High Performance Computing in Life Sciences

Part I
HPC Introduction

PartII
BioComputing Sofware Introduction

Oleksandr Moskalenko
om@ufl.edu

Matt Gitzendanner
magitz@ufl.edu

UF Research Computing
Information Technology

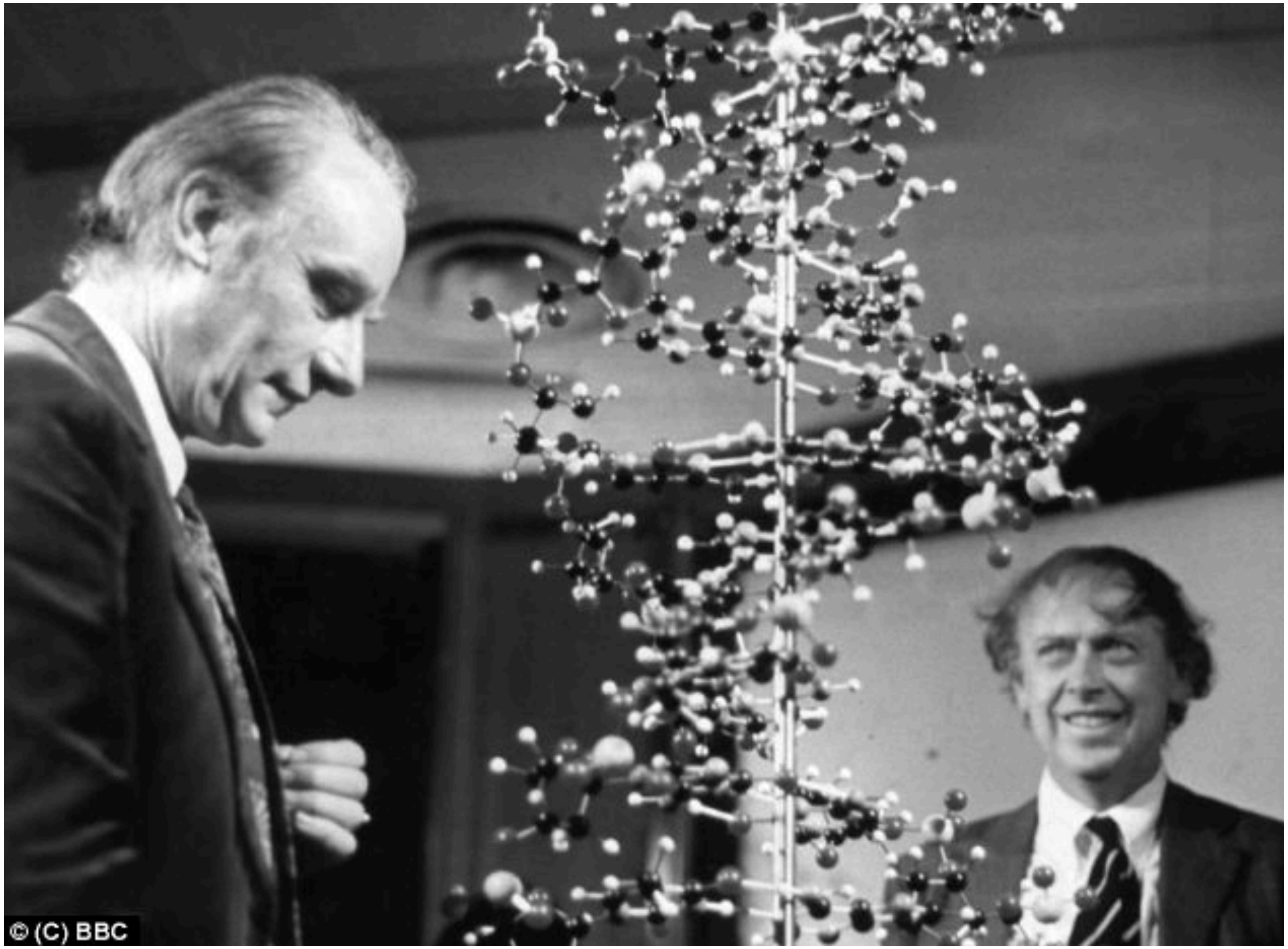
Summary

- The scale of biocomputing challenges
- The evolution of High-Performance Computing
- Current state of the traditional computing
- Parallelizing analyses
  - Traditional multiprocessing
  - Hadoop
  - Specialized approaches
- The interfaces
  - GUI vs. Web vs. Batch (comman-line)
- Biocomputing Software (Part II)

