Application Note

Long-read Genome Sequencing at Scale

Advanced Method for Cost-Effective, High-Throughput Long-Read Sequencing in Population Genomics

Implementing long-read DNA sequencing at scale has traditionally been cost-prohibitive, time-consuming, and required substantial expertise for data analysis. Our innovative method offers a cost-effective, streamlined solution for large-scale long-read sequencing, compatible with both Oxford Nanopore and Pacific Biosciences technologies. This end-to-end solution includes advanced bioinformatics and analytics, ensuring efficient and accurate sequencing results, making it accessible for a wide range of genomic research applications.

INTRODUCTION

Next-generation sequencing (NGS) technology has revolutionized genomic research, allowing for rapid and cost-effective sequencing of DNA. NGS technologies have advanced significantly, enabling researchers to sequence entire genomes, transcriptomes, and other genetic materials at an unprecedented scale. Traditional NGS methods, known as short-read sequencing, produce vast amounts of data by reading short fragments of DNA. This approach has been widely used due to its accuracy, efficiency, and lower costs, making it ideal for applications like whole-genome sequencing, gene expression profiling, and genetic variant detection.

However, short-read sequencing has limitations, particularly in resolving complex genomic regions, repetitive sequences, and large structural variants. Long-read sequencing, offered by technologies such as Oxford Nanopore and Pacific Biosciences, overcomes these challenges by producing longer reads that span these difficult regions, providing a more comprehensive view of the genome. The benefits of implementing long-read NGS at scale include improved genome assembly, better detection of structural variants, and enhanced accuracy in identifying genetic variants associated with diseases. Despite these advantages, scaling up long-read sequencing is challenging due to the high costs, time-consuming nature of the process, and the significant expertise required for data analysis. This highlights the need for a streamlined, cost-effective, and comprehensive solution that can facilitate the widespread adoption of long-read NGS in population genomics research.

In this application note, we demonstrate how our method overcomes the challenges associated with scaling long-read sequencing for population genomics. We provide a detailed overview of our streamlined, cost-effective approach that integrates advanced bioinformatics and analytics, making it feasible for large-scale genomic studies. By leveraging both Oxford Nanopore and Pacific Biosciences technologies, our method ensures accurate and comprehensive sequencing results while significantly reducing the expertise required for data analysis. We showcase how this new method can be implemented efficiently, offering a robust solution for researchers aiming to unlock the full potential of long-read sequencing in population genomics.

METHOD OVERVIEW

Our approach enables the efficient multiplexing of 10 samples in a single sequencing run, applicable to both PacBio and Oxford Nanopore technologies. In this method, samples are combined into sets of 10, and each set is sequenced using either a PacBio SMRTcell or an Oxford Nanopore PromethION Flowcell. Following sequencing, the data undergoes a comprehensive pipeline on the Sequegenics platform, which includes error correction, haplotype calling, and detailed data analysis. This streamlined process ensures cost-effective, high-throughput sequencing with accurate results, making it ideal for large-scale genomic studies.

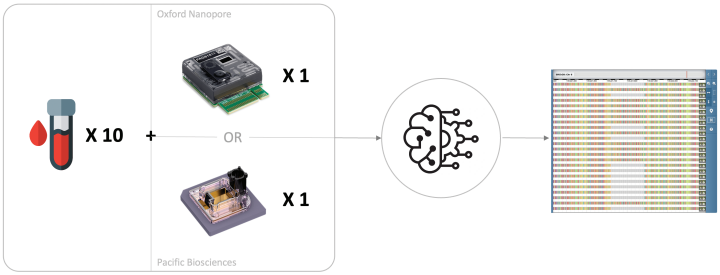

Figure 1. Squematic diagram of the method. Blood samples are barcoded and sequenced in a 10-plex design, either in a PacBio SMRTcell or a ONT flowcel to produce combinted raw sequencing data that is processed in Sequegenics platform

VALIDATION

Study population. In our experimental design, we included blood samples from 90 individuals diagnosed with early-onset coronary artery disease and 20 control samples.

Sample preparation. Genomic DNA was extracted using the Circulomics kit (Pacific Biosciences). The extracted gDNA was subjected to shearing using Megaruptor3 (Diagenode) and size selection using BluePippin (Sage Sciences). A library was prepared using the SMRTbell kit (Pacific Biosciences) in a 10-plex design. Each library was loaded in a single SMRT cell (Pacific Biosciences) and sequenced in the Sequel IIe instrument (Pacific Biosciences) with a total of 10 SMRT cells utilized for the entire project.

