PACIFIC BIOSCIENCES®

Targeted Sequencing – SNP Detection and Validation

Experimental Design

Introduction

Single-nucleotide polymorphism ("SNP") detection and validation is critical to identifying biological variations that lead to actionable results. In the context of next-generation sequencing ("NGS"), SNP detection can be defined as the identification of variants relative to a known reference sequence. SNP validation refers to the comparison of the detected variants with a known set of SNPs, and is used to evaluate the ability of a technology to identify known SNPs and discover novel ones. A high level of accuracy in determining true positives and true negatives is needed. The PacBio® *RS* system provides a high positive predictive value in SNP detection and validation, with flexibility in amplicon design, using a simple low-cost workflow.

In this document, we explain the scientific goals, experimental design, and workflow and analysis protocols for detecting and validating true SNPs from amplicons.

SNP Detection and Next Generation Sequencing

NGS relies mainly on sequencing depth, or coverage, to identify potential variant locations. Deep sequencing approaches are used to detect mutations at very low levels. Given a type of sequencing data, the goal is to achieve enough coverage at the position of interest to accurately disambiguate true variants from sequencing error.

The PacBio *RS* has several important features which make it an attractive and cost-effective platform for various SNP detection applications. Specifically:

- The protocols are relatively insensitive to insert size; the PacBio RS delivers customized solutions for different SNP detection applications

- Sequencing longer amplicons (>1 kb) as single molecules allows the empirical phasing of mutations hundreds of base pairs apart

- Long reads have a higher mapping specificity to the reference; so SNPs in repetitive regions can be investigated by including unique flanking regions as part of the amplicon

- The random nature of mis-calls in PacBio reads reduces the likelihood that systematic machine errors are classified as novel variants

- Consensus calls from multiple reads "wash out" random errors, leading to consensus accuracy >99.9% at coverage of 10X

PacBio® *RS* Data Types

The ability to select the insert size, movie time, and number of SMRT® Cell sets acquired allows the user to optimize the generation of either single pass Continuous Long Read ("CLR") or multi-pass Circular Consensus Sequencing Read ("CCS") data types for different SNP detection requirements. Table 1 summarizes the characteristics of CLR and CCS data most pertinent to targeted sequencing applications for variant detection.

Experiment Design ▸ **Isolate DNA** ▸ **Template Preparation** ▸ **Sequencing** ▸ **DATA Analysis**

PACIFIC
BIOSCIENCES®

Table 1: Characteristics of CLR and CCS Reads

	Continuous Long Read (CLR)	Circular Consensus Sequencing (CCS)
Description	Generated from a single pass (subread) of the polymerase across a single DNA template in a single Zero Mode Waveguide ("ZMW")	Generated from ≥2 pass subreads from the same ZMW or molecule
Read Accuracy	85-90%	≥97% (2 pass) ≥98% (3 pass) ≥99% (≥5 pass)
Per Base QVs	8.24-10	≥13 (2 pass) ≥17 (3 pass) ≥20 (≥5 pass)
Maximum Mean Readlength	2 kb	1 kb
Average # Usable Reads* per SMRT Cell (2X45 min protocol)	250K (<500 bp insert) 100K (1 kb insert) 50K (2 kb insert)	40K (<500 bp insert, ≥3 pass) 30K (500 bp insert, ≥3 pass) 15K (1 kb insert, ≥3 pass)
10X Coverage QV	≥30	≥50

*Usable reads in this table refers to full-pass subreads for CLR and ≥ 3 pass CCS reads.

Estimating Per Base Coverage Requirements for SNP Detection

Coverage is defined as the number of independent base calls generated for a particular position in a known reference alignment. The amount of coverage required to accurately call a SNP depends on:

- The type of variation being queried (e.g., somatic or germ-line, haploid or diploid)

- The SNP identification program being used

- The required confidence in the SNP call

- The QV or accuracy of the base calls at that position

SNP identification programs have varied prerequisites for SNP calling. Higher confidence and low frequency SNPs require more supporting evidence or coverage. In general, base QV and required base coverage are inversely related. Because of these factors, minimum coverage requirements for SNP detection necessarily vary by the specific use-case.

However, it is possible to model the SNP detection problem statistically. As stated above, the goal of SNP detection is distinguishing true variation from sequencing error.

