Pacific Biosciences Develops Transformative DNA Sequencing Technology

- Single Molecule Real Time (SMRT™) DNA Sequencing
- Long Reads, Short Run Time, and High Quality Sequence Data at Lower Cost

OVERVIEW

Pacific Biosciences was funded in 2004 with the goal of developing Single Molecule Real Time (SMRT) DNA sequencing technology. This technology enables for the first time, the observation of natural DNA synthesis by a DNA polymerase as it occurs.[1] Pacific Biosciences' (PacBio's) approach is based on eavesdropping on a single DNA polymerase molecule working in a continuous, processive manner. Distinguished by its long reads, short cycle time, and high quality sequence data with less effort and cost, SMRT™ DNA sequencing promises to be a transformative technology that will enable a new paradigm in genomic analysis.

PacBio's SMRT technology is built upon two key innovations that overcome major challenges facing the field of DNA sequencing:

- The SMRT chip, which enables observation of individual fluorophores against a dense background of labeled nucleotides by maintaining a high signal-to-noise ratio
- Phospholinked nucleotides, which produce a completely natural DNA strand through fast, accurate, and processive DNA synthesis

SMRT Technology At-a-Glance

DNA sequencing is performed on SMRT chips, each containing thousands of zero-mode waveguides (ZMWs). Utilizing the latest geometries available in semiconductor manufacturing, a ZMW is a hole, tens of nanometers in diameter, fabricated in a 100nm 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 activity of a single molecule can be detected amongst a background of thousands of labeled nucleotides.

The ZMW provides a window for watching DNA polymerase as it performs sequencing by synthesis. Within each chamber, a single DNA polymerase molecule is attached to the bottom surface such that it permanently resides within the detection volume. Phospholinked nucleotides, each type labeled with a different colored fluorophore, are then introduced into the reaction solution at high concentrations which promote

enzyme speed, accuracy, and processivity. Due to the small size of the ZMW, even at these high, biologically relevant concentrations, the detection volume is occupied by nucleotides only a small fraction of the time. In addition, visits to the detection volume are fast, lasting only a few microseconds, due to the very small distance that diffusion has to carry the nucleotides. The result is a very low background.

As the DNA polymerase incorporates complementary nucleotides, each base is held 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 holding 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 bases at a speed of tens per second. In this way, a completely natural long chain of DNA is produced in minutes. Simultaneous and continuous detection occurs across all of the thousands of ZMWs on the SMRT chip in real time. Researchers at PacBio have demonstrated this approach has the capability to produce reads thousands of nucleotides in length.[4]

Looking Ahead

SMRT DNA sequencing technology will offer a completely new performance envelope with its combination of long readlength, short cycle time, low cost, and high quality sequence data. Furthermore, the technology has been designed with massive scalability to enable ongoing advancement in performance and capabilities over time. Because the SMRT chip forms a literal barrier between the wet and dry system components, upgrades can be implemented seamlessly and independently on either set of components. For example, by enhancing polymerase characteristics, readlength and speed will continue to improve without requiring hardware changes. Similarly, as detector technology advances in the future per Moore's Law, increases in throughput will be possible without changes to the assay.

SMRT technology is expected to have advantages for many applications. Specifically, long reads provide simplified and improved sequence assembly with less fold coverage, and enable characterization of structural variation. Numerous applications of whole genome sequencing have been envisioned by the community that will for the first time become feasible. In addition, increasing sequence performance by orders of magnitude will lead to future uses of sequencing that today are inconceivable, analogous to what has taken place with digital computers.

PACIFIC BIOSCIENCES SMRT™ TECHNOLOGY

Creating the Observation Window

The SMRT Chip

Exploiting DNA polymerase as a sequencing engine requires single molecule detection. DNA polymerization is a stochastic process, where intervals between incorporation events typically vary. Thus a population of polymerases even acting on the same template would quickly become out of phase with each other. Existing single molecule detection techniques are limited to low nanomolar concentrations, in order to reduce the background fluorescence of other nucleotides in solution. At higher concentrations, the detection volumes of these microscope systems are flooded with hundreds or thousands of labeled molecules **(Figure 1)**. This creates a high background noise level, against which it is not possible to detect individual fluorophores. However, polymerases require this high concentration level, without which the speed, accuracy and processivity of the enzyme all suffer.

