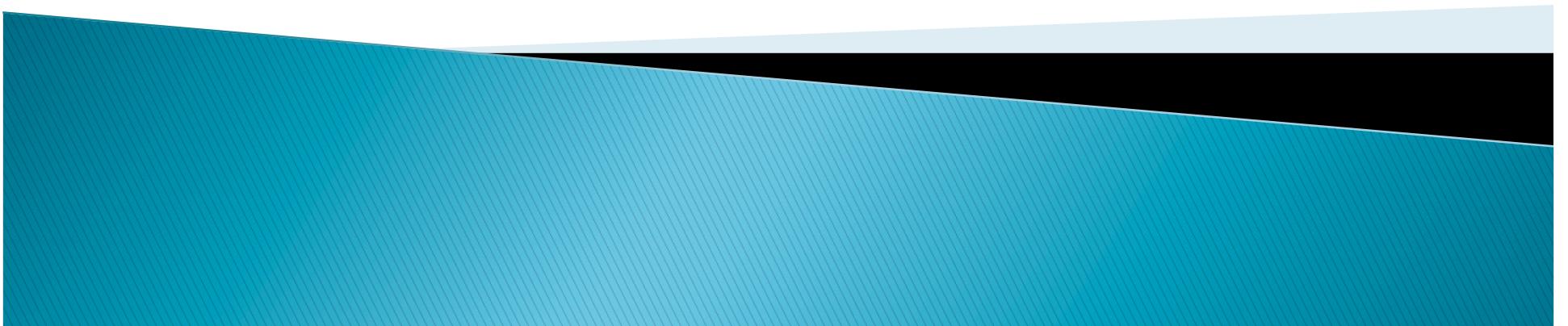


Blast on the lani cluster



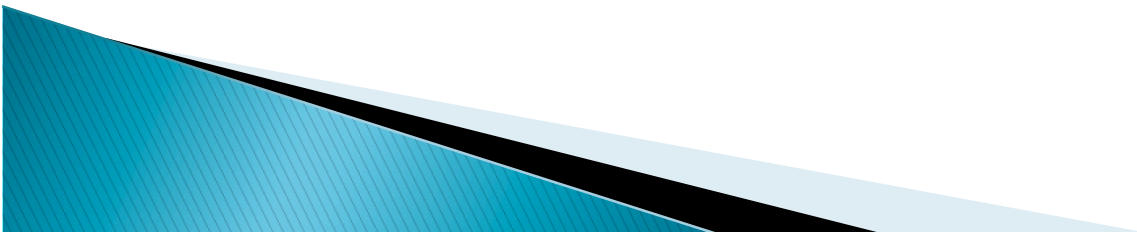
What do I need to connect to the cluster?

- ▶ Need to have SSH software installed
 - Windows: SSH 3.2.9 from ITS
 - <http://www.hawaii.edu/askus/778>
 - Mac: Built in for Mac OS X or newer
 - SSH: via terminal on the Mac
 - `ssh <username>@<IP Address>`
 - GUI for SFTP/SCP: fugu, filezilla
 - <http://rsug.itd.umich.edu/software/fugu/>
 - <http://filezilla-project.org/>
 - Linux: Normally pre-installed
 - openSSH via terminal
 - File transfer via scp or filezilla
- ▶ This guide will focus on Windows and SSH 3.2.9



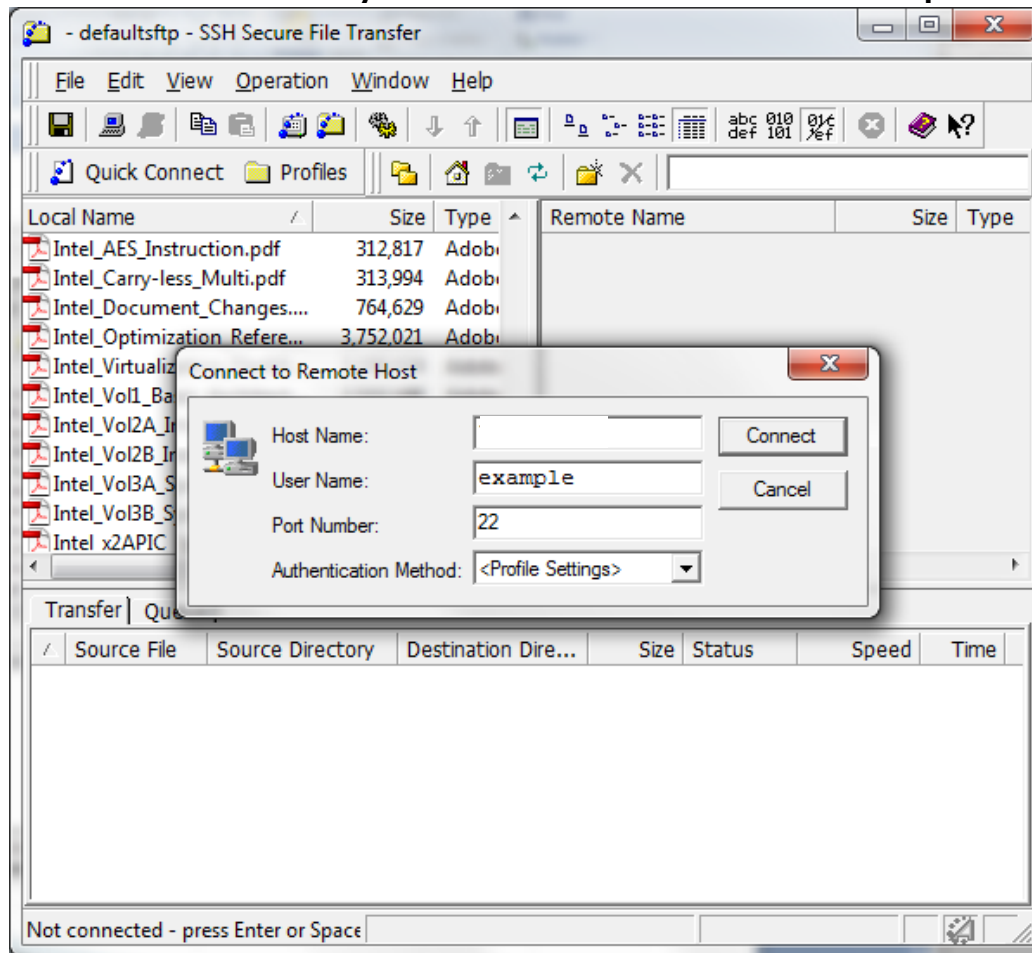
How do I connect to the cluster?

- ▶ Cluster is named lani
 - IP Address: <ip given during account creation process >
 - Consists of 55 machines for processing
 - Is shared amongst multiple users/labs
- ▶ Account information
 - Name: <Assigned upon request>
 - Password should be changed on first login using **passwd**
 - Password requirements:
 - 8 characters, using 3 out of 4 following groups
 - upper case, lower case, digits, symbols
 - Upper case as the first character or a digit as the last character does not fulfill the group requirements



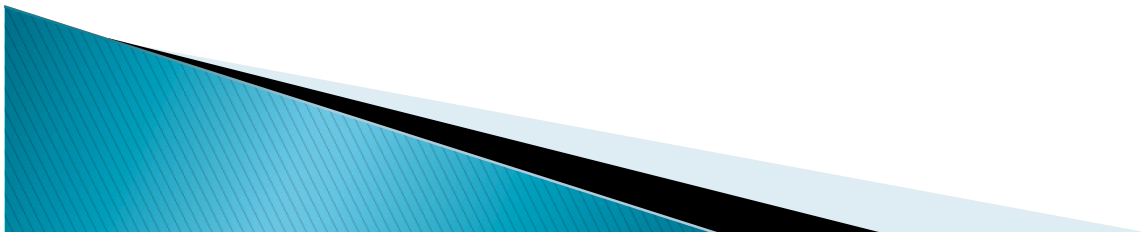
How do I connect to the cluster?

