the Pathologist

DIAGNOSTICS |
Genetics and epigenetics, Biochemistry and molecular biology, Omics, Technology and innovation

SMRT Long-Read Sequencing Solves Genetic Mysteries

Solving rare disease with single molecule, real-time sequencing

Luke Hickey | 05/08/2020 | Longer Read

Technologies for diagnosing rare genetic disorders are rapidly advancing. Next-generation sequencing can identify many, but not all such disorders. A new approach – SMRT sequencing – uses longer reads and can identify previously undetectable mutations.

We are in a golden age of rare disease research. Never before have our laboratory techniques been so successful at identifying rare diseases and elucidating their underlying biological causes. The knowledge we obtain today opens the door to new treatments, giving hope to people who suffer from these rare disorders.

Many of the recent advances in rare disease research stem from technology innovations in DNA sequencing. Falling costs have increased access to whole exome and whole

genome sequencing as tools to assess the genetic basis of individual rare disease cases. And in a relatively short time, the genome-scale data these methods produce has transformed the community's understanding of how these diseases arise through rare genetic mutations. There are now more than 7,000 known rare and Mendelian genetic diseases identified – with more added to the databases every year – providing an invaluable information resource for genome-wide screening and exploration.

> ***Solve rates range from 25 to 50 percent, leaving many individuals and their clinical teams without an answer to end the diagnostic odyssey.***

But even with these next-generation sequencing (NGS) tools, clinical research teams have been unable to explain the genetic basis behind a large percentage of rare disease cases. Solve rates range from 25 to 50 percent, leaving many individuals and their clinical teams without an answer to end the diagnostic odyssey. With an estimated 400 million people worldwide affected by a rare disease, there is still a large underserved population and a pressing need to improve our diagnostic yield.

To that end, scientists have begun deploying a higher-resolution DNA sequencing technology known as single molecule, real-time (SMRT) sequencing. SMRT sequencing differs from previous NGS tools by providing longer reads and even higher accuracy (1). In just the past few years, researchers have used SMRT whole genome sequencing to solve previously intractable rare diseases – and other significant efforts are now underway.

The long and the short of it

NGS tools use a variety of approaches to generate sequence data. What they have in common, though, is that they all produce short-read data. Massively parallel short-read sequencing platforms have a low cost per run, but generate sequence reads that are typically only 50–350 base pairs long. To identify genetic abnormalities, these short reads are either mapped to a reference genome or bioinformatically "stitched back together" in a complex assembly process.

Short reads are useful for detecting certain variant types known to occur in the human genome, such as single nucleotide variants (SNVs) and insertions or deletions (indels) less than 10 base pairs long. But, for larger variants, short reads are of limited utility.

The challenge lies in mapping across larger structural variants and indels; for instance, those that occur in repeat expansion disorders, such as fragile X syndrome, amyotrophic lateral sclerosis, and schizophrenia. Mapping issues associated with read length also limit short-read sequencing's ability to call variants across an entire genome. This includes variants in 193 medically relevant genes in the exome. For example, a 200-base sequence read may align to many different regions of the reference genome, leading to sequence gaps and conflating similar regions, such as pseudogenes, repetitive regions, and mobile elements (see Figure 1).

