Contents

1. Connecting to Biowulf
   1. With a Mac computer
   2. With a Windows PC
2. The command line
3. Handy unix commands
4. Unix tips and tricks
5. File transfer connections to Biowulf
   1. Globus.org
   2. Mount a drive
   3. WinSCP (windows)
   4. scp at the command line (Mac)
   5. FileZilla (only download from a trusted source)
6. File formats
   1. FASTA, FASTQ
   2. SAM, BAM
7. Modules (fastqc, bowtie2, samtools)
8. IGV (Integrated Genome Viewer)
9. Additional resources
10. Connecting to Biowulf
    1. On Mac, open the “Terminal” program and type

“ssh [username@biowulf.nih.gov](mailto:username@biowulf.nih.gov)” (without the quotes)

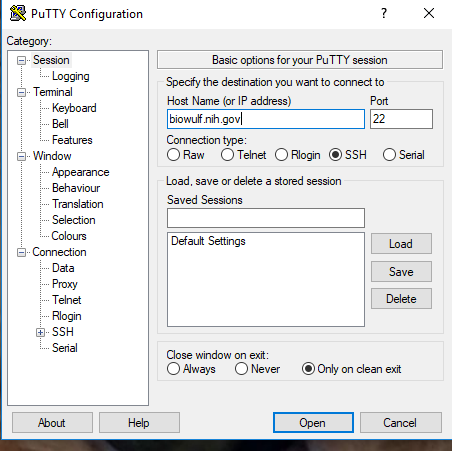
* 1. On PC, you will need to download and install PuTTY from:

<https://www.chiark.greenend.org.uk/~sgtatham/putty/latest.html>

Host Name: biowulf.nih.gov

Connection Type: SSH

Figure #: the PuTTY interface



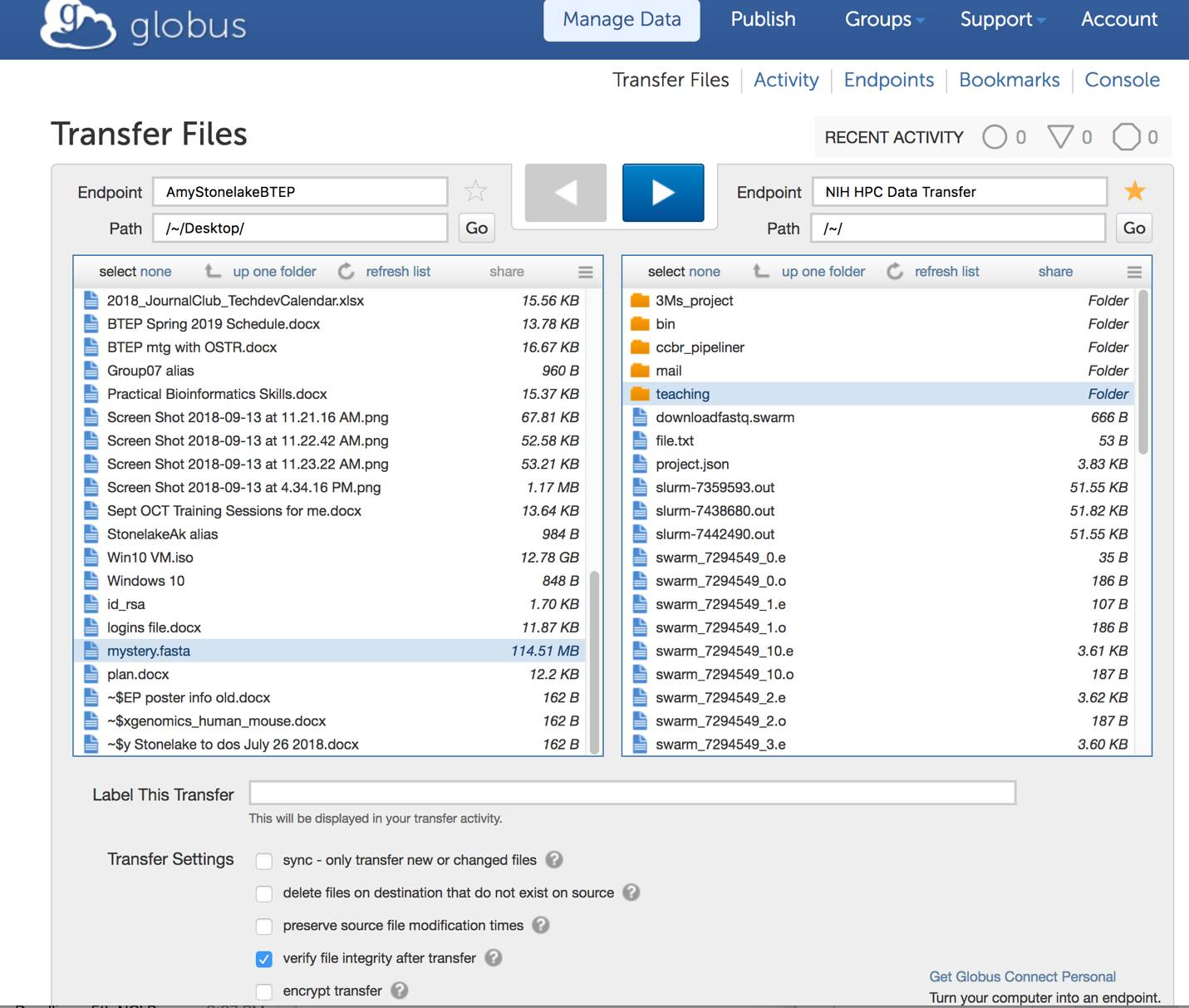
1. The Command line

[username@biowulf ~] $

* 1. “username” is your username
  2. @biowulf means you are logged into biowulf
  3. “~” indicates your home directory
  4. When you see the dollar sign “$”, you know you are at the command line
  5. If you don’t see the dollar sign, something is going on (running a program)