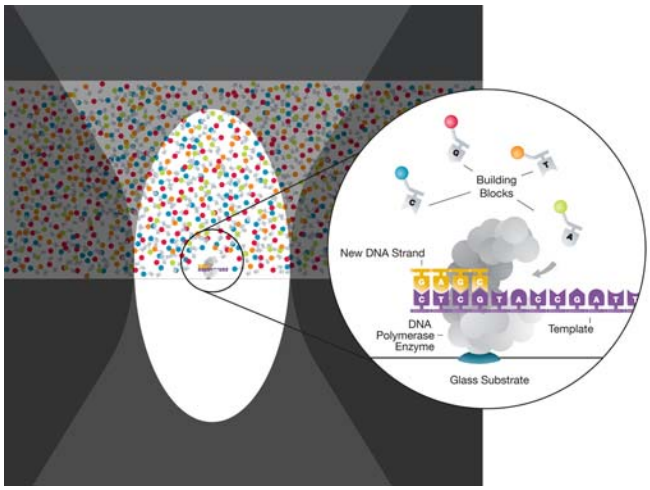


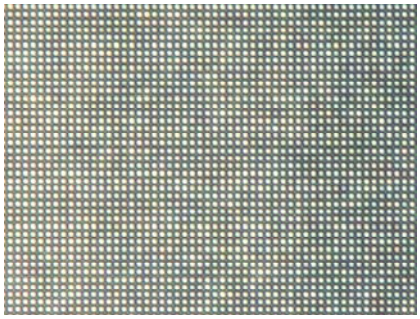
Figure 1. Problem of background interference.

For proper functioning, DNA polymerase requires a high concentration of labeled nucleotides, which creates a fluorescent background thousands of times brighter than the signal of a single incorporation event.

Sequencing approaches that circumvent this problem by step-wise addition of base-labeled nucleotides followed by washing, scanning and removal of the label, severely limit the capabilities of the polymerase, making it significantly slower and drastically reducing readlength. The

need for washing between bases also dramatically increases the consumption of reagents.

PacBio has solved this problem with the SMRT™ chip, which contains thousands of zero-mode waveguides (ZMWs) **(Figure 2)**. The ZMW provides the world's smallest detection volume, representing a 1000-fold improvement over existing single-molecule detection technology. Because the detection volume is so dramatically reduced, a single incorporation event can be observed against the background created by the high concentration of fluorescently labeled nucleotides. It makes possible the real-time observation of a single molecule of DNA polymerase as it synthesizes DNA.



Scale: 43.5 µm wide x 32.8 µm tall

Figure 2. SMRT™ chip

Each SMRT™ chip contains thousands of ZMWs

The ZMW operates using the same principle as the metallic screen of a microwave oven door. The screen is perforated with holes, much smaller than the wavelength of the electromagnetic waves. Because of their relative size, the holes prevent the microwaves from passing through. However, smaller wavelength visible light is able to pass through, allowing food to be visible.

Shrinking this idea to the nanoscale, the ZMW is a cylindrical hole just a few tens of nanometers in diameter, perforating a thin metal film supported by a transparent substrate **(Figure 3)**. When the ZMW is illuminated through the transparent substrate by laser light, the wavelength of the light is too large to pass through the waveguide's aperture.

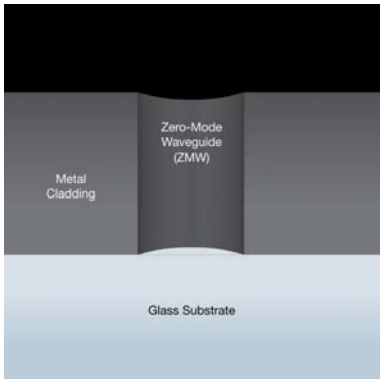


Figure 3. Individual ZMW

Each ZMW is a cylindrical hole tens of nanometers in diameter, perforating a thin metal film supported by a transparent

However, the light does not stop precisely at the waveguide entrance. Attenuated light from the excitation beam penetrates the lower 20-30nm of each waveguide, creating a detection volume of only 20 zeptoliters (10^{-21} liters) **(Figure 4)**. This dramatic reduction in the detection volume provides the needed 1000-fold improvement in rejection of background fluorescence.

