Research Update

Edmund Miller

2021-10-13 Wed

Internship with Element Biosciences

Element Biosciences

- Startup creating a new DNA sequencing platform
- Worked with the Bioinfomatics group
- Met Bryan Lajoie through nf-core

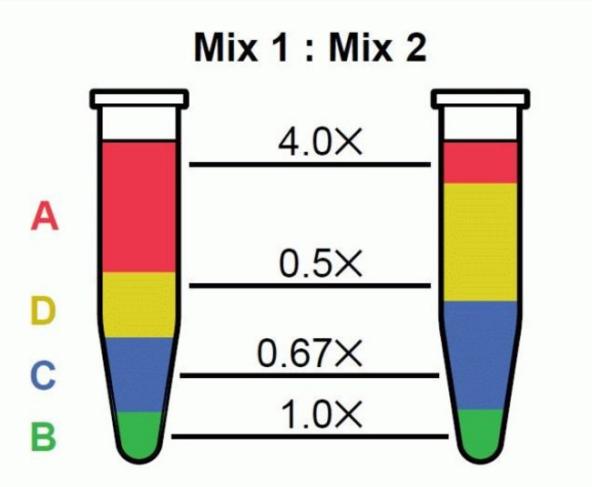
Overview

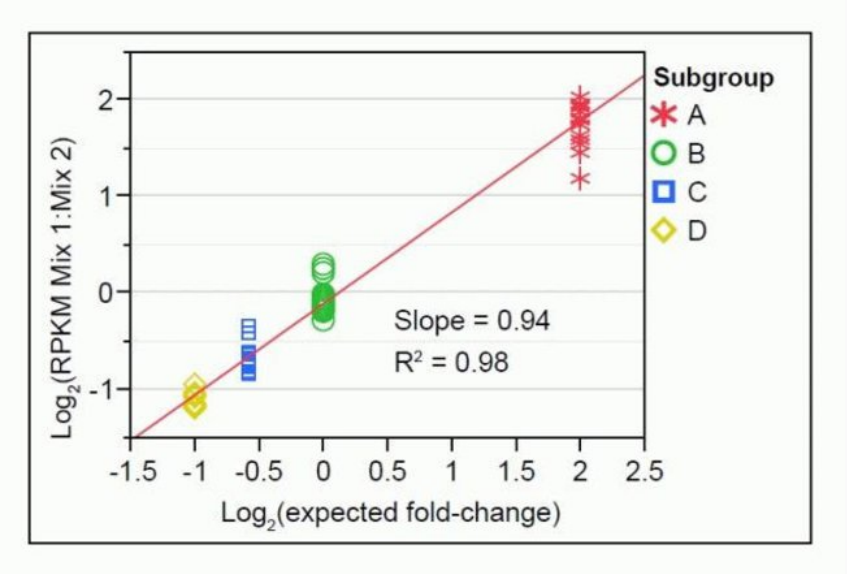
- ERCC Analysis
- COVID Assay analysis
- Secondary Analysis Infrastructure

ERCC Analysis - External RNA Controls Consortium

ERCC Analysis - External RNA Controls Consortium

- Evaluation of multiple performance characteristics
  - Linear performance of individual controls
  - Signal response within dynamic range pools of controls
  - Ratio detection between pairs of dynamic range pools.



