Variant calling of E. coli using illumina paired end read data

Date: 16/08/2020

Conducted by Pawan Verma

Code compilation by Maruf Ahmed Bhuiyan

```
# setup the initial conda environment
##########################################
##########################################
####-PACKAGE INSTALLATION INSTRUCTIONS-#####
##########-Variant Calling-################
##########################################
##########################################

NOTE: We will be installing all packages in a conda environment

#-----For those of you who have conda installed, need not follow the conda
installation instructions-----#

#--In your home directory, enter--#

Command 1: wget https://repo.anaconda.com/miniconda/Miniconda3-latest-
Linux-x86_64.sh

Command 2: sh Miniconda3-latest-Linux-x86_64.sh

#---Follow on-screen instructions until the installation is complete---#

**NOTE: When asked to add conda_init , enter YES**

##----Add conda to PATH environment---#

Command 3: source ~/.bashrc

##---If the installation is successful, you should see a list of installed
packages with---#
```

```
Command 4: conda list

#---If the command cannot be found,add conda to PATH environment, open the
.bashrc file and add the export PATH command to the end of the file and
save it---#

Command 5: sudo nano ~/.bashrc

#------Paste the below command at the end of the bashrc file, and save
using CTRL+o and ENTER and CTRL+X to exit

export PATH=~/miniconda3/bin:$PATH

#--- To check if it was installed correctly----#

Command 6: conda -V

#Adding the required channels to conda for seamless installation

Command 7: conda config --add channels defaults
Command 8: conda config --add channels bioconda
Command 9: conda config --add channels conda-forge



#-----For those of you who have java installed, need not follow the java
installation instructions-----#

#---To check if java is installed---#

Command 10: java --version

#---if command not found then---#

Command 11: sudo apt-get install default-jre
Command 12: sudo apt-get install default-jdk


############### PACKAGE INSTALLATION FOR VARINAT CALLING ##################

### For ease of package management, create a new conda environment ###

Command 13: conda create -n <your-env name>
Command 14: conda activate <your-env-name>
```

```
#### PACKAGE 1: SRA Tools #####

Command 15: conda install sra-tools

#### PACKAGE 2: fastqc #####

Command 16: conda install fastqc

#### PACKAGE 3: MultiQC #####

Command 17: conda install multiqc

#### PACKAGE 4: Bowtie2 #####

Command 18: conda install bowtie2

#### PACKAGE 5: SAMTools #####

Command 19: conda install samtools

#### PACKAGE 6: Integrative Genome Viewer: IGV #####

Command 20: conda install igv

#### PACKAGE 7: BCFTools #####

Command 21: conda install bcftools




################################################################################
############################
################################################################################
############################
################################################################################
############################

# After setting up the initail environment run the following commands in
terminal
# Just copy pase the following command in terminal to follow

# unzip the reference genome
# we are using this reference genome
```

```
https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/005/845/GCF_000005845.2_AS
M584v2/GCF_000005845.2_ASM584v2_genomic.fna.gz

# we can download it using browser or the command line

# curl
https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/005/845/GCF_000005845.2_AS
M584v2/GCF_000005845.2_ASM584v2_genomic.fna.gz

or

# wget
https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/005/845/GCF_000005845.2_AS
M584v2/GCF_000005845.2_ASM584v2_genomic.fna.gz

gunzip GCF_000005845.2_ASM584v2_genomic.fna.gz

# Download the SRR file using prefetch
prefetch SRR11866736

# let's see if any contaminant plasmids are there
grep '^>' GCF_000005845.2_ASM584v2_genomic.fna

# split the paired read files from the SRR file
fastq-dump --gzip  --defline-qual '+' --split-files SRR11866736

# create a fastqc report
fastqc *fastq.gz

# generate a multiqc report from the existing 2 fastqc html files
multiqc .

# let's open the report in browser
open multiqc_report.html

# let's index the reference genome for easier alignment
bowtie2-build GCF_000005845.2_ASM584v2_genomic.fna ecoli_k12

# lets's do alignment with the reference genome and create sam file
# this is a unsorted file
bowtie2 -x ecoli_k12 -1 SRR11866736_1.fastq.gz -2 SRR11866736_2.fastq.gz -S
SRR11866736.sam

# let's compress it into a bam file
samtools view -b -o SRR11866736.bam SRR11866736.sam
```

```
# sam file is very large. we won't need it again. so let's delete it
rm SRR11866736.sam

# now let's sort it
samtools sort -o SRR11866736.sort.bam SRR11866736.bam

# let's index the bam file now
samtools view SRR11866736.sort.bam > SRR11866736.sort.bam.bai

# this is a very long file. so let's look at the file using less commmand
samtools view SRR11866736.sort.bam | less

# let's now identify the variant calls and store it in a VCF file
bcftools mpileup -Ou -f GCF_000005845.2_ASM584v2_genomic.fna -o
SRR11866736.pileup.bcf SRR11866736.sort.bam

# now let's see the create a variant dile in binary format
bcftools call -m -v -Ou -o SRR11866736.call.bcf SRR11866736.pileup.bcf

# let's remove the duplicates
bcftools norm -Ou -f GCF_000005845.2_ASM584v2_genomic.fna -d all -o
SRR11866736.norm.bcf SRR11866736.call.bcf

# Let's filter it for various parameters
bcftools filter -Ob -e 'QUAL<40 || DP<10 || GT!="1/1"' -o
SRR11866736.variants.bcf SRR11866736.norm.bcf

# let's convert this file to a non-binary file & and replace the referece
genome version
bcftools view -Ov SRR11866736.variants.bcf | sed
's/NC_000913.3/NC_000913/g' > SRR11866736.variants.vcf

# open the file in IGV
```