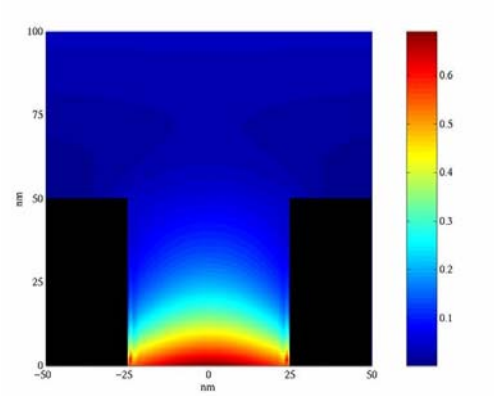


Figure 4. Detection volume

Attenuated light from the excitation beam penetrates only the lower 20-30 nm of each waveguide, creating a detection volume of 20 zeptoliters (10^{-21} liters).

A single DNA polymerase molecule is attached to the bottom of each waveguide. Biased immobilization of the enzyme is made possible through proprietary surface coatings that direct the protein attachment to the transparent floor of the ZMW **(Figure 5)** [3].

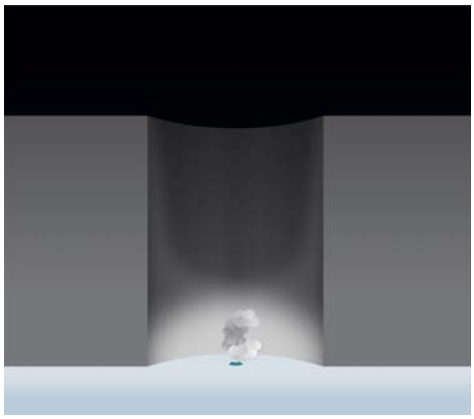


Figure 5. *ZMW with DNA polymerase*

A single DNA polymerase molecule is attached to the bottom of the ZMW using a proprietary biased immobilization process.

In this way, each ZMW provides a window that enables the real-time observation of a single molecule of DNA polymerase as it synthesizes DNA, with the ability to detect a single incorporation event against the background of fluorescently labeled nucleotides at biologically relevant concentrations.

2. SMRT™ Synthesis of Long DNA

Phospholinked Nucleotides

With an active polymerase immobilized at the bottom of each ZMW, phospholinked nucleotides are introduced into the ZMW chamber **(Figure 6).** In order to detect incorporation events and discriminate base identity, each of the four nucleotides A, C, G and T are labeled with a different colored fluorophore.

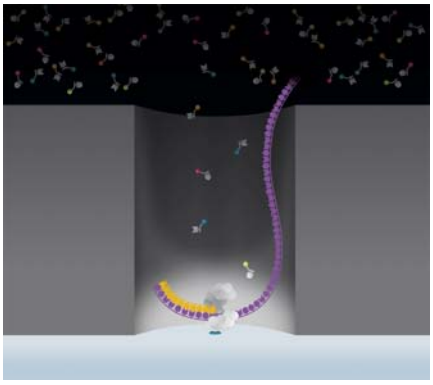


Figure 6. ZMW with DNA polymerase and phospholinked nucleotides

Phospholinked nucleotides are added into the ZMW at the high concentrations required for proper enzyme functioning.

Most sequencing-by-synthesis approaches utilize nucleotides with fluorophores attached directly to the base **(Figure 7)**. With this labeling approach, as each nucleotide is incorporated, its fluorophore becomes a permanent part of the DNA strand.