Enter Lani's IP address and your lab user name and press connect



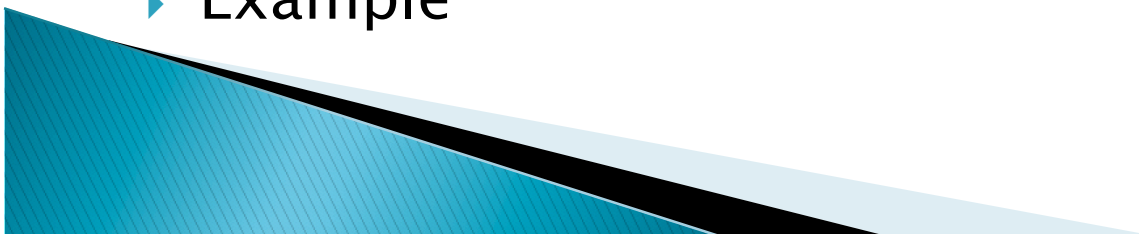
How do I upload data to process?

- ▶ Once you have connected using the SSH Secure File Transfer Client you are able to upload files.
- ▶ Uploading files is as simple as dragging files from your local computer onto to the SSH client under where it says “Remote name”



How do I submit a blast job?

- ▶ Blast jobs should be submitted using
 - **biohadoop_job_submit**
- ▶ Non-graphical question/answer process
 - Answer in the form of a number or plain text
- ▶ Terminology:
 - **Home directory**
 - The directory you are in first in when connecting to lani
 - **Working directory**
 - The directory your terminal is currently in.
- ▶ Example



How do I submit a blast job?

▶ Terminal commands of interest

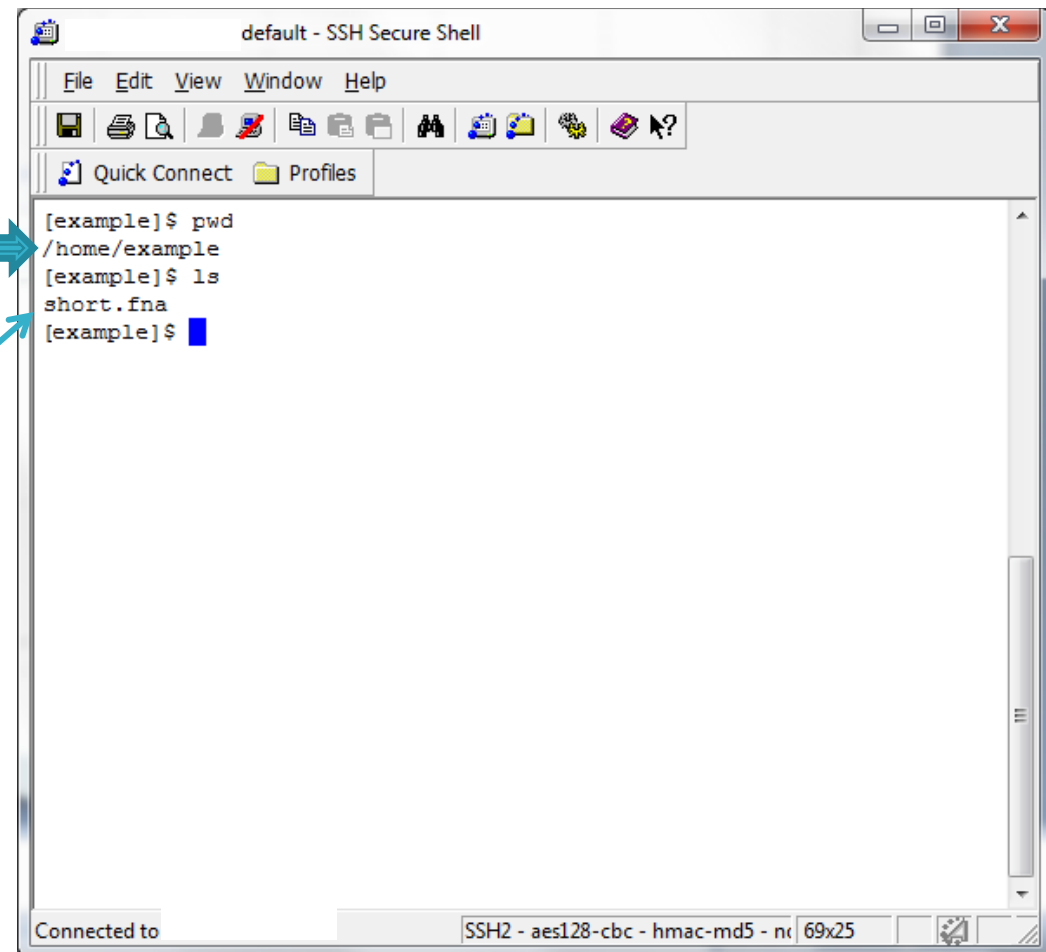
- General commands
 - **cd** – Change Directory
 - **ls** – List Directory
 - **mkdir** – Make directory
 - **passwd** – Change password
 - **pwd** – Print the working directory
 - [Control/Command-c] **Ctrl + c** – Aborts a program running in the current terminal
- Lani specific
 - **showq** – Shows currently queued and running jobs
 - Shows jobs for all users on Lani
 - **biohadoop_job_submit** – Create and submit a job
 - **biohadoop_job_status** – View known information about your jobs
 - **biohadoop_kill_job** – List your jobs and provides a way to cancel a job
 - **biohadoop_blastdb** – Create or remove custom blast databases



How do I submit a blast job?

