Computational Biology Applications Suite for High Performance Computing (BioHPC)

Jaroslaw Pillardy

CBSU,
Life Sciences Core Laboratories Center
Cornell University

Microsoft External Research Symposium

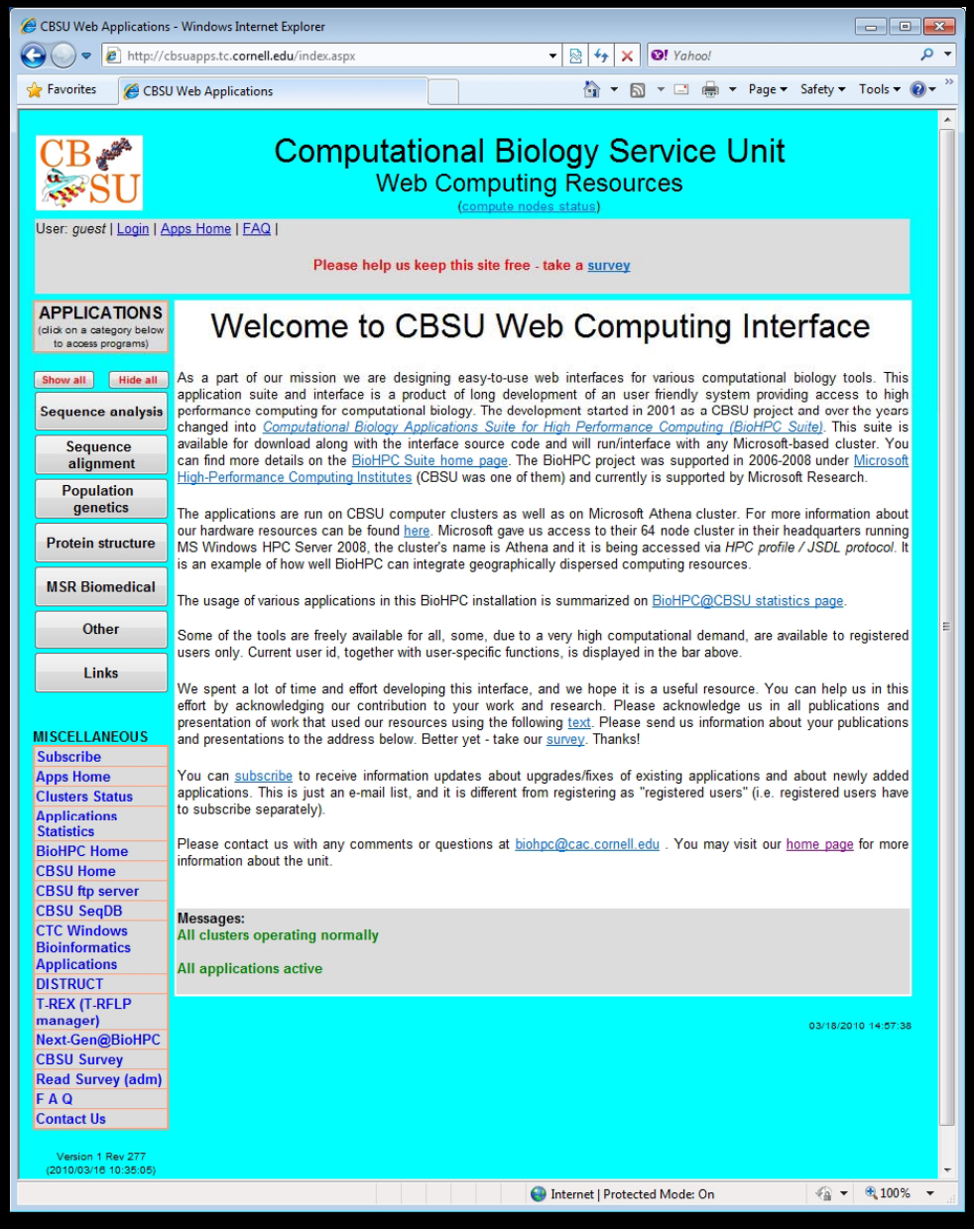
April 6-7, 2010

Redmond, Washington

Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)



- User accessibility is major problem in HPC and bio-computing
- Using a parallel cluster or remote resources required knowledge of the operating system, queuing system and parallel programming
- BioHPC suite provides easy access to standardized applications on Windows platform
- BioHPC suite provides easy way manage and integrate distributed computational resources
- Written in C# / ASP.NET

Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)

CB SU

What it does:

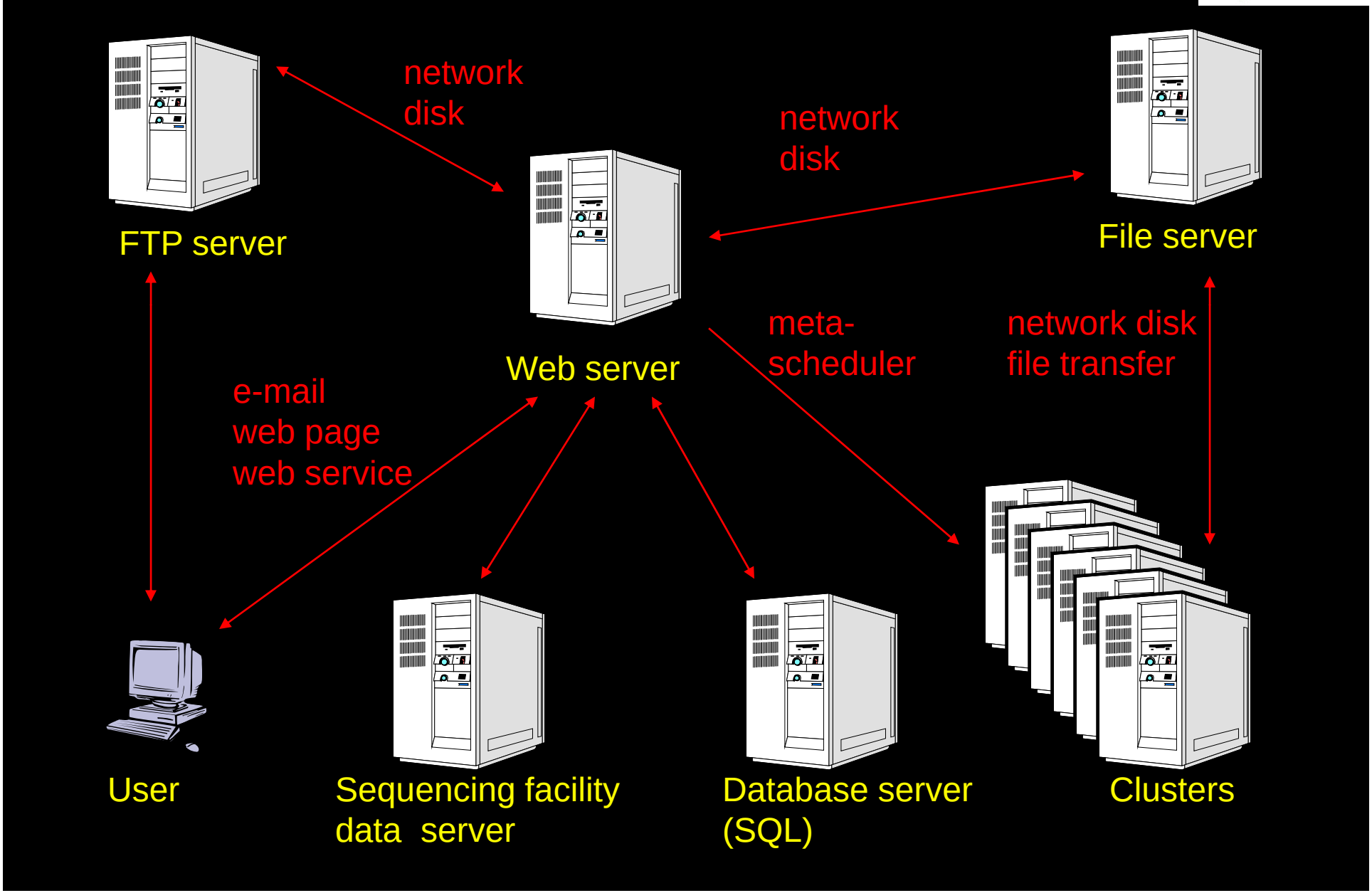
- Web-based, point-and-click access to a variety of bioinformatics applications with the underlying structure of computational platform transparent to the user
- Enhancement of standard applications through parallelization, transparent to the user
- Integration and simplified access to geographically dispersed hardware resources
- Web-based administration of users, jobs, applications, and clusters within the suite
- Standardized access to and maintenance of bioinformatics databases
- Next generation data management and distribution – from sequencing facility to users.
- Next generation sequencing pipelines and applications – with internal data management or optional external data upload

Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)

CB SU



Computational Biology Applications Suite for High Performance Computing (BioHPC)

ARCHITECHTURE

- Web server running the interface (ASP.NET C#)
- Microsoft SQL server (ADO.NET)
- Compute clusters running Microsoft Windows
- Ftp server / file server
- Two local compute cluster schedulers are supported (CCS and HPC Server 2008)
- Remote clusters can be used via JSDL/HPC Profile

Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)

37 applications available

- Data mining / sequence analysis (BLAST, HMMER, InterProScan, GIMSAN, SLIM)
- Protein structure prediction and modeling (LOOPP, Modeller)
- Population genetics (BEAST, BEST, Clumpp, IM, IMa, IMa2, InStruct, LAMARC, MDIV, Migrate, MKPRF, MSVAR, OmegaMap, Parentage, SFS_CODE, Structurama, Structure, TESS)
- Phylogenetics (MrBayes, ClustalW, Stretcher, T-COFFEE)
- Association analysis / statistics (PLINK, R)
- MSR Biomedical (CreateEpitome, Epipred, FalseDiscoveryRate, HlaAssignment, HlaCompletion, PhyloD)

The system is flexible and can be easily customized to include other software. The interface to each application is standardized, users can choose the cluster, number of nodes or allow the interface to determine it based on the best load balance and node availability

Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)

CB SU

The most popular applications

Job submission from 6/13/2003 to 3/18/2010

LOOPP	20,385	protein structure prediction
MDIV	20,965	population genetics
P-BLAST	4,504	sequence analysis / data mining
MrBayes	18,799	population genetics
IM/IMa/IMa2	22,567	population genetics
STRUCTURE	17,968	population genetics

All applications **140,141** (average **20,090** per year, **49,738** last year)

LOOPP	parallel, uses 5-20 cores for 3-10 hours
MDIV	serial, uses 1 core from few hours to two weeks (average: 2-5 days)
P-BLAST	parallel, restricted resource, uses 10 – 100 cores for a few days to a week (average: few days)
MrBayes	parallel, uses 8-20 cores for a few hours to two weeks (average: a week)

