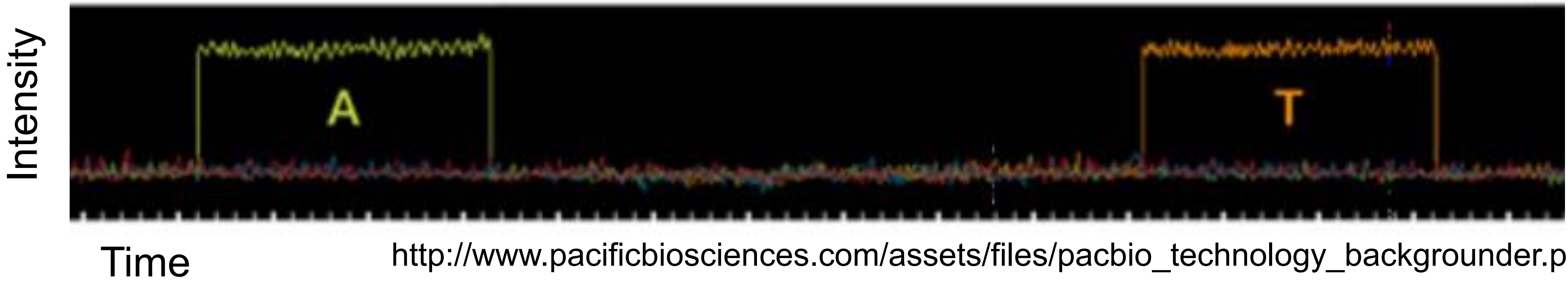
**Answering the demands of digital genomics**
Titmus, M.A.., Schatz, M.C.. (2011) *Under Review*

# Pacific Biosciences RS

## Single Molecule Real Time (SMRT) Sequencing

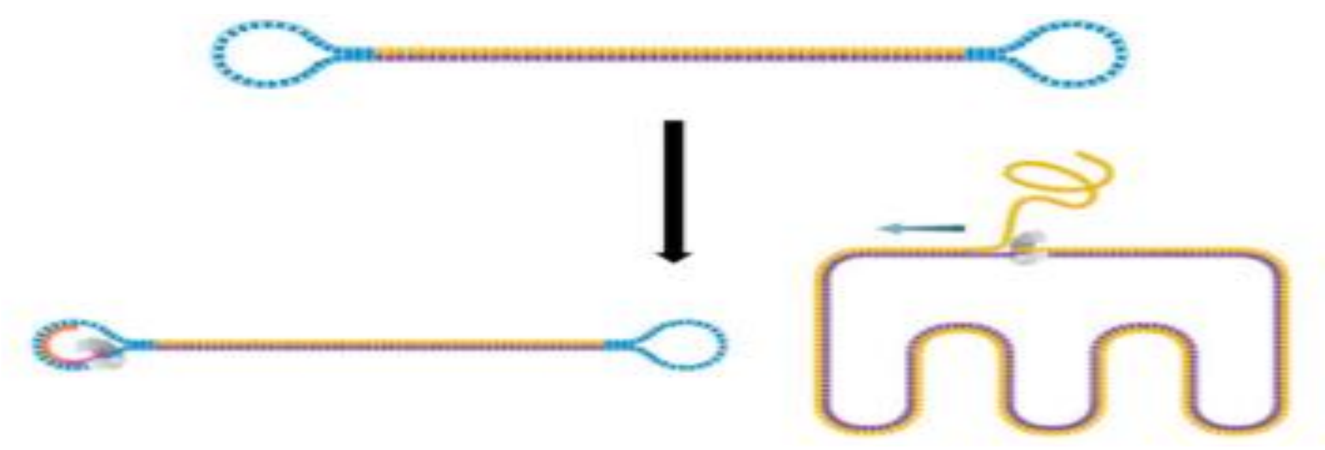
Imaging of florescent phospholinked labeled nucleotides as they are incorporated by a polymerase anchored to a Zero-Mode Waveguide (ZMW).





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

# SMRT Read Types



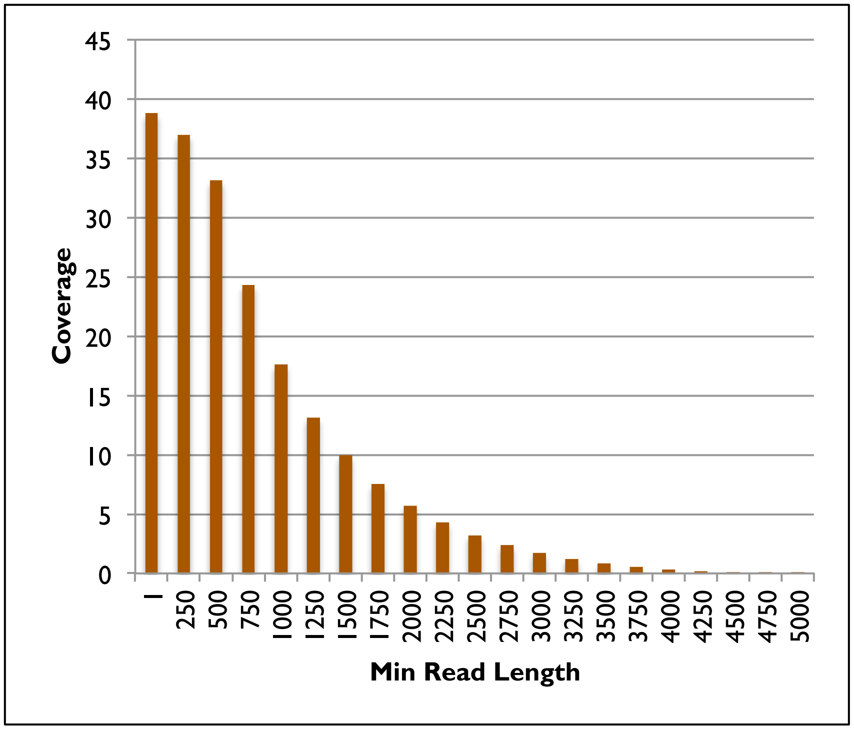
- ***Standard sequencing***
  - Long inserts so that the polymerase can synthesize along a single strand
- ***Circular consensus sequencing***
  - Short inserts, so polymerase can continue around the entire SMRTbell multiple times and generate multiple sub-reads from the same single molecule.
