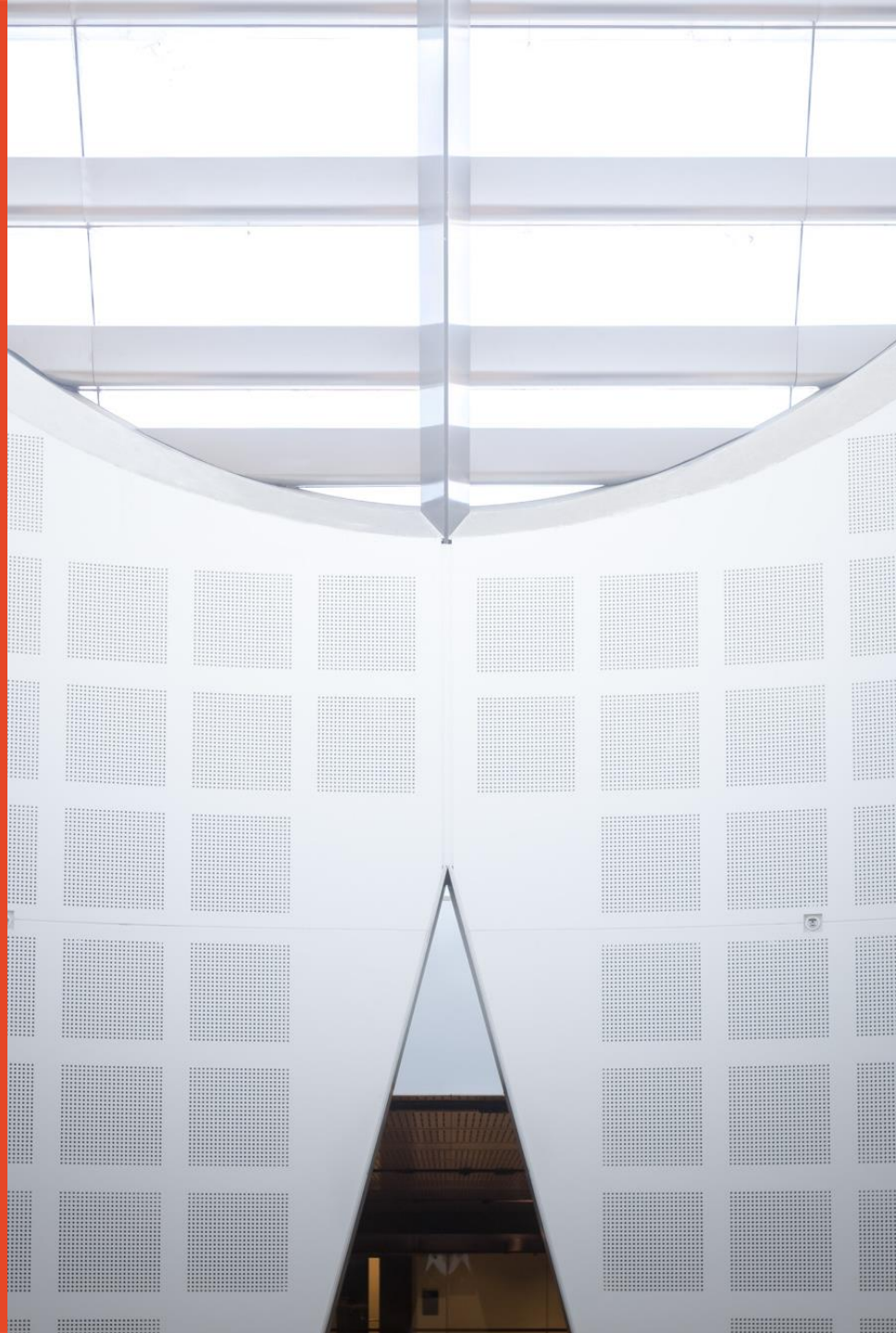
A Cloud-Based Testing Framework for Genomic Medicine Software

Presented by
Michael Troup
For fulfilment of BCST (Adv)(Hons)



