

Accelerate Your Genomic Research with Portable and Reproducible Workflows



We have reached a tipping point in bioinformatics. The resources needed for a major revolution are in place, including high speed, next-generation sequencers and standards-based tools and workflows that enable fast, accurate analysis of genomic data. At the same time, an explosion in computing capability is transforming the nature of what's possible and what's affordable in genomics, opening the door to new research models and high-volume clinical applications.

Just as important as these developments is the emergence of the cloud as a simple, secure, and cost-effective way to store, analyze, and share genomic data. By providing tools, storage, and compute power on demand, cloud-based genomic solutions simplify usage models and transform cost models, so researchers and clinicians can work and collaborate more efficiently.

With these innovations, small organizations can now access world-class genomic resources and supercomputing-class processing capability without the cost and complexity of onsite deployments.

Conducting Bioinformatics in the Cloud

Seven Bridges and Intel offer a complete solution for implementing cloud-based bioinformatics. The Seven Bridges Platform provides a central hub for teams to store, analyze, and jointly interpret their bioinformatics data. This platform provides a growing library of more than 200 tools and workflows for bioinformatics, including advanced, interactive visualization tools that are built for collaborative investigations among distributed research teams. Each tool and workflow is described using advanced methods for simple reproducibility and replicable analyses. To optimize processing on the Seven Bridges Platform, the largest datasets are co-located with the analysis workflows.

The Seven Bridges Platform is available from today's most popular cloud providers, including Amazon Web Services (AWS) and Google Cloud Platform (GCP). Seven Bridges technologies are also capable of on-premise deployment and for hybrid solutions that allow organizations to efficiently orchestrate the use of on-site and cloud resources. All these solutions are designed to provide strong data protection and to facilitate compliance with common bioinformatics regulatory requirements, including HIPAA, dbGaP, EU data protection, and CLIA and CAP standards.

The Seven Bridges software environment is consistent across all usage models, whether graphical or programmatic, and the underlying complex infrastructure is made accessible to researchers. The high performance and ubiquity of Intel® architecture adds to these advantages. Intel works extensively with commercial and open source developers to optimize the most popular genomics applications

for high performance on Intel architecture. Intel is also developing new technologies and solutions for High Performance Computing (HPC) that help to break down the barriers to fast, affordable analysis of massive datasets (see the sidebar, Breakthrough Computing Performance for Faster Analysis). As a result, organizations can count on fast, scalable biomedical analysis on site and across the widest range of cloud environments.

Achieving Higher Value with Portable Pipelines

It is one thing to have tools and workflows available for analyzing data on-premise and in the cloud. It is another thing to ensure that an existing workflow will perform efficiently and accurately when moved from one computing environment to another.

Bioinformatics workflows are typically built using Linux* command line tools that are combined to carry out complex analyses. Each tool does one job and has its own unique parameters, settings, inputs, and outputs. Most tools come in multiple versions with different dependencies, which typically evolve over time, creating interoperability issues with other tools and systems.

A complex workflow can contain dozens of individual tools, and most workflows must be re-built from scratch to run them any place other than where they were originally developed. This means developers must rebuild workflows even if they just want to deploy them from a laptop to the cloud. In most cases, this work is difficult and time-consuming, and a single error in reproducing a command line execution can potentially produce very different results.

To solve the challenges of workflow portability, the Seven Bridges Platform uses the Common Workflow Language (CWL), a language for developing complete, replicable, reproducible, and fully-documented workflows. Developers can use CWL to easily codify a complete workflow and all its settings. Seven Bridges also integrates file metadata with analysis, which enables researchers to use metadata as a component or variable during computation, allowing them to quickly scale an analysis across an entire cohort based on their attributes.

The Seven Bridges Platform also supports Docker, so an individual tool or groups of tools can be packaged as an independent software environment that runs in its own, lightweight virtual container. Unlike a traditional virtual machine, a virtual container can be spun up in a fraction of a second and imposes little or no performance penalty. All necessary software dependencies can be included within each container, which makes it relatively easy to resolve software compatibility issues.

With this approach, individual tools, and whole workflows, can be moved easily from one computing environment to another. No porting or other software modification is required, and there is no need to duplicate a complex computing environment. Developers can be confident their workflows will run quickly and efficiently and provide accurate results on any Intel architecture-based system.

Breakthrough Computing Performance for Faster Analysis

The Intel® Scalable Systems Framework for HPC is delivering platform-wide innovations in High Performance Computing. In combination with critical new software advances, such as the Seven Bridges Platform, these innovations can help organizations achieve new levels of performance with better cost models, whether they are running their workflows on local systems or in public clouds. Given the massive amount of raw data generated per genome, and the complexity of analysis, these innovations will be instrumental in powering the next wave of bioinformatics innovation.

