Genetic association studies

Cavan Reilly

September 24, 2015

Table of contents

Population based association studies

There are several types of studies that are distinguished:

1. Candidate polymorphism studies-test if a previously identified polymorphism influences some trait
2. Candidate gene studies-test if a gene has markers that are within it or near it that are associated with a trait (these markers needn't be functional)
3. Fine mapping studies-determine the location of a variant that influences some trait
4. Genome-wide association studies (GWAS)-a candidate gene study that examines the entire genome, or at least a large portion of it.

Population based association studies

While candidate gene studies and GWAS are similar in spirit, candidate gene studies are more focused and typically try to sort out other risk factors

GWAS are more computationally demanding due to the large number of variants examined.

GWAS also suffer from the problem of multiple testing (more on this latter).

Genotype

In association studies we examine if DNA sequences can predict a trait of interest.

So, for us, the DNA sequence is the *explanatory variable* and the trait is the *response variable*.

The *genotype* of someone is the specific DNA sequence that someone has at some location on the genome.

As humans generally have 2 copies of each chromosome (one from each parent), when we discuss a genotype we refer to the information on both copies.

Hence when we consider a SNP, which are mostly biallelic, there are 3 distinct genotypes (2 homozygous and 1 heterozygous).

Family versus population based studies

We focus on population based studies: family based studies have their own considerations.

Family based studies are difficult to design as one needs to recruit entire families into the study.

Also, very difficult to design family based studies for late onset diseases.

However family based studies do allow one to study rare variants.

They also allow one to estimate the *phase* with greater certainty.

Haplotype

When we consider 2 loci simultaneously, if know which pairs of alleles are on the same chromosome then we know the *haplotype* for these loci.

Knowing the haplotype for a set of loci is equivalent to knowing the phase for those loci.

For example, consider 2 genes (denoted by a letter) where each gene has 2 alleles (which we distinguish by case).

If someone has genotype AA and Bb, then we know that 1 chromosome must have alleles AB and the other chromosome must have alleles Ab-in this case we know the haplotype.

Haplotype

In contrast, if someone has genotype Aa and Bb then 2 configurations are possible: one chromosome has alleles AB while the other has ab or one chromosome has alleles Ab while the other has alleles aB-here haplotype is unknown.

Conventional assays for genotyping a subject (i.e. determining the genotype for many loci simultaneously) do not give information about haplotypes, they only provide information about genotypes for each locus separately.

Data structure

There are 3 classes of information that we will deal with

1. genotypic data
2. trait data
3. covariate data

Genotypes

Currently SNPs are the most common form of genotypic data, but there are other forms for genotypic data such as *indels* and microsattelite markers.

In all cases, given a sample from a population, there will be a collection of locations that differ among subjects (a set of loci that exhibit variation) and a finite number of possible values for these variants (the alleles found in this sample).

Sometimes we use the term *multilocus genotype* when we refer to the genotypes at multiple locations. These are distinct from haplotypes.

For SNPs, as there are only 2 alleles, the *minor allele* is the allele that is less frequent in the population and the *minor allele frequency* is the probability that one observes the minor allele.

Genotype frequency and allele frequency

Consider a biallelic locus with alleles a and A.

Suppose that there is a population with the following proportions of genotypes

$$P(AA) = 0.75 \quad P(Aa) = 0.20 \quad P(aa) = 0.05$$

To compute the allele frequency: pretend that the population has 100 subjects so that the proportions count numbers of people with a genotype.

Then our population of alleles is a population of 200, and of this 200, 75+75 are A alleles that are involved in a homozygous genotype and 20 come from the heterozygous subjects, thus the major allele frequency is

$$(75 + 75 + 20)/200 = 0.85.$$

Traits and covariates

Trait, phenotype and outcome are all used interchangeably in this course.

Continuous traits, e.g. HIV viral load, frequently need to be transformed prior to conducting any analysis.

The main problem with continuous traits is that sometimes they are more variable for higher levels of the variable: for example, if you dichotomize the data and compare the variance you will often notice this.

When this occurs with the residuals in a regression setting, this phenomenon is referred to as *heteroscedasticity*.

Traits and covariates

This will make methods designed for normally distributed data perform poorly.

Frequently just taking the logarithm (to any base) will get rid of this problem.

Using log transformations in conjunction with methods for normally distributed data usually outperforms strictly nonparametric approaches.

Traits and covariates

A covariate is a variable that may be relevant for understanding the relationship between the trait of interest and the genetic data.

We will discuss the role of covariates later, for now let's note that sometimes it makes sense to transform them too.

For example, in the context of multiple linear regression, if I use a log transformation on all of the variables then the regression coefficients give the percent change in the response variable for a 1 percent change in the explanatory variable.

Data examples

We will work with 3 data examples throughout this component of the course:

One has SNP data from a complex trait study of humans.

One has HIV sequence information.

