Best Practices Terra Workspace for Processing PacBio Long Reads Sequenced on the Revio system: Single Sample, Trio, and Cohort Modes

Mark Fleharty[1], Shadi Zaheri[1], William J Rowell[2], Michelle Cippicho[1], Emma Walker[1], Lisa Anderson[1], Kiran Garimella[1], Heather Rooke[1], Namrata Gupta[1], Sean Hofherr[1] and Niall Lennon[1].
Broad Clinical Labs[1], Pacific Biosciences[2]

Introduction

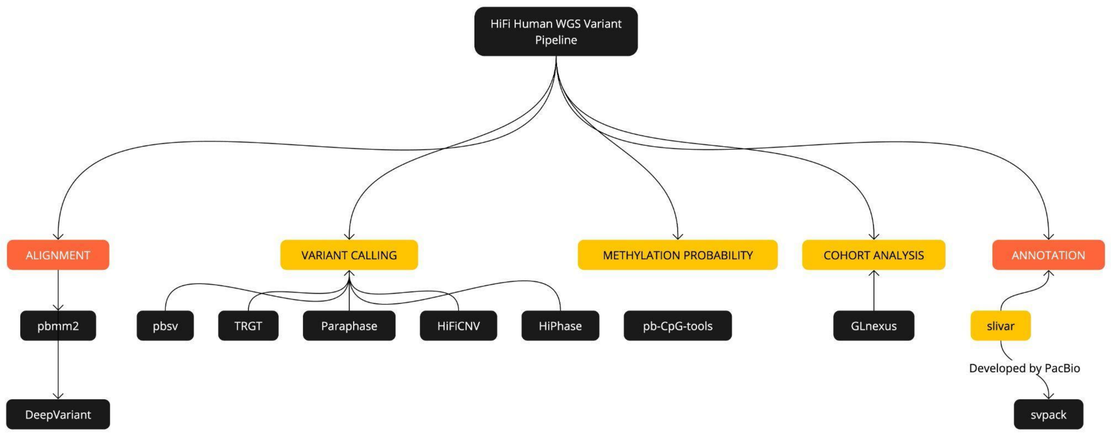
The latest generation of long-read sequencing technologies bring the promise of more comprehensive variation characterization in every sample. However, most secondary analysis infrastructure has been built to process short reads and cannot handle long read data. We present a best practices pipeline (hosted in a Terra Workspace) optimized for processing long reads human whole-genome sequencing (WGS) from the Pacific Biosciences Revio platform. This workspace makes use of a WDL-based pipeline developed by PacBio and available from GitHub. Our workspace supports single sample, trio, and cohort modes. Key steps include read preprocessing, error correction, alignment, and variant calling.

An optional tertiary analysis step includes advanced variant filtering by joint calling samples in either trio or cohort mode. Outputs are generated at both the sample level (e.g., BAM statistics, small variant VCFs, ROH outputs, phased variant VCFs, HiFiCNV outputs, CpG methylation data) and the cohort level (e.g., jointly called SV and small variant VCFs, haplotype phasing statistics). Tertiary analysis outputs include filtered and compound heterozygous variants, detailed TSV files, and filtered structural variant data from svpack. The workspace features conditional execution based on available data and parameters, along with customizable options for backend environments and computational resources, ensuring robust performance for diverse genomic studies.

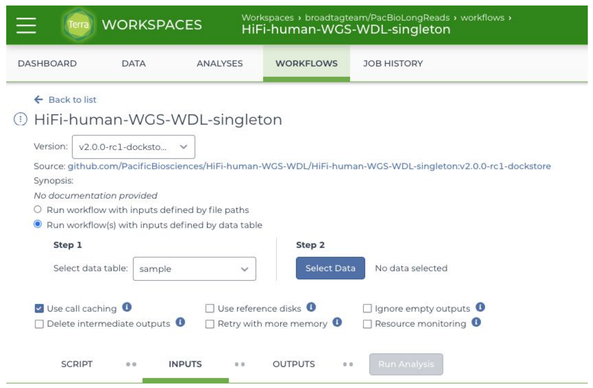
We validated the performance of this pipeline using the Ashkenazim trio sample HG002, HG003, and HG004. These samples were run in both single sample and trio modes, leveraging extensive truth data available for HG002 for SNP and indels.

This pipeline provides a robust, scalable platform to perform complete human LR-WGS for research, translational, and ultimately clinical intended uses.

