QC Functions v1.2 User Manual

Chen Chun-Yu

April 25, 2017

Contents

1 Introduction

This manual shows you how to use the quality control functions and give an example for demonstration.

2 Prerequest

To make the functions run successfully, please ensure you have installed PLINK in previous.
After you installed PLINK, please copy it from defult folder and paste it to /usr/local/bin.

```
$ cp ~/Bin/plink /usr/local/bin
```

3 Caution

Do not modify any content arbitrarily of ouput files generated from PLINK!

4 QC Functions

4.1 Per-individual QC

- sex_check(input.name, output.name)
  Place the name of your PLINK binary files(BED/BIM/FAM) at "input.name" argument, and give the output name at "output.name" argument. After the analysis, you will get a plot("sex_distribution.pdf") which shows the distribution of individual's sex and a "output.name_sex_problem.list" file which records individuals with discordant sex information. Defult homozygosity rates of identifying individual as male is above 0.8, and recognizing as female is below 0.2.

- missing_het_ind(input.name, pop.list, output.name)
  You still need to give input and ouput names for the analysis. You also need to give a list of individuals with their populations at "pop.list" argument, the example format shows below:

1	TDC13	Paiwan
2	TDC117	Amis
3	TDC18	Bunun
4	TDC129	Amis
5	TDC49	Amis
6	TDC497	Puyuma

  You will get a plot("imiss-vs-het.pdf") which shows the distribution of missingness and heterozygosity scores and a "output.name_miss_het_problem.list" file which records individuals do not pass criteria. Default cuttoffs of genotype failure rates are equal or larger than 0.03 and heterozygosity rates deviate more or less 3 s.d. from the mean.

- IBD(input.name, output.name)
  You still need to give input and ouput names for the analysis. After the analysis, you will get a plot("IBD.pdf") which shows the propotion of the different IBD between pairs of individuals. You will also get a "output.name_ibd_problem.list" that records the individuals do not pass the criterion. Default value of IBD we intend to remove is higher than 0.1875.

- ind_qc_rm(input.name, output.name)
  After you finish the steps of per-individual QC, you can use ind_qc_rm function to output the

list(”output.name_fail_ind_qc.txt”) which contains all the problem lists that are generated by previous QC steps. You can use following PLINK command to remove the individuals easily:

```
plink --bfile your.PLINK.bfile --remove output.name_fail_ind_QC.txt --make-bed --out output.name
```

- ind_qc_info(input.name, output.name)
  After you removing the individuals who failed to pass the QC steps, you can use this function to see the ID of individuals that removed in each step. You also can find the individual's ID which didn't pass two of three steps or even all three steps. The output file will be a text file which record all the information.

- List_to_DF(input.list)
  This function will be used inside of ind_qc_info function, so it is necessarily to load this function before you using ind_qc_info function. List_to_DF can also be used independently to convert the list structure to data frame struture in R.

4.2 Per-SNP QC

- missing_snp(input.name, output.name)
  You still need to give input and ouput names for the analysis. You will get a plot(”snpmiss_plot.pdf”) which shows the distribution of missing genotype rate and a threshold for extreme genotype failure rate. Default missing genotype rate threshold is equal to 0.03.

- hwe_test(input.name, output.name)
  You still need to give input and ouput names for the analysis. You will get a plot(”hwe_p_value.pdf”) which shows the distribution of Hardy-Weinberg Equilibrium test's p-value and a threshold for extreme high p-value. Default extreme p-value threshold is equal to 0.00001.