- ***Strobe sequencing***
  - Very long inserts, alternate the lasers in the instrument between on and off. On periods generate strobe sub-reads and the off periods determine the length of the spacing between, known as strobe advance

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

# Read Quality

**Yeast**
**(12 Mbp)**

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



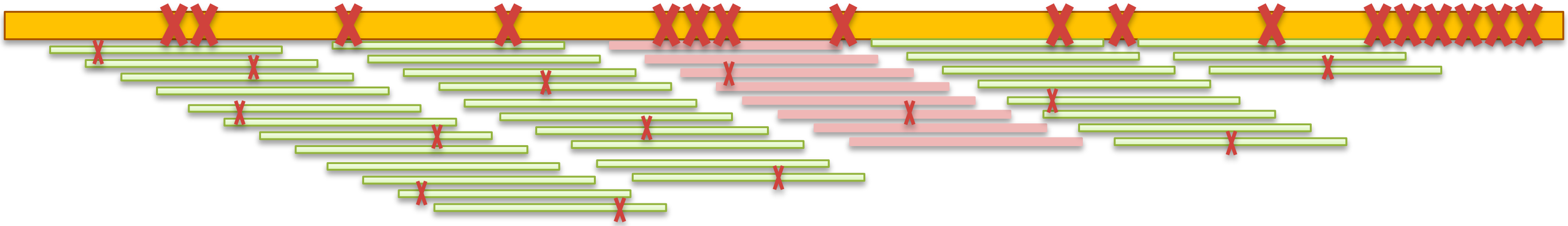
```
TTGTAAGCAGTTGAAAACTATGTGTGGATTTAGAATAAAGAACATGAAAG
||||||||||||||||||||||||| ||||||| |||||||||||| |||
TTGTAAGCAGTTGAAAACTATGTGT-GATTTAG-ATAAAGAACATGGAAG

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 & Assembly

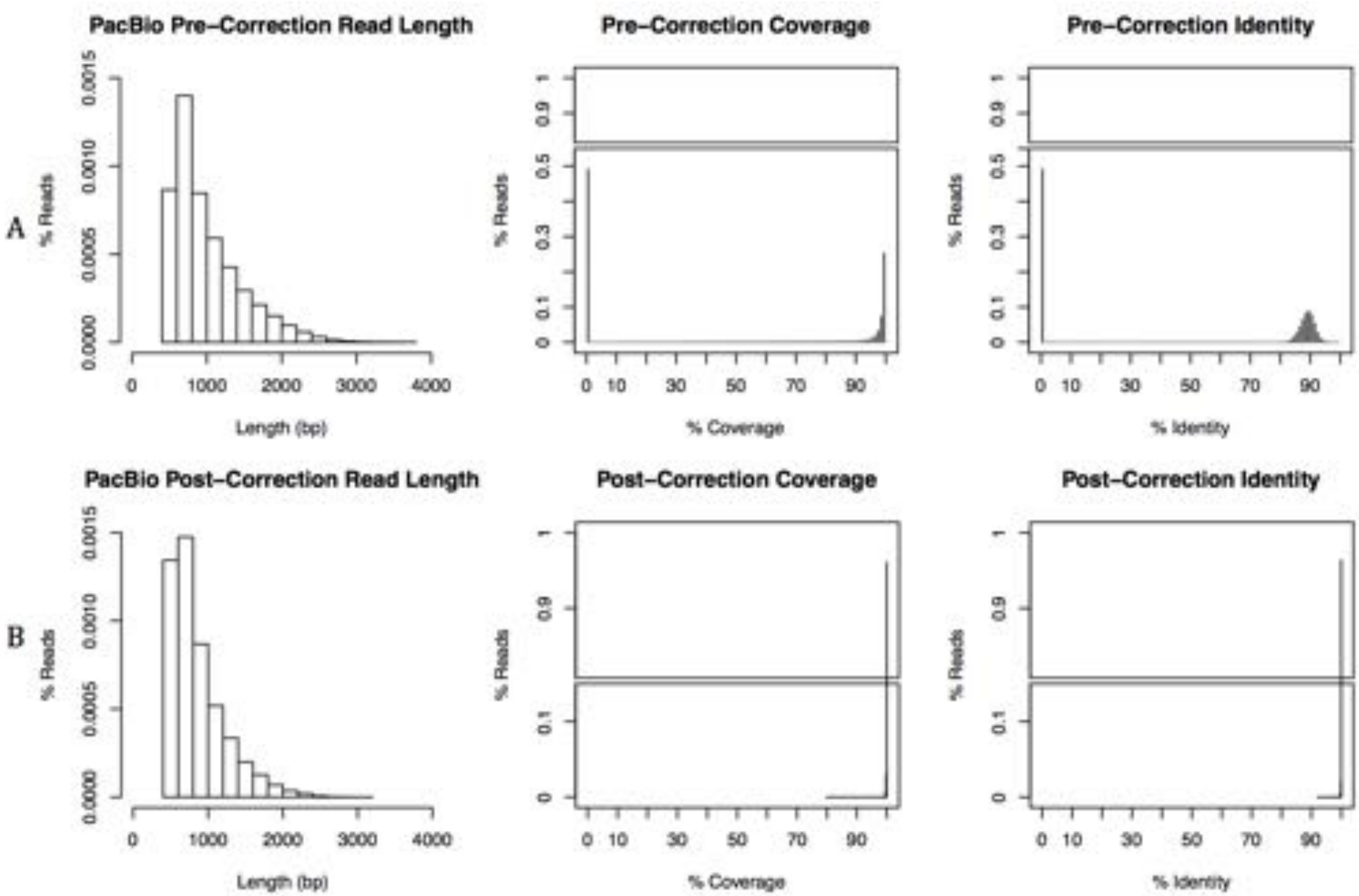
http://wgs-assembler.sf.net

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. Co-assemble corrected LRs and SRs
   – Celera Assembler enhanced to support 32 Kbp reads
3. Error corrected reads can be easily assembled, aligned



**Hybrid error correction and de novo assembly of single-molecule sequencing reads.**
Koren, S, Schatz, MC, Walenz, BP, Martin, J, Howard, J, Ganapathy, G, Wang, Z, Rasko, DA, McCombie, WR, Jarvis, ED, Phillippy, AM. (2011) *Under Review*

# Error Correction Results

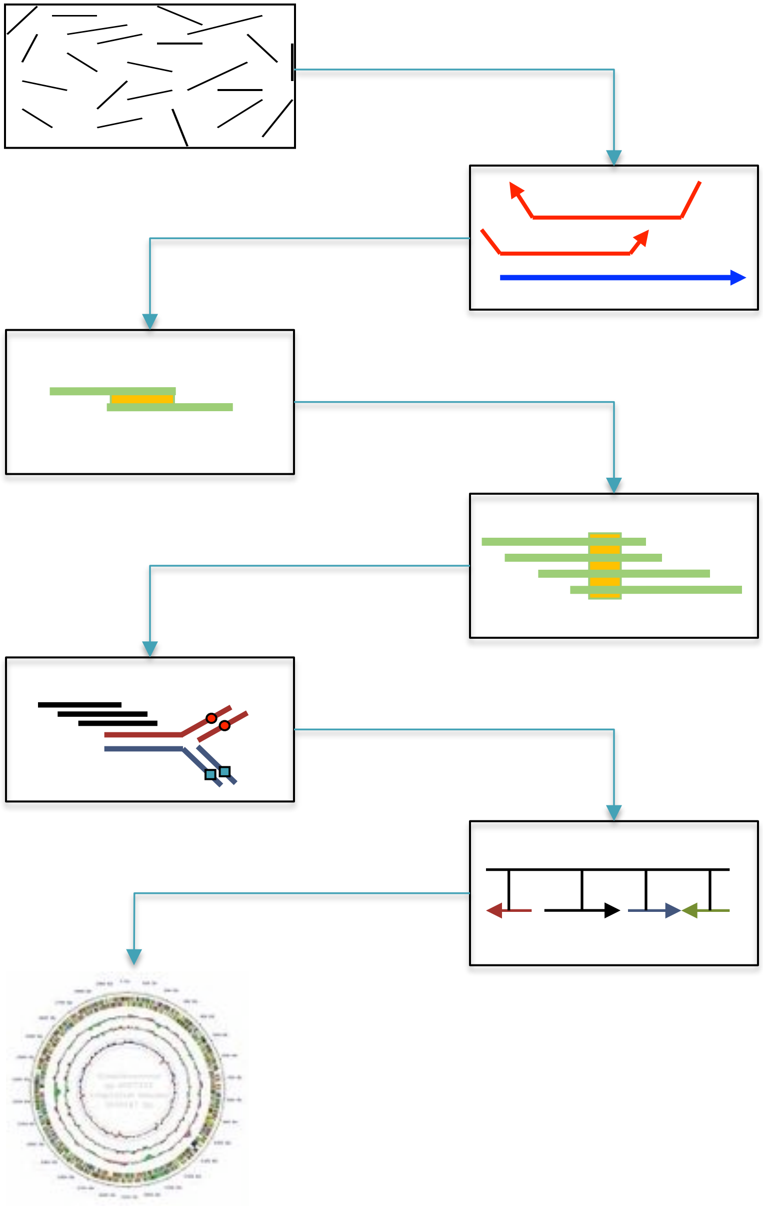


Correction results of 20x PacBio coverage of E. coli K12 corrected using 50x Illumina

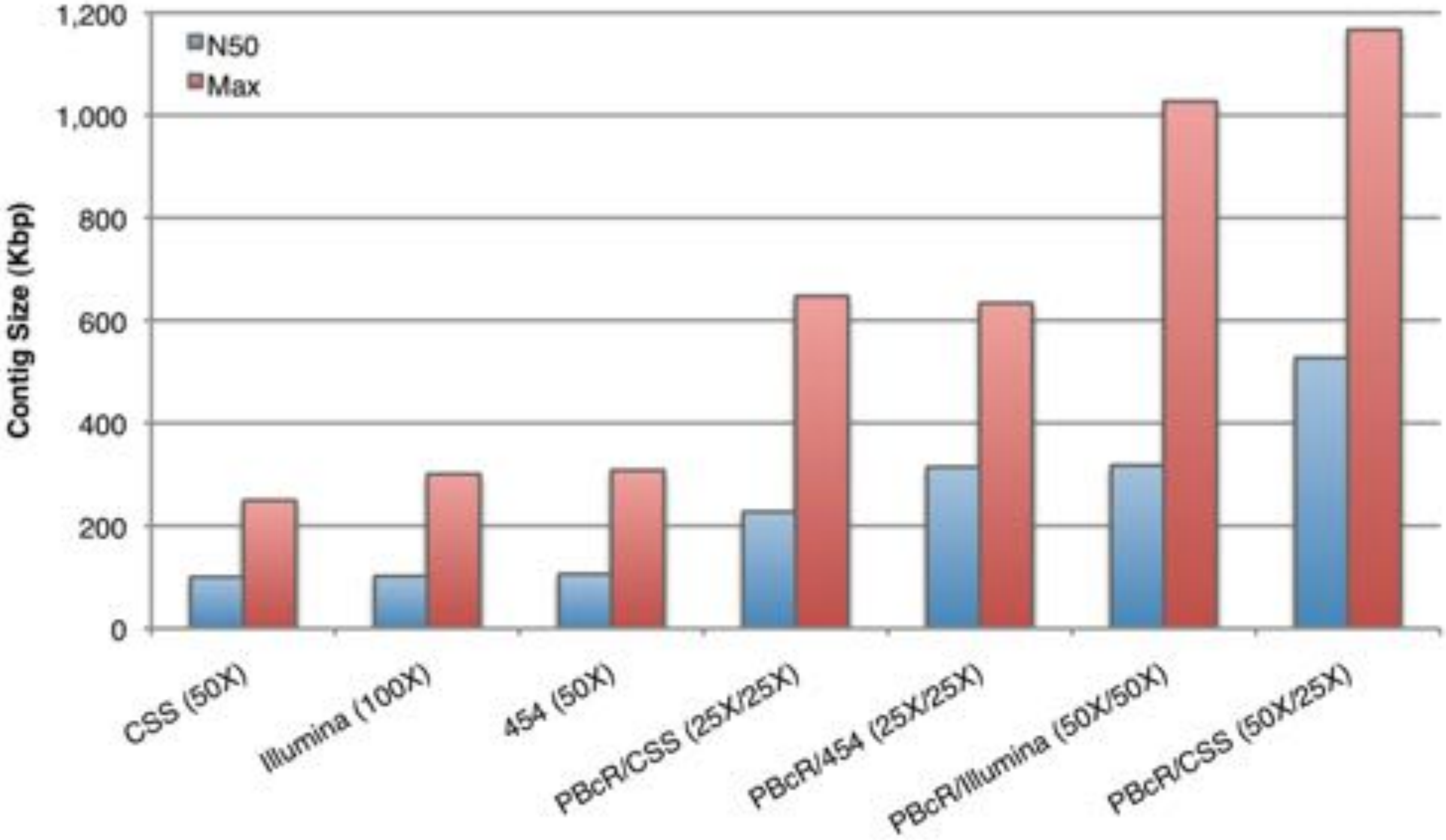
# Celera Assembler

*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

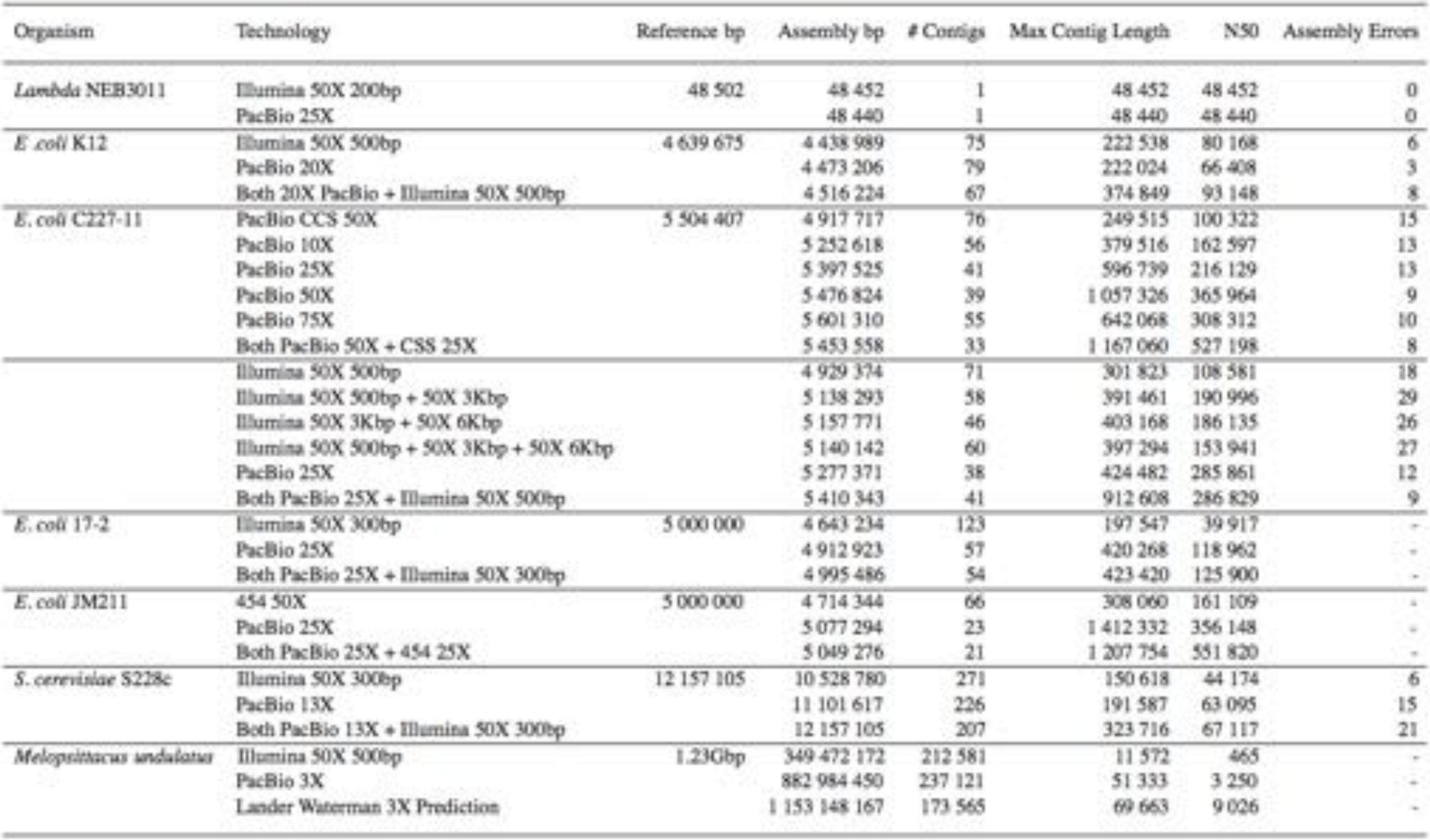


# Assembly Results



SMRT-hybrid assembly results of 50x PacBio corrected coverage of E. coli K12
Long reads lead to ***contigs*** over 1Mbp

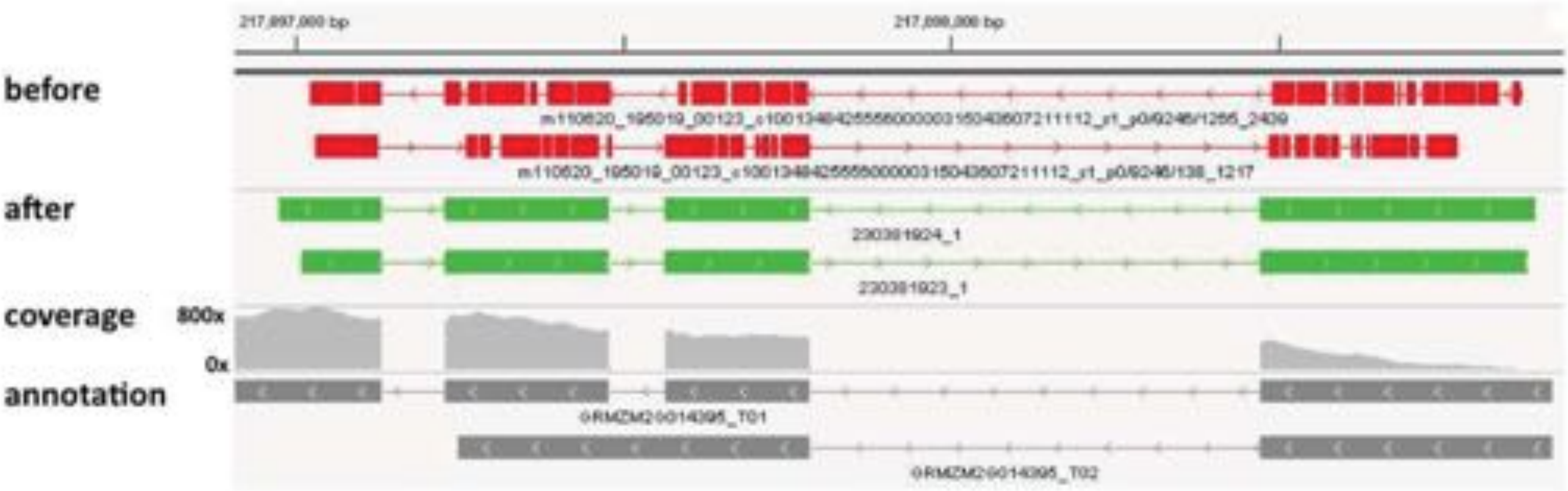