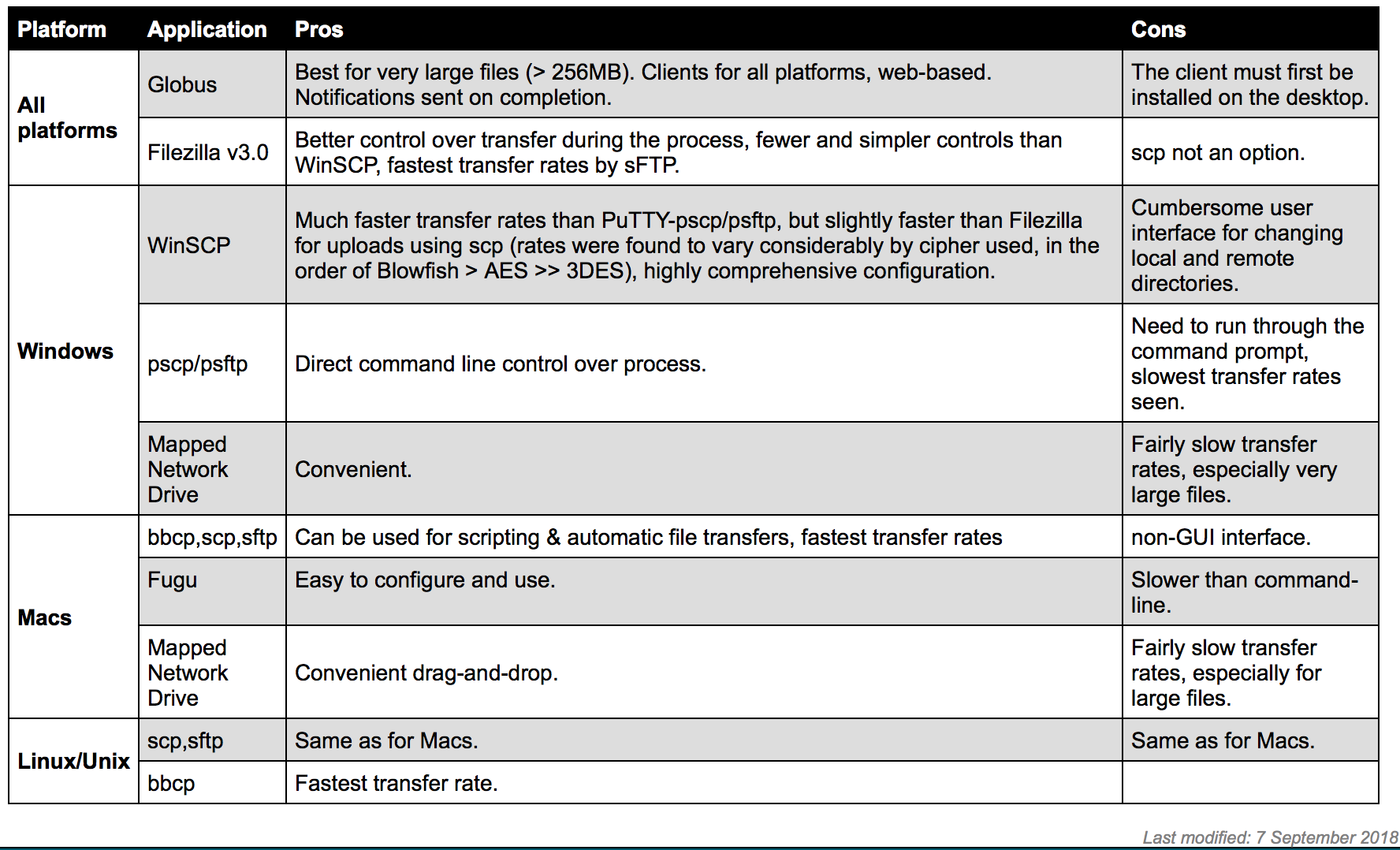
1. Handy unix commands
   1. pwd (print working directory)
   2. ls (list contents)
   3. cd (change directory)
   4. cd .. (go to home directory)
   5. cd ../..
   6. cd /home/pathtofile
   7. less (peek inside a file, press “q” to quit)
   8. ls -l (list details)
   9. ls -a (list hidden . files)
   10. ls -alt
   11. rm (remove file)
   12. rmdir (remove directory)
   13. mkdir (make directory)
   14. nano file.txt (nano editor for creating files)
   15. mv (move or rename files) mv oldfilename.txt newfilename.txt
   16. touch filename.txt (creates an empty file)
2. Unix tips and tricks
   1. Use the “up arrow” to go back to the previous command
   2. Use “tab complete” to finish typing a uniquely named file, directory or program
3. File transfer connections to Biowulf
   1. Globus.org
      1. Setup your Globus endpoint (only need to do this one time)
      2. Open Globus Connect Personal (need to do this every time)
      3. Go to globus.org
      4. Choose your personal endpoint
      5. Choose a folder on biowulf
      6. Click the blue arrow
      7. You get an e-mail when it’s done!
      8. Need to have helix/biowulf account to get globus

