# SMRT® Sequencing Solutions for Large Genomes and Transcriptomes

Jenny Gu, Jason Chin, Paul Peluso, David Rank, Kristi Kim, Jane Landolin, Elizabeth Tseng, Susana Wang, Primo Baybayan

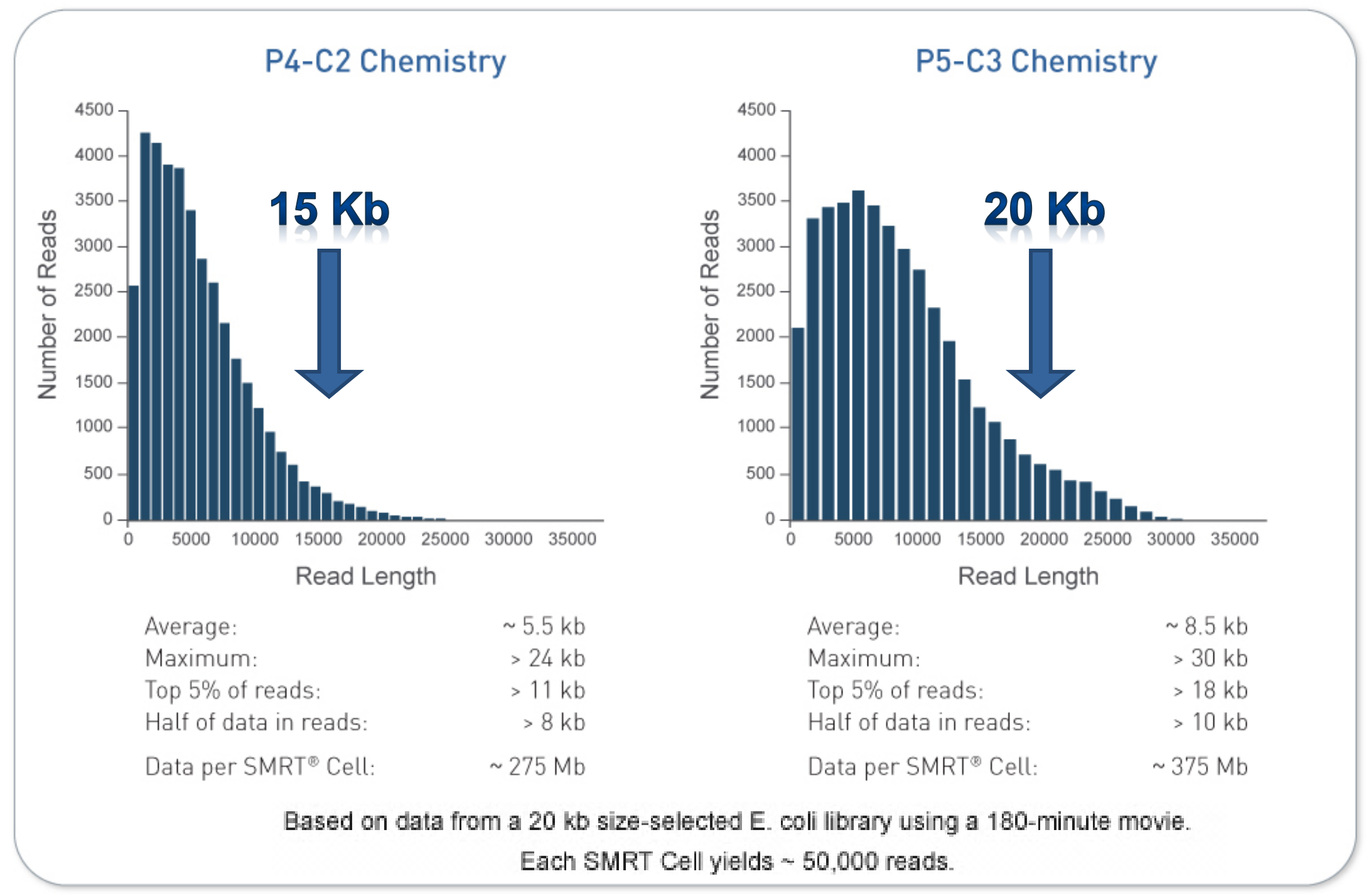
Pacific Biosciences, Menlo Park, CA

## Introduction

Single Molecule, Real-Time (SMRT) Sequencing holds promise for addressing new frontiers in large genome complexities, such as long, highly repetitive, low-complexity regions and duplication events, and differentiating between transcript isoforms that are difficult to resolve with short-read technologies. We present solutions available for both reference genome improvement (>100 MB) and transcriptome research to best leverage long reads that have exceeded 20 Kb in length. Benefits for these applications are further realized with consistent use of size-selection of input sample using the BluePippin™ device from