Optimizing CNV Workflow:
Bridging R&D to Production for Efficiency

Ramyashree H J, Krishna Kolla, Pratibha Chitipothu,
Mahesh Nagarajan, Anand Janakiraman; **Strand Life Sciences, Bangalore, India**

VISIT OUR **WEBSITE**

strand

Contact

Pratibha Chitipothu
Senior Director - Software

pratibha@strandls.com

Introduction

Copy number variations (CNVs) play a crucial role in the diagnosis of neurological and other rare diseases (RDs), contributing to 6-7% of positive cases in Whole Exome Sequencing (WES) tests. With the rapid development of whole exome sequencing, an increasing number of tools are being proposed for copy number variation detection. Also, verifying whether the variant is genuine or not before reporting is critical.

R&D: CNV Workflow

The CNV workflow uses R&D scripts to verify CNVs called by Dragen in the main pipeline. Dragen's static profile method may cause false positives/negatives due to inter-sample coverage variability. To improve accuracy, we developed an in-house CNV pipeline that creates dynamic profiles using previously sequenced samples, selecting the top 50 most similar ones. The pipeline then compares normalized coverage of samples across the selected profile samples to calculate copy number and z-scores. By combining two CNV detection methods, this approach increases call confidence.

Problem Statement

Taking bioinformatics scripts from R&D to production presents challenges due to differing goals. In R&D, scripts prioritize flexibility and rapid experimentation, while in production, efficiency, scalability, and reliability are key.

Opportunities for improvement include enhancing reproducibility, optimizing error handling, and managing large datasets more effectively. By focusing on performance, documentation, error management, and automation, we can ensure smooth deployment and efficient production operation.

Concordance Test

A thorough concordance test on a 100-sample run confirms that the new Nextflow pipeline maintains the integrity of results while delivering substantial time savings. A comparative bar graph—illustrating overall execution times between the R&D and Production approaches—clearly demonstrates the performance gains of our new approach.

Achievements

- **Developed a fully automated Nextflow-based CNV pipeline to streamline** analysis and eliminate manual intervention.
- **Integrated the pipeline into the WES workflow**, enabling seamless, high-throughput processing.
- **Enhanced scalability and efficiency**, allowing larger datasets to be processed with minimal effort.
- **Overall reduction of CNV pipeline time by ~80%**
- **Reduced AWS costs from $1.10 to $0.18 per 1 TB of data.**

Production Pipeline

The automation of the CNV pipeline workflow through a Nextflow-based pipeline (Figure 1) has significantly reduced processing time.

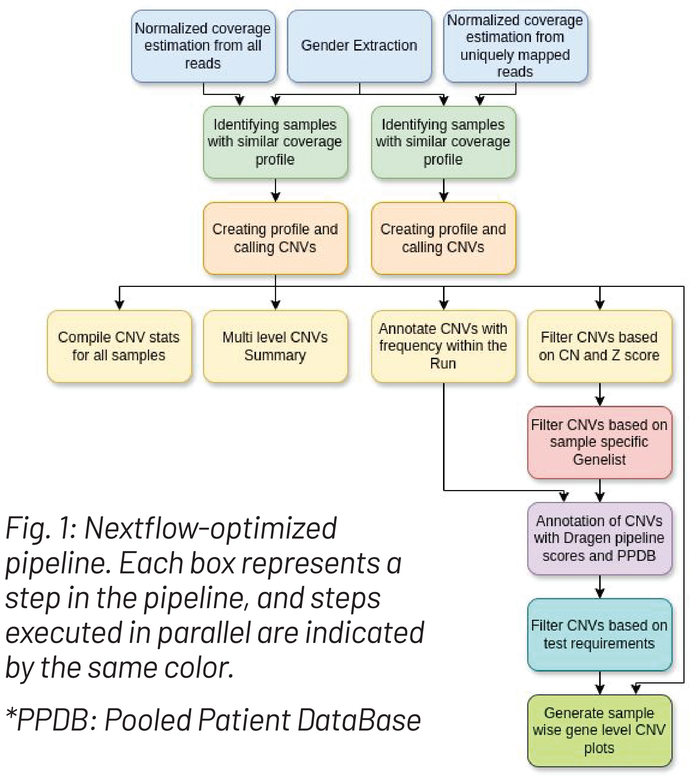
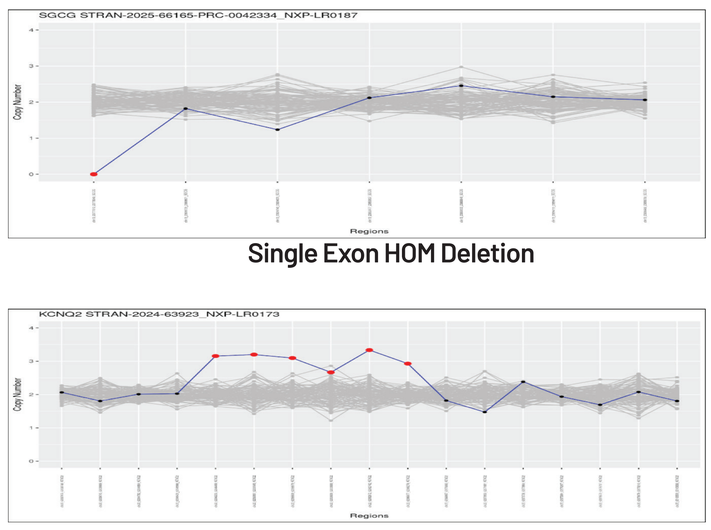


