This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n° 289592

# Standardization, integration and easy analysis of genomic data

Ezequiel L. Nicolazzi
Fondazione Parco Tecnologico Padano
ezequiel.nicolazzi@ptp.it
@KATAphone

# Outline

This supplementary material refers to a presentation given during the Gene2farm Winter School 2015, held in Piacenza (Italy), Nov 9th 2015.

All the software, help files and example data are available at the links provided here. Although this examples runs on a Linux Operative System (Ubuntu), it is compatible to any computer with Linux/Unix and Mac OSs.

Please note that, except for the “reduction” of data (e.g. small number of markers and individuals), this is a pseudo-real case scenario.

# The software used

All the software shown here is completely open-source, meaning you have access to the code.

Written in python 2.7 and for G2F/GenHome purposes, but extended to the general public.

Full information, including manuals, can be found at:

https://github.com/nicolazzie/SNPConvert.git (software and reference manual)

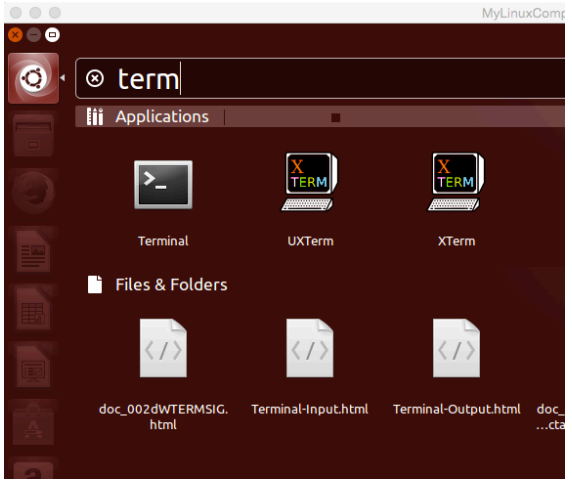
And

https://github.com/bioinformatics-ptp/Zanardi.git (software)
https://github.com/bioinformatics-ptp/Zanardi/wiki (wiki and full documentation)

# First: get the data

To obtain the data, use GIT, a version control software that is commonly used by developers working in teams.

