Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA

ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

ICSC

Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

Centro Nazionale di Ricerca in HPC, Big Data e Quantum Computing

Elisa Rossi, Xhulio Dhori

Plenaria Spoke 8 - 24 e 2

What is OMICS?

CINECA

CINECA has built OMICS, an **user-friendly web interface** designed for seamless **uploading** and **processing of genomic** data through NVIDIA Clara Parabricks, a software suite of GPU-accelerated genome analysis tools for genomic **read alignment**, **variant calling**, and GVCF filtering and post-processing.

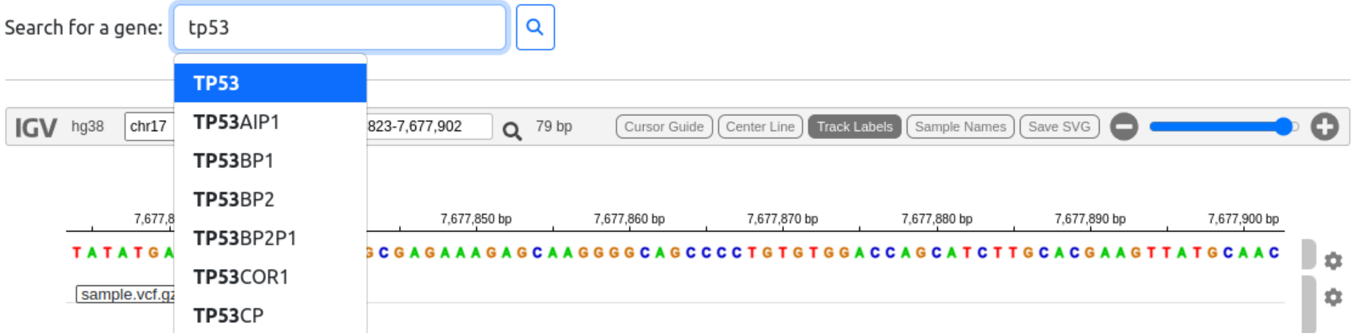
Main Features

- ✓ Data security and privacy compliance with EU GDPR regulations are maintained through a secure encrypted cloud Virtual Machine (VM)
- ✓ VM is equipped with GPUs to make the analyses increasingly scalable

What is OMICS?

CINECA

- OMICS main dashboard allows users to **effortlessly upload their raw data** and **customize advanced options** for three containerized pipelines
- Capability to swiftly upload and download large datasets (e.g. raw sequence data)



Search for a gene: tp53

- OMICS aims to streamline WGS data analysis **while prioritizing user experience**, data **privacy**, and **accelerated** processing through **GPU** technology, aiming at supporting activities of Spoke 0 and Spoke 8 of ICSC

Il gruppo di lavoro CINECA

- Giuseppe Melfi
- Silvia Gioiosa
- Alessandro Grottesi
- Juan Mata Naranjo
- Xhulio Dhori