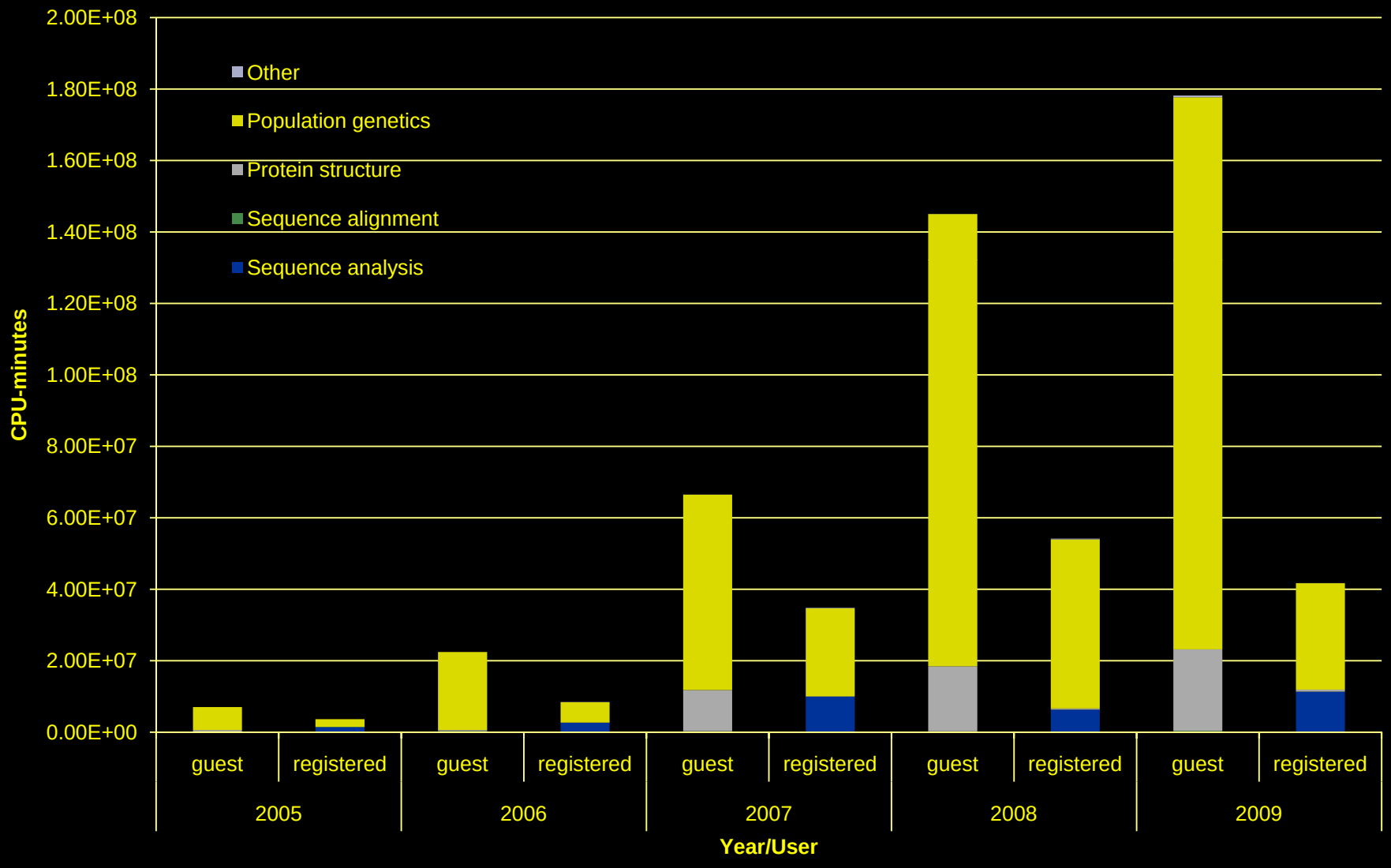
Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)

CB SU

Dynamics of BioHPC Utilization



Computational Biology Applications Suite for High Performance Computing (BioHPC)

The jobs were submitted by 11,471 unique users from 83 countries

The majority (57% by CPU time used) coming from the USA

52% of the USA utilized CPU time coming from New York

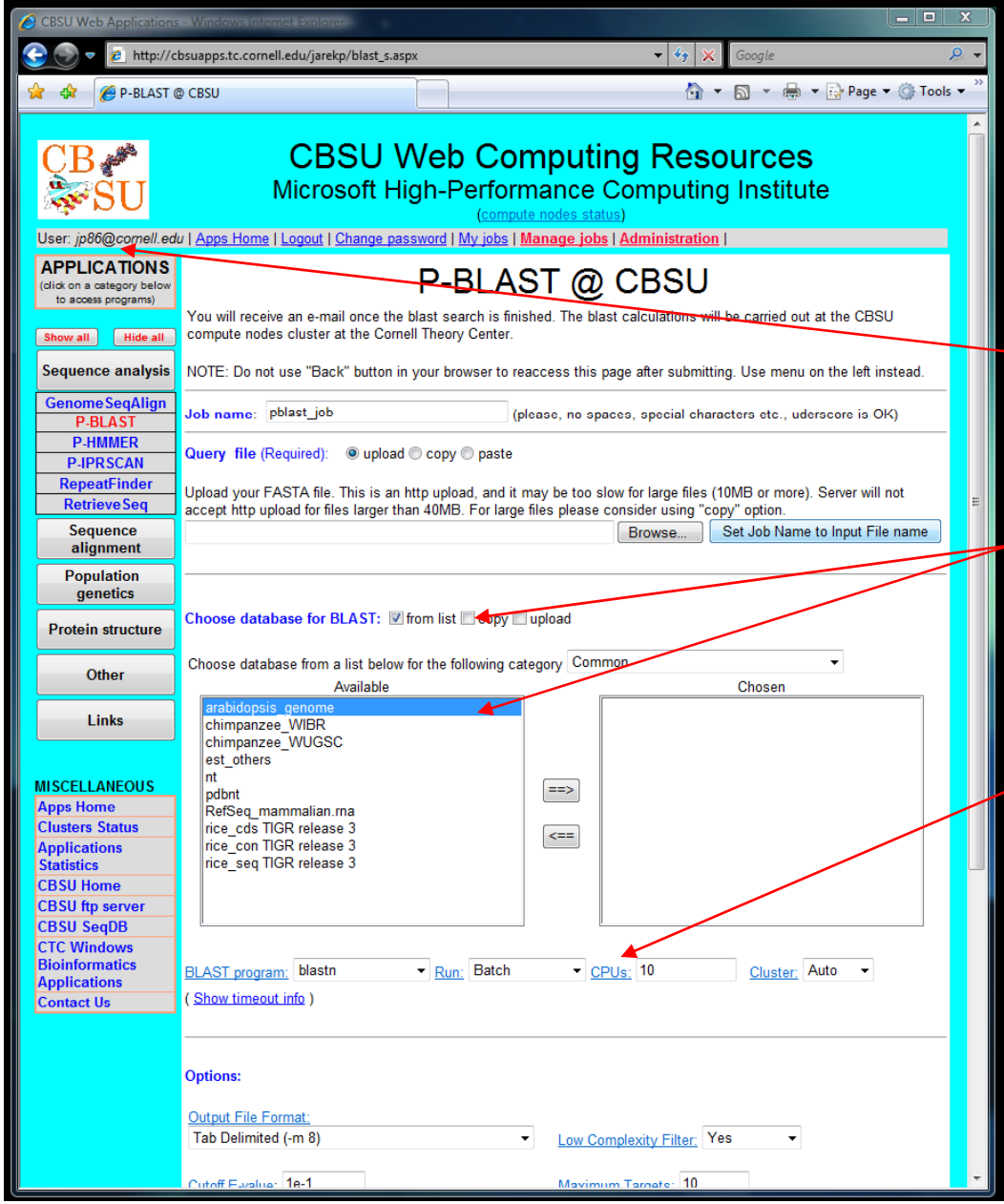
Among them there are

257 unique Cornell users,
2,580 users from *.edu* domains
426 unique *.edu* institutions
4,813 users from *.com* domains
4,191 users with Yahoo, Gmail and Hotmail e-mail addresses

Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)

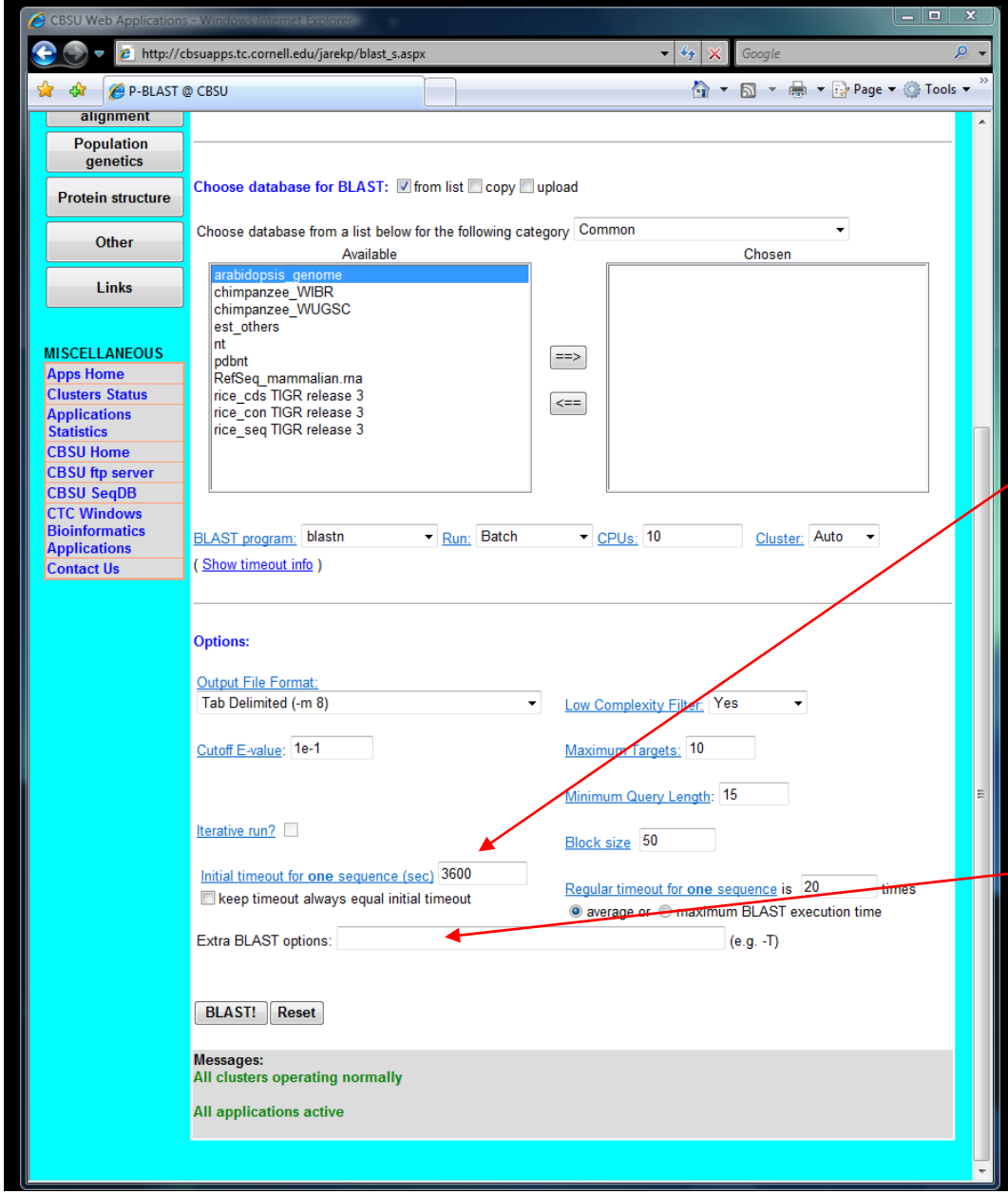


- Each application interface is standardized as much as practical
- Some applications can be used only by registered users
- Users can upload their data files via http, place them on our ftp server, or use their local network drive
- In addition to application-specific options, users can choose number of nodes, scheduler, and cluster

Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)

