Innovative Genomics Institute × LatchBio

Automating SARS-Cov-2 Variant Calling

for the Innovative Genomics Institute

Overview

During the peak of the Covid pandemic, the Innovative Genomics Institute (IGI) launched their first clinical laboratory to test COVID samples across the Bay Area for different variants and clades of COVID-19. To deliver results more rapidly they automated the entire bioinformatics process of calling & reporting variants using LatchBio.

IGI

Petros Giannikopoulos

Matthew McElroy

Netravathi Krishnappa

Iman Sylvain

Erica Moehle

Alina Minikhanova

LatchBio

Aidan Abdulali

Alfredo Andere

Kenny Workman

Kyle Giffin

Nathan Manske

Max Smolin

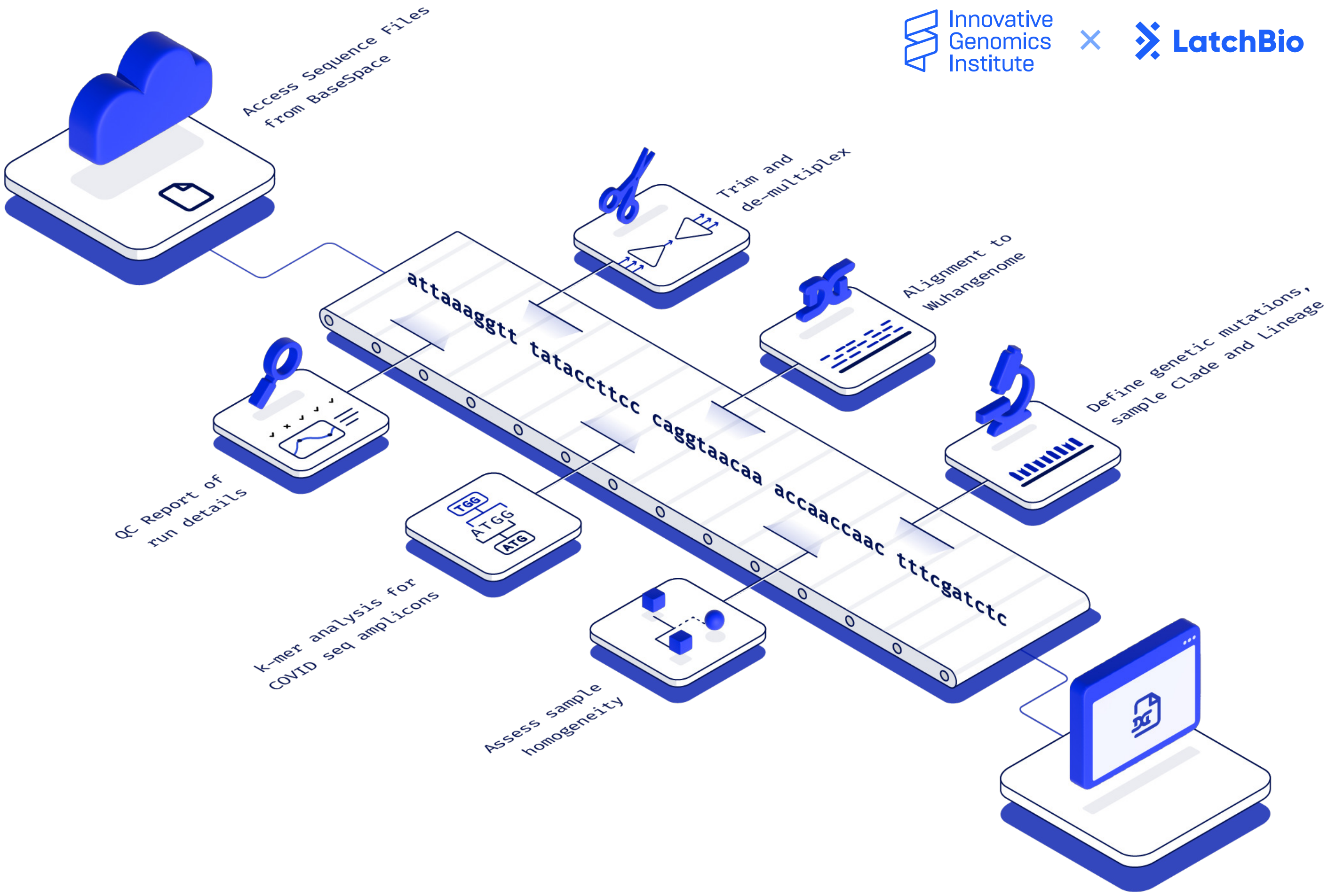
Overview

The team at the IGI Clinical Laboratory identifies 100s of positive Sars-Cov-2 patient samples every week but didn't have the bandwidth to build up the robust infrastructure to sequence these samples.

To address this, we teamed up with them and built a robust computational pipeline on the Latch platform which processes batches of several hundred samples in parallel. The result is a version-controlled and CLIA-compatible output that the testing personnel can use directly for patient care, or share with local public health departments for epidemiological use.