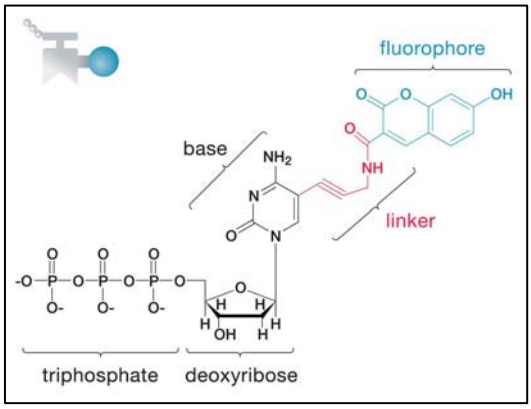


Figure 7. Base-labeled nucleotide.
Base-labeled nucleotides have fluorophores chemically attached directly to the base.

It is not possible to achieve processive synthesis with base-labeled nucleotides because if multiple bases are incorporated, the physical bulk of multiple dye molecules would create steric hindrance, limiting enzyme activity **(Figure 8)**.

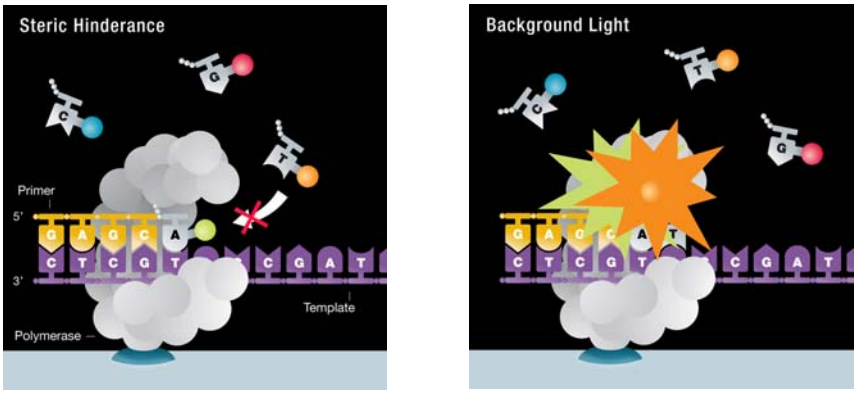


Figure 8. Issues caused by base-labeled nucleotides.

Incorporation of base-labeled nucleotides into a growing DNA chain creates: 1) steric hindrance that inhibits enzymatic activity and, 2) contributes to an increase in background.

To work around this issue, most sequencing approaches that use base-labeled nucleotides synthesize DNA a single nucleotide at a time, starting and stopping the reaction after each incorporation. The constant disruption of the reaction is time consuming, requires high reagent volumes, and severely limits the processivity of the polymerase.

In contrast, PacBio's SMRT™ sequencing is a real-time approach that uses alternatively labeled phospholinked nucleotides [4]. With this strategy, the fluorescent dye is attached to the phosphate chain of the nucleotide rather than to the base (**Figure 9**). As a natural step in the synthesis process, the phosphate chain is cleaved when the nucleotide is incorporated into the DNA strand. Thus, upon incorporation of a phospholinked nucleotide, the DNA polymerase naturally cleaves the dye molecule from the nucleotide when it cleaves the

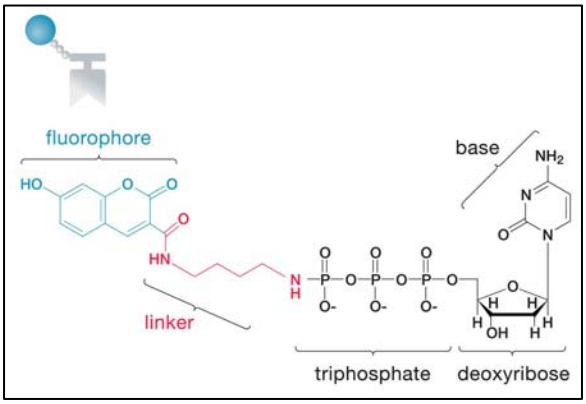


Figure 9. Phospholinked nucleotides

Phospholinked nucleotides have fluorophores attached to the triphosphate chain, which is naturally cleaved when the nucleotide is incorporated.

phosphate chain. The phosphate chain-dye complex quickly diffuses the short distance out of the detection volume, ensuring the background signal remains at the same low level.

