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

Mark Fleharty, Shadi Zaheri, William J Rowell, Michelle Cippicho, Emma Walker, Lisa Anderson, Kiran Garimella, Heather Rooke, Namrata Gupta, Sean Hofherr and Niall Lennon

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. Additionally, we used T2T-ACE, a tool we have created for the purpose of evaluating the positive predictive value (PPV) for copy number variant (CNV) detection.

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