

Detecting Genetic Variations using the PacBio® RS Targeted Sequencing

Introduction

The PacBio® RS provides long reads and low systematic errors for fully characterizing genetic complexity, including: rare SNPs, indels, structural variants, and mutation phasing. Single molecule resolution allows comprehensive characterization of heterogeneous samples and identification of variations invisible to multi-molecule sequencing technologies.

In this application note, several case studies from oncology and human genetic diseases are presented to illustrate the power of the PacBio RS, and Single Molecule, Real-Time (SMRT®) sequencing, in discovering and detecting genetic variants. A brief overview of the PacBio data types and analyses is also provided.

The PacBio RS Provides

- Extra-long read lengths and low systematic errors, which facilitate maximum mapping accuracy
- Highest accuracy for variant detection including: SNPs, indels, and structural variants
- Long reads provide phasing information
- True *De Novo* capability
- Easy integration with third party enrichment and analysis methods

Understanding PacBio® Data

The information content of a read is a combination of its read length and accuracy. Since the errors from SMRT sequencing are stochastic, they are naturally unbiased. The unbiased error profile allows for high consensus accuracy at low coverage (see Figure 1). In fact, only about 10 molecules-worth of coverage is needed to achieve 99.9% consensus accuracy.

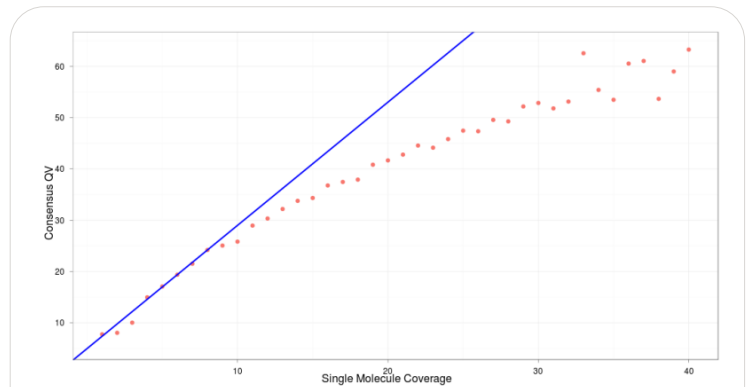
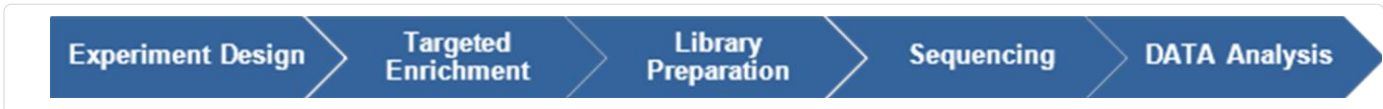


Figure 1. Coverage vs. Consensus QV (logarithmic accuracy transformation¹) for PacBio® RS (red points).

Long reads increase consensus accuracy because they map more accurately than short reads. Repetitive sequences in genomes such as low-complexity regions, microsatellites, tandem repeats, transposable elements, etc., make read length the main determinant in the ability to map sequence reads uniquely (see Figure 2). Thus, an 85% accurate 1,000 bp read will map as well as a 100% accurate 700 bp read.



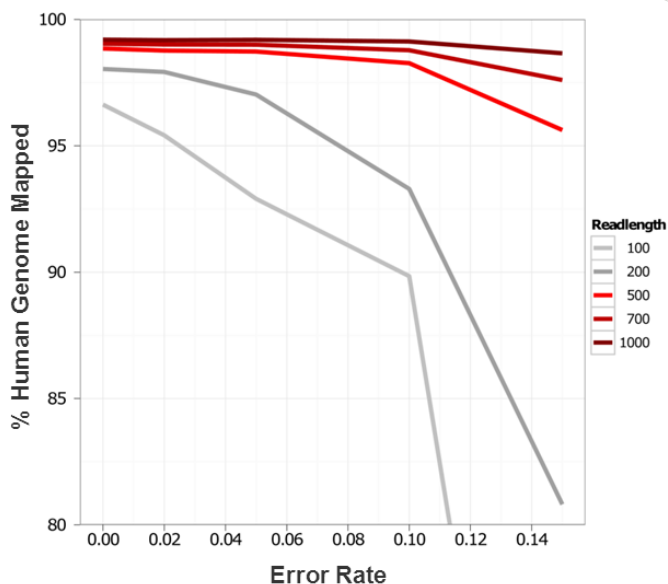


Figure 2: Percentage of the human genome that can be uniquely mapped as a function of read length and error rate. (Sorenson, Jon. AGBT Commercial Session. Marco Island, 2010.)