Working directory

Fasta file in the
working directory



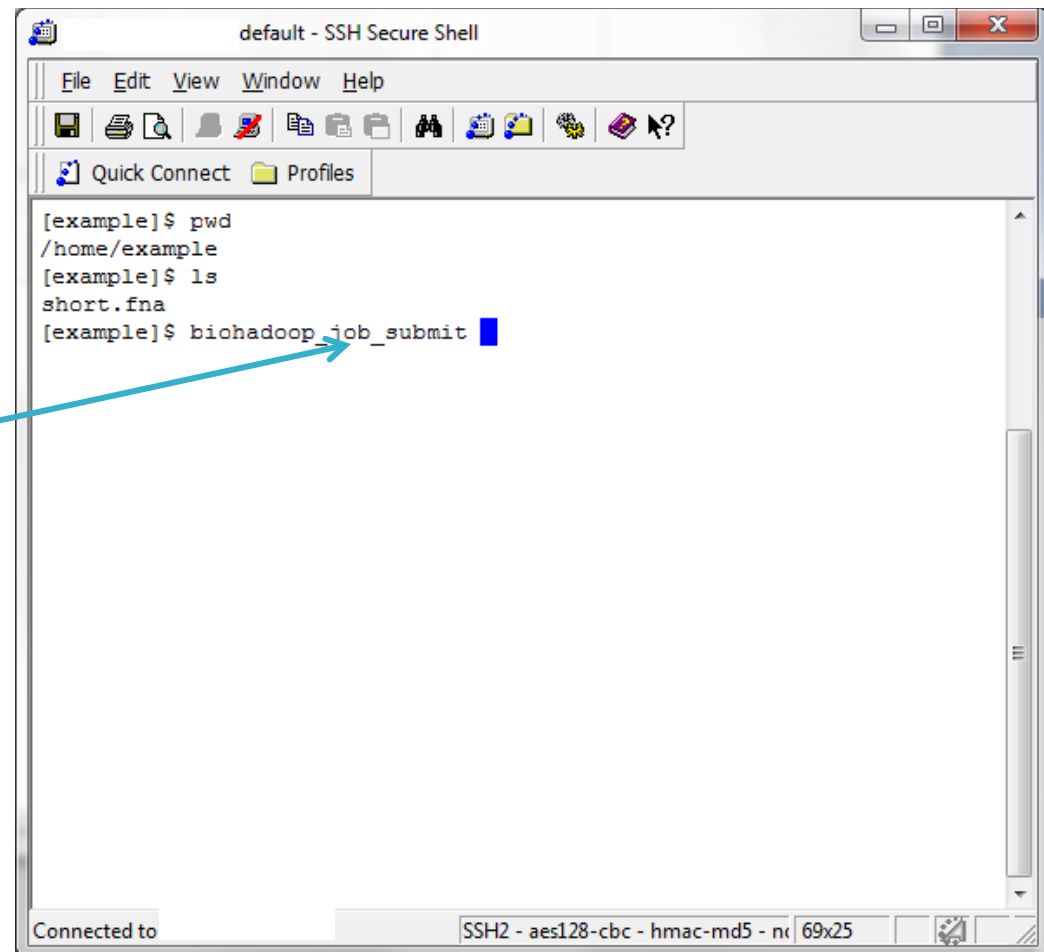
```
default - SSH Secure Shell
File Edit View Window Help
[example]$ pwd
/home/example
[example]$ ls
short.fna
[example]$
```

The screenshot shows a terminal window titled "default - SSH Secure Shell". It has a menu bar with "File", "Edit", "View", "Window", and "Help". Below the menu is a toolbar with various icons. The main area shows a command prompt where the user has entered "pwd" and "ls". The output of "pwd" is "/home/example" and the output of "ls" is "short.fna". A blue cursor is visible after the second prompt. At the bottom, a status bar indicates "Connected to" and "SSH2 - aes128-cbc - hmac-md5 - n 69x25".

How do I submit a blast job?

Job submission command

Execute this from the same directory as your input to simplify future steps

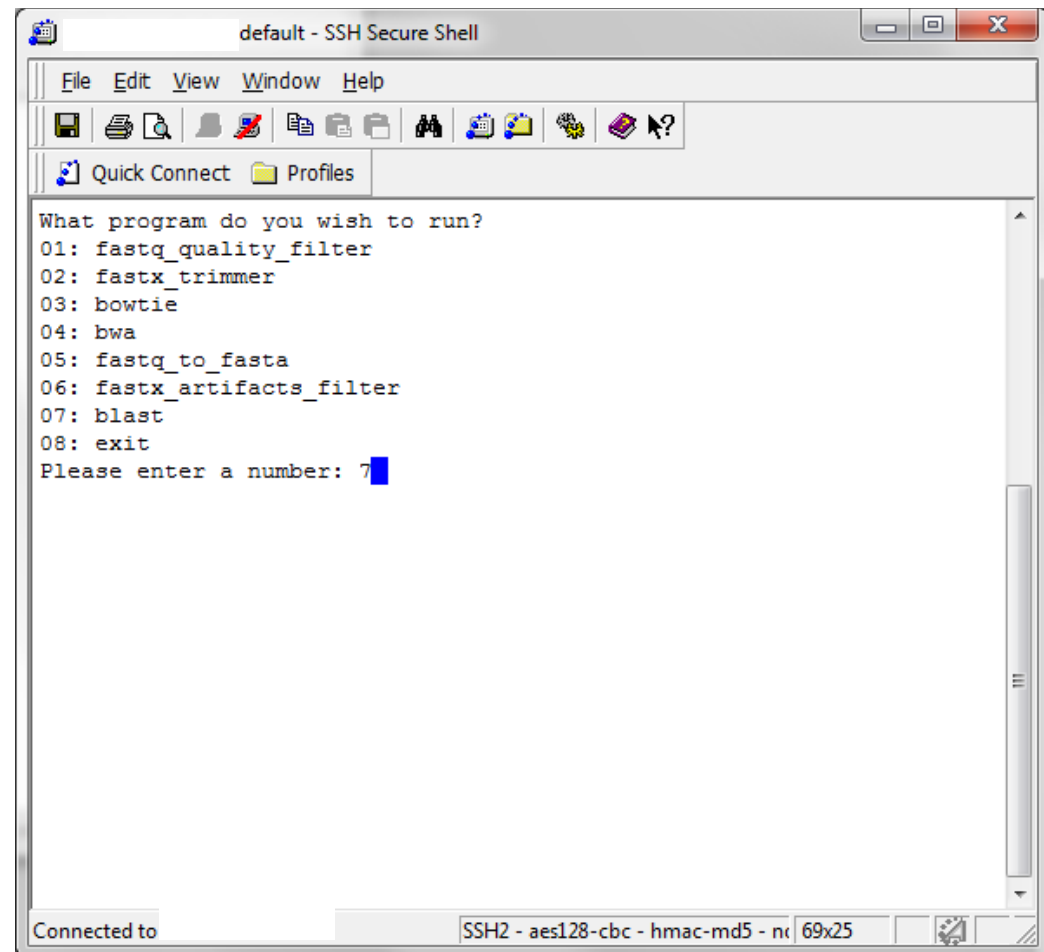


The screenshot shows a terminal window titled "default - SSH Secure Shell". The window has a menu bar with "File", "Edit", "View", "Window", and "Help". Below the menu bar is a toolbar with various icons. The main area of the window displays the following commands and their outputs:

```
[example]$ pwd
/home/example
[example]$ ls
short.fna
[example]$ biohadoop_job_submit
```

