Optimizing Methylation Pipeline:
Migration from Snakemake to Nextflow and Deployment on Seqera Platform for Efficiency and Cost Benefits.

Pavan Kotha, Neha Bhojani, Nishant Shekhar, Mayur Saini, Juhi Pandey, Ruthvik Bobba, Jaya Singh

Strand Life Sciences, Bangalore, India

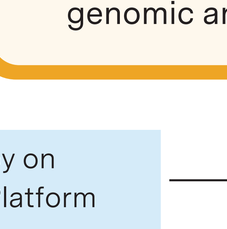
strand
WE SOLVE PROBLEMS IN GENOMICS.

VISIT OUR **WEBSITE**

Conference Name
Bio-IT World Conference & Expo 2024

Dates - April 15th to 17th

Place - Boston, MA, USA



Contact

Jaya Singh
Senior Director, Variant Science Solutions

+1 (916) 873-4726

jaya@strandls.com

Introduction

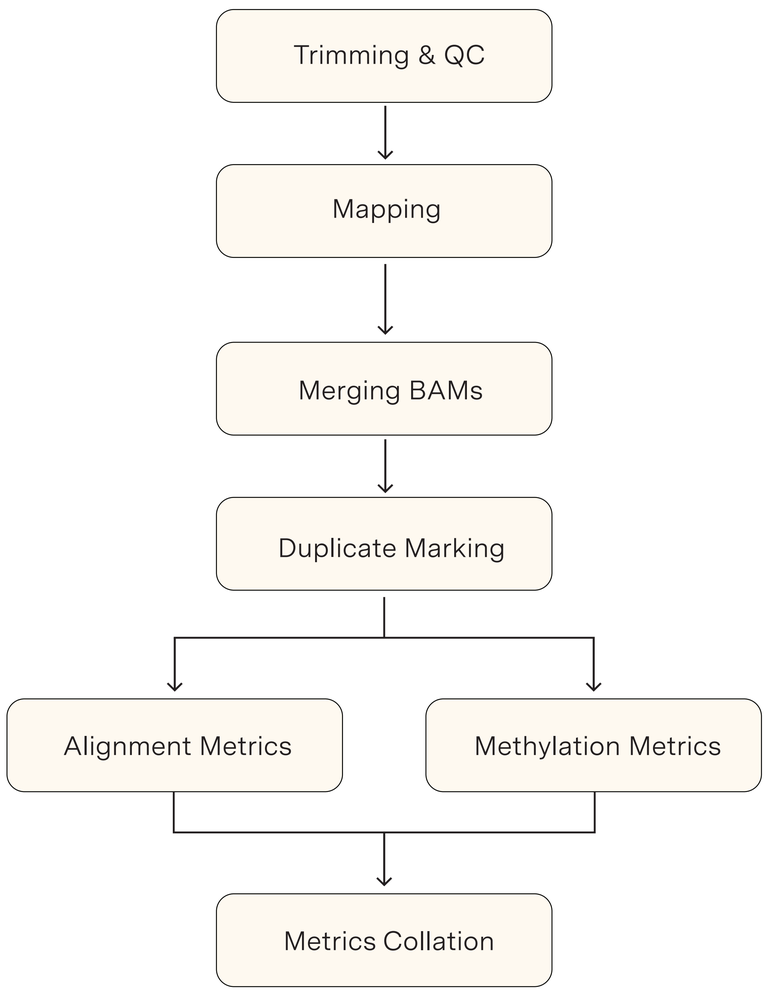
1. Reproducibility, scalability, and standardization are essential aspects of computational pipelines.
2. **Reproducibility** ensures the validity and transparency of research findings, **scalability** enables pipelines to handle increasing computational demands efficiently, and **standardization** promotes consistency.
3. Additionally, **interoperability** across diverse environments (such as local workstations, high performance computing clusters, or cloud infrastructure), fosters easy collaboration and ensures that pipelines behave consistently with **minimum compatibility issues** and **maximum flexibility** for further development and research.