The pipeline takes raw sequencing data directly from the sequencer hub and reports COVID variant information, along with a slew of useful statistics.

SARS Cov-2 Pipeline


Innovative Genomics Institute
×
LatchBio
Access Sequence Files from BaseSpace
Trim and de-multiplex
Alignment to Wuhangenome
Define genetic mutations, sample Clade and Lineage
QC Report of run details
k-mer analysis for COVID seq amplicons
Assess sample homogeneity
attaaaggtt tataccttcc caggtaacaa accaaccaac tttcgatctc
TGG
ATGG
ATG

Motivation

RNA viruses, such as SARS-CoV-2, mutate over time as they spread between individuals. The viruses that become better at surviving and spreading can become more common than their non-mutated counterparts. Viruses with these mutations are known as “variants”. They are genetically distinct from the original strain.

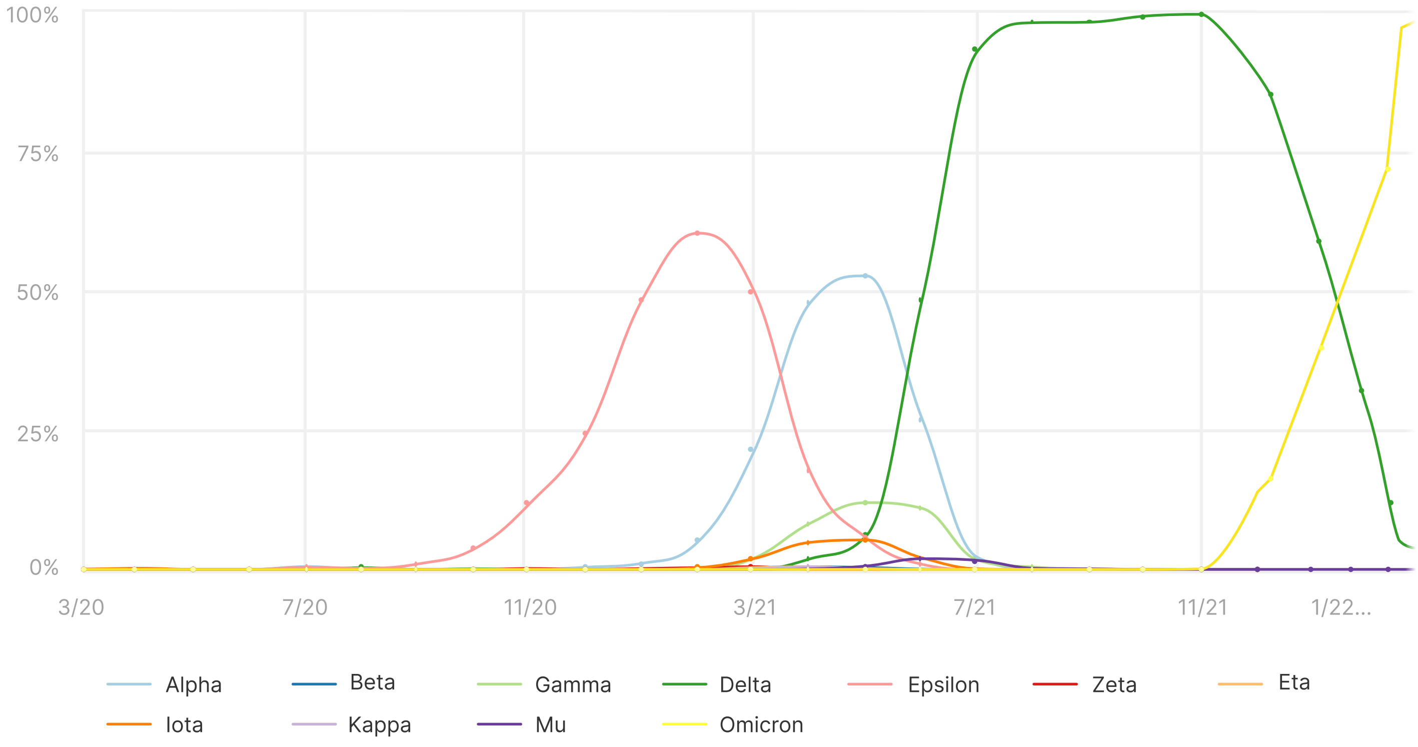
The Omicron variant of SARS-CoV-2 is one such example. With a mutation on its spike protein, it is more easily transmitted and more contagious than previous strains. It became the dominant variant in the US over the winter.

