



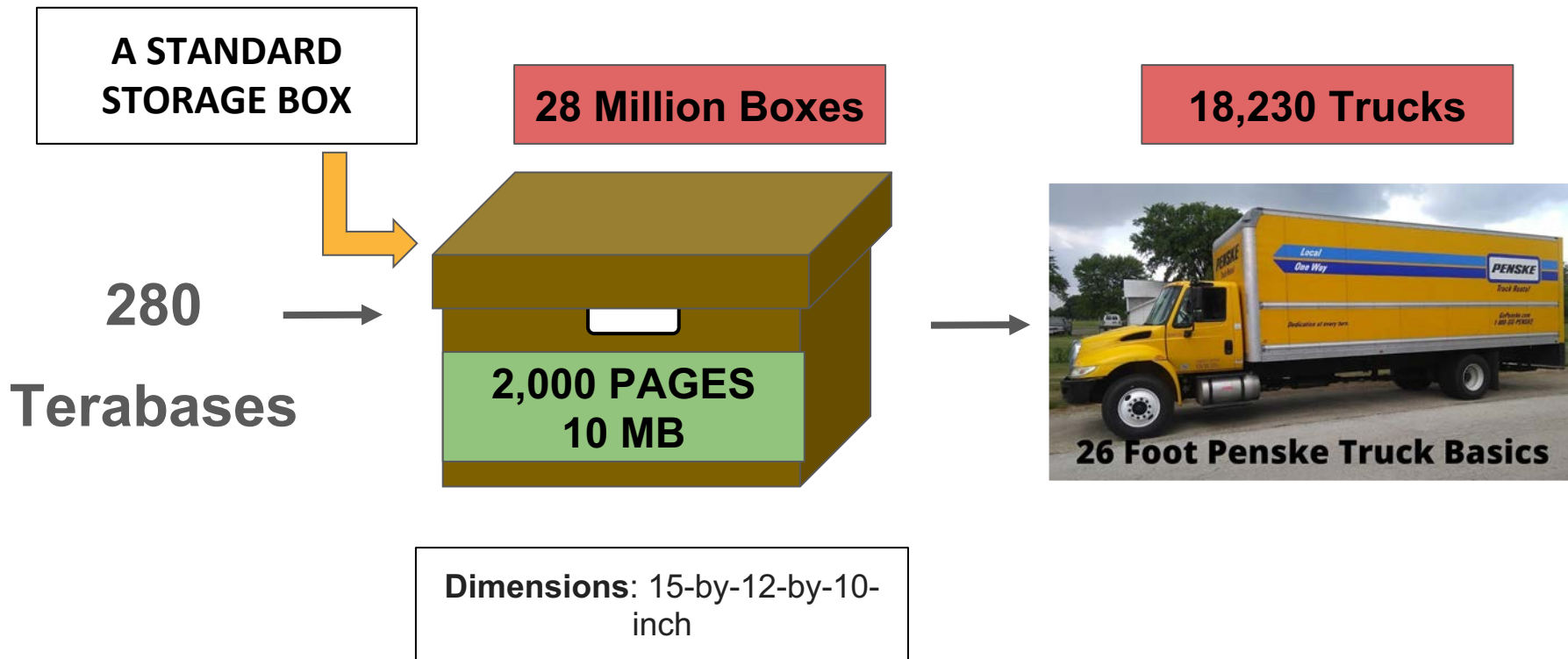
Implementing a Scalable JGI-RQC Pipeline on the Cloud

