PACBIO®

# Whole Genome Sequencing for Understanding Rare Diseases

Rare diseases are defined as diseases that affect a small number of people – fewer than 1 in 2,000 in the *European Union* and fewer than 200,000 total people (about 1 in 1,500) in the *United States*. Though individual rare diseases affect very few people, collectively they are common and affect over 300 million people worldwide.

### Rare Diseases — Individually Rare but Collectively Common

| Rare Disease | 1 case per |
|---|---|
| Hajdu-Cheney Syndrome | 1,000,000 |
| Tay-Sachs Disease | 300,000 |
| Gaucher Disease | 100,000 |
| Huntington Disease | 10,000 |
| Cyctic Fibrosis | 10,000 |

**HiFi sequencing is helping scientists better understand the genetic origins of rare diseases**

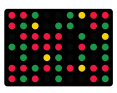
## Advances in Sequencing Technology for Improved Understanding of Rare Diseases

With more than **70% of rare diseases being genetic in origin**, scientists around the world have deployed genomic technologies to identify their causal mechanisms. Improvement in the technologies for identifying genetic variation have increased scientists' ability to understand rare diseases. Learn more about the **evolution of DNA sequencing tools**.

Karyotyping was the first technology to provide a view of the genome, revealing diseases due to chromosomal abnormalities such as Turner Syndrome (1 chromosome X instead of 2 in a female). Later, microarray provided a higher-resolution view, identifying large copy number variants, as in DiGeorge Syndrome (caused by a deletion of around 2.5 Mb on Chromosome 22). Exome or whole genome sequencing based on short-read sequencing platforms enabled even more progress by detecting single nucleotide variants (SNVs), insertions and deletions, and some larger variants.

But even whole genome sequencing with short reads finds a genetic cause in less than half of all instances of rare disease–leaving the causes of many rare diseases unknown. This, in part, is because even whole genome sequencing with short reads does not provide a comprehensive view of variation.

Fortunately, more recent advancements have led to the introduction of long-read sequencing, which has enabled sequencing of the **whole human genome** – every single base – so that all types of variants can be detected from SNV up to large structural variants (SVs). Ultimately, by detecting more variants, long-read sequencing provides a more complete picture of the genome and any abnormalities that may exist.



| Karyotype | Microarrays | Short-Read Sequencing | | Long-Read Sequencing |
|---|---|---|---|---|
| | | Exome | Partial Genome | Whole Genome |
| Chromosomal abnormalities | Copy-number variants >50 kb | SNVs, indels, CNVs, some large exonic variants | SNVs, indels, CNVs, some large variants | **SNVs, indels, SVs, CNVs, phasing, translocations, inversions, repeat expansions** |

**More Variants Detected = More Explanations**

*Technologies that provide better variant detection deliver more explanations in rare disease research.*

# What's the Difference Between Short-Read Sequencing and Long-Read Sequencing?

Like their names suggest, short-read sequencing looks at DNA in short snippets (100-350 base pairs) while long-read sequencing measures long fragments of DNA (tens of thousands of base pairs). Why does that matter? Well, when trying to characterize a human genome that has two copies (one maternal and one paternal), each 3.2 billion base pairs in length, having longer snippets of DNA means you:

- Need fewer snippets to make up the length of the **whole genome** with no unknown sequence gaps
- Can more **easily map** how one region of the genome is connected to another region
- Have the **ability to phase** or determine which copy of a gene, maternal or paternal, a mutation occurs in

As it turns out, the genetic variants underlying many of these diseases are exactly the types that short-read sequencers are least able to detect. From repeat expansions to large deletions or insertions, pathogenic variants are often large and complex structural elements that cannot be spanned by short reads of just a few hundred bases. Representing these variants accurately – and capturing all types of variants – requires much longer sequence reads that cover the entire variant in a single stretch.