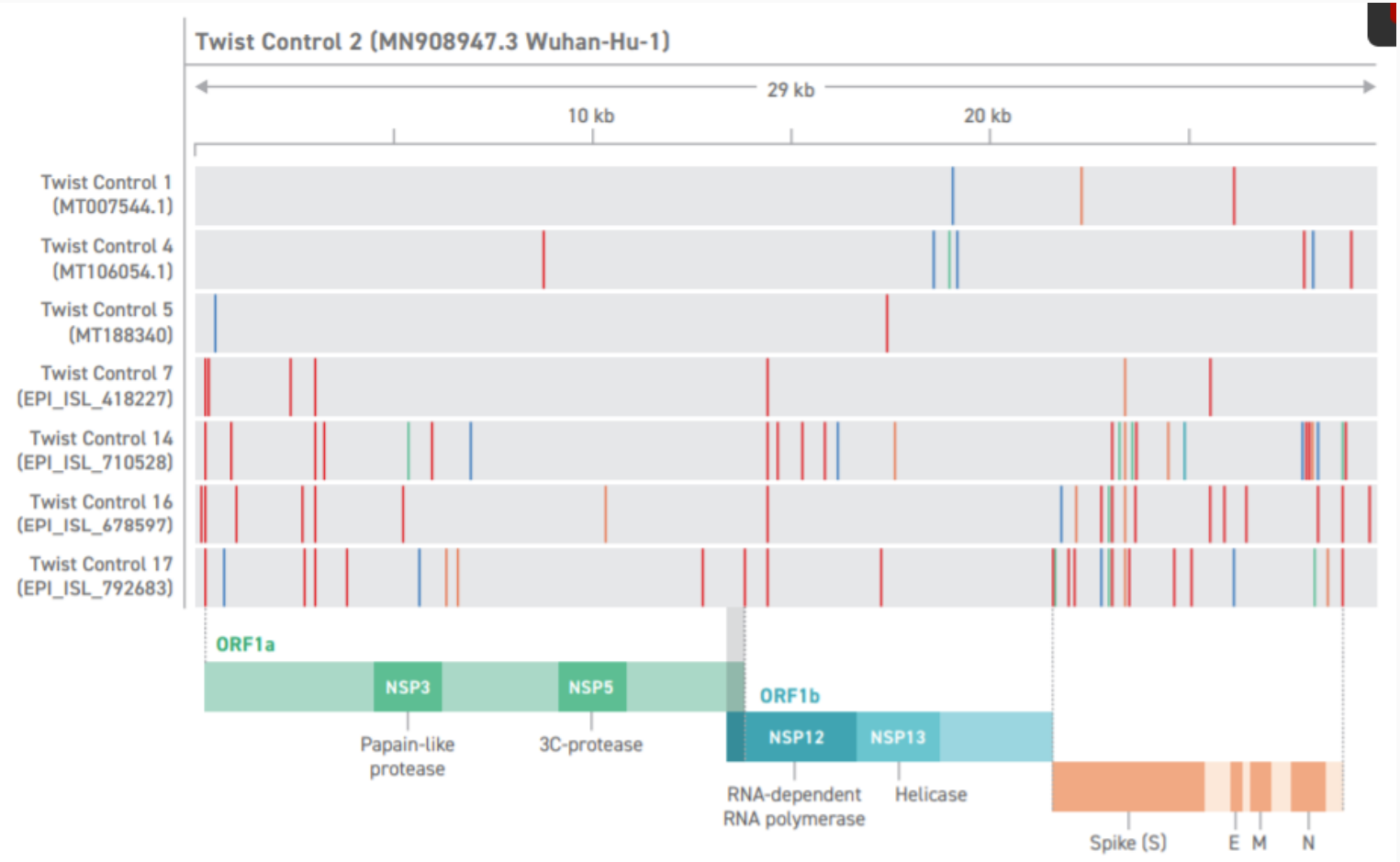


ERCC Analysis

- Allows for estimation of Lab to Lab (instrument to instrument) variation.
- Used erccdashboard to create a standardized analysis.