Date: July 30th, 2021

Presenter: Sandra Chacon

Mentors: Bryce Foster & Zhong Wang

JGI is Producing Big Sequencing Data

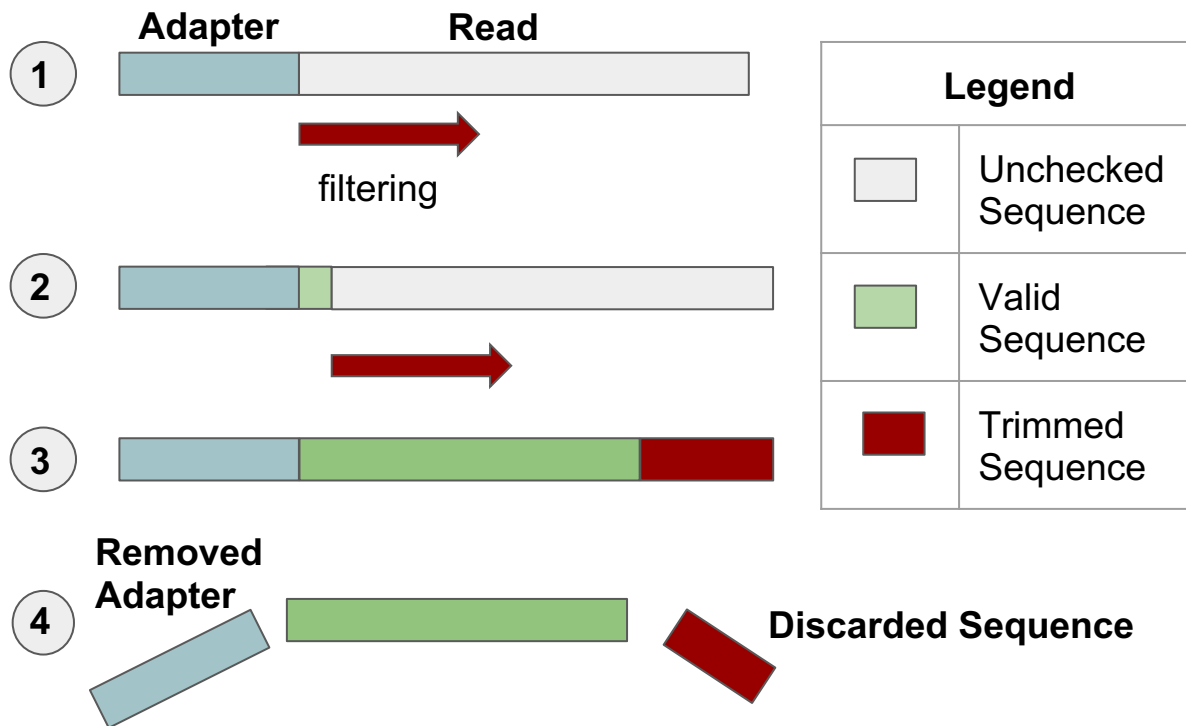


What is RQC Filtering and why is it important?