THE UNIVERSITY OF SYDNEY

Acknowledgement

Sydney University Supervisor

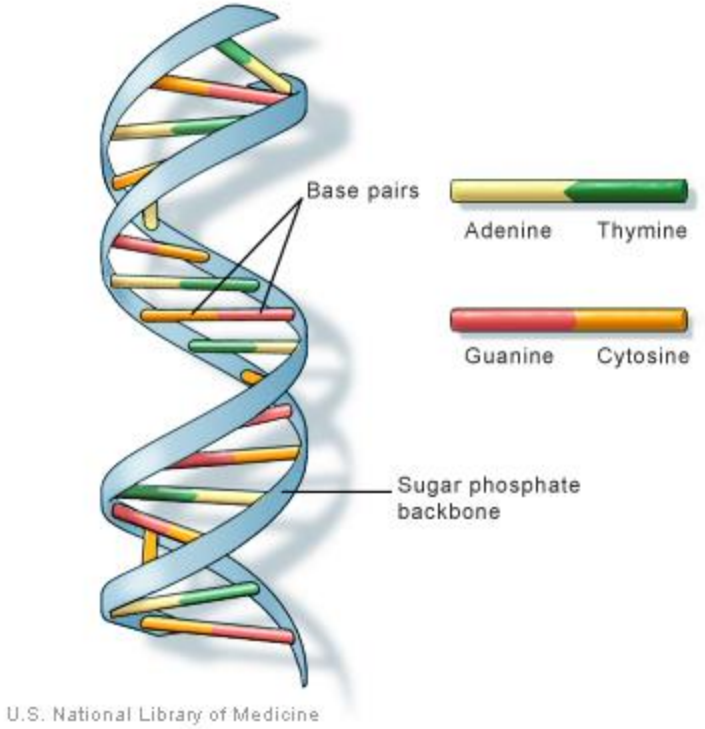
– Associate professor Bing Zhou

This research is undertaken on behalf of the Victor Chang Cardiac Research Institute

– Dr Joshua Ho

Motivation

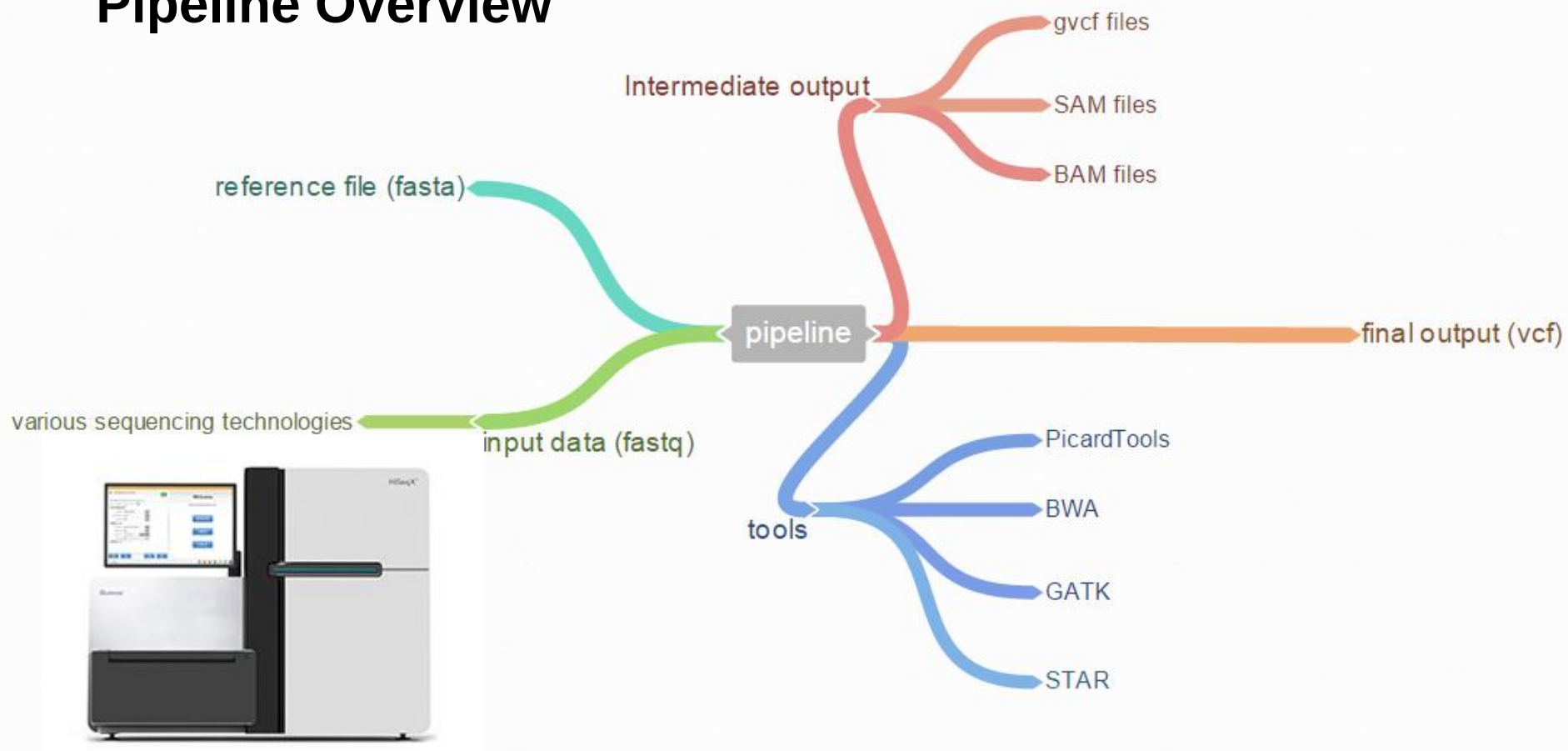
- DNA – ~3 billion base pairs
- Genomic sequencing
- Variants
- Uses of variant calling



