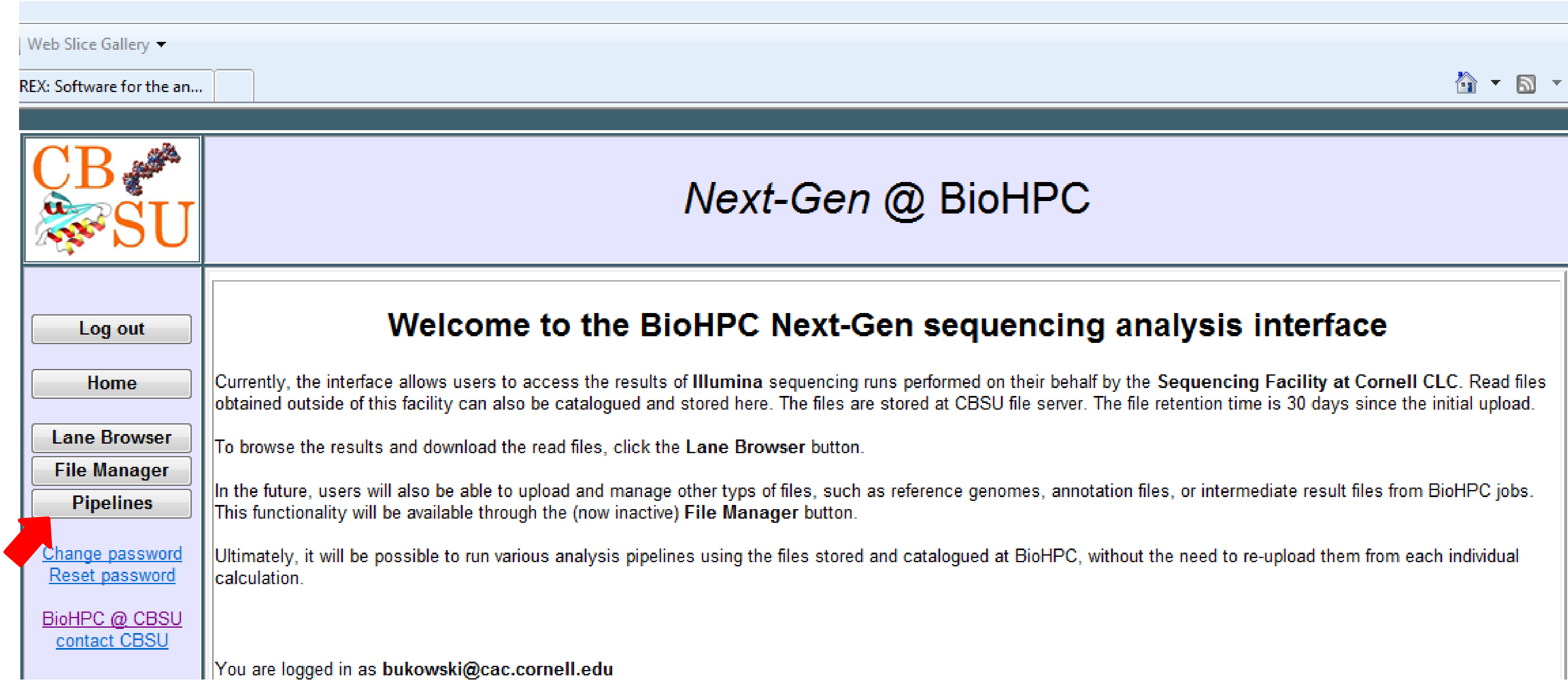
BioHPC next generation sequencing pipelines tutorial

This is a simple example illustrating the use of the BioHPC Pipeline Manager. The objective is to create and execute a somewhat artificial yet illustrative pipeline consisting of 5 steps:

1. **FASTX Trimmer**: trim a set of sequencing reads (from e-coli) to 30 bp
2. **BowtieBuild**: create an index of a reference (e-coli) genome in preparation for alignment
3. **Bowtie**: perform an alignment of reads trimmed in step 1 to the genome indexed in step 2
4. **SamTools**: create an alignment file in BAM format out of the result of step 3 (created in SAM format)
5. **SamTools**: create a "pileup" file from the BAM alignment obtained in step 4

Please note: the BioHPC Pipeline Manager is "work in progress". In particular, graphical details of actual working web pages may differ slightly from what is shown in this tutorial.

With questions/suggestions please contact biohpc@cac.cornell.edu.



Next-Gen @ BioHPC

Welcome to the BioHPC Next-Gen sequencing analysis interface

Currently, the interface 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. The files are stored at CBSU file server. The file retention time is 30 days since the initial upload.

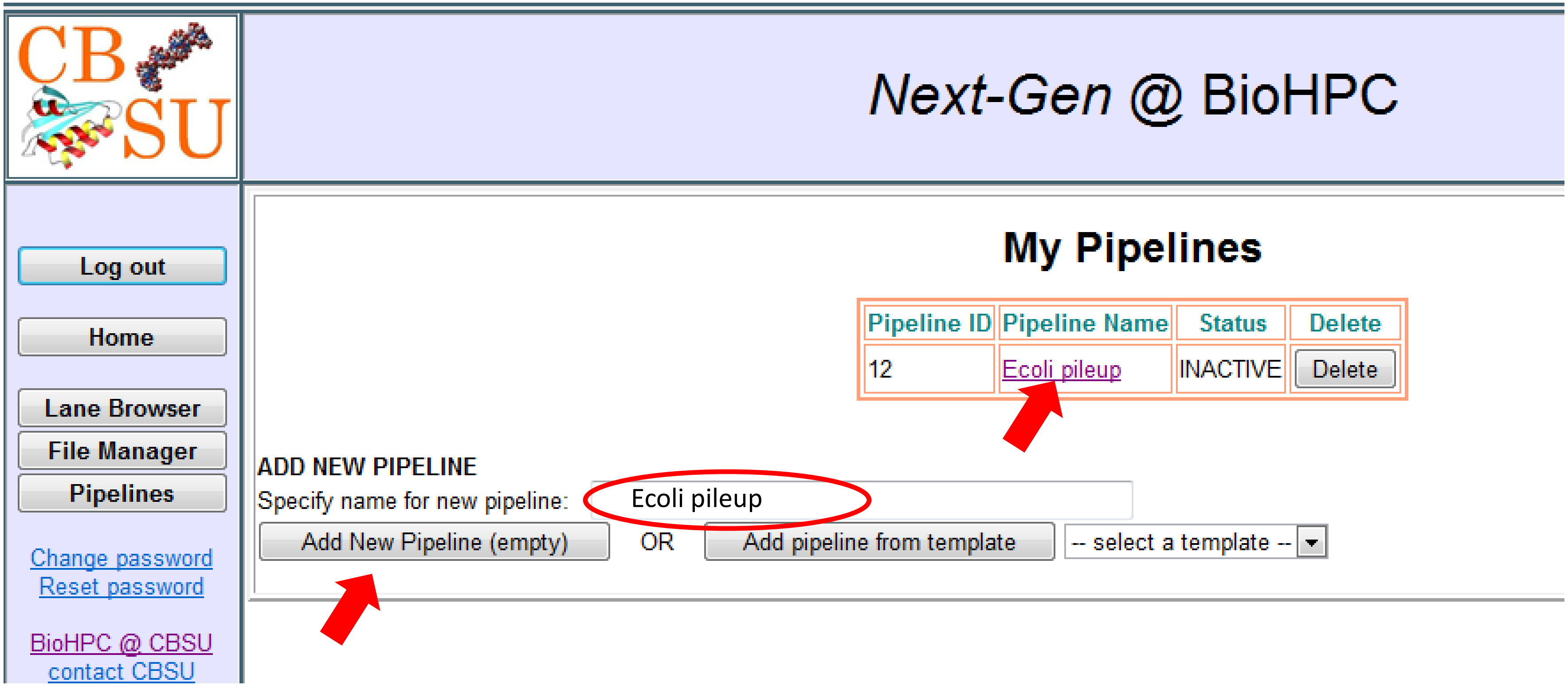
To browse the results and download the read files, click the **Lane Browser** button.

In the future, users will also be able to upload and manage other typs of files, such as reference genomes, annotation files, or intermediate result files from BioHPC jobs. This functionality will be available through the (now inactive) **File Manager** button.

Ultimately, it will be possible to run various analysis pipelines using the files stored and catalogued at BioHPC, without the need to re-upload them from each individual calculation.

You are logged in as **bukowski@cac.cornell.edu**

To access the BioHPC next generation sequencing pipeline manager, navigate to http://cbsuapps.tc.cornell.edu/Sequencing/seqmain.aspx, log in using your registered e-mail address as a login ID, and click on the **Pipelines** button.



Next-Gen @ BioHPC

My Pipelines

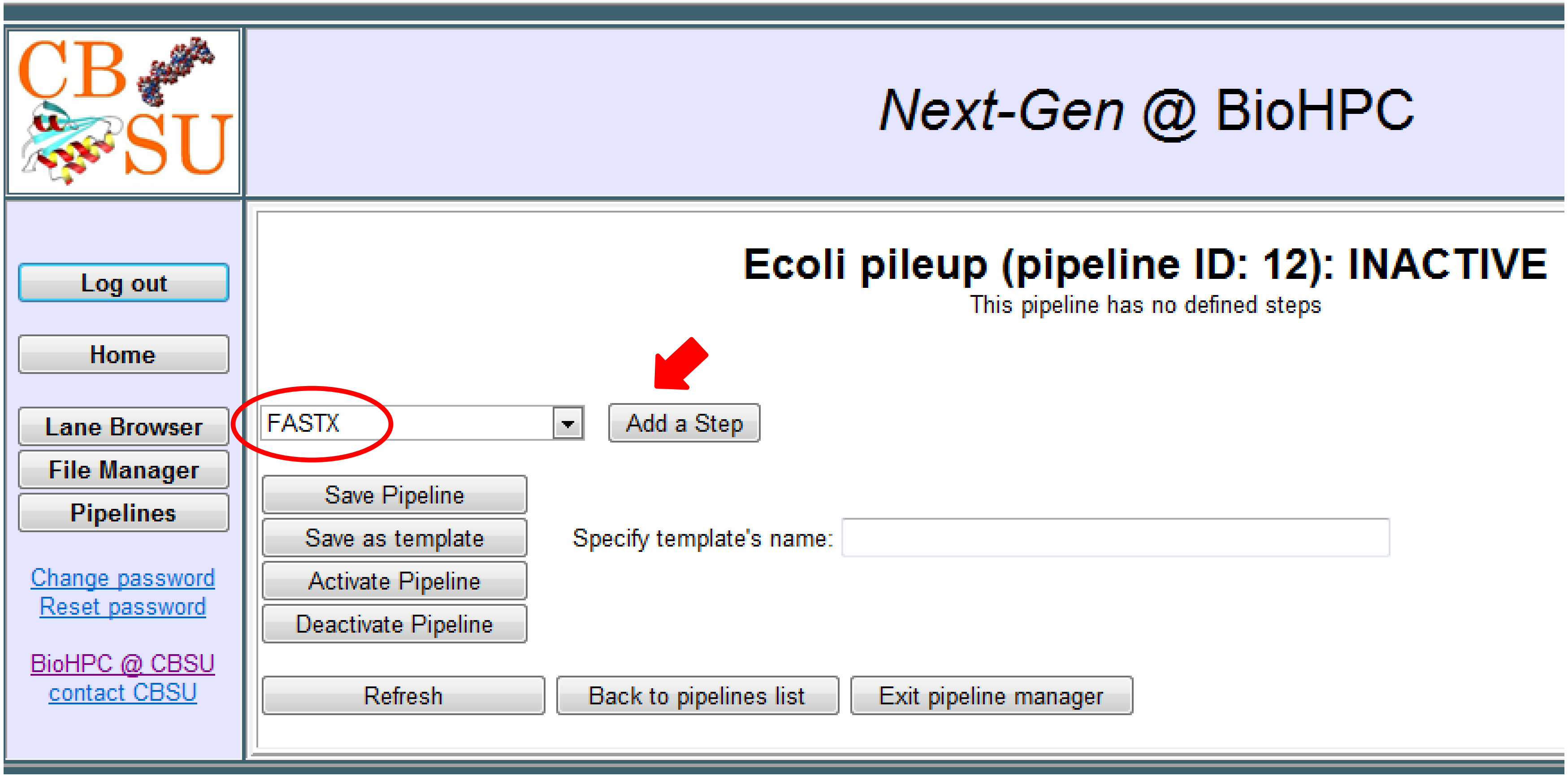
Pipeline ID	Pipeline Name	Status	Delete
12	Ecoli pileup	INACTIVE	Delete

ADD NEW PIPELINE

Specify name for new pipeline: Ecoli pileup

To add a new pipeline, specify its name and click **Add New Pipeline (empty)**. You can also create a new pipeline from a template, if any are defined (see the end of this presentation).

When the new pipeline shows up on the list, click on its name to add and configure steps.



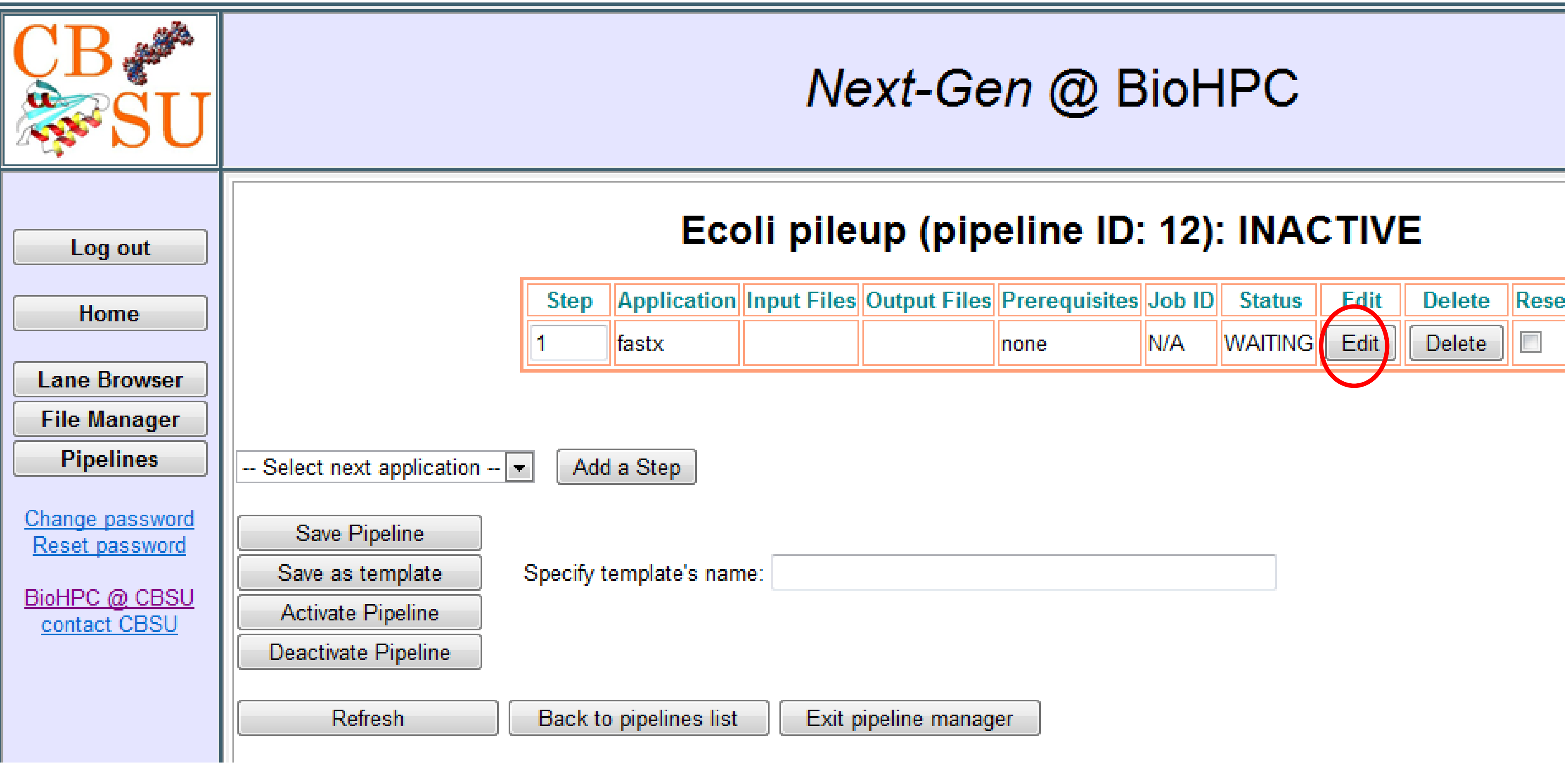
Ecoli pileup (pipeline ID: 12): INACTIVE

This pipeline has no defined steps

Specify template's name:

The new pipeline is still "empty" – there are no steps defined in it. The pipeline's status is shown as **INACTIVE**, i.e., no attempts will be made by BioHPC to submit it until the construction is completed.