1. Open a terminal

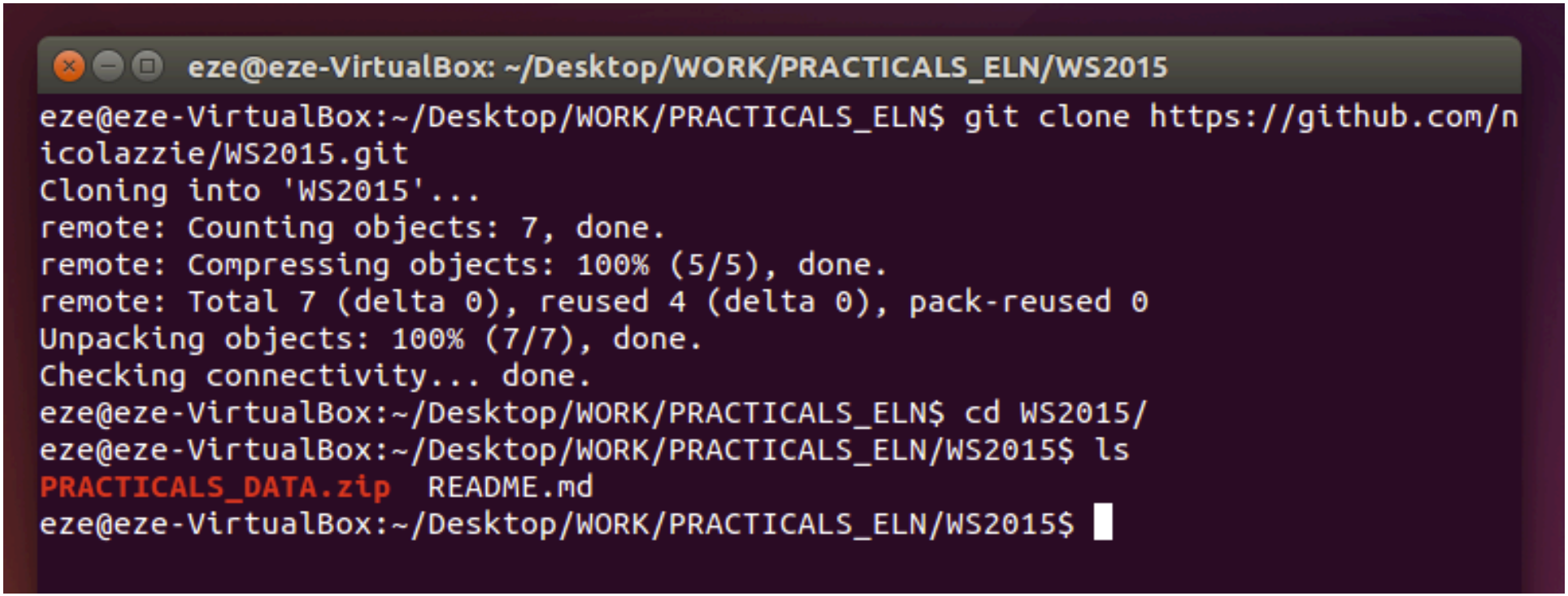



2. Go to Desktop folder, then go to WORK folder (previously created), then create a directory named “PRACTICALS_ELN” and go into it.

```
eze@eze-VirtualBox: ~/Desktop/WORK/PRACTICALS_ELN
eze@eze-VirtualBox:~$ cd Desktop/
eze@eze-VirtualBox:~/Desktop$ cd WORK
eze@eze-VirtualBox:~/Desktop/WORK$ mkdir PRACTICALS_ELN
eze@eze-VirtualBox:~/Desktop/WORK$ cd PRACTICALS_ELN/
eze@eze-VirtualBox:~/Desktop/WORK/PRACTICALS_ELN$
```

# First: Let's get the data

```
git clone https://github.com/nicolazzie/WS2015.git
```




**Once downloaded, go to the WS2015 folder that contains 2 files. The RED one is the one with the data.**

```
unzip PRACTICALS_DATA.zip
```

# CASE SCENARIO

The user is asked to standardize, impute and analyse data from 3 projects that were analyzed by 3 different groups using 3 different chips (e.g. different densities).
The data is in 3 different formats and with 3 different separators:

- Illumina ROW format, allele coding FORWARD or TOP, separator is “tab”
- Illumina MATRIX format, allele coding A/B, separator is “ comma”
- PLINK (ped/map) file, allele coding A/B, separator is “space”

**All data should be standardized, merged and analysed using TOP allele coding.**

# Files in PRACTICALS_DATA.zip

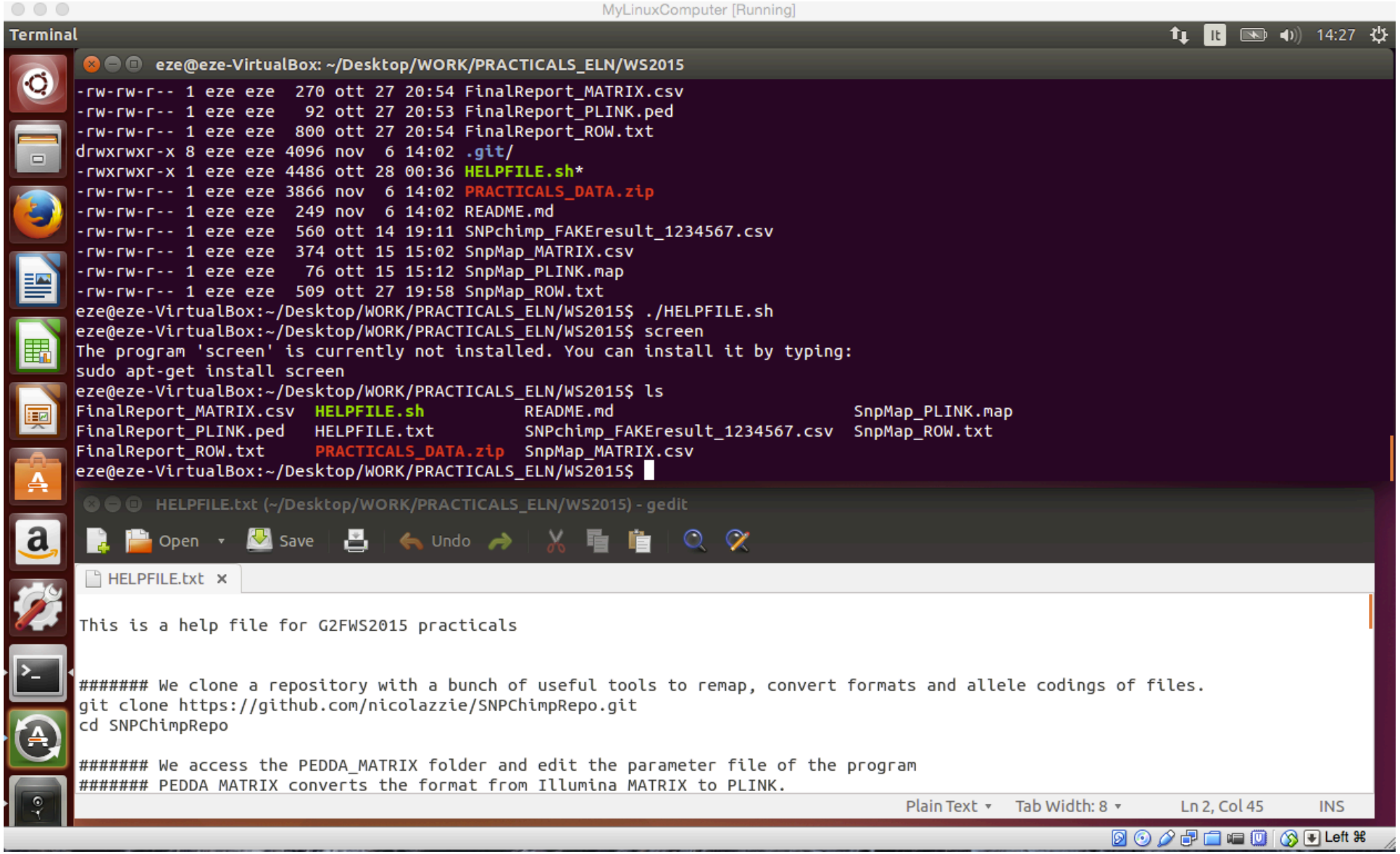
- **FinalReport_MATRIX.csv** ="Lab" file, contains genotypes in **matrix** format for 8 SNPs , sep is ","
- **FinalReport_ROW.txt** ="Lab" file, contains genotypes in **row** format for 10 SNPs , sep is "tab"
- **FinalReport_PLINK.ped** ="Lab" file, contains genotypes in **PLINK** format for 6 SNPs, sep is "space"
- **SnpMap_*** = Map and SNP coordinates and information (they come with the gen files)
- **SNPchimp_*** = See meaning of this file later.
- **HELPFILE.sh** = Helper file to create paths.

