Application note

Consolidated analysis with the PacBio WGS Variant Pipeline

Introduction

The increased throughput of PacBio® HiFi sequencers improves the cost and operational efficiency for human whole genome sequencing. This resulting growth in data yield has enabled breakthrough discoveries in human and population genomics yet underlines the need for the available analysis tools to pace with data generation.

This Application note introduces the PacBio WGS Variant Pipeline that consolidates several state-of-the-art analysis tools into a single user-friendly workflow. This piece also serves as a best practice guide for streamlining the analysis of HiFi data to generate compelling biological insight.

The WGS Variant Pipeline is powered by the Workflow Description Language (WDL), a programming language for genome analysis workflows created by the Broad Institute and maintained by the OpenWDL community. WDL defines the structure of a data processing workflow with a human-readable and writable syntax. Built from this widely adopted language, the WGS Variant Pipeline also offers the flexibility to use at the command line or through platforms provided by PacBio Compatible partners so that users with different levels of bioinformatics experience may access the pipeline.

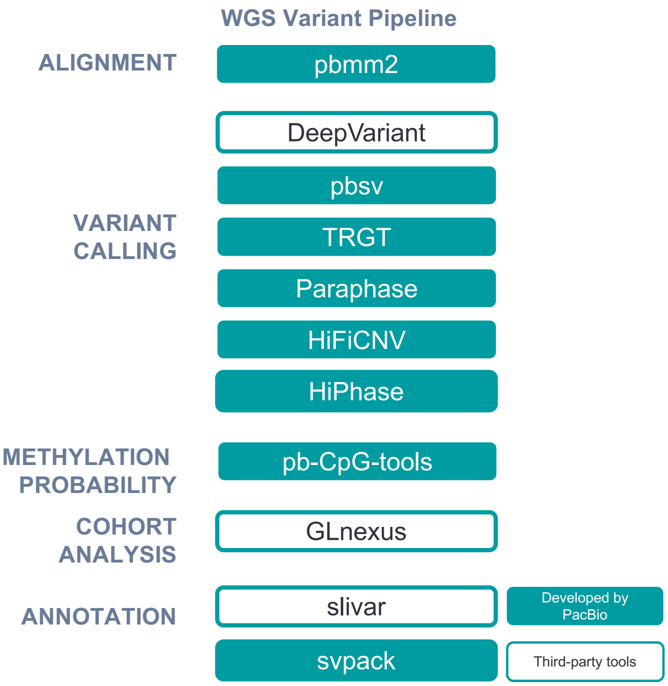

Figure 1. Overview of the WGS Variant Pipeline for secondary and tertiary sequencing analysis.

The WGS Variant Pipeline consolidates secondary and tertiary analysis tools spanning from alignment to annotation (Figure 1) developed by PacBio and third-party groups.

This pipeline packages tools for HiFi alignment, comprehensive variant callers for single nucleotide variants (SNVs), small insertions and deletions, copy number variants (CNVs), and structural variants (SVs), tandem repeat genotyping, typing of genes within segmental duplications, phasing of small variants and large insertions and deletions, and summarization of 5-methylcytosine-pG (5mCpG) methylation probabilities. For multi-sample cohorts, joint calling is included for small and structural variants. Finally, annotation tools are included for small and structural variants (Table 1).

Analysis	Tool	Description
Alignment	pbmm2	PacBio compatible minimap2 wrapper for alignment to reference
Variant calling	DeepVariant	SNV and small indel caller
	HiFiCNV	CNV caller
	pbsv	Structural variant caller
	Paraphase	Variant caller for genes within segmental duplications
	TRGT	Tandem repeat genotyper
	HiPhase	Read phasing for small variants from DeepVariant and structural variants from pbsv
Methylation probability	pb-CpG-tools	Summarize site probabilities for 5mCpG methylation
Cohort analysis	GLnexus	Joint calling for small variants
	pbsv	Joint calling for SVs
Annotation	slivar	Annotation tool for small variants
	svpack	Annotation tool for structural variants

Table 1. Analysis tools available in the WGS Variant Pipeline

Workflow setup and execution

For users with all levels of bioinformatics experience, this workflow can be executed through compatible software platforms, including those provided by DNAnexus, FormBio, Terra, and DNAstack. These platforms have been vetted by PacBio and allow users to run bundled workflows in a managed system. This option provides additional support for compliance and security without the use of internal resources needed for compute requirements.

For bioinformaticians who can negotiate command line tool installations and hardware configuration, the pipeline can also be installed from scratch onto one's preferred compute system. The pipeline setup requires configuration of execution engines and input according to the selected backend environment (Table 2).

Choose a backend	• Pre-configuration templates are available for the following environments: HPCS, AWS, Azure, and GCP.
Configure a workflow execution engine	• Cromwell is supported on HPC, AWS, Azure, and GCP. • Miniwdl is supported on HPC and AWS.
Clone the repo and configure input	• The PacBio WGS Variant Pipeline is publicly available on PacBio GitHub. Users can clone the repo to their corresponding environment. • Reference data used in the analysis is available for download. A copy is also available at each supported cloud environment. • An input JSON template file is present in the workflow repo. The template should be updated for any customization and point to the input data.
Run analysis	• The workflow can be directly run on the machine from which the workflow has been deployed. • Depending on the configuration, workflows may also be submitted via schedulers or other methods.

