

# De novo assembly of complex genomes

Michael Schatz

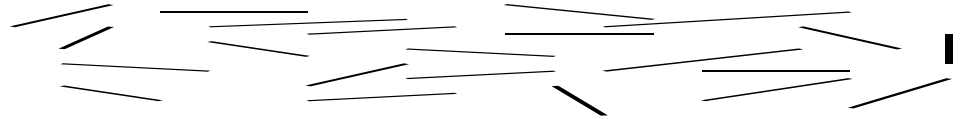
Oct 3, 2013  
Beyond the Genome



@mike\_schatz

# Assembling a Genome

## 1. Shear & Sequence DNA



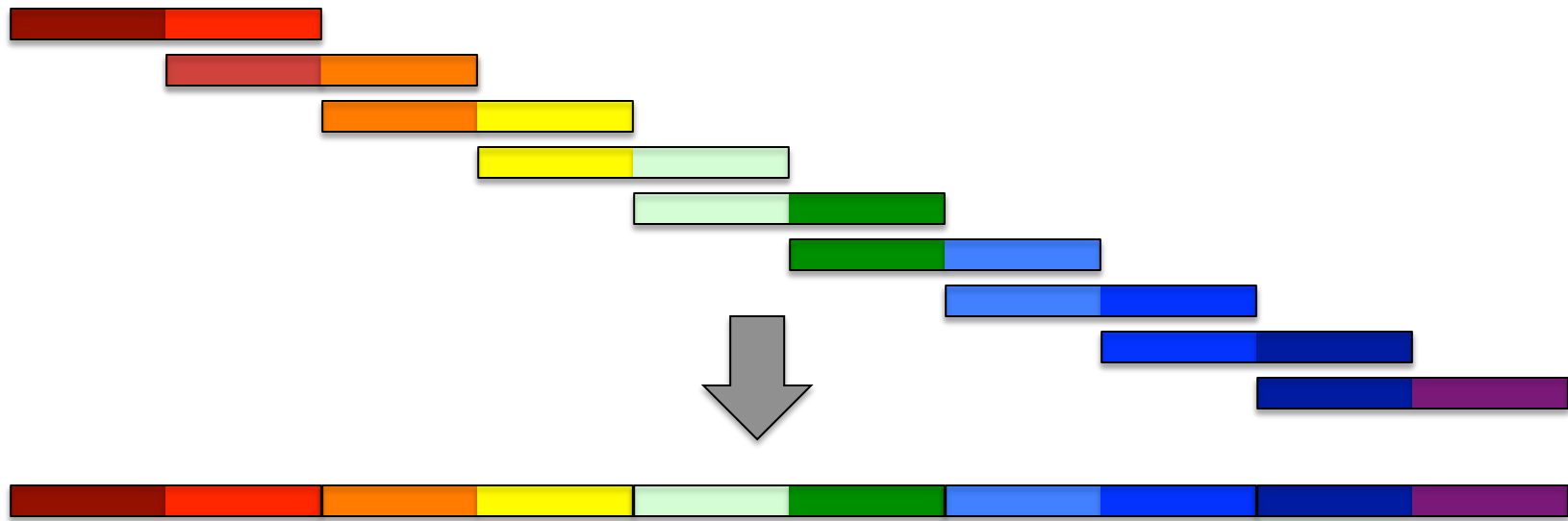
## 2. Construct assembly graph from overlapping reads

...AGCCTAGGGATGCGCGACACGT

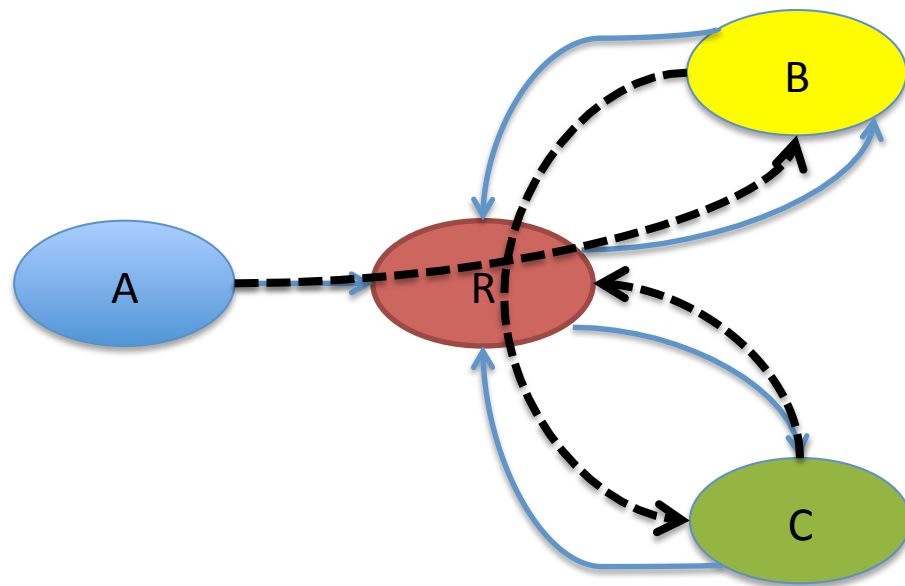
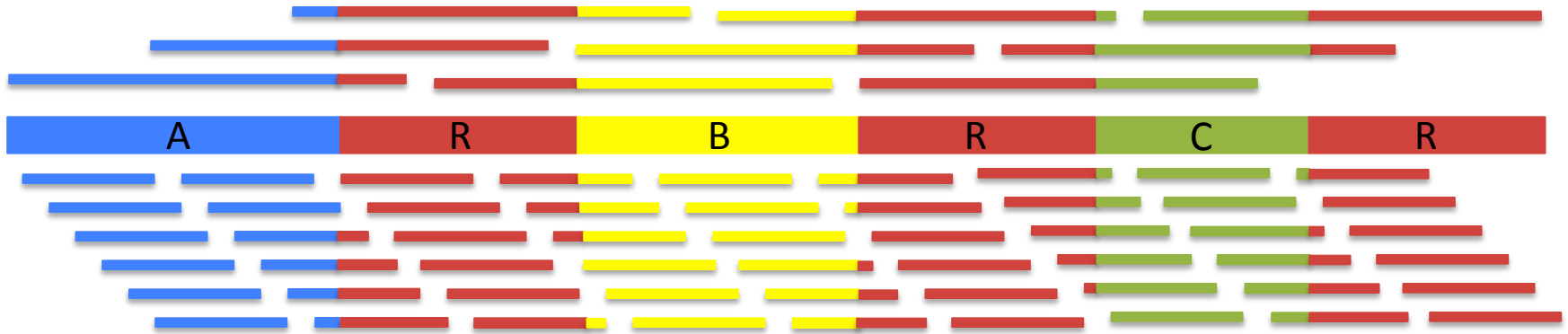
GGATGCGCGACACGTGCATATCCGGTTTGGTCAACCTCGGACGGAC

CAACCTCGGACGGACCTCAGCGAA...

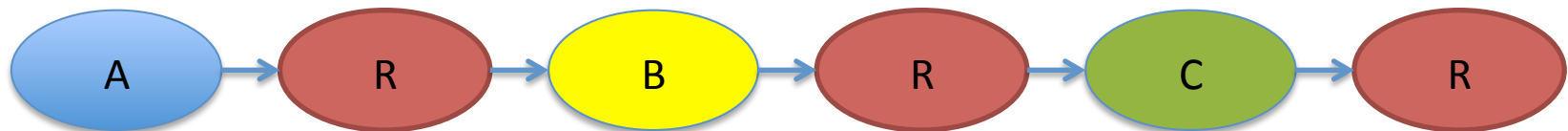
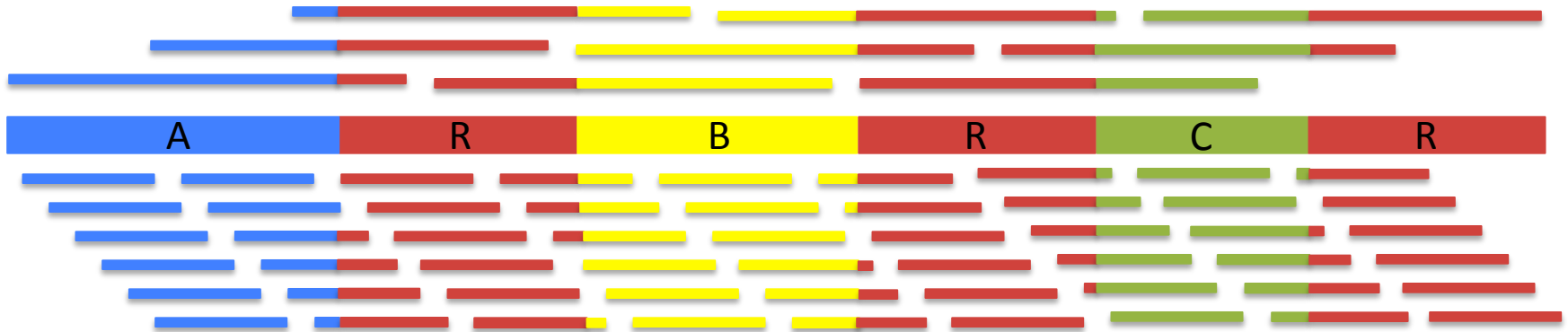
## 3. Simplify assembly graph



