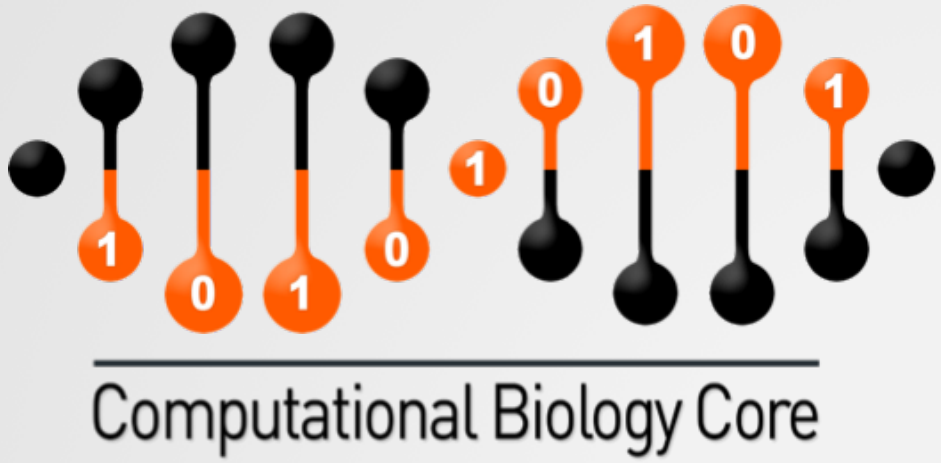
Institute for Systems Genomics

Computational Biology Core Open House 2018

Jill Wegrzyn

Vijender Singh
Neranjan Perera



Overview of Informatics Services

- Access to HPC
  - Software/Hardware Support
- Bioinformatics
  - Data Therapy
  - Consultations
  - Workshops
  - Full Project Support

CBC Team

- **Vijender Singh** – Lead Bioinformatics Scientist
- **Neranjan Perera** – Postdoctoral Scholar
- **Mike Wilson** – HPC and Software Support

- **Ion Moraru – Director UCH HPC Facility Director**
- **Stephen King** – HPC Administrator

HPC Access

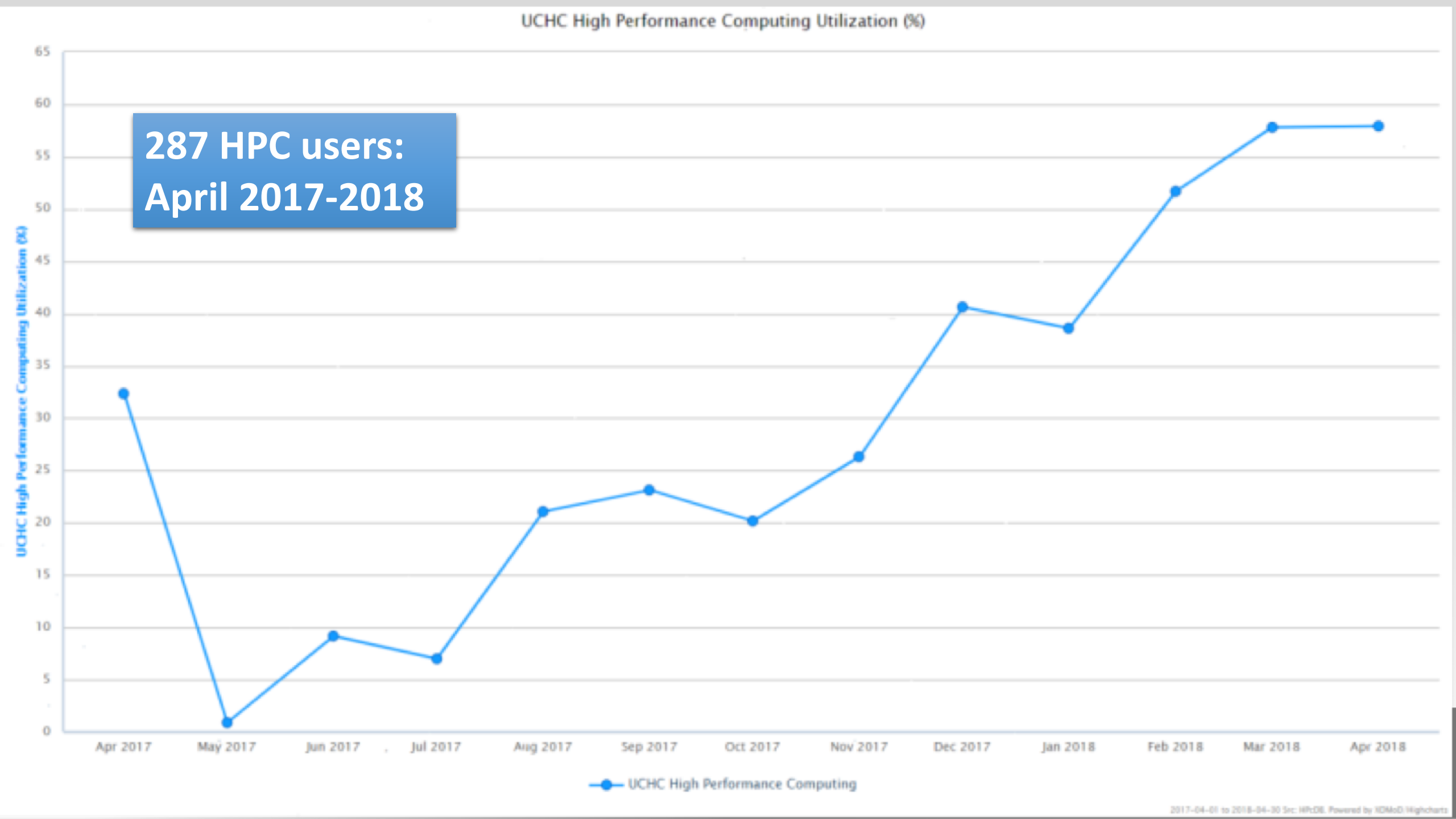
Type the commands

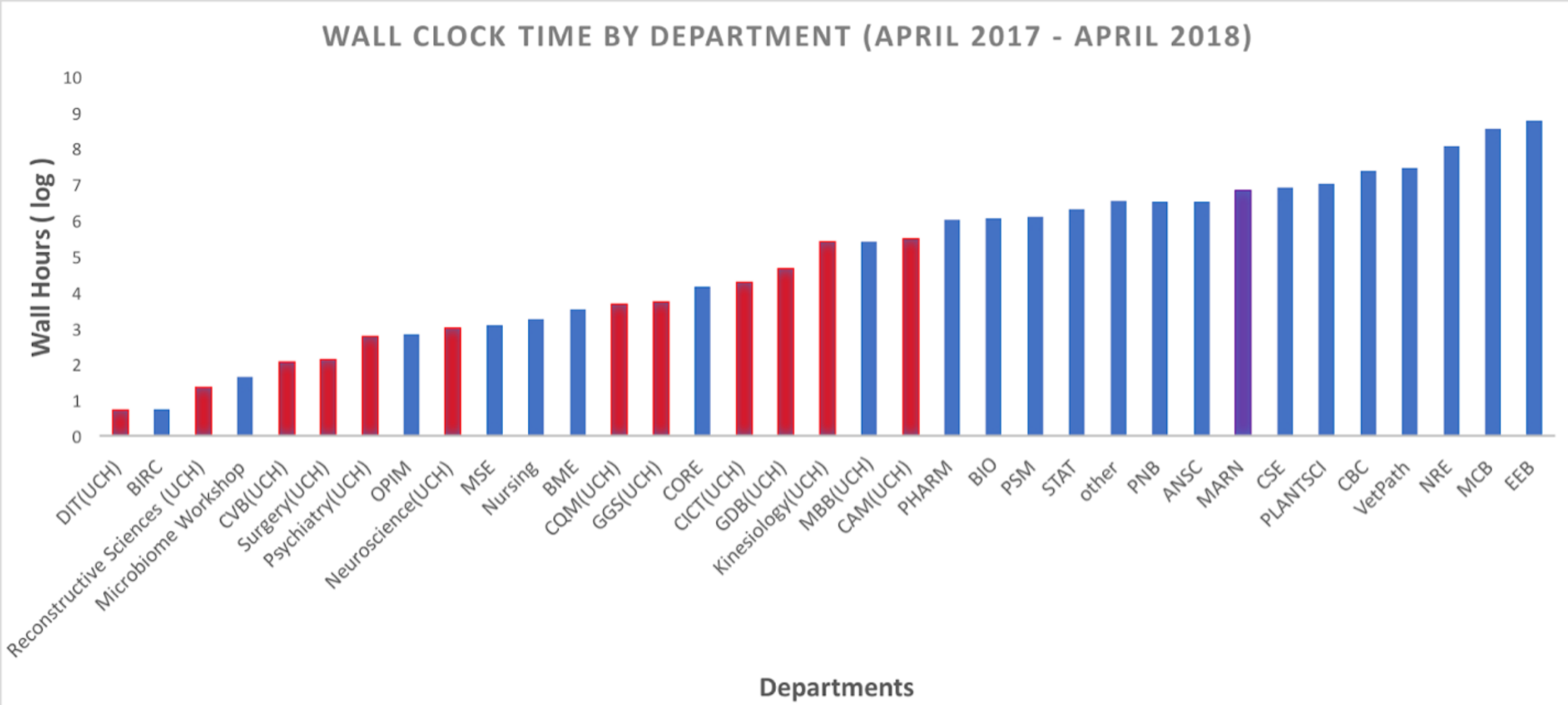
Actual work being done

- Large data is often too much for your small laptop or desktop to handle
- Move to large computers or clusters of computers
  - ie “remote” computers
- Vocabulary
  - “High performance computing” (HPC)