**HELPFILE.sh** is a simple program created ad-hoc for the G2F WinterSchool 2015. It will create a *HELP* file for your convenience during this practical. It contains the series of commands, including the long paths you have to use in the parameter files.

**How to access it:** `./HELPFILE.sh`

**It creates a HELPFILE.txt file, so if you accidentally close it, you can still open it using the command** `gedit HELPFILE.txt`

# Advice..


MyLinuxComputer [Running]
Terminal
14:27
eze@eze-VirtualBox: ~/Desktop/WORK/PRACTICALS_ELN/WS2015
-rw-rw-r-- 1 eze eze 270 ott 27 20:54 FinalReport_MATRIX.csv
-rw-rw-r-- 1 eze eze 92 ott 27 20:53 FinalReport_PLINK.ped
-rw-rw-r-- 1 eze eze 800 ott 27 20:54 FinalReport_ROW.txt
drwxrwxr-x 8 eze eze 4096 nov 6 14:02 .git/
-rwxrwxr-x 1 eze eze 4486 ott 28 00:36 HELPFILE.sh*
-rw-rw-r-- 1 eze eze 3866 nov 6 14:02 PRACTICALS_DATA.zip
-rw-rw-r-- 1 eze eze 249 nov 6 14:02 README.md
-rw-rw-r-- 1 eze eze 560 ott 14 19:11 SNPchimp_FAKEresult_1234567.csv
-rw-rw-r-- 1 eze eze 374 ott 15 15:02 SnpMap_MATRIX.csv
-rw-rw-r-- 1 eze eze 76 ott 15 15:12 SnpMap_PLINK.map
-rw-rw-r-- 1 eze eze 509 ott 27 19:58 SnpMap_ROW.txt
eze@eze-VirtualBox:~/Desktop/WORK/PRACTICALS_ELN/WS2015$ ./HELPFILE.sh
eze@eze-VirtualBox:~/Desktop/WORK/PRACTICALS_ELN/WS2015$ screen
The program 'screen' is currently not installed. You can install it by typing:
sudo apt-get install screen
eze@eze-VirtualBox:~/Desktop/WORK/PRACTICALS_ELN/WS2015$ ls
FinalReport_MATRIX.csv HELPFILE.sh README.md SnpMap_PLINK.map
FinalReport_PLINK.ped HELPFILE.txt SNPchimp_FAKEresult_1234567.csv SnpMap_ROW.txt
FinalReport_ROW.txt PRACTICALS_DATA.zip SnpMap_MATRIX.csv
eze@eze-VirtualBox:~/Desktop/WORK/PRACTICALS_ELN/WS2015$
HELPFILE.txt (~/Desktop/WORK/PRACTICALS_ELN/WS2015) - gedit
Open
Save
Undo
HELPFILE.txt
This is a help file for G2FWS2015 practicals
####### We clone a repository with a bunch of useful tools to remap, convert formats and allele codings of files.
git clone https://github.com/nicolazzie/SNPChimpRepo.git
cd SNPChimpRepo
####### We access the PEDDA_MATRIX folder and edit the parameter file of the program
####### PEDDA MATRIX converts the format from Illumina MATRIX to PLINK.
Plain Text
Tab Width: 8
Ln 2, Col 45
INS
Left ⌘

# Our PLAN:

1) **Standardize data formats** . We cannot handle 3 different format of files contemporarily! We'll transform EVERY genotype file into a PLINK format.

2) Deal with **allele conversion** and **map information**.

3) **Merge** all files

4) **Analyze** the data.

How? Using some tools developed in Gene2farm!

REALLY easy to use (will become easier), once you know how to!

