Understanding Accuracy in SMRT® Sequencing

Jonas Korlach, Chief Scientific Officer, Pacific Biosciences

Introduction

Single Molecule, Real-Time (SMRT®) DNA sequencing achieves highly accurate sequencing results, exceeding 99.999% (Q50) accuracy, regardless of the DNA's sequence context or GC content. This is possible because SMRT Sequencing excels in all three categories that are relevant when considering accuracy in DNA sequencing:

1. Consensus accuracy
2. Sequence context bias
3. Mappability of sequence reads

This paper details how SMRT Sequencing performs in each of these three areas, and compares its performance to other sequencing technologies.

1. Consensus Accuracy

A typical sequencing project consists of three basic steps: (i) generating sequence reads, (ii) mapping sequence reads to a known reference sequence, and (iii) generating consensus for producing the final sequence result. If the DNA sample is of unknown origin, step (ii) is replaced by a *de novo* genome assembly to generate a new reference, and the final step comprises mapping the original sequence reads to this assembly to verify that the consensus built from the reads is identical to the new assembly sequence.

To understand how SMRT Sequencing achieves results with >99.999% accuracy, **Figure 1** reviews how sequencing results are obtained in second-generation sequencing systems.

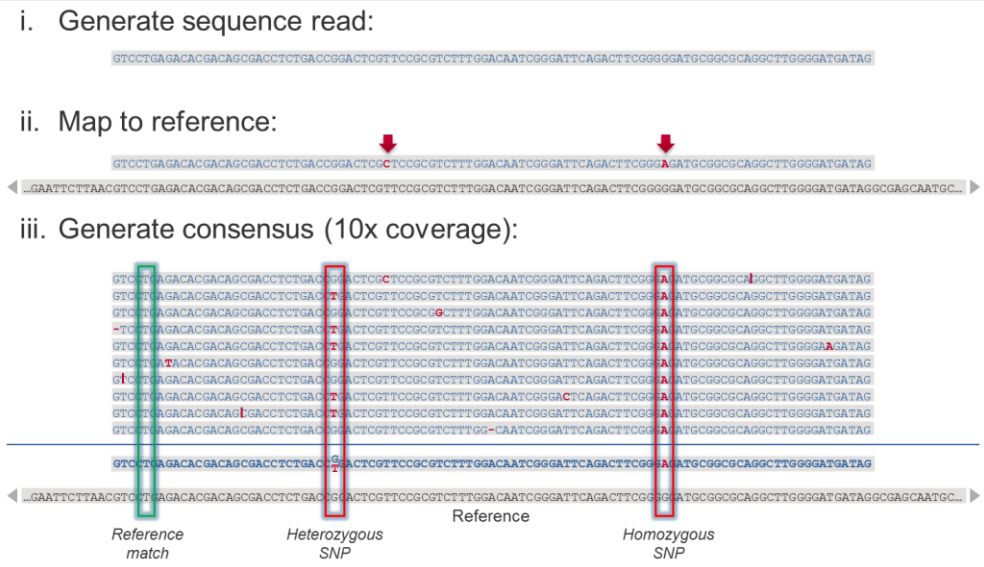

Figure 1. Obtaining sequencing results using second-generation sequencing technologies.

In this example, an individual sequence read of ~120 bases in length is mapped to the chosen reference genome, and the red arrows indicate discordant bases with respect to this reference. Biological conclusions cannot be drawn from this individual read because it is not possible to tell whether those discordant bases represent true biological variation or are caused by sequencing errors. Similarly, it is impossible to call heterozygous SNPs from single reads, as for such variation at least one read from the maternal and one read from the paternal chromosome are necessary. Therefore, **biological insight is gained through averaging the sequences from multiple reads that map to the same region in the reference, in other words, by building consensus.**

In this example, the average sequence information from 10 reads (i.e.,10x coverage) is used to infer for all reference positions whether they match with the reference, represent homozygous SNPs, or represent heterozygous SNPs. **The same principle applies for SMRT Sequencing** (see **Figure 2**).