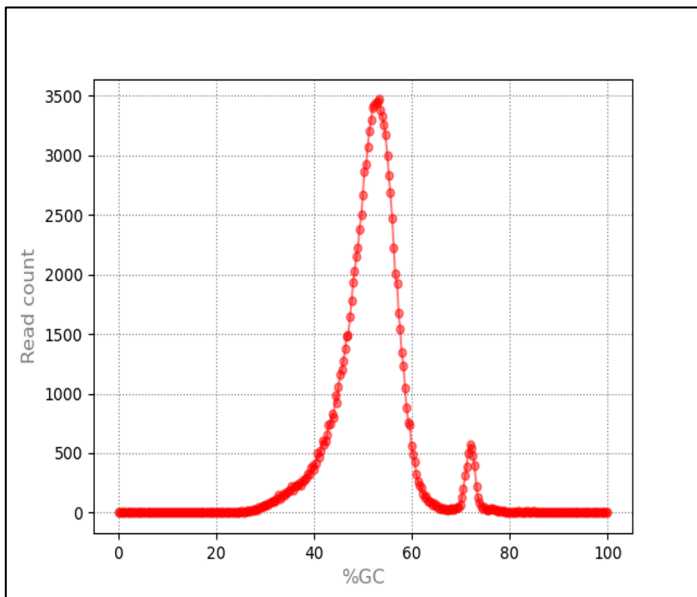
Short answer: garbage in, garbage out



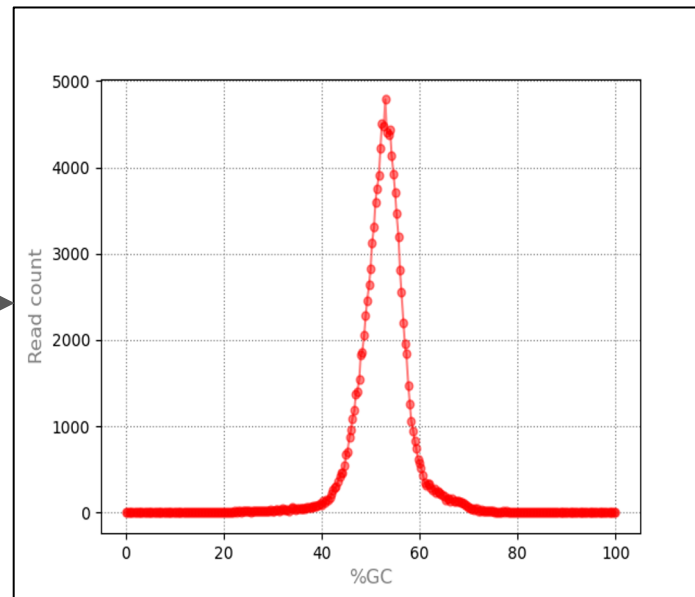
By Filtering



An example of contamination



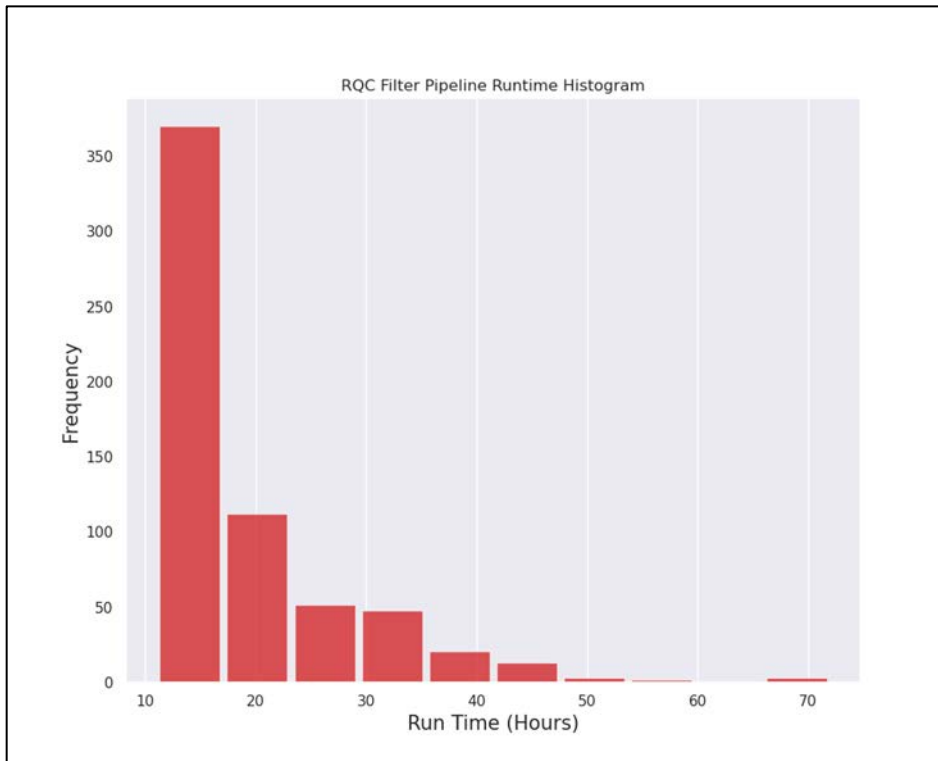
Unfiltered Sequencing Data



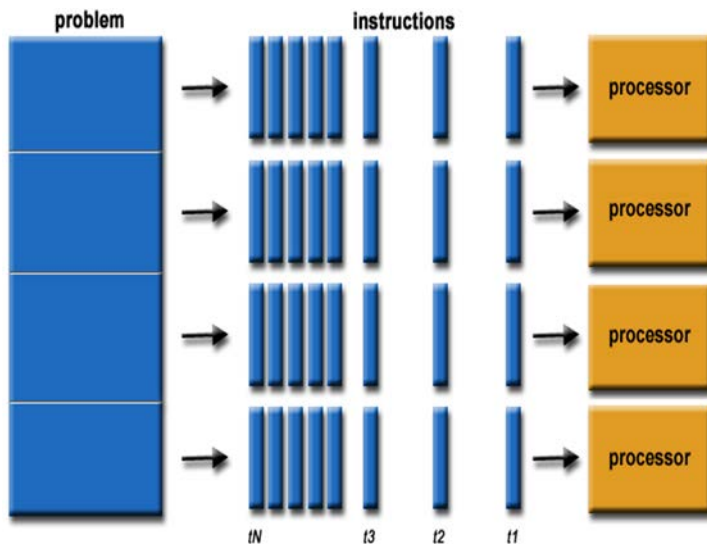
Filtered Sequencing Data

Current RQC Pipeline Runs on a Single Server

For large datasets it will take a long while




Data Parallel Processing



How Does Our Python Program Work?

Test_file:

 @H100 read 1: ACCATCTC
@H100 read 2: CATGCATG
@H101 read 1: TTCGAGTC

...

@H5000 read 2: CATCATGA

worker evaluates sequence:
H100 read 1

quality test

n test

kmer test

@H100 read 1: ACCATCTC
@H100 read 2: CATGCATG
@H101 read 1: TTCGAGTC

...

