# SMRT® Sequencing Solutions for Investigative Studies to Understand Evolutionary Processes

Pacific Biosciences, 1380 Willow Road, Menlo Park, CA  94025

Jenny Gu, jgu@pacificbiosciences.com

## Abstract

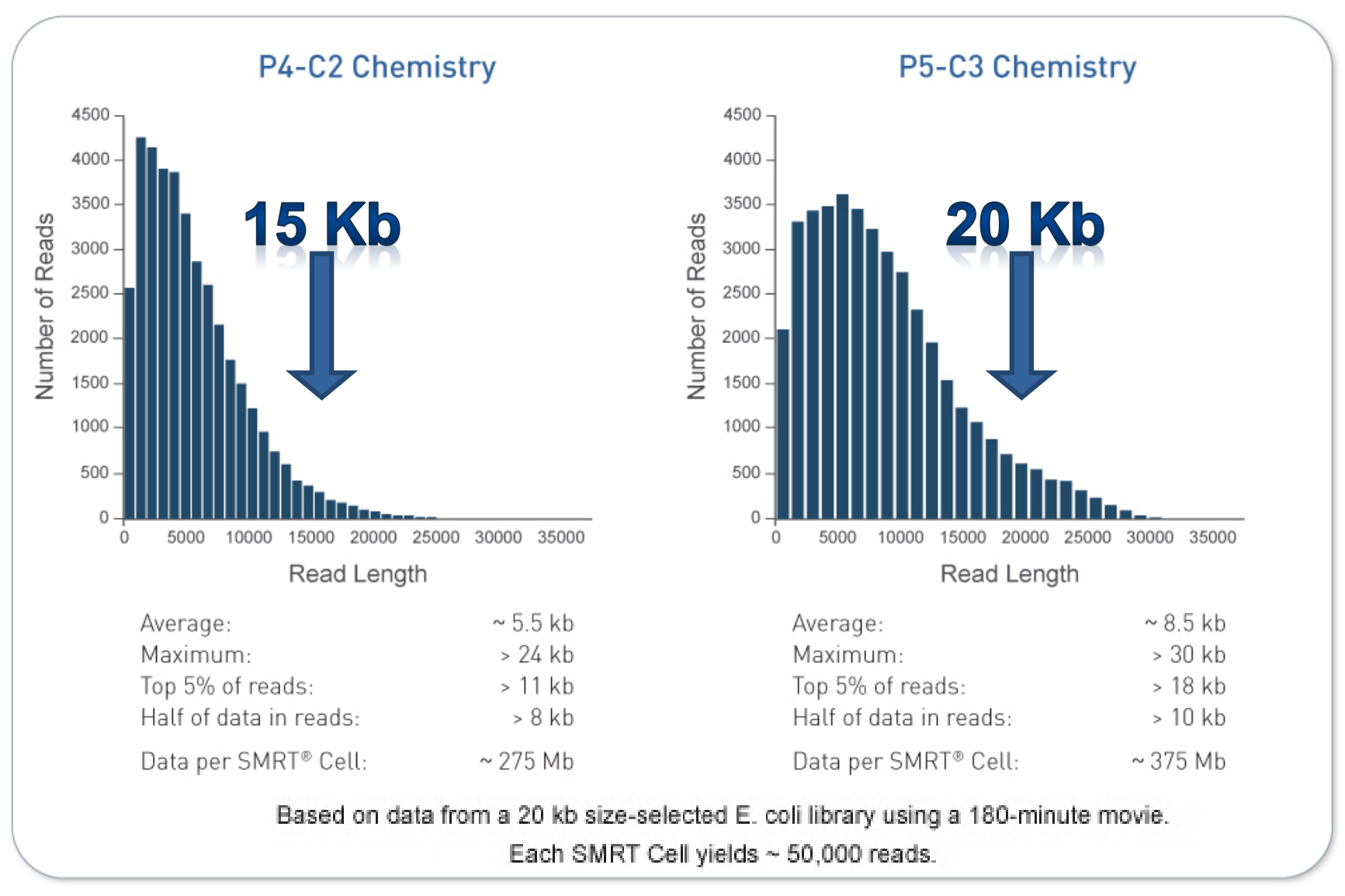
Single Molecule, Real-Time (SMRT®) Sequencing holds promise for addressing new frontiers to understand molecular mechanisms in evolution and gain insight into adaptive strategies. With read lengths exceeding 10 kb, we are able to sequence high-quality, closed microbial genomes with associated plasmids, and investigate large genome complexities, such as long, highly repetitive, low-complexity regions and multiple tandem-duplication events. Improved genome quality, observed at 99.9999% (QV60) consensus accuracy, and significant reduction of gap regions in reference genomes (up to and beyond 50%) allow researchers to better understand coding sequences with high confidence, investigate potential regulatory mechanisms in noncoding regions, and make inferences about evolutionary strategies that are otherwise missed by the coverage biases associated with short-read sequencing technologies.

Additional benefits afforded by SMRT Sequencing include the simultaneous capability to detect epigenomic modifications and obtain full-length cDNA transcripts that obsolete the need for assembly. With direct sequencing of DNA in real-time, this has resulted in the identification of numerous base modifications and motifs, which genome-wide profiles have linked to specific methyltransferase activities. Our new offering, the Iso-Seq™ Application, allows for the accurate differentiation between transcript isoforms that are difficult to resolve with short-read technologies. PacBio® reads easily span transcripts such that both 5'/3' primers for cDNA library generation and the poly-A tail are observed. As such, exon configuration and intron retention events can be analyzed without ambiguity. This technological advance is useful for characterizing transcript diversity and improving gene structure annotations in reference genomes.

