

# Basics in command line handling, with the examples on NGS data analysis

*Ivars Silamikelis, COST Basic training on NGS data analysis, Riga 2015*

## 1 Basic linux commands

Here some basic linux commands will be briefly explained. For information about more commands, feel free to consult with the internet.

### 1.1 Getting around

Command	Explanation
<code>ls</code>	list contents of current directory
<code>ls <i>dirname</i></code>	list contents of directory named <i>dirname</i>
<code>cd <i>dirname</i></code>	change current directory to <i>dirname</i>
<code>cd ~</code>	go to home directory (default directory)
<code>cd ..</code>	go one directory up
<code>pwd</code>	print name of current directory
<code>mkdir <i>dirname</i></code>	make a new directory named <i>dirname</i>
<code>man <i>command</i></code>	show manual for <i>command</i>

### 1.2 File manipulation

Command	Explanation
<code>cp <i>filename1 filename2</i></code>	copy file
<code>cp -r <i>dirname1 dirname2</i></code>	copy directory
<code>mv <i>name1 name2</i></code>	move file or directory (can be used for renaming)
<code>rm <i>filename</i></code>	remove file (you won't be able to recover removed file)
<code>rm -r <i>dirname</i></code>	remove directory and its contents
<code>more <i>filename</i></code>	command for paging through text file one screenful at a time
<code>less <i>filename</i></code>	similar as <code>more</code> . Better for viewing large text files ( <code>q</code> to quit)
<code>cat <i>filename1 filename2 filenameN</i></code>	concatenate text files in print output to terminal window
<code>head <i>filename</i></code>	print first 10 lines of a text file to terminal window
<code>head -n 20 <i>filename</i></code>	print first 20 lines of a text file to terminal window
<code>tail</code>	opposite of <code>head</code>
<code>wc <i>filename</i></code>	show the number of lines, words and characters in a text file
<code>cut -f 2 <i>tabDelimitedFilename</i></code>	extract 2nd column from a tab delimited file
<code>cut -f 3 -d , <i>comaSeperatedFilename</i></code>	extract 3rd column from a coma seperated file
<code>nano <i>filename</i></code>	open <i>filename</i> in text editor <code>nano</code>

## 1.3 Archiving and unarchiving of files

Note that for the `tar` utility, option `c` stands for compress, `x` - for uncompress or extract, `z` - for dealing with *tar.gz*, and `j` - for dealing with *tar.bz2*

### 1.3.1 Compressing

Command	Explanation
<code>tar -cvf filename.tar filename</code>	<b>compress</b> file to <b>.tar</b> format
<code>tar -zcvf filename.tar.gz filename</code>	<b>compress</b> file to <b>.tar.gz</b> format
<code>tar -jcvf filename.tar.bz2 filename</code>	<b>compress</b> file to <b>.tar.bz2</b> format
<code>zip filename.zip filename</code>	<b>compress</b> file to <b>.zip</b> format
<code>gzip filename</code>	<b>compress</b> file to <b>.gz</b> format

### 1.3.2 Uncompressing

Command	Explanation
<code>tar -xvf filename.tar</code>	<b>uncompress</b> from <b>.tar</b> format
<code>tar -zxvf filename.tar.gz</code>	<b>uncompress</b> from <b>.tar.gz</b> format
<code>tar -jxvf filename.tar.bz2</code>	<b>uncompress</b> from <b>.tar.bz2</b> format
<code>unzip filename.zip</code>	<b>uncompress</b> from <b>.zip</b> format
<code>gzip -d file.gz</code>	<b>uncompress</b> from <b>.gz</b> format

## 1.4 Input/Output redirection

Command	Explanation
<code>command &gt; filename</code>	output of <i>command</i> is saved to <i>filename</i> , overwriting it
<code>command &gt;&gt; filename</code>	output of <i>command</i> is appended at the end of <i>filename</i>
<code>command &lt; filename</code>	<i>command</i> reads input from <i>filename</i>
<code>command1   command2</code>	<i>command2</i> takes the output of <i>command1</i> and produces result

## 1.5 Lists of commands

Command	Explanation
<code>command1 ; command2</code>	<i>command2</i> is executed after <i>command1</i>
<code>command1 &amp;&amp; command2</code>	<i>command2</i> is executed if <i>command1</i> was successful
<code>command1    command2</code>	<i>command2</i> is executed if <i>command1</i> has failed

## 1.6 Filters

Command	Explanation
<code>grep text filename</code>	Prints every line in <i>filename</i> containing <i>text</i>
<code>sed 's/red/green/' filename</code>	Prints every line in <i>filename</i> substituting word <i>red</i> with word <i>green</i>

## 1.7 Pattern matching

Pattern	Explanation
<code>*</code>	matches zero or more characters
<code>?</code>	matches one character
<code>~</code>	refers to user's default (home) directory