Long reads offer the additional benefit of allowing for phasing of mutations. Once the long reads are uniquely mapped, both the allele fraction and haplotypes can be determined by correlating the distribution of calls at any given position.¹ Similarly, haplotypes can be discerned by correlating variations found in one location of the sequence to those found in another. For more information about the accuracy of PacBio data, refer to the *Pacific Biosciences® White Paper – Single Molecule Accuracy*.

Using CCS vs CLR Data in Heterozygote SNP Detection

There are two data types generated on the PacBio RS which accommodate different variant detection requirements:

1. Continuous Long Reads (“CLR”) are generated by splitting the raw sequence from a ZMW by the adapters, and is also referred to as a “subread”. A full pass subread is a subread with two observed adjacent adapters.
2. Circular Consensus Sequencing (“CCS”) read is a sequence generated by collapsing multiple CLRs or subreads from a single ZMW to form a single high-accuracy read. CCS data is generated when ≥ 2 full pass subreads are present.

Both CCS and CLR data can be used for accurate variant detection. Figure 3 shows both CCS and CLR data for a ~250 bp amplicon of exon 13 of the EGFR gene. Both CCS and CLR data accurately detect the SNP at position 104. Further, the variant fraction level of the false positive positions (i.e., all template positions that do not include position 104) is < 5% for both data types.

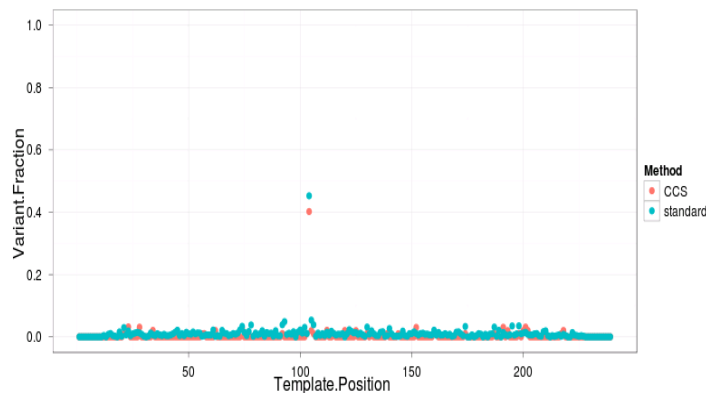


Figure 3. Heterozygous SNP detection at position 104 in EGFR exon 13 using either 100 CCS reads or the 540 CLR reads that comprise them.

CCS data from shorter amplicons (up to 2 kb) are preferred for minor (< 20%) variant detection, as it provides higher per-read accuracy. This is an important consideration in cancer samples, where tumor cellularity can be fairly low and/or the variants heterogeneous with important mutations represented at very low frequencies.

However, reads that do not include at least 2 full passes are discarded in CCS data, but are included in the CLR data. Therefore, CLR data provide more usable reads per SMRT Cell. Additionally, CLR data allow for longer reads from larger amplicons, which can be useful for discerning novel structural variants and haplotyping.

For more information on data types, refer to the *Pacific Biosciences Technical Note – Experimental Design for Targeted Sequencing – SNP Detection and Validation*.

Amplicon Sequencing Workflow

Because the end-products of various enrichment strategies are easily integrated into the standard, automated library construction workflow, SMRT

technology offers a unique opportunity to simultaneously sequence large numbers of amplicons. Input samples can be enriched for specific targets using a variety of methods, from traditional PCR amplification to commercially available approaches, including offerings from Fluidigm, Agilent Technologies, and Raindance (see Table 1). In every case, amplicon purification is followed by less than a few hours of standardized SMRTbell™ library preparation². These amplicon SMRTbell libraries can be put directly into the sequencing workflows appropriate to their insert size.