Figure #: the globus interface



* 1. Mount a drive
     1. Mac – “Go” -> ”Connect to server”
     2. PC - ”Computer”, “Tools” then “Map Network Drive” tab
     3. Be sure to set host as “smb://biowulf.nih.gov””
     4. See instructions on hpc.nih.gov (Biowulf) – “How To – Transfer Files”, “Transferring data to/from the NIH HPC systems

Figure #: File transfer options on Biowulf (hpc.nih.gov)



* 1. Secure ftp (sftp) or secure copy protocol (scp)
     1. FileZilla – be sure to get a clean copy!
     2. Mac OSX:

<http://packages.partek.com/bin/filezilla/fz-osx.app.tar.bz2>

* + 1. Windows 32-bit:

<http://packages.partek.com/bin/filezilla/fz-win32.exe>

* + 1. Windows 64-bit:

<http://packages.partek.com/bin/filezilla/fz-win64.exe>

Figure #: logging into Biowulf with Filezilla

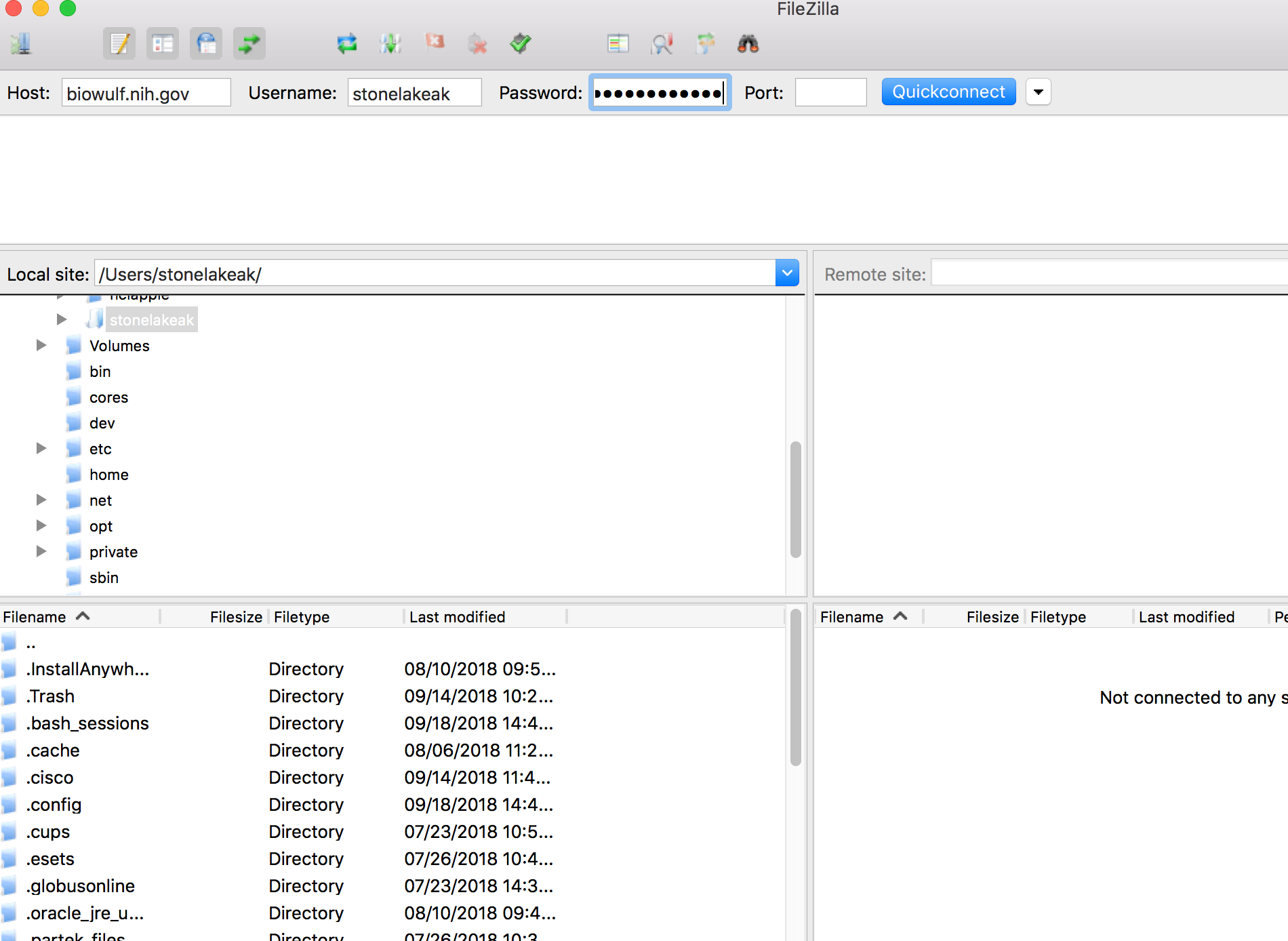


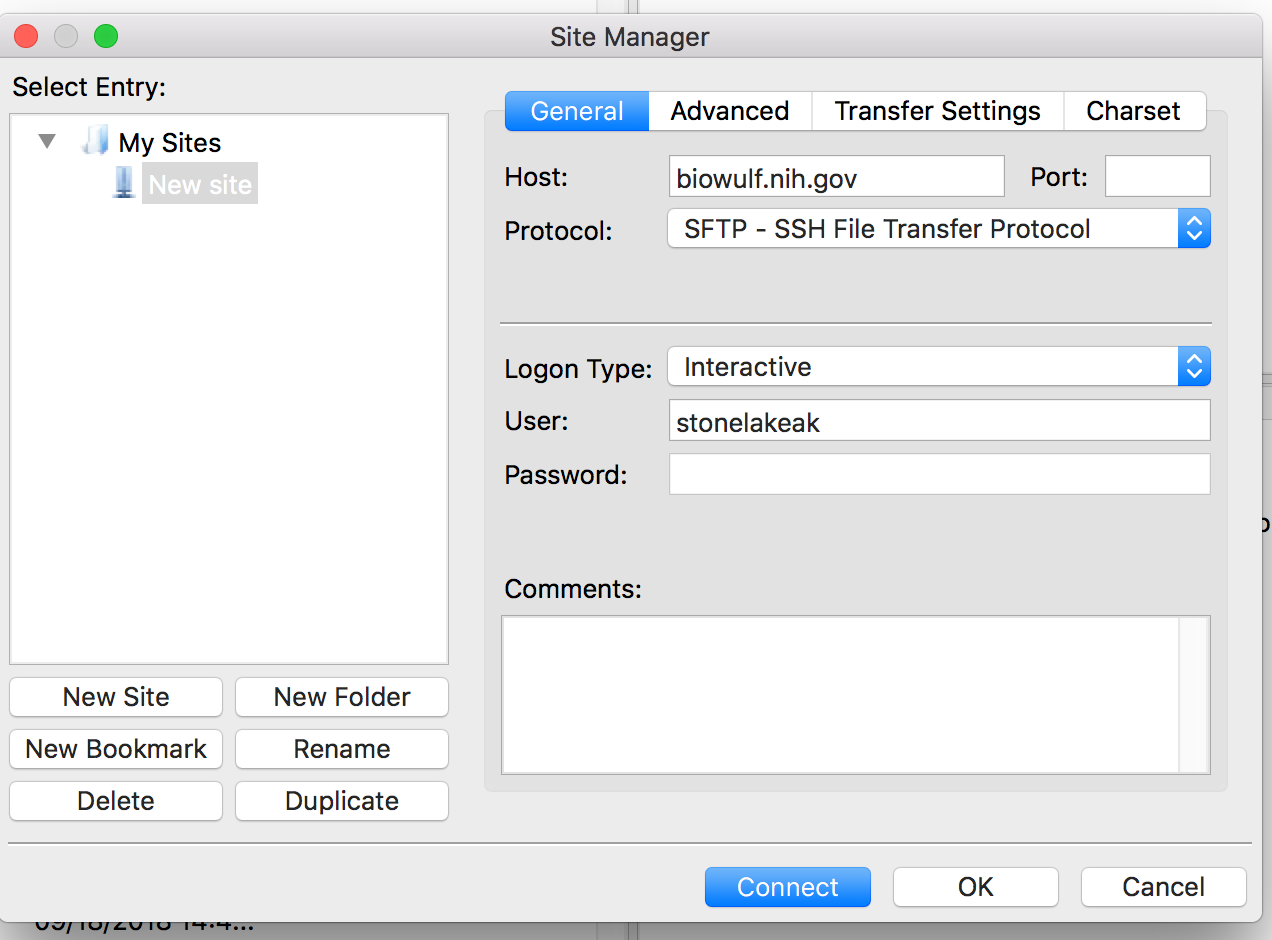
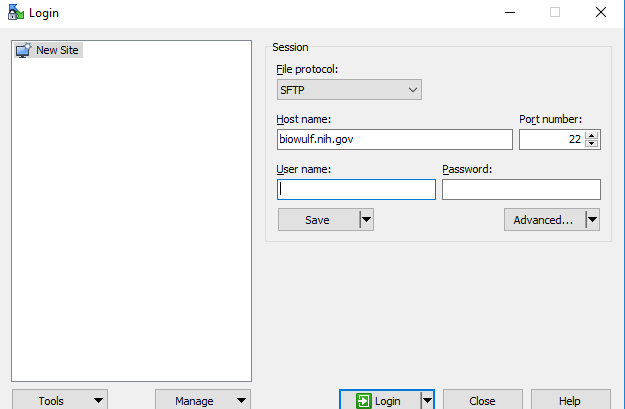
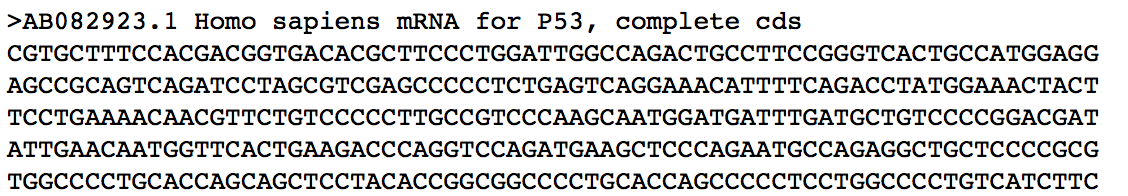
Figure #: setting up your connection to Biowulf with FileZilla

Figure #: WinSCP SFTP login interface



1. File Formats
   1. FASTA format has a header line followed by sequence data

Figure #: FASTA formatted sequence



* 1. FASTQ files contain sequence data that includes a quality (Phred) score for each nucleotide. Quality (Q) scores are integers usually between 2-40.
     1. line 1 starts with '@' character followed by a sequence identifier and optional description that may include instrument ID, run number, lane number, tile number, X and Y coordinates of cluster, read number, sample number from sample sheet.
     2. line 2 is the raw sequence letters (ACTG)
     3. line 3 begins with a '+' character and is *optionally* followed by the same sequence identifier (and any description) again
     4. line 4 has quality (Q) scores for each base (aka Phred score). Q scores are integer values that represent the estimated probability of an error. If the Q score is 0, the probability that it is incorrect is 1.0, and the ASCII symbol is “!”.

