PacBio

Application note

HiFi targeted sequencing and comprehensive analysis of *SMN1/2* with Paraphase

Introduction

Spinal muscular atrophy (SMA) refers to a group of inherited neuromuscular disorders characterized by loss of nerve cells in the spinal cord called lower motor neurons or anterior horn cells and resulting muscle atrophy and weakness. SMA is a leading genetic cause of death in children (Wirth, 2021) and is caused by biallelic mutations of the *SMN1* gene.

SMN1 and its paralog *SMN2* are found in a highly complex genomic region on chromosomal band 5q13 where gene conversion and unequal crossing-over frequently occurs, resulting in variable copy numbers of *SMN1* and *SMN2*. Both genes are nearly identical in sequence with just one functionally different base.

High sequence similarity between *SMN1* and *SMN2* makes analysis difficult and both genes have variable copy numbers across populations. It is currently not possible to identify silent carriers (2+0) with two copies of *SMN1* on one chromosome and zero gene copies on the other without pedigree information (Chen et al., 2023).

102-326-566 REV01 MAR2023

PacBio

In this Application Note, we demonstrate targeted HiFi sequencing assays combined with a new informatics method for the combined analysis of *SMN1* and *SMN2*. The software tool, Paraphase, is an informatics method that identifies full-length *SMN1* and *SMN2* haplotypes, determines gene copy numbers, and calls phased variants using PacBio® HiFi data.

Your advantages

PacBio HiFi reads are long (up to 25 kb) and accurate (99.9%). Targeted HiFi sequencing has the ability to span large portions of the *SMN1*/*SMN2* genes, allowing for haplotype construction, detection of structural variants or copy number variants, in addition to phased SNVs and indels. These attributes make HiFi sequencing well suited for cost-effective analysis of both rare and common variants of *SMN1*/*SMN2*. Paraphase performs per-haplotype variant calling in each sample. In combination with a scalable targeted sequencing assay, this method allows for population-wide analysis to identify genetic markers enabling haplotype-based screening of silent carriers.

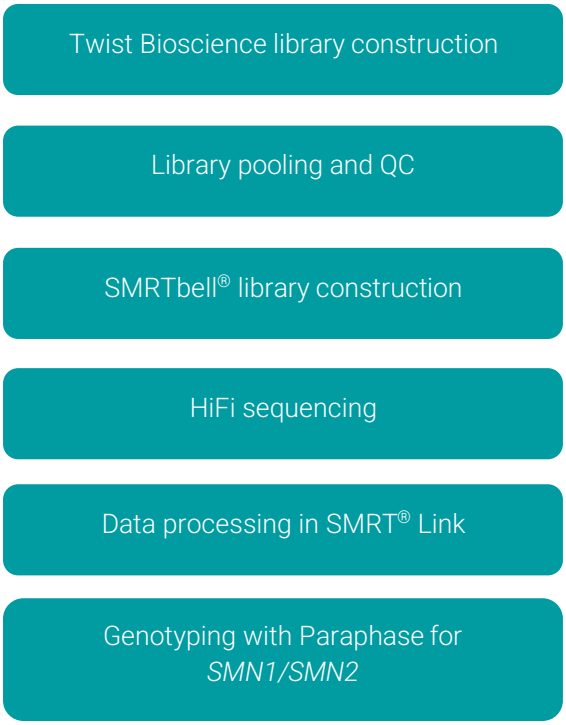

Figure 1. General overview of targeted enrichment workflow and analysis.

HiFi target enrichment

HiFi sequencing of targeted gene panels is a cost-effective way to sequence regions of interest at scale. *SMN1* and *SMN2* are included in the Twist Alliance Dark Genes panel which has been optimized for HiFi sequencing. The probe design targets the full gene body of *SMN1* and *SMN2* with sparsely tiled 120 bp probes that include intronic and exonic sequence.

The Dark Genes panel includes 389 genes with total target length of 22 Mb. Up to 12 samples can be barcoded and pooled per Revio™ SMRT® Cell and up to 4 samples per Sequel® IIe SMRT Cell 8M.