Sage Science. Highlights from our genome assembly projects using the latest P5-C3 chemistry on model organisms will be shared. Assembly contig N50 have exceeded 6 Mb and we observed longest contig exceeding 12.5 Mb with an average base quality of QV50. Additionally, the value of long, intact reads to provide a no-assembly approach to investigate transcript isoforms using our Iso-Seq™ Application will be presented.

## Methods

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

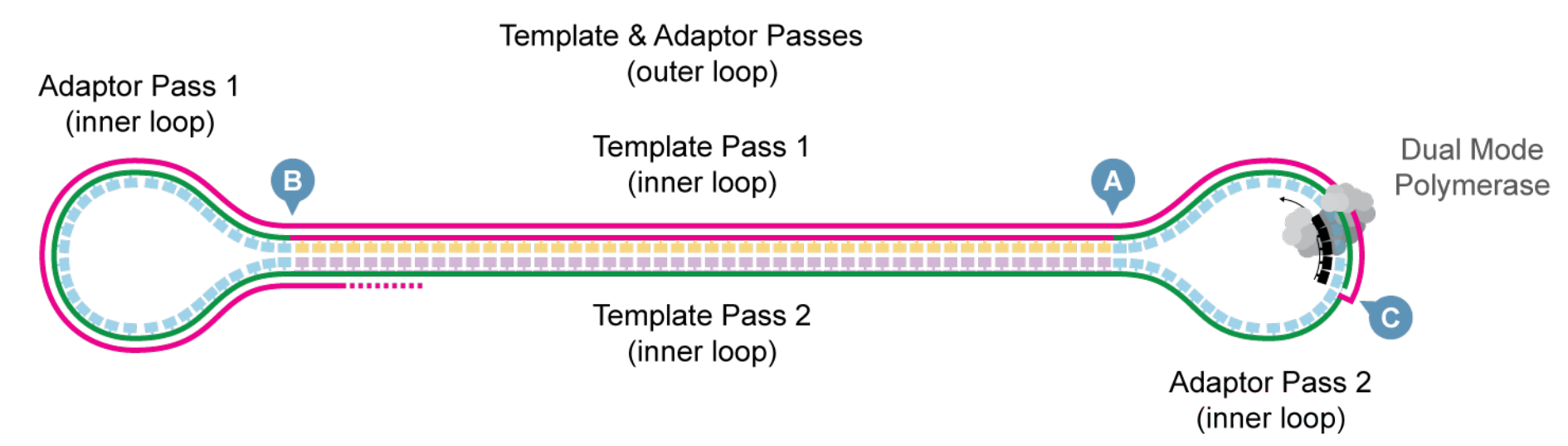


| P4-C2 Chemistry | | P5-C3 Chemistry | |
|---|---|---|---|
| Average: | ~ 5.5 kb | Average: | ~ 8.5 kb |
| Maximum: | > 24 kb | Maximum: | > 30 kb |
| Top 5% of reads: | > 11 kb | Top 5% of reads: | > 18 kb |
| Half of data in reads: | > 8 kb | Half of data in reads: | > 10 kb |
| Data per SMRT® Cell: | ~ 275 Mb | 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.