Figure #: The 4 lines of information in a FASTQ file (drive5.com)

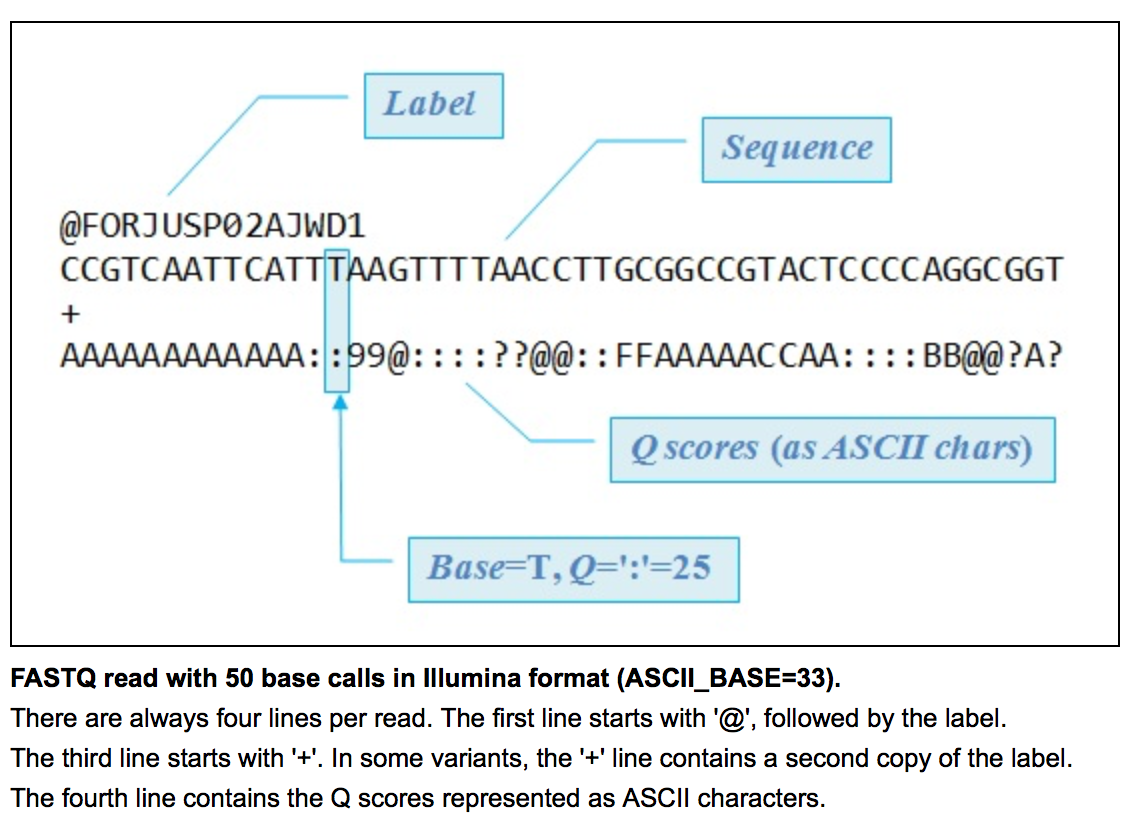
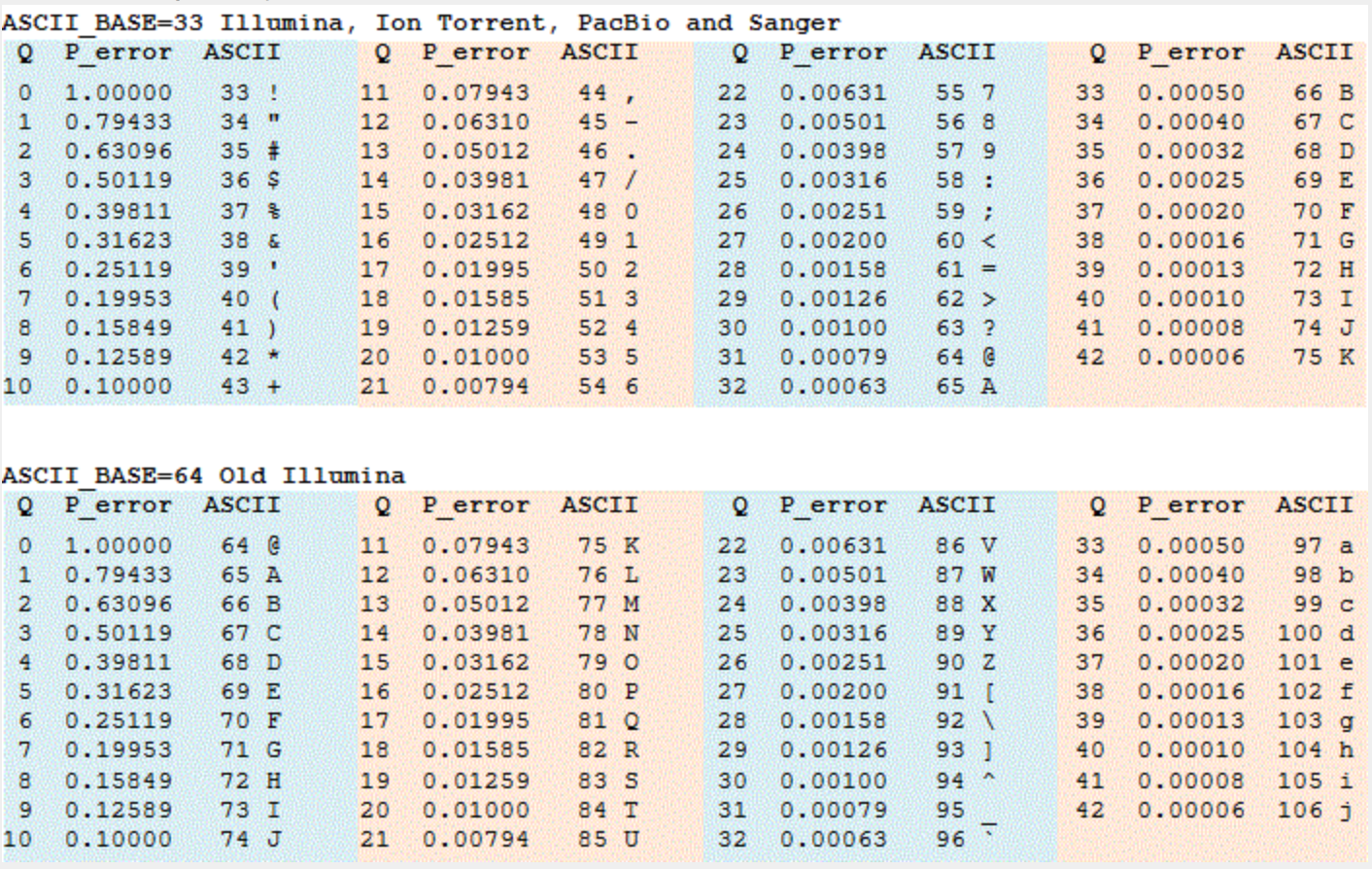
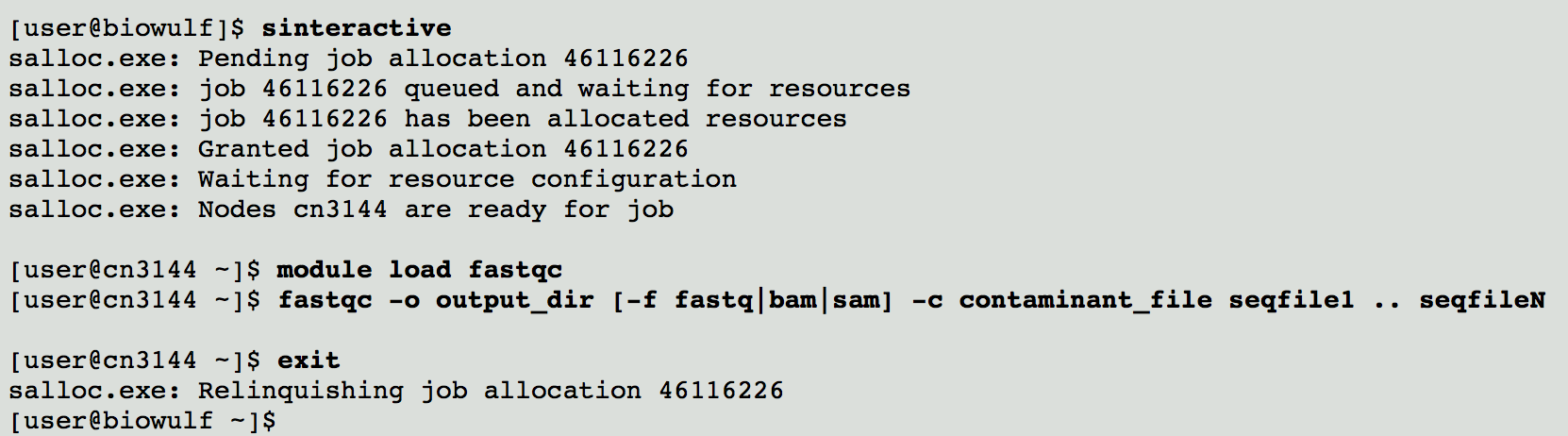