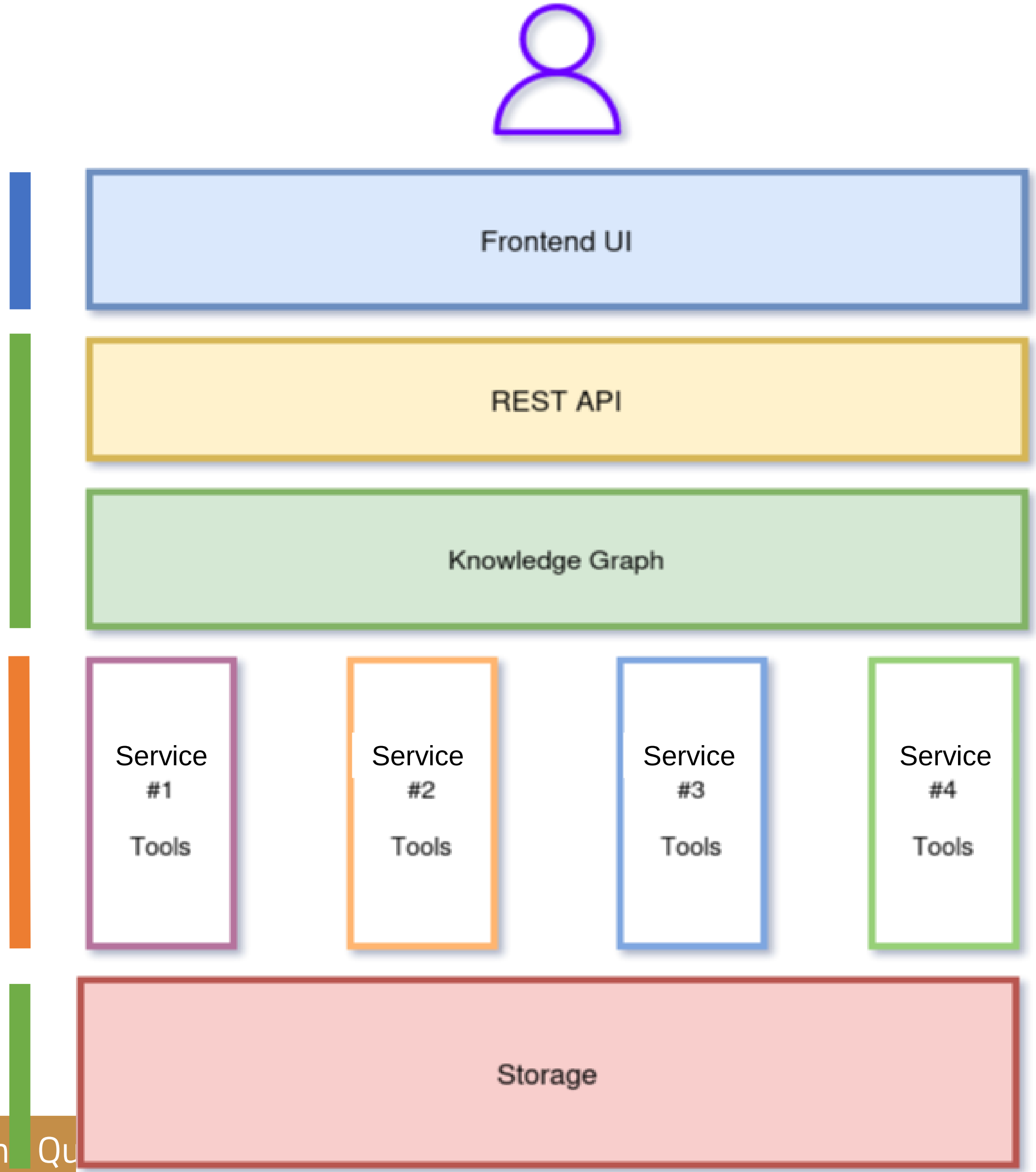
Photo by Tim Mossholder on Unsplash

Under the hood

Let's look at what's going on behind the scenes

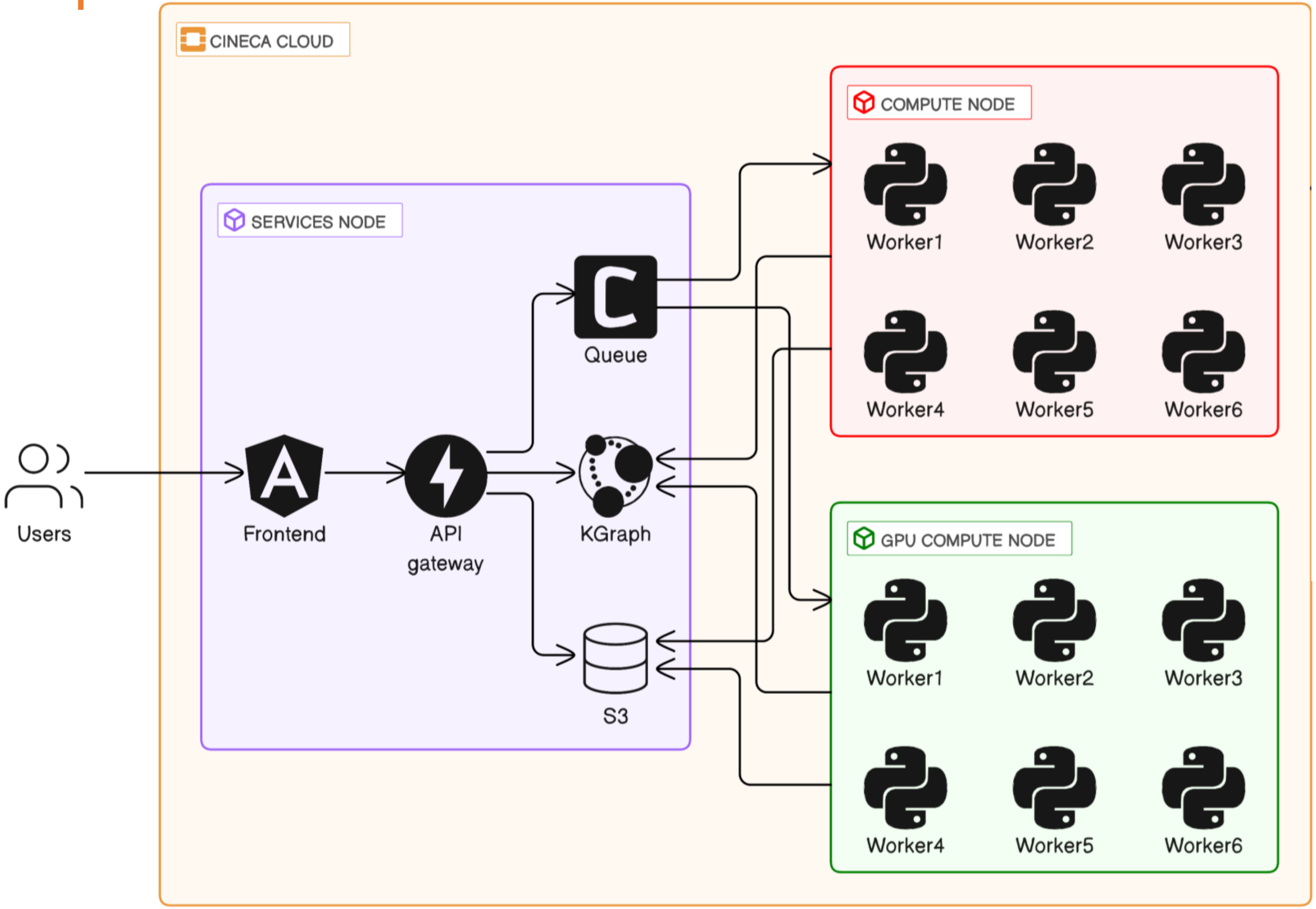
High level architecture

- Simplified/unified access
- Unified data layers
  - **APIs**
  - **Knowledge Graph**
  - **Storage**
- Vertical custom tools



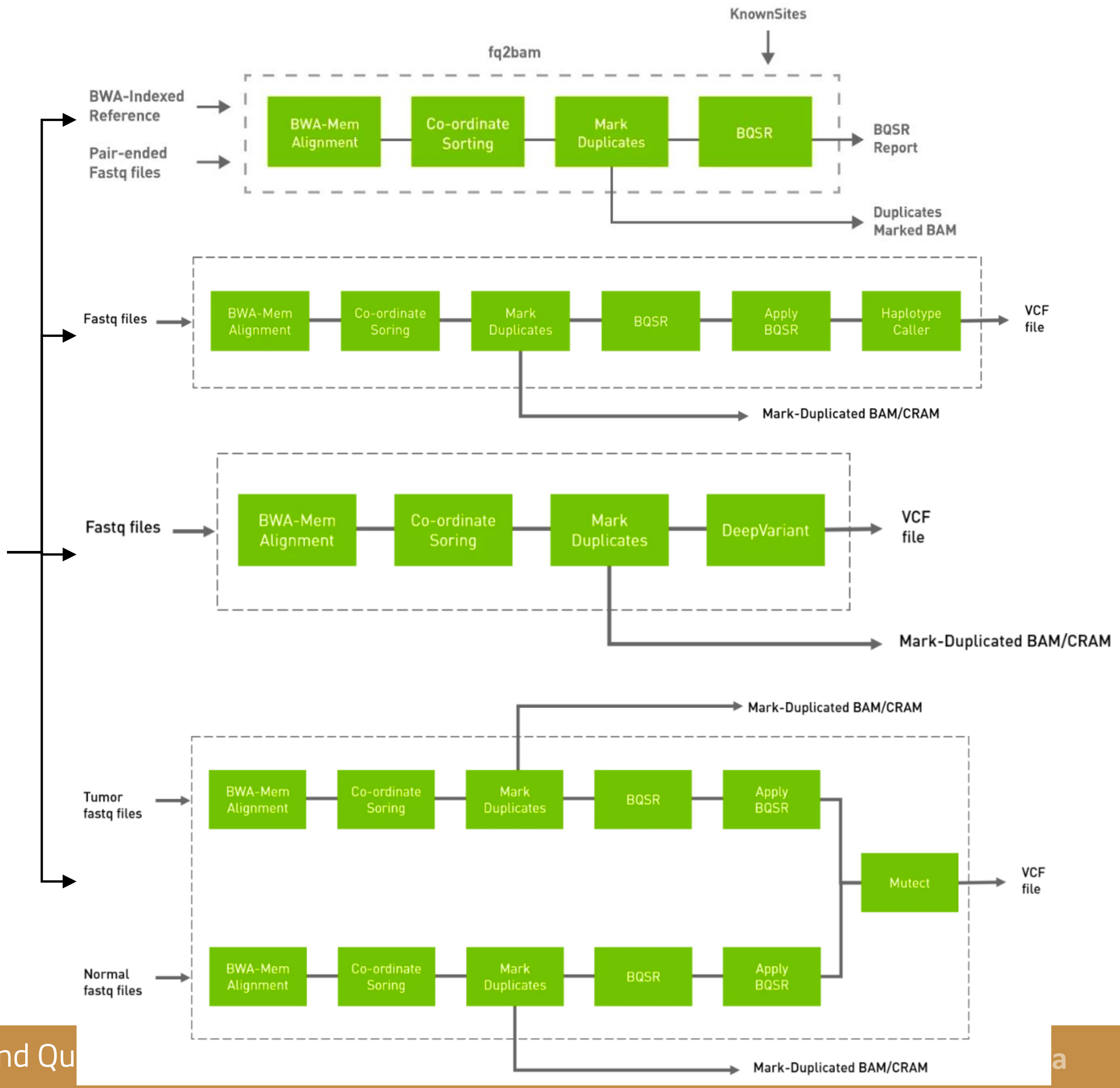
Design properties

- **Open Source** techs
- Microservices oriented
- Cloud ready
- Easy portability
- Easy scalability
- **Private Cloud** resources