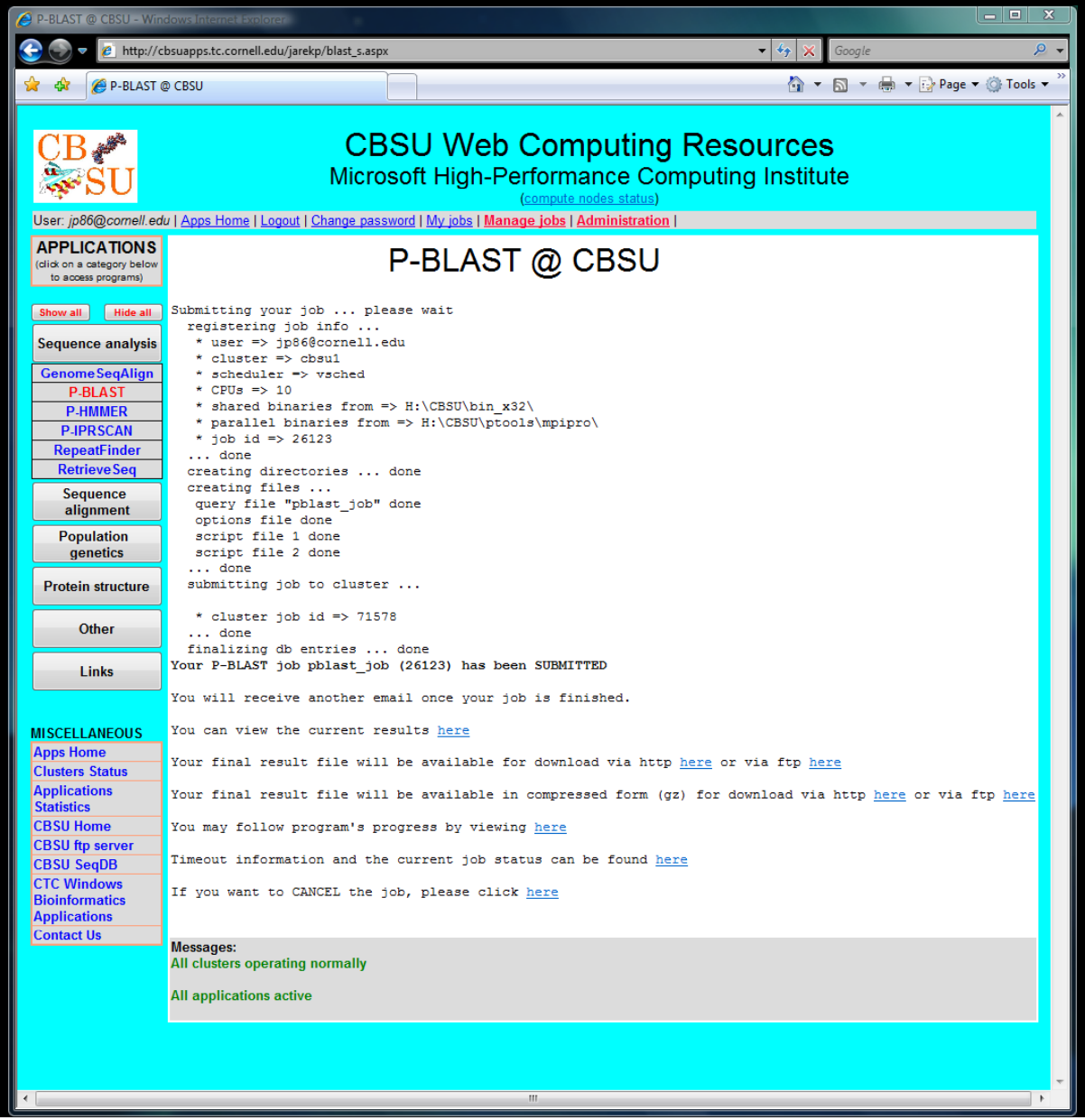


- For trivially parallelized programs extensive control over task performance is provided, preventing waste of computational resources in a case of errors in input
- Some application-specific options are available on the interface, some rarely used ones can be entered as string

Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)



A job is now submitted

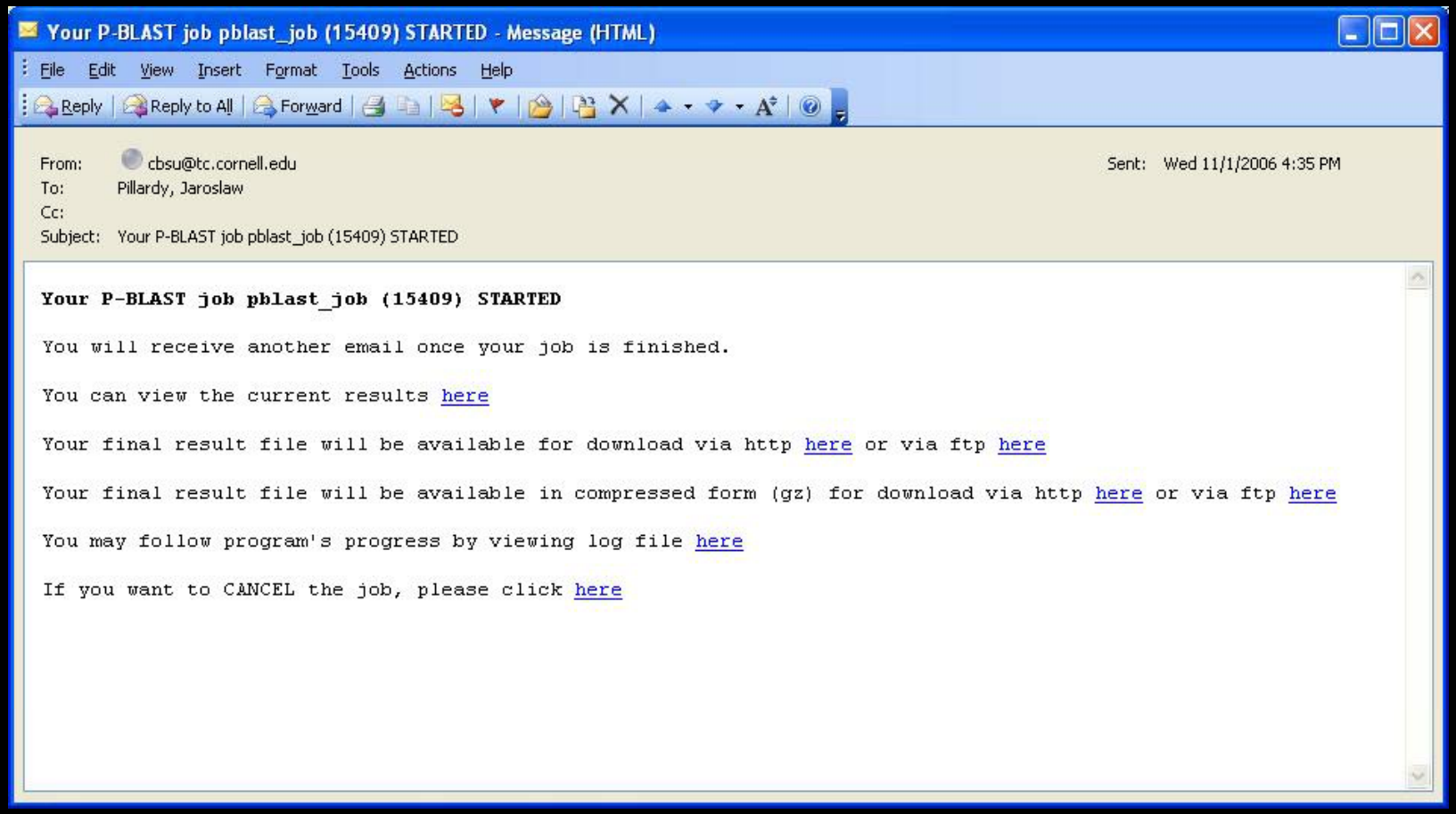
An email has been dispatched with links to output files and job control functions. These links are also available on this page along with submit log.

Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)

CB SU

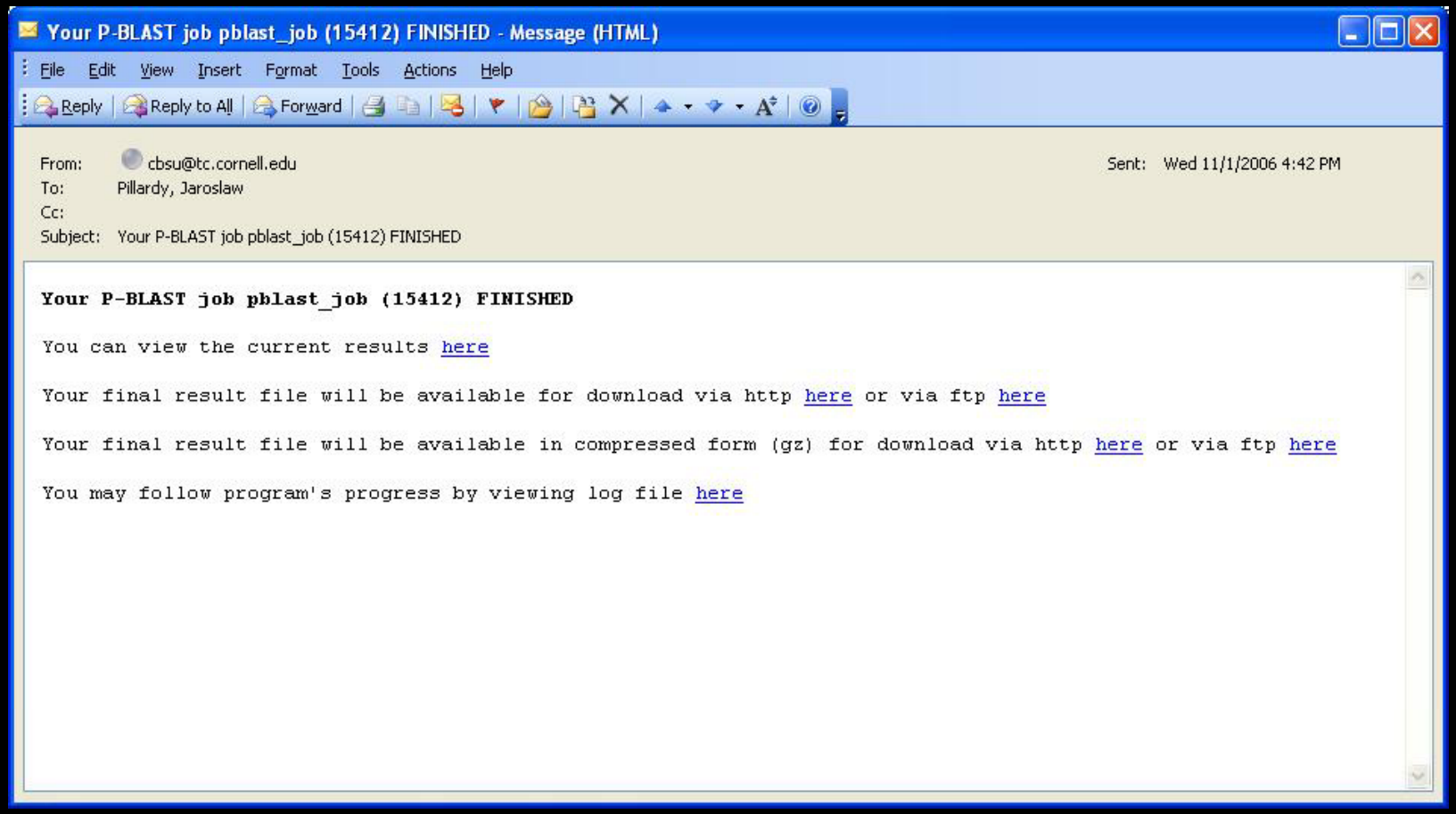


Jobs interact with user via e-mails. Links in the e-mail allow for viewing current results, computations progress (log) as well as cancelling the job if necessary.

Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)