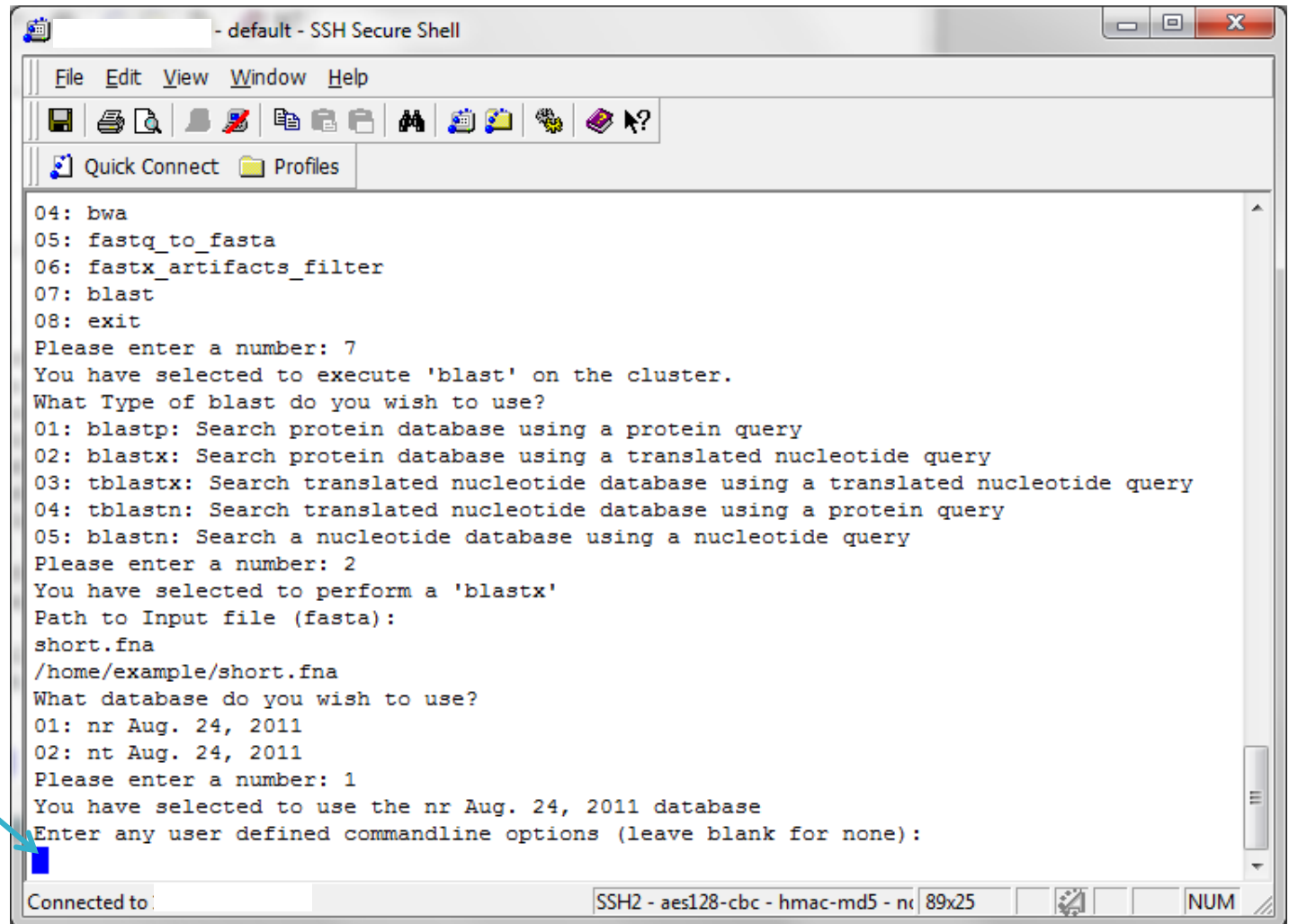
A blue arrow points from the text "Job submission command" to the command `biohadoop_job_submit` in the terminal. The status bar at the bottom of the window shows "Connected to" followed by "SSH2 - aes128-cbc - hmac-md5 - n" and "69x25".

How do I submit a blast job?



How do I submit a blast job?

Able to add
additional options
that are available
to the application



```
- default - SSH Secure Shell

File Edit View Window Help

Quick Connect Profiles

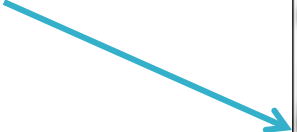
04: bwa
05: fastq_to_fasta
06: fastx_artifacts_filter
07: blast
08: exit
Please enter a number: 7
You have selected to execute 'blast' on the cluster.
What Type of blast do you wish to use?
01: blastp: Search protein database using a protein query
02: blastx: Search protein database using a translated nucleotide query
03: tblastx: Search translated nucleotide database using a translated nucleotide query
04: tblastn: Search translated nucleotide database using a protein query
05: blastn: Search a nucleotide database using a nucleotide query
Please enter a number: 2
You have selected to perform a 'blastx'
Path to Input file (fasta):
short.fna
/home/example/short.fna
What database do you wish to use?
01: nr Aug. 24, 2011
02: nt Aug. 24, 2011
Please enter a number: 1
You have selected to use the nr Aug. 24, 2011 database
Enter any user defined commandline options (leave blank for none):

```

Connected to: SSH2 - aes128-cbc - hmac-md5 - nc 89x25 NUM

How do I submit a blast job?

This depends
on how many records
you plan to blast



```
- default - SSH Secure Shell

File Edit View Window Help

Quick Connect Profiles

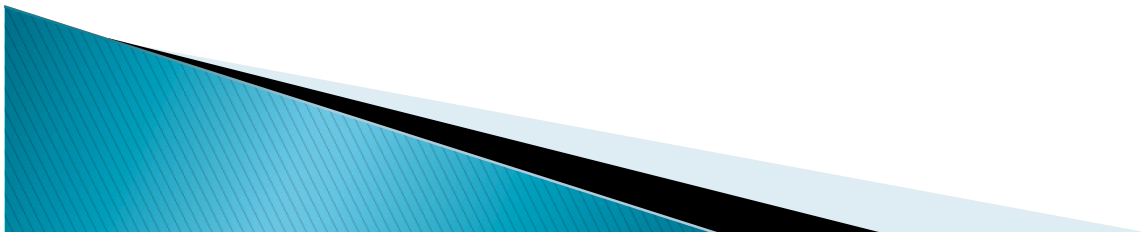
06: fastx_artifacts_filter
07: blast
08: exit
Please enter a number: 7
You have selected to execute 'blast' on the cluster.
What Type of blast do you wish to use?
01: blastp: Search protein database using a protein query
02: blastx: Search protein database using a translated nucleotide query
03: tblastx: Search translated nucleotide database using a translated nucleotide query
04: tblastn: Search translated nucleotide database using a protein query
05: blastn: Search a nucleotide database using a nucleotide query
Please enter a number: 2
You have selected to perform a 'blastx'
Path to Input file (fasta):
short.fna
/home/example/short.fna
What database do you wish to use?
01: nr Aug. 24, 2011
02: nt Aug. 24, 2011
Please enter a number: 1
You have selected to use the nr Aug. 24, 2011 database
Enter any user defined commandline options (leave blank for none):

Enter number of nodes to use (max: 55):
█

Connected to: SSH2 - aes128-cbc - hmac-md5 - nc 89x25 NUM
```

How do I submit a blast job?