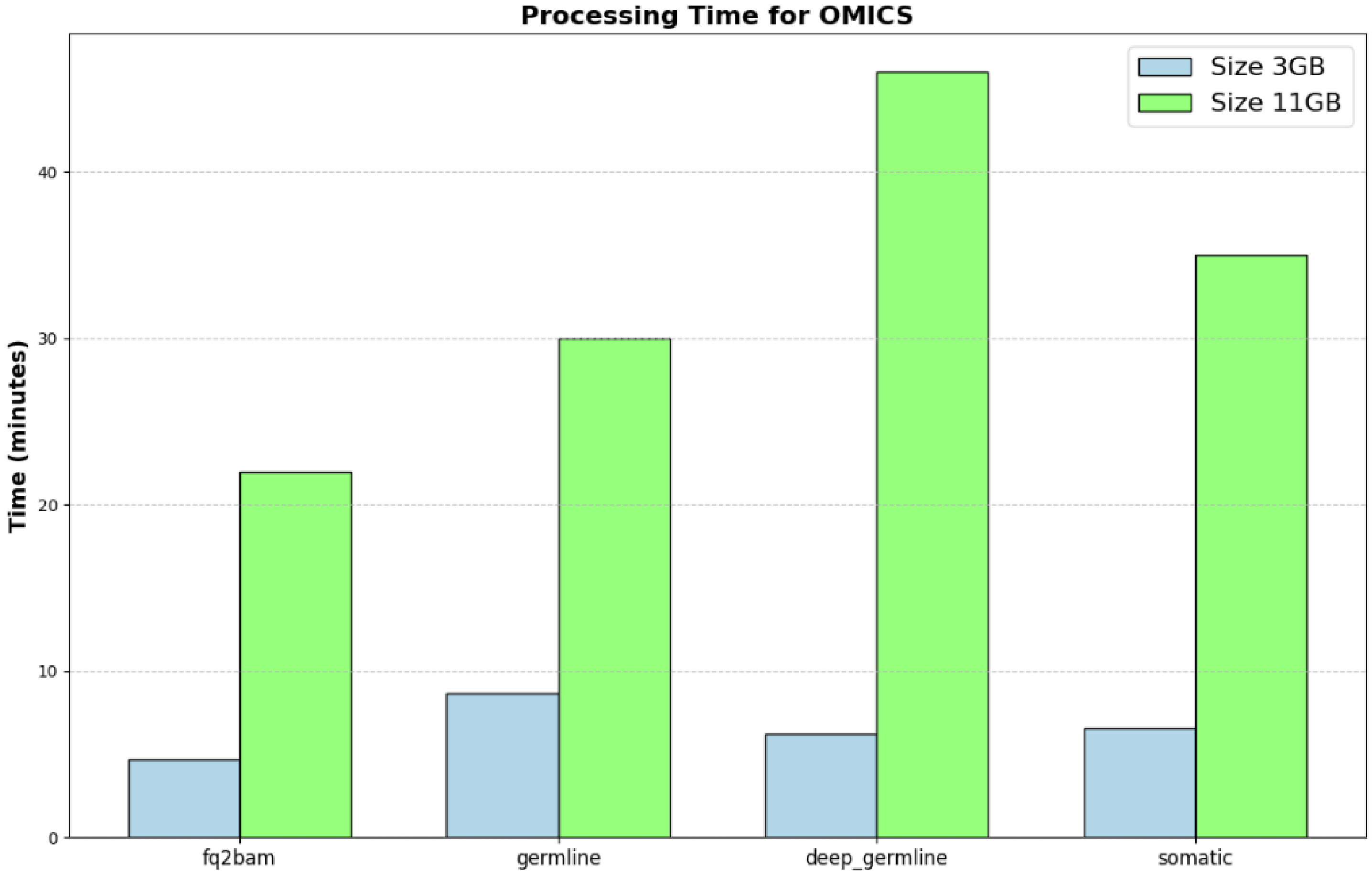


Current Features

- **Multichunk Upload** (User can upload large fastq files directly from the interface)
- **Quality Check** (Fastqc and Multiqc report)
- **Ready to use pipelines** on GPUs:
  - Fq2bam
  - Germline
  - DeepGermline
  - Somatic
- **Annotation** tools (VEP)
- **Visualization** (IGV Browser)
- **User Guide**



BENCHMARKS

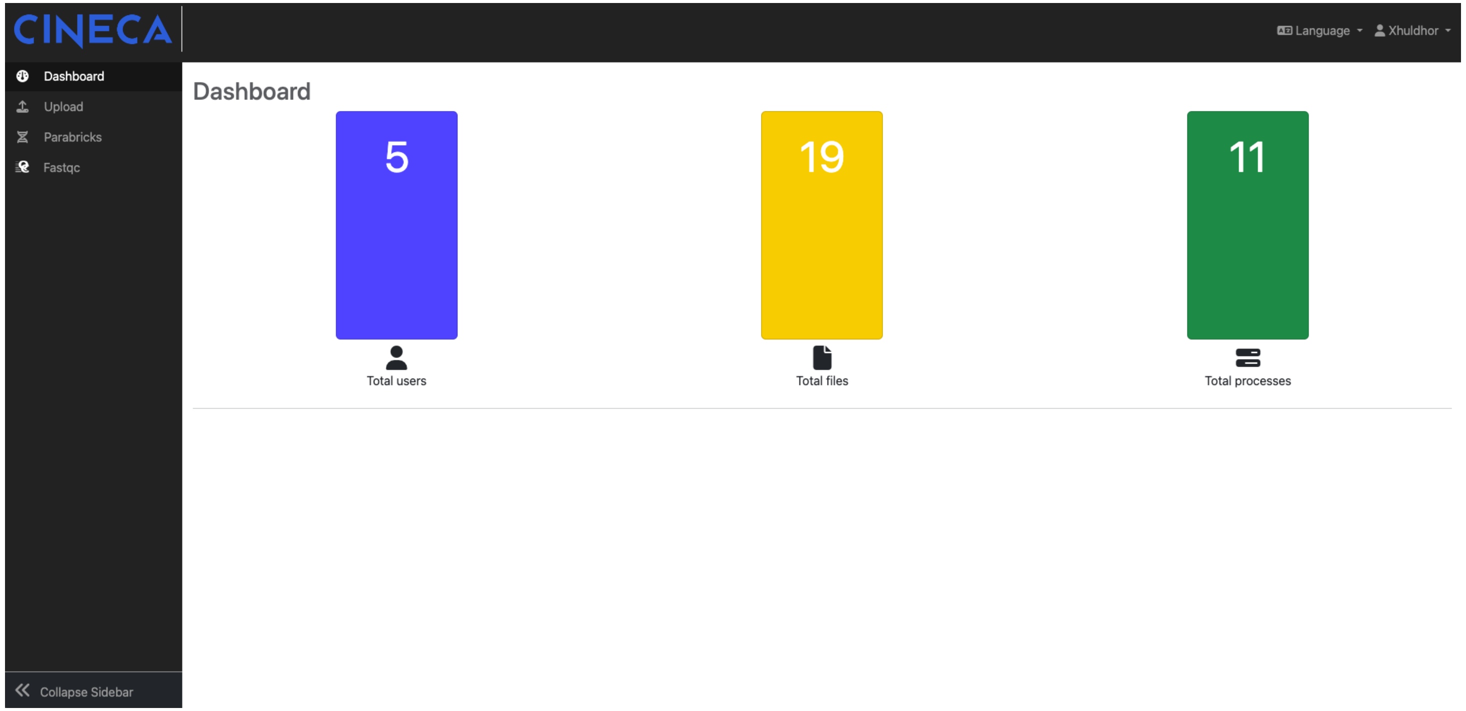

Processing Time for OMICS
Size 3GB
Size 11GB
Time (minutes)
0
10
20
30
40
fq2bam
germline
deep_germline
somatic

NEXT STEPS

- Beta tester **feedbacks**
- Partners **feedback** on **data management** and permissions, **sharing** results within own organization
- **Batch analysis** on multiple samples
- **Re-run** analysis with the same parameters
- **Upload** with **command line**
- **Preprocess** your data (trimming…)
- Cloud **upgrade** (more GPUs available)
- **UX/UI** improvements

