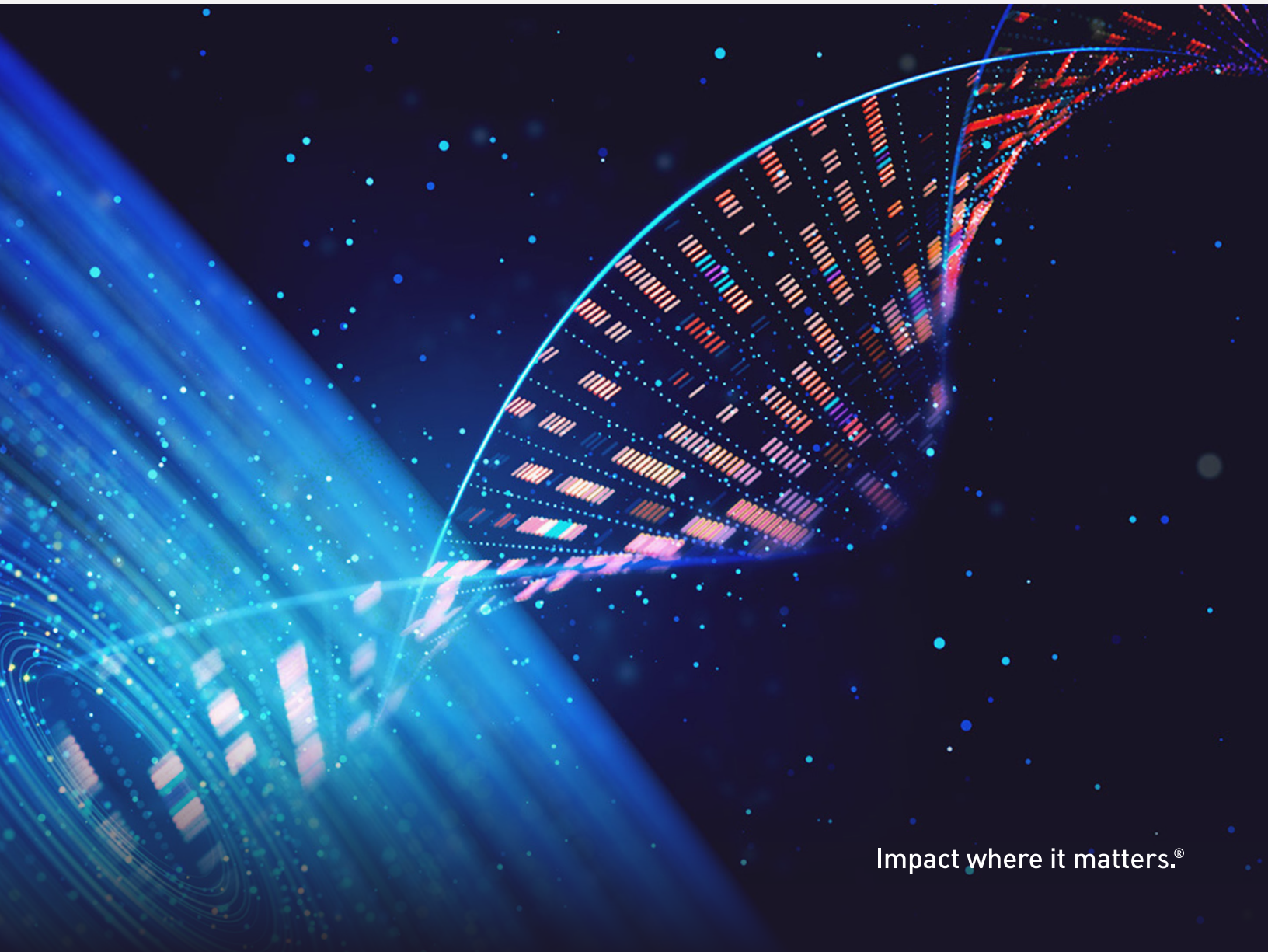
WHITE PAPER

Building robust and reproducible bioinformatics pipelines

Leveraging flexibility, portability and community

By Bruce Press, Brice Sarver, Federico De Masi, Maria Cecilia Argibay, Juan Sendoya, Francisco Avila Cobos, Milena Vujović, Felipe Almeida and Nicolas A. Schcolnicov



Impact where it matters.®

“Implementing robust and reproducible bioinformatics pipelines is a priority for us at Boehringer Ingelheim. Partnering with ZS Discovery to build efficient and scalable workflows using Nextflow has been a game-changer for our target discovery initiatives. This has saved us valuable resources, reducing the hands-on time and streamlining the interpretation of our data.”

