High Throughput Genomics: parallel read mapping and variant calling

Open Science Grid User School
July 29, 2016
University of Wisconsin - Madison, WI

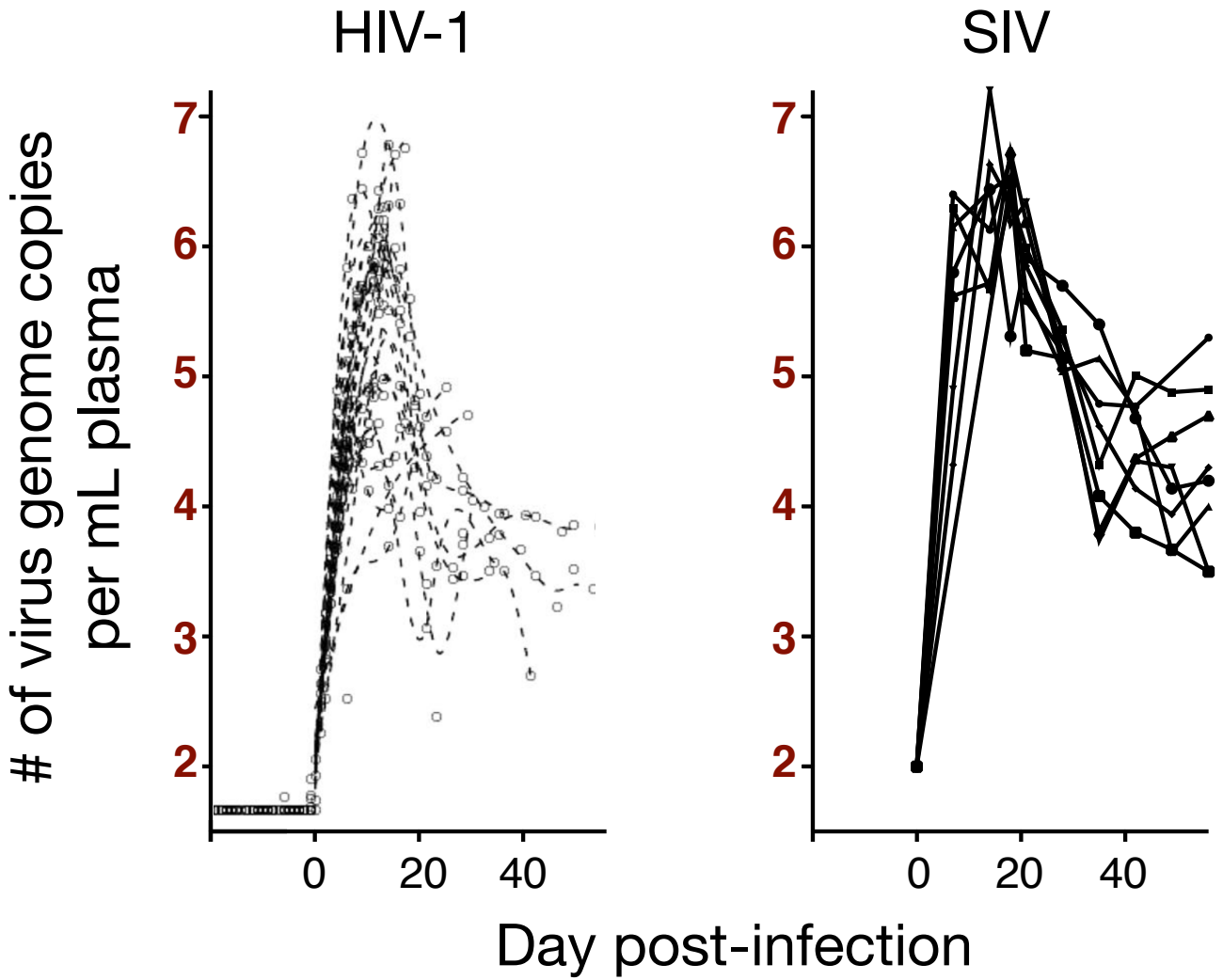
Outline

- Brief introduction into genomics
- Data processing bottleneck
- High-throughput computing as a solution

Introduction

- Laboratory of David O'Connor, Department of Pathology and Laboratory Medicine
- Macaque genomics:
  - Immune gene discovery/ genotyping
  - Disease kinetics of HIV/ SIV, Zika, & other viruses
- 2015 OSG user school student