We review solutions available with SMRT Sequencing, from targeted sequencing efforts to obtaining reference genomes (>100 Mb). This includes strategies for identifying microsatellites and conducting phylogenetic comparisons with targeted gene families. We highlight how to best leverage our long reads that have exceeded 20 kb in length for research investigations, as well as currently available bioinformatics strategies for analysis. Benefits for these applications are further realized with current use of size selection of input sample using the BluePippin™ device from Sage Science as demonstrated in our genome improvement projects. Using the latest P5-C3 chemistry on model organisms, these efforts have yielded an observed contig N50 of ~6 Mb, with the longest contig exceeding 12.5 Mb and an average base quality of QV50.

## Single Molecule Real Time (SMRT®) Sequencing

### PacBio® RS II Sequencing Chemistries Provide Long Read Lengths >20 Kb



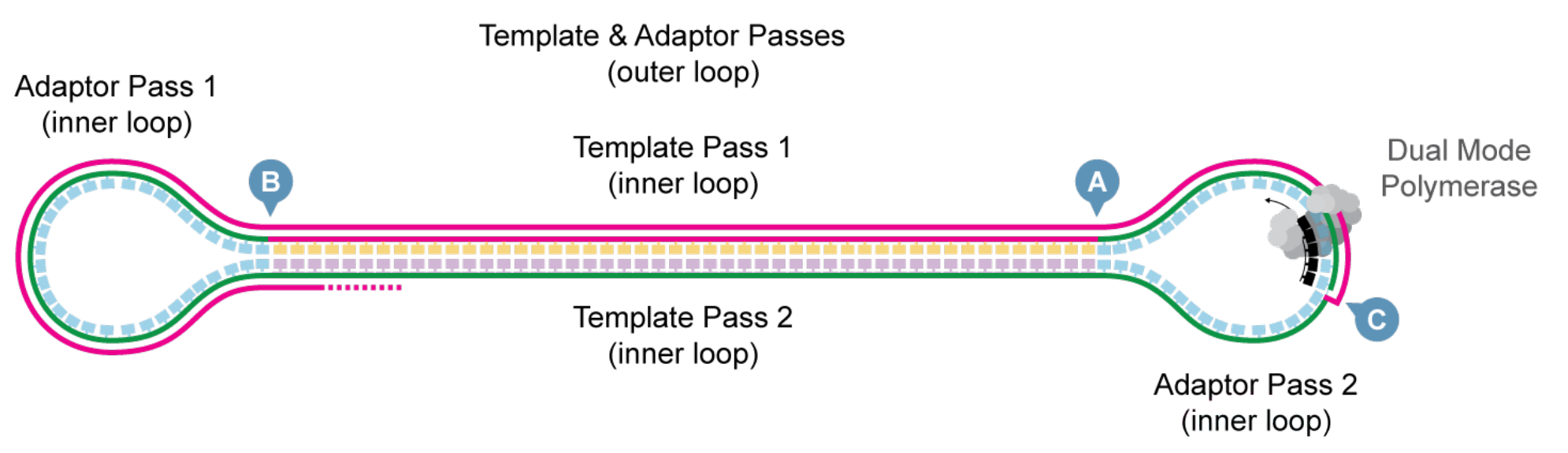
| P4-C2 Chemistry | |
|---|---|
| Average: | ~ 5.5 kb |
| Maximum: | > 24 kb |
| Top 5% of reads: | > 11 kb |
| Half of data in reads: | > 8 kb |
| Data per SMRT® Cell: | ~ 275 Mb |

| P5-C3 Chemistry | |
|---|---|
| Average: | ~ 8.5 kb |
| Maximum: | > 30 kb |
| Top 5% of reads: | > 18 kb |
| Half of data in reads: | > 10 kb |
| Data per SMRT® Cell: | ~ 375 Mb |

Based on data from a 20 kb size-selected E. coli library using a 180-minute movie.
Each SMRT Cell yields ~ 50,000 reads.

Example read length distribution from a SMRT Sequencing run with 20 kb size-selected *E. coli* library using a 180-min movie. Average throughput of 350 Mb per SMRT Cell with ~50,000 reads.

### Universal SMRTbell™ Template



- Mode change induced by template change
- Single-Stranded Mode
- Strand-Displacement Mode