[1]

Motivation

Pipeline Overview

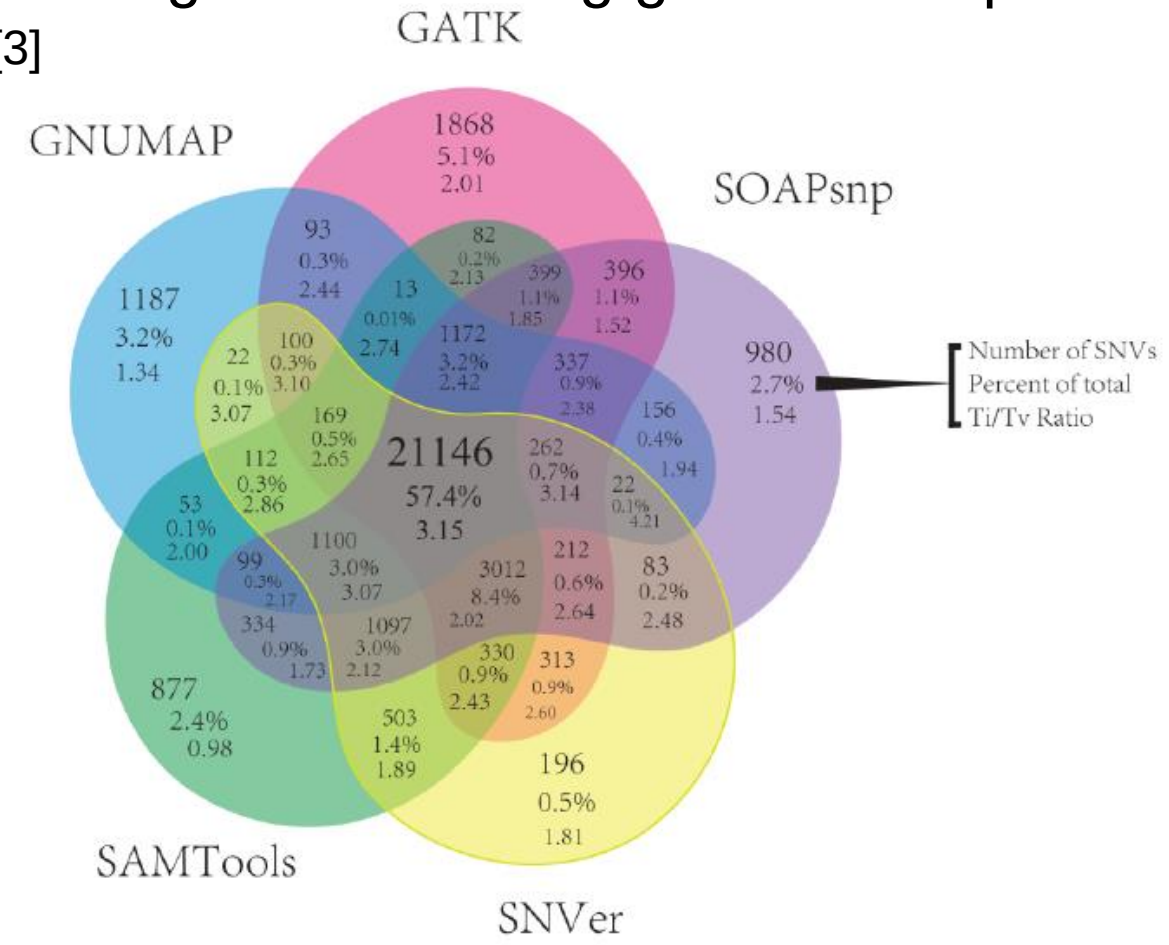


Sequencing Machine [2]

Motivation

Low concordance

– Between existing variant-calling genomic sequencing pipelines [3]



Motivation

Pipeline Testing Difficulties

- Size of input data
- Diverse components of a pipeline
  - Alignment phase – string matching
  - Variant calling – machine learning / classification
- There are many different ways to construct a pipeline
  - Open-source
  - Write your own
  - Many different offerings
- There is no easy way to decide if a pipeline has given the correct result
  - Oracle problem
  - False Negative results particularly difficult to detect

