HIGH-PERFORMANCE BIOLOGICAL COMPUTING
University of Illinois at Urbana Champaign

Instrumenting
Human Variant Calling Workflow

Liudmila Sergeevna Mainzer
Blue Waters Symposium
May 11-13, 2015

CompGen Initiative at UIUC

INSTITUTE FOR GENOMIC BIOLOGY

Victor Jongeneel,
Director of HPCBio

CSL: COORDINATED SCIENCE LAB

Ravi Iyer,
Professor of ECE

- Architecture:

  What kind of computer architecture is best suited for bioinformatics work?

- Performance bottlenecks:

  What are the performance bottlenecks for bioinformatics work, on different architectures?

- Future:

  How to structure the bioinformatics workflows for best performance on the architectures upcoming in the next 1, 3, 5 years?

Presentation Plan

Part 1: motivation and context

What is variant calling and why it is important

Part 2: work in progress

Computational challenges in variant calling

Part 3: outlook

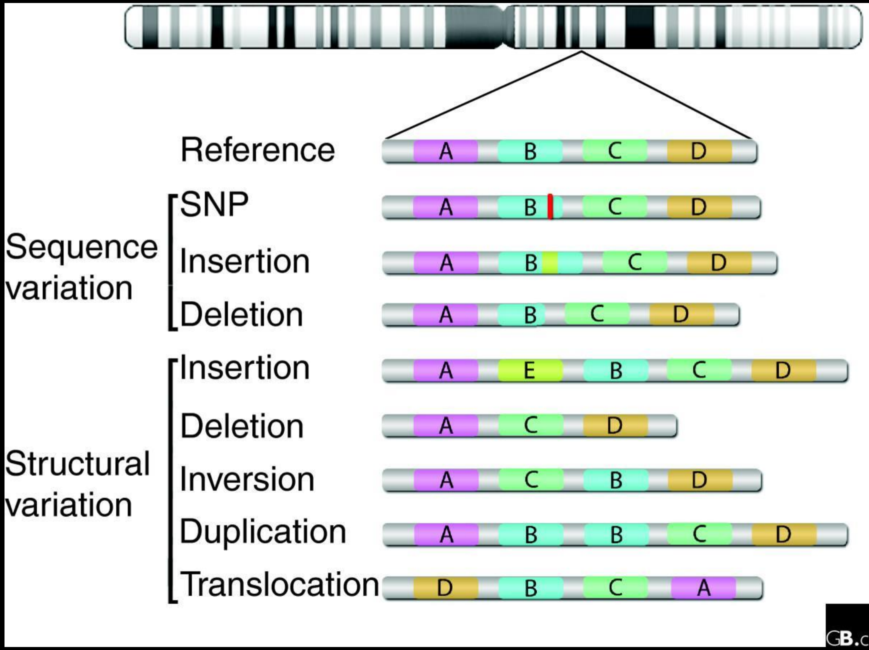
alternative solutions and potential production cases

Part 1:

What is Genomic Variant Calling and why we think it is important

Genomic Variant = a difference in the genetic code

```
goodnightgoodnightpartingissuchsweetsorrow
       htg-odnigh                Oetsorro
    nightg-od                  swOetsorr
 oodnightg     ghtpartingi  uchswOets
Goodnigh     nightparti  issuchswO
         g-odnightp
      ghtg-odnig
   dnightg-od
```



Rahim *et al. Genome Biology* 2008 **9**:215

Genomic Variation can affect phenotype

Mexican corn varieties. Imgarcade.com

A blond-haired Solomon Island child;
Credit: © Sean Myles

Cystic fibrosis
Sickle cell anemia
Huntington disease
Color blindness
Bloom’s syndrome
Down’s syndrome
Haemophilia
Cancer

Purebreddairycattle.com

Red & White
Holstein
Jersey
Milking Shorthorn
Ayrshire
Brown Swiss
Guernsey

How are genomic variants identified?

```
goodnightgoodnightpartingissuchsweetsorrow
       htg-odnigh                Oetsorro
    nightg-od                  swOetsorr
 oodnightg     ghtpartingi  uchswOets
Goodnigh     nightparti  issuchswO
         g-odnightp
      ghtg-odnig
   dnightg-od
```

1. Well studied diseases with known variants
   - "Run a panel"
   - Which of the known variants are present in this individual?
   - Not a computational challenge

2. Recalcitrant cancers, uncharacterized and rare diseases
   - Identify variants de-novo
   - Whole genome sequencing
   - Whole exome sequencing
   - Could be a computational challenge: variant calling workflow
   - Few days – 2 weeks on a small cluster

Neither of these are good cases for Blue Waters. What is?