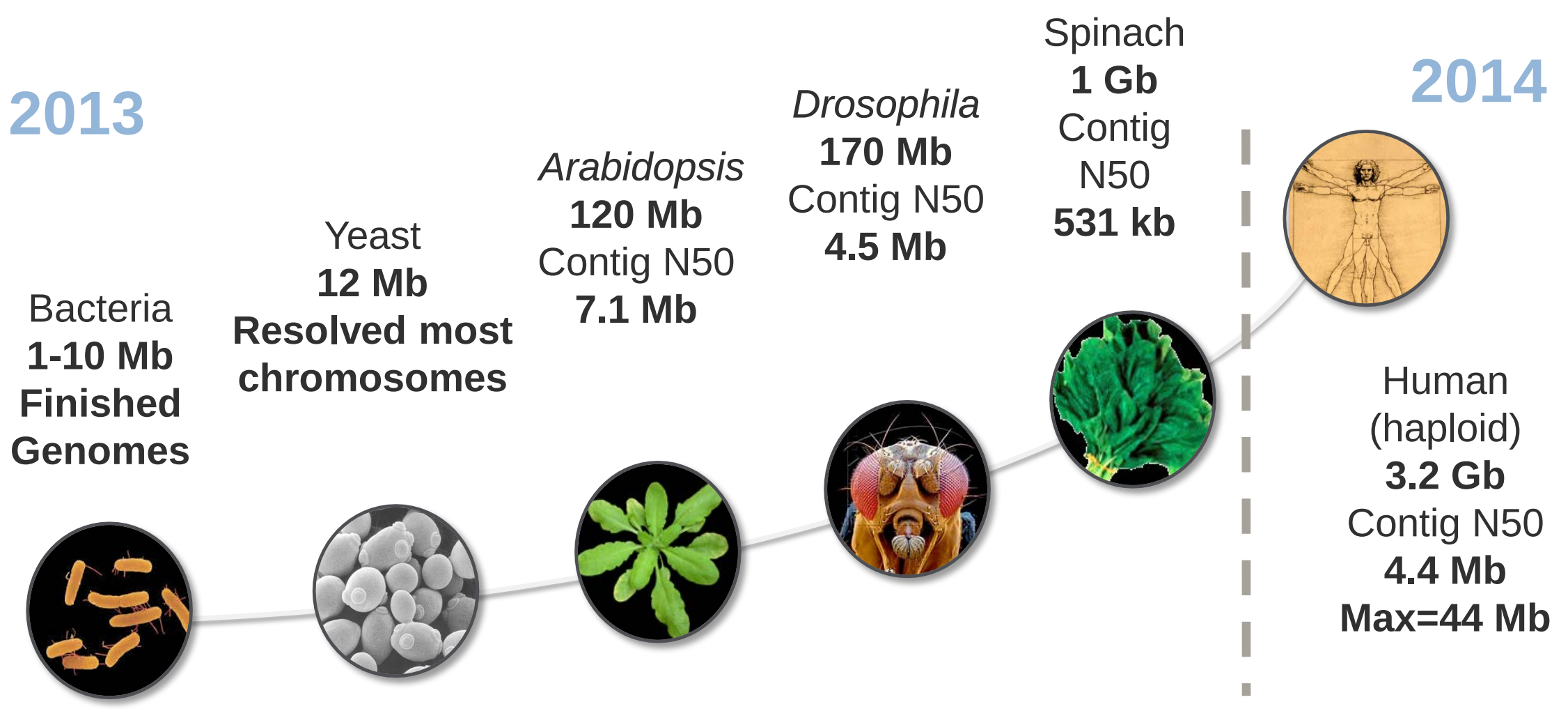
- Insert size: 2 kb to > 10 kb
- Movie collection time: 180 min

**A: Template Pass 1** = Strand-Displacement Mode
**B: Adaptor Pass 1** = Single-Stranded Mode
**Template Pass 2** = Single-Stranded Mode
**C: Adaptor Pass 2** = Single-Stranded (some ds) Mode
**Template & Adaptor Passes** = Strand-Displacement Mode

Schematic of SMRTbell Template Sequencing

## Top 10 Applications for Evolutionary Biology

### (1) PacBio-only *de novo* Genome Assemblies



Bacteria
**1-10 Mb**
**Finished Genomes**

Yeast
**12 Mb**
**Resolved most chromosomes**

*Arabidopsis*
**120 Mb**
Contig N50
**7.1 Mb**

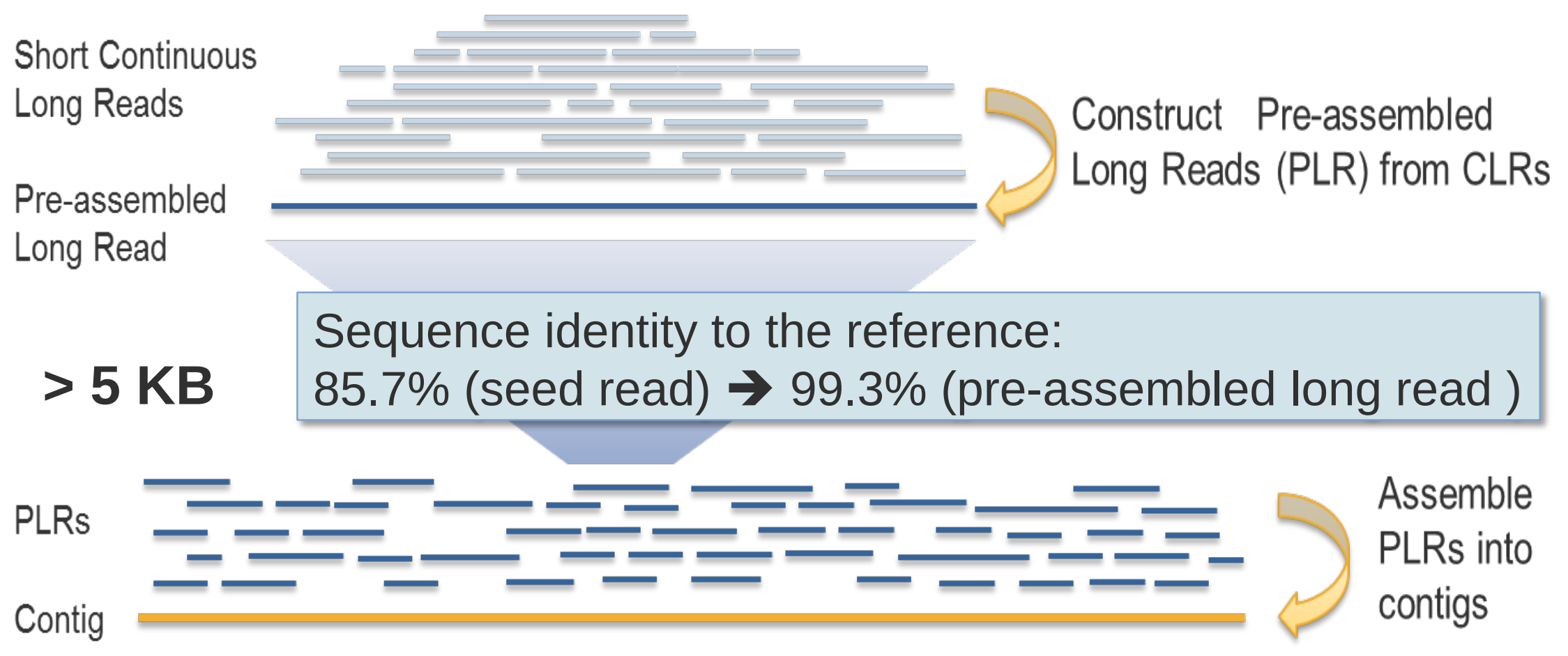
*Drosophila*
**170 Mb**
Contig N50
**4.5 Mb**

