PacBio

Application note

HiFi amplicon sequencing and isoform analysis for Gaucher disease research

Introduction

Gaucher disease (GD) is an autosomal recessive disorder with varied features that affects many parts of the body, and increases the risk of developing Parkinson disease (Do et al., 2019). GD is one of several disorders of sphingolipid degradation, known as sphingolipidoses. Each sphingolipidosis is associated with defects of a specific lysosomal enzyme or other protein involved in sphingolipid degradation. This occurs with subsequent accumulation of substrate (lipids) in one or more organs, resulting in a lysosomal storage disorder.

More than 495 variants related to the development of Gaucher disease have been identified in the *GBA* gene. Most of the *GBA* gene variants responsible for Gaucher disease, including single base changes, splicing alterations, partial and total deletions, insertions, and gene/pseudogene rearrangements, alter the structure of beta-glucocerebrosidase, preventing it from working normally. *GBA* contains 11 exons and 10 introns over a length of 7.6 kb. Located 16 kb downstream is the shorter (5.7 kb), non-processed pseudogene, *GBAP1*, with 96% homology to *GBA* (Zampieri et al., 2017). Currently, variant analysis on *GBA* is typically performed through targeted (short-read) exon sequencing. In general, this method presents technical limitations in detecting variants such as large rearrangements, structural variants, copy number variations (CNVs), epigenetic modifications, variants in repetitive or GC-rich regions, and variants in highly homologous genomic regions. Moreover, visual inspection and Sanger validation are required to confidently detect/confirm recombinant variants. In this Application note, we focus on an amplicon-sequencing based approach for the analysis of the entire *GBA* gene and its pseudogene with PacBio® HiFi sequencing. We also highlight the benefits of RNA isoform sequencing (Iso-Seq® method) to generate insights into the transcriptional impact of *GBAP1*.

Your advantages

PacBio HiFi reads are long (up to 25 kb) and accurate (99.9%). Targeted HiFi sequencing of amplicons has the ability to span entire genes, allowing for straightforward haplotype construction, detection of structural variants or copy number variants in addition to single-nucleotide variants (SNVs) and indels. These attributes make HiFi amplicon sequencing well suited to cost-effectively analyze both rare and common variants of all types.

102-326-551 REV01 JAN2023

PacBio

Workflow overview

Here, we present an example primer design for the amplification of full-length *GBA* and *GBAP1* from human genomic DNA (gDNA) samples (Figure 1). The workflow uses a single round of PCR using barcoded primers. The gene and pseudogene can be coamplified in the same PCR step, producing amplicons of ~12.5 kb and ~ 16 kb, respectively (Figure 2). Samples are then pooled, a SMRTbell® library is prepared, then sequencing is performed. HiFi reads are then demultiplexed by barcode pairs and analyzed to generate consensus sequence or variant calls for the gene and pseudogene.

Multiplexing approach

This document provides primer designs for barcoding up to 24 samples using 8 forward and 3 reverse barcoded primers for each gene. Table 1 lists a total of 22 primer designs. This approach can be extended to achieve a higher multiplexing level by designing and validating additional barcoded primers. See Available Materials below for more information.

PCR, library construction, and sequencing

Table 1 lists sequences for barcoded forward and reverse primers for *GBA* and *GBAP1*. For each sample, the two genes can be co-amplified in a single PCR by selecting a pair of barcoded forward and reverse primers for each locus (four primers total). This workflow has been demonstrated with 50 ng of DNA (90% of fragments over 25 kb) using the Platinum SuperFi II PCR Master Mix (Thermo Fisher) with the addition of 5M Betaine (Sigma-Aldrich) as an enhancer for the long range PCR (final concentration of 0.80M).

The resulting PCR product can be pooled and used to generate a SMRTbell library according to this protocol.