Historical Perspective

From a molecule to millions of genomes

The Beginning

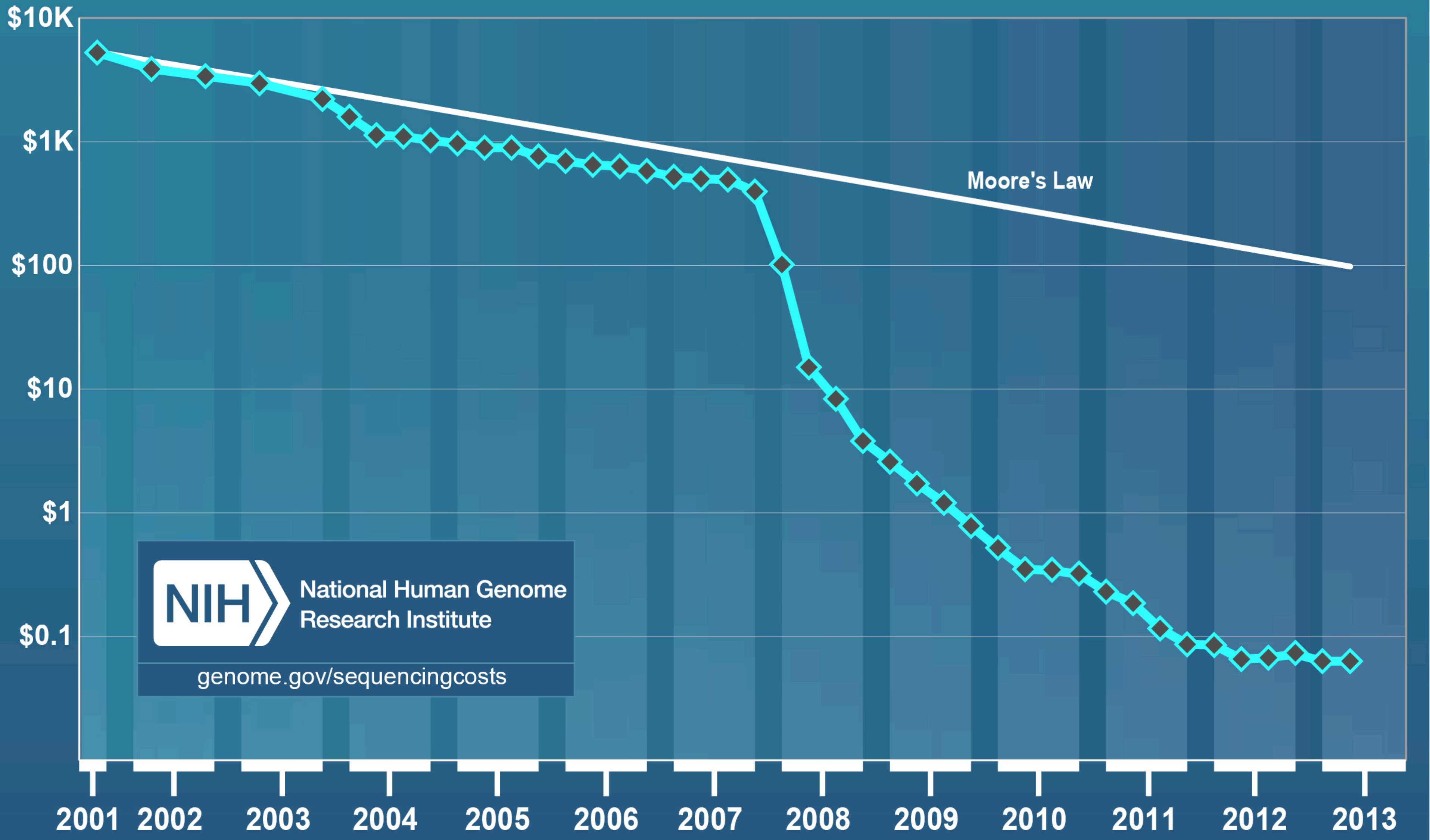


© (C) BBC

Sequencing Data Scaling

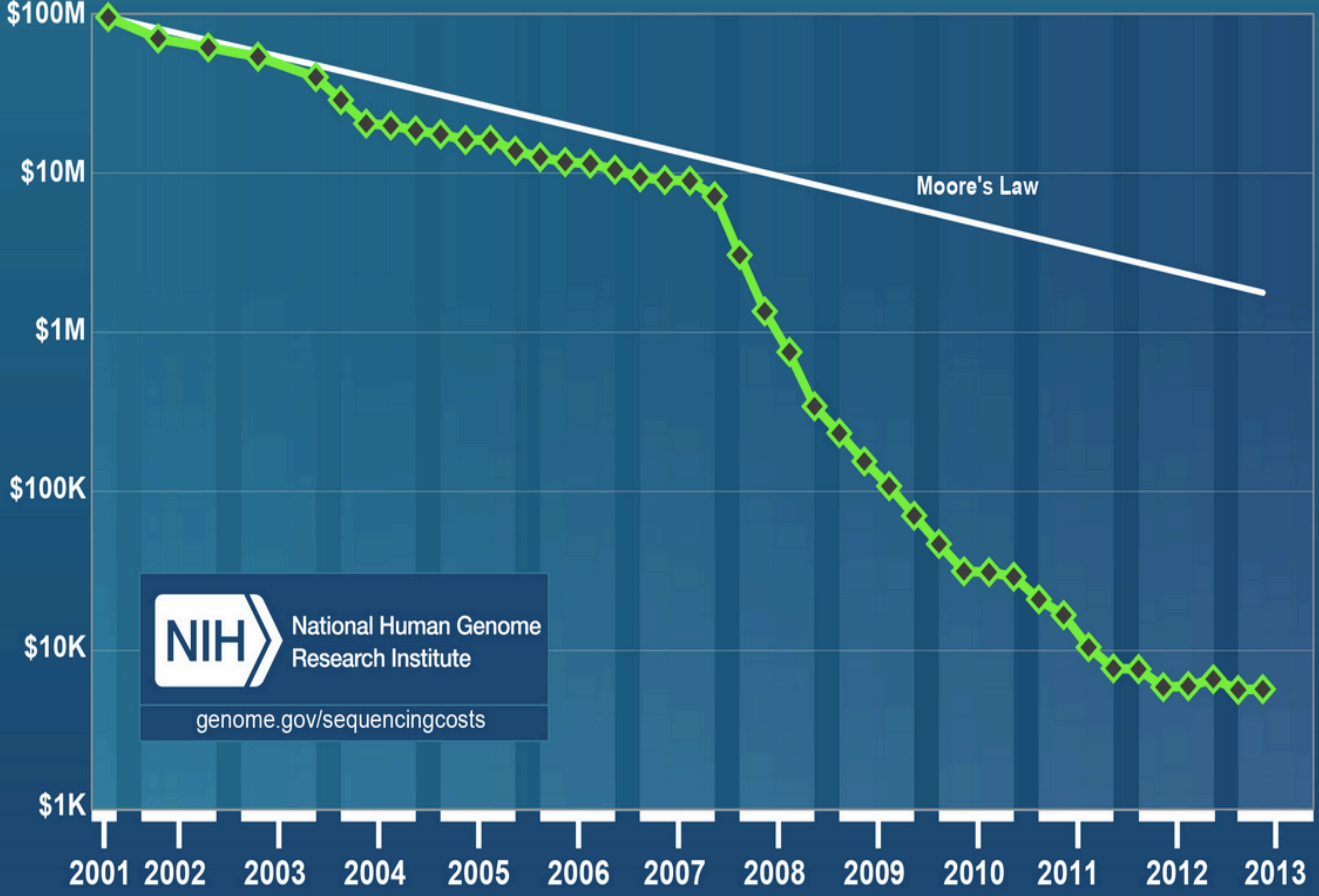
- Genome Size * Coverage
  - Viral – 1-100kbp
  - Bacteria, Archaea – 1-10Mbp
  - Simple Eukaryotes – 10-100 Mbp
  - Animals, Plants – 100Mbp - > 100Gbp
- Sequencing Coverage
  - ~10x in the Sanger Shotgun WGS times
  - ~30x for an average analysis
  - ~100x for metagenomic studies
  - Up to ~1000x for low-frequency SNP analysis in mixed samples

Cost per Raw Megabase of DNA Sequence



National Human Genome Research Institute

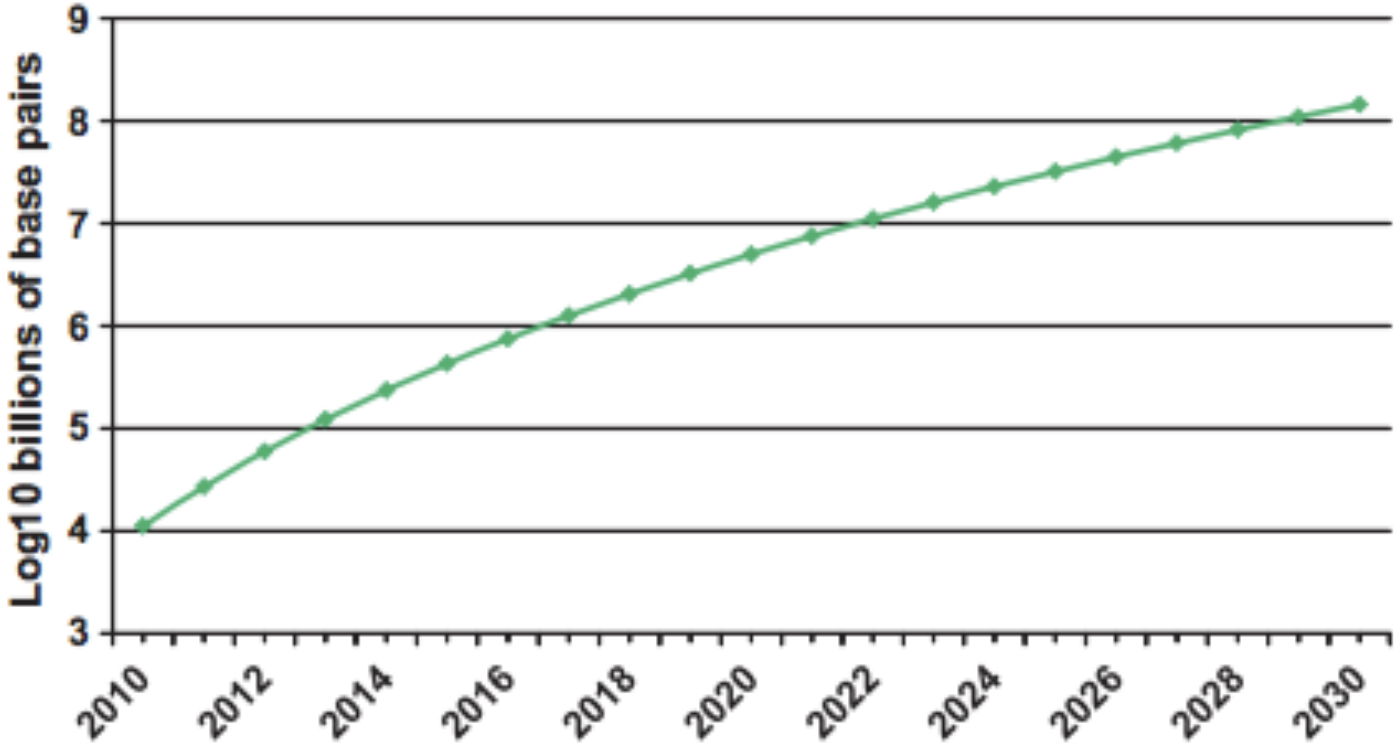
genome.gov/sequencingcosts



NIH National Human Genome Research Institute

genome.gov/sequencingcosts

Growth of Sequencing Data



10^6 (Mb) -> 10^9 (Gb) -> 10^{12} (Tb) -> 10^{15} (Pb) -> 10^{18} (Eb) -> 10^{21} (Zb)

Grossman et al. (2011)

