The Advantages of Long-Read Sequencing

The scientific community appears to be in agreement that NGS, the next step in sequencing, has the potential to provide vast and varied medical advantages

Jonas Korlach at Pacific Biosciences

DNA sequencing technologies have contributed to drug discovery and development pipelines for decades, but more recent innovations offer to significantly expand the applications for which sequencing is an appropriate choice.

While next-generation sequencing (NGS) instruments were essential for making sequencing more affordable, the short reads they produce are not universally useful. They have been excellent for applications such as genotyping or discovering single nucleotide variants associated with disease. However, for many applications, scientists need reads long enough to span whole genes or complex genomic elements such as disease-causing repeat expansions or structural variants.

Improvements in long-read sequencing technology have addressed this unmet need. Single-molecule, real-time (SMRT) sequencing can produce reads that are tens of thousands of bases long, allowing for complete coverage of large amplicons or genomic elements, alternative splicing isoforms, and more.

The impact of long-read sequencing was first seen in genome assemblies. Whether these reads were used alone or in combination with short-read data, they allowed scientists to boost the contiguity and completeness of draft assemblies. Today, long-read sequencing platforms are considered the gold standard for *de novo* assembly of microbial genomes and are also becoming the preferred option for plant and animal genomes. These technologies have produced the most contiguous de novo human genome assemblies available.

Now, long-read sequencing has been implemented for a wide range of applications, including many that are relevant to drug discovery and development scientists. From CYP2D6 and human leukocyte antigen (HLA) typing to discovering disease-causing elements, to providing quality control for gene therapy or CRISPR pipelines, long-read sequencing has much to offer the pharmaceutical field.

Drug Metabolism Profiling

About a quarter of all commonly used drugs are metabolised by the enzyme encoded by CYP2D6, giving this gene enormous importance for pharmacological purposes. Variants found in CYP2D6 can affect how patients will respond to medications such as antipsychotics and painkillers; therefore, understanding each person's natural variation in this gene can guide dosing decisions and inform analysis of clinical trial results.

While the importance of analysing CYP2D6 variants is clear, doing so has proved challenging. It is not only highly polymorphic – researchers have already characterised more than 100 allelic variations – but it also has a nearby pseudogene with high homology. Most of the interrogation of the CYP2D6 gene is performed with genotyping assays, but these detect a fraction of possible alleles and cannot be used for novel findings. The genetic complexity of CYP2D6 also makes it very difficult to resolve using short-read or Sanger sequencing.

Scientists have turned to SMRT sequencing to resolve the CYP2D6 gene. The long sequencing reads obtained from this sequencing technology can fully span the gene and its pseudogene, providing complete information that cannot be obtained with other methods. One group of scientists evaluating this approach demonstrated that not only did it accurately characterise the full gene – in some cases even generating information that allowed them to correct previous analyses of the same sample using other technologies – but it also proved useful for discovering novel variants that will make it possible to expand the catalogue of known CYP2D6 alleles (1). Such information should ultimately lead to improved prediction of a patient's drug metabolism profile.

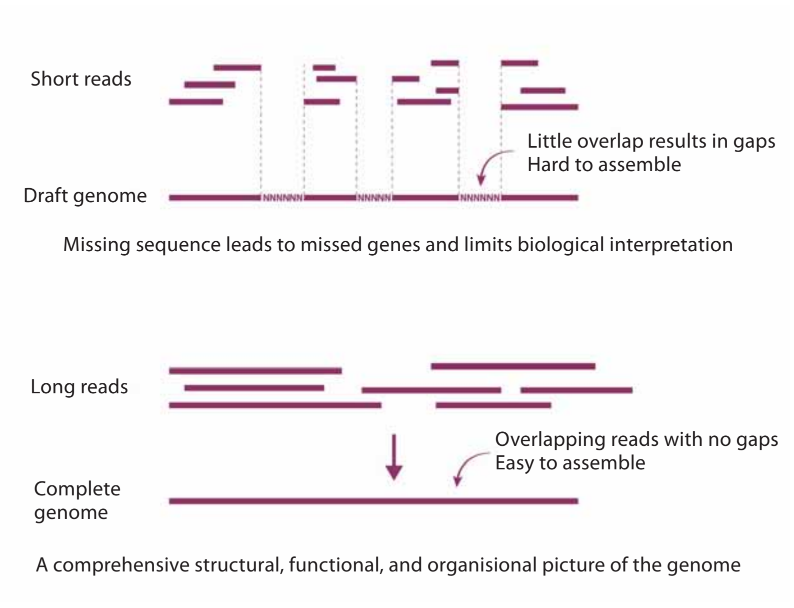

Short reads
Little overlap results in gaps
Hard to assemble
Draft genome
NNNNNN
NNNNN
NNNNNN
Missing sequence leads to missed genes and limits biological interpretation
Long reads
Overlapping reads with no gaps
Easy to assemble
Complete genome
A comprehensive structural, functional, and organisional picture of the genome