Literature

Traditional Testing Methodologies

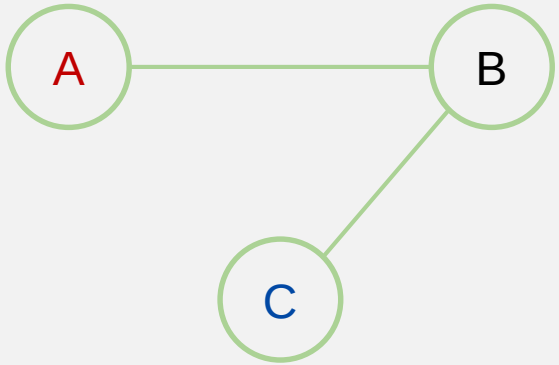
- Use reference input & compare with “gold standard”
  - US National Institute of Standards and Technology
  - Genome in a Bottle Consortium
  - Some sequencing manufacturers also provide Gold Standard
- Small results: Sanger Sequencing
  - Disadvantage: cost
- Simulated Data
  - Open-source software available
  - Produces simulated input reads
  - Also produces a “truth” output value

Literature

Metamorphic Testing

- Examine outputs from multiple related inputs
- Metamorphic Relation (MR)
- Applications in machine learning, web services, and bioinformatics

Example: Graph theory - length shortest path



MR1: SP(A,C) = SP(C,A)

Literature

Other testing frameworks

- Giannoulatou [4]
  - Metamorphic Relations for alignment part of pipeline
  - Not cloud-based or built for large data
- A number of cloud-based pipelines
  - For analysis – not testing
- Hignam – GCAT [5]
  - Web-based tool
  - Upload own results
  - Compare with gold standard & others
- General absence in the literature of pipeline testing frameworks

Research Objectives

For variant-calling genetic sequencing pipelines:

- Provide an automated, cloud-based testing framework
  - Using state-of-the-art testing techniques
  - Minimise the technical exposure to the user
  - Able to handle large-scale data
- Apply the framework to an industry-standard pipeline