- For more information, visit www.intel.com/content/www/us/en/high-performance-computing/product-solutions.html

Automating Pipeline Execution

Workflows written in CWL can be run using Rabix, an open source executor initially developed by Seven Bridges. The Rabix executor deciphers CWL applications and maps out each job that needs to be completed. Rabix can also be used to distribute individual jobs to multiple Intel architecture-based systems, whether those systems are in the cloud, local, or both. In this way, each CWL application can be run across the best available computational resources. All of this is accomplished in a scalable way that supports reproducibility. Seven Bridges is using this technology today to implement distributed computing for large-scale genomics initiatives, such as the Million Veteran Program (MVP).

A Flexible Platform for Genomic Innovation

Developing and running CWL-based workflows on the Seven Bridges Platform offers major advantages for laboratories and research organizations. The hundreds to thousands of parameters and settings in a typical workflow can be reconstructed automatically from the simple text of a CWL file, so complex experiments can be repeated easily, exactly, and virtually anywhere.

As a result, organizations can:

- **Run every workload on the best-fit computing resource** to achieve an optimal balance of speed and cost. If on-site HPC resources are running at capacity, for example, all or part of an analysis can be moved immediately into the cloud.
- **Share and repeat experiments among distributed teams and organizations**, while controlling access, use, and data handling based on legal and regulatory requirements. There is little if any effort or delay in replicating an experiment, and experiments can be duplicated exactly and run almost anywhere.

- **Create workflows** on workstations, or other small computing systems, then easily amplify them to analyze larger datasets and to run on larger systems or clusters. With this approach, initial development and testing takes place in a simple environment, and more costly computing resources are reserved for more demanding and time-sensitive workloads.

Conclusion

Portable and reproducible workflows are transforming bioinformatics research. Portability eliminates much of the cost and delay associated with recreating complex workflows on alternative systems and clouds. The Seven Bridges Platform running on Intel architecture provides an ideal foundation that allows organizations to run their analyses efficiently and almost anywhere. It can be deployed on site or accessed on demand through AWS or Google Cloud.

The Seven Bridges software environment provides a unified experience for researchers, while the use of CWL simplifies the development of portable and reproducible workflows. The high performance and ubiquity of Intel architecture adds to these advantages, helping to ensure fast, efficient processing across the widest range of infrastructure options.

Learn More

Get more information about CWL, Rabix, the Seven Bridges Platform and Intel architecture-based solutions from the links below. Then join the community to help fuel the next wave of bioinformatics innovation.

From Seven Bridges

- Contact Seven Bridges at team@sevenbridges.com
- [The Seven Bridges Platform](#)
- CommonWL.org
- [Seven Bridges documentation on CWL](#)
- [The Cancer Genomics Cloud for the NCI](#)

From Intel

- [Optimized Genomics on Intel Architecture](#)
- [Intel in High Performance Computing](#)

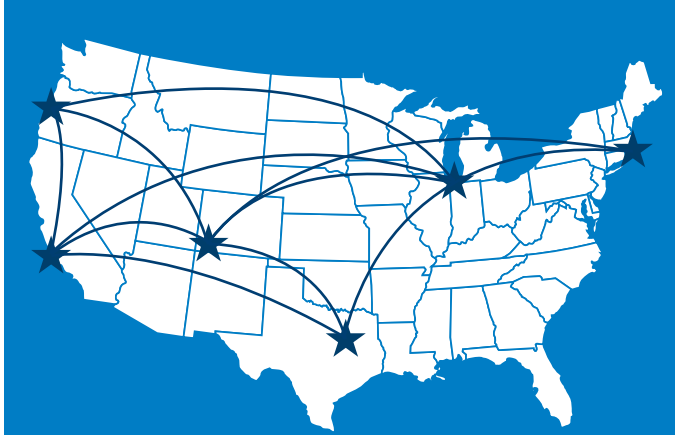


No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit <http://www.intel.com/performance>.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at <http://www.intel.com/content/www/us/en/benchmarks/intel-product-performance.html>.

Copyright © 2017 Intel Corporation. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.



Million Veteran Program and Seven Bridges

Seven Bridges is privileged to participate in the Million Veteran Program with the U.S. Department of Veterans Affairs (VA). As part of this mission, we will be using Rabix to enable reproducible analysis of biomedical data to benefit patient outcomes. By deploying our technology on the VA infrastructure, we allow researchers to execute workflows in exactly the same way across any environment.