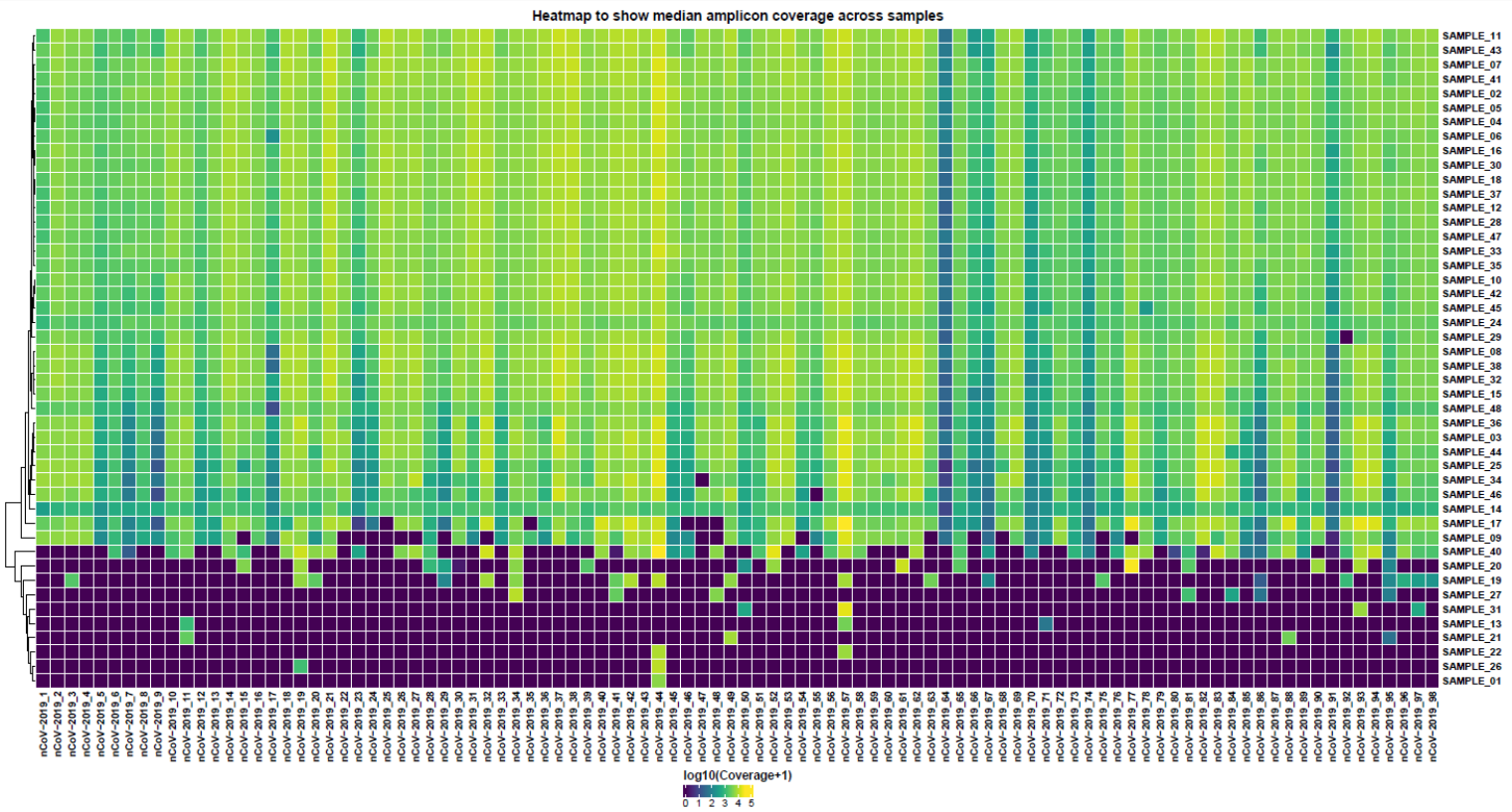
Amplicon Analysis for Covid assay

Amplicon Analysis for Covid assay

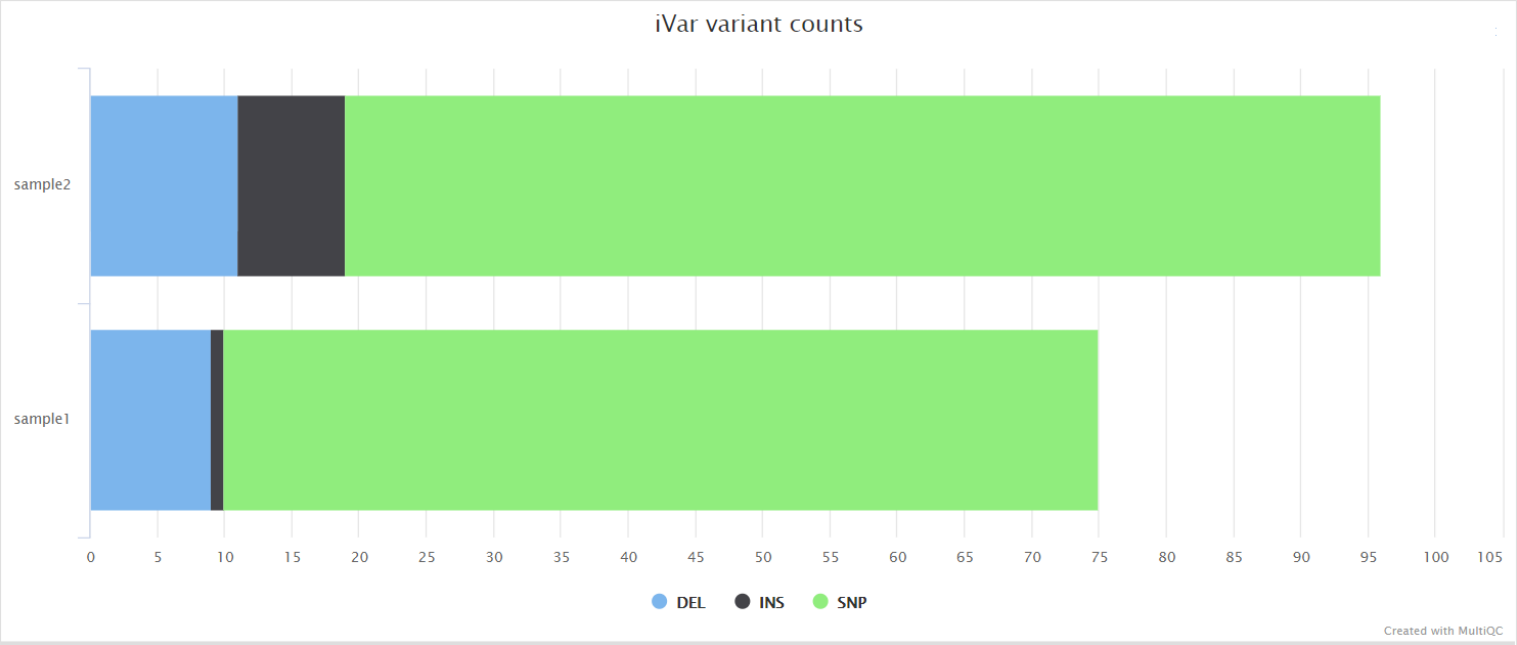


nf-core/viralrecon

nf-core/viralrecon



nf-core/viralrecon



Analysis of Covid Variants

Secondary Analysis Infrastructure

What is Secondary Analysis?

Primary Analysis - Specific steps needed to transform images into base-calls and compute quality scores for those bases

Secondary Analysis – Alignment of these short sequencing reads onto a reference genome and variant calling

Tertiary Analysis – Interpreting the secondary analysis data (annotation, qc metrics, filtering, benchmarking)