## 1.8 Miscellaneous

Command	Explanation
<code>echo text</code>	display a line of text
<code>history</code>	view your command line history
<code>wget someWebAddress</code>	download contents of <i>someWebAddress</i> to current directory
<code>make</code>	tool that is used to compile source code creating executables
<code>export name=value</code>	sets <i>value</i> to <i>name</i> . Type <code>echo \$name</code> to view <i>value</i>
<code>source filename</code>	read and execute commands from the <i>filename</i> argument

## 1.9 Syntax

Syntax element	Explanation
<code>\</code>	allows to split command over multiple lines

### 1.10 for loop

for loop allows to iterate over a list of items and apply commands on them:

```
for item in item1 item2 item3 item4; do
    echo $item
done
```

Note that we are assigning to `item` each value in list of items (`item1 item2 item3 item4`) in the first line, and we are retrieving the value of `item` by putting `$` before it.

We can also write for loop in a single line:

```
for item in item1 item2 item3 item4; do echo $item; done
```

## 2 Setup of the working environment

In our home folders there is only one directory named `data/day1`. You will find there *IonTorrent* sequencing reads that we are going to analyze and some other necessary resources.

Let's create two separate directories: one directory where we will perform all analysis for sequencing data, and other - where we will download and compile necessary software:

```
mkdir ngs_work
mkdir programs
```

Since we do not have the administrator's rights on this server, we can't install software on the system. However, we can still install software locally in our home directories. We will create

a special directory where all our executables will be stored. Let's create this directory and name it **binaries**. In linux terminal type:

```
mkdir binaries
```

Let's make this directory special - every executable file we put there, we will be able to easily execute, just by typing the executable's file name from anywhere in the system. To achieve this, we will add `~/binaries` folder to the `$PATH` system's environment variable by editing a text file named `.bashrc` using text editor **nano**. In linux terminal type:

```
nano ~/.bashrc
```

Text editor will open the file and you can edit it. Navigate to the bottom of the text file using arrows and type in following text:

```
export PATH=~/binaries:$PATH
```

Hit **Ctrl o** and **Enter** to save file and **Ctrl x** to exit. Reload `.bashrc`:

```
source ~/.bashrc
```

`$PATH` variable contains a list of directories that the system will look in, when we are entering a command. To view contents of `$PATH`, type in terminal:

```
echo $PATH
```

To install software we will simply have to copy executable to our special directory **binaries**.

Many of the open source tools are deposited in the <https://github.com> repository. To download software from <https://github.com> easily, we will use a tool called **git**. **git** is already preinstalled on our servers, however, on your own Ubuntu servers you can install it by typing:

```
sudo apt-get install git
```

### 3 *De-novo* assembly of sequenced reads

#### 3.1 Installation of *de-novo* assembler

For *de-novo* assembly we will use **mira** assembler. You can download it from <http://sourceforge.net/projects/mira-assembler/>. Click on **Files** → **MIRA** → **stable**. Rightclick on `mira_4.0.2_linux-gnu_x86_64_static.tar.bz2` and **Copy link address**. To download it on our linux server, we will use command **wget**. In linux terminal type:

```
cd ~/programs
```

```
wget -O mira.tar.bz2
```

paste the copied location and hit **Enter**. The program will be downloaded in our `~/programs` directory. The downloaded software is archived in `.tar.bz2` format, therefore we need to extract it from archive. To extract it from archive, type in terminal:

```
tar -jxvf mira.tar.bz2
```

A new folder named `mira_4.0.2_linux-gnu_x86_64_static` will appear. This is our extracted software. MIRA is already precompiled for us, so we just need to find the compiled executable files in the folder and install them copy them to our binaries:

```
cd mira_4.0.2_linux-gnu_x86_64_static
cd bin
cp mira ~/binaries
cd ~
```

Check if MIRA was installed successfully:

```
mira
```

If you see

The program 'mira' is currently not installed.

then reload `.bashrc` again:

```
source ~/.bashrc
```

and check the installation again.

The sequenced reads we obtain from sequencing platform usually are in *FASTQ* format. Each record in *FASTQ* file consists of four entries:

1. read ID, beginning with symbol @
2. DNA sequence of read
3. symbol +
4. *ASCII* encoded quality of each nucleotide in read

Let's create a separate directory for our *de-novo* assembly project and copy the reads in it:

```
cd ngs_work
mkdir denovo
cd denovo
cp ~/data/day1/denovo_reads.fastq .
```

MIRA needs a manifest file for performing *de-novo* assembly. We will create a basic manifest file. Our manifest file will consist of 5 entries. From MIRA's manual, these entries are:

- **project** - name of our assemblies project. The project name will be used by MIRA in project's directory naming
- **job** - tells the assembler whether
  1. we want to perform *de-novo* assembly or map reads against reference genome
  2. genomic DNA or transcripts were sequenced
  3. we want accurate (slow) or draft (fast) assembly
- **readgroup** - tells assembler which reads can be pooled together when assembling reads from multiple sequencing technologies
- **data** - tells the assembler where are our reads
- **technology** - tells the assembler what sequencing technology was used for generating reads

To create manifest file, we will use text editor **nano**. To start the text editor, type

```
nano
```

and the editor will open. Now, to create the manifest file, in the text editor type:

```
project=denovo_reads
job=denovo,genome,draft
readgroup
data=denovo_reads.fastq
technology=iontor
```

To save the text file hit **Ctrl o**, enter the name of the file (e.g. **denovo\_reads.mnfst**), hit **Enter** to save and **Ctrl x** to quit **nano**. To launch MIRA, type:

```
mira denovo_reads.mnfst
```

The obtained results are explained in MIRA's manual, and if you wish to gain finer control of some aspects of the assembling process, then, also refer to the MIRA's manual (<http://mira-assembler.sourceforge.net/docs/DefinitiveGuideToMIRA.html>).

## 4 Building variant calling pipeline

We have *IonTorrent* targeted resequencing data from human chromosomes 13., 17. and 20. Our task is to find all nonsynonymous and stop mutations present in the data and to automatize this process by building data analysis pipeline. To accomplish this task we can divide our work in following subtasks:

- Installation of relevant tools
- Obtaining of reference sequences
- Read mapping against reference sequence
- SNP calling

- SNP annotation and filtering of nonsynonymous and stop mutation variants
- Pipeline building

## 4.1 Tool installation

### 4.1.1 Installation of *IonTorrent* mapping software **tmap**

To detect variants present in the data we need to map sequencing reads against reference sequence. For reads generated with *IonTorrent* sequencing platform we will use program **tmap**.

**tmap** and its installation instructions can be found at <https://github.com/iontorrent/TS/tree/master/Analysis/TMAP>. Since we do not have the administrator's rights on this server, we can't install software on the server. However, we can still use it locally. Compilation instructions:

```
cd ~/programs
git clone git://github.com/iontorrent/TMAP.git
cd TMAP
git submodule init
git submodule update
sh autogen.sh
./configure
make
```

If you are installing **tmap** on your own Ubuntu servers, then you probably will encounter some errors about **aclocal** or **git** missing. If you have administrator's rights on your server, you can counter these errors by installing missing components as follows: in your linux terminal type

```
sudo apt-get install git
sudo apt-get install automake
```

Your best bet for countering other errors is just searching for the solution on the internet.

Lets test the program to confirm that it was compiled successfully:

```
./tmap
```

If you see the programs interface, the program was compiled successfully. Move the compiled binary file to our **binaries** folder:

```
cp tmap ~/binaries
```

### 4.1.2 Installation of **samtools** and **bcftools**

We will also need two tools named **samtools** and **bcftools** which are used for manipulation of mapped reads and variation calling. You can obtain these tools from <http://sourceforge.net/projects/samtools/>. Click on Files → **samtools** → 1.2

Rightclick on `samtools-1.2.tar.bz2` and choose `Copy link address`. We will download these tools in directory `programs`:

```
cd ~/programs
wget -O samtools.tar.bz2
```

paste the copied location and hit `Enter`. Repeat this process for `bcftools` - `Copy link address` and download it with `wget`:

```
wget -O bcftools.tar.bz2
```

paste the copied location and hit `Enter`.

The tools are compressed in `tar.bz2` format, so we need to extract them:

```
tar -jxvf samtools.tar.bz2
tar -jxvf bcftools.tar.bz2
```

Note that if we had a lot more files to extract and it would be too time consuming to manually extract them, we could use a `for` loop and pattern matching to extract archives automatically:

```
for archive in *.tar.bz2; do
    tar -jxvf $archive
done
```

We have downloaded and extracted source code of the tools, but to make these tools usable, we need to compile the source code. Source code compiling is performed with command `make`:

```
cd samtools-1.2
make
```

If (hopefully) no errors were encountered, then `samtools` was compiled correctly. Type:

```
./samtools
```

to test the tool. If you see the program's interface, then the program was compiled successfully. Move compiled binary file to a directory where we are storing our compiled software:

```
cp samtools ~/binaries
cd ..
```

Repeat the same process for `bcftools` (go to `bcftools` source code directory, compile it using `make` and copy resulting binary file to `~/binaries`).



## 4.2 Obtaining reference sequences

We will download reference sequences in a separate directory to avoid file cluttering. Let's make a new directory in our `ngs_work` directory named `reseq`, and there we will create a separate folder `ref` for our reference sequences:

```
cd ~
cd ngs_work
mkdir reseq
cd reseq
mkdir ref
cd ref
```

Our reference sequences can be accessed from a database made by University of California, Santa Cruz. The web address of the database is <http://genome.ucsc.edu/>. To find the necessary reference sequences for chromosomes 13., 17. and 20. click on **Downloads** → **human** → **Data set by chromosome**. Right click on `chr13.fa.gz` and choose **Copy link location**. In terminal type

```
wget
```

paste the copied location and hit **Enter**. The reference sequence is compressed in `.gz` format, so we need to extract it:

```
gzip -d chr13.fa.gz
```

File `chr13.fa` will appear in our folder.

