

# De novo assembly of complex genomes using 3rd generation sequencing

Michael Schatz

Jan 15, 2012

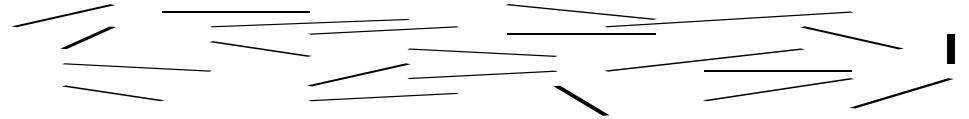
PAG-XX: Sequencing Complex Genomes



@mike\_schatz / #PAGXX

# Assembling a Genome

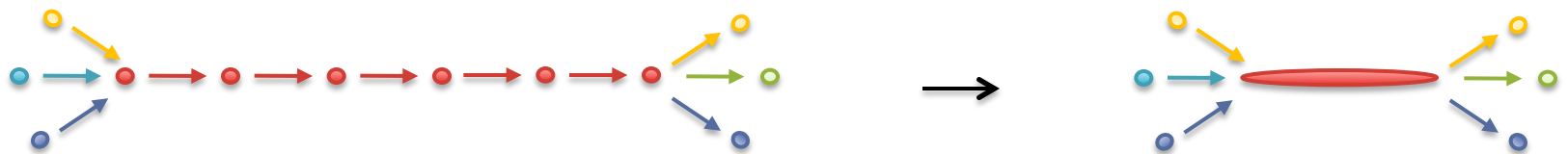
1. Shear & Sequence DNA



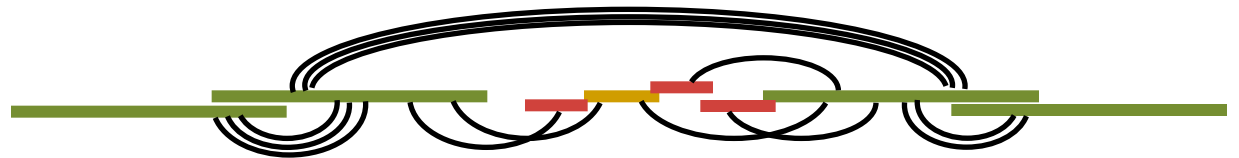
2. Construct assembly graph from overlapping reads

...AGCCTAGACCTACAGGATGCGCGACACGT  
GGATGCGCGACACGTTCGCATATCCGGT...

3. Simplify assembly graph



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



# Why are PAG genomes hard to assemble?



## **1. Instrumentation:**

- (Very) large genomes, imperfect sequencing

## **2. Biological:**

- (Very) High ploidy, heterozygosity, repeat content

## **3. Computational:**

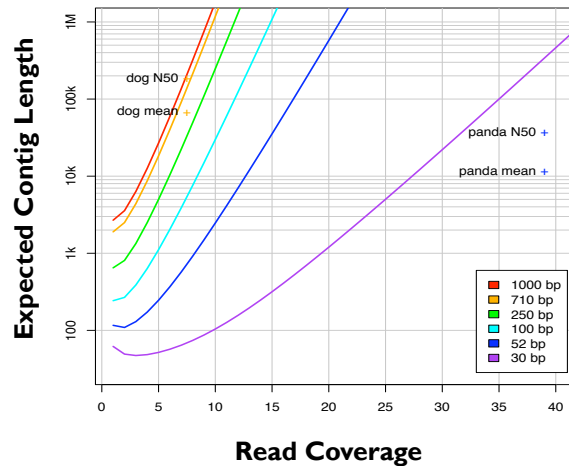
- (Very) Large genomes, complex structure