Types of Secondary Analysis

- WGS
  - Human, ecoli, phix, covid
- WES
  - Exome, panel, amplicon
- Single Cell
  - 10x scRNA-Seq, 10x spatial, 10x scATAC-Seq
- RNA-Seq
  - Bulk RNA
- MetaGenomics
  - Stool sample

Whole Genome Sequencing

- Allows us to compare with the "truth"
- Genome in a Bottle
    - Leveraging multiple instrument platforms to create truth datasets
    - Truth is available for HG001-HG005 with diverse genetic backgrounds
- Allowed us to provide feedback to the rest of the teams
    - Context Errors

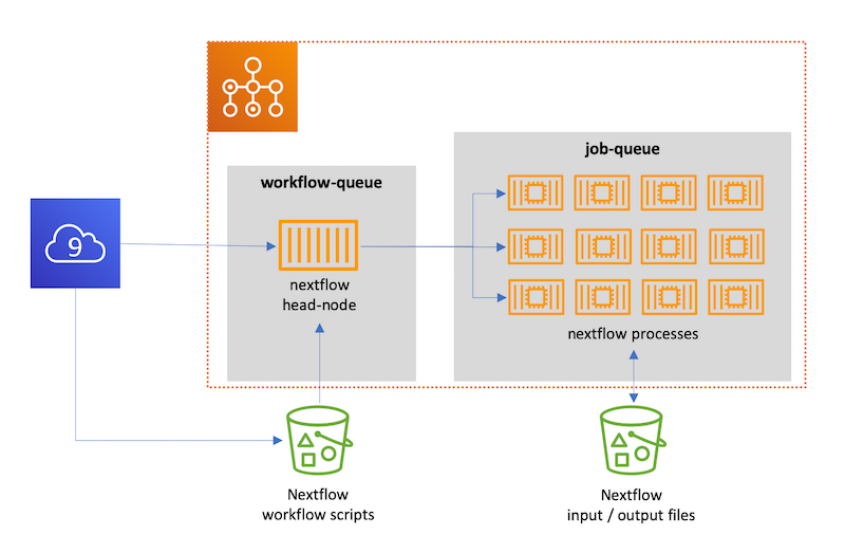
Goals of the Secondary Analysis Infrastructure