**Figure 1.** 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



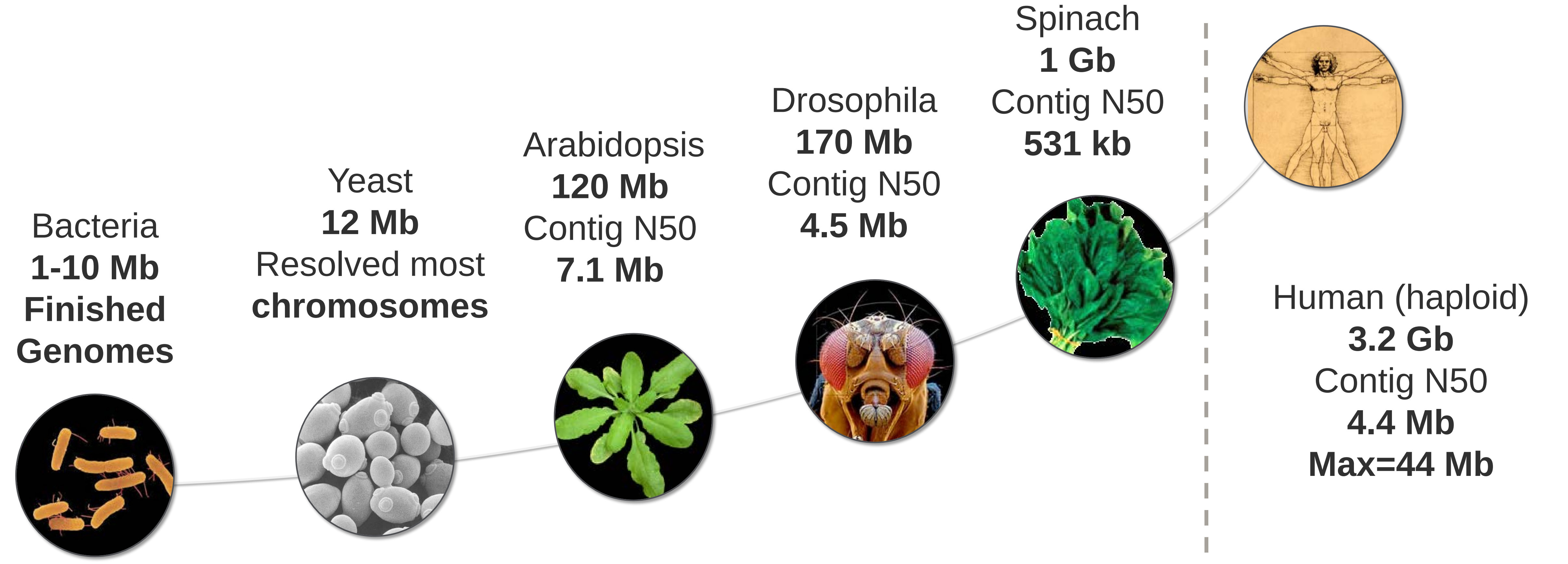
- Mode change induced by template change
- Single-Stranded Mode
- Strand-Displacement Mode

- 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

**Figure 2**. Schematic of SMRTbell Sequencing

## Progress of PacBio-Only *de novo* Large Genome Applications



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/**

**Figure 3**. Genomes completed using only PacBio Sequencing. Basic assembly stats provided.

## Genome

### One Contig Assembly = One Chromosome

Example: *D. melanogaster* PacBio-Only Assembly

| | Haploid<br>(Celera Assembler PBcR)<br>University of Maryland | Diploid<br>(FALCON)<br>Pacific Biosciences |
|---|---|---|
| # Contigs | 128 | 434 |
| N50 | 15.29 MB | 5.00 MB |
| Max | 24.62 MB | 21.34 MB |
| Total | 138.36 MB | 153.34 MB |