The first step we will add to the pipeline is **FASTX Trimmer** application. To do this, select **FASTX** from the applications menu and click **Add a Step**.

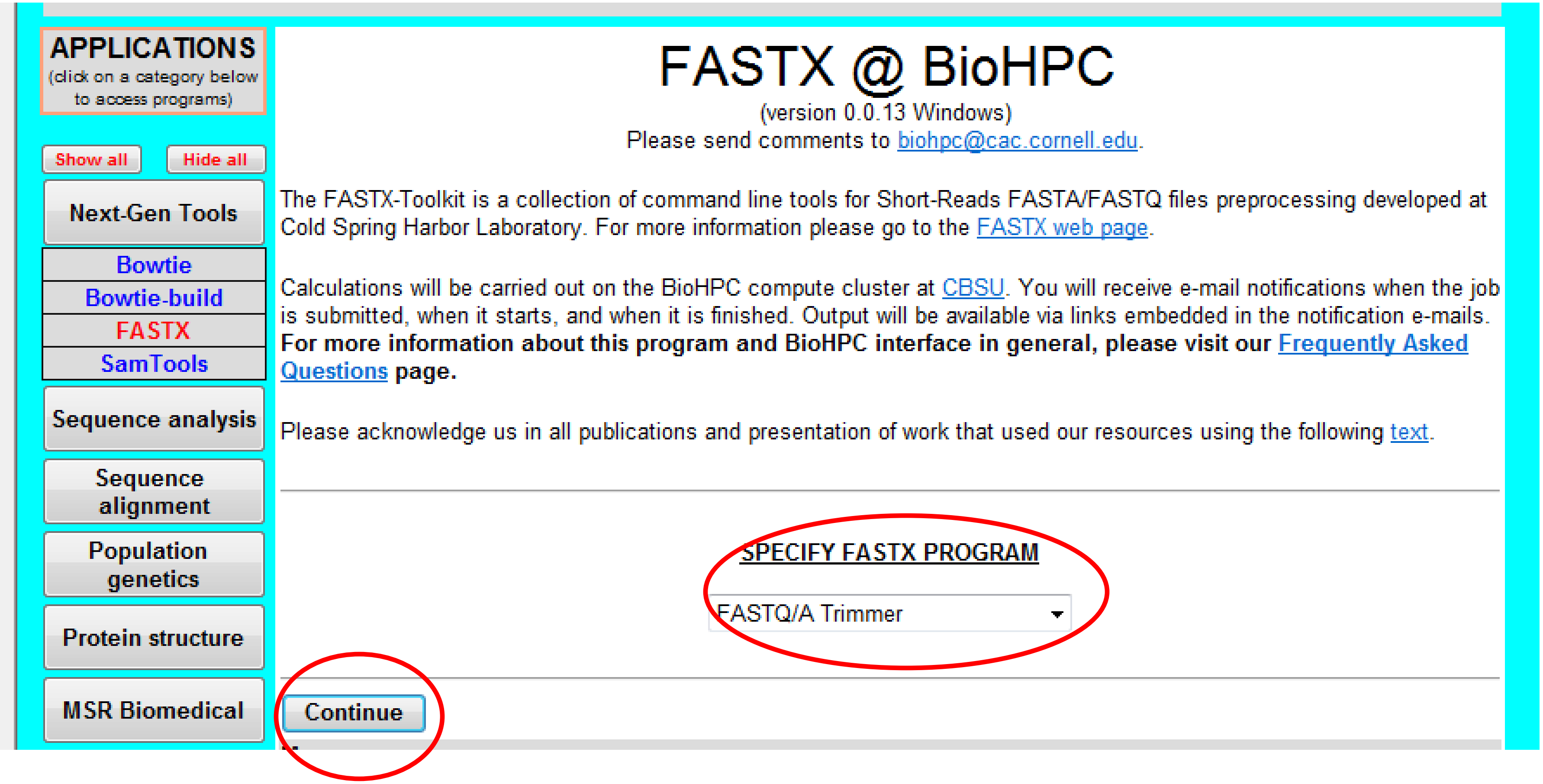


Next-Gen @ BioHPC

Ecoli pileup (pipeline ID: 12): INACTIVE

Step	Application	Input Files	Output Files	Prerequisites	Job ID	Status	Edit	Delete	Rese
1	fastx			none	N/A	WAITING	Edit	Delete	□

An entry for the FASTX step will be added, ready to be configured. To do this, click on the **Edit** button. This will redirect us to the FASTX submission page.



FASTX @ BioHPC

(version 0.0.13 Windows)
Please send comments to biohpc@cac.cornell.edu.

The FASTX-Toolkit is a collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing developed at Cold Spring Harbor Laboratory. For more information please go to the FASTX web page.

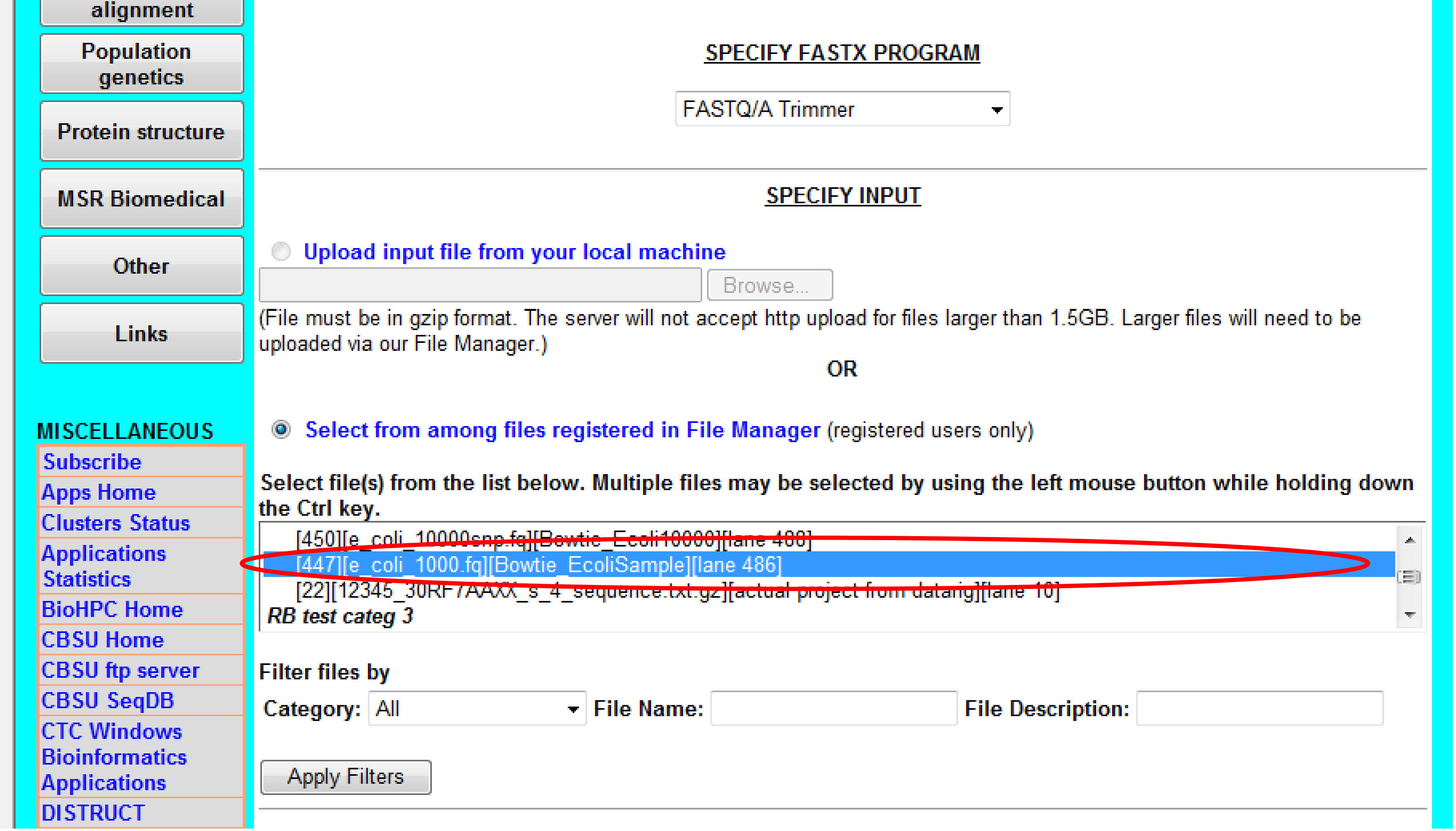
Calculations will be carried out on the BioHPC compute cluster at CBSU. You will receive e-mail notifications when the job is submitted, when it starts, and when it is finished. Output will be available via links embedded in the notification e-mails. **For more information about this program and BioHPC interface in general, please visit our Frequently Asked Questions page.**

Please acknowledge us in all publications and presentation of work that used our resources using the following text.

SPECIFY FASTX PROGRAM

FASTQ/A Trimmer

On the FASTX page, select the **trimmer** subprogram and click **Continue**



SPECIFY FASTX PROGRAM

FASTQ/A Trimmer

SPECIFY INPUT

○ Upload input file from your local machine

(File must be in gzip format. The server will not accept http upload for files larger than 1.5GB. Larger files will need to be uploaded via our File Manager.)

OR

● Select from among files registered in File Manager (registered users only)

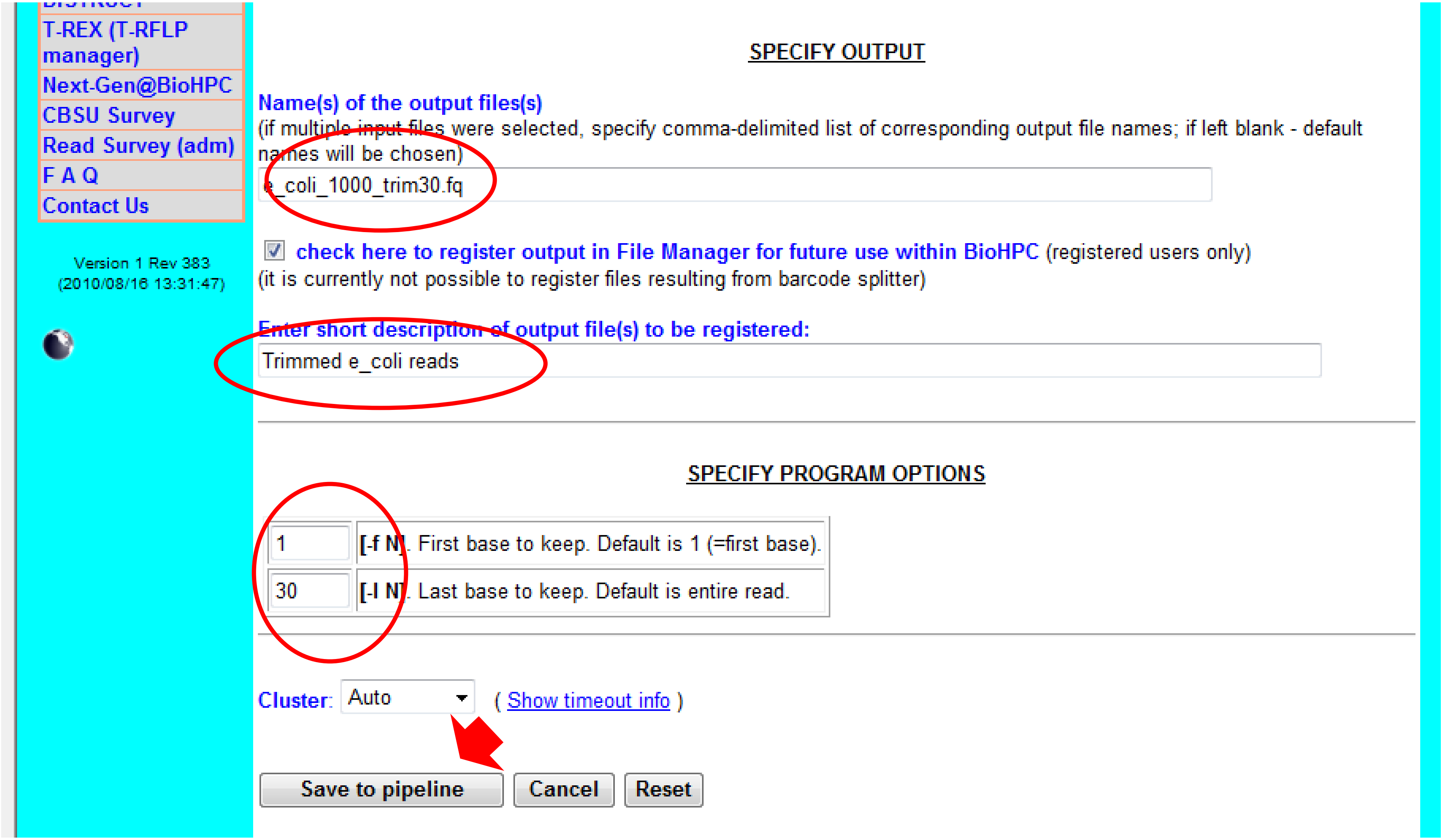
Select file(s) from the list below. Multiple files may be selected by using the left mouse button while holding down the Ctrl key.

[450][e_coli_10000snp.fq][Bowtie_Ecoli10000][lane 488]
[447][e_coli_1000.fq][Bowtie_EcoliSample][lane 486]
[22][12345_30RF7AAXX_s_4_sequence.txt.gz][actual project from datarig][lane 10]
RB test categ 3

Filter files by

Category: All File Name: File Description:

On the **Trimmer** subprogram page, the file dropdown menu will show all your FASTA and FASTQ files registered in File Manager (if needed, you can customize the file list using the filter controls under the dropdown). Scroll to the desired file to be trimmed (here: **e_coli_1000.fq**) and select it with left mouse click.



