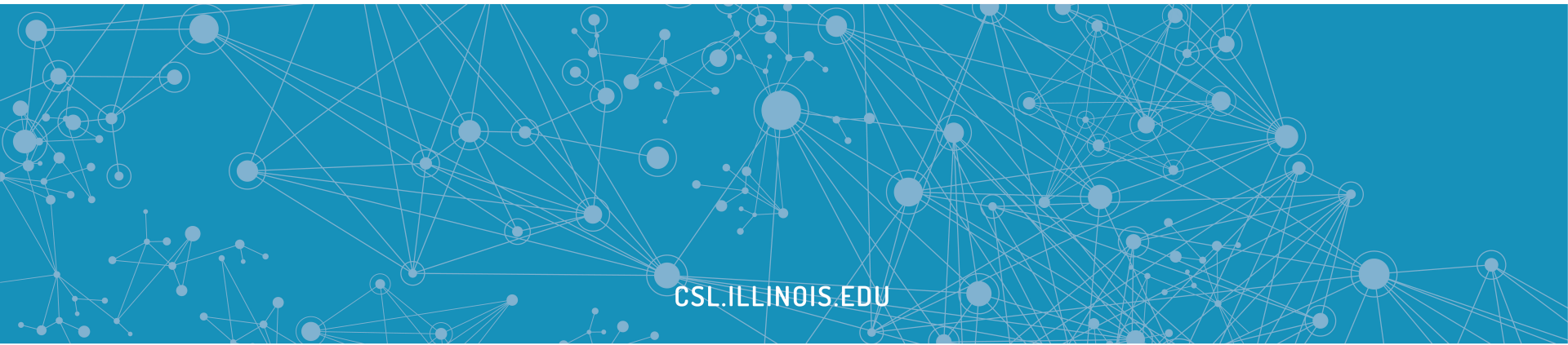
CSL: COORDINATED SCIENCE LAB

# Efficient and Scalable Workflows for Genomic Analyses

Subho S. Banerjee, Arjun P. Athreya, Liudmila S. Mainzer, C. Victor Jongeneel, Wen-Mei Hwu, Zbigniew T. Kalbarczyk, Ravishankar K. Iyer



# Summary

Contributions:

- **Common mathematical kernels**: Static analysis of genomic analyses algorithms
- **Performance Pathologies**: Measurement driven diagnosis of performance bottlenecks
- **IGen**: A scalable genomic data analytics framework which overcomes observed inefficiencies

"Variant Calling and Genotyping" Workflow as the driver

| | Baseline Runtime | IGen Accelerated Runtime ** | Speedup |
|---|---|---|---|
| Blue Waters – Single Node (CPU) | 59 hr | 28 hr | 2.1x |
| IBM Power 8 – Single Node (CPU + GPU + FPGA) | 36 hr | 11 hr | 5.3x, 3.2x |
| Blue Waters – 10 Nodes (CPU) | - | 2 hr | 22x |

# Outline

- Genomics Primer: Variant Detection
- Kernels for Genomics
- Performance Pathologies in State of the Art Genomics Pipelines
- IGen: The **I**llinois **G**enomics Execution **En**vironment

# Overview: Variant Calling and Genotyping

Detecting and characterizing mutations in a sample genome

# Overview: Variant Calling and Genotyping

Detecting and characterizing mutations in a sample genome

Reference Genome 

A C A G G T A T A T C T T T G A T A …

Sample Genome 

???

NGS Sequencing Technology 

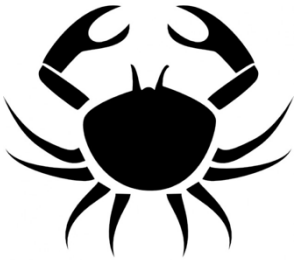
A C A T G T G A T A T A

C A T G T A T C T T T C

T G T A T C T T C G A T

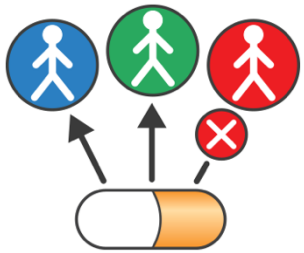
# Overview: Variant Calling and Genotyping