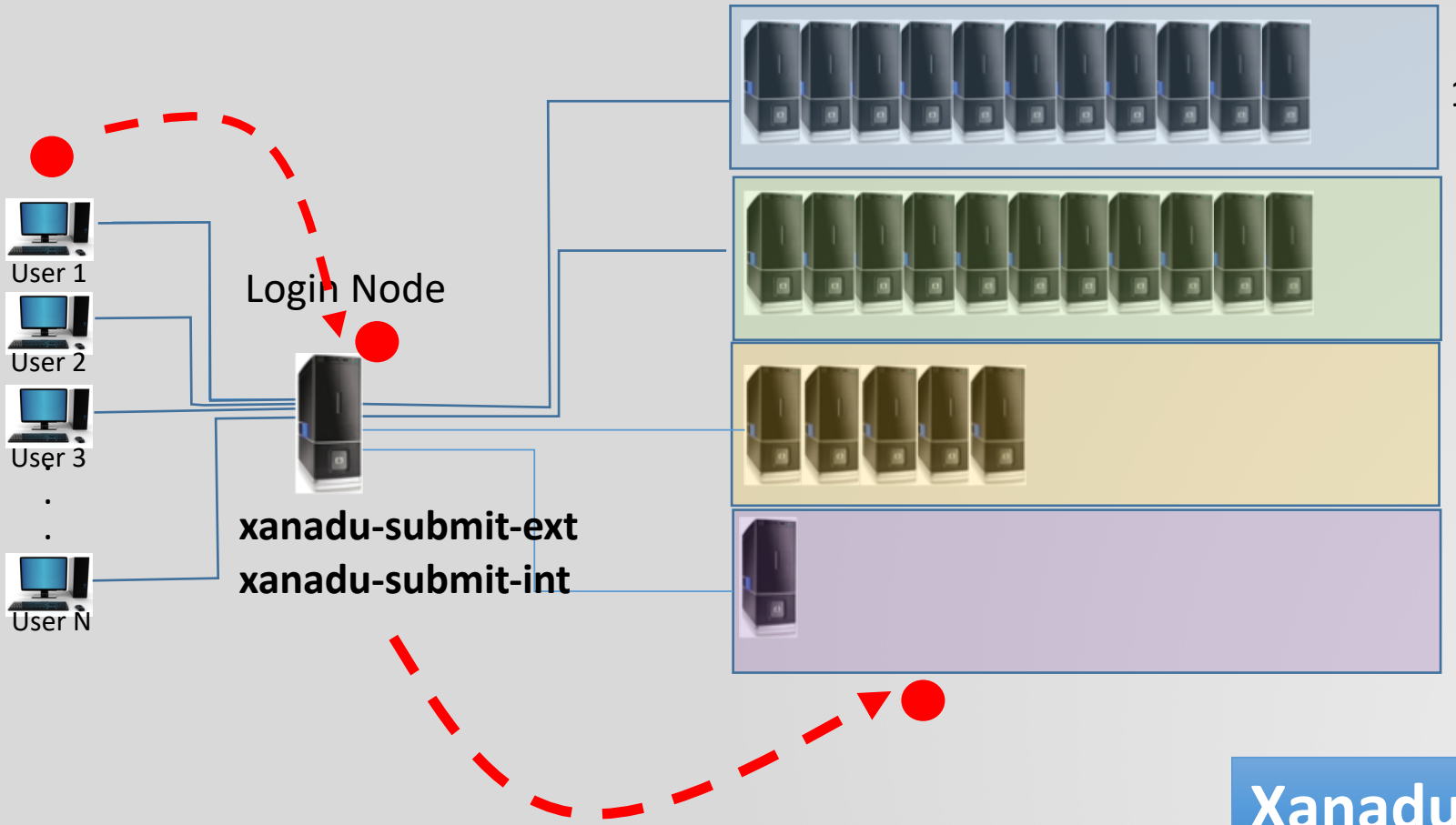
HPC Access

- CBC provides FREE, unlimited access to HPC computing
  - Unlimited account access for students, staff, postdocs, and faculty
  - Unlimited (and redundant) storage
- Provided through the Xanadu cluster
  - Administered by the UCH HPC Facility
  - Transition from BBC Cluster in 2018


UCHC High Performance Computing Utilization (%)
287 HPC users:
April 2017-2018
UCHC High Performance Computing Utilization (%)
65
60
55
50
45
40
35
30
25
20
15
10
5
0
Apr 2017
May 2017
Jun 2017
Jul 2017
Aug 2017
Sep 2017
Oct 2017
Nov 2017
Dec 2017
Jan 2018
Feb 2018
Mar 2018
Apr 2018
UCHC High Performance Computing


WALL CLOCK TIME BY DEPARTMENT (APRIL 2017 - APRIL 2018)
Wall Hours (log)
10
9
8
7
6
5
4
3
2
1
0
DIT(UCH)
BIRC
Reconstructive Sciences (UCH)
Microbiome Workshop
CVB(UCH)
Surgery(UCH)
Psychiatry(UCH)
OPIM
Neuroscience(UCH)
MSE
Nursing
BME
CQM(UCH)
GGS(UCH)
CORE
CICT(UCH)
GDB(UCH)
Kinesiology(UCH)
MBB(UCH)
CAM(UCH)
PHARM
BIO
PSM
STAT
other
PNB
ANSC
MARN
CSE
PLANTSCI
CBC
VetPath
NRE
MCB
EEB
Departments