# Hybrid Assembly Results



| Organism | Technology | Reference bp | Assembly bp | # Contigs | Max Contig Length | N50 | Assembly Errors |
|---|---|---|---|---|---|---|---|
| *Lambda* NEB3011 | Illumina 50X 200bp | 48 502 | 48 452 | 1 | 48 452 | 48 452 | 0 |
| | PacBio 25X | | 48 440 | 1 | 48 440 | 48 440 | 0 |
| *E. coli* K12 | Illumina 50X 500bp | 4 639 675 | 4 438 989 | 75 | 222 538 | 80 168 | 6 |
| | PacBio 20X | | 4 473 206 | 79 | 222 024 | 66 408 | 3 |
| | Both 20X PacBio + Illumina 50X 500bp | | 4 516 224 | 67 | 374 849 | 93 148 | 8 |
| *E. coli* C227-11 | PacBio CCS 50X | 5 504 407 | 4 917 717 | 76 | 249 515 | 100 322 | 15 |
| | PacBio 10X | | 5 252 618 | 56 | 379 516 | 162 597 | 13 |
| | PacBio 25X | | 5 397 525 | 41 | 596 739 | 216 129 | 13 |
| | PacBio 50X | | 5 476 824 | 39 | 1 057 326 | 365 964 | 9 |
| | PacBio 75X | | 5 601 310 | 55 | 642 068 | 308 312 | 10 |
| | Both PacBio 50X + CSS 25X | | 5 453 558 | 33 | 1 167 060 | 527 198 | 8 |
| | Illumina 50X 500bp | | 4 929 374 | 71 | 301 823 | 108 581 | 18 |
| | Illumina 50X 500bp + 50X 3Kbp | | 5 138 293 | 58 | 391 461 | 190 996 | 29 |
| | Illumina 50X 3Kbp + 50X 6Kbp | | 5 157 771 | 46 | 403 168 | 186 135 | 26 |
| | Illumina 50X 500bp + 50X 3Kbp + 50X 6Kbp | | 5 140 142 | 60 | 397 294 | 153 941 | 27 |
| | PacBio 25X | | 5 277 371 | 38 | 424 482 | 285 861 | 12 |
| | Both PacBio 25X + Illumina 50X 500bp | | 5 410 343 | 41 | 912 608 | 286 829 | 9 |
| *E. coli* 17-2 | Illumina 50X 300bp | 5 000 000 | 4 643 234 | 123 | 197 547 | 39 917 | - |
| | PacBio 25X | | 4 912 923 | 57 | 420 268 | 118 962 | - |
| | Both PacBio 25X + Illumina 50X 300bp | | 4 995 486 | 54 | 423 420 | 125 900 | - |