Macaques show similar disease kinetics to humans

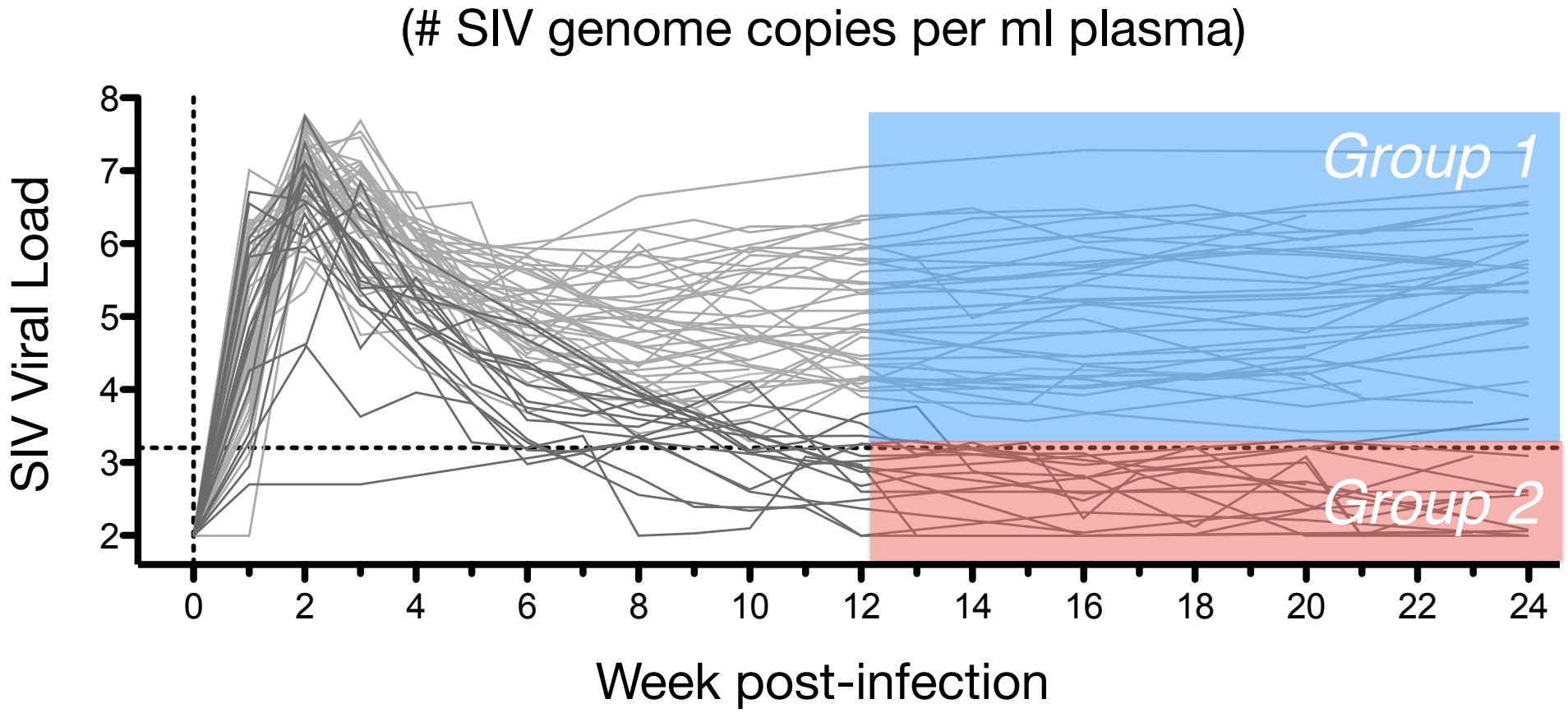


Cynomolgus macaque (island of Mauritius)

(Stacey et al. JV, 2009)