Sequencing-by-Synthesis

Phospholinked nucleotides enable the polymerase to synthesize DNA in a fast and processive manner. When the DNA polymerase encounters the nucleotide complementary to the next base in the template, it is incorporated into the growing DNA chain. During incorporation, the enzyme holds the nucleotide in the ZMW's detection volume for tens of milliseconds, orders of magnitude longer than the average diffusing nucleotide. The system detects this as a flash of bright light because the background is very low. The polymerase advances to the next base and the process continues to repeat **(Figure 10)**.

Figure 10. Processive Synthesis with Phospholinked Nucleotides.

Step 1: Fluorescent phospholinked labeled nucleotides are introduced into the ZMW.
Step 2: The base being incorporated is held in the detection volume for tens of milliseconds, producing a bright flash of light.
Step 3: The phosphate chain is cleaved, releasing the attached dye molecule.
Step 4-5: The process repeats.

With the use of phospholinked nucleotides, a long, natural strand of DNA is produced **(Figure 11)**. This processive synthesis is extremely efficient, consuming only one molecule per base sequenced in reagents. Researchers at PacBio have demonstrated this approach has the capability to produce reads tens of thousands of nucleotides in length[4].

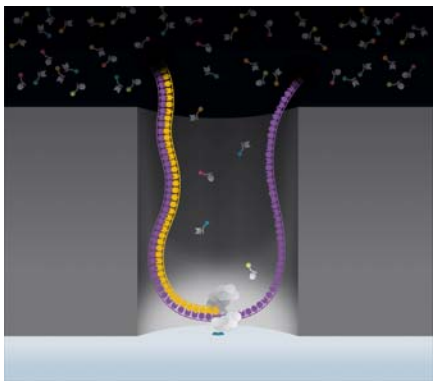


Figure 11. Synthesis of long DNA.

DNA polymerase processively incorporates nucleotides producing long, natural DNA.

Real-Time Detection

Processive synthesis is observed as it occurs in real-time across thousands of ZMWs simultaneously. To accomplish this, PacBio has developed several improvements to the state-of-the-art in single molecule detection and discrimination. In the PacBio instrument, both excitation and detection occur simultaneously through the transparent substrate of the SMRT™ chip **(Figure 12)**.

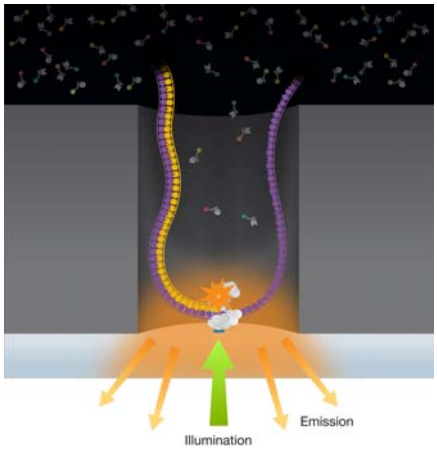


Figure 12. Continuous and simultaneous excitation and detection.

Both excitation and detection occur without interruption through the transparent glass bottom of the SMRT™ chip.

The light emitted by fluorophores is collected by a high numerical aperture objective lens and brought to a focus on a single-photon sensitive CCD array. Before reaching the array, the light passes through a prism dispersive element that deflects the fluorescent light according to its color, creating an individual rainbow for each ZMW **(Figure 13)**. This allows the position of the deflected light to encode the identity of the base that produced the signal. In this way a single high-sensitivity detector can be used to both identify and discriminate the pulses according to the position they strike the array. This process is repeated thousands of times over the area of the CCD array, enabling the DNA sequence to be read in real time in each ZMW across the entire SMRT chip.