## **4. Accuracy:**

- (Very) Hard to assess correctness

# Ingredients for a good assembly

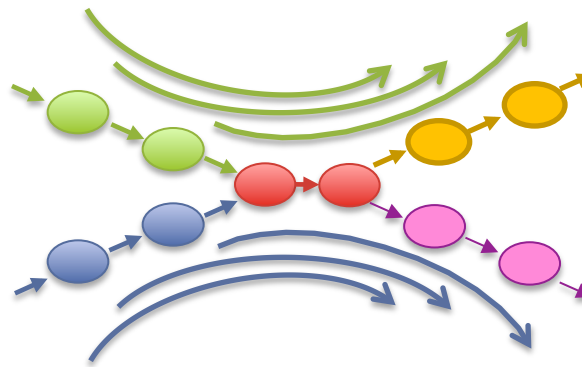
## Coverage



### High coverage is required

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

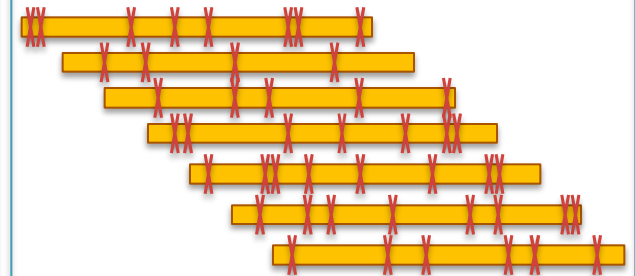
## Read Length



### 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

## Assembly of Large Genomes using Second Generation Sequencing

Schatz MC, Delcher AL, Salzberg SL (2010) *Genome Research*. 20:1165-1173.

# Hybrid Sequencing



## **Illumina**

*Sequencing by Synthesis*

High throughput (60Gbp/day)

High accuracy (~99%)

Short reads (~100bp)



## **Pacific Biosciences**

*SMRT Sequencing*

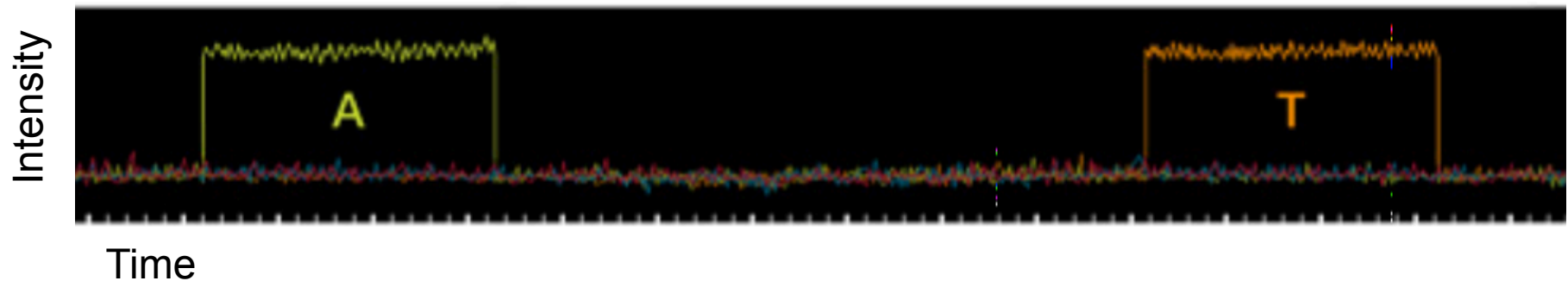
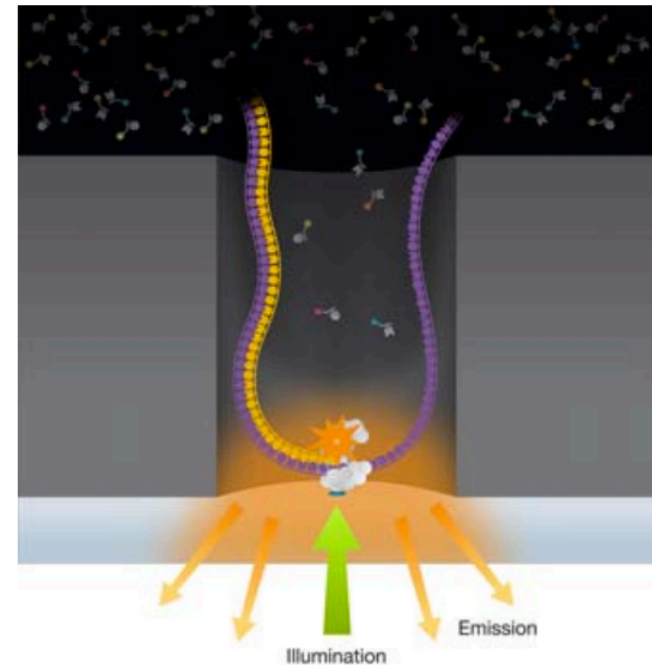
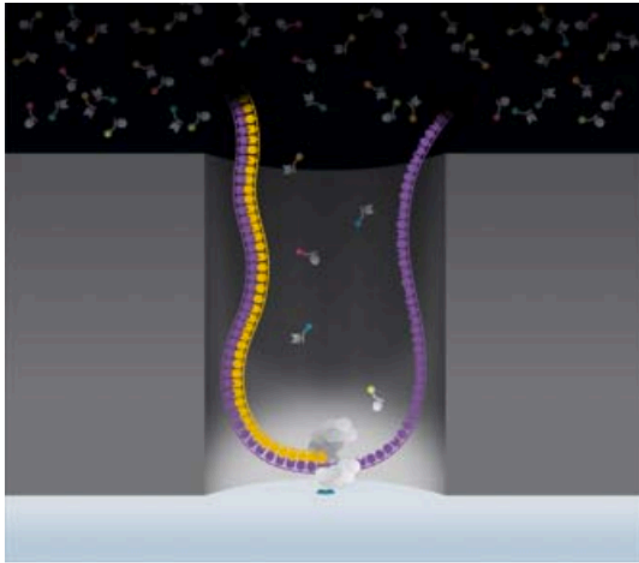
Lower throughput (600Mbp/day)

