PACBIO®

Understanding Accuracy in DNA Sequencing

For scientists who utilize DNA sequencing in their research but are not experts in the underlying technology, it can be difficult to determine the accuracy of sequencing results – and even harder to compare accuracy across sequencing platforms. Furthermore, accuracy differs not only between technologies but also across genomic regions as some stretches of the genome are inherently more difficult to read.

It is critically important to understand accuracy in DNA sequencing to distinguish important biological information from sequencing errors.

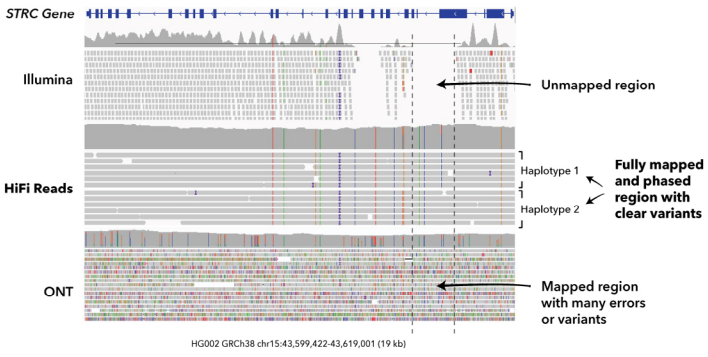
What are the Types of Sequencing Accuracy?

There are two key types of accuracy in DNA sequencing technologies: read accuracy and consensus accuracy.

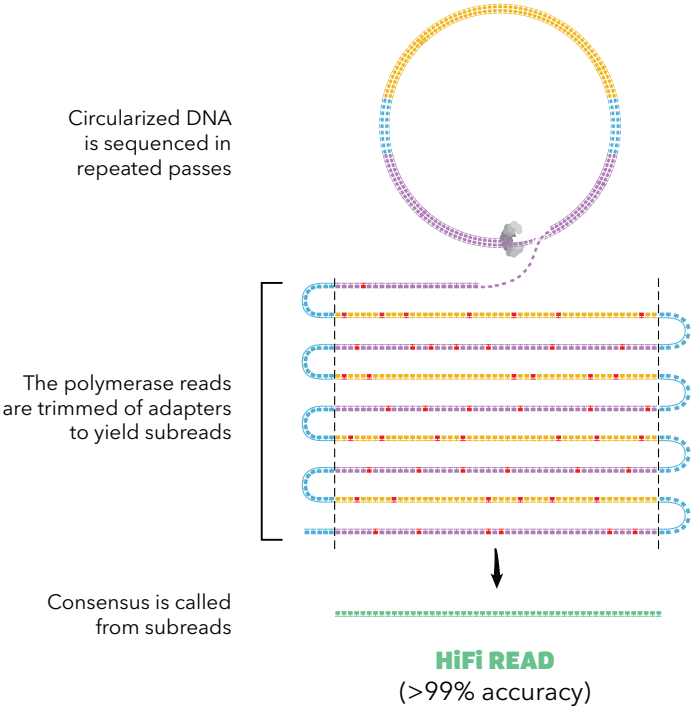
Read Accuracy is the inherent error rate of individual measurements (reads) from a DNA sequencing technology. Typical read accuracy ranges from ~90% for traditional long reads to >99% for short reads and HiFi reads.

Consensus Accuracy, on the other hand, is determined by combining information from multiple reads in a data set, which eliminates any random errors in individual reads. Deeper coverage, meaning more reads from which to build a consensus, generally increases the accuracy of results. However, there are still limitations to calling consensus from multiple reads. Consensus calculation is a complicated and computationally expensive process, and it cannot overcome systematic errors. If a sequencing platform consistently makes the same mistake, then it will not be erased by generating more sequencing coverage.

To sidestep this problem, it is common to "polish" long reads that have systematic errors with high accuracy short reads. However, because of their read length, short reads cannot always map to the long reads unambiguously, limiting their ability to improve accuracy. In general, consensus is improved – and vastly simplified – by starting with highly accurate reads with no systematic biases.



HiFi reads provide the accuracy needed to call single nucleotide variants, while improving mappability and enabling phasing with no systematic bias. STRC gene alignments from Genome in a Bottle (GIAB), HG002_NA24385_son. (IGV settings)



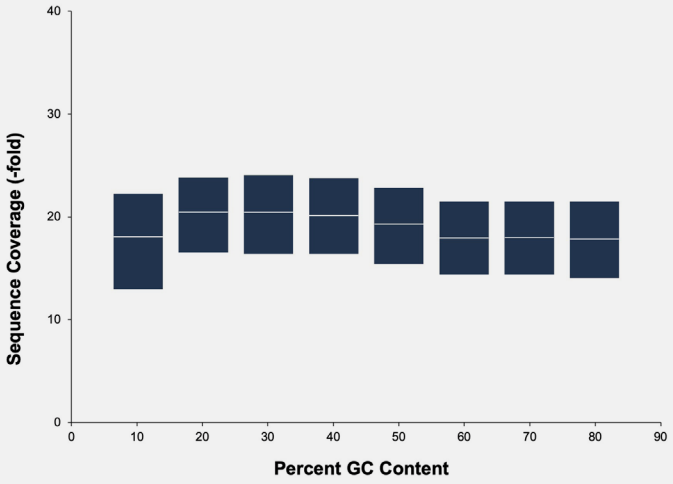
HiFi reads are generated by combining multiple consecutive observations of a DNA molecule (subreads), driving the accuracy of individual HiFi reads over 99%.

