PacBio

Whitepaper

Improving solve rates in rare disease research with HiFi long-read sequencing

Rare diseases affect an estimated 300 million people worldwide (*Lancet Global Health*, 2024), yet up to half of these cases remain unsolved with traditional diagnostic approaches (Graessner et al., 2021). Rare diseases present specific challenges for determining their genetic causes, as they are often caused by low frequency variants or complex genetic variation and exhibit genetic heterogeneity.

PacBio® long-read sequencing offers significant advantages for rare disease research by allowing for highly accurate, comprehensive coverage of the genome. HiFi sequencing can capture large and challenging genomic regions and resolve complex structural variants (SVs), repetitive sequences, and segmental duplications, as well as provide phasing information that short reads often miss. Additionally, simultaneous methylation detection provides a layer of epigenetic information without any additional library prep. This enhanced resolution into the genome is critical for identifying rare or novel pathogenic variants in rare disease, where small changes can have significant impact on disease etiology or progression. Long reads can sequence through difficult parts of the genome, such as GC-rich regions or repetitive elements, enabling researchers to uncover novel mutations and SVs that may be missed by short reads. This can lead to a more accurate understanding of disease mechanisms and rare disease solves.

102-326-624 REV01 OCT2024

PacBio

In addition to single nucleotide variants (SNVs) and SVs, HiFi sequencing excels in phasing variants. In rare disease research, this can help distinguish parental origin of mutations, enabling differentiation of inherited and *de novo* mutations that may guide clinical decisions.

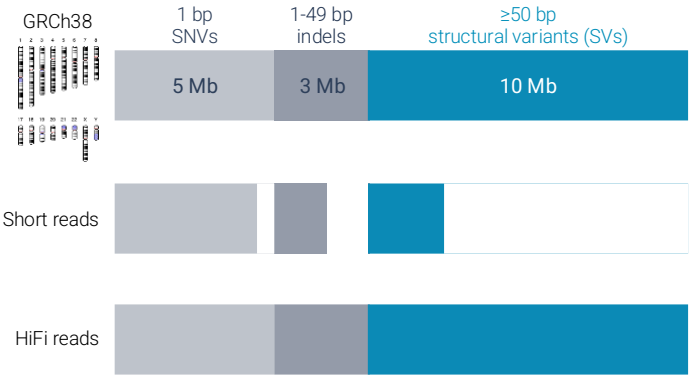

Figure 1. Long reads capture more complex variant types in more regions of the genome. Structural variants comprise more of the genome than SNVs and indels but are difficult to capture using short-read sequencing. PacBio HiFi sequencing provides comprehensive detection of more complex variant types.

Technology	Genetic variation captured	Explanation rate	References
Karyotype	Chromosomal abnormalities	~5%	De Vries et al. 2005
Chromosomal microarray analysis	Copy number variants (CNVs) >50 kb	~10%	Clark et al. 2018
Short-read whole exome	SNVs and indels, some large exonic variants	~30–40%	De Ligt et al. 2012; Chung et al. 2023
Short-read whole genome	SNVs, indels, some large variants	~40%	Gilissen et al. 2014; Chung et al. 2023
HiFi long-read whole genome	SNVs, SVs, CNVs, phasing, translocations, inversions, repeat expansions, methylation	>50%	Farrow et al. 2024*

Table 1. Rare disease solve rates of testing-naïve cases are improved with long-read sequencing. Legacy genetic testing technologies have limited ability to detect variations in underlying rare disease. *Platform presentation at the 2024 ACMG Annual Clinical Genetics Meeting.

HiFi sequencing enables rare disease research studies

Long-read whole genome sequencing has emerged as a transformative technology for increasing rare disease solve rates due to its unprecedented resolution and high accuracy, granting insights into genomic regions that were previously inaccessible by legacy technologies.

Improvements in automation and scalability have increasingly enabled researchers to adopt PacBio HiFi sequencing for rare disease cohorts, demonstrating the potential to significantly improve solve rates for rare disease, particularly for cases that remain unresolved after traditional short-read sequencing.