- ▶ How many nodes do I need for my job?
 - This depends largely on how big your input is.
- ▶ The cluster is designed for large datasets
 - Large is defined as **at least 5,000 sequences**
 - It will work on smaller datasets but may not be optimal
- ▶ The cluster is a **shared resource**
 - The amount of nodes requested should reflect this
 - Recommended max is about $\frac{1}{2}$ the total nodes per job
 - More nodes maybe used if the job is short lived (1–2 days)



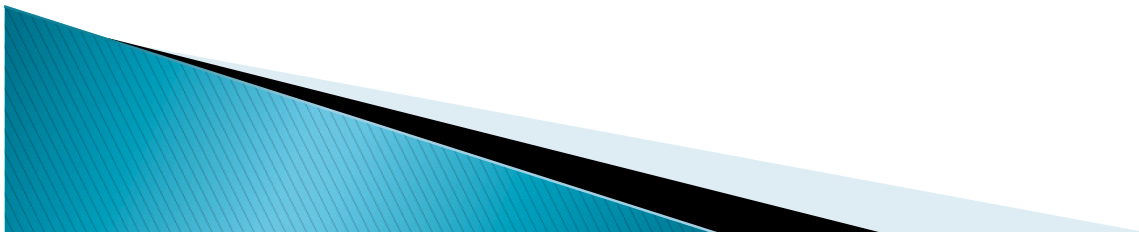
How do I submit a blast job?

▶ Runtime with a real input

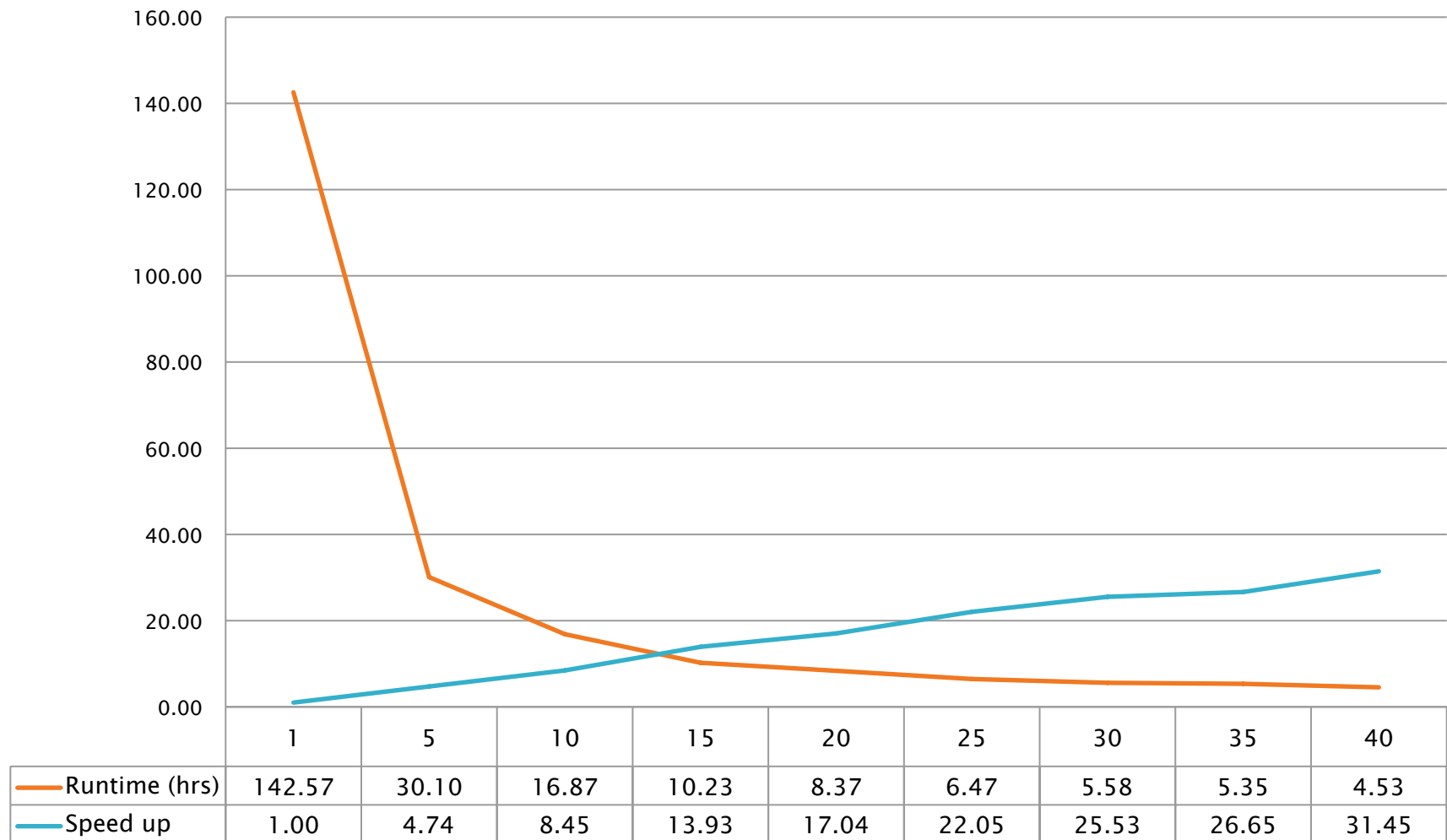
- Blastx
- Database: nr
- e-value cut off: $1e-05$
- Input Size: 247,586 sequences
- Nodes: 35 nodes

▶ Results

- Execution time: 2 days 11 hrs and 44 min
- 69.08 sequences per minute
- 1.97 sequences per minute per node