HPC Access – Xanadu Cluster – 40 nodes > 1500 cores



11 **GENERAL/**AMD

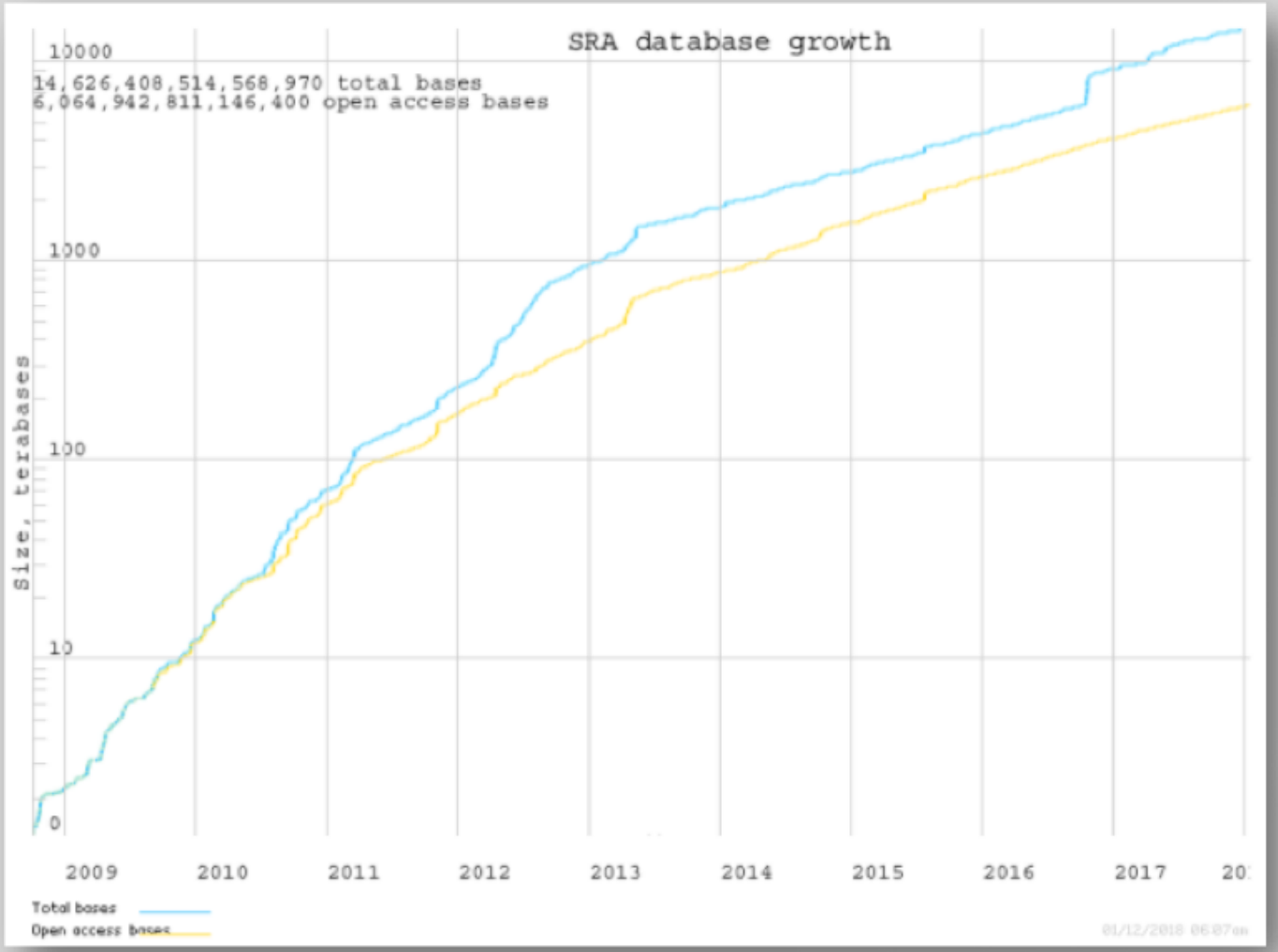
23 **GENERAL/**XEON

5 **HIMEM**

1 **GPU/XEON**

Xanadu Session at 12:30pm today!

HPC Access – Genomics Data



Big Data: Astronomical or Genomical?

Zachary D. Stephens, Skylar Y. Lee, Faraz Faghri, Roy H. Campbell, Chengxiang Zhai, Miles J. Efron, Ravishankar Iyer, Michael C. Schatz, Saurabh Sinha, Gene E. Robinson

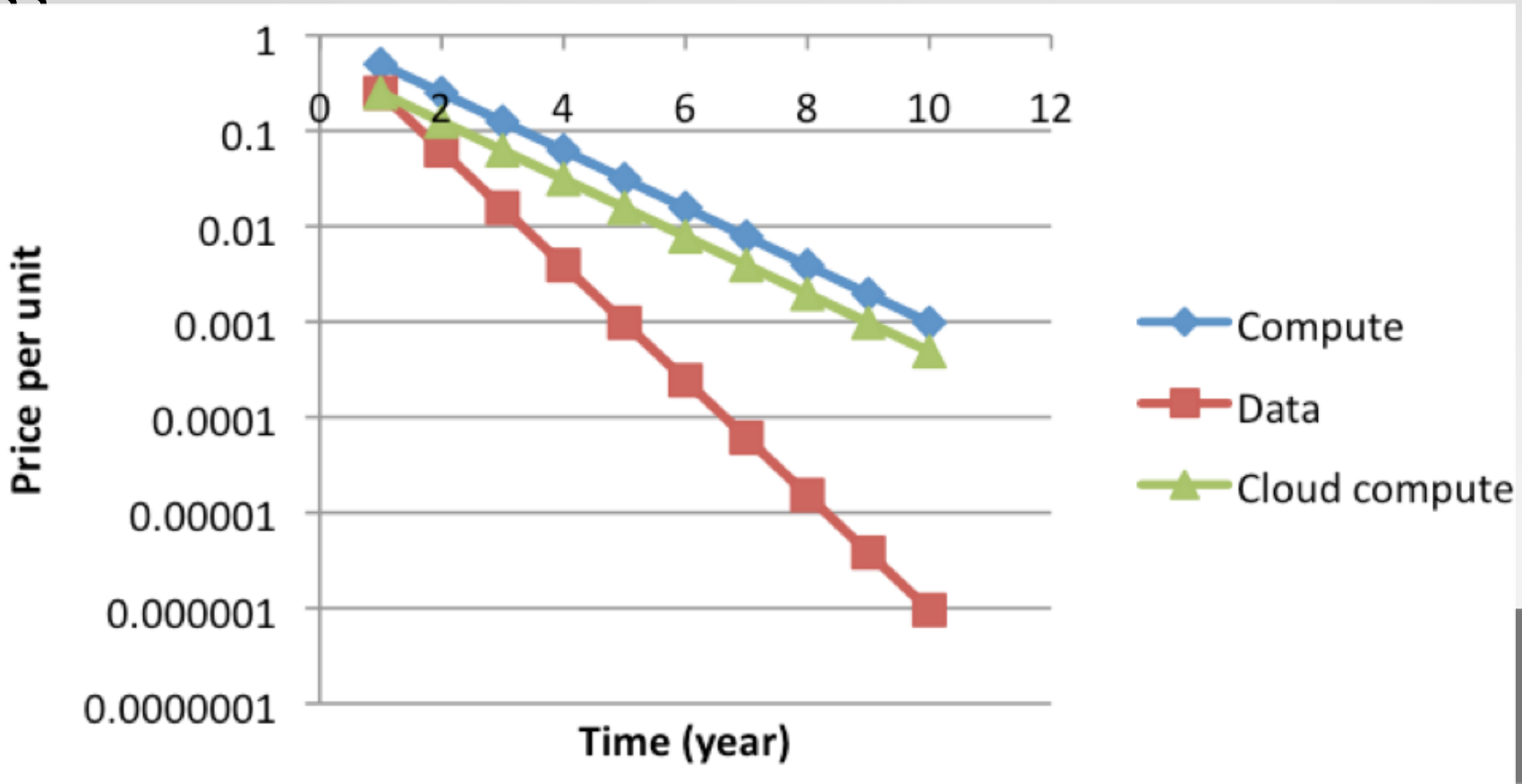
Published: July 7, 2015 • https://doi.org/10.1371/journal.pbio.1002195

By 2025! "Our estimates show that genomics is a "four-headed beast"—it is either on par with or the most demanding of the domains analyzed here in terms of data acquisition, storage, distribution, and analysis"

- YouTube
- Astronomy
- Twitter

HPC Access - Storage

- /linuxshare is available on Xanadu to provide archival storage
  - Not fast access
  - Redundant a
  - **Compressed**



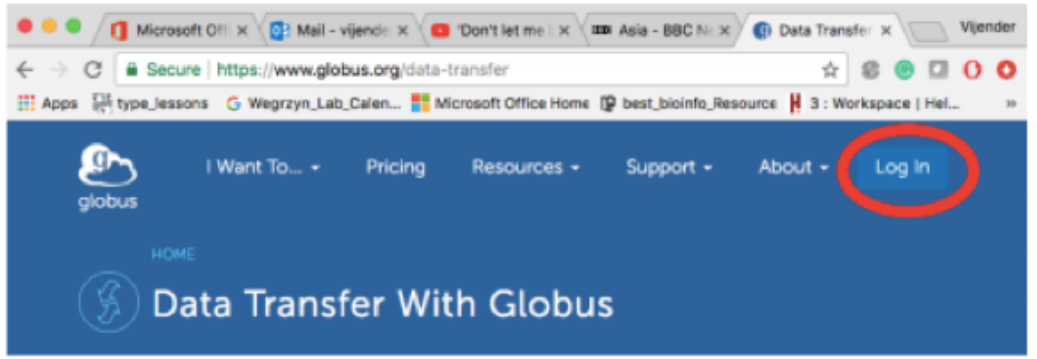
HPC Access – File Transfer

- File transfer can be challenging!
  - Globus endpoint at UCH
  - Globus endpoint on HPC Storrs
  - Drag and Drop
    - Local computer
    - FTP sites
    - Other servers
  - Not compatible?
    - Details on the website