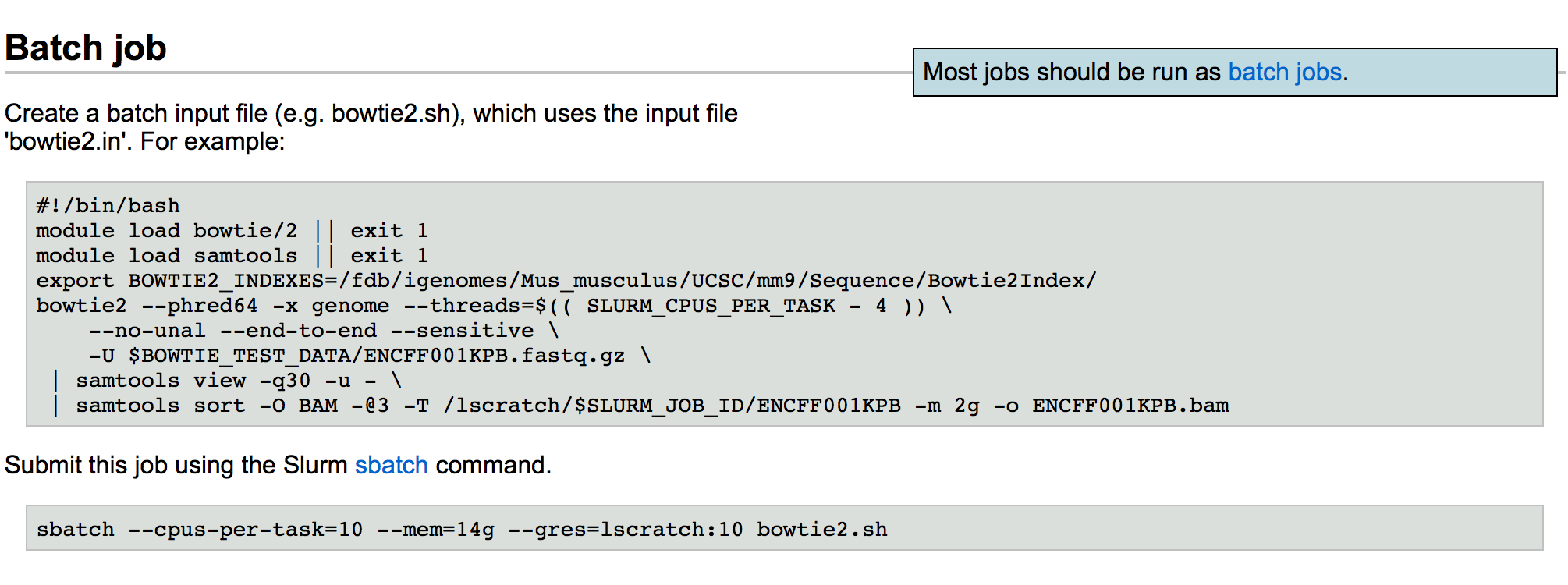
Figure #: Quality scores as represented by ASCII characters (drive5.com)