| *E. coli* JM211 | 454 50X | 5 000 000 | 4 714 344 | 66 | 308 060 | 161 109 | - |
| | PacBio 25X | | 5 077 294 | 23 | 1 412 332 | 356 148 | - |
| | Both PacBio 25X + 454 25X | | 5 049 276 | 21 | 1 207 754 | 551 820 | - |
| *S. cerevisiae* S228c | Illumina 50X 300bp | 12 157 105 | 10 528 780 | 271 | 150 618 | 44 174 | 6 |
| | PacBio 13X | | 11 101 617 | 226 | 191 587 | 63 095 | 15 |
| | Both PacBio 13X + Illumina 50X 300bp | | 12 157 105 | 207 | 323 716 | 67 117 | 21 |
| *Melopsittacus undulatus* | Illumina 50X 500bp | 1.23Gbp | 349 472 172 | 212 581 | 11 572 | 465 | - |
| | PacBio 3X | | 882 984 450 | 237 121 | 51 333 | 3 250 | - |
| | Lander Waterman 3X Prediction | | 1 153 148 167 | 173 565 | 69 663 | 9 026 | - |

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

# Transcript Alignment


217,897,000 bp
217,898,000 bp
before
m110620_195019_00123_c100134842555500000315043607211112_s1_p0/9246/1265_2439
m110620_195019_00123_c100134842555500000315043607211112_s1_p0/9246/138_1217
after
230381924_1
230381923_1
coverage
800x
0x
annotation
GRMZM2G014395_T01
GRMZM2G014395_T02

- 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

# meta-

# 2011: Year of the Assembly Bakeoff

THE ASSEMBLATHON

- Simulated genome distantly related to human chr13
- 17 labs, 50+ assemblies

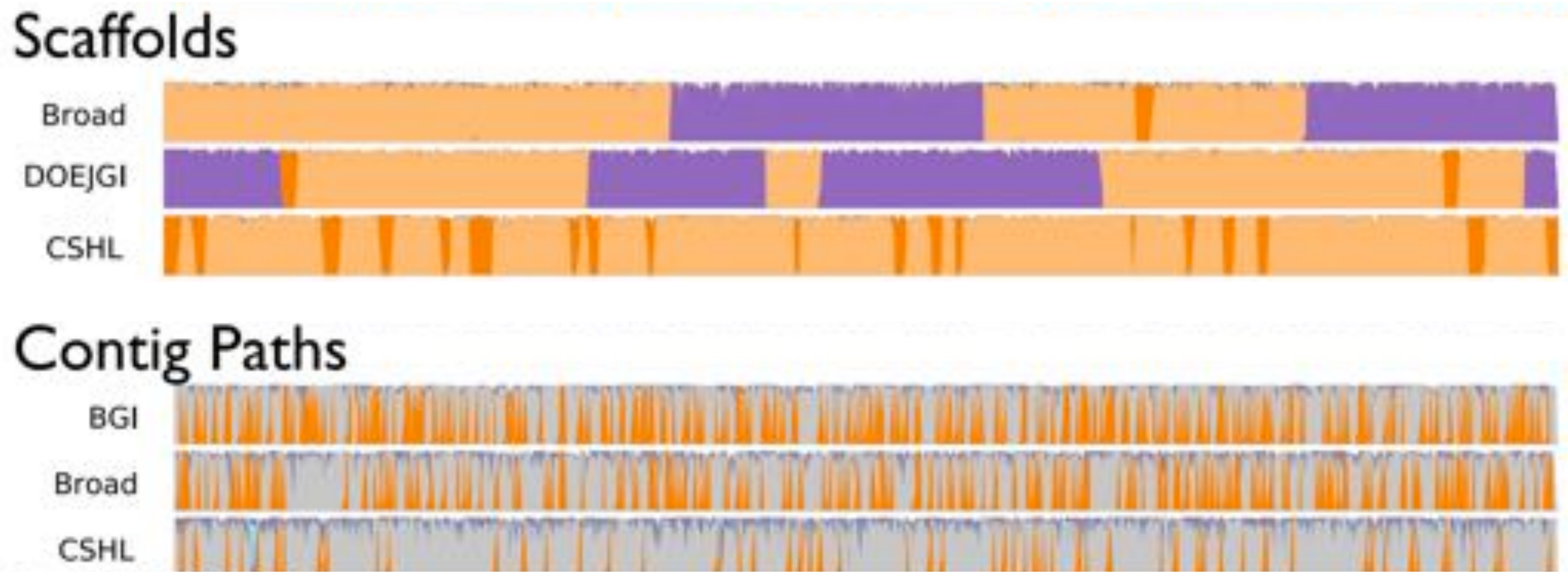
Genome Assembly Gold-Standard Evaluations

- 4 real genomes ranging from bacteria to individual human chromosome
- Internal evaluation of 8 assemblers

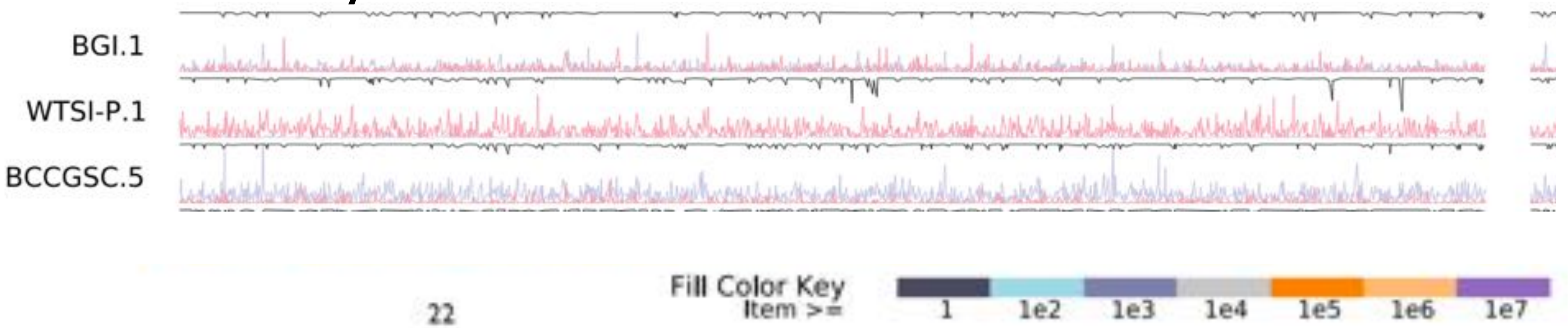
**Assemblathon 1: A competitive assessment of de novo short read assembly methods.**
Earl, DA *et al.* (2011) *Genome Research*. In press.