First, let's download (part of) the tools:

```
git clone https://github.com/nicolazzie/SNPConvert.git
```

# You have the command below on your “HELPFILE” (copy and paste it)

**git clone https://github.com/nicolazzie/SNPConvert.git**

```
eze@eze-VirtualBox:~/Desktop/WORK/PRACTICALS_ELN/WS2015$ cd SNPChimpRepo/
eze@eze-VirtualBox:~/Desktop/WORK/PRACTICALS_ELN/WS2015/SNPChimpRepo$ ls
iConvert  PEDDA_MATRIX  PEDDA_ROW  README.md
eze@eze-VirtualBox:~/Desktop/WORK/PRACTICALS_ELN/WS2015/SNPChimpRepo$ ll
total 36
drwxrwxr-x 6 eze eze 4096 ott 15 15:21 ./
drwxrwxr-x 4 eze eze 4096 ott 15 15:21 ../
drwxrwxr-x 8 eze eze 4096 ott 15 15:21 .git/
drwxrwxr-x 3 eze eze 4096 ott 15 15:21 iConvert/
drwxrwxr-x 3 eze eze 4096 ott 15 15:21 PEDDA_MATRIX/
drwxrwxr-x 3 eze eze 4096 ott 15 15:21 PEDDA_ROW/
-rw-rw-r-- 1 eze eze 8293 ott 15 15:21 README.md
eze@eze-VirtualBox:~/Desktop/WORK/PRACTICALS_ELN/WS2015/SNPChimpRepo$
```

3 tools: PEDDA_MATRIX, PEDDA_ROW, iConvert

# Gene2farm software we'll use

Now, SNPConvert GUI

**PEDDA_matrix.py**

Transforms Illumina MATRIX format files into PLINK. Maintains allele coding as is.

**PEDDA_row.py**

Transforms Illumina ROW format files into PLINK. If more than one allele codes are present, user gets to choose which allele format to output.

**iConvert.py**

Converts allele coding, using a SNPchimp file (e.g. downloaded from the website) to do the conversion. If required, it also updates map information.

**Zanardi.py**

Multi-purpose software that automatically downloads software, merges files and streamlines multiple G-analyses.

PTP SCIENCE PARK

# PEDDA_MATRIX

```
cd PEDDA_MATRIX
gedit peddam.param
```

```
eze@eze-VirtualBox:~/Desktop/WORK/
[Header]
GSGT Version 1.1.1
Processing Date,1/1/1000 0:00 PM
Content testfile.bpm
Num SNPs,8
Total SNPs,8
Num Samples,2
Total Samples,2
[Data]
,MATRIX_1,MATRIX_2
snp1,AB,BB
snp2,AB,AB
snp3,AA,AA
snp4,AA,AA
snp7,AB,AA
snp8,AA,--
snp9,--,AB
snp10,AB,BB
```

(a screen will open, scroll down with your down arrow or mouse)

```
finrep='Example_files/test_FinalReport.txt'

snpmap='Example_files/test_SnpMap.txt'

outname='test_outputfile'
brdcode='TEST'

sep=','
```

You only have to modify the 5 parameters

**finrep**: Final report in matrix fmt
**snpmap**: SNP map
**outname**: name of output file
**brdcode**: Name of breed
**sep**: separator

PTP SCIENCE PARK

# PEDDA_MATRIX

```
eze@eze-VirtualBox:~/Desktop/WORK/P
[Header]
GSGT Version 1.1.1
Processing Date,1/1/1000 0:00 PM
Content testfile.bpm
Num SNPs,8
Total SNPs,8
Num Samples,2
Total Samples,2
[Data]
,MATRIX_1,MATRIX_2
snp1,AB,BB
snp2,AB,AB
snp3,AA,AA
snp4,AA,AA
snp7,AB,AA
snp8,AA,--
snp9,--,AB
snp10,AB,BB
```

Use your own name!

```
finrep='/home/eze/Desktop/WORK/PRACTICALS_ELN/WS2015/FinalReport_MATRIX.csv'
snpmap='/home/eze/Desktop/WORK/PRACTICALS_ELN/WS2015/SnpMap_MATRIX.csv'
outname='/home/eze/Desktop/WORK/PRACTICALS_ELN/WS2015/PLINK_MATRIX_AB'
brdcode='BSW'
sep=','
```

Let me show you how to do this using the HELP file…

# PEDDA_MATRIX

python pedda_matrix.py

```
eze@eze-VirtualBox:~/Desktop/WORK/PRACTICALS_ELN/WS2015/SNPChimpRepo/PEDDA_MATRIX$ python pedda_matrix.py
############################################################
###                                                      ###
###               PEDDA MATRIX software                  ###
###      Converts Illumina row fmt into PLINK fmt        ###
###                                                      ###
###                              Coded by: E.L.Nicolazzi ###
############################################################
### Rough check of parameters
====> Parameters check: OK

### Processing SNPmap file
====> Total number of SNPs processed: 8

### Processing MATRIX FinalReport file:
====> Total number of INDIVIDUALS processed: 2

### Writing output ped file
====> PED FILE PRODUCED: /home/eze/Desktop/WORK/PRACTICALS_ELN/WS2015/PLINK_MATRIX_AB.ped
====> MAP FILE PRODUCED: /home/eze/Desktop/WORK/PRACTICALS_ELN/WS2015/PLINK_MATRIX_AB.map

BAZINGA! I'm done!
```

