Quality control report for Human Omni2.5 Exome8 SNP chip data

Chun-Yu Chen

May 15, 2017

Contents

1 Introduction

This report encompasses the quality control summary for Human Omni2.5 Exome8 SNP chip data. A total of 96 samples were genotyped for 2583651 SNPs.

2 Materials and Methods

2.1 Software

- PLINK (Purcell, 2007)
- Scripts for processing results files
- R (Statistical Software) for plotting results

2.2 Per-individual QC

Per-individual QC of GWA data consists of at least four steps: (i) identification of individuals with discordant sex information, (ii) identification of individuals with outlying missing genotype or heterozygosity rates, (iii) identification of duplicated or related individuals and (iv) identification of individuals of divergent ancestry.

2.2.1 Creation of BED files

The format in which genotype data are returned to investigators varies among genome-wide SNP platforms and genotyping centers. The format of the 96 samples data used in this report are provided in standard PED and MAP file formats. To facilitate the analysis of large-scale data sets, we use PLINK to create BED, BIM and FAM files. Type the following PLINK command in terminal:

```
plink --file 96samples --make-bed --out new
```

2.2.2 Merge the files in PLINK (Optional)

Because of 8 samples without genotyping records in our dataset were detected in the previous qulity control, here we showed the steps to remove those samples and merge with the right one by using PLINK. Remove the 8 samples.

```
plink --bfile 96samples --remove list.txt --make-bed --out extract
```

Merge the 8 samples file which contains right information with whole samples file.

```
plink --bfile 96samples --remove list.txt --make-bed --out extract
```

```
plink --bfile 8samples --bmerge extract.bed extract.bim extract.fam --recode --out new
```

2.2.3 Identification of individuals with discordant sex information

The best way to detect discrepancies between genotype information and ascertained sex is to calculate homozygosity rates across all X-chromosome SNPs for each individual in the sample and compare these with the expected rates. Because of our sample files didn't record any ascertained sex, we needed to add the sex information which based on the other file. Here, we used R to complete this work.

```
fam = read.table("~/QC/test/new.fam")
sex = read.csv("~/QC/test/96samples_sex_list.csv",header = F)
# To find out the ID names whether consist between two files or not
fam$V5 = sex$V3
fam$V2 == sex$V2
# Write out new fam file which record the right sex information
levels(fam$V5)[1] = 2
levels(fam$V5)[2] = 1
write.table(fam,"~/QC/test/new.fam",col.names = F,row.names = F,quote = F)
```

We firstly use PLINK to calculate the mean homozygosity rate across X-chromosome markers for each individual in the study:

```
plink --bfile new --check-sex --out new
```

Then use R to make a plot for showing the distribution of individual's sex:

```
sexcheck = read.table("~/QC/test/new.sexcheck",header=T)
colors = densCols(sexcheck$F)
#pdf("sex_distribution.pdf")
plot(sexcheck$PEDSEX,sexcheck$F,pch=19, xlim = c(0,2),
     col=colors,axes = F, xlab = "PED_Sex", ylab = "Homozygote rate")
axis(1,at=seq(0,2,1),labels = c("Undetermined","Male","Female"))
axis(2,at=seq(0,1,0.1))
abline(h=0.8,lty=2,col="green",lwd=3)
abline(h=0.2,lty=2,col="green",lwd=3)
title("Distribution of individual's sex")
#dev.off()
```

Finally, we use R to identify individuals with discordant sex information.

```
fail_sex = which(sexcheck$STATUS == "PROBLEM")
fail_sex_id = sexcheck[c(fail_sex),c(1,2)]
write.table(fail_sex_id,"~/QC/test/fail_sex.txt",col.names = F, row.names = F)
```

2.2.4 Identification of individuals with elevated missing data rates or outlying heterozygosity rate

Samples of low DNA quality or concentration often have below average call rates and genotype accuracy. The genotype failure rate and heterozygosity rate per individual are both measures of DNA sample quality. First, we use PLINK to calculate missingness score for each individual:

```
plink --bfile new --missing --out new
```

Second, we use PLINK to calculate heterozygosity score for each individual:

```
plink --bfile new --het --out new
```

