PACIFIC BIOSCIENCES™

Whitepaper
SMRT™ DNA Sequencing
November 2009

Beyond Next Generation DNA Sequencing
Single Molecule Real Time (SMRT™) Technology

OVERVIEW

Pacific Biosciences Single Molecule Real Time (SMRT) DNA sequencing technology enables, for the first time, observation of natural DNA synthesis by a DNA polymerase as it occurs. The approach is based on "eavesdropping" on a single DNA polymerase molecule working in a continuous, processive manner. Distinguished by its long reads, fast time to result, more informative data, and lower overall costs, SMRT DNA sequencing promises to be a transformative technology that will enable a new paradigm in genomic analysis.

INTRODUCTION

DNA sequencing has undergone several evolutions in the past decade. The most widely used DNA sequencing platform – the "1st generation" Sanger technology used in the Human Genome Project, has been supplemented by "2nd generation" systems promising higher throughput at reduced costs.

While 2nd generation technologies provided enormous improvements in throughput and dramatically lowered the cost per sequenced base, they are reaching their twilight in potential future performance enhancements.

Second generation technologies can be thought of as 'brute-force' systems that deliver high throughput at the expense of readlength and speed. In addition, several other inherent aspects of 2nd generation systems limit their utility. For example, because they are designed for high volume runs, they require customers to wait until enough samples have been collected to run at capacity, and often require complicated molecular barcoding methods to allow for maximum productivity. This increases costs significantly for smaller sized projects and severely reduces the flexibility of these systems.

Further, because of their short reads, 2nd generation technologies are not capable of addressing all of the relevant types of variation in the genome. Rare genetic variants, complex structural rearrangements, and other sources of variation (such as differentially methylated DNA sites) have recently been proven to play a larger role in explaining disease risk and progression, and may be more medically important than SNPs. Ultimately, it is necessary to look comprehensively across all types of variation to understand the fundamental complexity of the genome.

Currently, organizations are using a combination of 1st and 2nd generation approaches depending on the application. For example, 2nd generation systems are used primarily for resequencing and counting or tagging applications. Sanger sequencing continues to be the platform of choice for *de novo* sequencing, validation studies, and projects requiring a fast time to result (such as infectious disease monitoring and molecular diagnostics).

What is required is a breakthrough technology capable of offering a new performance envelope with improvements across applications and the ability to ultimately drive down the cost and time required for human genome sequencing to make it feasible for personalized medicine.

At a cost of more than $250 million, PacBio has developed that breakthrough technology. Because of the dramatic changes in all aspects of its usability and performance, the technique is now commonly referred to as "3rd Generation Sequencing."

THIRD GENERATION TECHNOLOGY

Third generation technologies are differentiated by **single molecule resolution**, **very long reads**, **fast time to results**, and **lower overall cost**, including the **flexibility** to cost-effectively perform both small and large projects.

PacBio's Single Molecule Real Time (SMRT) System is a 3rd generation DNA sequencing technology that enables a much wider range of applications when compared to 2nd generation technologies. Enzyme processivity enables much longer readlength while the speed of synthesis drives fast time to results. In addition, by monitoring the enzyme in real time, SMRT sequencing provides richer data, including kinetic information.

Together, these capabilities open new opportunities for disease research, including infectious disease studies, detection of rare variants, understanding the genomic complexity of cancer, and conducting epigenetic studies. Real-time detection is also critical to quickly and efficiently identifying and subtyping pathogens.

SMRT technology eliminates the current bottlenecks inherent in 2nd generation technologies by using DNA polymerase as a real-time sequencing engine. By observing the natural process of DNA

synthesis in real-time without interruption, the system harnesses the power of the DNA polymerase, thereby capitalizing on the performance increases derived from millions of years of natural evolution. In order to enable "eavesdropping" on DNA synthesis as it occurs, PacBio developed three key innovations that overcame the challenges faced in previous attempts to conduct real-time single molecule sequencing:

1) The SMRT Cell, which enables single molecule, real-time observation of individual fluorophores against a dense background of labeled nucleotides while maintaining a high signal-to-noise ratio,

2) Phospholinked nucleotides, which enable long readlengths by producing a completely natural DNA strand through fast, accurate, and processive DNA synthesis, and

3) A novel detection platform that enables single molecule, real-time detection as well as flexibility in run configurations and applications.