Demo live

DASHBOARD



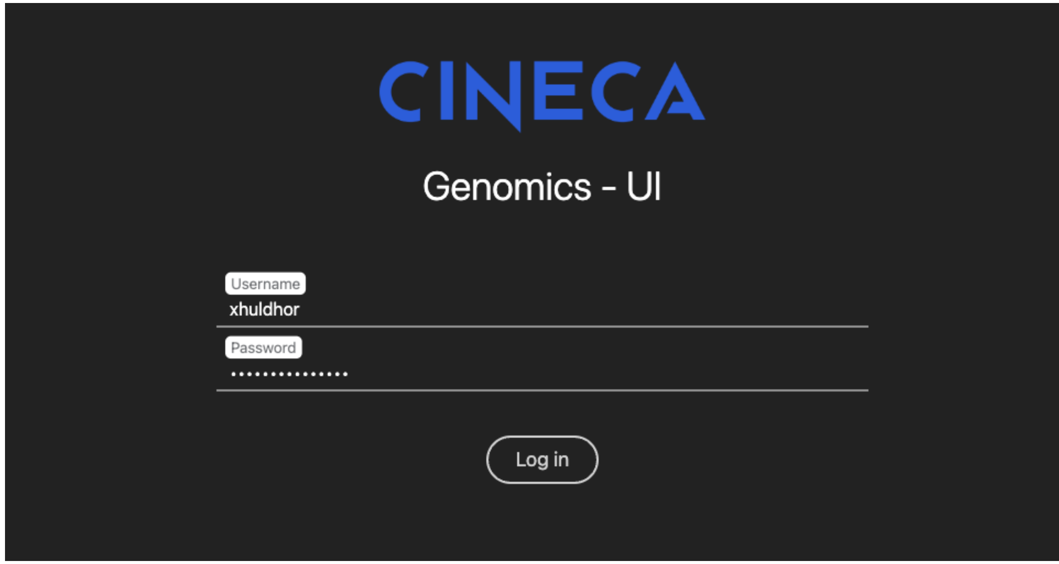
Dashboard

5 Total users

19 Total files

11 Total processes

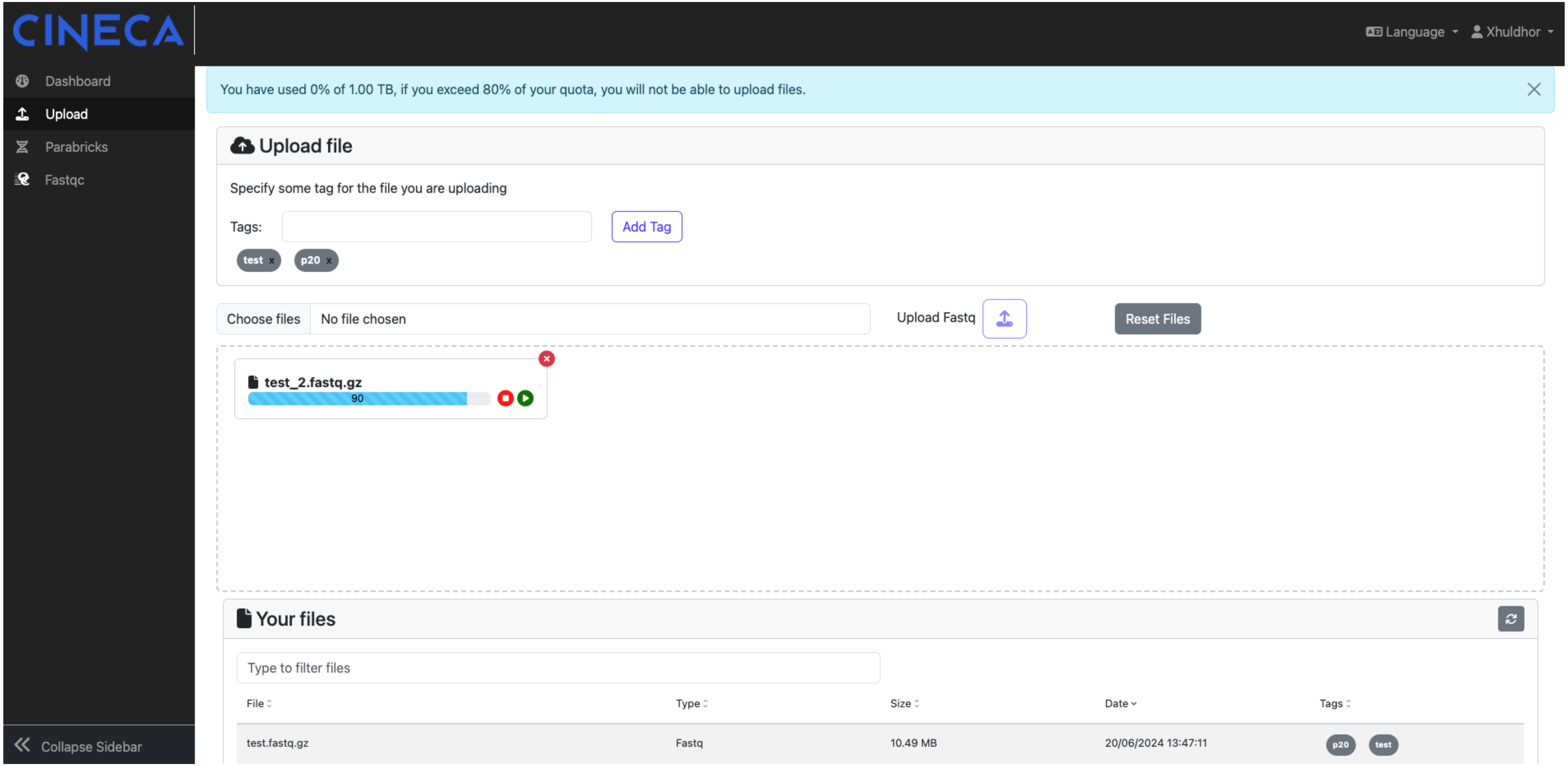
AND LOGIN



Username: xhuldhor

Password: •••••••••••••••

UPLOAD FILES



You have used 0% of 1.00 TB, if you exceed 80% of your quota, you will not be able to upload files.

Upload file

Specify some tag for the file you are uploading

Tags:

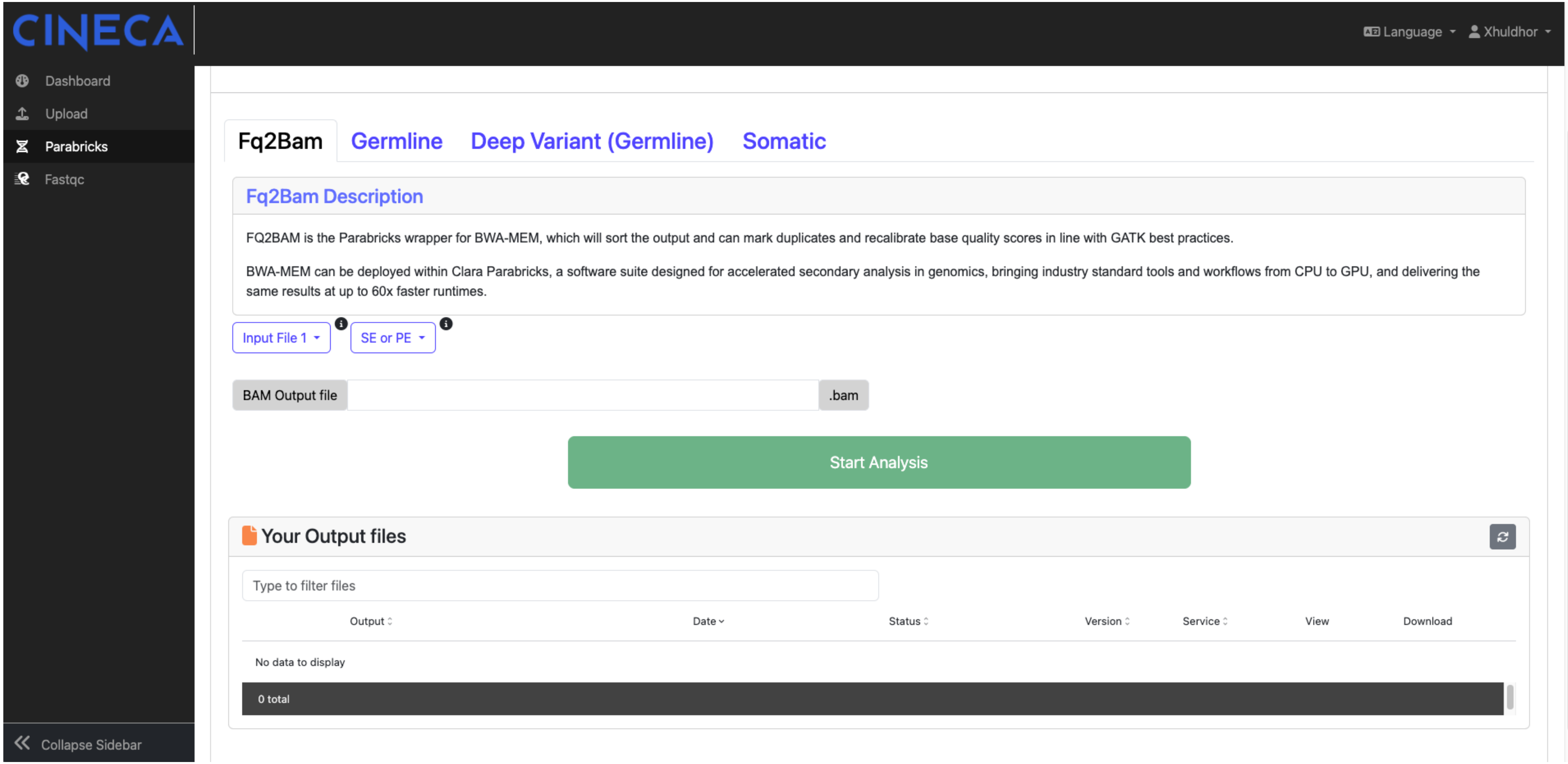
No file chosen

test_2.fastq.gz 90

Your files

File	Type	Size	Date	Tags
test.fastq.gz	Fastq	10.49 MB	20/06/2024 13:47:11	p20 test

START GPU ACCELERATED ANALYSIS



Fq2Bam Description

FQ2BAM is the Parabricks wrapper for BWA-MEM, which will sort the output and can mark duplicates and recalibrate base quality scores in line with GATK best practices.