SPECIFY OUTPUT

Name(s) of the output files(s)
(if multiple input files were selected, specify comma-delimited list of corresponding output file names; if left blank - default names will be chosen)

e_coli_1000_trim30.fq

☑ check here to register output in File Manager for future use within BioHPC (registered users only)
(it is currently not possible to register files resulting from barcode splitter)

Enter short description of output file(s) to be registered:

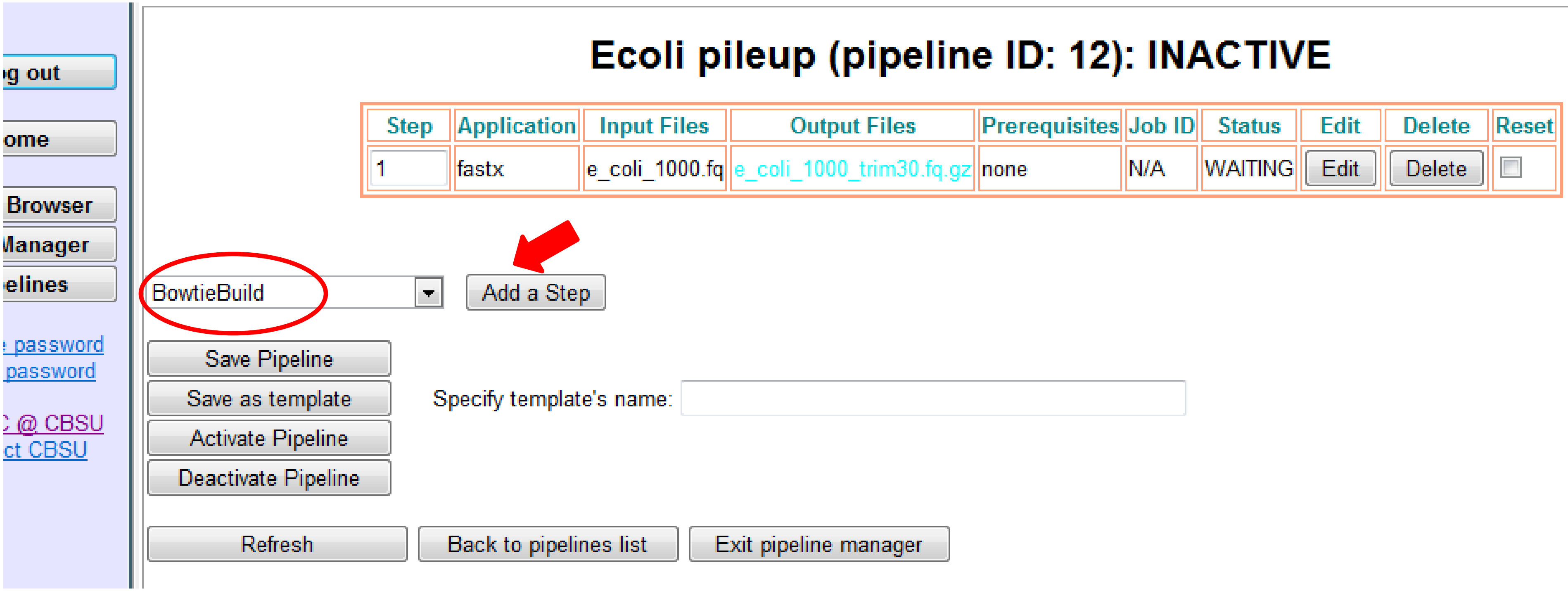
Trimmed e_coli reads

SPECIFY PROGRAM OPTIONS

1	[-f N]. First base to keep. Default is 1 (=first base).
30	[-l N]. Last base to keep. Default is entire read.

Cluster: Auto (Show timeout info)

Complete **Trimmer** submission form by specifying output options (in particular: output file name), then click **Save to pipeline**. NOTE: this will just configure the selected options in the pipeline, but will NOT yet submit an actual FASTX job. You will be redirected back to the list of steps.



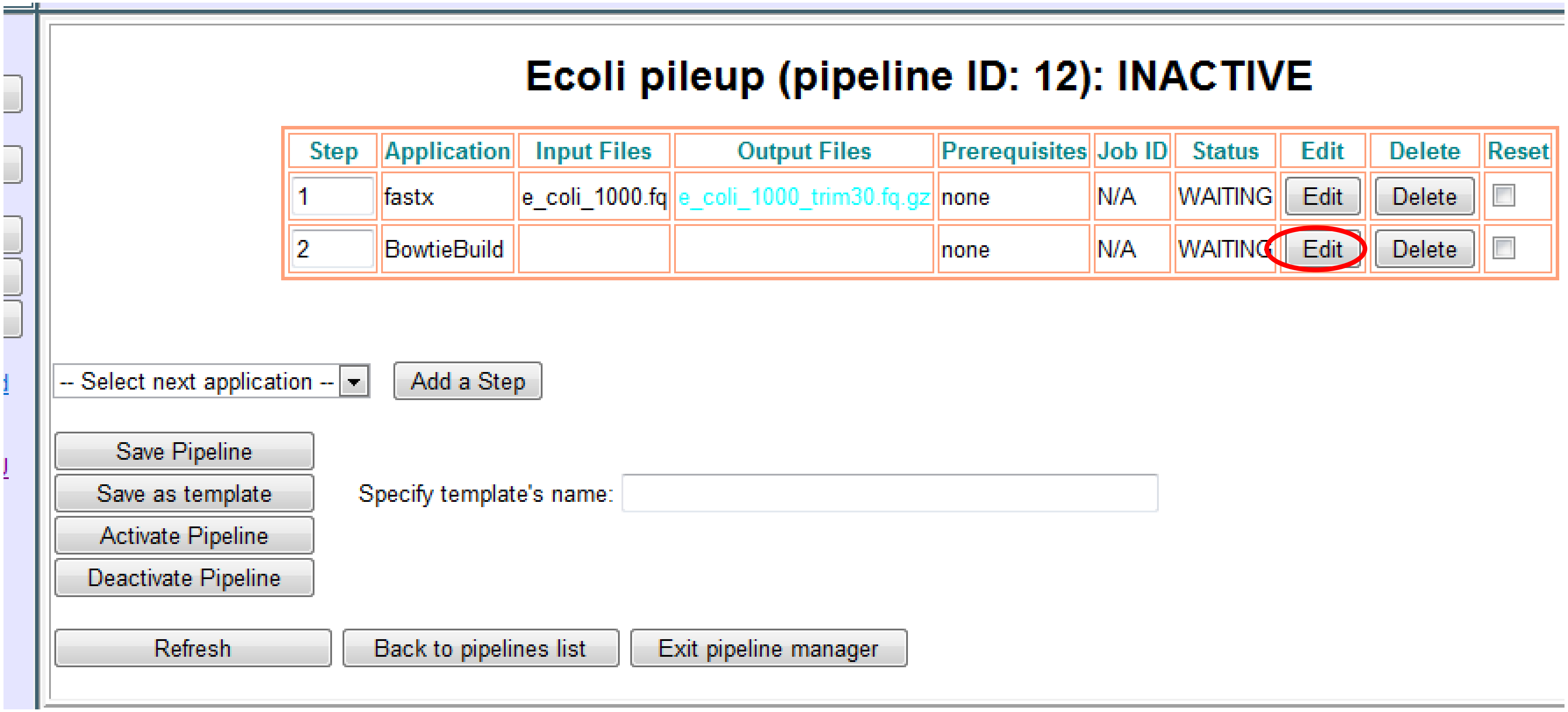
Ecoli pileup (pipeline ID: 12): INACTIVE

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	N/A	WAITING	Edit	Delete	□

Specify template's name:

The input and output files for the first step are now displayed in the table. The files already in File Manager are displayed in **black**. The output files to be produced by the programs in the pipeline are displayed in other colors (each step in different color). The **WAITING** status of the step and the still unavailable **job ID** indicate that the step has not yet been submitted to the cluster. There are no **prerequisites** to this step, so, once the pipeline is activated, it can be submitted without waiting for completion of any other steps.

To add the next step to the pipeline, select BowtieBuild from application dropdown and click **Add Step**.

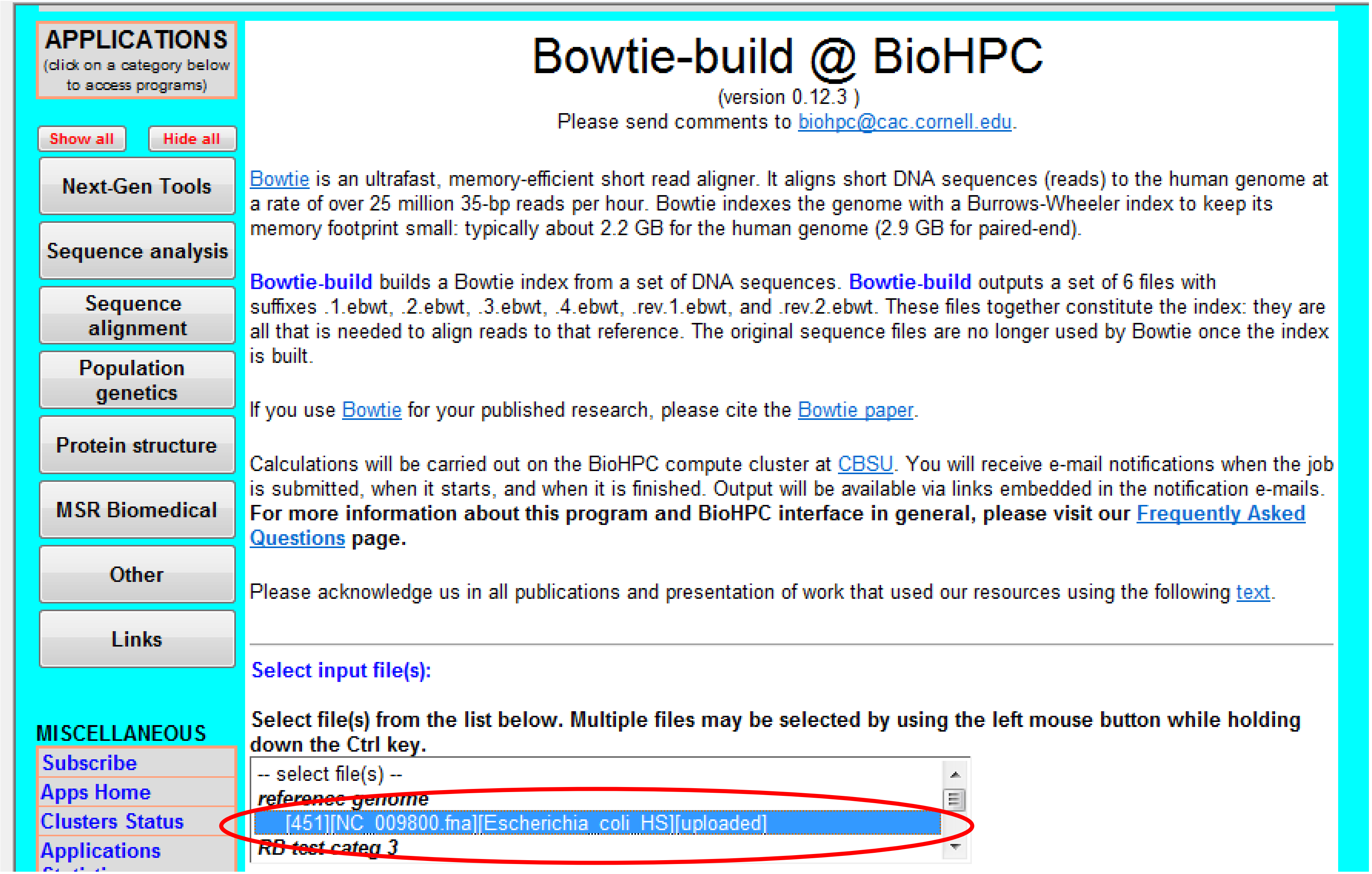


Ecoli pileup (pipeline ID: 12): INACTIVE

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	N/A	WAITING	Edit	Delete	□
2	BowtieBuild			none	N/A	WAITING	Edit	Delete	□

An empty BowtieBuild entry will be displayed. To configure BowtieBuild, click on **Edit** to be redirected to the BowtieBuild submission page.

We want to create and index from an E-coli genome file, NC_009800.fna (already registered in File Manager).



Bowtie-build @ BioHPC

(version 0.12.3)
Please send comments to biohpc@cac.cornell.edu.

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 builds a Bowtie index from a set of DNA sequences. **Bowtie-build** outputs a set of 6 files with suffixes .1.ebwt, .2.ebwt, .3.ebwt, .4.ebwt, .rev.1.ebwt, and .rev.2.ebwt. These files together constitute the index: they are all that is needed to align reads to that reference. The original sequence files are no longer used by Bowtie once the index is built.

If you use Bowtie for your published research, please cite the Bowtie paper.

Calculations will be carried out on the BioHPC compute cluster at CBSU. 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 Frequently Asked Questions page.**

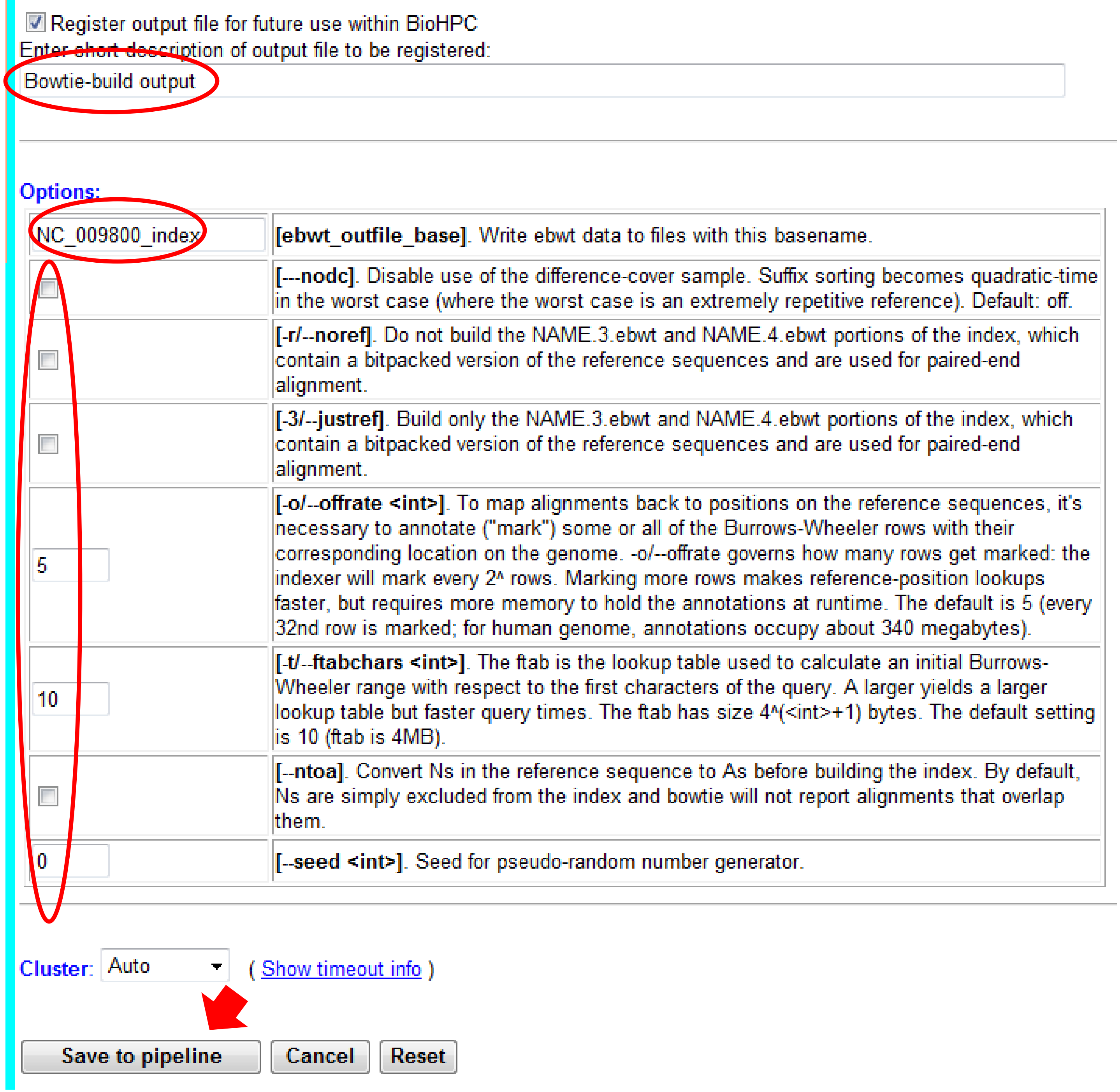
Please acknowledge us in all publications and presentation of work that used our resources using the following text.

Select input file(s):

Select file(s) from the list below. Multiple files may be selected by using the left mouse button while holding down the Ctrl key.

-- select file(s) --
reference genome
[451][NC_009800.fna][Escherichia_coli_HS][uploaded]
RB test categ 3

On **BowtieBuild** page, select the genome file to be indexed. The dropdown will show all your FASTA files you have in File Manager. If needed, you can select multiple files.



