User Group Meeting Agenda – September 7, 2017

Hosted by the Puglisi Laboratory
Stanford University School of Medicine—Paul Berg Hall

THE LEADER IN LONG-READ SEQUENCING

PACBIO®

8:00 – 9:00 a.m. **Registration and Continental Breakfast**

9:00 – 9:10 a.m. **Welcome Remarks**
Jonas Korlach, Ph.D., Chief Scientific Officer, PacBio

9:10 – 9:35 a.m. **Sequel System Updates**
Kevin Corcoran, Senior Vice President, PacBio

9:35 – 10:00 a.m. **Generating Complex High Quality Plant Reference Genomes with PacBio**
Jeremy Schmutz, Faculty Investigator, HudsonAlpha Institute of Biotechnology

The Department of Energy Joint Genome Institute has been producing plant reference genomes for just over eleven years. These genomes form the bedrock on which public plant comparative genomics have flourished. The majority of the Plant Flagship Genomes had been previously generated with Sanger-based paired-end sequencing and followed with laborious targeted finishing efforts. This was followed up with short-read genomes, based on Roche 454, and then Illumina and hybrid genomes. However, the results have always been somewhat less quality and completeness than the genomes that came before. During the previous two years, after the introduction of P6/C4 chemistry on the Pacific Biosciences RS II, we have made substantial efforts to make PacBio data the raw material for the genomes that we produce because of the advantages for assembly of the long-read data. We combine this raw data with Illumina short-read data for error correction and false join detection and then intersect it with mapping, long-range linking, and synteny data to produce high quality chromosome scale assemblies. At HudsonAlpha, we have added PacBio based sequencing to produce genomes of importance for food and fiber, targeting minor crops that are poised to take advantage of genomic accelerated breeding. For this talk, we will present three topics: 1) Upgrading genomes of importance with PacBio, including JGI Flagship Plant Genomes, 2) Sequencing and assembly of complex allotetraploid genomes, and 3) Introducing the results on the Sequel platform of recently produced plant genome references. Along the way, we will share guidelines and our latest best practices for data collection and high quality genome assembly that will be of use to the PacBio user community.

10:00 – 10:25 a.m. **From Hot Springs to Soil - Using Single Molecule Sequencing to Improve Metagenome Assembly**
Alicia Clum, Genome Assembly Group Lead, Lawrence Berkeley National Laboratory

The decrease in sequencing cost in recent years has resulted in an increase in the number of samples and depth to which metagenomes are sequenced. Assemblies from these Illumina datasets, of which JGI produce hundreds of a year, tend to be highly fragmented and incomplete. Challenges with metagenomics in general include quality and quantity of DNA. JGI has generated PacBio data from a variety of community types and complexities. Since nearly all the metagenome samples we've processed are of low DNA quality, the libraries and resulting read lengths are shorter than what the platform is capable of. Despite this limitation, PacBio reads can be longer than assembled contigs from Illumina only data. This talk will discuss how JGI applies PacBio data to metagenomics and current results including improvements in read lengths with the new Sequel platform.

10:25 – 11:10 a.m. **Coffee Break**

11:10 – 11:35 a.m. **The Genetic and Epigenetic Basis for Virulence and Antibiotic Resistance in *Mycobacterium tuberculosis (Mtb)***
Faramarz Valafar, Ph.D., Director and Professor, Biomedical Informatics Research Center, San Diego State University

Tuberculosis (TB) is one of the oldest human diseases with widespread public health burden throughout history. Today, over nine million new TB cases and nearly 1.5 million deaths are reported by the WHO, annually, surpassing AIDS as the infectious disease with highest mortality. While the incidence of TB is slowly coming down after significant global effort, the incidence of drug resistant cases is on the rise. Single Nucleotide Polymorphisms (SNPs) are considered to be the basis for the majority of drug resistant cases. In this talk we will take a broader look at the genetic and epigenetic basis for drug resistance and virulence in Mtb. We will discuss how a narrow perspective of this basis could lead to newly imposed selection for unusual and much more difficult to characterize strains causing difficult to manage outbreaks.

11:35 – 12:00 p.m. **SMRT Sequencing Substantially Improves the Assembly of Leishmania Genomes and Detection of Modified DNA base J.**
Peter J. Myler, Ph.D., Professor & Director of Core Services, Center for Infectious Disease Research

Leishmania is a genus of protozoan parasites that cause several different human (and animal) diseases. Their (~33 Mb) genomes (and those of the related trypanosomatids) have several unique features; most notably polycistronic transcription, absence of transcriptional regulation and presence of a hypermodified nucleobase call base J (β-D-glucopyranosyloxymethyluracil). Because of several factors, assembly of Illumina-based NGS reads usually results in genomes with thousands of contigs, which include collapsed repeats and/or allelic variants, comprising their utility as reference genomes for subsequent phylogenetic and expression analyses. We have used SMRT sequencing of several Leishmania (as well as Crithidia) species to generate reference genomes with most chromosomes having a single contig and the others having only one or two gaps. We report the results obtained using several assembly algorithms that employ different combinations of short and long reads. We have also used SMRT sequencing to identify the exact location of base J within these genomes, as well as exogenously added plasmids, which are maintained as episomes in Leishmania.

12:00 – 1:30 p.m. **Lunch**

1:30 – 1:55 p.m.

