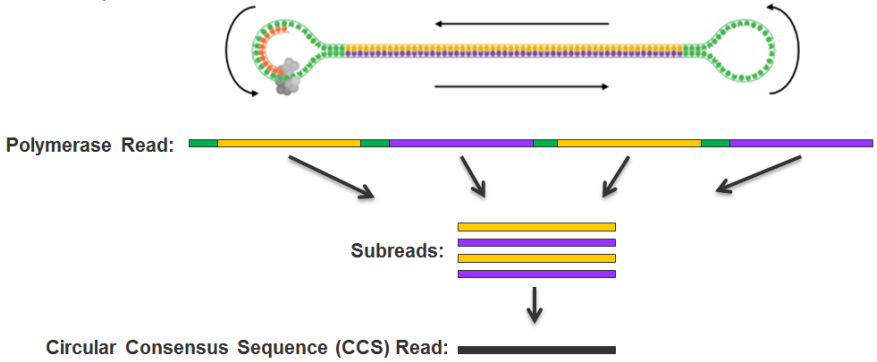
Pacific Biosciences Glossary of Terms

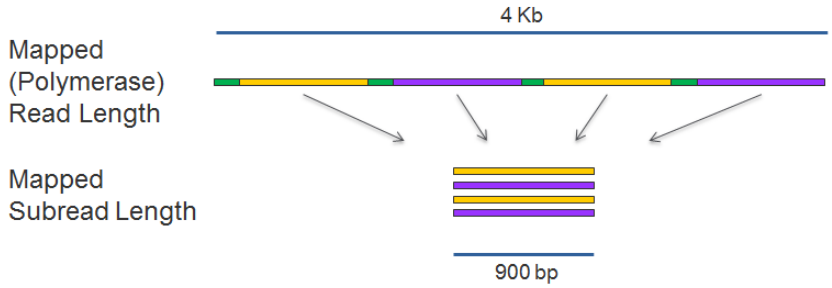
General Technology

- **Auto Analysis:** Allows a specific analysis to be **automatically** run after a sequencing run has finished and the data is transferred to the SMRT Link server. The analysis can include demultiplexed.
  - Sequel II System: Works with Sequel data, CCS data, and demultiplexed data.
  - Sequel System: Can perform Auto Analysis or Pre Analysis, but not both.
- **barcode score**: The alignment score between a read and an ideal barcode sequence. The maximum barcode score is twice the length of the ideal barcode sequence.
- **base yield density**: The total number of bases read in the run, as a function of the length of the read that contained them. While some binning is done to smooth the curve, the data should be interpreted as the number of bases per unit read length (every single integer read length spanned by the x-axis).
- **circular consensus sequencing (CCS) read**: The consensus sequence resulting from alignment between subreads taken from a single ZMW. Generating a CCS read does **not** include or require alignment against a reference sequence but **does** require at least two full-pass subreads from the insert. CCS reads are advantageous for amplicon and RNA sequencing projects and are highly accurate (>99% accuracy, Q≥20).



- **continuous long reads** (CLR): Reads with a subread length approximately equivalent to the polymerase read length indicating that the sequence is generated from a single continuous template from start to finish. The CLR sequencing mode emphasizes the **longest** possible reads.
- **full-pass subread**: A subread that begins at one adapter sequence and ends at another adapter sequence. A full-pass subread does **not** begin or end in the middle of an insert sequence.
- **high-fidelity long reads** (hifi reads): long reads with >Q20 (99%) single-molecule accuracy generated using the circular consensus sequencing analysis method.
- **longest subread length**: The mean of the maximum subread length per ZMW.

- **mapped polymerase read length**: Approximates the sequence produced by a polymerase in a ZMW. The total number of bases along a read from the first adapter of aligned subread to the last adapter or aligned subread.
- **mapped subread length**: The length of the subread alignment to a target reference sequence. This does **not** include the adapter sequence.