Spinach
**1 Gb**
Contig N50
**531 kb**

Human (haploid)
**3.2 Gb**
Contig N50
**4.4 Mb**
**Max=44 Mb**

Public datasets, SMRT Analysis and compatible third party software are available from PacBio DevNet: **http://pacbiodevnet.com/**

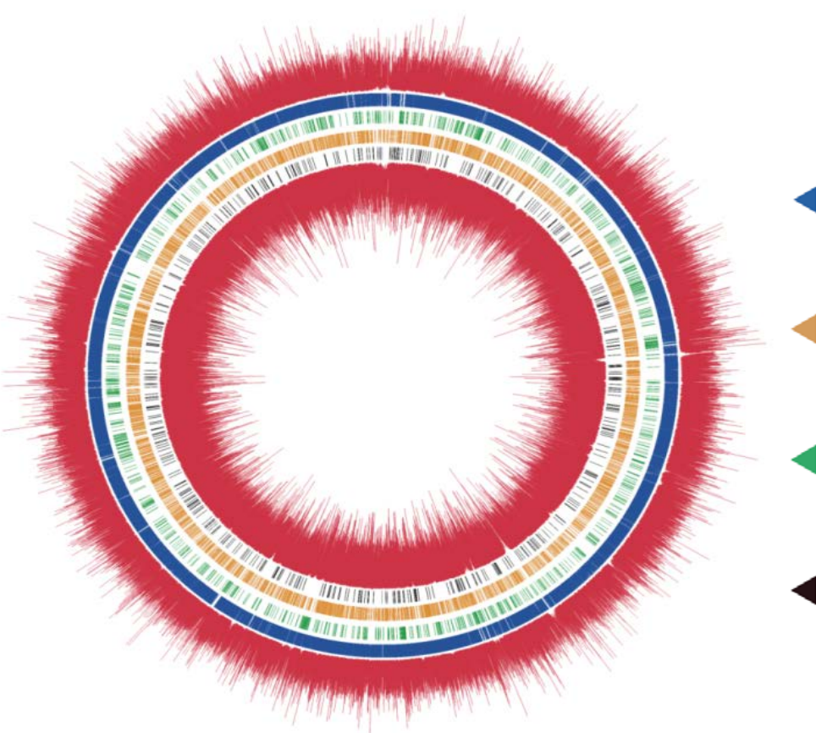
### (2) Hierarchical Genome Assembly Process (HGAP)



Sequence identity to the reference:
85.7% (seed read) ➔ 99.3% (pre-assembled long read )

A bioinformatics solution to make best use of all PacBio reads for improved consensus accuracy and genome assemblies.

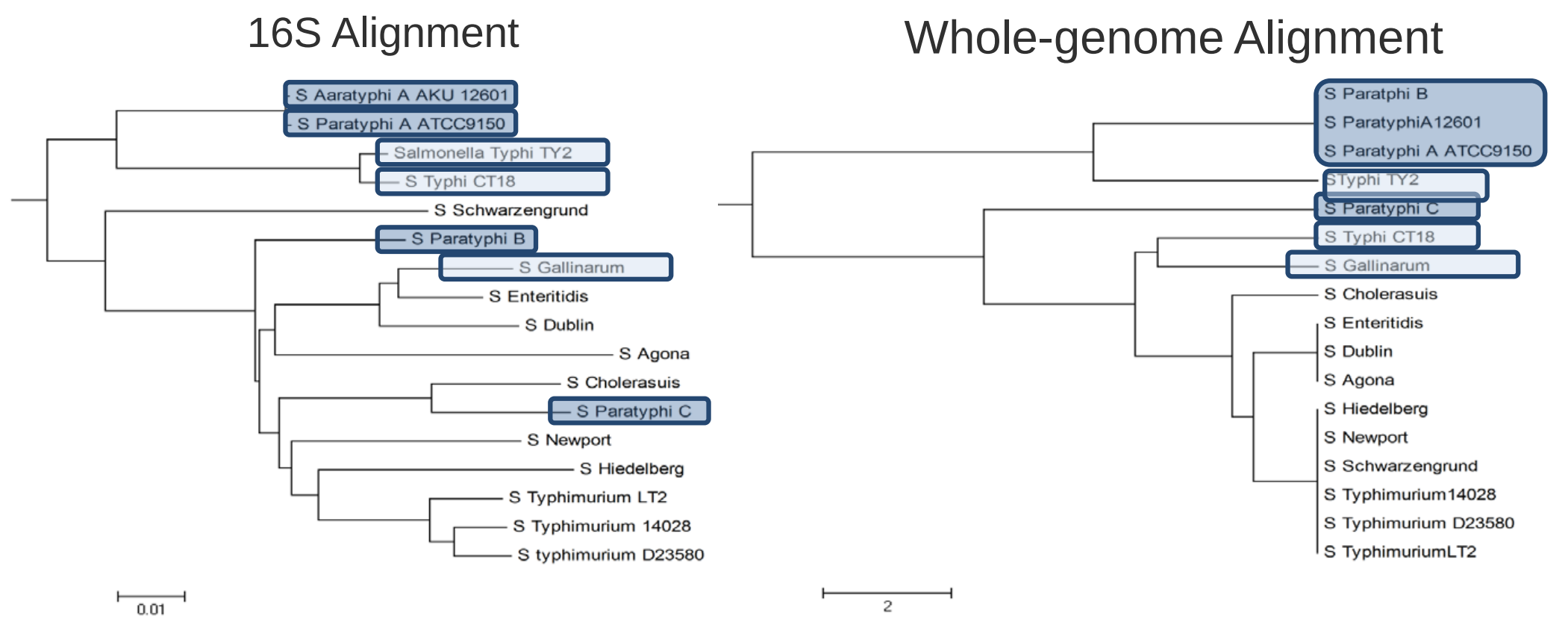
### (3) Obtain Complete Microbial Genomes with Epigenetic Modifications and Associated Plasmids



