Nextflow-Powered MOA-Seq Pipeline Enables Scalable Analysis of Regulatory Elements

Jacopo Tartaglia[1]; Mario Giorgioni[2]; Wolfgang De Salvador[3]; Primetta Faccioli[1]
[1]CREA - Research Centre for Genomics and Bioinformatics; CREA - Research Centre for Forestry and Wood; [3]Microsoft

Introduction

Plant phenotypic variation largely depends on transcriptional regulation, driven by transcription factors (TFs) binding to specific DNA sites. Mapping these sites genome-wide reveals active transcriptional programs and identifies *cis*-regulatory elements (CREs), which are keys to understanding how plants adapt to environmental changes, for example in tolerating biotic and abiotic stress (*Liang et al., 2022*). However, the annotation of these elements remains challenging, making comprehensive maps of TF binding sites essential for linking genotype to phenotype (*Engelhorn et al., 2025*). Existing methods typically focus on pinpointing binding regions for individual TFs using techniques such as ChIP-seq, or broadly identifying accessible chromatin regions (ACRs) through approaches like ATAC-seq. However, these strategies often lack the fine-scale resolution and consistency required to accurately delineate precise TF binding sites across the whole genome (*Savadel et al., 2021*). To overcome this limitation, MNase-defined cistrome-Occupancy Analysis (MOA-seq) was developed as a high-resolution, high-throughput, genome-wide method specifically designed to accurately identify putative TF binding sites, called MOA-footprints (MFs), and to reveal accessible chromatin regions (*Savadel et al., 2021*). MOA-seq produces massive datasets, making analysis complex and resource-intensive. To address this, the publicly available pipeline (*Liang et al., 2022*) was redesigned using Nextflow and tested on Azure cloud, creating a modular, automated, and portable solution. The results presented here can be considered as a proof of concept for the application on future works.

Materials and Methods

The Nextflow-powered MOAseq pipeline provides a portable workflow for MOA-seq data analysis, based on the pipeline developed by *Liang et al., 2022*, and built upon a wide range of *state-of-the-art* software and data resources. The pipeline is implemented using Nextflow (*Di Tommaso et al., 2017*) with DSL2 syntax, a widely adopted workflow management system for bioinformatics applications. The whole pipeline is organized into modular processes, ensuring clarity, reusability, replicability and ease of implementation. Main Nextflow benefits are shown in **Fig. 1**. The pipeline consists of three main phases: data pre-processing (orange), alignment (purple), and peak calling (blue). Replicates can be kept separate (default case) or merged (double red/green line in **Fig. 2**). If replicates are merged, samples from different experimental conditions (e.g., treatment and control) cannot be run at the same time. To date, software used in the MOAseq pipeline are run and containerized with Docker.

Fig. 1. Main benefits of Nextflow scientific workflow system.

Results and Discussion

The samples from the article of *Liang et al. (2022)* were used to test the Nextflow-powered MOAseq pipeline. Metrics used to assess pipeline performance included: the number of merged inputs reads with the percentage of uniquely mapped reads; the number of MOA peaks identified using full-length reads and reads trimmed to 20 bp (MFs); and the median MOA peak length for both full-length and shortened reads. **Table 1** reports inputs reads metrics. Results obtained with Nextflow-powered MOAseq, while not identical, closely match those reported in the original study. The deviation in input reads per replicate ranged from 0.09% to a maximum of 0.20% fewer reads. Unique alignments rates were also comparable. **Table 2** shows the number of MFs per sample. The results were highly consistent with the original ones, with differences in peak counts ranging from 0.02% to 0.065% more MFs. Median peak lengths were consistent: 180 bp for full-length peaks versus 179 bp reported, and 34 bp for shortened peaks, matching the original value.
The Nextflow-powered MOAseq pipeline has been successfully tested both on a local machine and on cloud environment (Microsoft Azure). On Azure, 87.1 GB of raw data were processed in 2 hours and 11 minutes, (280.5 core-hours).

Table 1. Number of input merged reads (unique alignment rate).

	Liang et al., 2022	MOAseq
Control rep. 1	126.939.893 (45%)	126.716.389 (46%)
Control rep. 2	166.432.747 (44%)	166.101.382 (45%)
Control rep.3	138.693.816 (40%)	138.569.295 (45%)
Stress rep. 1	158.734.914 (45%)	158.500.699 (49%)
Stress rep. 2	131.979.910 (46%)	131.850.844 (50%)
Stress rep. 3	160.081.276 (44%)	159.861.675 (47%)

Table 2. Number of MOA footprints (MFs, shortened peaks).