- **N50 read length metric**: The read length at which 50% of the bases are in reads longer than, or equal to, this value.
- **paired barcodes**: Barcode sequences that are different (asymmetric) on either end of an insert present in a SMRTbell® template. The barcoding analysis software uses unique pairs of barcodes to separate and analyze reads.
- **predictive loading**: A software feature that uses active monitoring of the ZMW loading process to predict a favorable loading end point.
- **pre-extension**: Pre-extension is a software feature that allows SMRTbell molecules to reach rolling circle replication (when the polymerase is most stable) before movie collection is initiated. Pre-extension is highly recommended for all insert sizes except size-selected large insert libraries.
- **Pre Analysis:** The process of CCS analysis and/or demultiplexing of Sequel basecalled data. Pre Analysis occurs **before** Auto Analysis.
- **polymerase read**: A sequence of nucleotides incorporated by the DNA polymerase while reading a template, such as a circular SMRTbell template. They can include sequences from adapters and from one or multiple passes around a circular template, which includes the insert of interest. Polymerase reads are most useful for quality control of the instrument run. Polymerase read metrics primarily reflect movie length and other run parameters rather than insert size distribution. Polymerase reads are trimmed to include only the high-quality region.
  Note: Sample quality is a major factor in polymerase read metrics.
- **polymerase read length**: The total number of bases produced from a ZMW after trimming the low-quality regions. It may include the adapter sequences.
- **polymerase read quality**: A trained prediction of a read's mapped accuracy based on its pulse and base file characteristics (peak signal-to-noise ratio, average base QV, inter-pulse duration, and so on).
- **preassembled long read (PLR)**: A generated read that has been output from the preassembly step of HGAP.

- **productivity**: A measure of the reads from a ZMW. `P=1` means that there is a polymerase read from that ZMW. `P=0` means that a ZMW did **not** produce a read and is presumed to be lacking a polymerase. `P=2` means "other" and the signal collected from the ZMW was not conducive to efficient base calling, possibly due to multiple template-polymerase complexes bound in the ZMW, high background signal, and so on.
- **read length**: The number of contiguous bases incorporated into a nascent strand during template-directed synthesis. Read length can be reported as **mean**, **N50**, **95th percentile**, and **max**.
- **read quality (RQ)**: The *de novo* prediction of the accuracy of subreads from a single ZMW. Sometimes also referred to as **QC Score** or **Read Score**.
- **sequence coverage** (polishing coverage): The total number of bases divided by the genome size. This includes multiple reads from same library molecules.
- **SMRT® Sequencing**: The process of nucleic acid sequencing using Pacific Biosciences' single molecule, real-time sequencing technology.
- **stage start**: When specified using RS Remote software, the sequencing reaction starts at the optical stage, just before data acquisition begins. This captures sequence data at the 5' end of the SMRTbell template that would otherwise be missed, and is especially important for long inserts where multi-pass CCS reads will not occur and the longest subreads are critical.
- **subread**: Each polymerase read is partitioned to form one or more **subreads**, which contain sequence from a single pass of a polymerase on a single strand of an insert within a SMRTbell template and **no** adapter sequences. The subreads contain the full set of quality values and kinetic measurements. Subreads are useful for applications such as *de novo* assembly, resequencing, base modification analysis, etc.
- **symmetric barcodes**: Barcode sequences that are identical on both ends of an insert present in a SMRTbell template.
- **unique molecular coverage** (physical coverage): The number of bases from unique library molecules (or ZMWs or sequencing reactions) divided by the genome size.
- **unique molecular yield**: The sum total length of unique single molecules that were sequenced. It is calculated as the sum of per-ZMW median subread lengths.
- **zero-mode waveguide (ZMW)**: A nanophotonic device for confining light to a small observation volume. This can be, for example, a small hole in a conductive layer whose diameter is too small to permit the propagation of light in the wavelength range used for detection. Part of a SMRT® Cell.

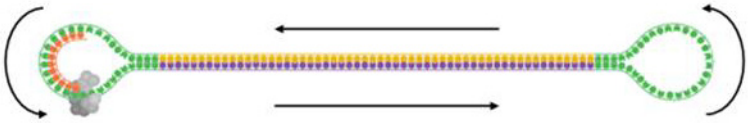
Consumables

- **DNA Controls:** DNA control template-polymerase complexes that are ready for loading on the instrument. There are two types of controls: **Spike-in** controls and **stand-alone** controls. Spike-in controls are added into your samples for sequencing. Stand-alone controls are run on a single SMRT Cell for instrument troubleshooting purposes.
- **MagBead**: Small paramagnetic bead, 2-3 µm in size. The DNA-polymerase complexes are attached to the magnetic beads, which can then be pulled down for easy removal of contaminants from the supernatant during the binding step. The DNA-polymerase complex/ bead mixture can be used for the on-instrument immobilization step. See also **MagBead loading**.
- **SMRT® Cells**: Consumable substrates comprising arrays of zero-mode waveguide nanostructures. SMRT Cells are used in conjunction with the **DNA Sequencing Kit** for on-instrument DNA sequencing.
- **SMRT® Cells LR**: A SMRT Cell version supporting movie collection times up to twenty hours.
- **SMRT Cells 8M**: SMRT Cells used with the Sequel II instrument, containing eight million zero mode waveguides, and supporting movie collection times up to 30 hours.

