Cornell University
Institute of Biotechnology

Biotechnology Resource Center

Bioinformatics Facility

ABSTRACT

The Bioinformatics Facility of the Biotechnology Resource Center (BRC) provides state-of-the-art computational resources and analysis tools, and expertise in a their applications, to the university community and to outside investigators. The core facility offers research collaboration; bioinformatics support for next generation sequencing; software, database and website development; computational resources access; consultation; seminars, educational workshops and hands-on training.

OVERVIEW

Services Provided: The BRC Bioinformatics Facility (formerly called the Computational Biology Service Unit or CBSU) offers computational solutions for biological research, including software development and computational hardware infrastructure, research collaboration, consulting and training.

History: The facility was founded in 2001 as a computational resource for the Tri-institutional Collaboration of Cornell University / Weill Cornell Medical College, Rockefeller University, and Memorial Sloan-Kettering Cancer Center. The facility became part of the Cornell University Biotechnology Resource Center (BRC) in 2006. The facility was designated one of ten Microsoft High-Performance Computing Institutes worldwide in 2006. The Bioinformatics Facility has been a Microsoft Biology Initiative partner since 2010.

Administration: The Bioinformatics Facility is part of the Biotechnology Resource Center of the Cornell University Institute of Biotechnology.

Open to all: The resources and services of the facility are open to all investigators at Cornell University and Cornell-affiliated institutions. The facility also provides services to external investigators at both academic institutions and commercial enterprises. In association with the NY State Center for Advanced Technology (CAT) in Life Science Enterprise, NY State companies receive a discount on facility services.

RESOURCES

Facility computing resources are available via BioHPC Web Computing, BioHPC Computing Laboratory, or through collaborations.

Computing Nodes: 1 x 1024 GB RAM 64 core, 8 x 512 GB RAM 64 core, 2 x 512 GB RAM 96 core, 16 x 128 GB RAM 12 core, 1 x 64 GB RAM 12 core, 32 x 16 GB RAM 8 core, 60 x 4 GB RAM 4 core

Storage: 772 TB Lustre/Gluster file system, 30 TB Windows storage, 200 TB distributed storage on various Linux servers

Servers: 2 Windows web servers, 2 Linux web/SQL servers, 2 MS SQL servers, 6 Gluster component servers, 9 Lustre component servers

Computing Servers

- 3 Gen2 512 GB machines: 64 core (4 AMD CPUs), 512 GB RAM, 1 TB SSD fast storage, 9 TB regular SATA storage (4 x 3 TB RAID5)
- 5 Gen1 512 GB machines: 64 core (4 AMD CPUs), 512 GB RAM, 12 TB regular SATA storage (6 x 3 TB RAID6)
- 1 Gen3 512 GB machines: 64 core (4 Intel E5 CPUs), 1024 GB RAM, 16 TB regular SATA storage (6 x 4 TB RAID6)
- 2 Gen4 512 GB machines: 96 core (4 Intel E7 CPUs), 512 GB RAM, 1 TB SSD fast storage, 12 TB regular SATA storage (6 x 4 TB RAID10)
- 16 128 GB RAM machines: each 12 core (2 Intel CPUs), 128 GB RAM, 1 TB SSD fast storage, 4 TB regular SATA storage
- 4 workstations dedicated to sequence data processing: 1 MS Windows (16 GB RAM), 3 Linux (24 GB RAM)

Other Servers

- 5 file servers and 1 ftp server. File servers: 1 MS Windows (total 30 TB of storage) and 5 Linux (total of 200 TB of storage). Ftp server capacity is 6.5 TB.
- 3 general purpose MS Windows servers for web and data processing
- 2 MS SQL database servers. MS SQL Server, 1-8 TB HDD storage each
- 7 general purpose Linux servers: three 8 GB RAM machines, two 16 GB RAM machines, one 32 GB machine, one 48 GB RAM machine