| Motif | Occurrence in Genome | Modified in Genome | % Modified |
|---|---|---|---|
| 5'-GATC-3'<br>3'-CTAG-5' | 42,992 | 41,969 | 97.6% |
| 5'-ACCACC-3'<br>3'-TGGTGG-5' | 4,569 | 4,492 | 98.3% |
| 5'-CTGCAG-3'<br>3'-GACGTC-5' | 2,746 | 2,678 | 97.5% |
| 5'-CCACN$_8$TGAY-3'<br>3'-GGTGN$_8$ACTR-5' | 492<br>492 | 478<br>484 | 97.2%<br>98.4% |

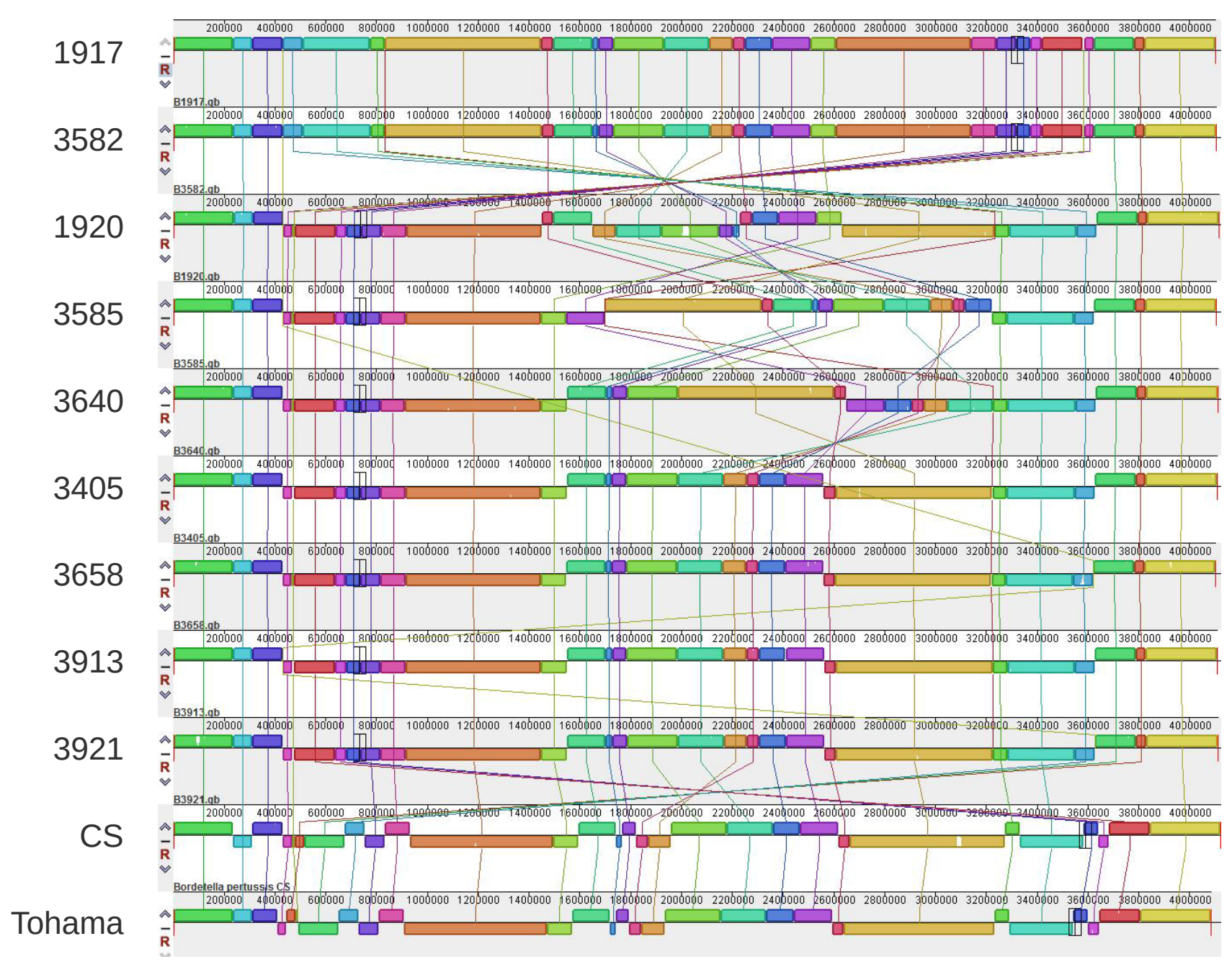
Genome-wide detection of methylation for the German *E. coli* outbreak strain.