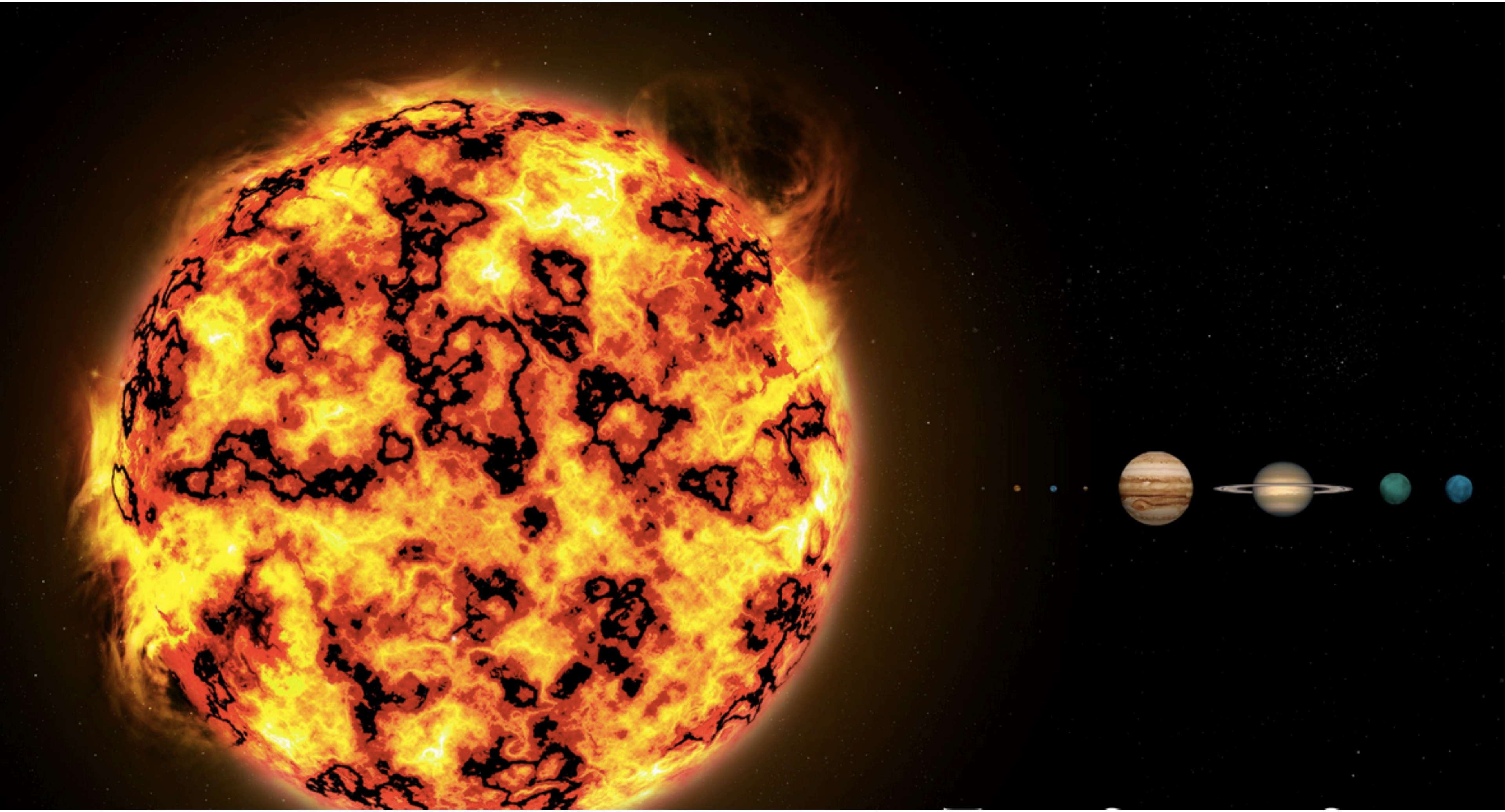
Growth of Sequencing Data

- 1 Gigabyte: A pickup truck filled with paper OR A symphony in high-fidelity sound OR A movie at TV quality
- 10 Terabytes: The printed collection of the US Library of Congress
- 2 Petabytes: All US academic research libraries
- 5 Exabytes: All words ever spoken by human beings.
- 2.7 Zettabytes: the total amount of global data in 2012 (IDC).

10^6 (Mb) -> 10^9 (Gb) -> 10^{12} (Tb) -> 10^{15} (Pb) -> 10^{18} (Eb) -> 10^{21} (Zb)

Grossman et al. (2011)

BioComputing Growth - NGS



Evolution of HPC

From Local to Global

“Local” BioComputing

Early Grid BioComputing

UNIVERSITY OF FLORIDA | High-Performance Computing

HiPerGator

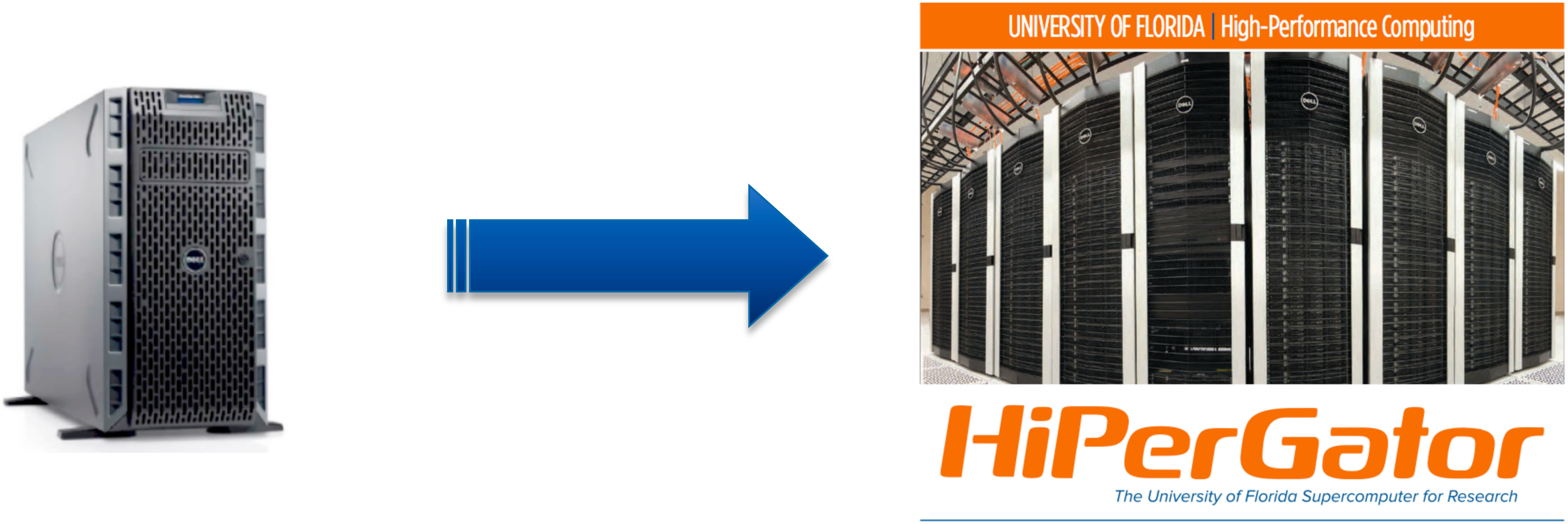
The University of Florida Supercomputer for Research

Contemporary Cluster Specs

- Storage and Networking:
  - 2Pb – Lustre parallel file system
  - 100Gbit networking, Infiniband Fabric
- Computing nodes:
  - 64 x 2.4GHz AMD Abu Dhabi cores
  - 254gb of usable memory
  - 1TB of local storage
- Big memory nodes:
  - 512Gb and 1TB memory with 48-80 cores
- GPU nodes:
  - Tesla, Fermi, Kepler GPU classes

HPC Considerations

- Scale


UNIVERSITY OF FLORIDA | High-Performance Computing
HiPerGator
The University of Florida Supercomputer for Research

HPC Considerations

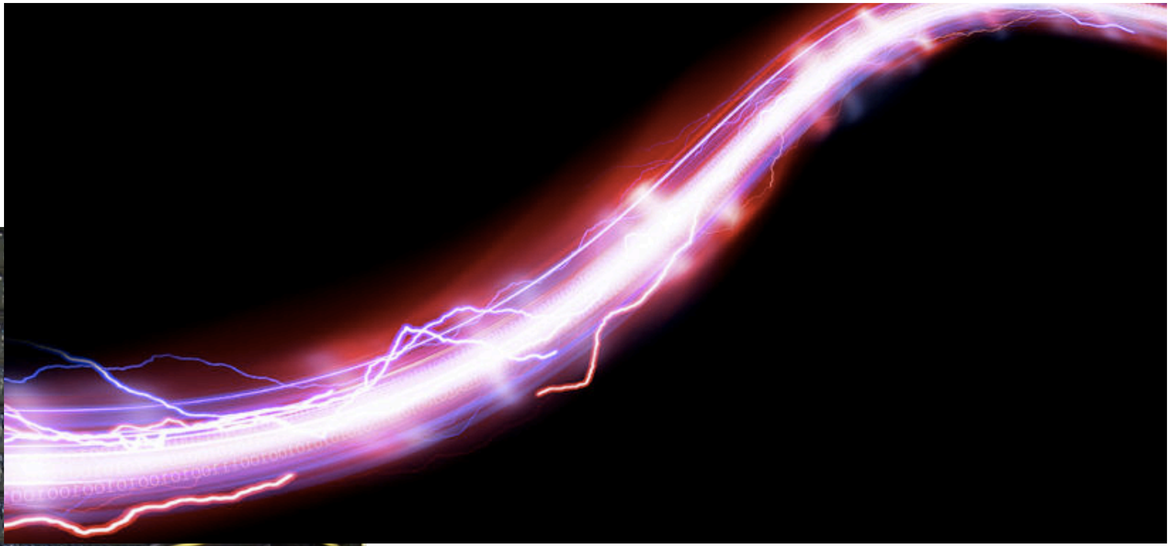
- Computational capacity vs.
  power and cooling

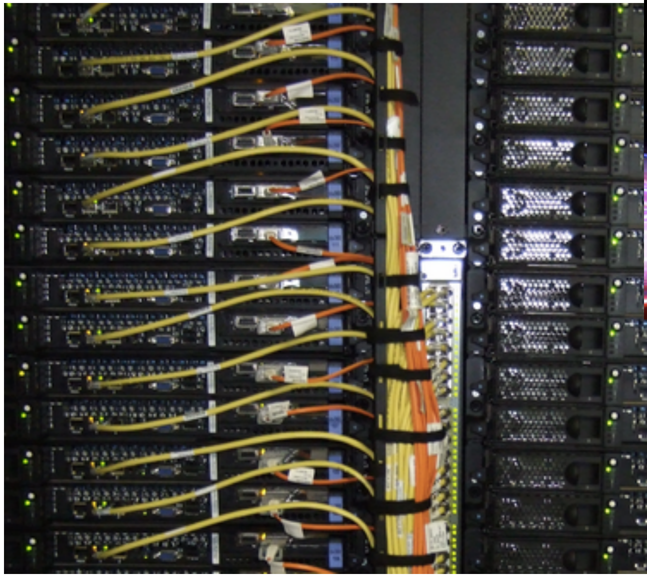
UF Data Center

- UF Data Center on Eastside Campus
  - 10,000 sq.ft and 1.75 MW total
  - 5,000 sq. ft. space for Research Computing