New SARS-Cov-2 variants are detected every week and classified into different categories by the WHO and CDC. Variants of interest have greater transmissibility by effectively evading our immunity.

Variants of concern can evade vaccines, making a prior vaccine ineffective to new variants. Variants of high consequence completely evade prior vaccines, natural immunity, and other existing protections.

Identifying these variants early is critical to ensure that testing, treatment, and vaccine strategies are targeting the correct virus and not being bypassed by novel variants. This information is also key for patients and doctors to understand which variant of the virus they have and how to react accordingly.

Variant Percentage of SARS-CoV-2 Specimens Sequenced in California



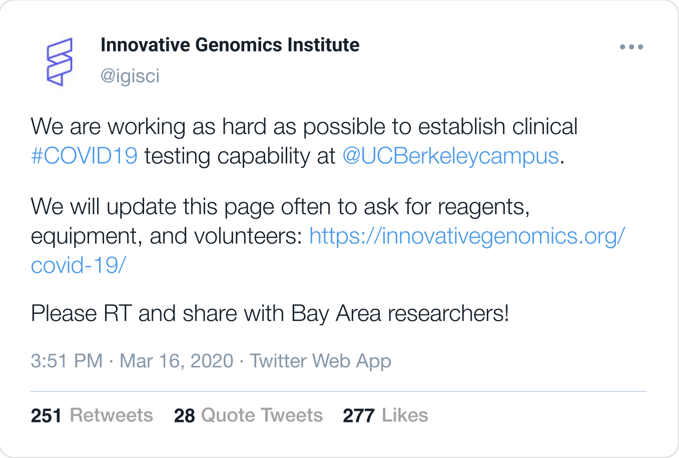
Source: www.cdph.ca.gov/Programs/CID/DCDC/Pages/COVID-19/COVID-Variants.aspx

Innovative Genomics Institute Clinical Laboratory

During the beginning of the pandemic, the IGI launched an operation to quickly provide COVID test results for patients across the Bay Area, focusing on the UC Berkeley Campus, first responders, and underserved populations in the area.