— A Boehringer Ingelheim principal investigator

Pipelines are the recipes for analyzing data, and they were originally purpose-built for specific tasks. Individual labs and research cores made their own workflows with their institution's computational equipment—tying them to the software packages and runtime environment used. In the past, researchers seldom shared their analysis pipelines because of a lack of portability—why share a recipe if nobody else can use it? Worse yet, frequently the originator couldn't reproduce their own work because a piece of the pipeline had received an update. In science, reproducibility is king, and a discovery made once isn't truly a discovery.

"Solving complex problems is done best with simple solutions. We've seen how bioinformatics has evolved, and we believe that accessible, portable and shareable tools for workflows will unlock novel discoveries and innovations in the 'omics age,' ultimately improving health outcomes."

— Bruce Press, ZS principal

Dependence on a specific computational environment also hinders scaling a workflow, which typically needs to be deployed on high-performance computing infrastructure, such as a computing cluster. However, each cluster may have a unique scheduling, hardware and software environment. As a result, developing pipelines requires different tests to ensure compatibility with different platforms, which can slow down deployment.

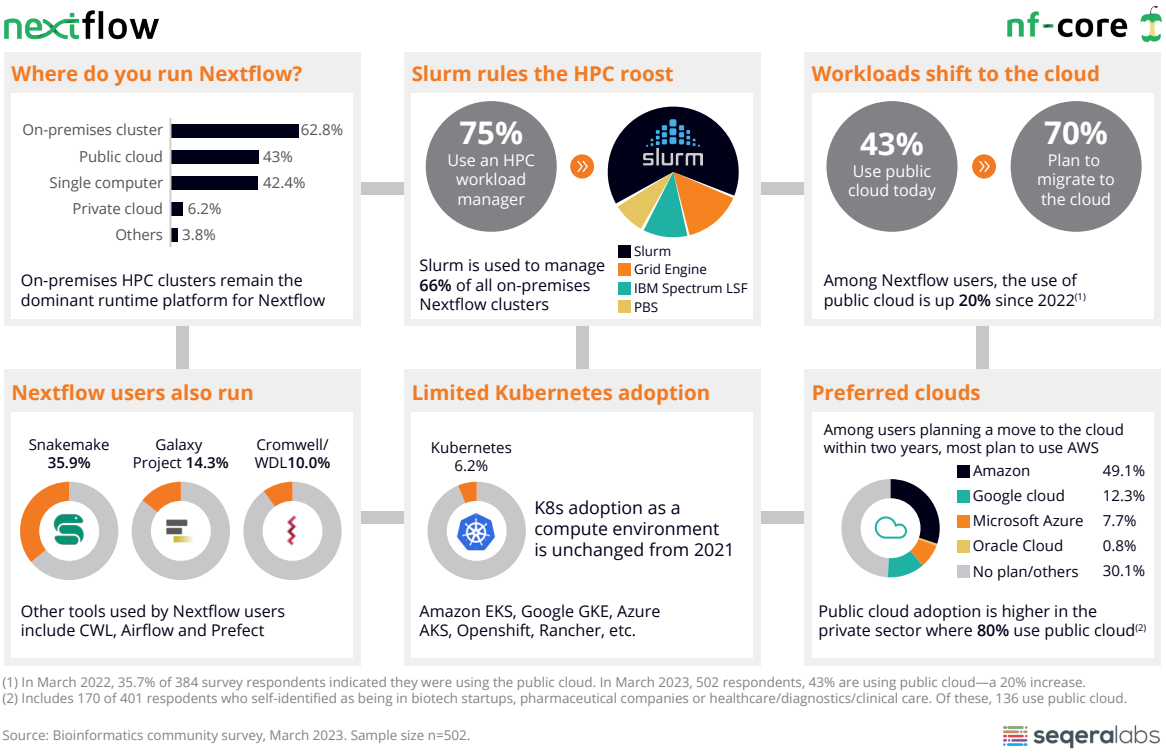
Computational architectures have evolved as new technologies become commonplace. Virtualization allows us to create reproducible environments that can be kept consistent while the host machine's software and hardware change. However, virtualizing an entire pipeline requires extensive storage and is computationally expensive, with each workflow required to be "frozen" along with the operating system and runtime environment.