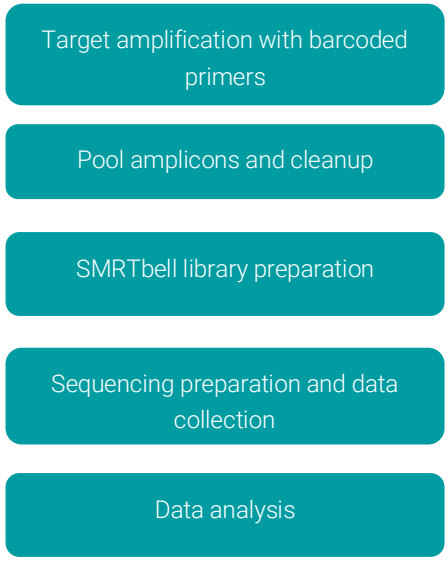

Figure 2. General overview of targeted sequencing workflow on PacBio systems with barcoded primers.

The resulting library can be prepared for sequencing on PacBio systems following instructions in the Sample Setup module of SMRT® Link.

Note that the PCR assay design, including primer sequences and PCR conditions, is intended as an example workflow, and is neither optimized nor supported by PacBio.

Recommended analysis tools

HiFi reads may be produced and demultiplexed on the instrument or in SMRT® Analysis off-instrument. HiFi reads for each sample can then be analyzed using pbaa, which generates phased consensus sequences by clustering HiFi reads and enables haplotype phasing for accurate diplotype calls using downstream community tools.

Figure 3 (page 4) illustrates how the paired primer approach enables the phasing of variants and identification of gene/pseudogene fusions.

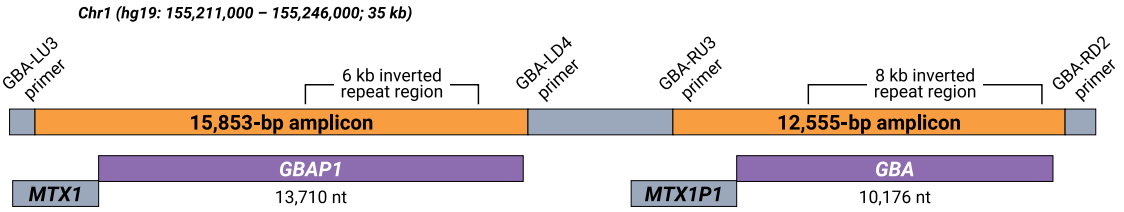

Figure 1. Primer design for co-amplification of GBA and GBAP1.

Forward barcoded primers for GBA (0.08 µM)	
GBA-RU3_BC1017_Forward	/5Phos/GCATC**CACACGCGCGCTATAT**CTGTCAGCAGGCATTGTTC
GBA-RU3_BC1018_Forward	/5Phos/GCATC**TCACGTGCTCACTGTG**CTGTCAGCAGGCATTGTTC
GBA-RU3_BC1019_Forward	/5Phos/GCATC**ACACACTCTATCAGAT**CTGTCAGCAGGCATTGTTC
GBA-RU3_BC1020_Forward	/5Phos/GCATC**CACGACACGACGATGT**CTGTCAGCAGGCATTGTTC
GBA-RU3_BC1021_Forward	/5Phos/GCATC**CTATACATAGTGATGT**CTGTCAGCAGGCATTGTTC
GBA-RU3_BC1022_Forward	/5Phos/GCATC**CACTCACGTGTGATAT**CTGTCAGCAGGCATTGTTC
GBA-RU3_BC1023_Forward	/5Phos/GCATC**CAGAGAGATATCTCTG**CTGTCAGCAGGCATTGTTC
GBA-RU3_BC1024_Forward	/5Phos/GCATC**CATGTAGAGCAGAGAG**CTGTCAGCAGGCATTGTTC
Reverse barcoded primers for GBA (0.05 µM)	
GBA-RD2_BC1057_Reverse	/5Phos/GCATC**CTCTCAGACGCTCGTC**GGTTAGGAATCCTCTGAGCTTC
GBA-RD2_BC1058_Reverse	/5Phos/GCATC**TACTCTCAGAGACACA**GGTTAGGAATCCTCTGAGCTTC
GBA-RD2_BC1059_Reverse	/5Phos/GCATC**TATCTCAGTGCGTGTG**GGTTAGGAATCCTCTGAGCTTC
Forward barcoded primers for GBAP1 (0.10 µM)	
GBA-LU3_BC1001_Forward	/5Phos/GCATC**CACATATCAGAGTGCG**ACTGTTAGAAAGTTGCCGTCA
GBA-LU3_BC1002_Forward	/5Phos/GCATC**ACACACAGACTGTGAG**ACTGTTAGAAAGTTGCCGTCA
GBA-LU3_BC1003_Forward	/5Phos/GCATC**ACACATCTCGTGAGAG**ACTGTTAGAAAGTTGCCGTCA
GBA-LU3_BC1004_Forward	/5Phos/GCATC**CACGCACACACGCGCG**ACTGTTAGAAAGTTGCCGTCA
GBA-LU3_BC1005_Forward	/5Phos/GCATC**CACTCGACTCTCGCGT**ACTGTTAGAAAGTTGCCGTCA
GBA-LU3_BC1006_Forward	/5Phos/GCATC**CATATATATCAGCTGT**ACTGTTAGAAAGTTGCCGTCA
GBA-LU3_BC1007_Forward	/5Phos/GCATC**TCTGTATCTCTATGTG**ACTGTTAGAAAGTTGCCGTCA
GBA-LU3_BC1008_Forward	/5Phos/GCATC**ACAGTCGAGCGCTGCG**ACTGTTAGAAAGTTGCCGTCA
Reverse barcoded primers for GBAP1 (0.20 µM)	
GBA-LD4_BC1033_Reverse	/5Phos/GCATC**AGAGACTGCGACGAGA**GGAGAGTTGTCCGTCTTCAA
GBA-LD4_BC1034_Reverse	/5Phos/GCATC**ATATCTATATACACAT**GGAGAGTTGTCCGTCTTCAA
GBA-LD4_BC1035_Reverse	/5Phos/GCATC**CAGAGAGTGCGCGCGC**GGAGAGTTGTCCGTCTTCAA