Obama announces Precision Medicine Initiative

" to bring us closer to curing diseases like cancer and diabetes – and to give all of us access to the personalized information we need to keep ourselves and our families healthier."

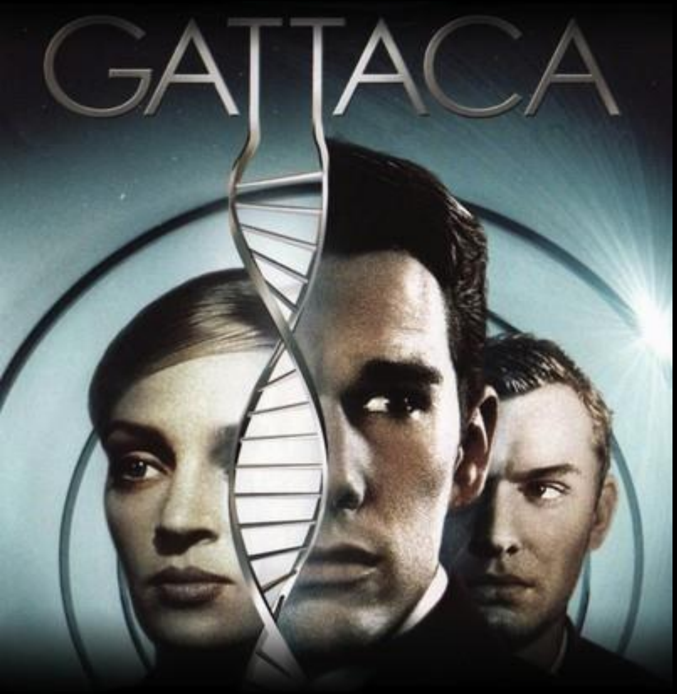
"I want the country that eliminated polio and mapped the human genome to lead a new era of medicine – one that delivers the right treatment at the right time,"

U.S. President Barack Obama delivers his State of the Union address to a joint session of the U.S. Congress on Capitol Hill in Washington, January 20, 2015. Reuters/Jonathan Ernst

NIH http://www.nih.gov/precisionmedicine/

Precision medicine is an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person.

Variant Calling: hypothetical case



What if we had to genotype every baby being born?
= 500 genomes/day in the state of Illinois

NERVE CONDITION - PROBABILITY 60%,
MANIC DEPRESSION - 42%,
OBESITY - 66%,
ATTENTION DEFICIT DISORDER - 89%
HEART DISORDER - 99%
EARLY FATAL POTENTIAL
LIFE EXPECTANCY - 33 YEARS

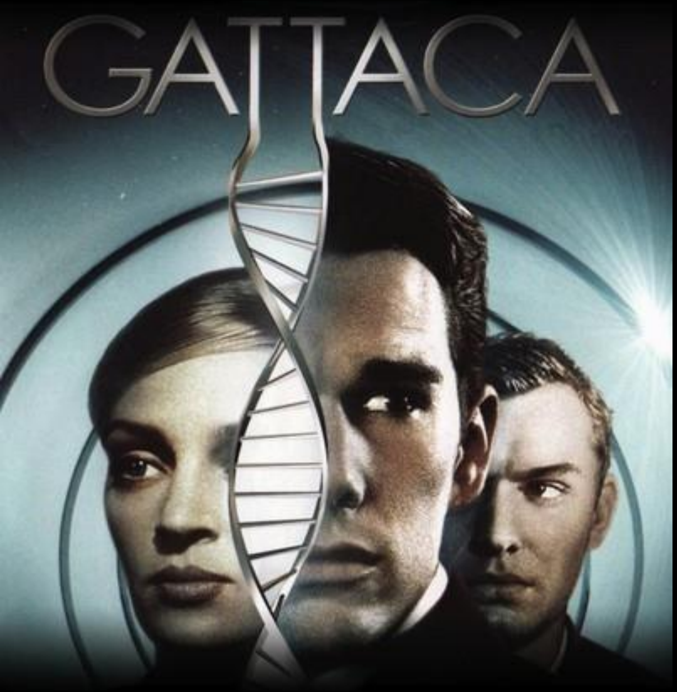
Image from http://fcw.com

Part 2:

Computational challenges in sustained high-throughput genomic variant calling

Big data, big compute on a sustained basis

Genotyping every baby being born? 500 genomes/day in the state of Illinois result in:



Input

- 300-600 GB/genome
- 150-300 TB/day
- 2 files/genome = 1000 files

Intermediary

- 1-3 TB per sample
- 0.3-1.5 PB/day total
- 525 files/sample = 262,500 files total

Output

- < 500 M per sample
- 26 files/sample = 13,000 files total

Computational cost

- 100,000 – 300,000 node-hours per day

Why need Blue Waters? … and the BW team!

What kind of facility will be able to sustain this kind of throughput?

Our goals on Blue Waters:

- Set up workflow
- Prove function on test cases
- Demonstrate readiness for high throughput
- Profile performance
- Determine and eliminate bottlenecks
- Make recommendations for a computational facility appropriate for genomic variant calling, for the future

Kinds of challenges

1. Large total data footprint
2. Large number of files

Data Management

3. Large number of simultaneous but independent non-mpi computations
4. Keeping track of what was done to the data: large amount of Metadata
5. Workflow bottlenecks: fans and merges, followed by fans

Workflow management

Data management

Incoming data

- auto-md5
- auto-archive
- stream directly into the workflow

Output data

- auto-check for correctness at every step
- auto-archive during/after computation
- auto-stream to the recipient

Solved problems
in some other areas of science;

hope to learn, borrow and adapt solutions

Identifying potential i/o bottlenecks

- uneven file distribution
- simultaneous file access
- saturating i/o in certain steps of the workflow
- impact on metadata servers