Table 1: Target Enrichment Strategies Compatible with SMRT Sequencing

Enrichment Technology	Enrichment Method	Target Insert Size	Targeted Region
Traditional PCR	PCR	Up to 10 kb	Flexible
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	≤50 Mb

For more information on the Amplicon sequencing workflow, including preparation and handling of amplicons and sequencing on the PacBio RS, refer to:

- *Pacific Biosciences Technical Note – Experimental Design for Targeted Sequencing – SNP Detection and Validation*
- *Pacific Biosciences Technical Note – Targeted Sequencing on the PacBio RS using Agilent Technologies SureSelect Target Enrichment*
- *Pacific Biosciences Technical Note – Targeted Resequencing EGFR MET Genes*

Analyzing Targeted Sequencing Data

In addition to several third party tools available through the Developers' Network (<http://www.pacbiodevnet.com>),

PacBio offers several tools for targeted sequencing data analysis through our SMRT Portal. There, sequencing reads are aligned to the reference using BLASR (Basic Local Alignment with Successive Refinement) software. In addition, SNP detection and validation can be facilitated using the Broad Institutes' open source Genome Analysis Toolkit ("GATK"). GATK is the Broad Institute's unified genotyper for Bayesian diploid and haploid SNP calling. It is available at http://www.broadinstitute.org/gsa/wiki/index.php/Variant_quality_score_recalibration. GATK has also been specifically modified for integration into the PacBio routine analysis pipeline: the SMRT Portal v1.3 software protocol *RS_Resequencing_GATK* to detect variants.

For more information on analysis tools, see the *Pacific Biosciences Technical Note – Experimental Design for Targeted Sequencing – SNP Detection and Validation*.

Case Studies

Detection and Phasing of Rare Mutations in BCR-ABL Fusion Protein

Secondary Kinase Domain ("KD") mutations are a well-recognized resistance mechanism to Tyrosine Kinase Inhibitors ("TKI") in Chronic Myeloid Leukemia ("CML") and other cancers. In some cases, multiple drug resistant KD mutations can coexist in an individual patient ("polyclonality"). Alternatively, more than one mutation can occur in tandem on a single allele (compound mutations) following response-and-relapse to sequentially administered TKI therapy.

Distinguishing between these two scenarios can inform the clinical choice of subsequent TKI treatment. There is currently no clinically adaptable methodology that offers the ability to distinguish polyclonal from compound mutations. Due to the size of the BCR-ABL KD where TKI-resistant mutations are detected, second-generation platforms are unable to generate reads of sufficient length to determine if two mutations separated by 500 nucleotides reside on the same allele.

The PacBio RS is capable of rapidly and reliably achieving average read lengths of ~1000 bp and frequently beyond 3000 bp, allowing sequencing of the entire ABL KD on a single strand of DNA. This allows for distinguishing polyclonal from compound mutations in

clinical samples obtained from patients who have relapsed on BCR-ABL TKI treatment.

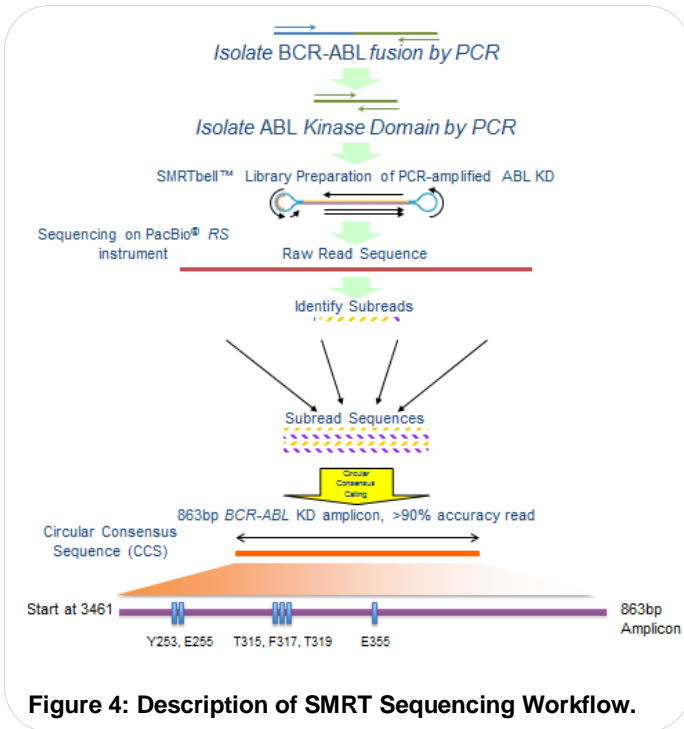


Figure 4: Description of SMRT Sequencing Workflow.