Detecting and characterizing mutations in a sample genome



Diagnosis

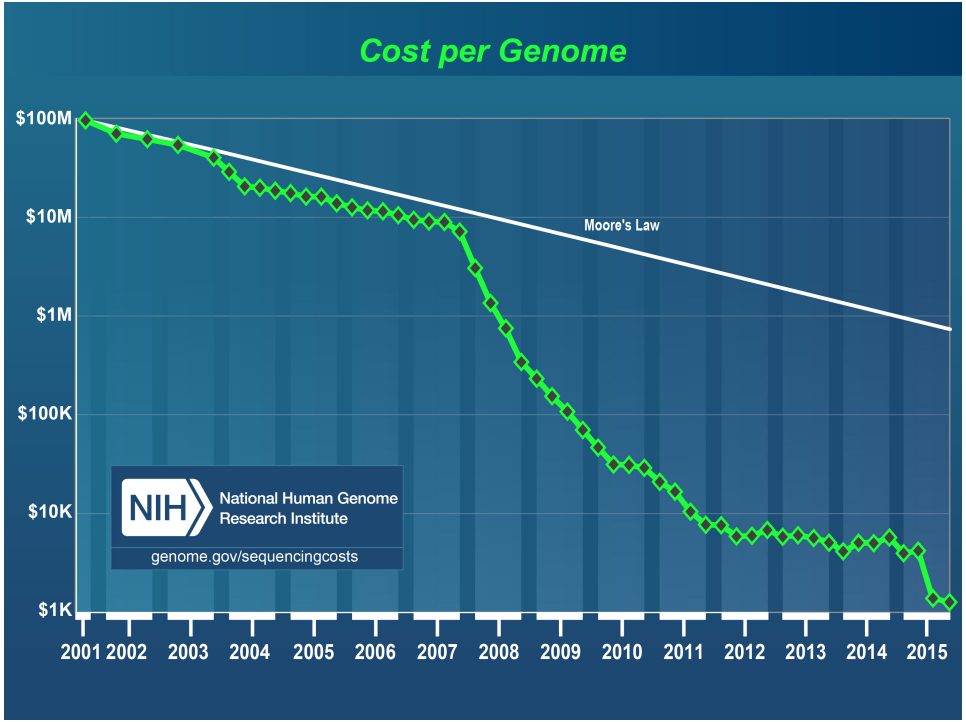
e.g., Cancer



Personalized Medicine

Noisy Data  Polyploid Samples

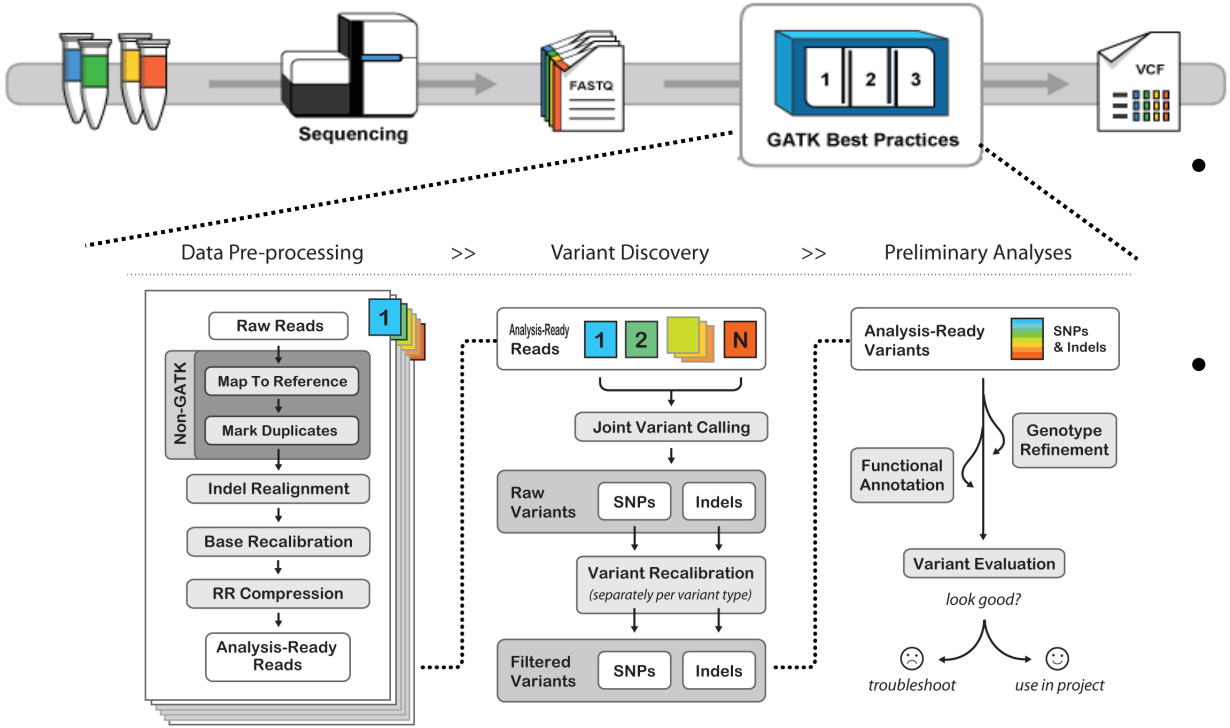
# The Genomic Data Deluge




http://www.genome.gov/sequencingcosts

- Falling costs
  - **Capex**: Cost of buying sequencing machines
  - **Opex**: Cost of sequencing genomes
- Potential for large amounts of sequence data to be generated over a short span of time
- Societally important problem
  - Scope for personalized medicine changing healthcare delivery

# Variant Discovery as a Workflow




https://www.broadinstitute.org/gatk/guide/best-practices.php

- The Broad Institute Best Practices Guidelines
- Tools come from disparate sources
  - Designed for workstations