A NEW PERFORMANCE ENVELOPE

The SMRT System offers a completely new performance envelope with dramatically improved economics that will continue to expand over time. With readlengths as long as, *or longer*, than traditional Sanger reads and very fast cycle times, constraints associated with the short reads of 2nd generation technologies are not an issue. Furthermore, because much of the performance is delivered by the polymerase enzyme, significant performance increases are seamless upgrades to reagent kits and software, without requiring changes in hardware.

The wet chemistry and the hardware can be independently enhanced because the SMRT Cell forms a literal barrier between the two. This provides for both simple upgrades to leverage future performance enhancements and the ability to customize and implement multiple application-specific assays on the same instrument. For instance, since readlength is a function of the enzyme, it will continue to increase with time as improvements are made to PacBio's proprietary polymerase.

Further, at initial commercial release, sequencing speed will be 1-3 bases/second. However, synthesis will become faster as it is driven by the enzyme turnover rate and the hardware is designed to support a sequencing speed of up to 10-15 bases/second. That represents a minimum of a 5-fold throughput increase with no changes in hardware. (*See Performance Trajectory, Page 4)*.

Other parameters such as yield and density are also expected to increase, resulting in even higher performance. Headroom has been built into the hardware in anticipation of these expected performance improvements. For example, the system has been designed to monitor all of the ZMWs continuously in real time regardless of whether they are "productive" (loaded with one and only one polymerase). Over time, with loading strategies that improve the number of productive ZMWs from the initial 30% to much higher yields, up to 3-fold performance gains will be possible, again with no changes to the hardware system required. (*For more information about the ZMW technology, see "SMRT ZMWs" on page 4)*.

The first commercial version of the detection platform has significant multiplex with the ability to monitor 80,000 ZMWs simultaneously. In addition, the platform will have the capability to successively look at multiple sets of 80,000 ZMWs on the same SMRT Cell, further increasing the throughput per Cell. At initial commercial launch, the system will be able to look at two sets of 80,000 ZMWs. Over time the number of sets of ZMWs per Cell will increase. This ability to successively increase the amount of data captured per SMRT Cell directly reduces the cost per base, while increasing the sequence output and number of reads possible.

The combination of the significant headroom embedded in the initial instrument design to support downstream chemistry changes, and the many performance increases possible with a simple reagent kit upgrade, means that the useful life of the PacBio system is expected to be much longer than any of the current, relatively static systems. Longer useful life translates to dramatically lower cost of ownership.

LONG READLENGTHS

Long reads are the only way to obtain a comprehensive view of genomes, as they can reveal all types of genetic variation - from SNPs to segmental duplications. They provide access to novel, medically relevant features found in repeat regions of the genome. As a result, longer readlengths offer the potential to generate far more clinically useful models that could then be more routinely applied in the clinical and consumer genomics setting.

Moreover, long readlengths allow insights into biology that are not possible with 2nd generation technology, such as the ability to map complex rearrangements, identify haplotypes, non-coding RNAs, different isoforms of genes, and allelic imbalances between those isoforms – all of which play a role in predisposition to disease or drug response.

At launch, the SMRT System is expected to routinely generate readlengths at or greater than what is achieved with Sanger sequencing, with some reads going well beyond the limits of Sanger sequencing.

FAST TIME TO RESULT

Using DNA polymerase as a real-time sequencing machine, SMRT sequencing offers unprecedented sequencing speed. At initial commercial release, sequencing will be 20,000 times faster than the most popular 2nd generation technology on a per nucleotide basis. To monitor base incorporation at these speeds, PacBio developed a detection technology that is 1,000 times more sensitive than existing microscopes. This enables the SMRT System to discriminate signals against background noise when reading the individual bases of DNA as close as possible to the speed in which they are naturally synthesized.

At these speeds, run times are very short. A typical run can be as short as 15 minutes. These fast run times, combined with simple and streamlined sample preparation and concurrent base calling, translate to very fast time to results: sample prep to sequencing results in < 1 day. In contrast, 2nd generation systems can take as much as 12 days for the equivalent process.

FLEXIBLE SEQUENCING MODES

The SMRT System offers the flexibility to conduct sequencing projects in different modes depending on the need. Short run times and a flexible consumables format support both small and large projects cost effectively. Furthermore, within each protocol, there are user-adjustable parameters such as collection time, for additional tuning.