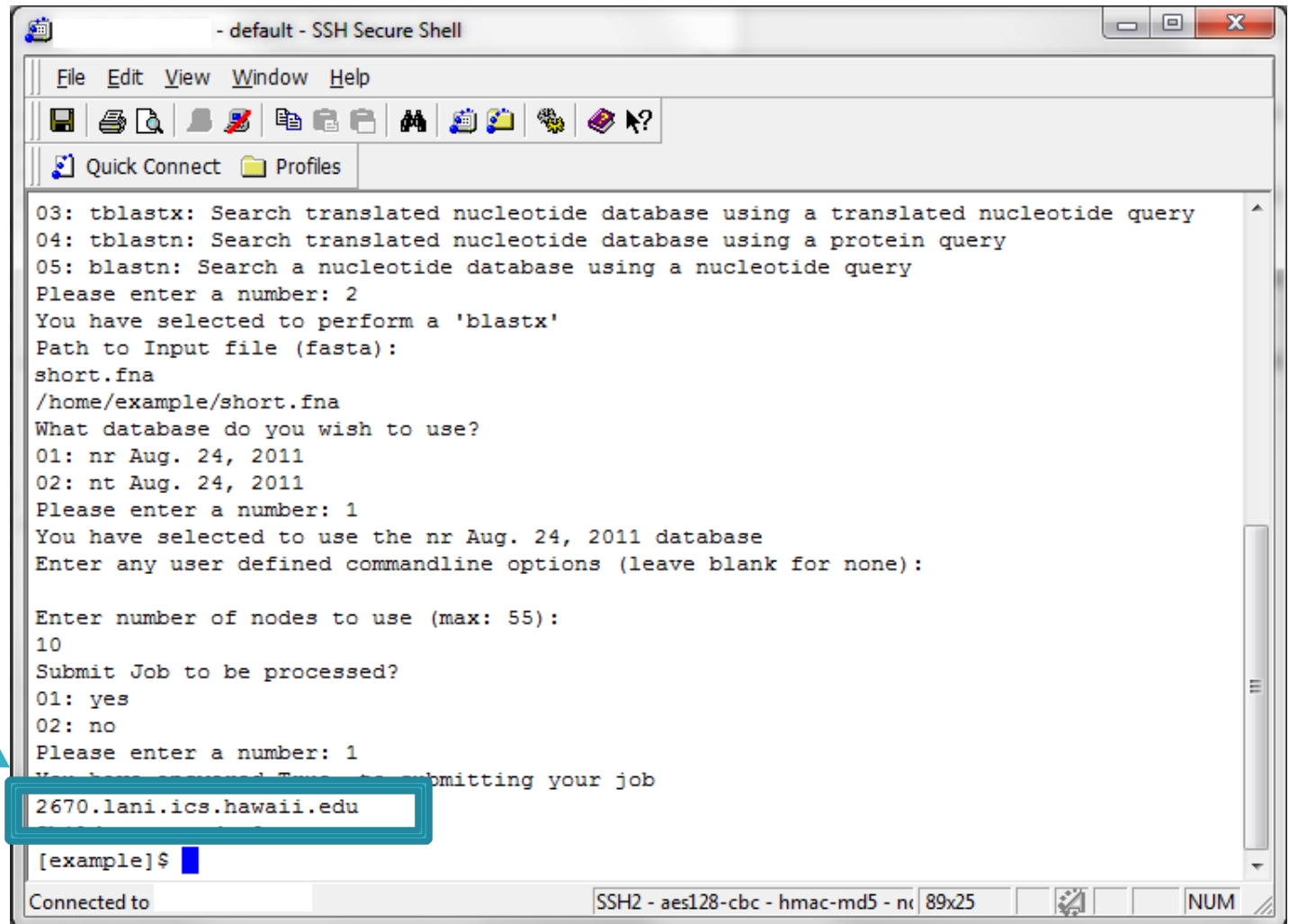
Runtime vs. # of nodes for blastx



Blastx vs. nr , e-value= 1e-05
using 4318 sequences as input

How do I submit a blast job?

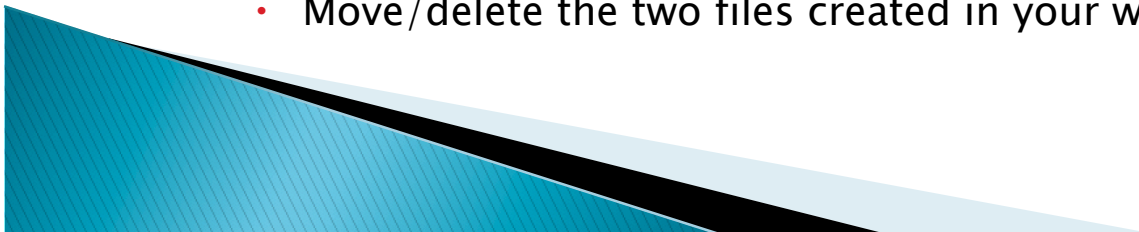
Copy this line down.
This is your **Job ID**



```
- default - SSH Secure Shell
File Edit View Window Help
Quick Connect Profiles
03: tblastx: Search translated nucleotide database using a translated nucleotide query
04: tblastn: Search translated nucleotide database using a protein query
05: blastn: Search a nucleotide database using a nucleotide query
Please enter a number: 2
You have selected to perform a 'blastx'
Path to Input file (fasta):
short.fna
/home/example/short.fna
What database do you wish to use?
01: nr Aug. 24, 2011
02: nt Aug. 24, 2011
Please enter a number: 1
You have selected to use the nr Aug. 24, 2011 database
Enter any user defined commandline options (leave blank for none):
Enter number of nodes to use (max: 55):
10
Submit Job to be processed?
01: yes
02: no
Please enter a number: 1
You have answered 'Yes' to submitting your job
2670.lani.ics.hawaii.edu
[example]$
Connected to SSH2 - aes128-cbc - hmac-md5 - nc 89x25 NUM
```

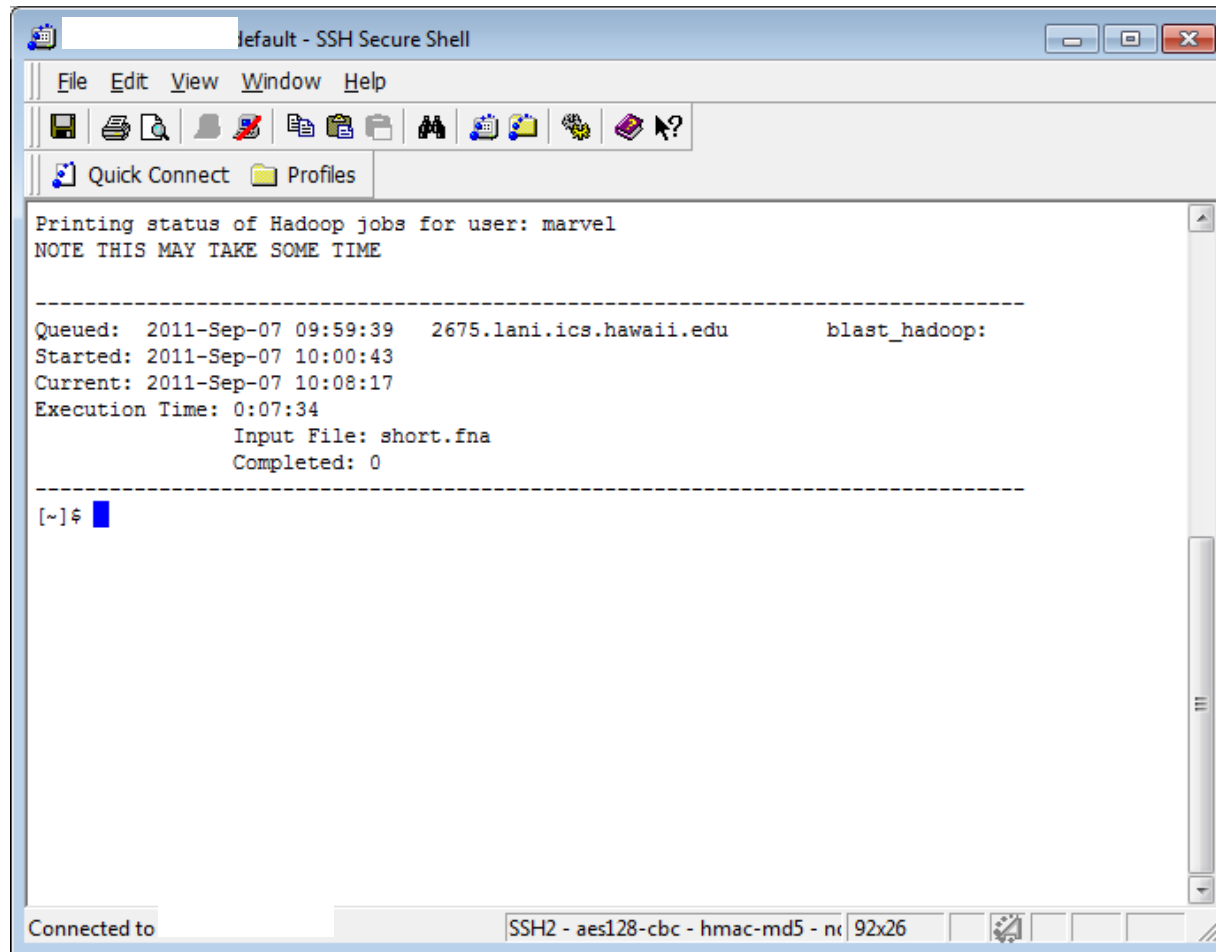

What happens after I submit?