	Liang et al., 2022	MOAseq
Control rep. 1	142.815	142.908
Control rep. 2	184.536	184.539
Control rep.3	131.872	131.877
Stress rep. 1	202.906	202.920
Stress rep. 2	159.774	159.744
Stress rep. 3	172.688	172.722

Conclusion

Although further refinement are needed, the current implementation shows that results are reliable. It should be emphasized that the implementation using Nextflow and the containerization system (in this case, Docker) makes the use of the pipeline extremely straightforward both locally and in the cloud, with the sole requirement of having Nextflow and Docker installed in the respective environments. Future developments will focus on differential analysis of MOA footprints and motif discovery.

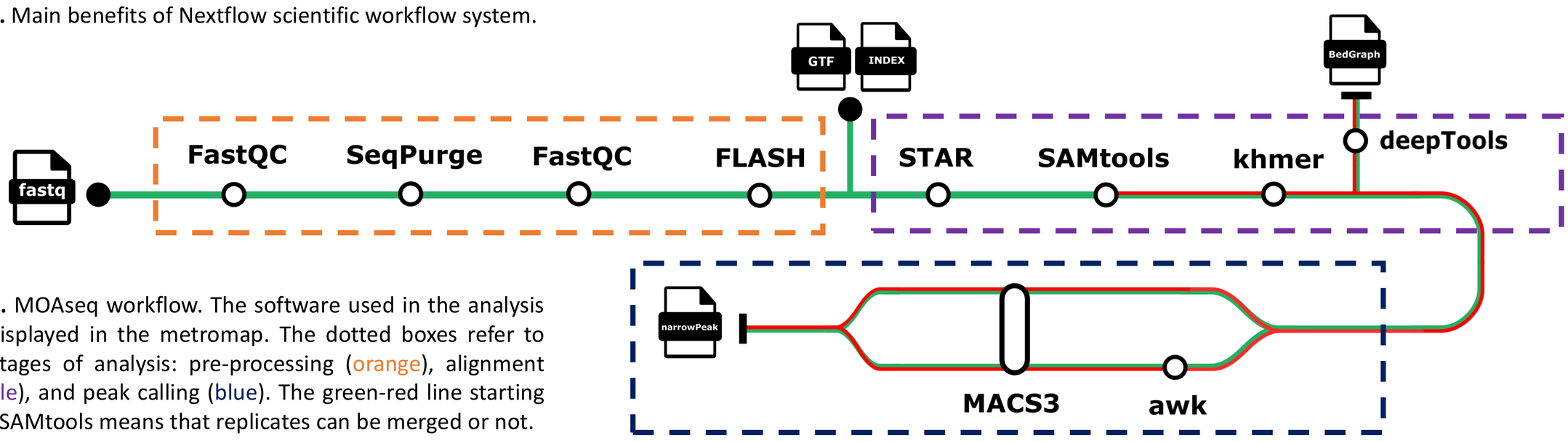

Fig. 2. MOAseq workflow. The software used in the analysis are displayed in the metromap. The dotted boxes refer to the stages of analysis: pre-processing (orange), alignment (purple), and peak calling (blue). The green-red line starting from SAMtools means that replicates can be merged or not.

Contact

Jacopo Tartaglia
CREA - Research Centre for Genomics and Bioinformatics
Via S. Protaso 69, 29017 Fiorenzuola d'Arda (PC), Italy
jacopo.tartaglia@crea.gov.it

References

- Di Tommaso, Paolo, et al. "Nextflow enables reproducible computational workflows." *Nature biotechnology* 35.4 (2017): 316-319.
- Engelhorn, Julia, et al. "Genetic variation at transcription factor binding sites largely explains phenotypic heritability in maize." Nature Genetics (2025): 1-10.
- Liang, Zhikai, et al. "Mapping responsive genomic elements to heat stress in a maize diversity panel." *Genome biology* 23.1 (2022): 234.
- Savadel, Savannah D., et al. "The native cistrome and sequence motif families of the maize ear." *PLoS genetics* 17.8 (2021): e1009689.