Figure 1: Short versus long reads

HLA Typing

The HLA genes play a pivotal role in immune response. Like CYP2D6, these genes are highly polymorphic and difficult to analyse using many standard genomic technologies. HLA genes are used to assess donor-recipient tissue matches prior to organ transplantation, as well as other immune-related traits; therefore, detecting the variation they harbour is critical.

Historically, HLA typing has been performed using genotyping and other methods that scan for specific, known variants in this gene family. Recent studies have shown that fully sequencing these genes – a feat that is now possible with long sequencing reads that cover the entire region – offers a higher-resolution view of the genes, which is expected to result in better matches between organ donors and recipients in the future.

For example, scientists at the Anthony Nolan Research Institute demonstrated that fully sequencing the HLA genes made it possible to identify stronger matches associated with longer survival rates in patients who received haematopoietic cell transplants (2). Their retrospective study covered nearly 900 donor/recipient pairs that had previously been rated as a perfect match, or a 12/12 score for all six HLA genes. Reanalysis of these pairs with SMRT sequencing detected additional variation, finding that 29% of the pairs were, in fact, not perfect matches. The pairs whose perfect matches were confirmed by long-read sequencing had much better five-year overall survival rates (54.8%) compared to the pairs that were newly determined to have been imperfectly matched (30.1%).

This information may prove critical for transplant patients. According to the study, patients with a perfect HLA match were less likely to die from transplant-related complications, such as graft-versus-host disease, in the first year than patients with any degree of mismatch.

Structural Variant Discovery

NGS technologies have been instrumental in improving the diagnosis rate for rare diseases, primarily by identifying disease-causing single nucleotide variants. Despite these successes, the solve rate from exome sequencing for many different cohorts and disease types remains around 40%. There is a considerable amount that must still be learned to shed light on the 60% of cases that cannot yet be explained.

Some of that genomic dark matter can be attributed to structural variants (SVs) – large and often complex genomic elements, at least 50 base pairs long, that can escape detection by short-read sequencing tools. Accurate SV detection requires reads long enough to span the element, whether it's an inversion, a translocation, duplication, or a large indel. SVs also include repeat expansions, which can grow to hundreds of bases or longer.

Long-read sequencing has enabled the reliable detection of SVs, and researchers have now found many examples of pathogenic SVs to explain diseases that have long gone unsolved. Efforts are underway around the world to establish more comprehensive catalogues of disease-causing SVs. In addition to their potential diagnostic utility some of these variants could have potential as new drug targets.

Newly discovered pathogenic SVs have been found to cause a range of diseases and conditions, including certain types of epilepsy, Carney complex, myotonic dystrophy, X-linked intellectual disability, and many others (3-7). In one particularly exciting project, scientists analysed variants associated with a common form of Alzheimer's disease; their findings not only led to the first proof of gene recombination in the brain, but also suggested that existing antiretroviral therapies may help protect against the neurodegenerative disease (8).

Quality Control

At various points in the drug discovery and development pipeline, long-read sequencing can be used for quality control to help scientists ensure that experimental results are as expected, or that certain types of therapies are optimally manufactured for patients.

> Scientists have now demonstrated the utility of SMRT sequencing for evaluating results of CRISPR editing experiments for a more accurate and comprehensive view of the edits made

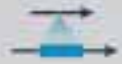
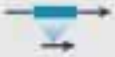
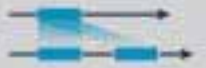
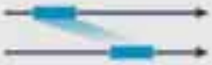
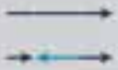
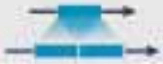
Structural variant	Disease examples
Insertion 	Charcot-Marie Tooth disease, Tay-Sachs disease
Deletion 	Williams syndrome, Duchenne muscular dystrophy, Smith-Magenis syndrome, Carney Complex
Interspersed Duplication 	APP in Alzheimer's disease, Potocki-Lupski syndrome, Prader-Willi syndrome, Angelman syndrome
Translocation 	Down syndrome, XX male syndrome (SRY), schizophrenia (chr11), Burkitt's Lymphoma
Inversion 	Hemophilia A, Hunter Syndrome, Emery-Dreifuss muscular dystrophy
Tandem Duplication 	FMR1 in Fragile-X, Huntington's disease, Spinocerebellar ataxia

