User Group Meeting Agenda – June 28, 2017

Hosted by the Genomics Resource Center,
University of Maryland School of Medicine

THE LEADER IN LONG-READ SEQUENCING

PACBIO®

8:00 - 8:55 a.m. **Registration and Continental Breakfast**

9:00 – 9:35 a.m.

Welcome and SMRT Sequencing Updates

Kevin Corcoran, Senior Vice President, PacBio

9:35 – 10:05 a.m.

In Pursuit of Perfect Genome Sequencing

Michael Schatz, Ph.D., Bloomberg Distinguished Associate Professor of Computer Science and Biology, Johns Hopkins University and Cold Spring Harbor Laboratory

Genome sequencing is an essential component to many studies of disease, development, or evolution. The community strives for a perfect result, although it is not always clear what this means or how it could be achieved. In this presentation, I will outline the major reasons for why perfect genome sequencing is important and the key metrics for assessing genome quality, especially Correctness, Contiguity, and Completeness – the "three C's" of genome sequencing. Based on several recent projects in a variety of plant and animal species, I will present a critical analysis of the current capabilities of the major sequencing and mapping approaches, including PacBio, Illumina, Oxford Nanopore, 10X Genomics, and Hi-C-based matepairs. From these results, I will argue for the need for multiple approaches, especially PacBio long read sequencing plus a long-range scaffolding technology for constructing nearly perfect chromosome-scale results.

10:05 – 10:35 a.m.

Detecting Structural Variants in PacBio Reads – Tools and Applications

Aaron Wenger, Ph.D., Staff Scientist, PacBio

Most of the basepairs that differ between two human genomes are in intermediate-sized structural variants (50 bp to 5 kb), which are too small to detect with array comparative genomic hybridization but too large to reliably discover with short-read DNA sequencing. SMRT Sequencing fills this technology gap. SMRT Sequencing detects tens of thousands of structural variants in a human genome, approximately five times the sensitivity of short-read DNA sequencing.

Effective application of SMRT Sequencing to detect structural variants requires quality bioinformatics tools. To discover variants, we have developed pbsv, which is available in version 5.0 of the PacBio SMRT Link software suite. The pbsv algorithm applies a sequence of stages: 1) identify reads with signatures of structural variation, 2) cluster nearby reads with similar signatures, 3) summarize each cluster into a consensus variant, and 4) filter for variants with sufficient read support. For visualization, we have extended the popular genome browser IGV to better support structural variants and PacBio long reads. The improvements are available in IGV 3.0.

To evaluate pbsv, we generated high coverage of a diploid human genome and then titrated to lower coverage levels. The false discovery rate for pbsv is low at all coverage levels. Sensitivity is high even at modest coverage, above 85% at 10-fold and 95% at 20-fold. We also applied pbsv to identify structural variants in an individual with Carney complex for whom short-read whole genome sequencing was non-diagnostic. Filtering for rare, genic structural variants left six candidates, one of which was determined to be likely causative. These applications demonstrate the ability of pbsv to detect structural variants in low-coverage PacBio sequencing and suggest the importance of considering structural variants in any study of human genetic variation.

10:35 - 11:20 a.m. **Coffee Break**

11:20 – 11:50 a.m.

Full-length 16S rRNA Gene Sequencing (and Longer) for Microbiome Analysis

Jethro Johnson, Ph.D., Postdoctoral Associate, The Weinstock Lab, The Jackson Laboratory for Genomic Medicine

Sequencing the 16S rRNA gene is a mainstay of microbiome research, enabling robust quantification of bacterial taxa. However, short-read sequencing has proven unable to produce contiguous sequences long enough to span an entire 16S gene (~1500bp). Consequently current high-throughput approaches target only sub-regions of the 16S gene and are therefore limited in their ability to distinguish between potentially important bacterial taxa. In contrast, long-read sequencing technologies, such as PacBio SMRT Sequencing, for the first time present an opportunity for high-throughput sequencing of the entire 16S gene and its neighboring regions.

Using a combination of *in silico*, *in vitro* and *in vivo* studies we set out to evaluate the theoretical and practical advantages of high-thoughput, full-length 16S gene sequencing for microbiome analysis. *In silico* analysis demonstrated the benefit of capturing the full 16S gene over sub-regions. Sequencing a mock microbial community demonstrated that PacBio CCS reads are accurate enough to detect polymorphisms between

multiple copies of the 16S gene that occur within the same bacterial strain. *In vivo* studies demonstrated that full-length 16S gene sequencing resulted in accurate detection of bacterial taxa at the species-level, as verified by metagenomic whole genome shotgun sequencing.

