CASE STUDY

Standardizing Portable and Reproducible Genomics Data Analysis Pipelines

THE ELI LILLY EXPERIENCE

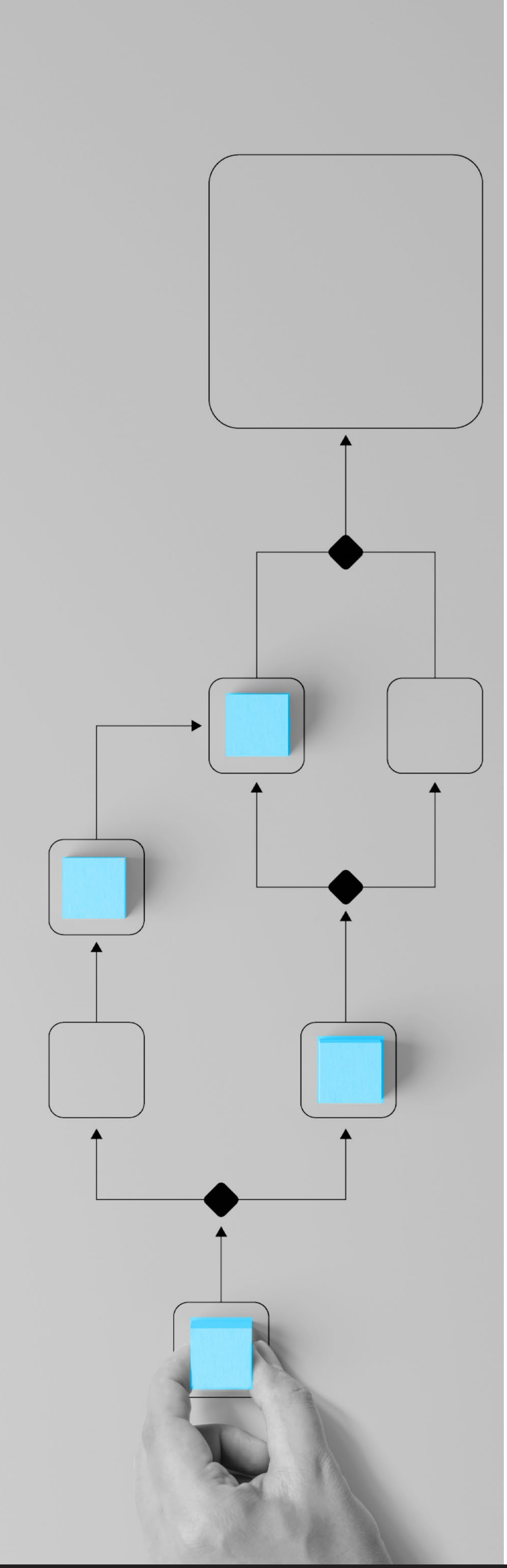
DNAnexus®

Contents

It was Heraclitus who said "The only constant is change." But it is the savvy bioinformaticians at Eli Lilly who say "bring it on!"

New tools and technologies, new algorithms, and new questions posed by scientists mean that there's a constant rhythm of change. For bioinformaticians in pharma, dealing with change is part of the job. But what happens when you compound change with a growing amount of data and the need to be able to run pipelines across global regions, all of which have different infrastructure?

Just ask bioinformatician Michael Neylon, and the group he works with in the Research Data Science and Engineering group at Eli Lilly. He and his team addressed these challenges by partnering with DNAnexus and using open-source standards to tackle the changing bioinformatics landscape with on-demand compute and portable, reproducible pipelines to run in any region.



Multiple Regions, Growing Data, and Custom Requests Create a Complex Environment

Eli Lilly has multiple therapeutic areas that span global regions. Historically, bioinformaticians developed pipelines in support of these areas on an ad hoc basis using the infrastructure that was available on local premises. The pipelines were designed soundly but didn't always translate across infrastructures located in different regions.

Additionally, the group received frequent requests for custom development in their pipelines. They had the expertise to handle custom requests but lacked a systematic way to recover from failures, which is common when developing in a custom environment. The fixing of errors, if any, often necessitated re-rerunning the entire workflow. It wouldn't be unusual for implementation of a small modification to take 1-2 weeks!

Finally, Eli Lilly's R&D teams were challenged by the sheer amount of data they had to manage. Already consuming petabytes of space, the data being generated continued to grow. Since the team at Eli Lilly weren't planning to expand their compute and storage capacity on-premise, they were interested in solutions that would enable them to take advantage of on-demand cloud computing.

Establishing a Solid Foundation for Change, Data Growth, and Portability

The data growth, demand for multi-region functionality, and custom development requests necessitated an IT-supported and methodical approach to their bioinformatics pipelines, which is why they partnered with DNAnexus and launched an initiative to redesign several of their legacy pipelines.

The first workflow the group focused on was their whole exome pipeline. This pipeline was a logical first choice because scientists at Eli Lilly wanted to update it with new variant callers and other bioinformatic tools. Additionally, it was developed on an older infrastructure and had what could be considered "legacy artifacts." There were hard-coded references to data and resource requirements, and in general, the workflow was entangled with the infrastructure on which it ran — so much so that it was difficult for it to be run elsewhere.