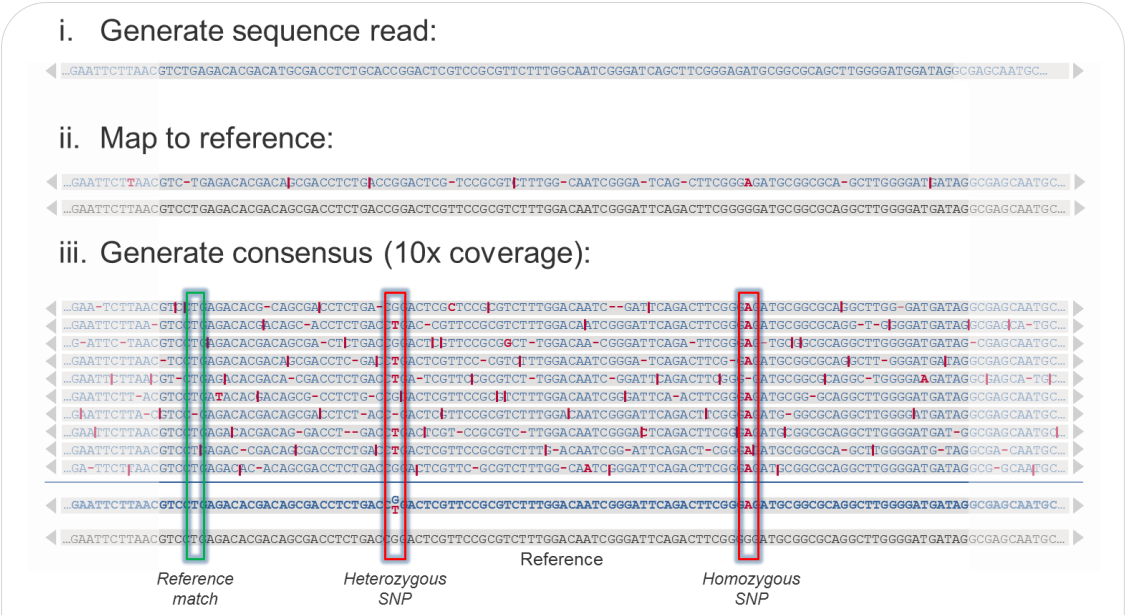

Figure 2. Obtaining sequencing results using SMRT Sequencing.

Reads in SMRT Sequencing are much longer, typically several kilobases, but for the purpose of this discussion the figure highlights the same ~120 bases shown in Figure 1. Single-pass sequence reads in SMRT Sequencing are more error-prone, with a median error of ~11%, and are predominantly either deletions (indicated by red dashes) or insertions (indicated by vertical red lines). Taking into account these characteristics of SMRT-sequencing reads, we have developed a mapping tool called BLASR[1], which has been optimized for mapping SMRT-sequencing reads (for read length effects on mapping, see Mappability of Sequence Reads below). Using BLASR allows confident mapping of SMRT-sequencing reads to their respective locations in the chosen reference, despite the higher single-pass error rate. Then, as is the case for second-generation sequencing systems, results are generated through consensus analysis, again taking 10 reads that map to the same region in the reference, and averaging the sequence information for each reference position vertically. The same accuracy that characterizes an individual read 'horizontally' now applies 'vertically' when building consensus: on average, 9 out of 10 reads will contain a correctly sequenced base, **making it straightforward to arrive at the correct sequencing result**.

Using these characteristic features of SMRT-sequencing reads, we have developed a consensus tool called Quiver that delivers the high-quality consensus sequence (www.pacbiodevnet.com/Quiver).

Systematic error in a sequencing method will affect whether the consensus sequence can be determined correctly – if a base is systematically read incorrectly, it will be called incorrectly in consensus, and this cannot be overcome by adding more sequencing coverage. A key feature in SMRT Sequencing that allows the determination of a >99.999% accurate consensus lies in the fact that the **single-pass errors are distributed *randomly*, which means that they wash out very rapidly upon building consensus**. This has been theoretically and experimentally verified in multiple independent publications[2,3].

Figure 3 shows how accuracy in SMRT Sequencing builds as a function of sequencing coverage.