How Does Accuracy Impact the Utility of Sequencing Data

1 Coverage Uniformity

It is commonly known that certain genomic regions are more difficult for sequencers to get through than others. Centromeres and telomeres are notoriously tough because of the highly repetitive sequence they contain. Regions that are AT-rich or GC-rich are similarly difficult because they respond poorly to the amplification protocols required by some platforms. Palindromic sequences or hairpin structures are difficult to denature, making such regions challenging for sequencing tools that include a denaturation step.

Many scientists avoid these problems by opting for a single-molecule sequencing method that does not require amplification or denaturation, such as PacBio's SMRT Sequencing technology. Because SMRT Sequencing can process even difficult regions, performing uniformly regardless of sequence context, it generates accurate results even in regions that would flummox other platforms. Selecting a platform without systematic bias, like the Sequel Systems, is important to producing the most accurate sequence data.

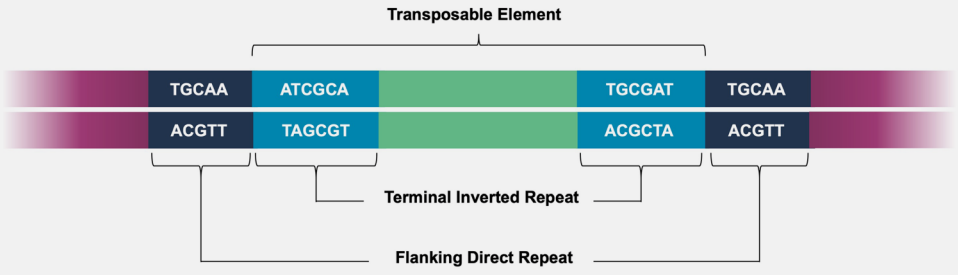


HiFi reads generated with Single Molecule, Real-Time (SMRT) Sequencing technology provide uniform coverage regardless of GC content.

2 Mappability

The accuracy of a genome assembly goes beyond the accuracy of each individual base. Even perfect reads can contribute to poor accuracy if they are not ordered and oriented correctly in the assembly. This question of where to place the read is called mappability.

Reads containing only a piece of a large structural element, or consisting of highly repetitive sequences, can be very difficult to align, mapping ambiguously to many different locations in a reference. This is where short reads really struggle; because of their size, there is a greater chance that they will not contain enough unique sequence data to anchor them properly in a genome. Since HiFi reads stretch across many kilobases of DNA, they almost always contain unique flanking sequences that can be used to map them accurately in an assembly.



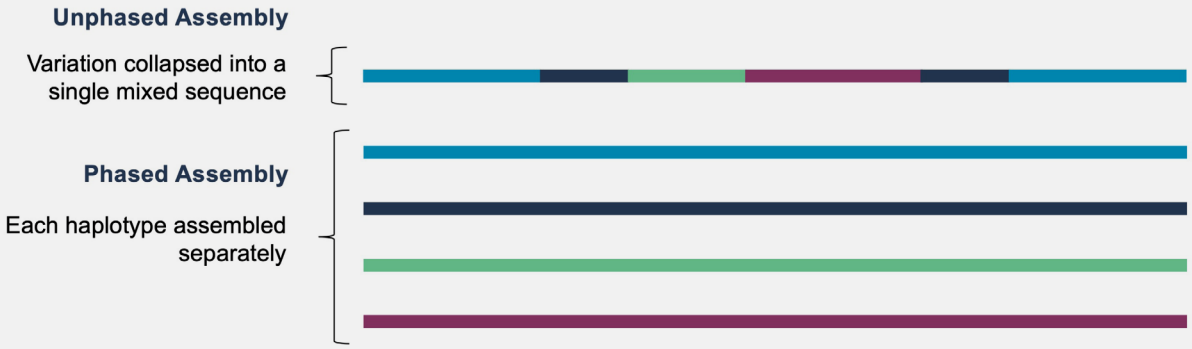
HiFi reads span repetitive repeats increasing mappability.

3 Phasing

When exploring diploid or polyploid genomes, phasing means separating the different copies of each chromosome (e.g. maternal and paternal for diploid), known as haplotypes. With sufficient accuracy, the identity of nucleotides at each position in the genome can be compared with a reference sequence to identify SNVs, with a heterozygous locus indicating a difference in sequence between a homologous chromosome pair. This is where the inherent low accuracy of traditional error prone long reads becomes a limitation – with a high error rate, it makes it impossible to decide whether a disagreement between a reference and data set is a variant or a sequence error.

Another approach to obtain phase information is to also sequence the parents of the individual whose genome you need phased. However, in many wild species where the parents aren't available, a highly accurate long-read sequencing approach, like HiFi sequencing, would be simpler. There are also computational methods (learn about Nighthawk) or the use of population haplotype frequency information to infer phasing.

Overall, phased genomes or variant calls are higher quality than haplotype collapsed versions as they provide allelic information, which can be important for studying human diseases, crop improvement, evolution, and more. HiFi reads, with accuracy high enough to detect SNVs and read lengths to detect these SNVs over many kilobases, generate larger phased haplotype blocks.



Phasing involves separating maternally and paternally inherited copies of each chromosome.

As scientists analyze more and more genomic data, the role of sequence accuracy will likely only become more important. HiFi reads offer the benefits of high accuracy equivalent to short-read sequencing data, but with the length necessary for complex genome assemblies and phasing of variants across large swaths of the genome.

Learn more about HiFi sequencing at pacb.com/HiFi

Get in touch with a PacBio scientist to scope out your project.

Connect with a PacBio Scientist

PN: IG102-091820