Third, we use R to plot the distribution of missingness and heterozygosity scores:

```r
imiss = read.table("~/QC/test/new.imiss",header=T)
het = read.table("~/QC/test/new.het",header=T)

#Calculate call rate, and add a column
imiss$CALL_RATE = 1-imiss$F_MISS
#Log 10 the F_MISS column, abd a column
imiss$logF_MISS = log10(imiss[,6])
#Calculate heterozygosity rate, and add a column
het$Het_propo = (het$N.NM. - het$O.HOM.)/het$N.NM.
#Find out 'NaN' value in meanHet column
#het£Het_propo = ifelse(het£Het_propo=="NaN", 0,het£Het_propo)

pop = names(table(sex$V4))
pop_col = rainbow(length(pop))
colors = densCols(imiss$logF_MISS,het$Het_propo)
for(i in 1:length(sex$V2)){
    if(sex$V2[i] == het$IID[i]){
        colors[i] = pop_col[sex$V4[i]]
    }
}
#pdf("imiss-vs-het.pdf")
plot(imiss$logF_MISS,het$Het_propo, col=colors,pch=20, xlim=c(-3,0),
     ylim=c(0,1),xlab="Proportion of missing genotypes",
     ylab="Heterozygosity rate",axes=FALSE)
axis(2,at=seq(0,1,0.1),tick=T)
axis(1,at=c(-3,-2,-1,0),labels=c(0.001,0.01,0.1,1))
legend("bottomright", legend=pop, col=pop_col, pch=16, ncol=2, bty="n", cex=0.8)
#Heterozygosity thresholds (Horizontal Line) +-3 s.d from the mean
abline(h=mean(het$Het_propo)-(3*sd(het$Het_propo)),col="RED",lty=2)
abline(h=mean(het$Het_propo)+(3*sd(het$Het_propo)),col="RED",lty=2)
#Missing Data Thresholds (Vertical Line)
abline(v=-1.522879, col="BLUE", lty=2) #THRESHOLD=0.03
title("Distribution of missingness and heterozygosity scores")
#dev.off()
```

Finally, we use R to identify individuals with high missingness and/or outlier heterozygosity based on preselected cutoff:

```r
imiss_het = merge(het,imiss,by = "FID")
fail_imisshet = imiss_het$Het_propo < mean(het$Het_propo)-(3*sd(het$Het_propo))
| imiss_het$Het_propo > mean(het$Het_propo)+(3*sd(het$Het_propo))
| imiss_het$F_MISS >= 0.03
which_bad = which(fail_imisshet,TRUE)
fail_imisshet_id = imiss_het[c(which_bad),c(1,2)]
colnames(fail_imisshet_id) = c("FID","IID")
write.table(fail_imisshet_id,file = "~/QC/test/fail_miss_het.txt",
            row.names = FALSE, col.names = FALSE)
```

2.2.5 Identification of duplicated or related individuals

In population-based case-control studies, all efforts should be made to limit the number of duplicate and related individuals in the design phase of the study. To minimize computational complexity, reduce the number of SNPs used to create the IBS matrix by pruning the data set so that no pair of SNPs (within a given number of base pairs) has an r^2 value greater than a given threshold (typically, 0.2).

```
plink --file new --indep-pairwise 50 5 0.2 --out new
```

Create a file new.genome containing pairwise IBS for all pairs of individuals:

```
plink --bfile new --extract new.prune.in --genome --out new
```

We use R to make a plot to show the propotion of the different IBD:

```
ibd_data = read.table("~/QC/test/new.genome",header = T)

# If there are any 'NaN' in data.
#bad = which(is.nan(ibd_data£PI_HAT))
#ibd_new = ibd_data£PI_HAT[-bad]

pro = data.frame()
total = length(ibd_data$PI_HAT)
for(i in 1:length(unique(ibd_data$PI_HAT))){
    pro[i,1] = unique(ibd_data$PI_HAT)[i]
    pro[i,2] = sum(ibd_data$PI_HAT == unique(ibd_data$PI_HAT)[i])
    pro[i,3] = sum(ibd_data$PI_HAT == unique(ibd_data$PI_HAT)[i])/total
}
colnames(pro) = c("ibd","total","proportion")
#pdf("ibd.pdf")
plot(pro$ibd, pro$proportion, type="h", ylim=c(0,0.8),
     xlim=c(0,1),ylab="probability",xlab="IBD", lwd=8, col="orange")
abline(v=0.1875, lty=2, col="red")
text(x=0.2280, y=0.4, labels=paste("cutoff=",0.1875, sep=""), col="red")
title("Propotion of the different IBD")
#dev.off()
```