GLOBUS FILE TRANSFER TUTORIAL

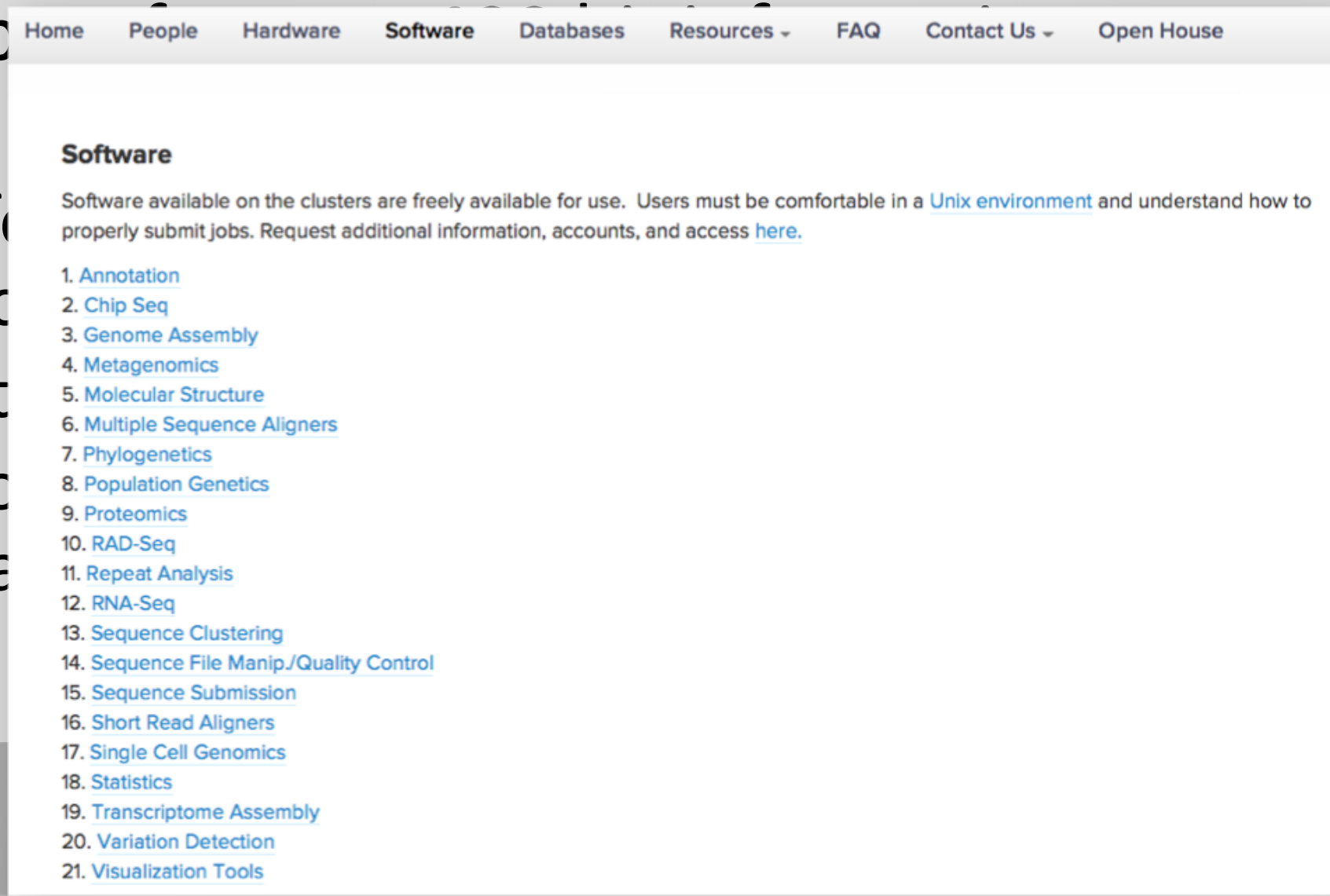
STEPS:

1. Go to https://www.globus.org/data-transfer and click on login option (right top).



Software Support

- CBC provides supp
  packages
  - Majority of bioinf
    command line acc
  - Users can request
  - **Module** system tc
    knowledge of loca

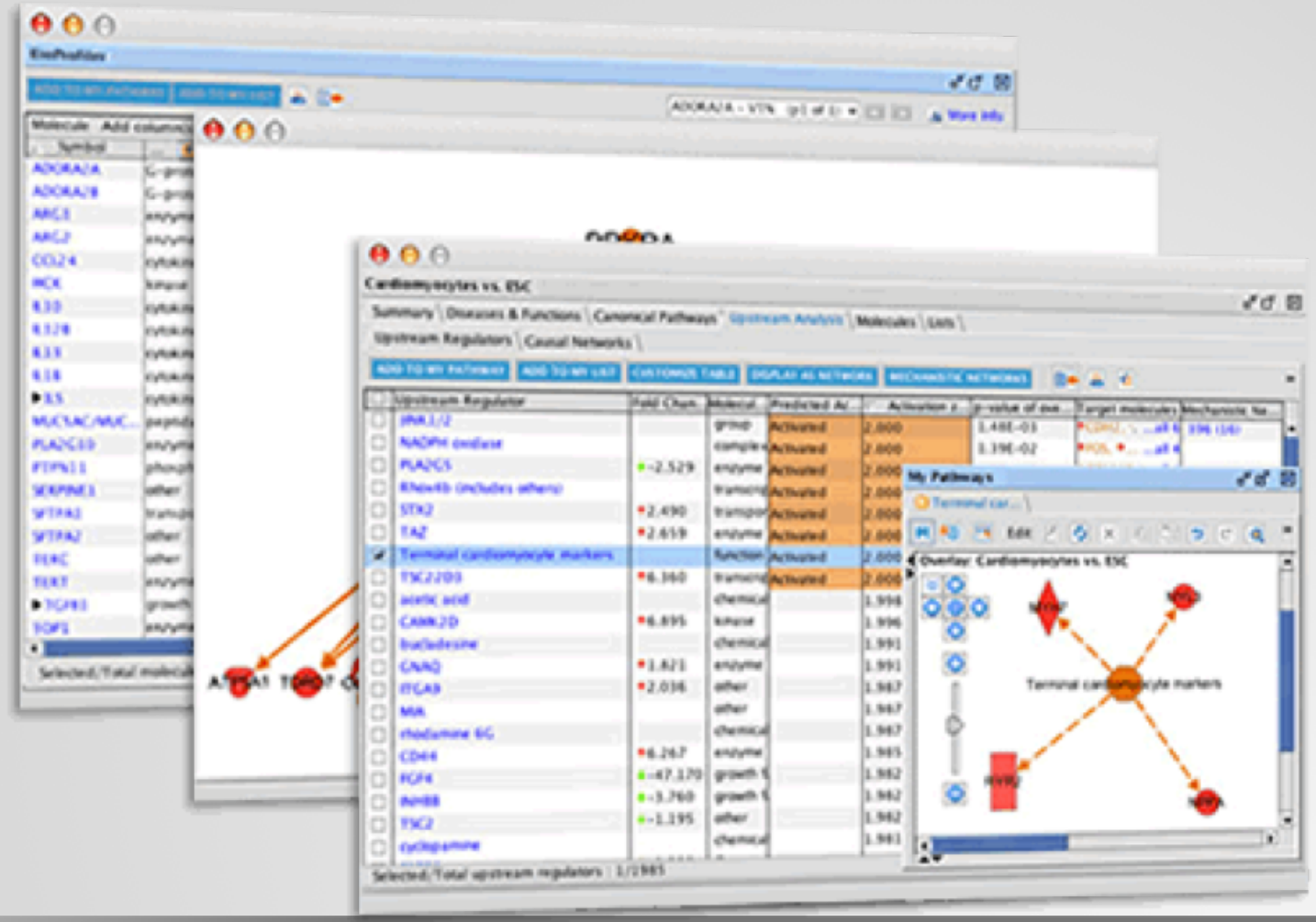


Software Support (Commercial Packages)

- CBC provides support for two commercial packages
  - **Geneious** package (server license)
    - Local instance or dedicated node with increased resources
    - Geneious Training Today!
  - **IPA**
    - Qiagen's Pathway Analysis Package
    - RNA-Seq to Pathway
    - Custom databases
    - Limited license (1 concurrent user + 5 licenses)

Software Support (Commercial Packages)

- Mechanistic Pathways
- Upstream Regulators
- Integrated Expression Results



Database Support

- Pre-Indexed Databases!
- NCBI
  - RefSeq (protein, rna, genomic)
  - NR
  - NT
- Ensembl
  - SwissProt
- Diamond (Protein)
- Genome Indexes
  - Short read aligners

Databases

NCBI BLAST database indexes are updated on a bi-weekly basis and available to users with an account on the cluster. Genome indexes for short read aligners are updated on request. Older versions are not maintained but if you require a specific version of a database or one not listed here, please contact us.

BBC cluster database path is indicated in **black** text while the paths relevant to the **Xanadu cluster** database is indicated in **purple** text.