☑ Register output file for future use within BioHPC

Enter short description of output file to be registered:

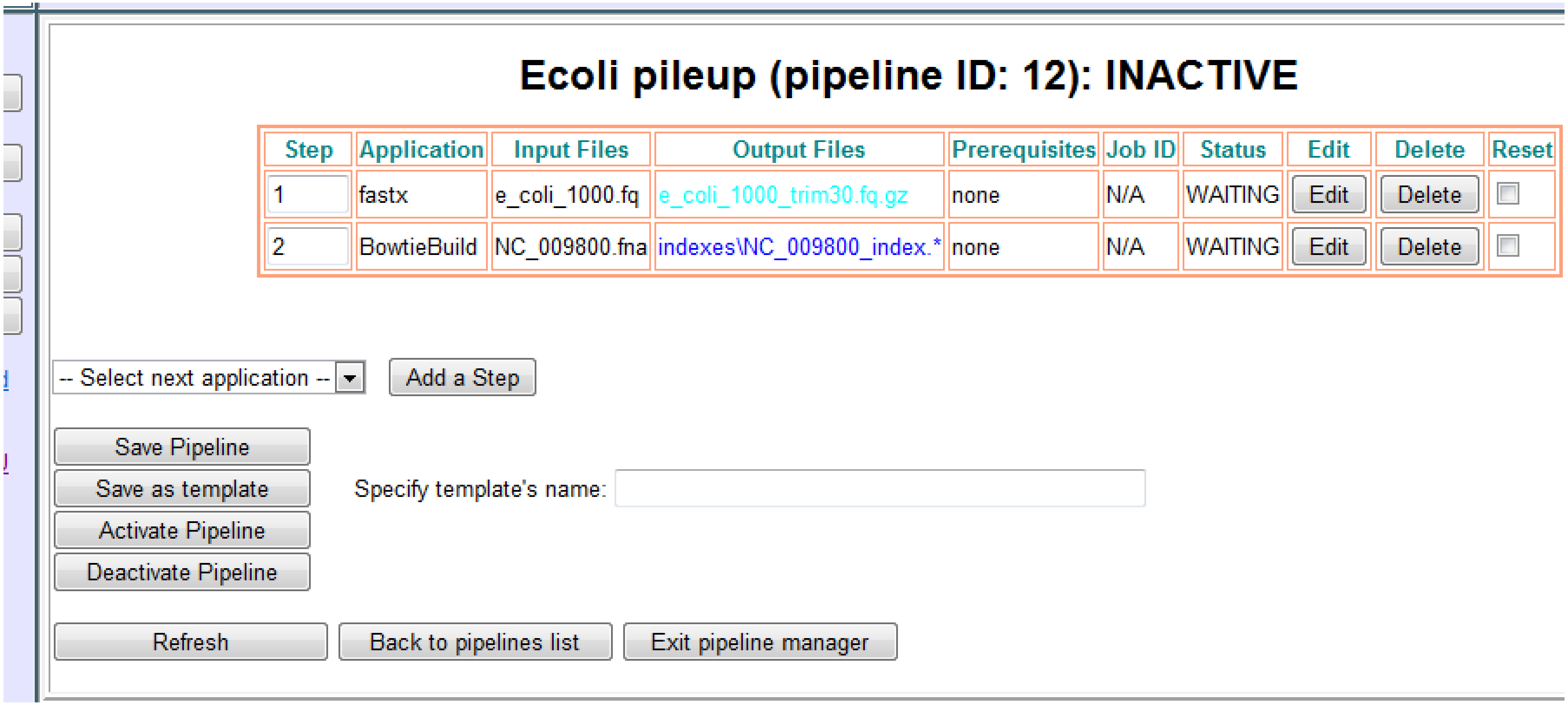
Bowtie-build output

Options:

NC_009800_index	**[ebwt_outfile_base]**. Write ebwt data to files with this basename.
□	**[---nodc]**. Disable use of the difference-cover sample. Suffix sorting becomes quadratic-time in the worst case (where the worst case is an extremely repetitive reference). Default: off.
□	**[-r/--noref]**. Do not build the NAME.3.ebwt and NAME.4.ebwt portions of the index, which contain a bitpacked version of the reference sequences and are used for paired-end alignment.
□	**[-3/--justref]**. Build only the NAME.3.ebwt and NAME.4.ebwt portions of the index, which contain a bitpacked version of the reference sequences and are used for paired-end alignment.
5	**[-o/--offrate <int>]**. To map alignments back to positions on the reference sequences, it's necessary to annotate ("mark") some or all of the Burrows-Wheeler rows with their corresponding location on the genome. -o/--offrate governs how many rows get marked: the indexer will mark every 2^ rows. Marking more rows makes reference-position lookups faster, but requires more memory to hold the annotations at runtime. The default is 5 (every 32nd row is marked; for human genome, annotations occupy about 340 megabytes).
10	**[-t/--ftabchars <int>]**. The ftab is the lookup table used to calculate an initial Burrows-Wheeler range with respect to the first characters of the query. A larger yields a larger lookup table but faster query times. The ftab has size 4^(<int>+1) bytes. The default setting is 10 (ftab is 4MB).
□	**[--ntoa]**. Convert Ns in the reference sequence to As before building the index. By default, Ns are simply excluded from the index and bowtie will not report alignments that overlap them.
0	**[--seed <int>]**. Seed for pseudo-random number generator.

Cluster: Auto ▾ (Show timeout info)

Complete **BowtieBuild** submission page by specifying program options, output parameters (in particular: base name for index files to be created), and then click **Save to pipeline**.

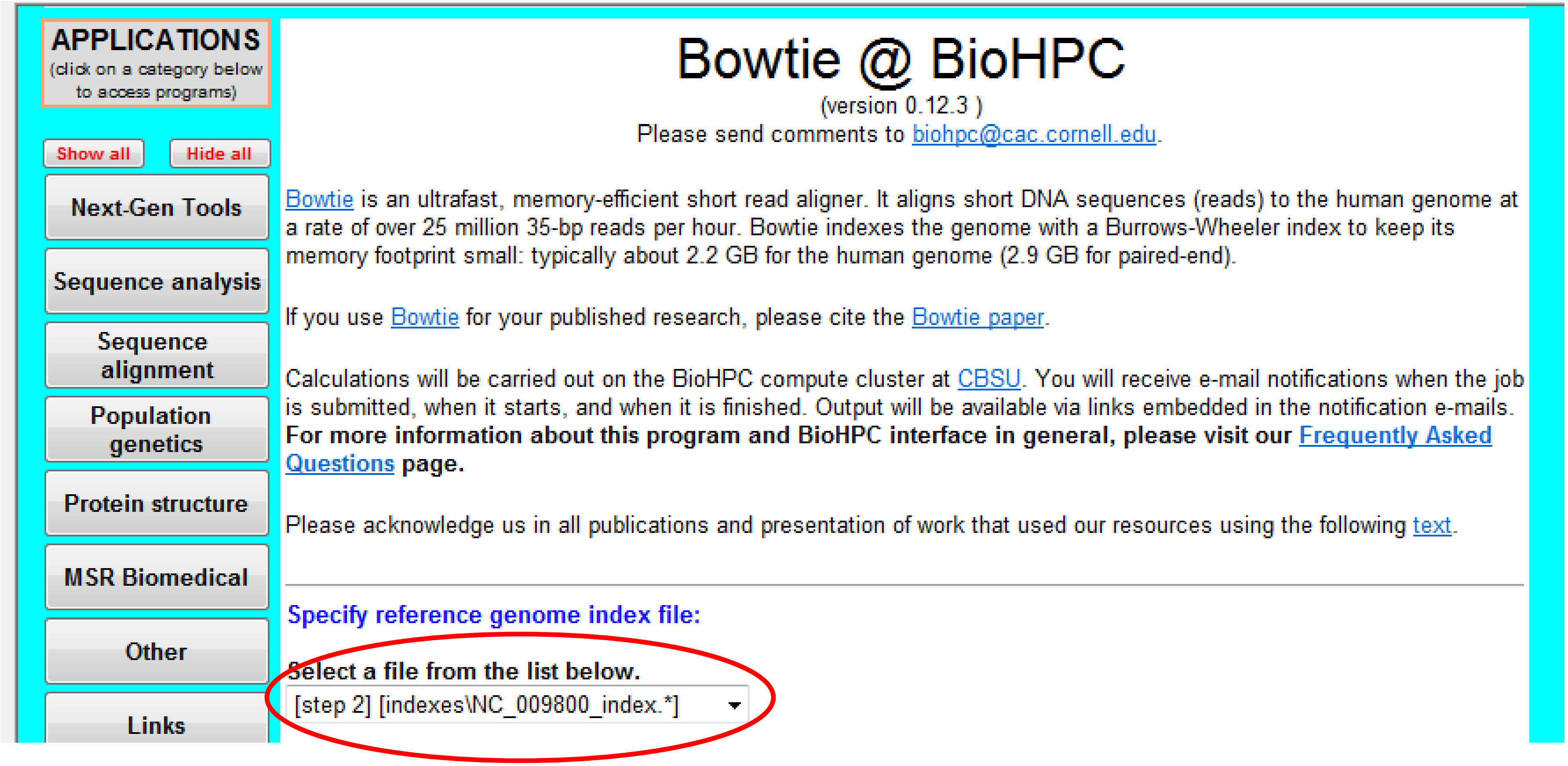


Ecoli pileup (pipeline ID: 12): INACTIVE

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	N/A	WAITING	Edit	Delete	☐
2	BowtieBuild	NC_009800.fna	indexes\NC_009800_index.*	none	N/A	WAITING	Edit	Delete	☐

The input and output files for the BowtieBuild step are now displayed in the table. The anticipated output file is color-coded. Similarly as for step 1, step 2 has no prerequisites. Therefore, once the pipeline is activated, BioHPC **may** decide to execute both steps 1 and 2 **simultaneously**, depending on resource availability.

The next step will run Bowtie to align the trimmed reads generated in step 1 to the genome indexed in step 2. As with previous steps, we select Bowtie from the dropdown, click **Add a Step**, and then click the new step's **Edit** button to access the Bowtie submission page.



Bowtie @ BioHPC

(version 0.12.3)

Please send comments to biohpc@cac.cornell.edu.

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).

If you use Bowtie for your published research, please cite the Bowtie paper.

Calculations will be carried out on the BioHPC compute cluster at CBSU. 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 Frequently Asked Questions page.**

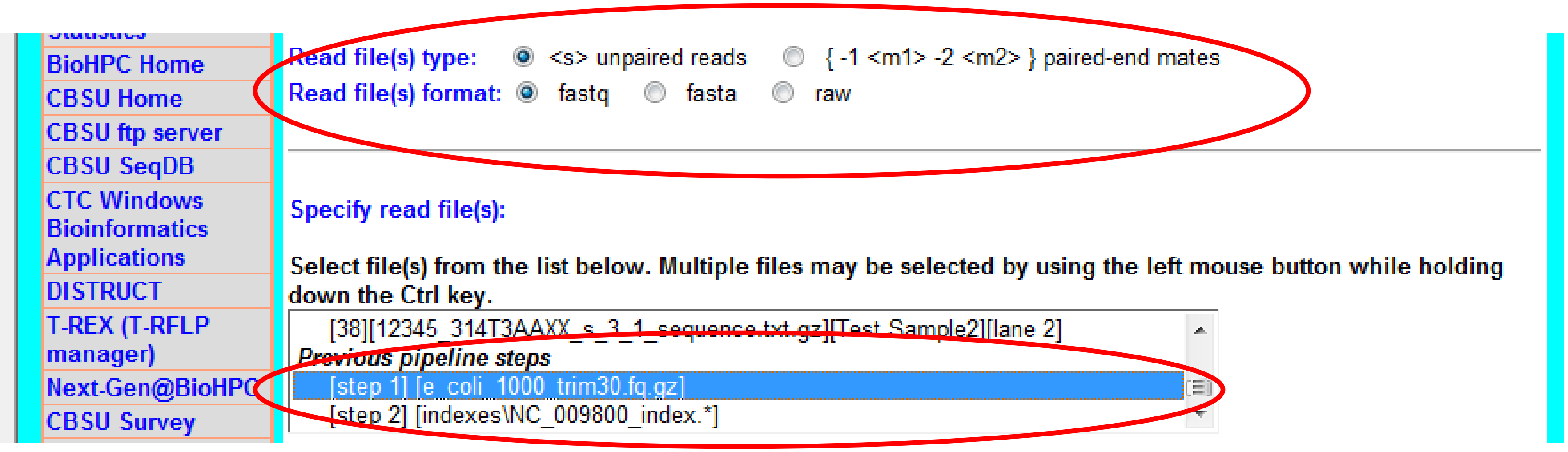
Please acknowledge us in all publications and presentation of work that used our resources using the following text.

Specify reference genome index file:

Select a file from the list below.

[step 2] [indexes\NC_009800_index.*]

Specify the reference genome index files to be used by Bowtie. Here we use the index obtained in step 2 of our pipeline. **Output files anticipated from previous steps of the pipeline can be found on the bottom of the file selection dropdown**. If you do not see them, scroll down the dropdown.



Read file(s) type: ● <s> unpaired reads ○ { -1 <m1> -2 <m2> } paired-end mates

Read file(s) format: ● fastq ○ fasta ○ raw

Specify read file(s):

Select file(s) from the list below. Multiple files may be selected by using the left mouse button while holding down the Ctrl key.

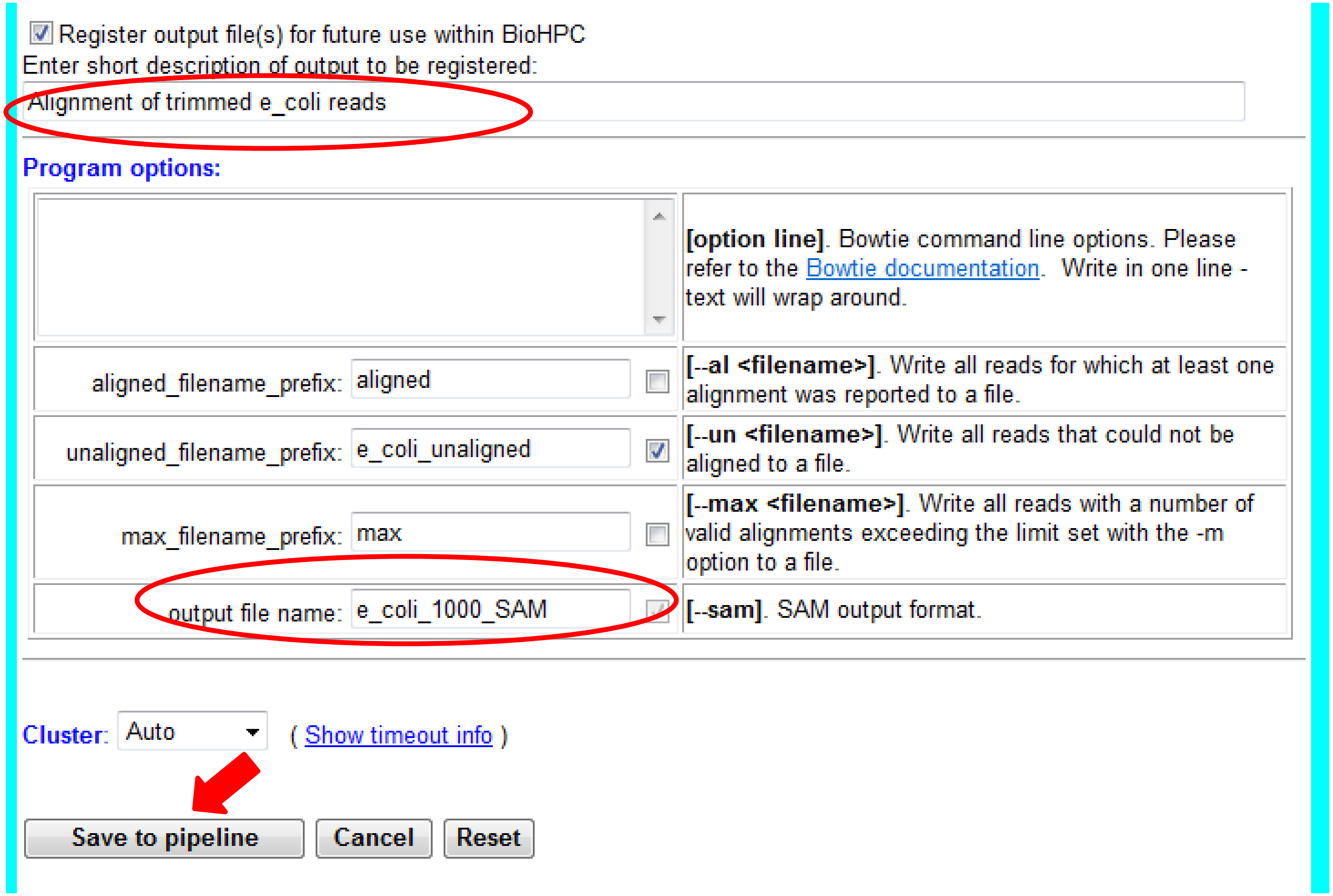
[38][12345_314T3AAXX_s_3_1_sequence.txt.gz][Test Sample2][lane 2]

Previous pipeline steps

[step 1] [e_coli_1000_trim30.fq.gz]

[step 2] [indexes\NC_009800_index.*]

Specify the type of the read file(s) (paired end or non-paired end), the format of the read file(s), and the read files themselves. Here, we want to use the trimmed FASTQ read file obtained in step 1 of the pipeline (FASTX Trimmer). Again, the output files anticipated from previous pipeline steps are found at the bottom of the file selection dropdown list.