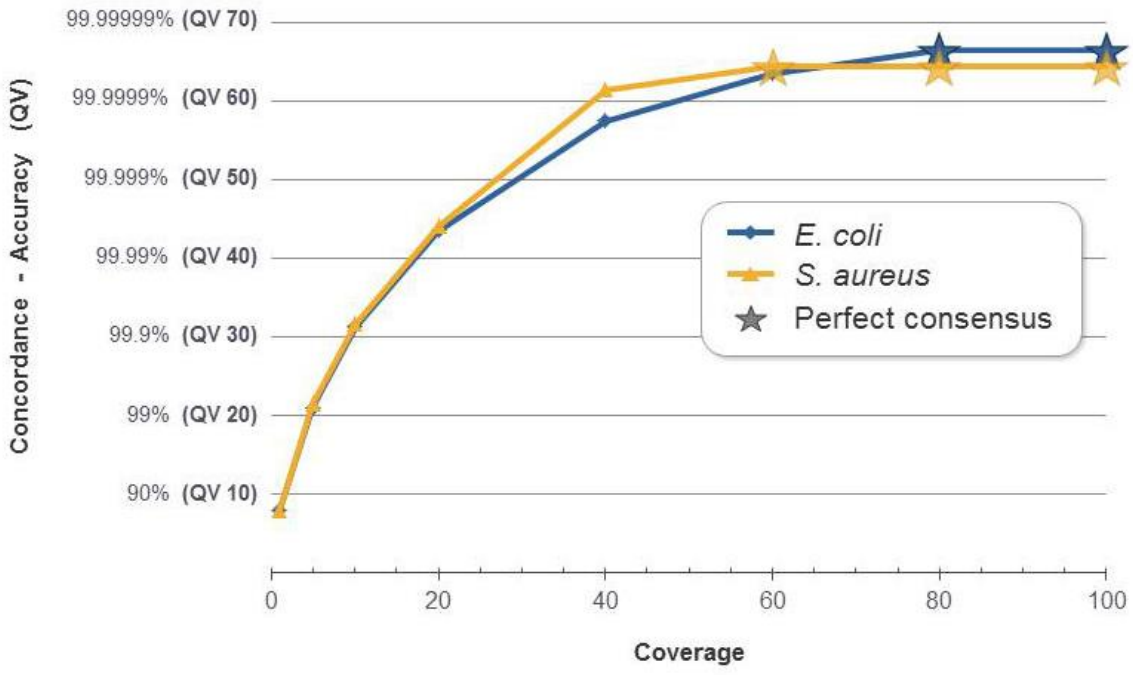

Figure 3. Consensus accuracy as a function of sequencing coverage. The stars indicate 100% identity with the reference.

To judge levels of accuracy in the QV50 range and beyond, the obtained consensus has to be compared to a reference sequence that is precisely known (i.e., having 'gold standard' references). Certain bacterial genomes that have been extensively sequenced with Sanger sequencing can provide such a standard. The figure demonstrates that although SMRT-sequencing reads have lower single-pass accuracy than other methods, very high accuracy is achieved rapidly upon building sequencing coverage, and perfect identity to the reference genome is achieved in many cases. Accordingly, accuracies in SMRT sequencing of less than 1 sequence error in a million bases (>QV60) have been described in the literature (Table 1).

Because of the random nature of the errors, **SMRT Sequencing in many cases exceeds the consensus accuracy achieved by other sequencing methods**. This is the underlying reason why several studies have used SMRT Sequencing to *validate* SNPs that were discovered on other platforms[2,4] – highlighting that it is the consensus result that matters, rather than single-pass error rates.