**GAGE: A critical evaluation of genome assemblies and assembly algorithms.**
Salzberg, SL *et al.* (2011) *Genome Research*. In press.

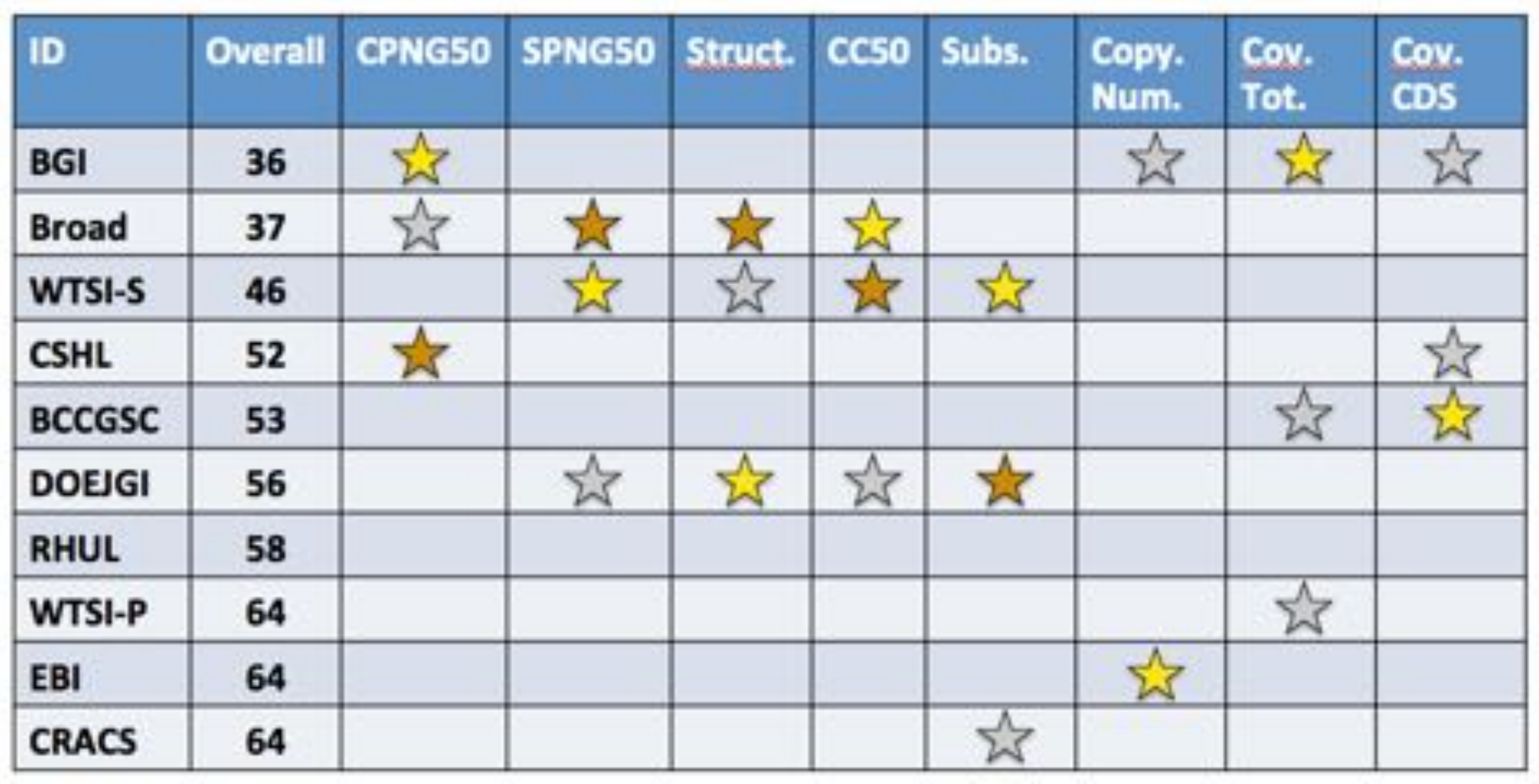
# Assemblathon Results



## Mis-assembly Markers



# 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, and ABySS

- My recommendation for "typical" short read assembly is to use ALLPATHS

# Assemblathon 2

- Real sequence data, *de novo* assembly

- Step 1: Apply best practices from Assemblathon 1
- Step 2: Add secret weapon for winning…

Images from Assemblathon

Forensics

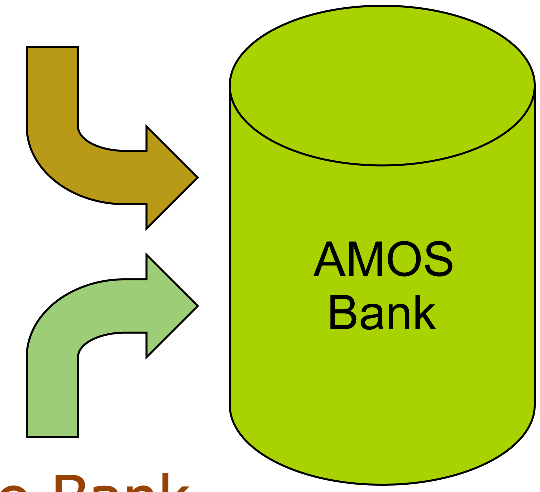
# Assembly Forensics

Computationally scan an assembly for mis-assemblies.

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

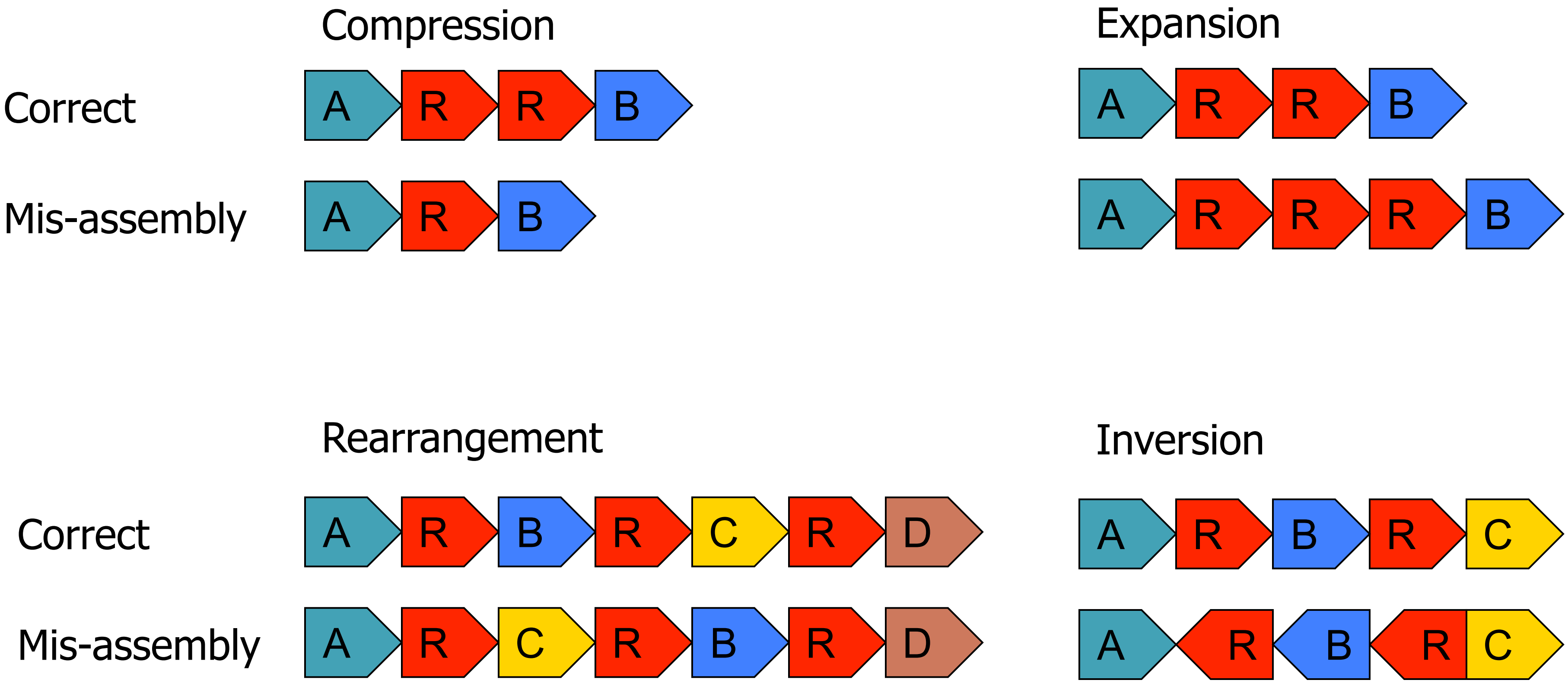
## AMOSvalidate

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



**Genome Assembly forensics: finding the elusive mis-assembly.**
Phillippy, AM, Schatz, MC, Pop, M. (2008) Genome Biology 9:R55.