All databases listed are current as of 04/02/2018

NCBI BLAST Databases

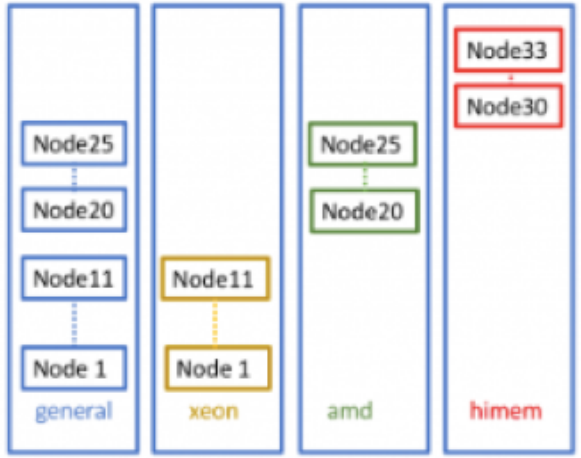
Database	Type	Path on Server	Description
nt	Nucleic	/common/blast/data /isg/shared/databases/blast	Nucleotide sequence database, with entries from all traditional divisions of GenBank, EMBL, and DDBJ excluding bulk divisions (gss, sts, pat, est, and htg divisions. wgs entries are also excluded. Not non-redundant.
nr	Protein	/common/blast/data /isg/shared/databases/blast	Non-redundant protein squence database with entries from GenPept, Swissprot, PIR, PDF, PDB and NCBI RefSeq
swissprot	Protein	/common/blast/data /isg/shared/databases/blast	Swiss-prot sequence databases (last major update), it's parent database is nr.
env_nt	Nucleic	/common/blast/data /isg/shared/databases/Blast	Nucleotide sequences for metagenome
env_nr	Protein	/common/blast/data /isg/shared/databases/blast	Nucleotide sequences for metagenome
tsa_nt	Nucleic	/common/blast/data /isg/shared/databases/blast	Sequences from the TSA division of GenBank, EMBL, and DDBJ
16SMicrobial	Nucleic	/common/blast/data /isg/shared/databases/blast	Bacterial and Archaeal 16S rRNA sequences from BioProjects 33175 and 33117
refseq_rna	Nucleic	/common/blast/data /isg/shared/databases/blast	NCBI Transcript reference sequences
refseq_protein	Protein	/common/blast/data /isg/shared/databases/blast	NCBI protein reference sequences.
refseq_genomic	Nucleic	/common/blast/data /isg/shared/databases/blast	NCBI genomic reference sequences
taxdb	Phylogeny	/common/blast/data /isg/shared/databases/blast	Taxonomy
cdd_delta	Nucleic	/common/blast/data /isg/shared/databases/blast	Conserved Domain Database sequences for use with stand alone deltablast

Accounts and Guides

- Request Account
  - Account needed for commercial software as well
- Xanadu Guide
  - What resources do I need?
  - Memory – fast access storage close to the CPU
  - CPU – processing units
  - Parallel applications?

HPC resources and limits

Xanadu cluster uses the Slurm, which is a highly scalable cluster management and job scheduling system for large and small Linux clusters. The nodes (individual nodes within the cluster) are divided into groups which are called partitions. Xanadu has several partitions available: general, xeon, amd, himem1, himem2, himem3, himem4.



Partitions on Xanadu

To look up the available partition information you can use 'sinfo -s' which will give you the current list:

```
$ sinfo -s
PARTITION AVAIL  TIMELIMIT   NODES(A/I/O/T)  NODELIST
general*     up   infinite       19/15/1/35  shangrila[01-18],xanadu-[01-11,20-25]
xeon         up   infinite        0/11/0/11  xanadu-[01-11]
amd          up   infinite        19/4/1/24  shangrila[01-18],xanadu-[20-25]
himem1       up   infinite          0/1/0/1  xanadu-30
himem2       up   infinite          0/1/0/1  xanadu-31
himem3       up   infinite          0/1/0/1  xanadu-32
himem4       up   infinite          0/1/0/1  xanadu-33
himem5       up   infinite          0/1/0/1  xanadu-29
```

In the above the general* is the default partition for the users. Where NODES(A/I/O/T) are a count of a particular configuration by node state in the form of "Available / Idle / Other / Total ".

Xanadu cluster is divided into six partitions:

- general partition
- himem1 partition
- himem2 partition
- himem3 partition
- himem4 partition
- himem5 partition

Consultations

- CBC offers FREE consultations to members of the research community (UCH on Tuesdays!)
  - Connecting to the server for the first time?
  - Troubleshooting submission scripts
  - Configuring software
  - Basic shell scripting to assist with large jobs (moving over hundreds of files)
  - Recommending workflows and approaches
  - Installing software on the cluster
  - Troubleshooting commands and determining parameters
  - Indexing databases

Consultations

- Form available on the website to ***submit a request to meet*** or troubleshooting on specific scripts
- **SLACK channel**
  - Anyone with a uchc.edu or uconn.edu address can join!
  - #General channel allows community response and feedback
  - Good place to check on server status
  - Ask general bioinformatics questions to the community (recommended parameters)
  - Possibly not for troubleshooting more complex issues – best to send a request
- ***All requests go into the Help Desk as a ticketed/trackable item***

Request Support for your Project

- CBC approved a **rate plan** in January 2018
  - Community members interested in full/partail support for their project
    - Data storage – compress and archive
    - Data quality control and submission to Genbank
    - Alignment -> ChiP-Seq or RNA-Seq
    - Variant detection -> GBS or RAD-Seq
    - Genome assembly and annotation (microbial or eukaryotic)
  - Custom bioinformatics plans developed as well

CBC: Tutorials

CBC Provides access to step by step tutorials!

- RNA-Seq
- CHiP-Seq
- Genome Assembly
- Variant Detection (coming soon!)

Tutorials

Tutorial	Last Updated	Description
Server Access		
UConn Health Cluster (PBS)	June 2015	Understanding the UConn Health Cluster
BBC Cluster (SGE)	June 2015	Understanding the BBC Cluster
UNIX and R		
Unix Basics	November 2013	Introduction to Command Line Operations
VIM	December 2013	Introduction to VIM UNIX Editor
Unix Examples	November 2013	Basic Bioinformatic Exercises
Introduction to R	November 2013	Basic Analysis and Plots
RNA-Seq Guides		
Prokaryote RNA-Seq (EDGE-pro/DESeq2)	July 2015	EDGE-pro tutorial (with Listeria reference genome)
Model Plant RNA-Seq (STAR/DESeq2)	July 2015	RNASeq tutorial (with Glycine max reference genome)
Non-Model Plant RNA-Seq (Bowtie2/eXpress/DESeq2	August 2015	RNA-Seq tutorial (with Picea rubens reference transcriptome)
Human RNA-Seq (no replicates) (STAR/DESeq2)	March 2015	Introduction to RNASeq
Model Insect RNASeq (Web-based Galaxy)	July 2015	RNASeq tutorial (with Drosophila reference genome)
Genome Assembly		
Genome Size Estimation Tutorial	January 2017	Genome Size Estimation Tutorial
Bacterial Genome Assembly Tutorial	September 2015	Genome Assembly tutorial

CBC: Tutorials

CBC Provides access to step by step tutorials!

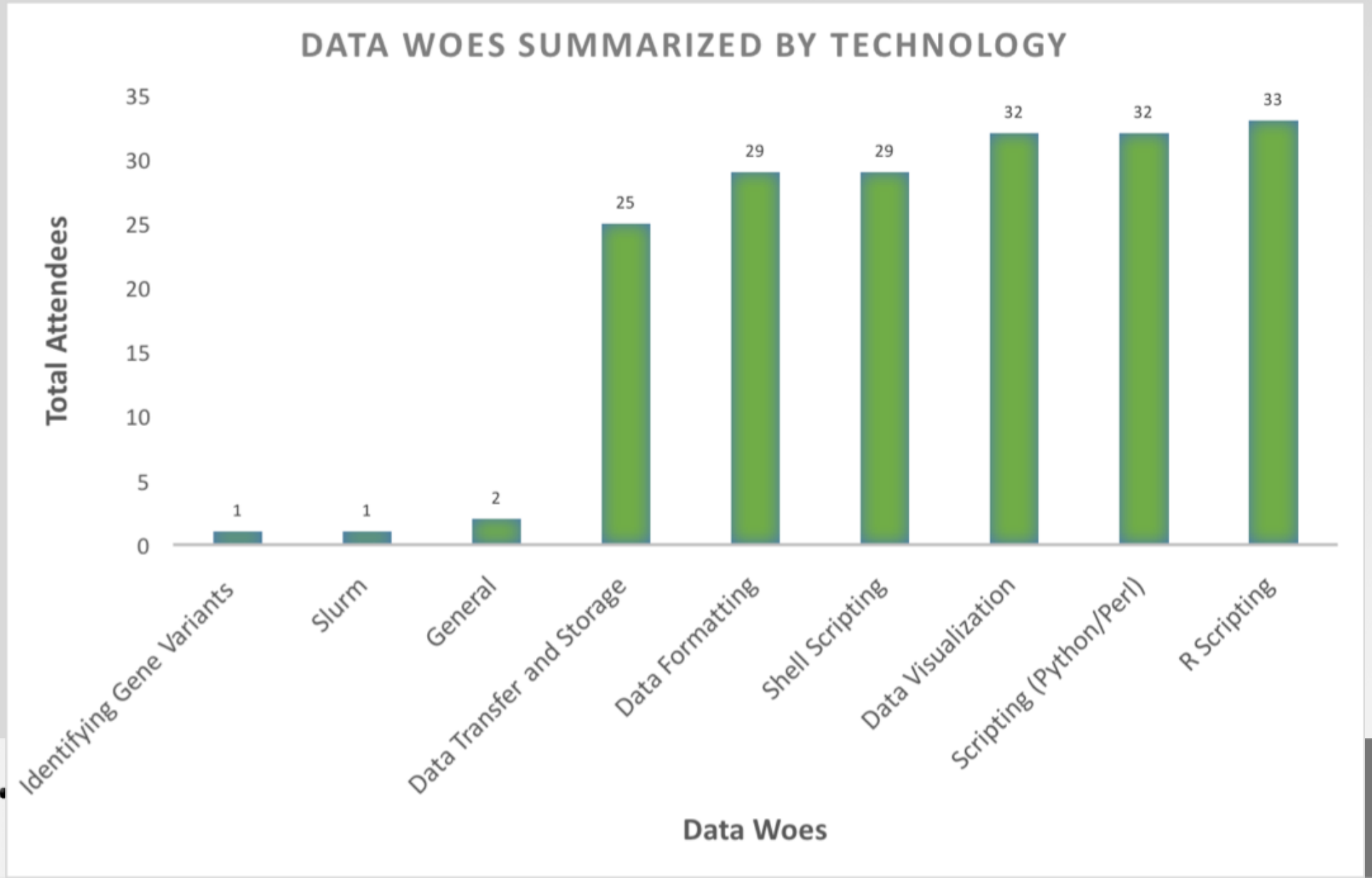
- RNA-Seq
- CHiP-Seq
- Genome Assembly
- Variant Detection (coming soon!)