# PEDDA_ROW

```
cd ../PEDDA_MATRIX
gedit peddar.param
```

(a screen will open, scroll down with your down arrow or mouse)

```
eze@eze-VirtualBox:~/Desktop/WORK/PRACTICALS_ELN/WS2015$ more FinalReport_ROW.txt
[Header]
GSGT    Version 1.1.1
Processing      Date    1/1/1000        0:00    PM
Content testfile.bpm
Num     SNPs    10
Total   SNPs    10
Num     Samples 2
Total   Samples 2
[Data]
SNP Name        Sample ID       Allele1 - Forward       Allele2 - Forward       Allele1 - Top
Allele2 - Top   GC Score
snp1    ROW_38  A       G       A       G       0.6853
snp2    ROW_38  T       T       T       T       0.5647
snp3    ROW_38  A       A       A       A       0.5620
snp4    ROW_38  A       A       A       A       0.4777
snp5    ROW_38  G       G       G       G       0.7372
snp6    ROW_38  T       G       A       C       0.5613
snp7    ROW_38  A       G       A       G       0.6411
```

```
finrep='Example_files/test_FinalReport.txt'

snpmap='Example_files/test_SnpMap.txt'

allele='top'


SNPid_pos='1'

INDid_pos='2'


outname='test_outputfile'
brdcode='TEST'
sep='\t'
```

Same 5 as before + 3 new params:

**allele**: allele format desired (must be in the FinalReport)
**SNPid_pos**: Column of the SNP id
**INDid_pos**: Column of the ID

# PEDDA_ROW

Use your own name!

```
finrep='/home/eze/Desktop/WORK/PRACTICALS_ELN/WS2015/FinalReport_ROW.txt'
snpmap='/home/eze/Desktop/WORK/PRACTICALS_ELN/WS2015/SnpMap_ROW.txt'
allele='top'
SNPid_pos='1'
INDid_pos='2'
outname='/home/eze/Desktop/WORK/PRACTICALS_ELN/WS2015/PLINK_ROW_TOP'
brdcode='FLK'
sep='\t'
```

| SNP Name | Sample ID | Allele1 - Forward | | Allele2 - Forward | | Allele1 - Top | Allele2 - Top | GC Score |
|---|---|---|---|---|---|---|---|---|
| snp1 | ROW_38 | A | G | A | G | 0.6853 | | |
| snp2 | ROW_38 | T | T | T | T | 0.5647 | | |
| snp3 | ROW_38 | A | A | A | A | 0.5620 | | |
| snp4 | ROW_38 | A | A | A | A | 0.4777 | | |
| snp5 | ROW_38 | G | G | G | G | 0.7372 | | |

# PEDDA_ROW

python pedda_row.py

```
eze@eze-VirtualBox:~/Desktop/WORK/PRACTICALS_ELN/WS2015/SNPChimpRepo/PEDDA_ROW$ python pedda_row.py
#############################################################
###                                                       ###
###                  PEDDA software                       ###
###      Converts Illumina row fmt into PLINK fmt         ###
###                                                       ###
###                              Coded by: E.L.Nicolazzi ###
#############################################################
### Rough check of parameters
====> Parameters check: OK

### Processing FinalReport file:
Finshed processing individual: ROW_38  - Total SNPs: 10
Finshed processing individual: ROW_39  - Total SNPs: 10
====> Total number of INDIVIDUALS processed: 2

### Processing SNPmap file
====> Total number of SNPs processed: 10

### Writing output map file
====> PED FILE PRODUCED: /home/eze//Desktop/WORK/PRACTICALS_ELN/WS2015/PLINK_ROW_TOP.ped
====> MAP FILE PRODUCED: /home/eze//Desktop/WORK/PRACTICALS_ELN/WS2015/PLINK_ROW_TOP.map

BAZINGA! I'm done!
```

# Step 1 (standardizing formats): DONE

We have successfully translated all the formats into 3 single PLINK (ped/map) files.

However, they are in 3 different allele codings!

**PLINK_MATRIX_AB.ped is in AB (NOT OK)**

**PLINK_ROW_TOP.ped is in TOP (OK)**

**FinalReport_PLINK.ped is in FORWARD (NOT OK)**

We need to standardize that!!!

Fortunately, there is a tool to do that using *SNPchimp* data as source for allele conversion. Its name is *iConvert*, and allows you to convert allele codings and update the map information in the process.

# iConvert

I convert is a more user-friendly program. You do not have to enter the program to set the parameters: there is a parameter file.

First, let's go to the right folder:

```
cd ../iConvert
```

There you'll find the iConvert.py program and a convert.param file.

