MBI6013

Research proposal

Dockerized computational pipeline for SMRT sequence data analysis

Faculty of Medicine,Chinese University of Hong Kong, Shenzhen

Name: <<\* Xu Hang \*>>

Student ID: <<\* 223050032 \*>>

Advisor : Professer Sun Hao

Background and Significance

SMRT (Single Molecule Real-Time) sequencing is a third-generation sequencing technology developed by Pacific Biosciences (PacBio). The fundamental concept is that DNA polymerase serves as an exceptional sequencing tool. Guided by the DNA template strand, it can synthesize a new strand with remarkable accuracy, achieving an error rate of 10-5, and the process is unbiased, eliminating issues like GC bias. Furthermore, the enzyme reaction is not only swift, reaching speeds of 750bp/s, but it is also efficient, requiring minimal ATPs for energy and occupying a small volume. To observe the enzyme reaction, we can utilize dNTPs with distinct fluorescent signals to capture snapshots of the synthesis process. This approach allows us to record the polymerization reaction of the enzyme in detail. Moreover, the fluorescent signals are labeled on the polyphosphate moiety of dNTPs, ensuring that they have no impact on the enzyme reaction. Following the synthesis process, the fluorophores are cleaved from the strand, eliminating any fluorophore presence within the strand. This approach also prevents steric hindrance and background signal accumulation, ensuring accurate and reliable observation of the enzyme reaction.[1]

SMRT is a high-throughput system, posing challenges in accurately recording the reaction process of individual polymerases. To overcome this, the ZMW Hole (Zero-Mode Waveguide Hole) is utilized to facilitate the observation of a single DNA polymerase at a given time. [2]The physics behind this is the zero-mode waveguide effect, a phenomenon where, if the aperture is smaller than the wavelength of the laser light, the laser is unable to pass through the aperture. Instead, it creates a tiny high-energy zone at the aperture's bottom, an area where the laser light can reach. By positioning the polymerase within this confined space, we create a microscopic environment conducive to observing the polymerase's polymerization reaction. The laser can excite the fluorescence of dNTPs.

Compared to the existing second-generation sequencing (NGS) technologies, SMRT sequencing offers several advantages. NGS sequencing technologies typically have short library insert sizes of around 300 base pairs (bp), resulting in challenges related to capturing long-range genomic information. Both library preparation and sequencing in NGS require PCR amplification, resulting in uneven genome coverage and difficulties in amplifying regions with high GC content. Consequently, certain regions of the genome are challenging to sequence effectively using NGS. Furthermore, NGS involves cluster generation, and errors that occur during this process are amplified. The sequencing errors in NGS are non-random, and increasing sequencing depth does not effectively improve accuracy. As a result, NGS struggles to correct errors in regions that are prone to frequent errors.

In contrast, SMRT sequencing addresses these limitations by allowing the construction of libraries with longer fragments ranging from 1 to 100 kilobases (kb). This is achieved by attaching hairpin adapters to both ends of the DNA double strand, forming a single-stranded circular structure. The sequencing process occurs in a rolling-circle way, with DNA polymerase initiating from the hairpin adapters within the zero-mode waveguide (ZMW) holes. Moreover, SMRT sequencing eliminates the need for PCR amplification during library preparation. This results in improved library uniformity and eliminates GC bias, enhancing the accuracy and representation of the entire genome. SMRT sequencing utilizes phi-29 DNA polymerase instead of PCR polymerase, enabling a rolling-circle replication mode. The random nature of errors introduced by the polymerase during its activity can be reduced by increasing sequencing depth, leading to higher accuracy. In fact, SMRT sequencing can achieve accuracy levels up to Q50 or Q60 for shorter read lengths, surpassing the accuracy attainable by NGS technologies.

