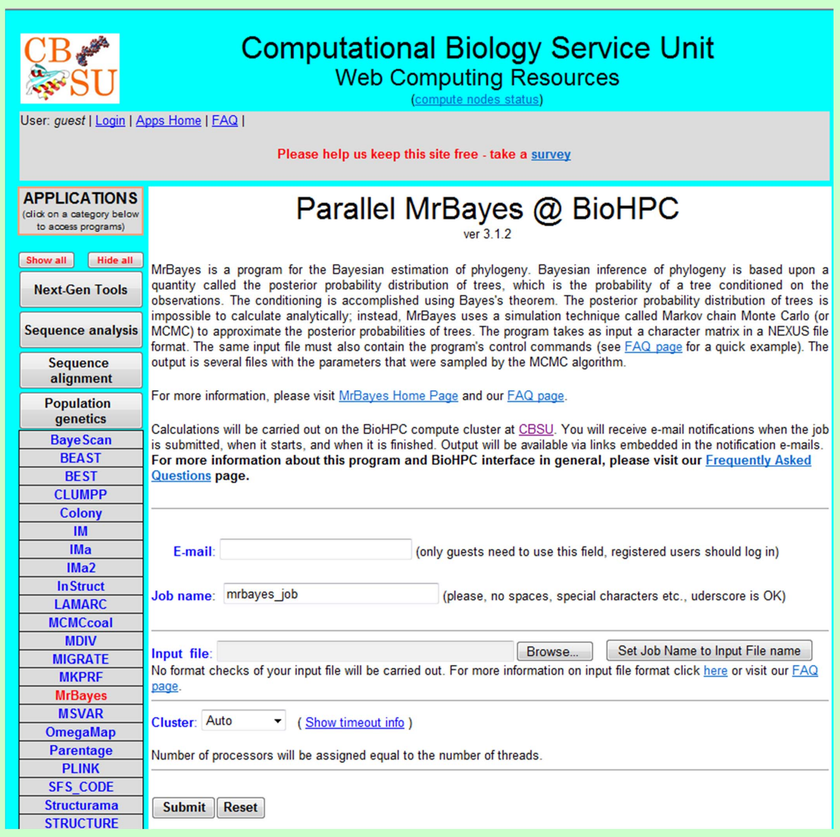
Next Generation Sequencing Module for Computational Biology Application Suite (http://BioHPC.org)

Robert Bukowski, Jaroslaw Pillardy, Mary Howard, Qi Sun

Computational Biology Service Unit, Cornell University, Ithaca, NY

INTRODUCTION: BioHPC

One of the challenges of **High Performance Computing (HPC)** is the user accessibility. At the **Cornell University Computational Biology Service Unit**, which is also a **Microsoft HPC institute**, we have developed a suite of computational biology applications for HPC (**BioHPC**) that allows researchers from biological laboratories to submit their jobs to parallel clusters and retrieve results through an easy-to-use web interface. Knowing the application, parameters and input is all that is required. Recently, a **web service** layer has been added which allows job control through other clients, such as MS Excel or perl scripts. Through web services, BioHPC is being integrated with the **Microsoft Biology Foundation** platform.



BioHPC: sample submission page

FEATURES OF BioHPC

- ❑ Users interact with their jobs and data primarily by a **web browser** (ASP.NET, Javascript) and **e-mail**. **Web service interface** also available for most applications.
- ❑ Automatic e-mail notifications sent upon changes in job status containing links for job control and results retrieval (by http or ftp).
- ❑ User-transparent **integration of distributed HPC cluster resources**
- ❑ Jobs and data files are **private** – they can be accessed only by a user who submitted a given job.
- ❑ Built-in **user and data management system** to configure user access to software and/or data.
- ❑ **Administrative interface** for easy management of jobs, clusters, applications, and data with automatic e-mail notification of possible problems.
- ❑ 52 applications covering various aspects of computational biology: **data mining/sequence, protein structure prediction and modeling, population genetics, phylogenetics, association analysis/statistic, MSR Biomedical applications, next generation sequencing data analysis.**
- ❑ The system is flexible and can be easily customized to include other software.
- ❑ Over 70,000 job submissions a year, many of them **parallel,** typically **several hours to several days long**
- ❑ Over 15,000 users from 83 countries (51% CPU time used from within USA).
- ❑ **TAIR**, the major database of the plant model organism Arabidopsis, and **SGN**, the international tomato genome database, are both using our system for data analysis.
- ❑ **BioHPC** source code is **freely available** from **BioHPC.net**. It can be installed locally with any Microsoft CCS or HPC 2008 cluster.