- Mimicking a Customer environment
- Internal Data discoverability
- Automation

Design Decisions

nextflow

nf-core



nextflow *tower*

Nextflow

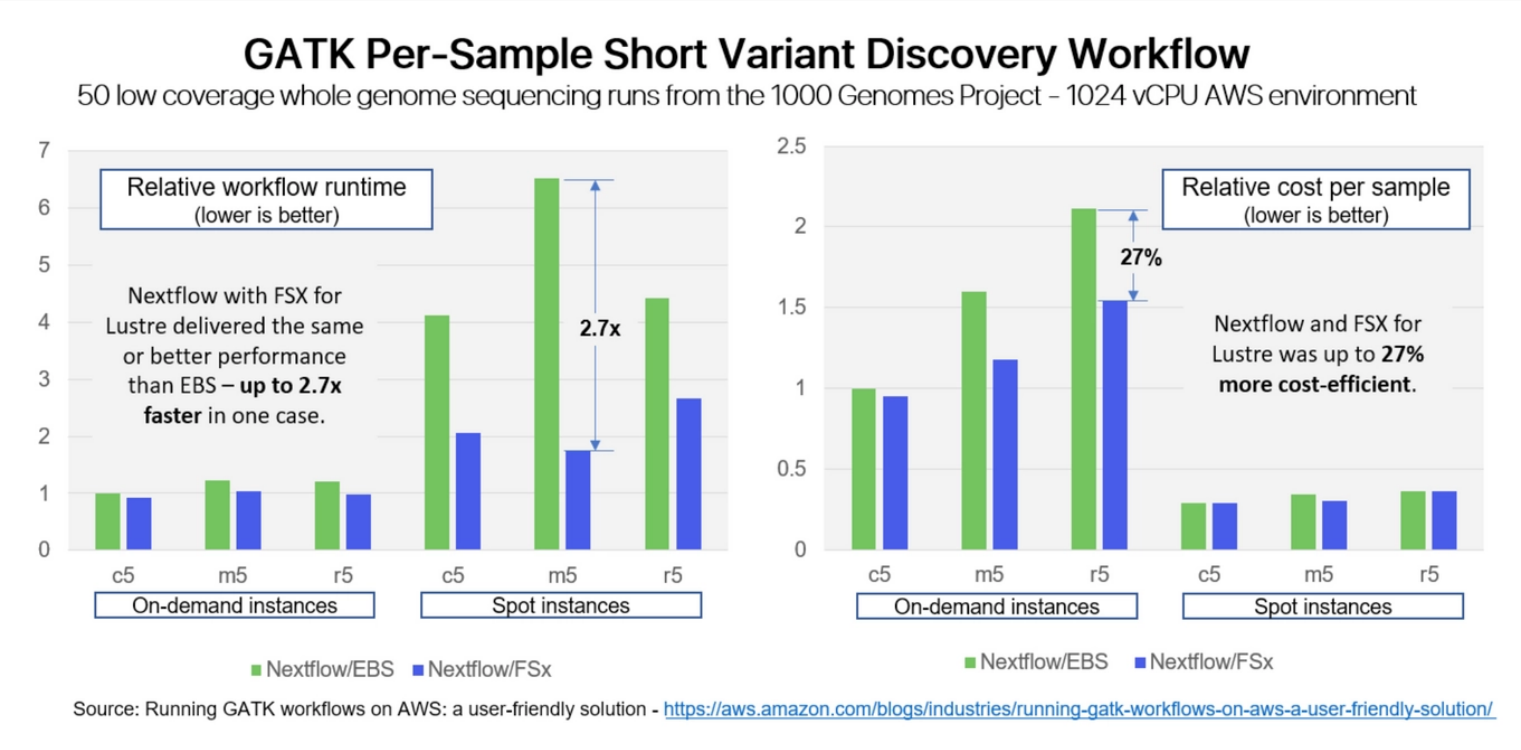
- Opensource
  - Supported by Seqera
- Platform independent
  - Runs locally, Cloud, SLURM, hybrid
- Reproducibility

nf-core

- Common bioinformatics software modules make creating new workflows quickly
- Curated set of best practice pipelines to avoid reinventing the wheel for secondary analysis.
- Template to quickly start new pipelines