Migration to Nextflow

1. Nextflow is a domain-specific language (DSL) for scalable and reproducible scientific workflows across various computing environments, incorporating features like data and task parallelism for efficient computational analyses.
2. Nextflow provides a lot of resource metrics that help in optimizing the resource allocation
3. Along with this Nextflow offers better tracing and visualization of the pipeline execution that helps analyze the pipeline performance and behavior.
4. There are several cloud native platforms that are optimized for nextflow pipelines:
   - Seqera Platform
   - AWS HealthOmics

We migrated our existing pipeline from Snakemake to Nextflow to leverage these robust features for workflow management and execution.

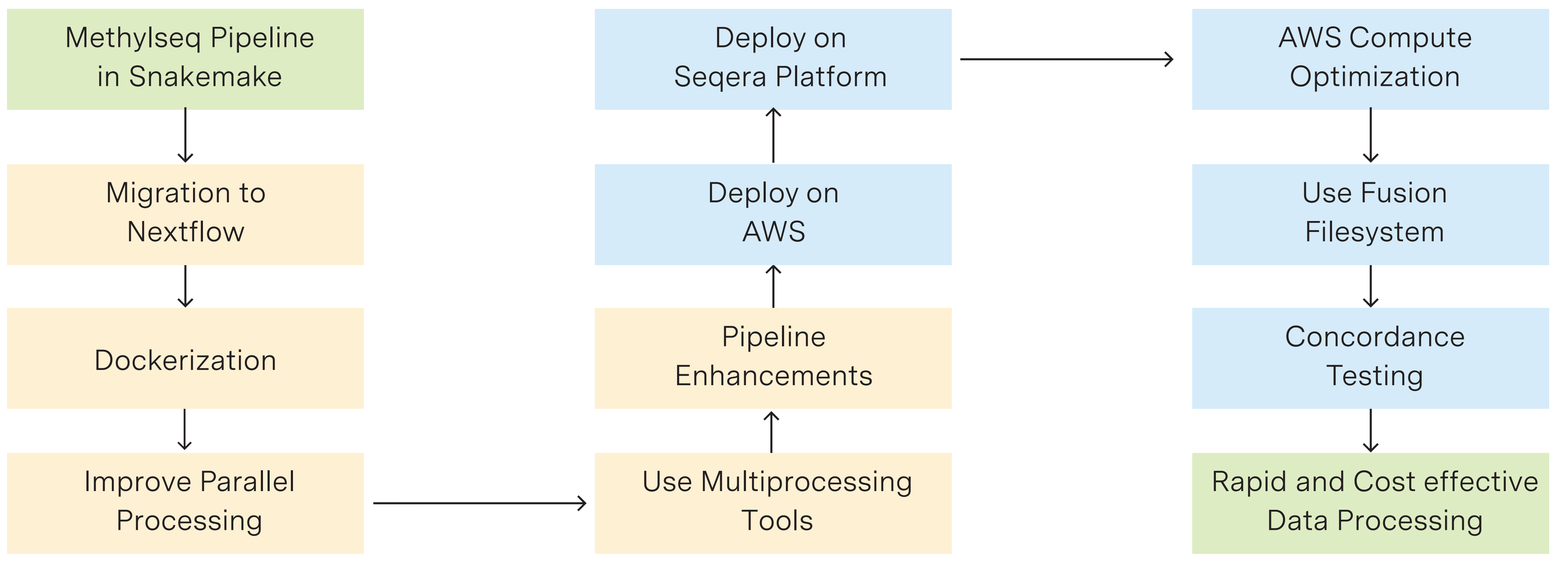
Fig1: Methylseq pipeline



Advantages of Seqera Platform* (Nextflow Tower)

1. Simplifies workflow management with an intuitive web interface, provides interactive visualization and detailed job monitoring for efficient workflow tracking.
2. Facilitates collaborative workflow development and sharing, enhancing team productivity, enables non-technical users to run pipelines via an intuitive interface.
3. Seamlessly integrates with cloud-based compute environments like AWS Batch for automated workflow execution.

