SAMTOOLS

You will need to know:

* What a sam/bam file is
* Less/head/tail/grep
* File paths

Data files (found in the data folder):

* smo.bam
* arm.bam
* fl1-1.bam
* f11-2.bam
* header.sam
* dnaSeq1.bam
* dnaSeq2.bam
* intervalFile.bed

Website:

<http://samtools.sourceforge.net/>

<http://samtools.sourceforge.net/samtools-c.shtml>

<http://samtools.sourceforge.net/samtools.shtml>

samtools is a set of scripts (a toolbox so to say) that can be used to manipulate and view sam/bam files. In particular you can: sort, index, merge, view these files.

**Getting help**

To print a list of all the tools available in the samtools suite, simply type samtools on the command line

samtools

To print a list of the parameters required and options available with each tool in the suite, simply type samtools followed by the name of the tool on the command line. For example, to get a list of the options available for the “view” tool in the samtools suite, simply type:

samtools view

**View bam files:**

smo.bam is a bam file and is not human readable. To make it human readable you can convert it to a sam file.

samtools view –h –o smo.sam smo.bam

*What does –h and –o do in the above example?*

*Convert arm.bam into a sam file.*

**Viewing the header of a bam file**

In some cases you might only want to see or generate the header of a bam file.

samtools view –H smo.bam

Now try:

samtools view –H smo.bam > smo.header.sam

*What does the “> smo.header.sam” of the above statement do?*

*What information is stored in the header of the sam file?*

*From the header of the file, can you tell which alignment program was used to generate the bam file*

*TO DO: Ask more specific question about header in file e.g. what alignment program was used to generate alignment*

*TO DO: View a specific region of a file*

**Count the number of alignments in a bam file**

samtools view –c smo.bam

*Can you use samtools to count the number of alignments above a quality score of 20 in your file?*

*What does the quality score of an alignment indicate?*

*How many alignments are in the arm.bam above a quality score of 50*

*TODO: other flags/filtering options*

**Create a bam index**

Use samtools index to create an index of smo.bam

*What does creating an bam index mean? Why would one want to create a bam index?*

**Sorting the sam file**

samtools –o sort smo.bam

*What does –o in the above command do?*

*Change the command to sort by read names rather than chromosomal locations.*

*Change the above command so that the sorted reads are outputted to smo.sorted.bam*

**Merging sam files**

Fl1-1.bam and FL1-2.bam are two technical replicates of the one control sample FL1. We would like to merge these two sam files.

samtools merge –h header.sam FL1-merged.bam FL1-1.bam FL1-2.bam

*What does –h in the above command do?*

*Does this command work? Why not? (TODO: The files are sorted wrongly). Change the command so that the files can be merged.*

**Get summary statistics**

Use “samtools idxstats” to get summary statistics for the aligned file

**Create a fasta file index:**

samtools faidx dros\_BD5.25.fa

*What is the benefit of creating an index of a fasta file?*

**Create a pileup**

samtools mpileup is a very useful utility for calling variants in alignment files. Read the help documentation carefully.

samtools mpileup –g –l intervalFile.bed –I –D –q 20 **-f** *dros\_BD5.25.fa dnaSeq1.bam dnaSeq2.bam*

Note: dros\_BD5.25.fa needs to be indexed

**Getting bored? Good to know:**

**Working with the stream**

You can take the output of one command from the “standard stream” and pipe it as input to an(other) samtools command.

samtools view –u dnaSeq1.bam chr1 | samtools pileup –cf dros\_BD5.25.fa -

Use samtools and awk to count the number of mapped reads in your file.

samtools idxstats smo.bam | awk '{s+=$3} END {print s}'

*Is this the same number as with*

*samtools view –c smo.bam*

*What is the awk command doing in the above?*