The final one has SNP data from populations all over the world.

HIV carries its genetic information in a single RNA molecule whereas humans have 2 copies of the (haploid) human genome.

Data examples

HIV is a virus that has a much higher error rate (compared to mammals) when it replicates its genome.

The various strains (or *quasi-species*) that one finds depends on a variety of factors, e.g. administration of antiretroviral drugs (ART or HAART).

The primary mechanisms whereby genetic diversity is introduced into humans are recombination and independent assortment of the chromosomes (although mutation plays a role).

Human genetics

Independent assortment ensures that pairs of alleles at loci on distinct chromosomes are equally likely to end up in the same sex cell.

Recombination goes further and mixes up pairs of alleles on the same chromosome, resulting in chromosomes that are not found in either parent.

If 2 loci are very close on a chromosome it is unlikely that a recombination will separate the loci and switch pairs of alleles.

Human genetics

Hence pairs of alleles that are at loci that are very close tend to be transmitted together.

Such loci are said to be in *linkage disequilibrium*.

While we intuitively think of there being a disease causing mutation in a coding region of a gene, it is not uncommon for a variant that is in a noncoding region to be in complete linkage disequilibrium with that mutation.

Human genetics

In practice this implies that the difference between causal variants and tagging variants (i.e. those that mark the causal variants due to strong linkage disequilibrium) is actually not so clear.

In fact, it's possible that the variant in the noncoding region is the actual causal variant (e.g. it interferes with transcription factor binding) and the variant in the coding region is not harmful.

HIV genetics

HIV creates further copies of itself by binding to certain cells and inserting its genetic material into the cell.

Upon gaining entry to a cell, it uses reverse transcriptase to make a DNA copy of the material it stores in RNA.

The enzyme integrase then inserts this DNA into the DNA of the cell, and the cell then makes the proteins HIV needs to assemble new infectious particles.

One of these proteins is protease and this protein is essential for creating some of the other proteins that are necessary for HIV to replicate itself.

HIV genetics

Modern ART use multiple drugs so that different drugs target different components of the HIV life cycle.

These treatment strategies have been very effective at prolonging the life of HIV positive patients, nonetheless these patients are at greater risk for many negative health outcomes (e.g. cardiovascular events) even though no virus can be measured in their blood.

Due to the high error rate of replication there is a substantial amount of genetic diversity within each patient. An amino acid needs to be present in at least 20% of the virions to be detectable as a variant using contemporary methods.

Data example

We will use data from the Functional SNPs associated with muscle size and strength (FAMuSSS) study.

This study obtained data on 225 SNPs from 1397 college students.

The study had the subjects participate in a 12 week exercise program and measured a number of variables related to muscle and metabolic syndrome.

We will read this data into R by writing in the URL and then using the read.delim function:

```
> fmsURL <- "http://people.umass.edu/foulkes/asg/data/FMS_data.txt"
> fms <- read.delim(file=fmsURL)
```

Data example

We can see all of the variables which were measured by looking at

```
> names(fms)
```

There are over 340 of them, so if we just want to look at the first 20, we can do

```
> names(fms)[1:20]
 [1] "id"                    "acdc_rs1501299"        "ace_id"
 [4] "actn3_r577x"           "actn3_rs540874"        "actn3_rs1815739"
 [7] "actn3_1671064"         "ardb1_1801253"         "adrb2_1042713"
[10] "adrb2_1042714"         "adrb2_rs1042718"       "adrb3_4994"
[13] "agrp_5030980"          "akt1_t22932c"          "akt1_g15129a"
[16] "akt1_g14803t"          "akt1_c10744t_c12886t"  "akt1_t10726c_t12868c"
[19] "akt1_t10598a_t12740a"  "akt1_c9756a_c11898t"
```

Data example

If we want a table of the gender of the subjects we can issue the command

```
> table(fms$Gender)
Female   Male
   607    426
```

Or we can first attach our data frame, then just refer to the gender variable

```
> attach(fms)
> table(Gender)
Gender
Female   Male
   607    426
```

Computing the minor allele frequency

As an example we will compute the minor allele frequency for the SNP rs540874.

Note that if you search on the name of this SNP you will find a link to dbSNP which has lots of useful information, for example it is on chromosome 11 in a gene called ACTN3, alpha-actinin-3.