BWA-MEM can be deployed within Clara Parabricks, a software suite designed for accelerated secondary analysis in genomics, bringing industry standard tools and workflows from CPU to GPU, and delivering the same results at up to 60x faster runtimes.

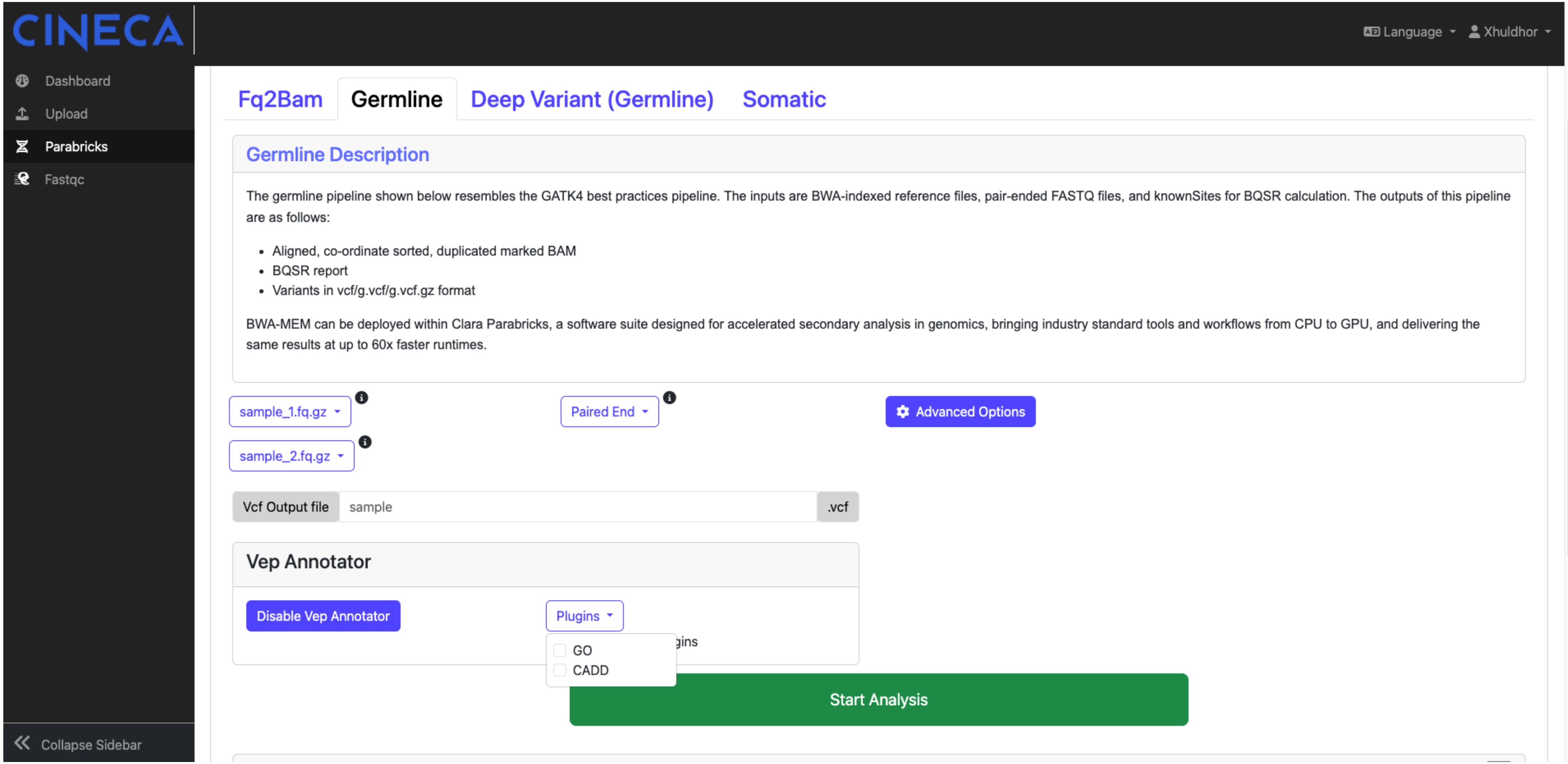
Your Output files

Type to filter files

Output	Date	Status	Version	Service	View	Download
No data to display						

0 total

START GPU ACCELERATED ANALYSIS



Germline Description

The germline pipeline shown below resembles the GATK4 best practices pipeline. The inputs are BWA-indexed reference files, pair-ended FASTQ files, and knownSites for BQSR calculation. The outputs of this pipeline are as follows:

- Aligned, co-ordinate sorted, duplicated marked BAM
- BQSR report
- Variants in vcf/g.vcf/g.vcf.gz format

BWA-MEM can be deployed within Clara Parabricks, a software suite designed for accelerated secondary analysis in genomics, bringing industry standard tools and workflows from CPU to GPU, and delivering the same results at up to 60x faster runtimes.