AWS Batch

- Abstracts away the cluster provisioning
- Spot Instances
- Utilizing High Performance systems



Nextflow Tower

- Handling AWS batch environment
- Monitoring, logging & observability
- Automation
- Smoothing out Customer Experience

Things learned from this Internship

- Exposure to Cloud computing for bioinformatics
- Improved my tertiary analysis skills
- Exposure to primary analysis
- Exposure to a greater variety of assays
- Better understanding of job titles and roles that are out there
- Skills Seymon looks for when hiring(in order):
    - Ability to write production level code
    - Developing novel algorithms
    - Tertiary analysis skills

Notebook Template

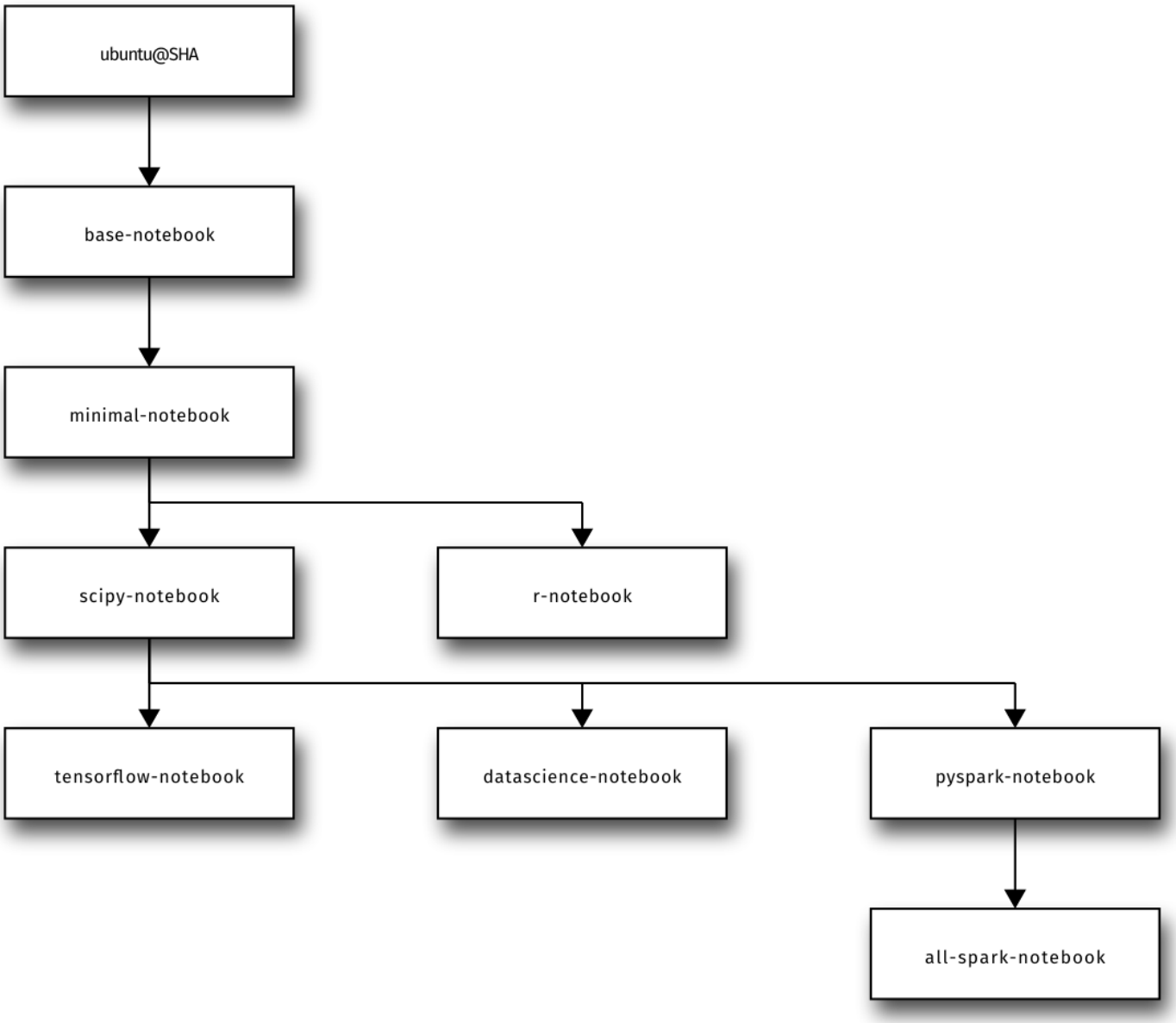
Notebook Template Goals

- Creating a separation between secondary and tertiary analysis
- Ingesting the expected results from secondary analysis
- Environment is easily reproducible but flexible for moving quickly
- Avoid being tied to one language
- Data science instead of data engineering

Getting started