HPC Considerations

- Interconnects
- Networking





Internet2 Network

UF was 1st in the nation

- Internet2 Innovation Platform
  - 100 Gpbs connectivity
  - Campus Research Network now 200 Gbps

HPC Considerations

- Storage
- Parallel file systems
- High I/O storage
- Distributed storage

Scaling the HPC

The power of many

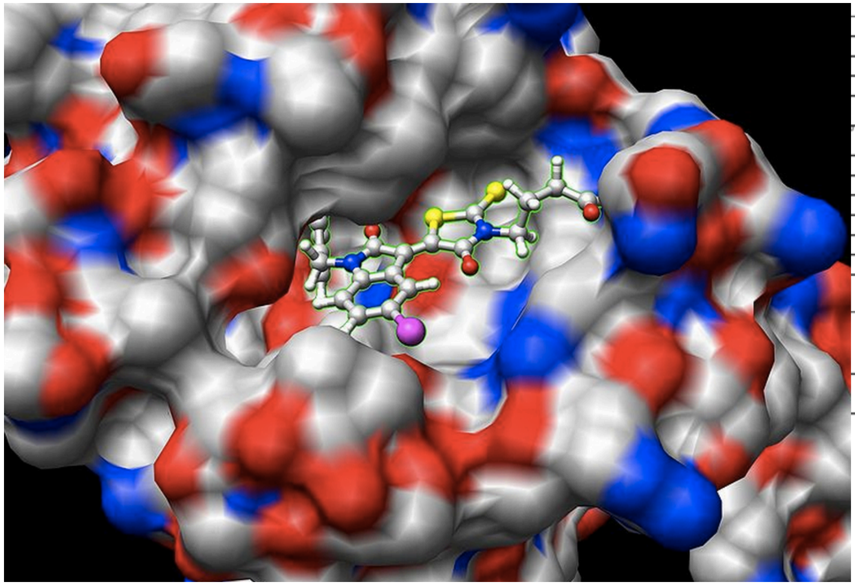
UNIVERSITY OF FLORIDA | High-Performance Computing

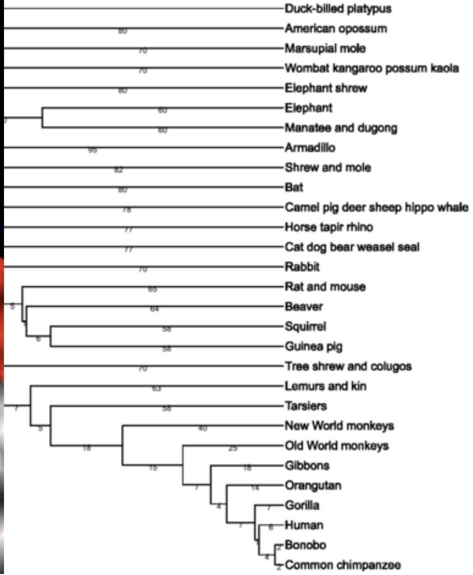
HiPerGator

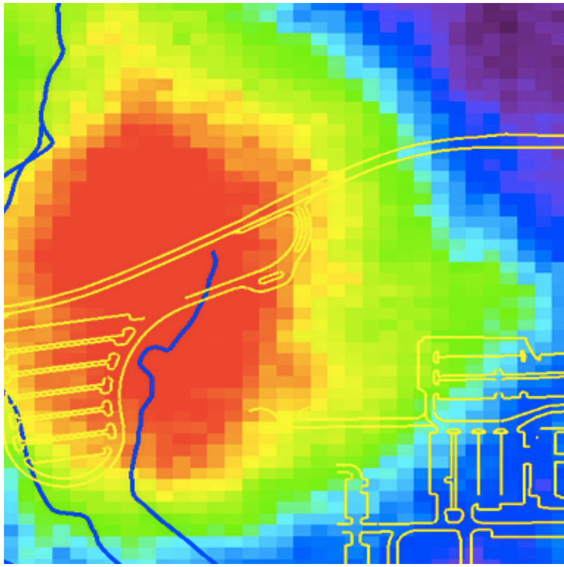
The University of Florida Supercomputer for Research

Computational Power

- Modeling, phylogenetics, simulations







Traditional Computation

- *De*-novo genome assembly
- Short-read mapping
- RNA-Seq
- BS-Seq
- CHIP-Seq
- SNP calling
- Pathway analysis
- …

- Why? Poor parallelization

Circumventing the Moore's Law

Divide and conquer

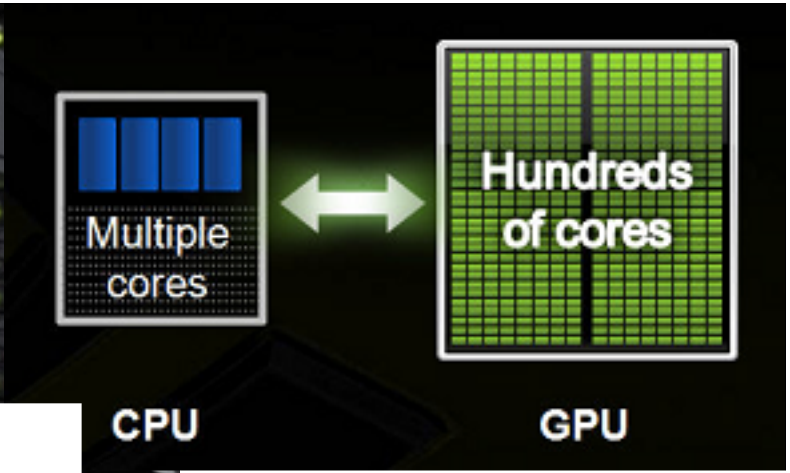
Traditional Parallel Computing

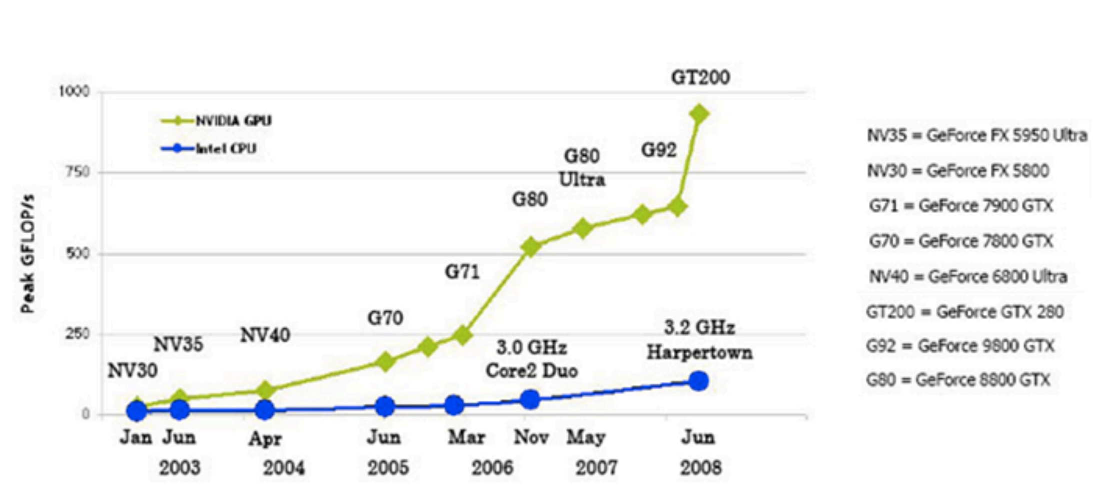
- Split analyses manually, run separately
- Multi-core (SMP) analyses with enabled software
- Multi-node (MPI) analyses with specially constructed software

GPU Computing

- Highly Parallelizable

Tesla™ GPU Parallel High Performance Computing (HPC) Solutions
Accelerate Your Performance - With Supermicro Massively Multithreaded.