# Assembly Complexity

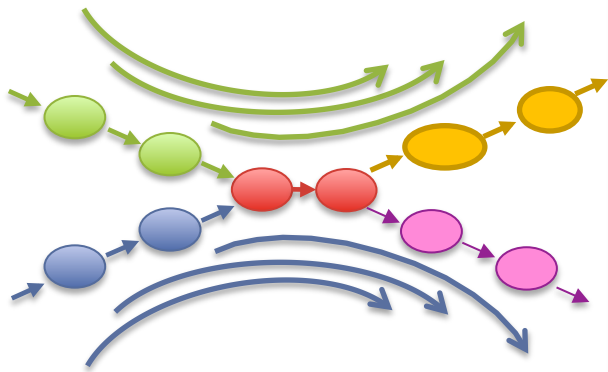


# Assembly Complexity



# Ingredients for a good assembly

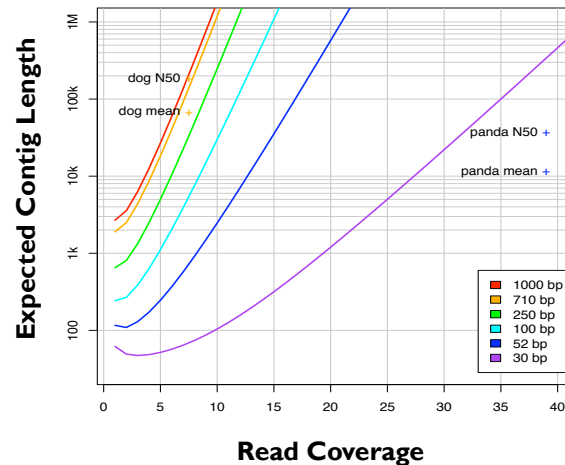
## Read Length



### **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

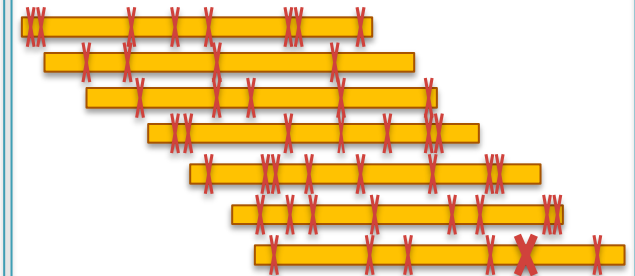
## Coverage



### **High coverage is required**

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

## 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, WVR (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*

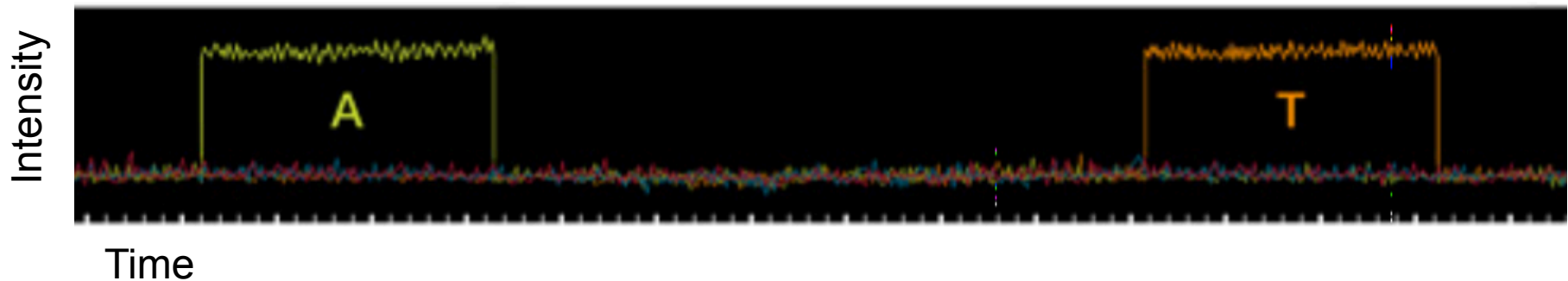
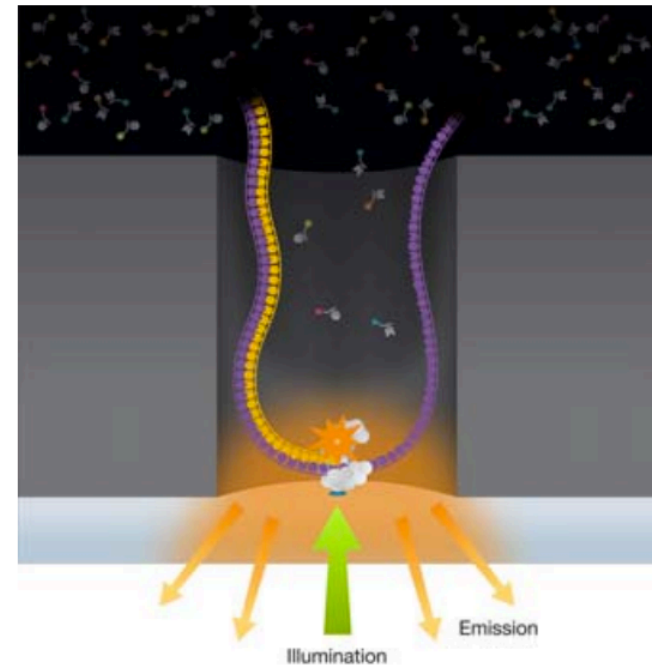
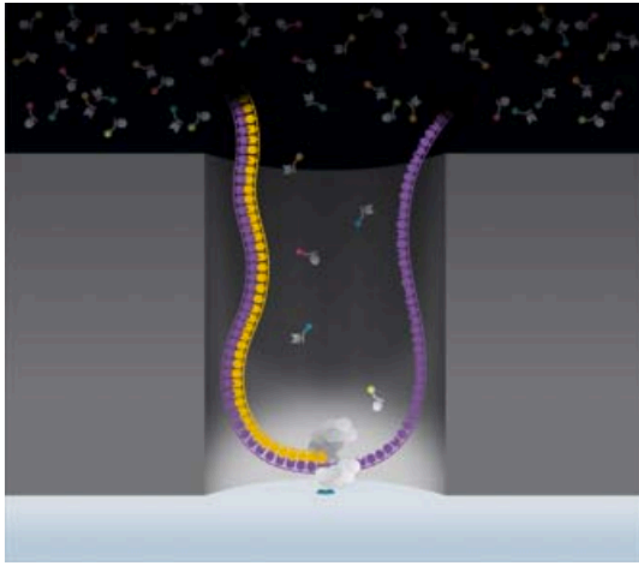
Lower throughput (1Gbp/day)

Lower accuracy (~85%)

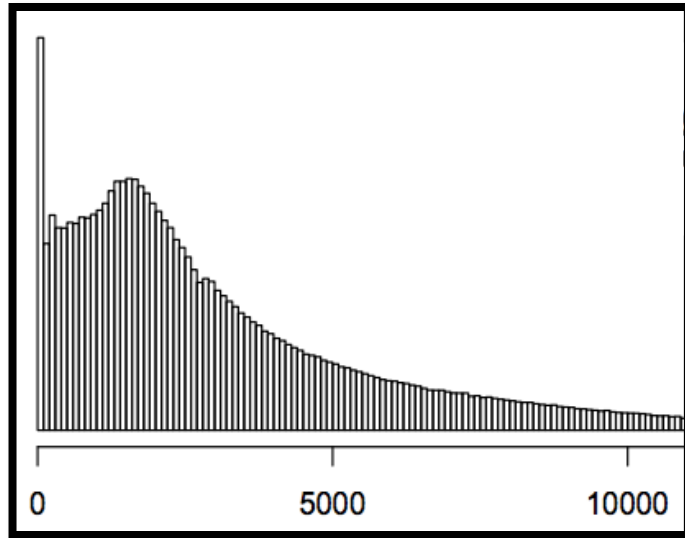
Long reads (5kbp+)