  - **Few** are performance tuned
  - Do not fit well in traditional HPC

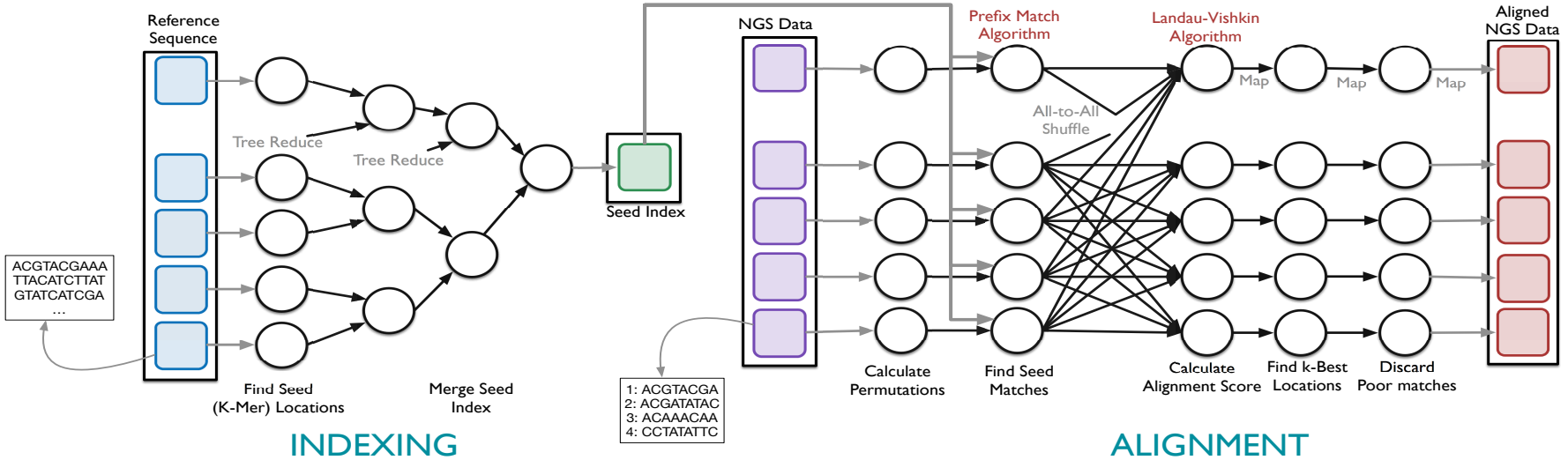
# Outline

- Genomics Primer: Variant Detection
- Kernels for Genomics
- Performance Pathologies in State of the Art Genomics Pipelines
- IGen: The **I**llinois **G**enomics Execution **En**vironment

# Computational Kernels

- **Computational Kernels**: Basic Mathematical Operations common to large number of bioinformatics analysis
- Kernels enable system level optimizations effecting a large number of tools
- Clearly show commonalities between different tools performing the same analysis
- Provide an interface between algorithm designers and system designers
  - Future benchmarks for data-intensive HPC machines
- Defines a simple data-flow abstraction for non-expert programmers (biologists)