- ▶ Two files are created in the working directory
 - *.submit – Specifies the work flow of the job
 - *.yaml – Information/parameters specific to the job
 - Both files are moved into the **job folder** once the job becomes active
- ▶ A new folder is made in the working directory when the job starts to run
 - The new folder is named after your Job ID (known as the **job folder**)
 - Contains important job information during execution
 - Contains the results upon the completion of the job
 - *.pbsout – Information returned about the job
 - error – The error output from the job
 - output – The output or results from the job
 - input – Contains the name of the original input
- ▶ While a job is running:
 - Do not:
 - Move/delete the input file
 - Move/delete the job folder
 - Move/delete the two files created in your working directory



How do I verify the status of my jobs?

Run the command **biohadoop_job_status**



The screenshot shows a terminal window titled "default - SSH Secure Shell". The window has a menu bar with "File", "Edit", "View", "Window", and "Help". Below the menu bar is a toolbar with various icons. The main text area displays the output of the `biohadoop_job_status` command for user "marvel". The output includes a header, a note about execution time, a table of job details, and a footer with a prompt.

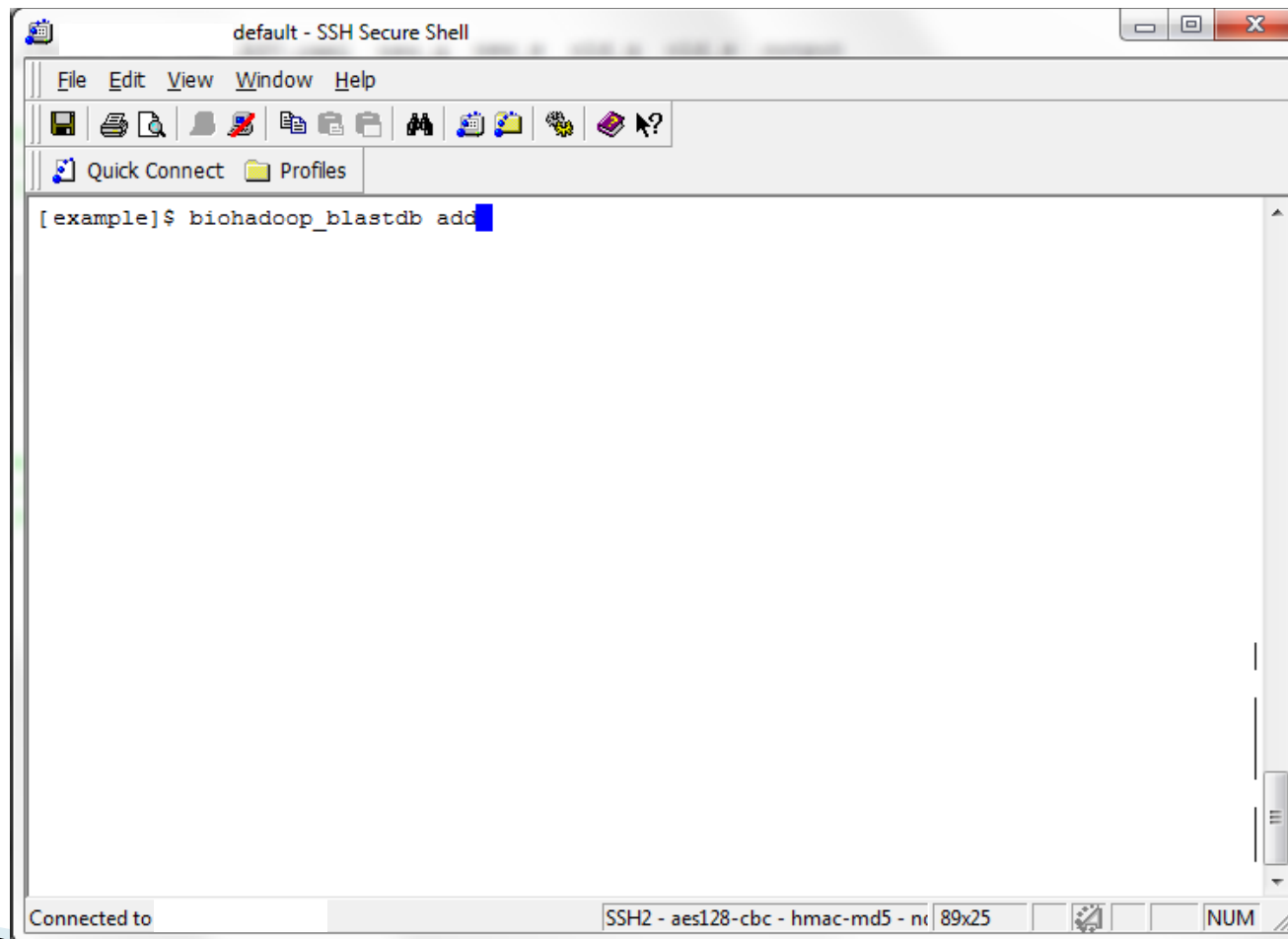
```
Printing status of Hadoop jobs for user: marvel
NOTE THIS MAY TAKE SOME TIME

-----
Queued:  2011-Sep-07 09:59:39    2675.lani.ics.hawaii.edu    blast_hadoop:
Started: 2011-Sep-07 10:00:43
Current: 2011-Sep-07 10:08:17
Execution Time: 0:07:34
           Input File: short.fna
           Completed: 0
-----

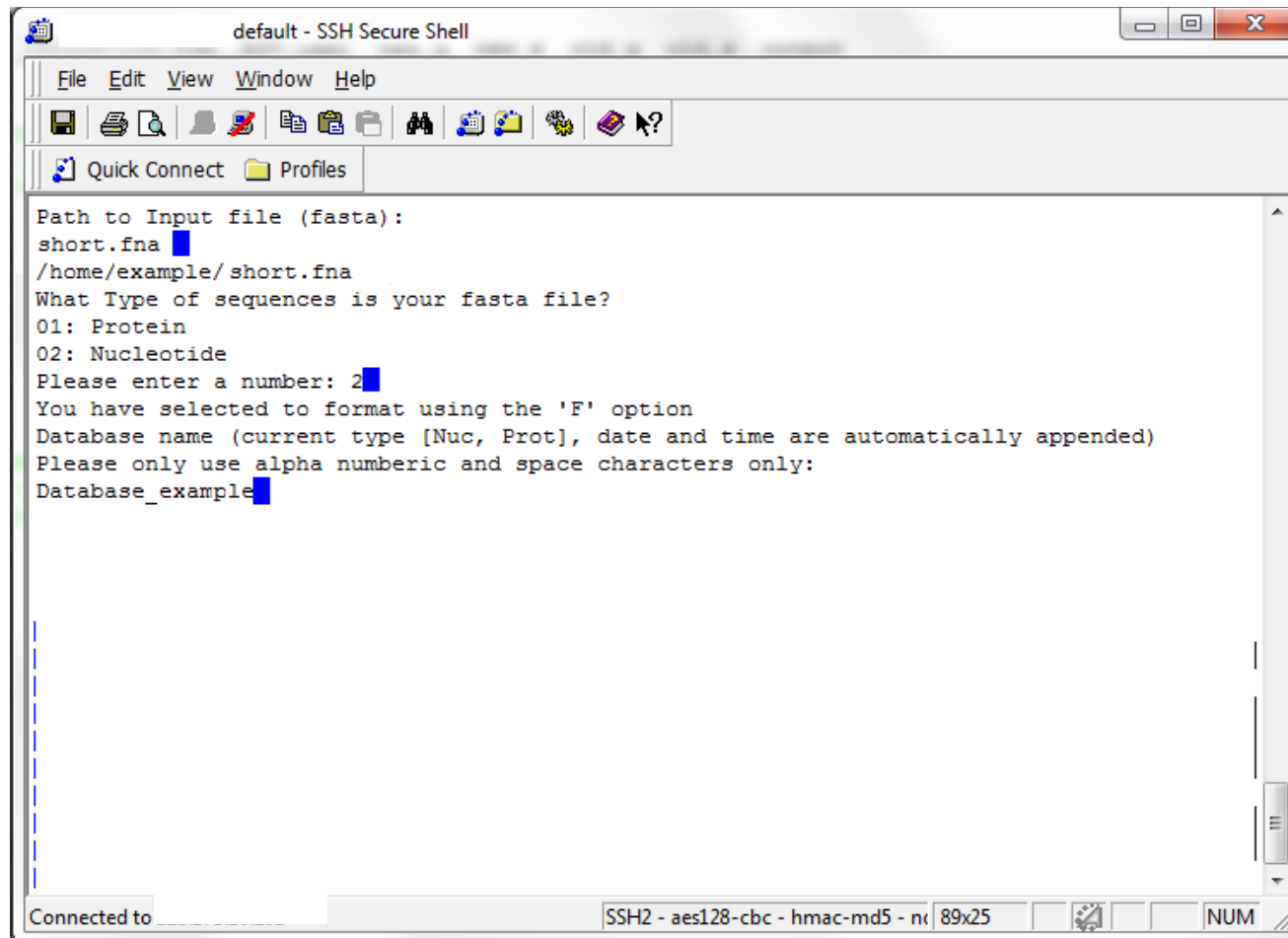
[~]$
```