This whitepaper highlights key studies that apply long-read sequencing in rare disease research, focusing on solve rates and methodologies used in each study for improving the detection and understanding of rare disease.

Long-read sequencing in the pan-European Solve-RD program

The pan-European rare disease research program Solve-RD aims to solve undiagnosed rare and inherited disease (RID) cases in the European Rare Disease Network (ERN). In this study, authors split families into two phenotype buckets: phenotypes clinically recognized as “unsolvable” due to complex biology or presentation heterogeneity, and “unsolved” phenotypes where disease etiology was suspected but not yet found. In total, 293 individuals, including probands and healthy relatives, from 114 families were sequenced with PacBio technology.

Using HiFi sequencing, a number of pathogenic variants were identified in “unsolved” families where the likely causal gene fit the phenotype. Additionally, HiFi sequencing revealed novel, likely pathogenic aberrations in 4 additional families.

The authors noted the challenge with "unsolvable" cases, which might be addressed with further methylation analysis, full-length isoform sequencing, or deeper sequencing coverage. Data from this study will be made publicly available.

Key findings:

- Total solve rate (all short-read negative): 11.4%
  - Previously "unsolved" cases: 13.0%
  - Previously "unsolvable" cases (93 families): 4.8%
- Coverage: mean 10x
- Phenotypes: "Unsolved" – neurological/, neuromuscular disease, and/or epilepsy; "Unsolvable" (21 families) –: clinically well-recognized "unsolvable" syndromes

"Now that these [long-read sequencing] technologies produce high-quality sequencing reads at steadily dropping costs, researchers can evaluate the hypothesis that part of the genetically undiagnosed [rare disease] are caused by variants that remain hidden from previously used technologies."

Analysis of long-read sequencing in pediatric rare disease cohort

This study of the Genomic Answers for Kids (GA4K) program generated an 11% total solve rate of 584 affected individuals who lacked a diagnosis despite previous short-read exome, genome, or panel testing with Illumina or MGI sequencing. Recently, a subsequent analysis of this dataset examined over 1,500 HiFi genomes from families with rare disease for variant calling validation and repeat expansion testing using TRGT. Altogether, the authors find that the implementation of HiFi sequencing in the exome/short-read negative cohort results in a >10% increase in solve rate and improved variant calling from short reads at 99.74% sensitivity and 99.99% specificity with the PacBio WGS Variant Pipeline.

The authors note the promise of HiFi WGS as a potential comprehensive approach to boost solve rates and reduce costs and time:

"The implementation of HiFi-GS as a first-line genetic test paves the path for a transformative era of genomic testing."

Key findings:

- Total solve rate of short-read negative cohort: 11%
  - 13% of explanations in previously unsolved cases were due to SVs revealed by HiFi sequencing
  - Characterization of >4x more rare coding SVs compared to short-read whole genomes
- Solve rate of cohort with no prior testing: 34.5%
- Coverage: ~10x, >25x

Improving solve rates in rare neurodevelopmental disorders

In a recent analysis of a neurogenerative disease cohort, researchers sequenced 96 singleton probands using HiFi WGS that were negative after short-read whole exome or whole genome trio analysis. The authors state "While re-analysis of older data clearly increases diagnostic yield, we find that [long-read genome sequencing] allows for substantial additional yield beyond [short-read genome sequencing]". Seven of the 16 total explanations with HiFi sequencing included variants that could only have been found with long reads, which include complex variation such as copy number variants, inversions, a mobile element insertion, low complexity repeat-expansions, and a deletion.

"Long-read genome sequencing represents the next phase of…acceleration by facilitating a substantial increase in variant comprehensiveness and accuracy."

Key findings:

- Total solve rate (all short-read negative): 16.7%
  - Of these, 44% included variants that were only detectable by long-read sequencing
- Coverage: mean 26.1x, singletons only
- Phenotypes: neurodegenerative, suspected to be genetic

Advantages of long reads over exome sequencing for autosomal recessive disorders

In this study, authors explored the utility of HiFi sequencing in cases that remained undiagnosed after short-read whole exome sequencing and reanalysis of the proband. The study included 34 families with suspected autosomal recessive diseases, with index cases sequenced to an average depth of 10x.