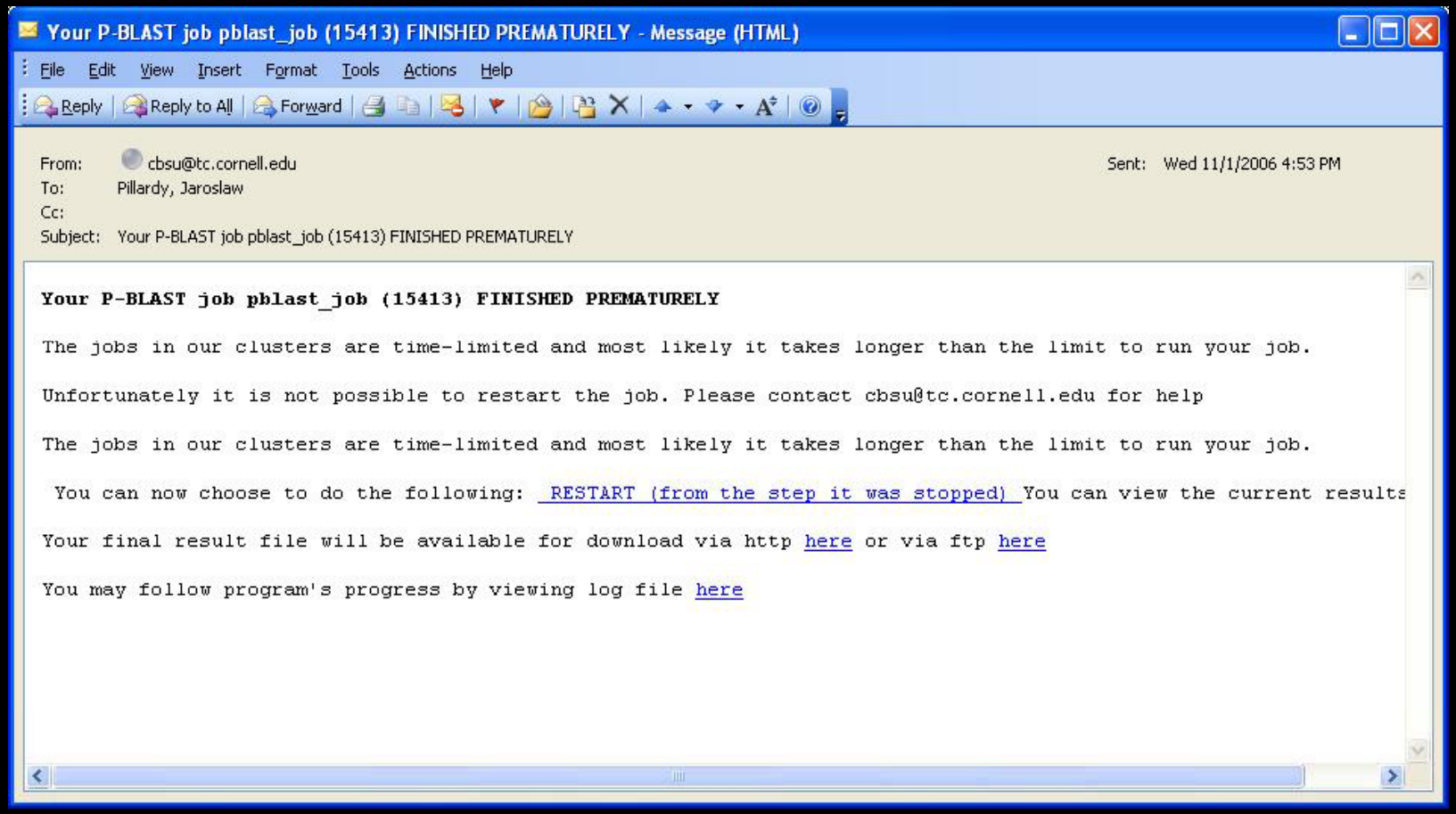
When job finishes, another e-mail is sent.

Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)

CB SU



Your P-BLAST job pblast_job (15413) FINISHED PREMATURELY - Message (HTML)

File Edit View Insert Format Tools Actions Help

Reply | Reply to All | Forward

From: cbsu@tc.cornell.edu
Sent: Wed 11/1/2006 4:53 PM
To: Pillardy, Jaroslaw
Cc:
Subject: Your P-BLAST job pblast_job (15413) FINISHED PREMATURELY

Your P-BLAST job pblast_job (15413) FINISHED PREMATURELY

The jobs in our clusters are time-limited and most likely it takes longer than the limit to run your job.

Unfortunately it is not possible to restart the job. Please contact cbsu@tc.cornell.edu for help

The jobs in our clusters are time-limited and most likely it takes longer than the limit to run your job.

You can now choose to do the following: RESTART (from the step it was stopped) You can view the current results

Your final result file will be available for download via http here or via ftp here

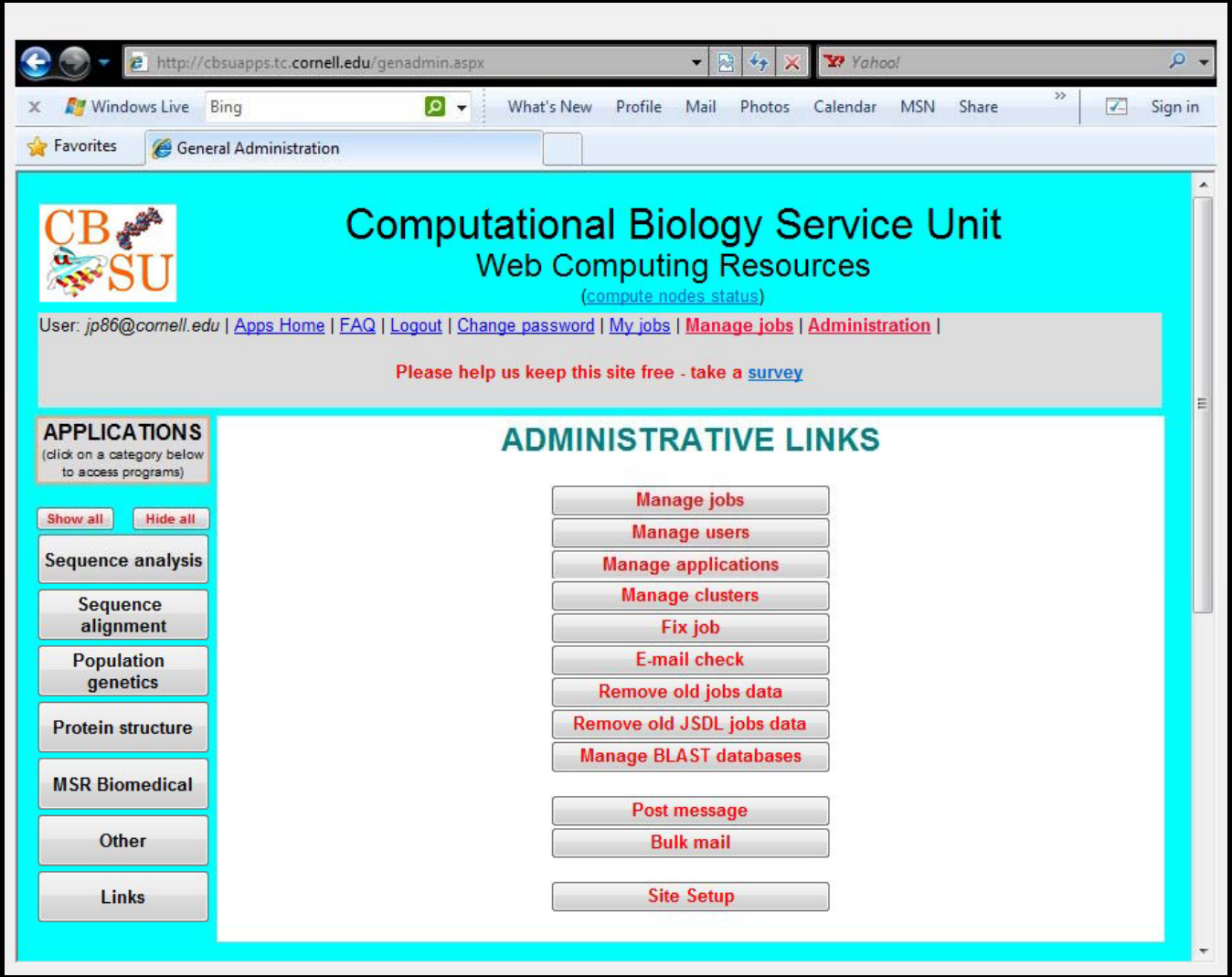
You may follow program's progress by viewing log file here

Sometimes a job finishes prematurely. Usually it happens for a very long jobs run on relatively small number of nodes. Many applications can be restarted and continued from the stopping point via a link.

Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)



Links to administration pages.

Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)

Job Monitoring - Windows Internet Explorer

http://cbsuapps.tc.cornell.edu/checkinfo.aspx

Computational Biology Service Unit
Web Computing Resources
(compute nodes status)

User: *jp86@cornell.edu* | Apps Home | FAQ | Logout | Change password | My jobs | Manage jobs | Administration |

Please help us keep this site free - take a survey

APPLICATIONS
(click on a category below to access programs)

Show all | Hide all

- Sequence analysis
- Sequence alignment
- Population genetics
- Protein structure
- MSR Biomedical
- Other
- Links

MISCELLANEOUS

- Subscribe
- Apps Home
- Clusters Status
- Applications Statistics
- BioHPC Home
- CBSU Home
- CBSU ftp server
- CBSU SeqDB
- CTC Windows Bioinformatics Applications
- DISTRUCT
- T-REX (T-RFLP manager)
- CBSU Survey
- Read Survey (adm)
- F A Q
- Contact Us

Job admin

cache last updated between 10/28/2009 4:01:00 PM and 10/28/2009 4:14:40 PM cache hits 99.3% over last 1000 calls

On this page: total 198 jobs, 9 queued, 133 running, 41 finished, 9 failed, 6 other

Display | Clear Cache & Display | Display & update | up to 600 jobs (300 per page). Last refreshed: 10/28/2009 4:19:58 PM | Autorefresh in 14min | Unlock Admin

S	No	ID	Appname	Submitted	User	Cluster	Nodes/ (CPU)	Cluster jobid	Timeout	Started	Ending	Job name	Status	Active?	Files	Action
		>0 <999999	*	after 12/31/2004 before 12/31/2010	*	*	>0	*	>0	after 12/31/2004 before 12/31/2010	after 12/31/2004 before 12/31/2010	*	ANY	Yes		
Page:1																
	1	102826	MrBayes	10/28/2009 4:09:57 PM	dtanner@biology.usu.edu	Athena (Auto)	(4)	16088	10080	10/28/2009 4:09:57 PM	10/28/2009 4:16:52 PM	odonmeli_its1	FINISHED CORRECTLY	Yes	input: show|http|ftp output: show|http|ftp log file: show DIR	ARCHIVE DEL RESTART NOTIFY:S 0 1 E P STAT: SC FL FU PM
	2	102825	MrBayes	10/28/2009 4:00:59 PM	chris.yesson@ioz.ac.uk	biosim (Auto)	(4)	24920	10080	10/28/2009 4:02:10 PM	11/4/2009 4:02:10 PM	nemertesiaalignmentmega_trim_mb_constrained	RUNNING	Yes	input: show|http|ftp output: show|http|ftp log file: show DIR	STOP RESTART NOTIFY:S 0 1 E P STAT: SC FL FU PM
	3	102824	InStruct	10/28/2009 3:48:47 PM	alvarosolergarzon@gmail.com	Athena (Auto)	(1)	16087	7200	10/28/2009 3:51:50 PM	11/2/2009 3:51:50 PM	Silvestres_MATRIX_INSTRUCT.txt	RUNNING	Yes	input: show|http|ftp output: show|http|ftp log file: show DIR	STOP RESTART NOTIFY:S 0 1 E P STAT: SC FL FU PM
	4	102823	BEAST	10/28/2009 2:57:02 PM	chuav2@unlv.nevada.edu	cbsum2 (cbsum2)	(2)	4016	7200			Picoides_Run_4	QUEUED	Yes	input: show|http|ftp output: show|http|ftp log file: show DIR	STOP RESTART NOTIFY:S 0 1 E P STAT: SC FL FU PM
	5	102822	P-CLUSTALW	10/28/2009 2:49:51 PM	bukowski@cac.cornell.edu	cbsum2k8 (cbsum2k8)	(1)	1623	4320	10/28/2009 2:49:56 PM	10/28/2009 2:50:04 PM	clustalw_job	FINISHED CORRECTLY	Yes	input: show|http|ftp output: show|http|ftp log file: show DIR	ARCHIVE DEL RESTART NOTIFY:S 0 1 E P STAT: SC FL FU PM
	6	102821	MrBayes	10/28/2009 2:44:38 PM	wksavage@ucdavis.edu	biosim (Auto)	(4)	24919	14400	10/28/2009 2:47:10 PM	11/7/2009 2:47:10 PM	adelpha_ef1a	RUNNING	Yes	input: show|http|ftp output: show|http|ftp log file: show DIR	STOP RESTART NOTIFY:S 0 1 E P STAT: SC FL FU PM
	7	102820	MrBayes	10/28/2009 2:44:28 PM	wksavage@ucdavis.edu	biosim (Auto)	(4)	24918	14400	10/28/2009 2:47:09 PM	11/7/2009 2:47:09 PM	adelpha_coi	RUNNING	Yes	input: show|http|ftp output: show|http|ftp log file: show DIR	STOP RESTART NOTIFY:S 0 1 E P STAT: SC FL FU PM
	8	102819	P-CLUSTALW	10/28/2009 2:22:45 PM	bukowski@cac.cornell.edu	cbsusrv05 (Auto)	(2)	886	180	10/28/2009 2:22:50 PM	10/28/2009 2:22:52 PM	clustalw_job	FINISHED CORRECTLY	Yes	input: show|http|ftp output: show|http|ftp log file: show DIR	ARCHIVE DEL RESTART NOTIFY:S 0 1 E P STAT: SC FL FU PM
	9	102818	IMa	10/28/2009 2:22:26 PM	forister@gmail.com	cbsum1 (Auto)	(1)	16239	4320	10/28/2009 2:22:51 PM	10/31/2009 2:22:51 PM	kkbbIM_3_H_numfixed_4.txt	RUNNING	Yes	input: show|http|ftp output: show|http|ftp log file: show DIR	STOP RESTART NOTIFY:S 0 1 E P STAT: SC FL FU PM
	10	102817	IMa	10/28/2009 2:13:17 PM	forister@gmail.com	cbsum1 (Auto)	(1)	16238	4320	10/28/2009 2:13:42 PM	10/28/2009 2:18:06 PM	kkbbIM_3_H_numfixed_4.txt	FINISHED CORRECTLY	Yes	input: show|http|ftp output: show|http|ftp log file: show DIR	ARCHIVE DEL RESTART NOTIFY:S 0 1 E P STAT: SC FL FU PM
	11	102816	MrBayes	10/28/2009 2:00:46 PM	wksavage@ucdavis.edu	cbsum1 (Auto)	(4)	16237	14400	10/28/2009 2:01:10 PM	11/7/2009 2:01:10 PM	adelpha_by_gene	RUNNING	Yes	input: show|http|ftp output: show|http|ftp log file: show DIR	STOP RESTART NOTIFY:S 0 1 E P STAT: SC FL FU PM
	12	102815	STRUCTURE	10/28/2009 1:59:01 PM	g62mudlm@uco.es	cbsum2k8 (Auto)	(11)	1622	1440	10/28/2009 2:10:49 PM	10/29/2009 2:10:49 PM	LOC.str	RUNNING	Yes	input: show|http|ftp output: show|http|ftp log file: show DIR	STOP RESTART NOTIFY:S 0 1 E P STAT: SC FL FU PM
	13	102814	IM	10/28/2009 1:57:31 PM	felipemartins@hotmail.com	cbsum1 (Auto)	(1)	16236	7200	10/28/2009 1:57:54 PM	11/2/2009 1:57:54 PM	desmodusIM.txt	RUNNING	Yes	input: show|http|ftp output: show|http|ftp log file: show DIR	STOP RESTART NOTIFY:S 0 1 E P STAT: SC FL FU PM
	14	102813	IM	10/28/2009 1:56:27 PM	felipemartins@hotmail.com	cbsum1 (Auto)	(1)	16235	7200	10/28/2009 1:56:51 PM	11/2/2009 1:56:51 PM	desmodusIM.txt	RUNNING	Yes	input: show|http|ftp output: show|http|ftp log file: show DIR	STOP RESTART NOTIFY:S 0 1 E P STAT: SC FL FU PM