Organism	Assembled with	Assembly bp	Contigs	N50	LAP	Discordant Bases	QV
E. coli K12	MiSeq 100X 2x150bp 300bp (MaSuRCA iCORN)	4682345	139	113852	9.68E + 07	28	52.23
	454 50X	4569757	93	117490	−9.73E + 07	17	54.29
	PBcR 200X	**4653486**	**1**	**4653486**	**−9.64E + 07**	**3**	**>60**
E .coli O157:H7	MiSeq 100X 2x150bp 500bp (SPAdes iCORN)	5433737	413	133641	−3.67E + 07	62	49.43
	454 22X + 8X 5Kbp + 10X 10Kbp	5347050	409	133665	−3.73E + 07	66	49.09
	PBcR 200X	**5611389**	**9**	**4324437**	**−3.66E + 07**	**0**	**>60**
B. trehalosi	MiSeq 100X 2x150bp 500bp (SPAdes iCORN)	2377594	83	222446	−3.31E + 07	10	53.76
	454 50X	2364704	66	117742	−3.32E + 07	9	54.20
	PBcR 200X	**2411068**	**1**	**2411068**	**−3.27E + 07**	**0**	**>60**
M. haemolytica	MiSeq 100X 2x150bp 500bp (MaSuRCA iCORN)	2721965	89	84094	−3.33E + 07	47	47.63
	PBcR 200X	**2736037**	**1**	**2736037**	**−3.31E + 07**	**0**	**>60**
F. tularensis	MiSeq 100X 2x250bp 500bp (SPAdes iCORN)	1825374	130	24065	−1.33E + 07	0	>60
	454 50X	1655657	326	7316	−1.33E + 07	28	47.72
	PBcR 300X	**1877407**	**3**	**573021**	**−1.33E + 07**	**0**	**>60**
S. enterica Newport	MiSeq 56X 2x150bp 500bp (MaSuRCA iCORN)	5187269	114	195780	−2.24E + 07	360	41.59
	454 23X + 2X 10Kbp	5005089	172	372513	−2.25E + 07	39	51.08
	PBcR 200X	**5029197**	**2**	**4919684**	**−2.24E + 07**	**2**	**>60**

Table 1. Accuracy comparison for different sequencing technologies (last column, PacBio® assembly accuracy in bold, and compared to Illumina® and 454® data, respectively). From Koren et al. (2013) *Genome Biology* 14:R101[5].

Figure 4 shows a representative example for such validation, directly comparing a SNP call made on an Illumina® platform with the corresponding results obtained from SMRT Sequencing, clearly demonstrating the ability of SMRT Sequencing to identify biological variation with confidence.

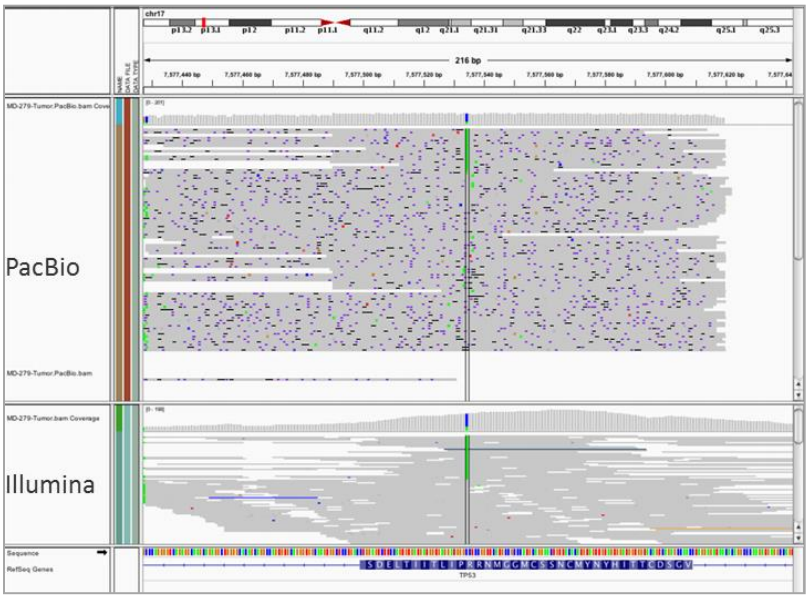

Figure 4. Validation of SNPs using SMRT Sequencing. From Pugh et al. (2012) *Nature* 488: 106-110[4].

2. Sequence Context Bias

The comparisons discussed above only apply to DNA sequences that are actually amenable to sequencing. Many sequencing methods have limitations with respect to certain sequence contexts or extreme base composition that prevents sequencing such regions altogether, and the sequencing accuracy by those technologies through these regions is 0% by definition.