Using containerization for portability

A similar but lighter-weight approach is containerization. Instead of putting the entire pipeline in a virtual machine, each pipeline step can execute in a container, enhancing reproducibility and portability. Over the past decade, containerization has gained popularity with the release of container management and orchestration platforms.

FIGURE 1:

Nextflow related infrastructure, platform, and workflow usage patterns in 2023



(1) In March 2022, 35.7% of 384 survey respondents indicated they were using the public cloud. In March 2023, 502 respondents, 43% are using public cloud—a 20% increase.
(2) Includes 170 of 401 respodents who self-identified as being in biotech startups, pharmaceutical companies or healthcare/diagnostics/clinical care. Of these, 136 use public cloud.

Source: Bioinformatics community survey, March 2023. Sample size n=502.

seqeralabs

We're firm believers in the advantages of containerized bioinformatics workflow development and execution. Based on our own experience working with these platforms, combined with a partnership with Seqera and its containerized scientific workflow system, Nextflow, we've seen the advantages of containerized bioinformatics workflow development and execution.

Containerized bioinformatics workflows are:

Lightweight. A container holds an application and all its dependencies. It's independent, involves minimal installation and doesn't require specialized external software or environments. Containers are instantiated, loaded and unloaded as needed. Virtual machines, by comparison, need time and resources to load the virtual operating system and runtime environment. This consumes resources throughout the entire workflow process.

Portable and scalable. Since containers hold all the information needed to run a process, applications can be written in any language and moved from one machine to another, regardless of the computing platform. Instead of testing and optimizing the workflow on production environments, Nextflow workflows can be developed on a personal device and be easily moved to cloud or computing clusters. Such workflows produce the same results when run in different environments, such as on macOS and other Unix-based systems. This portability means that established workflows can be easily deployed at new institutions or sites.

Shareable and reusable. Each container in a workflow has a specific function related to a particular analytical step. Since these functions can be used in multiple workflows, containers are reusable in different contexts and with different data without needing to "reinvent the wheel" or manually copy code into a noncontainerized workflow. Individual containers and whole workflows are readily shared online via container registries and code repositories.

"The era and need of shareable and portable workflows across infrastructures is already a reality. Bioinformaticians need technology that helps them focus on making their analyses right, without worrying about porting it to a particular cluster or cloud."

— Felipe Almeida, ZS senior bioinformatician

Putting it together with Nextflow

Nextflow is a scientific workflow ystem that creates and organizes bioinformatics workflows using containers operating sequentially through message passing, where the pipeline orchestration and data processing can be scripted. The output from one container is passed to one or more downstream containers to create a functional pipeline. This allows containers to be assembled in a flexible, scalable, adaptable workflow.

At the same time, Nextflow provides a very efficient resource management system that, together with the reactive asynchronous programming paradigm, makes the most out of available resources through better process parallelization. Nextflow uses this message-passing paradigm to run continuous checkpoints. Intermediate results are tracked and pipeline execution can be resumed from the last successfully executed step, no matter why it was stopped.