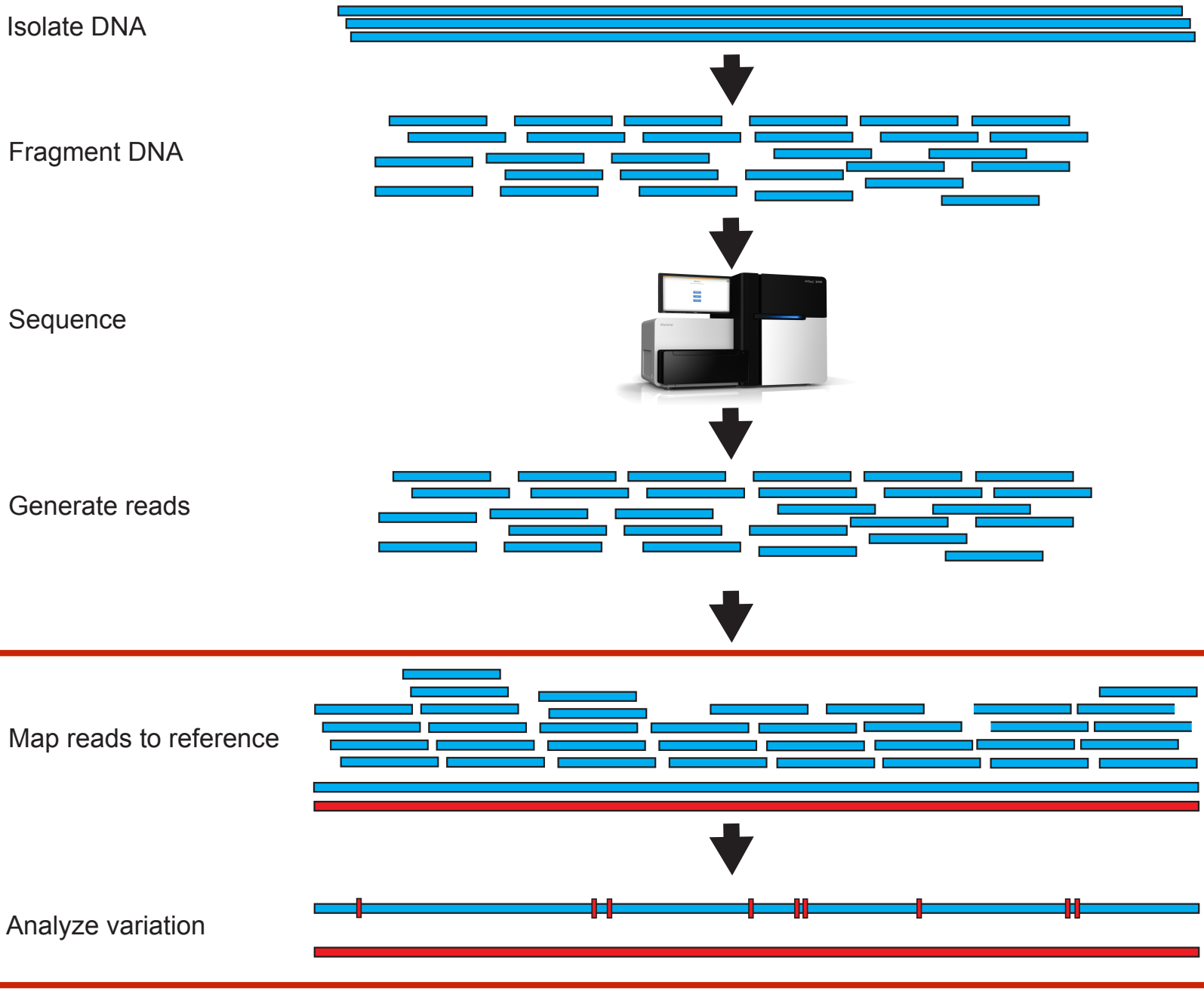
(y-axis is Log_{10} transformed)

Longitudinal data segregate phenotypic groups



(y-axis is Log_{10} transformed)