# SMRT Sequencing

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



# SMRT Sequencing Data



Match	83.7%
Insertions	11.5%
Deletions	3.4%
Mismatch	1.4%

TTGTAAGCAGTTGAAAACATATGTGTGGATTTAGAATAAAGAACATGAAAG  
 |||  
 TTGTAAGCAGTTGAAAACATATGTGT-GATTTAG-ATAAAGAACATGGAAG

ATTATAAA-CAGTTGATCCATT-AGAAGA-AAACGCAAAGGC GGCTAGG  
 |  
 A-TATAAATCAGTTGATCCATTAGAA-AGAAACGC-AAAGGC-GCTAGG

CAACCTTGAATGTAATCGCACTTGAAGAACAAGATTTTATTCCGCGCCCG  
 |  
 C-ACCTTG-ATGT-AT--CACTTGAAGAACAAGATTTTATTCCGCGCCCG

TACGAATCAAGATTCTGAAAACACAT-ATAACAACCTCCAAAA-CACAA  
 |  
 T-ACGAATC-AGATTCTGAAAACA-ATGAT----ACCTCCAAAAGCACAA

-AGGAGGGGAAA GGGGGGAATATCT-ATAAAGATTACAAATTAGA-TGA  
 |||  
 GAGGAGG---AA-----GAATATCTGAT-AAAGATTACAAATT-GAGTGA

ACT-AATTCACAATA-AATAACACTTTTA-ACAGAATTGAT-GGAA-GTT  
 |||  
 ACTAAATTCACAA-ATAATAACACTTTTAGACAA AATTGATGGGAAGGTT

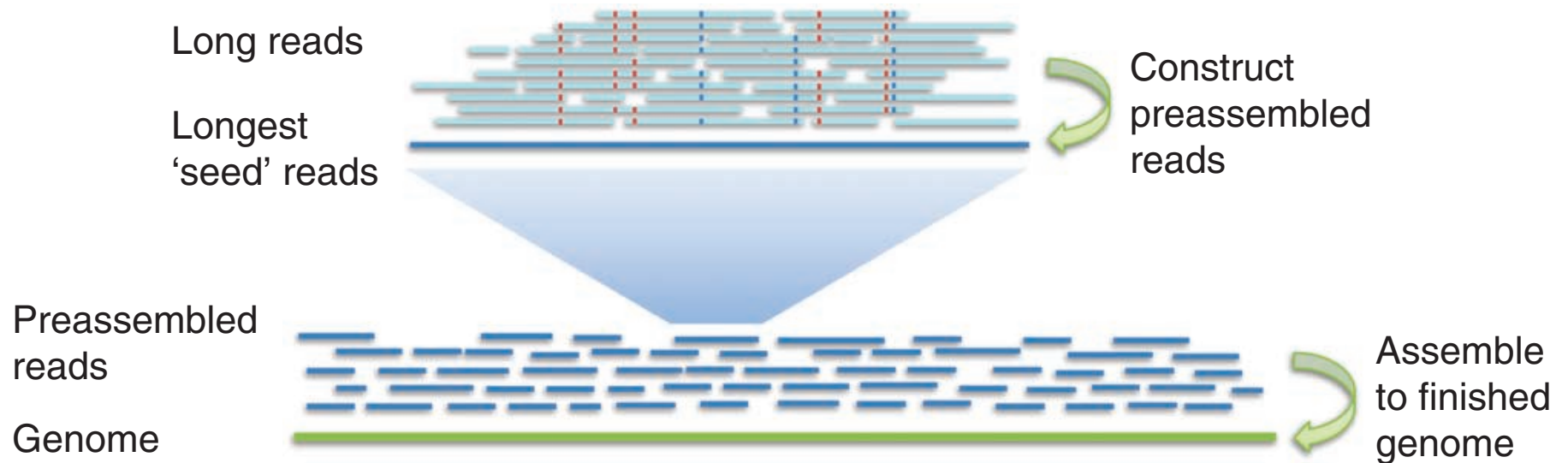
TCGGAGAGATCCAAAACAATGGGC-ATCGCCTTTGA-GTTAC-AATCAAA  
 |||  
 TC-GAGAGATCC-AAACAAT-GGCGATCG-CCTTGCAGTTACAAATCAAA

ATCCAGTGGAAAATATAATTTATGCAATCCAGGAACTTATTCACAATTAG  
 |||  
 ATCCAGT-GAAAATATA--TTATGC-ATCCA-GAACTTATTCACAATTAG

Sample of 100k reads aligned with BLASR requiring >100bp alignment



# PacBio Error Correction: HGAP



- With 50-100x of Pacbio coverage, virtually all of the errors can be eliminated
  - Works well for Microbial genomes: single contig per chromosome routinely achieved
  - Difficult to scale up for use with eukaryotic genomes

**Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data**  
Chin, CS *et al.* (2013) *Nature Methods*. 10: 563-569

# Hybrid Error Correction: PacBioToCA

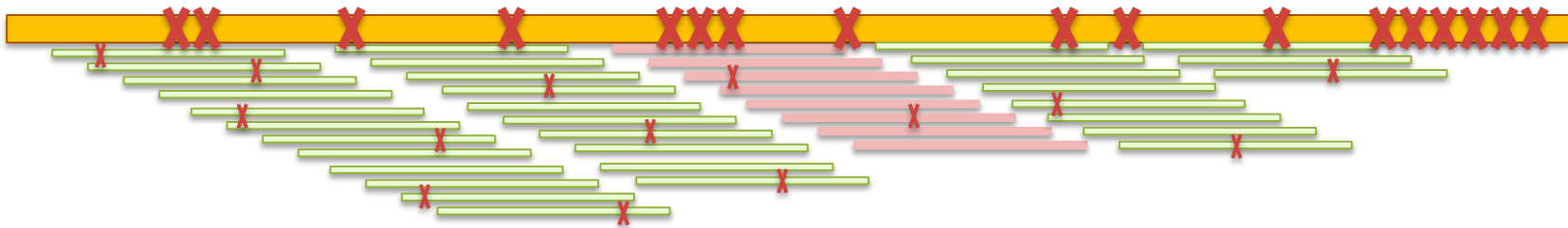
<http://wgs-assembler.sf.net>

## I. 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

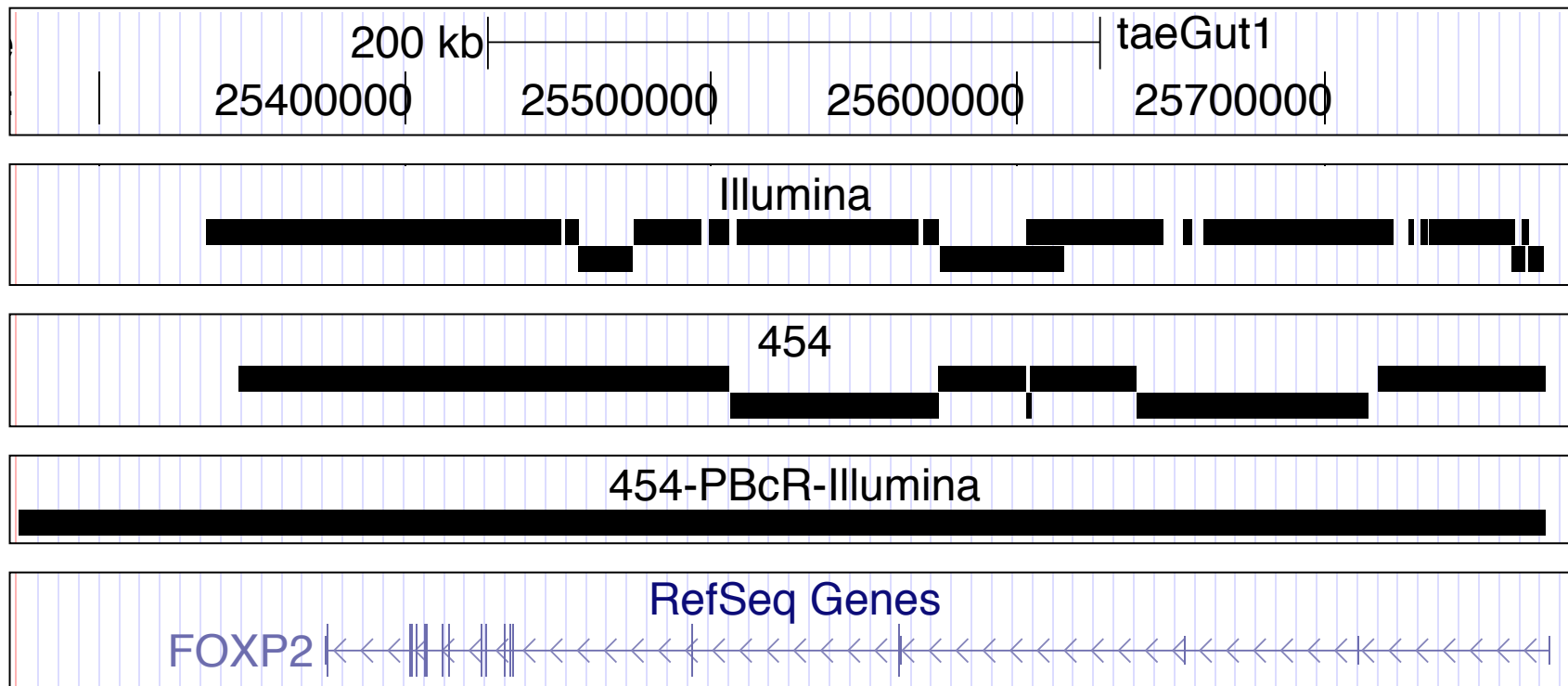


**Hybrid error correction and de novo assembly of single-molecule sequencing reads.**

Koren, S, Schatz, MC, et al. (2012) *Nature Biotechnology*. doi:10.1038/nbt.2280

# Improved Gene Reconstruction

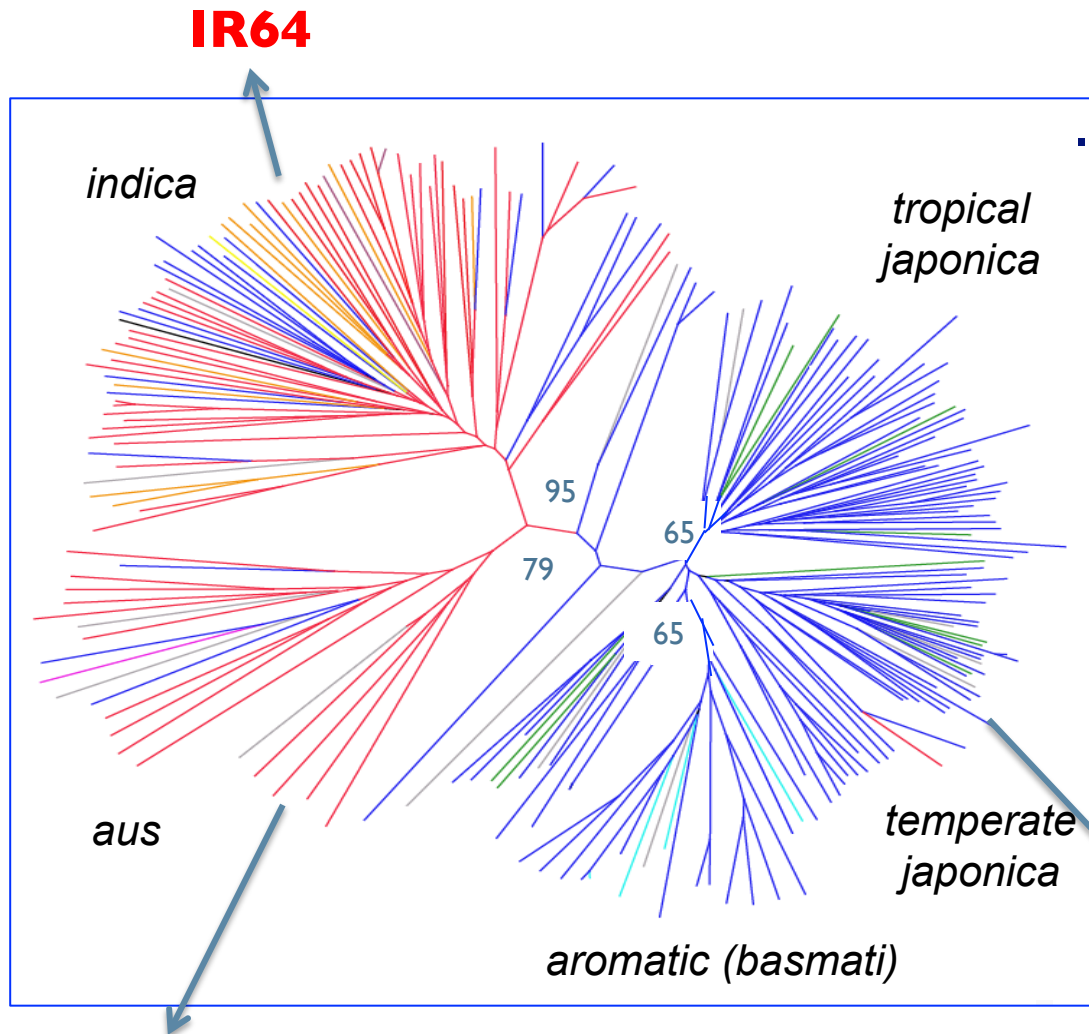
FOXP2 assembled in a single contig in the PacBio parrot assembly



**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

# Population structure in *Oryza sativa*

3 varieties selected for *de novo* sequencing



High quality BAC-by-BAC reference

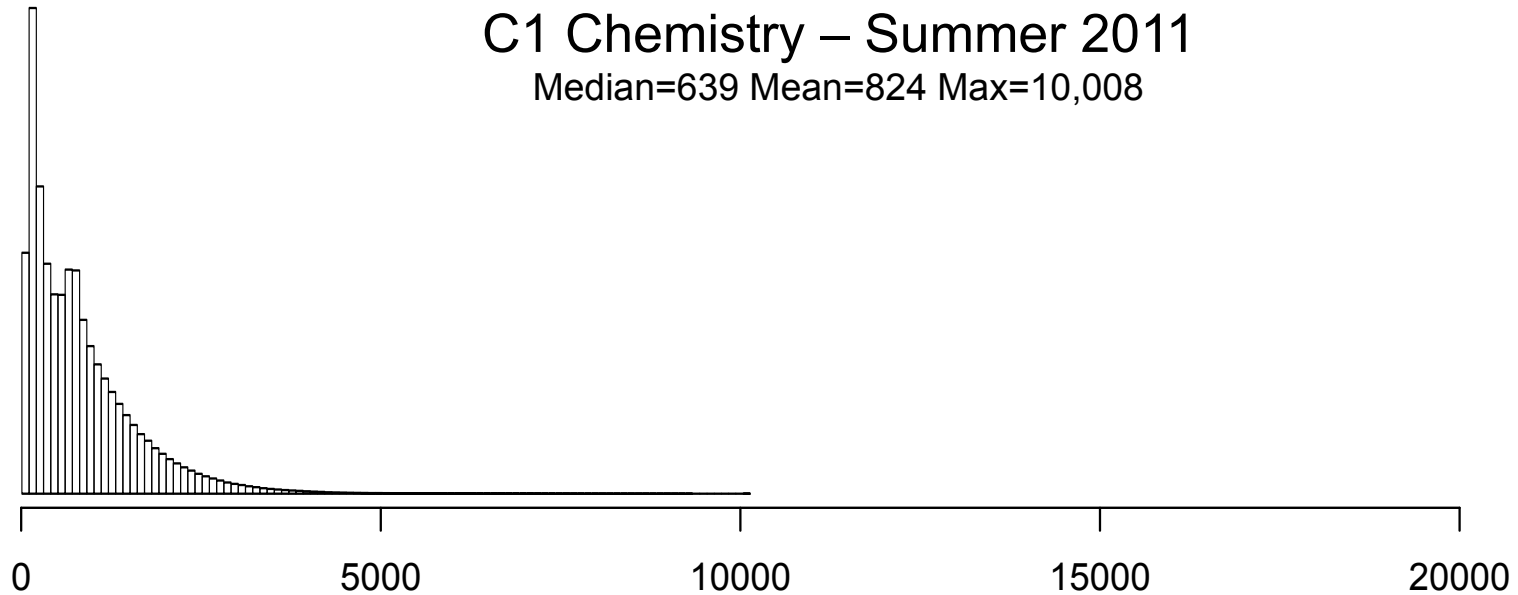
- ~370 Mbp genome in 12 chromosomes
- About 40% repeats:
  - Many 4-8kbp repeats
  - 300kbp max high identity repeat (99.99%)
- Useful model for other cereal genomes