Repeat the process for chromosomes 17. and 20.

After we have obtained and extracted our reference sequences, we need to concatenate them. To accomplish this task we will use command `cat`:

```
cat chr13.fa chr17.fa chr20.fa > chr_merged.fa
```

And we have the needed reference for further data processing steps.

## 4.3 Read mapping against reference sequence

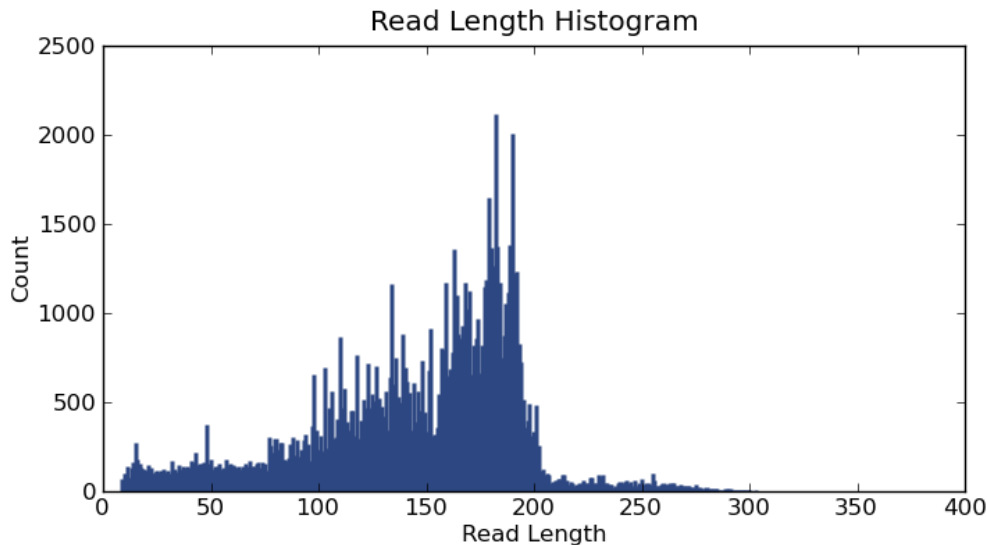
To enable read mapper to efficiently read and process reference sequence, we need to index reference sequence using mapping software's provided indexing function. Often generated indexes are incompatible between different read mappers. As a consequence, almost every read mapping software has their own indexing algorithms.

To perform reference sequence indexing with `tmap` software, type in linux terminal:

```
tmap index -f chr_merged.fa
cd ..
```

To perform the actual read mapping against our reference there are available six different mapping algorithms (see `tmap`'s manual (<https://github.com/iontorrent/TS/blob/master/Analysis/TMAP/doc/tmap-book.pdf>). Some of them are suitable for short reads, some - for longer reads. In the figure below, histogram of read lengths generated by *IonTorrent* server is

shown. By looking at read length distribution of our sample and it seems that most of the reads are longer than 150 bp.



We will choose from one of the algorithms that are suited for longer reads, and one of such algorithms is `map3`. To perform the read mapping we need to:

- copy reads from `~/data/day1` directory to current one
- provide it with reference sequence (`-f ref/chr_merged.fa`)
- provide it with sequencing reads (`-r reseq_reads.fastq`)
- tell that sequencing reads are in *FASTQ* format (`-i fastq`)
- tell that we want alignment to be in compressed *BAM* format (`-o 1`)
- tell that we want output to be saved in a file named `reseq_reads_mapped.bam` (`-s reseq_reads_mapped.bam`)

In linux terminal type:

```
cd ~/ngs_work/reseq
cp ~/data/day1/reseq_reads.fastq .
tmap map3 -f ref/chr_merged.fa \
  -r reseq_reads.fastq \
  -i fastq \
  -o 1 \
  -s reseq_reads_mapped.bam
```

We have obtained *BAM* file. *BAM* stands for **B**inary **S**AM file. *SAM*, in turn, stands for **S**equencing **A**lignment/**M**apping format.

Detailed information about *SAM/BAM* file format, see <http://samtools.github.io/hts-specs/SAMv1.pdf>.