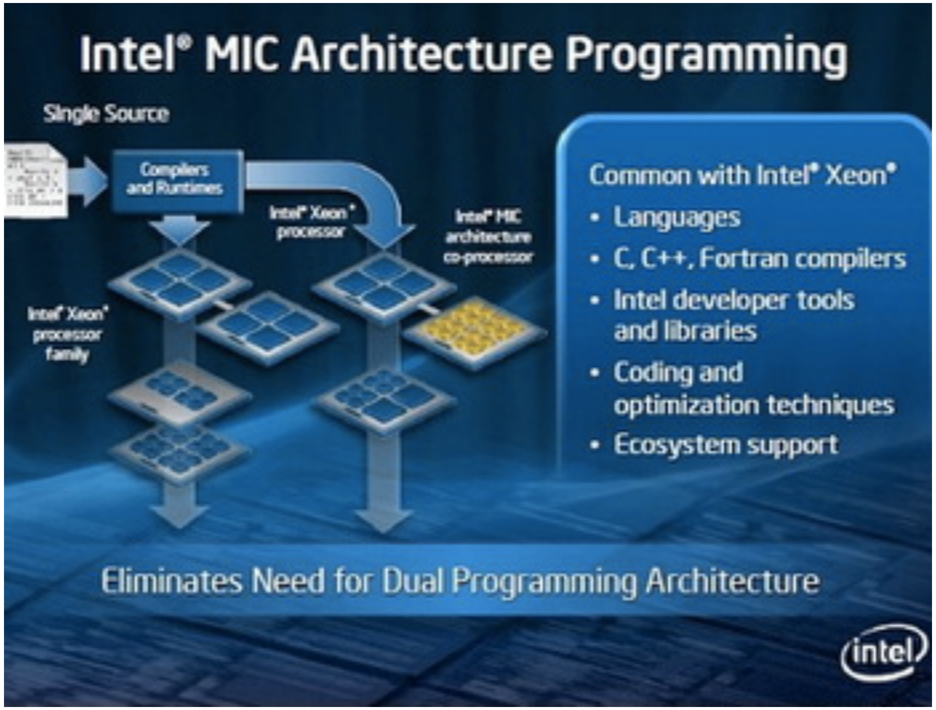


Need the code!

CUDA

MIC Computing

- Highly Parallelizable
- Standard x86 cores
- No need for learning a different programming paradigm ???



Intel® MIC Architecture Programming

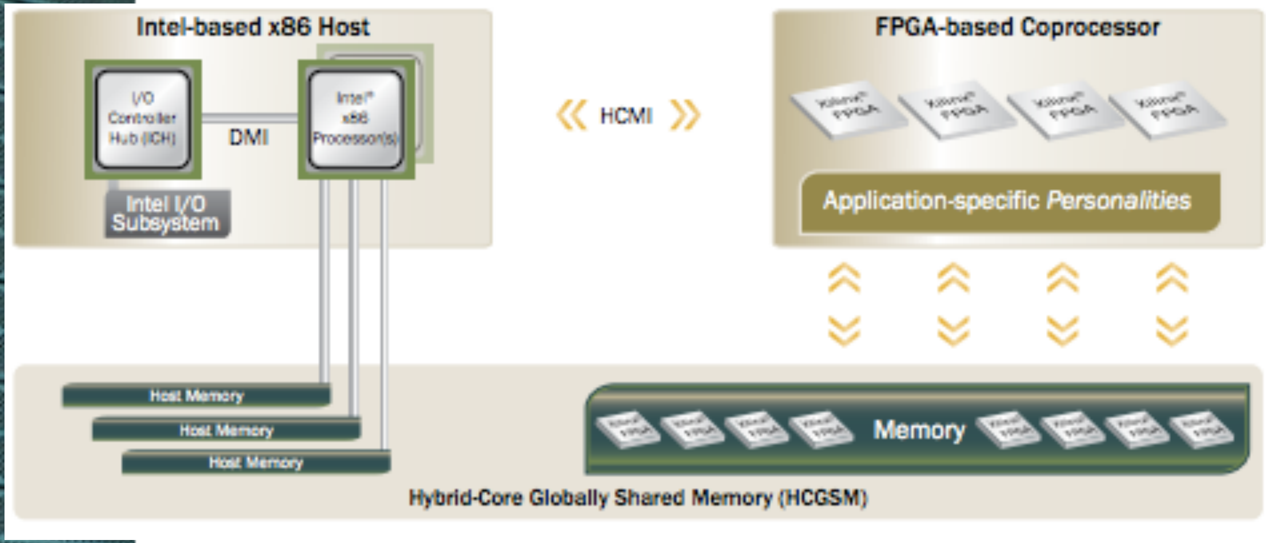
Single Source
Compilers and Runtimes
Intel® Xeon® processor
Intel® MIC architecture co-processor
Intel® Xeon® processor family
intel

Common with Intel® Xeon®

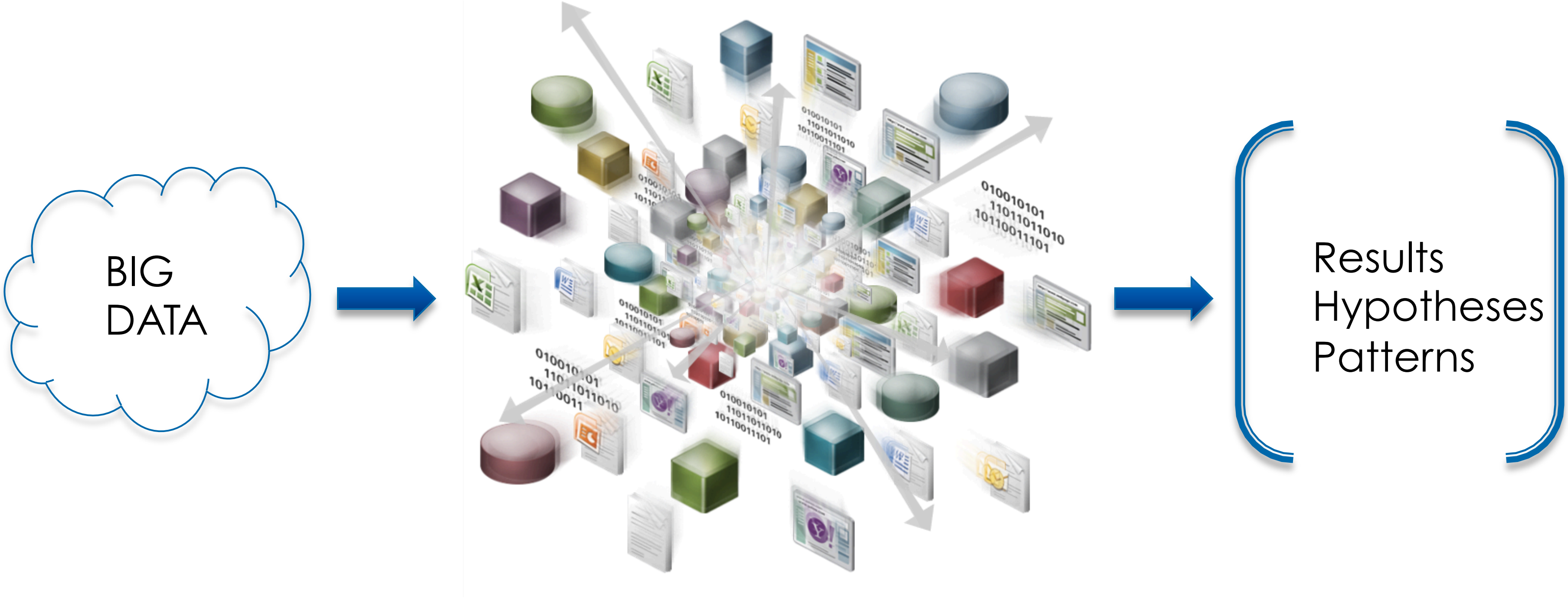
- Languages
- C, C++, Fortran compilers
- Intel developer tools and libraries
- Coding and optimization techniques
- Ecosystem support

Eliminates Need for Dual Programming Architecture

Specialized Processing



Distributed Computation (Hadoop)

