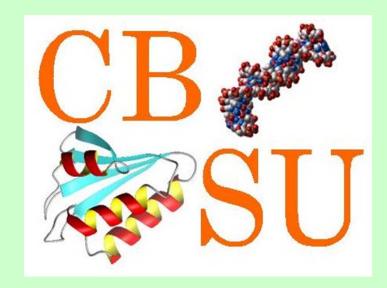


Cornell University Life Sciences **Core Laboratories Center**

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



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.

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

	Next-Gen @ BioHPC
Log out	My Pipelines
Home	Pipeline ID Pipeline Name Status Delete 12 Ecoli pileup INACTIVE Delete
Lane Browser File Manager Pipelines	ADD NEW PIPELINE Specify name for new pipeline:
Change password Reset password	Specify name for new pipeline: Add New Pipeline (empty) OR Add pipeline from template select a template
BioHPC @ CBSU contact CBSU	

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 easyto-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.

User: guest | Login | Apps Home | FAQ

Computational Biology Service Unit Web Computing Resources

Please help us keep this site free - take a survey



Log out Home

Run Manager

Lane Browser

File Manager Pipelines

Change password Reset password

CB Se SU

Log out

Lane Browse

File Manager Pipelines

Next-Gen @ BioHPC

APPLICATIONS (click on a category below	Parallel MrBayes @ BioHPC
to access programs)	ver 3.1.2
	Vol 5.1.2
Show all Hide all	MrBayes is a program for the Bayesian estimation of phylogeny. Bayesian inference of phylogeny is based upon a
Next-Gen Tools	quantity called the posterior probability distribution of trees, which is the probability of a tree conditioned on the observations. The conditioning is accomplished using Bayes's theorem. The posterior probability distribution of trees is
Sequence analysis	impossible to calculate analytically; instead, MrBayes uses a simulation technique called Markov chain Monte Carlo (or MCMC) to approximate the posterior probabilities of trees. The program takes as input a character matrix in a NEXUS file format. The same input file must also contain the program's control commands (see <u>FAQ page</u> for a quick example). The
Sequence	output is several files with the parameters that were sampled by the MCMC algorithm.
alignment	For more information, places visit McRayes, Harry Days and ave EAO appro-
Population	For more information, please visit <u>MrBayes Home Page</u> and our <u>FAQ page</u> .
genetics	Calculations will be carried out on the BioHPC compute cluster at <u>CBSU</u> . You will receive e-mail notifications when the job
BayeScan	is submitted, when it starts, and when it is finished. Output will be available via links embedded in the notification e-mails.
BEAST	For more information about this program and BioHPC interface in general, please visit our Frequently Asked
BEST	Questions page.
CLUMPP	
Colony	
IM	
IMa	E-mail: (only guests need to use this field, registered users should log in)
IMa2	
InStruct	Job name: mrbayes_job (please, no spaces, special characters etc., uderscore is OK)
LAMARC	
MCMCcoal	
MDIV MIGRATE	Input_file: Browse Set Job Name to Input File name
MIGRATE	No format checks of your input file will be carried out. For more information on input file format click here or visit our FAQ
MrBayes	page.
MSVAR	
OmegaMap	Cluster: Auto (Show timeout info)
Parentage	Number of processory will be precised a surplus the number of these de-
PLINK	Number of processors will be assigned equal to the number of threads.
SFS CODE	
Structurama	Submit Reset
STRUCTURE	

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).