### 4.3.1 BAM file viewing and manipulation

Let's find out what is in the first 20 lines of *BAM* file using `samtools` and `head`:

```
samtools view reseq_reads_mapped.bam | head -n 20
```

To interpret `FLAG` field (2nd column in *BAM* file), you can use `samtools flags` command providing it with the flag you are interested in:

```
samtools flags 4
```

To view *BAM* file header, use `-H` option in `samtools view` command:

```
samtools view -H reseq_reads_mapped.bam
```

As you can see, we have not defined our read group identifier (`ID:NOID`) and sample name (`SM:NOSM`). We can correct it using `samtools` command `reheader`. It takes a *BAM* header and *BAM* file as input. It replaces *BAM* file's header with the provided header and outputs new *BAM* file.

We will read our *BAM* files header and modify it using `sed` utility. The modified header will be given to `samtools reheader` along with the original bam file. The output will be saved using `>` operator:

```
samtools view -H reseq_reads_mapped.bam \
| sed 's/ID:NOID/SM:SomeID/;s/SM:NOSM/SM:Sample1/' \
| samtools reheader - reseq_reads_mapped.bam \
> reseq_reads_mapped.reheaded.bam
```

Alternatively, we could have simply defined the read group ID and sample name during mapping process, using `tmap`'s option `-R`:

```
tmap map3 -f ref/chr_merged.fa \
-r reseq_reads.fastq \
-i fastq \
-R ID:SomeID \
-R SM:Sample1 \
-o 1 \
-s reseq_reads_mapped.reheaded.bam
```

We can use `samtools` command `tvview` to view alignment in a more human readable format. `tvview` accepts only sorted and indexed *BAM* files. Fortunately, we can do this easily using `samtools` commands `sort` and `index`:

```
samtools sort reseq_reads_mapped.reheaded.bam reseq_reads_mapped.reheaded.sorted
samtools index reseq_reads_mapped.reheaded.sorted.bam
```

*BAM* file indexing allows programs to efficiently access any part of the *BAM* file. *BAM* index file names usually end with `.bam.bai`.

After sorting and indexing, we can view our *BAM* file. In linux terminal type:

```
samtools tview reseq_reads_mapped.reheaded.sorted.bam ref/chr_merged.fa
```

Alignment will open in terminal. To get help on viewing alignment, type `?`. Type `q` to quit viewer. A good place to view mapped reads and test the viewer is `chr13:27925825`. To go to this position, type `g` and enter this coordinate as it is shown.

It is also possible to extract some specific region from the *BAM* file with `samtools view` by adding coordinates in form `chrN:start-end` at the end of command:

```
samtools view reseq_reads_mapped.reheaded.sorted.bam chr13:27925800-27925900
```

To get overall *BAM* file statistics and number of mapped reads for each chromosome, you can use `samtools` commands `flagstat` and `idxstats`:

```
samtools flagstat reseq_reads_mapped.reheaded.sorted.bam
samtools idxstats reseq_reads_mapped.reheaded.sorted.bam
```

To use these commands, *BAM* file must be sorted and indexed.

## 4.4 SNP calling

To call SNPs we will use `samtools` and `bcftools`. Using `samtools` we will generate our data in *pileup* format which is then accessed by `bcftools` to call SNPs. *Pileup* format is text based description of mapped bases at each position of reference sequence. You can think of it as a vertical version of alignment viewing. Variant callers need sorted by position *BAM* file to successfully detect variants and we have already done that, so we can move to variant calling.

There are two choices of variant callers in `bcftools`: `--consensus-caller` or `--multiallelic-caller`. `consensus-caller` is older variant calling model and assumes only biallelic sites. It can miss somatic mutations. `multiallelic-caller` does not have such an assumption and is more sensitive than `consensus-caller`. `multiallelic-caller` is recommended for most tasks, and we will also use it. We will use option `--variants-only` to only detect sites where mapped reads differ from reference sequence. Using option `--skip-variants indels` we will skip insertions/deletions. If you do want to see indels, then you can skip this option. To call SNPs only, in linux terminal type:

```
samtools mpileup -uf ref/chr_merged.fa \
    reseq_reads_mapped.reheaded.sorted.bam \
    | bcftools call --multiallelic-caller --variants-only --skip-variants indels \
    > reseq_reads.bcftools_snps.vcf
```

This will generate file named `reseq_reads.bcftools_snps.vcf` which is in *VCF* format. See <http://samtools.github.io/hts-specs/VCFv4.2.pdf> for explanation of *VCF* format.

## 4.5 SNP annotation and filtering with `snpEff` and `SnpSift`

After generating a list of SNPs in *VCF* format, we will use a tool `snpEff` to predict whether they are causing amino acid change and `SnpSift` to filter nonsynonymous and stop gained SNPs.

#### 4.5.1 Setting up and configuring snpEff and SnpSift

snpEff and SnpSift are bundled together and are available at <http://snpeff.sourceforge.net>. To download and install them, type in linux terminal:

```
cd ~/programs
wget http://sourceforge.net/projects/snpeff/files/snpEff_latest_core.zip
```

To unzip snpEff type:

```
unzip snpEff_latest_core.zip
```

Unfortunately we will not be able to launch both programs from `binaries` folder and we can leave the software in `programs` directory.

