Advancing the pace of OMICS research and discovery with a full-stack genomics analysis platform based on open science

Evan Floden, Paolo Di Tommaso, Phil Ewels, and Harshil Patel — Seqera, Barcelona, Spain

Abstract

Increasingly, multiple omics disciplines are being combined to study biological processes in a more comprehensive manner. Multi-omics techniques are used in everything from personalized medicine and dosing control to population-level studies to developing new therapeutics.

Traditionally, analysis in research and clinical settings has been plagued with technical and procedural challenges, including complexity, reproducibility, auditability, ensuring data provenance, and integrating with public datasets.

Based on an open-science foundation, Seqera enables researchers to accelerate the pace of omics research and discovery by leveraging reusable scientific pipelines in a full-stack analysis environment. Researchers can improve run-times by **40%** and reduce cloud spending by up to **85%**.[1]

Challenges in scaling bioinformatics analysis

Bioinformaticians and clinicians face practical challenges running genomics analysis at scale:

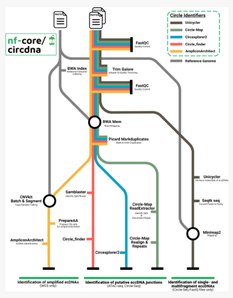
- Orchestrating complex multi-step pipelines to make them robust, scalable, and reproducible.
- Ensuring pipeline portability across computing environments (on-prem, cloud, and multi-cloud).
- Launching, monitoring, and managing pipelines, especially for non-technical users.
- Deploying complex cloud computing infrastructure and on-premises HPC environments.
- Collaborating and securely sharing research and results among local and distributed teams.

Leveraging best practices and foundational open technologies

Researchers need a modern platform that embraces open tools, data sets, and modern software engineering approaches to ensure flexibility and avoid lock-in.

- Support for modern container formats and runtimes
- Interoperability with popular container registries
- Integration with popular source code managers
- Support for all major clouds and HPC environments
- Easy access to public datasets
- Integrations with major sequencing platforms
- Easy access to third-party algorithms & IP

Data-driven computational pipelines fo reproducible workflows



- Fast prototyping
- Unified parallelism
- Stream-oriented
- Resilient, reproducible workflows
- Continuous checkpointing
- **170K+** monthly downloads
- Freely available at https://nextflow.io

A distributed, lightweight file system for cloud-native data pipelines

- A POSIX interface over cloud object storage
- Seamless integration
- **2.2x** throughput gain vs. Amazon S3[2]
- Reduce combined pipeline and storage costs by up to **76%**[2]

seqera
the modern biotech stack

nextflow
workflow management

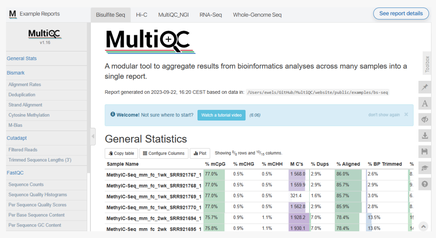
multiqc
reporting and analytics

wave
container provisioning

fusion
cloud native file-system

Aggregate results from bioinformatics analyses across many samples

- Visualize statistics from your pipeline(s)
- Plot all your samples together
- Support for **100+** bioinformatics tools
- Extensible and documented
- Freely available at https://multiqc.info



Next-generation container provisioning and management for data pipelines

- Build containers on the fly
- Dockerfile and Conda recipe support
- Access private container registries
- Deploy containers across multiple clouds
- Augment existing containers

DOCKERFILE

CONDA

Features & supported platforms

- A modern, reactive domain-specific language (DSL)
- Your choice of scripting languages for ease of integration (Python, R, bash, Perl, etc..)
- Enable non-technical users to easily run pipelines via the intuitive Launchpad interface
- Run pipelines on any environment – from workstations to on-prem clusters to private, hybrid, and public clouds
- Collaborate and share data securely among local and remote teams with rich organization and workspace management features
- Connect to external data sources, including open datasets, instruments, LIMS, and databases

Compute Environments

- AWS Batch, Azure Batch, Bridge. Flux Framework Executor, GA4GHTES, Google Cloud Batch, Google Life Sciences, HyperQueue, HTCondor, Ignite, Kubernetes, IBM Spectrum LSF, Moab. NQSII, OAR. PBS/Torque, SGE/Altair Grid Engine, Altair PBS Pro, SLURM

Source code managers

- GitHub, GitLab, Gitea, Azure Repos, AWS CodeCommit

Container technologies

- Docker, Singularity, Shifter, Charliecloud, Podman, Sarus

Curated community pipelines

nf-core

nf-core is a community effort to collect a curated set of analysis pipelines built using Nextflow that is freely available to the bioinformatics community.

Samples of popular nf-core pipelines are below. Visit https://nf-co.re for a complete list.

- **nf-core/rnaseq**—RNA sequencing analysis pipeline using STAR, RSEM, HISAT2 or Salmon with gene/isoform counts and extensive quality control
- **nf-core/sarek**—Analysis pipeline to detect germline or somatic variants from WGS / targeted sequencing
- **nf-core/chipseq**—ChIP-seq peak-calling, QC and differential analysis pipeline
- **nf-core/atacseq**—ATAC-seq peak-calling and QC analysis pipeline
- **nf-core/maq**—Assembly and binning of metagenomes
- **nf-core/ampliseq**—Amplicon sequencing analysis workflow using DADA2 and QIIME2
- **nf-core/nanoseq**—Nanopore demultiplexing, QC and alignment pipeline

Compelling benefits for research

- **Improve productivity**—With a unified view of data, pipelines, results, and compute resources, users can collaborate more effectively and streamline analysis, data generation, and reporting.
- **Reduce costs**—Optimize compute and storage costs, avoid expensive errors, and manage spending across projects and teams more effectively.
- **Reduce complexity**—Research teams can automate tasks, streamline operations, and focus on the science that matters, rather than infrastructure engineering.
- **Simplify compliance**—With reliable, predictable, reproducible, and auditable pipeline execution.

1. How to slash cloud spending with the Seqera platform. https://seqera.io/ebooks/slash-spending-with-tower/
2. Breakthrough performance and cost-efficiency with the new Fusion file system. https://seqera.io/whitepapers/breakthrough-performance-and-cost-efficiency-with-the-new-fusion-file-system/