Open the convert.param file

```
gedit convert.param
```

# iConvert parameter file

gedit convert.param

```
#### Name of PED (plink style) file
PEDfile       = EXAMPLEFILES/Bovine_54k1.ped

#### Name of MAP (plink style) file
MAPfile       = EXAMPLEFILES/Bovine_54k1.map

#### Missing value for genotypes in PLINK files
MISSING       = 0

#### INPUT allele format
####    - If Affymetrix chip: aff_forward, aff_ab, NO
####    - If Illumina or Illumina-based chip: ill_top, ill_forward, ill_ab, NO
IN_format     = ill_forward

#### Update your map (chrom and position) information with data in SNPchimp? (Y/N)
UPDATE_map    = Y

#### Name of SNPchimp file containing allele conversions
SNPchimp_file = EXAMPLEFILES/SNPchimp_result_685452573.csv

#### OUTPUT allele format
####    - If Affymetrix chip: aff_forward, aff_ab, NO
####    - If Illumina or Illumina-based chip: ill_top, ill_forward, ill_ab, NO
OUT_format    = ill_top
```

# Modifying the parameter file

Use your own name!

```
#### Name of PED (plink style) file
PEDfile       = /home/eze/Desktop/WORK/PRACTICALS_ELN/WS2015/PLINK_MATRIX_AB.ped

#### Name of MAP (plink style) file
MAPfile       = /home/eze/Desktop/WORK/PRACTICALS_ELN/WS2015/PLINK_MATRIX_AB.map

#### Missing value for genotypes in PLINK files
MISSING       = 0

#### INPUT allele format
####    - If Affymetrix chip: aff_forward, aff_ab, NO
####    - If Illumina or Illumina-based chip: ill_top, ill_forward, ill_ab, NO
IN_format     = ill_ab

#### Update your map (chrom and position) information with data in SNPchimp? (Y/N)
UPDATE_map    = Y

#### Name of SNPchimp file containing allele conversions
SNPchimp_file = /home/eze/Desktop/WORK/PRACTICALS_ELN/WS2015/SNPchimp_FAKEresult_1234567.csv

#### OUTPUT allele format
####    - If Affymetrix chip: aff_forward, aff_ab, NO
####    - If Illumina or Illumina-based chip: ill_top, ill_forward, ill_ab, NO
OUT_format    = ill_top
```

# First run…done

python iConvert.py

```
eze@eze-VirtualBox:~/Desktop/WORK/PRACTICALS_ELN/WS2015/SNPChimpRepo/iConvert$ python iConvert.py

*******************************************************************************
==> PROGRAM STARTS: October 15, 2015 -  4:22PM CEST
*******************************************************************************
Parameters read from: ./convert.param
    - PLINK PED file         : /home/eze/Desktop/WORK/PRACTICALS_ELN/WS2015/PLINK_MATRIX_AB.ped
    - PLINK MAP file         : /home/eze/Desktop/WORK/PRACTICALS_ELN/WS2015/PLINK_MATRIX_AB.map
    - Input allele format    : ill_ab
    - Update map information : y
    - SNPchimp file          : /home/eze/Desktop/WORK/PRACTICALS_ELN/WS2015/SNPchimp_FAKEresult_1234567.csv
    - Output allele format   : ill_top

Reading SNPchimp file and checking available options
    - COMMAS used as separator for SNPchimp file
    - Rows read in SNP chimp file                      : 10
    - SNPs with allele conversion information available : 10

Reading MAP file and converting map info (if required)
    - SPACES used as separator for PLINK MAP file
    - Rows read in PLINK MAP file (number of SNPs)     : 8
    - MAP file updated, please see file                : /home/eze/Desktop/WORK/PRACTICALS_ELN/WS2015/PLINK_MATRIX_AB_updated.map

Reading PED file and converting alleles as required
    - SPACES used as separator for PLINK PED file
    - Rows read in PLINK PED file (number of animals)  : 2
    - PED file updated, please see file                : /home/eze/Desktop/WORK/PRACTICALS_ELN/WS2015/PLINK_MATRIX_AB_updated.map

*******************************************************************************
==> PROGRAM ENDS: October 15, 2015 -  4:22PM CEST
*******************************************************************************

eze@eze-VirtualBox:~/Desktop/WORK/PRACTICALS_ELN/WS2015/SNPChimpRepo/iConvert$
```

# Modifying the parameter file

`gedit convert.param`

Use your own name!

```
#### Name of PED (plink style) file
PEDfile       = /home/eze/Desktop/WORK/PRACTICALS_ELN/WS2015/FinalReport_PLINK.ped

#### Name of MAP (plink style) file
MAPfile       = /home/eze/Desktop/WORK/PRACTICALS_ELN/WS2015/SnpMap_PLINK.map

#### Missing value for genotypes in PLINK files
MISSING       = 0

#### INPUT allele format
####    - If Affymetrix chip: aff_forward, aff_ab, NO
####    - If Illumina or Illumina-based chip: ill_top, ill_forward, ill_ab, NO
IN_format     = ill_ab

#### Update your map (chrom and position) information with data in SNPchimp? (Y/N)
UPDATE_map    = Y

#### Name of SNPchimp file containing allele conversions
SNPchimp_file = /home/eze/Desktop/WORK/PRACTICALS_ELN/WS2015/SNPchimp_FAKEresult_1234567.csv

#### OUTPUT allele format
####    - If Affymetrix chip: aff_forward, aff_ab, NO
####    - If Illumina or Illumina-based chip: ill_top, ill_forward, ill_ab, NO
OUT_format    = ill_top
```