# Kernels




Single Ended NGS Read Alignment as a DFG

# Kernels

Repeated kernel usage across tools/stages

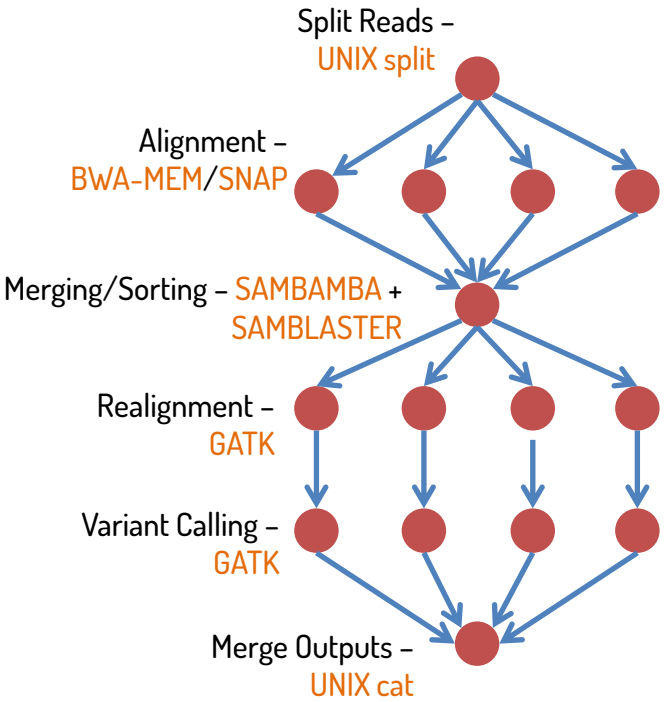
| Workflow stage | Kernels |
|---|---|
| Error Correction | K-mer computation |
| Alignment | K-mer computation, Prefix Tree, Edit-distance computation |
| Indel Re-Alignment | Edit-distance computation |
| Re-Calibration | Yates correction |
| Variant Calling | Entropy, Convolution, Assembly, Edit-distance, Pair-HMM, Bayesian inference |

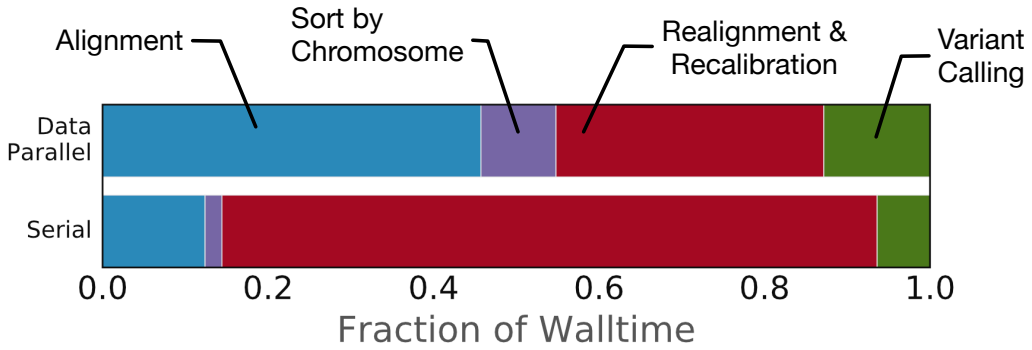
See paper for common kernels across multi-sequence alignment, metagenomics and phylogeny