Moreover, SMRT sequencing has the unique capability to detect DNA modifications during the sequencing process. [3]It utilizes DNA polymerase kinetics to measure the fluorescence signals and inter-pulse duration (IPD) between two consecutive bases. By analyzing the time differences, it can identify potential base modifications. Modified bases exhibit slower kinetics due to steric hindrance, resulting in longer time intervals between incorporation events. This temporal information allows for the identification and characterization of various types of base modifications.[4]

HiFi Reads, a remarkable feature of SMRT sequencing are used to sequence long fragments. These long reads, typically spanning 10 to 25 kilobases (kb), provide single-molecular resolution, allowing for the direct sequencing of DNA or RNA (cDNA) without the need for amplification. With individual molecular accuracy exceeding Q20 (99%), HiFi Reads deliver highly reliable and precise sequence information. What sets them apart is their unbiased nature, as HiFi Reads are obtained without DNA amplification, minimizing biases associated with GC content and sequence complexity. This unbiased approach ensures a comprehensive and representative view of the sequenced genome or transcriptome. possesses strong strand displacement and continuous synthesis capabilities, allowing it to unwind and replicate complex DNA structures. It enables isothermal DNA amplification in vitro without the need for thermal cycling, resulting in a high yield of high-molecular-weight DNA. With 3' to 5' exonuclease proofreading activity, it exhibits superior fidelity compared to most other DNA polymerases. Rolling Circle Amplification (RCA) is a nucleic acid detection method that emulates the natural circular DNA rolling circle replication process. Using circular DNA as a template, RCA utilizes a short DNA primer that extends and displaces the previously synthesized strand, facilitated by Phi29 DNA polymerase. This process leads to the generation of repeated long single-stranded DNA products. Additionally, HiFi Reads harness the power of Circular Consensus Sequencing (CCS), utilizing Phi29 DNA polymerasea (rolling-circle DNA synthesis method) to generate multiple subreads during repetitive sequencing. By performing error correction within the same molecule, these subreads collectively produce exceptionally accurate consensus reads. [5]

SMRT sequencing offers the advantage of longer read lengths compared to NGS, as mentioned previously. In comparison to ONT sequencing, SMRT sequencing provides additional benefits. In the first place, SMRT sequencing exhibits higher raw accuracy than ONT sequencing. This is achieved through circular consensus sequencing (CCS), which generates multiple passes of the same DNA molecule, allowing for error correction. In contrast, ONT sequencing may have higher error rates due to the nature of nanopore technology. Secondly, SMRT sequencing excels in identifying DNA base modifications, such as DNA methylation, by directly detecting the kinetic changes caused by modified bases. This capability is particularly valuable for epigenetic studies and understanding gene expression regulation. While ONT sequencing can also detect base modifications, it may have limitations in accurately distinguishing between different types of modifications.

SMRT sequencing data analysis holds profound applications across diverse areas of genomics research.[6] In whole-genome sequencing (WGS), its long read lengths and high accuracy enable the comprehensive analysis of genetic variation, encompassing single nucleotide polymorphisms (SNPs), insertions/deletions (Indels), and structural variants. This comprehensive approach is invaluable in studying human genetic diseases, cancer genomics, and population genetics. Furthermore, SMRT sequencing's application in RNA sequencing contributes significantly to our understanding of dynamic gene expression processes. By directly measuring the full length of transcripts, it reveals the complexity of transcript isoforms, alternative splicing variants, and non-coding RNAs, crucial for gene regulation studies, the identification of novel functional transcripts, and disease mechanism investigations. In DNA methylation research, SMRT sequencing stands out for its unique ability to directly detect dynamic changes caused by modified bases, enabling high-sensitivity identification of DNA methylation sites. This is paramount in epigenetics studies, genome stability assessments, and understanding the mechanisms underlying disease occurrence and development. Moreover, SMRT sequencing's ability to differentiate between different types of DNA methylation modifications, such as 5-methylcytosine (5mC) and 6-methyladenine (6mA), offers powerful tools for delving into the functional roles of these modifications in gene regulation.

Especially，SMRT sequencing demonstrates significant advantages in analyzing the methylation of circulating free DNA (cfDNA) in plasma, particularly for non-invasive early cancer screening. [7]While traditional NGS techniques face challenges in detecting long cfDNA fragments, SMRT sequencing overcomes this obstacle with its unique capabilities. Firstly, the ultra-long read lengths of SMRT sequencing allow for the coverage of longer DNA fragments, enabling more comprehensive capture of methylation information. In contrast, traditional NGS techniques often struggle with long cfDNA fragments, potentially leading to the omission of important methylation information. SMRT sequencing, on the other hand, can accurately reveal changes in methylation patterns, which often serve as early signals for cancer occurrence and development. Secondly, SMRT sequencing offers high accuracy and sensitivity, precisely identifying methylation sites and their alterations. This precision is crucial for distinguishing normal cells from cancer cells, assessing tumor burden and metastasis. In comparison, NGS technologies may be limited in accurately identifying methylation sites when detecting long cfDNA fragments due to read length constraints. SMRT sequencing provides more reliable methylation analysis results, offering a more accurate basis for personalized treatment strategies. Additionally, SMRT sequencing has the ability to directly detect methylation without the need for additional chemical processing or conversion steps. This simplifies the experimental workflow and reduces potential errors and biases. In contrast, NGS techniques may require complex preprocessing and conversion processes when detecting methylation, increasing experimental complexity and potential sources of error.