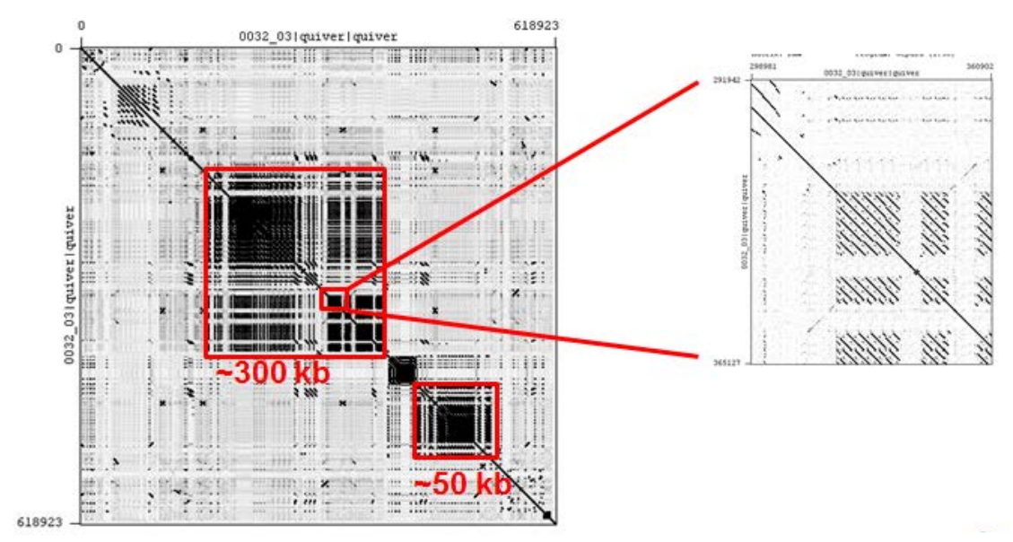
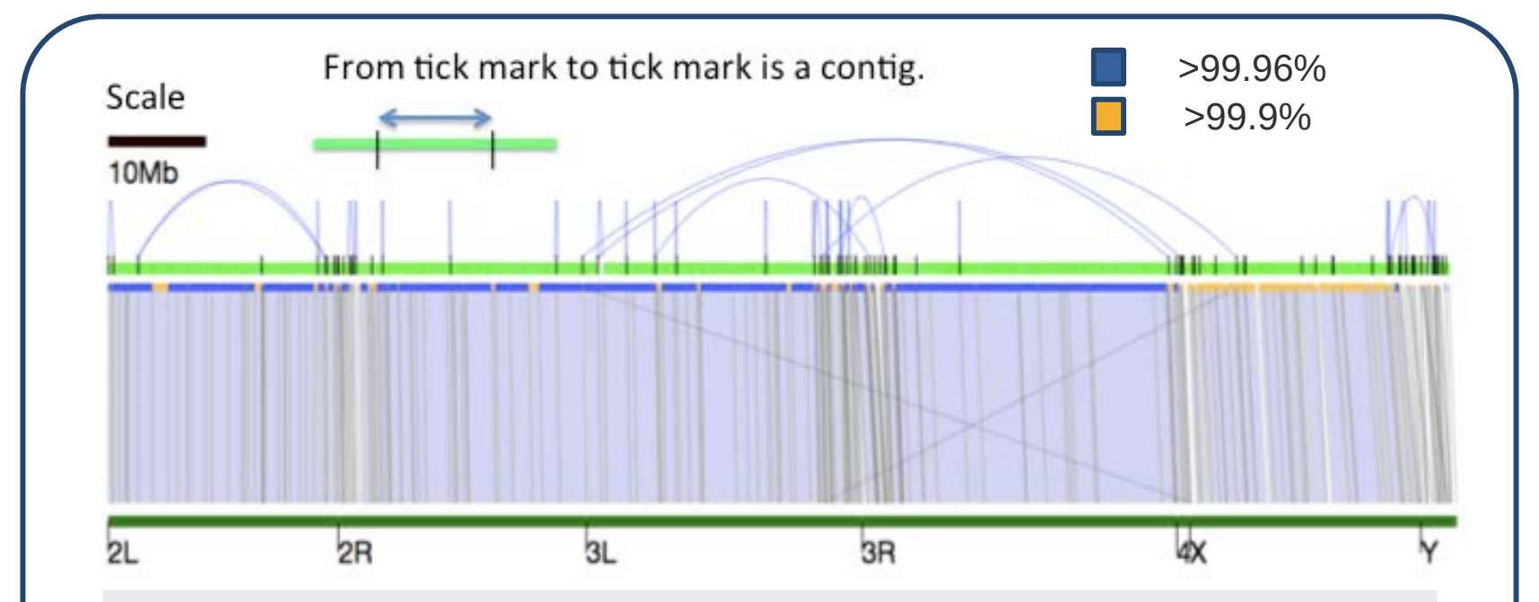


**Figure 4**. Preliminary results of two assemblies for a PacBio-only *D. melanogaster* genome (Above, Left). Assembly of Y-chromosome (50%) missing in Reference Genome (1%) with complex repeat regions spanning multi-kilobases (Above, Right).



Number of Contigs

| | RELEASE 5 Reference Genome | U of Maryland *de novo* Haploid Assembly |
|---|---|---|
| Chr 2L | 6 | 6 |
| Chr 2R | 27 | 2 |
| Chr 3L | 22 | 1 |
| Chr 3R | 15 | 3 |
| Chr 4 | 2 | 3 |
| Chr X | 3 | 42 |

**Figure 5**.

Comparison of PacBio-Only *D. melanogaster* preliminary assembly using FALCON aligned to current RELEASE 5 reference genome (Left).