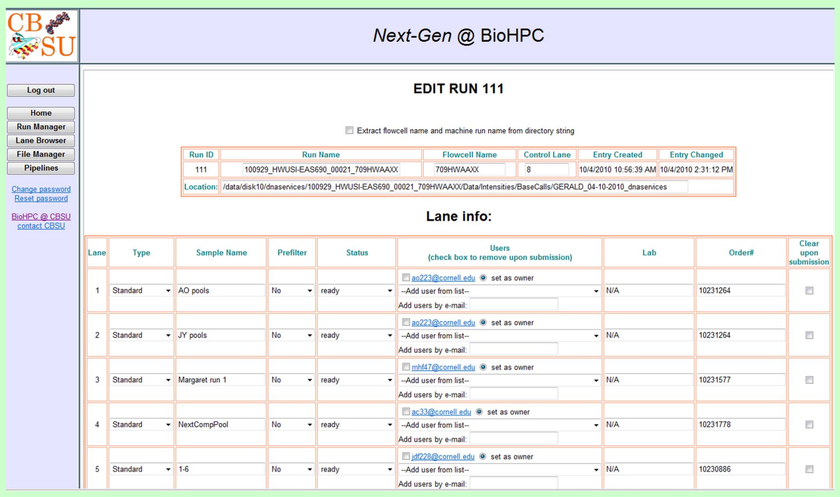
ARCHITECTURE

The system consists of a **web server** running the interface (ASP.NET C#), **Microsoft SQL server** (ADO.NET), **compute clusters** running Microsoft Windows or Linux, **ftp server** and **file server**. Two local Windows compute cluster schedulers are supported (CCS and HPC Server 2008), remote clusters can be used via **JSDL/HPC Basic Profile**. Linux clusters are accessed via ssh with the SGE scheduler supported. The BioHPC installation at CBSU is currently using 5 Microsoft Windows based local compute clusters totaling 976 cores and an experimental Linux cluster. The local nodes use Microsoft Server 2003 with CCS and Microsoft Server 2008 with HPC Server 2008. 80 CPU cores of the remote cluster Athena (located in Redmond, WA) are also available via JSDL, courtesy of Microsoft.

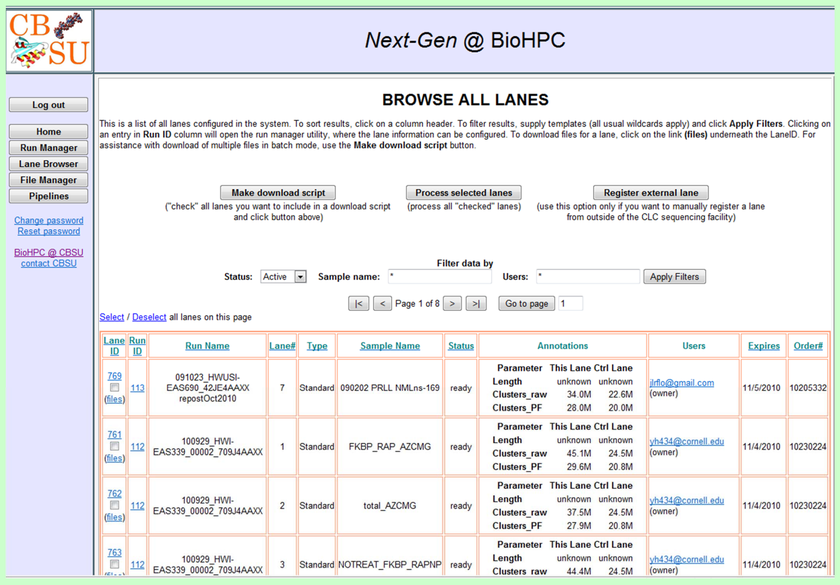
NEXT GENERATION SEQUENCING DATA ANALYSIS SUPPORT

We are currently implementing a new module of BioHPC designed to support analysis of **next generation sequencing** results. The module consists of a number of analysis applications (currently 7 and growing) plus several other components:

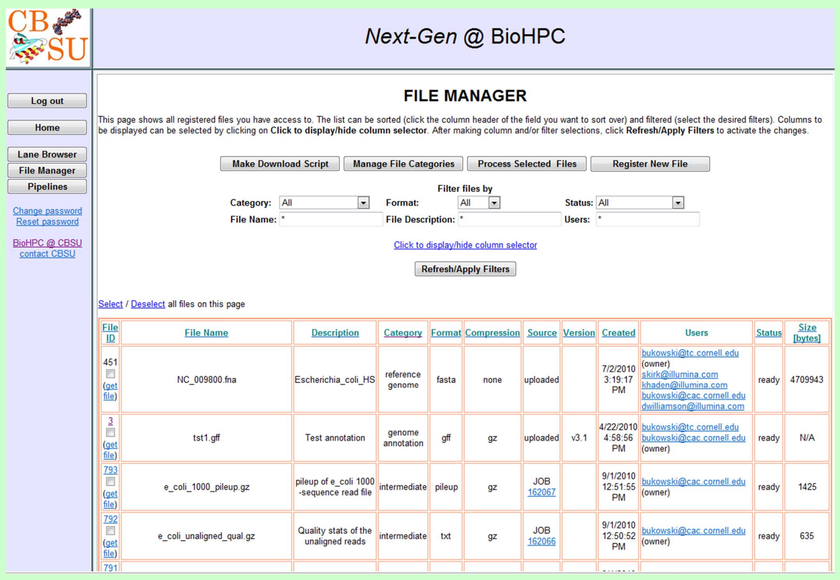
Run Manager: connects to the sequencing facility and automatically detects finished sequencing runs for which base calling has been completed. It then configures the run in BioHPC database and sends an invitation to the facility manager to approve the results for distribution to users. Once approved, the results (read files) are asynchronously transferred to BioHPC file server and catalogued there for further use. Once the transfer is complete, all users assigned to distributed lanes are automatically notified by an e-mail message containing download links. Run Manager is currently geared to handle mainly **Illumina** sequencing results, but extensions are possible.



Lane Browser: allows users to browse their sequencing read files (Illumina lanes) catalogued at BioHPC. The browser displays lane annotation information and allows the file owner to grant additional users access to a file. Read files obtained outside of the Cornell sequencing facility can also be uploaded and catalogued at BioHPC.



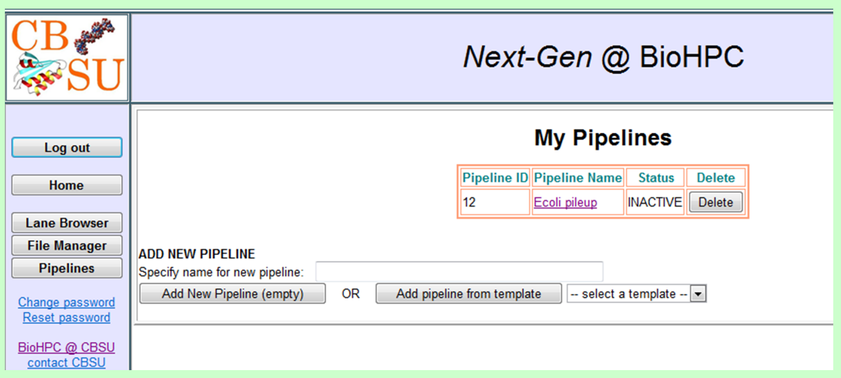
File Manager: allows users to upload and catalogue reference genome files, annotation files, and all other files needed in downstream data analysis. Intermediate files generated by BioHPC jobs can also be registered and re-used later on without the need for the user to download them to his local machine.