In particular, many sequencing systems have difficulties sequencing through extremely AT-rich or GC-rich DNA regions, highly repetitive sequences, or long homonucleotide stretches, either yielding no sequence or sequence of poor quality. Similarly, palindromic sequences often represent unsequenceable regions on other sequencing systems because this DNA is lost during amplification steps in sample preparation[6]. Lack of coverage for certain genomic regions translates into incomplete sequencing results and fragmented contigs in genome assemblies, sometimes missing as much as 10% of the genome[7,8]. It precludes a comprehensive determination of the DNA sequence in a given sample and the construction of complete, finished genomes.

SMRT Sequencing does not exhibit such sequence context bias and performs very uniformly, even through regions previously considered difficult to sequence. This advantage has been used to close gaps in genomes that were called 'hard stops' for other sequencing systems[9]. An extreme example for demonstrating the lack of bias in SMRT Sequencing was demonstrated by sequencing through thousands of bases of 100% GC content: CGG trinucleotide repeat expansions which cause fragile X syndrome[10] (see **Figure 5**). Similarly, because there is no amplification step in the sample preparation, palindromes can be sequenced without difficulty.

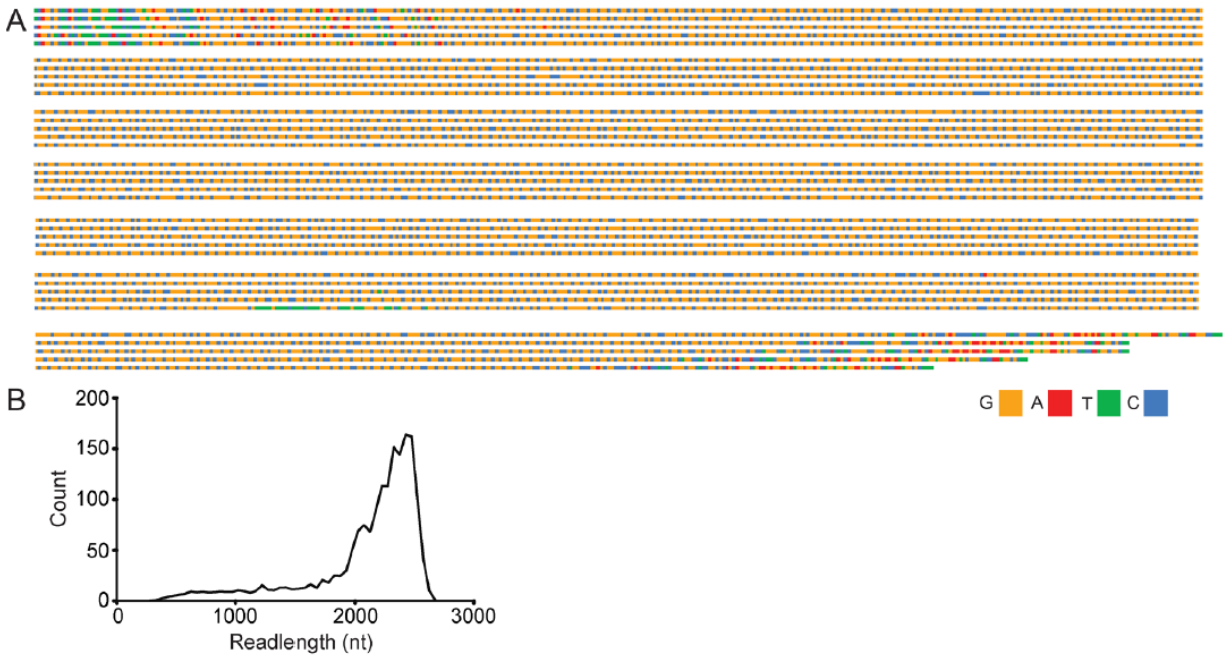

Figure 5. Sequencing the 'unsequencable' with SMRT Sequencing: 100% GC sequence contexts. From Loomis et al. (2012) *Genome Research* 23: 121-128[10].