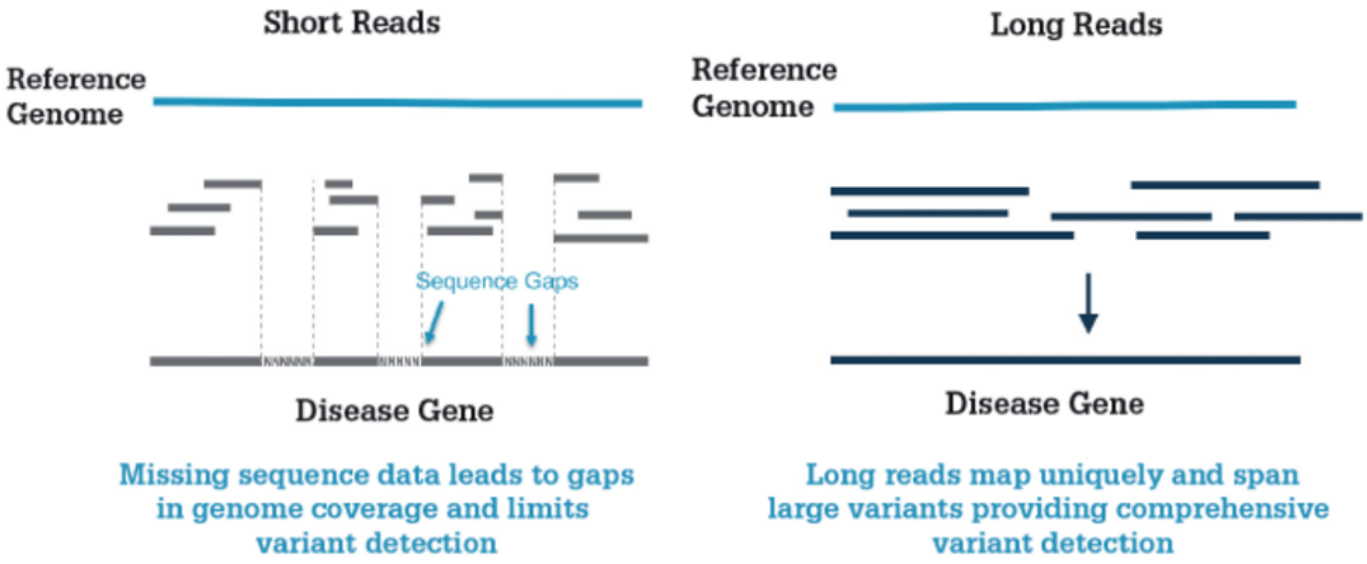

Figure 1. Short-read sequencing produces reads of 50–350 bp, which can lead to sequence gaps and incomplete coverage of disease-causing gene regions. Long-read sequencing produces reads tens of kilobases long, providing high-quality mapping across a genome for comprehensive variant detection.

Short reads also tend to defy variant phasing efforts, making it impossible to distinguish maternally from paternally derived haplotypes. Why does this matter? It can be important when determining whether an individual has a functional copy of a gene in cases where multiple mutations are present. Perhaps most importantly, short-read sequence information often misses the structural variants (50 base pairs or longer; see Figure 2) that comprise most of the sequence variation between any two individuals'

genomes (2). These larger variants can be called incorrectly or excluded entirely from genomes sequenced with short-read data alone.

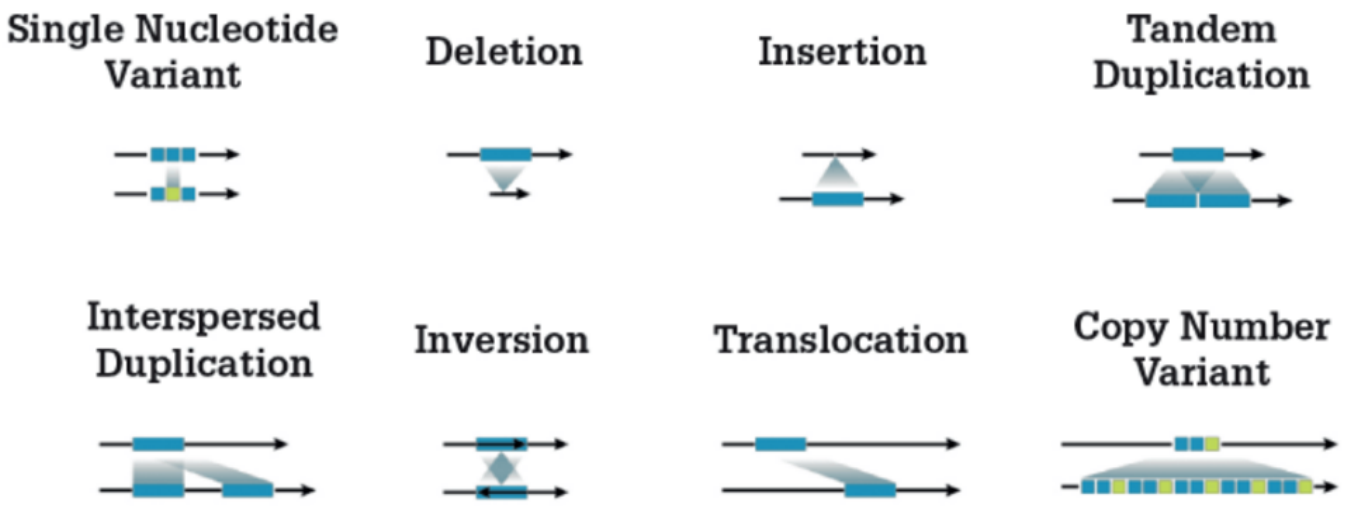

Figure 2. The types of sequence variants found in a human genome. Variants range in size from 1 bp (single nucleotide variant), to >50 bp for larger structural variants such as deletions, insertions, duplications, inversions translocations, and copy number variants (3).