*Short-read sequencing produces reads of 50-350 base pairs, which can lead to sequence gaps and incomplete coverage of disease-causing gene regions. Long-read sequencing, produces reads tens of kilobases in length, providing high-quality mapping across a genome for comprehensive variant detection.*

# HiFi Sequencing – the Key to Seeing All Variant Types Involved in Rare Disease

Unlike the data produced by short-read sequencing platforms, highly accurate long-read sequencing, known as **HiFi sequencing**, generates extremely long reads (>25 kb) that span even the largest structural variants. HiFi sequencing provides the most comprehensive view of variation in a genome, identifying the variation found with short reads and detecting the larger and more complex variants that short reads miss.

The long reads and **high accuracy** (>99.9%) of HiFi sequencing provide very complete genome assemblies, **comprehensive variant detection** with base-pair resolution, and phasing to represent maternal and paternal haplotypes.

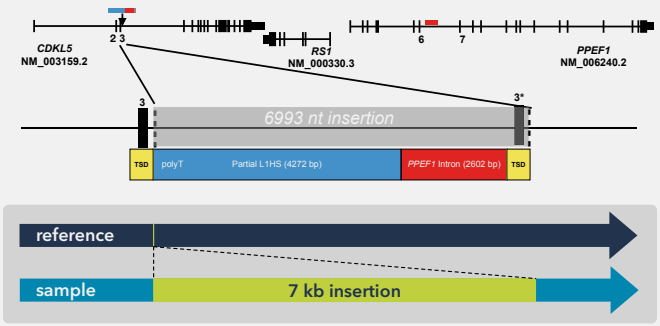

*PacBio highly accurate long reads provide comprehensive detection of all variant types, from single nucleotide variants (SNVs) to insertions and deletions and structural variants.*

# Unlocking the Secrets of Rare Diseases with HiFi Sequencing

HiFi sequencing has already made a substantial difference in rare disease research by identifying variants that were missed by short-read sequencing and other technologies. Check out these research studies of undiagnosed rare diseases and the types of pathogenic variants underlying them.

## Structural Variant Calling in Rare Disease Studies

One of the earliest examples of how PacBio sequencing technology could play a role in rare disease research came from the Stanford lab of cardiologist Euan Ashley and a young man who had suffered a series of tumors in his heart and glands. Eight years of genetic analyses had produced no firm answers. Ashley's team **used a novel method** of PacBio whole genome sequencing to find a novel structural variant in a gene associated with Carney syndrome, which was later validated as the correct mutation and finding.




More recently, a group at HudsonAlpha **found new evidence** in the study of a young girl with intellectual disabilities, seizures, and speech delay. With HiFi sequencing, the scientists at HudsonAlpha identified a de novo heterozygous insertion of nearly 7,000 bases in an intron of the *CDKL5* gene that they deemed likely pathogenic. Since *CDKL5* has been associated with early infantile epileptic encephalopathy 2, a condition characterized by many symptoms experienced by the proband, "we prioritized this event as the most interesting candidate variant," the authors reported.

*Analysis of HiFi reads in a proband indicated a de novo structural variant within the CDKL5 gene. Results supported the authors' theory that the variant has a loss-of-function effect for the individual (**Hiatt et al., (2021) HGG Advances**).*

*Structural variants are generally classified as being ≥50 bp in length and include insertions, deletions, duplications, copy-number variants, inversions, and translocations.*

**pacb.com/variant**

In Japan, researchers deployed HiFi sequencing to find the cause of an undiagnosed syndrome in twin 12-year-old girls. Clinical symptoms matched Dravet syndrome, but no molecular evidence was available to confirm that finding. They sequenced one of the twins and both parents, identifying a **novel 12 kb inversion** in a region that had previously been associated with the same symptoms affecting the girls.


*HiFi sequencing of a trio identifies a pathogenic heterozygous 12 kb de novo inversion that disrupts the gene BRPF1. SNVs (marked with "*") show that the inversion occurred on the maternal allele #3 (**Mizuguchi et al. (2021) Genomics**).*

In one last structural variant example, Kristen Sund from Cincinnati Children's Hospital identified a **13 Mb complex rearrangement** that appears to be responsible for a movement disorder in a 17-year-old with chorea, myoclonus, anxiety, and hypothyroidism. The variant was found in the *NKX2-1 gene*.