Standard Sequencing

The standard SMRT sequencing protocol is designed to generate single pass long reads. The protocol uses long insert lengths so that the polymerase can continuously and processively synthesize along a single strand. As with all protocols, this process is parallelized across thousands of ZMWs in a single SMRT Cell at the same time. This protocol has utility for a range of both resequencing and *de novo* applications.

Circular Consensus Sequencing

PacBio's innovative SMRTbell™ sample preparation creates a circularized template which, due to the system's long readlengths and strand displacing enzyme, can be read multiple times to achieve unprecedented accuracy on a single molecule. Furthermore, this approach provides reads on both the forward and reverse strands. This type of accuracy at a single molecule level is novel, as it is fundamentally impossible to achieve with ensemble-based approaches. This method also offers advantages for the discovery and confirmation of rare variants.

Strobe Sequencing

Because the chief source of read termination is currently due to polymerase damage from laser illumination, the physical coverage and effective readlength of the system can be increased by "strobing" the illumination on and off. When the illumination is on, sequence data is collected, but when the illumination is off, the polymerase continues to polymerize in the dark at a predictable speed for thousands of bases. After a user defined interval, the illumination can be turned back on and the system can resume collecting data.

The result is analogous to paired-end or mate-pair methods, but is more flexible. Multiple sub-reads at varying sequence advance lengths can be generated from a single molecule. The length of the strobe sub-reads and advances can be controlled dynamically as a run parameter, thereby eliminating the need to create multiple libraries of different sizes. This method will be extremely useful for scaffolding and for identifying and resolving structural variation.

Combining Sequencing Modes

The SMRT System's flexibility also offers the ability to approach a problem in multiple phases. For example, users may first want to scaffold the sequence (using the strobe sequencing protocol to generate physical readlengths of thousands of base pairs), then generate long linear single molecule reads, and finally, apply circular consensus sequencing to achieve unprecedented accuracy rates.

GRANULARITY

Another unique benefit of the SMRT System is the ability to scale the throughput and cost of sequencing to the needs of the experiment across a range of small and large projects. We call this granularity, and it results in part from our high speed chemistry coupled with a flexible consumables format.

A single experiment can be run for as little as $99 for the SMRT Cell and sequencing reagents, giving users a very low cost entry barrier to experimentation. The ability to run a single SMRT Cell in as little as 15 minutes, or a batch of up to 96 Cells in a single job for up to 12 hours without operator intervention, provides enormous flexibility in

experimental design and implementation. This flexibility in project investment is unique compared to the long, multiday run times and resultant, single, massive data output of 2nd generation systems.

Over time, the SMRT Cell and reagent cost is expected to remain relatively fixed, while users will continue to experience dramatic increases in the amount of sequencing data as upgrades and enhancements are implemented (see Figure 1).

EASE OF USE

Ease of use is a key feature of the SMRT sequencing system. The sample prep protocol is simple and fast, and can be completed in less than a day. When sufficient template material is available, amplification is not required. The same protocol can be used with a variety of input types, including purified genomic DNA, BACs, cDNA libraries or PCR products, and can generate a variety of library sizes. The instrument is designed with a simple touchscreen interface and requires minimal user intervention. As the instrument is sequencing, signal processing, base-calling and quality assessment is performed in real-time to provide immediate feedback as well as reduce time to result and IT requirements. Alignments and assemblies are simplified as the software enables an automated pipeline starting with data automatically streaming from the instrument as sequential SMRT Cells are processed.

CONCLUSION

PacBio has developed a transformative technology platform for real-time detection of biological events at single molecule resolution. The first commercial application for this platform is DNA sequencing (available in the second half of 2010).

We believe our 3rd generation sequencing system offers an entirely new performance envelope and a new economic paradigm that will expand the market for DNA analysis by enabling new applications beyond what has been possible with 2nd generation technology.

PacBio has begun expanding internal research programs and developing collaborations for additional ‘SMRT Biology’ applications, such as simpler and more direct solutions for RNA sequencing and methylation sequencing. This will allow scientists to acquire new, fundamental knowledge about the molecular dynamics of life.