Lower accuracy (~85%)

Long reads (1-2kbp+)

# SMRT Sequencing

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





# PacBio Error Correction

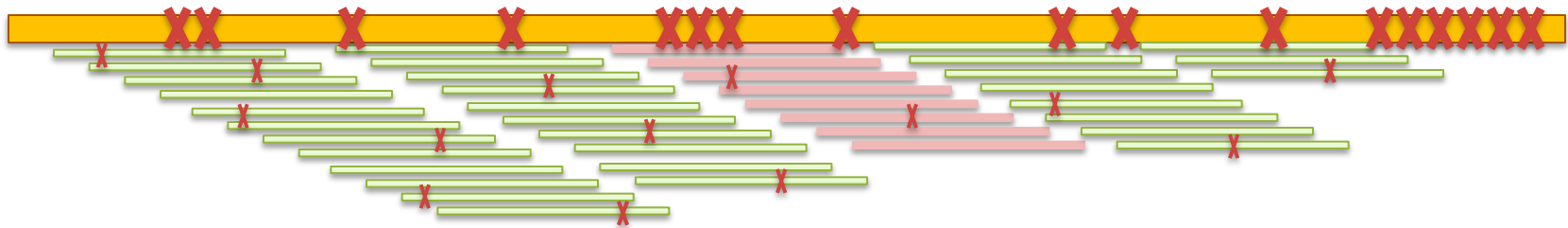
<http://wgs-assembler.sf.net>



## I. Correction Pipeline

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

## 2. Error corrected reads can be easily assembled, aligned

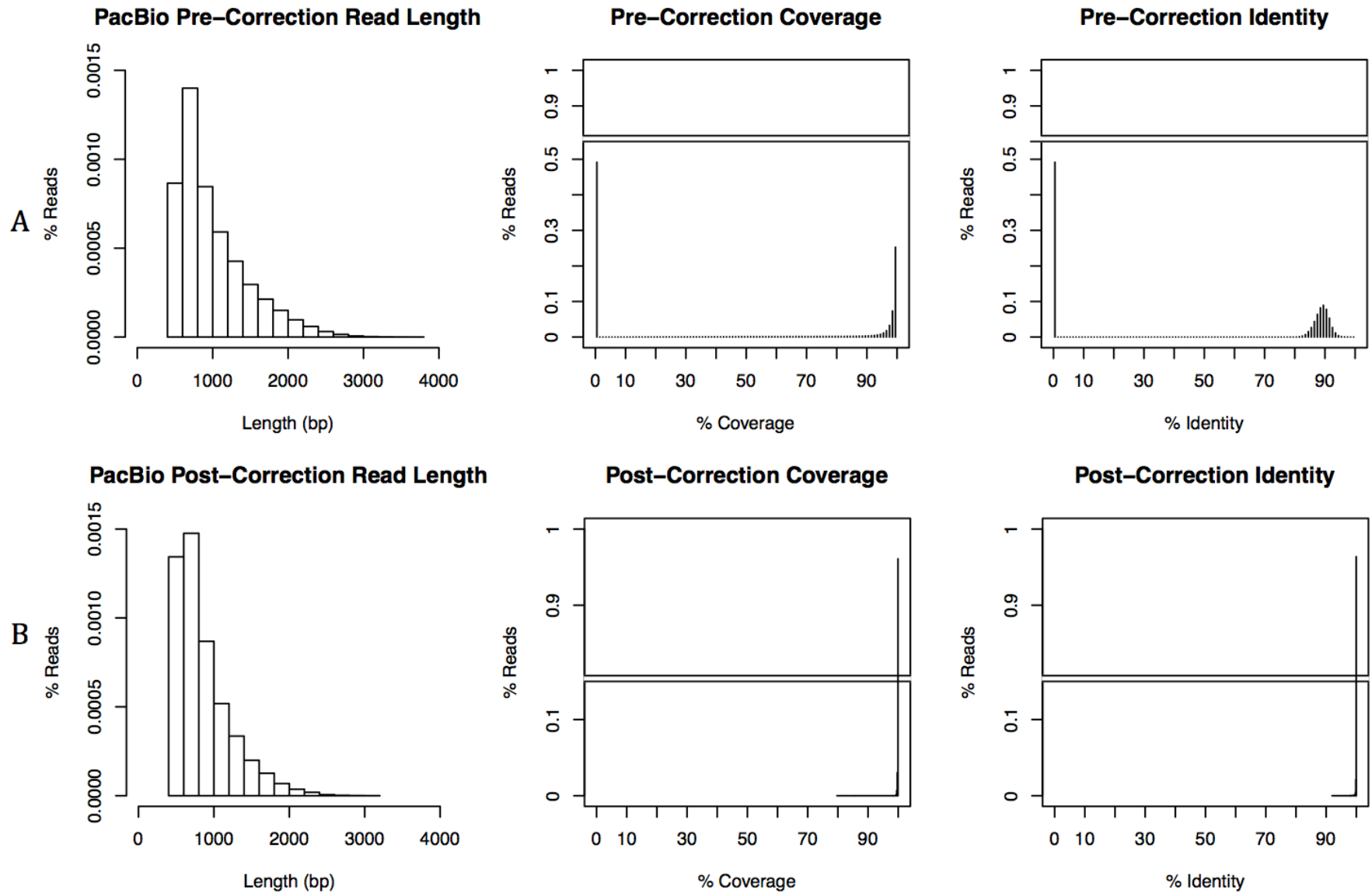


**Hybrid error correction and de novo assembly of single-molecule sequencing reads.**

Koren, S, Schatz, MC, Walenz, BP, Martin, J, Howard, J, Ganapathy, G, Wang, Z, Rasko, DA, McCombie, WR, Jarvis, ED, Phillippy, AM. (2012) *Under Review*