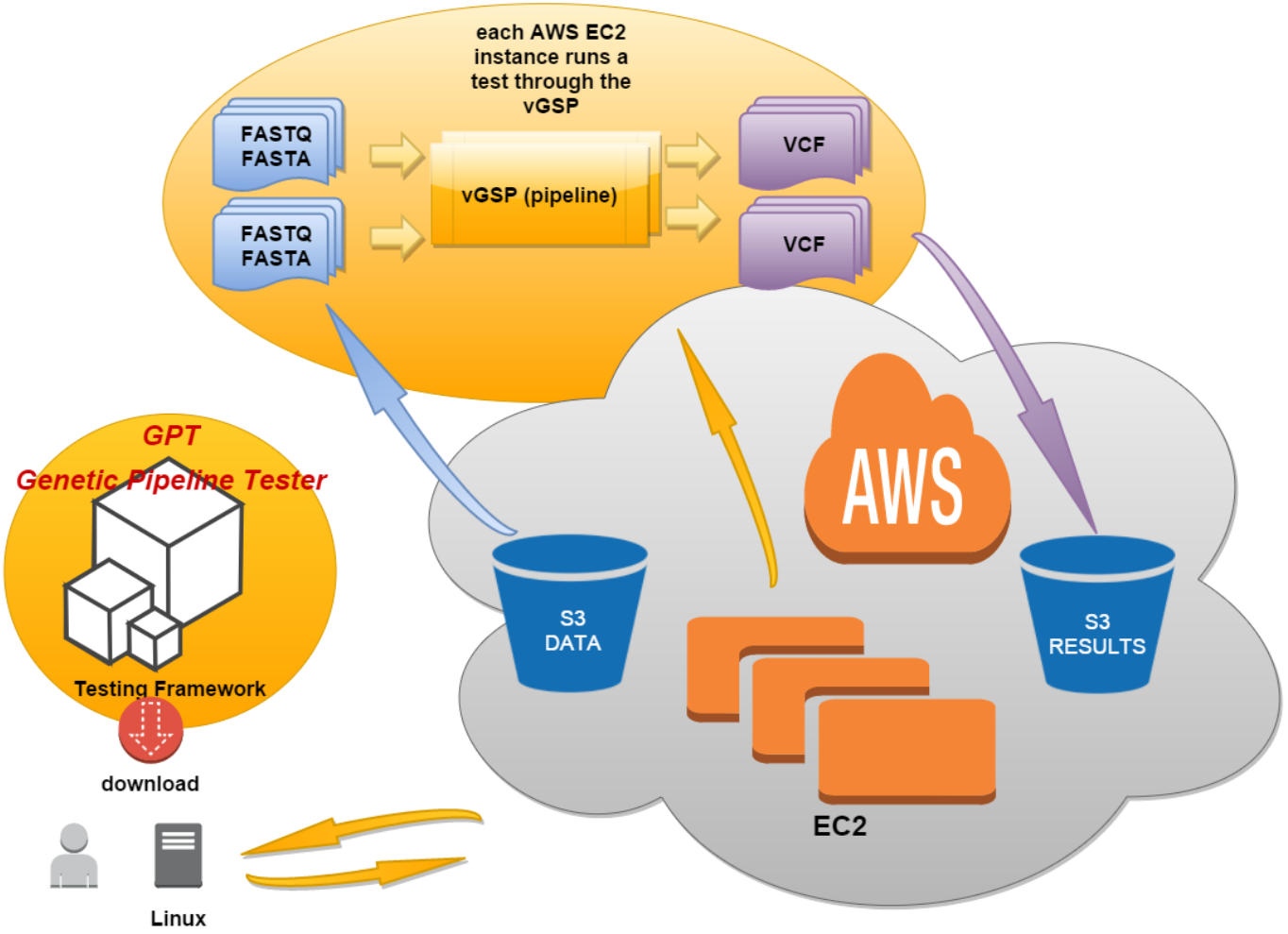
Solution

- Framework downloaded to local resource
  - called GPT – Genomic Pipeline Tester
  - Tests run on AWS cloud resources
  - Analysis of results
  - Reporting
- Fills in a configuration file
  - Data file names
  - Names of pipeline files – execute & install
  - AWS details: number & type of instances, spot price

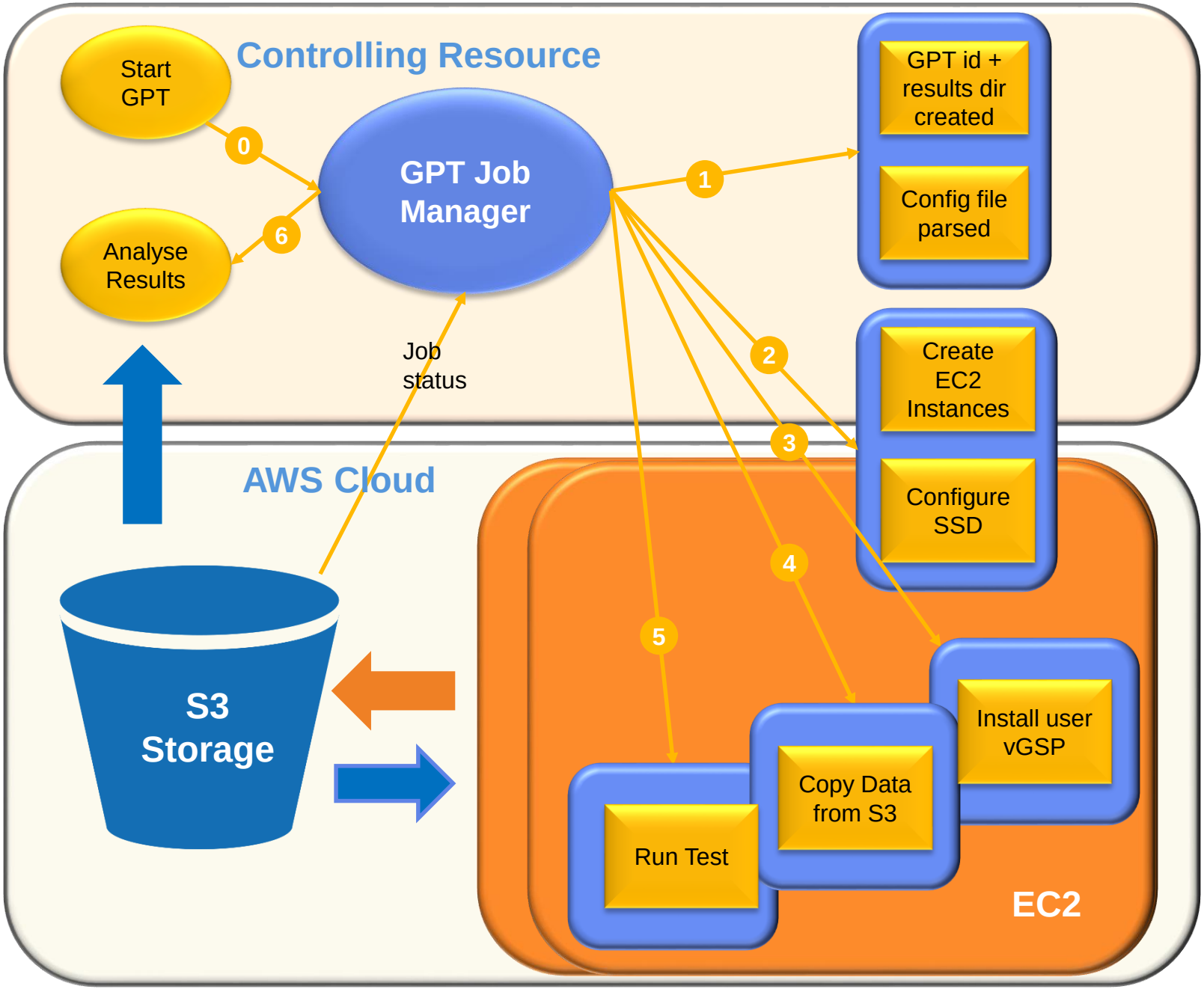
```
[aws]
region=us-west-2

[aws-instances]
#instance-type=r3.large
instance-type=c4.2xlarge
count=1
user-data=gpt-system/user-data-ssd.sh
```