# Mis-assembly types



Compression

Expansion

Correct

Mis-assembly

Rearrangement

Inversion

Correct

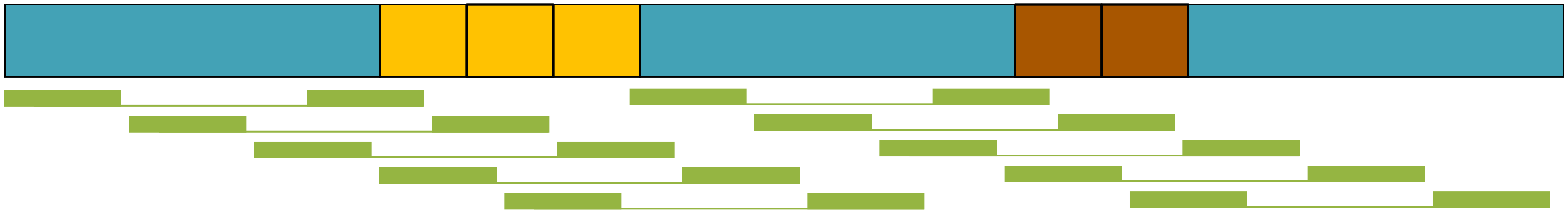
Mis-assembly

Basic mis-assemblies can be combined into more complicated patterns:
Insertions, Deletions, Giant Hairballs

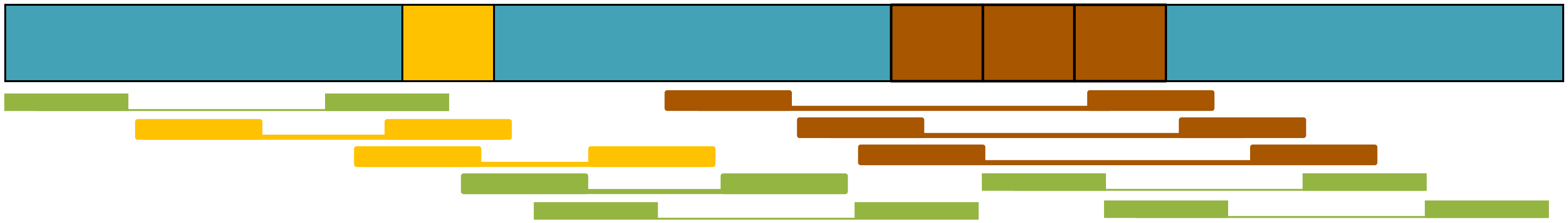
Forensics

# Mate Evaluation

- Correct: mates have expected orientation and separation



- Mis-assembled: mates have incorrect orientation and separation

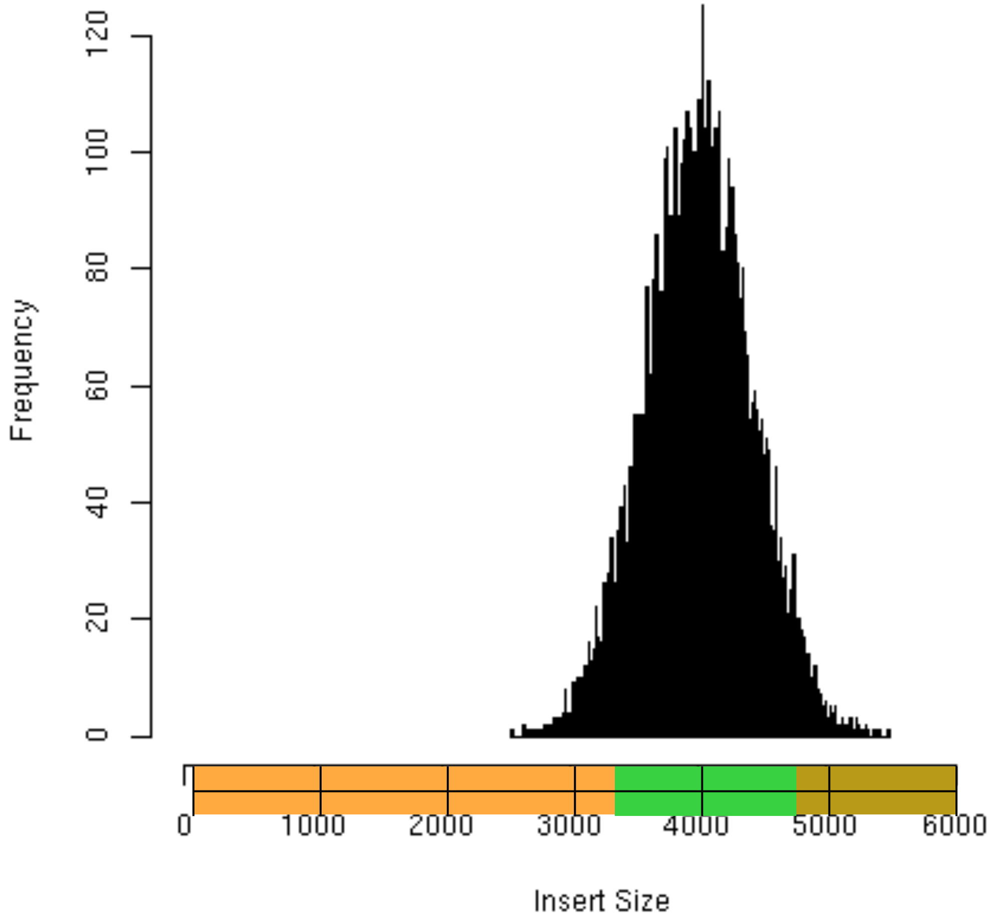


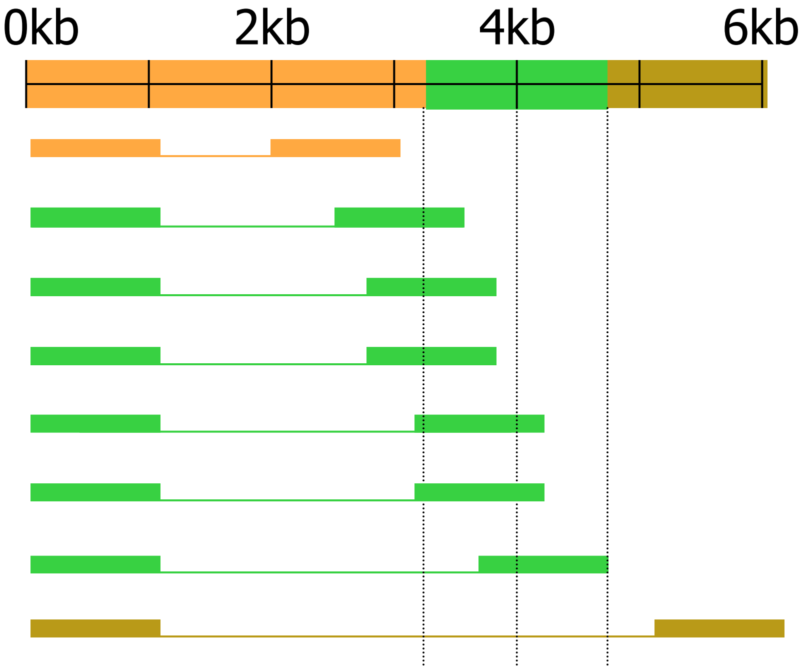
- Slightly compressed/expanded mates are expected because mates are sampled from a distribution of fragments

# Compression/Expansion Statistic

Library size distribution

Mean: 4000, SD: 400