Improved Transcriptome Analysis Using Iso-Seq2 Analysis

Elizabeth Tseng, Ph.D., Senior Staff Scientist, Bioinformatics, PacBio

The existing Iso-Seq bioinformatics pipeline, also known as ToFU, has been available through PacBio's official software suite and online through GitHub since 2015. ToFU is distinct from traditional short read transcript assembly algorithms in several ways. First, it does not require a reference genome or annotation, making it suitable for organisms with little or poor reference genomes. Second, it utilizes PacBio's long read length to identify and cluster full-length single molecules. Finally, it uses PacBio's consensus calling software to polish clustered results to high-quality sequences. The output from Iso-Seq are full-length, high-quality isoform sequences that give accurate start and end sites as well as unambiguous connectivity between exons.

I will first talk about how Iso-Seq (ToFU) has been applied for annotating genomes, identifying novel isoforms in disease genes, characterizing cancer fusion genes, and matching against proteomics data. The second part of my talk will focus on improvements to Iso-Seq. The new Iso-Seq2 (ToFU2) pipeline is designed for Sequel-level throughput with improvements in speed, performance, and cross-platform usability. We evaluated ToFU2's performance on Sequel Iso-Seq runs of human UHRR samples with spiked-in synthetic RNA controls and show that it can recover both more human and synthetic isoforms while reducing the number of false positives.

ToFU2 is currently available as an unsupported, developers version at https://github.com/PacificBiosciences/IsoSeq_SA3nUP/, with official release through SMRT Link/SMRT Analysis in the works.

1:55 – 2:20 p.m.

Parkinson's Disease Associated with Pure ATXN10 Repeat Expansion

Birgitt Schuele, MD, Associate Professor, Program Director of Gene Discovery and Stem Cell Modeling, Parkinson's Institute and Clinical Center

2:20 – 3:05 p.m. **Coffee Break**

3:05 – 3:30 p.m.

From Contigs to Chromosomes: High-throughput Crop Reference Genomes

Kevin Fengler, Research Scientist, Data Science and Informatics, DuPont Pioneer

Reference genomes serve as a platform for genetic discovery, gene-editing and pan-genome analysis. However, high-quality reference assemblies for crops plants are not easy to come by. PacBio SMRT sequencing has enabled many plant genomes to readily join the 1 Mb Contig Club. Yet, the accessibility and utility of a contig assembly is maximized when it is elevated to the status of a finished, "nearly perfect" reference genome. How can you upgrade your membership? How can it be done high-throughput?

3:30 – 3:55 p.m.

Solving the Puzzle of the Octoploid Strawberry Genome

Tom Poorten, Ph.D., Postdoctoral Researcher, Knapp Lab, Department of Plant Sciences, University of California, Davis

Garden strawberry (*Fragaria* × *ananassa*), a synthetic allo-octoploid (2n = 8x = 56), has a comparatively small, albeit complex genome (850 Mb), and a peculiar early origin and domestication history. *F.* × *ananassa* originated in European botanical gardens in the early 1700s from spontaneous interspecific hybrids between geographically isolated wild octoploid species imported from the New World. While the earliest cultivars (ca. 1760-1900) were hybrids between *F. virginiana* subsp. *virginiana* and *F. chiloensis* subsp. *chiloensis*, the genomes of present-day cultivars are mosaics of the genomes of four inter-fertile wild octoploid taxa that evolved 0.5-4.0 Mya and have supplied diversity for strawberry breeding. Our laboratory is engaged in genomic-enabled breeding applications that hinge on the development of a high-quality reference genome sequence and other genomic resources that have been lacking in octoploid strawberry. We recently resequenced the genome of *F. vesca* (2n = 2x = 14; 219 Mp), one of the putative diploid progenitors of *F.* × *ananassa*, with an approach involving long-read PacBio Single Molecule, Real-Time (SMRT) Sequencing and optical mapping that yielded an assembly with chromosome-scale contiguity. Here, we describe the assembly of the octoploid genome using long-read DNA sequencing coupled with other approaches. We sequenced a non-inbred individual of *F.* × *ananassa* to 78× depth with PacBio. Using Falcon, we assembled the reads into 6,450 contigs spanning 1.1 Gb with a contig N50 of 465 kb. The current assembly is longer than expected because haplotypes were frequently assembled into separate contigs thereby producing a haploid-diploid assembly. We are investigating approaches to separate haplotypes from homeologs and will present our latest findings. Despite the haploid-diploid conundrum, the current assembly provides a strong foundation for breeding and genetics applications in strawberry.

3:55 – 4:20 p.m.

Exploring Animal Diversity with Genomics

Prof. Daniel Rokhsar, University of California, Berkeley; US DOE Joint Genome Institute; and Okinawa Institute of Science and Technology

Genomics provides a window into the origin and diversification of animals, events that occurred nearly half a billion years ago. To make inferences about early events in animal evolution requires comparison of deeply divergent genomes from all branches of the animal tree. In contrast to the inbred populations of many model organisms, samples from non-model systems must be collected from the wild, and can be characterized by considerable heterozygosity. I will describe how we have used PacBio to begin to tame these wild animal genomes, and discuss some insights that are emerging from chromosome-scale comparisons.

4:20 – 4:45 p.m.

Concluding Remarks

Jonas Korlach, Ph.D., Chief Scientific Officer, PacBio

5:00 – 6:30 p.m. **Cocktail Reception**

Thanks to our Partners:

Advanced Analytical — Automating genomic discovery

Covaris

diagenode — Innovating Epigenetic Solutions

DNAnexus

sage science