Table 2. WGS Variant Pipeline workflow setup.

The workflow can be run on Azure, AWS, GCP, and HPC. The choice of backend is largely determined by data location. The open-source workflow engines Cromwell and miniwdl are supported for execution.

Following engine configuration, the repo is cloned to the chosen environment and input template files are configured. Backend-specific documentation and templates are available at the WGS Variant Pipeline GitHub. The workflow can be directly run on the machine from which it was deployed (see Figure 2 for sample run summary).

Conclusion

The consolidation of these analysis tools into the WGS Variant Pipeline streamlines the analysis of HiFi sequencing data into a single accessible, scalable workflow. This simplified process enables users to focus on biological insight and discovery through the interpretation of human whole genome datasets with state-of-the-art tools.

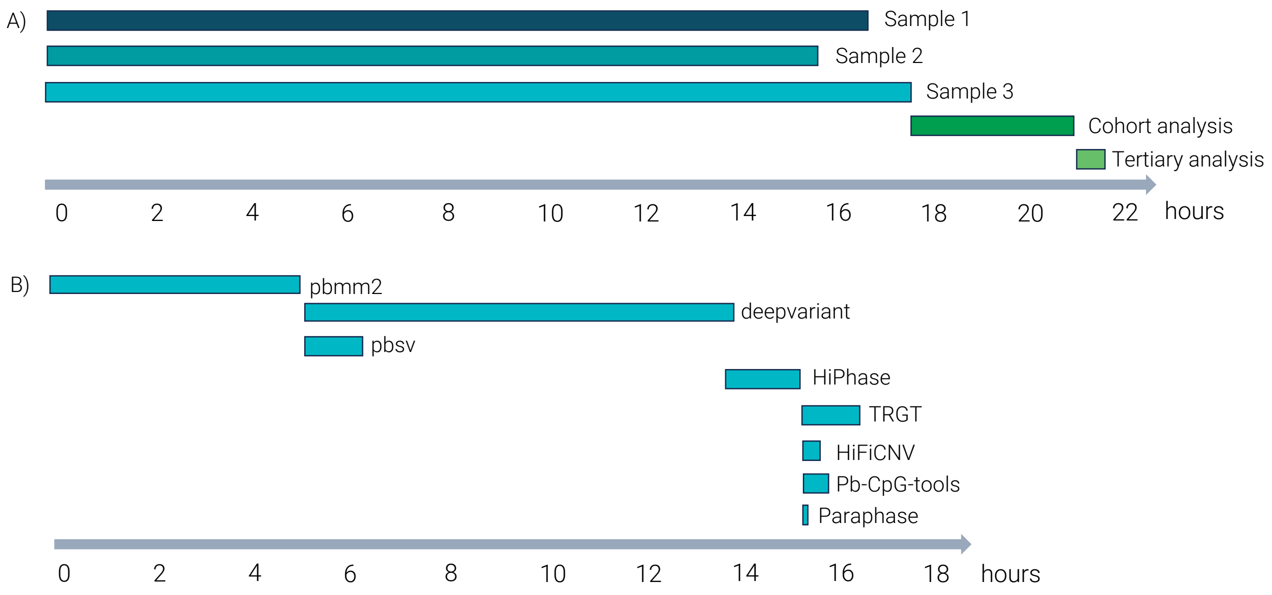

Figure 2. Run time summary for WGS Variant Pipeline performed on HG002/HG003/HG004 trio data. This analysis was performed on a large commercial HPC infrastructure without queuing time. A) Overall summary for the trio analysis. B) Run time for each major component in a per-sample analysis.

Resources

WGS Variant Pipeline GitHub

Open WDL community

Reference datasets

Partner information

PacBio Compatible program

FormBio

DNAnexus

Terra

DNAstack

Technical resources

Cromwell

Miniwdl

DeepVariant

GLnexus

slivar

For additional information on computational tools and workflows for HiFi applications visit the PacBio Computational tools webpage.

Note: Command line tools are considered in development, can be updated frequently, and are not supported as part of the PacBio software product suite.

Research use only. Not for use in diagnostic procedures. © 2023 Pacific Biosciences of California, Inc. ("PacBio"). All rights reserved. Information in this document is subject to change without notice. PacBio assumes no responsibility for any errors or omissions in this document. Certain notices, terms, conditions and/or use restrictions may pertain to your use of PacBio products and/or third-party products. Refer to the applicable PacBio terms and conditions of sale and to the applicable license terms at pacb.com/license. Pacific Biosciences, the PacBio logo, PacBio, Circulomics, Omniome, SMRT, SMRTbell, Iso-Seq, Sequel, Nanobind, SBB, Revio, and Onso are trademarks of PacBio.