8 inserts: 3kb-6kb

Local Mean: 4048

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

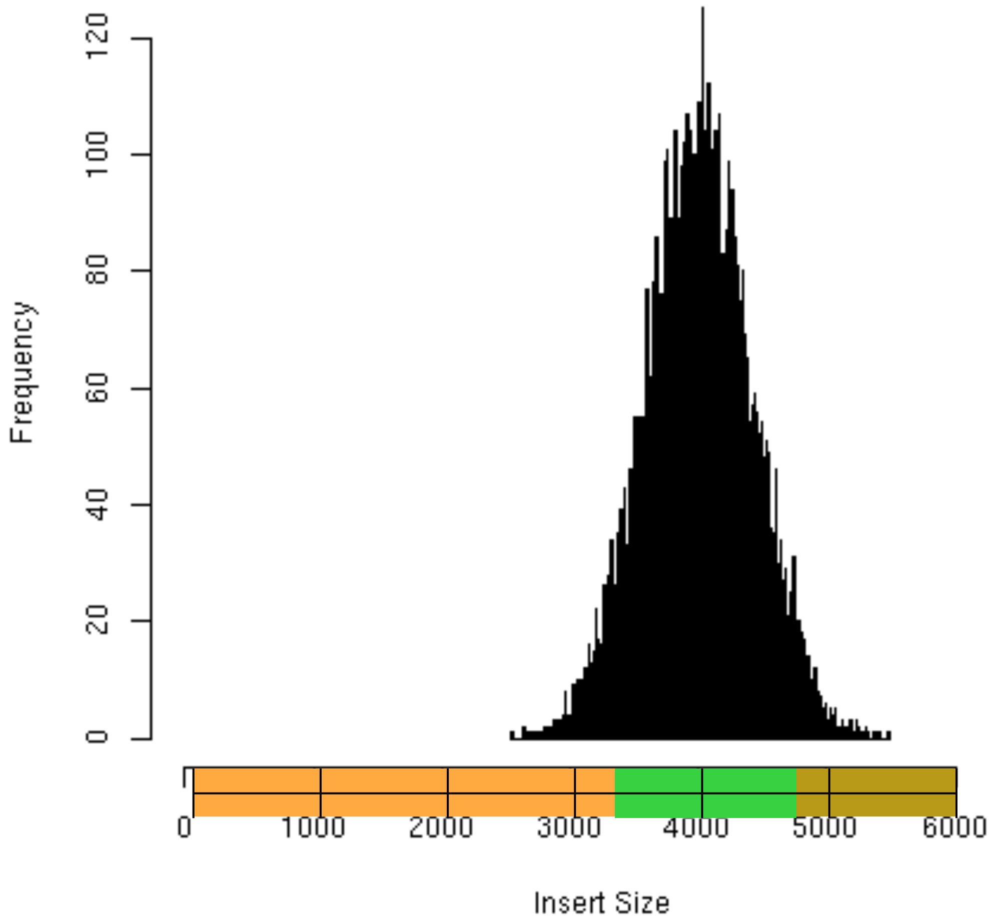
Near 0 indicates overall happiness

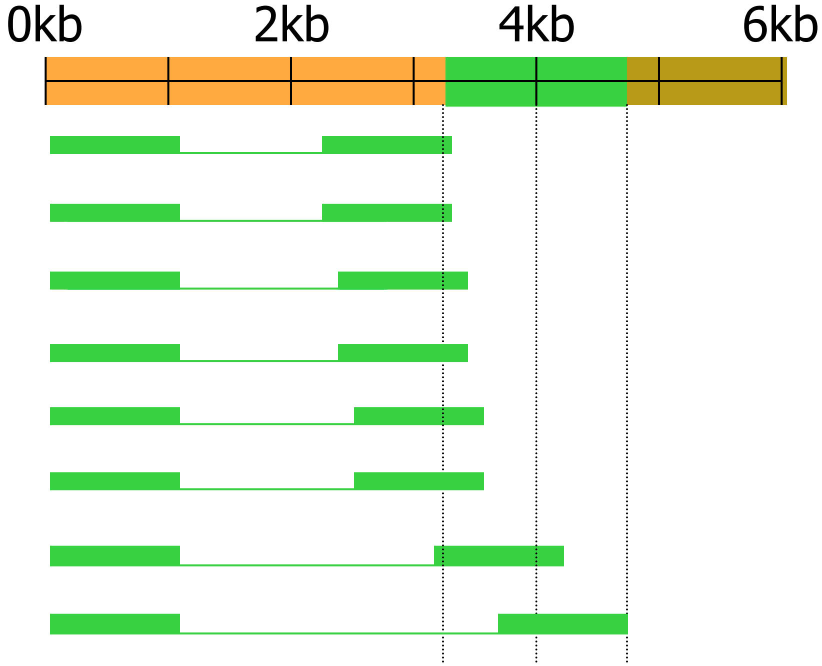
Forensics

# Hidden Compression

Library size distribution

Mean: 4000, SD: 400





8 inserts: 3.2 kb-4.8kb

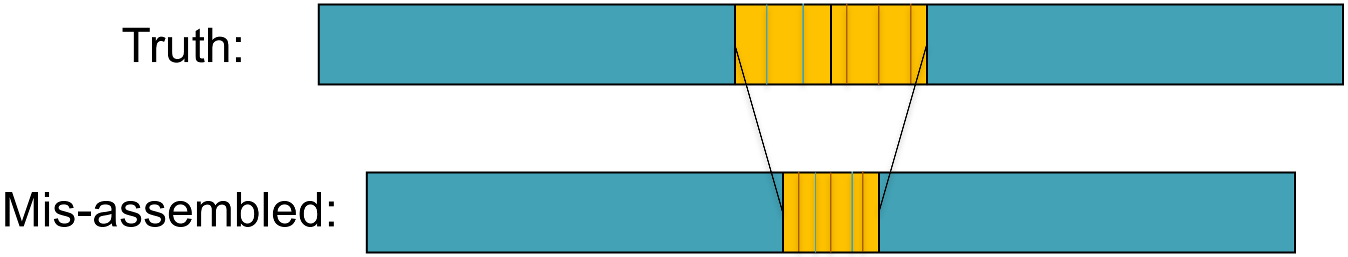
Local Mean: 3488

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

C/E Stat ≤ -3.0 indicates Compression

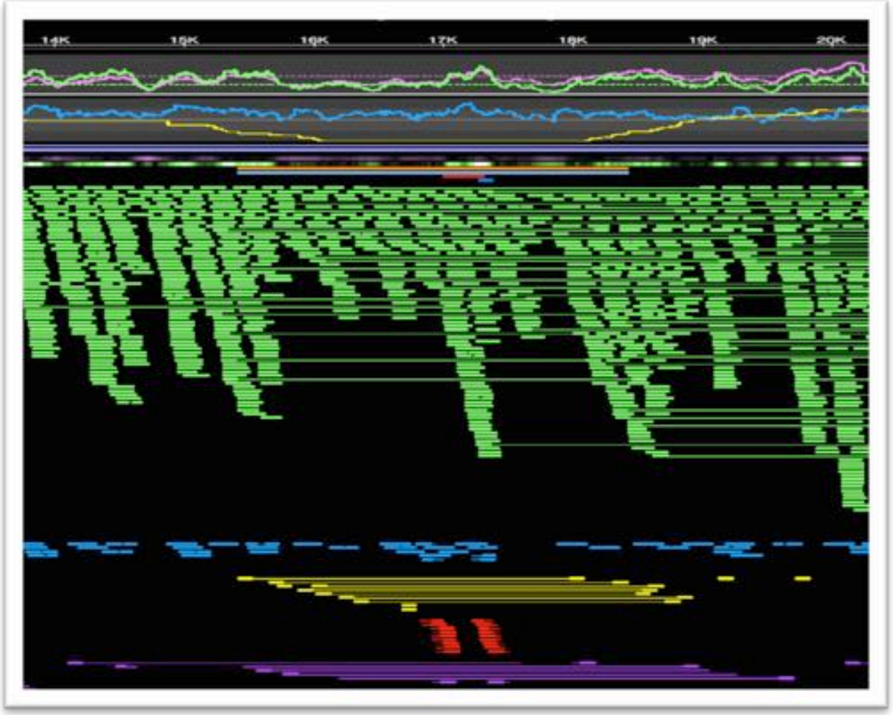
Forensics

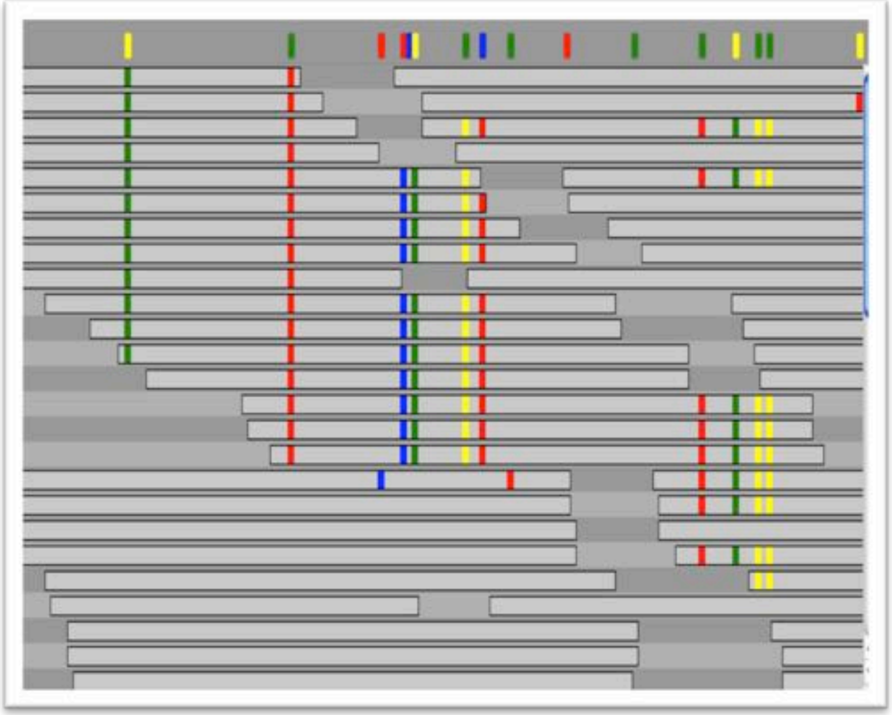
# Real Mis-assembly



Truth:

Mis-assembled:





**Hawkeye & AMOS: Visualizing and assessing the quality of genome assemblies**
Schatz, M.C. *et al.* (2011) *Briefings in Bioinformatics.* In Press.