SnpSift and snpEff are `.jar` files which means that these are java applications. To run java applications through linux terminal, use command `java -jar filename.jar`. snpEff requires databases to predict effects of SNPs. We will use human genome's (version GRCh38.76) database to annotate SNPs and snpEff's database dbNSFP. The needed databases are already downloaded for us and resides in `~/data/day1/snpEffData`. Our task is to configure snpEff to tell it where these databases can be found.

After downloading and unzipping of snpEff we will need to edit snpEff's configuration file named `snpEff.config` and change entry `data.dir = ./data/` to `data.dir = ~/data/day1/snpEffData/`. We will also add line `database.local.dbnsfp = ~/data/day1/snpEffData/dbNSFP2.7.txt.gz` to tell snpEff to look for databases in these locations.

#### 4.5.2 SNP annotation and filtering

To successfully perform SNP annotation, we need to provide snpEff with the command we are executing (`ann` - telling snpEff that we want to annotate SNPs), its configuration file (`-c snpEff.config`), database we are annotating against (GRCh38.76) and our *VCF* file. Since snpEff by default prints output to terminal, we can save it using `>` operator. To annotate called variants, type:

```
cd ~/ngs_work/reseq
java -jar ~/programs/snpEff/snpEff.jar \
  ann \
  -c ~/programs/snpEff/snpEff.config \
  GRCh38.76 \
  reseq_reads.bcftools_snps.vcf \
  > reseq_reads.bcftools_snps.annotated.vcf
```

If you see an error saying `GC overhead limit exceeded` then you will have to give this program more memory. To do this, type:

```
java -jar -Xmx3g ~/programs/snpEff/snpEff.jar \
  ann \
  -c ~/programs/snpEff/snpEff.config \
```

```
GRCh38.76 \  
reseq_reads.bcftools_snps.vcf \  
> reseq_reads.bcftools_snps.annotated.vcf
```

In the resulting *VCF* file **snpeff** adds a new entry in the INFO field - ANN. Explanation of this field can be found in *VCF* file's header. See [http://snpeff.sourceforge.net/SnpEff\\_manual.html](http://snpeff.sourceforge.net/SnpEff_manual.html) for more information about this tool and generated output files.

We can also annotate our *VCF* file by using information from another *VCF* file. To do this, we will use **SnpSift**'s command **annotate**. We need to supply **SnpSift**'s **annotate** with the database *VCF* file from which the program will acquire information and *VCF* file that we are going to annotate. We can choose whether to annotate ID, INFO or both fields. In the example we will annotate only ID fields by giving **-id** option to program. To annotate called variants with **SnpSift**'s **annotate**, type:

```
java -jar ~/programs/snpEff/SnpSift.jar \  
  annotate \  
  ~/data/day1/snpEffData/00-All.vcf \  
  -id \  
  reseq_reads.bcftools_snps.annotated.vcf \  
> reseq_reads.bcftools_snps.annotated.id.vcf
```

**SnpSift** can also give information about SNP's: frequency in 1000 Genomes database, *SIFT*, *Polyphen*, *GERP* scores, etc. To perform this kind of annotation, we will use **SnpSift**'s command **dbnsfp**. To annotate with **SnpSift**'s **dbnsfp**, type:

```
java -jar ~/programs/snpEff/SnpSift.jar \  
  dbnsfp \  
  reseq_reads.bcftools_snps.annotated.id.vcf \  
> reseq_reads.bcftools_snps.annotated.id.dbnsfp.vcf
```

To filter only those SNPs that are marked as nonsynonymous or stop gained we are going to use **SnpSift**'s command **filter**. **SnpSift**'s **filter** command needs *VCF* file annotated by **snpeff** and filtering expression. We will use expression "(ANN[\*].EFFECT = 'missense\_variant') || (ANN[\*].EFFECT = 'stop\_gained')" which can be translated as: if any "ANN" field for this variant, has effect of missense or stop gain, then print this variant to terminal. For more examples on filtering expressions, see <http://snpeff.sourceforge.net/SnpSift.html>. To perform variant filtering, type:

```
java -jar ~/programs/snpEff/SnpSift.jar \  
  filter \  
  -f reseq_reads.bcftools_snps.annotated.id.dbnsfp.vcf \  
  "(ANN[*].EFFECT = 'missense_variant') || (ANN[*].EFFECT = 'stop_gained')" \  
> reseq_reads.bcftools_snps.annotated.id.dbnsfp.nonsyn_stop.vcf
```

A new *VCF* file named **reseq\_reads.bcftools\_snps.annotated.id.dbnsfp.nonsyn\_stop.vcf** will appear, in which there will be only missense and stop gain SNPs.