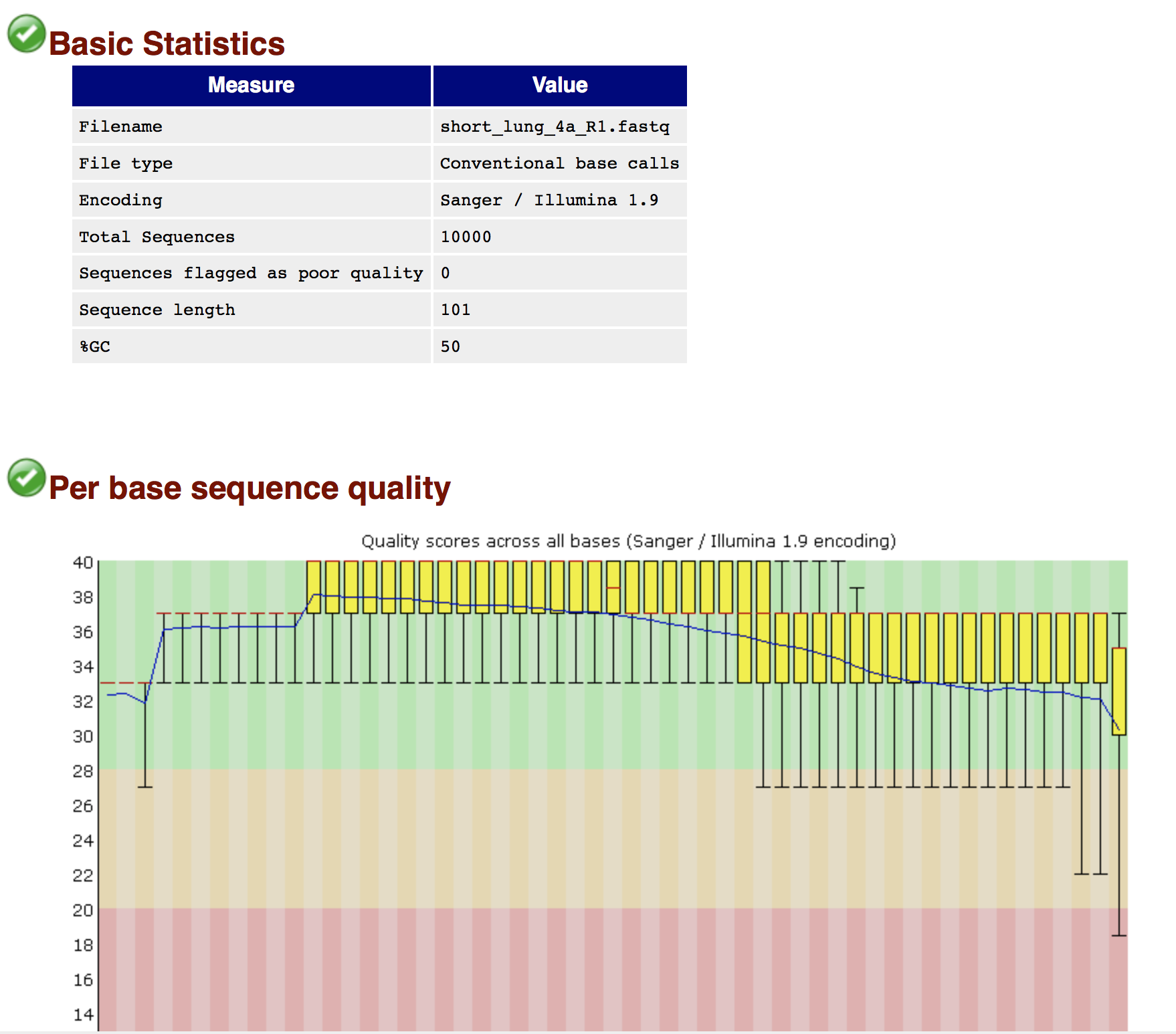
* 1. SAM – a tab-delimited text file with optional header and alignment sections, use program SAMtools, each alignment has 11 mandatory fields
  2. BAM – binary format of SAM, can be sorted and indexed