Job administration.

BioHPC as a web service

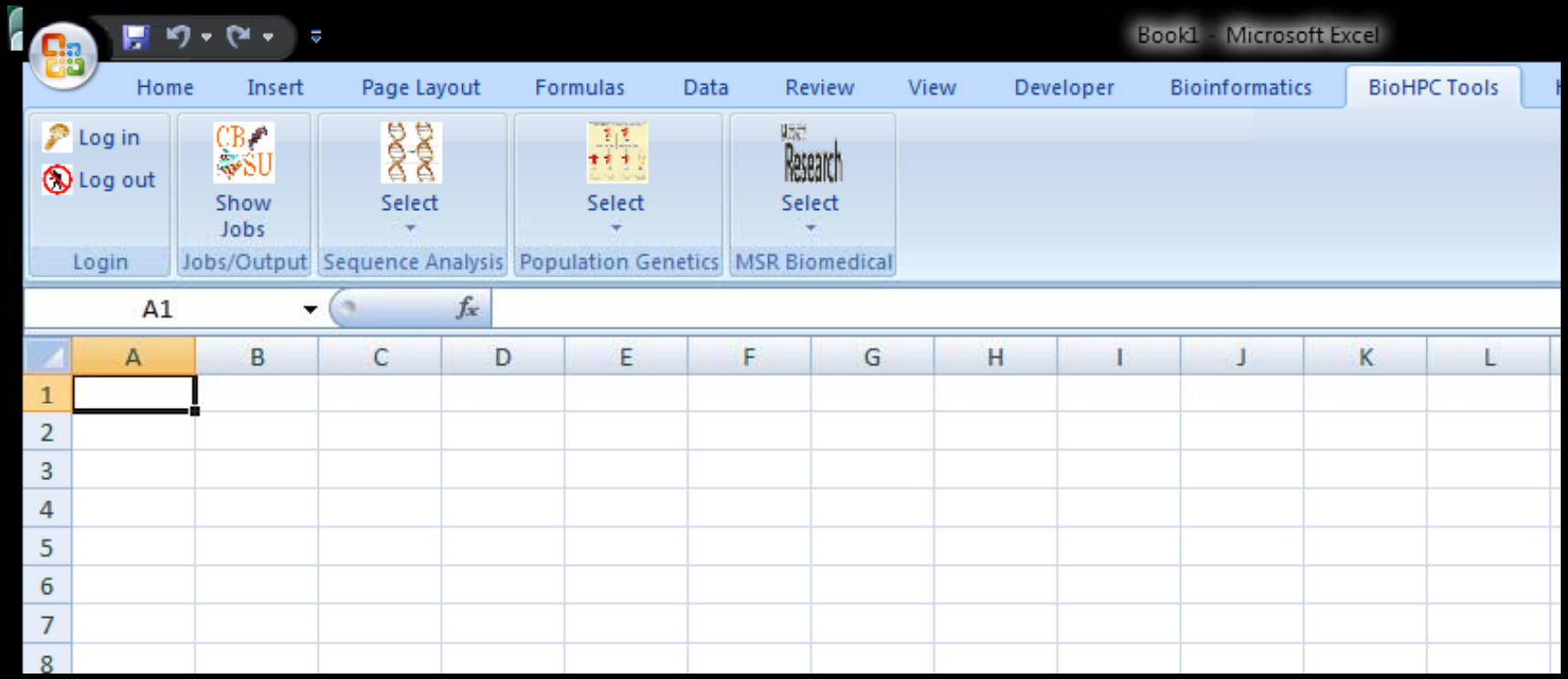
- Convenient user interfaces other than web forms, e.g. Excel
  - Job submission with immediate results visualization / analysis
- Incorporate HPC applications in automated analysis pipelines
  - Especially important in the context of Next Generation Sequencing pipelines
- BioHPC resources available through Microsoft Biology Foundation for command-line utilization

Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)

Web service application example: BioHPC Excel add-in

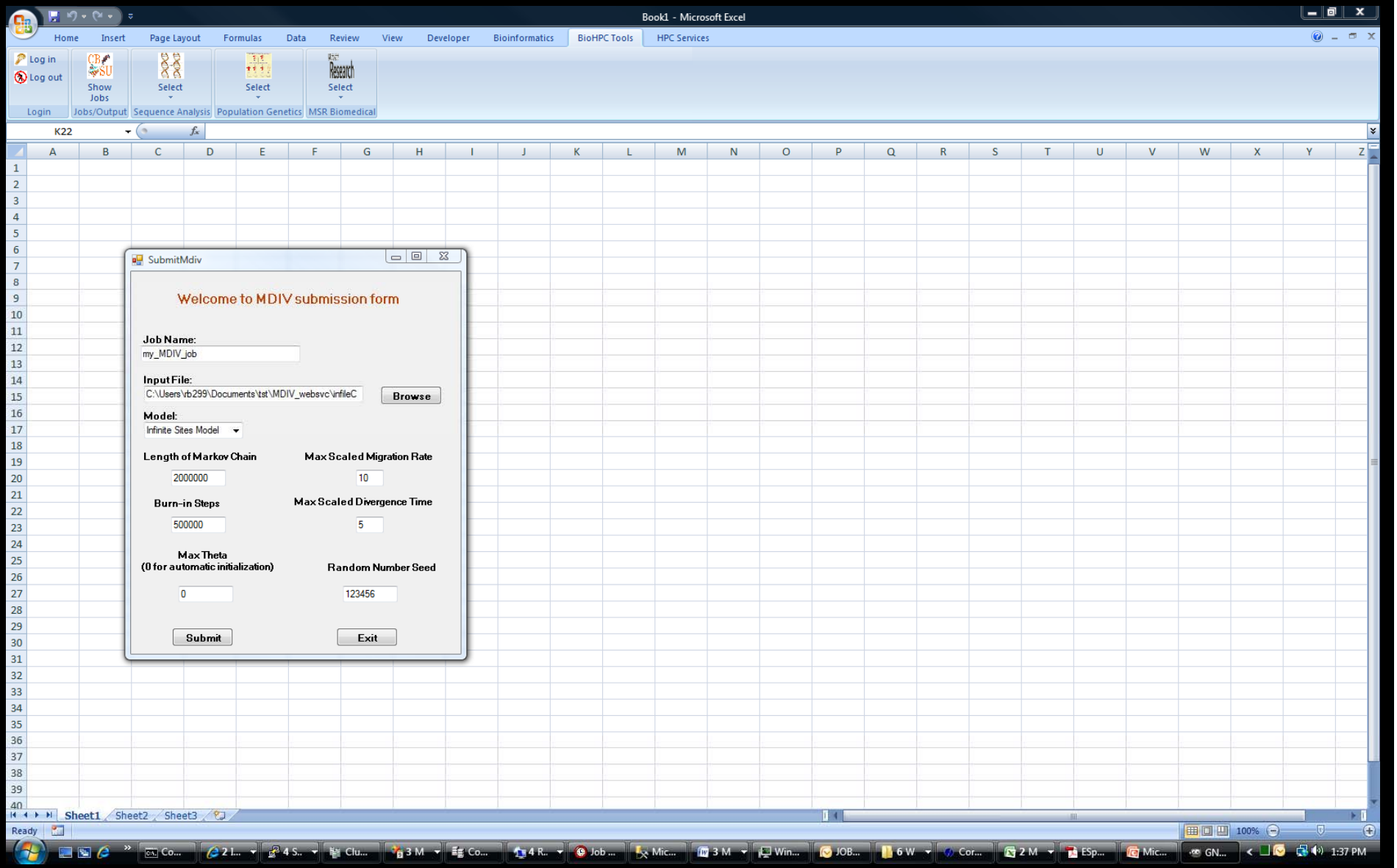


Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)