RNA-Seq: Reference Genome, Differential Expression, and Functional Annotation

This repository is a usable, publicly available differential expression and functional annotation tutorial. All steps have been provided for the UConn CBC Xanadu cluster here with appropriate headers for the Slurm scheduler that can be modified simply to run. Commands should never be executed on the submit nodes of any HPC machine. If working on the Xanadu cluster, you should use sbatch scriptname after modifying the script for each stage. Basic editing of all scripts can be performed on the server with tools, such as nano, vim, or emacs. If you are new to Linux, please use this handy guide for the operating system commands. In this guide, you will be working with common bioinformatic file formats, such as FASTA, FASTQ, SAM/BAM, and GFF3/GTF. You can learn even more about each file format here. If you do not have a Xanadu account and are an affiliate of UConn/UCHC, please apply for one here.

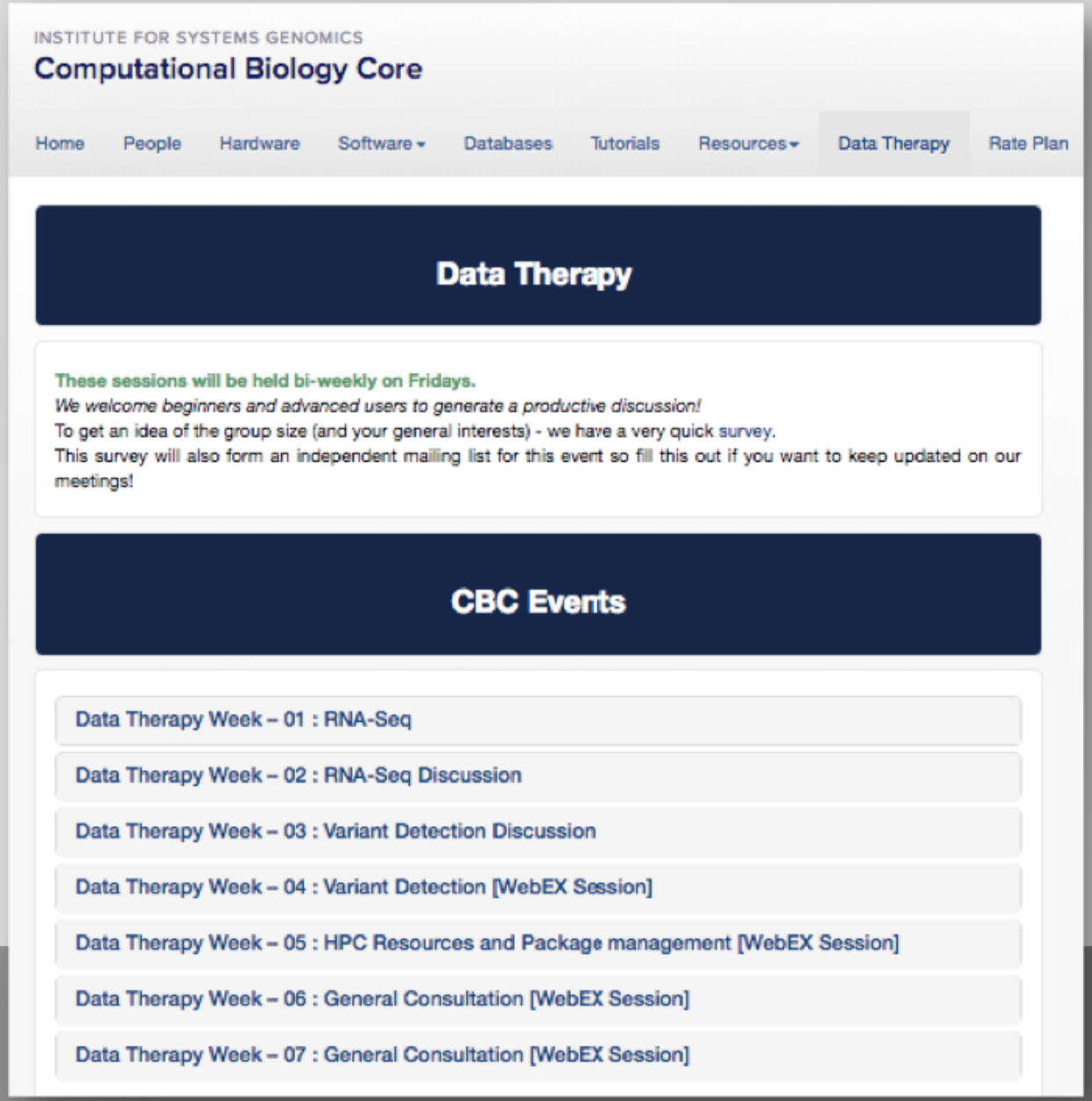
Contents

Data Therapy Sessions



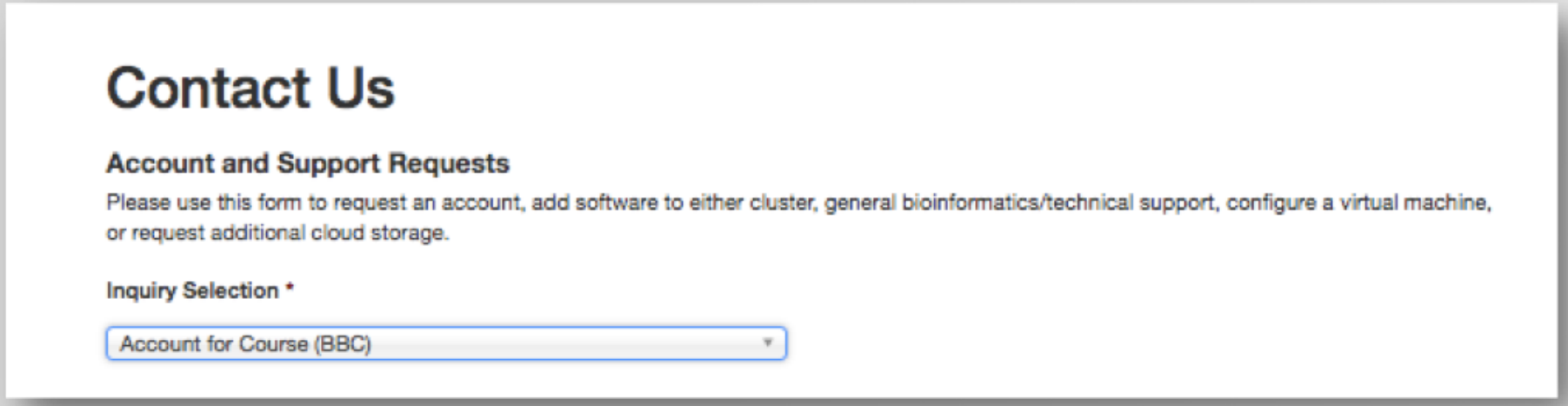
Data Therapy Sessions

- Bi-weekly discussion group at 10:30am in ESB 304
- Rotating Topics or Open Sessions
- Aimed at all levels
  - Encourage novice and advanced members for a productive discussion
- Metagenomics, Variant Detection, RNA-Seq