Have done a lot of profiling,
Identified corner cases,
worst case scenarios

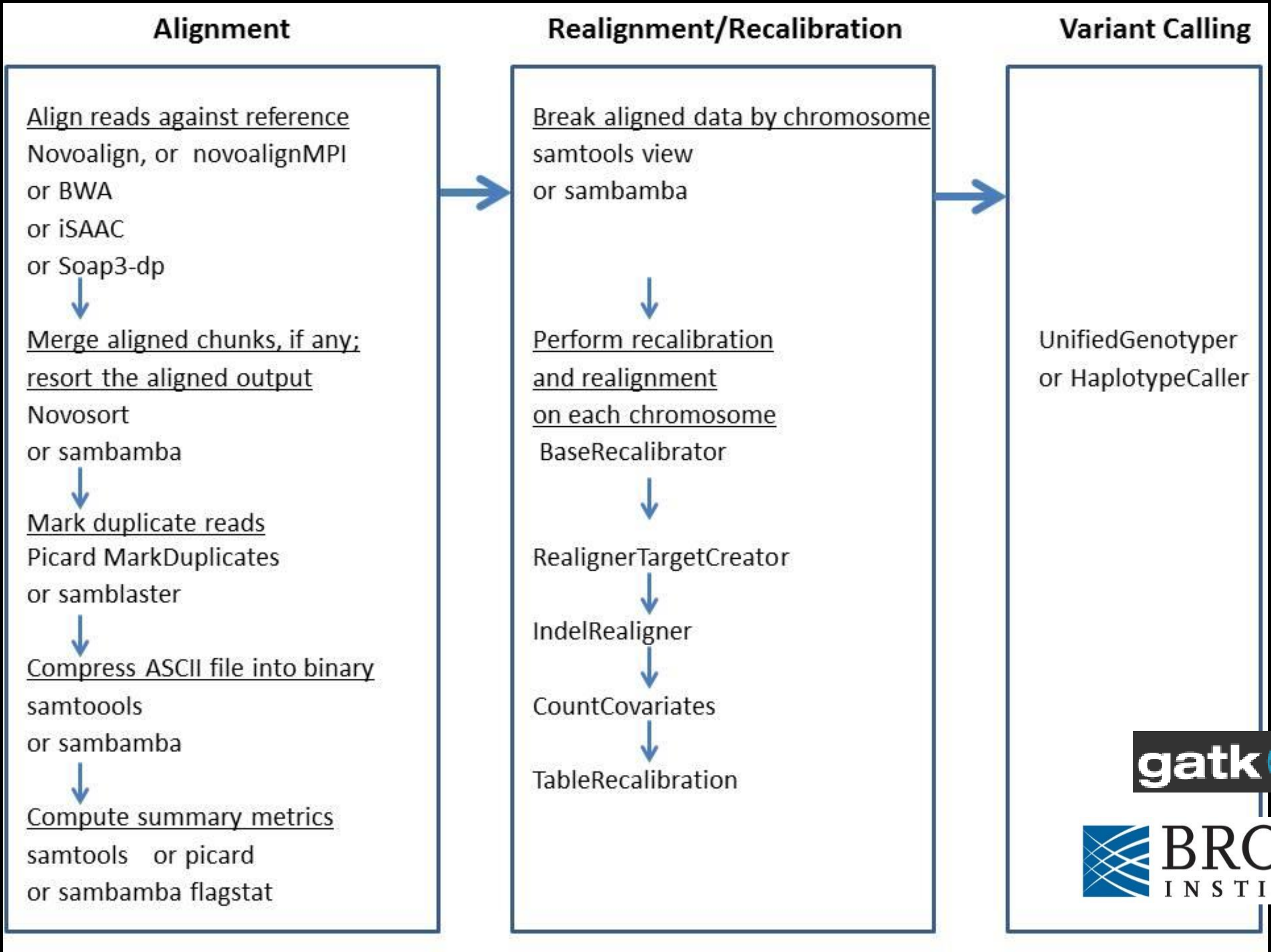
Blue Waters:

Craig Steffen, Jeremy Enos, Ryan Mokos, Jason Alt, Galen Arnold, Greg Bauer

CSL:

Subho Banerjee, Arjun Athreya, Zachary Stephens, Dr. Ravi Iyer

Workflow management and scheduling



Alignment

Align reads against reference
Novoalign, or novoalignMPI
or BWA
or iSAAC
or Soap3-dp

