This short tutorial is meant to work through the five scripts starting in this Github repository using three samples as an example. It is not an extensive tutorial but will go through the steps starting with cleaning raw NGS data to obtaining a concatenated SNP alignment for phylogenetic reconstructions. This tutorial has been developed and validated for unix using an Ubuntu operating system. Some basic unix commands are expected to be known.

**Step 1:** Download raw fastq-files for PRJNA506987, PRJNA506987, and PRJNA506987 from Short Read Archive (<https://www.ncbi.nlm.nih.gov/sra>) and save in folder called raw\_data\_tutorial.

**Step 2:** Using script “1\_Cleaned\_Data.sh” clean the raw data.

>>> Open 1\_Cleaned\_Data.sh and edit the two paths to the raw data and the location of NGS\_QC\_Toolkit. If needed edit the file extensions for the raw data (e.g. \_R1.fastq.gz and \_R2.fastq.gz, if following this tutorial there is no need). Save the script in the folder you want to execute it in. Make the script executable.

chmod +x 1\_Cleaned\_Data.sh

>>> Make a list of your sample names and save in text file called “samples.txt” (if following this tutorial you can used the supplied file) in the folder where you want to execute the 1\_Cleaned\_Data.sh script. The file can be made automatically by executing the following command in the raw\_data\_tutorial golder

ls \*.gz | sed 's/\_\_R[1|2].fastq.gz//g' | sort | uniq > /folder\_to\_save\_in/samples.txt

>>> Run the 1\_Cleaned\_Data.sh script

./1\_Cleaned\_Data.sh

>>> The cleaned read file will be in the folder “cleaned\_reads”

**Step 3:** Download the reference genome (in this tutorial we will use the olive reference genome Oe6, which can be found here <http://denovo.cnag.cat/genomes/olive/>). Save the downloaded fasta-file in a folder olive\_refgenome\_Oe6

**Step 4:** Index the reference genome with bowtie2 and SAMtools.

>>> Execute the following commands in the folder where you saved the reference genome-wide

bowtie2-build Oe6.scaffolds.fa Oe6.scaffolds.fa

samtools faidx Oe6.scaffolds.fa

**Step 5:** Map the cleaned reads to the reference genome using script “2\_bowtie\_map\_to\_a\_reference.sh”

>>> Open 2\_bowtie\_map\_to\_a\_reference.sh with a text editor and edit the path given to the reference and the location of the cleaned data. If needed edit the file extension of the cleaned data (eg. \_R1.fastq.gz and \_R2.fastq.gz; if following this tutorial no need to change) as well as give the correct name of the reference (Oe6.scaffolds.fa; if following this tutorial no need to change). Save the script in the folder you want to execute it in.

>>> Make the script executable

chmod +x 2\_bowtie\_map\_to\_a\_reference.sh

>>> make sure your sample.txt file from above is also in the same folder

>>> Execute the script

./2\_bowtie\_map\_to\_a\_reference.sh

>>> The raw mapping files will be in the folder “map”

>>> The filtered mapping files will be in the folder “map/1\_proper\_pairs”

>>> The “raw” SNP calls will be in the folder “map/2\_SNPs”

>>> At this point it is possible to delete the raw mapping files as all further steps are using the reads mapped in proper pairs.

**Step 6:** Merge the raw vcf-file, re-call all SNPs in all the samples, and filter the SNPs using the script “3\_merge\_vcf\_covfilt\_SNPcall.sh”.

>>> Open the script and change the paths to the reference genome and the data. Change file extensions and reference genome as needed (no need to do this if following this tutorial). Save the script and make it executable.

chmod +x 3\_merge\_vcf\_covfilt\_SNPcall.sh

>>> Make sure the file samples.txt is in the folder you want to execute the script in.

>>> Execute the script

./3\_merge\_vcf\_covfilt\_SNPcall.sh

>>> File “Oe6\_coverag\_counts\_allcov\_50per.txt” contains information about coverage etc. The columns are as follows: total\_counts; number\_of\_SNP; mean\_coverage; median\_coverage; min\_coverage; max\_coverage.

>>> Add sample names to the coverage file

paste samples.txt Oe6\_coverag\_counts\_allcov\_50per.txt > temp && mv temp Oe6\_coverag\_counts\_allcov\_50per.txt

>>> Open Oe6\_coverag\_counts\_allcov\_50per.txt in for example Excel/LibreCalc (columns separated by TAB) and calculate 0.5 and 2 times the median coverage for each sample. Sort the file according to 0.5\*median and 2\*median in ascending order.

>>> For each category of coverages (eg. median\_1-4, median\_1-6) make new “sample files”, eg. Oe6\_med1-4.txt and Oe6\_med1-4.txt. For this tutorial there might just be one sample per category as we are only dealing with three samples in total. Save these files in the same folder as you have executed the script in.

**Step 7:** Filter the SNP based on individual coverages, merge the filtered SNPs and then filter the merged file only keeping SNPs which have less than 80% missing data and are present in at least 2 samples using the script “4\_merge\_vcf\_covfilt\_filt.sh”.

>>> Open the script and change the paths to the reference genome and the data. Change file extensions as needed (no need to do this if following this tutorial). Remove or add filter categories as needed. Save the script and make it executable.

chmod +x 4\_merge\_vcf\_covfilt\_filt.sh

>>> IF playing around with amount of missing data and minor allele counts change the following settings in Step 3 in the script:

--max-missing 0.2

--mac 3

>>> Execute the script

./4\_merge\_vcf\_covfilt\_filt.sh

>>> File “merged\_Oe6\_covfilt\_allcov\_80per\_mac3.bed” contains position information of all SNPs which will be re-called bioinformatically by counting alleles at each position in the mapping files using the next script.

Step 8: Bioinformatically re-call all SNPs using allelic information from mapped reads using script “5\_Gsk\_to\_GT.sh”

>>> Open the script and change the paths to the bam-files, reference genome, bed file (the one above), sample file, and accessory scripts. Change file extensions and reference genome as needed (no need to do this if following this tutorial). Save the script and make it executable.

chmod +x 5\_Gsk\_to\_GT.sh

>>> Execute the script

./5\_Gsk\_to\_GT.sh

>>> No intermediate files are saved in by this script. If you want to save intermediate files then hash out (put # in front) all rm commands.

>>> End files are \*.vertical.fasta which are sample files for all re-called SNPs in a “vertical” fasta format.

>>> The last command in this script take the vertical fasta-files and make one multi-fasta file of aligned and concatenated SNP calls. This is the file that can now be used for phylogenetic analyses, given that it is converted to the appropriate file formate for the appropriate phylogenetic programs.