Consider two populations of base calls at the same genomic position. All base calls have constant QVs. One population contains 100% wild-type base calls while the other contains base calls with a defined wildtype:variant frequency (e.g., 50/50 or 80:20). Base calls are sampled from both populations randomly, and the sample size is equivalent to the coverage at that position. A contingency table can be constructed from the two samples, and Fisher's exact test can be used to test the hypotheses that the two samples came from different read populations, with the null hypotheses being that the wild-type:variant frequency in the populations are equal.

Figure 1 shows the results of a Monte Carlo simulation from sampling base calls from a 50/50 population (e.g., heterozygote germ-line variant) and a 100/0 population (e.g., homozygote wild-type). Fisher's exact test is used to calculate a p-value for whether the resulting variant base call samples are associated to the wild-type base calls: a p-value ≤ 0.05 represents a successful identification of the variant population, while a p-value > 0.05 represents the failure to do so. The testing was done 1000 times for each coverage level 1-50 and the probability of detection was calculated as (# p-value observations ≤ 0.05) / 1000. The simulation was done at four Quality Value ("QV") levels: Q8.24, Q10, Q20, and Q30. Here, Q8.24 is representative of the 85% accuracy of PacBio CLR, Q17 is representative of PacBio CCS accuracy at 3 passes, and Q20 is representative of PacBio CCS accuracy at ≥ 5 passes. The Q10 (90% accuracy) and Q30 (99.9% accuracy) data are included for comparison. This simulation is also a conservative upper-bound estimate for required coverage for 95% probability of detection.

Base calls from all levels of QV have low probabilities of detecting the variant population at coverage levels < 6. A 95% detection rate (the yellow horizontal solid line) was achieved at 14X, 17X and 29X coverage for the Q20, Q17, and Q8.24 base calls, respectively. This agrees with the notion that high accuracy base calls require less coverage to observe a variant position than lower accuracy base calls.
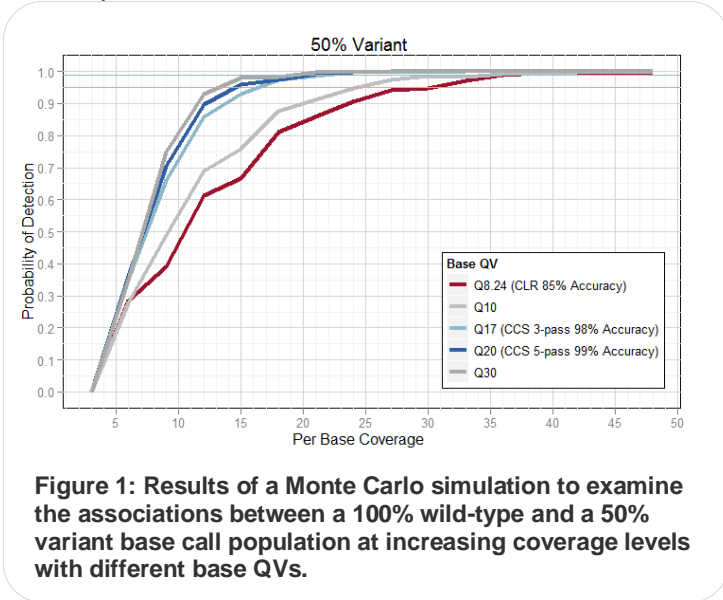

Figure 1: Results of a Monte Carlo simulation to examine the associations between a 100% wild-type and a 50% variant base call population at increasing coverage levels with different base QVs.