Figure 2: Structural variants of all types are known to cause Mendelian disease and contributes to complex diseases

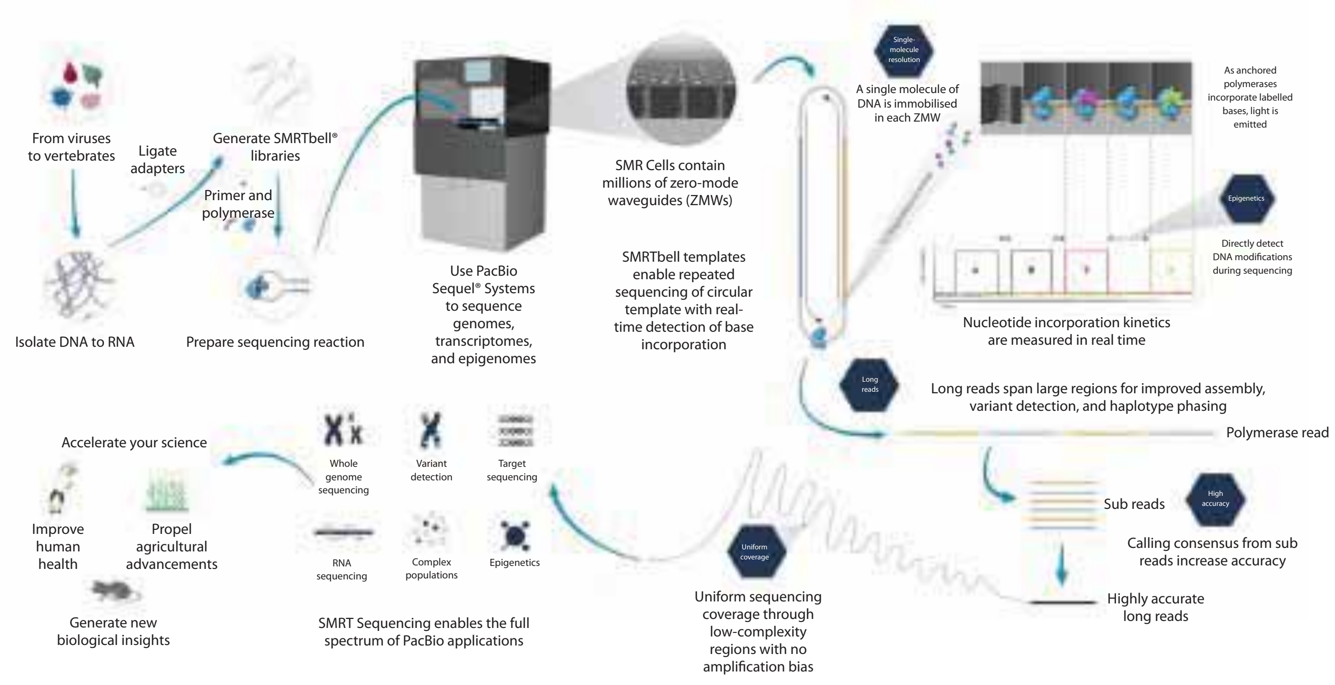

From viruses to vertebrates
Ligate adapters
Generate SMRTbell® libraries
Primer and polymerase
Isolate DNA to RNA
Prepare sequencing reaction
Use PacBio Sequel® Systems to sequence genomes, transcriptomes, and epigenomes
SMR Cells contain millions of zero-mode waveguides (ZMWs)
SMRTbell templates enable repeated sequencing of circular template with real-time detection of base incorporation
Single-molecule resolution
A single molecule of DNA is immobilised in each ZMW
As anchored polymerases incorporate labelled bases, light is emitted
Epigenetics
Directly detect DNA modifications during sequencing
Nucleotide incorporation kinetics are measured in real time
Long reads
Long reads span large regions for improved assembly, variant detection, and haplotype phasing
Polymerase read
Sub reads
High accuracy
Calling consensus from sub reads increase accuracy
Highly accurate long reads
Uniform coverage
Uniform sequencing coverage through low-complexity regions with no amplification bias
Accelerate your science
Improve human health
Propel agricultural advancements
Generate new biological insights
Whole genome sequencing
Variant detection
Target sequencing
RNA sequencing
Complex populations
Epigenetics
SMRT Sequencing enables the full spectrum of PacBio applications