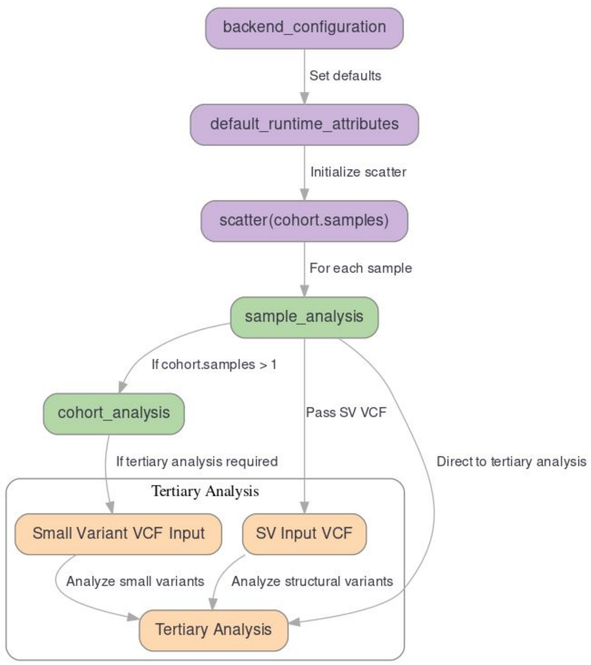
PacBio WGS Variant Calling Pipeline for Secondary and Tertiary Analysis



Workflow as Implemented in Terra



Workflow Diagram



Performance of Pipeline on HG002

The PacBio long-read sequencing pipeline, implemented on the Terra platform, demonstrated high accuracy and robustness in variant detection on the benchmark Ashkenazim trio (HG002, HG003, HG004). Hosted in a Terra Workspace, the pipeline achieved excellent sensitivity and precision for single nucleotide polymorphisms (SNPs), insertions/deletions (indels), and structural variants (SVs), particularly in complex genomic regions. Utilizing tools like pbsv and HiFiCNV, it excelled in detecting structural and copy number variants (CNVs) with high positive predictive value, while HiPhase supported accurate phasing and haplotype analysis. The Terra environment allowed for scalable and efficient processing, with trio analysis further enhancing variant interpretation by confirming inheritance patterns, crucial for clinical research. The pipeline also produced CpG methylation data for epigenetic insights, showcasing its versatility and reliability for both research and translational genomics applications.

Discussion

This Best Practices Workspace for processing PacBio long reads sequenced on the Revio platform represents an important resource for researchers working with long-read sequencing data. This workspace makes use of the PacBio WGS Variant Pipeline v1.2.0 and provides an easy and robust way to process long-read data.

The pipeline supports single samples, trios, and cohorts, making it adaptable for various research goals. It enhances the potential for large-scale genomic studies by offering joint calling capabilities in trio and cohort modes. The pipeline's integration within the Terra Workspace further simplifies access, enabling users to manage complex datasets and perform high-throughput analyses efficiently. The dual-level output (sample and cohort) provides both individual and population-level insights, which is invaluable for variant interpretation.

The optional tertiary analysis step, which includes joint variant filtering and structural variant refinement, adds precision to the data processing workflow. This step facilitates the identification of clinically relevant genetic variations, such as compound heterozygous and structural variants, which are critical for understanding disease phenotypes and inheritance patterns. The pipeline's ability to execute conditionally based on input parameters and resource availability demonstrates adaptability across different research needs and computing infrastructures, making it a scalable solution for genomics research that requires rigorous accuracy and reliability.

The validation of this workspace using the Ashkenazim trio, including HG002 with its comprehensive truth dataset, underscores the reliability of the approach in detecting SNPs and indels. These validations establish the utility of the workspace for both research and translational applications, providing a strong foundation for future clinical adaptation.

Conclusion

This long-read WGS pipeline offers a powerful and adaptable solution for comprehensive genomic analysis, addressing the specific challenges associated with long-read data by incorporating optimized processing steps. By supporting diverse sample structures, offering robust error correction and variant detection, and providing customizable computational resources, this pipeline enhances human genomic studies and clinical investigations. Notably, its flexibility allows it to be used effectively in both small-scale familial studies and large-scale population research. The validation results, including metrics such as high sensitivity and precision in SNP and indel detection, affirm its reliability and precision, positioning it as a valuable resource for future genomics research. Consequently, this pipeline not only advances research capabilities but also holds significant promise for translational and clinical applications, including the identification of rare variants, inheritance pattern analysis, and comprehensive methylation profiling, promoting broader utilization of long-read sequencing technologies in genomics.

References

1. https://github.com/PacificBiosciences/HiFi-human-WGS-WDL/releases/tag/v2.0.0-rc2
2. https://terra.bio/

Contact

Mark Fleharty
fleharty@broadinstitute.org