Library construction

To prepare a targeted sequencing long-read panel, samples are first prepared according to Long Read Library Preparation and Standard Hyb v2 Enrichment with reagents provided by Twist Bioscience (see Figure 1). Samples are barcoded during Twist library preparation with up to 400 sample barcodes available. Sample multiplexing recommendations depend on panel size and sequencing platform.

Twist libraries are then pooled and a SMRTbell library is constructed using SMRTbell prep kit 3.0. Libraries are prepared for sequencing using Binding Kit 3.2 on the Sequel IIe system or the Revio polymerase kit for sequencing on the Revio system.

PCR amplicons

Targeted HiFi sequencing of *SMN1* and *SMN2* has also been demonstrated with amplicon approaches. In the study published by Li et al. (2022), the authors describe a long-range multiplex PCR strategy with two pairs of primers to cover the full-length and downstream regions of *SMN1* and *SMN2* genes.

Data processing and QC

In SMRT Link, HiFi reads can be processed for analysis by demultiplexing, marking PCR duplicates, and mapping reads to a reference genome. For best genotyping results for *SMN1* and *SMN2*, Paraphase can be used to determine complete haplotypes, make phased variant calls, and call copy number variation.

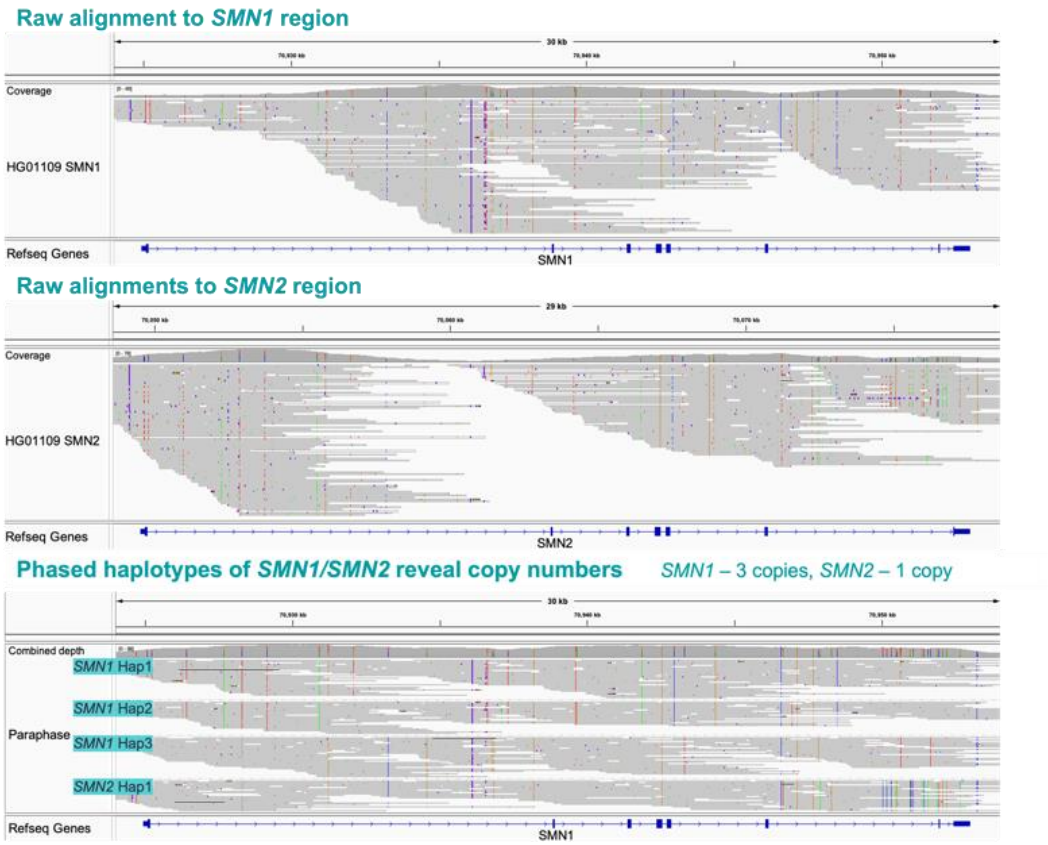

Figure 2. Paraphase resolves *SMN1* and *SMN2* using enrichment data in sample HG01109. Raw HiFi read alignments to *SMN1* and *SMN2* are ambiguous due to high sequence similarity, resulting in low mapping quality, mismapped reads, uneven depth, inaccurate variant calls and failed phasing. Paraphase realigns all reads to *SMN1* and resolves four haplotypes, among which three are *SMN1* copies and one is an *SMN2* copy.

Genotyping with Paraphase

Paraphase takes the aligned BAM file from targeted sequencing (we recommend target enrichment data) or whole-genome sequencing data as input and performs haplotype phasing, copy number calling and phased variant calling. It outputs a summary JSON file with copy number calls and VCF files for phased haplotypes. It also includes a small BAM file of the region where all reads from either *SMN1* or *SMN2* are realigned to *SMN1* and grouped by the haplotypes that they originate from, enabling users to QC the results (Figure 2).

Advanced analysis with Paraphase