Docker is an open-source containerization platform widely used in cluster computing environments. It provides a lightweight, portable, and scalable containerization solution that allows developers to package applications and their dependencies into independent containers, enabling fast deployment and portability.[8]

In cluster computing, Docker offers several advantages. Firstly, Docker containers provide isolated runtime environments, allowing different applications to run independently in the same cluster without interference. This isolation enhances deployment flexibility and reduces failures and instability caused by conflicts between applications. Secondly, Docker containers are lightweight and have fast startup times. As containers share resources with the operating system kernel, they generally have much faster startup times compared to traditional virtual machines. This enables efficient dynamic scheduling and scaling of containers in the cluster, better adapting to changes in computational workloads. Additionally, Docker containers possess portability and repeatability. Containers package applications and their dependencies into a self-contained unit, simplifying and ensuring reliability in application deployment and migration. Developers can create and test containers in their local development environment and seamlessly deploy them to any node in the cluster, without concerns about environment configurations and dependencies. Lastly, the Docker ecosystem is rich and active. Docker Hub provides a vast repository of public images, allowing developers to easily access and share commonly used container images. Moreover, Docker supports container orchestration tools such as Docker Compose and Kubernetes, facilitating convenient and flexible management and orchestration of large-scale containerized applications in the cluster.[9]

Specific Aims

The objective of this research is to establish a comprehensive analysis pipeline tailored for SMRT sequencing data, encompassing Whole Genome Sequencing (WGS), RNA sequencing, and DNA methylation analysis, utilizing Docker images and cluster computing. This pipeline aims to streamline the analysis process, enabling researchers to efficiently analyze SMRT sequencing data with minimal hassle related to environmental dependencies and configurations.

Specifically, the research aims to accomplish three key goals. Firstly, we intend to create a Docker container specifically designed for SMRT sequencing data analysis. This container will incorporate essential tools and dependencies necessary for genomic assembly, gene expression analysis, and DNA methylation analysis. By packaging these resources into a Docker image, we aim to simplify deployment and execution for researchers, eliminating the need for manual installation and configuration of individual tools. Secondly, we aim to develop a distributed computing workflow based on cluster computing for SMRT sequencing data analysis. This workflow will leverage task scheduling and container orchestration to automate task allocation and container deployment, maximizing computational efficiency and scalability. Additionally, we will utilize data parallel processing and distributed storage techniques to handle and analyze large-scale SMRT sequencing data effectively. Finally, we plan to validate and apply the established analysis pipeline using real SMRT sequencing data. This validation process will assess the pipeline's performance, accuracy, and reliability by conducting genomic assembly on WGS data, expression analysis on RNA sequencing data, and methylation analysis on DNA samples. Through practical application, we aim to demonstrate the utility and effectiveness of our developed pipeline in SMRT sequencing data analysis.

Furthermore, we propose to leverage our established analysis pipeline for conducting comprehensive studies on whole-genome sequencing (WGS), RNA sequencing, and DNA methylation analysis. Utilizing this pipeline, we aim to delve into the intricacies of structural variations for cancers. In RNA sequencing, our focus will be on alternative splicing and full-length RNA sequencing, which are crucial in elucidating the complexity of gene expression regulation and transcriptional diversity. These studies will enhance our understanding of gene transcription and expression mechanisms, providing novel insights into the etiology and progression of diseases. Additionally, our pipeline will be employed to investigate methylation patterns in cancer, a crucial epigenetic modification in cancer development. Through methylation analysis, we seek to reveal alterations in methylation patterns in cancerous cells and explore their association with cancer occurrence, progression, and prognosis. This will facilitate a deeper understanding of the molecular mechanisms underlying cancer and pave the way for novel diagnostic, therapeutic, and preventive strategies. By applying our analysis pipeline across these domains, we anticipate revealing the complexity of the genome, diversity of transcriptomes, and methylation alterations in cancer, thereby contributing novel perspectives and solutions to related research fields and propelling scientific advancements. Furthermore, we hope that our Docker and cluster computing-based analysis pipeline will contribute to more efficient and scalable solutions for processing and analyzing large-scale SMRT sequencing data.