IGI Founder and Nobel Laureate Jennifer Doudna initially turned to Fyodor Urnov to command the operation, a gene-editing wizard who had already been leading multiple efforts at the IGI, including efforts to develop affordable cures for sickle cell disease and gene-editing techniques to make soldiers immune to radiation poisoning.

They put out a call to action and within two days 860 people had responded to volunteer. They converted a 2,500-square-foot space on the ground floor of The IGI Building on the UC Berkeley Campus into a coronavirus testing facility.


Innovative Genomics Institute
@igisci

We are working as hard as possible to establish clinical #COVID19 testing capability at @UCBerkeleycampus.

We will update this page often to ask for reagents, equipment, and volunteers: https://innovativegenomics.org/covid-19/

Please RT and share with Bay Area researchers!

3:51 PM · Mar 16, 2020 · Twitter Web App

251 Retweets **28** Quote Tweets **277** Likes

Variant Operation

As they increased their capacity to 1000s of tests per week, they realized a simple "positive or negative" COVID result was not sufficient. Public health officials needed genomic information to track the spread of new variants, information which would also be valuable to clinicians and patients wanting to know what strain they have. Thus the IGI knew it was important to start identifying COVID variants as quickly as possible. During a time of global uncertainty, The IGI assembled a team of scientists and health experts to build a COVID variant identification program. That team was Petros Giannikopoulos, Matthew McElroy, Erica Moehle, Alina Minikhanova, Iman Sylvain, and Netravathi Krishnappa.

They set up an end-to-end operation to sequence, quality control and classify the variants of Sars-Cov-2 RNA from clinical test samples.

Sequencing machines have notoriously massive data footprints. Each run can generate 100s of gigabytes of raw genomic code. The size of this data alone can crash laptops and cause day-long wait times uploading and downloading from more robust storage solutions. Typical cloud-storage websites (Google Drive, Dropbox) are hard to integrate with and lack proper compliances for clinical data.

The data storage issue is compounded by more painful downstream analysis: A scientist cannot do much with a raw file of cryptic "A"s and "T"s repeated millions of times. Analysis methods require bespoke logic known by a lucky few dubbed "bioinformaticians." These people write thousands of lines of code in different languages to process and analyze the data off of the sequencer. However, the tools available are extremely clunky and time-consuming to set up.

For the IGI Clinical Laboratory, analyzing sequenced samples requires a custom bioinformatics pipeline, well-orchestrated cloud computing infrastructure and an interface all working in sync to process sequence datasets to generate accurate results every week.

Yet, like most scientists in biotech, the CLIA lab members are too busy at the bench to wrangle cloud computing, bioinformatics tooling, and complex infrastructure provisioning. They are busy trying to do the actual biology. Thus the IGI team was blocked, looking for a way to accurately identify COVID variants quickly and efficiently, without having to spend precious time piecing together different bioinformatics processes.

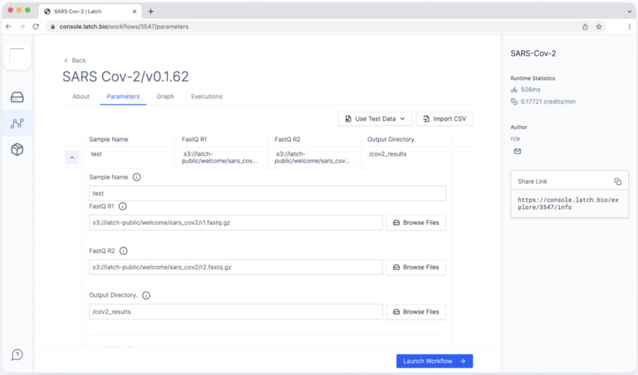
The Solution

To address this bottleneck, the IGI called upon the LatchBio team to build a biocompute workflow on Latch to parallelize sample processing and produce clean results for scientists through a UI.

LatchBio first constructed a pipeline to take raw sequencing data and accurately report COVID variant information, along with a load of useful statistics.

