hla\_analysis

David Martino

2 July 2015

**stage 1: preparation of data**

library(reshape)  
library(plyr)

##   
## Attaching package: 'plyr'  
##   
## The following objects are masked from 'package:reshape':  
##   
## rename, round\_any

setwd("/Users/david/Documents/projects/healthnuts/hla/imputationData/imp2.v2/HN\_hla")  
  
source('hlaAnalysis.R')  
rawdata=read.delim('HN2.hla.csv',header=T,sep=",")  
imputes=read.delim('HN2.hla.txt',header=T,sep="")  
phenotypes=read.delim('PEANUT\_ALLERGY.4PLINK.phen',sep="",header=T)  
phenotypes$Phenotype=ifelse(phenotypes$PEANUT\_ALLERGY==2,'cases',   
 ifelse(phenotypes$PEANUT\_ALLERGY==1,'controls', 'missing'))  
  
head(rawdata)

## IndividualID Chromosome Gene Allele Posterior  
## 1 2678 1 A 201 1.0000000  
## 2 2678 2 A 201 1.0000000  
## 3 3583 1 A 101 1.0000000  
## 4 3583 2 A 201 1.0000000  
## 5 127 1 A 6801 1.0000000  
## 6 127 2 A 3303 0.9933333

tail(rawdata)

## IndividualID Chromosome Gene Allele Posterior  
## 7931 2876 1 DRB1 101 1.0000000  
## 7932 2876 2 DRB1 101 1.0000000  
## 7933 781 1 DRB1 1101 1.0000000  
## 7934 781 2 DRB1 1102 0.9666667  
## 7935 2653 1 DRB1 1001 1.0000000  
## 7936 2653 2 DRB1 405 0.9800000

table(rawdata$Gene)

##   
## A B C DPA1 DPB1 DQA1 DQB1 DRB1   
## 992 992 992 992 992 992 992 992

table(rawdata$Chromosome)

##   
## 1 2   
## 3968 3968

summary(rawdata$Posterior)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.1067 0.9800 1.0000 0.9512 1.0000 1.0000

table(phenotypes$Phenotype,phenotypes$ANCESTRY)

##   
## Asian European MixedEuropeanAsian Other  
## cases 14 51 8 0  
## controls 10 119 19 0  
## missing 34 188 39 14

dat=merge(rawdata,phenotypes,by.x="IndividualID",by.y='IID')  
dat=dat[!dat$Phenotype=='missing',]

**stage2: Summaries and data structures to represent HLA allele data.**

Counts the number of cases and controls for each allele. Creates a matrix representation which can be used for regression modelling.

The input should be a data frame obtained by reading IMP1 or IMP2 imputation calls, with an extra column attached called "Phenotype" which takes two possible values: "cases" and "controls".

hlaCounts(dat,callThreshold=0.7)

## hlaCounts: 8 HLA genes, 93 alleles, 221 individuals  
## First few alleles:  
## Gene Allele numHap.cases numHap.controls numHap.total  
## 1 A A.101 23 54 77  
## 2 A A.1101 14 21 35  
## 3 A A.1102 2 0 2  
## 4 A A.201 31 71 102  
## 5 A A.203 0 1 1  
## 6 A A.205 1 1 2  
## alleleFreq.controls alleleFreq.cases alleleFreq.total  
## 1 0.190140845 0.163120567 0.181176471  
## 2 0.073943662 0.099290780 0.082352941  
## 3 0.000000000 0.014184397 0.004705882  
## 4 0.250000000 0.219858156 0.240000000  
## 5 0.003521127 0.000000000 0.002352941  
## 6 0.003521127 0.007092199 0.004705882