Cirrus-NGS: Cloud-optimized compute infrastructure for next generation sequencing analysis

Guorong Xu[1], Mustafa Guler[1], Mengyi Liu[1], Roman Sasik[1], Amanda Birmingham[1], Kathleen M. Fisch[1]

[1]*Center for Computational Biology & Bioinformatics, Department of Medicine, University of California, San Diego, La Jolla, CA 92093. Email: g1xu@ucsd.edu*

Source Code: https://github.com/ucsd-ccbb/cirrus-ngs
License: MIT License

Bioinformatics analysis of large-scale next-generation sequencing (NGS) data requires significant compute resources. While cloud computing makes such processing power available on-demand, the administration of dynamic compute clusters is a daunting task for most working biologists. To address this pain point, we have developed cirrus-ngs, an open-source turn-key solution for common NGS analyses leveraging the open-source CfnCluster on Amazon Web Services (AWS) (Figure 1). Cirrus-ngs allows users to deploy and manage elastic HPC clusters and run NGS pipelines from within a web browser on their local machine using flexible but light-weight Jupyter notebooks. Cirrus-ngs users need not have any bioinformatics tools for these pipelines installed on their local machine, as the HPC clusters are created dynamically in AWS based on a custom EBS snapshot with all of the necessary software pre-installed. All computation is performed using AWS EC2 compute instances and all results are uploaded to AWS S3 storage. Cirrus-ngs currently supports primary analysis pipelines for RNA-Seq, miRNA-Seq, ChIP-Seq, and whole-genome/whole-exome sequencing data. These pipelines have been optimized for use on AWS HPC clusters, which dynamically scale depending on the compute resources required for the various steps in the pipelines to minimize the per sample compute cost. For example, cirrus-ngs can call variants from a human whole genome (30x coverage) in 2 days for a cost of ~$15 using AWS spot instances, all controlled through a Jupyter Notebook running on a local machine. Cirrus-ngs is a lightweight, reproducible tool to perform scalable NGS primary analyses on the cloud.

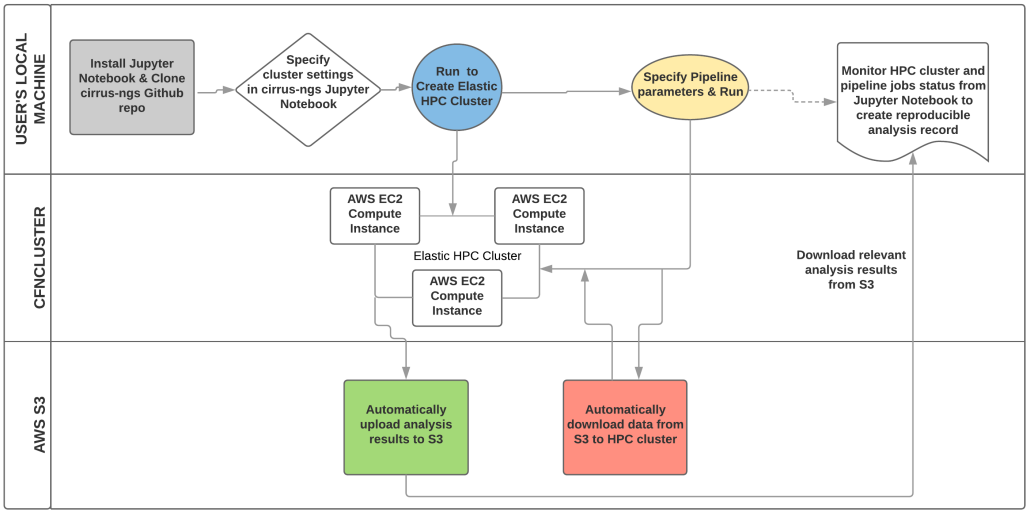

Figure 1. Schematic of cirrus-ngs workflow.