Chr3L was assembled into 1 contig using only PacBio sequencing.

Only adult male flies were sequenced. The poor assembly of ChrX is likely a result of lower coverage given the expected 50:50 ratio of sex chromosomes in the library.

For more details about *D. melanogaster* effort, See PacBio Blog Entry: http://blog.pacificbiosciences.com/2014/01/data-release-preliminary-de-novo.html

## Transcriptome

### Iso-Seq™ Application: Full-length Intact Transcripts No Assembly Required

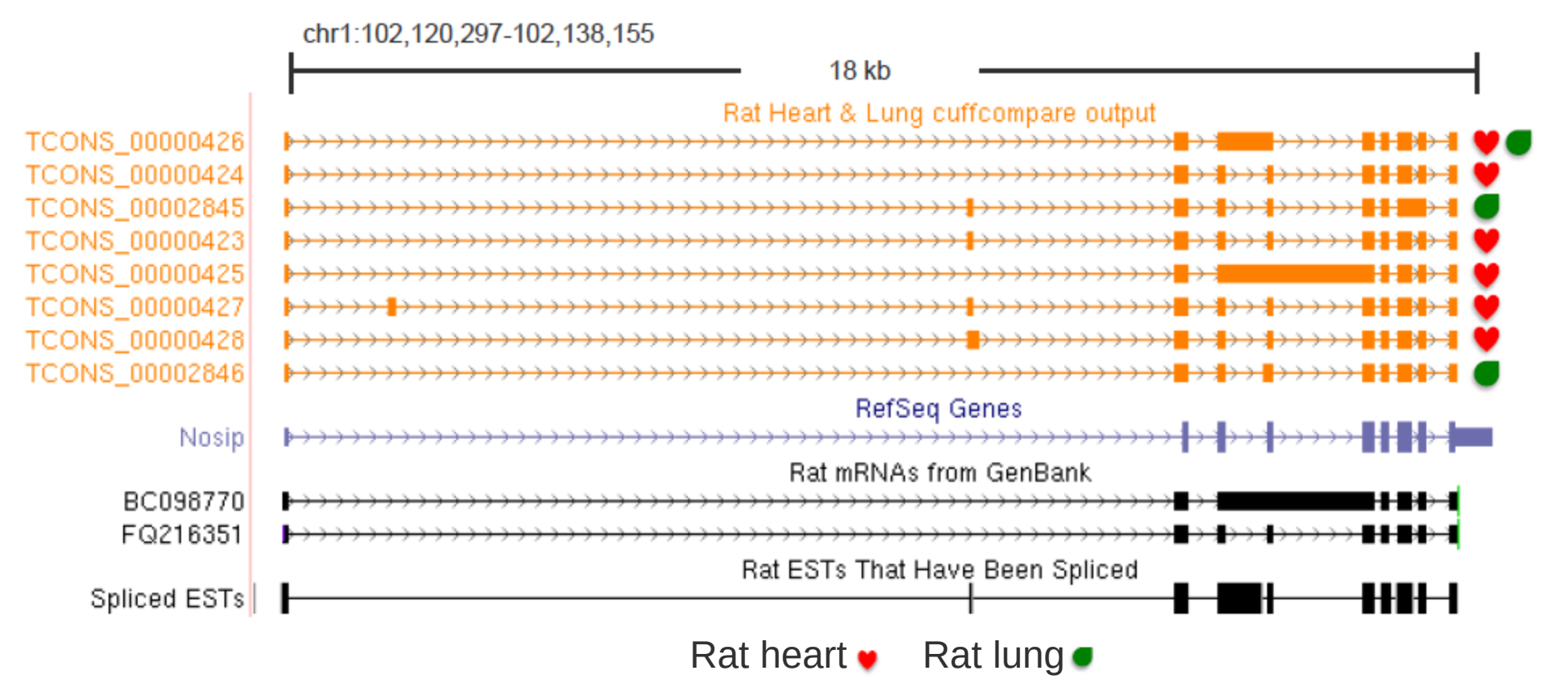
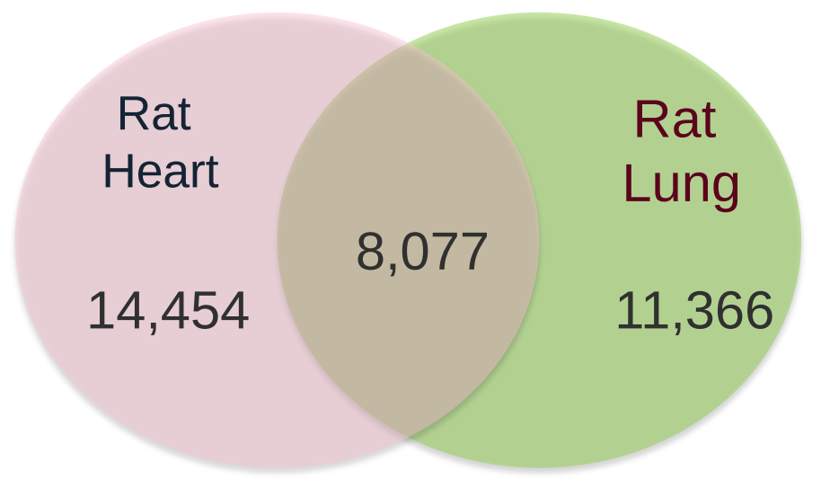