Figure 2 shows the simulation results for a 20% variant population (e.g., somatic mutation in solid tumor biopsy) and a 100% population (e.g., normal issue). As expected, a higher amount of coverage is required to detect a lower frequency variation at equivalent base QVs. With CLR reads of Q8.24 coverage in excess of 135X may be required to detect a 20% variant base population with 95% probability. These simulation estimates can be either conservative or extreme given different SNP detection requirements. Base QV recalibration methods such as the Broad Institute's open source Genome Analysis Toolkit ("GATK")[1] can effectively raise the QV for CLR to Q20 levels, reducing the coverage needed to call variants at this frequency. This is a function of the low mismatch error rate for CLR reads, making the empirical SNP QV higher than 8.24.
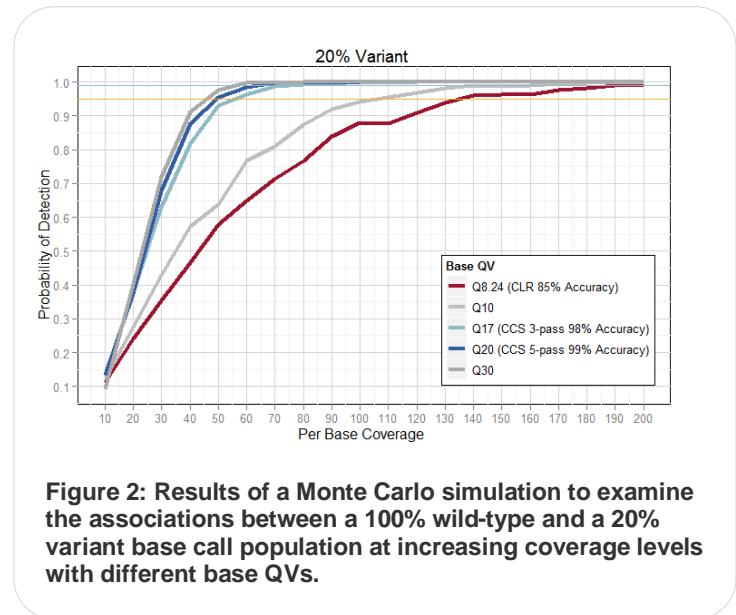

Figure 2: Results of a Monte Carlo simulation to examine the associations between a 100% wild-type and a 20% variant base call population at increasing coverage levels with different base QVs.

Selecting the Data Type

The coverage simulations above show that both CLR and CCS data can be effective in detecting variants with differing levels of coverage. Higher accuracy CCS data allow for increased sensitivity in detecting minor variants, but CLRs allow for larger amplicon sizes as well as more usable reads per SMRT Cell. The following table highlights the major considerations taken into account when choosing between the two.

Table 2: Considerations for Choosing CLR or CCS

	Continuous Long Reads (CLR)	Circular Consensus Sequencing (CCS)
Amplicon Size	>1 kb	<1 kb
Variant Frequency	20-50%	1-50%
Base QV Recalibration	Beneficial and can reduce coverage requirements by increasing per base QVs	Not currently recommended
Barcoding For Multiplexing	≥20-mer barcode design	≥16-mer barcode design
Ploidy	Haploid	Diploid and Polyploid

Generally speaking, CCS reads should be used with amplicon sizes < 1 kb and CLRs for amplicon sizes 1 kb to 3 kb. Although insert size is the major determinant of data type, there can be considerable overlap in the utility of the two data types. For example, either data type may be used for 1 kb reads at 50%, and the choice to use CLR or CCS may be driven by other considerations (e.g., use of barcodes, multiplex composition, confidence required, base recalibration, or requirements of a specific downstream data-processing pipeline). Further, as Q20 reads can be achieved with ≥ 5 passes across the template in CCS reads, CCS yield and accuracy can be optimized by generating inserts that are an appropriate fraction of the raw mean read length.

Estimating the Number of SMRT® Cells Required

The number of SMRT Cells needed for a particular experiment is mainly contingent upon the number of reads required to achieve a specific level of coverage and accuracy. However, there are several other factors that influence the uniformity of coverage across a library.

Most notably, multiplexed libraries require special consideration and constituent amplicons should be uniform in both size and frequency. Specifically, unequal representation of amplicons comprising a multiplexed SMRTbell™ library may result in uneven coverage among amplicons. In this case, the desired coverage should be calculated to accommodate the rarest constituent and can, therefore, be defined as a *minimum* value not a *mean* value. Otherwise, while the mean coverage for all targets may be above a desired

threshold, the minimum coverage for certain targets may fall below the threshold.

Libraries containing a broad range of amplicon sizes require higher *mean* coverage to ensure adequate *minimum* coverage across the larger amplicons. Variability of coverage across multiple amplicons can be reduced if the amplicon sizes are +/-10% of each other. Additionally, consistent amplicon sizes also allow for the generation of a single data type for analyses (i.e. mixing 300 bp and 1.5 kb amplicons will necessitate the use of CLRs for the 1.5 kb amplicon, while the 300 bp amplicon will have sufficient data for both CLR and CCS analysis).

The following formula can be used to estimate the number of SMRT Cells required for a targeted sequencing experiment.

$$\#required\ reads\ N = \frac{t\ \ c\ \ \beta}{r} \qquad (1)$$

Where t = (# targets), c = (required coverage per target), β = (sample bias), and r = (number of usable reads). The # usable reads refers to full-pass subreads for CLR and ≥3 pass CCS reads. Sample bias, β, represents the fold difference in read counts between the lowest frequency amplicon in the sample pool and the expected read count for the amplicon. N can then be divided by the number of reads per SMRT Cell given the appropriate data type and insert size (see Table 1) to get the estimated required number of SMRT Cells.