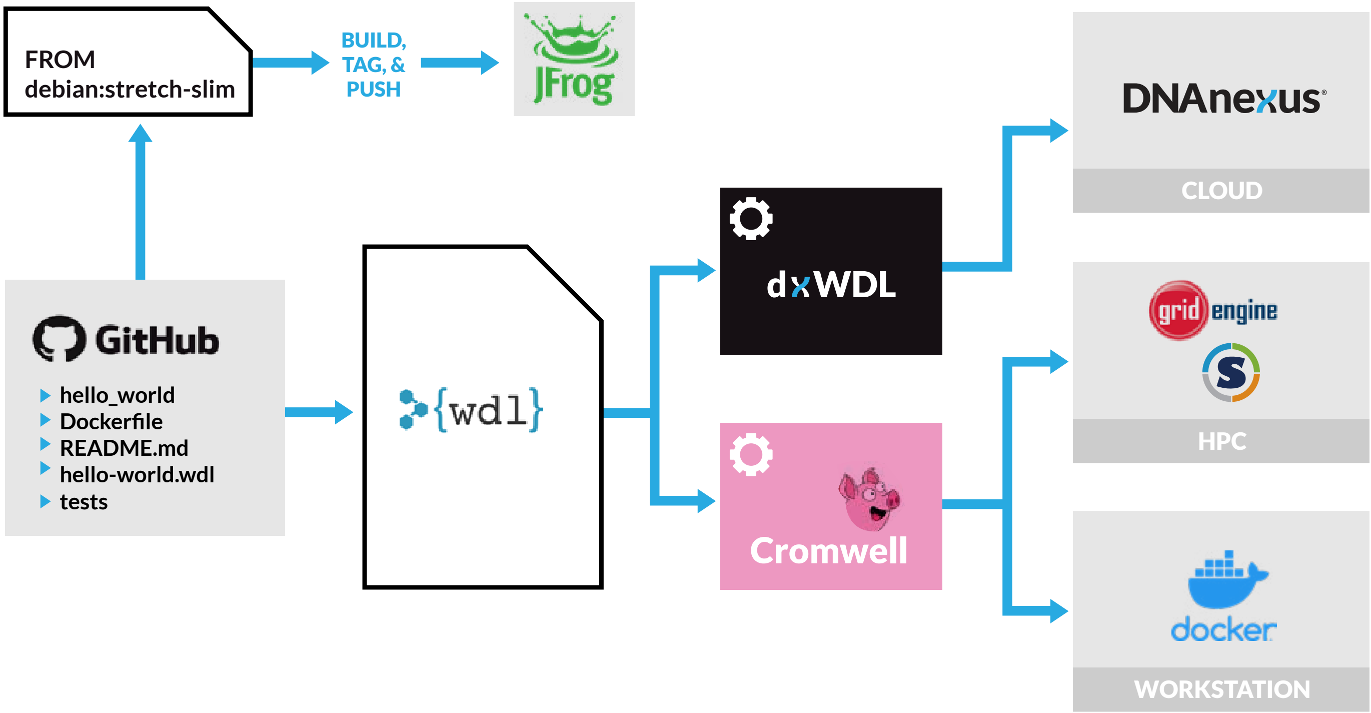
Before development began, Eli Lilly codified a primary requirement: Deploying in the cloud was important for accessing on-demand compute resources, but the team also needed to run on individual workstations and on HPC clusters, and ideally, be able to maintain the pipeline externally.

Building their own custom workflow language would have addressed that requirement, but the company was strongly opposed to the idea of maintaining proprietary tools. Instead, they decided to use open standards – Workflow Design Language (WDL), Docker, and Singularity – and they partnered with DNAnexus because this combination enabled them to meet all of their requirements.

Taking an *Open*-Minded Approach

The decision to go with open standards wasn't immediate. Eli Lilly worked with the DNAnexus xVantage Professional Services Team to fully understand how the open-source software would work on the DNAnexus Platform. Since the team at Eli Lilly hadn't seen many examples of portable pipelines, the exercise required them to build out certain components of the pipeline, deploy them in different environments, and validate that they would work. The team arrived at the following high-level architecture.

High-level Architecture



- To make the pipelines reproducible and remove infrastructure dependencies, Eli Lilly uses Docker to package software tools for individual workflow steps. Docker is a containerization and virtualization technology that enables tools to be bundled with all of their dependencies (including the operating system) so that they can be run agnostically with regards to the host system.
- To manage workflows in a constant and readable way, Eli Lilly adopts WDL to define the logic steps of workflows. WDL tasks define their inputs, outputs, resource requirements, and processing steps (written as unix shell commands), while WDL workflows define the dependencies between tasks and enable parallelization of independent tasks.
- To run on DNAnexus, Eli Lilly uses dxWDL to translate WDL tasks and workflows into native DNAnexus apps and workflows. dxWDL is a command-line tool written in Java and based on the same code used by the mature Cromwell workflow engine that is developed by the Broad Institute.
- To run on their HPC cluster, they use Cromwell to orchestrate job scheduling, and they convert their Docker images to Singularity to satisfy their internal security requirements (since Docker requires root access to the host system while Singularity does not). Running on the HPC cluster requires using a Cromwell configuration file that describes how to convert resource requests and tasks into Grid Engine submission scripts.
- To run on individual workstations, Lilly uses Cromwell and Docker.