Figure 3: How SMRT sequencing works

> **“Long-read technologies continue to improve, delivering increased read lengths and accuracy as software, chemistry, and hardware are upgraded”**

The rise of CRISPR gene-editing methods has been of tremendous interest in drug discovery, but their reliability has been limited by unwanted on-target and off-target edits that can be made by the nuclease. Scientists have now demonstrated the utility of SMRT sequencing for evaluating results of CRISPR editing experiments for a more accurate and comprehensive view of the edits made. In one study, they performed amplicon sequencing of the CRISPR-targeted region, including significant lengths of flanking sequence, and discovered far more undesired on-target changes than expected – including large deletions, insertions, and complex rearrangements (9). Performing post-experiment quality control with genotyping or short-read sequencing technologies would likely have missed these unwanted edits because of their complex nature. The use of long-read sequencing, on the other hand, would allow scientists to analyse a population of cells to determine which ones had only the anticipated edits for follow-on testing.

Downstream in the drug development workflow, the resurgent interest in gene therapies has created a need to ensure that each therapy contains the specific gene sequence of interest. Unfortunately, investigations have shown that the viral vectors used to deliver gene therapies can carry unexpected cargo. One team of scientists used SMRT sequencing to analyse the adeno-associated virus vectors produced by a gene therapy pipeline and discovered a host of problems: many vectors contained less than half of the needed DNA sequence, while others harboured chimeras and other genetic errors (10). Together, these issues would have reduced the likelihood of a positive response to the therapy. Researchers involved in this study released their SMRT sequencing protocol and encouraged anyone involved in creating gene therapies to implement some kind of long-read-based quality control step.

Looking Ahead

Long-read sequencing is already useful at many points along the continuum of drug discovery and development. In some cases, it serves as a helpful complement to Sanger or short-read sequencing platforms; in other situations, it creates entirely new opportunities. Long-read technologies continue to improve, delivering increased read lengths and accuracy as software, chemistry, and hardware are upgraded. Between technical innovation and scientific creativity, there is no doubt that long-read sequencing tools will ultimately be adopted for many other applications relevant to biopharma scientists.

References

1. Qiao W *et al*, Long-read single molecule real-time full gene sequencing of cytochrome P450-2D6, *Hum Mutat* 37(3): pp315-23, 2015
2. Mayor N *et al*, Recipients receiving better HLA-matched hematopoietic cell transplantation grafts, uncovered by a novel HLA typing method, have superior survival: A retrospective study, *Bio Blood Marrow Transplan* 25(3): pp443-50, 2019
3. Mizuguchi T *et al*, Detecting a long insertion variant in SAMD12 by SMRT sequencing: implications of long-read whole-genome sequencing for repeat expansion diseases, *J Hum Genet* 64(3): pp191-97, 2018
4. Mizuguchi T *et al*, A 12-kb structural variation in progressive myoclonic epilepsy was newly identified by long-read whole-genome sequencing, *J Hum Genet* 64(5): pp359-68, 2019
5. Merker J *et al*, Long-read whole genome sequencing identifies causal ttructural variation in a mendelian disease, *Genet Med* 20(1): pp159-63, 2016
6. Cumming S *et al*, De novo repeat interruptions are associated with reduced somatic instability and mild or absent clinical features in myotonic dystrophy type 1, *Eur Hum Gene* 26(11): pp1635-647, 2018
7. Zablotskaya A *et al*, Mapping the landscape of tandem repeat variability by targeted long read single molecule sequencing in familial X-linked intellectual disability, *BMC Medical Genomics* 11(1):p123, 2018
8. Lee, MH *et al*, Somatic APP gene recombination in Alzheimer’s disease and normal neurons, *Nature* 563(7733): pp639-45, 2018
9. Kosick M *et al*, Repair of double-strand breaks induced by CRISPR–Cas9 leads to large deletions and complex rearrangements, *Nat Biotechnol* 36(8):pp765-71, 2018
10. Tai P *et al*, Adeno-associated virus genome population sequencing achieves full vector genome resolution and reveals human-vector chimeras, *Mol Ther - Methods Clin Dev* (9):pp130-41, 2018

About the author

Jonas Korlach PhD is Chief Scientific Officer of Pacific Biosciences. He is the recipient of multiple grants, an inventor on 70 issued US patents and 61 international patents, and an author of more than 100 scientific studies on the principles and applications of SMRT technology, including publications in Nature, Science, and PNAS. He received both his PhD and his MS degrees in Biochemistry, Molecular and Cell Biology from Cornell University, and received MS and BA degrees in Biological Sciences from Humboldt University in Berlin.