Then, we use R to identify individuals with high IBD, and if there were missing values in the data.

```
imiss_data = read.table("~/QC/test/new.imiss",header=T)
first = ibd_data[,1][which(ibd_data$PI_HAT > 0.1875)]
second = ibd_data[,3][which(ibd_data$PI_HAT > 0.1875)]

fail_ibd = c()
for(n in 1:length(first)){
    if(imiss_data[first[n],6] > imiss_data[second[n],6]){
        fail_ibd[n] = imiss_data[first[n],1]
    }else if(imiss_data[first[n],6] < imiss_data[second[n],6]){
        fail_ibd[n] = imiss_data[second[n],1]
    }else if(imiss_data[first[n],6] == imiss_data[second[n],6]){
        fail_ibd[n] = imiss_data[first[n],1]
    }
}
fail_ibd = sort(unique(fail_ibd))
fail_ibd_id = imiss_data[c(fail_ibd),c(1,2)]
write.table(fail_ibd_id,file = "~/QC/test/fail_IBD.txt",col.names = F,row.names = F)
```

2.2.6 Identification of individuals of divergent ancestry

Skipped ancestry calculations due to time limitation.

2.2.7 Removal of all individuals failing QC

To concatenate all files listing individuals who fail the previous QC steps into a single file, type in R:

```
mg1 = rbind(fail_sex_id,fail_ibd_id)
mg2 = rbind(mg1,fail_imisshet_id)
fail_qc_inds = write.table(unique(mg2[order(mg2$FID),])
            ,"~/QC/test/fail_qc_inds.txt",row.names = F,col.names = F,quote = F)
```

To remove all the individuals failed in the previous QC steps, type:

```
plink --bfile new --remove fail_qc_inds.txt --make-bed --out clean_inds_data
```

2.3 Per-marker QC

Per-marker QC of GWA data consists of at least four steps: (i) identification of SNPs with an excessive missing genotype, (ii) identification of SNPs showing a significant deviation from Hardy-Weinberg equilibrium (HWE), (iii) identification of SNPs with significantly different missing genotype rates between cases and controls and (iv) the removal of all markers with a very low minor allele frequency (MAF). Here, we skip the step of identification of SNPs with different missing genotype rates between cases and controls due to the expermenatal purpose.

2.3.1 Identification of all markers with an excessive missing data rate

To calculate the missing genotype rate for each marker, type the following PLINK's command in terminal:

```
plink --bfile new --missing --out new
```

After getting the ouput file new.lmiss, we plot a histogram of the missing genotype rate to identify a threshold for extreme genotype failure rate.

```
# Histogram of the missing genotype rate
lmiss = read.table("~/QC/test/new.lmiss",header=T)
ylabels=c("0","20K","40K","60K","80K","500K")
xlabels=c("0.001","0.01","0.1","1")

#pdf("hist_missing_geno.pdf")
hist(log10(lmiss$F_MISS),axes=F,xlim=c(-3,0),
     col="RED",ylab="Number of SNPs",xlab="Fraction of missing data",
     main="All SNPs",ylim=c(0,500000))
```

```r
axis(side=2,labels=F)
mtext(ylabels,side=2,las=2, at=c(0,20000,40000,60000,80000,500000),line=1)
axis(side=1,labels=F)
mtext(xlabels,side=1,at=c(-3,-2,-1,0),line=1)
abline(v=log10(0.03),lty=2,col="blue")
#dev.off()

# Pie chart
#pdf("pie_missing_geno.pdf")

small_table = data.frame(c(sum(lmiss$F_MISS == 0),sum(lmiss$F_MISS < 0.03)-
            sum(lmiss$F_MISS == 0),sum(lmiss$F_MISS>0.03))
            ,row.names = c("missingness = 0","0 < missingness < 0.03",
                           "missingness > 0.03"))
colnames(small_table) = "counts"


lbls = paste(row.names(small_table),"\n",small_table$counts,sep = "")
pie(small_table$counts,labels = lbls,col = rainbow(length(lbls)))
#dev.off()
```

2.3.2 Removal of all markers failing QC

This step will exclude all the SNPs which are not pass the criteria that were described previously. The criteria include the SNPs are significant deviation from Hardy-Weinberg equilibrium, or have very low minor allele frequency.

```
plink --bfile new -maf 0.01 --geno 0.03 --hwe 0.00001 --make-bed --out new_clean
```

3 Results and Discussion

3.1 Per-individual QC

3.1.1 Distribution of individual's sex

As mentioned in previous, we calculated homozygosity rates across X-chromosome SNPs for each individual in the sample to find out the individuals that were not consistent in genotype information and ascertained sex. In our results, we identified 38 male individuals, 47 female individuals, and 11 individuals with not consistent data.