☑ Register output file(s) for future use within BioHPC

Enter short description of output to be registered:

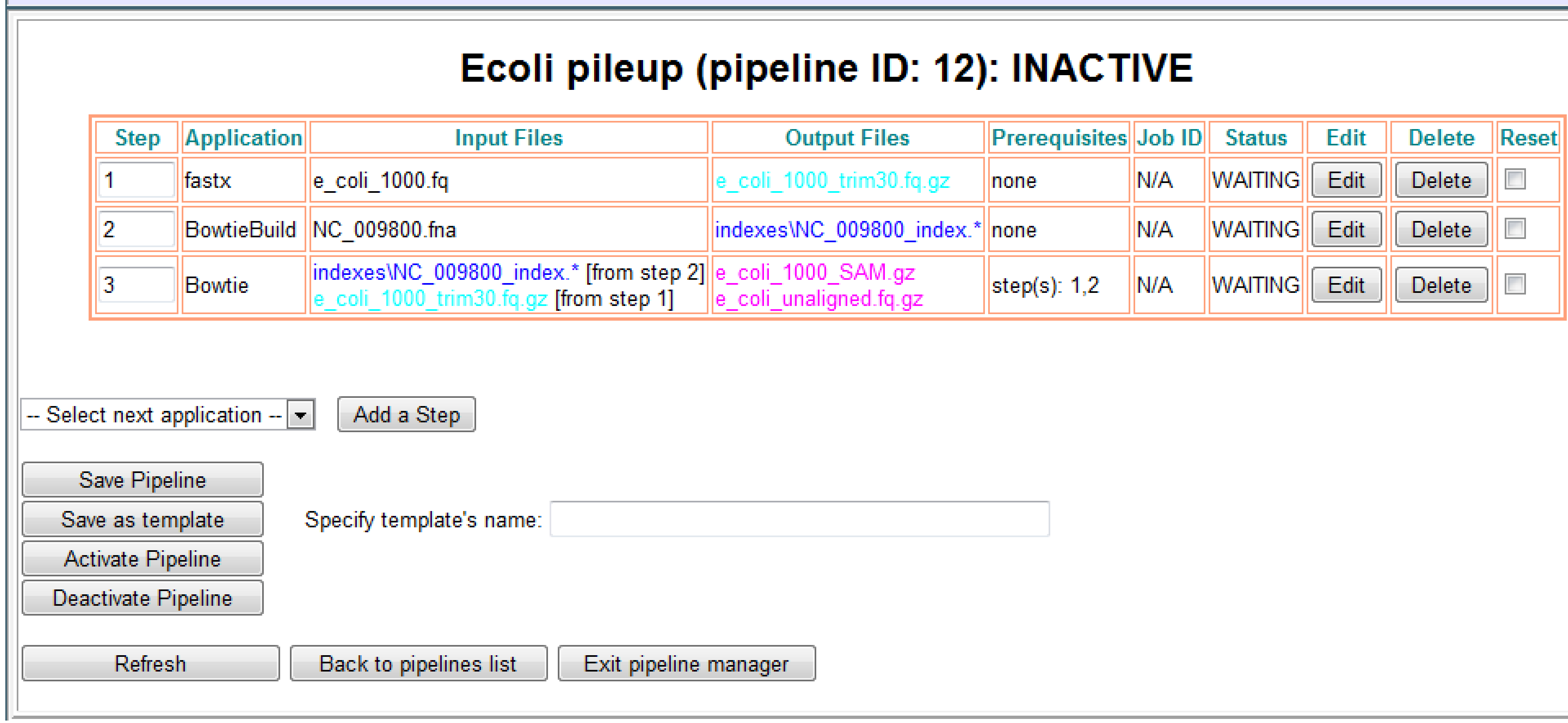
Alignment of trimmed e_coli reads

Program options:

			[option line]. Bowtie command line options. Please refer to the Bowtie documentation. Write in one line - text will wrap around.
aligned_filename_prefix:	aligned	□	**[--al <filename>]**. Write all reads for which at least one alignment was reported to a file.
unaligned_filename_prefix:	e_coli_unaligned	☑	**[--un <filename>]**. Write all reads that could not be aligned to a file.
max_filename_prefix:	max	□	**[--max <filename>]**. Write all reads with a number of valid alignments exceeding the limit set with the -m option to a file.
output file name:	e_coli_1000_SAM	☑	**[--sam]**. SAM output format.

Cluster: Auto (Show timeout info)

Specify other Bowtie options and any associated file names. Output file name (here: e_coli_1000_SAM) is required. In this example we also request the file with unaligned reads to be generated and have a file name starting with "e_coli_unaligned". When the form is complete, click **Save to pipeline**.



Ecoli pileup (pipeline ID: 12): INACTIVE

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	N/A	WAITING	Edit	Delete	☐
2	BowtieBuild	NC_009800.fna	indexes\NC_009800_index.*	none	N/A	WAITING	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	N/A	WAITING	Edit	Delete	☐

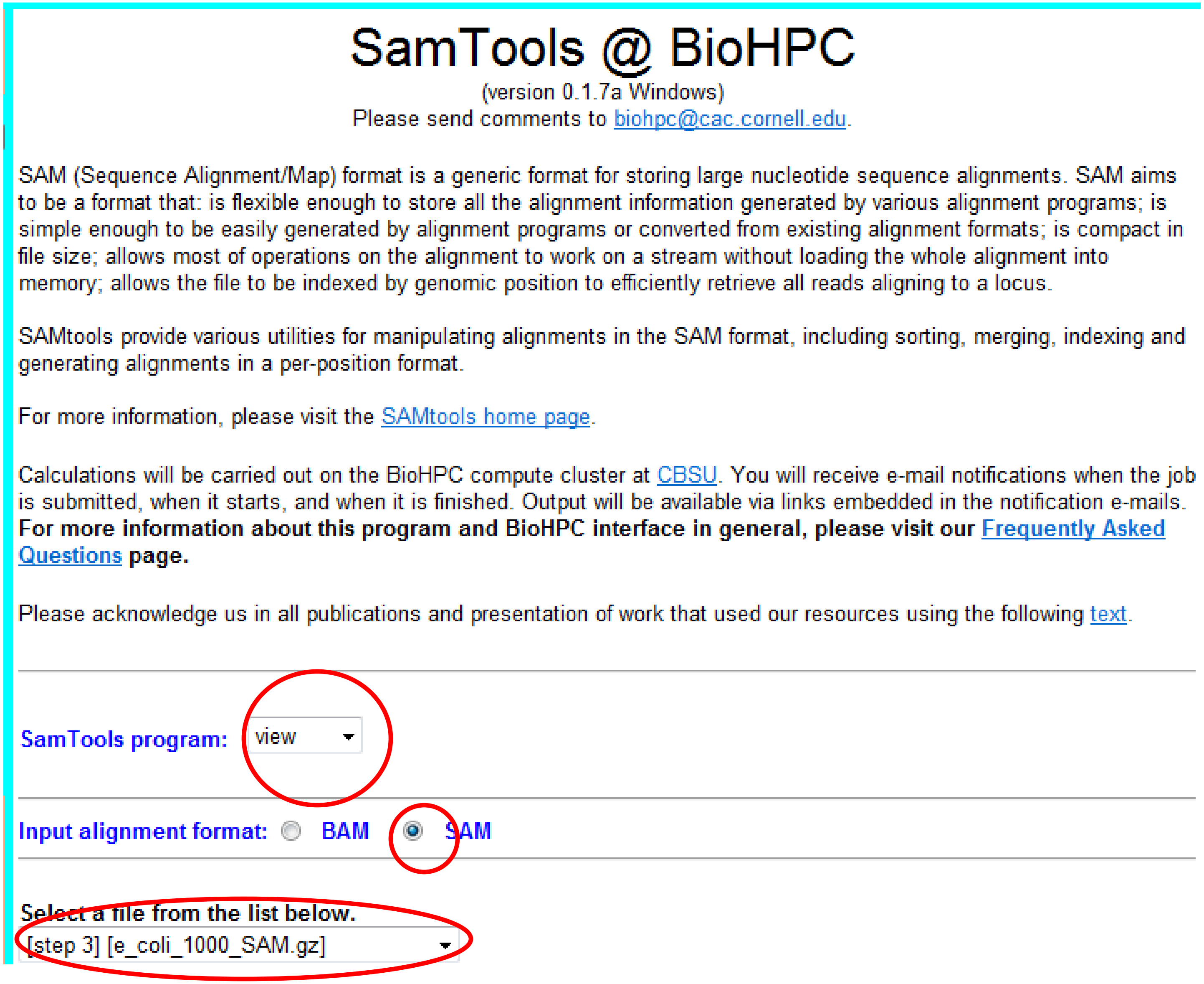
Specify template's name:

The resulting table shows that the input files to Bowtie are not yet in File Manager – they are to be generated by previous steps of the pipeline. Color coding helps to quickly assess which previous steps (here: 1 and 2) are involved. This dependence of step 3 on not yet existing files is reflected in the **Prerequisites** column. Once the pipeline is active, BioHPC will make sure that step 3 is not submitted until both steps 1 and 2 are completed.

We will now add two more SamTools steps:

The first SamTools step (step 4 of the pipeline) will produce a BAM file from the SAM file created in step 3. To do this, we will configure SamTools to run with the “view –b –S” options.

The second SamTools step (step 5 of the pipeline) will produce a “pileup” version of the alignment. We will configure SamTools to run with the pileup option, taking the BAM file created in step 4 as input.



SamTools @ BioHPC

(version 0.1.7a Windows)
Please send comments to biohpc@cac.cornell.edu.

SAM (Sequence Alignment/Map) format is a generic format for storing large nucleotide sequence alignments. SAM aims to be a format that: is flexible enough to store all the alignment information generated by various alignment programs; is simple enough to be easily generated by alignment programs or converted from existing alignment formats; is compact in file size; allows most of operations on the alignment to work on a stream without loading the whole alignment into memory; allows the file to be indexed by genomic position to efficiently retrieve all reads aligning to a locus.

SAMtools provide various utilities for manipulating alignments in the SAM format, including sorting, merging, indexing and generating alignments in a per-position format.

For more information, please visit the SAMtools home page.

Calculations will be carried out on the BioHPC compute cluster at CBSU. 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 Frequently Asked Questions page.**

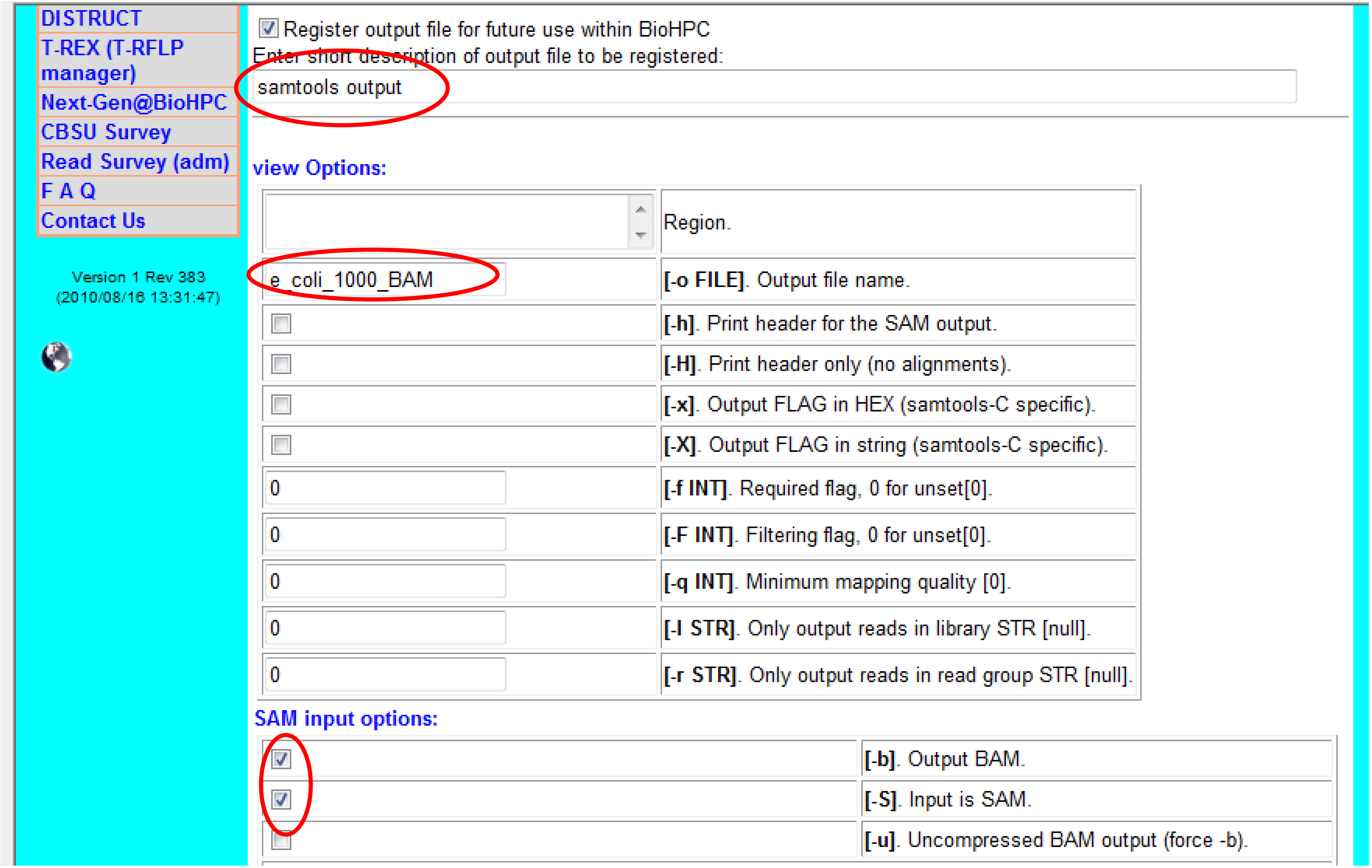
Please acknowledge us in all publications and presentation of work that used our resources using the following text.