Merge aligned chunks, if any;
resort the aligned output
Novosort
or sambamba

Mark duplicate reads
Picard MarkDuplicates
or samblaster

Compress ASCII file into binary
samtoools
or sambamba

Compute summary metrics
samtools or picard
or sambamba flagstat

Realignment/Recalibration

Break aligned data by chromosome
samtools view
or sambamba

Perform recalibration
and realignment
on each chromosome
BaseRecalibrator

RealignerTargetCreator

IndelRealigner

CountCovariates

TableRecalibration

Variant Calling

UnifiedGenotyper
or HaplotypeCaller

Hypothetical job pattern: 500 genomes run

1. Alignment
   500 jobs for BWA
   10 chunks * 500 genomes = 5,000 jobs for Novoalign

2. Split data by chromosome
   25 chromosomes * 500 genomes = 12,500 jobs

3. Realignment/Recalibration
   25 chromosomes * 500 genomes = 12,500 jobs

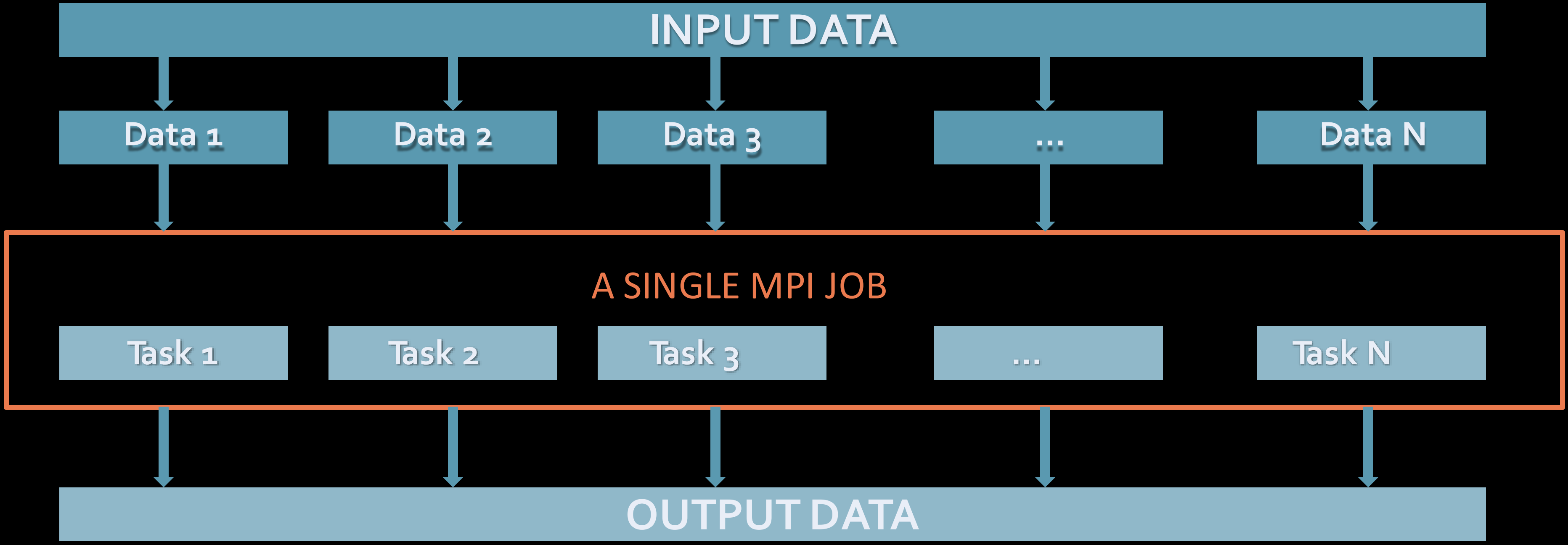
4. Variant calling
   25 chromosomes * 500 genomes = 12,500 jobs