In a joint study with the Department of Medicine, Division of Hematology/Oncology at the University of California San Francisco, approximately 1400 bp of the BCR-ABL fusion protein was amplified by traditional PCR from patient-derived cDNA and isolated by gel-purification. An 863 bp fragment containing the ABL KD was subsequently amplified by nested PCR, and converted into a SMRTbell library using the standard PacBio library preparation process² (Figure 4).

SMRT sequencing accurately, and with high sensitivity, identified BCR-ABL KD mutations in patient samples (see Table 2). SMRT sequencing identified the same mutations found in CLIA-certified direct sequencing assays with equal or improved sensitivity. Frequencies of known mutants in SMRT sequences were all above 15%, suggesting this is the sensitivity of direct sequencing by capillary electrophoresis. By contrast, direct readouts of haplotypes from the full length (863 bp) amplicons revealed minor alleles were detected at frequencies < 5%.

Moreover, SMRT sequencing distinguished polyclonal and compound mutations in individual patient samples not detectable by direct sequencing³ (Figure 5).

Specifically, multiple BCR-ABL KD mutations at distinct AA residues were observed in 3 samples⁴. For each sample, the number of times different mutations were observed together in a single-molecule read were tallied and the simple fractional abundances of each of the mutation combination was calculated.

Although direct sequencing can identify mutations at individual nucleotide positions, it is not quantitative, cannot resolve multiple nucleotide substitutions at a single position, and provides no information about whether mutations occur in cis or in trans.

Table 2: ABL KD Mutations Identified by SMRT Sequencing in Ph+ Patients and Normal Controls at a Frequency of >1%.⁵

Sample	Known Mutant	Detected Variants	Freq	#CCS Reads	CCS Read length
Pt 1	F359I	F359I F359V	47% 6%	10,149	786
Pt 2	None	None	-	8,237	804
Pt 3	T315I F359V	T315I F359V	47% 49%	7,878	762
Pt 4	T315I	T315I	99%	9,411	766
Pt 5	T315I	T315I	99%	7,094	722
Pt 6	None	None	-	10,075	785
Pt 7	F317L	F317L	15%	10,754	762
Pt 8	T315I	T315I	96%	13,444	644
Pt 9	F317L	F317L	96%	13,870	674
Pt 10	T315? E355G	Y253H T315A T315F T315I T319A E355G	8% 60% 19% 11% 7% 78%	13,172	798
Pt 11	E255V T315I	E255V T315I	47% 45%	14,689	769
N 1	None	None	-	11,123	810
N 2	None	None	-	14,510	752
N 3	None	None	-	12482	787