# Outline

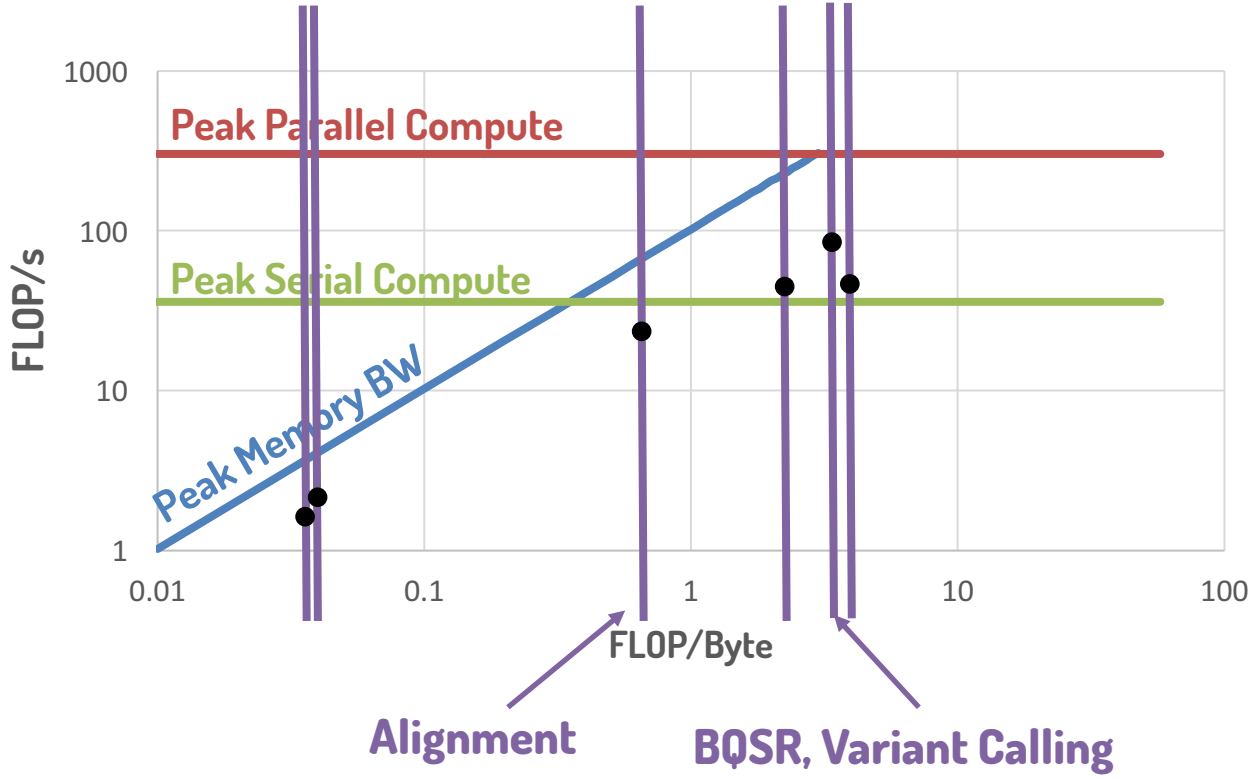
- Genomics Primer: Variant Detection
- Kernels for Genomics
- Performance Pathologies in State of the Art Genomics Pipelines
- IGen: The **I**llinois **G**enomics Execution **En**vironment

