







1

Accessing Servers

Accessing servers via the command line (ssh and sftp) ssh....secure shell. usage.... \$ ssh username@server_address

A common method for downloading data from a public server (ftp)

ftp....file transfer protocols. usage.... \$ ftp ftp.someaddress.org

Transferring data between computers/servers (scp) scp....secure copy protocol. usage.... \$ scp /path/to/local/file username@hostname:/path/to/remote/file

Run multiple sessions in one window (GNU screen) (to keep processes running on a remote server without a job scheduler) usage.... \$ screen

See Unix cheat sheet and exercises for additional information.





Read vs Sequence	
10 reads, 6 sequences What is the difference between a read and a sequence?	

Quality Control

Assessing read quality

 $\label{eq:Q} \begin{array}{ll} \mbox{Phred quality score: a measure of the quality of base calling:} \\ \mbox{Q} = -10 \mbox{ log(P)} & \mbox{where P is the error probability} \end{array}$

Phred Quality Score	Probability of incorrect base call	Base call accuracy	
10	1 in 10	90%	
20	1 in 100	99%	
30	1 in 1000	99.9%	
40	1 in 10,000	99.99%	
50	1 in 100,000	99.999%	
60	1 in 1,000,000	99.9999%	









Filtering and Trimming Reads

Trimmomatic: a Java program for trimming adapter sequences and low-quality bases from sequencing reads and for filtering out low quality reads.



Table 1	A selection of short-read analysis softwa	re		
Program	Website	Open source?	Handles ABI color space?	Maximum read length
Bowtie	http://bowtie.cbcb.umd.edu	Yes	No	None
BWA	http://maq.sourceforge.net/bwa-man.shtml	Yes	Yes	None
Maq	http://maq.sourceforge.net	Yes	Yes	127
Mosaik	http://bioinformatics.bc.edu/marthlab/Mosaik	No	Yes	None
Novoalign	http://www.novocraft.com	No	No	None
SOAP2	http://soap.genomics.org.cn	No	No	60
ZOOM	http://www.bioinfor.com	No	Yes	240





SAM/BAM Format

SAMtools: a software package for mining NextGen sequencing data after alignment. SAM (Sequence Alignment Map): A widely use format for storing alignment data for highthroughput sequencing reads.

BAM (binary SAM): Compressed SAM (binary format).

The file is broken down into two sections:

- 1. Header section (optional): contains general information about the data such as alignment software used, reference genome aligned againsts, etc. Header lines start with $@. \ensuremath{\mathcal{B}}$
- 2. Alignment section: contains much of the same information as a fastq file, such as sequence and base quality scores, as well as information about alignment to reference sequence.

For a more complete description, see <u>https://genome.sph.umich.edu/wiki/SAM</u>

SAM/BAM Format

11 fields + optional 12th TAGs field (not shown)

Col	Field	Туре	Brief Description	
1	QNAME	String	Query template NAME	
2	FLAG	Int	bitwise FLAG	
3	RNAME	String	References sequence NAME	
4	POS	Int	1- based leftmost mapping POSition	
5	MAPQ	Int	MAPping Quality	
6	CIGAR	String	CIGAR String	
7	RNEXT	String	Ref. name of the mate/next read	
	-			

- 8
 PNEXT
 Int
 Position of the mate/next read

 9
 TLEN
 Int
 observed Template LENgth

 10
 SEQ
 String
 segment SEQuence

 11
 QUAL
 String
 ASCII of Phred-scaled base QUALity+33