### (4) Conduct Whole-Genome Phylogeny



Collaboration with P. Desai, B. Weimer (UC Davis)

### (5) Compare Genome Organization between Strains



Collaboration with A. Zeddeman, H. van der Heide, M. Bart & F. Mooi
National Institute for Public Health and the Environment (RIVM), Netherlands

### (6) Discover *de novo* Microsatellites

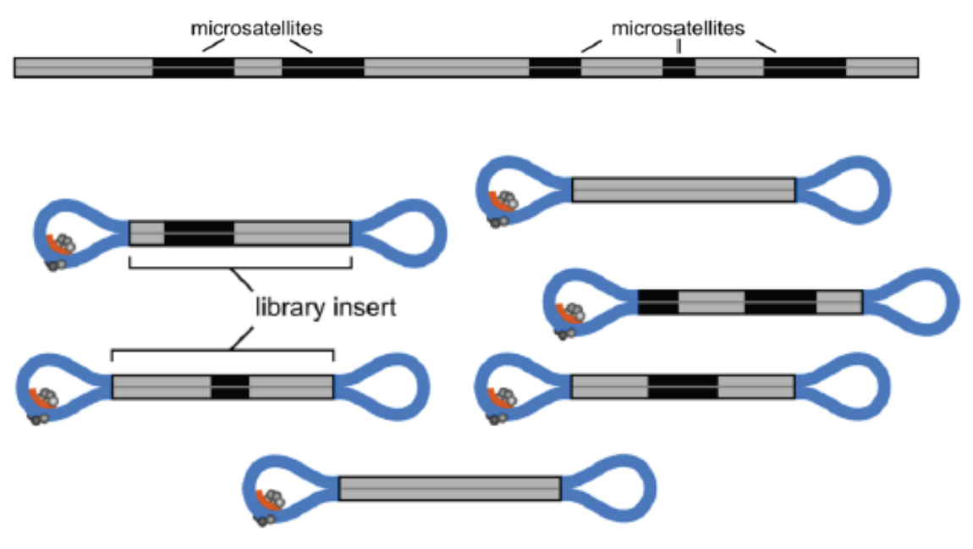
316 microsatellite loci in a single-genome shotgun sequencing

Handle large insert sizes (3 kb) with high-quality results

Primer design advantage with long reads and ~90% conversion

Economically viable

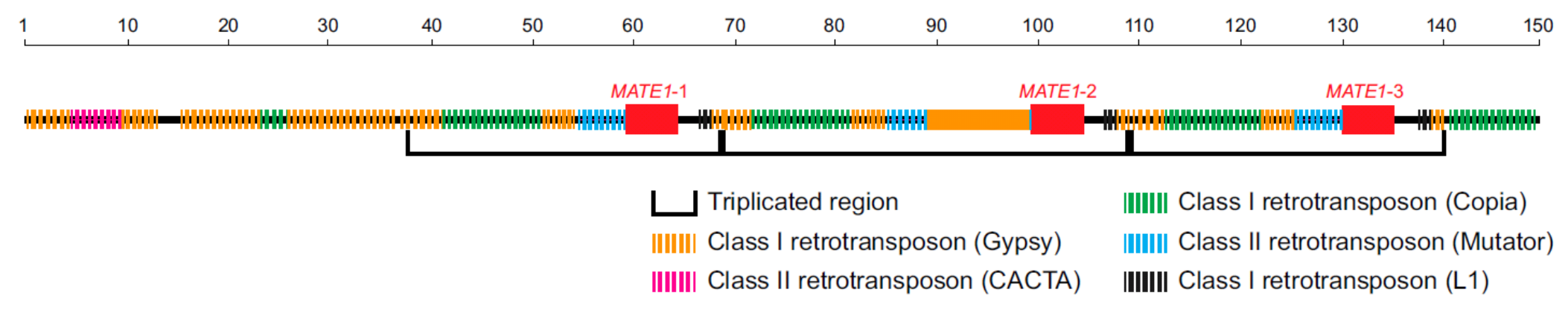
Grohme M.A., et al. (2013) *Biotechniques* 55(5), 253–256.



| Motifs Identified | |
|---|---|
| 213 | dinucleotide |
| 50 | trinucleotide |
| 25 | pentanucleotide |

### (7) Detect Complex Duplication &Triplication Events

**Genomic organization of the MATE1 locus**


1
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
MATE1-1
MATE1-2
MATE1-3
Triplicated region
Class I retrotransposon (Gypsy)
Class II retrotransposon (CACTA)
Class I retrotransposon (Copia)
Class II retrotransposon (Mutator)
Class I retrotransposon (L1)