Having demonstrated practical advantages of sequencing the full 16S gene in high-throughput studies of the microbiome, we are now exploring the benefits of extending our approach to incorporate adjacent internal transcribed spacer regions.

11:50 – 12:20 p.m.

SMRT-cappable-seq Reveals the Complex Operome of Bacteria

Bo Yan, Ph.D. Postdoc, New England Biolabs (NEB)

Gene expression in bacteria is organized into operons, a functional unit of genomic DNA containing a cluster of genes under the control of a single promoter. Operon structures in bacteria are difficult to tackle because the accurate identification of transcripts start and end is challenging. Additionally, short read sequencing cannot establish the connectivity between the transcription start and end. As a result, and despite overwhelming mass of transcriptome information, accurate operon structures have so far been solved only for a handful of cases.

Here we describe SMRT-cappable-seq as the first experimental methodology to identify operon structures genome-wide in prokaryotes. It combines the isolation of full length primary transcripts with SMRT Sequencing. Applied to *E. coli*, SMRT-cappable-seq identifies a total of 2300 operons from which around 900 are novel. Importantly, our result reveals a pervasive read-through of previous experimentally validated transcription termination sites. Termination read-through represents a powerful strategy to control gene expression and establish operon polarity. Taken together this data provides a first glance at the complexity of the 'operome' in bacteria and presents an invaluable resource for understanding gene regulation and function in bacteria.

12:20 – 1:50 p.m.

Lunch

1:50 – 2:20 p.m.

Quantification of CAG Repeat Instability in Huntington's Disease

Ricardo Mouro Pinto, Ph.D., Instructor in Neurology, Center for Genomic Medicine, Massachusetts General Hospital, Harvard Medical School

Huntington's disease (HD) is a neurodegenerative disorder caused by the hyper-expansion of a CAG repeat in the HTT gene. The length of the mutant CAG is inversely correlated with HD onset, with earlier disease onset being associated with longer repeats. Interestingly, the HTT CAG repeat undergoes further tissue-specific somatic expansion, particularly in affected brain regions. This raises the hypothesis that somatic CAG length increases contribute to HD pathogenesis and that factors that modify somatic instability may also modify disease.

The methodology used for the quantification of CAG repeat instability is often dependent on PCR amplification which, due to the potential for very long sequences, their repetitive nature as well as high GC content, constitutes a challenge for PCR to faithfully reproduce the native genomic sequence. This results in a strong bias towards the amplification of shorter repeats and therefore an underrepresentation of expanded alleles post PCR amplification. As far as sequencing-based methodology is concerned, these are also some of the most challenging regions of the genome, particularly due to the limited read length of regular next generation sequencing platforms.

We have developed a novel, amplification-free enrichment technique that employs the CRISPR/Cas9 system for specific targeting of HTT alleles. This method, in conjunction with PacBio's long reads and unbiased coverage, enables sequencing of complex genomic regions such as CAG repeat tracts, providing a better characterization of the repeat size distribution in somatic tissues.

We will present preliminary results obtained by this approach using HD patient postmortem brain samples. Ultimately, this data would allow us to search for genetic modifiers of instability. In addition, this novel methodology will also provide us with a unique characterization of polymorphisms present at the HTT CAG repeat locus itself, including repeat interruptions, allowing us to identify potential cis-modifiers of repeat instability.

2:20 – 2:50 p.m.

Adeno-associated Virus Genome Population Sequencing and Implications for Detecting Undesirable Genomes in Therapeutic Vectors

Phillip Tai, Ph.D., Horae Gene Therapy Center, Department of Microbiology and Physiological Systems, University of Massachusetts Medical School

Recombinant adeno-associated virus (rAAV)-based gene therapy has entered a phase of clinical translation and commercialization for multiple therapeutic applications. Despite this new era for gene therapy, the integrity of packaged vector genomes during rAAV production is often overlooked. It has long been shown that fragmented or truncated wild-type AAV genomes can be packaged to form defective interfering particles. However, the extent and frequency of occurrence during rAAV production is unknown. Encapsidation of partial vector genomes can negatively impact therapeutic efficacy and potential safety. Using single molecule, real-time (SMRT) sequencing, we can comprehensively profile packaged genomes as a single intact molecule and directly assess vector integrity without extensive library preparation steps. We have used this methodology to profile all heterogenic populations of self-complementary AAV genomes, and have coined this approach AAV-genome population sequencing (AAV-GPseq). We show that this method can reveal the relative distribution of truncated genomes versus full-length genomes in vector preparations. We discovered that highly complex

recombination events occur to yield chimeric rAAV genomes containing sequences originating from multiple sources. These particles are detected with frequencies as high as 5% of vector preparations. Most puzzling is the observation that many chimeras that contain host-genomic sequences, originate from gene promoter sequences. These discoveries pose new concerns for how we define the purity and homogeneity of viral vector preparations, and how undesired packaged sequences may impact rAAV-based clinical modalities. Our work introduces unique next generation QC pipelines to the field of gene therapy, and in many respects, responds to a world of rapid clinical development and commercialization for rAAV gene therapy.

