Long-Read Variant Calling HPC Pipeline:

Scalable and Efficient Genomic Analysis using Nextflow + SLURM

Priyanshu Agarwal, Sai Deepak Gorla, Sudhanva HR; **Strand Life Sciences, Bangalore, India**

VISIT OUR **WEBSITE**

strand

Contact

Priyanshu Agarwal
Senior Associate Director - Solutions

priyanshu@strandls.com

Introduction

With a growing interest in long-read sequencing, there is a need for fast and efficient variant calling pipelines. Our Solutions team has worked to streamline the execution of PacBio tools using Nextflow and the SLURM job scheduler, ensuring scalability, reproducibility, and efficient resource utilization in high-performance computing (HPC) environments.

In a conventional setup, it takes up to **3.5 hrs** to process a WGS with **~7 million** reads. Our setup has been able to achieve the same outcome in under **2 hrs**.

Long-read Variant Calling

Long-reads have **higher error rates but provide better resolution for structural variants**, making specialized tools necessary to accurately detect large insertions, deletions, and complex genomic rearrangements.

Our pipeline uses specific tools for each category of variants, as below:

Tool Name	Type of Variant Detected
PBSV	Structural Variants (SV)
HiFi-CNV	Copy Number Variations (CNV)
Paraphase	Phasing of variants – Haplotyping and allele-specific variant calling
TRGT	Tandem Repeat Variations

Achievements & Highlights

- **>50% Improved Computational Efficiency** by distributing workloads across multiple nodes.
- **Processing capacity** increased 3.7x from 6.4k samples/mo in standard setup to 24k samples/mo in a 4-node setup.
- **Reproducibility & Automation** to ensure consistency across multiple runs.
- **Enhanced Data Processing with Parallelism** through the integration of **Nextflow** and **SLURM**, enabling parallel execution of multiple tools.

Workflow Overview

Input Handling: BAM data processing.

Variant Calling: Running tools like PBSV, HiFi-CNV, Paraphase, and TRGT.

Parallel Execution: Nextflow distributes jobs across SLURM nodes.

Pipeline Architecture

- **Nextflow [Workflow Engine]:** Automates execution, manages dependencies, enables parallel execution, and handles failures through automated retries and checkpoints.
- **SLURM [Job Scheduling]:** Distributes workload across multiple computing nodes, optimizing performance by ensuring resource-aware scheduling.
- **High Scalability:** The pipeline is designed to scale seamlessly across multiple compute nodes.
- **Data Management:** Structured handling of intermediate & final results; easily track and analyze outputs while maintaining reproducibility.

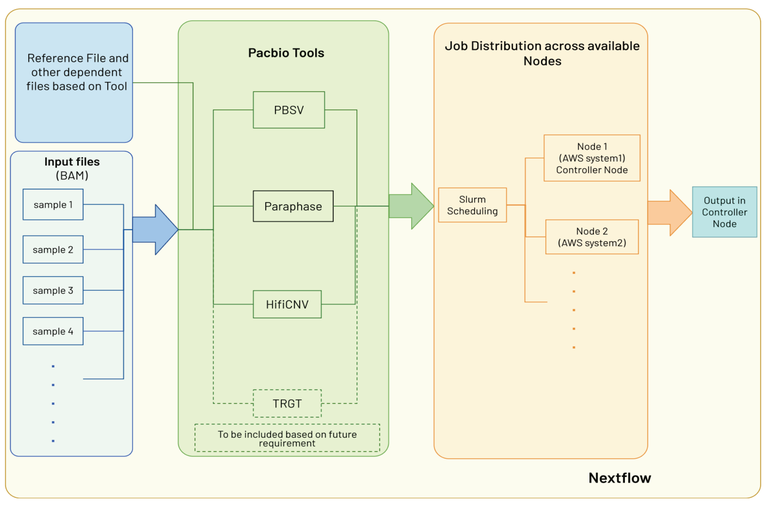

Fig 1: Pipeline describing the input files and reference files provided to Pacbio Tools in the Nextflow platform which are distributed to multiple computational nodes by SLURM Scheduler and the output is stored in Controller Node.