Fig 2. Cloud Migration



Deployment on Seqera Platform

We deployed our pipeline to Seqera Platform for several reasons.

Firstly, we migrated to the cloud, leveraging AWS Batch with AWS ECS, which dynamically provisions compute instances as needed, reducing costs by over 50% compared to static EC2 instances. This setup easily scales to over thousands of vCPUs and processes ~100 samples simultaneously, an accomplishment that was challenging with static instances or cloud HPC.

Secondly, we utilize faster AWS EC2 instances, specifically the 6th generation c6id, which completes tasks 50% faster. Additionally, we integrated the Seqera Fusion file system to optimize job execution, improve efficiency, and minimize data transfer time, resulting in pipelines executing twice as fast and at half the overall cost when combined with the 6th gen EC2 instances.

Dockerization of the Pipeline

By dockerizing our pipeline, we encapsulated dependencies, ensuring reproducibility across various environments. This optimization not only improved portability but also slashed runtime and costs by optimizing resource utilization. Docker's flexibility also streamlines deployment, enabling easy image transfer across platforms, enhancing scalability, and minimizing overhead.

Multiprocessing / Parallelization

Large files underwent parallel processing by breaking them into smaller chunks, enabling simultaneous processing of these segments. This approach replaced tools like Picard, lacking multiprocessing support, with Sentieon toolkit, offering parallelized versions of the same tools, yielding over 20 times faster processing speeds.

Concordance Test

To validate the effectiveness of our optimizations, we conducted a concordance test comparing the results of the optimized pipeline with those from the original implementation. Our findings demonstrated very high concordance, confirming the reliability and accuracy of the optimized pipeline.

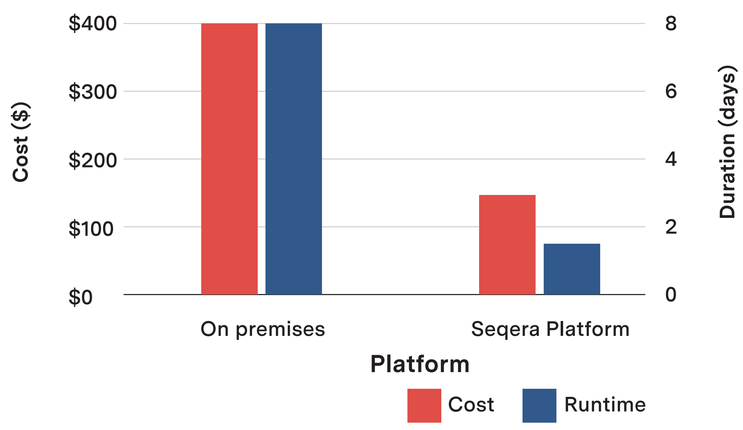
Summary

1. By porting our pipeline to **Nextflow**, dockerizing the workflow steps, deploying it on the **Seqera** platform, and optimizing performance by enabling multiprocessing with toolkits like **Sentieon**, we achieved significant improvements in scalability, efficiency, reproducibility, and cost effectiveness.
2. These enhancements helped **standardize the code,** paving the way for robust and **reliable computational pipelines** in genomic analysis and beyond.

Achievements

1. Reduced pipeline runtime from **8 days to just 1.5 days**, boosting efficiency and accelerating data analysis. This drastic reduction in TAT enabled our customer to increase processing from **1 batch/week** (10 samples, 0.5 TB) to **20 batches/week** (200 samples, 10TB) ensuring that the pipeline is not a bottleneck.
2. Achieved a remarkable **~66% reduction in costs** by streamlining resource utilization and minimizing overheads.

Cost and Runtime Comparison



Seqera Platform* previously known as Nextflow Tower, is the centralized command post for data management and workflows.