## Small Variants in Challenging Regions of the Genome

For an individual with lissencephaly (lack of folds in brain), developmental delay, and seizures, scientists at Children's Mercy Kansas City used HiFi sequencing to reveal a pathogenic variant in a region that proved difficult for short reads to represent accurately. HiFi sequencing provided even coverage – unlike the coverage dropout seen with short-read data for the same region – which spotted the key variant.


*A pathogenic variant in the CEP85L gene was detected by HiFi reads even though it was missed by short reads due to coverage dropout in that region (**Farrow, E. et al. (2021) ACMG Presentation**).*

## Capturing the Full Length and Sequence of Repeat Expansions

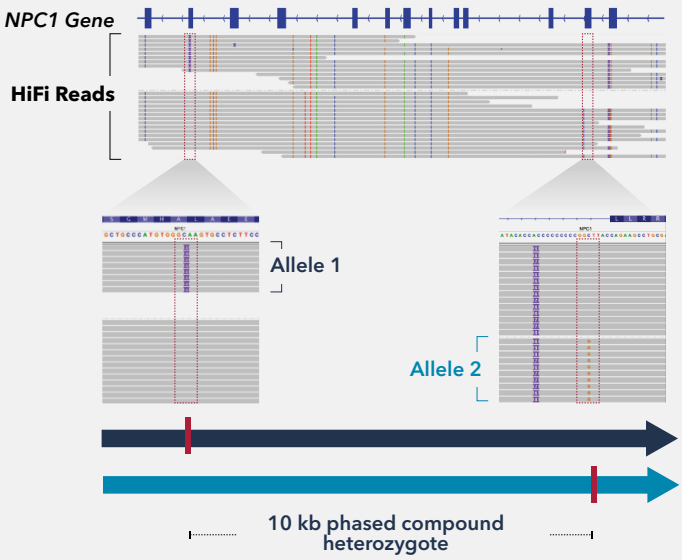
Repeat expansions have previously been shown to cause a range of diseases and can be tough to characterize accurately with short-read sequencing tools. HiFi sequencing can get through even very long expansions. Recently, scientists from Adelaide Medical School and the Robinson Research Institute linked the expansion of an ATTTC repeat in the first intron of *STARD7* with **familial adult myoclonic epilepsy**.

*Repeat expansions are mutations that result in repeating sequence that may extend for hundreds to thousands of bases. For example, the trinucleotide repeat expansion that causes Huntington's disease, consists of hundreds of CAG repeats.*


*ATTTC repeat identified in the first intron of STARD7 with familial adult myoclonic epilepsy (**Corbett et al. (2019) Nature Communications**).*

## Phasing Rare Disease Variants Across Alleles

Back at Children's Mercy Kansas City, researchers analyzed the genome of a four-year-old girl with hepatosplenomegaly whose parental genomes were not available. The individual was believed to have Niemann Pick disease Type C, but more data was needed to support the theory. HiFi reads showed two key variants located on different alleles of the relevant gene; with the **phased variants**, scientists were able to confirm the original finding.




*HiFi sequencing revealed a compound heterozygous variant in the NPC1 gene explained hepatosplenomegaly in a young female (**Farrow, E. et al. (2021) ACMG Presentation**).*

*Phasing involves separating maternally and paternally inherited copies of each chromosome into haplotypes to get a complete picture of genetic variation.*

## The Future of Rare Disease Research is Bright

Scientists around the world are striving to improve the lives of those affected by rare diseases, translating the latest research approaches and high-quality genomic data into insights that could enable the development of improved diagnostics for rare diseases. As HiFi sequencing continues to shed light on more areas of the genome, it should have a profound effect on our ability to diagnose, understand and ultimately improve treatment for the rare disease community.

Learn more about using HiFi sequencing to understand rare diseases at **pacb.com/rare-disease**

**Get in touch with a PacBio scientist to discuss your project or instrument needs**

**Connect with a PacBio Scientist**



PN: IG103-072121