Whole Genome Sequencing



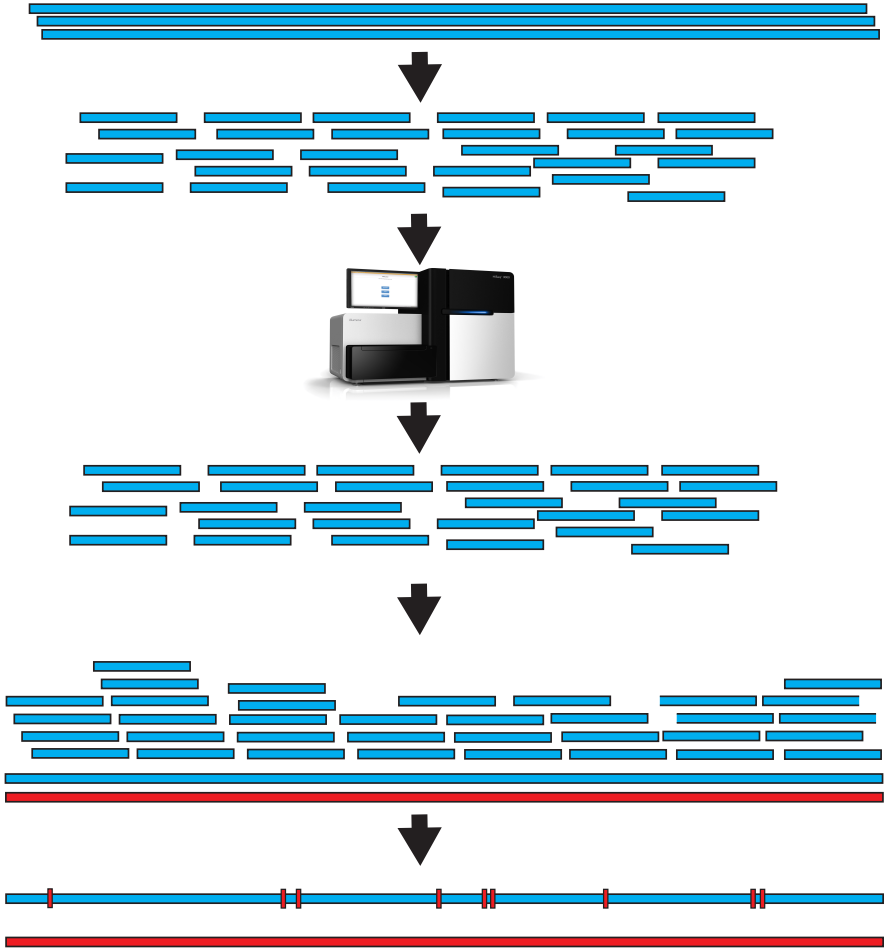
The Problem

- Genomics datasets are rapidly increasing in size and complex - more and more information at lower costs
- Whole Genome reads ~ 120 GB per sample
- Strength of associations increases with more samples
- Processing one dataset on a single high end machine takes 2-4 days