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

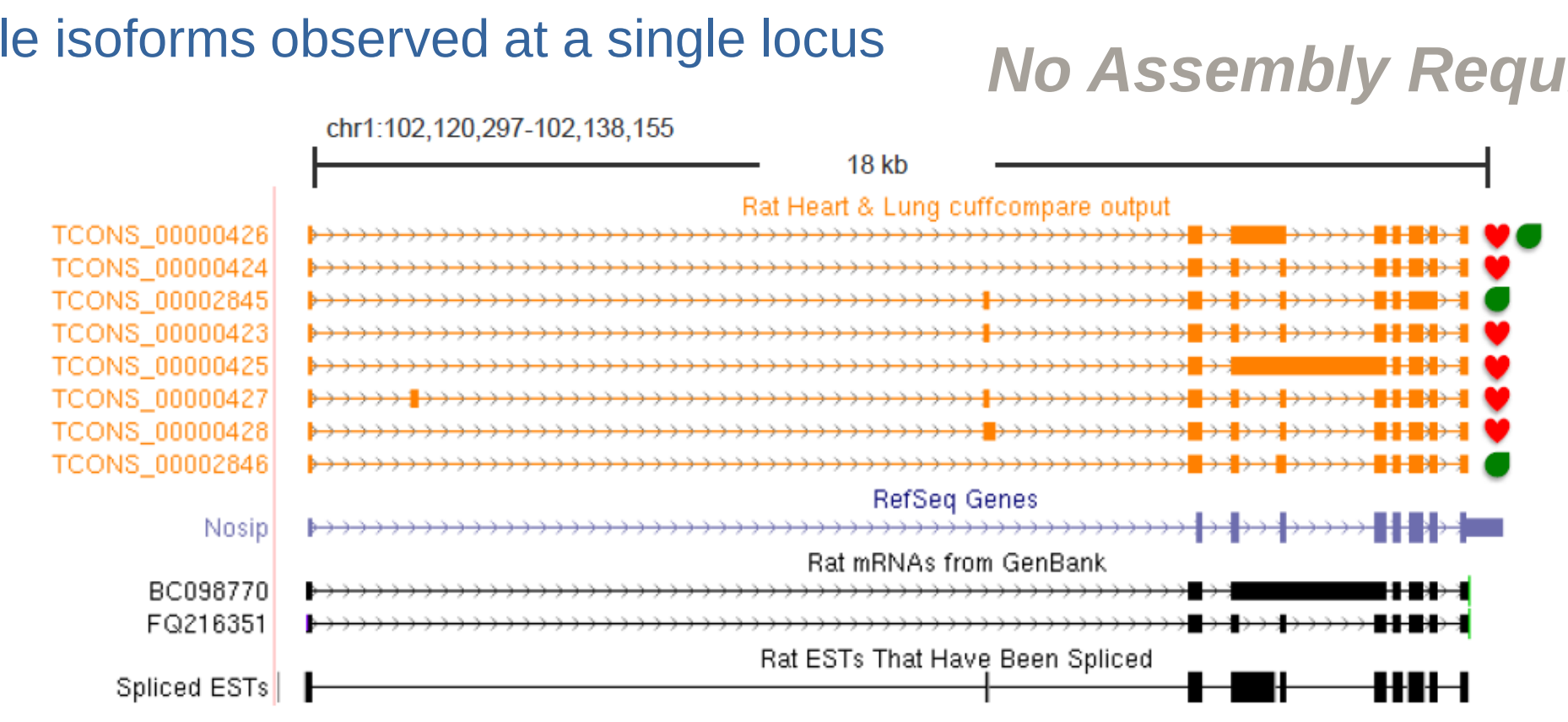
- Segregating population localized a QTL on a BAC, but unable to genotype with short-read sequencing due to high repeat content and GC skew
- BAC assembly with PacBio long reads revealed a triplication of the ZnMATE1 membrane transporter

Maron, LG et al. (2013) *PNAS* 110(13):5241-6

### (8) Direct Full-length Isoform Sequencing

Multiple isoforms observed at a single locus

**No Assembly Required!**



Tseng, E. (2014) PAG conference Poster

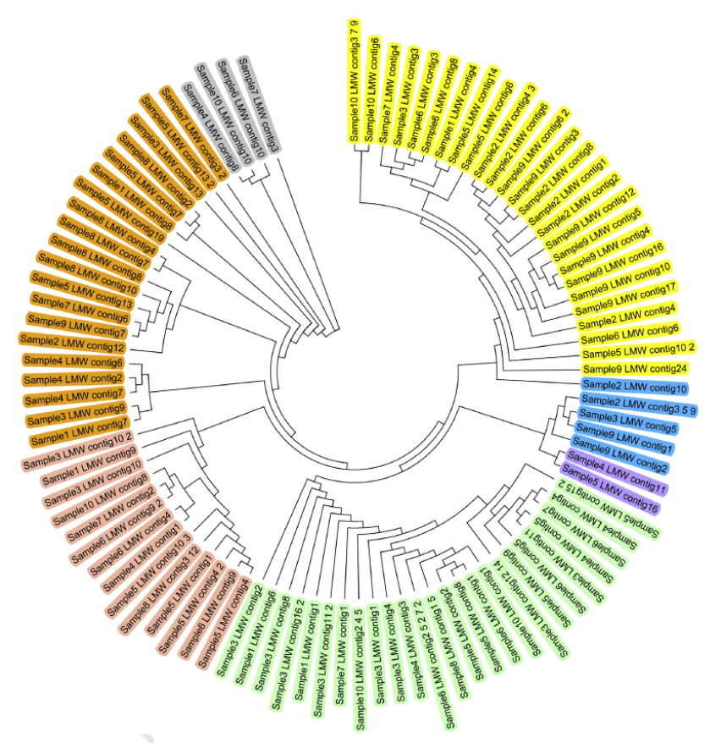
Rat heart ♥ Rat lung ●

### (9) Survey Targeted Gene Families

Detailed **gluten** content and composition **overview** from **10 wheat cultivars** in **single SMRT Cell**