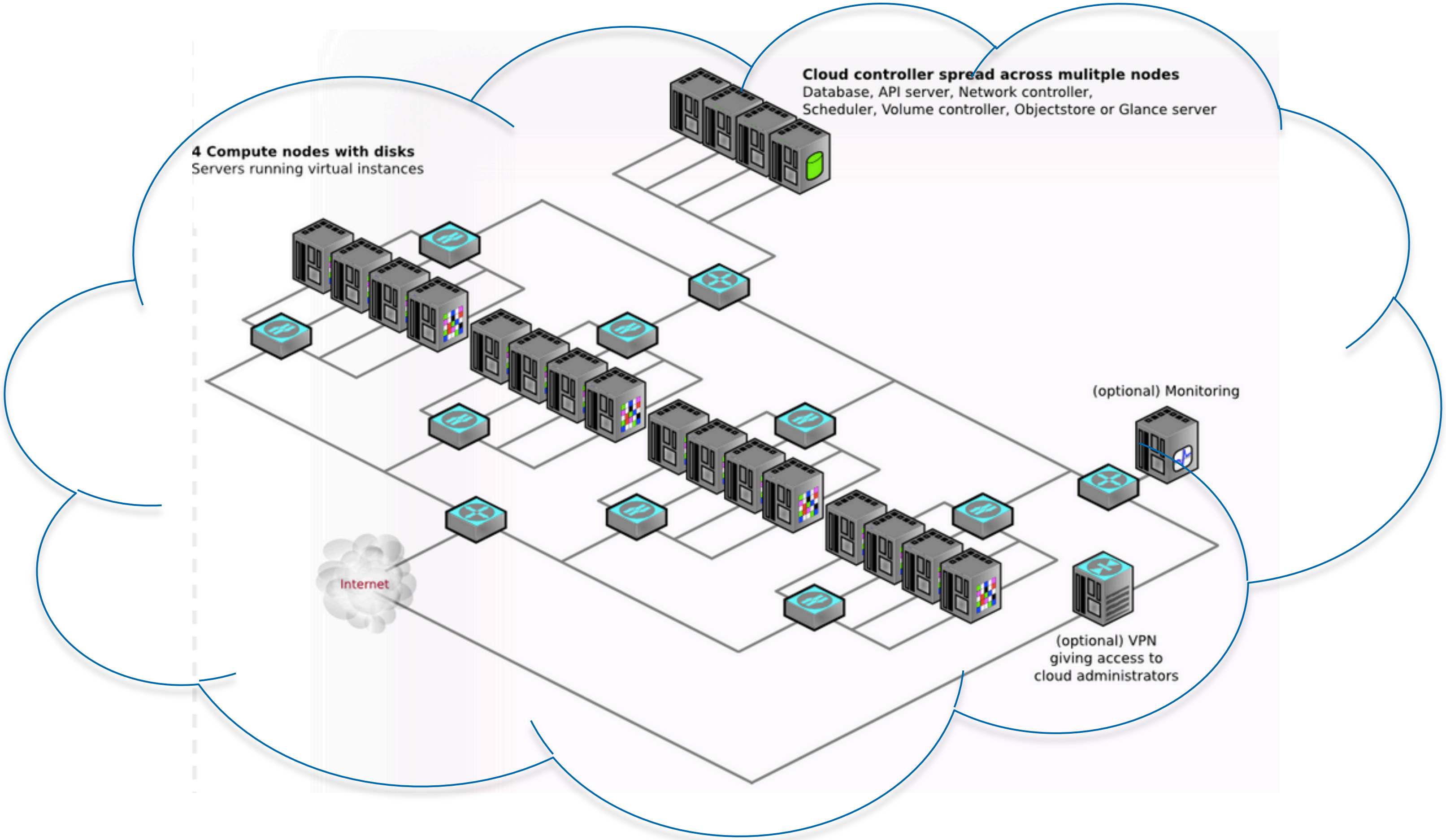


BIG
DATA

Results
Hypotheses
Patterns

Map-Reduce Approach

Biocomputing Cloud 9 ???

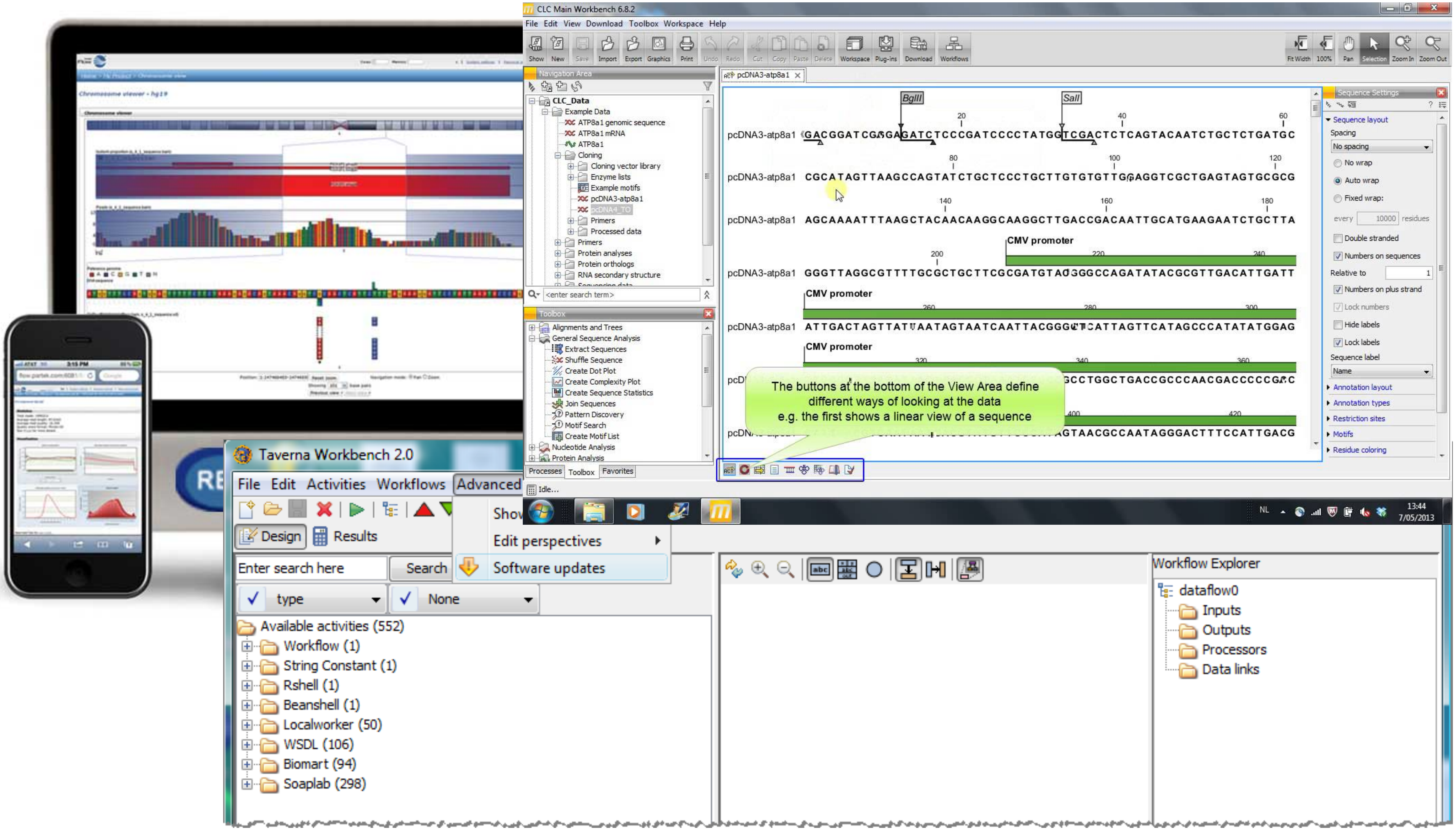

Cloud controller spread across mulitple nodes
Database, API server, Network controller,
Scheduler, Volume controller, Objectstore or Glance server
4 Compute nodes with disks
Servers running virtual instances
(optional) Monitoring
Internet
(optional) VPN
giving access to
cloud administrators

Interfaces

Interfaces, Interfaces, Interfaces!!!

What the Future May Bring

Graphical User Interfaces



The buttons at the bottom of the View Area define different ways of looking at the data e.g. the first shows a linear view of a sequence

Graphical User Interfaces

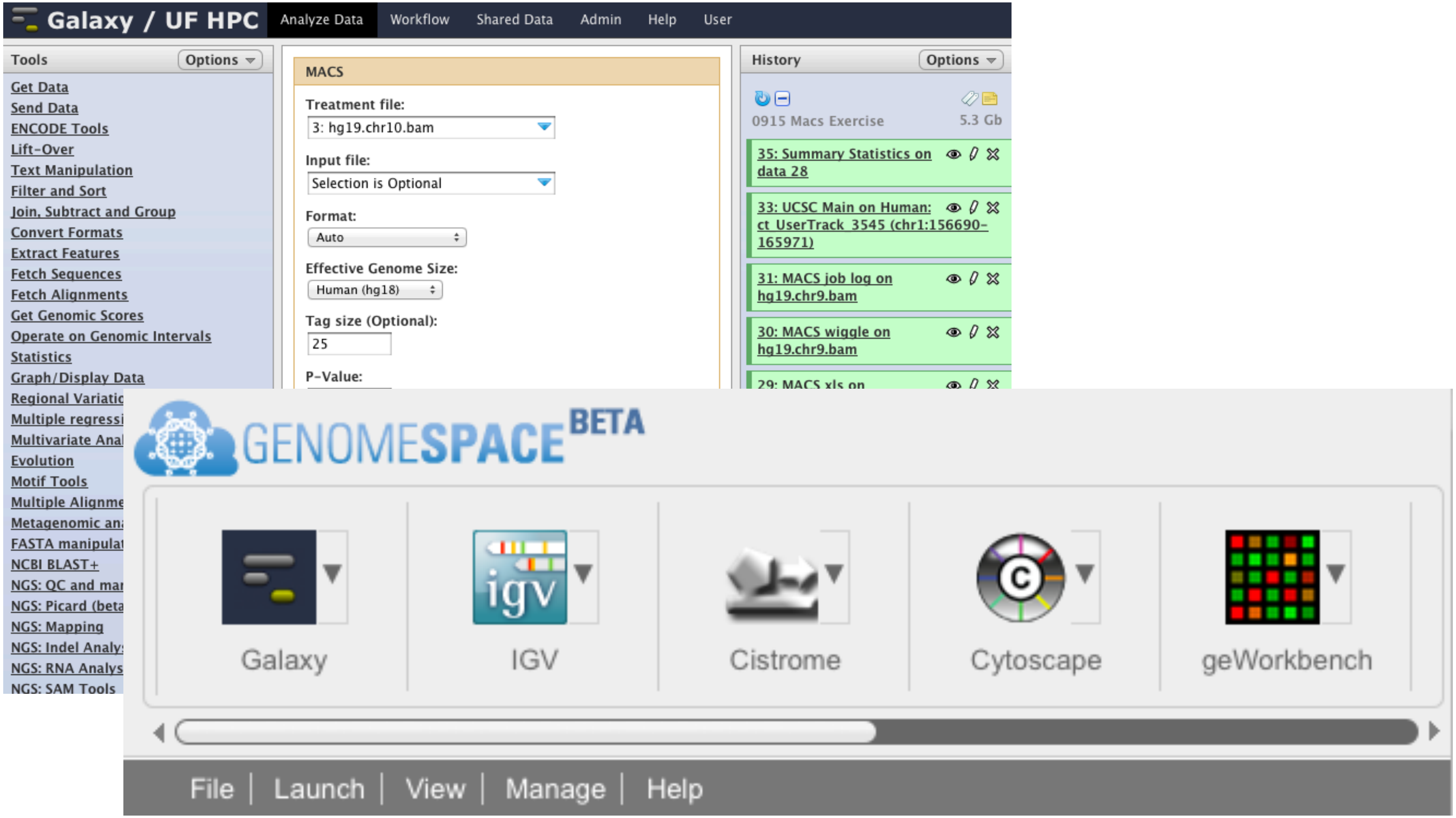
- Proprietary applications
  - Graphical User Interface
  - Integrate multiple tools, pipelines
  - User friendly-wizards for analyses
  - Many can tie into servers or clusters
  - Often highly optimized
  - Expensive
  - Limited flexibility
  - Limited scalability
  - Proprietary algorithms

