code ocean

CASE STUDY

How the Seattle Hub for Synthetic Biology is Leveraging Code Ocean to Push the Boundaries of Synthetic Biology

Seattle Hub for SYNTHETIC BIOLOGY

Simmi Mourya[1], Jake Valsamis[2], Florence Chardon[3]

1.Software Engineer II, Computational Biology Team, Seattle Hub for Synthetic Biology; 2.Solutions Architect, Code Ocean; 3.Scientist II, Computational Biology Team, Seattle Hub for Synthetic Biology

Background

The Seattle Hub for Synthetic Biology aims to advance the DNA Typewriter [1] and ENGRAM [2] technologies to enable organism-scale in vivo biological recording. These technologies have the potential to usher in a new era of synthetic biology in which cells are engineered to record their own histories, enabling new insights into developmental biology and new types of diagnostics and therapeutics.

Progress on these fronts relies on wet lab experiments in cells and mouse models, and the primary data output from these experiments is both short- and long-read next-generation sequencing data. The Seattle Hub for Synthetic Biology is using Code Ocean as its primary computational platform to develop pipelines that automate cloud-based data transfer and perform preliminary bioinformatics analysis of generated sequencing data.

Key Focus Areas:

1. **Data Transfer Bottlenecks:** Sequencing data is saved to BaseSpace, a cloud platform for managing Illumina genomic data, and manually transferred to AWS S3 for downstream processing. This manual process is time-consuming, increases the risk of data loss, and poses security concerns. Automating data transfer between cloud platforms would enhance efficiency, security, and reliability.

2. **Manual, Redundant Analysis Steps:** The Computational Biology team manually performs pre-processing steps such as paired-end read merging, low-quality read filtering, and CRISPR edit analysis using custom scripts tailored to each experiment. While effective, this approach is resource-intensive and introduces redundant work for repetitive tasks, highlighting the need for a unified, automated workflow to reduce effort and improve turnaround time.

3. **Fragmented Codebase and Maintenance Overhead:** Custom scripts tailored to individual experiments, while flexible, have led to a fragmented and sprawling codebase. As with any growing codebase, this makes maintenance, updates, and standardization more difficult and time consuming.

1. Choi, J. et al. A time-resolved, multi-symbol molecular recorder via sequential genome editing. Nature 608, 98–107 (2022).

2. Chen, W. et al. Symbolic recording of signalling and cis-regulatory element activity to DNA. Nature 632, 1073–1081 (2024).

code ocean

The solution

Recognizing these challenges, the Computational Biology team adopted **Code Ocean** to standardize and automate NGS data processing, enabling consistent, reusable results across experiments. By deploying a **unified workflow** for preliminary analysis, the team reduces code redundancy and frees up computational scientists and engineers to focus on experiment-specific insights rather than rebuilding foundational steps each time.

Key features used

App Builder

Capsules

Data

Key integrations

Amazon S3

Amazon Event Bridge

Illumina Basespace CLI

AWS Lambda

AWS Cloudwatch

The results

Automatic Data Transfer

Using Code Ocean Capsules, along with BaseSpace and S3 connectors, the transfer of sequencing data from instruments to the AWS ecosystem is fully automated. This significantly reduces risk of data loss while **saving hours of manual work per experiment**.

Efficient Data Packaging for Analysis

After data transfer, the sequencing data is automatically packaged into a Code Ocean Data Asset, which can be used as input data in a Code Ocean Capsule. This packaged data serves as input for the next stage of the short-read analysis pipeline.

Seamless Preliminary Analysis

A second Code Ocean Capsule uses the App Panel to deliver a JSON based parameterized interface that allows computational scientists and engineers to conduct preliminary analyses of the sequencing data. This allows a reusable Python script to be adapted across sequencing runs, **saving 15 hours of maintenance per month**. The Capsule **generates reproducible and traceable results, which is essential for tracking experimental progress**.

Scalability for long read sequences

As the scope of the group's work expands to include larger datasets, the scale of data processing will grow from tens to hundreds of GBs. With Code Ocean's automated resource allocation, data transfer and analysis workflows will **scale seamlessly as the projects evolve**.

By integrating automation, scalability, and reproducibility, the Computational Biology team's pipeline offers rapid sequencing data analysis—a key step toward advancing in vivo biological recording.

code ocean

"Code Ocean is a great platform that keeps all of our computational workflows highly organized, reproducible, and traceable. With its intuitive and easy-to-use interface, it is an ideal platform for our team that includes both software engineers and computational biologists."

Florence Chardon, Ph.D., Computational Biology team lead and Scientist II at the Seattle Hub for Synthetic Biology

Background on DNA Typewriter: DNA Typewriter is a groundbreaking genome-editing technology that enables sequential genome editing along a synthetic array. Leveraging the precision of CRISPR, this innovative approach functions much like an old-fashioned typewriter. Each edited section of DNA activates the next target site, creating a continuous "DNA tape" that records cellular events with remarkable precision. This method provides a powerful tool for tracing cellular histories, having the potential to unlock profound insights into complex biological processes.

About The Seattle Hub for Synthetic Biology

The Seattle Hub for Synthetic Biology, established by the Allen Institute, Chan Zuckerberg Initiative, and the University of Washington, is at the forefront of transforming how we study cellular biology through the development of innovative genomic recording technologies. More information can be found at https://alleninstitute.org/division/seattle-hub-for-synthetic-biology/.

About Code Ocean

Code Ocean is a Computational Science platform for life science R&D teams who want a fast and efficient way to start, scale, collaborate, and reproduce computational research. It helps Computational Scientists set up and scale their workflows, work closer together, and lets them support non-coding scientists with accessible, intuitive applications. Built on FAIR data principles, it helps avoid technical debt, improves data architecture, and improves organizational compliance and quality.