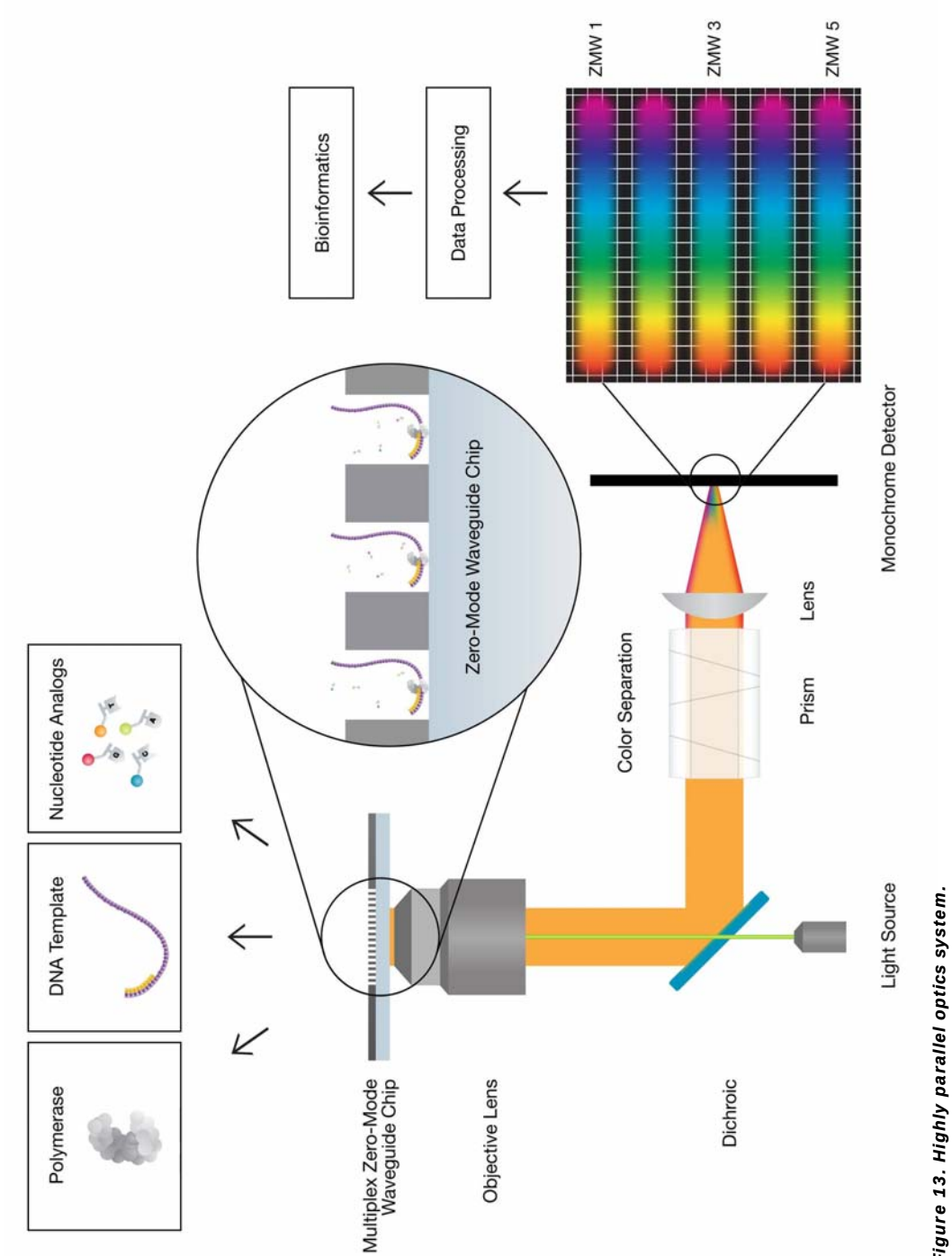


Figure 13. Highly parallel optics system.

The detected flash of light is separated into a spatial array, from which the identity of the incorporated base is determined.

High Quality Sequence Data

An optimized set of algorithms is used to translate the information that is captured by the optics system. Using the recorded spectral information and pulse characteristics, signals are converted into base calls with associated quality metrics **(Figures 14)**. To generate consensus sequence from the data, an assembly process aligns the different fragments based on common sequences.