Solution - GPT



GPT Workflow



Solution

Tests

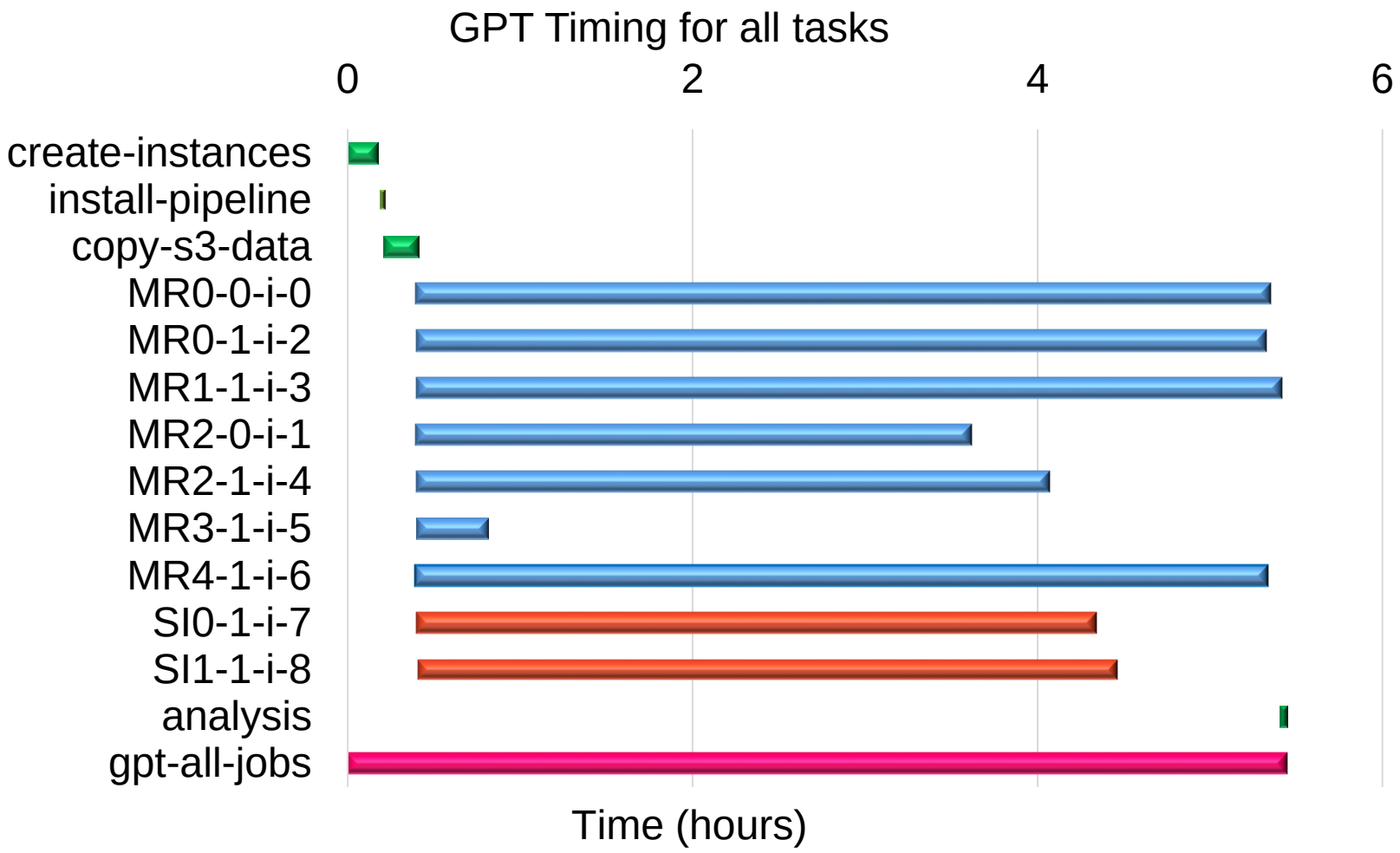
Test	Description
MR0	Deterministic Output
MR1	random permutation of input
MR2	duplication of reads
MR3	unmapped reads
MR4	mapped reads
SI0	simulated reads – no mutations
SI1	simulated reads – mutations