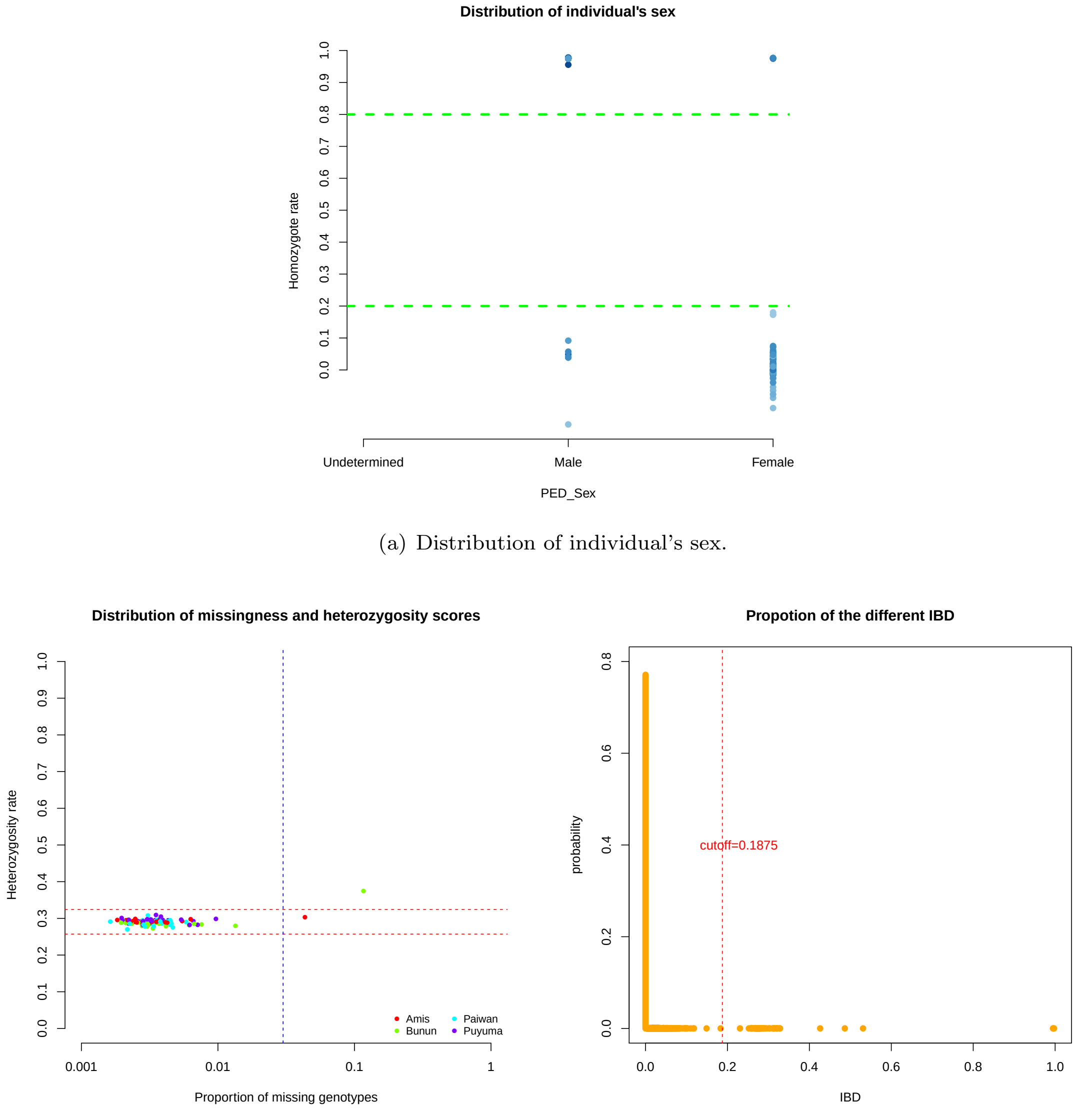
Two per-SNP QC steps show above are aim to show the distribution of missing genotype rate and Hardy-Weinberg Equilibrium test's p-value. If you want to further exclude those SNPs, please use following PLINK commands:

```
plink --bfile your.PLINK.bfile --maf 0.01 --geno 0.03 --hwe 0.00001 --make-bed --out output.name
```

5 Example

Here we give you an example of applying the functions we have illustrated. The example data is consists of 96 individuals and about 2.5 millions of SNPs.

5.1 Per-individual QC



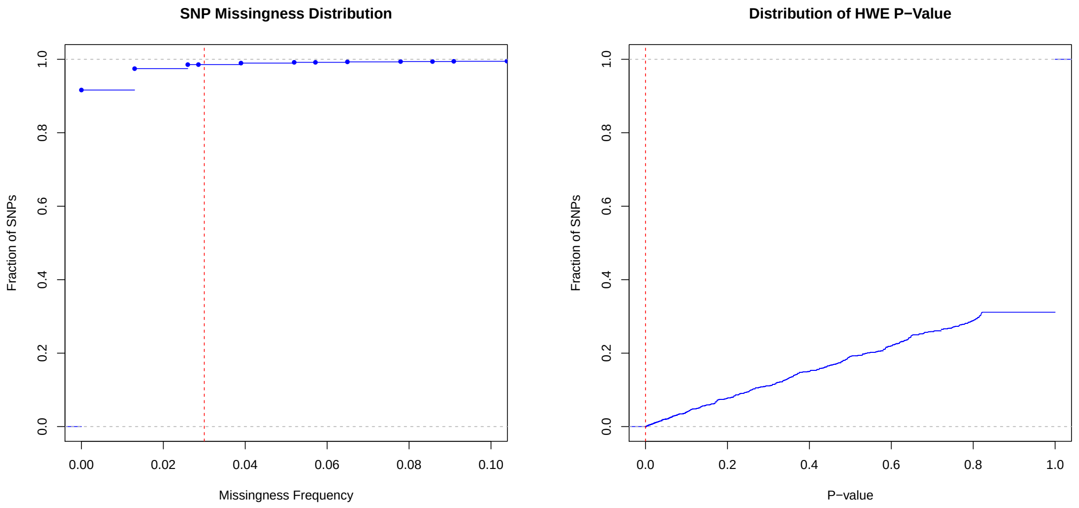
(a) Distribution of individual's sex.

(b) Distribution of missingness and heterozygosity scores.

(c) Propotion of the different IBD

Figure 1: Three analysis in per-individual QC.

5.2 Per-SNP QC



(a) Distribution of missing genotype rate.

(b) Distribution of Hardy-Weinberg Equilibrium test's p-value.

Figure 2: Two analysis in per-SNP QC.

References

[1] Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. *Nature protocols.* 2010;5(9):1564-1573. doi:10.1038/nprot.2010.116.

[2] Purcell S, Neale B, Todd-Brown K, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *American Journal of Human Genetics.* 2007;81(3):559-575.