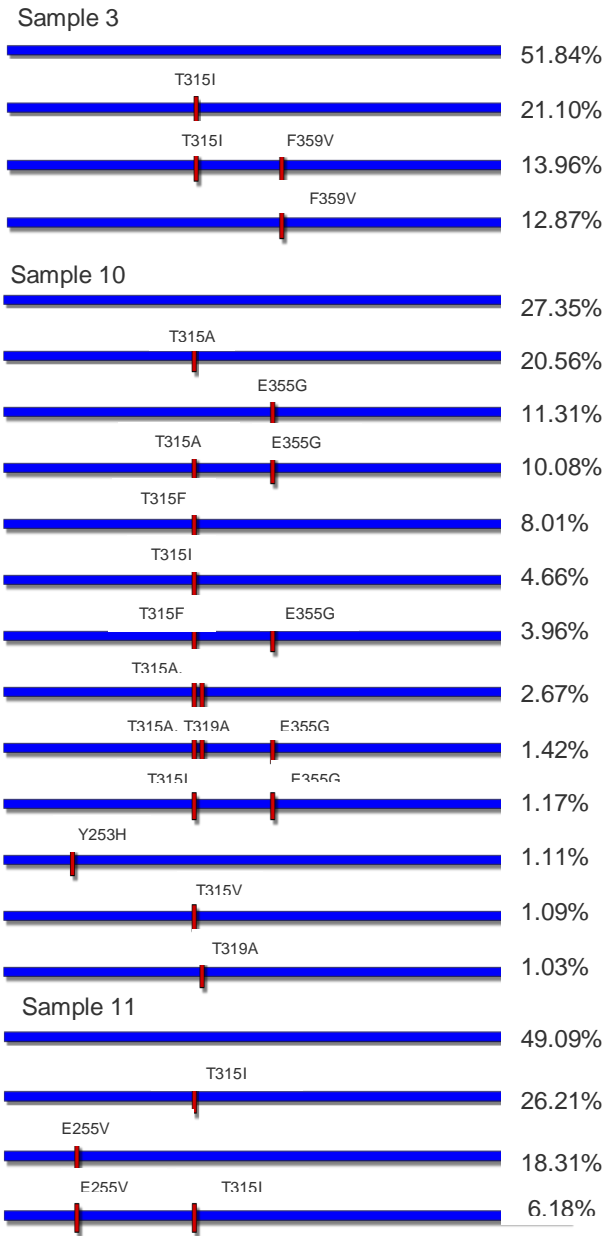


Figure 5: SMRT Sequencing distinguishes Polyclonal and Compound Mutations in Patients⁵.

Sequencing through Highly Repetitive DNA: Expansions of the CGG Repeat in FMR1

Clinical outcomes (phenotypes) related to FMR1 mutations (genotypes) are associated with a trinucleotide repeat expansion (structural variation). The molecular etiology of the disease is due to decreases in Fragile X Mental Retardation 1 (“FMR1”) mRNA

expression levels resulting from the expansion of a CGG repeat within the gene⁶. Alleles are classified as either “premutation” (containing >100 CGG repeats) or “full mutation” (containing >200 – 1500 CGG repeats) genotypes (see Figure 6).

Full mutation genotypes are clinically associated with Fragile X syndrome. Neither premutation nor full mutation allelic genotypes have ever been sequenced by other technologies because of the high GC content of the FMR1 repeat element.

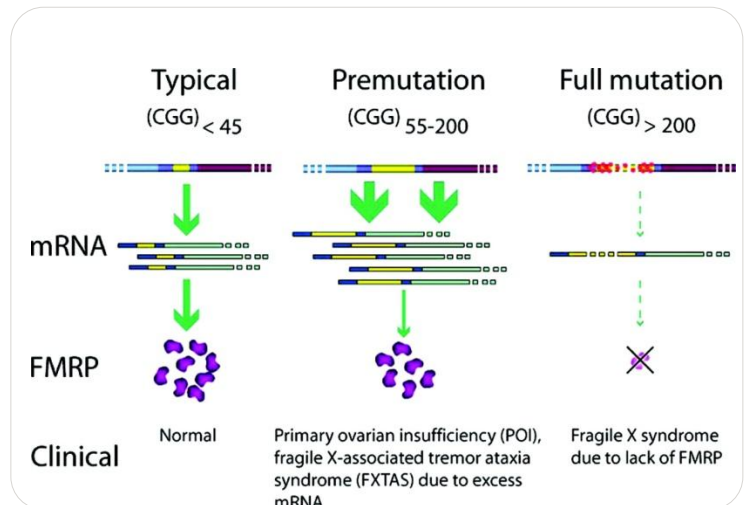


Figure 6: Expansion of CGG Repeat and Link to Clinical Phenotype⁷.

In a joint study with the Department of Biochemistry and Molecular Medicine, at the University of California at Davis, amplicons spanning the FMR1 repetitive region (i.e., including unique sequences flanking both sides of the repeat) were generated by standard PCR methods (Figure 7). The purified amplicons were converted into SMRTbell libraries using the standard PacBio library preparation process². CLR data was generated spanning across the full amplicon and the repeat regions, as well as the unique mapping reference sequences on either side.

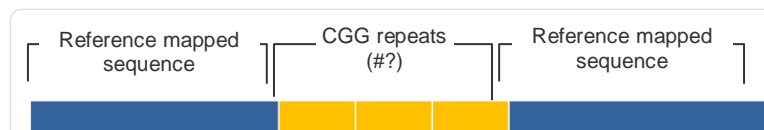


Figure 7: Amplicon structure

The number of tri-nucleotide repeats comprising the repeat region was analyzed to determine pre-mutation and full mutation states for each of the genotypes being studied. Finished sequence reads for each amplicon quantified the exact number of tri-nucleotide repeats. Accurate counting of tri-nucleotide repeats was accomplished by reference-based CCS⁸, using the unique sequences flanking the repeats as anchors mapped to the genomic reference (Figure 8).