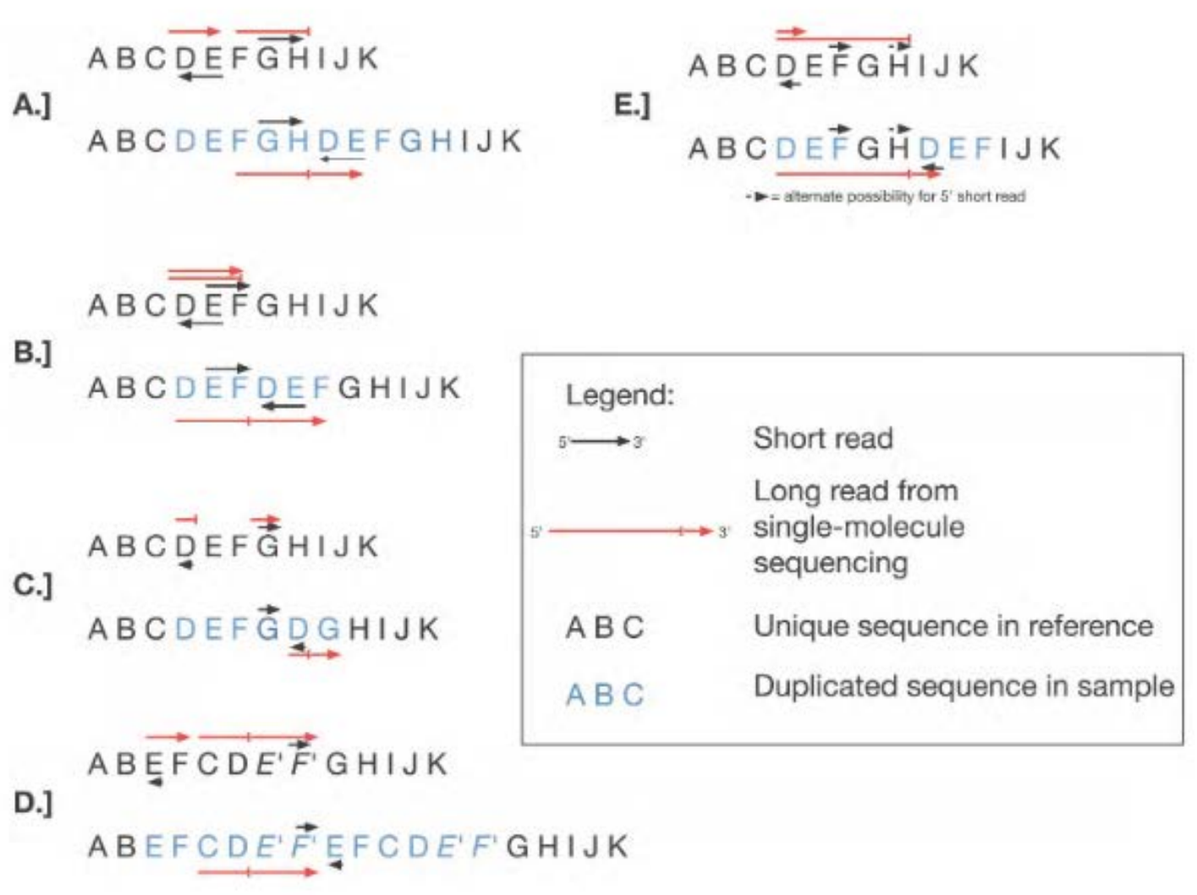
(424 proteins)

**Validated with 99% Concordance**



Zhang W et al. (2013) *Gene* doi:10.1016/j.gene.2013.10.009

### (10) Whole Genome Survey of Duplication Events



5 example scenarios of duplication events where **Pacbio® long reads resolved breakpoint ambiguities** in two whole genome surveys. (*D. yakuba and D. simulans*)

Confirmed 661 of 668 mutations with low genome coverage.

Breakpoints confirmed:
PacBio: 96.1%
Split short read mapping: 12.7%

Rogers RL, et. al. (2014) *Mol Biol Evol* doi: 10.1093/molbev/msu124

## Recent Publications

- PacBio Blog: *A. thaliana* genome assembly effort
- PacBio Blog : *D. melanogaster* genome assembly effort
- PacBio Blog: PAG PacBio Workshop Showcases User's Large Genome Efforts
- Chin CS., et. al. "Nonhybrid , finished microbial genome assemblies from long-read SMRT sequencing data." *Nat Methods*. Jun;10(6):563-9 (2013).
- Case Study: Beyond Four Bases: Epigenetic Modifications Prove Critical to Understanding *E. Coli* Outbreak (2012)
- Cooper KK, et. al. "Comparative genomics of enterohemorrhagic Escherichia coli O145:H28 demonstrates a common evolutionary lineage with Escherichia coli O156:H7 (2014)
- Grohme M.A., et al. "Microsatellite marker discovery using single molecule real-time circular consensus sequencing on the Pacific Biosciences RS" *Biotechniques* 55(5), 253–256 (2013).
- Maron, LG et al. (2013) A rare gene copy-number variant that contributes to maize aluminum tolerance and adaptation to acid soils. *PNAS* 110(13):5241-6
- Poster: Tseng, E. et. al. " Isoform Sequencing: Unveiling the Complex Landscape of the Eukaryotic Transcriptome on the PacBio® RS II" (PAG conference 2014)
- Zhang W et. al. (2013) "PacBio sequencing of gene families – A case study with wheat gluten genes." *Gene* doi:10.1016/j.gene.2013.10.009
- Rogers RL, et. al. (2014) "Landscape of Standing Variation for Tandem Duplications in Drosophila yakuba and Drosophila simulans." *Mol Biol Evol* doi: 10.1093/molbev/msu124