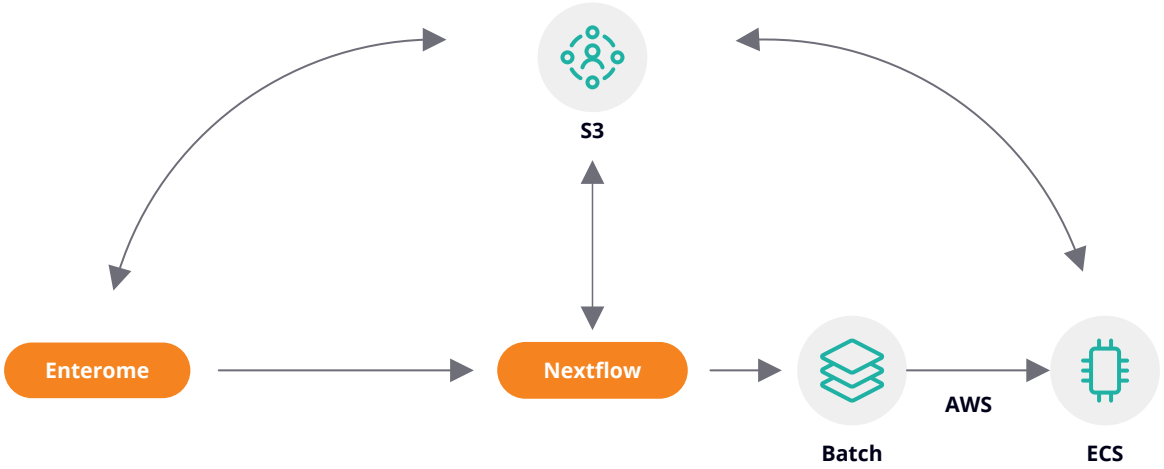
"Remarkably, our streamlined analytics processes have substantially reduced timeframes while considering cost-efficiency. Our expert teams have guided numerous projects with a diverse range of clients to realize the benefits of containerized bioinformatics workflow development. This is made possible by Nextflow's interoperability and the availability of shared components through the nf-core community."

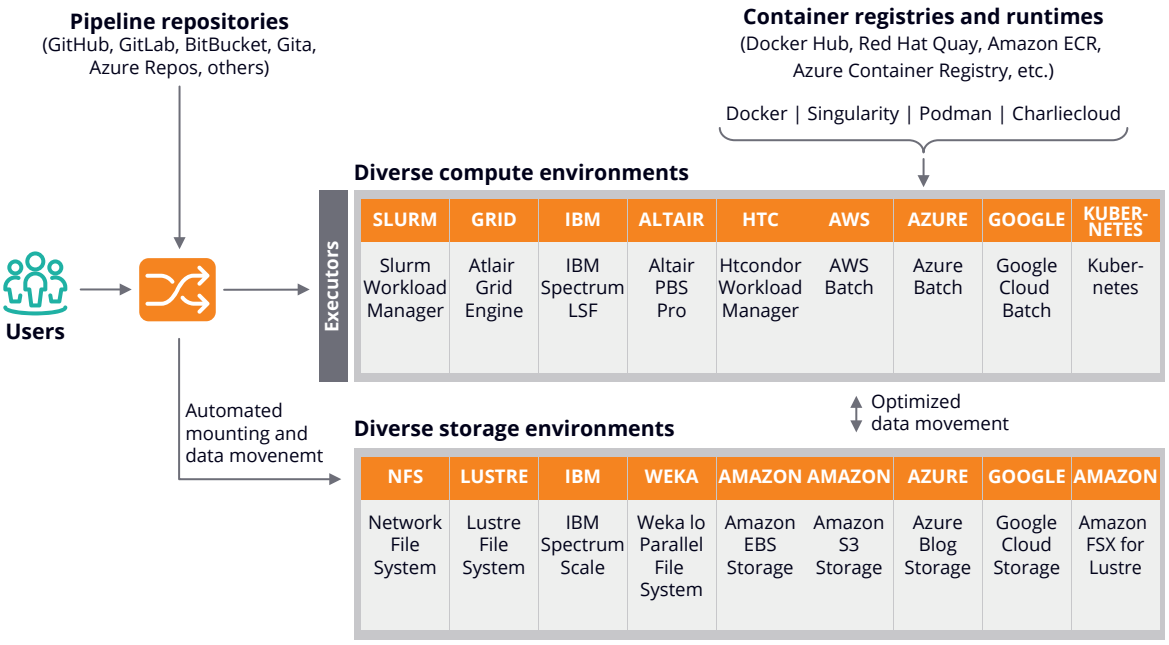
— Juan Sendoya, ZS senior lead bioinformatician