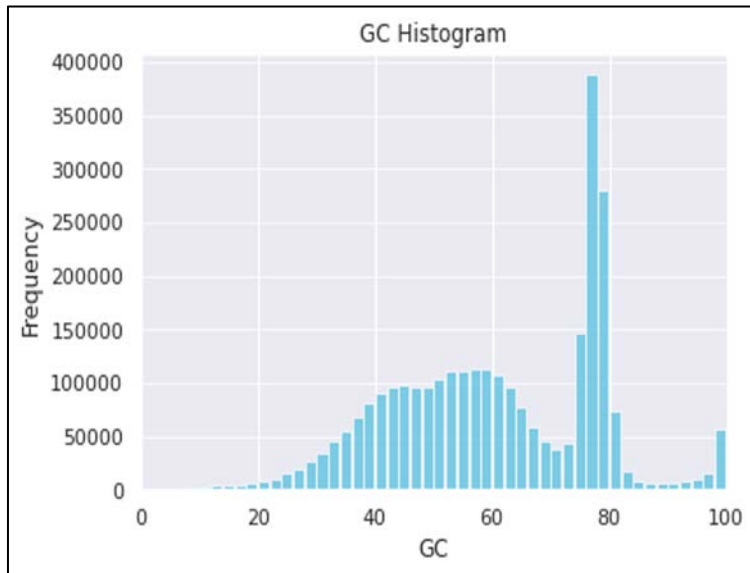
Workers

Computer #1

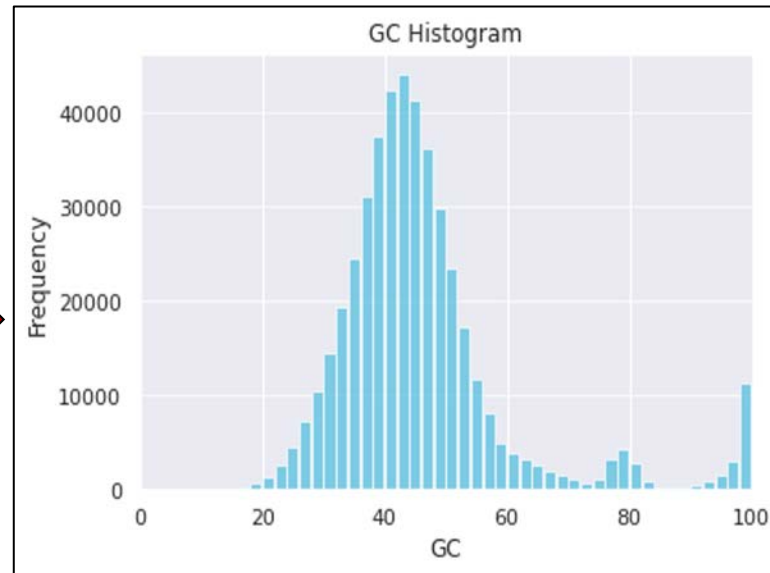
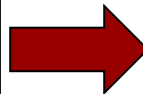


*reads one line at a time
processes one test at time

Big File: HCTUP2.fastq.gz



Raw Sequence Data



Filtered Sequence Data

Test_file:

@H100 read 1: ACCATCTC
@H100 read 2: CATGCATG
@H101 read 1: TTCGAGTC

...

@H5000 read 2: CATCATGA



Spark Data Frame

Index	DNA Sequence	Quality	Q Test
@H100 read 1	ACCATCTC.. .	FFFF:;,;,; ...	PASS
@H100 read 2	CATGCATG. ..	FF,FF:FF; F...	PASS
@H101 read 1	TTCGAGTC ...	FF::FF;;, ...	FAIL
...			

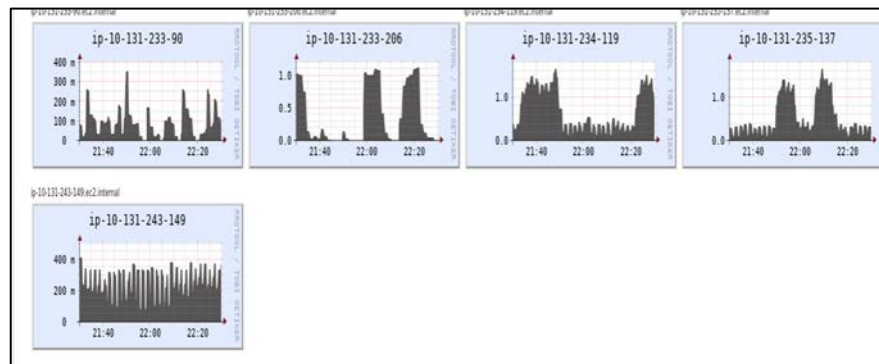
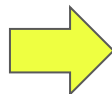
quality test