Education Mission

- Supporting Courses
- Tutorials in the Classroom
- Guest Lectures

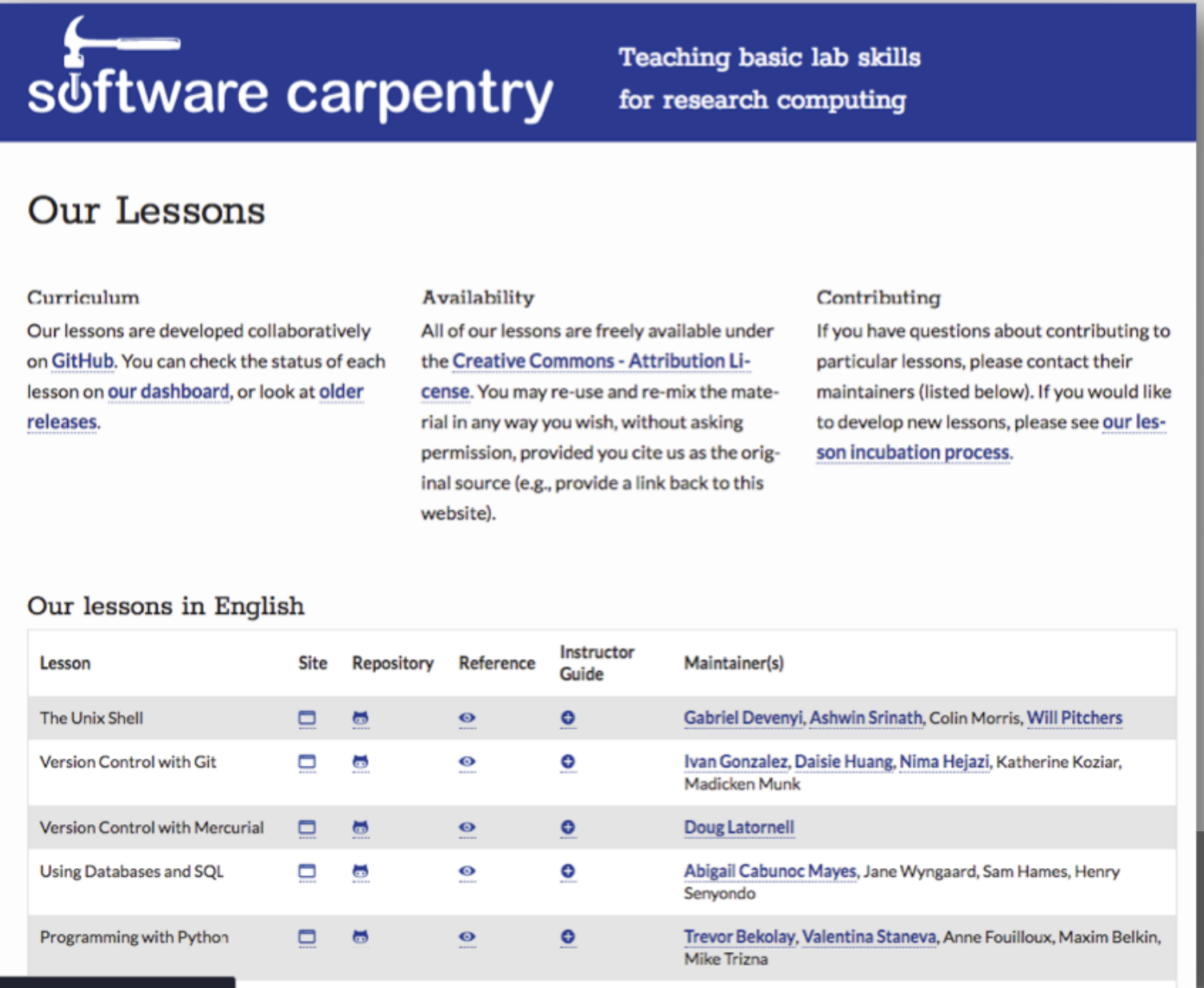


Workshop Offerings

CBC Workshops are offered as 2-3 day intensive sessions on topics of interest to researchers

- **RNA-Seq**
  - Experimental design
  - Quality control considerations
  - Alignment and de novo assembly strategies
- Next Workshop: **July 26-27, 2018 – Storrs Campus**

Software Carpentry



software carpentry

Teaching basic lab skills for research computing

Our Lessons

Curriculum

Our lessons are developed collaboratively on GitHub. You can check the status of each lesson on our dashboard, or look at older releases.

Availability

All of our lessons are freely available under the Creative Commons - Attribution License. You may re-use and re-mix the material in any way you wish, without asking permission, provided you cite us as the original source (e.g., provide a link back to this website).

Contributing

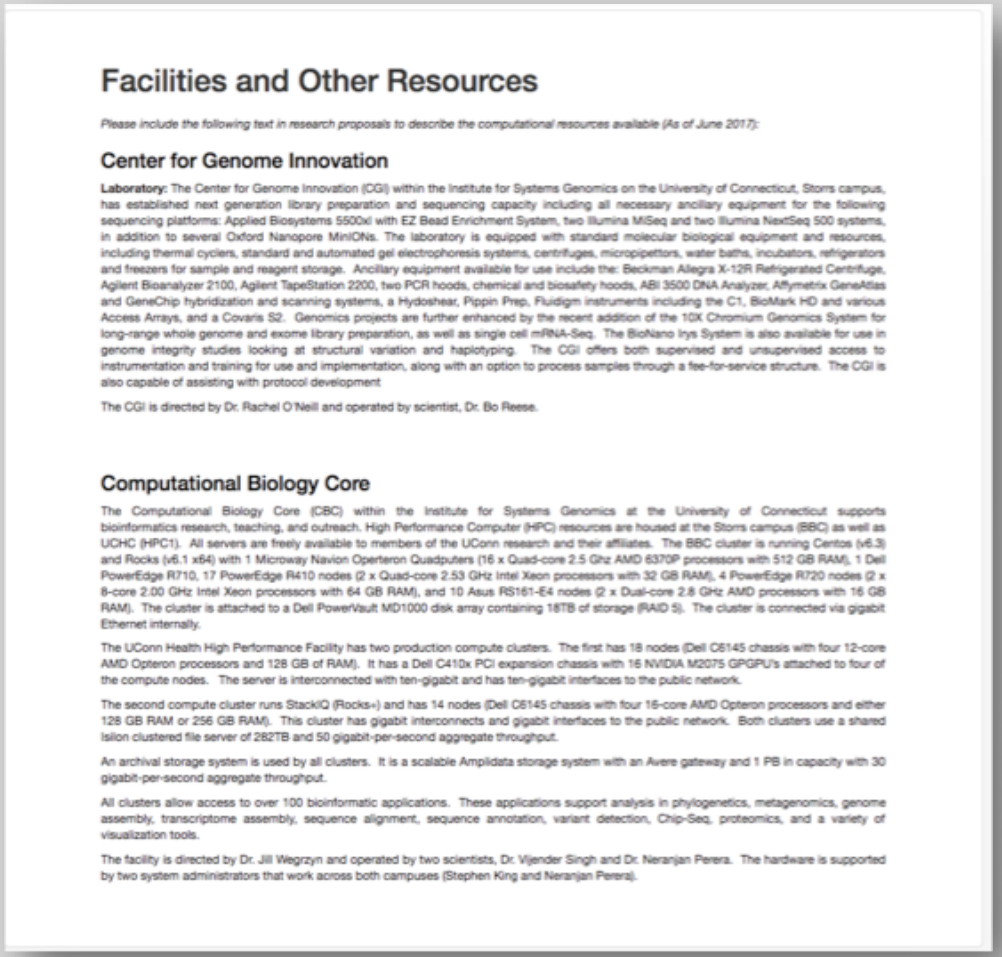
If you have questions about contributing to particular lessons, please contact their maintainers (listed below). If you would like to develop new lessons, please see our lesson incubation process.

Our lessons in English

Lesson	Site	Repository	Reference	Instructor Guide	Maintainer(s)
The Unix Shell					Gabriel Devenyi, Ashwin Srinath, Colin Morris, Will Pitchers
Version Control with Git					Ivan Gonzalez, Daisie Huang, Nima Hejazi, Katherine Koziar, Madicken Munk
Version Control with Mercurial					Doug Latornell
Using Databases and SQL					Abigail Cabunoc Mayes, Jane Wyngaard, Sam Hames, Henry Senyondo
Programming with Python					Trevor Bekolay, Valentina Staneva, Anne Fouilloux, Maxim Belkin, Mike Trizna

- Open Source Curriculum
- Community Developed (and taught!)
  - Git
  - R
  - Python
  - SQL
- Frequently offered on campus
- Almost always free!