PACIFIC
BIOSCIENCES®

Continuous Long Read (Multi-Molecule Consensus)

The number of SMRT Cells required for CLR sequencing can be calculated using formula (1) when using 2 kb or smaller inserts.

In Table 3, a sample bias of 3-fold is used to estimate the number of reads and SMRT Cells. The variation in each experiment may be different and is affected by factors such as molar variation when pooling samples, PCR amplification bias, and size differences between amplicons.

Prior sample experience or independent validation tests should be used to estimate the true amplicon variation. The number of CLR per SMRT Cell in the denominator corresponds to the number of full-pass subreads (spanning the complete amplicon length). For simplicity, partial-pass subreads are excluded from the estimations and the required number of reads may be an overestimate.

Table 3: Example estimates of SMRT Cells using CLR

	50% Variant			20% Variant		
Target Size	500 bp	1,000 bp	2,000 bp	500 bp	1,000 bp	2,000 bp
# Targets	500	2,000	1,000	500	2,000	1,000
Minimum Coverage	30X	30X	30X	140X	140X	140X
Required # Reads	45,000	180,000	90,000	210,000	840,000	420,000
Estimated CLR per SMRT Cell	250,000	100,000	50,000	250,000	100,000	50,000
Estimated # SMRT Cells	1	2	2	1	9	9

Circular Consensus Sequencing

The number of SMRT Cells required for CCS can be calculated using formula (1).

In Table 4, 3-pass CCS accuracy and yields were used to estimate SMRT Cell requirements. The 17X and 55X minimum coverage values are used (rounded from the intercepts in figures 1 and 2) for simplicity of calculation. Again, a 3-fold sample bias is assumed for estimation purposes.

Table 4. Example estimates of SMRT Cells using CCS

	50% Variant		20% Variant	
Target Size	500 bp	1,000 bp	500 bp	1,000 bp
# Targets	500	2,000	500	2,000
Minimum Coverage	17X	17X	55X	55X
Required # Reads	25,500	102,000	82,500	330,000
Estimated CCS Reads per SMRT Cell	30,000	15,000	30,000	15,000
Estimated # SMRT Cells	1	7	2	22

As evident in Tables 3 and 4, SMRT Cell usage is fairly consistent for 500 bp amplicons across both data types. However, 1000 bp amplicons have a much higher CCS cost since the number of CCS reads, for that amplicon size, decreases by more than half from the 500 bp amplicon. Therefore, for a 2 kb insert, CLRs are the better option.

Preparing the Samples

Samples can be enriched for various targets using multiple methods including traditional PCR and commercialized hybridization-based approaches (see Table 5). The creation and processing of amplicons introduces damage to DNA molecules, which should be mitigated prior to SMRTbell library prep.

Specifically, during the denaturation and extension steps of PCR amplification, PCR templates are exposed to high temperatures which can cause damage to the amplicons as they are being created. Longer amplicons (> 2 kb) are generated with protocols using more cycles and longer incubations. This greater exposure to damaging temperatures means a higher likelihood of DNA nicking across the length of the amplicon.

Damage to amplicons manifests as low yields during SMRTbell library preparation. Degraded PCR primers may also contribute to low SMRTbell library yields. Degraded primers may contain hydrolyzed bases, which are incorporated into amplicons, and lower SMRTbell library yields. Proper storage and treatment of primers is critical to successful SMRT Sequencing. Specifically, primers should be re-suspended, at high concentrations in buffered solutions, and aliquot to minimize freeze thaw cycles. Additionally, we recommend purifying PCR products using methods that do not require visualization with UV light (as even brief exposure to UV severely damages PCR amplicons). Should amplicons require gel purification, we recommend using SYBR® GOLD and a blue light.

Since it is impossible to be sure that PCR products are completely undamaged, a DNA damage repair step should be included in all amplicon re-sequencing projects (especially when the insert size is 2 kb or larger). See our recommendations in the *Pacific Biosciences® Template Preparation and Sequencing Guide*. In addition to avoiding DNA damage, primers used for targeting should not have modified dyes or other labels so that the ends can be compatible for SMRT template preparation.