Template Preparation

- **AT ligation**: The library construction protocol option by which an adapter with a single-nucleotide T overhang is ligated to an insert with a single-nucleotide A overhang. The workflow that uses this ligation option also contains an A-tailing step.
- **barcode padding:** An optional 5 base pair constant sequence appended to unique barcode sequences. Can be used to normalize ligation of adapters during template preparation.
- **barcoded adapter:** A SMRTbell adapter with a barcode sequence appended to the end of the stem region. When using barcoded adapters, SMRTbell templates will have a symmetric barcode structure.
- **barcoded SMRTbell® template**: A SMRTbell template with two barcoded adapters.
- **blunt ligation**: The library construction protocol option by which an adapter lacking any overhangs is ligated to an insert also lacking any overhangs. The workflow that uses this ligation option also lacks the A-tailing step.
- **diffusion loading**: Immobilization of DNA-polymerase complex into the ZMWs on the SMRT Cell via diffusion. Smaller inserts load preferentially compared to larger inserts.
- **DNA damage repair**: A step in the SMRTbell template preparation that repairs a variety of types of DNA damage, including pyrimidine dimers, abasic sites, and nicks.
- **DNA end repair**: A step in the SMRTbell template preparation that removes 5’ and 3’ overhangs, and phosphorylates 5’ ends.
- **DNA fragmentation**: The generation of smaller DNA fragments. Multiple methods may be used to fragment DNA, including hydrodynamic shearing, mechanical shearing, sonication, and enzymatic digestion.

- **input DNA**: Purified DNA used as an input to the **DNA Template Prep Kit**.
- **insert size**: The length of the double-stranded nucleic acid fragment in a SMRTbell template, excluding the hairpin adapters.
- **MagBead loading**: Immobilization of large DNA molecules into the ZMWs on the SMRT Cell chip via MagBeads. The smallest inserts, hairpin dimers, and excess polymerase are washed out in the initial MagBead binding and washing steps. As a result, medium and larger size inserts load better and have a higher sequencing accuracy (compared to diffusion loading of similar-sized inserts).
- **polymerase binding**: The binding of the sequencing polymerase to an appropriate binding site on a nucleic acid template.
- **primed template**: Refers to a template molecule that is annealed with primer. A primed template is a product of the template prep protocol and an input to the binding protocol.
- **primer annealing**: The hybridization of a sequencing primer to an appropriate binding site on a template.
- **size selection**: The removal of unwanted fragments from a mixture based on size. This can refer to the removal of only the shortest fragments, such as adapter dimers, or to the isolation of a very narrow range of insert sizes. Depending on the size range of interest and the equipment available, size selection can be accomplished with AMPure PB beads, manual isolation from an Agarose-gel, automated gel isolation, or BluePippin™ or SageELF™ size-selection systems.
- **SMRTbell® template**: A double-stranded DNA template capped by hairpin adapters (i.e., SMRTbell adapters) at both ends. A SMRTbell template is topologically circular and structurally linear, and is the library format created by the **DNA Template Prep Kit**.



- **template**: Molecules to be sequenced; the **DNA Template Prep Kit** produces SMRTbell templates.
- **template annealing**: Process of hybridizing primer(s) to nucleic acid templates.
- **template library**: A set of nucleic acid molecules to be sequenced; the DNA Template Prep Kit process generates template libraries.
- **template-polymerase complex**: Primed template bound to DNA polymerase; the output of the DNA/Polymerase Binding Kit process.

General Software

- **PacBio® DevNet** (http://www.pacb.com/products-and-services/analytical-software/devnet/): Resource for informatics researchers, and life scientists; includes data sets, source code, application programming interfaces, and documentation.
- **SMRT® Link**: Web-based end-to-end workflow manager for the Sequel® System. It includes software applications for setting up samples, designing and monitoring sequencing runs, and analyzing and managing sequence data.
  - **Data Management**: SMRT Link software module used to create Projects and Data Sets, and manage access permissions for Projects and users.
  - **SMRT® Analysis**: SMRT Link software module used to perform secondary analysis of data including sequence alignment, variant detection, *de novo* assembly, RNA analysis, and detecting epigenetic DNA modifications.
- **SMRT® View**: Java-based genome browser used to visualize aligned or assembled reads. Part of the SMRT Analysis suite.