Table 1. Paired primers capture full-length *GBA* and its nearby pseudogene *GBAP1*, including flanking regions. Outer primer pairs amplify fusion alleles generated by deletions in the target region. The concentrations in parentheses represent primer concentrations used in this example workflow and sequences in bold represent the barcodes.

Pseudogene analysis through isoform sequencing

The impact of the *GBAP1* pseudogene can be investigated via full-length isoform sequencing with HiFi sequencing, either in bulk brain tissue or from iPSC-derived purified cell samples. In a recent example, a new preprint entitled "Pseudogenes limit the identification of novel common transcripts generated by their parent genes" studies the impact of pseudogenes on transcriptome analyses, using the parent-pseudogene pair *GBA-GBAP1* (Gustavsson et al., 2022). The authors observe that only 42% of short reads map uniquely to *GBA*, with the remaining reads mapping primarily to *GBAP1*. This resulted in a significant misestimation of the relative expression of *GBA* to *GBAP1*. To obtain clearer insights into *GBA-GBAP1* transcription, the PacBio Iso-Seq method was used. By using the targeted Iso-Seq method on 12 human brain regions, the researchers identified 18 *GBA* transcripts that had a novel open reading frame (ORF) and 7 *GBAP1* transcripts predicted to encode a protein, despite *GBAP1* being classified as a pseudogene. The researchers identified significant differences in *GBAP1* ORF usage across different brain cell types.