EDIT RUN 111

Extract flowcell name and machine run name from directory string

Run ID	Run Name	Flowcell Name	Control Lane	Entry Created	Entry Changed					
111	100929_HWUSI-EAS690_00021_709HWAAXX	709HWAAXX	8	10/4/2010 10:56:39 AM	10/4/2010 2:31:12 PM					
Location:	/data/disk10/dnaservices/100929_HWUSI-EAS690_00021_709HWAAXX/Data/Intensities/BaseCalls/GERALD_04-10-2010_dnaservices									

BioHPC @ CBSU contact CBSU		Lane info:														
	Lane	Туре	Sample Name	Prefilter	Status	Users (check box to remove upon submission)	Lab	Order#	Clea upor submiss							
	1	Standard	✓ AO pools	No 🗸	ready	 ✓ ao223@cornell.edu set as owner -Add user from list- Add users by e-mail: 	• N/A	10231264								
	2	Standard	▼ JY pools	No 🗸	ready	 ao223@cornell.edu set as owner Add user from list Add users by e-mail: 	▼ N/A	10231264								
	3	Standard	✓ Margaret run 1	No 🗸	ready	 ✓ mhf47@cornell.edu set as owner -Add user from list Add users by e-mail: 	• N/A	10231577								
	4	Standard	- NextCompPool	No 🗸	ready	 ✓ ac33@cornell.edu ● set as owner ✓Add user from list Add users by e-mail: 	• N/A	10231778								
	5	Standard	▼ 1-6	No 🗸	ready	 Add user from list- Add users by e-mail: 	▼ N/A	10230886								

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.

CB SSU	Next-Gen @ BioHPC											
Log out	BROWSE ALL LANES											
Log out												
Home	his is a list of all lanes configured in the system. To sort results, click on a column header. To filter results, supply templates (all usual wildcards apply) and click Apply Filters. Clicking on n entry in Run ID column will open the run manager utility, where the lane information can be configured. To download files for a lane, click on the link (files) underneath the LaneID. For											
Run Manager	assistance with download of multiple files in batch mode, use the Make download script button.											
Lane Browser												
File Manager												
Pipelines	Make download script Process selected lanes Register external lane											
Change password Reset password	("check" all lanes you want to include in a download script (process all "checked" lanes) (use this option only if you want to manually register a lane and click button above) from outside of the CLC sequencing facility)	("check" all lanes you want to include in a download script (process all "checked" lanes) (use this option only if you want to manually register a lane										
BioHPC @ CBSU contact CBSU	Filter data by Status: Active Sample name: * Users: * Apply Filters											
	Select / Deselect all lanes on this page I I Go to page I Go to page											
	Lane Run Run Name Lane# Type Sample Name Status Annotations Users Expires Order	#										

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.

(click on a category below to access programs)	Bowtie @ BioHPC									
Show all Hide all	(version 0.12.3) Please send comments to <u>biohpc@cornell.edu</u> .									
Next-Gen Tools Bowtie	Bowtie is an ultrafast, memory-efficient short read aligner. It aligns short DNA sequences (reads) to the human genome at a rate of over 25 million 35-bp reads per hour. Bowtie indexes the genome with a Burrows-Wheeler index to keep its memory footprint small: typically about 2.2 GB for the human genome (2.9 GB for paired-end).									
Bowtie-build BWA	If you use <u>Bowtie</u> for your published research, please cite the <u>Bowtie paper</u> .									
BWA index Cufflinks FASTX RNASeq	Calculations will be carried out on the BioHPC compute cluster at <u>CBSU</u> . You will receive e-mail notifications when the job is submitted, when it starts, and when it is finished. Output will be available via links embedded in the notification e-mails. For more information about this program and BioHPC interface in general, please visit our <u>Frequently Asked</u> <u>Questions</u> page.									
SamTools TopHat	Please acknowledge us in all publications and presentation of work that used our resources using the following text.									
Sequence analysis	Specify reference genome index file:									
Sequence alignment										
Population genetics	step 2] [indexes\NC_009800_index.*] filter file list more file details									
Protein structure	tead file(s) type: <s> unpaired reads</s> <m1> -2 <m2> } paired-end mates</m2></m1> 									
MSR Biomedical	Read file(s) format: fastq fasta raw 									
Other	Specify read file(s):									
Links	Multiple files may be selected by using the left mouse button while holding down the Ctrl key.									
MISCELLANEOUS Subscribe	Illumina lane Image: Constraint of the second sec									
Apps Home Clusters Status	filter file list more file details									
Applications Statistics	Register output file(s) for future use within BioHPC Enter short description of output to be registered:									
BioHPC Home CBSU Home	Alignment of trimmed e_coli reads									
CBSU ftp server CBSU SeqDB	Program options:									
CDSU SeqUE CTC Windows Bioinformatics Applications	[option line]. Bowtie command line options. Please refer to the <u>Bowtie documentation</u> . Write in one line - text will wrap around.									