CB SU

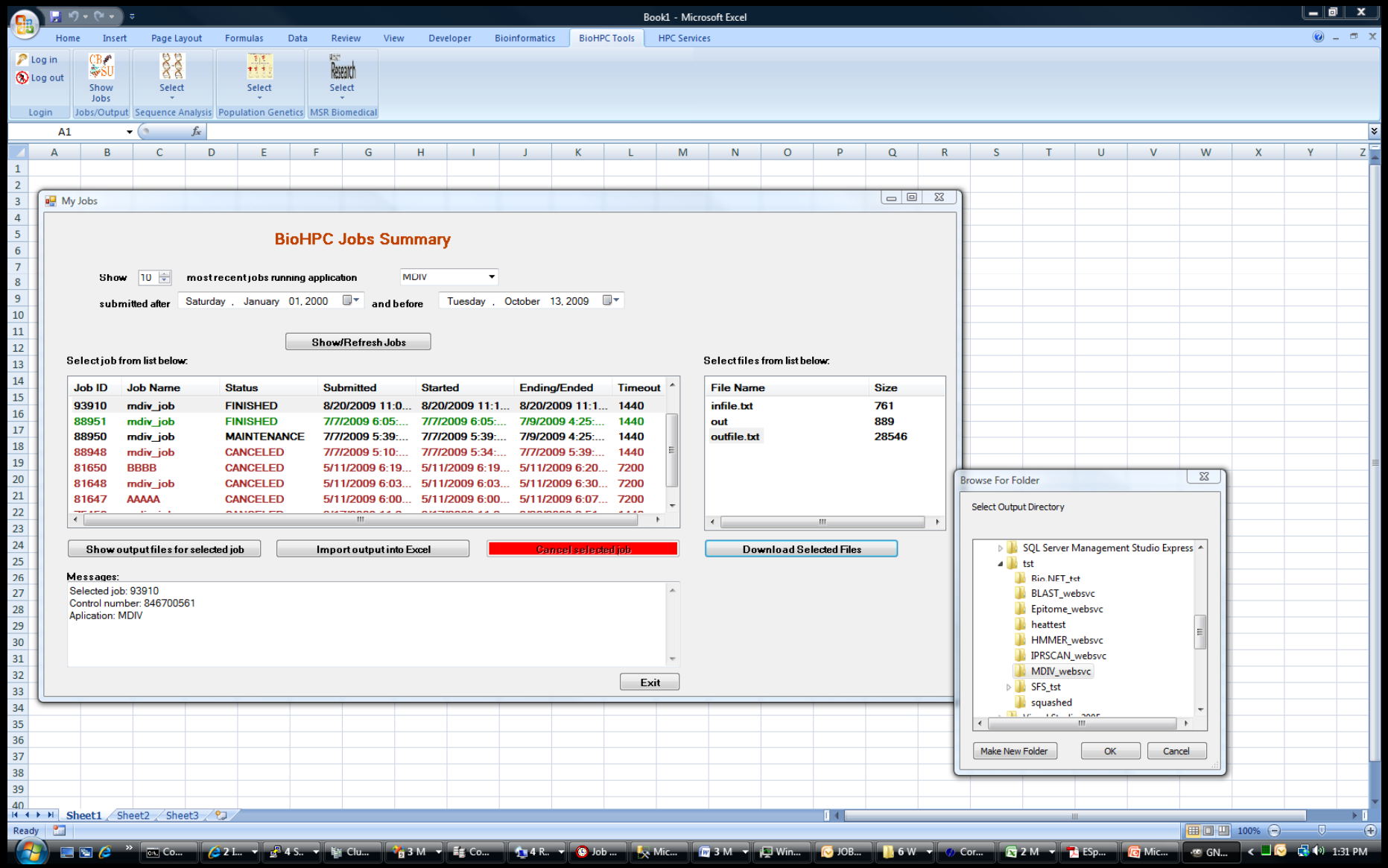


Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)

CB SU

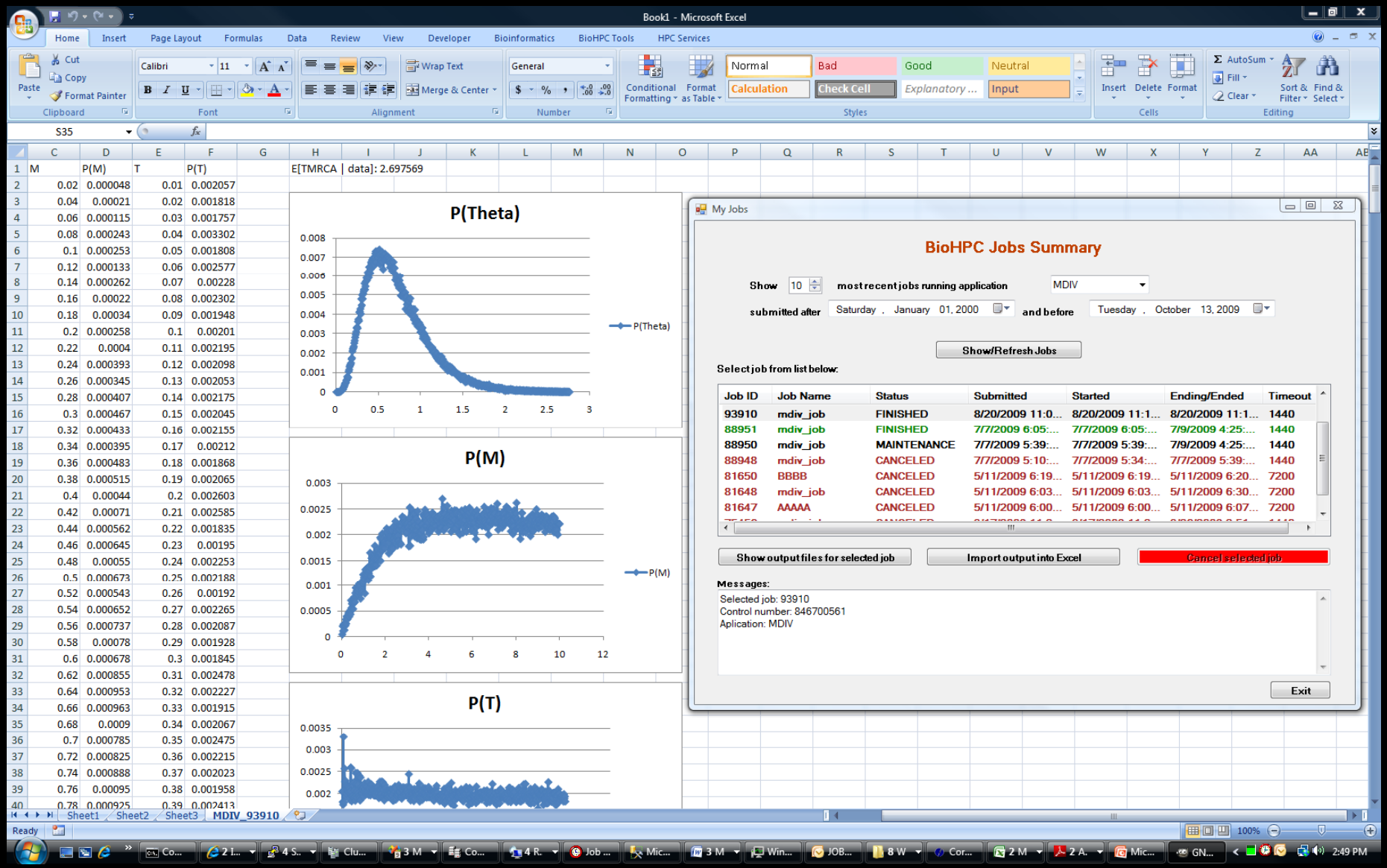


Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)

CB SU



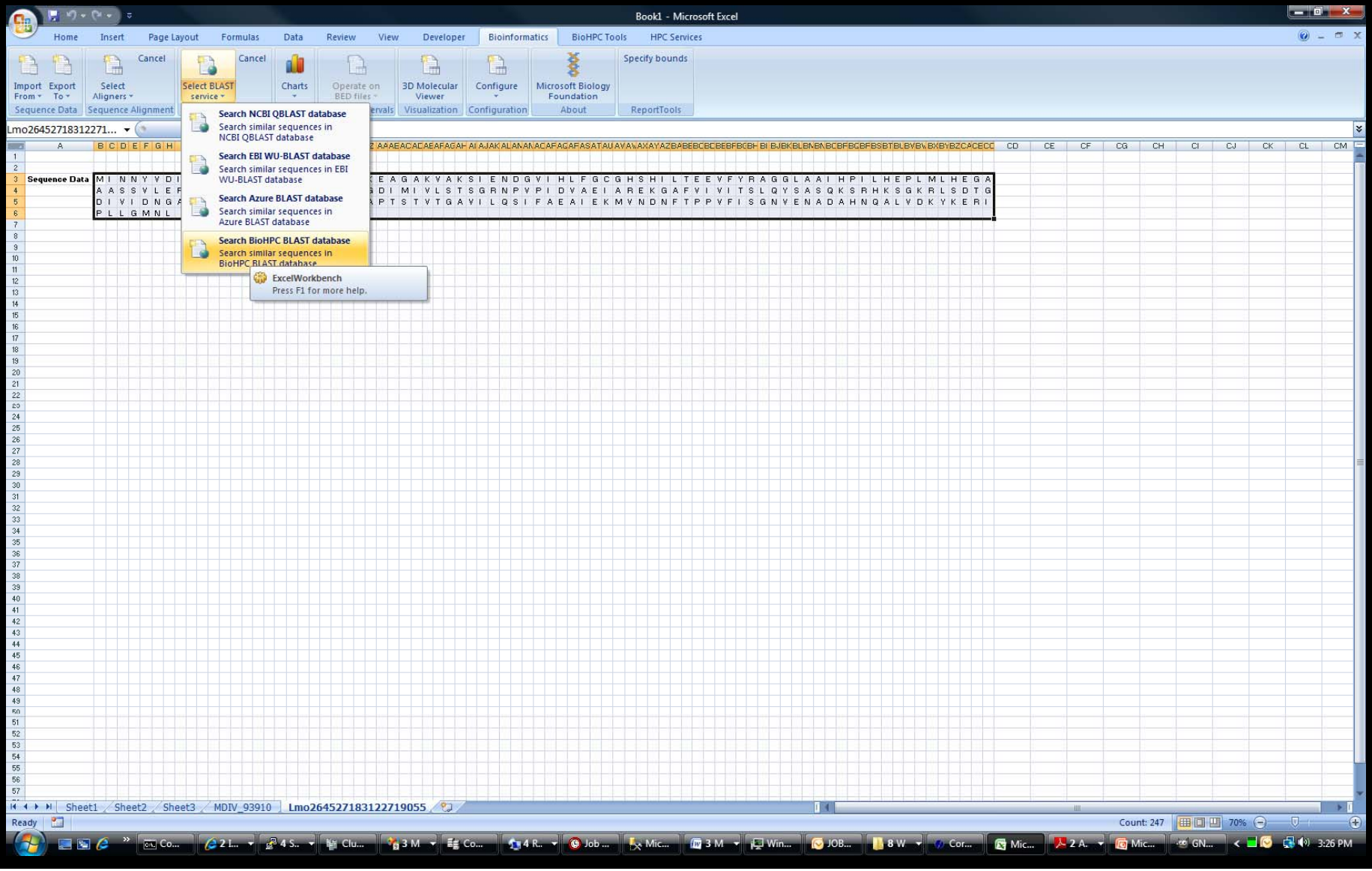
Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)

CB SU

BioHPC BLAST web service in Microsoft Biology Foundation client



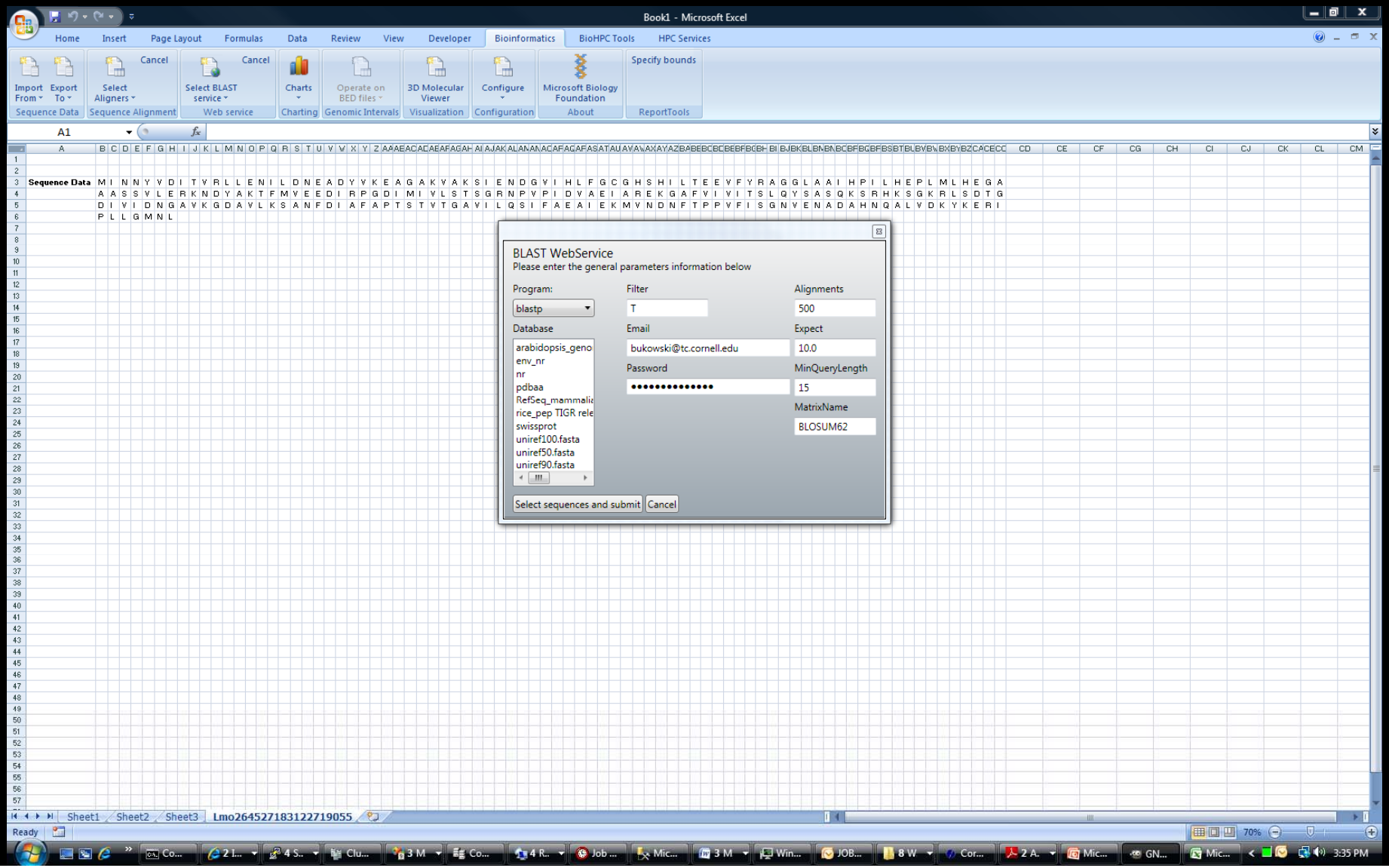
Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)

CB SU

BioHPC BLAST web service in Microsoft Biology Foundation client



Next Generation Sequencing and BioHPC

There are a few LIMS-like systems available for next generation sequencing, but none has HPC bio computing implemented and none uses Windows

Computational Biology Applications Suite for High Performance Computing (BioHPC)

Next Generation Sequencing @ BioHPC

- Data management

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.