**Figure 6**.

Full-length transcript sequencing, defined by the observance of 5'/3' primers and poly-A tails, allows for the differentiation of isoforms without assembly. Tissue-specific isoforms have been identified from rat heart and lung RNA. Improved transcript sequencing leads to better gene model annotation.



Quality of transcripts from Iso-Seq Method summarized below.

| Sample | Number of transcripts | Aligned transcript coverage | | Base differences against reference genome | | | |
|---|---|---|---|---|---|---|---|
| | | 95-99% | 100% | Sub | Ins | Del | Total |
| Heart | 51,043 | 11,262<br>(22%) | 37,105<br>(73%) | 166,359<br>(0.17%) | 91,146<br>(0.09%) | 113,190<br>(0.12%) | 370,695<br>(0.39%) |
| Lung | 44,083 | 7,908<br>(18%) | 33,474<br>(76%) | 172,871<br>(0.25%) | 64,189<br>(0.09%) | 101,014<br>(0.15%) | 338,074<br>(0.49%) |

Iso-Seq™ Application webinar for additional information: http://j.mp/1kK2Cy5

## Best Practice Corner

### Considerations for Success:

HMW DNA and BluePippin™ Size Selection

Electropherogram of SMRTbell™ library

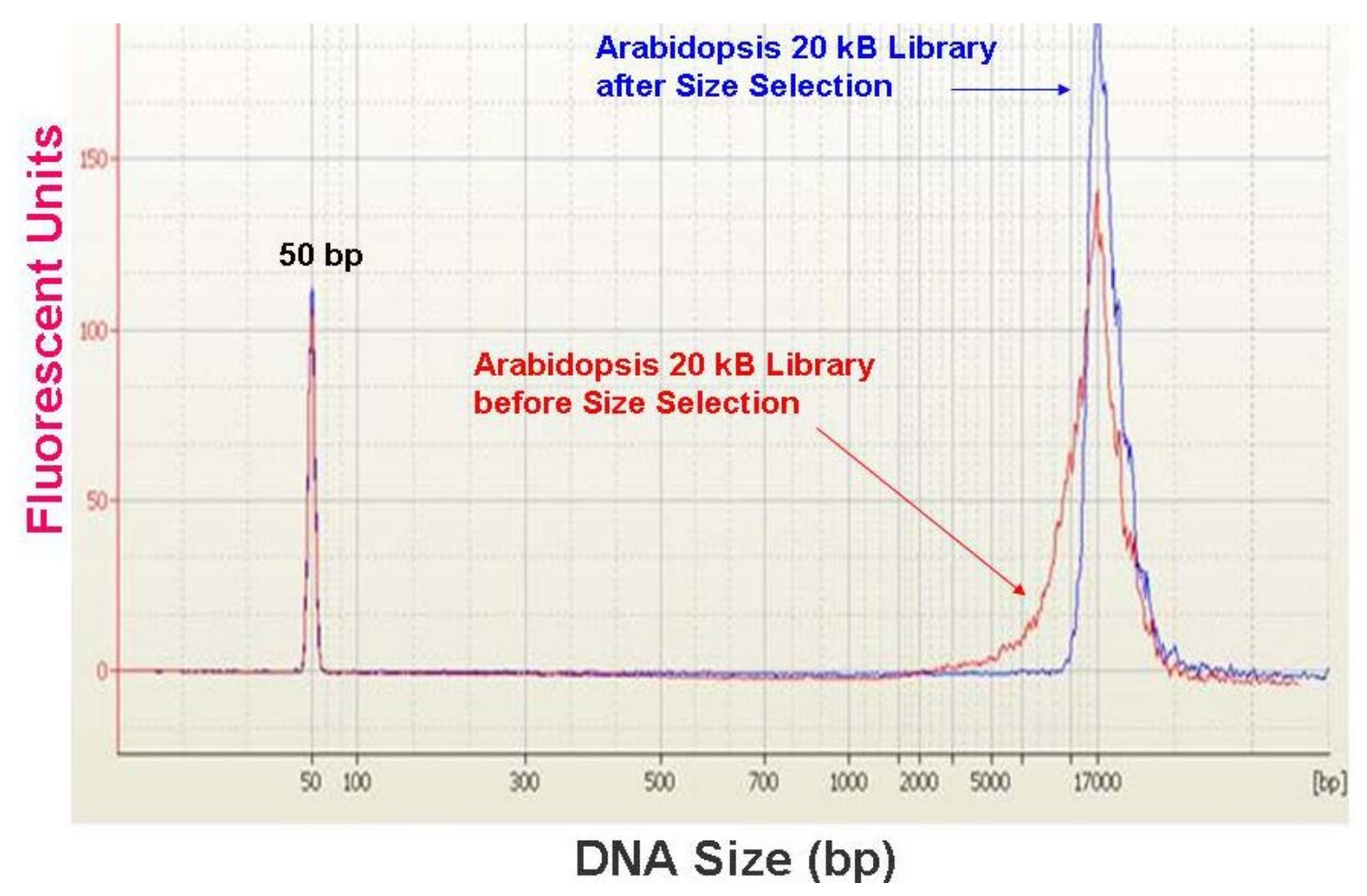


**Figure 4.** Size distribution of SMRTbell library before and after size selection using the BluePippin™ system from Sage Science. The size-selected library (blue) was used for sequencing an *Arabidopsis* genome using only PacBio reads.

## Conclusions

PacBio provides complete solutions for large genome *de novo* sequencing and reference improvement efforts. This is enabled by long-read sequencing to span complex genomic regions, and full-length transcripts for improved gene model annotations that delineate isoforms. The Iso-Seq application also offers a solution to fully characterize transcript isoforms to improve gene-expression quantification that cannot be resolved with short-read technologies (data not shown, Au, et. al PNAS 2013).