Bioinformatics and data analysis. Following sequencing, error correction, haplotype calling, and data analysis were performed using the Sequegenics platform. Briefly, raw sequencing data is mapped against the Human Genome reference GRCh38, and haplotype-aware error correction and variant calling is performed on a subset of genes medically relevant to cardiovascular disease. Additionally, genes that contained ethnicity-related SNPs were included in the analysis as well as internal controls.

Table 1. List of medically relevant genes included in the validation study.

ACTC1, ACVR2B, ALMS1, ANGPTL4, APOA5, APOB, APOC3, BRAF, CASZ1, CBL, CHD7, CRELD1, ELN, FOXH1, GATA4, GATA5, GATA6, GDF1, GJA1, HAND1, HAND2, HRAS, JAG1, KMT2D, KRAS, LEFTY2, LPA, MAP2K1, MAP2K2, MEIS2, MESP1, MYH6, NFATC1, NKX2-5, NKX2-6, NODAL, NOS3, NOTCH1, NR2F2, NRAS, NSD1, PCSK9, PLD1, PTPN11, RAF1, RIT1, SHOC2, SMAD6, SOS1, TAB2, TBX1, TBX20, TBX5, TFAP2B.

RESULTS

Sample preparation and sizing. Using the circulomics HMW DNA extraction kit, an minimum of 12ug were extracted from each sample. Each sample was sheared in the Megaruptor3 at speed 29 and fragments smaller than 15kb were removed using a cassette BPLUS10 in the BluePipping system. An example size profile obtained with the femtoPulse instrument is shown in Figure 2.

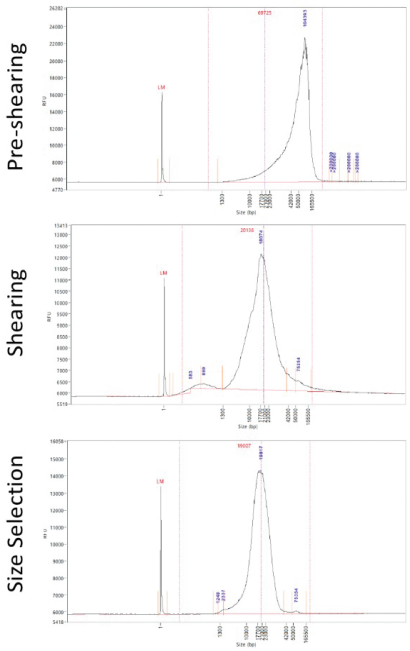

Figure 2. DNA sizing and QC. Femto-pulse profile of gDNA utilized for library preparation and sequencing.

Long-read PacBio sequencing. Sets of 10 library preps (10-plex design) were sequenced in a single PacBio Sequel II2 SMRTcell, each. Raw subread data was produced and uploaded to the Sequegenics Platform as separate biosamples. Next, variant calling was initiated for the set of genes listed in Table 1. AWS Athena was used to query the list of variants of interest which fell into three categories: Potentially pathogenic variants, splice variants, structural variants (INDELs > 50bps) and small INDELs in coding regions.