# Assemblathon 2

- Real sequence data, *de novo* assembly

- Step 1: Apply best practices from Assemblathon 1
- Step 2: Add secret weapon for winning…

Images from Assemblathon

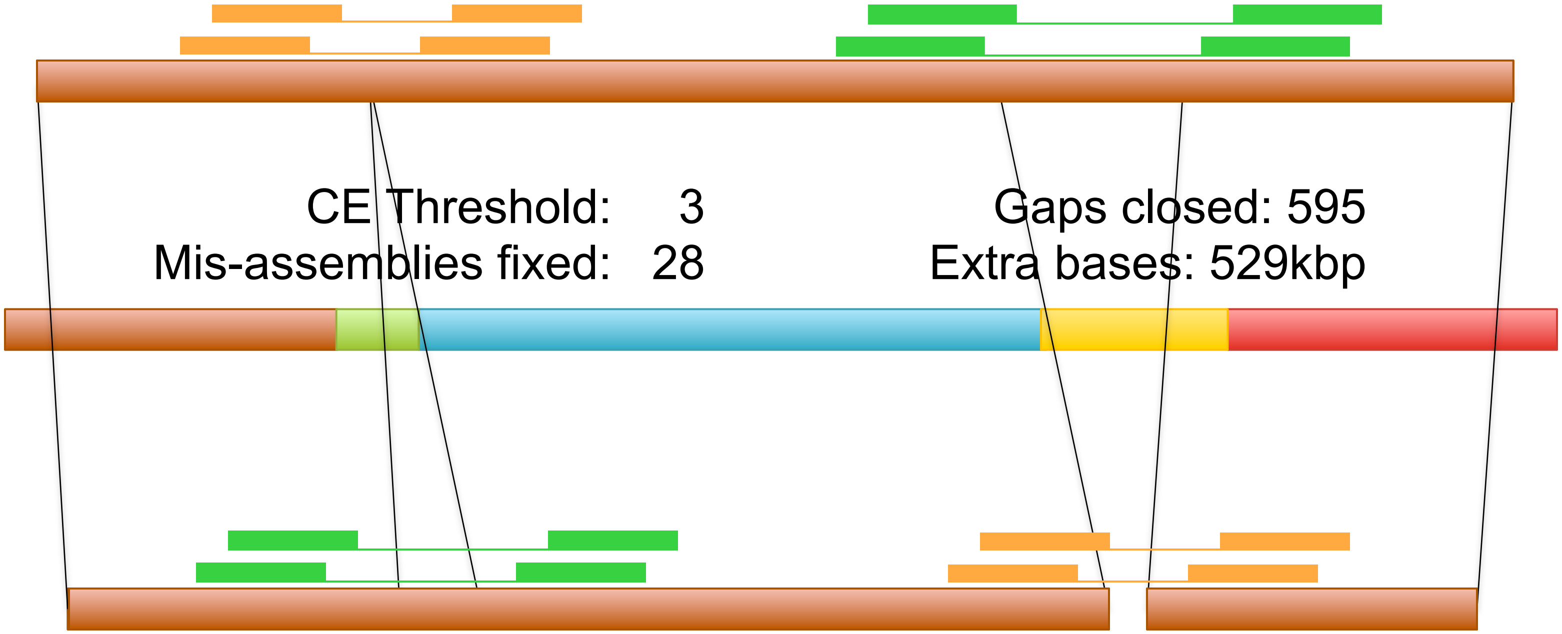
# Assemblathon 2: Metassembly

Paul Baranay, Scott Emrich, Michael Schatz

ALLPATHS-LG

Scaffold N50: 3,710,017
#>1000: 2,791

Contig N50: 20,183
#>1000: 68,591



CE Threshold: 3
Mis-assemblies fixed: 28

Gaps closed: 595
Extra bases: 529kbp

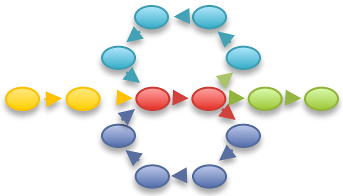
SOAPdenovo
+ FLASH
+ Quake
+ AMOS

Scaffold N50: 285,413
#>1000: 29,119

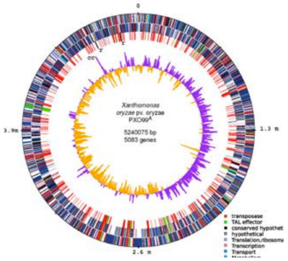
Contig N50: 1,607
#>1000: 218,643

# Summary

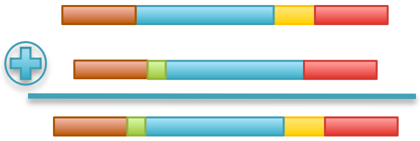
Assembly is a powerful tool for analyzing sequences, and is moving to increasingly more complex genomes and data types.



- Microassembly is a powerful tool needed to fully understand the genetics of autism and other diseases.

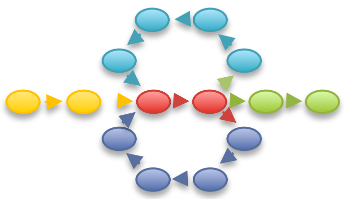


- A global analysis of the genome requires new statistics and computational methods to understand the patterns that we observe.



- Metassembly lets us maximize connectivity without sacrificing the quality of a de novo assembly.

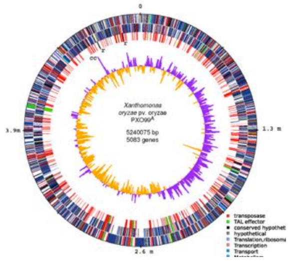
# Acknowledgements



Mitch Bekritsky
Giuseppe Narzisi

Ivan Iossifov
Wigler Lab

SFARI
SIMONS FOUNDATION
AUTISM RESEARCH INITIATIVE

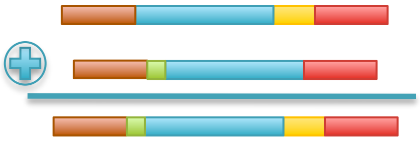


Hayan Lee
Matt Titmus
James Gurtowski

Ware Lab
McCombie Lab

Adam Phillippy (NBACC)
Sergey Koren (NBACC)

KBASE
predictive biology
DOE Systems Biology Knowledgebase



Paul Baranay (CSHL/ND)

Scott Emrich (ND)
Steven Salzberg (JHU)
Mihai Pop (UMD)

National Human
Genome Research
Institute

# Thank You

http://schatzlab.cshl.edu
@mike_schatz