# Deploying genomics workflows on parallel systems

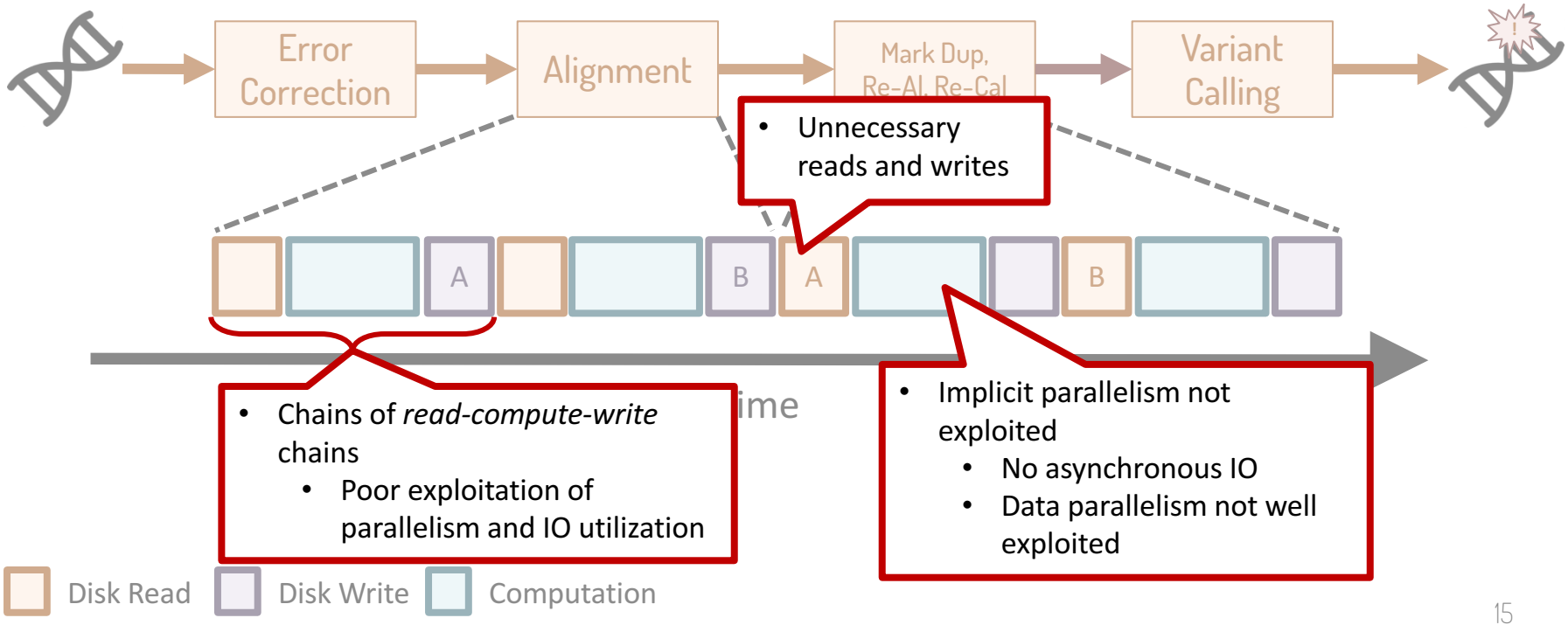


# Tools are not well suited for HPC machines



# Understanding Performance Issues

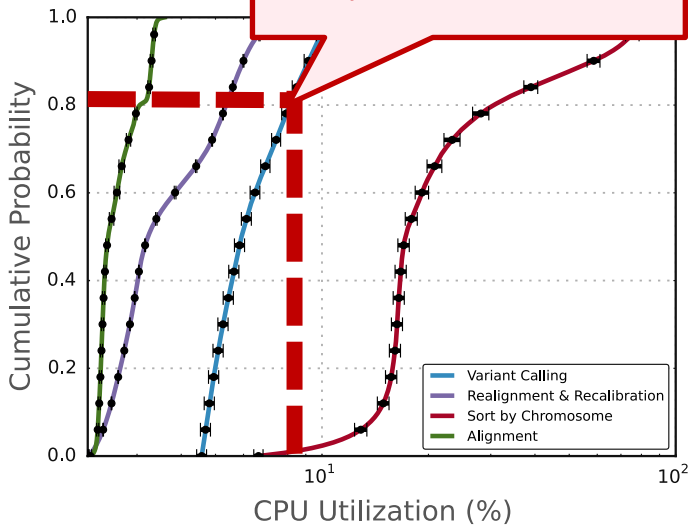


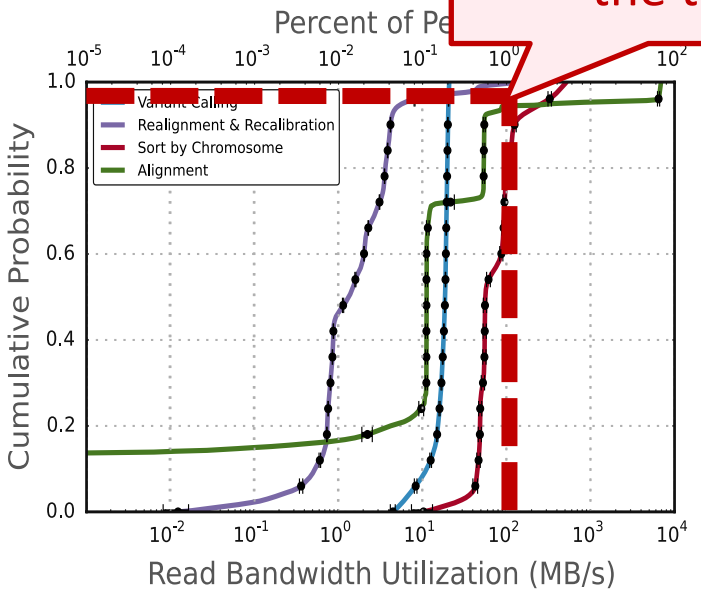
# Understanding Performance

surement on Blue-Waters

Less than 10% CPU Utilization 80% of the time

Less than 1% IO utilization 95% of the time



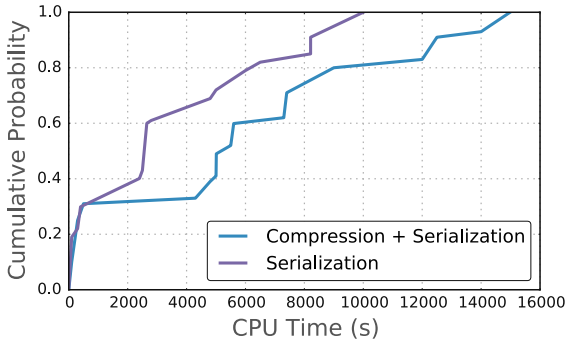




System resource utilization for phases of the Broad Institute Best Practices Guide Workflow