Clusters

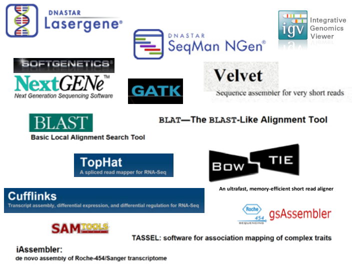
- 60 node 240 core Windows cluster. Dell PowerEdge 1855 nodes with two x64 Pentium 4 Xeon 3.4 GHz, 4 GB RAM and 144 GB HD
- 32 node 128 core Linux cluster. Dell PowerEdge M600 blade nodes with two quad-core Intel Xeon E5420 2.5 GHz, 16 GB RAM and 300 GB HD

BioHPC Computing Laboratory

Cloud Computing Resource for Life Sciences Data Analysis

Remote or On-site Access

Available software includes:


DNASTAR Lasergene
DNASTAR SeqMan NGen
igv Integrative Genomics Viewer
SOFTGENETICS NextGENe Next Generation Sequencing Software
GATK
Velvet
Sequence assembler for very short reads
BLAST
Basic Local Alignment Search Tool
BLAT—The BLAST-Like Alignment Tool
TopHat
A spliced read mapper for RNA-Seq
BOW TIE
An ultrafast, memory-efficient short read aligner
Cufflinks
Transcript assembly, differential expression, and differential regulation for RNA-Seq
Roche 454 SEQUENCING gsAssembler
SAMTOOLS
TASSEL: software for association mapping of complex traits
iAssembler:
de novo assembly of Roche-454/Sanger transcriptome

The BRC Bioinformatics Facility's BioHPC Computing Laboratory is a cloud computational resource configured for biologists. The Lab is targeted for biologists who want to learn Linux or Windows operating system, and to do bioinformatics data analysis themselves.

The Lab provides investigators with direct access to a wide range of bioinformatics data analysis software tools on appropriate hardware platforms. The available hardware ranges from small 8-core 16 GB RAM blades to large 96-core 512 GB RAM machines to 64-core 1024 GB RAM server.

There are over 280 bioinformatics software titles installed and ready for BioHPC Lab users. New software can be added by request. Development tools are also available for deploying custom programs.

Workstations are reserved using an online calendar-based scheduler; only registered Lab users can make reservations. All workstations are accessible remotely with ssh or VNC; selected workstations are accessible on-site.

BioHPC Computing Lab has over 770 TB of network storage. Each registered active user gets 200 GB storage space; users can purchase additional storage. Data can be transferred in and out of the Lab with sftp/scp or Globus. Data can be shared with external users using Globus.

The Bioinformatics Facility offers workshops introducing biologists to the Linux computing environment. Users can also talk to us during our office hours (Monday and Thursday, 1:00-3:00 pm, appointment required) for consultation and training.

Research groups can also have their own workstations hosted in the lab. Hosted workstations are maintained by the facility staff but are accessible only to the host group's members.

Next Generation Sequencing Data Analysis and Data Management Module

NGS data automatically transferred from Genomics Facility and made available to users for analysis

Users can upload and use their reference data for sequence analysis

- **Data Management System** imports large sequence data files directly from the sequencing instrument pipeline, cataloguing and storing them for subsequent use
- **Analysis Software** for next generation sequence alignment and post-processing
- **Analysis Pipeline Management System** allows users to execute a series of software applications on imported sequence data without additional data transfers

Bioinformatics Support for Next Generation Sequencing

Next generation sequencing (NGS) data analysis and data management services include support for :

- whole genome sequencing
- targeted region sequencing
- genotyping-by-sequencing (GBS)
- whole genome shotgun genotyping
- transcriptome profiling
- digital gene expression
- small RNA discovery and analysis
- epigenomics profiling (ChIP-Seq and DNA methylation)

Customized support for:

- data analysis
- development of customized software tools
- building customized data analysis pipelines

The facility staff can help investigators with NGS data analysis using available software and can also develop customized software tools.

Software, Database & Website Development

Analysis Software and Pipelines for Diverse Data Types

Laboratory Information Management Systems (LIMS) for Research Projects:

Integrate diverse data sets
such as genomics, proteomics, & imaging information

Examples of this type of service include:

- Plant Proteome Database is both a LIMS and a public outreach /publication tool for *Arabidopsis thaliana* and maize (*Zea mays*).
  Developed in collaboration with Klass Van Wijk at Cornell University.