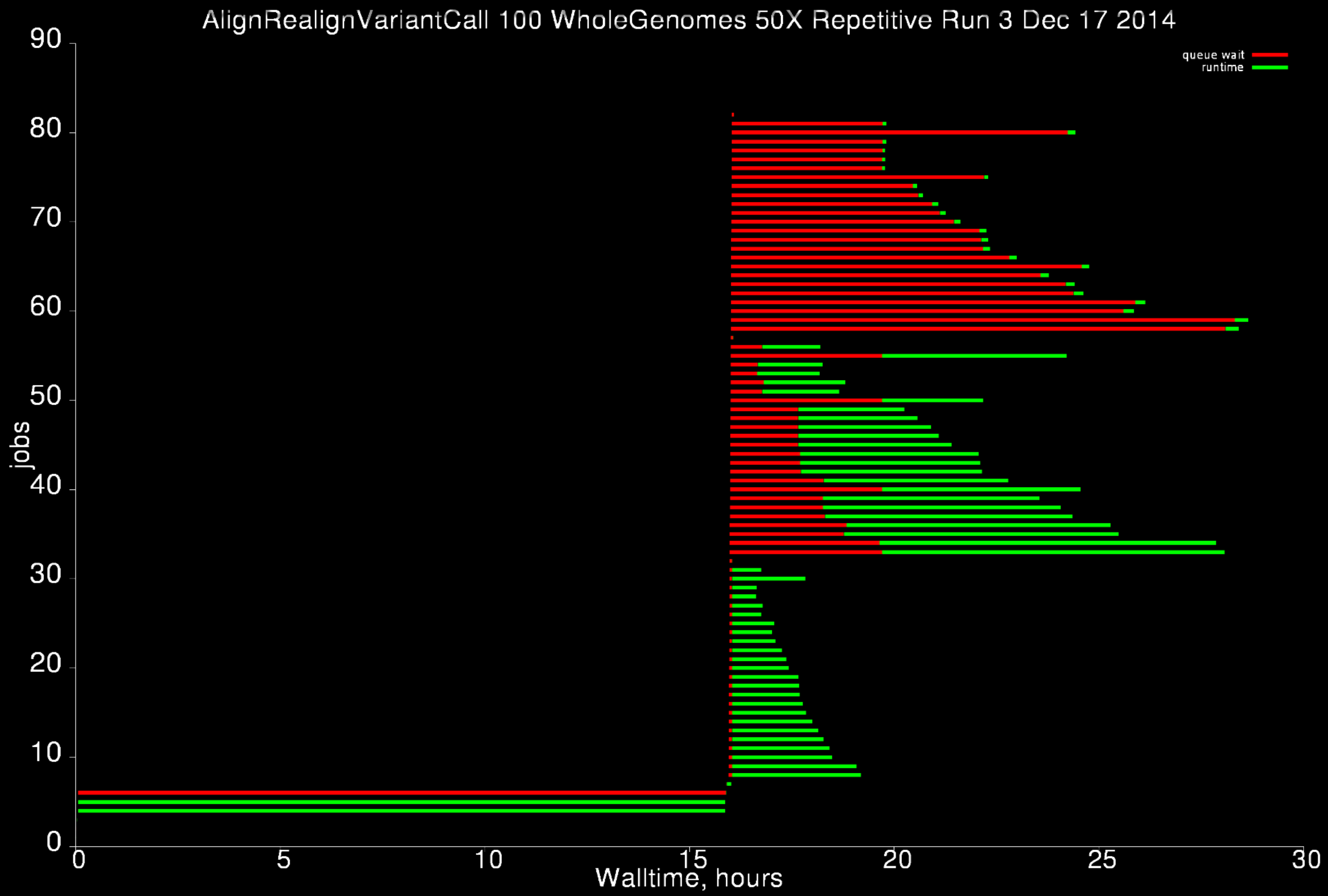
Job management

Solution: wrap multiple SMP jobs with a launcher, turning them into a single MPI job

- A single multi-node reservation is made on the cluster
- Launcher is started within that reservation
- It launches each task within this reservation
- As tasks complete, it launches new ones, until the list of tasks is exhausted

Victor Anisimov, NCSA
Blue Waters support group




AlignRealignVariantCall 100 WholeGenomes 50X Repetitive Run 3 Dec 17 2014
queue wait
runtime
jobs
90
80
70
60
50
40
30
20
10
0
0
5
10
15
20
25
30
Walltime, hours

Part 3:

Outlook
Alternative solutions
Production cases

Making big data be small data
Making big compute be small compute

Ultrafast => no need for checkpointing, only 2 output files

Monolithic => only 1-2 jobs, no workflow management needs

Making big data be small data =>

- Changing encoding protocols: letters to bits
- Compressing the data
- Computing on compressed data
- Changing the contents of the output files to encode the same information with fewer bits

bina TECHNOLOGIES GENALICE TECHNOLOGY FOR PEOPLE & SCIENCE BALSA illumina®

Variant calling: a production case

H3

wellcome trust
sanger
institute

Baylor
College of
Medicine

- Human Heredity and Health in Africa
- A massively collaborative project
- To profile the genotypic diversity across the African continent
    - Help cure diseases
    - Help understand human evolution

- > 2,000 genomes total
- ~350 genomes sequenced at 30X depth, at Baylor
- To arrive in batches of 50 genomes

Acknowledgements

INSTITUTE FOR
GENOMIC BIOLOGY

HPCBio

Victor Jongeneel
Gloria Rendon
Chris Fields

CompGen
INITIATIVE

CompGen

Ravi Iyer
Subho Banerjee
Arjun Athreya
Zachary Stephens

NCSA

Private Sector Program

Evan Burness
Jim Long
Wayne Hoyenga

Innovative Systems Lab

Volodymyr Kindratenko

Blue Waters support team

Greg Bauer
Victor Anisimov
Ryan Mokos
Kalyana Chadalavada
Alex Parga
Jeremy Enos
Andriy Kot
Jason Alt
Craig Steffen

CRAY

Cray

Bob Fiedler
Carlos Sosa
Pierre Carrier
Richard Walsh
Bill Long
Jef Dawson