3. Mappability of Sequence Reads

Even if a sequence read is 100% accurate, it can still be uninformative or even misleading if it cannot be mapped correctly onto the reference genome. This is why read length in sequencing directly impacts sequencing accuracy. If a read is not long enough to span a repetitive region in the genome with at least one unique flanking sequence, the origin of the read cannot be determined unequivocally, and thus any variation observed in this read is ambiguous with respect to where this variation occurred in the genome.

Such incorrect mapping can lead to false assignments of biological variation[2]. **Figure 6** compares the mapping of Illumina reads and SMRT-sequencing reads to a repetitive region in the human genome. In this example demonstrating systematic error during mapping, a population of Illumina reads was mismapped due to their short read lengths, thereby causing false positive SNP calls.

The **long SMRT-sequencing reads avoid mismapping by providing long, multi-kilobase reads that can stretch through repetitive genomic regions and anchor those reads to their correct location**, thereby overturning the false SNP assignments.

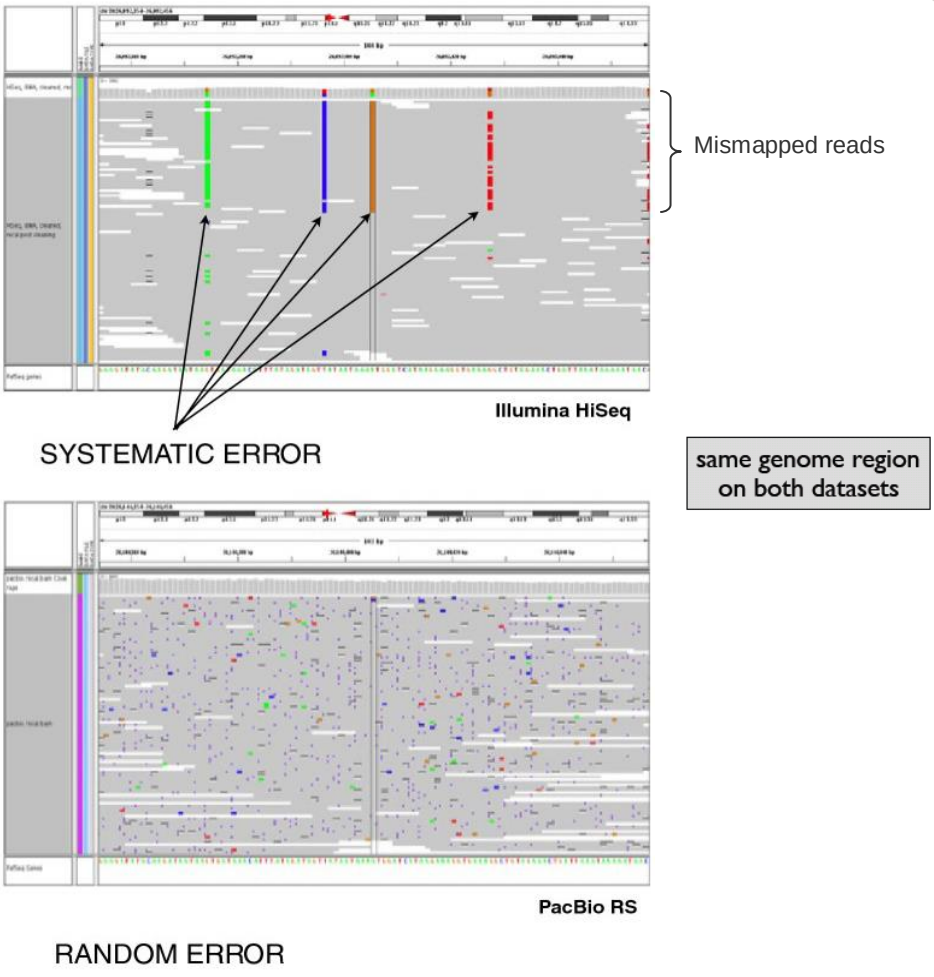

Figure 6. The importance of long read lengths for accurate mapping in sequencing. Adapted from Carneiro et al. (2012) *BMC Genomics* 13: 375-383[2].