SamTools program: view

Input alignment format: ○ **BAM** ● **SAM**

Select a file from the list below.
[step 3] [e_coli_1000_SAM.gz]

In the first SamTools run, we will create a BAM version of the alignment file obtained (in SAM format) in step 3 of the pipeline (Bowtie). We specify the SamTools subprogram as **view**, input format as **SAM**, and select the SAM file obtained in step 3 from file selection dropdown.



☑ Register output file for future use within BioHPC

Enter short description of output file to be registered:

samtools output

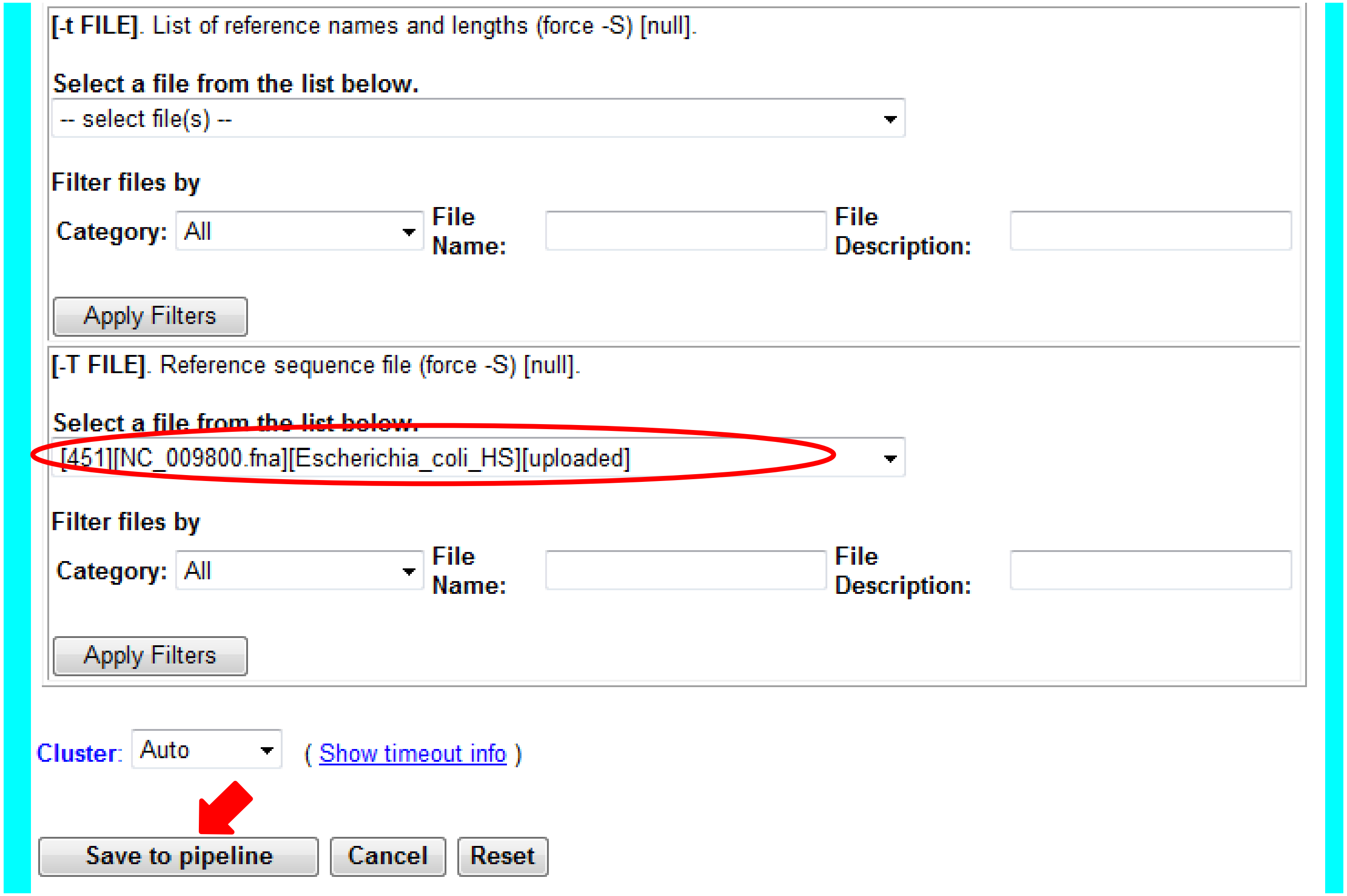
view Options:

	Region.
e_coli_1000_BAM	**[-o FILE]**. Output file name.
□	**[-h]**. Print header for the SAM output.
□	**[-H]**. Print header only (no alignments).
□	**[-x]**. Output FLAG in HEX (samtools-C specific).
□	**[-X]**. Output FLAG in string (samtools-C specific).
0	**[-f INT]**. Required flag, 0 for unset[0].
0	**[-F INT]**. Filtering flag, 0 for unset[0].
0	**[-q INT]**. Minimum mapping quality [0].
0	**[-l STR]**. Only output reads in library STR [null].
0	**[-r STR]**. Only output reads in read group STR [null].

SAM input options:

☑	**[-b]**. Output BAM.
☑	**[-S]**. Input is SAM.
□	**[-u]**. Uncompressed BAM output (force -b).

Specify the name and description of the output file, and other SamTools **view** options. Note that “–b” and “–S” must be checked if a BAM file is to be created.



[-t FILE]. List of reference names and lengths (force -S) [null].

Select a file from the list below.

-- select file(s) --

Filter files by

Category: All File Name: File Description:

[-T FILE]. Reference sequence file (force -S) [null].

Select a file from the list below.

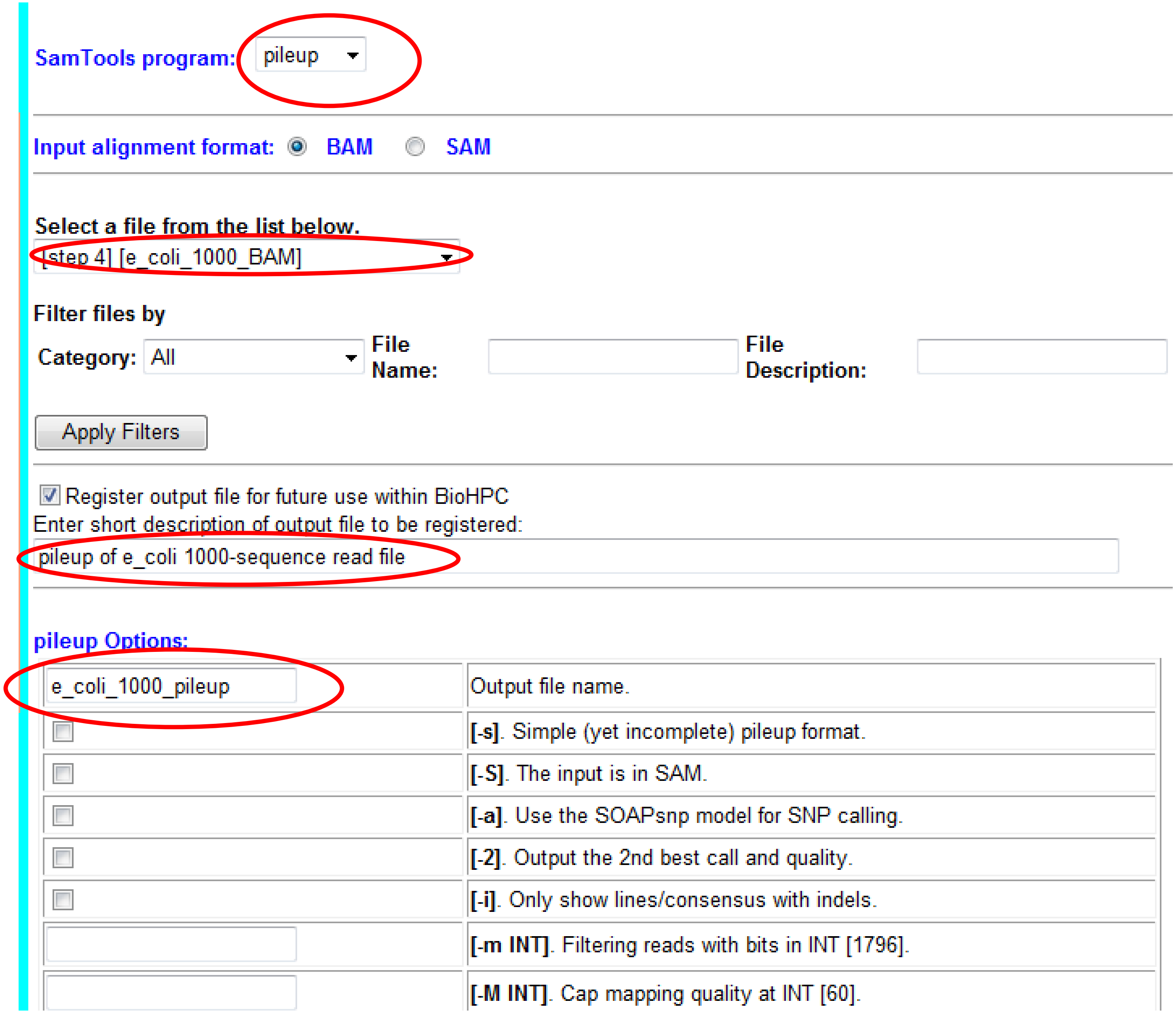
[451][NC_009800.fna][Escherichia_coli_HS][uploaded]

Filter files by

Category: All File Name: File Description:

Cluster: Auto (Show timeout info)

Some SamTools **view** options include file specifications. Here, we specify the reference genome file NC_009800.fna (which we also used in Step 2 to created index for Bowtie). When the form is completed, click **Save to pipeline**.



SamTools program: pileup

Input alignment format: ● BAM ○ SAM

Select a file from the list below.

[step 4] [e_coli_1000_BAM]

Filter files by

Category: All | File Name: | File Description:

☑ Register output file for future use within BioHPC

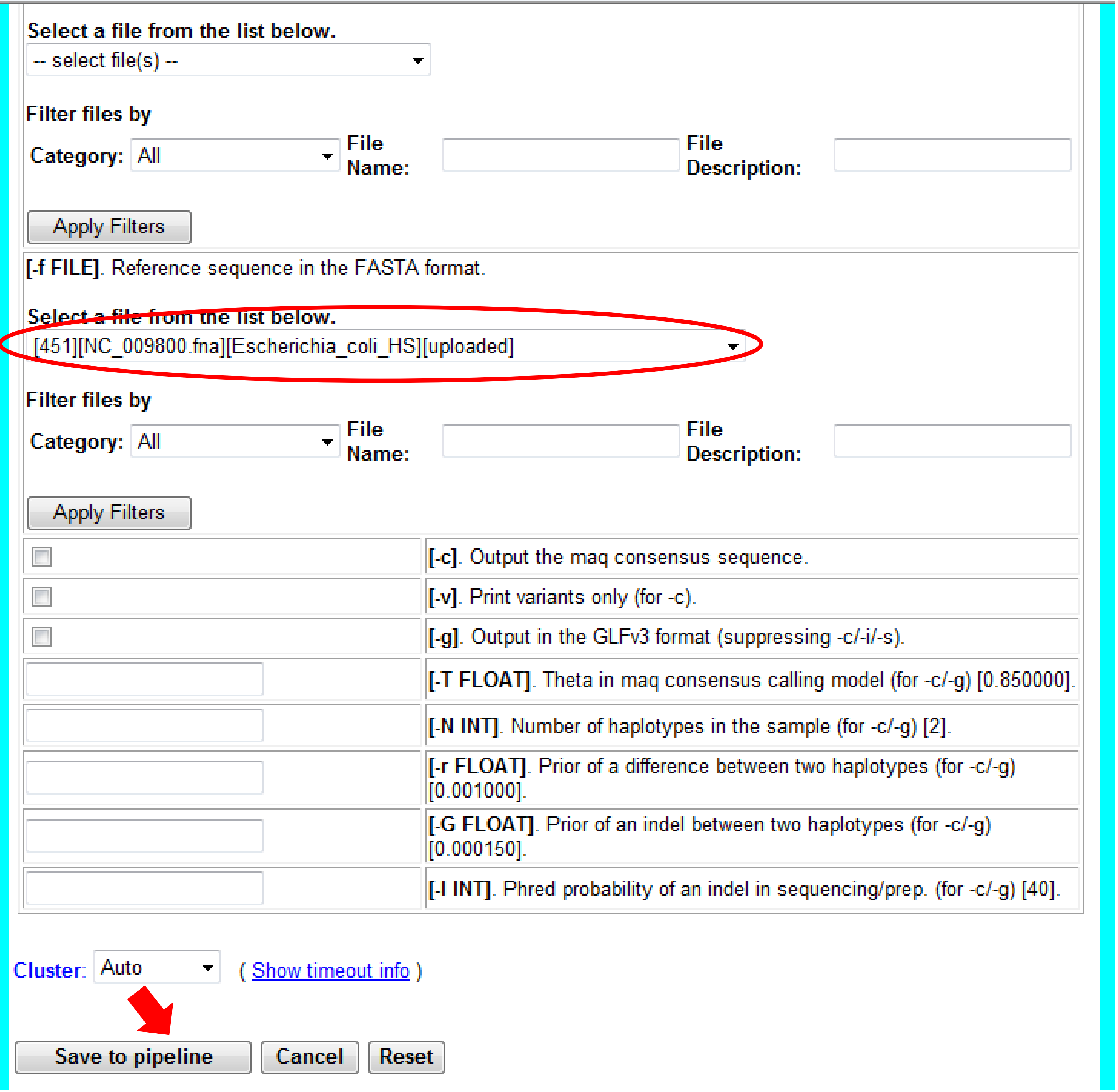
Enter short description of output file to be registered:

pileup of e_coli 1000-sequence read file

pileup Options:

e_coli_1000_pileup	Output file name.
□	[-s]. Simple (yet incomplete) pileup format.
□	[-S]. The input is in SAM.
□	[-a]. Use the SOAPsnp model for SNP calling.
□	[-2]. Output the 2nd best call and quality.
□	[-i]. Only show lines/consensus with indels.
	[-m INT]. Filtering reads with bits in INT [1796].
	[-M INT]. Cap mapping quality at INT [60].

We select the **pileup** subprogram, and the BAM file from step 4 of the pipeline. We set some program options, including the name we want for the resulting pileup file.



Select a file from the list below.

-- select file(s) --

Filter files by

Category:	All	**File Name:**		**File Description:**	

Apply Filters

[-f FILE]. Reference sequence in the FASTA format.

Select a file from the list below.