Resources

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

Additional multiplexing resources:
https://www.pacb.com/multiplexing/

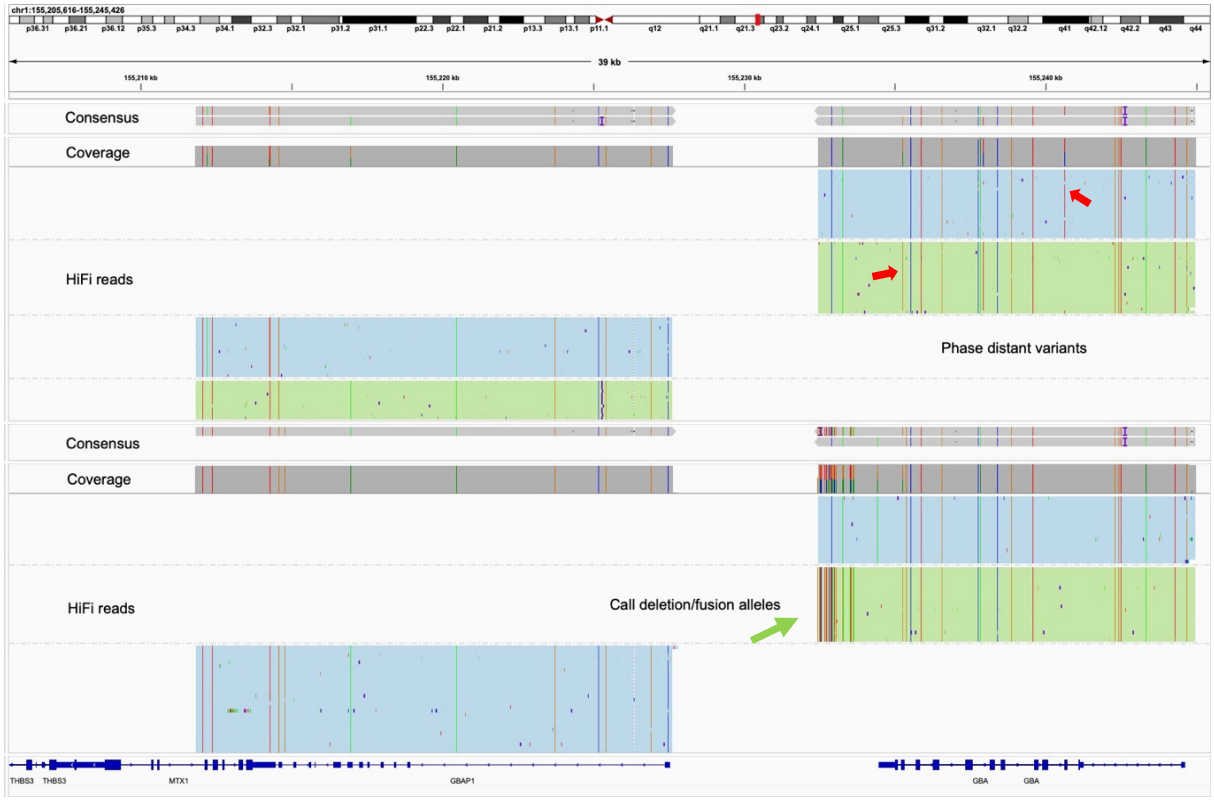

Figure 3. Top panel: Coriell sample NA20270 – Long amplicons sequenced with HiFi reads allow phasing of variants to identify compound heterozygote variants across the entire coding region. Two pathogenic variants (red arrows) are separated by 5 kb of sequence. Bottom panel: Coriell sample NA08753 – Gene/pseudogene fusions are easily identifiable with a paired primer design (most distant primers amplify a fusion allele). The cluster of variants (green arrow) is indicative of a fusion of *GBAP1* and *GBA*.

Available materials

1. Targeted sequencing for amplicons – Application Brief
2. Preparing multiplexed amplicon libraries using SMRTbell prep kit 3.0

Customer feedback

"It has been incredible to find how little we really understood about GBA gene transcription in the human brain. PacBio sequencing was key to the work."
Mina Ryten, MD, PhD, UCL Great Ormond Street Institute of Child Health, UK

References

Do, J., McKinney, C., Sharma, P., & Sidransky, E. (2019). Glucocerebrosidase and its relevance to Parkinson disease. Molecular Neurodegeneration, 14(1), 1-16. https://doi.org/10.1186/s13024-019-0336-2

Zampieri, S., Cattarossi, S., Bembi, B., & Dardis, A. (2017). GBA analysis in next-generation era: pitfalls, challenges, and possible solutions. *The Journal of Molecular Diagnostics*, *19*(5), 733-741. https://doi.org/10.1016/j.jmoldx.2017.05.005

Gustavsson, E. K., et al. (2022). Pseudogenes limit the identification of novel common transcripts generated by their parent genes. bioRxiv. https://doi.org/10.1101/2022.10.21.513169

102-326-551 REV01 JAN2023