**DJ123**

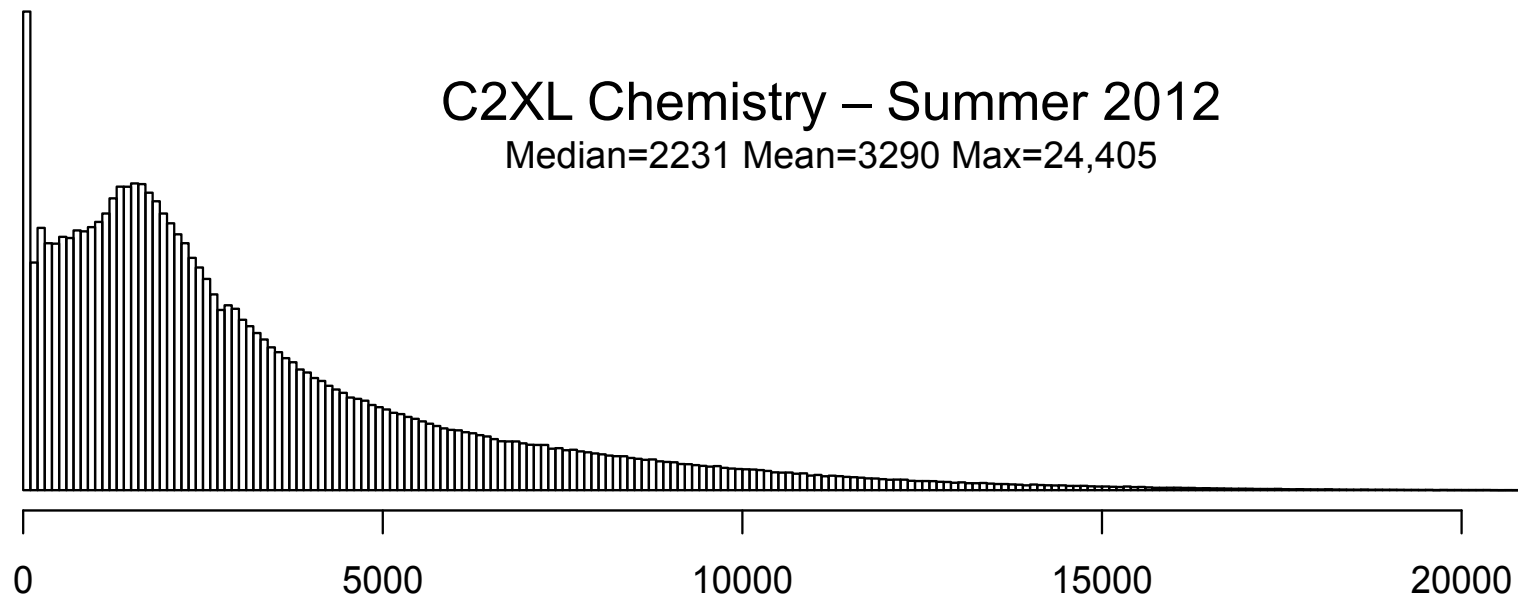
Garris et al. (2005)  
Genetics 169: 1631–1638