[451][NC_009800.fna][Escherichia_coli_HS][uploaded]

Filter files by

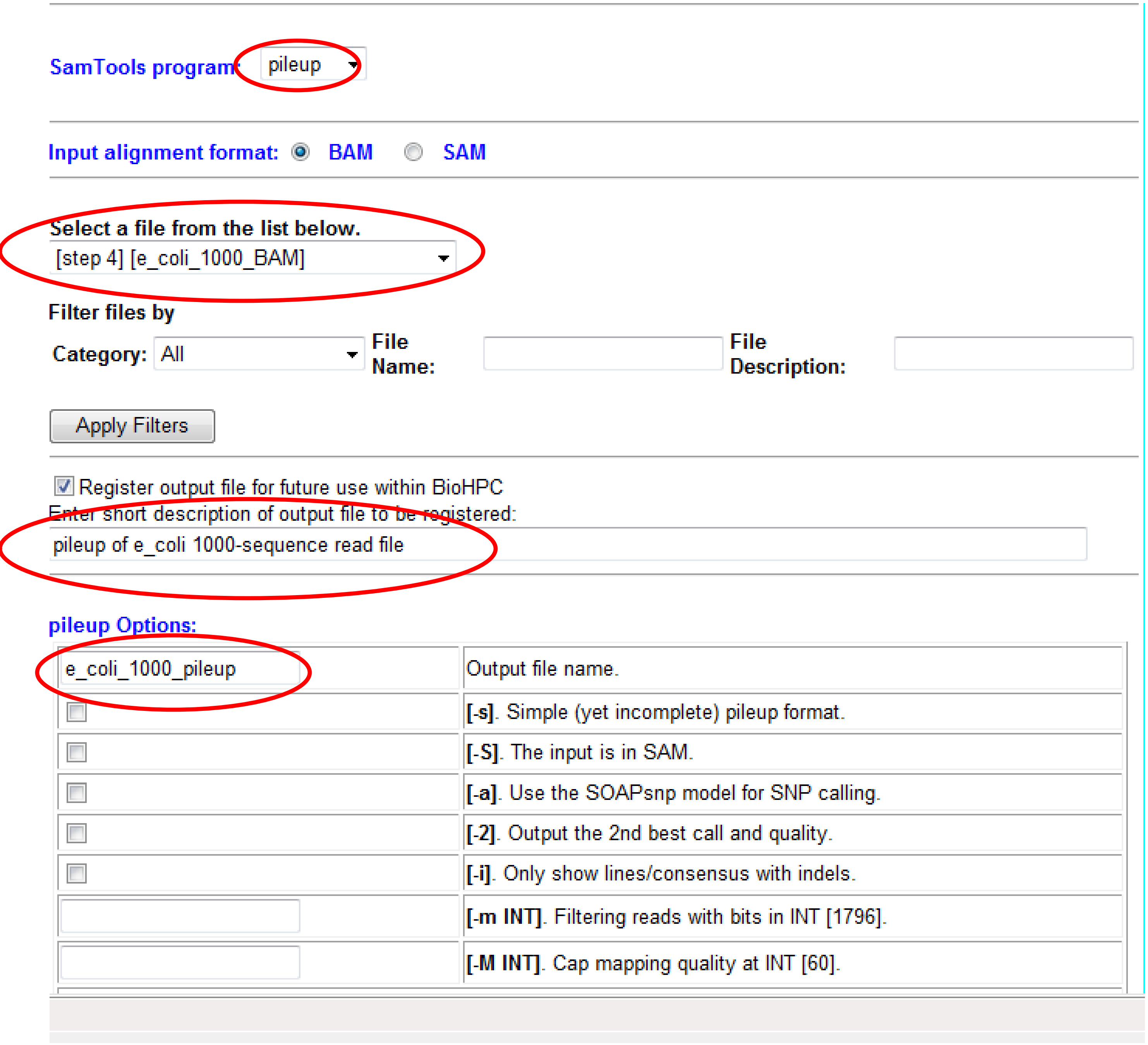
Category:	All	**File Name:**		**File Description:**	

Apply Filters

☐	**[-c]**. Output the maq consensus sequence.
☐	**[-v]**. Print variants only (for -c).
☐	**[-g]**. Output in the GLFv3 format (suppressing -c/-i/-s).
	[-T FLOAT]. Theta in maq consensus calling model (for -c/-g) [0.850000].
	[-N INT]. Number of haplotypes in the sample (for -c/-g) [2].
	[-r FLOAT]. Prior of a difference between two haplotypes (for -c/-g) [0.001000].
	[-G FLOAT]. Prior of an indel between two haplotypes (for -c/-g) [0.000150].
	[-I INT]. Phred probability of an indel in sequencing/prep. (for -c/-g) [40].

Cluster: Auto (Show timeout info)

We leave other options at default values, except for "-f", for which we specify the reference genome FASTA file (the same file was used before to produce Bowtie index in step 2, and the BAM file in step 4). After the form is complete, click **Save to pipeline**.



SamTools program: pileup

Input alignment format: ● BAM ○ SAM

Select a file from the list below.
[step 4] [e_coli_1000_BAM]

Filter files by

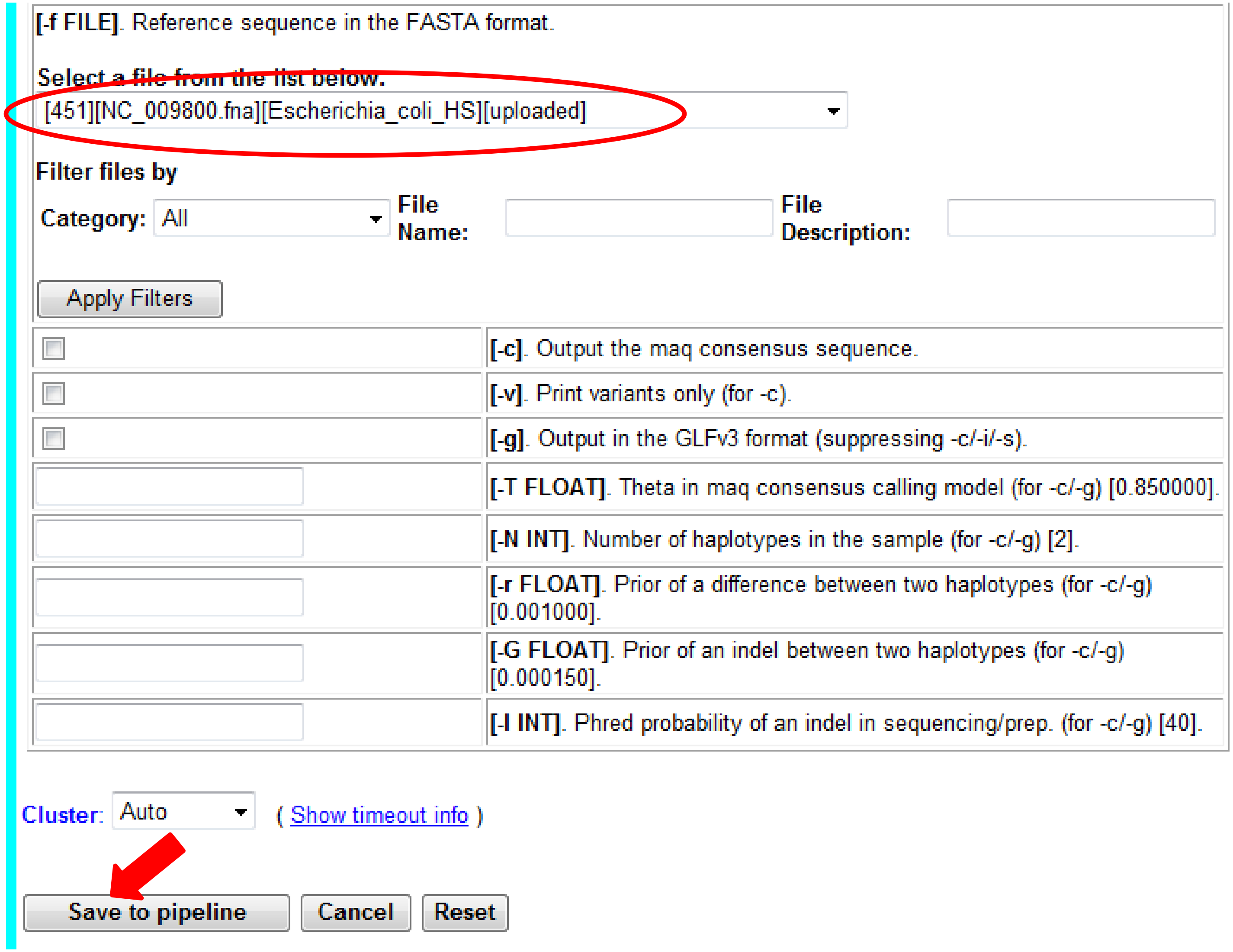
Category: All   File Name: File Description:

☑ Register output file for future use within BioHPC
Enter short description of output file to be registered:
pileup of e_coli 1000-sequence read file

pileup Options:

e_coli_1000_pileup	Output file name.
□	[-s]. Simple (yet incomplete) pileup format.
□	[-S]. The input is in SAM.
□	[-a]. Use the SOAPsnp model for SNP calling.
□	[-2]. Output the 2nd best call and quality.
□	[-i]. Only show lines/consensus with indels.
	[-m INT]. Filtering reads with bits in INT [1796].
	[-M INT]. Cap mapping quality at INT [60].

Select **pileup** subprogram, specify the BAM file of step 4 as input, provide **output file name** and other program options.



[-f FILE]. Reference sequence in the FASTA format.

Select a file from the list below.

[451][NC_009800.fna][Escherichia_coli_HS][uploaded]

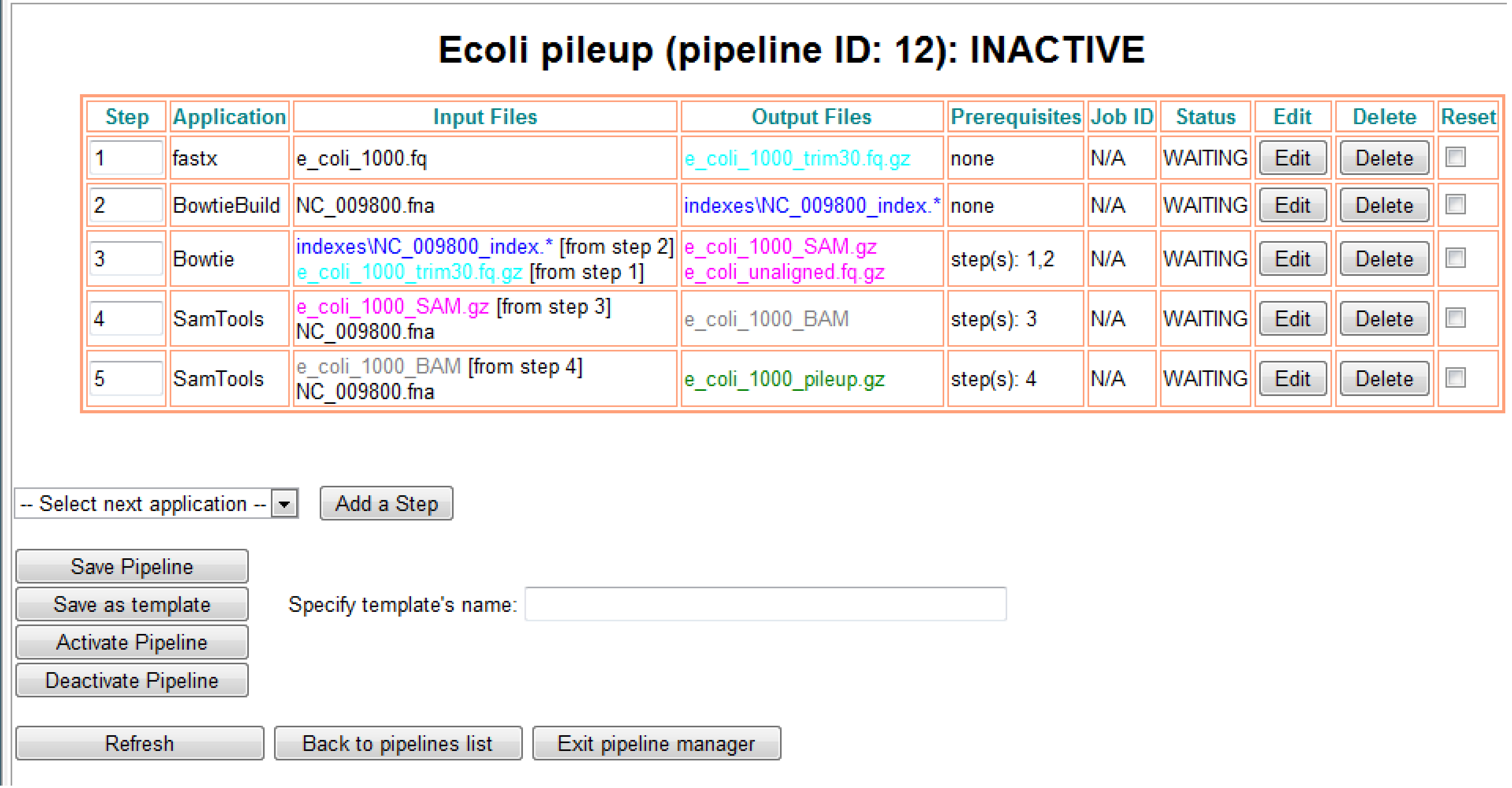
Filter files by

Category: All **File Name:** **File Description:**

□	**[-c]**. Output the maq consensus sequence.
□	**[-v]**. Print variants only (for -c).
□	**[-g]**. Output in the GLFv3 format (suppressing -c/-i/-s).
	[-T FLOAT]. Theta in maq consensus calling model (for -c/-g) [0.850000].
	[-N INT]. Number of haplotypes in the sample (for -c/-g) [2].
	[-r FLOAT]. Prior of a difference between two haplotypes (for -c/-g) [0.001000].
	[-G FLOAT]. Prior of an indel between two haplotypes (for -c/-g) [0.000150].
	[-I INT]. Phred probability of an indel in sequencing/prep. (for -c/-g) [40].

Cluster: Auto (Show timeout info)

As "-f" option, specify the reference genome FASTA file we used throughout the pipeline. If needed, specify the remaining run options, then click **Save to pipeline**.



Ecoli pileup (pipeline ID: 12): INACTIVE

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	N/A	WAITING	Edit	Delete	☐
2	BowtieBuild	NC_009800.fna	indexes\NC_009800_index.*	none	N/A	WAITING	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	N/A	WAITING	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	☐

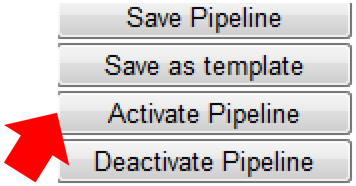
Specify template's name:

All 5 steps of the pipeline have been added and configured. A few notes:

- Step 4 is dependent on steps 3 (through the file e_coli_SAM.gz) – BioHPC will submit step 4 only after step 3 is finished. Similarly, step 5 is dependent on step 4.
- Usually, the output files listed in Output Files column are used as inputs in subsequent pipeline steps, but it does not always have to be the case. For example, file **e_coli_unaligned.fq.gz** generated in step 3 will be available upon completion of step 3 of the pipeline, but it won't be used in further pipeline steps.