SIDEBAR: SMRT ZMWs

A ZMW is a hole, tens of nanometers in diameter, fabricated in a 100 nm metal film deposited on a silicon dioxide substrate. Each ZMW becomes a nanophotonic visualization chamber providing a detection volume of just 20 zeptoliters (10-21 liters). At this volume, the technology detects the activity of a single molecule among a background of thousands of labeled nucleotides.

DNA polymerase molecules are attached to the bottom surface such that they permanently reside within the detection volume. Phospholinked nucleotides, each type labeled with a different colored fluorophore, are then introduced into the reaction solution at high concentrations that promote enzyme speed, accuracy, and processivity.

Through directed attachment strategies, over time, the number of ZMWs with a single active polymerase can be increased, delivering higher and higher yields. When DNA polymerase incorporates complementary nucleotides, the enzyme holds each nucleotide within the detection volume for tens of milliseconds—orders of magnitude longer than the amount of time it takes a nucleotide to diffuse in and out of the detection volume.

During this time, the engaged fluorophore emits fluorescent light whose color corresponds to the base identity. Then, as part of the natural incorporation cycle, the polymerase cleaves the bond that previously held the fluorophore in place and the dye diffuses out of the detection volume. Following incorporation, the signal immediately returns to baseline and the process repeats.

Unhampered and uninterrupted, the DNA polymerase continues incorporating multiple bases per second. In this way, the SMRT approach produces a completely natural long chain of DNA in minutes. Simultaneous and continuous excitation and detection occurs across all of the thousands of ZMWs in the SMRT Cell in real time.

PACIFIC BIOSCIENCES™

Whitepaper
SMRT™ DNA Sequencing
November 2009

SMRT Performance Trajectory

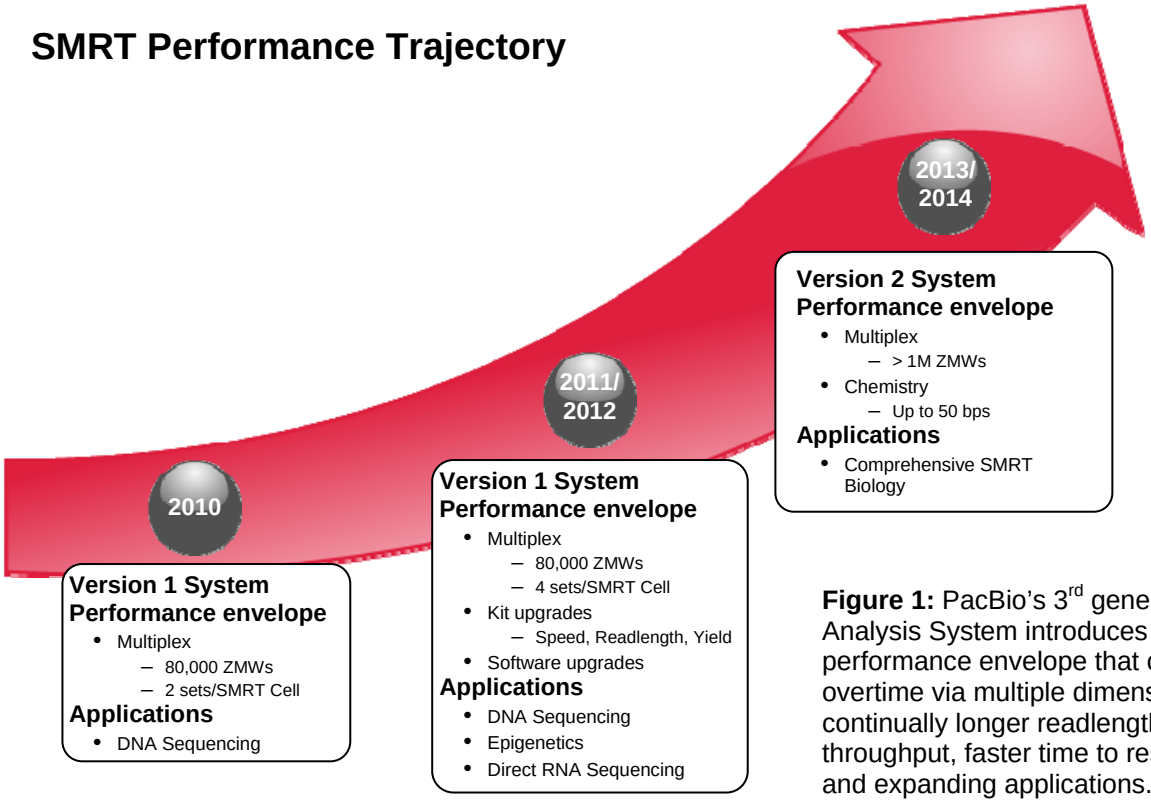

Figure 1: PacBio's 3rd generation DNA Analysis System introduces an entirely new performance envelope that can be increased overtime via multiple dimensions leading to continually longer readlength, higher throughput, faster time to result, lower costs and expanding applications.