In contrast, longer individual reads can fully cover even large structural variants, removing assembly ambiguity problems and revealing several times more structural variants, with higher precision and recall, than short reads (3).

Genomic dark matter

As short-read systems became more affordable, scientists were eager to apply these new genetic tools to rare disease cases. Exome sequencing and, eventually, whole genome sequencing with these platforms turned out to be a game-changer. Diseases that had long resisted explanation were suddenly understandable thanks to DNA sequence data. It seemed that there was finally an approach that could discover as-yet unknown pathogenic variants in disease-causing genes, giving affected families long-sought answers.

Certainly, short-read NGS tools have significantly increased the diagnostic yield for rare disease cases, but they cannot provide answers for every situation. In fact, scientists estimate that NGS platforms leave more than half of rare disease cases still unsolved.

Could these remaining diseases be caused by something other than genetic mutations? Unlikely, given what we already know about rare diseases; based in part on hereditary patterns and syndromic features, the vast majority appear to be driven by genetic mechanisms (4). What seems more likely is that over 50 percent of genetic mutations accountable for these diseases are invisible to short-read technologies. We do know that several genetic variants can be pathogenic – including repeat expansions, large deletions, complex rearrangements, transposable elements, and more. Now, with growing awareness that short-read NGS tools cannot accurately detect most pathogenic structural variants, scientists are turning to long-read sequencing – and answers have begun to emerge.

Identifying pathogenic structural variants

In one of the first examples of SMRT sequencing on a rare disease, scientists from Stanford University reported the discovery of a disease-causing mutation in an individual who had suffered a series of benign tumors over the course of two decades (5). Although the patient met the clinical criteria for Carney complex, a rare genetic disorder, experts had spent eight years performing various types of genetic analyses without success. Because they could not find the underlying mutation, they were unable to provide a confirmed diagnosis.

Ultimately, Stanford scientists decided to try SMRT sequencing, which led to the answer – a disease-causing deletion, stretching more than two kilobases, that affects the gene associated with Carney complex. The team sequenced the same region in the individual's parents, finding that neither carried the same mutation, allowing them to classify the mutation as *de novo* in the affected patient.

> ***"As the number of reported successes ramps up, scientists engaged in rare disease research are adopting SMRT sequencing technologies more readily."***

In another case, researchers at Yokohama City University and other institutes in Japan deployed SMRT sequencing to investigate the genetic mechanism responsible for the progressive myoclonic epilepsy affecting two siblings in a family (6). But it was not their first attempt to find an answer; previous efforts, including exome sequencing with short-read sequencing tools, had proven unsuccessful. SMRT sequencing allowed the scientists to focus on potentially causative structural variants. A quick filter of the over 17,000 structural variants found across the genome led to a homozygous 12.4 kilobase deletion in a gene known to be associated with a disease that causes similar clinical symptoms to those found in the siblings. Notably, the deletion fell in a region with high GC content, which poses processing challenges to short-read sequencers. Follow-up testing confirmed the deletion and proved that it was pathogenic.

There have been many other rare disease diagnostic victories based on SMRT sequencing (see Figure 3). Among them, advances in repeat expansion disorders stand out. These large runs of repetitive sequence are associated with a wide range of conditions, and the number of repeats is often closely tied to the severity of disease. Long-read sequence data can fully span these large regions and deliver direct, countable results – an approach that has been used for ataxias, fragile X syndrome, myotonic dystrophy, and other disorders (7, 8, 9, 10, 11).