# PacBio Long Read Rice Sequencing

C1 Chemistry – Summer 2011  
Median=639 Mean=824 Max=10,008

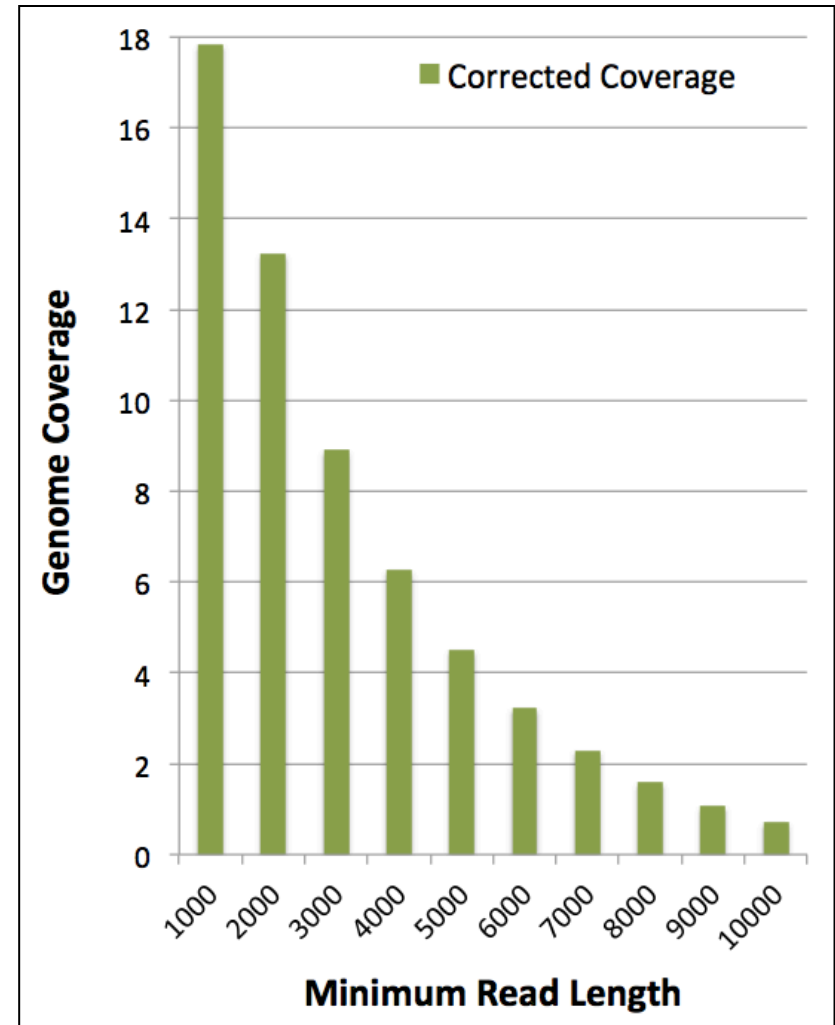


C2XL Chemistry – Summer 2012  
Median=2231 Mean=3290 Max=24,405



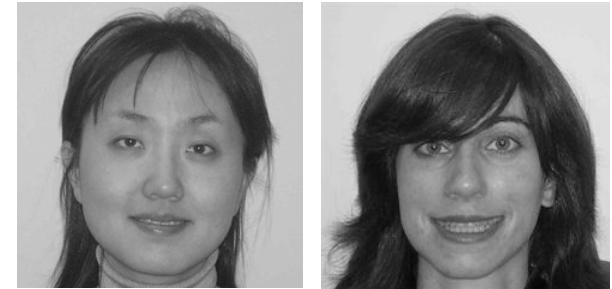
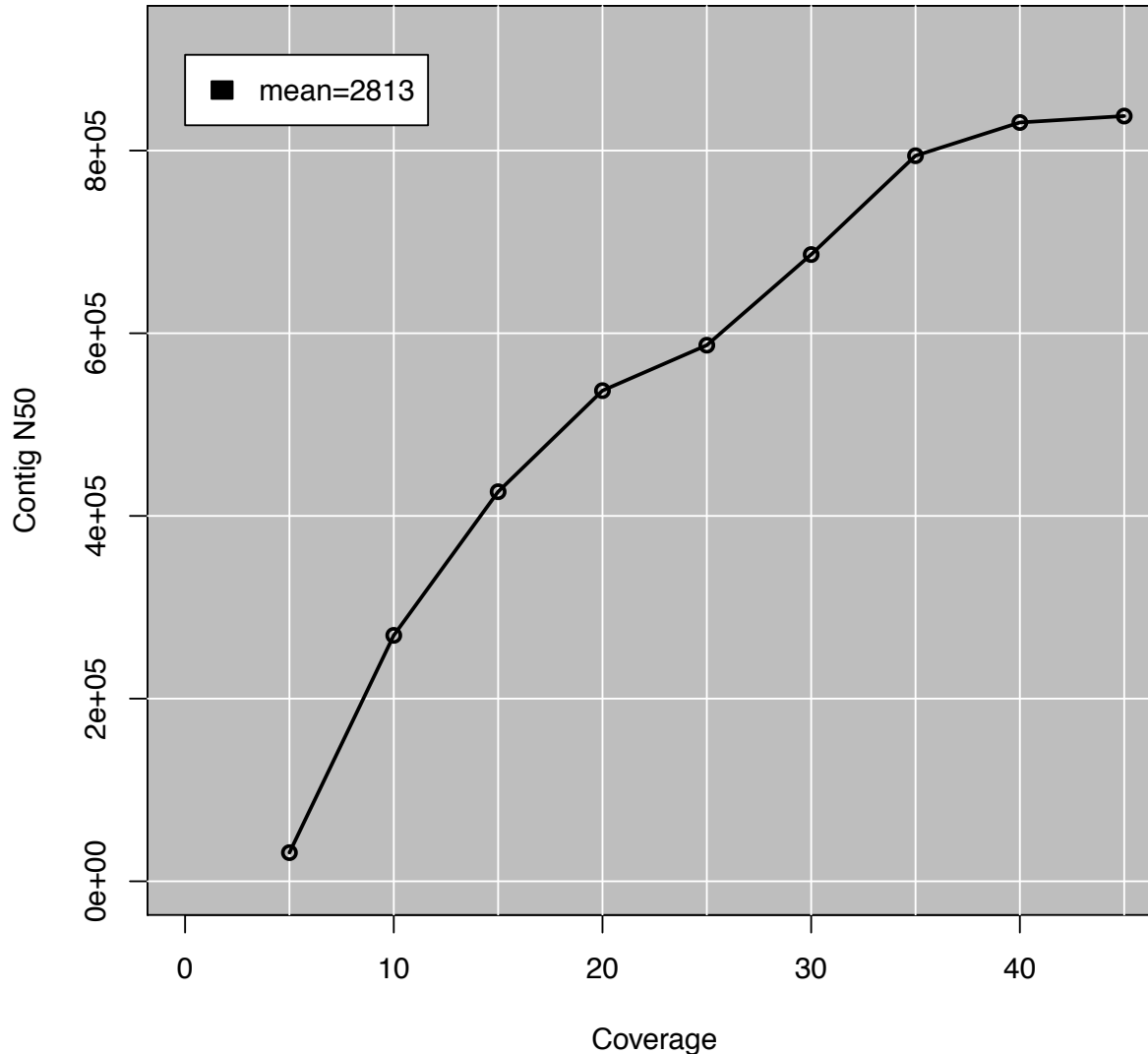
# Preliminary Rice Assemblies

Assembly	Contig NG50
HiSeq Fragments 50x 2x100bp @ 180	3,925
MiSeq Fragments 23x 459bp 8x 2x251bp @ 450	6,332
“ALLPATHS-recipe” 50x 2x100bp @ 180 36x 2x50bp @ 2100 51x 2x50bp @ 4800	18,248
PBeCR Reads 19x @ 3500 ** MiSeq for correction	50,995



In collaboration with McCombie & Ware labs @ CSHL

# Assembly Coverage Model



Simulate PacBio-like reads to predict how the assembly will improve as we add additional coverage

Only 8x coverage is needed to sequence every base in the genome, but 40x improves the chances repeats will be spanned by the longest reads

## Assembly complexity of long read sequencing

Lee, H\*, Gurtowski, J\*, Yoo, S, Marcus, S, McCombie, VWR, Schatz MC et al. (2013) *In preparation*

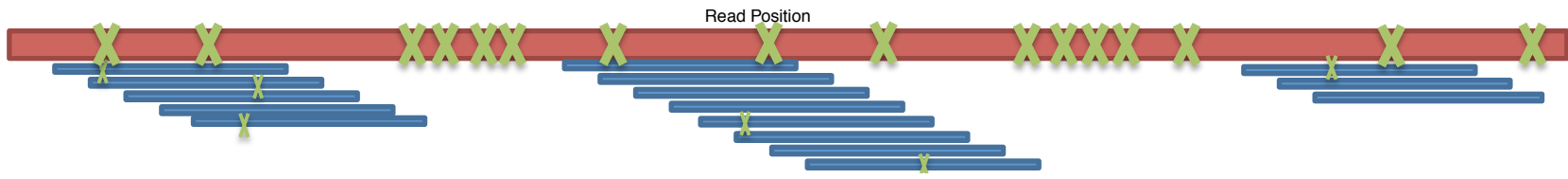
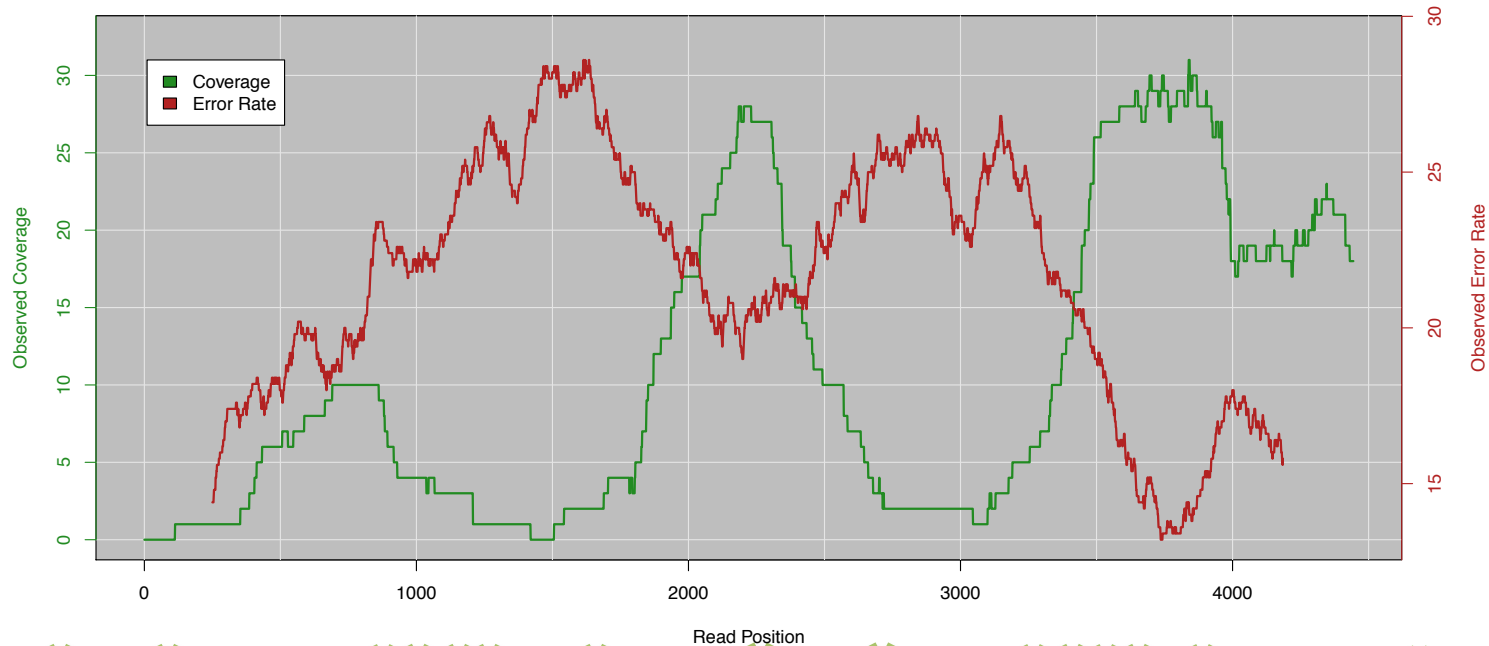
# Enhanced PacBio Error Correction

## PacBioToCA fails in complex regions

1. Simple Repeats – Kmer Frequency Too High to Seed Overlaps
2. Error Dense Regions – Difficult to compute overlaps with many errors
3. Extreme GC – Lacks Illumina Coverage