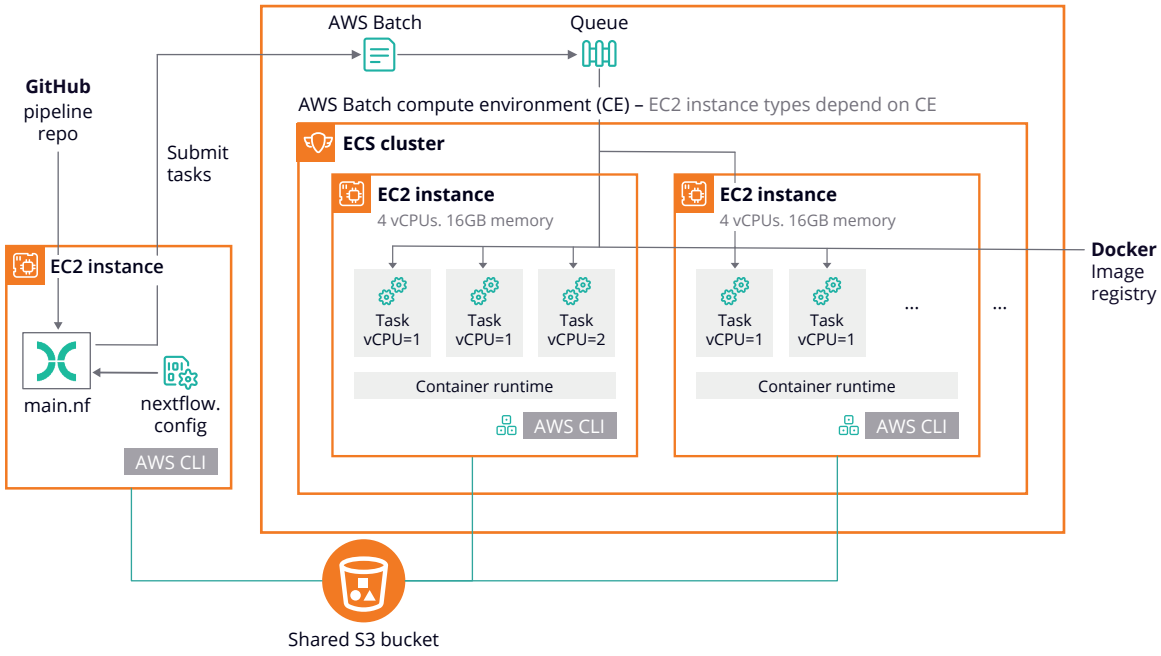
Figure 2 illustrates the behind-the-scenes operations of a Nextflow pipeline execution, highlighting Nextflow's ability to simplify complex bioinformatics workflows for the user.

FIGURE 2:

Nextflow pipeline example schematics







Despite the intricate background processes, Nextflow abstracts the complexity, providing the user with a straightforward interpretation of the pipeline execution. Nextflow automatically manages the interactions between data requests, data placement across databases or processes and the client's computing infrastructure, typically leveraging technologies such as Batch and S3.

Nextflow isn't limited to AWS services but can integrate with a wide array of computing infrastructures and schedulers, offering the same seamless automated interactions. This includes support for diverse container technologies, enabling users to deploy their pipelines across various computing environments seamlessly.

The final subfigure above illustrates the complex connections that occur in the background. Nextflow, based on the processes initiated by the pipeline, can dynamically select the appropriate resources required by the scheduler (in this case, Batch). The scheduler then interacts with the computing environment to launch the job, with all the necessary resources requested by Nextflow. This allows the user to focus on the pipeline execution without the need to manage the underlying infrastructure details.

Flexible scripting and operation, combined with the wealth of shared workflows and containers available on GitHub and nf-core—a community initiative for developing curated open-source Nextflow pipelines and modules—provide a vast toolbox that can shorten the time required to build and deploy a pipeline. Nextflow encourages adopting existing modules, reusing scripts for rapidly building and prototyping new modules and sharing workflow innovations.