# Variant Calling on Spark

## ADAM and Avocado

- Best performance – Best place to start!
- ADAM: Extremely efficient data formats for parallel compute




**Time Spent in Serialization for ADAM based file formats**

- Several Problems
  - Serialization takes a lot of time
  - Easy to program ≠ Good performance
  - Single Node performance quite poor, Great Scalability
  - Non-trivial (12.3 %) amount of time spent in fault-tolerance related computation/messaging
  - JVM - Garbage collection

# Outline

- Genomics Primer: Variant Detection
- Kernels for Genomics
- Performance Pathologies in State of the Art Genomics Pipelines
- IGen: The **I**llinois **G**enomics Execution **En**vironment

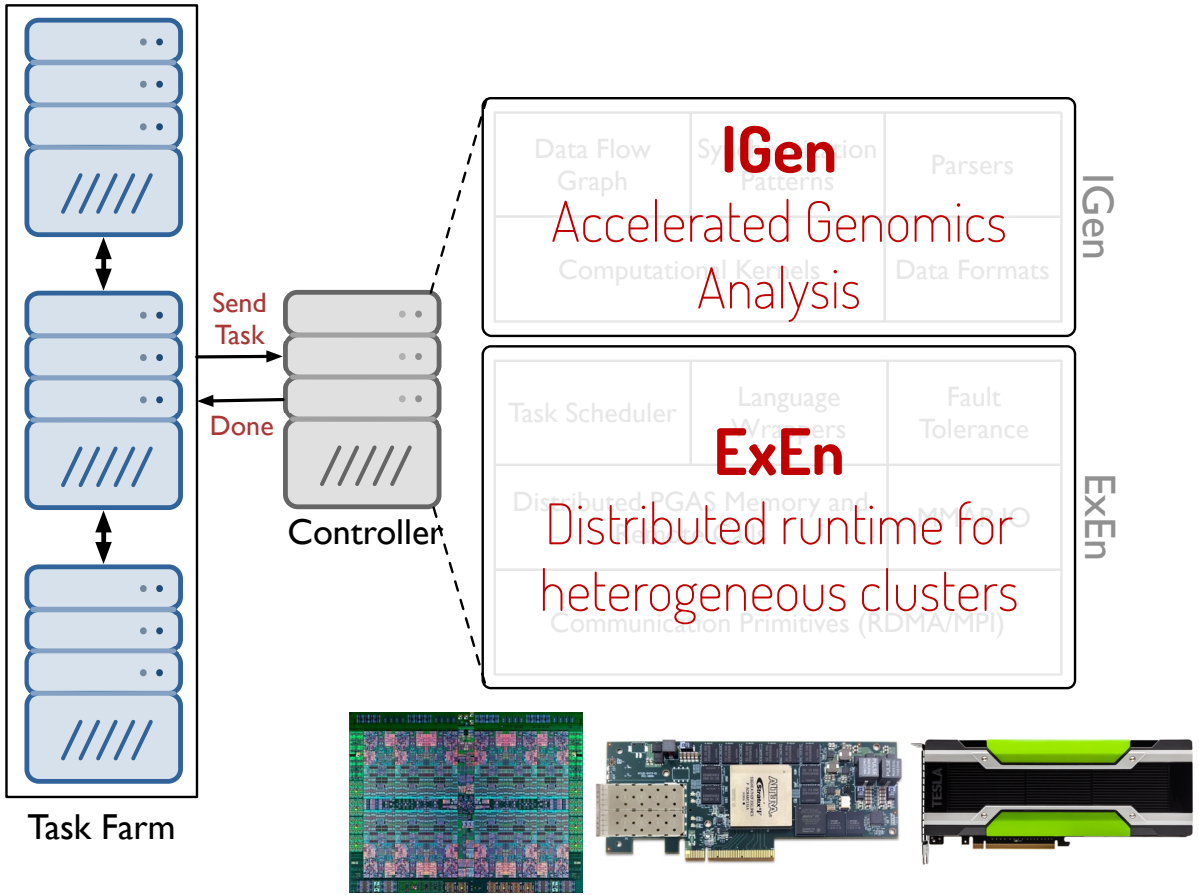
# Sequences to Systems

**Key Idea**: Decouple **algorithms**, **schedule** and **accelerators**

- Algorithm: What is computed
- Schedule: Where and when it is computed
- Accelerators: How it is computed