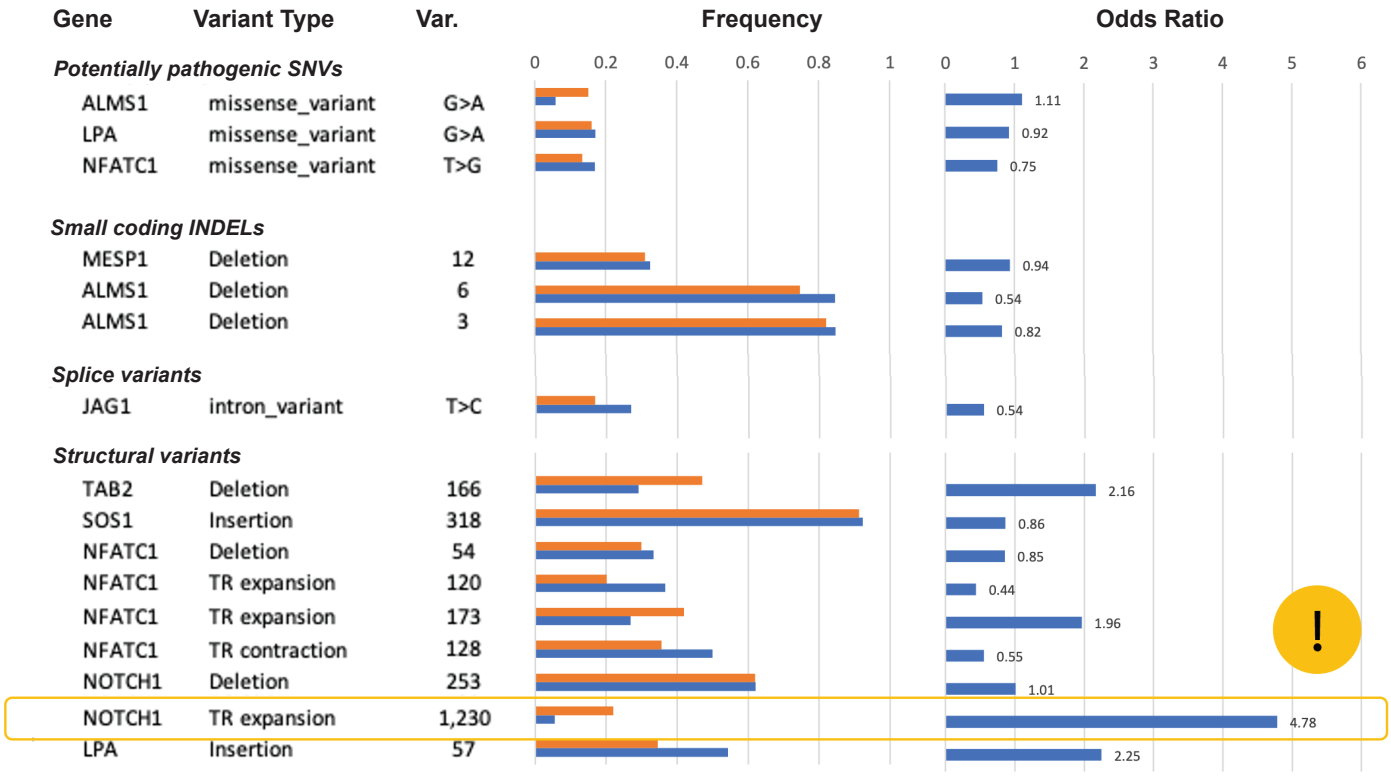

Figure 3. Comparative analysis between cases and controls. Frequencies of variants of interest are shown for cases (orange bars) and controls (blue bars). ORs are provided as well. Highlighted is a variant of particular interest which is a large expansion repeat more frequently present among cases in the gene NOTCH1.

On Figure 4, a detail of three of the nine structural variants identified is provided. It is possible to appreciate the hidden diversity of this previously innaccessible types of genetic variants that is underlying study population.

CONCLUSIONS

In this application note, we demonstrated that our innovative method can produce comprehensive insights that only long-read technologies can offer at a fraction of the cost and through a fully automated end-to-end solution. This method can be particularly relevant to idetnfify hidden correlates to disease as it allows for more detailed and accurate investigations into genetic variations, ultimately advancing our understanding of complex genetic traits and improving personalized medicine.

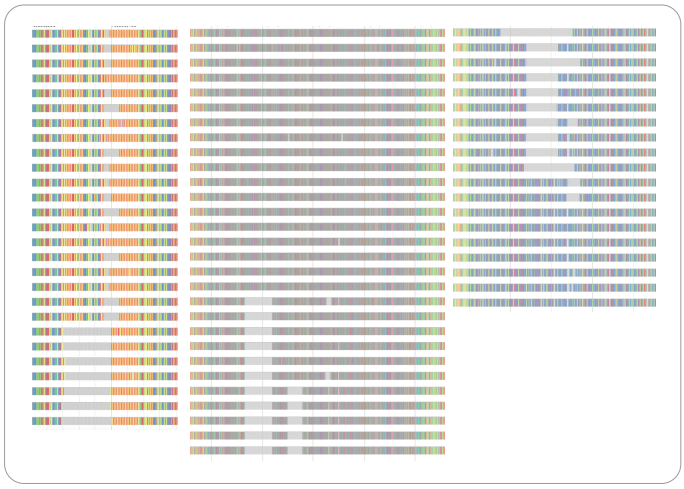

Figure 4. Detail of three structural variants found in the population. Our approach was able to identify not only locations in the genome containing structural variants but it was also able to derive the diversity of this type of variants that can be found when long-read NGS is implemented at scale.

Published by:

Sequegenics, Inc
1860 Montreal Rd
Tucker, GA 30084.

sequegenics.com