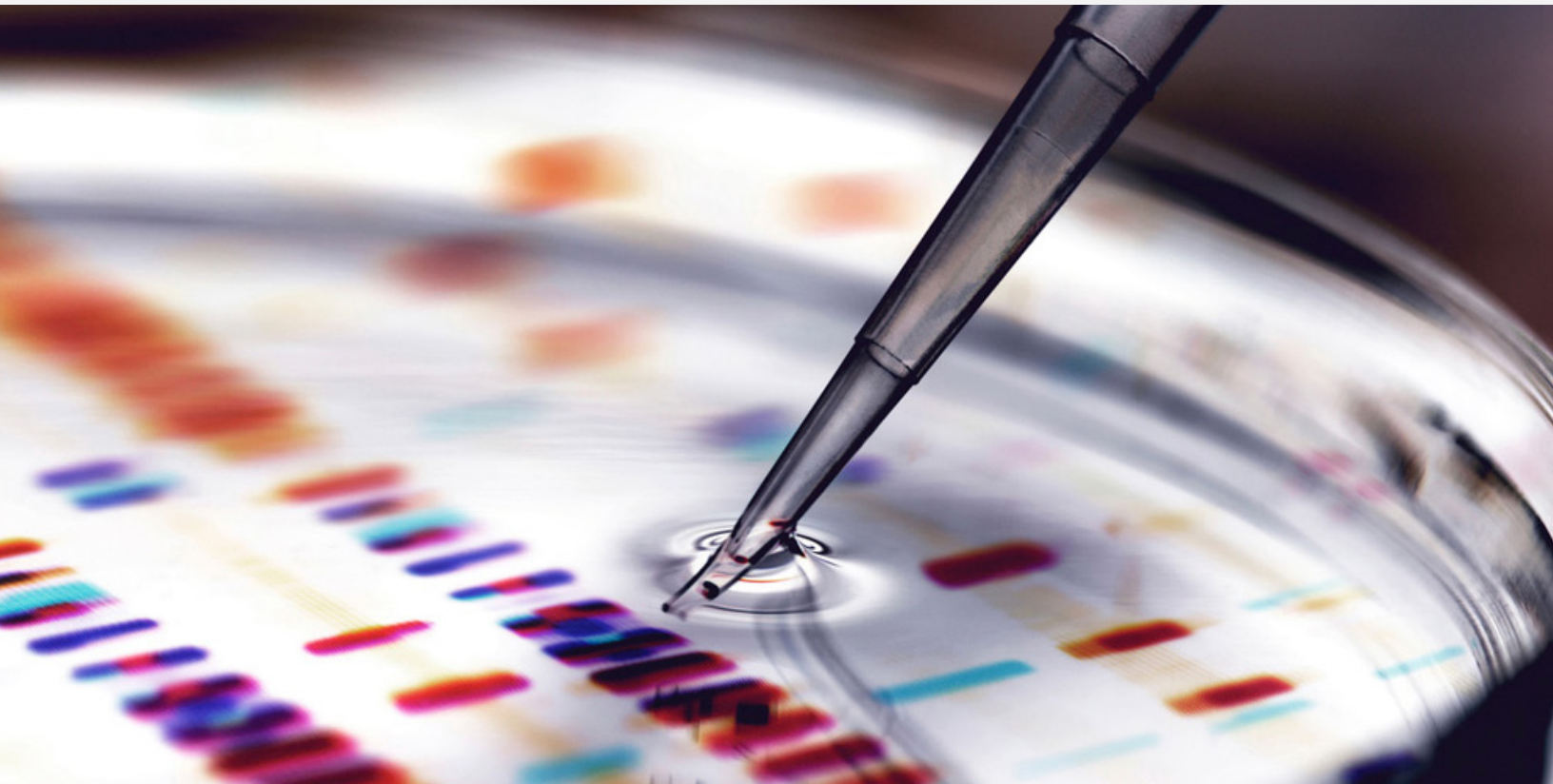
Built on sharing, a robust community reinforces this strength: over 1,000 open-source Nextflow pipelines were developed from scratch from 2018 to 2023, and the number of open-source modules and subworkflows doubled from 2021 to 2023. Shared containers from nf-core were reused in significantly more bioinformatics workflows than containers originating outside nf-core.