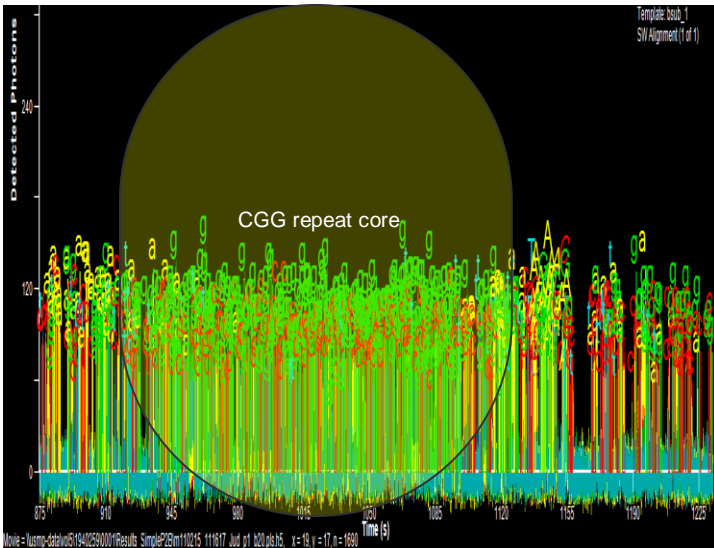


Figure 7: The Entire CGG Repeat Core of the Fragile X Gene (FMR1), Flanked on Either Side by Unique Mappable Sequences.

Conclusion

Long read lengths, high accuracy, simple workflows, and quick turn-around time from sample to sequencing makes SMRT sequencing on the PacBio RS extremely suitable for targeted sequencing applications, including discovery of novel and rare variants, sequencing through highly repetitive regions, and phasing of mutations.

References

1. White Paper, Pacific Biosciences, Utilizing Single Molecule Accuracy, (2011).
2. Kevin J. Travers, Chen-Shan Chin, David R. Rank, John S. Eid, Stephen W. Turner. "Flexible and efficient template format for circular consensus sequencing and SNP detection." *Nucleic Acids Research*, (2010): e159.
3. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences, U S A* 74: 5463-7 (1977).
4. Blood (ASH Annual Meeting Abstracts) 2011 118: Abstract 3752, American Society of Hematology, © 2011.
5. Smith et al; "Single Molecule Real Time (SMRT) Sequencing Sensitivity Detects Polyclonal and Compound BCR-ABL in Patients Who Relapse on Kinase Inhibitor Therapy." *ASH Annual Meeting Abstracts 2011 118: 3752.*
6. Loomis et al; "Sequencing of expanded CGG repeats in the FMR1 gene." Poster session presented at: Presented at the 12th International Congress of Human Genetics/61st Annual Meeting of The American Society of Human Genetics, (October 13, 2011, Montreal, Canada).
7. Hagerman et al; "Advances in the treatment of fragile X syndrome" *Pediatrics* Vol. 123 No. 1 January 1, 2009, pp. 378 - 390.
8. Chin, Chen-Shan. "Reference CCS Algorithm." Internal Document. Menlo Park, CA, February 18, 2010.

For Research Use Only. Not for use in diagnostic procedures. © Copyright 2012, Pacific Biosciences of California, Inc. All rights reserved. Information in this document is subject to change without notice. Pacific Biosciences assumes no responsibility for any errors or omissions in this document. Certain notices, terms, conditions and/or use restrictions may pertain to your use of Pacific Biosciences products and/or third party products. Please refer to the applicable Pacific Biosciences Terms and Conditions of Sale and to the applicable license terms at <http://www.pacificbiosciences.com/licenses.html>.

Pacific Biosciences, the Pacific Biosciences logo, PacBio, SMRT and SMRTbell are trademarks of Pacific Biosciences in the United States and/or certain other countries. All other trademarks are the sole property of their respective owners.

PN 100-092-200-01