Proposal Preparation



Facilities and Other Resources

Please include the following text in research proposals to describe the computational resources available (As of June 2017):

Center for Genome Innovation

Laboratory: The Center for Genome Innovation (CGI) within the Institute for Systems Genomics on the University of Connecticut, Storrs campus, has established next generation library preparation and sequencing capacity including all necessary ancillary equipment for the following sequencing platforms: Applied Biosystems 5500xl with EZ Bead Enrichment System, two Illumina MiSeq and two Illumina NextSeq 500 systems, in addition to several Oxford Nanopore MinIONs. The laboratory is equipped with standard molecular biological equipment and resources, including thermal cyclers, standard and automated gel electrophoresis systems, centrifuges, micropipettors, water baths, incubators, refrigerators and freezers for sample and reagent storage. Ancillary equipment available for use include the: Beckman Allegra X-12R Refrigerated Centrifuge, Agilent Bioanalyzer 2100, Agilent TapeStation 2200, two PCR hoods, chemical and biosafety hoods, ABI 3500 DNA Analyzer, Affymetrix GeneAtlas and GeneChip hybridization and scanning systems, a Hydoshear, Pippin Prep, Fluidigm instruments including the C1, BioMark HD and various Access Arrays, and a Covaris S2. Genomics projects are further enhanced by the recent addition of the 10X Chromium Genomics System for long-range whole genome and exome library preparation, as well as single cell mRNA-Seq. The BioNano Irys System is also available for use in genome integrity studies looking at structural variation and haplotyping. The CGI offers both supervised and unsupervised access to instrumentation and training for use and implementation, along with an option to process samples through a fee-for-service structure. The CGI is also capable of assisting with protocol development

The CGI is directed by Dr. Rachel O'Neill and operated by scientist, Dr. Bo Reese.

Computational Biology Core

The Computational Biology Core (CBC) within the Institute for Systems Genomics at the University of Connecticut supports bioinformatics research, teaching, and outreach. High Performance Computer (HPC) resources are housed at the Storrs campus (BBC) as well as UCHC (HPC1). All servers are freely available to members of the UConn research and their affiliates. The BBC cluster is running Centos (v6.3) and Rocks (v6.1 x64) with 1 Microway Navion Operteron Quadputers (16 x Quad-core 2.5 Ghz AMD 6370P processors with 512 GB RAM), 1 Dell PowerEdge R710, 17 PowerEdge R410 nodes (2 x Quad-core 2.53 GHz Intel Xeon processors with 32 GB RAM), 4 PowerEdge R720 nodes (2 x 8-core 2.00 GHz Intel Xeon processors with 64 GB RAM), and 10 Asus RS161-E4 nodes (2 x Dual-core 2.8 GHz AMD processors with 16 GB RAM). The cluster is attached to a Dell PowerVault MD1000 disk array containing 18TB of storage (RAID 5). The cluster is connected via gigabit Ethernet internally.

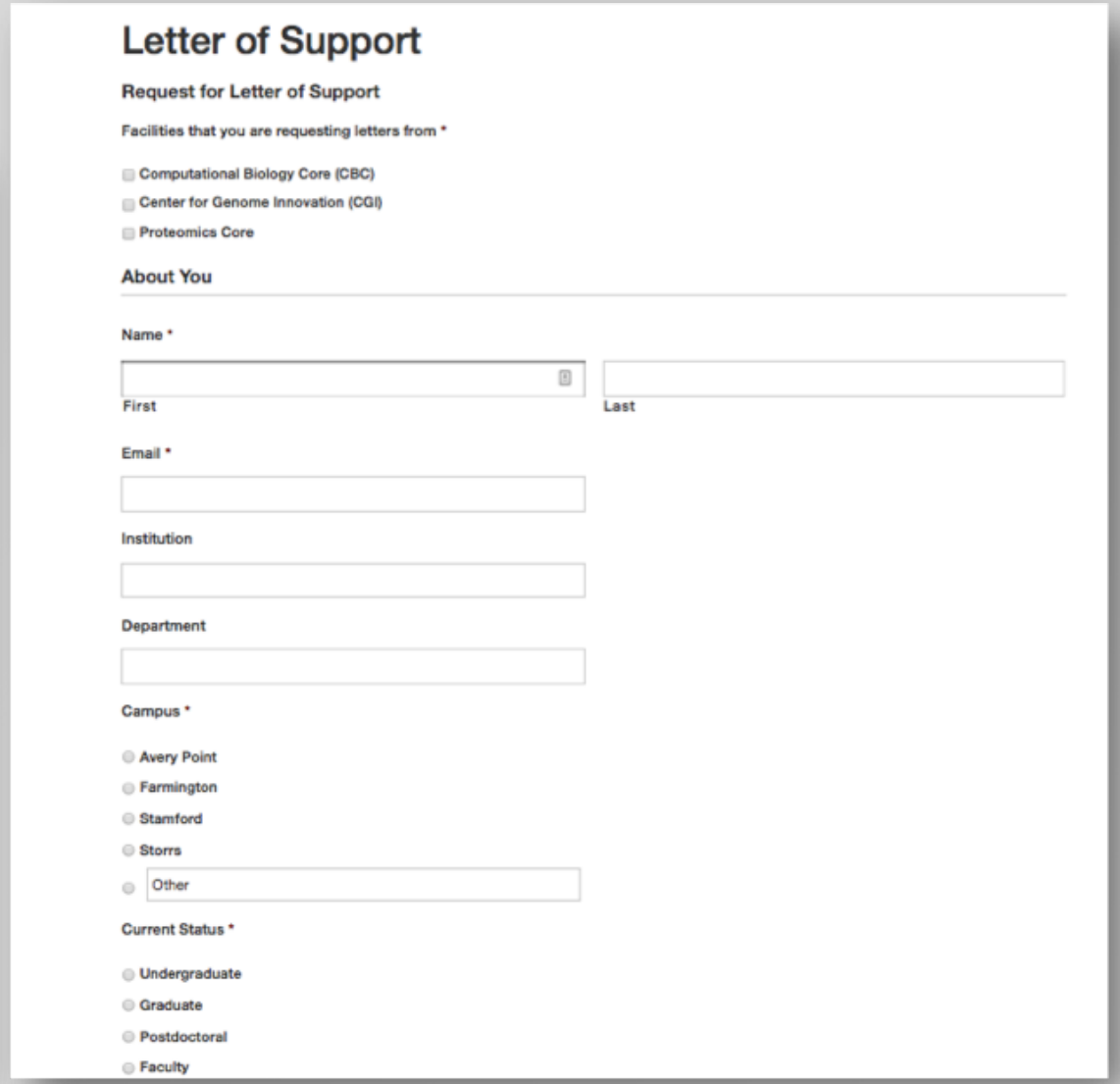
The UConn Health High Performance Facility has two production compute clusters. The first has 18 nodes (Dell C6145 chassis with four 12-core AMD Opteron processors and 128 GB of RAM). It has a Dell C410x PCI expansion chassis with 16 NVIDIA M2075 GPGPU's attached to four of the compute nodes. The server is interconnected with ten-gigabit and has ten-gigabit interfaces to the public network.

The second compute cluster runs StackIQ (Rocks+) and has 14 nodes (Dell C6145 chassis with four 16-core AMD Opteron processors and either 128 GB RAM or 256 GB RAM). This cluster has gigabit interconnects and gigabit interfaces to the public network. Both clusters use a shared Isilon clustered file server of 282TB and 50 gigabit-per-second aggregate throughput.

An archival storage system is used by all clusters. It is a scalable Amplidata storage system with an Avere gateway and 1 PB in capacity with 30 gigabit-per-second aggregate throughput.

All clusters allow access to over 100 bioinformatic applications. These applications support analysis in phylogenetics, metagenomics, genome assembly, transcriptome assembly, sequence alignment, sequence annotation, variant detection, Chip-Seq, proteomics, and a variety of visualization tools.

The facility is directed by Dr. Jill Wegrzyn and operated by two scientists, Dr. Vijender Singh and Dr. Neranjan Perera. The hardware is supported by two system administrators that work across both campuses (Stephen King and Neranjan Perera).



Letter of Support

Request for Letter of Support

Facilities that you are requesting letters from *

- ☐ Computational Biology Core (CBC)
- ☐ Center for Genome Innovation (CGI)
- ☐ Proteomics Core

About You

Name *

First | Last

Email *

Institution

Department

Campus *

- ○ Avery Point
- ○ Farmington
- ○ Stamford
- ○ Storrs
- ○ Other

Current Status *

- ○ Undergraduate
- ○ Graduate
- ○ Postdoctoral
- ○ Faculty

Stay Connected!

Online Support

Connect to the CBC team through

slack

#bioinformatics_help

uconn-cbc.slack.com

@Uconn_Bioinfo

Bioinformatics-l@Listserv.uconn.edu

Computational Biology Core

UCONN

UNIVERSITY OF CONNECTICUT