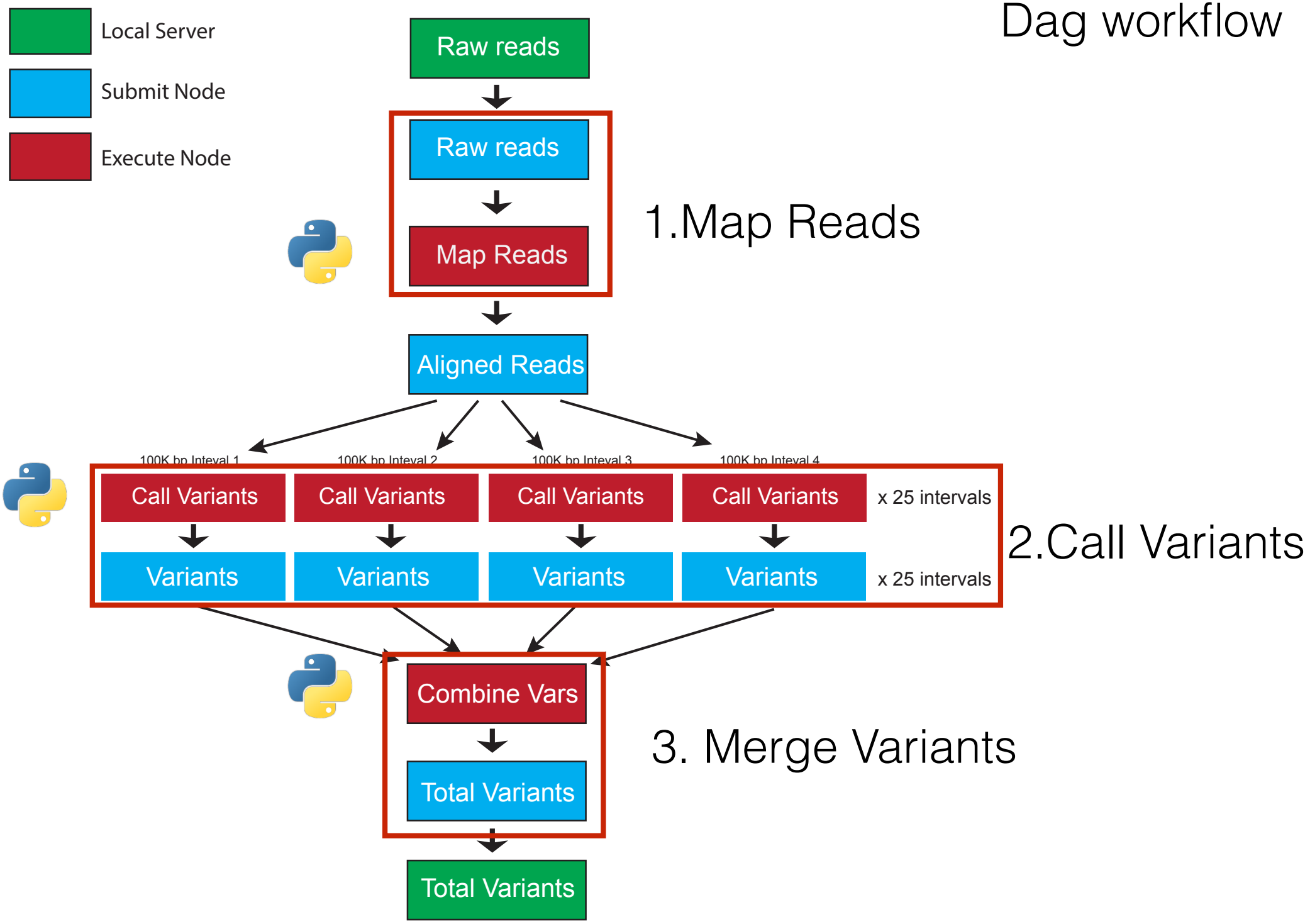
The Solution

- Next Generation "shutgun" sequencing is a high throughput approach to processing an extremely large dataset

- Apply the same approach to re-constructing the genome sequence from reads



Dag workflow



1. Map reads to reference (BWA MEM)

- runs on single execute node
- request 8CPU, 16GB RAM, 20GB Disk
- WGS: 18-24 hrs

```
CY0164

    5922659328  -  Run Bytes Sent By Job
    9974532  -  Run Bytes Received By Job
    5922659328  -  Total Bytes Sent By Job
    9974532  -  Total Bytes Received By Job
    Partitionable Resources :    Usage   Request Allocated
       Cpus                 :                8         8
       Disk (KB)            : 12909024 209715200 212300286
       Memory (MB)          :    14281     16384     16384

CY0165

    5510552576  -  Run Bytes Sent By Job
    9974532  -  Run Bytes Received By Job
    5510552576  -  Total Bytes Sent By Job
    9974532  -  Total Bytes Received By Job
    Partitionable Resources :    Usage   Request Allocated
       Cpus                 :                8         8
       Disk (KB)            : 5391158 209715200 212317578
       Memory (MB)          :    14229     16384     16384

CY0166

    8003021312  -  Run Bytes Sent By Job
    9974525  -  Run Bytes Received By Job
    8003021312  -  Total Bytes Sent By Job
    9974525  -  Total Bytes Received By Job
    Partitionable Resources :    Usage   Request Allocated
       Cpus                 :                8         8
       Disk (KB)            : 12572118 209715200 212300286
       Memory (MB)          :     9622     16384     16384
```

2. Call variants for intervals (GATK haplotypecaller)

- request 2CPU, 16GB RAM, 2GB Disk
- ~18-24 hrs per sample

```
CY0165-hc1

    673367936  -  Run Bytes Sent By Job
    12949015  -  Run Bytes Received By Job
    673367936  -  Total Bytes Sent By Job
    12949015  -  Total Bytes Received By Job
    Partitionable Resources :    Usage   Request Allocated
       Cpus                 :                1         1
       Disk (KB)            :   674223 209715200 222320761
       Memory (MB)          :    11654     16384     16384

CY0165-hc2

    942119040  -  Run Bytes Sent By Job
    12949015  -  Run Bytes Received By Job
    942119040  -  Total Bytes Sent By Job
    12949015  -  Total Bytes Received By Job
    Partitionable Resources :    Usage   Request Allocated
       Cpus                 :                1         1
       Disk (KB)            :   935924 209715200 210835104
       Memory (MB)          :    11726     16384     16384
```