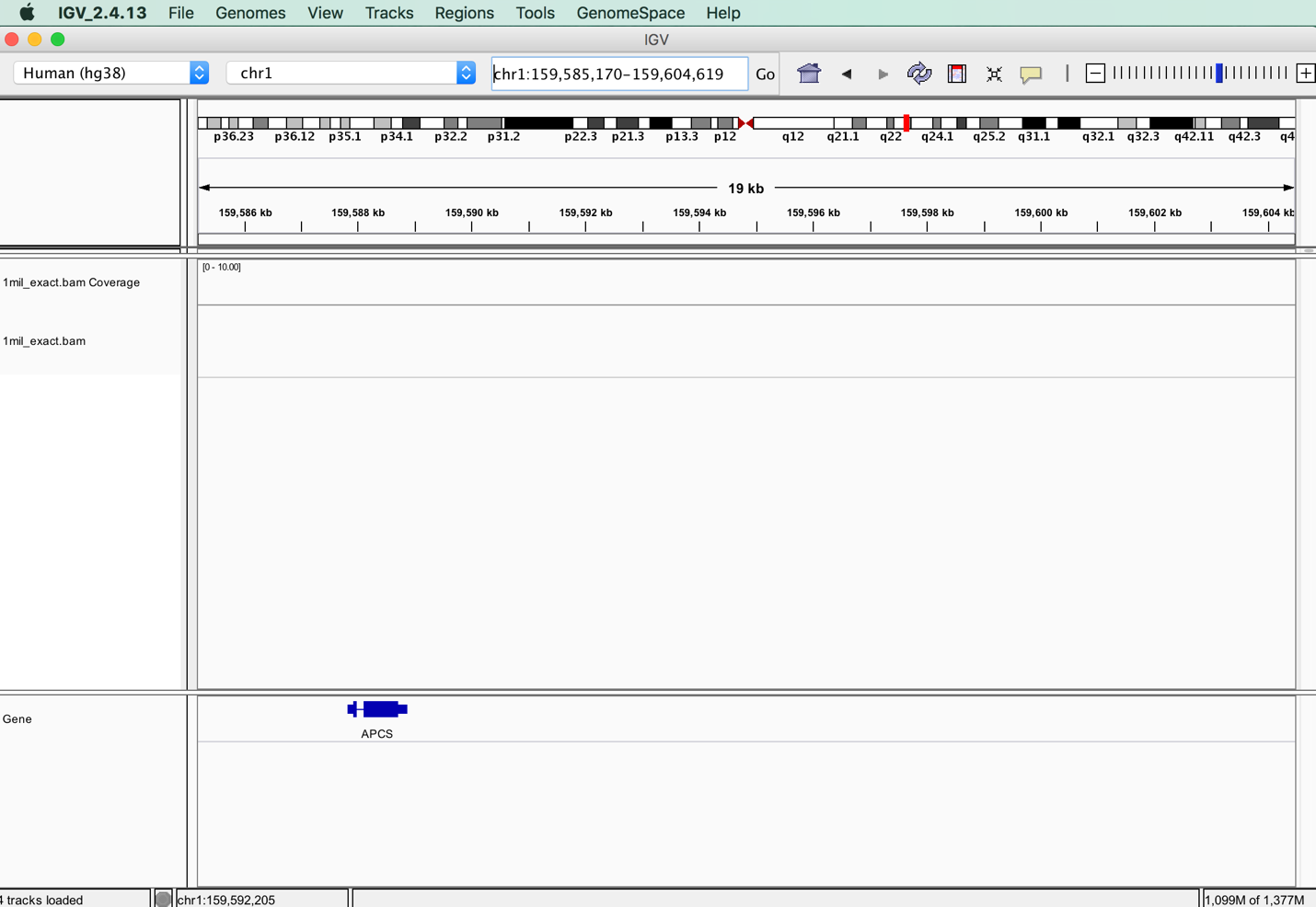
1. Modules (fastqc, multiqc, bowtie2, samtools)
   1. Use “module load” command
   2. “module avail” to see list of all modules
   3. “module spider” to do text matching on module name
2. Running jobs on Biowulf – all jobs on Biowulf must be run as **interactive, batch or swarm! This is extremely important! Do not run jobs on the Biowulf login node.**





FastQC report



****

**Downloading RNA Seq data from the Human Protein Atlas**

1. Go to The Human Protein Atlas (proteinatlas.org)
2. Click on MENU and then click on DOWNLOADABLE DATA
3. Scroll down to number 4. RNA Gene data
4. Click on “RNA Sequencing Data for Human Tissue”
5. This takes you to the “Array Express” web site (<https://www.ebi.ac.uk/arrayexpress>)
6. You should see “E-MTAB-2836 – RNA-seq of coding RNA from tissue samples of 122 human individuals representing 32 different tissues”
7. Click on “Click for detailed sample information and links to data”
8. Click on “Display full-sample data table”
9. Scroll across to see that this is PAIRED, Transcriptomic, RNA seq data, and ftp address
10. We are going to retrieve some of these data using the “wget” command on Biowulf
11. Copy the ftp address of the first line – lung\_4a
12. At the command line (on Biowulf), type “wget” and then paste in the ftp address

wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR315/ERR315326/ERR315326\_1.fastq.gz

1. to get the first file of lung\_4a data. You will see the file getting retrieved and downloaded.
2. When it has finished downloading and you get your command line back, retrieve the next set of lung data by copying and pasting the ftp address into the command line.

wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR315/ERR315326/ERR315326\_2.fastq.gz

1. Use the “ls” command to verify that the files are there