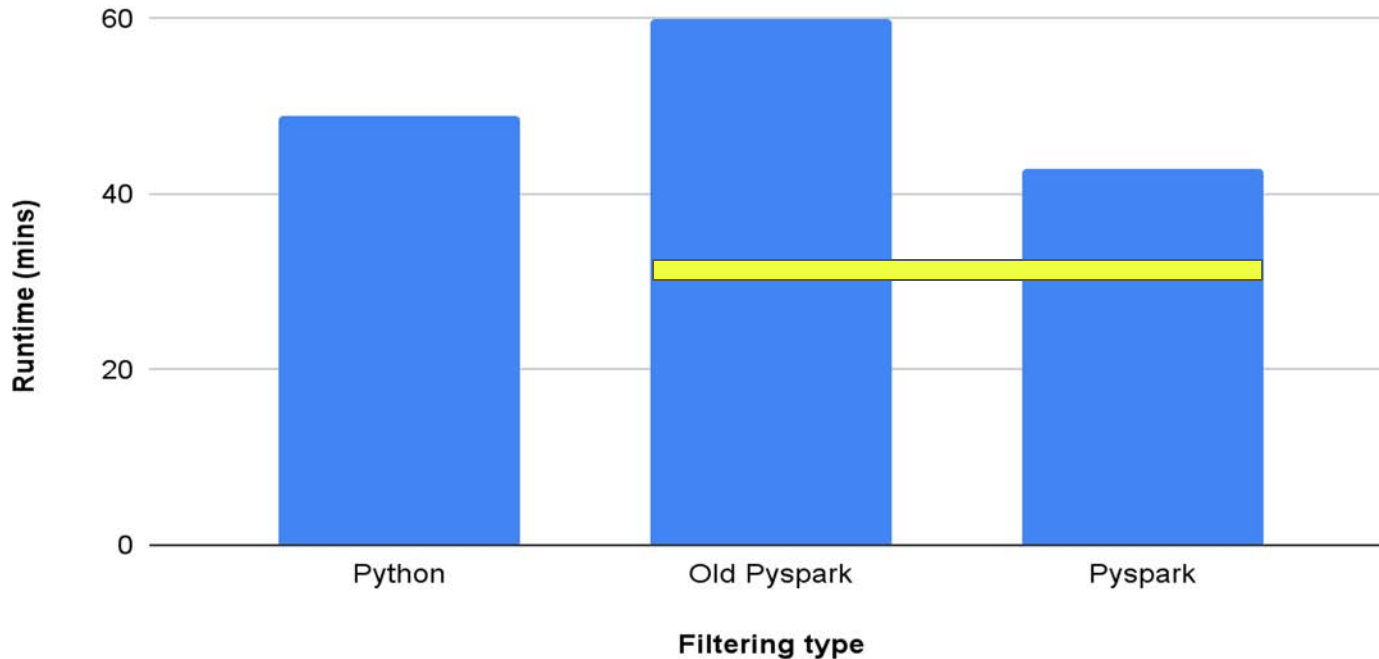
n test

kmer test

Example of tasks
being distributed
between all available
CPUs



Comparison of Filtering Programs



What happened with Scala?



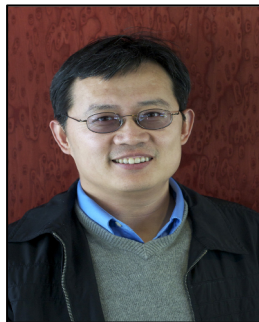
ERROR: Failed with Error
...ClassNotFound



Pipeline Improvements

- Improvement in how contamination filtering done
- Research how to make the contamination reference file quick to load for each run.
- Convert more user defined functions into PySpark functions
- PySpark filter pipeline write to local disk

Acknowledgements:



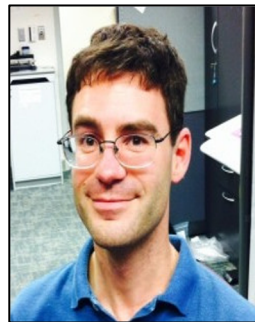
Dr. Zhong Wang

Department of
Energy Joint
Genome Institute



Bryce Foster

Department of
Energy Joint
Genome Institute



Brian Bushnell

Department of
Energy Joint
Genome Institute



Chen Zhang

Shanghai University
(Graduate Student)

THANK YOU TO OUR SPONSORS



THANK YOU FOR LISTENING

QUESTIONS? COMMENTS?