3. Merge Variants (GATK CatVariants)

- runs on single execute node
- request 1CPU, 1GB RAM, 5GB disk
- > 3 min runtime

```
CY0164

    1146748160  -  Run Bytes Sent By Job
    12948765  -  Run Bytes Received By Job
    1146748160  -  Total Bytes Sent By Job
    12948765  -  Total Bytes Received By Job
    Partitionable Resources :    Usage   Request Allocated
       Cpus                 :                1         1
       Disk (KB)            :  1132534 209715200 215233128
       Memory (MB)          :      196     16384     16384

CY0165

    1615156352  -  Run Bytes Sent By Job
    12948765  -  Run Bytes Received By Job
    1615156352  -  Total Bytes Sent By Job
    12948765  -  Total Bytes Received By Job
    Partitionable Resources :    Usage   Request Allocated
       Cpus                 :                1         1
       Disk (KB)            :  1589964 209715200 340811632
       Memory (MB)          :        2     16384     16384

CY0166

    1274070272  -  Run Bytes Sent By Job
    12948765  -  Run Bytes Received By Job
    1274070272  -  Total Bytes Sent By Job
    12948765  -  Total Bytes Received By Job
    Partitionable Resources :    Usage   Request Allocated
       Cpus                 :                1         1
       Disk (KB)            :  1256872 209715200 213003620
       Memory (MB)          :      228     16384     16384
```

Runtime Comparison

High Throughput

- runtime: 32-48 hrs per sample
  - map reads: ~20 hrs
  - call var: ~ 24 hrs
  - merge: 3 min

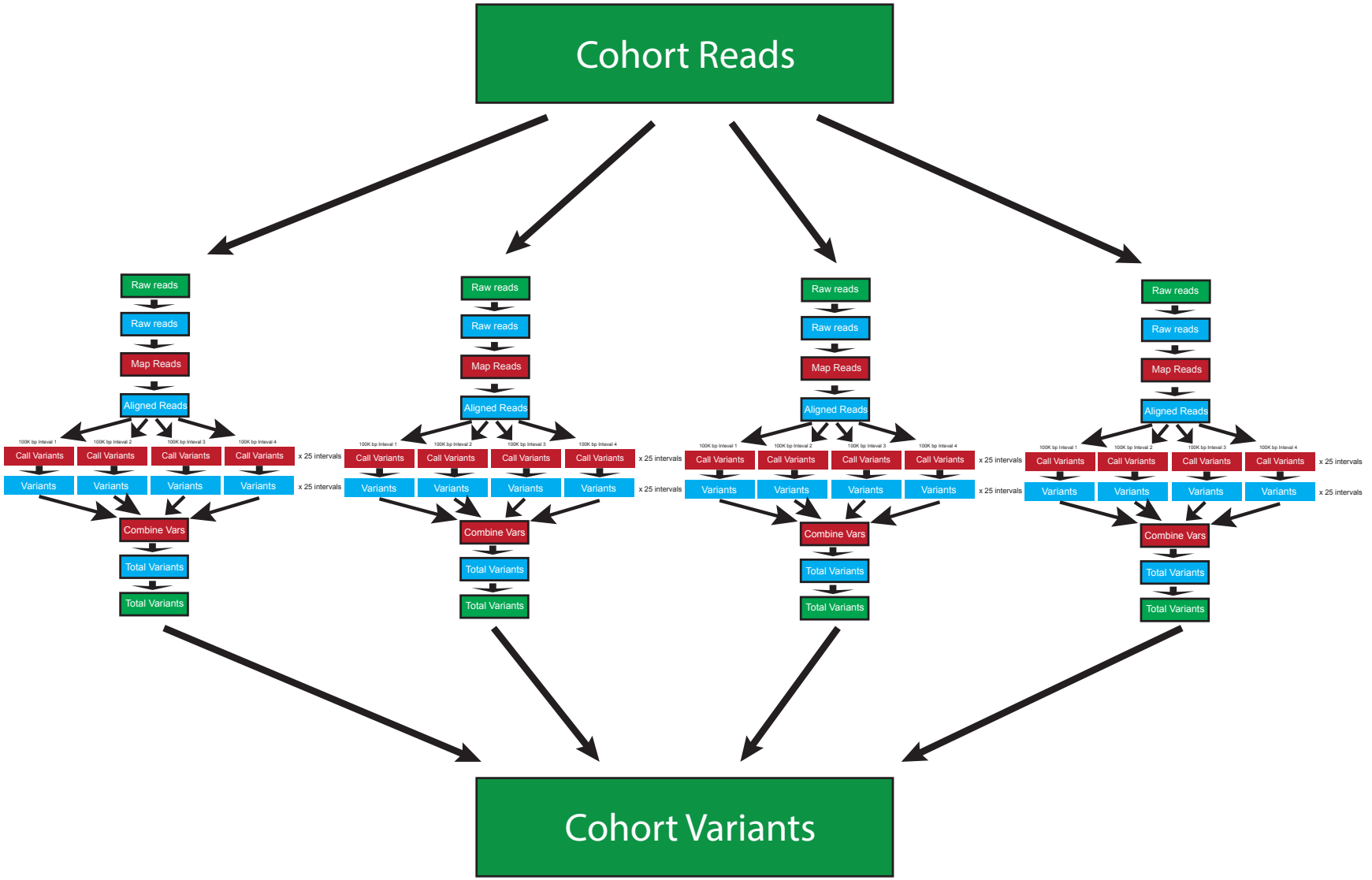
That's it…?