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.

Computational Biology Applications Suite for High Performance Computing (BioHPC)

Next Generation Sequencing @ BioHPC

- Data analysis

Reference Manager: allows users to upload and catalogue reference genome files and annotation files needed in downstream data analysis.

Pipeline Manager (under development): allows users to construct and run various analysis pipelines using sequencing reads and reference files stored at BioHPC as input. While default parameters are provided, steps of each pipeline will be individually configurable by a user. Users interface with pipeline manager using our specially constructed web interface or using a web service layer. Computationally intensive steps run on clusters linked to BioHPC.

The web service interface 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.

Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)

CB SU

http://cbsuapps.tc.cornell.edu/Sequencing/runmanager.aspx

File Edit View Favorites Tools Help

Run Manager

Next-Gen @ BioHPC

Log out

Home
Run Manager
Lane Browser
Reference Mgt

Change password
Reset password

BioHPC @ CBSU
contact CBSU

RUN MANAGER

This page lists all runs configured in the system. Use the **Create New Run** button to configure a new run. To sort the data by a certain column, click on that column's header. To filter the data, supply templates (all usual wildcards work) and click **Apply Filters**. To configure lane information for a run, click on that run's entry in the **Run ID** column.

Create new run

Filter data by

Cell name: * Directory: * Users: * Apply Filters

|< < Page 2 of 4 > >| Go to page 2

Run ID	Run Name	Flowcell Name	Directory	Entry Created	Entry Changed	Users
23	100218_HWI-EAS83_0004_614ABAAXX	614ABAAXX	/data/disk2/dnaservices/100218_HWI-EAS83_0004_614ABAAXX/Data/C1-86_Firecrest1.6.0_23-02-2010_dnaservices/Bustard1.6.0_23-02-2010_dnaservices/GERALD_25-02-2010_dnaservices	2/25/2010 1:34:21 PM	2/25/2010 2:04:01 PM	okv2@cornell.edu pas48@cornell.edu rlt8@cornell.edu arboyko@gmail.com led8@cornell.edu nds25@cornell.edu tjs11@cornell.edu
22	100211_HWUSI-EAS690_0001_617TKAAXX	617TKAAXX	/data/disk5/dnaservices/100211_HWUSI-EAS690_0001_617TKAAXX/Data/Intensities/BaseCalls/GERALD_25-02-2010_dnaservices	2/25/2010 11:22:01 AM	3/1/2010 10:47:47 AM	smb31@cornell.edu pas48@cornell.edu sen26@cornell.edu ceg2004@med.cornell.edu jch63@cornell.edu jdw25@cornell.edu cwt6@cornell.edu clo5@cornell.edu tjs11@cornell.edu
21	100210_HWI-EAS339_0001_613V2AAXX	613V2AAXX	/data/disk4/dnaservices/100210_HWI-EAS339_0001_613V2AAXX/Data/Intensities/BaseCalls/GERALD_23-02-2010_dnaservices	2/25/2010 10:31:22 AM	2/25/2010 11:41:59 AM	pas48@cornell.edu cwt6@cornell.edu tjs11@cornell.edu
20	100211_HWI-EAS83_0002_6100DAAXX	6100DAAXX	/data/disk3/dnaservices/100211_HWI-EAS83_0002_6100DAAXX/Data/Intensities/BaseCalls/GERALD_14-02-2010_dnaservices	2/14/2010 10:14:23 AM	2/24/2010 1:51:15 PM	jjv2@cornell.edu fac2003@med.cornell.edu pas48@cornell.edu smm225@cornell.edu ps27@cornell.edu tjs11@cornell.edu
						esk72@cornell.edu

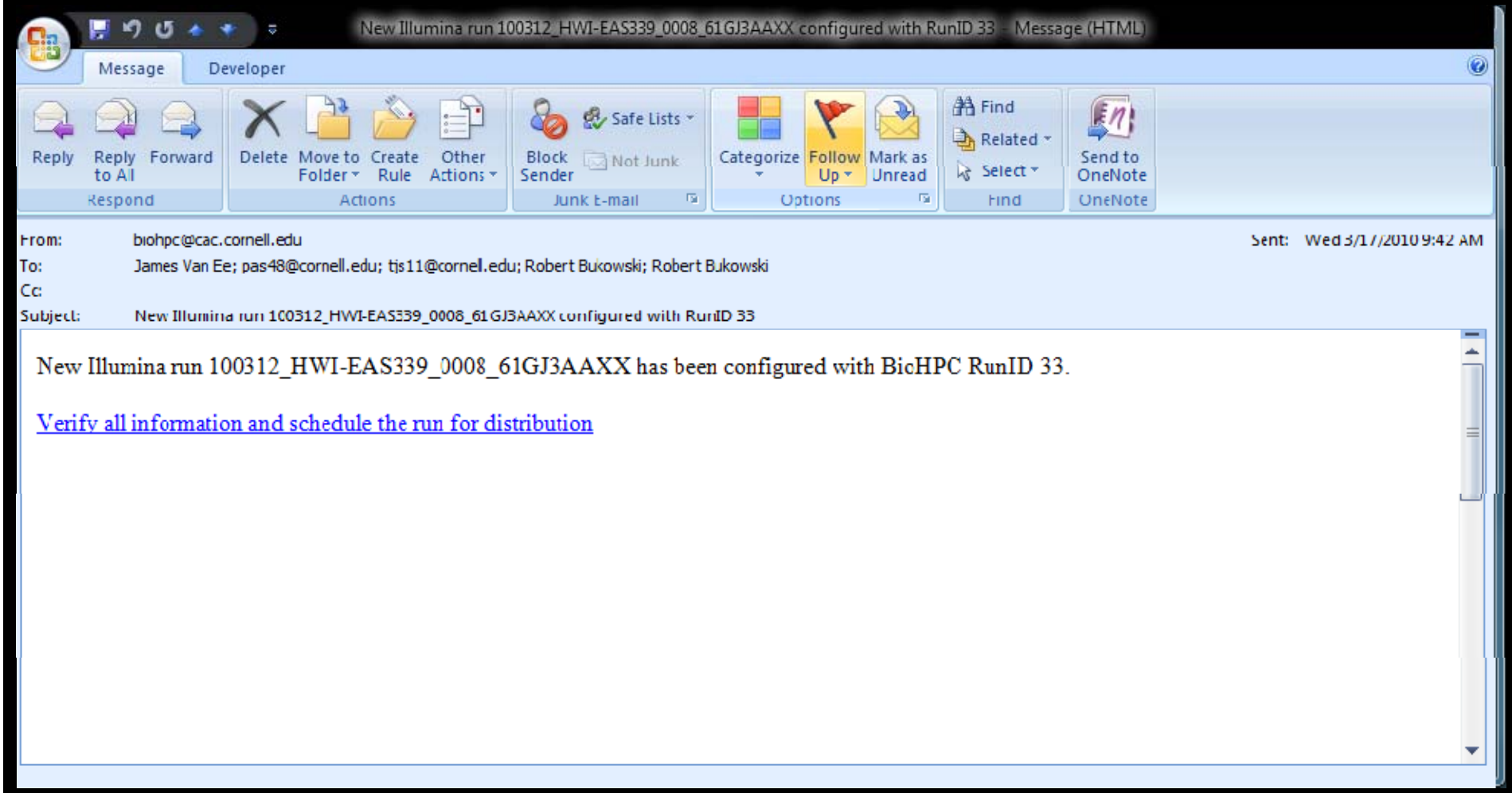
Internet | Protected Mode: On 100%

Intercept finished sequencing runs and configure them in BioHPC data manager.

Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)



Notify sequencing facility administrators about the new results to be approved for distribution to users.

Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)

CB SU

http://cbsuapps.tc.cornell.edu/Sequencing/runmanager.aspx?runid=33

File Edit View Favorites Tools Help

Run Manager

Log out

Home
Run Manager
Lane Browser
Reference Mgt

Change password
Reset password

BioHPC @ CBSU
contact CBSU

EDIT RUN 33

Extract flowcell name and machine run name from directory string

Run ID	Run Name	Flowcell Name	Control Lane	Entry Created	Entry Changed
33	100312_HWI-EAS339_0008_61GJ3AAXX	61GJ3AAXX	8	3/17/2010 9:41:47 AM	3/18/2010 9:28:52 AM
Location:	/data/disk3/dnaservices/100312_HWI-EAS339_0008_61GJ3AAXX/Data/Intensities/BaseCalls/GERALD_17-03-2010_dnaservices				

Lane info:

Lane	Type	Sample Name	Prefilter	Status	Users (check box to remove upon submission)	Lab	Order#	Clear upon submission
1	Standard	Par	No	ready	jr286@cornell.edu set as owner --Add user from list-- Add users by e-mail:	N/A	10214774	
2	Standard	Var	No	ready	jr286@cornell.edu set as owner --Add user from list-- Add users by e-mail:	N/A	10214774	
3	Standard	End	No	ready	jr286@cornell.edu set as owner --Add user from list-- Add users by e-mail:	N/A	10214774	
4	Standard	V-1	No	ready	vnm1@cornell.edu set as owner --Add user from list-- Add users by e-mail:	N/A	10216342	
5	Standard	mel2 sim2 yak2 ana2	No	processing(zip)	jl2434@cornell.edu set as owner --Add user from list-- Add users by e-mail:	N/A	10216484	
6	Standard	Pooled samples	No	approved	samantha.brooks@cornell.edu set as owner --Add user from list-- Add users by e-mail:	N/A	10216429	
7	Standard	Pooled samples	No	approved	ac347@cornell.edu set as owner --Add user from list-- Add users by e-mail:	N/A	10216609	

runmanager.aspx?runid=33 Internet | Protected Mode: On 100%

Approval page for sequencing facility. Transfer (asynchronous) to BioHPC will start after a lane status is changed to approved.

Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)

CBSU

http://cbsuapps.tc.cornell.edu/Sequencing/lanebrowser.aspx

Lane Browser

Log out

Home
Run Manager
Lane Browser
Reference Mgt

Change password
Reset password

BioHPC @ CBSU
contact CBSU

BROWSE ALL LANES

This 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 an 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 assistance with download of multiple files in batch mode, use the **Make download script** button.

