QC Analysis

From Cox Lab Projects

Contents

Order of QC Analysis

First, conduct analyses that are not dependent on having marker allele frequencies correct: call rate filter/flag, gender check, Mendelian incompatibility checking.

Flag/filter SNPs with poor call rates, too many Mendelian incompatibilities and filter individuals with mis-specified gender low call rates

Note, you may have to circle back to this point AFTER homogeneous subgroups are established.

Within each homogeneous subgroup:

1. Calculate HWE statistics
2. Heterozygosity (across all SNPs) -- look at that distribution across individuals to check for outliers
3. Relationship checking (including looking for duplicates)
4. Plate effects

Flag/filter as needed given above results

Then EIGENSTRAT analyses. First with HapMap samples, then within homogeneous group to get final Eigenvectors. With genetically defined subgroups (if anyone changed), you may have to repeat steps 1-4 above.

Per SNP call-rate check

To calculate Call-Rate in a file-set some-file.map and some-file.ped one can run PLINK with following options:

```
plink --file some-file --missing --out some-file
```

This command will create two files with per subject and per SNP call-rate, some-file.imiss and some-file.lmiss correspondingly.

Gender check

Mendelian incompatibility check

Relationship check

Description:

Some genotyping errors (contamination, conversion errors, etc.) can be detected by looking for impossiblerelation between subjects.

Methods:

Relationship checking can be performed using PLINK with "--genome" and "--mind" options to calculate pairwise IBD for all subjects: plink --ped ped.ped --map ped.map --genome --mind (add additional dataset specific options)

Possible relationships:				
Relationship	PI-HAT	Z0	Z1	Z2
unrelated	0	1	0	0
identical-twins	1	0	0	1
parent-child	0.5	0	1	0
full siblings	0.5	0.25	0.5	0.25
half-siblings	0.25	0.5	0.5	0
grandparent-grandchild	0.25	0.5	0.5	0
avuncular	0.25	0.5	0.5	0
half-avuncular	0.125	0.75	0.25	0
first-cousin	0.125	0.75	0.25	0
half-first-cousin	0.0625	0.875	0.125	0
half-sibling-plus-first-cousin	0.375	0.375	0.5	0.125

Population stratification

Description:

Some subjects could be misclassified by declared race.

Input data

SMARTPCA needs three files:

1. Genotype file
   should be a regular PLINK ped file
2. SNP file
   should be a regular PLINK map file
3. Pedigree file
   should be a regular PLINK tfam file
   should have extension '.pedind'

Unrelated ids from Hapmap

Media:CEU.tnr.spca.fam

Media:CHB.tnr.spca.fam

Media:GIH.tnr.spca.fam

Media:JPT.tnr.spca.fam

Media:MEX.tnr.spca.fam

Media:YRI.tnr.spca.fam

Data preparation

All that is required to run SMARTPCA and check ancestry are regular PLINK .ped and .map files.

Converting tped/tfam files to ped/map format:

```
plink --recode --tab --tfile filename --out filename
```

Exclude SNPs in on chromosomes X and Y, on mitochondrial DNA, and SNP with unknown chromosome:

Converting bed/bim/fam files to ped/map format:

```
plink --recode --tab --bfile filename --out filename --extract plink.regiones.chr1t22.txt --range
```

Regions file contains chromosomes 1 - 22:

plink.regiones.chr1t22.txt --Anuar 13:09, 23 December 2009 (CST)

Prepare files for SMARTPCA:

```
cp newfilename.map newfilename.pedsnp
cut -f 1-6 newfilename.ped > newfilename.pedind
```

Running Eigenstrat

Create a parameters file newfilename.par:

```
genotypename: newfilename.ped
snpname: newfilename.map
indivname: newfilename.pedind
evecoutname: newfilename.evec
evaloutname: newfilename.eval
outliername: newfilename.outlier
numoutevec: 10
numoutlieriter: 0
numoutlierevec: 2
outliersigmathresh: 6
```

Parameter explanation:

genotypename - PLINK ped-file with genotypes

snpname - PLINK map-file

indivname - PLINK fam-file

evecoutname - principal components output name

evaloutname - output eigenvector filename

outliername - outlier filename

numoutevec: - number of principal components to write as an output

numoutlieriter - maximum number of outlier removal iterations. To turn off outlier removal, set 0.

numoutlierevec - number of principal components along which to remove outliers during each outlier removal iteration.

outliersigmathresh - (Default is 6.0) number of standard deviations which an individual must exceed, along one of topk top principal components, in order to be removed as an outlier.

Run SMARTPCA:

```
smartpca -p newfilename.par
```

Could fail if IDs are too long

Output files:

newfilename.evec	contains 10 first principal componnints
newfilename.eval	do not know right now
newfilename.evec	outliers, samples excluded from analysis

Visualization example

R command to print thyroid PCA analysis:

```
plot(a[,5],a[,6],col=as.matrix(a[,3]),pch=16,main="Thyroid samples",xlab="PC 1",ylab="PC 2")
text(a[1:19,5]-0.01,a[1:19,6],labels=labs)     # put labels on samples
points(a[1:19,5],a[1:19,6],col="black",pch=16)
text(popsx,popsy,labels=popsn)                 # put labels on population clusters
for(i in 1:length(evecs[,2])){if(evecs[i,2]>0.0){text(evecs[i,2]-0.015,evecs[i,3],label=unlist(strsplit(as.ch
```

Plate effect analysis

Description:

Some SNPs are difficult to genotype, these SNPs can be highly associated with some of the plates.

Methods:

Plate effect analysis can be done using PLINK with "--assoc" option using plate id for phenotype.

plink --ped.ped --map ped.map --pheno plate_phenos.txt --pheno-name plate_id --assoc [--remove already_detected_outliers.txt [add additional dataset specific options]]

plate_phenos.txt should have one for each plate with 2 for all subjects on the plate and ones for all those that are not on the plate, first three columns are family_id, subject_id and plate_id:

FAMID	SUBJID	Plate_name	plate1	plate2
fam1	subj1	plate1	2	1
fam2	subj2	plate1	2	1
fam3	subj3	plate2	1	2
fam4	subj4	plate2	1	2

Retrieved from "http://genemed.uchicago.edu/projects/prjwiki/index.php?title=QC_Analysis&oldid=3193"

- This page was last modified on 9 March 2011, at 10:01.