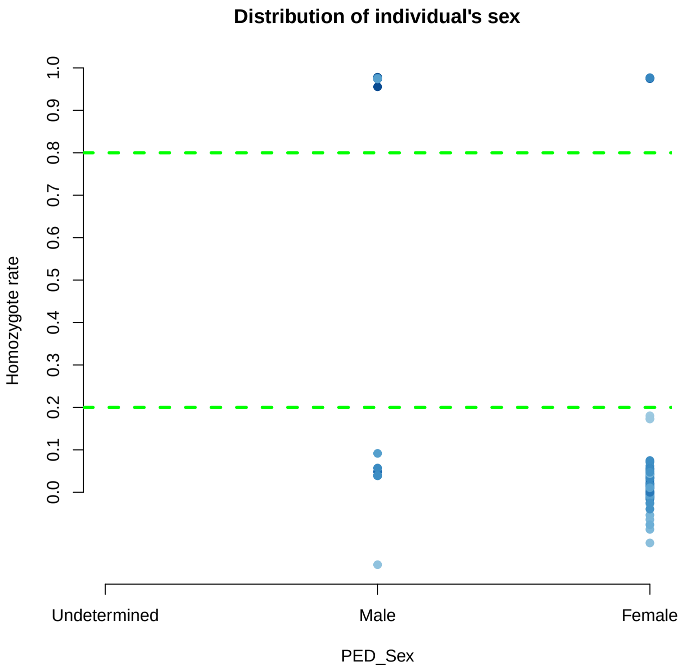

Figure 1: Distribution of individual's sex. When the homozygosity rate is more than 0.8, the sex of individuals will regard as male; in the other hand, indviduals will regard as female when the homozygosity rate is less than 0.2. When the homozygosity rate is more than 0.2 but less than 0.8, the genotype data are inconclusive regarding the sex of an individual.

Table 1: Sex information of 96 samples

Sex	By SNP	By record
Male	38	45
Female	47	51
PROBLEM	11	0

Table 2: Individuals with discordant sex information

Family ID	Individual ID
6	TDC497
8	TDC486
13	TDC495
18	TDC494
29	TDC475
31	TDC213
58	TDC465
65	TDC545
79	TDC506
80	TDC22
82	TDC451

3.1.2 Distribution of missingness and heterozygosity scores

Missing data rates and heterozygosity rate were measured to identify individuals with low DNA sample quality. In our results, there were 2 individuals out of our threholds. Specifically, sample 18 was reported by genome center that it might contaminate during genotyping.

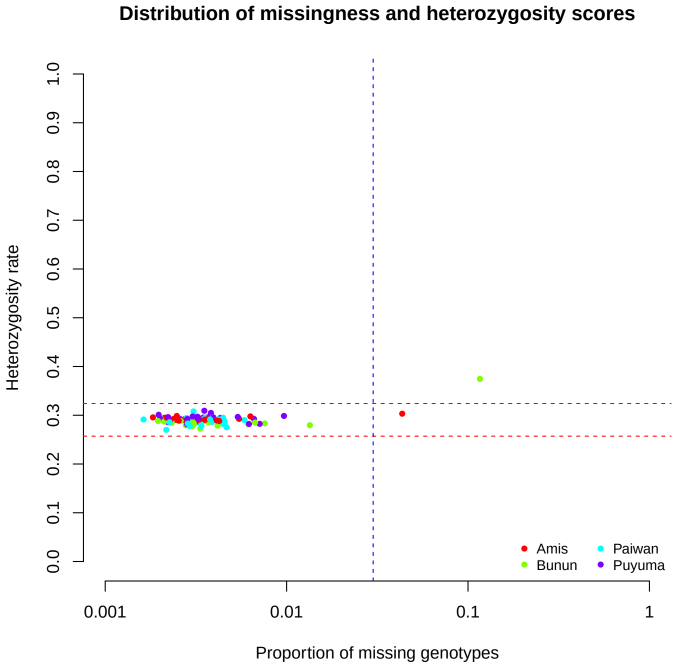

Figure 2: Distribution of missingness and heterozygosity scores. We chose to exclude all individuals with a genotype failure rate 0.03 and/or a heterozygosity rate 3 s.d. from the mean.

Table 3: Individuals with high missingness and/or outlier heterozygosity

Family ID	Individual ID
18	TDC494
73	TDC489

3.1.3 Propotion of the different IBD

To remove related individuals, we set IBD value of 0.18575 as criteria. The IBD value of 0.1875 was calculated from (0.125 + 0.25)/2 which 0.125 refered to third-degree relatives and 0.25 represented for second-degree relatives. In our results, there were 8 individuals which IBD were lager than 0.1875. Those individuals not only related to some individuals in samples but with lower call rates which stored in 96smples.imiss.