To filter out samples without COVID RNA early in the pipeline, the sequence fragments are first compared against a fragment database of known human and COVID origin. Those samples that surpass a threshold of “hits” with known COVID sequence fragments advance.

The millions of short read fragments are then assembled against the reference Wuhan COVID genome and differences between the resulting consensus sequence and the reference are recorded. Additionally, more statistics about the depth and coverage of the sequencing reads are collected to further ensure the result is of high quality. The pipeline then feeds the sample into Pangolin to classify the sample as a known COVID variant and clade.



SARS Cov-2/v0.1.62

Sample Name	FastQ R1	FastQ R2	Output Directory.
test	s3://latch-public/welcome/sars_cov...	s3://latch-public/welcome/sars_cov...	/cov2_results

Sample Name: test

FastQ R1: s3://latch-public/welcome/sars_cov2/r1.fastq.gz

FastQ R2: s3://latch-public/welcome/sars_cov2/r2.fastq.gz

Output Directory.: /cov2_results

SARS-Cov-2

Runtime Statistics
506ms
0.17721 credits/min

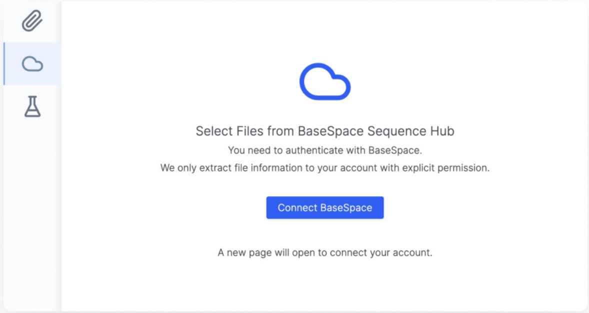
Author
n/a

Share Link
https://console.latch.bio/explore/3547/info

The Sars-Cov-2 pipeline on Latch

After the first iteration, CLIA scientists could simply drag and drop genomic sequence data and click “Launch” — this would initiate the whole SARS-Cov-2 pipeline to run end-to-end on the cloud.

To reduce upload times, the LatchBio team set up BaseSpace Integration, which allowed IGI to connect directly to the data coming from the sequencer. This meant their team no longer had to wait for data to upload, a process which can often take hours with large sequencing datasets.



Select Files from BaseSpace Sequence Hub
You need to authenticate with BaseSpace.
We only extract file information to your account with explicit permission.

A new page will open to connect your account.

Sequencer runs can be pulled directly into Latch through the intergration with BaseSpace

After integration, the CLIA team leverages a custom-built SARS-Cov-2 aggregator workflow to automatically generate a CSV from all of the BaseSpace files. This CSV serves as a perfect template as it contains the appropriate sample names and metadata necessary to match the parameters of the SARS-Cov-2 workflow. From there the team simply clicks “import CSV” and then “Launch”, a trigger which spins up 100s executions on AWS to run in parallel rather than sequentially. This parallelization turns what is normally a multi-hour job into 3 minutes of work.

In minutes, a flood of successive results are generated within Latch data, a collaborative file system engineered to support biological sequences. This enables scientists to navigate the results of the pipeline in real time by accessing shared files on the cloud as opposed to siloed data on local computers. If they desire to share the results with a colleague, the “share” button generates a secure link to that file in an instant.

The next and final step is simply uploading the results to The IGI’s lab inventory management system (LIMS). From there the results can be shared with officials at local hospitals and Bay Area health clinics to help patients understand the state of their infection.

Conclusion

The IGI is now testing hundreds of samples per week using Latch, allowing them to expand the speed, scope, and scale of their operation and save months on bioinformatics setup. What was previously a bottleneck is now an automated, end-to-end solution; one which is enabling precise results to patients throughout the Bay Area. We are grateful for the chance to help with the IGI as they work around the clock to keep us all safe.

Learn More

To learn more about the IGI:
innovativegenomics.org

To learn more about LatchBio:
latch.bio