CLC bio
Accelerating Scientific Research

NextGENe®
2nd Generation Sequence Analysis Software

Partek®
Genomics Suite™

Web Interfaces



Galaxy / UF HPC

Analyze Data | Workflow | Shared Data | Admin | Help | User

Tools — Options

- Get Data
- Send Data
- ENCODE Tools
- Lift-Over
- Text Manipulation
- Filter and Sort
- Join, Subtract and Group
- Convert Formats
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- Statistics
- Graph/Display Data
- Regional Variatic
- Multiple regressi
- Multivariate Anal
- Evolution
- Motif Tools
- Multiple Alignme
- Metagenomic ana
- FASTA manipulat
- NCBI BLAST+
- NGS: QC and mai
- NGS: Picard (beta
- NGS: Mapping
- NGS: Indel Analy:
- NGS: RNA Analys
- NGS: SAM Tools

MACS

Treatment file: 3: hg19.chr10.bam

Input file: Selection is Optional

Format: Auto

Effective Genome Size: Human (hg18)

Tag size (Optional): 25

P-Value:

History — Options

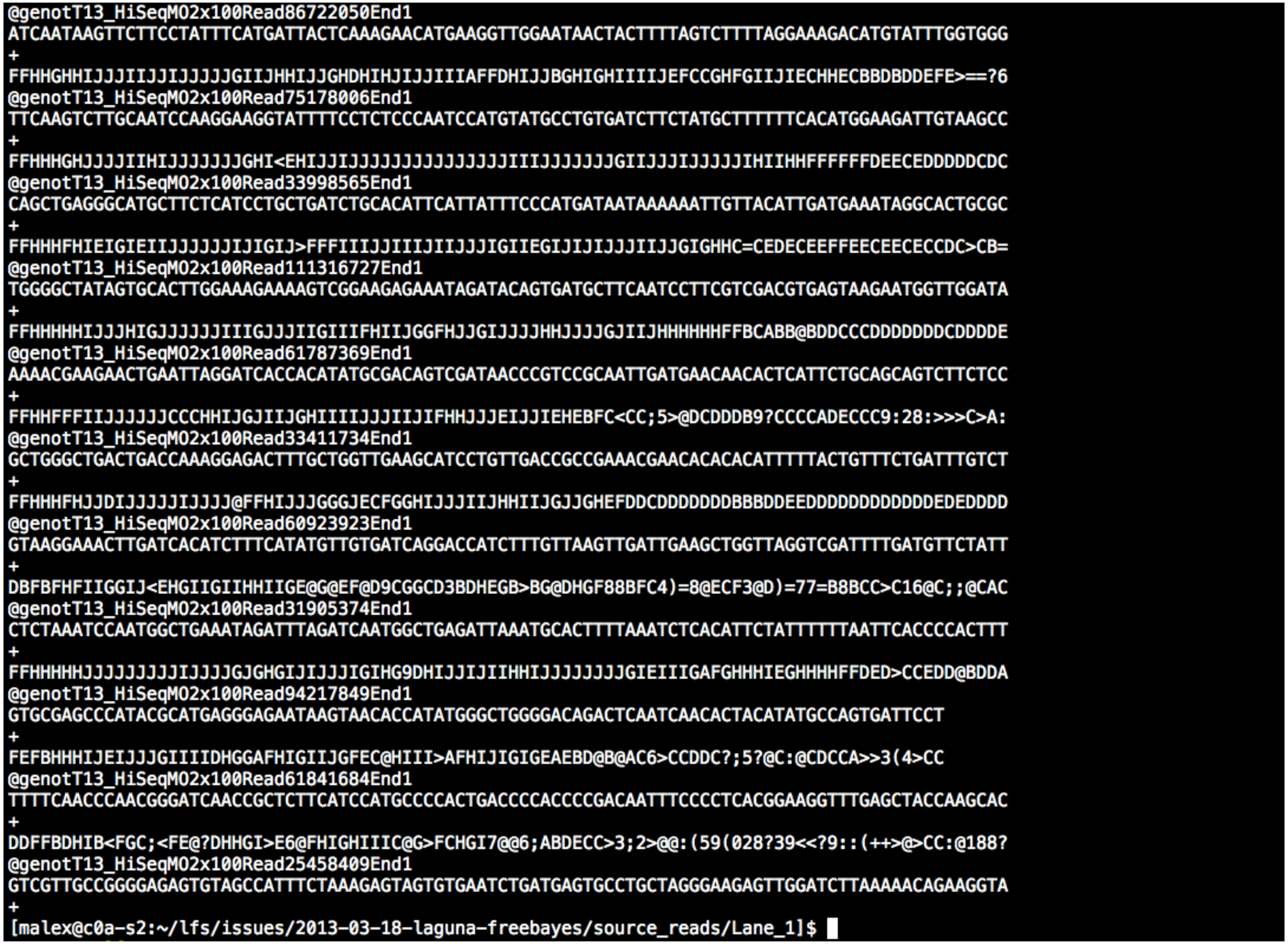
0915 Macs Exercise — 5.3 Gb

- 35: Summary Statistics on data 28
- 33: UCSC Main on Human: ct_UserTrack_3545 (chr1:156690-165971)
- 31: MACS job log on hg19.chr9.bam
- 30: MACS wiggle on hg19.chr9.bam
- 29: MACS xls on

Web Interfaces

- Galaxy
  - Free, Open Source
  - Public or private instance, physical or cloud-based
  - Web interface
  - Most applications can be integrated
  - User made pipelines
  - Moderately scalable
  - Integrating applications time consuming
  - User made pipelines—where to start? reliability?