Success for both applications is highly dependent on the quality of input libraries. Size selection is highly recommended to eliminate shorter inserts in the SMRTbell library to maximize and capitalize on multi-kilobase reads >20 KB.

## References and Resources

**Genome:**

Chin CS., et. al. “Nonhybrid , finished microbial genome assemblies from long-read SMRT sequencing data.” *Nat Methods*. Jun;10(6):563-9 (2013).

English et al. (2012) Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. *PLoS One*.

Koren S., et. al. (2013) Reducing assembly complexity of microbial genomes with single molecule sequencing. *Genome Biology*, 14:R101

PAG 2013: Michael Schatz, “De novo assembly of complex genomes using single molecule sequencing”

**Transcriptome:**

Au et al. (2013) Characterization of the human ESC transcriptome by hybrid sequencing. *PNAS* doi: 10.1038/pnas.1320101110.

Sharon et al. (2013) A single-molecule long-read survey of the human transcriptome. *Nat Biotechnol* doi: 10.1038/nbt.2705

Tseng, PAG 2014, “ Isoform Sequencing: Unveiling the Complex Landscape of the Eukaryotic Transcriptome on the PacBio® RS II” (Poster).

Acknowledgements

The authors would like to thank everyone who helped generate data for the poster, and the community for their interest in Pacific Biosciences. We also thank Casey Bergman, Susan Celniker, Roger Hoskins, Sergey Koren and Adam M. Phillippy for their contributions to the *D. melanogaster* genome assembly (University of Maryland).