# Second run…done

python iConvert.py

```
eze@eze-VirtualBox:~/Desktop/WORK/PRACTICALS_ELN/WS2015/SNPChimpRepo/iConvert$ python iConvert.py

*******************************************************************************
==> PROGRAM STARTS: October 15, 2015 -  4:33PM CEST
*******************************************************************************
Parameters read from: ./convert.param
    - PLINK PED file         : /home/eze/Desktop/WORK/PRACTICALS_ELN/WS2015/FinalReport_PLINK.ped
    - PLINK MAP file         : /home/eze/Desktop/WORK/PRACTICALS_ELN/WS2015/SnpMap_PLINK.map
    - Input allele format    : ill_ab
    - Update map information : y
    - SNPchimp file          : /home/eze/Desktop/WORK/PRACTICALS_ELN/WS2015/SNPchimp_FAKEresult_1234567.csv
    - Output allele format   : ill_top

Reading SNPchimp file and checking available options
    - COMMAS used as separator for SNPchimp file
    - Rows read in SNP chimp file                      : 10
    - SNPs with allele conversion information available : 10

Reading MAP file and converting map info (if required)
    - SPACES used as separator for PLINK MAP file
    - Rows read in PLINK MAP file (number of SNPs)     : 6
    - MAP file updated, please see file                : /home/eze/Desktop/WORK/PRACTICALS_ELN/WS2015/SnpMap_PLINK_updated.map

Reading PED file and converting alleles as required
    - SPACES used as separator for PLINK PED file
    - Rows read in PLINK PED file (number of animals)  : 2
    - PED file updated, please see file                : /home/eze/Desktop/WORK/PRACTICALS_ELN/WS2015/FinalReport_PLINK_updated.map

*******************************************************************************
==> PROGRAM ENDS: October 15, 2015 -  4:33PM CEST
*******************************************************************************
```

# What we have (cleaning up)

```
cd ../..
ls
```

```
eze@eze-VirtualBox:~/Desktop/WORK/PRACTICALS_ELN/WS2015$ ls
FinalReport_MATRIX.csv          PLINK_MATRIX_AB.map          SNPchimp_FAKEresult_1234567.csv
FinalReport_PLINK.ped           PLINK_MATRIX_AB.ped          SNPChimpRepo
FinalReport_PLINK_updated.ped   PLINK_MATRIX_AB_updated.map  SnpMap_MATRIX.csv
FinalReport_ROW.txt             PLINK_MATRIX_AB_updated.ped  SnpMap_PLINK.map
#HELPFILE.sh#                   PLINK_ROW_TOP.map            SnpMap_PLINK_updated.map
HELPFILE.sh                     PLINK_ROW_TOP.ped            SnpMap_ROW.txt
HELPFILE.txt                    PRACTICALS_DATA.zip
ORIG_TEMP_FILES                 README.md
```

```
mkdir FINAL_FILES
mv *_updated* PLINK_ROW_TOP.* FINAL_FILES/
```

FinalReport_PLINK_updated.ped
SnpMap_PLINK_updated.map
PLINK_MATRIX_AB_updated.map
PLINK_MATRIX_AB_updated.ped
PLINK_ROW_TOP.map
PLINK_ROW_TOP.ped

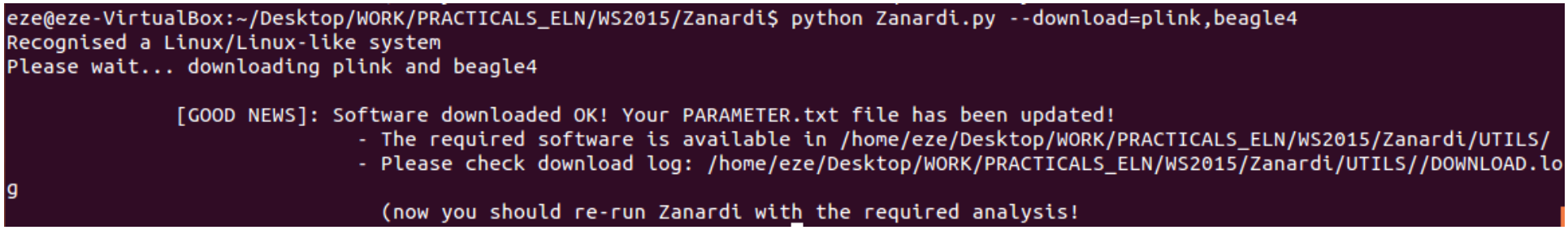
# Now we download and run Zanardi

```
git clone https://github.com/bioinformatics-ptp/Zanardi.git
```

We will try to do an imputation process using beagle v.4, so we will need to download PLINK (to manage the datasets) and BEAGLE v.4.