Authors were able to identify candidate variants in 13 of the 34 families (38%), including discovery of novel gene-disease relationships. The authors note that for those explained with long-read sequencing, "nearly half of the identified candidate variants were SVs and SNVs that were missed by exome." They also discuss limitations of low-depth sequencing, citing at least one case of a novel large insertion that they predict would have been identified with higher-depth sequencing.

Key findings:

- Total solve rate (negative cases after proband exome + reanalysis): 38%
- Coverage: mean 10x
- Phenotypes: various

Increasing solve rates for rare pediatric sensorineural hearing loss

Researchers report on a cohort of 19 pediatric cases with rare sensorineural hearing loss (SNHL) of unknown etiology and variable phenotype that remained unexplained after extensive genetic testing. For this overall phenotype, researchers and clinicians tend to employ panel-based or exome sequencing. However, authors note that noncoding or complex variation is often missed, and certain associated genes like *OTOA* and *STRC* are difficult to resolve due to high homology and segmental duplications.

After HiFi whole genome sequencing, authors identified causative or suspected causative variants in 4 of 19 cases (21.1%), which were orthogonally validated in a CLIA laboratory. Variants detected by long-read sequencing include: a hemizygous deletion in *trans* with a pathogenic missense variant in *OTOA* and two loss-of-function single nucleotide variants in *trans* with a known copy number loss for the associated *STRC* gene, which has a highly homologous pseudogene, as well as a complex copy neutral inversion in *MITF*.

"[Long-read genome sequencing] provided significantly improved resolution for complex structural variation and, in this cohort, substantially improved diagnostic yield over [exome] and [short-read genome sequencing]."

Key findings:

- Total solve rate (negative cases after prior short read panels, WES, WGS): 21.1%
- Coverage: mean 24–32x
- Phenotypes: congenital sensorineural hearing loss

Additional studies demonstrate HiFi sequencing utility for rare disease

Assessing HiFi genomes for first-tier testing of clinically-relevant variants

In this study, researchers assessed the utility of long-read sequencing to capture complex, clinically relevant germline variants that are difficult to detect using short reads, and typically require multiple test modalities. This positive control study encompassed 100 samples with 145 known pathogenic, clinically relevant variants. Known variants included short tandem repeats, pseudogenes or regions of high homology, complex SVs, mtDNA variation, and imprinting loci.

Sequencing at mean 30x coverage on a Revio system detected 83% of the 145 variants with a fully automated pipeline, while 10% were detected with manual inspection or data visualization. Compared to a subset of 70 variants that were also benchmarked with short-read genomes, HiFi long reads could detect more (90% vs 41% with short reads), suggesting that "**HiFi genomes may be a more attractive first tier, generic assay" for germline rare disease testing**.

A multigenerational truth set pedigree reveals higher de novo mutation rates

With this study, researchers expanded on previous benchmarking initiatives (like Genome in a Bottle and Platinum Genomes, Eberle et al., 2017) with a 28-member four-generation CEPH pedigree. While 5 total technologies were utilized, HiFi sequencing was particularly successful for analysis of *de novo* mutations (DNM), and expanded the high-confidence "truth set" region of the human genome by ~244 Mb.

Based on the inclusion of long-read sequencing data, authors estimate 128–259 DNMs per generation, significantly more than the 60-70 DNMs per generation previously thought based on short-read sequencing benchmarking. Given that *de novo* variants are often causal in rare disease cases, short-read sequencing alone may miss >50% of DNMs in probands. By including HiFi sequencing in an analysis of a large pedigree, authors were able to create "**the most comprehensive, publicly available 'truth set' of all classes of genomic variants**". This Platinum Pedigree resource is further described in this study.