Now, suppose that we want to extract only chromosome, position, reference and alternate alleles, *SIFT* prediction and genotype from *VCF* file. We can achieve it with **SnpSift**'s

`extractFields` command, which takes a *VCF* file and a list of fields it needs extract. Type in terminal:

```
java -jar ~/programs/snpEff/SnpSift.jar    extractFields \  
    reseq_reads.bcftools_snps.annotated.id.dbnsfp.nonsyn_stop.vcf \  
    CHROM POS ID REF ALT dbNSFP_SIFT_pred GEN[*].GT \  
> reseq_reads.bcftools_snps.annotated.id.dbnsfp.nonsyn_stop.txt
```

A tab delimited file named `reseq_reads.bcftools_snps.annotated.id.dbnsfp.nonsyn_stop.txt` will appear which can be easily viewed in spreadsheet software.

## 4.6 Pipeline building

Writing all these commands by hand is time consuming, tedious and error prone and we need a better way to execute these commands. One way how to solve this problem is to save all our commands in a text file and, when the time comes to analyze our data, copy these commands in the terminal. While this method can save our time on looking in manuals for the correct commands, it is still time consuming to wait for each command to end, so that we can copy the next one in. We need a better way how to automatically execute these commands and here *bash* scripting becomes handy.

*bash* stands for *Bourne-Again Shell* and is one of the most popular choices for performing shell scripting. Whenever we typed a command in linux terminal it was shell that was performing all the commands and giving us output. So we have been using *bash* every time we typed a command.

Starting a *bash* script is really easy - open a text editor, on the first line write `#!/bin/bash`, on the next lines some commands, and save the file. You can tell *bash* to launch script by typing in terminal:

```
bash yourFileName
```

To build SNP calling pipeline we will use exactly this approach - open text editor and copy following lines (note that text starting with `#` is a comment line where you can write anything you want to make the code more readable (except the first line)):

```
#!/bin/bash  
  
#Read mapping  
tmap map3 -f ref/chr_merged.fa \  
    -r reseq_reads.fastq \  
    -i fastq \  
    -R ID:SomeID \  
    -R SM:Sample1 \  
    -o 1 \  
    -s reseq_reads_mapped.reheaded.bam  
  
#Read sorting  
samtools sort reseq_reads_mapped.reheaded.bam \  
reseq_reads_mapped.reheaded.sorted  
  
#SNP calling with bcftools
```

```
samtools mpileup -uf ref/chr_merged.fa \
  reseq_reads_mapped.reheaded.sorted.bam \
  | bcftools call --multiallelic-caller --variants-only --skip-variants indels \
  > reseq_reads.bcftools_snps.vcf
```

```
#SNP annotating with snpEff and SnpSift
java -jar ~/programs/snpEff/snpEff.jar \
  ann \
  -c ~/programs/snpEff/snpEff.config \
  GRCh38.76 \
  reseq_reads.bcftools_snps.vcf \
  > reseq_reads.bcftools_snps.annotated.vcf
```

```
java -jar ~/programs/snpEff/SnpSift.jar \
  annotate \
  -id \
  ~/data/day1/snpEffData/00-All.vcf \
  reseq_reads.bcftools_snps.annotated.vcf \
  > reseq_reads.bcftools_snps.annotated.id.vcf
```

```
java -jar ~/programs/snpEff/SnpSift.jar \
  dbnsfp \
  reseq_reads.bcftools_snps.annotated.id.vcf \
  > reseq_reads.bcftools_snps.annotated.id.dbnsfp.vcf
```

```
#filtering of missense and stop SNPs
java -jar ~/programs/snpEff/SnpSift.jar \
  filter \
  -f reseq_reads.bcftools_snps.annotated.id.dbnsfp.vcf \
  "(ANN[*].EFFECT='missense_variant')||(ANN[*].EFFECT='stop_gained')" \
  > reseq_reads.bcftools_snps.annotated.id.dbnsfp.nonsyn_stop.vcf
```

```
#Table creation
java -jar ~/programs/snpEff/SnpSift.jar
extractFields \
reseq_reads.bcftools_snps.annotated.id.dbnsfp.nonsyn_stop.vcf \
CHROM POS ID REF ALT dbNSFP_SIFT_pred GEN[*].GT \
> reseq_reads.bcftools_snps.annotated.id.dbnsfp.nonsyn_stop.txt
```

Save it as `snp_pipeline.sh` and launch it by typing in terminal (note that we still are in `~/ngs_work/reseq` directory):

```
bash snp_pipeline.sh
```