Zanardi has an automatic routine to download, install and auto-compile the required input on the parameter file.

```
cd Zanardi
python Zanardi.py --download=plink,beagle4
```



# Managing Zanardi: ALL is on the parameter file



```
gedit PARAMETER.txt
```

PARAMFILE.txt ×

```
## This is a parameter file for Zanardi pipeline. Follow the specifics in Zanardi WIKI.
##
## NOTES: ALL lines beginning with "#" will be skipped!
##        DO NOT delete or modify ANY variable name (not even if you're not using the pgm!)
######
## History:
## Coded originally: Gabriele Marras & Ezequiel L. Nicolazzi (2015)
##    |   Oct. 2015 : Release of stable version with params for options still not released (e.g. phenotype)
######

### ----------------------
### --- SOFTWARE PATHS ---
### ----------------------
### --- *ABSOLUTE* path to the required software for all the analyses within Zanardi
### --- If you don't have a required program installed please see WIKI (specifically, the --download option)
PGM_PLINK=/home/eze/Desktop/WORK/PRACTICALS_ELN/WS2015/Zanardi/UTILS//PLINK
PGM_FCGENE=
PGM_BEAGLE3=
PGM_BEAGLE4=/home/eze/Desktop/WORK/PRACTICALS_ELN/WS2015/Zanardi/UTILS//BEAGLE
PGM_ADMIXTURE=
```

Compiled by the --download option

Zanardi can automatically merge and analyze multiple PLINK files.

What we're going to do is to include the 3 PLINK files already standardized, edit the merged file, run a multidimensional scaling plot analysis and impute the whole dataset using beagle v.4.

All this can be done editing a single parameter file!

# We're going to modify 4 sections of the parameter file: INPUT, PLINKQC, MDS and BEAGLE

```
### ----------------------------------------
### --- GENOTYPE & PEDIGREE INPUT FILES ---
### ----------------------------------------
### --- *ABSOLUTE* path preferred
SPECIES= COW
INPUT_PED= Example_data/example_PLINK_snp2000.ped,Example_data/example_PLINK_snp4000.ped
INPUT_MAP= Example_data/example_PLINK_snp2000.map,Example_data/example_PLINK_snp4000.map
INPUT_705=
INPUT_705_MAP=
INPUT_PEDIG= Example_data/example_pedigree.txt
INPUT_PHENO= Example_data/example_phenotype.txt
OUTPUT_NAME= EXAMPLEDATA

### ----------------------------------------
### --- PARAMETERS FOR plinkqc OPTION   ---
### ----------------------------------------
QCMISS_IND= 0.10
QCMISS_SNP= 0.05
QCMAF=
QCHWE=
QC_OTHOPT= --autosome
```

Meaning: Individuals with > 10% missing will be discarded, SNP with > 5% missing calls or not in the autosomal chromosomes will be discarded

```
### ----------------------------------------
### --- PARAMETERS FOR mds OPTION (Y/N) ---
### ----------------------------------------
MDSGROUPop='N'
```

```
### ----------------------------------------------
### --- PARAMETERS FOR IMPUTATION - BEAGLE  ---
### ----------------------------------------------
BGMEMORY=4000
BG3_MISSING=0
BG_OTHOPT=
```

Meaning:
MDS option: Do not group individuals by population
BEAGLE option: RAM memory usage: 4Gb
Missing values: 0
Other options: none

# gedit PARAMETER.txt

Modify INPUT_PED and INPUT_MAP

Delete the variables in INPUT_PEDIG and INPUT_PHENO

Edit the variable OUTPUT_NAME to “WINTERSCHOOL”

```
### ---------------------------------------
### --- GENOTYPE & PEDIGREE INPUT FILES ---
### ---------------------------------------
### --- *ABSOLUTE* path preferred
SPECIES= COW
INPUT_PED= /home/eze/Desktop/WORK/PRACTICALS_ELN/WS2015/FINAL_FILES/FinalReport_PLINK_updated.ped,/home/eze/Desktop/
WORK/PRACTICALS_ELN/WS2015/FINAL_FILES/PLINK_MATRIX_AB_updated.ped,/home/eze/Desktop/WORK/PRACTICALS_ELN/WS2015/
FINAL_FILES/PLINK_ROW_TOP.ped
INPUT_MAP= /home/eze/Desktop/WORK/PRACTICALS_ELN/WS2015/FINAL_FILES/SnpMap_PLINK_updated.map,/home/eze/Desktop/WORK/
PRACTICALS_ELN/WS2015/FINAL_FILES/PLINK_MATRIX_AB_updated.map,/home/eze/Desktop/WORK/PRACTICALS_ELN/WS2015/
FINAL_FILES/PLINK_ROW_TOP.map
INPUT_705=
INPUT_705_MAP=
INPUT_PEDIG=
INPUT_PHENO=
OUTPUT_NAME=WINTERSCHOOL
```

Keep the other variables as default

# Now run the program..

```
python Zanardi.py --plinkqc --mds --beagle4
```

A step-by-step log will be printed on screen (and on a "Zanardi.log" file).

An output folder will be created containing all output files.

You have just edited, run an MDS and imputed 3 different chips.

**Congrats!**

(Know that many other analyses are available…)

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n° 289592

GENE 2 FARM

PTP
SCIENCE
PARK

adding value **from research**

**follow us**

**www.ptp.it**