Fig. 1: Nextflow-optimized pipeline. Each box represents a step in the pipeline, and steps executed in parallel are indicated by the same color.

**PPDB: Pooled Patient DataBase*

The above optimized pipeline enabled the rapid generation of output files (See the plots in Figure 2) ensuring accurate and fast identification of clinically significant CNV events which are essential for clinical interpretation.



Single Exon HOM Deletion

Multi Exon Amplification

Fig: 2. A Homozygous deletion of exon -1 in the SGCG gene, may be associated with autosomal recessive limb-girdle muscular dystrophy-5 (LGMDR5). 2. A duplication of exons 8-13 in the KCNQ2 gene, may be associated with developmental and epileptic encephalopathy-7 (DEE7).

R&D to Production Transition

Hands-Free Execution

Transitioning to Nextflow has revolutionized our pipeline by fully automating the entire workflow, streamlining processes from start to finish.

Dockerization

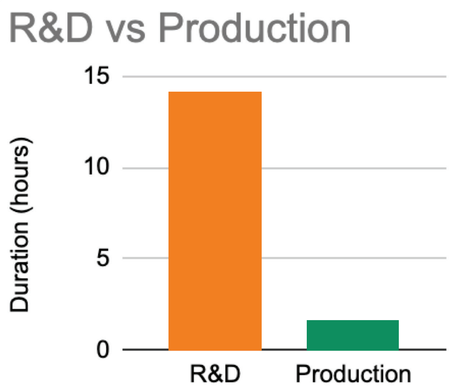
Containerized the pipeline to ensure consistent results across environments, eliminating system-related issues and speeding up deployments for more reliable releases.

Multi-Processing

Leveraging Nextflow's native support for multiprocessing, we run tasks in parallel at multiple levels, processing each sample simultaneously. This speeds up processing and reduces manual intervention.

Optimization

Optimized pipeline scripts, particularly those for generating plots, by utilizing more efficient data structures and refining the code to minimize memory usage.



Batch Processing

To manage large datasets more efficiently, we've implemented batch processing, optimizing data-intensive operations and improving scalability.

Fault Tolerance

The pipeline is designed with robust fault tolerance, allowing it to resume from failures and support reruns, ensuring continuous operation even in the event of interruptions.

Improved Logging

Our enhanced logging system captures detailed script logs and tracks infrastructure usage through execution traces, timelines, and Nextflow reports, improving monitoring, troubleshooting, and performance analysis.

These enhancements collectively have significantly improved performance, increased efficiency, and ensured a smooth transition from R&D to production, offering a more reliable and scalable solution.