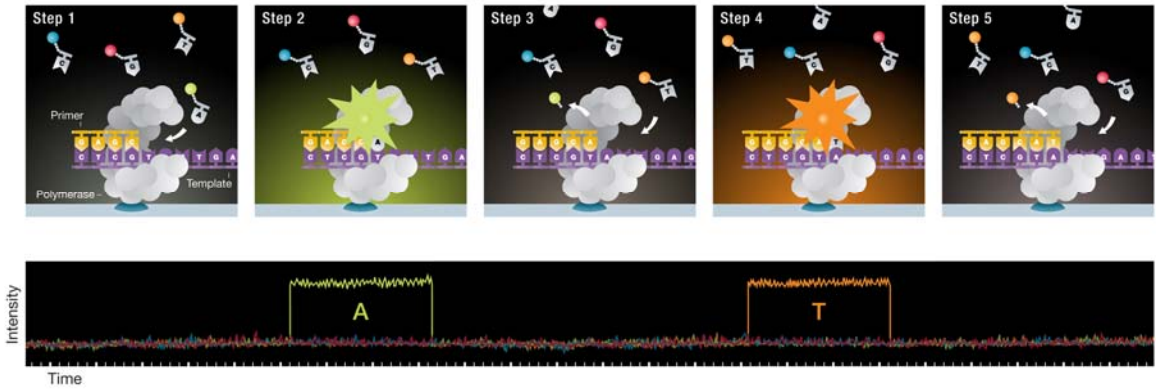


Figure 14. Generation of sequence data.

Enzymatic incorporation of the labeled nucleotide creates a flash of light, which is converted into a base call using optimized algorithms.

Alternative sequencing approaches produce short fragments of DNA, typically 30 – 50 bases in length. These are challenging and time consuming to assemble because they do not provide enough context to confidently determine their location within a genome. High coverage of the sequence target has to be used to compensate and increase confidence in assembly **(Figure 15)**.

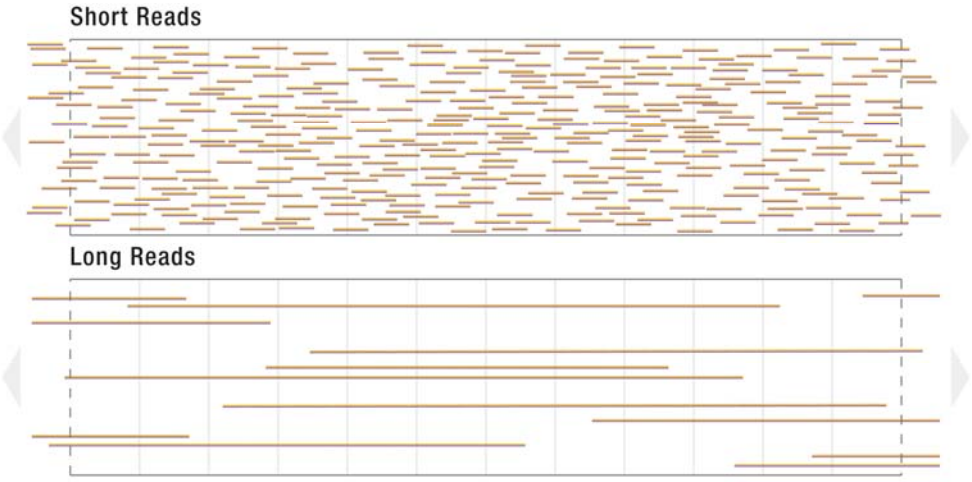


Figure 15. Sequence Assembly of Short versus Long Readlengths.

Long reads greatly simplify the assembly process, provide higher quality alignments, and generate finished sequence.

However, even with high coverage, it is impossible to generate a finished sequence with short fragments in highly repetitive areas of the genome.

The long reads generated by SMRT™ technology, greatly simplify this process. Because long reads can fully span repetitive or structurally varied genomic regions, assembly is simpler, alignments are of higher quality, and a finished sequence can be generated efficiently.