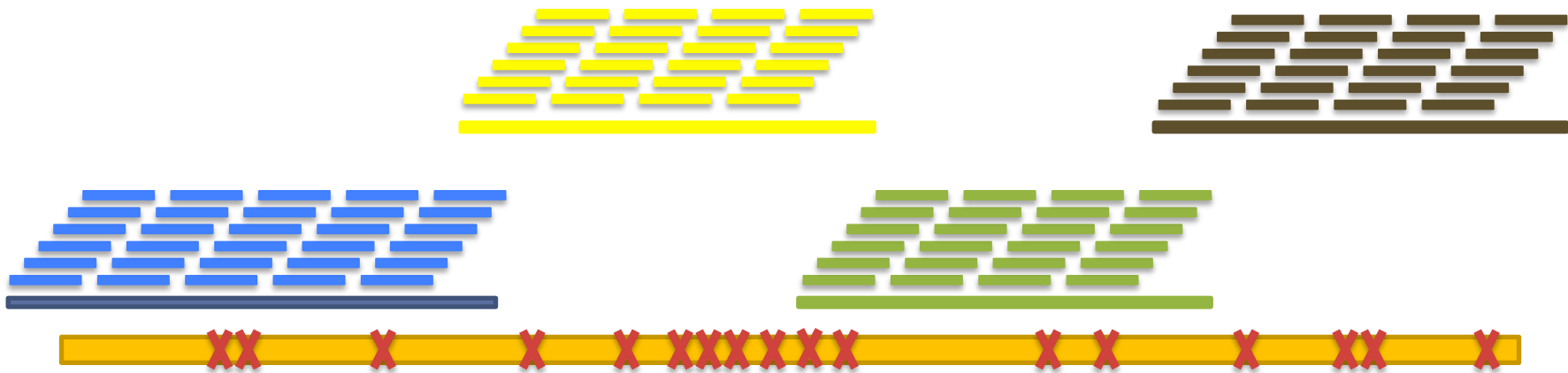
Position Specific Coverage and Error Rate





# Error Correction with pre-assembled Illumina reads

<https://github.com/jgurtowski/pbtools>



**Short Reads -> Assemble Unitigs -> Align & Select -> Error Correct**

## **Unitigs:**

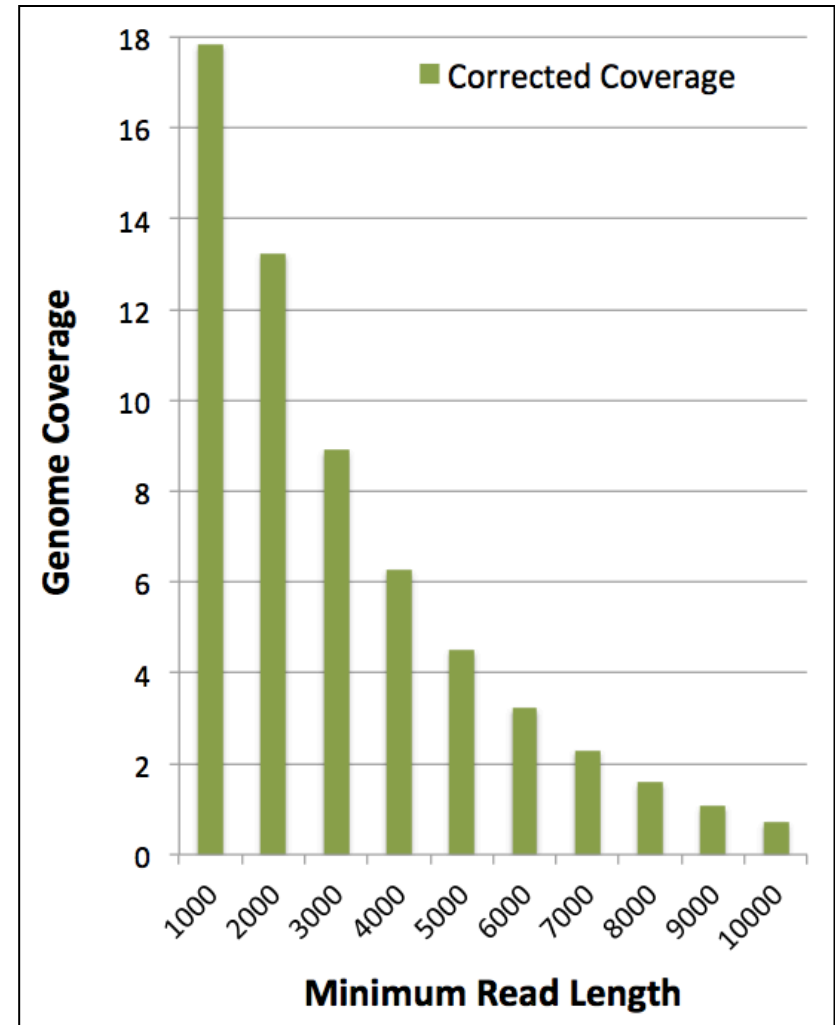
High quality contigs formed from unambiguous, unique overlaps of reads  
Each read is placed into a single unitig

Can Help us overcome:

- 1. Simple Repeats – Kmer Frequency Too High to Seed Overlaps**
- 2. Error Dense Regions – Difficult to compute overlaps with many errors**

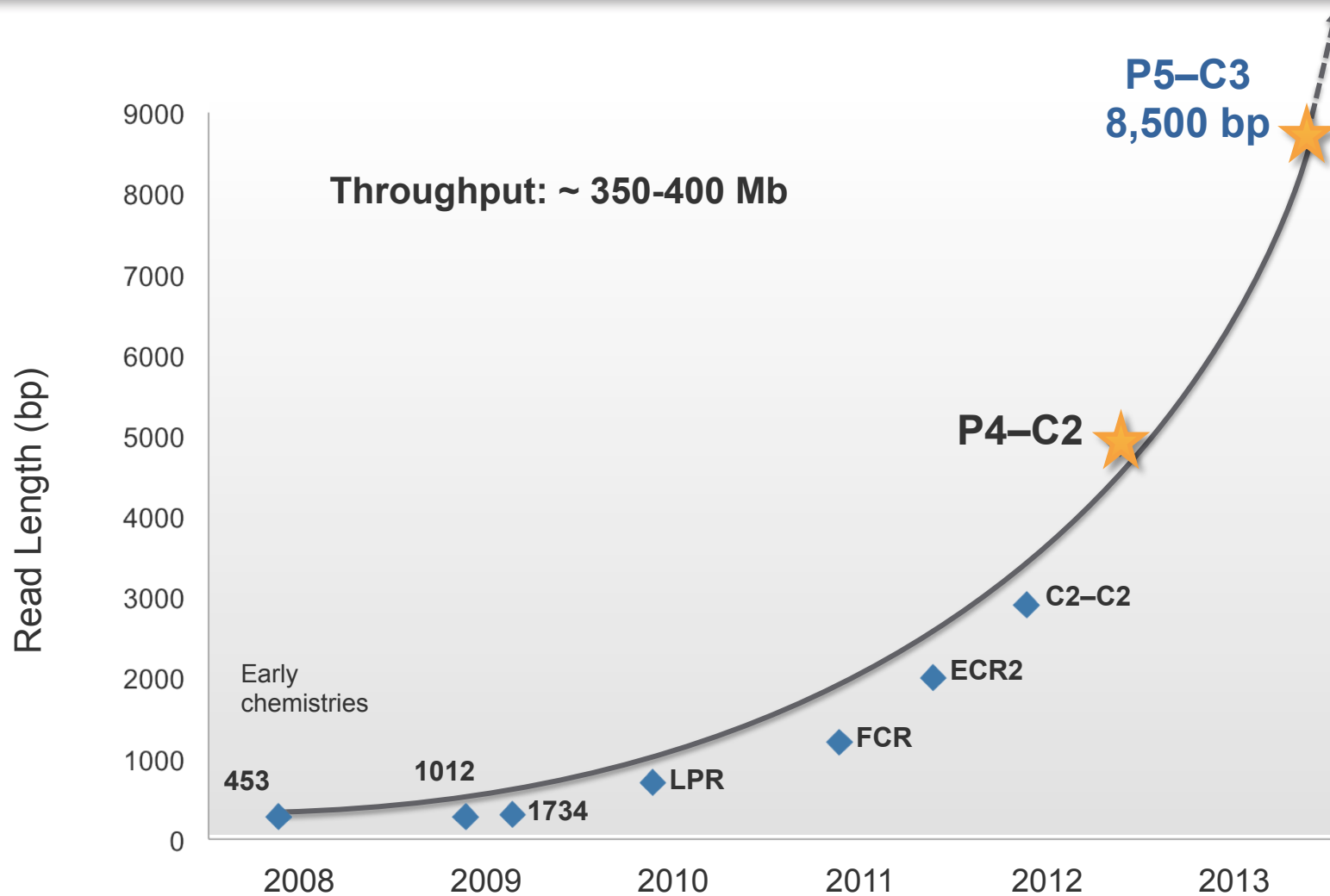
# Preliminary Rice Assemblies

Assembly	Contig NG50
HiSeq Fragments 50x 2x100bp @ 180	3,925
MiSeq Fragments 23x 459bp 8x 2x251bp @ 450	6,332
“ALLPATHS-recipe” 50x 2x100bp @ 180 36x 2x50bp @ 2100 51x 2x50bp @ 4800	18,248
PBeCR Reads 19x @ 3500 ** MiSeq for correction	50,995
Enhanced PBeCR 19x @ 3500 ** MiSeq for correction	155,695