Research (Experimental) Design and Methods

The research design and methods of this project are centered on comprehensive data collection, analysis, and interpretation using SMRT sequencing technology. Initially, we will acquire whole-genome sequencing (WGS) data, RNA sequencing and DNA methylation data from SMRT sequencing instruments, followed by rigorous quality control and preprocessing to ensure data accuracy and reliability. For data analysis, we have established a Docker-based framework with a distributed computing pipeline. This will enable us to call hifi reads with kinetics for base calling. HiFi reads are called from the original subreads. Subsequently, multiple subreads are aligned to the same starting position and stacked or "piled up" together to form a "pileup" for generating CCS. The base at each position in the consensus sequence is determined based on a specific algorithm. And then align hifi reads to reference genome. Moreover, the downstream analysis can be performed next like DNA methylation calling,variant detection, annotation, gene expression analysis, differential expression analysis, and functional enrichment analysis. The interpretation of these results will involve comparing and validating the discovered key gene variants, expression patterns, and methylation patterns against existing knowledge. Additionally, we will utilize appropriate visualization tools and charts to present the analysis outcomes, enabling researchers to gain a deeper understanding of the data.

One of the innovative aspects of this study is the utilization of Docker images and cluster computing, which offers flexibility and portability. This allows researchers to easily deploy and run the analysis pipeline in various computing environments. Furthermore, the distributed computing and data parallel processing approach significantly enhance analysis efficiency and scalability, making it suitable for handling large-scale SMRT sequencing data.

Bibliography or Reference List

[1]Ardui S, Ameur A, Vermeesch JR, Hestand MS. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. Nucleic Acids Res. 2018 Mar 16;46(5):2159-2168. doi: 10.1093/nar/gky066. PMID: 29401301; PMCID: PMC5861413.

[2]Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG, Webb WW. Zero-mode waveguides for single-molecule analysis at high concentrations. Science. 2003 Jan 31;299(5607):682-6. doi: 10.1126/science.1079700. PMID: 12560545.

[3]Ni, P., Nie, F., Zhong, Z. et al. DNA 5-methylcytosine detection and methylation phasing using PacBio circular consensus sequencing. Nat Commun 14, 4054 (2023). https://doi.org/10.1038/s41467-023-39784-9

1. Feng L, Lou J. DNA Methylation Analysis. Methods Mol Biol. 2019;1894:181-227. doi: 10.1007/978-1-4939-8916-4\_12. PMID: 30547463.

[5]van Dijk EL, Naquin D, Gorrichon K, Jaszczyszyn Y, Ouazahrou R, Thermes C, Hernandez C. Genomics in the long-read sequencing era. Trends Genet. 2023 Sep;39(9):649-671. doi: 10.1016/j.tig.2023.04.006. Epub 2023 May 23. PMID: 37230864.

[6]Hestand MS, Ameur A. The Versatility of SMRT Sequencing. Genes (Basel). 2019 Jan 4;10(1):24. doi: 10.3390/genes10010024. PMID: 30621217; PMCID: PMC6357146.

[7]Choy LYL, Peng W, Jiang P, Cheng SH, Yu SCY, Shang H, Olivia Tse OY, Wong J, Wong VWS, Wong GLH, Lam WKJ, Chan SL, Chiu RWK, Chan KCA, Lo YMD. Single-Molecule Sequencing Enables Long Cell-Free DNA Detection and Direct Methylation Analysis for Cancer Patients. Clin Chem. 2022 Sep 1;68(9):1151-1163. doi: 10.1093/clinchem/hvac086. PMID: 35587130.

[8]Merenda, M., & Turra, A. (2019). Docker containers for scientific computing. Journal of Physics: Conference Series, 1198(6), 062003. doi: 10.1088/1742-6596/1198/6/062003

[9]Radchenko, I., & Breiter, G. (2016). Docker-based container approach for bioinformatics software deployment. Bioinformatics, 32(17), 2712-2714. doi: 10.1093/bioinformatics/btw307