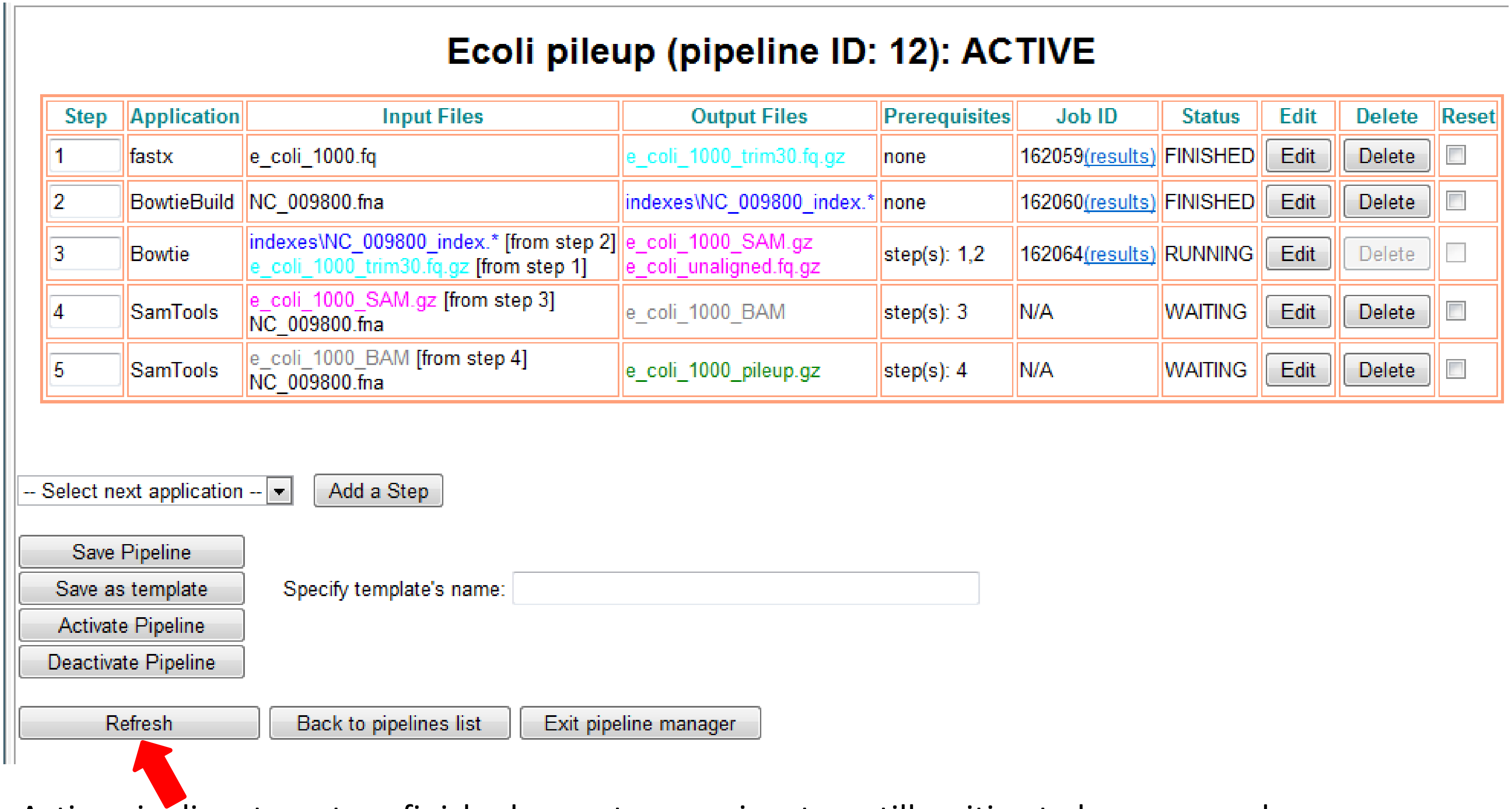
A few more notes:

- Run options for any step can be adjusted by clicking on this step's **Edit** button – this will take you back to the appropriate application page. This is also a good method of checking what options a given application has been configured with.
- Any step can be deleted using this step's **Delete** button.
    - **After deleting a step, you have to reconfigure the remaining steps' input files** (using the **Edit** buttons).
    - After deleting a step and reconfiguring applications – click **Save Pipeline**
- Upon successful completion of the pipeline, all files listed in **Output Files** column will be available in File Manager and thus also in applications' file selection menus. You can use them in other BioHPC pipelines, or you can download them to your local machine (see further slides) for use outside of BioHPC.



The pipeline is now configured.

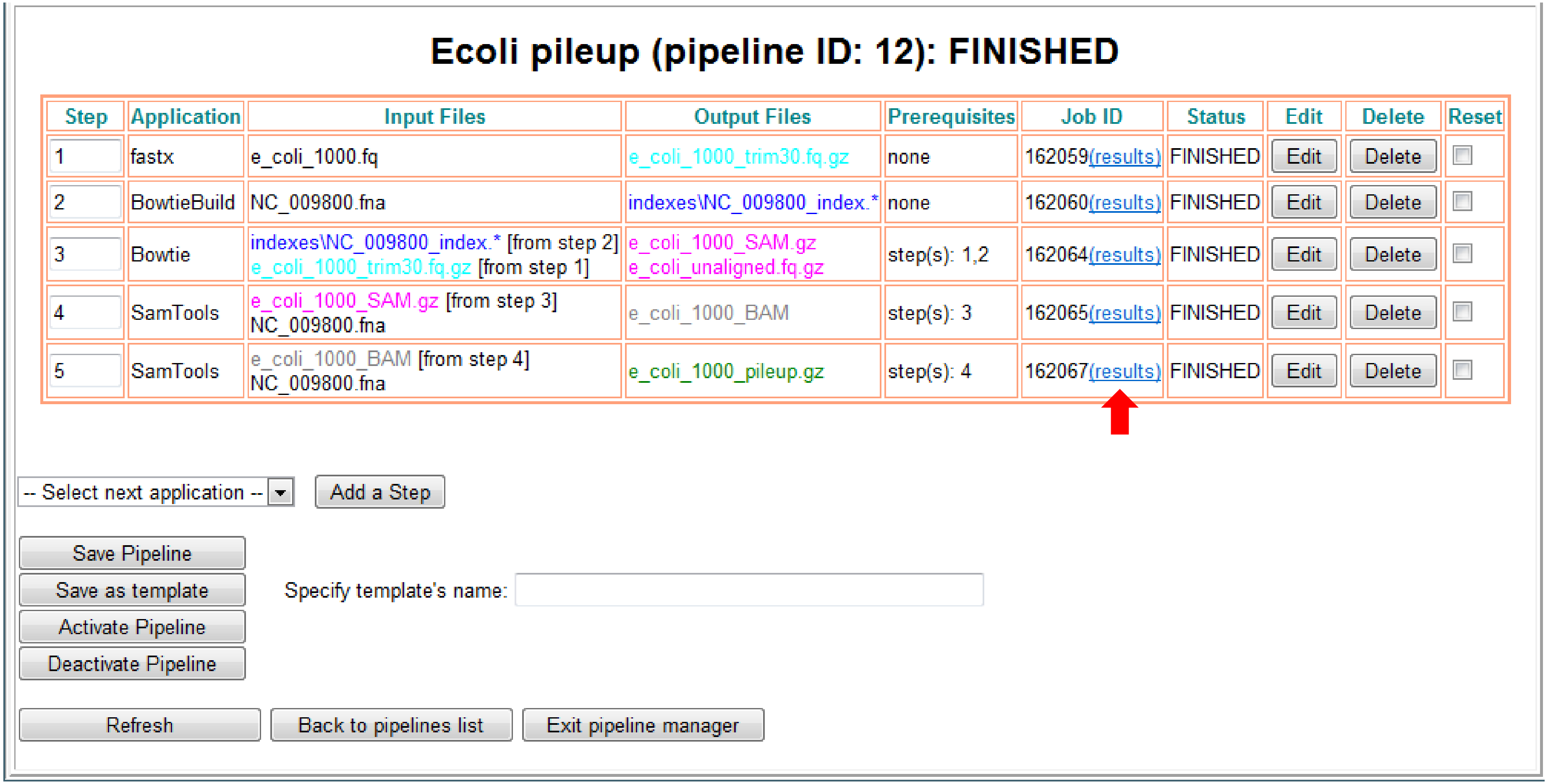
- **IMPORTANT**: Click on **Save Pipeline** to save all settings.
- Then click on **Activate Pipeline** to tell BioHPC to start the processing.



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(results)	FINISHED	Edit	Delete	☐
2	BowtieBuild	NC_009800.fna	indexes\NC_009800_index.*	none	162060(results)	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(results)	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	☐

- Active pipeline: two steps finished, one step running, two still waiting to be processed.
- Click **Refresh** for most current step status.
- Results from finished jobs and partial results from running jobs are available (click **results**).
- Some jobs may be long (hours to days). You can **Exit pipeline manager** and return to your pipeline later.
- If for any reason you want to stop the processing of the pipeline, click **Deactivate Pipeline**. Note: this will **NOT** cancel any steps which have already been submitted. Unless the corresponding jobs are canceled explicitly, they will be allowed to run to completion.

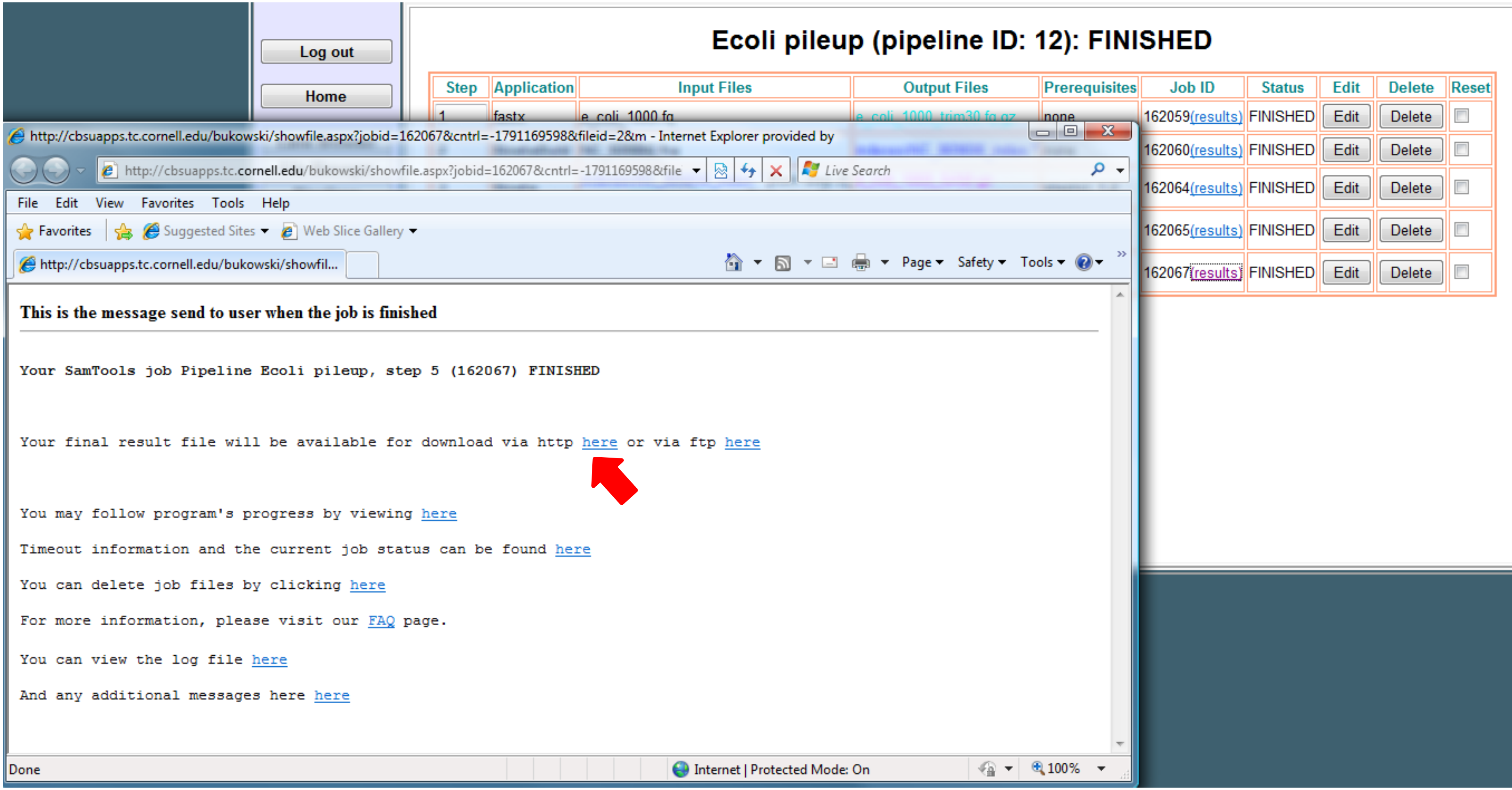


Ecoli pileup (pipeline ID: 12): FINISHED

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(results)	FINISHED	Edit	Delete	□
2	BowtieBuild	NC_009800.fna	indexes\NC_009800_index.*	none	162060(results)	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(results)	FINISHED	Edit	Delete	□
4	SamTools	e_coli_1000_SAM.gz [from step 3] NC_009800.fna	e_coli_1000_BAM	step(s): 3	162065(results)	FINISHED	Edit	Delete	□
5	SamTools	e_coli_1000_BAM [from step 4] NC_009800.fna	e_coli_1000_pileup.gz	step(s): 4	162067(results)	FINISHED	Edit	Delete	□

Specify template's name:

Pipeline completed. Click on **results** link of any job to access this job's results


Log out
Home
http://cbsuapps.tc.cornell.edu/bukowski/showfile.aspx?jobid=162067&cntrl=-1791169598&fileid=2&m - Internet Explorer provided by
http://cbsuapps.tc.cornell.edu/bukowski/showfile.aspx?jobid=162067&cntrl=-1791169598&file
Live Search
File Edit View Favorites Tools Help
Favorites
Suggested Sites
Web Slice Gallery
http://cbsuapps.tc.cornell.edu/bukowski/showfil...
Page Safety Tools
Done
Internet | Protected Mode: On
100%

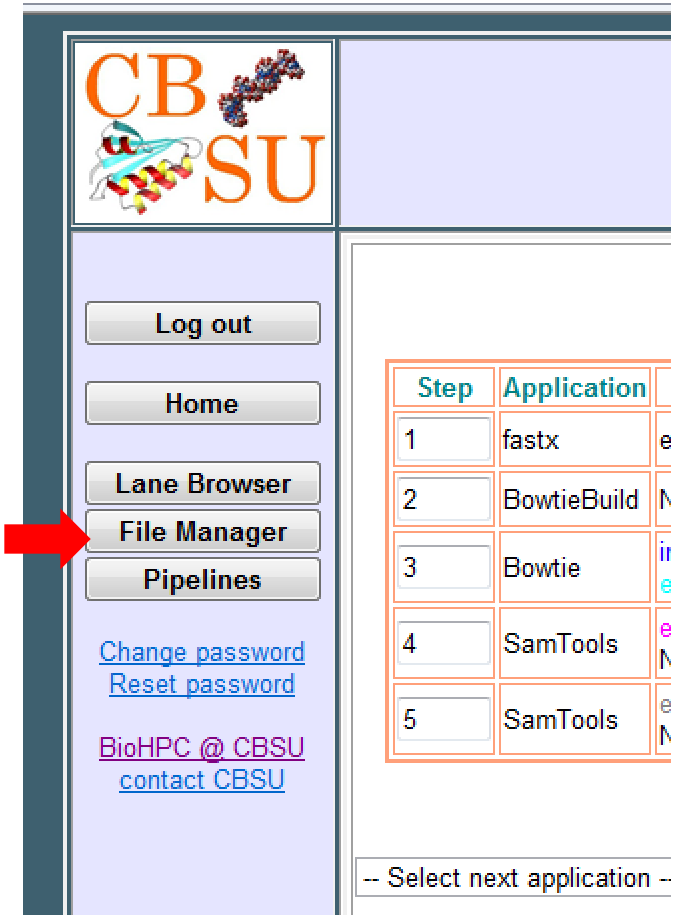
Ecoli pileup (pipeline ID: 12): FINISHED

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(results)	FINISHED	Edit	Delete	□
					162060(results)	FINISHED	Edit	Delete	□
					162064(results)	FINISHED	Edit	Delete	□
					162065(results)	FINISHED	Edit	Delete	□
					162067(results)	FINISHED	Edit	Delete	□

This is the message send to user when the job is finished

```
Your SamTools job Pipeline Ecoli pileup, step 5 (162067) FINISHED


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


You may follow program's progress by viewing here

Timeout information and the current job status can be found here

You can delete job files by clicking here

For more information, please visit our FAQ page.

You can view the log file here

And any additional messages here here
```

Clicking on **results** link will take you to a standard BioHPC job completion notification page containing links to result files. The notification pages depend slightly on the applications they serve, but are generally similar to one another.



Step	Application	
1	fastx	e
2	BowtieBuild	N
3	Bowtie	in e
4	SamTools	e N
5	SamTools	e N

Another way of accessing output files is to use File Manager functions.

Click on **File Manager** button.

Other considerations

- Column **Reset**: if you want to re-run any steps of the pipeline (e.g., after changing run options), check the corresponding boxes in the **Reset** column, and then click **Save pipeline**. The job IDs for the steps involved will be set to “N/A” and status will be set to “WAITING”. Clicking on **Activate pipeline** will signal BioHPC to resume processing of the “WAITING” steps.
- **Pipeline templates**: Once a pipeline has been configured and saved with **Save pipeline**, it can be used as a template. Clicking on **Save as template** (after specifying the template name) will make the pipeline available on the dropdown list of templates you can choose from when creating a new pipeline. When a new pipeline is created from a template, its steps can be modified via the application submission pages reached by clicking on **Edit** buttons provided in the pipeline steps table. NOTE: **Save as template** cannot be used instead of **Save Pipeline** – these two buttons have different functionality!