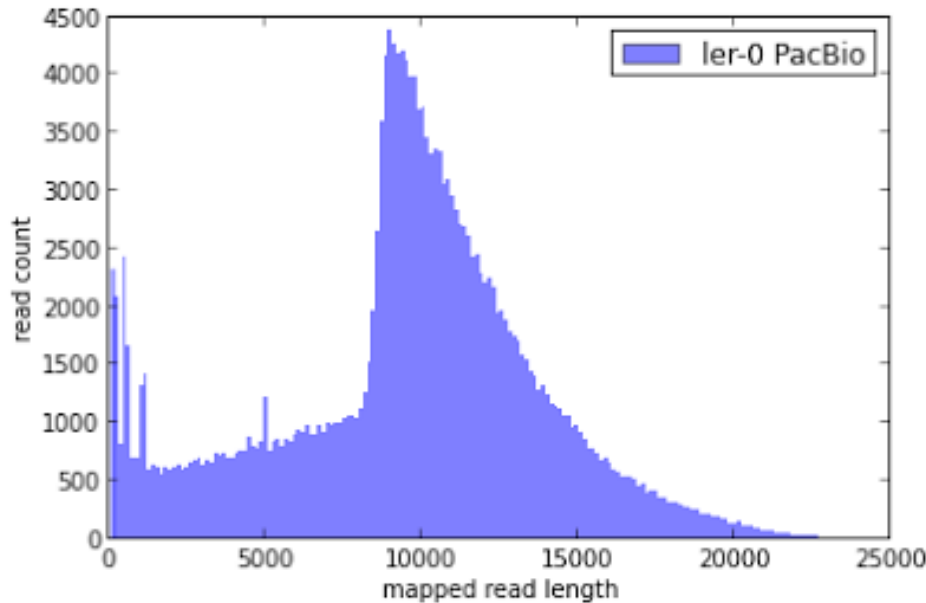
In collaboration with McCombie & Ware labs @ CSHL

# P5-C3 Chemistry Read Lengths



# De novo assembly of Arabidopsis

<http://blog.pacificbiosciences.com/2013/08/new-data-release-arabidopsis-assembly.html>



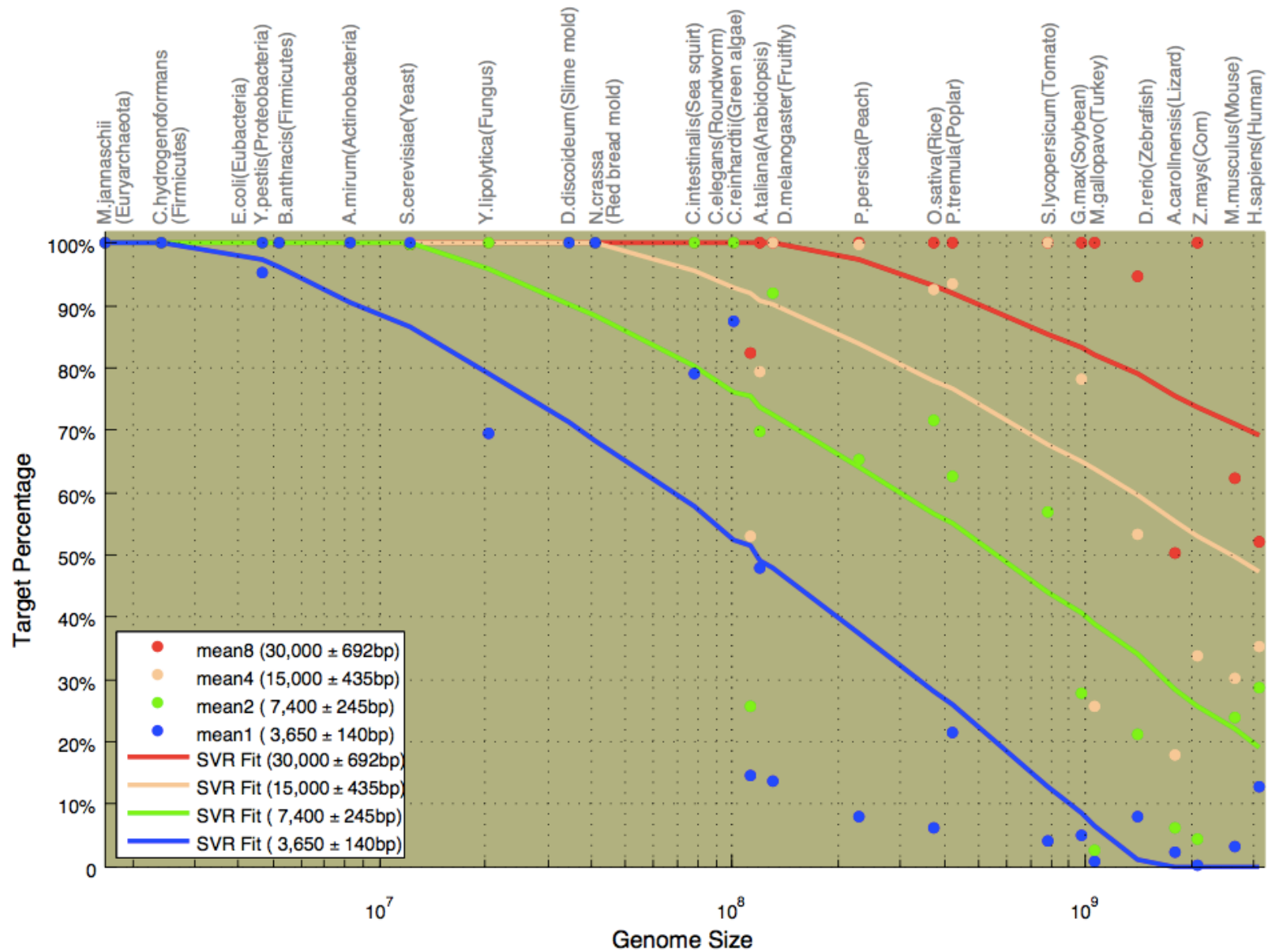
## *A. thaliana* Ler-0 sequenced at PacBio

- Sequenced using the latest P4 enzyme and C2 chemistry
- Size selection using an 8 Kb to 50 Kb elution window on a BluePippin™ device from Sage Science
- Total coverage >100x

Genome size: 124.6 Mb  
GC content: 33.92%  
Raw data: 11 Gb  
Assembly coverage: 15x over 9kbp

Sum of Contig Lengths: 149.5Mb  
Number of Contigs: 1788  
Max Contig Length: 12.4 Mb  
N50 Contig Length: 8.4 Mb

# Assembly Complexity of Long Reads



# Summary

- Hybrid assembly let us combine the best characteristics of 2<sup>nd</sup> and 3<sup>rd</sup> gen sequencing
  - Better repeat resolution and error correction by pre-assembling Illumina reads into unitigs
- Long reads and good coverage are the keys to a high quality de novo assembly
  - Single contig de novo assemblies of entire microbial chromosomes are now routine
  - Single contig de novo assemblies of entire plant and animal chromosomes on the horizon
- We are starting to apply these technologies to discover significant biology that is otherwise impossible to measure
  - Expect to see results in smaller genomes scale up over the next few years

# Acknowledgements

## Schatz Lab

Giuseppe Narzisi  
Shoshana Marcus  
James Gurtowski  
Srividya  
Ramakrishnan  
Hayan Lee  
Rob Aboukhalil  
Mitch Bekritsky  
Charles Underwood  
Tyler Gavin  
Alejandro Wences  
Greg Vulture  
Eric Biggers  
Aspyn Palatnick

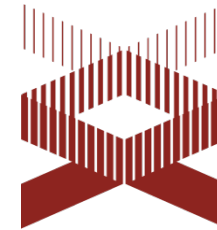
## CSHL

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

IT Department

## NBACC

Adam Phillippy  
Sergey Koren



National Human  
Genome Research  
Institute



U.S. DEPARTMENT OF  
**ENERGY**



# Thank You!

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

## Beyond the Genome 2013

1-3 October 2013

Mission Bay Conference Center, San Francisco, USA

[www.beyond-the-genome.com](http://www.beyond-the-genome.com)

A BioMed Central conference hosted by



**Genome Biology**  
Biology for the post-genomic era



**Genome Medicine**  
Medicine in the post-genomic era