Leveraging 5mC calling for detection of rare disease hypermethylation events

Leveraging the Genomic Answers 4 Kids (GA4K) program, authors of this study dove deeper into HiFi data to specifically investigate CpG methylation (mCpG) levels, often impacted by noncoding SNVs. Using simultaneous 5mC methylation detection inclusive to HiFi sequencing, authors examined genome-wide variant calling and mCpG levels of 276 GA4K participants from 152 families.

The authors found over 25,000 rare hypermethylation events, with 81% of these defined as allele-specific through phasing of HiFi sequencing data. An average of 117 hypermethylation events per sample were identified with HiFi sequencing compared to 8 with srWGBS. Such events can cause a loss of regulatory element activity, and when compared to rare disease genes, may provide context for variant prioritization and improve understanding of disease etiology.

Characterizing inversions in rare disease

Through an analysis of rare disease genomes from the 100,000 Genomes Project, this study focused on the contribution of inversions in genetic rare disease through haploinsufficiency. A pilot study with 9 samples sequenced on PacBio systems was conducted to demonstrate the utility of long-read sequencing for this application (Sequel® IIe with 2–4 samples per SMRT® Cell 8M, estimated 2.5–5x depth). Long-read sequencing data was able to confirm prior short-read SV interpretation or clarify ambiguity regarding breakpoints and SV configuration. For two cases with complex SVs in the final exon of *MECP2*, PacBio data "resolved the SVs, and these findings directly influenced clinical interpretation."

Resolving rare retinal dystrophies

In this study, researchers describe 3 unrelated cases of inherited retinal dystrophies (IRD) that remained unsolved after short-read sequencing. IRD is genetically heterogenous, with mutations in approximately 300 different genes known to be associated with disease. HiFi sequencing on the Revio system revealed likely pathogenic variants in all 3 cases, including a deep intronic variant, a 13 kb exon deletion, and a large complex SV (multiple exon deletion plus splice-altering intronic variant), highlighting the utility of long-read sequencing for addressing this rare phenotype.

Rare disease toolbox

Rare disease researchers in these studies leveraged tools unique to long-read whole genome sequencing to identify pathogenic variants that explain rare disease etiology or phenotypes. PacBio HiFi sequencing unlocks insight into many complex regions of the genome, which may explain disease biology and improve solve rates. To get the most value out of HiFi sequencing, many analysis tools and methods have been developed to extract information from long reads, which can be further filtered and interpreted in the context of rare disease cases.

Variation Type	Tool/Method	References
SNVs/indels	DeepVariant, Sentieon	Poplin et al. 2018; Freed et al. 2018
SVs	sawfish, pbsv	Saunders, et al. 2024
CNVs	HiFiCNV	PacBio GitHub
Mulitple variant types	Consolidated WGS variant calling pipeline	PacBio GitHub
Methylation	pb-CpG-tools	Cheung, et al. 2023, PacBio GitHub
Phasing	HiPhase	Holt, et al. 2024
Tandem repeats	TRGT, TRGTdenovo	Dolzhenko, et al. 2024, Mokveld, et al. 2024
Paralogous regions, Segmental duplications	Paraphase	Chen, et al. 2023, Chen et al. 2024
RNA variants	Kinnex	pacb.com/Kinnex
Chromatin	Fiber-seq	Jha et al. 2024
Multiomic variation	Multiple tools available	Vollger, et al. 2023
Long-read variation catalog (SVs, SNVs)	CoLoRS	CoLoRSDB.org

Table 2. Features of HiFi sequencing and relevant analysis tools. For additional documentation, visit github.com/PacificBiosciences.

Conclusion

These recent studies demonstrate how HiFi sequencing can be used to transform rare disease research. With more accurate, comprehensive coverage of the genome, HiFi sequencing has been shown to improve solve rates over short-read sequencing and other legacy technology across rare disease phenotypes. This marked improvement holds in studies even with coverage depths of less than 30x, further emphasizing the robust ability to identify pathogenic variants in rare disease cases. Broader adoption of HiFi sequencing with greater coverage promises to drive progress in rare disease research.

Resources

PacBio rare disease webpage
Brochure – Scale human disease research with HiFi sequencing

102-326-624 REV01 OCT2024