Addressing Minor Gaps to Create a Test-driven Environment

Because there are slight differences in how Cromwell and dxWDL process WDL tasks and workflows, Lilly wanted to create test cases to ensure that their pipelines would run identically in both environments. When they first started testing, they would write the tasks and copy the command block into a test script that would then be executed against the Docker container. That process gave them some coverage, but what they really needed to do was test the entire execution of the pipeline with a WDL execution engine. To address this gap, they collaborated with DNAnexus to create pytest-wdl, a plugin for the popular Python Pytest framework.

The pytest-wdl plugin enables Lilly to execute WDL tasks with defined inputs and to assert that the outputs generated by their workflows match the expected outputs. Eli Lilly open-sourced the plug-in, located here: https://github.com/EliLillyCo/pytest-wdl

Embracing the Changes

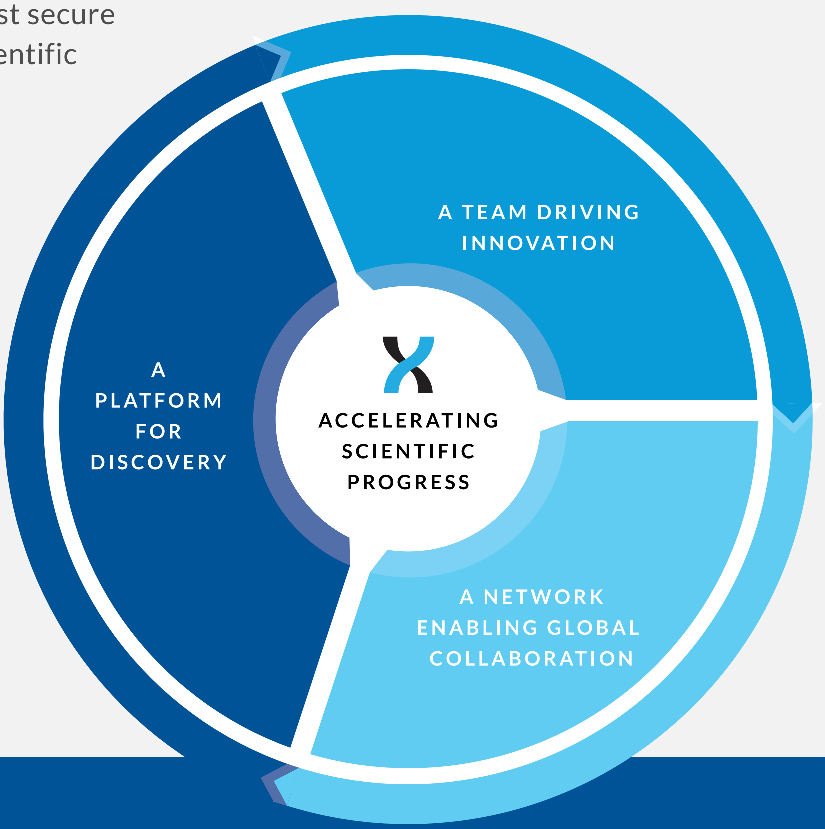
After moving to the DNAnexus Platform and employing open standards, the group at Eli Lilly realized immediate benefits. **They can now:**

- Port and deploy the pipelines in any region and they will successfully complete. Furthermore, they have been able to remove all hard-coded references and entanglement with infrastructure; inputs and outputs for tasks and workflows are all parameterized.
- Access on-demand cloud compute in the DNAnexus Platform to experience a more consistent and reliable computing experience. According to Neylon, the pipelines that access on-demand compute using the DNAnexus Platform run twice as fast.
- Accommodate changes and have confidence in the results because of the test-driven environment.

About DNAnexus®

DNAnexus® has built the world's most secure cloud platform and global network for scientific collaboration and accelerated discovery. We embrace challenges and partnership to tackle the world's most exciting opportunities in human health

We are scientists, engineers, cloud experts, compliance specialists, and thought leaders dedicated to the acceleration of scientific discovery. The work we do enables the world's most important breakthroughs in health science, the development of life-saving cures, and access to critical data needed for new discovery.



A TEAM DRIVING INNOVATION

A PLATFORM FOR DISCOVERY

ACCELERATING SCIENTIFIC PROGRESS

A NETWORK ENABLING GLOBAL COLLABORATION

Dedicated to Your Success.

**Learn how DNAnexus can help you accelerate your diagnostics programs.
Contact us for a brief scientific consultation:**

info@dnanexus.com

**For more information about DNAnexus solutions, visit
www.dnanexus.com**