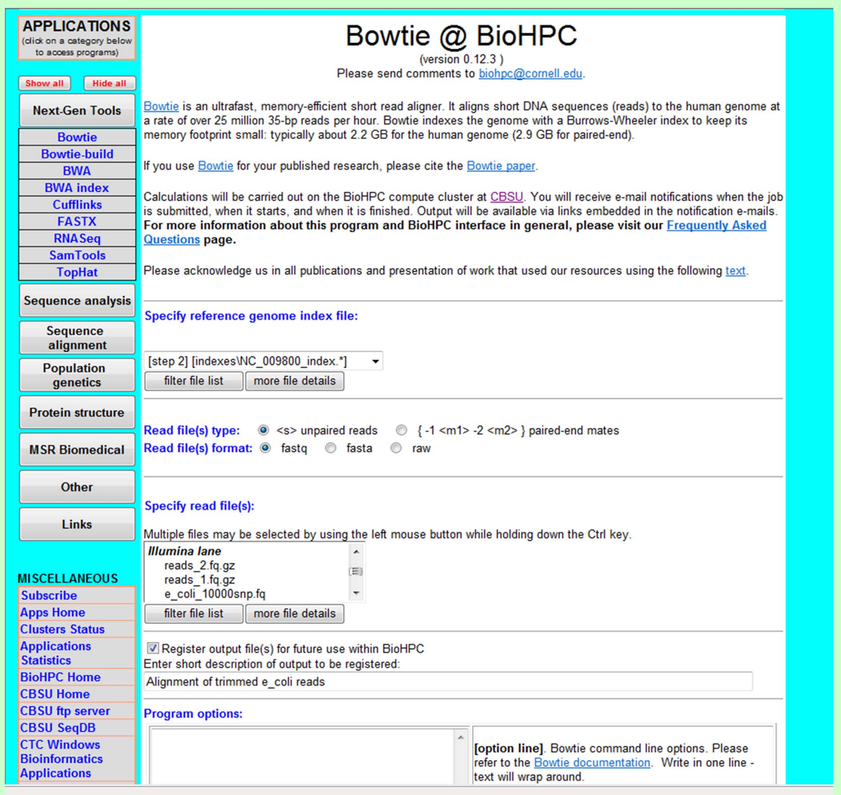
ABOUT CBSU

Computational Biology Service Unit (CBSU) of the Cornell University Life Sciences Core Laboratories Center was initiated by the Tri-Institutional collaboration among Cornell University, Weill Cornell Medical College, Rockefeller University, and Memorial Sloan-Kettering Cancer Center. In February 2006 CBSU became Microsoft HPC Institute charter member. CBSU is Cornell core facility for computational biology. BioHPC development is now partially funded by **Microsoft Research** and CBSU is a **Microsoft Biology Initiative** partner.

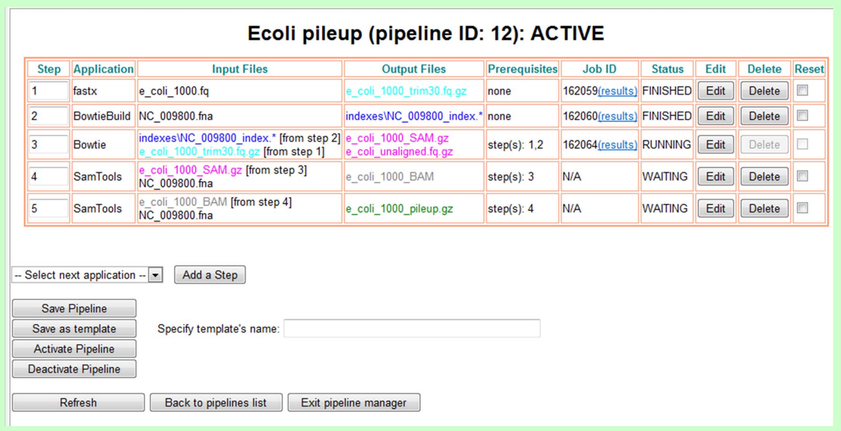
Pipeline Manager (under development): allows users to construct and run various analysis pipelines using sequencing reads, reference genomes, and other files stored at BioHPC as input.



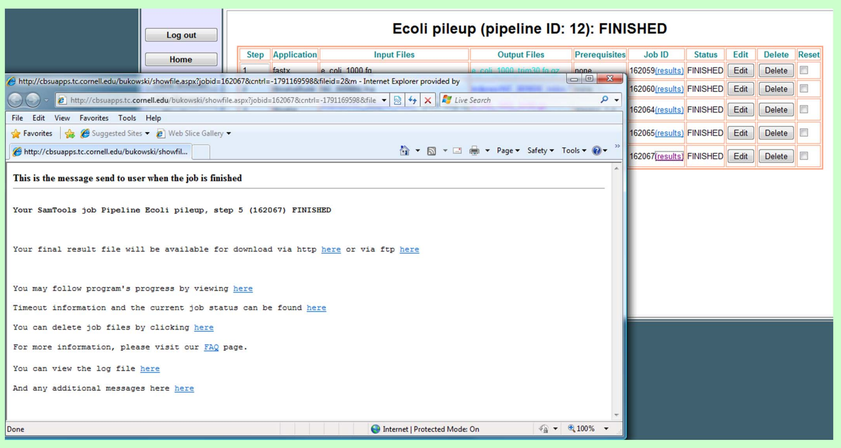
Steps of each pipeline are individually configurable using application submission pages. In these pages, input can be selected from among the files registered in File Manager as well as the ones anticipated from previous steps of the pipeline. This is how consecutive pipeline steps are connected.



Names of files involved in the pipeline are color-coded according to the step they come from. In the example below, BioHPC may decide to execute steps 1 and 2 simultaneously, since they have no prerequisites.



Pipeline steps are submitted and executed as "regular" BioHPC jobs, with job notification e-mails sent to the user. Additionally, the pipeline status can be seen instantly in the Pipeline Manager, as shown above. For every completed step, the output can be retrieved via links in notification e-mails, links in the Pipeline Manager table (as shown below), or via File Manager, where all step outputs are automatically registered.



The web service interface (planned) will allow pipelines to be controlled from any client application, such as the **MBF platform** or the **Illumina Genome Studio**, or **Trident** scientific workflow workbench.