Sequel® System only:

- ***Sequel® Instrument Control Software (ICS)****: Software that controls instrument functionality including: on-instrument operations, user interface, and primary analysis.*
- ***Run Design:*** *SMRT Link software module used to design and import sequencing runs into Sequel ICS.*
- ***Run QC****: SMRT Link software module that displays quality metrics for individual instruments and runs on the Sequel System.*
- ***Sample Setup****: SMRT Link software module used to calculate binding and annealing reactions for preparing DNA samples for use on the Sequel System.*

PacBio® RS II System only:

- ***Binding Calculator****: Web-based application used to calculate binding and annealing reactions for preparing DNA samples for use on the PacBio RS II.*
- ***RS Dashboard****: Web-based software that displays quality metrics for individual instruments, runs, and SMRT Cells.*
- ***RS Remote****: Windows-based client software used to design and monitor sequencing runs. Directs user to look at primary analysis data in-depth using RS Dashboard.*
- ***RS Touch****: Touchscreen user interface on the instruments to help the user load the instrument and start a run. Provides direct feedback on instrument status.*
- ***SMRT® Pipe****: Command-line interface of SMRT Analysis suite. SMRT Pipe is used to launch secondary analysis jobs.*
- ***SMRT® Portal****: Web-based software used to help set up secondary analysis jobs and view quality reports. Part of the SMRT Analysis suite.*

Primary Analysis

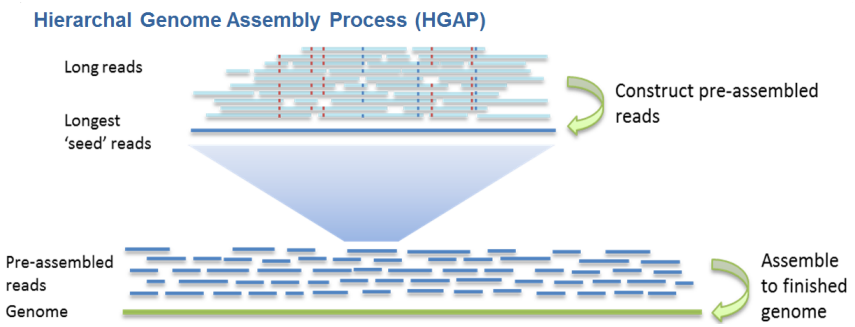
- **high-quality (HQ) region**: Annotates the high-quality sequencing regions of a read to be used during raw read trimming.
- **movie time**: The time specified for collecting data from a SMRT Cell.
- **primary analysis**: On-instrument analysis which includes signal processing of the movie, base calling of the traces and pulses, and quality assessment of the base calls. Subsequently, it trims the sequences to the high-quality (HQ) regions, identifies adapter, barcode (optional), and control sequence reads, assigns read scores, and outputs the subread data in a BAM file.
- **pulse**: The representation of an incorporation event derived from a trace. It includes metrics such as interpulse duration, pulse height, and pulse width.
- **QV Metric**: Phred-like scores that reflect the probability of a correct call for each base.
- **raw read trimming**: Removing a section of the raw read such as a low-quality region. Trimming of an unfiltered read produces a polymerase read.
- **reads/SMRT® Cell**: The number of reads generated per SMRT Cell.

Secondary Analysis

- **Arrow**: (**Note**: This replaces **Quiver**.) A highly accurate consensus and variant caller that can generate >99.999% accurate consensus sequences using local realignment and the full range of quality scores associated with reads. Part of SMRT Analysis.
- **barcode FASTA**: A FASTA-format file used by barcoding software to identify ideal barcode sequences. For symmetric barcodes, each barcode sequence identifies a single bin for demultiplexing reads. For paired barcodes, each unique pair of barcodes should be listed as two sequentially-named FASTA sequences.
- **BLASR**: (Note: This is replaced by minimap2.) Used for targeted sequencing. Maps reads against a reference; part of SMRT Analysis.
- **consensus accuracy**: Accuracy based on aligning multiple sequencing reads or subreads together.
- **circular consensus accuracy**: Accuracy based on consensus sequence from multiple sequencing passes around a single circular template molecule.
- **circular consensus sequence analysis**: Processing of sequencing data generated by circular consensus sequencing to create a circular consensus read.
- **circular consensus sequencing (CCS)**: Sequencing performed on a circular template in which a subread was generated during a sequencing pass around the template. These reads are aligned to each other to generate a single high-accuracy consensus read. CCS analysis to generate consensus reads requires CCS data with at least two full-pass subreads.
- **consensus sequence determination**: Generation of a consensus sequence from multiple individual reads or from multiple subreads of the same DNA molecule (CCS). Also termed “consensus calling”. Multi-molecule consensus for targeted applications can be generated using