Results

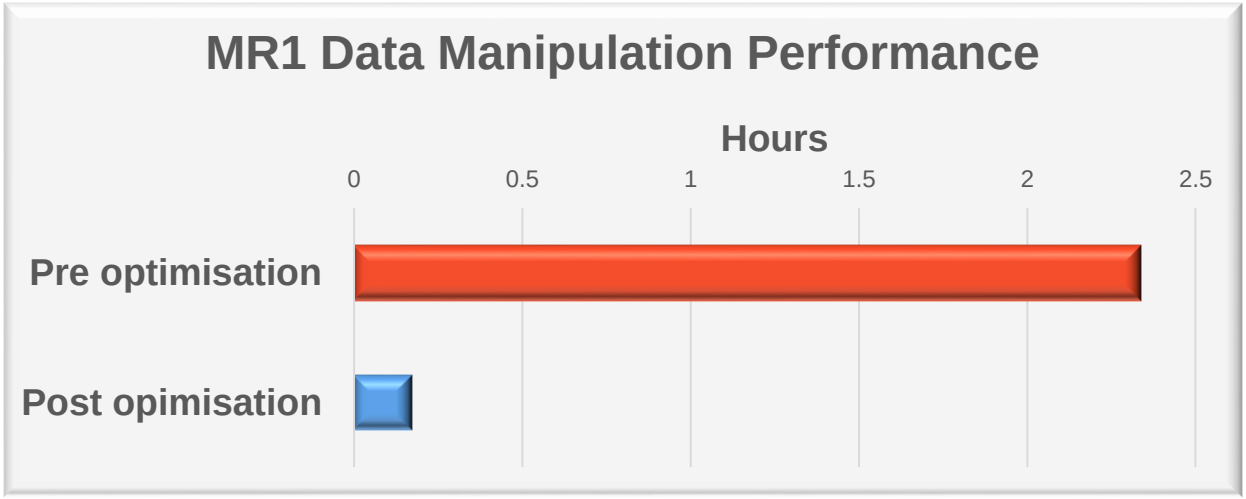
Test Configuration

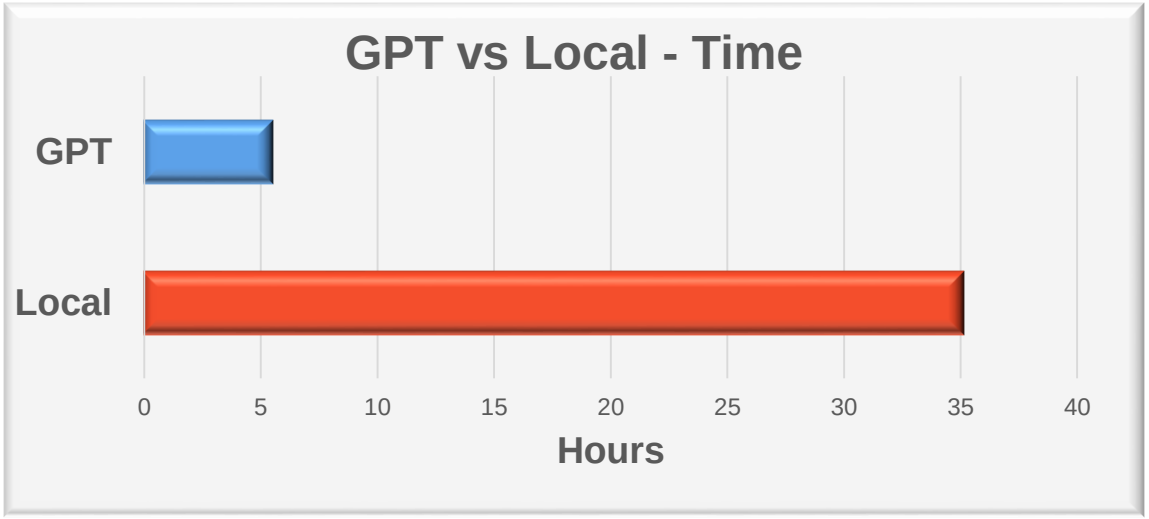
Category	Description
Input Data	Real input read sequence data Reference UCSC hg19 Simulated read data - chr11 of hg19 **Total input file size ~30GB**
AWS	EC2 Instance type: c3.8xlarge (32 CPU) Instance count: 9
Pipeline Under Test	BWA, SAMTools, PicardTools GATK

Results


GPT Timing for all tasks
0
2
4
6
create-instances
install-pipeline
copy-s3-data
MR0-0-i-0
MR0-1-i-2
MR1-1-i-3
MR2-0-i-1
MR2-1-i-4
MR3-1-i-5
MR4-1-i-6
SI0-1-i-7
SI1-1-i-8
analysis
gpt-all-jobs
Time (hours)

Results


MR1 Data Manipulation Performance
Hours
0
0.5
1
1.5
2
2.5
Pre optimisation
Post opimisation


GPT vs Local - Time
GPT
Local
0
5
10
15
20
25
30
35
40
Hours

Results

Cost: on-demand vs spot

- 9 x c3.8xlarge instances for 6 hours

On-Demand	Spot
$90.72	$21.60

- 76% saving using spot instances

Results

Test Statuses

Test Name	Status
MR0	Passed *
MR1	**Failed**
MR2	**Failed**
MR3	Passed
MR4	Passed
SI0	Passed
SI1	**Failed**

