Samtools on HPC Clusters:

What is Samtools?

Samtools is a collection of programs for interacting with high-throughput sequencing data. It is made up of three additional repositories:

- Samtools: for reading, writing, editing, and viewing SAM/BAM/CRAM format.
- BCFtools: reading/writing BCF2/VCF/gVCF file and calling/filtering/summarizing SNP (Single Nucleotide Polymorphism) and short indel sequence data
- HTSlibt: A C library for reading/writing high-throughput sequencing data

The documentation for the Samtools is found on:

https://www.htslib.org/

https://github.com/samtools/

Versions Available:

Samtools – 1.10, 1.9, 1.6, (1.2) **Module unavailable, code is very old.

How to load a version of Samtools?

To view readily available builds of Samtools on the HPC on current terminal session, use the following command:

module avail bio/samtools

The version will be listed. Select one which fits your project. For this example, I am loading:

module load bio/samtools /1.10

Verify by using this command:

module list

This will show all modules loaded and all the dependencies required to run.

How to use Samtools on the cluster?

Here is one example of how to use the Samtools on the cluster. To use simply execute a simple task

```
cd ~ && mkdir samtools_test # Make a test directory on home folder


cd samtools


wget http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwRepliSeq/wgEncodeUwRepliSeqBg02esG1bAlnRep1.bam          #download data


mv * test.bam                    # renames the .bam file to test file

# Now after loading module and downloaded test files, we can make a script to run Sam tools

# You can see the plots by using the display command
```

After executing the above command, we need to create a bash script which will submit the job to the scheduler.

How to Run Interactively:

Most of the easy tasks in samtools such as viewing and writing some file types can be done by running a bash session on compute node. Note: Do not use the head node!

```
#To go to a bash session in a compute node use the following command

srun -J test_session -n 1  -p main --qos main --mem 5G --pty bash


#Now we can load the module samtools and do the light tasks on the session.

#For this example, I am going to view a .bam file
```

```
samtools view test.bam

#Or we could store the status of a bam file

samtools stats test.bam > stats.txt        # save the status of file into a text file
```

Note to the user: You can also use commands form bcftools, hstlib, bgzip, tabix in addition to the samtools.

The Script:

```
#!/bin/bash
$SBATCH -J Jobname # Jobname
#SBATCH -n 1       # Nodes per task
#SBATCH -p main    # Partition
#SBATCH --qos main  # Quality of service
#SBATCH -o Samtools_sim_out-%J.txt  # STDOUT Out file
#SBATCH -e Samtools_sim_out-%J.txt  # Error file file



cd $SLURM_SUBMIT_DIR                #go to the submit directory

module load bio/samtools/1.10       # load the module

samtools view test.bam > bam_file.txt  #writes bam file on the text file
samtools stats test.bam > stats.txt    #writes the stats of the bam file to text file
mkdir plots_for_bam_file              #make directory to store plots
plot-bamstats -p /plots_for_bam_file/ stats.txt  #generate plots about the test.bam
#status
```

Now, you can schedule the job with sbatch command.

sbatch myscript.sh

You can use GNU parallel package to run samtools in parallel to process big files parallelly. This can substantially reduce the runtime.

See this for resources:

http://zvfak.blogspot.com/2012/02/samtools-in-parallel.html

Where to find help?

If you are stuck on some part or need help at any point, please contact us at the following address.

https://ua-app01.ua.edu/researchComputingPortal/public/oitHelp

Resources:

https://github.com/alekseyzimin/samtools