Paraphase makes phased variant calls for each copy of *SMN1* or *SMN2* in a sample. This allows high-throughput analysis of SMA samples to identify variants of interest and their disease-modifying effects. This also allows population-wide analysis to identify genetic markers for haplotype-based screening of complex traits such as silent carriers.

In a recent study published by Chen et al. (2023), the authors utilized Paraphase and HiFi sequencing data to conduct a first-of-its-kind population study across 438 samples from five ethnic populations, identifying the major *SMN1* and *SMN2* sequence haplogroups. Using pedigrees, the authors studied alleles with variable copies (0, 1 or 2) of *SMN1* or *SMN2* and characterized the co-segregation pattern of *SMN1* and *SMN2* haplogroups. Furthermore, the team identified two *SMN1* haplotypes forming a common two-copy *SMN1* allele in African populations. Testing positive for these two haplotypes in an individual with two copies of *SMN1* gives a silent carrier risk of 88.5%, which is significantly higher than the currently used SNP marker g.27134T>G (1.7%–3.0%). This study and its results demonstrate the potential of haplotype-based screening of silent carriers for SMA.

Resources and references

Resources

Paraphase is open source at:
https://github.com/PacificBiosciences/paraphase

For more details please visit:
https://www.pacb.com/products-and-services/analytical-software/targeted-sequencing/

Available materials

1. Application brief – HiFi Sequencing with Twist Bioscience Target Enrichment
   https://www.pacb.com/wp-content/uploads/Application-Brief-HiFi-Target-Enrichment-Best-Practices.pdf
2. Procedure & checklist – Preparing multiplexed amplicon libraries using SMRTbell prep kit 3.0
   https://www.pacb.com/wp-content/uploads/Procedure-checklist-Preparing-multiplexed-amplicon-libraries-using-SMRTbell-prep-kit-3.0.pdf

References

Chen, X., et al. (2023). Comprehensive SMN1 and SMN2 profiling for spinal muscular atrophy analysis using long-read PacBio HiFi sequencing. *The American Journal of Human Genetics*, 110(2), 240-250. https://doi.org/10.1016/j.ajhg.2023.01.001

Li, S., et al. (2022). Comprehensive analysis of spinal muscular atrophy: SMN1 copy number, intragenic mutation, and 2+ 0 carrier analysis by third-generation sequencing. *The Journal of Molecular Diagnostics*, 24(9), 1009-1020. https://doi.org/10.1016/j.jmoldx.2022.05.001.

Wirth, B. (2021). Spinal muscular atrophy: in the challenge lies a solution. *Trends in Neurosciences*, *44*(4), 306-322. https://doi.org/10.1016/j.tins.2020.11.009

102-326-566 REV01 MAR2023