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.

769 (files) <u>113</u>	091023_HWUSI- EAS690_42JE4AAXX repostOct2010	7	Standard	090202 PRLL NMLns-169	ready	Parameter This Lane Ctrl Lane Length unknown unknown Clusters_raw 34.0M 22.6M Clusters_PF 28.0M 20.0M	<u>ilrflo@gmail.com</u> (owner)	11/5/2010	1020533
761 (files)	100929_HWI- EAS339_00002_709J4AAXX	1	Standard	FKBP_RAP_AZCMG	ready	Parameter This Lane Ctrl Lane Length unknown unknown Clusters_raw 45.1M 24.5M Clusters_PF 29.6M 20.8M	<u>yh434@cornell.edu</u> (owner)	11/4/2010	1023022
762 (files)	100929_HWI- EAS339_00002_709J4AAXX	2	Standard	total_AZCMG	ready	Parameter This Lane Ctrl Lane Length unknown unknown Clusters_raw 37.5M 24.5M Clusters_PF 27.9M 20.8M	<u>yh434@cornell.edu</u> (owner)	11/4/2010	1023022
763	100929_HWI- EAS339_00002_709J4AAXX	3	Standard	NOTREAT_FKBP_RAPNP	ready	Parameter This Lane Ctrl Lane Length unknown unknown Clusters raw 44.4M 24.5M	<u>yh434@cornell.edu</u> (owner)	11/4/2010	1023022

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.

Ecoli pileup (pipeline ID: 12): ACTIVE

Step	Application	Input Files	Output Files	Prerequisites	Job ID	Status	Edit	Delete	Reset
1	fastx	e_coli_1000.fq	e_coli_1000_trim30.fq.gz	none	162059 <u>(results)</u>	FINISHED	Edit	Delete	
2	BowtieBuild	NC_009800.fna	indexes\NC_009800_index.*	none	162060 <u>(results)</u>	FINISHED	Edit	Delete	
3	Bowtie	indexes\NC_009800_index.* [from step 2] e_coli_1000_trim30.fq.gz [from step 1]	e_coli_1000_SAM.gz e_coli_unaligned.fq.gz	step(s): 1,2	162064 <u>(results)</u>	RUNNING	Edit	Delete	
4	SamTools	e_coli_1000_SAM.gz [from step 3] NC_009800.fna	e_coli_1000_BAM	step(s): 3	N/A	WAITING	Edit	Delete	
5	SamTools	e_coli_1000_BAM [from step 4] NC_009800.fna	e_coli_1000_pileup.gz	step(s): 4	N/A	WAITING	Edit	Delete	

Select next application	Add a Step		
Save Pipeline			
Save as template	Specify template's name:		
Activate Pipeline			
Deactivate Pipeline			
Refresh	Back to pipelines list	Exit pipeline manager	

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 emails, links in the Pipeline Manager table (as shown below), or via File Manager, where all step outputs are automatically registered.

	Log out				Ecoli pileu	ıp (pipeline ID: 12): FINISHED						
	Home		Step	Application	Input Files	Output Files	Prerequisites	Job ID	Status	Edit	Delete	Reset
					e coli 1000 fa	e coli 1000 trim30 fa az	none	162059 <u>(results)</u>	FINISHED	Edit	Delete	
http://cbsuapps.tc.cornell.edu/bukowski/showfile.aspx?jobid=162067&cntrl=-1791169598&fileid=2&m - Internet Explorer provided by							162060 <u>(results)</u>	FINISHED	Edit	Delete		
🖉 🗸 http://cbsuapps.tc.cornell.edu/bukowski/showfile.aspx?jobid=162067&cntrl=-1791169598&file 🔻 🗟 🐓 🗙 🧗 Live Search 👂 🔹								162064 <u>(results)</u>	FINISHED	Edit	Delete	

pelines to

the **MBF**

Trident

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 reused later on without the need for the user to download them to his local machine.