# Error Correction Results



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

# SMRT-assembly

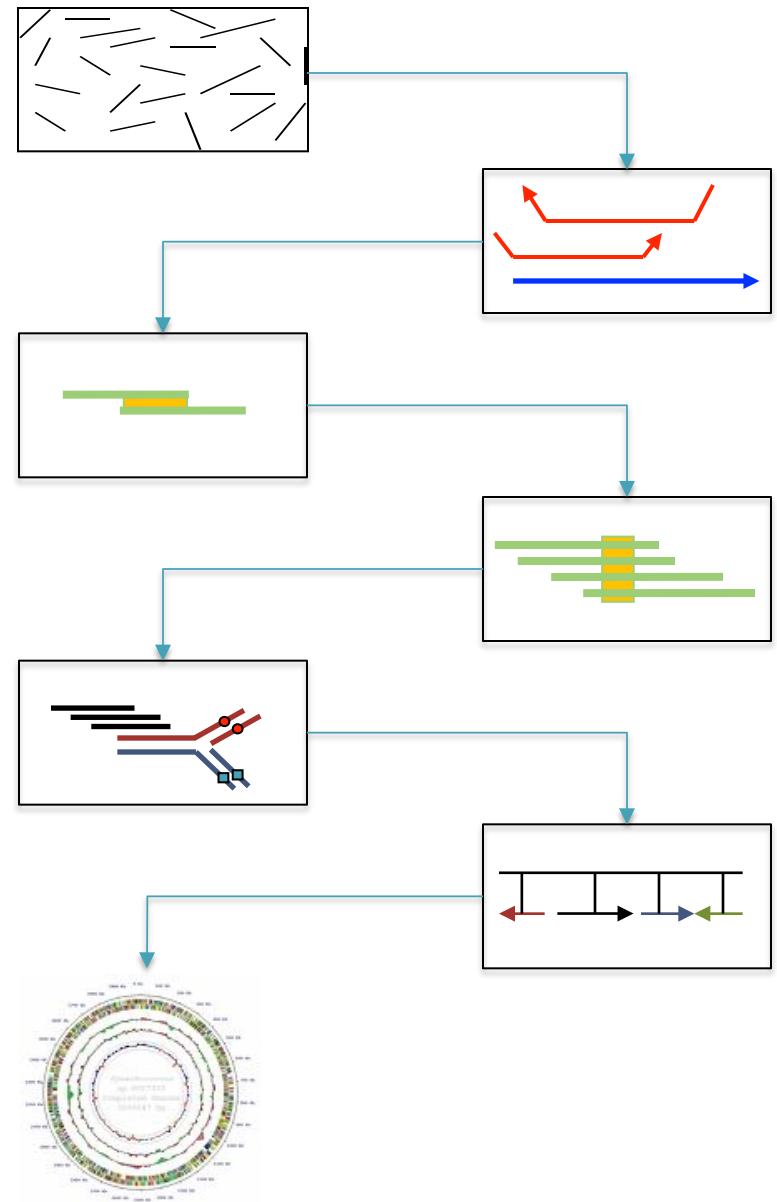


- Co-assemble error corrected long reads with short reads
  - Long reads natively span repeats (red)
  - Guards against mis-assemblies in draft assembly
  - Use all available data at once
- Challenges
  - Assembler must supports a wide mix of read lengths