At the bottom of the window, a status bar shows "Connected to" followed by "SSH2 - aes128-cbc - hmac-md5 - nc 92x26".

How do I make a database?



How do I make a database?

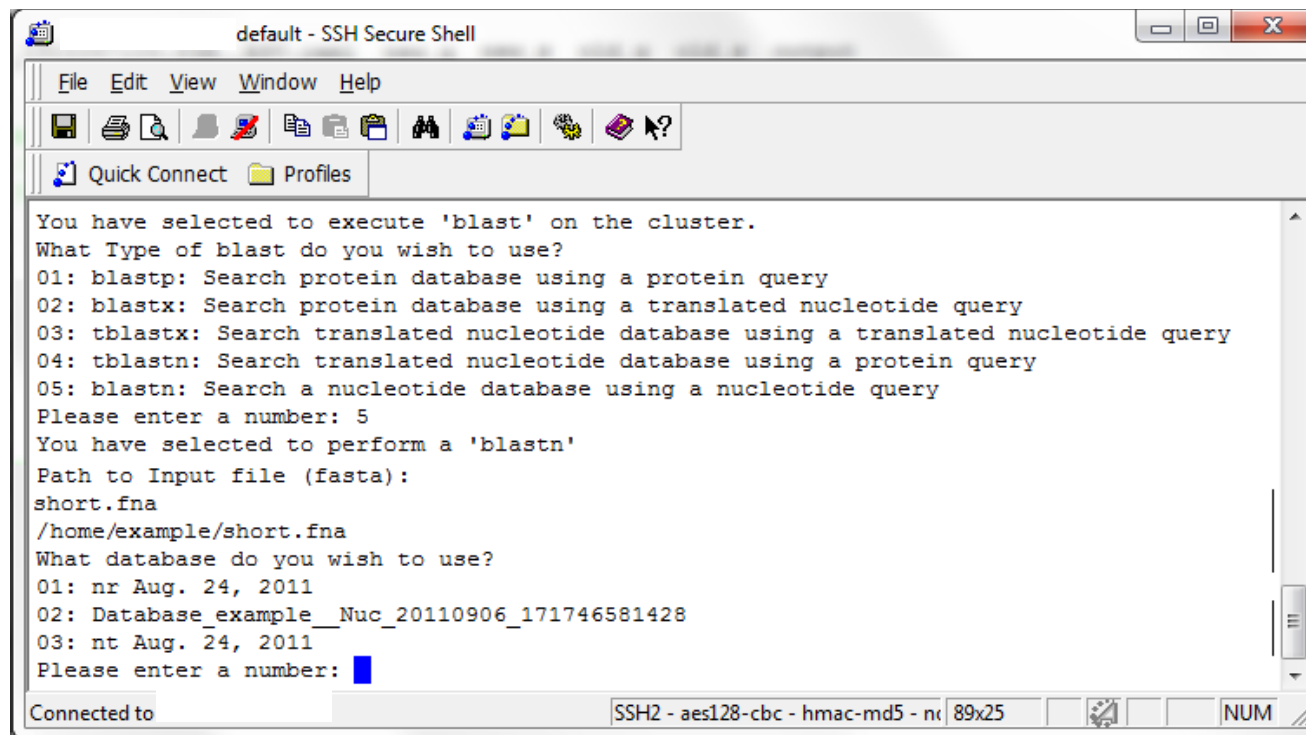


The screenshot shows a terminal window titled "default - SSH Secure Shell". It contains a menu bar with "File", "Edit", "View", "Window", and "Help". Below the menu is a toolbar with various icons. The main text area shows the following prompts and user input:

```
Path to Input file (fasta):
short.fna
/home/example/short.fna
What Type of sequences is your fasta file?
01: Protein
02: Nucleotide
Please enter a number: 2
You have selected to format using the 'F' option
Database name (current type [Nuc, Prot], date and time are automatically appended)
Please only use alpha numeric and space characters only:
Database_example
```

The status bar at the bottom indicates "Connected to" followed by a redacted address, "SSH2 - aes128-cbc - hmac-md5 - n", "89x25", and a "NUM" button.

How do I make a database?



The screenshot shows a terminal window titled "default - SSH Secure Shell". The window has a menu bar with "File", "Edit", "View", "Window", and "Help". Below the menu bar is a toolbar with various icons. The main text area displays the following content:

```
You have selected to execute 'blast' on the cluster.  
What Type of blast do you wish to use?  
01: blastp: Search protein database using a protein query  
02: blastx: Search protein database using a translated nucleotide query  
03: tblastx: Search translated nucleotide database using a translated nucleotide query  
04: tblastn: Search translated nucleotide database using a protein query  
05: blastn: Search a nucleotide database using a nucleotide query  
Please enter a number: 5  
You have selected to perform a 'blastn'  
Path to Input file (fasta):  
short.fna  
/home/example/short.fna  
What database do you wish to use?  
01: nr Aug. 24, 2011  
02: Database_example__Nuc_20110906_171746581428  
03: nt Aug. 24, 2011  
Please enter a number: █
```

The status bar at the bottom of the window shows "Connected to" followed by "SSH2 - aes128-cbc - hmac-md5 - nc 89x25" and a "NUM" button.