Conclusion

Through the combination of high consensus accuracy due to a random error profile, lack of sequence context bias, and very long reads that avoid mismapping artifacts, SMRT Sequencing produces comprehensive and highly accurate sequencing results. The higher rate of single-pass errors is not a problem because they wash out rapidly when building consensus. As highlighted above, it is important to use appropriate bioinformatic tools that take into account the specific characteristics of SMRT Sequencing to obtain the highest quality results.

Appendix: Minor-variant Detection

For applications where very closely related DNA molecules are present in the same sample at low frequencies (e.g., <10%), the highest possible single-read accuracies are required to distinguish them confidently. For these applications, we have developed a sequencing mode that generates *intramolecular* consensus, producing single-molecule sequencing reads that are as accurate, or more accurate than 1st and 2nd generation technologies at their respective read lengths. This is achieved through the circular nature of the SMRTbell™ DNA template[8], which allows the polymerase to sequence the same base of the same DNA molecule multiple times.

SMRT Sequencing is the only technology that allows generation of such an intramolecular consensus, yielding highly accurate base calls on long DNA molecules that are present as a very minor fraction in a sample. This capability has been utilized, for example, to determine low-frequency mutations in acute myeloid leukemia[9]. **Table 2** outlines when such single-molecule consensus mode is suggested, and compares it to all other applications where the standard multimolecule consensus described above should be used.

Multi-Molecule Consensus	Single-Molecule Consensus
Microbial *de novo* assembly	Rare variant detection
BAC *de novo* assembly	• Viral quasi-species
Gap closing/scaffolding	• Rare resistance mutations
Structural-variation detection	• Rare, closely related transcript isoforms
Amplicon sequencing	• Low-frequency heteroplasmy
Gene-panel sequencing	• 16S metagenomics
Transcript-isoform sequencing	• DNA-damage detection
Haplotype phasing	
Base-modification detection	

Table 2. Applications for multi-molecule and single-molecule consensus SMRT Sequencing.

References

1. Chaisson, M.J. and G. Tesler, *Mapping single molecule sequencing reads using Basic Local Alignment with Successive Refinement (BLASR): Theory and Application.* BMC Bioinformatics, 2012. **13**(1): p. 238.

2. Carneiro, M.O., et al., *Pacific biosciences sequencing technology for genotyping and variation discovery in human data.* BMC Genomics, 2012. **13**: p. 375.

3. Koren, S., et al., *Hybrid error correction and de novo assembly of single-molecule sequencing reads.* Nat Biotechnol, 2012. **30**(7): p. 693-700.

4. Pugh, T.J., et al., *Medulloblastoma exome sequencing uncovers subtype-specific somatic mutations.* Nature, 2012. **488**(7409): p. 106-110.

5. Koren, S. et al., *Reducing assembly complexity of microbial genomes with single-molecule sequencing.* Genome Biology, 2013. **14**:R101.

6. Rattray, A.J., *A method for cloning and sequencing long palindromic DNA junctions.* Nucleic Acids Res, 2004. **32**(19): p. e155.

7. Ferrarini, M., et al., *An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome*. BMC Genomics, 2013. 14: 670.

8. Shin, S.-C., et al., *Advantages of Single-Molecule Real-Time Sequencing in High-GC Content Genomes*. PLoS ONE, 2013. 8: e68824.

9. Zhang, X., et al., *Improving genome assemblies by sequencing PCR products with PacBio.* Biotechniques, 2012. **53**(1): p. 61-62.

10. Loomis, E.W., et al., *Sequencing the unsequenceable: Expanded CGG-repeat alleles of the fragile X gene.* Genome Res, 2012. **23**(1): p. 121-128.

11. Travers, K.J., et al., *A flexible and efficient template format for circular consensus sequencing and SNP detection.* Nucleic Acids Res, 2010. **38**(15): p. e159.

12. Smith, C.C., et al., *Validation of FLT3-ITD as a therapeutic target in human acute myeloid leukemia.* Nature. 2012. **485**: p. 260-263.