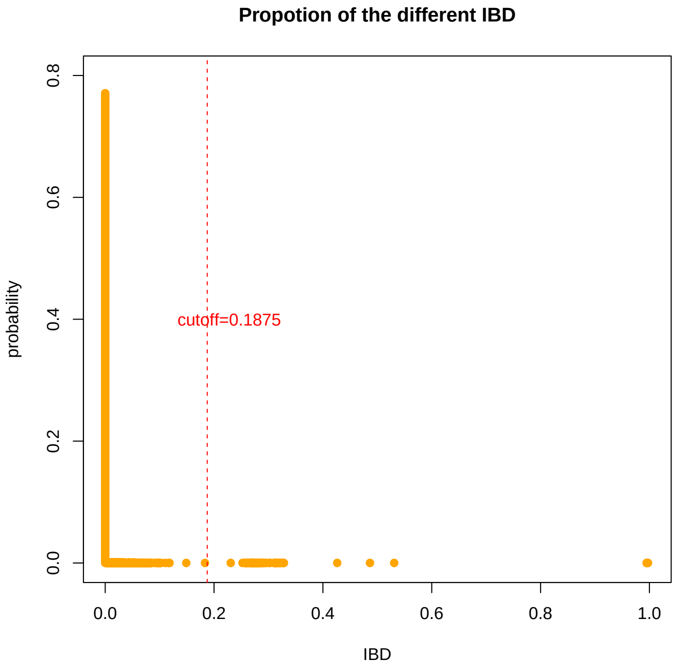

Figure 3: Propotion of the different IBD. Verticle lines show the threhold with IBD equals to 0.1875.

Table 4: Individuals with high IBD

Family ID	Individual ID
18	TDC494
22	TDC500
25	TDC165
28	TDC406
36	TDC478
56	TDC480-2
74	TDC212
89	TDC480

3.2 Per-marker QC

3.2.1 Distribution of excessive missing data rate

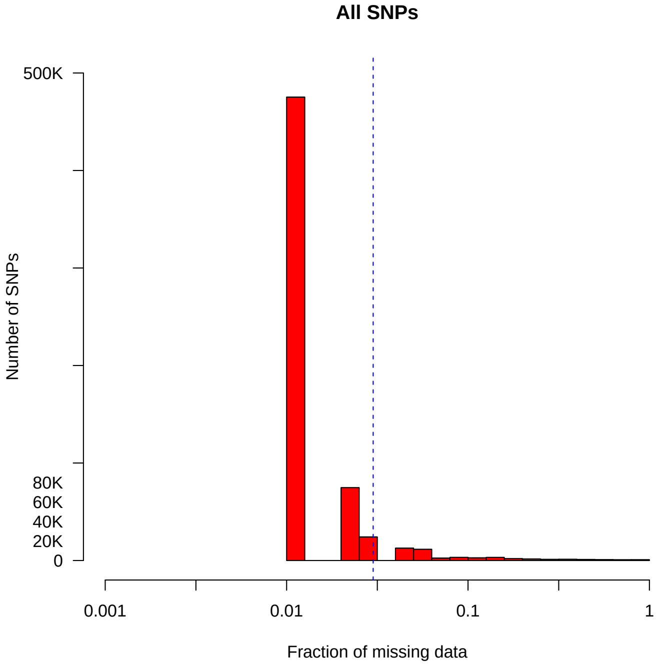

Figure 4: Histogram of missing data rate across all individuals passing per-individual quality control. The dashed vertical line represents the threshold (3%) at which SNPs were removed from further analysis because of an excess failure rate.

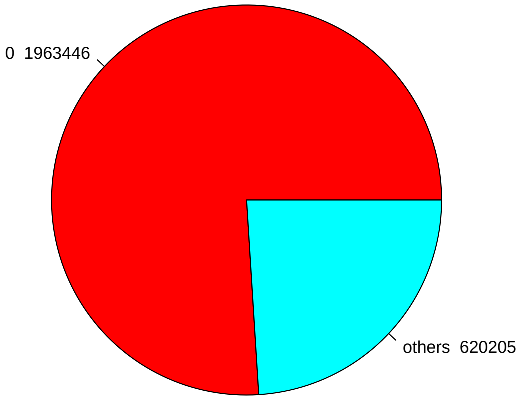

Figure 5: Pie chart shows the propotion of missing genotype rate = 0 and IBD>0. The proportions of IBD=0 is about 76%; IBD>0 is about 24%.

4 References

1. Anderson et al. Data quality control in genetic case-control association studies, Nature protocols, 2010.