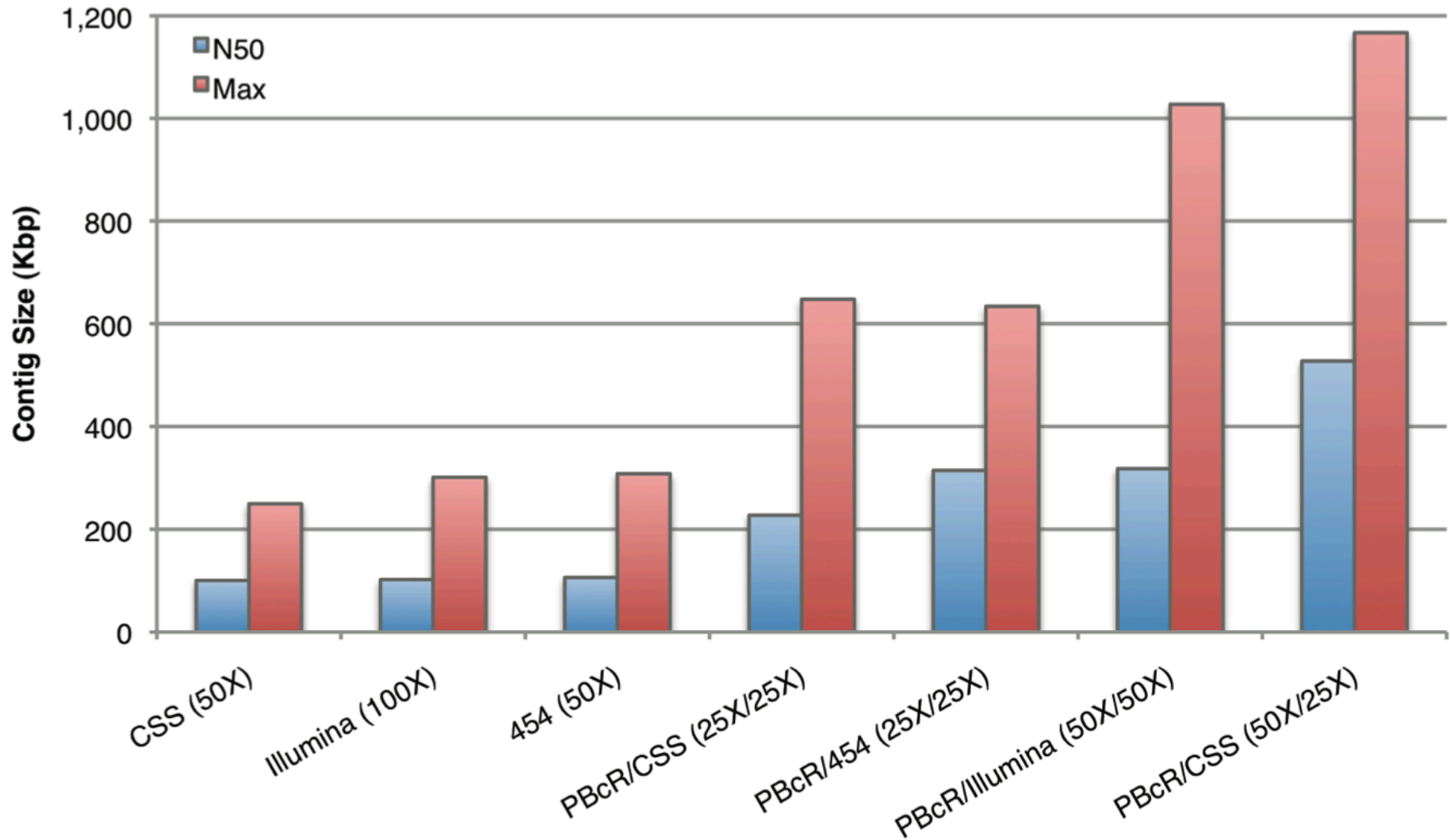
# Celera Assembler

<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

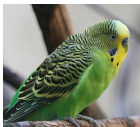
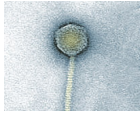


# Assembly Results



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

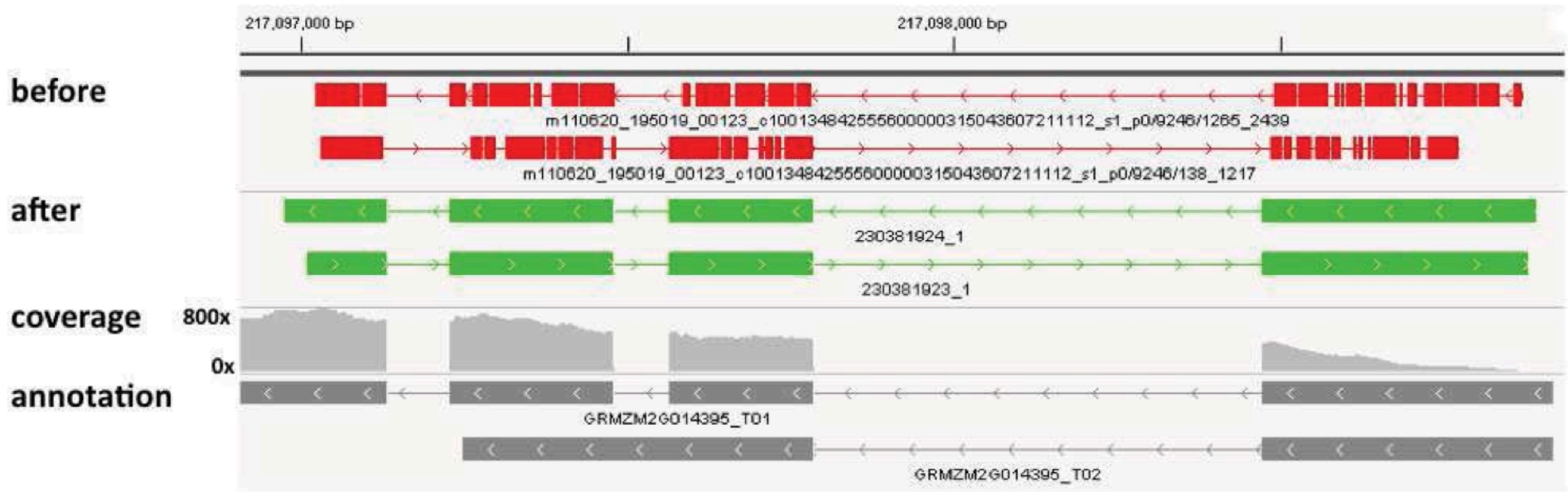
# SMRT-Assembly Results



Organism	Technology	Reference bp	Assembly bp	# Contigs	Max Contig Length	N50
<i>Lambda</i> 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%) *
<i>E. coli</i> 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%) *
<i>E. coli</i> 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 <sup>9</sup>		5 452 251	23	653 382	402 041
<i>S. cerevisiae</i> 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%) *
<i>Melopsittacus undulatus</i>	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