1. Go to GitHub - Functional-Genomics-Lab/notebook-template
2. Click "Use this Template"
3. `docker-compose up`
4. Copy the link to your local jupyter instance from the terminal and open it in your browser.

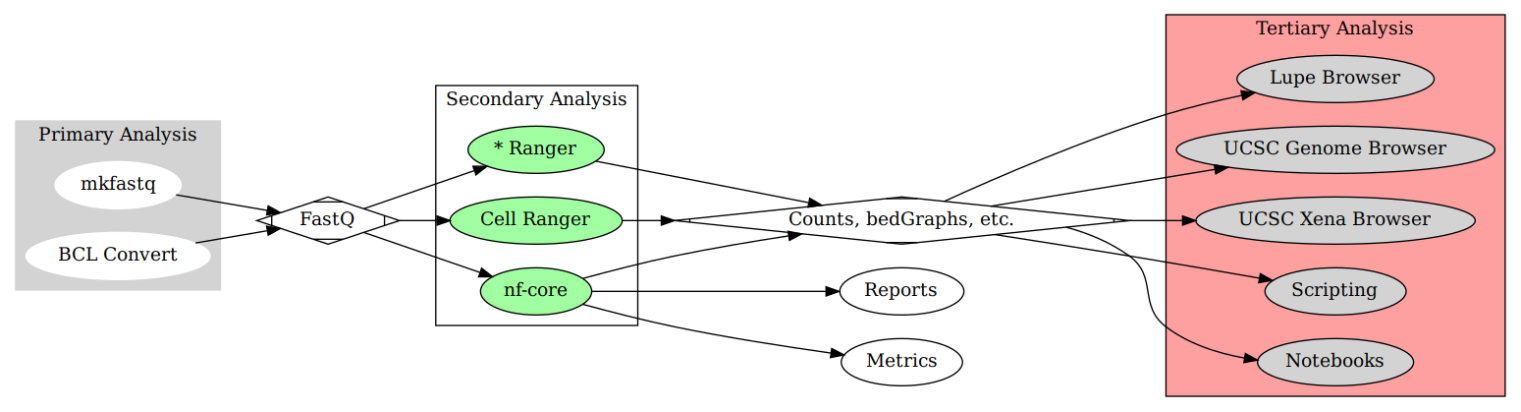
Selecting an Image



Quick Tour using GRO-Seq Analysis

- Dockerfile
- requirements.txt
- Notebooks

Inspiration from 10x



nf-core/nascent

nf-core/nascent

- Taking over an old repo to avoid duplication of work and fragmenting community
- Main purpose is going from FastQ to counts, nascent transcripts, and bedGraph/bigWigs
- The output files can be used in UCSC Genome Browser or in Notebooks

Conversion Progress

- Updated to the most recent nf-core template
- Rebased our Commits on top of the old repo (To preserve v1.0 for any legacy research)

Things left TODO

- Old nascent functionality added in a subworkflow
- Add test data to nf-core test data
- Refgenie nf-core infrastructure to use T2T-CHM13 reference

nf-core Hackathon

- October 27th-29th 2021
- Focus is going to be on converting pipelines to DSL2
- Sign up form