sample_1.fq.gz

sample_2.fq.gz

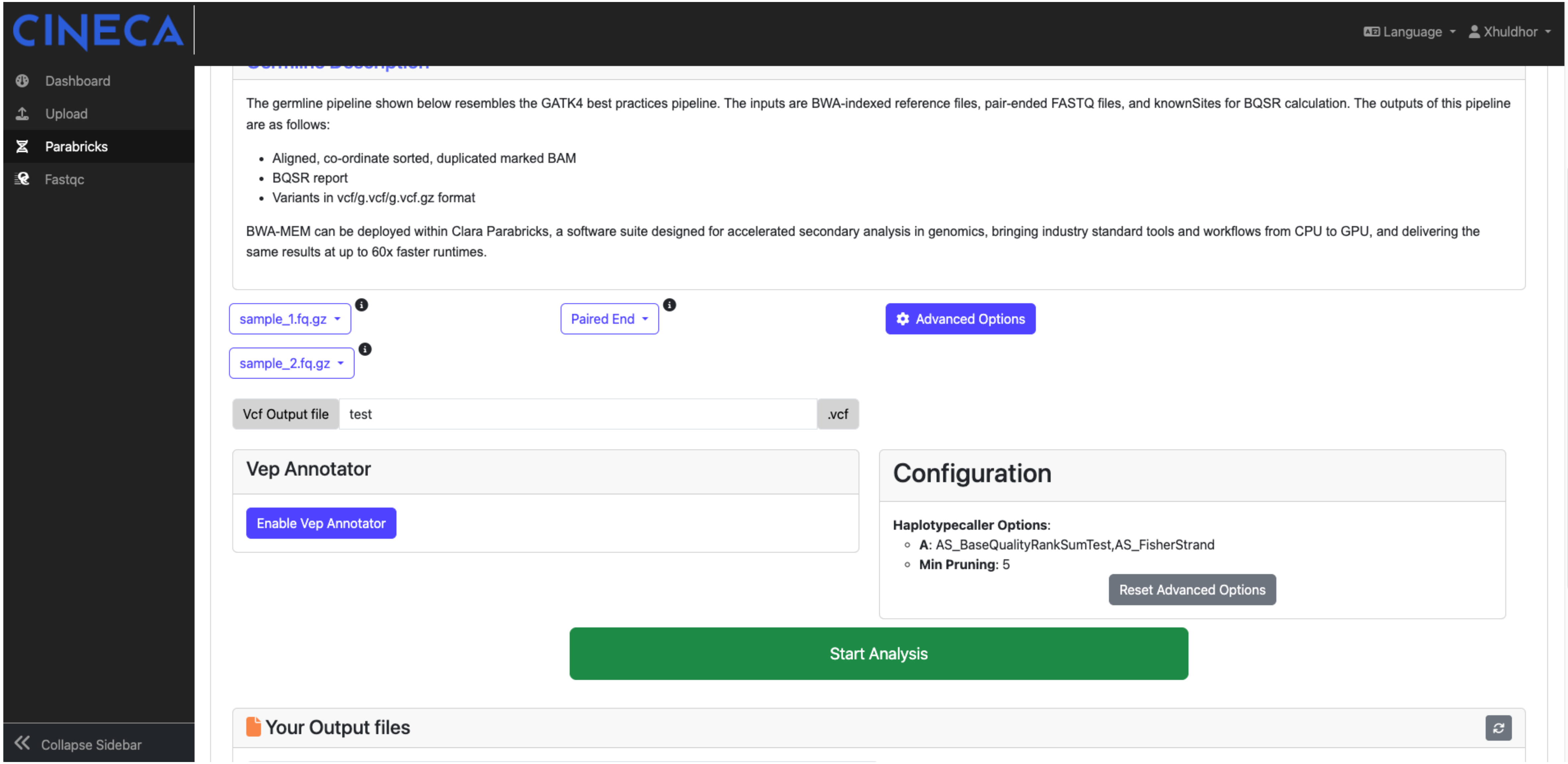
Paired End

Advanced Options

Vcf Output file: sample .vcf

Vep Annotator

CONFIGURE ADVANCED OPTIONS



The germline pipeline shown below resembles the GATK4 best practices pipeline. The inputs are BWA-indexed reference files, pair-ended FASTQ files, and knownSites for BQSR calculation. The outputs of this pipeline are as follows:

- Aligned, co-ordinate sorted, duplicated marked BAM
- BQSR report
- Variants in vcf/g.vcf/g.vcf.gz format

BWA-MEM can be deployed within Clara Parabricks, a software suite designed for accelerated secondary analysis in genomics, bringing industry standard tools and workflows from CPU to GPU, and delivering the same results at up to 60x faster runtimes.

Vcf Output file test .vcf

Vep Annotator

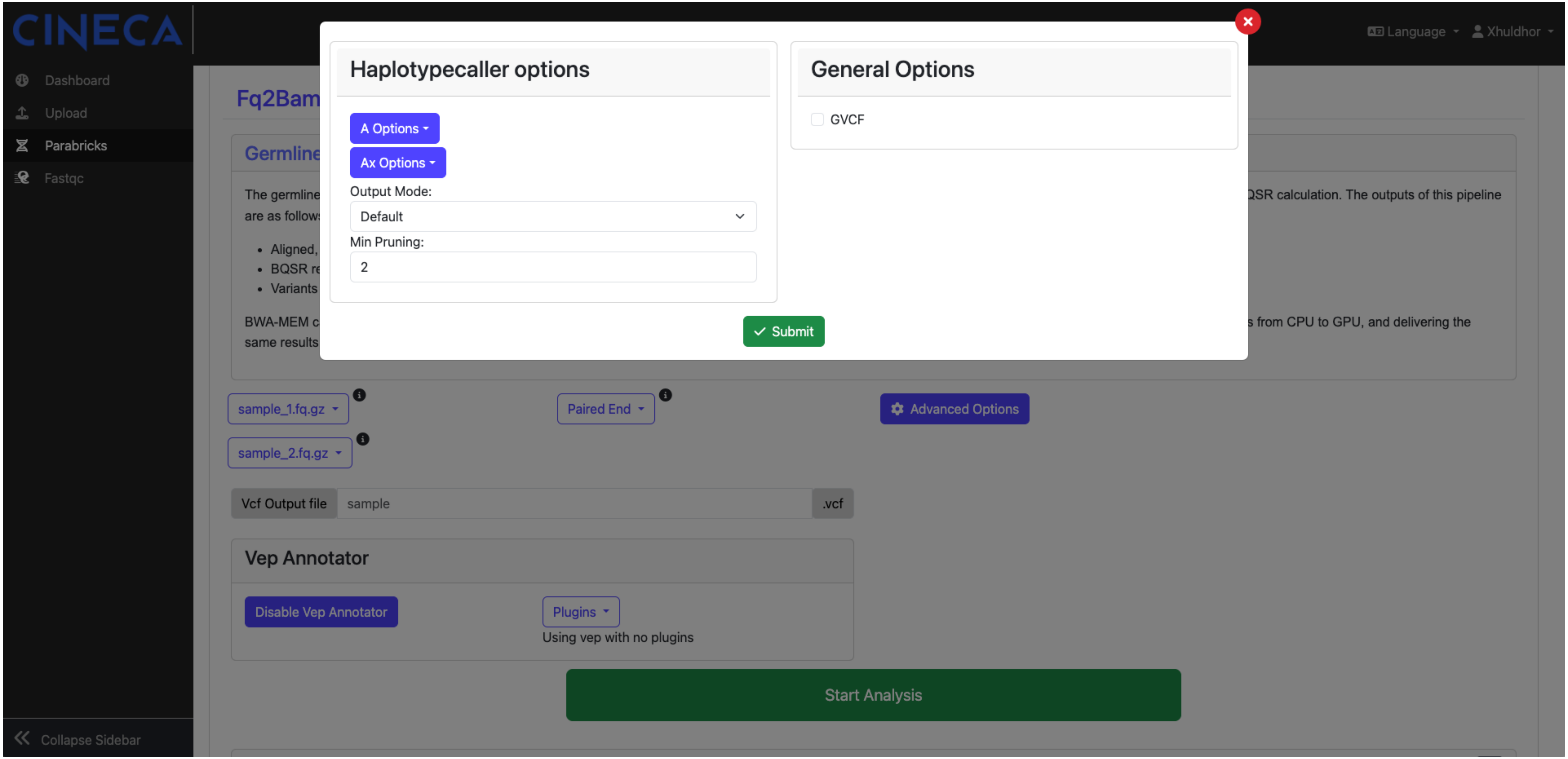
Configuration

Haplotypecaller Options:

- **A**: AS_BaseQualityRankSumTest,AS_FisherStrand
- **Min Pruning**: 5

Your Output files

CONFIGURE ADVANCED OPTIONS



The germline ... QSR calculation. The outputs of this pipeline are as follows

- Aligned, ...
- BQSR re...
- Variants ...

BWA-MEM c... s from CPU to GPU, and delivering the same results ...