* Variants called the same but other information different

Results

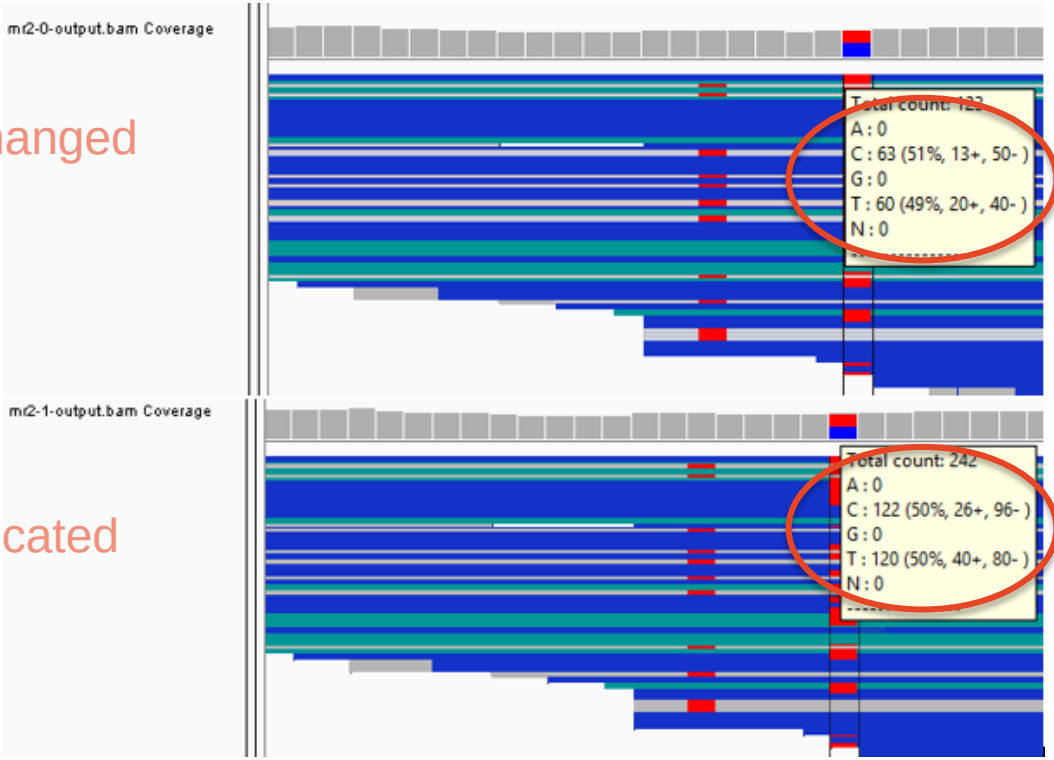
MR0 – Deterministic output

- Variants called the same
- 3% of variants called had different values for:
  - “Variant Confidence/Quality by Depth”
- In a particular example this value differed by 37%
- Could make a difference in borderline cases

Results

MR2 – Duplication of reads

- False negative rate of 23 per 100,000
- Discovered important example in protein coding region



Unchanged data

51% C, **49**% T

Duplicated data

50% C, **50**% T

Discussion

- Metamorphic Relations useful to deal with oracle problem
  - Useful results without gold standard
  - Identified false negatives
- Framework handles large data
- Reduces barriers to testing
- Framework overhead high for small data
- Configuration still requires some technical exposure

Future Work

- Add a browser-based interface
- Complete a larger study of pipelines
- Platform independence
- Develop more involved metamorphic relations
- Build a classifier
  - Identify failure-causing characteristics

Conclusion

- Need to improve software testing in genomic medicine
  - Reliance on domain testing with “gold-standards”
- First cloud-based fully self-contained pipeline testing framework
  - Allows testing on large scale data
  - Test real data without gold-standard
  - Applies Metamorphic Relations to whole pipeline
  - Created a new Metamorphic Relation: deterministic output
  - Combines traditional and metamorphic testing
  - Reduces technical and financial barriers
- Applied framework to an industry-standard pipeline
- Good scope for future work

Thank You

References

[1] “What is DNA?,” Genetics Home Reference, 12-Oct-2015. [Online]. Available: http://ghr.nlm.nih.gov/handbook/basics/dna. [Accessed: 19-Oct-2015].

[2] Illumina, “HiSeq X Series of Sequencing Systems.” [Online]. Available: http://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet-hiseq-x-ten.pdf. [Accessed: 21-Oct-2015].

[3] J. O’Rawe, et al., “Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing,” *Genome Med*., vol. 5, no. 3, p. 28, Mar. 2013.

[4] E. Giannoulatou, et al., “Verification and validation of bioinformatics software without a gold standard: a case study of BWA and Bowtie,” *BMC Bioinformatics*, vol. 15, no. Suppl 16, p. S15, Dec. 2014.

[5] G. Highnam, et al., “An analytical framework for optimizing variant discovery from personal genomes,” *Nat. Commun*., vol. 6, Feb. 2015.