SMRT Sequencing System Preliminary Specifications[1]

System Architecture
- Architecture to monitor 80,000 sets ZMWs at a time
- Initial commercial system will allow monitoring of 2 sets of ZMWs per SMRT Cell

Adjustable Run Parameters:
- Maximize readlength or throughput
- Multiple sequencing modes
- Adjust collection time (5 to 20 minutes)
- Run small or large projects (one SMRT Cell or multiple SMRT Cells)

Readlength
- At or greater than Sanger by introduction
- Strobe sequencing protocol will provide the ability to distribute reads across much longer inserts

Speed
- Polymerase speed will average 1-3 bases/second
- Minimum run time as short as 15 minutes for a single SMRT Cell, after initial setup

Variety of Template Types
- Linear or circular
- Genomic DNA, cDNA, and PCR products

SMRTbell DNA Sample Prep
- Read forward and reverse strands
- Get multiple sub-reads to generate circular consensus
- Can use a wide range of insert sizes
- Amplification not required (unless additional template required) or no routine amplification

Workflow
- Library preparation to sequencing results in <8 hours, with <4 hours hands-on time
- SMRT Cells can be batched for up to12 hours of unattended operation

Data Output Format & Size
- HD5F
- <100 gigabytes/24 hours for the standard pipeline

Data Conversion Tools
- FASTA, FASTQ, SRA, SRF

[1] PacBio's SMRT Sequencing System is currently in the development phase, and the specifications and performance of our commercial systems have not been finalized. The preliminary specifications provided herein are for information purposes only and do not constitute any commitment, promise, or other representation as to the final performance and specifications of those commercial systems.

REFERENCES

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., deWinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., and Turner, S. 2009. **Real-time dna sequencing from single polymerase molecules.** *Science* 323:133-138.

Korlach, J., Bibillo, A., Wegener, J., Peluso, P., Pham, T., Park, I., Clark, S., Otto, G., Turner, S. 2008. **Long, processive enzymatic dna synthesis using 100% dye-labeled terminal phosphate-linked nucleotides.** *Nucleosides, Nucleotides & Nucleic Acids* 27:1072-1082.

Lundquist, P. M., Zhong, C. F., Zhao, P., Tomaney, A. B., Peluso, P. S., Dixon, J., Bettman, B., Lacroix, Y., Kwo, D. P., Mccullough, E., Maxham, M., Hester, K., Mcnitt, P., Grey, D. M., Henriquez, C., Foquet, M., Turner, S. W., and Zaccarin, D. 2008. **Parallel confocal detection of single molecules in real time.** *Optics Letters* 33:1026-1028.

Korlach, J., Marks, P. J., Cicero, R. L., Gray, J. J., Murphy, D. L., Roitman, D. B., Pham, T. T., Otto, G. A., Foquet, M., and Turner, S. W. 2008. **Selective aluminum passivation for targeted immobilization of single dna polymerase molecules in zero-mode waveguide nanostructures.** *Proc Natl Acad Sci USA* 105:1176-1181.

Foquet, M., Samiee, K.T., Kong, X., Chaudhuri, B.P., Lundquist, P.M., Turner, S.W., Freudenthal, J., Roitman, D.B. 2008. **Improved fabrication of zero-mode waveguides for single-molecule detection.** *Journal of Applied Physics* 103:034301.

Levene, M. J., Korlach, J., Turner, S. W., Foquet, M., Craighead, H. G., and Webb, W. W. 2003. **Zero-mode waveguides for single-molecule analysis at high concentrations.** *Science* 299:682-686.

PACIFIC BIOSCIENCES
1505 Adams Drive, Menlo Park, California 94025

http://www.pacificbiosciences.com
General Inquiries:
info@pacificbiosciences.com
(650) 521-8000