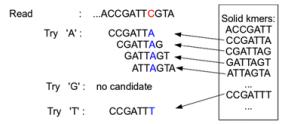


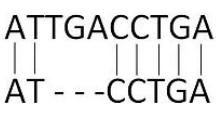

Hardware Layer
- Massively parallel processors
- Specialized hardware

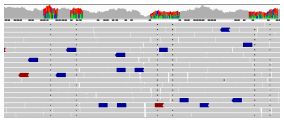

Send Task
Done
Controller
Task Farm
IGen
Accelerated Genomics Analysis
Data Flow Graph
Parsers
Computational Kernels
Data Formats
ExEn
Distributed runtime for heterogeneous clusters
Task Scheduler
Fault Tolerance
Communication Primitives (RDMA/MPI)

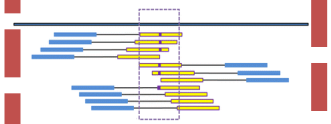
# Variant Calling and Genotyping in IGen

Detecting and characterizing mutations in a sample genome





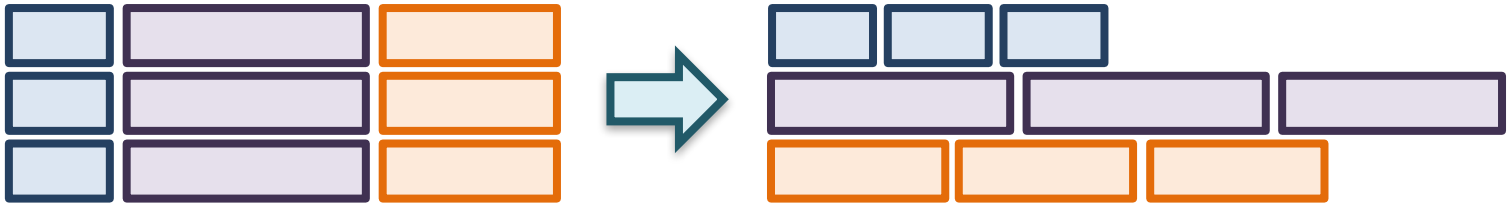
Error Correction → Alignment → Mark Dup, Re-Al, Re-Cal → Variant Calling

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Blue Waters – 22 Nodes (CPU) | - | 14.3 hr | 3.8 | 2.3 | 0.2 | 4.8 | 3.1 hr | Broad Inst. Best Practices |
| Blue Waters – 10 Nodes (CPU) | - | 48 min | 63 min* | | | | 12 min | IGen Accelerated |

* Whole Human Genome @ 60x coverage
** Default tool parameters

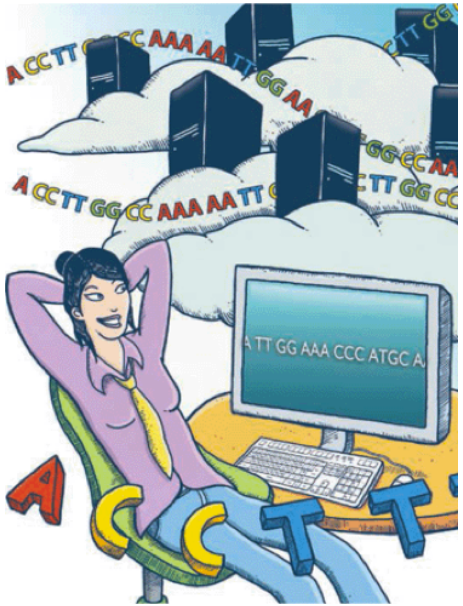
# Enhancements in IGen

- Succinct data representations
  - All tools use ASCII based in-memory representations
  - Use 2 and 4 bit representation for Nucelotides/CIGAR
- Asynchronous File IO
- Column based data-structures to improve locality and aid vectorization



- Compiler assisted and SIMD intrinsic based implementations of kernels

# Conclusions




- Bringing computer systems and analytics to precision medicine
  - ExEn and IGen for accelerated NGS analysis
  - NEAT and AssembleSV for quality control of NGS pipelines
  - Statistical analysis for deriving actionable intelligence



Image source: Google Images

# CompGen Machine

Medical Knowledge


Continuous Monitoring


*omics Data


Medical Devices


Patient Records

- Timely Diagnosis
- Personalized Drugs
- Model Drug Response
- New Biological Insight