How ZS Discovery helps clients drive success in bioinformatics

ZS Discovery provides in-depth scientific knowledge to enhance analytical capabilities and delivers scientific rigor and quality control in every project, delivering robust results. Our team of world-class scientists bring years of experience tackling complicated scientific questions across the R&D pipeline. This perspective, combined with deep technical expertise, makes ZS a trusted partner in end-to-end scientific pipeline development.

As part of ZS's R&D excellence group, ZS Discovery works closely with colleagues with expertise in data science, generative AI, real-world evidence, medical devices and clinical development, among many other groups. This cross-team collaboration ensures that pipeline-generated results have the greatest possible impact regardless of their position in the R&D life cycle.

We believe that flexibility and portability will drive success in bioinformatics discovery, where large and disparate data sets are becoming more crucial in health and biomedical R&D. ZS Discovery's approach eases existing bottlenecks in workflow development and saves time in analytical processing. It also encourages sharing that will drive progress in bioinformatics by aligning with the FAIR guiding principles of findability, accessibility, interoperability and reusability.



About the authors

Bruce Press is a technology and life sciences leader with 30-plus years of experience. He's helped organizations build and develop commercialization strategies as well as supported early-stage research and development data landscapes and multiomics analysis frameworks. As a member of the ZS Discovery leadership team, Bruce is focused on product development and bringing his industry expertise to drive meaningful decisions.

Brice Sarver has a background in nonmodel and model system genomics, bioinformatics and evolutionary biology. As a director in ZS Discovery, he leverages his experience in the biotech and pharmaceutical spaces to help clients answer scientific questions and derive impactful insights while considering appropriate platform and technology solutions.

Federico De Masi has 25 years of experience both in the lab and in bioinformatics. His biological expertise is mainly in gene regulatory networks, transcription factors and protein-DNA interactions. Fred is involved in leading ZS Discovery activities for a few major clients and coordinates various systems biology, bioinformatics, data infrastructure and data management activities within these client spaces.

Maria Cecilia Argibay is a senior research scientist at ZS and member of the ZS Discovery team. She supports clients' scientific needs by providing biological, data management and programming insights; specifically, she works in the design and implementation of different automated biological analysis in Nextflow.

Juan Sendoya brings over a decade of expertise in translational research, particularly in omics-based technologies and human health. At ZS, he serves as a senior lead bioinformatician within the Medical and Scientific Expertise Center, where he oversees the development of robust bioinformatic pipelines using Nextflow. These pipelines are pivotal in analyzing intricate biomedical data, advancing precision medicine and uncovering profound insights.

Francisco Avila Cobos is currently working on expanding the single-cell capability offering within ZS Discovery while undertaking various roles as a subject matter expert and project manager with different teams across the globe. Francisco also coordinates the scientific thought leadership initiative to disseminate scientific contributions of ZS and its ZSers in different R&D areas.

Milena Vujović is an expert in T-cell receptor repertoire analysis and immune-informatics with a diverse background and experience that transverses both wet and dry lab. She's been managing client projects as project manager and project coordinator and is involved in business development initiatives in the immune-informatics and immune-oncology space.

Felipe Almeida is an expert in the development of computational pipelines with Nextflow using containers. Felipe is working on sharing knowledge and giving support to the teams in at ZS to strengthen the internal capabilities in workflow development and management. Felipe also coordinates work in the implementation of a centralized Nextflow pipeline framework for the company to make work more efficient, reproducible and portable.

Nicolas A. Schcolnicov, an agrobiotechnology engineer, excels in bioinformatics, with a particular focus on workflow management, notably in Nextflow. At ZS, Nicolas serves as a decision analytics associate consultant within the ZS Discovery team, where he works as a bioinformatics engineer, developing and supporting Nextflow pipelines.

About ZS

ZS is a management consulting and technology firm focused on transforming global healthcare and beyond. We leverage our leading-edge analytics, plus the power of data, science and products, to help our clients make more intelligent decisions, deliver innovative solutions and improve outcomes for all. Founded in 1983, ZS has more than 13,000 employees in 35 offices worldwide.

Learn more: www.zs.com/discovery