the Long Amplicon Analysis (LAA) protocol. For other multi-molecule applications, such as whole-genome sequencing and *de novo* assembly, consensus is generated using the Arrow consensus caller as part of the resequencing protocol.

- **HGAP**: The Hierarchical Genome Assembly Process (HGAP) is an algorithmic approach for generating high-quality *de novo* assemblies using PacBio data. HGAP includes pre-assembly, *de novo* assembly and assembly polishing steps. HGAP3 utilizes an AssembleUnitig module for *de novo* assembly and Quiver for polishing. HGAP4 uses Falcon for *de novo* assembly and **Arrow** for polishing.



- **Quiver**: (**Note**: This is replaced by **Arrow**.) A highly accurate consensus and variant caller that can generate >99.999% accurate consensus sequences using local realignment and the full range of quality scores associated with reads. Part of SMRT Analysis.
- **secondary analysis**: Follows primary analysis and uses basecalled data. It is application-specific, and may include:
  - Filtering/selection of data that meets a desired criteria (such as quality, read length, etc.).
  - Comparison of reads to a reference or between each other for mapping and variant calling, consensus sequence determination, alignment and assembly (*de novo* or reference-based), variant identification, etc.
  - Quality evaluations for a sequencing run, consensus sequence, assembly, etc.
  - PacBio's SMRT Analysis contains a variety of secondary analysis applications including RNA and Epigenomics analysis tools.
- **secondary analysis application** (Formerly "Secondary analysis protocol"): A secondary analysis workflow that may include multiple analysis steps. Examples include *de novo* assembly, resequencing, RNA and epigenomics analysis.
- **SMRT® Analysis Suite**: Client/server software that performs automated and distributed analysis of sequencing data generated by the PacBio System.
- **tertiary analysis**: Analyses following secondary analysis, which includes comparisons of secondary analysis results across different samples, application-specific analyses, variant classification, and disease/gene annotations.

Base Modification

- **amplified control**: Control created by separately sequencing an amplified version of the sample of interest.
- **interpulse duration (IPD)**: Metric for the length of time between emission pulses indicative of base incorporation events. Base modifications in a template molecule can impact IPD, so changes in IPD are used to detect base modifications during SMRT Sequencing.
- **IPD Ratio**: The ratio of the mean IPD of a native sample to the mean IPD in a second sample or in silico control at a position of inquiry in a template.
- **i*n silico* control**: Computational model for predicting the mean IPD per given sequence context at the position of inquiry.

Instrument Terminology

PacBio RS II System only:

- ***Blade Center****: The podium containing computers used to process primary analysis and control the instrument.*
- ***carrier plate****: Metal plate with 12 slots to accommodate 8Pac strips.*
- ***drawer****: Each of the two compartments where consumables are placed by the user (including the door). The left drawer houses reagents and samples, and the right drawer houses SMRT Cells and tips.*
- ***Environmental Cabinet****: The right-most compartment of the instrument containing the nitrogen intake and chiller sub-module.*
- ***insert in drawer****: The removable aluminum assembly found in the (left) reagent drawer. The insert consists of a top unit with septa clamps and tube openings, and a bottom unit with receptacle areas for two reagent plates, one mixing plate, one sample plate, and four tubes. The insert is removable for cleaning.*
- ***tip station****: The area of the SMRT Cell and tip drawer that accommodates up to 6 boxes of pipette tips.*

Revision History (Description)	Version	Date
Added SMRT Cell LR.	08	January 2018
Added base yield density, continuous long reads, and high-fidelity long reads, sequence coverage, and unique molecular coverage.	09	October 2018
Added SMRT Cells 8M, auto analysis, unique modular yield, Auto Analysis, Pre Analysis, predictive loading. Modified CLR and CCS.	10	April 2019

PN 000-710-267 Version 10 (April 2019)