Haplotypecaller options

Output Mode: Default

Min Pruning: 2

General Options

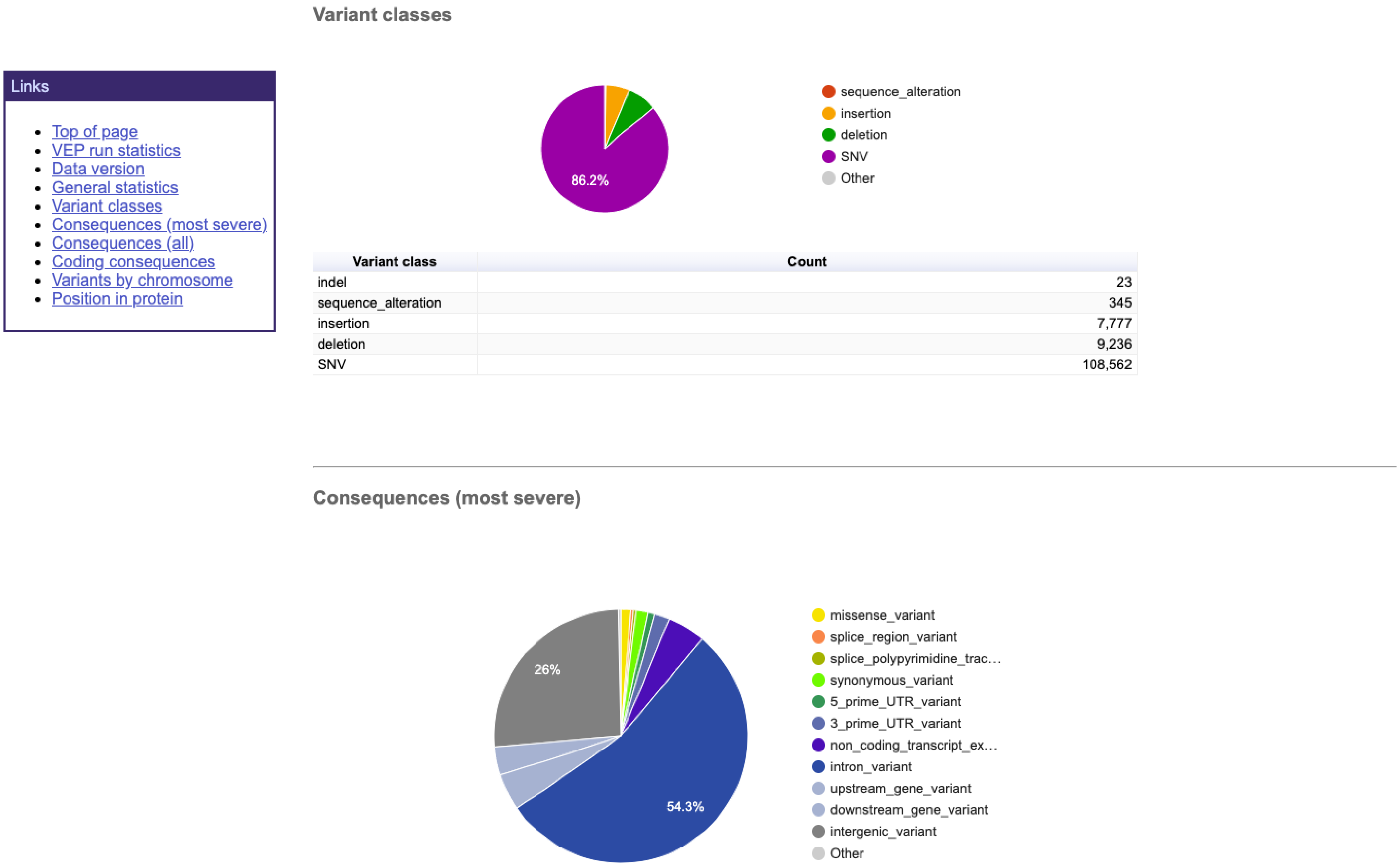
□ GVCF

Vcf Output file: sample .vcf

Vep Annotator

Using vep with no plugins

VEP ANNOTATION



Links

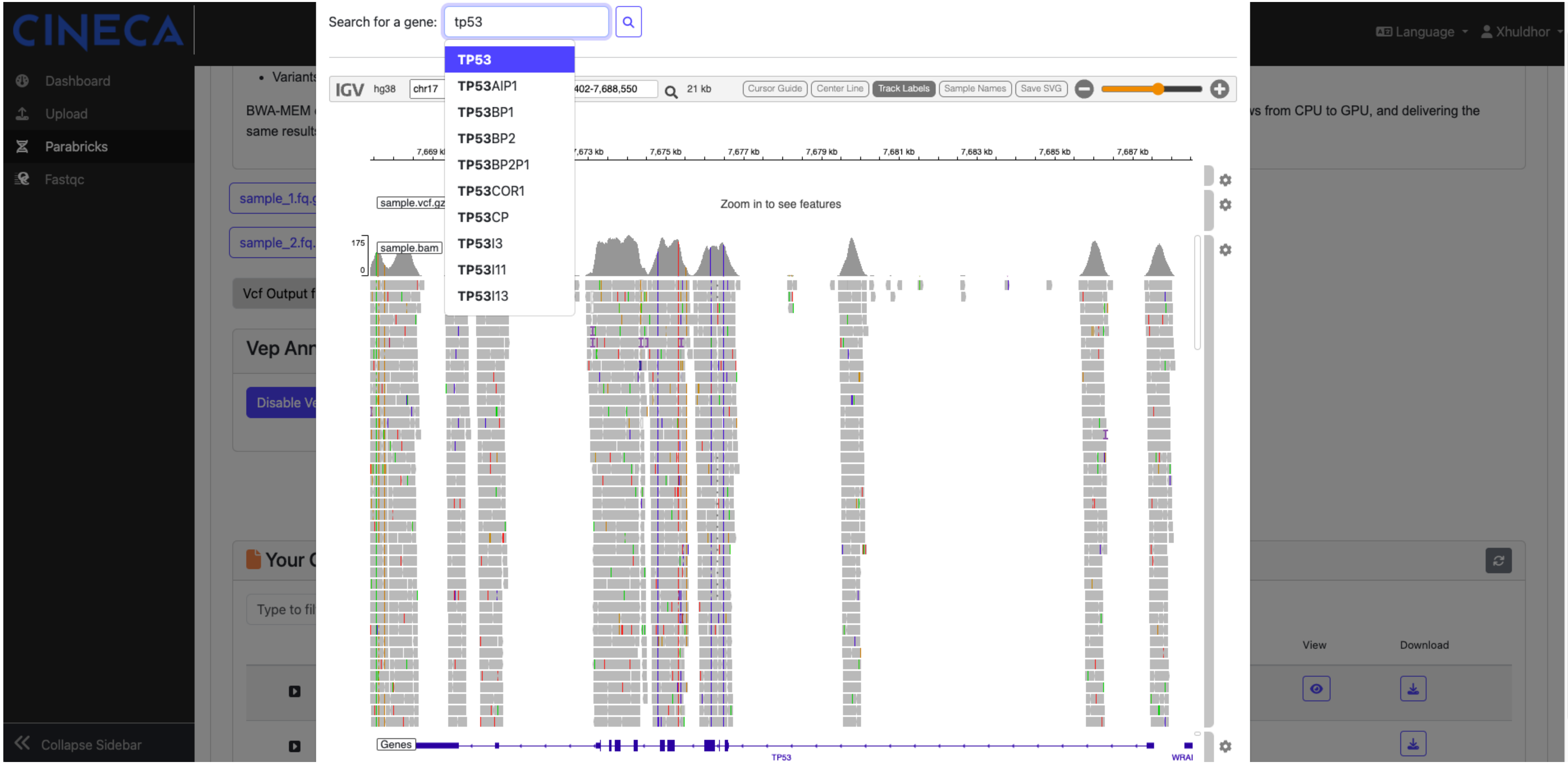
- Top of page
- VEP run statistics
- Data version
- General statistics
- Variant classes
- Consequences (most severe)
- Consequences (all)
- Coding consequences
- Variants by chromosome
- Position in protein

Variant classes

Variant class	Count
indel	23
sequence_alteration	345
insertion	7,777
deletion	9,236
SNV	108,562