2:50 – 3:35 p.m. **Coffee Break**

3:35 – 4:05 p.m.

Using SMRT Iso-Seq Sequencing to Dissect Polyploid Transcriptomes: Lesson Learned from Tetra- and Hexaploid Blueberries

Hamid Ashrafi, Ph.D., Assistant Professor, Department of Horticultural Science, North Carolina State University

Blueberry (*Vaccinium corymbosum*) is one of the most economically important crops in the dicot family of Ericaceae. It is a nutrient dense small fruit that is popular because of its flavor and well known health benefits. There are multiple lines of evidence that show the anthocyanin and other flavonoids in blueberry fruit have beneficial effects against several chronic diseases including, cardiovascular disorders, neurodegenerative diseases, diabetes, and cancer. However, factors affecting the biosynthesis and regulation of the various flavonoids in blueberries, including agronomic and genetic factors, and the possible pathways of biosynthesis of the major anthocyanin have not been fully elucidated partly due to lack of genomic resources. Sequencing the genes responsible for complex phenotypic traits has remained a major objective in modern biology. At NCSU, we have sequenced and assembled a diploid blueberry genome via phased genome assembly using SMRT sequencing. This great resource provided us the opportunity to further explore the genes that make a blueberry, a blueberry. Commercially available blueberry fruits are harvested from tetra- and hexaploid species which are called highbush and rabbiteye blueberries, respectively. We bar-coded cDNA libraries of 9 tissue types of O'Neal (4X) and Premier (6X) cultivars, and sequenced them using 87 RSII-SMRT cells. A total of 141,399 and 110,050 high quality (HQ) full length non-chimeric (FLNC) sequences were generated for O'Neal and Premier, respectively. Visualizing the alignments of iso-seq sequences to the genome sequence revealed a number of redundancies in HQ data but at the same time a number of obvious isoforms. The smaller redundant sequences of the same gene were originated from fragmented and short mRNAs during RNA extraction or cDNA library construction. To make a non-redundant set of isoforms for each cultivar we used CAP3 program to assemble the reads that were categorized as FLNC but not resulted from full length mRNA. In total the 141K and 110K HQ sequences of O'Neal and Premier reduced to 56,564 (48,127 singlets and 8437 contigs) and 46,280 (36,866 singlets and 9414 contigs), respectively. Despite 4X and to some extent 6X blueberries are believed to be autotetraploid, but a number of SNPs were observed between and among transcriptome sequences of the same gene, suggesting current blueberry cultivars may not be truly autopolyploids.

4:05 – 4:35 p.m.

SQANTI: Extensive Characterization of Long-read Transcript Sequences for Quality Control in Full-length Transcriptome Identification and Quantification

Manuel Tardaguila, Ph.D., Department of Microbiology and Cell Science, University of Florida

High-throughput sequencing of full-length transcripts using long reads has paved the way for the discovery of thousands of novel transcripts, even in very well annotated organisms as mice and humans. Nonetheless, there is a need for studies and tools that characterize these novel isoforms. Here we present SQANTI, an automated pipeline for the classification of long-read transcripts that computes over 30 descriptors, which can be used to assess the quality of the data and of the preprocessing pipelines. We applied SQANTI to a neuronal mouse transcriptome using PacBio long reads and illustrate how the tool is effective in readily describing the composition of and characterizing the full-length transcriptome. We perform extensive evaluation of ToFU PacBio transcripts by PCR to reveal that an important number of the novel transcripts are technical artifacts of the sequencing approach, and that SQANTI quality descriptors can be used to engineer a filtering strategy to remove them. A comparison of Iso-Seq over the classical RNA-seq approaches solely based on short-reads demonstrates that the PacBio transcriptome not only succeeds in capturing the most robustly expressed fraction of transcripts, but also avoids quantification errors caused by unaccounted 3' end variability in the reference. SQANTI allows the user to maximize the analytical outcome of long read technologies by providing the tools to deliver quality-evaluated and curated full-length transcriptomes. SQANTI is available at https://bitbucket.org/ConesaLab/sqanti.

4:35 – 4:45 p.m. **Closing Remarks**

Kevin Corcoran, Senior Vice President, PacBio

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

Thanks to our Partners:

University of Maryland School of Medicine Institute for Genome Sciences

Advanced Analytical — Automating genomic discovery

diagenode — Innovating Epigenetic Solutions

sage science