Batch Processing

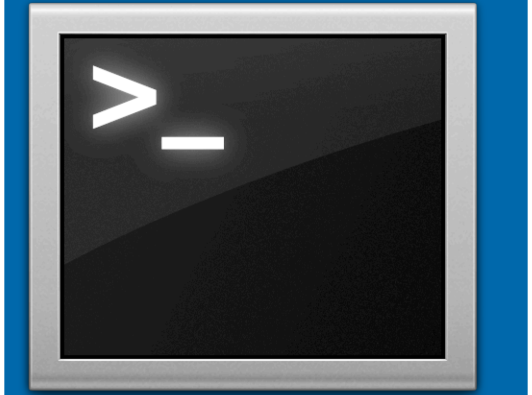


```
@genotT13_HiSeqM02x100Read86722050End1
ATCAATAAGTTCTTCCTATTTCATGATTACTCAAAGAACATGAAGGTTGGAATAACTACTTTTAGTCTTTTAGGAAAGACATGTATTTGGTGGG
+
FFHHGHHIJJJIIJJIJJJJJGIIJHHIJJGHDHIHJIJJIIIAFFDHIJJBGHIGHIIIIJEFCCGHFGIIJIECHHECBBDBDDEFE>==?6
@genotT13_HiSeqM02x100Read75178006End1
TTCAAGTCTTGCAATCCAAGGAAGGTATTTTCCTCTCCCAATCCATGTATGCCTGTGATCTTCTATGCTTTTTTCACATGGAAGATTGTAAGCC
+
FFHHHGHJJJJIIHIJJJJJJJGHI<EHIJJIJJJJJJJJJJJJJJJIIIJJJJJJGIIJJJIJJJJJIHIIHHFFFFFFDEECEDDDDDCDC
@genotT13_HiSeqM02x100Read33998565End1
CAGCTGAGGGCATGCTTCTCATCCTGCTGATCTGCACATTCATTATTTCCCATGATAATAAAAAATTGTTACATTGATGAAATAGGCACTGCGC
+
FFHHHFHIEIGIEIIJJJJJJIJIGIJ>FFFIIIJJIIIJIIJJJIGIIEGIJIJIJJJIIJJGIGHHC=CEDECEEFFEECEECECCDC>CB=
@genotT13_HiSeqM02x100Read111316727End1
TGGGGCTATAGTGCACTTGGAAAGAAAAGTCGGAAGAGAAATAGATACAGTGATGCTTCAATCCTTCGTCGACGTGAGTAAGAATGGTTGGATA
+
FFHHHHHIJJJHIGJJJJJJIIIGJJJIIGIIIFHIIJGGFHJJGIJJJJHHJJJJGJIIJHHHHHHFFBCABB@BDDCCCDDDDDDDCDDDDE
@genotT13_HiSeqM02x100Read61787369End1
AAAACGAAGAACTGAATTAGGATCACCACATATGCGACAGTCGATAACCCGTCCGCAATTGATGAACAACACTCATTCTGCAGCAGTCTTCTCC
+
FFHHFFFIIJJJJJJCCCHHIJGJIIJGHIIIIJJJIIJIFHHJJJEIJJIEHEBFC<CC;5>@DCDDDB9?CCCCADECCC9:28:>>>C>A:
@genotT13_HiSeqM02x100Read33411734End1
GCTGGGCTGACTGACCAAAGGAGACTTTGCTGGTTGAAGCATCCTGTTGACCGCCGAAACGAACACACACATTTTTACTGTTTCTGATTTGTCT
+
FFHHHFHJJDIJJJJJIJJJJ@FFHIJJJGGGJECFGGHIJJJIIJHHIIJGJJGHEFDDCDDDDDDDBBBDDEEDDDDDDDDDDDDEDEDDDD
@genotT13_HiSeqM02x100Read60923923End1
GTAAGGAAACTTGATCACATCTTTCATATGTTGTGATCAGGACCATCTTTGTTAAGTTGATTGAAGCTGGTTAGGTCGATTTTGATGTTCTATT
+
DBFBFHFIIGGIJ<EHGIIGIIHHIIGE@G@EF@D9CGGCD3BDHEGB>BG@DHGF88BFC4)=8@ECF3@D)=77=B8BCC>C16@C;;@CAC
@genotT13_HiSeqM02x100Read31905374End1
CTCTAAATCCAATGGCTGAAATAGATTTAGATCAATGGCTGAGATTAAATGCACTTTTAAATCTCACATTCTATTTTTTAATTCACCCCACTTT
+
FFHHHHHJJJJJJJJJIJJJJGJGHGIJIJJJIGIHG9DHIJJIJIIHHIJJJJJJJGIEIIIGAFGHHHIEGHHHHFFDED>CCEDD@BDDA
@genotT13_HiSeqM02x100Read94217849End1
GTGCGAGCCCATACGCATGAGGGAGAATAAGTAACACCATATGGGCTGGGGACAGACTCAATCAACACTACATATGCCAGTGATTCCT
+
FEFBHHHIJEIJJJGIIIIDHGGAFHIGIIJGFEC@HIII>AFHIJIGIGEAEBD@B@AC6>CCDDC?;5?@C:@CDCCA>>3(4>CC
@genotT13_HiSeqM02x100Read61841684End1
TTTTCAACCCAACGGGATCAACCGCTCTTCATCCATGCCCCACTGACCCCACCCCGACAATTTCCCCTCACGGAAGGTTTGAGCTACCAAGCAC
+
DDFFBDHIB<FGC;<FE@?DHHGI>E6@FHIGHIIIC@G>FCHGI7@@6;ABDECC>3;2>@@:(59(028?39<<?9::(++>@>CC:@188?
@genotT13_HiSeqM02x100Read25458409End1
GTCGTTGCCGGGGAGAGTGTAGCCATTTCTAAAGAGTAGTGTGAATCTGATGAGTGCCTGCTAGGGAAGAGTTGGATCTTAAAAACAGAAGGTA
+
[malex@c0a-s2:~/lfs/issues/2013-03-18-laguna-freebayes/source_reads/Lane_1]$ 
```

Batch Processing

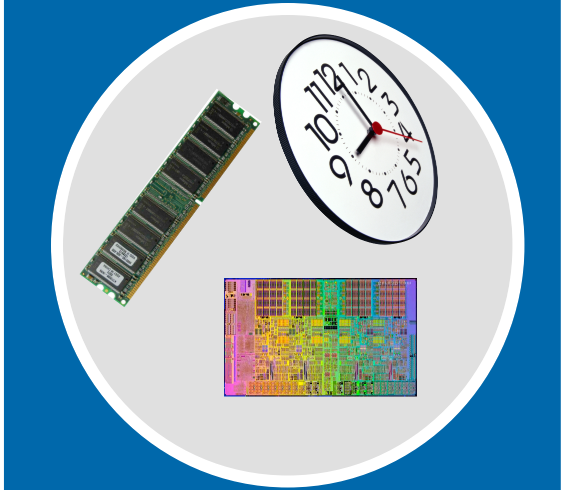
User interaction

Galaxy



Login node
(Head node)

Scheduler



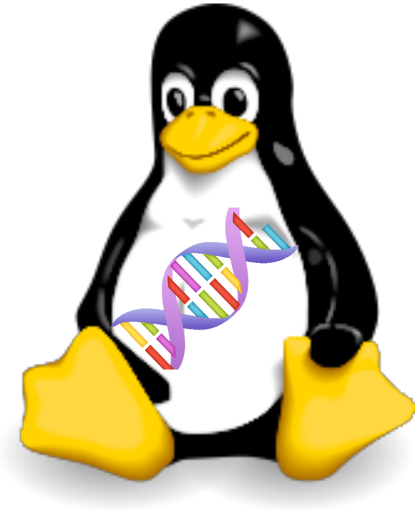
Tell the scheduler what you want to do

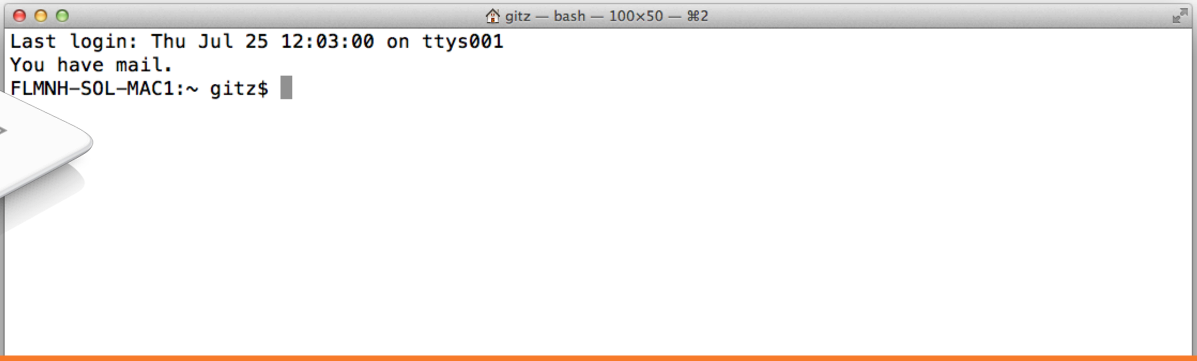
Compute resources

Your job runs on the cluster

Batch Processing

- The Linux Command Line
  - Maximum flexibility
  - Most informatics tools run under Linux
  - Write your own tool, or script
  - Maximum scalability
  - Learning barrier of entry





```
Last login: Thu Jul 25 12:03:00 on ttys001
You have mail.
FLMNH-SOL-MAC1:~ gitz$ 
```

Batch processing

- Submission Script

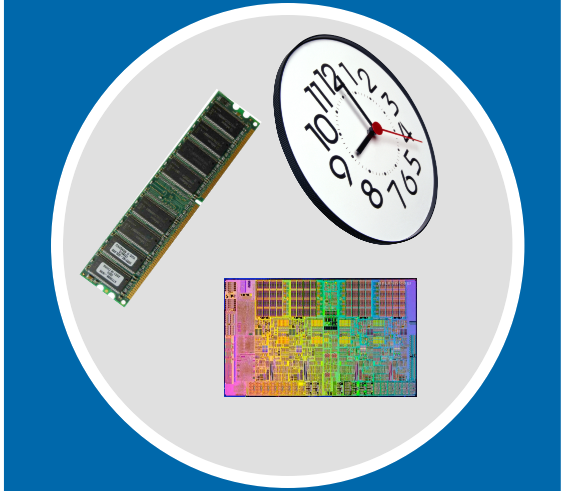
```
#!/bin/bash
#PBS -N My_Job_Name
#PBS -M Joe_Shmoe@ufl.edu
#PBS -m abe
#PBS -o My_Job.log
#PBS —e My_Job.err
#PBS -l nodes=1:ppn=1
#PBS -l walltime=00:05:00
#PBS —l pmem=900mb

cd $PBS_O_WORKDIR
date
module load test_app
test_app —i file.txt
```

Compute resources

Your job runs on the cluster

Scheduler



Tell the scheduler what you want to do

Accessing software via environment modules

- `module load trinity`

- Automatically:
  - Sets, `$HPC_TRINITY_DIR`
    - To run Inchworm, simply type

      ```
      inchworm --reads reads.fa --run_inchworm [opts]
      ```

  - Loads Bowtie and Allpaths, two Trinity dependencies
    - You don't need to hunt those down, or worry if they are in your path or not

It’s all in the software!

Matt Gitzendanner

UF Research Computing

Questions?

Thank you!