# 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

# Why are PAG genomes hard to assemble?



## **1. Instrumentation:**

- (Very) large genomes, imperfect sequencing

## **2. Biological:**

- (Very) High ploidy, heterozygosity, repeats

## **3. Computational:**

- (Very) Large genomes, complex structure

## **4. Accuracy:**

- (Very) Hard to assess correctness

With new sequencing technologies and improved algorithms we can address these challenges

=> Cautiously optimistic



# Acknowledgements

## Schatzlab

Giuseppe Narzisi

Mitch Bekritsky

Matt Titmus

Hayan Lee

James Gurtowski

Rohith Menon

Goutham Bhat

## CSHL

Dick McCombie

Melissa Kramer

Eric Antonio

Mike Wigler

Zach Lippman

Doreen Ware

Ivan Iossifov

## NBACC

Adam Phillipy

Sergey Koren

## JHU

Steven Salzberg

Ben Langmead

Jeff Leek

## Univ. of Maryland

Mihai Pop

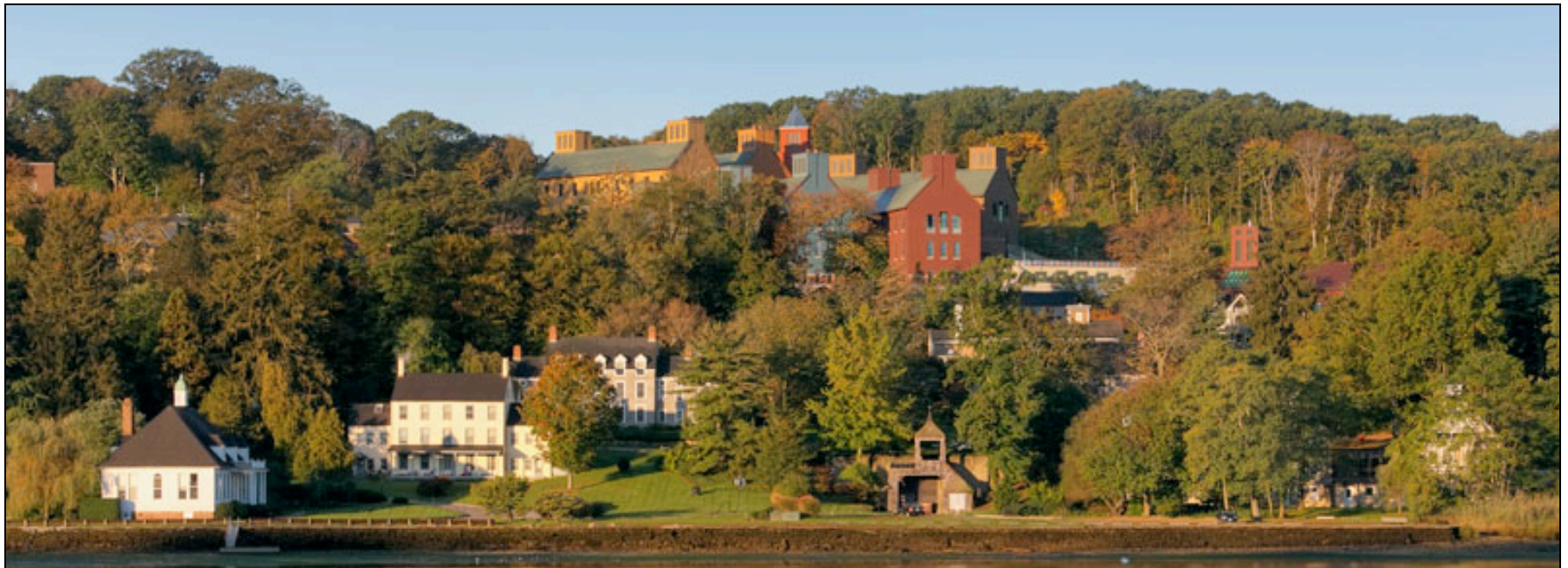
Art Delcher

Jimmy Lin

David Kelley

Dan Sommer

Cole Trapnell





# Thank You

<http://schatzlab.cshl.edu>  
[@mike\\_schatz](#) / [#PAGXX](#)

More Discussion @ PacBio  
Workshop 1:30 Tuesday