This is good if we want to repeat this analysis over and over again on this one sample. What if we have a new sample we want to analyse? We could replace filenames of previous sample, to filenames of the new sample by hand. However, there is a better way - we can save our sample name in a variable, and use that variable throughout the script. In *bash* we assign value to a variable with `=` operator, and retrieve value from variable by adding `$` at the beginning of variable's name:



```

#!/bin/bash

#Setting sample name
sample_name=reseq_reads.fastq

#Read mapping
tmap map3 -f ref/chr_merged.fa \
  -r $sample_name \
  -i fastq \
  -R ID:SomeID \
  -R SM:Sample1 \
  -o 1 \
  -s $sample_name.mapped.reheaded.bam

#Read sorting
samtools sort $sample_name.mapped.reheaded.bam \
  $sample_name.mapped.reheaded.sorted

#SNP calling with bcftools
samtools mpileup -uf ref/chr_merged.fa \
  $sample_name.mapped.reheaded.sorted.bam \
  | bcftools call --multiallelic-caller --variants-only --skip-variants indels \
  > $sample_name.bcftools_snps.vcf

#SNP annotating with snpEff and SnpSift
java -jar ~/programs/snpEff/snpEff.jar \
  ann \
  -c ~/programs/snpEff/snpEff.config \
  GRCh38.76 \
  $sample_name.bcftools_snps.vcf \
  > $sample_name.bcftools_snps.annotated.vcf

java -jar ~/programs/snpEff/SnpSift.jar \
  annotate \
  -id \
  ~/data/day1/snpEffData/00-All.vcf \
  $sample_name.bcftools_snps.annotated.vcf \
  > $sample_name.bcftools_snps.annotated.id.vcf

java -jar ~/programs/snpEff/SnpSift.jar \
  dbnsfp \
  $sample_name.bcftools_snps.annotated.id.vcf \
  > $sample_name.bcftools_snps.annotated.id.dbnsfp.vcf

#filtering of missense and stop SNPs
java -jar ~/programs/snpEff/SnpSift.jar \
  filter \
  -f $sample_name.bcftools_snps.annotated.id.dbnsfp.vcf \
  "(ANN[*].EFFECT='missense_variant')||(ANN[*].EFFECT='stop_gained')" \

```

```
> $sample_name.bcftools_snps.annotated.id.dbnsfp.nonsyn_stop.vcf
```

We can go even further - we can provide sample name as an argument to our pipeline. There is a predefined *bash* variable (\$1) that stores everything we supply as an argument for our script:

```
#!/bin/bash

#Setting sample name
sample_name=$1

#Read mapping
tmap map3 -f ref/chr_merged.fa \
  -r $sample_name \
  -i fastq \
  -R ID:SomeID \
  -R SM:Sample1 \
  -o 1 \
  -s $sample_name.mapped.reheaded.bam

#Read sorting
samtools sort $sample_name.mapped.reheaded.bam \
  $sample_name.mapped.reheaded.sorted

#SNP calling with bcftools
samtools mpileup -uf ref/chr_merged.fa \
  $sample_name.mapped.reheaded.sorted.bam \
  | bcftools call --multiallelic-caller --variants-only --skip-variants indels \
  > $sample_name.bcftools_snps.vcf

#SNP annotating with snpEff and SnpSift
java -jar ~/programs/snpEff/snpEff.jar \
  ann \
  -c ~/programs/snpEff/snpEff.config \
  GRCh38.76 \
  $sample_name.bcftools_snps.vcf \
  > $sample_name.bcftools_snps.annotated.vcf

java -jar ~/programs/snpEff/SnpSift.jar \
  annotate \
  -id \
  ~/data/day1/snpEffData/00-All.vcf \
  $sample_name.bcftools_snps.annotated.vcf \
  > $sample_name.bcftools_snps.annotated.id.vcf

java -jar ~/programs/snpEff/SnpSift.jar \
  dbnsfp \
  $sample_name.bcftools_snps.annotated.id.vcf \
  > $sample_name.bcftools_snps.annotated.id.dbnsfp.vcf

#filtering of missense and stop SNPs
```

```
java -jar ~/programs/snpEff/SnpSift.jar \
  filter \
  -f $sample_name.bcftools_snps.annotated.id.dbnsfp.vcf \
  "(ANN[*].EFFECT='missense_variant')||(ANN[*].EFFECT='stop_gained')" \
  > $sample_name.bcftools_snps.annotated.id.dbnsfp.nonsyn_stop.vcf
```

#Table creation

```
java -jar ~/programs/snpEff/SnpSift.jar
extractFields \
$sample_name.bcftools_snps.annotated.id.dbnsfp.nonsyn_stop.vcf \
CHROM POS ID REF ALT dbNSFP_SIFT_pred GEN[*].GT \
> $sample_name.bcftools_snps.annotated.id.dbnsfp.nonsyn_stop.txt
```

To launch it, type:

```
bash snp_pipeline.sh reseq_reads.fastq
```

Suppose that we have multiple *FASTQ* files that we want to analyze. Using pattern matching and `for` loop, we can automatically process them with our pipeline:

```
for fastqFile in *.fastq; do bash snp_pipeline.sh $fastqFile; done
```