Performance Optimization

To ensure efficient execution, we implemented various optimization strategies:

- **Dynamic Resource Allocation:** Each process requests only the necessary CPU and memory resources, preventing overuse and ensuring balanced workload distribution.
- **Intelligent Job Scheduling with SLURM:** Tasks are assigned to nodes based on resource availability, maximizing throughput and minimizing idle time.
- **Checkpointing & Auto-Retry Mechanism:** Failed jobs are automatically retried at the last successful checkpoint, reducing manual intervention and improving pipeline robustness.
- **Optimized Data I/O Management:** By leveraging parallel file systems and efficient data caching, the pipeline minimizes disk I/O bottlenecks and accelerates data processing.

Benchmark Results

Tool	Standalone Runs (Single Node, Single 7M reads Sample)	Standalone Runs (Single Node, 16 Sample each ~430k reads)	Pipeline (2 Nodes, Single 7M reads Sample)	Pipeline (2 Nodes, 16 Samples)	Pipeline (4 Nodes, 16 Samples)
PBSV	88 min	86 min			
HiFi-CNV	6 min	6 min			
Paraphase on 8 threads	20 min	16 min			
Total Time (min)	114	108	92	61	29

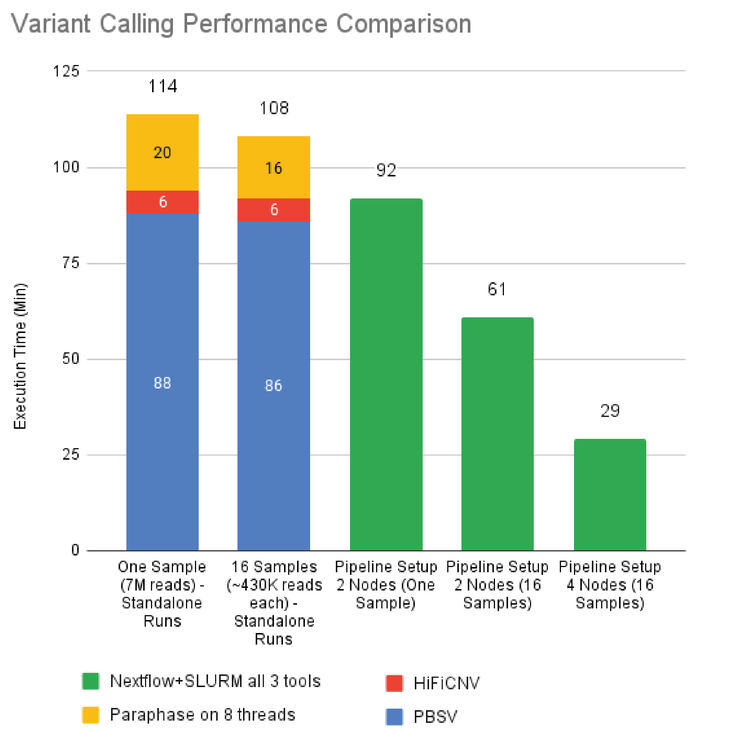

Fig 2: Execution time comparison of PacBio tools in sequential vs. parallel execution using Nextflow + SLURM, demonstrating significant speedup with increased compute nodes.

Conclusion

By integrating PacBio tools with **Nextflow and SLURM**, we significantly improved the **Turn Around Time (TAT)** for long-read variant calling. The pipeline provides a robust solution for large-scale genomic workflows, offering scalability, reproducibility, and efficient resource utilization.

Future Work

- Enhancing dynamic cloud scalability by integrating auto-scaling and cost-optimized resource allocation across cloud-based HPC environments.
- **Data Merging & Output Generation:** Combining results into a final VCF
- Further Optimization for Cost-Effective Execution
- Enhancing Monitoring with Nextflow Tower
- Extending Compatibility to Other Sequencing Platforms
- Automating Data Preprocessing and Quality Control