ASSEMBLY GUIDE AND VARIABLE CALLING FROM RAD SEQUENCES

Mauricio Peñuela mauricio.penuela@hotmail.com

1) For this practice, download the reference genome of *Oryza sativa* IRGSP-1.0 from this link: https://rapdb.dna.affrc.go.jp/download/irgsp1.html
2) We will use the fastq files available for download in the attached document.
3) Create a folder called reference and unzip the reference genome using the command:
   ```
   gunzip referencegenome
   ```
4) Create a folder called samples and introduce the downloaded fastq files.
5) Concatenate the files from the same sample using the command:
   ```
   cat file1 file2 file3 > newname.fastq.gz
   ```
6) Install the program bwa:
   ```
   sudo apt-get install bwa
   ```
7) Index the reference genome using the command:
   ```
   bwa index referencegenome
   ```
   This will create five documents necessary to the align process.
8) Align your fastq files to the reference genome using the command:
   ```
   bwa mem –t 12 referencegenome file.fastq.gz > newname.sam
   ```
   -t 12 corresponds to the number of threads we want to use for this action, will change according to your computer. Use the extension .sam for the output file. If you want to run this command for several files at the same time, we recommend create folders and use the loop:
   ```
   for samples in ./samples/*.fastq.gz ; do bwa mem –t 12
   ./reference/referencegenome $samples > $samples.sam ; done
   ```
9) Install the program samtools:
   ```
   sudo apt-get install samtools
   ```
10) Transform your .sam file to binary format, sort it and generate a new file with extension .bam:
    ```
    samtools view -b file.sam |samtools sort -t 12 -O bam -T tmp_ -o
    file.bam
    ```
    To run this command for several files at the same time, use the loop:
    ```
    for samples in *.sam ; do samtools view –b $samples | samtools sort –
    threads 12 > $samples.bam ; done
    ```
11) Index your .bam files using the command:
    ```
    samtools index file.bam
    ```
    This creates a .bai file which helps to the next steps. To do it for several files, use the loop:
    ```
    for samples in *.bam ; do samtools index $samples ; done
    ```
    Create a folder called aligned and move both .bam and .bai files to there.
12) Install the program bcftools
    ```
    sudo apt-get install bcftools
    ```
13) Call variants from your aligned files
    ```
    Bcftools mpileup --per-sample-mF –annotate
    FORMAT/AD,FORMAT/ADF,FORMAT/ADR,FORMAT/SP,INFO/AD,INFO/ADF,INFO/ADR
    FORMAT/DP -Ou -f ./reference/referencegenome ./aligned/*.bam |
    bcftools call -vmO z -o estudy.vcf.gz
    ```

Unzip the file, and you will found a .vcf file where are all the variants and tags for each sample of our study. Create a folder called vcf and move this file to there.

14) Install the program vcftools

```
sudo apt-get install vcftools
```

15) To filter this information and work in a more organized way use:

```
vcftools --vcf estudio.vcf --remove-indels --max-missing 1 --max-alleles 2 --minQ 30 --minDP 6 --maf 0.05 --recode --recode-INFO-all --out snpsq30dp6
```

This command removes indels and missing data, call maximum two alleles per locus with a Phred score quality of 30, a deep of 6, and a minor allele frequency of 0.05.

16) If you wish to know the number of homozygous and heterozygous markers use:

```
vcftools --vcf snpsq30dp6.vcf –het
```

17) To sort the information for analysis in R or another program use:

```
bcftools query  -H -f '%CHROM %POS [\t%TGT]\n' snpsq30dp6.vcf > R.vcf
```