The PacBio informatics architecture is designed to be an open system, enabling scalable customization and integration with existing infrastructure.

High Performance with Massive Scalability

By harnessing the natural capabilities of DNA polymerase, our SMRT technology will offer a completely new performance envelope. Enzyme processivity enables long readlength while the speed of synthesis drives short cycle times. Along with generating high quality sequence data with less effort, there is a significant reduction in reagent cost due to the small reagent requirements. In contrast to other approaches, there are no wasteful and time consuming washing, scanning, and reagent redeployment steps.

Furthermore, with SMRT™ sequencing, performance and capabilities will continue to advance over time. As readlength is a function of the processivity of the enzyme, it will continue to increase as enhanced polymerases are developed. Using the same approach, synthesis will become faster as it is driven by the enzyme turnover rate. Higher throughput will be achieved by designing SMRT™ chips with larger numbers of ZMWs, as CCD array technology and bandwidth improve.

Finally, the wet chemistry and the hardware can be independently enhanced because the SMRT™ chip forms a literal barrier between the two. This provides for both seamless upgrades to leverage future performance, and the ability to customize and implement multiple application-specific assays on the same instrument.

RANGE OF APPLICATIONS

Due to the combination of long reads, fast cycle times, low costs, and high quality data, SMRT™ sequencing will be applicable for a broad range of applications, beginning with *de novo* sequencing and whole genome resequencing.

De Novo Sequencing

With the extensive repeat regions and overall size of complex genomes, long readlength provides a critical advantage. Overall, less coverage is required and more complete sequence is produced. Long reads provide the context that is required to correctly place extensive, repeat regions within the genome. Similarly, for metagenomics, long reads provide the capability to determine organism identity within heterogeneous samples. Simpler genomes benefit from long reads as well primarily through an easier assembly process.

Resequencing

SMRT™ sequencing provides significant advantages for resequencing applications as well. The system design enables customization of the assay for specific applications. One example of this is a molecular redundant sequencing technique, or repeated sequencing of a single molecule, for mutation discovery and detection **(Figure 16)**.

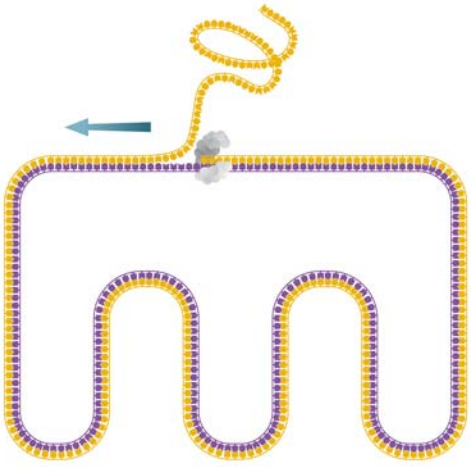


Figure 16. Molecular redundant sequencing

A strand displacing enzyme can be utilized on a circular template to generate independent reads of the same DNA molecule. The PHRED score increases linearly with the number of times the molecule is sequenced.

A strand displacing enzyme can be utilized on a circular template to generate multiple, independent reads of the same DNA molecule. The quality score increases linearly with the number of times the molecule is sequenced.

In addition, resequencing of complex genomes is simplified with long reads, as characterization of structural variation, including insertions, deletions, and rearrangements, is enabled. The SMRT™ approach also has the capability to provide haplotype information which is critical for medical resequencing applications and diagnostics in the future.

References

Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG, Webb WW, Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations, Science 299:682-686 (2003).

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. & Turner, S.W. (2008) Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. Proceedings of the National Academy of Sciences U.S.A. 105(4): 1176-1181.

Mathieu Foquet, Kevan T. Samiee, Xiangxu Kong, Bidhan P. Chaudhuri , Paul M. Lundquist, Stephen W. Turner, Jake Freudenthal, and Daniel B. Roitman, Improved fabrication of zero-mode waveguides for single-molecule detection, J. Appl. Phys. 103, 034301 (2008)