- PathogenTracker is a LIMS for bacterial biodiversity & strain diversity studies.
  Developed in collaboration with Kathryn Boor and Martin Wiedmann at Cornell University.

- Panzea Database contains genotypic and phenotypic data of maize and teosinte.
  Developed in collaboration with Ed Buckler at Cornell University and USDA-ARS.

- Human Clinical Database for Hepatitis C integrates clinical data from hospital databases, basic research, genotyping and biobanking, in support of medical research, with HIPPA compliance and with support for data collection by physicians at the point of care.
  Developed in collaboration with Andrew Talal at the Weill Cornell Medical College.

- T-REX is an online LIMS for storage, management, and analysis of Terminal Restriction Fragment Length Polymorphism (T-RFLP) data.
  Developed in collaboration with Steve Culman and Dan Buckley at Cornell University.

The Bioinformatics Facility offers software, database and website development for both basic research and clinical research. The facility supports analysis software and develops analysis pipelines for the diverse data types generated by all the BRC core facilities. The facility also provides design, development and hosting of Laboratory Information Management Systems (LIMS), including: (1) designing database systems for specific research problems; (2) developing interfaces for data access, storage and analysis; and (3) deploying these applications on the core facility's computing resources. The facility can help researchers integrate diverse data sets, such as genomics, proteomics, and imaging information.

Research Collaboration

Examples of research collaboration projects include:

- **Biology of rare alleles in maize and its wild relatives.** This project is studying the relationship of rare alleles to fitne related traits across a diverse range of Zea germplasm.

- **A systems approach to photosynthesis.** The goals of this project are to study C4 leaf developmental stages and to utilize novel computational approaches for data integration and regulatory network modeling.

- **Genome-wide impact of mPing transposition on rice phenotypic diversity.** The aim of this project is to develop a functional genomics approach to determine the contribution of rice transposable elements to gene and genome evolution.

- **Accelerating grape cultivar improvement via phenotyping centers and next generation markers.** This project is accelerating grape cultivar improvement by providing cutting-edge molecular marker technologies, rigorous centralized phenotyping and molecular breeding support.

- **A high-resolution map of recombination in maize.** The goal of this project is to generate the first comprehensive, high-density map of recombination in maize.

- **Food Microbe Tracker.** The facility has worked with Cornell Food Safety Laboratory to develop and support the web based system for information exchange on bacterial subtypes and strains and for studies on bacterial biodiversity and strain diversity.

Research collaboration services are available as either (a) short-term fee-based service contracts that provide full bioinformatics solutions for research projects or (b) long-term project-based collaborations funded through collaborative research grants. The facility can work with investigators to apply for joint funding. The facility provides collaborative support for multidisciplinary projects in close coordination with the BRC genomics, genomic diversity, proteomics and mass spectrometry, imaging, bio-IT, and advanced technology assessment facilities.

Consultation, Workshops and Training

Examples of recent workshops:

- Linux for biologists
- Perl for biologists
- Variant Calling
- Transcriptome assembly
- RNA-Seq Data Analysis
- Gene Function Annotation
- Genome Assembly
- ChIP-Seq Data Analysis and Motif Calling

Consultation on project design and data analysis available upon request. The BRC Bioinformatics Facility and Genomics Facility provide coordinated joint consultation services for next generation sequencing projects. This includes consultation on the selection and optimal application of software tools for data analysis.

Bioinformatics software analysis tools and training available through the facility.

Educational workshops and training on data analysis available through the facility.

Coordinated project design consultation and data analysis support available with the BRC genomics, genomic diversity, proteomics and mass spectrometry, imaging, bio-IT, and advanced technology assessment core facilities.

Contact Information

Biotechnology Resource Center (BRC)

Bioinformatics Facility
Jaroslaw Pillardy, Director
Qi Sun, Co-Director
BRC_bioinformatics@cornell.edu
www.biotech.cornell.edu/brc/bioinformatics

*For questions about the **Biotechnology Resource Center**, please contact George Grills at BRC_director@cornell.edu*

www.biotech.cornell.edu

NYSTAR designated
Center For Advanced Technology