For multiplexing samples, we have tested 96 16-mer barcodes that can be incorporated as tails within the PCR primers. See the *Pacific Biosciences Technical Note – Multiplexing Targeted Sequencing Using Barcodes*.

Table 5. Commercial Target Enrichment Options

Enrichment Technology	Enrichment Method	Target Insert Size	Targeted Region
Fluidigm Access Array™ System	PCR	Up to 10 kb	≤10 Mb
Agilent Technologies SureSelect Target Enrichment System	Hybridization	Up to 2 kb	≤50 Mb
Raindance™ Technologies Rainstorm™ Technology	PCR	Up to 1.5 kb	≤10 Mb

To minimize sample bias in pooled amplicons, amplicons comprising a multiplexed SMRTbell library should be equally represented in the pool, and their lengths within +/-10% of each other. Primer dimers should be removed to avoid their conversion into small inserts during library preparation.

SMRTbell libraries should be constructed using the Pacific Biosciences' commercial Template Prep Kit and accompanying procedures. For additional details about integrating various target enrichment strategies with the Pacbio *RS* sequencing protocols, see *Pacific Biosciences Technical Note – Targeted Sequencing on the PacBio RS using Agilent Technologies SureSelect Target Enrichment.*

Sequencing

The variety in input material leads to libraries with varied loading efficiencies. When running multiple SMRT Cells for a given SMRTbell library, an optimal loading concentration should be determined following Pacific Biosciences recommendations for loading titrations (see the *Pacific Biosciences Template Preparation and Sequencing Guide*).

Selecting the Collection Protocol

The 2X45 minute collection protocol is sufficient to generate CCS data from short inserts (500 bp to 1 kb) and CLR data for long inserts (1 kb to 3 kb). The 1X75 or 1X90 minute protocols may be appropriate when sequencing amplicons longer than 3 kb, or when CCS data is desired for long inserts (1 kb – 3 kb). However, the number of reads generated will be lower.

Analyzing the Data

Data analysis can be done in several ways. PacBio's SMRTAnalysis v1.3 has the following tools for variant detection.

PacBio v1.3 SMRT® Analysis: EviCons

EviCons produces the consensus sequence from a multiple sequence alignment corresponding to mapped reads (resequencing) or a contig (*de novo*). Using empirical conditional probabilities and a likelihood ratio test, EviCons separates the multiple sequence alignment into regions of certainty and regions of uncertainty.

For regions of uncertainty, EviCons uses base quality values and the Steiner framework to produce the best estimate of the local consensus sequence. EviCons is appropriate for SNP calling in haploid organisms. The *RS_Resequencing* protocol selection uses EviCons to detect variants.

GATK

GATK is the Broad Institute's unified genotyper for Bayesian diploid and haploid SNP calling and is available at: *http://www.broadinstitute.org/gsa/wiki/index.php/Variant_quality_score_recalibration*. It supports base quality score recalibration through known SNP data that has been shown to increase variant calling accuracy. The recalibration used by GATK uses data from the experiment to correct for variation in quality scores between machine cycles or sequence context. Variant confidence recalibration and other advanced operations are not currently supported through SMRT Pipe. However, it has been specifically modified for integration into the PacBio routine analysis pipeline (the SMRT Portal v1.3 software protocol *RS_Resequencing_GATK*) to detect variants.

Third-Party Tools

SAMtools

SAMtools is a set of utilities for working with alignments in the SAM/BAM format. SMRT Pipe outputs both SAM and BAM files for alignments which can be used with SAMtools utilities. The `mpileup`[2] option in `samtools` can be used in conjunction with `bcftools` to call SNPs and short indels. However, SAMtools and BCFtools cannot handle multi-allelic variants and indels.

Conclusion

The PacBio *RS* provides a robust platform for SNP detection and validation. When planning a targeted sequencing experiment for variant detection, it is important to consider the size of the amplicons, read types desired, level of coverage, and confidence of variant detection required. Shorter inserts (< 1 kb) can generate high accuracy CCS reads that can detect multi-allelic variants with relatively low coverage. Longer inserts can generate long CLRs that effectively phase variants on the same allele. Accurate haploid SNP calling can be achieved with CLRs with 20X coverage or

CCS reads with 10X coverage. Multi-allelic variant detection can also be achieved, with both read types, using the GATK unified genotyper. CLRs can benefit greatly from the base recalibration pipeline included with GATK.

References

1. http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_AnalysisToolkit

2. http://samtools.sourceforge.net/mpileup.shtml