Let's take a look at the frequency of the genotypes in our study

```
> GenoCount <- summary(actn3_rs540874)
> GenoCount
  AA   GA   GG NA's
 226  595  395  181
```

Computing the minor allele frequency

Then let's compute the genotype frequencies treating the NA's as missing at random.

```
NumbObs <- sum(!is.na(actn3_rs540874))
```

```
> GenoFreq <- as.vector(GenoCount/NumbObs)
> GenoFreq
[1] 0.1858553 0.4893092 0.3248355 0.1488487
```

then we can compute them directly as follows

```
> (2*GenoFreq[1]+GenoFreq[2])/2
[1] 0.4305099
> (2*GenoFreq[3]+GenoFreq[2])/2
[1] 0.5694901
```

Computing the minor allele frequency

We can also do this using prebuilt packages, in particular the genetics package, as follows.

```
> library(genetics)
> Geno <- genotype(actn3_rs540874,sep="")
> summary(Geno)
Number of samples typed:  1216 (87%)
Allele Frequency:  (2 alleles)
   Count Proportion
G   1385       0.57
A   1047       0.43
NA   362         NA
Genotype Frequency:
    Count Proportion
G/G   395       0.32
G/A   595       0.49
A/A   226       0.19
NA    181         NA
Heterozygosity (Hu)  = 0.4905439
Poly. Inf. Content   = 0.3701245
```

Computing the minor allele frequency

Note that we get the same allele frequencies along with some other summaries.

Both the *Heterozygosity* and the *Polymorphism information content* tells us how informative this marker is for localizing a disease gene.

The heterozygosity tells us the probability that a random individual will be heterozygous at this locus.

Computing the minor allele frequency

The polymorphism information content is the probability that marker genotype of an offspring will allow deduction of which of the 2 marker alleles it received from one of the parents.

As such it indicates the probability that one will be able to infer phase when one has data from related subjects.

This is a measure of marker usefulness that is more relevant to linkage analysis than association studies.

Human genome diversity project

Here is the updated website for the HGDP:

`http://hsblogs.stanford.edu/morrison/human-genome-diversity-project/`

This project was undertaken to collect cell lines for studying human diversity.

We will look at 4 SNPs in the v-akt murine thymoma oncogene homolog 1 gene.

We can read in the data using our previous strategy.

```
> hgdpURL <- "http://people.umass.edu/foulkes/asg/data/HGDP_AKT1.txt"
> hgdp <- read.delim(file=hgdpURL)
```

Human genome diversity project

Then we inspect the top of the data set

```
> head(hgdp)
  Well         ID Gender     Population        Geographic.origin Geographic.area
1  B12 HGDP00980      F Biaka Pygmies  Central African Republic  Central Africa
2  A12 HGDP01406      M          Bantu                   Kenya  Central Africa
3   E5 HGDP01266      M       Mozabite           Algeria (Mzab) Northern Africa
4   B9 HGDP01006      F      Karitiana                  Brazil   South America
5   E1 HGDP01220      M           Daur                   China           China
6   H2 HGDP01288      M            Han                   China           China
  AKT1.C0756A AKT1.C6024T AKT1.G2347T AKT1.G2375A
1          CA          CT          TT          AA
2          CA          CT          TT          AA
3          AA          TT          TT          AA
4          AA          TT          TT          AA
5          AA          TT          TT          AA
6          AA          TT          TT          AA
```

Human genome diversity project

and we can see there is data for 1064 subjects from 52 populations.

```
> dim(hgdp)
[1] 1064   10
> length(table(hgdp[,4]))
[1] 52
```

The virco data set

The virco data set has information on the protease sequence from 1066 HIV viral isolates.

For each virus we also have fold resistance measures for 8 protease inhibitors.

By fold resistance we mean the resistance of each viral strain to the protease inhibitors relative to the wild type strain.

The virco data set

We can read in the data using a strategy similar to what we have done thus far, but now use the `read.csv` function since commas separate the entries.

```
> vircoURL <- "http://people.umass.edu/foulkes/asg/data/Virco_data.csv"
> virco <- read.csv(file=vircoURL)
```

and we can see that the data for each isolate is represented by 124 variables.

```
> dim(virco)
[1] 1066  124
```

The virco data set

Here we look at a selection of these variables to see that the data is coded in 2 ways that are consistent

```
> virco[1:5,c(1,6,11,32,85,93,104,112,122)]
  SeqID IsolateName IDV.Fold P10 P63 P71 P82 P90
1  3852      CA3176     14.2   I   P   -   -   M
2  3865      CA3191     13.5   I   P   V   T   M
3  7430      CA9998     16.7   I   P   V   A   M
4  7459 Hertogs-Pt1      3.0   I   P   T   -   M
5  7460 Hertogs-Pt2      7.0   -   -   -   A   -

                                                                                CompMutList
1                                                                L10I, M46I, L63P, G73CS, V77I, L90M, I93L
2                                  L10I, R41K, K45R, M46I, L63P, A71V, G73S, V77I, V82T, I85V, L90M, I93L
3 L10I, I15V, K20M, E35D, M36I, I54V, R57K, I62V, L63P, A71V, G73S, V82A, L90M
4                                                L10I, L19Q, E35D, G48V, L63P, H69Y, A71T, L90M, I93L
5                                                                K14R, I15V, V32I, M36I, M46I, V82A
```

This is the usual nomenclature for discussing amino acid mutations.