Make download script Register new lane

("check" all lanes you want to include in a download script and click button above) (use this option only if you want to manually upload a lane from outside of the CLC sequencing facility)

Filter data by

Status: All Sample name: * Users: * Apply Filters

|< < Page 1 of 10 > >| Go to page 1

Lane ID	Run ID	Run Name	Lane#	Type	Sample Name	Status	Annotations	Users	Lab	Order#
191 (files)	33	100312_HWI-EAS339_0008_61GJ3AAXX	1	Standard	Par	ready	Parameter This Lane Ctrl Lane Length unknown unknown Clusters_raw 18.8M 27.6M Clusters_PF 11.9M 16.8M	jr286@cornell.edu (owner)	N/A	10214774
192 (files)	33	100312_HWI-EAS339_0008_61GJ3AAXX	2	Standard	Var	ready	Parameter This Lane Ctrl Lane Length unknown unknown Clusters_raw 18.7M 27.6M Clusters_PF 14.4M 16.8M	jr286@cornell.edu (owner)	N/A	10214774
193 (files)	33	100312_HWI-EAS339_0008_61GJ3AAXX	3	Standard	End	ready	Parameter This Lane Ctrl Lane Length unknown unknown Clusters_raw 25.2M 27.6M Clusters_PF 20.0M 16.8M	jr286@cornell.edu (owner)	N/A	10214774
194	33	100312_HWI-EAS339_0008_61GJ3AAXX	4	Standard	V-1	processing (transit)	--- NO ANNOTATIONS ---	vnm1@cornell.edu (owner)	N/A	10216342
195	33	100312_HWI-EAS339_0008_61GJ3AAXX	5	Standard	mel2 sim2 yak2 ana2	processing (zip)	--- NO ANNOTATIONS ---	jl2434@cornell.edu (owner)	N/A	10216484
196	33	100312_HWI-EAS339_0008_61GJ3AAXX	6	Standard	Pooled samples	approved	--- NO ANNOTATIONS ---	samantha.brooks@cornell.edu (owner)	N/A	10216429
197	33	100312_HWI-EAS339_0008_61GJ3AAXX	7	Standard	Pooled samples	approved	--- NO ANNOTATIONS ---	ac347@cornell.edu (owner)	N/A	10216609
198	33	100312_HWI-EAS339_0008_61GJ3AAXX	8	PhiX	PhiX	pending	--- NO ANNOTATIONS ---	tjs11@cornell.edu (owner)	N/A	NA
183	32	100312_HWI_EAS339_0008_61GJ3AAXX	1	Standard	Par	pending	--- NO ANNOTATIONS ---	jr286@cornell.edu (owner)	N/A	10214774
184	32	100312_HWI_EAS339_0008_61GJ3AAXX	2	Standard	Var	pending	--- NO ANNOTATIONS ---	jr286@cornell.edu (owner)	N/A	10214774

http://cbsuapps.tc.cornell.edu/resetpass.aspx

Main administration page for lanes. Users can only manage their own data.

Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)

CB SU

[BioHPC]: CLC Illumina Genome Analyzer Results: 100312_HWI-EAS339_0008_61GJ3AAXX / Order: 10214774 / Sample: Par - Message...

Message | Developer

Reply | Reply to All | Forward | Delete | Move to Folder | Create Rule | Other Actions | Block Sender | Safe Lists | Not Junk | Categorize | Follow Up | Mark as Unread | Find | Related | Select | Send to OneNote

Respond | Actions | Junk E-mail | Options | Find | OneNote

From: biohpc@cac.cornell.edu
Sent: Thu 3/18/2010 9:45 AM
To: jr286@cornell.edu
Cc: James Van Ee; pas48@cornell.edu; tjs11@cornell.edu; Robert Bukowski
Subject: [BioHPC]: CLC Illumina Genome Analyzer Results: 100312_HWI-EAS339_0008_61GJ3AAXX / Order: 10214774 / Sample: Par

The sequencing result for sample "**Par**" (**lane 1** from run **100312_HWI-EAS339_0008_61GJ3AAXX**) has been registered with the CBSU file manager as **Lane 191**. The files are ready for pickup using the following link:

http://cbsuapps.tc.cornell.edu/Sequencing/showseqfile.aspx?mode=outlist&cntrl=1758691306&laneid=191

This and other sequencing results may also be retrieved using the following methods:

- From the **Order Status/Results** section of the CLC **web site** (https://cores.lifesciences.cornell.edu/userdev)
- From the **BioHPC Next Generation Data Anlysis site** at **Computational Biology Service Unit (CBSU)** (http://cbsuapps.tc.cornell.edu/Sequencing/seqmain.aspx). To learn how to access the BioHPC site please see the explanation below.

About BioHPC:
The BioHPC Next Generation Sequencing Data Analysis site allows users to access the results of Illumina sequencing runs performed on their behalf by the Sequencing Facility at Cornell CLC. Read files obtained outside of this facility can also be catalogued and stored here. In the future, users will also be able to upload and manage reference genome and annotation files and run various analysis pipelines involving read and reference files stored and catalogued at CBSU. To access the site, navigate to

http://cbsuapps.tc.cornell.edu/Sequencing/seqmain.aspx

and log in using your e-mail address jr286@cornell.edu as login ID and your BioHPC password. Please note that your account and password at BioHPC are **separate** from your account at the CLC sequencing facility. If you do not yet know your BioHPC password or if you need to reset it, go to

http://cbsuapps.tc.cornell.edu/resetpass.aspx?userId=jr286@cornell.edu

If you would like to use the BioHPC site but are not yet registered, contact CBSU at http://cbsuapps.tc.cornell.edu/contactus.aspx.

File Formats and Data:
Data are distributed as gzipped "fastq" formatted files generated from the Illumina Genome Analyzer Analysis Pipeline. Unfiltered data in Illumina ".qseq" format or other intermediate data from the pipeline and data collection process are available by special arrangement with the CLC (dna_services@cornell.edu). Special data handling may result in additional processing fees.

Data Retention Policy:
Data distributed through the BioHPC web site is available for 30 days (once analysis pipelines become operational, retention time may be extended for lanes involved in calculations). After 30 days, the data may be reposted subject to a service fee. The CLC maintains an archive of all sequencing data, however, we cannot guarantee the availability of data after the initial distribution. Customers are responsible for retrieving and storing their data safely. CLC and BioHPC cannot serve as a backup or archive.

More Information:
Life Science Core Laboratories Center (CLC): http://cores.lifesciences.cornell.edu/

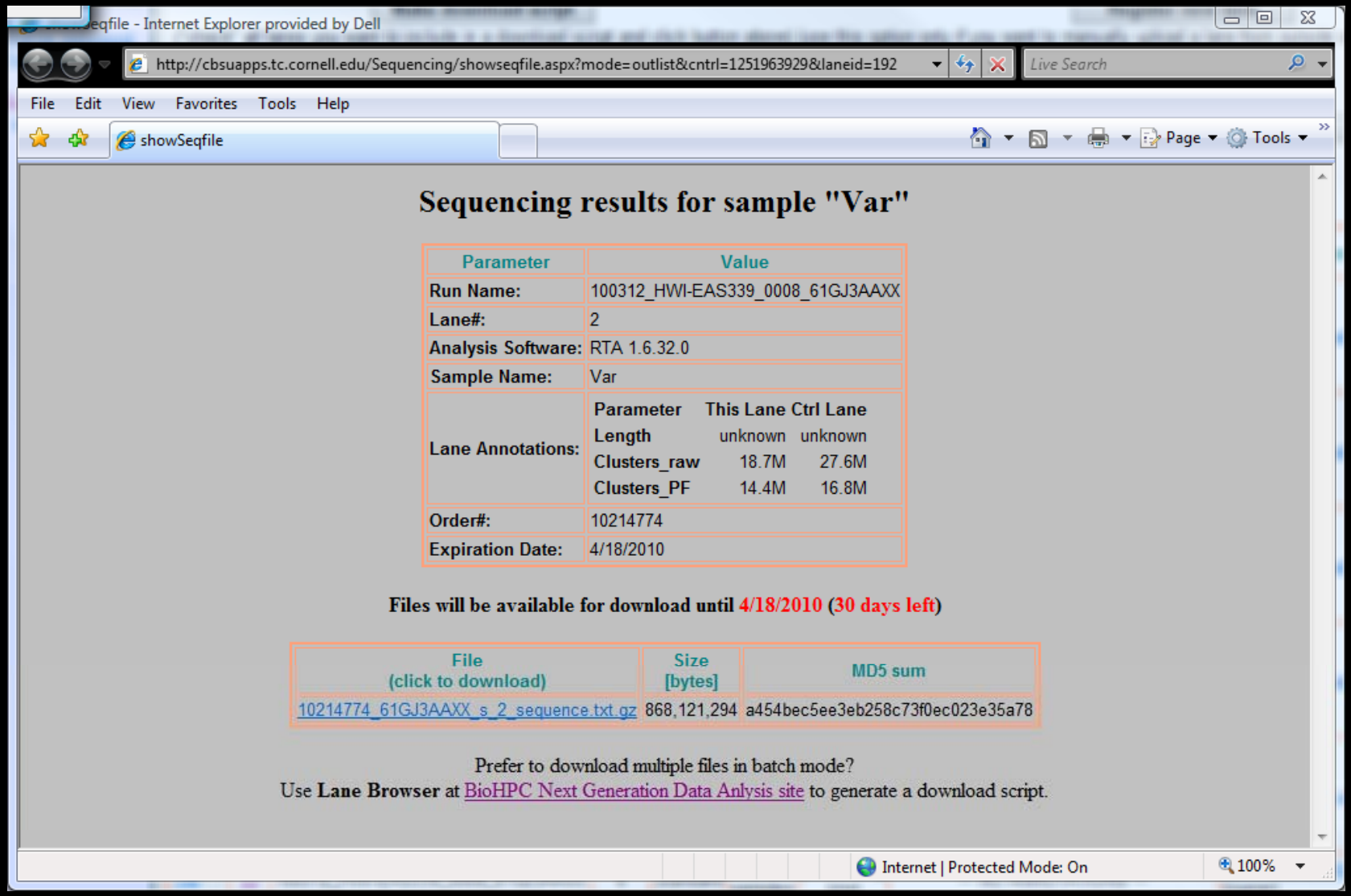
Once data files are transferred users obtain links to download them.

Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)

CB SU



User data download page.

Problems

- HPC basic profile / JSDL limitations in Windows Server 2008 HPC

  Only limited subset of commands and controls is implemented in the native HPC Basic Profile service. Need to develop BioHPC web service to control HPC scheduler.

- SUA and porting

  Porting applications to Windows environment and convincing original authors to keep them updated on Windows is still a challenge. SUA only supported 32 bit development, severely limiting memory usage – same with Cygwin. Direct native porting not an efficient choice for rapidly changing, not yet established software. Experimenting with MINGW 64 bit environment.

Computational Biology Applications Suite for High Performance Computing (BioHPC)

Progress to date summary

- Fully functional HPC computational biology application suite offering 37 applications, resource integration and management available as open source.
- Very popular service – massively utilized by Cornell and external users
- Web service access implemented for several applications, will be available for all suitable applications by the end of this year
- Integration with MBF via web services and Excel client implementation is in progress for several applications. Extensive participation in MBF development and testing.
- Fully implemented next generation sequencing data manager with asynchronous data transfer from sequencing facility and data distribution to users

Future effort and directions

- Keeping up to date with new developments in bio computing: applications and algorithm/software updates, especially in next generation sequencing
- Better integration with MBF, especially regarding data management
- Integrating with commercial applications (Ilumina, Real Time Genomics)
- Full implementation of next generation sequencing pipeline manager as web form, web services integrated into MBF, and available as Trident workflows
- Support for Azure cloud – will allow users to install and use BioHPC locally and utilize Azure as a remote HPC resource
- Improved internal maintenance tools (external authentication, better user group management, improved asynchronous data transfer)

Cornell University
Life Sciences
Core Laboratories Center

Computational Biology
Service Unit

Computational Biology Applications Suite for High Performance Computing (BioHPC)

CB SU

Computational Biology Service Unit

Cornell core facility for computational biology and bioinformatics

Part of Cornell Life Sciences Core Facilities Center

Provides bioinformatics and computational support for biological research at Cornell and beyond by means of research collaborations, consultations, software development and more.

Genomics/Proteomics
Qi Sun
Lalit Ponnala
Stefan Stefanov

HPC / Computing
Jaroslaw Pillardy (director)
Robert Bukowski
Mary Howard

System Biology
Chris Myers

Computational Biology Applications Suite for High Performance Computing (BioHPC)

Many thanks for Microsoft Research

for support that allowed us to

to develop our own local computing solutions

into a tool that can help others.

Without MSR BioHPC would be just a set of unorganized interface and admin tools useful only for us.