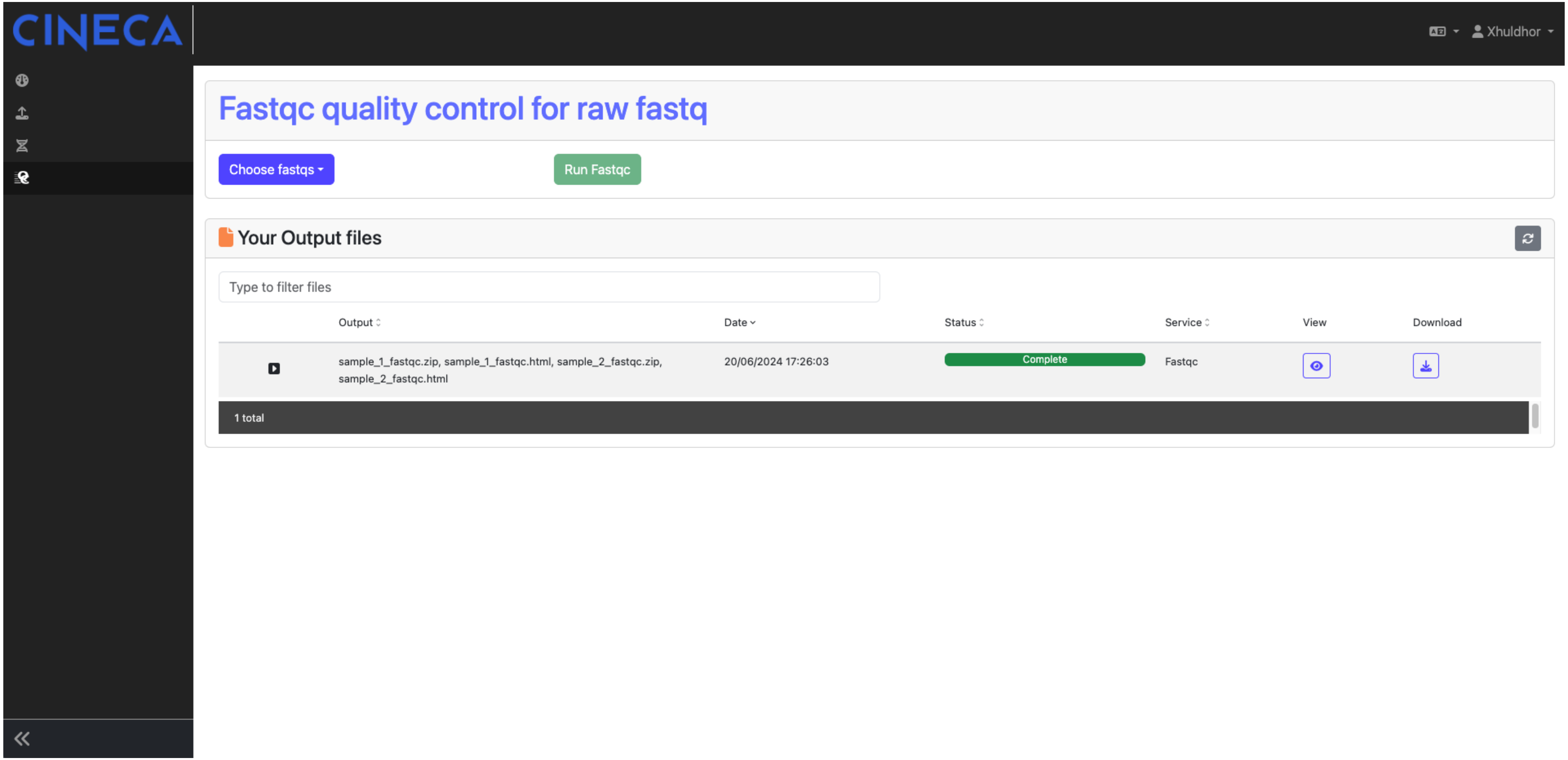
Consequences (most severe)

IGV VISUALIZATION FOR VCF AND BAM FILES



Search for a gene: tp53

FASTQC AND MULTIQC REPORT



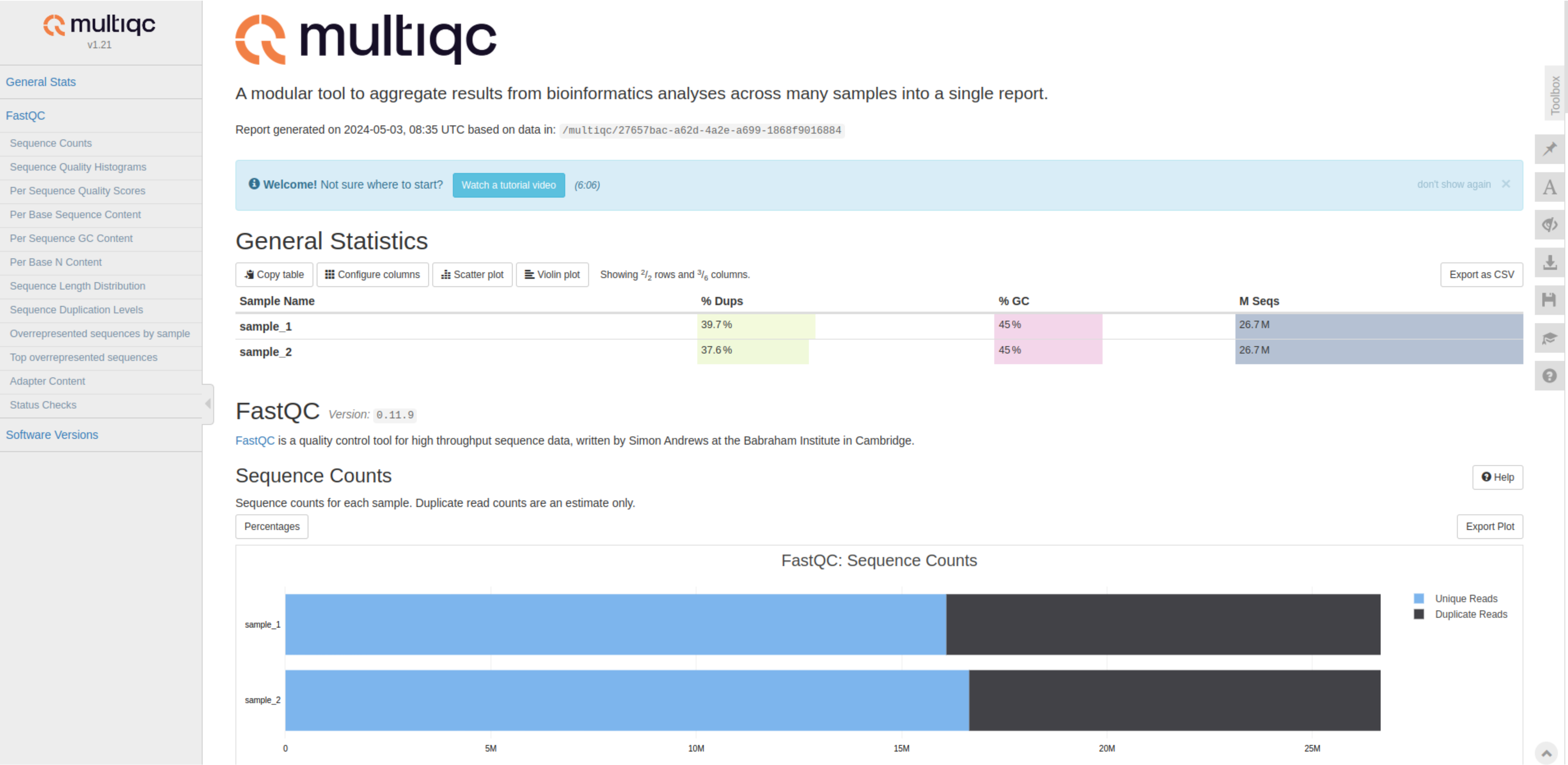
Fastqc quality control for raw fastq

Your Output files

	Output	Date	Status	Service	View	Download
	sample_1_fastqc.zip, sample_1_fastqc.html, sample_2_fastqc.zip, sample_2_fastqc.html	20/06/2024 17:26:03	Complete	Fastqc		

1 total

FASTQC AND MULTIQC REPORT



multiqc
v1.21

- General Stats
- FastQC
  - Sequence Counts
  - Sequence Quality Histograms
  - Per Sequence Quality Scores
  - Per Base Sequence Content
  - Per Sequence GC Content
  - Per Base N Content
  - Sequence Length Distribution
  - Sequence Duplication Levels
  - Overrepresented sequences by sample
  - Top overrepresented sequences
  - Adapter Content
  - Status Checks
- Software Versions

multiqc

A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

Report generated on 2024-05-03, 08:35 UTC based on data in: `/multiqc/27657bac-a62d-4a2e-a699-1868f9016884`

Welcome! Not sure where to start? Watch a tutorial video *(6:06)* don't show again ✕

General Statistics

Copy table | Configure columns | Scatter plot | Violin plot | Showing $^{2}/_{2}$ rows and $^{3}/_{6}$ columns. | Export as CSV

Sample Name	% Dups	% GC	M Seqs
sample_1	39.7 %	45 %	26.7 M
sample_2	37.6 %	45 %	26.7 M

FastQC *Version:* `0.11.9`

FastQC is a quality control tool for high throughput sequence data, written by Simon Andrews at the Babraham Institute in Cambridge.

Sequence Counts

Help

Sequence counts for each sample. Duplicate read counts are an estimate only.

Percentages | Export Plot

FastQC: Sequence Counts