FILE MANAGER

Next-Gen @ BioHPC

page shows all registered files you have access to. The list can be sorted (click the column header of the field you want to sort over) and filtered (select the desired filters). Columns Home by clicking on Click to display/hide column selector. After making column and/or filter selections, click Refresh/Apply Filters to activate the changes





	<u>Selec</u>	t / <u>Deselect</u> all files on this page										
	File ID	File Name	Description	Category	<u>Format</u>	Compression	Source	<u>Version</u>	Created	Users	<u>Status</u>	<u>Size</u> [bytes]
	451 (<u>get</u> <u>file</u>)	NC_009800.fna	Escherichia_coli_HS	reference genome	fasta	none	uploaded		7/2/2010 3:19:17 PM	bukowski@tc.cornell.edu (owner) skirk@illumina.com khaden@illumina.com bukowski@cac.cornell.edu dwilliamson@illumina.com	ready	4709943
	3 (get file)	tst1.gff	Test annotation	genome annotation	gff	gz	uploaded		4/22/2010 4:58:56 PM	bukowski@tc.cornell.edu bukowski@cac.cornell.edu (owner)	ready	N/A
	793 (get file)	e_coli_1000_pileup.gz	pileup of e_coli 1000 -sequence read file	intermediate	pileup	gz	JOB <u>162067</u>		9/1/2010 12:51:55 PM	<u>bukowski@cac.cornell.edu</u> (owner)	ready	1425
	792								0.11.000.10			

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 Cor used via **JSDL/HPC Basic Profile**. Linux clusters are accessed was via ssh with the SGE scheduler supported. The BioHPC Cor installation at CBSU is currently using 5 Microsoft Windows Uni based local compute clusters totaling 976 cores and an Feb experimental Linux cluster. The local nodes use Microsoft me Server 2003 with CCS and Microsoft Server 2008 with HPC bio Server 2008. 80 CPU cores of the remote cluster Athena Mic (located in Redmond, WA) are also available via JSDL, courtesy Init of Microsoft.

Image: Part of the state of the unaligned_qual.gz Quality stats of the unaligned reads intermediate txt gz JOB 162066 9/1/2010 12:50:52 PM bukowski@cac.cornell.edu (owner) ready 635 791 File File<	Inttp://cbsuapps.tc.cornell.edu/bukowski/showfile.aspx?jobid=162067&ccntrl=-1791169598&cfile File Edit View Favorites Tools Help File Edit View Favorites Tools Help Image: Application of the point of	Edit Delete
ABOUT CBSU Computational Biology Service Unit (CBSU) of the	Your SamTools job Pipeline Ecoli pileup, step 5 (162067) FINISHED Your final result file will be available for download via http <u>here</u> or via ftp <u>here</u>	
nell University Life Sciences Core Laboratories Center initiated by the Tri-Institutional collaboration among	You may follow program's progress by viewing <u>here</u> Timeout information and the current job status can be found <u>here</u> You can delete job files by clicking <u>here</u>	
rnell University, Weill Cornell Medical College, Rockefeller versity, and Memorial Sloan-Kettering Cancer Center. In	For more information, please visit our FAQ page. You can view the log file <u>here</u> And any additional messages here <u>here</u>	
oruary 2006 CBSU became Microsoft HPC Institute charter mber. CBSU is Cornell core facility for computational	Done 🕒 Internet Protected Mode: On 🖓 🕶 🔍 100% 🕶	
ogy. BioHPC development is now partially funded by prosoft Research and CBSU is a Microsoft Biology iative partner.	The web service interface (planned) will allow pipeli be controlled from any client application, such as the platform or the Illumina Genome Studio, or T scientific workflow workbench.	e ME