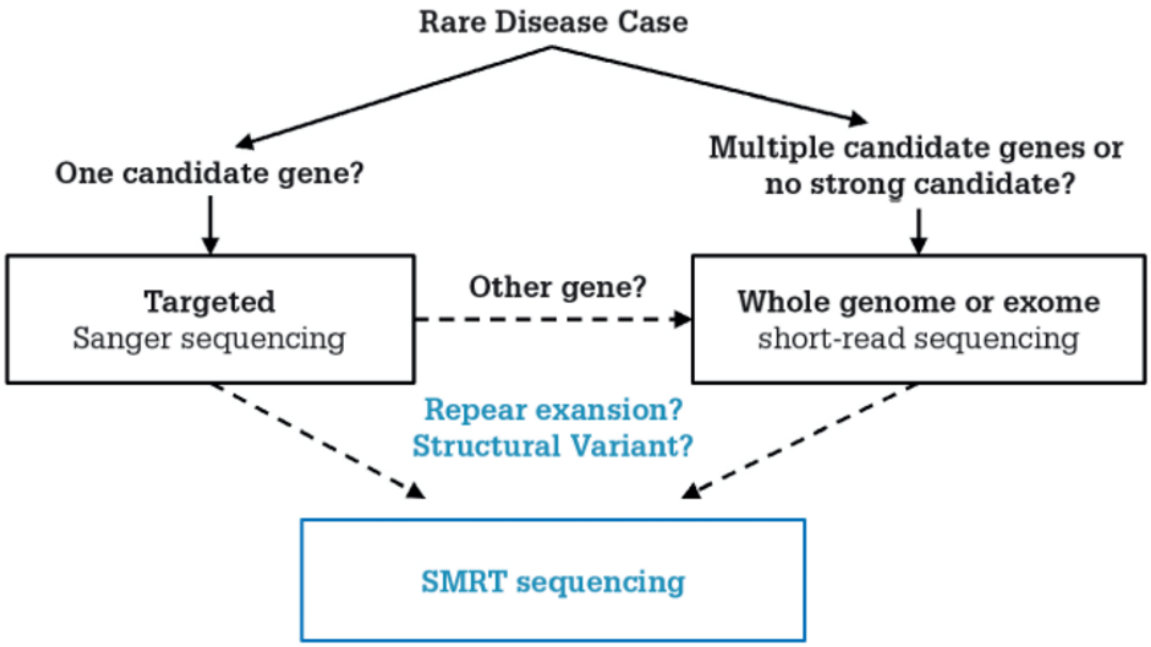

Figure 3. A workfiow for identifying pathogenic mutations in rare disease cases. Adapted from (5)

Large-scale efforts to solve rare disease

As the number of reported successes ramps up, scientists engaged in rare disease research are adopting SMRT sequencing technologies more readily.

In Europe, the SOLVE-RD consortium consists of nearly two dozen institutions in 10 countries. Funded by a €15 million award from the European Union, SOLVE-RD works to improve the diagnosis and treatment of rare diseases that have evaded explanation. The program will sequence 500 whole genomes with long-read sequencing tools to find disease-causing variants and increase solve rates.

In the USA, the National Institutes of Health-funded Clinical Sequencing Exploratory Research program uses SMRT sequencing as part of a large effort to increase the diagnostic success rate for pediatric cases that have proven challenging with other approaches. Scientists at the HudsonAlpha Institute for Biotechnology are generating whole genome sequences for hundreds of children with intellectual and developmental disabilities for which the genetic cause has not yet been found.

Large-scale programs like these should contribute a significant amount of new knowledge about the genetic mechanisms underlying rare disease, filling in many of the gaps in our understanding today. As SMRT sequencing helps to explain more rare diseases and increase overall diagnostic yield, it should have a profound effect on our ability to diagnose, understand, and ultimately improve treatment for rare disease cases.

SOLVE MORE GENETIC DISEASES WITH LONG-READ SEQUENCING

RARE DISEASES
affect 1 in 10 individuals

80%
are genetic in origin

>50%
of cases remain unsolved after short-read exome or WGS

MENDELIAN DISEASES
include over 8,500 known disorders

40%
have unknown genetic cause

DISEASES REMAIN UNSOLVED

STRUCTURAL VARIANTS ARE KNOWN TO CAUSE DISEASE
e.g. Schizophrenia, Carney Complex, Hereditary Breast & Ovarian Cancer

ACCESS THE FULL SPECTRUM OF GENETIC VARIATION

INCREASE SOLVE RATE

INCREASE DISEASE GENE DISCOVERY