6.881 Lab 1. Linkage disequilibrium

Yaping Liu, Manolis Kellis
Massachusetts Institute of Technology

February 8, 2016

1 Introduction

Consider two SNPs on the same chromosome, and that you have access to phased genotypes for both your maternal and paternal copy of that chromosome. According to Mendel's law of independent assortment, the probability of your carrying both SNPs on your maternal chromosome is equal to the product of their allele frequencies in the population. However, if the two SNPs are near each other, they may tend to be inhereted as a haplotype rather than independently, due to genetic linkage. We will explore this phenomenon by looking at SNPs around some interesting genes using phased genomes from the 1000 Genomes Project.

2 Downloading and exploring the data

First look at the *EDAR* gene on chromosome 2. Complete 1000 Genomes Project data are available for download, but are very large; in order to just get data at his locus, use the browser at `http://browser.1000genomes.org`. Enter `EDAR` in the search form, then select "Region in detail" next to the second search result (Gene ENSG0000013 5960). Two region views will appear, one zoomed out and one zoomed in. How many genes are in the 1-Mb region? How many in the 94.9-Kb EDAR locus? On which strand is EDAR encoded? Is the gene mostly exonic or intronic?

To download variation, click "Linkage data" in the menu on the left. Then click "Get VCF data", keep the defaults, and proceed to the download page.

Notice that the columns in this file correspond to individuals, and that each individual has a genotype of either 00, 01, 10, or 11. Population data for these individuals is available in the following file, which you should download:

ftp://ftp.1000genomes.ebi.ac.uk/vol1/
ftp/release/20130502/integrated_call_samples_v3.20130502.ALL.panel

We can perform some rather inelegant text parsing within R to arrive at matrices describing the genotypes within three of the populations:

```
vcffn = "2.109510927-109605828.ALL.chr2.phase3_shapeit2
            _mvncall_integrated_v5a.20130502.genotypes.vcf.gz"
fullvcf = read.delim(vcffn,sep="\t",header=T,skip=252)
panelinfo = read.delim("integrated_call_samples_v3.20130502.ALL.panel", header=T)


populationSlice = function(vcf, population, panelinfo){
  vcft<-vcf[grep("VT=SNP",vcf[,8]),]
  tmp<-vcft[, colnames(vcft) %in% panelinfo[grep(population, panelinfo[,2]),1]]
  rownames(tmp)<-vcft[,"ID"]
  tmp
}

vcfToGenotypeMatrix = function(vcf){
  result = matrix(nrow=nrow(vcf), ncol=ncol(vcf)*2)
  for (i in 1:nrow(vcf)){
    for (j in 1:ncol(vcf)){
      result[i,(j*2)-1] = as.integer(substr(vcf[i,j], 1, 1))
      result[i,(j*2)] = as.integer(substr(vcf[i,j], 3, 3))
    }
  }
  rownames(result)<-rownames(vcf)
  result = result[apply(result, 1, function(x) sum(x)) > 0,]
  result = result[apply(result, 1, function(x) sum(x)) < ncol(result),]
  return(result)
}

british = vcfToGenotypeMatrix(populationSlice(fullvcf, "GBR", panelinfo))
han = vcfToGenotypeMatrix(populationSlice(fullvcf, "CHB", panelinfo))
```

```
yoruba = vcfToGenotypeMatrix(populationSlice(fullvcf, "YRI", panelinfo))
```

Note that we are disassociating the two copies of the gene in each individual, and within each population we are discarding any sites that aren't polymorphic. How many individuals are in each population sample? How many chromosomes? How many variants?

We can write a function to visualize the mutations in each individual along the gene, clustering the individuals by similarity:

```
clusteredImage = function(gt){
  clustered.individuals = hclust(dist(t(gt), method="euclidean"),
  method="centroid")$order
  image(gt[,clustered.individuals], ylab="SNP",xlab="Samples")
}

 dev.new()

clusteredImage(british)
clusteredImage(han)
clusteredImage(yoruba)
```

Where does it look like recombination tends to occur? Where are the LD blocks?

One way of quantifying linkage disequilibrium between pairs of loci is with the deviation from independence as stated as in the introduction. Consider the first two SNPs in the British sample:

```
snpsample = british[1:2,]
snpsample
```

The allele frequencies of these two SNPs is:

```
> rowMeans(snpsample)
 rs1478517 rs10865025
 0.6208791  0.8681319
```

If these SNPs are independent, we would expect about $62\% \times 87\% = 54\%$ of individuals to carry both. We can see how many in fact carry both:

```
> mean(apply(snpsample, 2, function(x) sum(x)==2))
[1] 0.6208791
```

There is an excess of $D = 62\% - 54\% = 8\%$. By visualizing the pairs:

```
image(snpsample)
```

we see that all individuals with rs1478517 also have rs10865025. The opposite is not true. This configuration is compatible with recombination never having taken place between these two loci (the so-called four-gamete test.)

In practice, D is often scaled by the theoretical maximum value D_{max}, and reported as D'.

$$D' = \frac{D}{D_{max}} \tag{1}$$

Let us call rs1478517 site p with frequencies $p_1 = 0.62$ and $p_2 = 0.38$, and let us call rs10865025 site q with frequencies $q_1 = 0.87$ and $q_2 = 0.13$. The frequency of carriers of both, x_{11}, is expected to be $p_1 q_1 \approx 0.54$. The linkage disequilibrium $D = x_{11} - p_1 q_1 \approx 0.08$. The maximum value D_{max} is defined as:

$$D_{max} = \begin{cases} min\left(p_1 q_1, p_2 q_2\right) & \text{when } D < 0 \\ min\left(p_1 q_2, p_2 q_1\right) & \text{when } D > 0 \end{cases} \tag{2}$$

In this case $D_{max} \approx 0.08$ and $D' = \frac{D}{D_{max}} = 0.91$, reflecting the fact that the two variants are as genetically linked very often given their allele frequencies!

Another way is with the Pearson correlation coefficient r^2:

```
> cor(snpsample[1,], snpsample[2,])**2
[1] 0.2487617
```

We can calculate this between all pairs and visualize this quickly:

```
correlationImage = function(gt){
  result = matrix(nrow=nrow(gt), ncol=nrow(gt))
  for (a in 1:(nrow(gt)-1)){
    for (b in (a+1):nrow(gt)){
      result[a,b] = cor(gt[a,], gt[b,])**2
    }
  }
```

```
  image(result, col=colorRampPalette(c("black", "red"))(32))
}

correlationImage(british)
correlationImage(han)
correlationImage(yoruba)
```

3 Assignment

Visualize r^2 at a gene with a known disease-associated SNP. Explain the evidence for association and any mechanisms that have been proposed or demonstrated. Report the variant's frequency in the three populations. Optional: visualize D' in these three populations at *EDAR*.

6.881 Computational Personal Genomics: Making Sense of Complete Genomes
Spring 2016