Single Machine

- runtime: 2-4 days per sample
  - map reads: ~20 hrs
  - call var: ~ 48 hrs
  - merge: 3 min

lol nope

Automate everything!

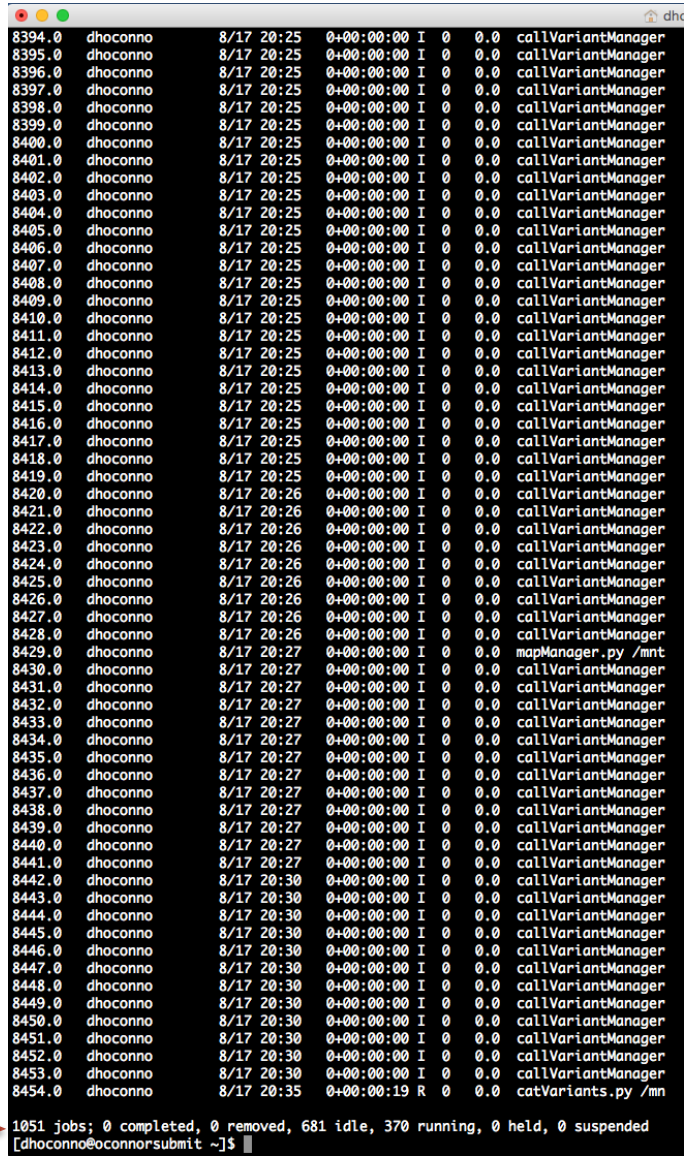


x N samples

Python wrapper

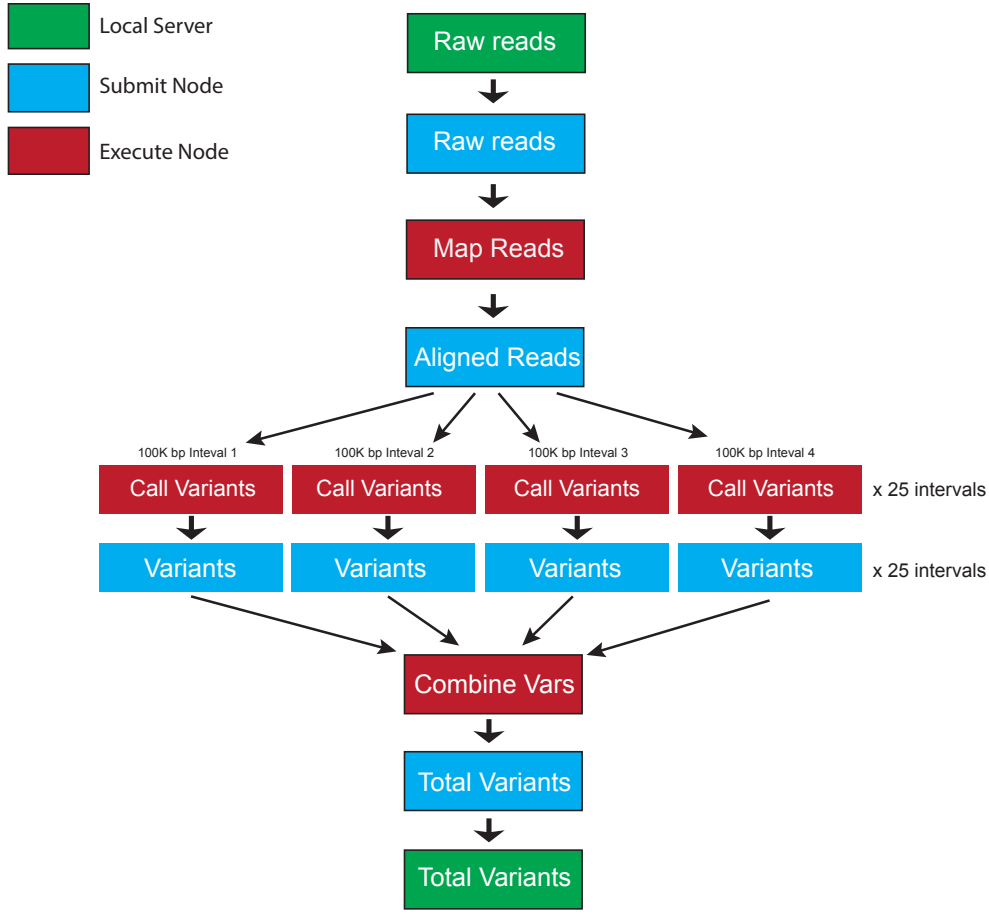
- Creates submit and dag files for each dataset staged in specified directory
- submits all jobs at once
- 174 datasets processed in > 2 weeks
- Would have taken around 1 year @ 48hrs per dataset

1051 jobs at once



Caveats

- We require a shared filesystem (Gluster) to avoid data transfer overhead (reads, reference file, ect.)
- Variant calling may still benefit from HTC approach regardless of data transfer time
- Not all processes can be split



Conclusions

- Divide and conquer where possible
- Automate, automate, automate
- Processed ~ 60TB of data in under a month
    - (174 datasets against 4 references)
    - Would have taken ~ 4 years on local servers

Acknowledgments

OC Lab Genomics Group

- David O'Connor
- Mike Graham
- Adam Ericsen
- Roger Wiseman

CENTER FOR HIGH THROUGHPUT COMPUTING

- Lauren Michael
- Christina Koch

THE UNIVERSITY of WISCONSIN MADISON

Wisconsin National Primate Research Center
UNIVERSITY OF WISCONSIN–MADISON